Proceedings
22nd IEEE International Parallel and Distributed Processing Symposium

IPDPS 2008 Advance Program Abstracts

Abstracts for both contributed papers and all workshops have been compiled to allow authors to check accuracy and so that visitors to this Website may preview the papers to be presented at the conference. Full proceedings of the conference will be published on a cdrom to be distributed to registrants at the conference.

Summary of Contents

IEEE International Parallel & Distributed Processing Symposium

IPDPS 2008

SESSION 1

Algorithms - Scheduling

Efficient Resource Management Using Advance Reservations for Heterogeneous Grids

Claris Castillo, George N. Rouskas and Khaled Harfoush
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
Email: {ccastil,rouskas,kaharfou}@ncsu.edu

Support for advance reservations of resources plays a key role in Grid resource management as it enables the system to meet user expectations with respect to time requirements and temporal dependence of applications, increases predictability of the system and enables co-allocation of resources. Despite these attractive features, adoption of advance reservations is limited mainly due to the fact that related algorithms are typically complex and fail to scale to large and loaded systems. In this work we consider two aspects of advance reservations. First, we investigate the impact of heterogeneity on Grid resource management when advance reservations are supported. Second, we employ techniques from computational geometry to develop an efficient heterogeneity-aware scheduling algorithm. Our main finding is that Grids may benefit from high levels of resource heterogeneity, independently of the total system capacity. Our results show that our algorithm performs well across several user and system performance and overcome the lack of scalability and adaptability of existing mechanisms.

Online Scheduling in Grids

Uwe Schwiegelshohn
Technische Universität Dortmund
Robotics Research Institute
44221 Dortmund, Germany
uwe.schwiegelshohn@udo.edu

Andrei Tchernykh
CICESE Research Center
22860 Ensenada, Baja California, Mexico
chernykh@cicese.mx

Ramin Yahyapour
Technische Universität Dortmund
IT and Media Center
44221 Dortmund, Germany
ramin.yahyapour@udo.edu

This paper addresses nonclairvoyant and nonpreemptive online job scheduling in Grids. In the applied basic model, the Grid system consists of a large number of identical processors that are divided into several machines. Jobs are independent, they have a fixed degree of parallelism, and they are submitted over time. Further, a job can only be executed on the processors belonging to the same machine. It is our goal to minimize the total makespan. We show that the performance of Garey and Graham's list scheduling algorithm is significantly worse in Grids than in multiprocessors. Then we present a Grid scheduling algorithm that guarantees a competitive factor of 5. This algorithm can be implemented using a "job stealing" approach and may be well suited to serve as a starting point for Grid scheduling algorithms in real systems.

Scheduling with Storage Constraints

Erik Saule, Pierre-François Dutot and Grégory Mounié
LIG - MOAIS Team
Grenoble Universités, France
{firstname.lastname}@imag.fr

Cumulative memory occupation is a quite intuitive but not so studied constraint in scheduling. The interest in such a constraint is present in multi-System-on-Chip, embedded systems for storing instruction code, or in scientific computation for storing results. Memory occupation seen as a constraint is impossible to solve with approximation algorithms. We believe that transforming the constraint into a second objective to optimize helps to deal with such constraints.

The problem addressed in this paper is to schedule tasks on identical processors in order to minimize both maximum completion time and maximum cumulative memory occupation. For independent tasks, a family of algorithms with good approximation ratios based on a PTAS is given. Several approximation ratios are proved to be impossible to achieve with any schedule. The precedence constrained case is then studied and a family of performance guaranteed algorithms based on List Scheduling is proposed. Finally, optimizing the mean completion time as a third objective is also studied and a tri-objective algorithm is given.

Portioned Static-Priority Scheduling on Multiprocessors

Shinpei Kato and Nobuyuki Yamasaki
School of Science for Open and Environmental Systems
Keio University, Japan
{shinpei,yamasaki}@ny.ics.keio.ac.jp

This paper proposes an efficient real-time scheduling algorithm for multiprocessor platforms. The algorithm is a derivative of the Rate Monotonic (RM) algorithm, with its basis on the portioned scheduling technique. The theoretical design of the algorithm is well implementable for practical use. The schedulability of the algorithm is also analyzed to guarantee the worst-case performance. The simulation results show that the algorithm achieves higher system utilizations, in which all tasks meet deadlines, with a small number of preemptions compared to traditional algorithms.

SESSION 2

Applications - General Applications

A Parallel Software Toolkit for Statistical 3-D Virus Reconstructions from Cryo Electron Microscopy Images Using Computer Clusters with Multi-core Shared-Memory Nodes

Yili Zheng and Peter C. Doerschuk
School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285 USA
Department of Biomedical Engineering and School of
Electrical and Computer Engineering, Cornell University
Ithaca, NY 14853-5401 USA

A statistical approach for computing 3-D reconstructions of virus particles from cryo electron microscope images and minimal prior information has been developed which can solve a range of specific problems. The statistical approach causes high computation and storage complexity in the software implementation. A parallel software toolkit is described which allows the construction of software targeted at commodity PC clusters which is modular, reusable, and user-transparent while at the same time delivering nearly linear speedup on practical problems.

Modeling and Predicting Application Performance on Parallel Computers Using HPC Challenge Benchmarks

Wayne Pfeiffer and Nicholas J. Wright
San Diego Supercomputer Center, La Jolla CA 92093-0505, USA
{pfeiffer, nwright}@sdsc.edu

A method is presented for modeling application performance on parallel computers in terms of the performance of microkernels from the HPC Challenge benchmarks. Specifically, the application run time is expressed as a linear combination of inverse speeds and latencies from microkernels or system characteristics. The model parameters are obtained by an automated series of least squares fits using backward elimination to ensure statistical significance. If necessary, outliers are deleted to ensure that the final fit is robust. Typically three or four terms appear in each model: at most one each for floating-point speed, memory bandwidth, interconnect bandwidth, and interconnect latency. Such models allow prediction of application performance on future computers from easier-to-make predictions of microkernel performance.

The method was used to build models for four benchmark problems involving the PARATEC and MILC scientific applications. These models not only describe performance well on the ten computers used to build the models, but also do a good job of predicting performance on three additional computers with newer design features. For the four application benchmark problems with six predictions each, the relative root mean squared error in the predicted run times varies between 13 and 16%.

The method was also used to build models for the HPL and G-FFTE benchmarks in HPCC, including functional dependences on problem size and core count from complexity analysis. The model for HPL predicts performance even better than the application models do, while the model for G-FFTE systematically underpredicts run times.

Optimizations in Financial Engineering: The Least-Squares Monte Carlo Method of Longstaff and Schwartz

Anamitra Roy Choudhury
IBM India Research Lab, Plot 4
Block-C, Institutional Area
Vasant Kunj, New Delhi, India

Alan King
IBM T. J. Watson Research Center,
10598 Yorktown Heights, NY,USA

Sunil Kumar
Dept. of Electrical Engineering,
University of Southern California,
Los Angeles, USA

Yogish Sabharwal
IBM India Research Lab, Plot 4
Block-C, Institutional Area
Vasant Kunj, New Delhi, India

In this paper we identify important opportunities for parallelization in the Least-Squares Monte Carlo (LSM) algorithm, due to Longstaff and Schwartz, for the pricing of American options. The LSM method can be divided into three phases: Path-simulation, Calibration and Valuation. We describe how each of these phases can be parallelized, with more focus on the Calibration phase, which is inherently more difficult to parallelize.

We implemented these parallelization techniques on Blue Gene using the Quantlib open source financial engineering package. We achieved up to factor of 9 speed-up for the Calibration phase and 18 for the complete LSM method on a 32 processor BG/P system using monomial basis functions.

Massively Parallel Cosmological Simulations with ChaNGa

Pritish Jetley, Filippo Gioachin, Celso Mendes and Laxmikant V. Kalé
Dep. of Computer Science, University of Illinois at Urbana-Champaign, USA
{pjetley2, gioachin, cmendes, kale}@uiuc.edu
Thomas Quinn
Dep. of Astronomy, University of Washington, USA
trq@astro.washington.edu

Cosmological simulators are an important component in the study of the formation of galaxies and large scale structures, and can help answer many important questions about the universe. Despite their utility, existing parallel simulators do not scale effectively on modern machines containing thousands of processors. In this paper we present ChaNGa, a recently released production simulator based on the Charm++ infrastructure. To achieve scalable performance, ChaNGa employs various optimizations that maximize the overlap between com- putation and communication. We present experimental results of ChaNGa simulations on machines with thousands of processors, including the IBM Blue Gene/L and the Cray XT3. The paper goes on to highlight efforts toward even more effcient and scalable cosmological simulations. In particular, novel load balancing schemes that base decisions on certain characteristics of tree-based particle codes are discussed. Further, the multistepping capabilities of ChaNGa are presented, as are solutions to the load imbalance that such multiphase simulations face. We outline key requirements for an effective practical implementation and conclude by discussing preliminary results from simulations run with our multiphase load balancer.

SESSION 3

Architecture - Input/Output

Sacrificing Reliability for Energy Saving: Is It Worthwhile for Disk Arrays?

Tao Xie and Yao Sun
Department of Computer Science
San Diego State University
{xie, calvin}@rohan.sdsu.edu

Mainstream energy conservation schemes for disk arrays inherently affect the reliability of disks. A thorough understanding of the relationship between energy saving techniques and disk reliability is still an open problem, which prevents effective design of new energy saving techniques and application of existing approaches in reliability-critical environments. As one step towards solving this problem, this paper presents an empirical reliability model, called Predictor of Reliability for Energy Saving Schemes (PRESS). Fed by three energy-saving-related reliability-affecting factors, operating temperature, utilization, and disk speed transition frequency, PRESS estimates the reliability of entire disk array. Further, a new energy saving strategy with reliability awareness called Reliability and Energy Aware Distribution (READ) is developed in the light of the insights provided by PRESS. Experimental results demonstrate that compared with existing energy saving schemes, MAID and PDC, READ consistently performs better in performance and reliability while achieving a comparable level of energy consumption.

User-Centric Data Migration in Networked Storage Systems

Sukwoo Kang
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA
sukwkang@us.ibm.com

A. L. Narasimha Reddy
Texas A&M University
College Station, TX 77843, USA
reddy@ece.tamu.edu

This paper considers the problem of balancing locality and load in networked storage systems with multiple storage devices (or bricks). Data distribution affects locality and load balance across the devices in a networked storage system. This paper proposes a user-centric data migration scheme which tries to balance locality and load in such networked storage systems. The presented approach automatically and transparently manages migration of data blocks among disks as data access patterns and loads change over time. We implemented a prototype system, embodying our ideas, on PCs running Linux. This paper presents the design of user-centric migration and an evaluation of it through realistic experiments.

I/O Performance on a Massively Parallel Cray XT3/XT4

Mark Fahey
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6008
faheymr@ornl.gov

Jeff Larkin
Cray Inc.
Oak Ridge, TN 37831
larkin@cray.com

Joylika Adams
Fisk University
Nashville, TN 37208
jadams49@fisk.edu

We present an overview of the current status of input/ output (I/O) on the Cray XT line of supercomputers and provide guidance to application developers and users for achieving efficient I/O. Many I/O benchmark results are presented, based on projected I/O requirements for some widely used scientific applications in the Department of Energy. Finally, we interpret and summarize these benchmark results to provide forward-looking guidance for I/O in large-scale application runs on a Cray XT3/XT4.

Performance Characterization and Optimization of Parallel I/O on the Cray XT

Weikuan Yu, Jeffrey S. Vetter and H. Sarp Oral
Oak Ridge National Laboratory
Oak Ridge, TN 37831
{wyu,vetter,oralhs}@ornl.gov

This paper presents an extensive characterization, tuning, and optimization of parallel I/O on the Cray XT supercomputer, named Jaguar, at Oak Ridge National Laboratory. We have characterized the performance and scalability for different levels of storage hierarchy including a single Lustre object storage target, a single S2A storage couplet, and the entire system. Our analysis covers both data- and metadata-intensive I/O patterns. In particular, for small, non-contiguous data-intensive I/O on Jaguar, we have evaluated several parallel I/O techniques, such as data sieving and two-phase collective I/O, and shed light on their effectiveness. Based on our characterization, we have demonstrated that it is possible, and often prudent, to improve the I/O performance of scientific benchmarks and applications by tuning and optimizing I/O. For example, we demonstrate that the I/O performance of the S3D combustion application can be improved at large scale by tuning the I/O system to avoid a bandwidth degradation of 49% with 8192 processes when compared to 4096 processes. We have also shown that the performance of Flash I/O can be improved by 34% by tuning the collective I/O parameters carefully.

SESSION 4

Software - Redundancy and Faults

On Utilization of Contributory Storage in Desktop Grids

Chreston Miller, Ali R. Butt, and Patrick Butler
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061
{chmille3, butta, pabutler}@cs.vt.edu

Modern desktop grid environments and shared computing platforms have popularized the use of contributory resources, such as desktop computers, as computing substrates for a variety of applications. However, addressing the exponentially growing storage demands of applications, especially in a contributory environment, remains a challenging research problem. In this paper, we propose a transparent distributed storage system that harnesses the storage contributed by desktop grid participants arranged in a peer-to-peer network to yield a scalable, robust, and self- organizing system. The novelty of our work lies in (i) design simplicity to facilitate actual use; (ii) support for easy integration with grid platforms; (iii) innovative use of striping and error coding techniques to support very large data files; and (iv) the use of multicast techniques for data replication. Experimental results through large-scale simulations, verification on PlanetLab, and an actual implementation show that our system can provide reliable and efficient storage with support for large files for desktop grid applications.

Supporting Fault-Tolerance in Streaming Grid Applications

Qian Zhu, Liang Chen and Gagan Agrawal
Department of Computer Science and Engineering
Ohio State University
Columbus, OH, 43210
{zhuq,chenlia,agrawal}@cse.ohio-state.edu

This paper considers the problem of supporting and efficiently implementing fault-tolerance for tightly-coupled and pipelined applications, especially streaming applications, in a grid environment. We provide an alternative to basic check-pointing and use the notion of Light-weight Summary Structure(LSS) to enable efficient failure-recovery. The idea behind LSS is that at certain points during the execution of a processing stage, the state of the program can be summarized by a small amount of memory. This allows us to store copies of LSS for enabling failure-recovery, which causes low overhead fault-tolerance. Our work can be viewed as an optimization and adaptation of the idea of application-level checkpointing to a different execution environment, and for a different class of applications.

Our implementation and evaluation of LSS based failure-recovery has been in the context of the GATES (Grid-based AdapTive Execution on Streams) middleware. An observation we use for providing very low overhead support for fault-tolerance is that algorithms analyzing data streams are only allowed to take a single pass over data, which means they only perform approximate processing. Therefore, we believe that in supporting fault-tolerant execution for these applications, it is acceptable to not analyze a small number of packets of data during failure-recovery. We show how we perform failure-recovery and also demonstrate how we could use additional buffers to limit data loss during the recovery procedure. We also present an efficient algorithm for allocating a new computation resource for failure-recovery at runtime. We have extensively evaluated our implementation using three stream data processing applications, and shown that the use of LSS allows effective and low-overhead failure-recovery.

A Game Theoretical Data Replication Technique for Mobile Ad Hoc Networks

Samee Ullah Khan[1], Anthony A. Maciejewski[1] and Howard Jay Siegel[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
Colorado State University
Fort Collins, CO 80523
{samee.khan, aam, hj}@colostate.edu
Ishfaq Ahmad
Department of Computer Science and Engineering
University of Texas
Arlington, TX 76019
iahmad@cse.uta.edu

Adaptive replication of data items on servers of a mobile ad hoc network can alleviate access delays. The selection of data items and servers requires solving a constrained optimization problem, that is in general NP-complete. The problem is further complicated by frequent partitions of the ad hoc network. In this paper, a mathematical model for data replication in ad hoc networks is formulated. We treat the mobile servers in the ad hoc network as self-interested entities, hence they have the capability to manipulate the outcome of a resource allocation mechanism by misrepresenting their valuations. We design a game theoretic "truthful" mechanism in which replicas are allocated to mobile servers based on reported valuations. We sketch the exact properties of the truthful mechanism and derive a payment scheme that suppresses the selfish behavior of the mobile servers. The proposed technique is extensively evaluated against three ad hoc network replica allocation methods: (a) extended static access frequency, (b) extended dynamic access frequency and neighborhood, and (c) extended dynamic connectivity grouping. The experimental results reveal that the proposed approach outperforms the three techniques in solution quality and has competitive execution times.

Model-Based Fault Localization in Large-Scale Computing Systems

Naoya Maruyama[1] and Satoshi Matsuoka[1,2]
[1] Tokyo Institute of Technology
[2] National Institute of Informatics
naoya.maruyama,matsu@is.titech.ac.jp

We propose a new fault localization technique for software bugs in large-scale computing systems. Our technique always collects per-process function call traces of a target system, and derives a concise execution model that reflects its normal function calling behaviors using the traces. To find the cause of a failure, we compare the derived model with the traces collected when the system failed, and compute a suspect score that quantifies how likely a particular part of call traces explains the failure. The execution model consists of a call probability of each function in the system that we estimate using the normal traces. Functions with low probabilities in the model give high anomaly scores when called upon a failure. Frequently-called functions in the model also give high scores when not called. Finally, we report the function call sequences ranked with the suspect scores to the human analyst, narrowing further manual localization down to a small part of the overall system. We have applied our proposed method to fault localization of a known non-deterministic bug in a distributed parallel job manager. Experimental results on a three-site, 78-node distributed environment demonstrate that our method quickly locates an anomalous event that is highly correlated with the bug, indicating the effectiveness of our approach.

SESSION 5

Algorithms - Numerical Algorithms

A Space- and Time-Efficient Hash Table Hierarchically Indexed by Bloom Filters

Heeyeol Yu and Rabi Mahapatra
Texas A&M University
College Station, TX 77843
Email: {hyyu,rabi}@cs.tamu.edu

Hash tables (HTs) are poorly designed for multiple memory accesses during IP lookup and this design flow critically affects their throughput in high-speed routers. Thus, a high capacity HT with a predictable lookup throughput is desirable. A recently proposed fast HT (FHT) has drawbacks like low on-chip memory utilization for a high-speed router and substantial memory overheads due to off-chip duplicate keys and pointers. Similarly, a Bloomier filter-based HT (BFHT), generating an index to a key table, suffers from setup failures and static membership testing for keys. In this paper, we propose a novel hash architecture which addresses these issues by using pipelined Bloom filters. The proposed scheme, a hierarchically indexed HT (HIHT), generates indexes to a key table for the given key, so that the on-chip memory size is reduced and the overhead of pointers in a linked list is removed. Secondly, an HIHT demonstrates approximately 5.1 and 2.3 times improvement in onchip space efficiency with at most one off-chip memory access, compared to an FHT and a BFHT, respectively. In addition to our analyses on access time and memory space, our simulation for IP lookup with 6 BGP tables shows that an HIHT exhibits 4.5 and 2.0 times on-chip memory efficiencies for 160Gbps router than an FHT and a BFHT, respectively.

Avoiding Communication in Sparse Matrix Computations

James Demmel, Mark Hoemmen, Marghoob Mohiyuddin and Katherine Yelick
Department of Electrical Engineering and Computer Science
University of California at Berkeley
{demmel,mhoemmen,marghoob,yelick}@eecs.berkeley.edu

The performance of sparse iterative solvers is typically limited by sparse matrix-vector multiplication, which is itself limited by memory system and network performance. As the gap between computation and communication speed continues to widen, these traditional sparse methods will suffer. In this paper we focus on an alternative building block for sparse iterative solvers, the "matrix powers kernel" $[x, Ax, A^2x, \cdots, A^kx]$, and show that by organizing computations around this kernel, we can achieve nearminimal communication costs. We consider communication very broadly as both network communication in parallel code and memory hierarchy access in sequential code. In particular, we introduce a parallel algorithm for which the number of messages (total latency cost) is independent of the power k, and a sequential algorithm, that reduces both the number and volume of accesses, so that it is independent of k in both latency and bandwidth costs. This is part of a larger project to develop "communication-avoiding Krylov subspace methods," which also addresses the numerical issues associated with these methods. Our algorithms work for general sparse matrices that "partition well".

On the Representation and Multiplication of Hypersparse Matrices

Aydın Buluç and John R. Gilbert
Department of Computer Science
University of California, Santa Barbara
{aydin,gilbert}@cs.ucsb.edu

Multicore processors are marking the beginning of a new era of computing where massive parallelism is available and necessary. Slightly slower but easy to parallelize kernels are becoming more valuable than sequentially faster kernels that are unscalable when parallelized. In this paper, we focus on the multiplication of sparse matrices (SpGEMM). We first present the issues with existing sparse matrix representations and multiplication algorithms that make them unscalable to thousands of processors. Then, we develop and analyze two new algorithms that overcome these limitations. We consider our algorithms first as the sequential kernel of a scalable parallel sparse matrix multiplication algorithm and second as part of a polyalgorithm for SpGEMM that would execute different kernels depending on the sparsity of the input matrices. Such a sequential kernel requires a new data structure that exploits the hypersparsity of the individual submatrices owned by a single processor after the 2D partitioning. We experimentally evaluate the performance and characteristics of our algorithms and show that they scale significantly better than existing kernels.

Evaluating the Role of Scratchpad Memories in Chip Multiprocessors for Sparse Matrix Computations

Aditya Yanamandra, Bryan Cover, Padma Raghavan, Mary Jane Irwin and Mahmut Kandemir
Department of Computer Science and Engineering
The Pennsylvania State University
{yanamand, cover, raghavan, mji, kandemir}@cse.psu.edu

Scratchpad memories (SPMs) have been shown to be more energy efficient and have faster access times than traditional hardware-managed caches. This, coupled with the predictability of data presence, makes SPMs an attractive alternative to cache for many scientific applications. In this work, we consider an SPM based system for increasing the performance and the energy efficiency of sparse matrix-vector multiplication on a chip multi-processor. We ensure the efficient utilization of the SPM by profiling the application for the data structures which do not perform well in traditional cache. We evaluate the impact of using an SPM at all levels of the on-chip memory hierarchy. Our experimental results show an average increase in performance by 13.5-15% and an average decrease in the energy consumption by 28-33% on an 8-core system depending on which level of the hierarchy the SPM is utilized.

SESSION 6

Applications - P2P Systems Architecture

Energy Efficient Media Streaming in Wireless Hybrid Peer-to-Peer Systems

MarkKai Ho Yeung
Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
khyeung@eee.hku.hk

Yu-Kwong Kwok
Electrical and Computer Engineering
Colorado State University
Fort Collins, CO 80523, USA
Ricky.Kwok@colostate.edu

With the proliferation of sophisticated wireless devices with more than one network interfaces, it is now possible for the devices to form hybrid wireless networks. Specifically, we consider a hybrid wireless networking scenario in which each device has two heterogeneous wireless network interfaces: a server interface (e.g., a CDMA2000 cellular interface) and a peer interface (e.g., a IEEE 802.11g WLAN interface). Our insight is that we could exploit the heterogeneity in energy consumption in such a dual-interface networking capability. In view of the higher energy consumption in using the server interface compared with using the client interface, we propose two novel protocols where neighboring clients form either a master-slave or peer-to-peer relationship to reduce their energy consumption. For the master-slave relationship, each master retrieves media packets from the server and sends them to its slaves via the peer interface. On the other hand, each peer-to-peer relationship consists of one coordinator and at least one helpers. Both coordinator and helpers are responsible for retrieving media packets from the server. Our analysis shows that the two proposed relationships reduce the energy consumption of participating clients. Furthermore, the relationships are stable where rational clients would not voluntarily leave and unilaterally deviate from the coalition. We evaluate their performance in homogeneous and heterogeneous client distributions. Simulation results indicate that both relationships improve streaming performance without violating the energy consumption constraints of clients.

Game-Theoretic Scalable Peer-to-Peer Media Streaming

MarkKai Ho Yeung
Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
khyeung@eee.hku.hk

Yu-Kwong Kwok
Electrical and Computer Engineering
Colorado State University
Fort Collins, CO 80523, USA
Ricky.Kwok@colostate.edu

Peer-to-peer media streaming framework has been widely considered as a promising platform for delivering high quality multimedia content on the global scale. A fundamental requirement is that each peer needs to contribute outgoing bandwidth to deliver media packets to its neighbors. Although most existing protocols mandate such contribution, misbehaving peers may still deliberately limit their outgoing bandwidth to conserve their own resources. This would inevitably lead to performance degradation of other well-behaving peers. It is crucial to have an effective incentive mechanism such that peers are encouraged to contribute. In this paper, we formulate two strategic games to model the interactions between server and its immediate peers and between neighboring peers, respectively. We have devised the equilibrium strategies which relate a peer's streaming performance to its contribution. Simulation results show that the proposed game-theoretical incentive mechanism protects well-behaving peers from being exploited by misbehaving counterparts.

ContinuStreaming: Achieving High Playback Continuity of Gossip-based Peer-to-Peer Streaming

Zhenhua Li
State Key Lab for Novel Software Technology
Nanjing University, Nanjing, P. R. China
lizhenhua@dislab.nju.edu.cn

Jiannong Cao
Internet and Mobile Computing Lab
Hong Kong Polytechnic University, Hong Kong
csjcao@comp.polyu.edu.hk

Guihai Chen
State Key Lab for Novel Software Technology
Nanjing University, Nanjing, P. R. China
gchen@nju.edu.cn

Gossip-based Peer-to-Peer(P2P) streaming has been proved to be an effective and resilient method to stream qualified media contents in dynamic and heterogeneous network environments. Because of the intrinsic randomness of gossiping, some data segments cannot be disseminated to every node in time, which seriously affects the media playback continuity. In this paper we describe the design of ContinuStreaming, a gossip-based P2P streaming system which can maintain high resilience and low overhead while bring a novel and critical property – full coverage of the data dissemination. With the help of DHT, data segments which are likely to be missed by the gossip-based data scheduling can be quickly fetched by the on-demand data retrieval so as to guarantee continuous playback. We discuss the results of both theoretical analysis and comprehensive simulations on various real-trace overlay topologies to demonstrate the effectiveness of our system. Simulation results show that ContinuStreaming outperforms the existing representative gossip-based P2P streaming systems by increasing the playback continuity very close to 1.0 with only 4% or less extra overhead.

PROD: Relayed File Retrieving in Overlay Networks

Zhiyong Xu, Dan Stefanescu and Honggang Zhang
Suffolk University
{zxu, dan, zhang}@mcs.suffolk.edu

Laxmi Bhuyan
University of California, Riverside
bhuyan@cs.ucr.edu

Jizhong Han
Chinese Academy of Science
jzhan@ict.ac.cn

To share and exchange the files among Internet users, Peer-to-Peer (P2P) applications build another layer of overlay networks on top of the Internet Infrastructure. In P2P file sharing systems, a file request takes two steps. First, a routing message is generated by the client (request initiator) and spread to the overlay network. After the process finishes, the location information of the requested file is returned to the client. In the second step, the client establishes direct connection(s) with the peer(s) who store a copy of that file to start the retrieving process.

While numerous research projects have been conducted to design efficient, high-performance routing algorithms, few work concentrated on file retrieving performance. In this paper, we propose a novel and efficient algorithm – PROD to improve the file retrieving performance in DHT based overlay networks. In PROD, when a file or a portion of a file is transferred from a source peer to the client, instead of creating just one direct link between these two peers, we build an application level connection chain. Along the chain, multiple network links are established. Each intermediate peer on this chain uses a store-and-forward mechanism for the data transfer. PROD also introduces a novel topological based strategy to choose these peers and guarantees the transmission delay of each intermediate link is much lower than the direct link. We conducted extensive simulation experiments and the results shown that PROD can greatly reduce the transfer time per file in DHT base P2P systems.

SESSION 7

Architecture - Multi-core

Intermediate Checkpointing with Conflicting Access Prediction in Transactional Memory Systems

M. M. Waliullah and Per Stenstrom
Department of Computer Science and Engineering
Chalmers University of Technology
SE 412-96 Göteborg, Sweden
Email: {waliulla, pers}@chalmers.se

Transactional memory systems promise to reduce the burden of exposing thread-level parallelism in programs by relieving programmers from analyzing complex inter-thread dependences in detail. By encapsulating large program code blocks and executing them as atomic blocks, dependence checking is deferred to run-time at which point one of many conflicting transactions will be committed whereas the others will have to roll-back and re-execute. In current proposals, a checkpoint is taken at the beginning of the atomic block and all execution can be wasted even if the conflicting access happens at the end of the atomic block.

In this paper, we propose a novel scheme that (1) predicts when the first conflicting access occurs and (2) inserts a checkpoint before it is executed. When the prediction is correct, the only execution discarded is the one that has to be redone. When the prediction is incorrect, the whole transaction has to be re-executed just as before. Overall, we find that our scheme manages to maintain high prediction accuracy and leads to a quite significant reduction in the number of lost cycles due to roll-backs; the geometric mean speedup across five applications is 16%.

An Interconnect-Aware Power Efficient Cache Coherence Protocol for CMPs

Hongbo Zeng, Jun Wang, Ge Zhang and Weiwu Hu
Key Laboratory of Computer System and Architecture
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{hbzeng, wangjun, gzhang, hww}@ict.ac.cn

The continuing shrinking of technology enables more and more processor cores to reside on a single chip. However, the power consumption and delay of global wires have presented a great challenge in designing future chip multiprocessors. With these overheads of wires properly accounted for, researchers have explored some efficient onchip network designs in the domain of larger scale caches. While in the paper, we attempt to reduce the interconnect power consumption with a novel cache coherence protocol. Conventional coherence protocols are kept independent from underlying networks for flexibility reasons. But in CMPs, processor cores and the on-chip network are tightly integrated. Exposing features of interconnect networks to protocols will unveil some optimization opportunities for power reduction. Specifically, by utilizing the location information of cores on a chip, the coherence protocol we propose in this work chooses to response the requester with the data copy in the closest sharer of the desired cache line, other than fetching it from distant L2 cache banks. This mechanism reduces the hops cache lines must travel and eliminates the power that would have incurred on the corresponding not-traveled links. To get accurate and detailed power information of interconnects, we extract wire power parameters by physical level simulation (HSPICE) and obtain router power by synthesizing RTL with actual ASIC libraries. We conduct experiments on a 16-core CMP simulator with a group of SPLASH2 benchmarks. The results demonstrate that an average of 16.3% L2 cache accesses could be optimized, resulting in an average 9.3% power reduction of data links with 19.2% as the most. This mechanism also yields a performance speedup of 1.4%.

DiCo-CMP: Efficient Cache Coherency in Tiled CMP Architectures

Alberto Ros, Manuel E. Acacio, José M. García
Departamento de Ingeniería y Tecnología de Computadores
Universidad de Murcia
Campus de Espinardo S/N, 30100 Murcia, Spain
{a.ros,meacacio,jmgarcia}@ditec.um.es

Future CMP designs that will integrate tens of processor cores on-chip will be constrained by area and power. Area constraints make impractical the use of a bus or a crossbar as the on-chip interconnection network, and tiled CMPs organized around a direct interconnection network will prob- ably be the architecture of choice. Power constraints make impractical to rely on broadcasts (as Token-CMP does) or any other brute-force method for keeping cache coherence, and directory-based cache coherence protocols are currently being employed. Unfortunately, directory protocols introduce indirection to access directory information, which negatively impacts performance. In this work, we present DiCo-CMP, a novel cache coherence protocol especially suited to future tiled CMP architectures. In DiCo-CMP the role of storing up-to-date sharing information and ensuring totally ordered accesses for every memory block is assigned to the cache that must provide the block on a miss. Therefore, DiCo-CMP reduces the miss latency compared to a directory protocol by sending coherence messages directly from the requesting caches to those that must observe them (as it would be done in brute-force protocols), and reduces the network traffic compared to Token-CMP (and consequently, power consumption in the interconnection network) by sending just one request message for each miss. Using an extended version of GEMS simulator we show that DiCo-CMP achieves improvements in execution time of up to 8% on average over a directory protocol, and reductions in terms of network traffic of up to 42% on average compared to Token-CMP.

HelperCore$_{DB}$: Exploiting Multicore Technology to Improve Database Performance

Kostas Papadopoulos, Kyriakos Stavrou and Pedro Trancoso
Department of Computer Science, University of Cyprus, Cyprus
Email: {csp5kp1,tsik,pedro}@cs.ucy.ac.cy
http://www.cs.ucy.ac.cy/carch/casper

Due to limitations in the traditional microprocessor design, such as high complexity and power, all current commercial high-end processors contain multiple cores on the same chip (multicore). This trend is expected to continue resulting in increasing number of cores on the chip. While these cores may be used to achieve higher throughput, improving the execution of a single application requires careful coding and usually results in smaller benefit as the scale of the system is increased.

In this work we propose an alternative way to exploit the multicore technology where certain cores execute code which indirectly improves the performance of the application. We call this approach the Helper Core approach. The main contribution of this work is the proposal and evaluation of HelperCore$_{DB}$, a Helper Core designed to improve the performance of database workloads by performing efficient data prefetching.

We validate the proposed approach using a HelperCore$_{DB}$ implementation for the PostgreSQL DBMS. The approach is evaluated with native execution on a dual core processor system using the standard TPC-H benchmark. The experimental results show a significant performance improvement specially considering that the baseline system is a modern optimized processor that includes advanced features such as hardware prefetching. In particular, for the 22 queries of TPC-H, our approach achieves a large reduction of the secondary cache misses, 75% on average, and an improvement in the execution time of up to 21.5%.

SESSION 8

Software - Implementing Message Passing

Receiver-initiated Message Passing Over RDMA Networks

Scott Pakin
pakin@lanl.gov
CCS-1 Performance and Architecture Lab (PAL)
Los Alamos National Laboratory

Providing point-to-point messaging-passing semantics atop Put/Get hardware traditionally involves implementing a protocol comprising three network latencies. In this paper, we analyze the performance of an alternative implementation approachłreceiver-initiated message passing–that eliminates one of the three network latencies. Performance measurements taken on the Cell Broadband Engine indicate that receiver-initiated message passing exhibits substantially lower latency than standard, sender-initiated message passing.

MVAPICH-Aptus: Scalable High-Performance Multi-Transport MPI Over InfiniBand

Matthew J. Koop[1,2], Terry Jones[2] and Dhabaleswar K. Panda[1]

[1] Network-Based Computing Laboratory
The Ohio State University
Columbus, OH 43210
{koop, panda}@cse.ohio-state.edu

[2]Lawrence Livermore National Laboratory
Livermore, CA 94550
trj@llnl.gov

The need for computational cycles continues to exceed availability, driving commodity clusters to increasing scales. With upcoming clusters containing tens-of-thousands of cores, InfiniBand is a popular interconnect on these clusters, due to its low latency (1.5μsec) and high bandwidth (1.5 GB/sec). Since most scientific applications running on these clusters are written using the Message Passing Interface (MPI) as the parallel programming model, the MPI library plays a key role in the performance and scalability of the system. Nearly all MPIs implemented over InfiniBand currently use the Reliable Connection (RC) transport of InfiniBand to implement message passing. Using this transport exclusively, however, has been shown to potentially reach a memory footprint of over 200MB/task at 16K tasks for the MPI library. The Unreliable Datagram (UD) transport, however, offers higher scalability, but at the cost of medium and large message performance.

In this paper we present a multi-transport MPI design, MVAPICH-Aptus, that uses both the RC and UD transports of InfiniBand to deliver scalability and performance higher than that of a single-transport MPI design. Evaluation of our hybrid design on 512 cores shows a 12% improvement over an RC-based design and 4% better than a UD-based design for the SMG2000 application benchmark. In addition, for the molecular dynamics application NAMD we show a 10% improvement over an RC-only design. To the best of our knowledge, this is the first such analysis and design of optimized MPI using both UD and RC.

Designing Passive Synchronization for MPI-2 One-Sided Communication to Maximize Overlap

Gopal Santhanaraman, Sundeep Narravula and Dhabaleswar K. Panda
Department of Computer Science and Engineering
The Ohio State University
{santhana, narravul, panda}@cse.ohio-state.edu

Scientific computing has seen an immense growth in recent years. MPI has become the de facto standard for parallel programming model for distributed memory systems. MPI-2 standard expanded MPI to include onesided communications. Computation and communication overlap is an important goal for one-sided applications. While the passive synchronization mechanism for MPI-2 one-sided communication allows for good overlap, the actual overlap achieved is often limited by the design of both the MPI library and the application.

In this paper we aim to improve the performance of MPI-2 one-sided communication. In particular, we focus on the following important aspects: (i) designing one-sided passive synchronization (Direct Passive) support using InfiniBand atomic operations to handle both exclusive as well as shared locks (ii) enhancing one-sided communication progress to provide scope for better overlap that one-sided applications can leverage. (iii) study the overlap potential of passive synchronization and its impact on applications. We demonstrate the possible benefits of our approaches for the MPI-2 SPLASH LU application benchmark. Our results show an improvement of up to 87% for a 64 process run over the existing design.

Scalable Group-based Checkpoint/Restart for Large-Scale Message-passing Systems

Justin C. Y. Ho, Cho-Li Wang and Francis C. M. Lau
Department of Computer Science,
The University of Hong Kong,
{cyho2,clwang,fcmlau}@cs.hku.hk

The ever increasing number of processors used in parallel computers is making fault tolerance support in large-scale parallel systems more and more important. We discuss the inadequacies of existing system-level checkpointing solutions for message-passing applications as the system scales up. We analyze the coordination cost and blocking behavior of two current MPI implementations with checkpointing support. A group-based solution combining coordinated checkpointing and message logging is then proposed. Experiment results demonstrate its better performance and scalability than LAM/MPI and MPICH-VCL. To assist group formation, a method to analyze the communication behaviors of the application is proposed.

SESSION 9

Algorithms - P2P and Overlay Networks

Self-Optimizing Distributed Trees

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

Asad Samar
Goldman Sachs International
London, UK
Asad.Samar@gs.com

Chenxi Wang
Forrester Research
Foster City, CA, USA
chenxi.wang@gmail.com

We present a novel protocol for restructuring a tree-based overlay network in response to the workload of the application running over it. Through low-cost restructuring operations, our protocol incrementally adapts the tree so as to bring nodes that tend to communicate with one another closer together in the tree. It achieves this while respecting degree bounds on nodes so that, e.g., no node degenerates into a "hub" for the overlay. Moreover, it limits restructuring to those parts of the tree over which communication takes place, avoiding restructuring other parts of the tree unnecessarily. We show via experiments on PlanetLab that our protocol can significantly reduce communication latencies in workloads dominated by clusters of communicating nodes.

Towards a Decentralized Architecture for Optimization

Marco Biazzini, Mauro Brunato and Alberto Montresor
University of Trento
Dipartimento di Ingegneria e Scienza dell'Informazione
via Sommarive 14, 38100 Trento, Italy
{biazzini,brunato,montresor}@disi.unitn.it

We introduce a generic framework for the distributed execution of combinatorial optimization tasks. Instead of relying on custom hardware (like dedicated parallel machines or clusters), our approach exploits, in a peer-to-peer fashion, the computing and storage power of existing, off-the-shelf desktops and servers. Contributions of this paper are a description of the generic framework, together with a first instantiation based on particle swarm optimization (PSO). Simulation results are shown, proving the efficacy of our distributed PSO algorithm in optimizing a large number of benchmark functions.

Heterogenous Dating Service with Application to Rumor Spreading

Olivier Beaumont and Philippe Duchon
LaBRI . University of Bordeaux, INRIA Futurs
351 cours de la Liberation,
FR-33405 Talence, France
{obeaumon,duchon}@labri.fr

Miroslaw Korzeniowski
Wroclaw University of Technology
ul. Wybrzeze Wyspianskiego 27,
PL-50370 Wroclaw, Poland
Miroslaw.Korzeniowski@pwr.wroc.pl

Peer-to-Peer overlay networks have proven their efficiency for storing and retrieving data at large scale, but new services are required to take the actual performances of resources into account. In this paper, we describe a fully decentralized algorithm, called “dating service” meant to organize communications in a fully heterogeneous network, that ensures that communication capabilities of the nodes are not exceeded. We prove that with high probability, this service ensures that a constant fraction of all possible communications is organized. Interestingly enough, this property holds true even if a node is not able to choose another node uniformly at random. In particular, the dating service can be implemented over existing DHT-based systems. In order to illustrate the expressiveness and the usefulness of proposed service, we also present a possible practical application of the dating service. As an illustration, we propose an algorithm for rumor spreading that enables to broadcast a unit-size message to all the nodes of a P2P system in logarithmic number of steps with high probability.

Self-stabilizing Population of Mobile Agents

Zbigniew Gołębiewski
Institute of Computer Science
Wrocłw University
ul. Joliot–Curie 15
50–83 Wrocłw, Poland
Zbigniew.Golebiewski@ii.uni.wroc.pl

Mirosłw Kutyłwski
Institute of Mathematics and Computer Science
Wrocłw University of Technology
ul. Wybrzeże Wyspiańskiego 27
50–70 Wrocłw, Poland
mirekk@im.pwr.wroc.pl

Tomasz Łuczak
Faculty of Mathematics and Computer Science
Adam Mickiewicz University
ul. Umultowska 87
61–14 Poznań, Poland
tomasz@amu.edu.pl

Filip Zagórski
Institute of Mathematics and Computer Science
Wrocłw University of Technology
ul. Wybrzeże Wyspiańskiego 27
50–70 Wrocłw, Poland
filipz@im.pwr.wroc.pl

We investigate a problem of maintaining a target population of mobile agents in a distributed system. The purpose of the agents is to perform certain activities, so the goal is to avoid overpopulation (leading to waste of resources) as well as underpopulation (resulting in a poor service). We assume that there must be no centralized control over the number of agents, since it might result in system’s vulnerability.

We analyze a simple protocol in which each node keeps at most one copy of an agent and if there is a single agent in a node, a new agent is born with a certain probability p. At each time step the agents migrate independently at random to chosen locations.

We show that during a protocol execution the number of agents stabilizes around a level depending on p. We derive analytically simple formulas that determine probability p based on the target fraction of nodes holding an agent. The previous proposals of this type were based on experimental data only.

SESSION 10

Applications - Grids

Decentralized Market-Based Resource Allocation in a Heterogeneous Computing System

James Smith[1,2], Edwin K. P. Chong[2],4, Anthony A. Maciejewski[2] and Howard Jay Siegel[2,3]

[1]DigitalGlobe
Longmont, CO 80503 USA
Email: jtsmith@digitalglobe.com

Colorado State University
[2]Dept. of Electrical and Computer Engineering
[3]Dept. of Computer Science
[4]Mathematics Department
Fort Collins, CO 80523C1373, USA
Email: {echong, aam, hj}@engr.colostate.edu

We present a decentralized market-based approach to resource allocation in a heterogeneous overlay network. The presented resource allocation strategy assigns overlay network resources to traffic dynamically based on current utilization, thus, enabling the system to accommodate fluctuating demand for its resources. We present a mathematical model of our resource allocation environment that treats the allocation of system resources as a constrained optimization problem. Our presented resource allocation strategy is based on solving the dual of this centralized optimization problem. The solution to the dual of our centralized optimization problem suggests a simple decentralized algorithm for resource allocation that is extremely efficient. Our results demonstrate the near optimality of the proposed approach through extensive simulation of a real-world environment. That is, the conducted simulation study utilizes components taken from a real-world middleware application environment and clearly demonstrates the practicality of the approach in a realistic setting.

SLA-based Resource Allocation in Cluster Computing Systems

Kaiqi Xiong and Harry Perros
Department of Computer Science
North Carolina State University
Raleigh, NC 27965-7534
{xiong,hp}@csc.ncsu.edu

Resource allocation is a fundamental but challenging problem due to the complexity of cluster computing systems. In enterprise service computing, resource allocation is often associated with a service level agreement (SLA) which is a set of quality of services and a price agreed between a customer and a service provider. The SLA plays an important role in an e-business application. A service provider uses a set of computer resources to support e-business applications subject to an SLA. In this paper, we present an approach for computer resource allocation in such an environment that minimizes the total cost of computer resources used by a service provider for an e-business application while satisfying the quality of service (QoS) defined in an SLA. These QoS metrics include percentile response time, cluster utilization, packet loss rate and cluster availability. Simulation results show the applicability of the approach and validate its accuracy.

Optimizing XML Processing for Grid Applications Using an Emulation Framework

Rajdeep Bhowmik, Chaitali Gupta and Madhusudhan Govindaraju
Department of Computer Science
State University of New York (SUNY) at Binghamton
NY, 13902
Aneesh Aggarwal
Electrical and Computer Engineering
State University of New York (SUNY) at Binghamton
NY, 13902

Chip multi-processors (CMPs), commonly referred to as multi-core processors, are being widely adopted for deployment as part of the grid infrastructure. This change in computer architecture requires corresponding design modifications in programming paradigms, including grid middleware tools, to harness the opportunities presented by multi-core processors. Simple and naive implementations of grid middleware on multi-core systems can severely impact performance. This is because programming for CMPs requires special consideration for issues such as limitations of shared bus bandwidth, cache size and coherency, and communication between threads. The goal of developing an optimized multi-threaded grid middleware for emerging multi-core processors will be realized only if researchers and developers have access to an in-depth analysis of the impact of several low level microarchitectural parameters on performance. None of the current grid simulators and emulators provide feedback at the microarchitectural level, which is essential for such an analysis. In earlier work we presented our initial results on the design and implementation of such an emulation framework, Multicore Grid (McGrid). In this paper we extend that work and present a performance study on the effect of cache coherency, scheduling of processing threads to take advantage of data available in the cache of each core, and read and write access patterns for shared data structures. We present the performance results, analysis, and recommendations based on experiments conducted using the McGrid framework for processing XML-based grid data and documents.

Overcoming Scaling Challenges in Biomolecular Simulations across Multiple Platforms

Abhinav Bhatelé[1], Sameer Kumar[2], Chao Mei[1], James C. Phillips[3], Gengbin Zheng[1] and Laxmikant V. Kalé[1]

[1] Department of Computer Science
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: {bhatele2, chaomei2, gzheng, kale}@uiuc.edu
[2] IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
E-mail: sameerk@us.ibm.com
[3] Beckman Institute
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: jim@ks.uiuc.edu

NAMD is a portable parallel application for biomolecular simulations. NAMD pioneered the use of hybrid spatial and force decomposition, a technique now used by most scalable programs for biomolecular simulations, including Blue Matter and Desmond developed by IBM and D. E. Shaw respectively. NAMD has been developed using Charm++ and benefits from its adaptive communication-computation overlap and dynamic load balancing. This paper focuses on new scalability challenges in biomolecular simulations: using much larger machines and simulating molecular systems with millions of atoms. We describe new techniques developed to overcome these challenges. Since our approach involves automatic adaptive runtime optimizations, one interesting issue involves dealing with harmful interaction between multiple adaptive strategies. NAMD runs on a wide variety of platforms, ranging from commodity clusters to supercomputers. It also scales to large machines: we present results for up to 65,536 processors on IBM's Blue Gene/L and 8,192 processors on Cray XT3/XT4. In addition, we present performance results on NCSA's Abe, SDSC's DataStar and TACC's LoneStar cluster, to demonstrate efficient portability. We also compare NAMD with Desmond and Blue Matter.

SESSION 11

Architecture - Supercomputing/SIMD

A Plug-and-Play Model for Evaluating Wavefront Computations on Parallel Architectures

Gihan R. Mudalige[1],Mary K. Vernon[2] and Stephen A. Jarvis[1]

[1]Dept. of Computer Science
University of Warwick
Coventry, CV4 7AL, U.K
{g.r.mudalige,saj}@dcs.warwick.ac.uk

[2]ept. of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706-1685
vernon@cs.wisc.edu

This paper develops a plug-and-play reusable LogGP model that can be used to predict the runtime and scaling behavior of different MPI-based pipelined wavefront applications running on modern parallel platforms with multicore nodes. A key new feature of the model is that it requires only a few simple input parameters to project performance for wavefront codes with different structure to the sweeps in each iteration as well as different behavior during each wavefront computation and/or between iterations. We apply the model to three key benchmark applications that are used in high performance computing procurement, illustrating that the model parameters yield insight into the key differences among the codes. We also develop new, simple and highly accurate models of MPI send, receive, and group communication primitives on the dual-core Cray XT system. We validate the reusable model applied to each benchmark on up to 8192 processors on the XT3/XT4. Results show excellent accuracy for all high performance application and platform configurations that we were able to measure. Finally we use the model to assess application and hardware configurations, develop new metrics for procurement and configuration, identify bottlenecks, and assess new application design modifications that, to our knowledge, have not previously been explored.

Massive Supercomputing Coping with Heterogeneity of Modern Accelerators

Toshio Endo and Satoshi Matsuoka
Tokyo Institute of Technology/JST, Japan

Heterogeneous supercomputers with combined general-purpose and accelerated CPUs promise to be the future major architecture due to their wide-ranging generality and superior performance / power ratio. However, developing applications that achieve effective scalability is still very difficult, and in fact unproven on large-scale machines in such combined setting. We show that an effective method for such heterogeneous systems so that the porting from applications written with homogeneous assumptions could be achieved. For this goal, we divide porting of applications into several steps, analyze performance of the kernel computation, create processes that virtualize the underlying processors, tune parameters with preferences to accelerators, and balance the load between heterogeneous nodes. We apply our method to the parallel Linpack benchmark on the TSUBAME heterogeneous supercomputer. We efficiently utilize both 10,000 general purpose CPU cores and 648 SIMD accelerators in a combined fashion—the resulting 56.43 TFlops utilized the entire machine, and not only ranked significantly on the Top500 supercomputer list, but also it is the highest Linpack performance on heterogeneous systems in the world.

DC-SIMD : Dynamic Communication for SIMD processors

Raymond Frijns, Hamed Fatemi, Bart Mesman and Henk Corporaal
Eindhoven University of Technology
Den Dolech 2, NL-5600 MB Eindhoven, The Netherlands
R.M.W.Frijns@student.tue.nl
{H.Fatemi, B.Mesman, H.Corporaal}@tue.nl

SIMD (single instruction multiple data)-type processors have been found very efficient in image processing applications, because their repetitive structure is able to exploit the huge amount of data-level parallelism in pixel-type operations, operating at a relatively low energy consumption rate. However, current SIMD architectures lack support for dynamic communication between processing elements, which is needed to efficiently map a set of non-linear algorithms. An architecture for dynamic communication support has been proposed, but this architecture needs large amounts of buffering to function properly. In this paper, three architectures supporting dynamic communication without the need of large amounts of buffering are presented, requiring 98% less buffer space. Cycle-true communication architecture simulators have been developed to accurately predict the performance of the different architectures. Simulations with several test algorithms have shown a performance improvement of up to 5x compared to a locally connected SIMD-processor. Also, detailed area models have been developed, estimating the three proposed architectures to have an area overhead of 30-70%compared to a locally connected SIMD architecture (like the IMAP).When memory is taken into account as well, the overhead is estimated to be 13-28%.

Analysis of Double Buffering on Two Different Multicore Architectures: Quad-core Opteron and the Cell-BE

José Carlos Sancho and Darren J. Kerbyson
Performance and Architecture Laboratory (PAL)
Computer Science for HPC (CCS-1)
Los Alamos National Laboratory, NM 87545, USA
{jcsancho,djk}@lanl.gov

In order to take full advantage of multi-core processors careful attention must be given to the way in which each core interacts with main memory. In data-rich parallel applications multiple transfers between the main memory and local memory (cache or other) of each core will be required. It will be increasingly important to overlap these data transfers with useful computation in order to achieve high performance. One approach to exploit this compute-transfer overlap is to use double-buffering techniques that require minimal resources in the local memory available to the cores. In this paper, we present optimized buffering techniques and evaluate them for two state-of-the-art multi-core architectures: quad-core Opteron and the Cell-BE. Experimental results show that using double buffering can substantially deliver higher performance for codes with data-parallel loop structures. Performance improvements of 1.4× and 2.2× can be achieved for the quad-core Opteron and Cell-BE respectively. Moreover, this study also provides insight into the application characteristics required for achieving improved performance when using double-buffering, and also the tuning that is required in order to achieve optimal performance.

Plenary Session

Best Papers

A New Diffusion-based Multilevel Algorithm for Computing Graph Partitions of Very High Quality

Henning Meyerhenke, Burkhard Monien and Thomas Sauerwald
University of Paderborn
Department of Computer Science
Fuerstenallee 11, D-33102 Paderborn
{henningm— bm — sauerwal} @ uni-paderborn.de

Graph partitioning requires the division of a graph's vertex set into k equally sized subsets such that some objective function is optimized. For many important objective functions, e.g., the number of edges incident to different partitions, the problem is NP-hard. Graph partitioning is an important task in many applications, so that a variety of algorithms and tools for its solution have been developed. Most state-of-the-art graph partitioning libraries use a variant of the Kernighan-Lin (KL) heuristic within a multilevel framework. While these libraries are very fast, their solutions do not always meet all requirements of the users. This includes the choice of the appropriate objective function and the shape of the computed partitions. Moreover, due to its sequential nature, the KL heuristic is not easy to parallelize. Thus, its use as a load balancer in parallel numerical applications requires complicated adaptations. That is why we have developed previously an inherently parallel algorithm, called Bubble-FOS/C (Meyerhenke et al., IPDPS'06), which optimizes the partition shapes by a diffusive mechanism. Yet, it is too slow to be of real practical use, despite its high solution quality.

In this paper, besides proving that Bubble-FOS/C converges towards a local optimum, we develop a much faster method for the improvement of partitionings. It is based on a different diffusive process, which is restricted to local areas of the graph and also contains a high degree of parallelism. By coupling this new technique with Bubble-FOS/C in a multilevel framework based on two different hierarchy construction methods, we obtain our new graph partitioning heuristic DibaP. Compared to Bubble-FOS/C, it shows a considerable acceleration, while retaining the positive properties of the slower algorithm.

Experiments with popular benchmark graphs show an extremely good behavior. First, DibaP computes consistently better results ?measured by the edge-cut and the number of boundary vertices in the summation and the maximum norm - than the state-of-the-art libraries METIS and JOSTLE. Second, with our new algorithm, we have improved the best known edge-cut values for a significant number of partitionings of six widely used benchmark graphs.

Parallel IP Lookup Using Multiple SRAM-based Pipelines

Weirong Jiang and Viktor K. Prasanna
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089, USA
{weirongj, prasanna}@usc.edu

Pipelined SRAM-based algorithmic solutions have become competitive alternatives to TCAMs (ternary content addressable memories) for high throughput IP lookup. Multiple pipelines can be utilized in parallel to improve the throughput further. However, several challenges must be addressed to make such solutions feasible. First, the memory distribution over different pipelines as well as across different stages of each pipeline must be balanced. Second, the traffic among these pipelines should be balanced. Third, the intra-flow packet order should be preserved. In this paper, we propose a parallel SRAM-based multi-pipeline architecture for IP lookup. A two-level mapping scheme is developed to balance the memory requirement among the pipelines as well as across the stages in a pipeline. To balance the traffic, we propose a flow pre-caching scheme to exploit the inherent caching in the architecture. Our technique uses neither a large reorder buffer nor complex reorder logic. Instead, a payload exchange scheme exploiting the pipeline delay is used to maintain the intra-flow packet order. Extensive simulation using real-life traffic traces shows that the proposed architecture with 8 pipelines can achieve a throughput of up to 10 billion packets per second (GPPS) while preserving intra-flow packet order.

Lattice Boltzmann Simulation Optimization on Leading Multicore Platforms

Samuel Williams[1,2], Jonathan Carter[1], Leonid Oliker[1], John Shalf[1] and Katherine Yelick[1,2]
[1]CRD/NERSC, Lawrence Berkeley National Laboratory Berkeley, CA 94720
[2]S Division, University of California at Berkeley, Berkeley, CA 94720
{SWWilliams, JTCarter, LOliker, JShalf, KAYelick}@lbl.gov

We present an auto-tuning approach to optimize application performance on emerging multicore architectures. The methodology extends the idea of searchbased performance optimizations, popular in linear algebra and FFT libraries, to application-specific computational kernels. Our work applies this strategy to a lattice Boltzmann application (LBMHD) that historically has made poor use of scalar microprocessors due to its complex data structures and memory access patterns. We explore one of the broadest sets of multicore architectures in the HPC literature, including the Intel Clovertown, AMD Opteron X2, Sun Niagara2, STI Cell, as well as the single core Intel Itanium2. Rather than hand-tuning LBMHD for each system, we develop a code generator that allows us identify a highly optimized version for each platform, while amortizing the human programming effort. Results show that our autotuned LBMHD application achieves up to a 14× improvement compared with the original code. Additionally, we present detailed analysis of each optimization, which reveal surprising hardware bottlenecks and software challenges for future multicore systems and applications.

A Helper Thread Based EDP Reduction Scheme for Adapting Application Execution in CMPs

Yang Ding, Mahmut Kandemir, Padma Raghavan and Mary Jane Irwin
Department of Computer Science & Engineering
Pennsylvania State University, University Park, PA 16802, USA
{yding, kandemir, raghavan, mji}@cse.psu.edu

In parallel to the changes in both the architecture domain C the move toward chip multiprocessors (CMPs) C and the application domain C the move toward increasingly data-intensive workloads C issues such as performance, energy efficiency and CPU availability are becoming increasingly critical. The CPU availability can change dynamically due to several reasons such as thermal overload, increase in transient errors, or operating system scheduling. An important question in this context is how to adapt, in a CMP, the execution of a given application to CPU availability change at runtime. Our paper studies this problem, targeting the energy-delay product (EDP) as the main metric to optimize. We first discuss that, in adapting the application execution to the varying CPU availability, one needs to consider the number of CPUs to use, the number of application threads to accommodate and the voltage/frequency levels to employ (if the CMP has this capability). We then propose to use helper threads to adapt the application execution to CPU availability change in general with the goal of minimizing the EDP. The helper thread runs parallel to the application execution threads and tries to determine the ideal number of CPUs, threads and voltage/frequency levels to employ at any given point in execution. We illustrate this idea using two applications (Fast Fourier Transform and MultiGrid) under different execution scenarios. The results collected through our experiments are very promising and indicate that significant EDP reductions are possible using helper threads. For example, we achieved up to 66.3% and 83.3% savings in EDP when adjusting all the parameters properly in applications FFT and MG, respectively.

SESSION 12

Algorithms - Communication Algorithms

A Predicate-Based Approach to Dynamic Protocol Update in Group Communication

Olivier Rütti and André Schiper
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
{Olivier.Rutti, Andre.Schiper}@epfl.ch

In this paper we study dynamic protocol updates (DPU), which consist in replacing, without interruption, a given protocol during execution. We focus especially on group communication protocols. The paper proposes a methodology to conveniently describe which protocols are correctly replaced by a given DPU algorithm. More precisely, our methodology characterizes DPU algorithms by a set of inference rules. To validate our approach, we illustrate our methodology with a new DPU algorithm.

SNAP, Small-world Network Analysis and Partitioning: An Open-source Parallel Graph Framework for the Exploration of Large-scale Networks

David A. Bader and Kamesh Madduri
College of Computing
Georgia Institute of Technology
{bader,kamesh}@cc.gatech.edu

We present SNAP (Small-world Network Analysis and Partitioning), an open-source graph framework for exploratory study and partitioning of large-scale networks. To illustrate the capability of SNAP, we discuss the design, implementation, and performance of three novel parallel community detection algorithms that optimize modularity, a popular measure for clustering quality in social network analysis. In order to achieve scalable parallel performance, we exploit typical network characteristics of small-world networks, such as the low graph diameter, sparse connectivity, and skewed degree distribution. We conduct an extensive experimental study on real-world graph instances and demonstrate that our parallel schemes, coupled with aggressive algorithm engineering for small-world networks, give significant running time improvements over existing modularity-based clustering heuristics, with little or no loss in clustering quality. For instance, our divisive clustering approach based on approximate edge betweenness centrality is more than two orders of magnitude faster than a competing greedy approach, for a variety of large graph instances on the Sun Fire T2000 multicore system. SNAP also contains parallel implementations of fundamental graph-theoretic kernels and topological analysis metrics (e.g., breadth-first search, connected components, vertex and edge centrality) that are optimized for small-world networks. The SNAP framework is extensible; the graph kernels are modular, portable across shared memory multicore and symmetric multiprocessor systems, and simplify the design of high-level domain-specific applications.

Efficient MPI_Bcast Across Different Process Arrival Patterns

Pitch Patarasuk and Xin Yuan
Department of Computer Science, Florida State University
Tallahassee, FL 32306
{patarasu, xyuan}@cs.fsu.edu

A Message Passing Interface (MPI) collective operation such as broadcast involves multiple processes. The process arrival pattern denotes the timing when each process arrives at a collective operation. It can have a profound impact on the performance since it decides the time when each process can start participating in the operation. In this paper, we investigate the broadcast operation with different process arrival patterns. We analyze commonly used broadcast algorithms and show that they cannot guarantee high performance for different process arrival patterns. We develop two process arrival pattern aware algorithms for broadcasting large messages. The performance of proposed algorithms is theoretically within a constant factor of the optimal for any given process arrival pattern. Our experimental evaluation confirms the analytical results: existing broadcast algorithms cannot achieve high performance for many process arrival patterns while the proposed algorithms are robust and efficient across different process arrival patterns.

Providing Flow Based Performance Guarantees for Buffered Crossbar Switches

Deng Pan
Dept. of Electrical & Computer Engineering
Florida International University
Miami, Florida 33174, USA
pand@fiu.edu

Yuanyuan Yang
Dept. of Electrical & Computer Engineering
State University of New York
Stony Brook, NY 11794, USA
yang@ece.sunysb.edu

Buffered crossbar switches are a special type of combined input-output queued switches with each crosspoint of the crossbar having small on-chip buffers. The introduction of crosspoint buffers greatly simplifies the scheduling process of buffered crossbar switches, and furthermore enables buffered crossbar switches with speedup of two to easily provide port based performance guarantees. However, recent research results have indicated that, in order to provide flow based performance guarantees, buffered crossbar switches have to either increase the speedup of the crossbar to three or greatly increase the total number of crosspoint buffers, both adding significant hardware complexity. In this paper, we present scheduling algorithms for buffered crossbar switches to achieve flow based performance guarantees with speedup of two and with only one or two buffers at each crosspoint. When there is no crosspoint blocking in a specific time slot, only the simple and distributed input scheduling and output scheduling are necessary. Otherwise, the special urgent matching is introduced to guarantee the on-time delivery of crosspoint blocked cells. With the proposed algorithms, buffered crossbar switches can provide flow based performance guarantees by emulating push-in-first-out output queued switches, and we use the counting method to formally prove the perfect emulation. For the special urgent matching, we present sequential and parallel matching algorithms. Both algorithms converge with N iterations in the worst case, and the latter needs less iterations in the average case. Finally, we discuss an alternative backup-buffer implementation scheme to the bypass path, and compare our algorithms with existing algorithms in the literature.

SESSION 13

Applications - P2P Structure

On Performance Bottleneck of Anonymous Communication Networks

Ryan Pries[1], Wei Yu[2], Steve Graham[1] and Xinwen Fu[1]
[1]Dakota State University, Madison, SD 57042
{priesr,Steve.Graham,xinwen.fu}@dsu.edu
[2]Texas A&M University, College Station, TX 77843
weiyu@cs.tamu.edu

Although a significant amount of effort has been directed at discovering attacks against anonymity communication networks and developing countermeasures to those attacks, there is little systematic analysis of the Quality of Service (QoS) for such privacy preserving systems. In this paper, we initiate the effort to study the QoS of Tor, a popular and representative implementation of anonymous communication networks on the Internet. We find that Tor suffers severe TCP performance degradation because of its random path selection strategy. Our investigation shows that Tor's bandwidth weighted path selection algorithm can only improve the performance to a very limited extent. We analyze this performance issue from the perspective of overlay networks and model the TCP throughput of Tor. We conduct extensive experiments on the real-world Tor network and the experimental results validate our theory. We also discuss possible remedies to this performance issue.

DHT-assisted Probabilistic Exhaustive Search in Unstructured P2P Networks

Xucheng Luo and Zhiguang Qin
School of Computer
Science and Engineering
University of Electronic
Science and Technology of China
Chengdu, 610054, China
xucheng, qinzg@uestc.edu.cn

Jinsong Han
Department of Computer
Science and Engineering
Hong Kong University
of Science and Technology
Kowloon, Hong Kong
jasonhan@cse.ust.hk

Hanhua Chen
School of Computer
Science and Technology
Huazhong University of
Science and Technology
Wuhan, 430074, China
chenhanhua@hust.edu.cn

Existing replication strategies in unstructured P2P networks, such as square-root principle based replication, can effectively improve search efficiency. How to get optimal replication strategy, however, is not trivial. In this paper we show, through mathematical proof, that random replication strategy achieves the optimal results. By randomly distributing rather small numbers of item and query replicas in the unstructured P2P network, we can guarantee perfect search success rate comparable to exhaustive search with high probability. Our analysis also shows that the cost for such replication strategy is determined by the network size of a P2P system. We propose a hybrid P2P architecture which combines a lightweight DHT with an unstructured P2P overlay to address the problems of network size estimating and random peer sampling. We conduct comprehensive simulation to evaluate this design. Results show that our scheme achieves perfect search success rate with quite small overhead.

An Efficient Hybrid Peer-to-Peer System for Distributed Data Sharing

Min Yang and Yuanyuan Yang
Department of Electrical & Computer Engineering,
State University of New York, Stony Brook, NY 11794, USA

Peer-to-peer overlay networks are widely used in distributed systems. Based on whether a regular topology is maintained among peers, peer-to-peer networks can be divided into two categories: structured peer-to-peer networks in which peers are connected by a regular topology, and unstructured peer-to-peer networks in which the topology is arbitrary. Structured peer-to-peer networks usually can provide efficient and accurate services but need to spend a lot of efforts in maintaining the regular topology. On the other hand, unstructured peer-to-peer networks are extremely resilient to the frequent peer joining and leaving but this is usually achieved at the expense of efficiency. The objective of this work is to design a hybrid peer-to-peer system for distributed data sharing which combines the advantages of both types of peer-to-peer networks and minimizes their disadvantages. The proposed hybrid peer-to-peer system is composed of two parts: the first part is a structured core network which forms the backbone of the hybrid system; the second part is multiple unstructured peer-to-peer networks each of which is attached to a node in the core network. The core structured network can narrow down the data lookup within a certain unstructured network accurately, while the unstructured networks provide a low cost mechanism for peers to join or leave the system freely. A data lookup operation first checks the local unstructured network and then the structured network. This two-tier hierarchy can decouple the flexibility of the system from the efficiency of the system. Our simulation results demonstrate that the hybrid peer-to-peer system can utilize both the efficiency of structured peer-to-peer network and the flexibility of the unstructured peer-to-peer network and achieve a good balance between the two types of networks.

Random Choices for Churn Resilient Load Balancing in Peer-to-Peer Networks

Song Fu and Cheng-Zhong Xu
Dept. of Electrical & Computer Engineering
Wayne State University, Detroit, MI
{song, czxu}@eng.wayne.edu

Haiying Shen
Dept. of Computer Science & Engineering
University of Arkansas, Fayetteville, AR
hshen@uark.edu

Peer-to-peer (P2P) networks based on consistent hashing functions have an inherent load uneven distribution problem. Things are even worse in unstructured P2P systems. The objective of load balancing in P2P networks is to balance the workload of the network nodes in proportion to their capacity so as to eliminate traffic bottleneck. It is challenging because of the dynamic nature of overlay networks and time-varying load characteristics. Random choices schemes can balance load effectively while incurring only a small overhead, making such schemes appealing for practical systems. Existing theoretical work analyzing properties of random choices algorithms can not be applied in the highly dynamic and heterogeneous P2P systems. In this paper, we characterize the behaviors of randomized search schemes in the general P2P environment. We extend the supermarket model by investigating the impact of node heterogeneity and churn to the load distribution in P2P networks. We prove that by using d-way random choices schemes, the length of the longest queue in P2P systems with heterogeneous nodal capacity and node churn for $d \geq 2$ is $c \log\log n / \log d + O(1)$ with high probability, where c is a constant.

SESSION 14

Architecture - Power/SMT/ILP

Efficient Resources Assignment Schemes for Clustered Multithreaded Processors

Fernando Latorre, José González and Antonio González
Intel Barcelona Research Center
Intel Labs - UPC
{fernando.latorre, pepe.gonzalez, antonio.gonzalez}@intel.com

New feature sizes provide larger number of transistors per chip that architects could use in order to further exploit instruction level parallelism. However, these technologies bring also new challenges that complicate conventional monolithic processor designs. On the one hand, exploiting instruction level parallelism is leading us to diminishing returns and therefore exploiting other sources of parallelism like thread level parallelism is needed in order to keep raising performance with a reasonable hardware complexity. On the other hand, clustering architectures have been widely studied in order to reduce the inherent complexity of current monolithic processors. This paper studies the synergies and trade-offs between two concepts, clustering and simultaneous multithreading (SMT), in order to understand the reasons why conventional SMT resource assignment schemes are not so effective in clustered processors. These trade-offs are used to propose a novel resource assignment scheme that gets and average speed up of 17.6% versus Icount improving fairness in 24%.

A Software-Hardware Hybrid Steering Mechanism for Clustered Microarchitectures

Qiong Cai, Josep M. Codina, José González and Antonio González
Intel Barcelona Research Centers, Intel-UPC
{qiongx.cai, josep.m.codina, pepe.gonzalez, antonio.gonzalez}@intel.com

Clustered microarchitectures provide a promising paradigm to solve or alleviate the problems of increasing microprocessor complexity and wire delays. High-performance out-of-order processors rely on hardware-only steering mechanisms to achieve balanced workload distribution among clusters. However, the additional steering logic results in a significant increase on complexity, which actually decreases the benefits of the clustered design. In this paper, we address this complexity issue and present a novel software-hardware hybrid steering mechanism for out-of-order processors. The proposed software-hardware cooperative scheme makes use of the concept of virtual clusters. Instructions are distributed to virtual clusters at compile time using static properties of the program such as data dependences. Then, at runtime, virtual clusters are mapped into physical clusters by considering workload information. Experiments using SPEC CPU2000 benchmarks show that our hybrid approach can achieve almost the same performance as a state-of-the-art hardware-only steering scheme, while requiring low hardware complexity. In addition, the proposed mechanism outperforms state-of-the-art software-only steering mechanisms by 5% and 10% on average for 2-cluster and 4-cluster machines, respectively.

The Impact of Out-of-Order Commit in Coarse-Grain, Fine-Grain and Simultaneous Multithreaded Architectures

R. Ubal, J. Sahuquillo, S. Petit, P. López and J. Duato
Dept. of Computer Engineering (DISCA)
Universidad Politécnica de Valencia, Spain
raurte@gap.upv.es

Multithreaded processors in their different organizations (simultaneous, coarse grain and fine grain) have been shown as effective architectures to reduce the issue waste. On the other hand, retiring instructions from the pipeline in an out-of-order fashion helps to unclog the ROB when a long latency instruction reaches its head. This further contributes to maintain a higher utilization of the available issue bandwidth. In this paper, we evaluate the impact of retiring instructions out of order on different multithreaded architectures and different instruction fetch policies, using the recently proposed Validation Buffer microarchitecture as baseline out-of-order commit technique. Experimental results show that, for the same performance, out-of-order commit permits to reduce multithread hardware complexity (e.g., fine grain multithreading with a lower number of supported threads).

Low Power/Area Branch Prediction Using Complementary Branch Predictors

Resit Sendag[1], Joshua J. Yi[2], Peng-fei Chuang[3] and David J. Lilja[3]
[1] Electrical and Computer Engineering
University of Rhode Island
Kingston, Rhode Island
[2] Networking and Multimedia Group
Freescale Semiconductor, Inc.
Austin, The Great State of Texas
[3] Electrical and Computer Engineering
University of Minnesota – Twin Cities
Minneapolis, Minnesota

Although high branch prediction accuracy is necessary for high performance, it typically comes at the cost of larger predictor tables and/or more complex prediction algorithms. Unfortunately, large predictor tables and complex algorithms require more chip area and have higher power consumption, which precludes their use in embedded processors. As an alternative to large, complex branch predictors, in this paper, we investigate adding complementary branch predictors (CBP) to embedded processors to reduce their power consumption and/or improve their branch prediction accuracy. A CBP differs from a conventional branch predictor in that it focuses only on frequently mispredicted branches while letting the conventional branch predictor predict the more predictable ones. Our results show that adding a small 16-entry (28 byte) CBP reduces the branch misprediction rate of static, bimodal, and gshare branch predictors by an average of 51.0%, 42.5%, and 39.8%, respectively, across 38 SPEC 2000 and MiBench benchmarks. Furthermore, a 256-entry CBP improves the energy-efficiency of the branch predictor and processor up to 97.8% and 23.6%, respectively. Finally, in addition to being very energy-efficient, a CBP can also improve the processor performance and, due to its simplicity, can be easily added to the pipeline of any processor.

SESSION 15

Software - Tuning and Performance

Scalable Methods for Monitoring and Detecting Behavioral Equivalence Classes in Scientific Codes

Todd Gamblin, Rob Fowler and Daniel A. Reed
Renaissance Computing Institute
University of North Carolina at Chapel Hill
tgamblin@cs.unc.edu, {rjf, dan_reed}@renci.org

Emerging petascale systems will have many hundreds of thousands of processors, but traditional task-level tracing tools already fail to scale to much smaller systems because the I/O backbones of these systems cannot handle the peak load offered by their cores. Complete event traces of all processes are thus infeasible. To retain the benefits of detailed performance measurement while reducing volume of collected data, we developed AMPL, a general-purpose toolkit that reduces data volume using stratified sampling. We adopt a scalable sampling strategy, since the sample size required to measure a system varies sub-linearly with process count. By grouping, or stratifying, processes that behave similarly, we can further reduce data overhead while also providing insight into an application's behavior. In this paper, we describe the AMPL toolkit and we report our experiences using it on large-scale scientific applications. We show that AMPL can successfully reduce the overhead of tracing scientific applications by an order of magnitude or more, and we show that our tool scales sub-linearly, so the improvement will be more dramatic on petascale machines. Finally, we illustrate the use of AMPL to monitor applications by performance-equivalent strata, and we show that this technique can allow for further reductions in trace data volume and traced execution time.

Understanding Tuning Complexity in Multithreaded and Hybrid Web Servers

Vicenç Beltran, Jordi Torres and Eduard Ayguadé
{vbeltran, jordi.torres, eduard.ayguade}@bsc.es
Barcelona Supercomputing Center (BSC) - Technical University of Catalonia (UPC)
Barcelona, Spain

Adequately setting up a multi-threaded web server is a challenging task because its performance is determined by a combination of configurable web server parameters and unsteady external factors like the workload type, workload intensity and machine resources available. Usually administrators set up web server parameters like the keep-alive timeout and number of worker threads based on their experience and judgment, expecting that this configuration will perform well for the guessed uncontrollable factors. The nontrivial interaction between the configuration parameters of a multi-threaded web server makes it a hard task to properly tune it for a given workload, but the burst nature of the Internet quickly change the uncontrollable factors and make it impossible to obtain an optimal configuration that will always perform well.

In this paper we show the complexity of optimally configuring a multi-threaded web server for different workloads with an exhaustive study of the interactions between the keep-alive timeout and the number of worker threads for a wide range of workloads. We also analyze the Hybrid web server architecture (multi-threaded and event-driven) as a feasible solution to simplify web server tuning and obtain the best performance for a wide range of workloads that can dynamically change in intensity and type. Finally, we compare the performance of the optimally tuned multithreaded web server and the hybrid web server with different workloads to validate our assertions. We conclude from our study that the hybrid architecture clearly outperforms the multi-threaded one, not only in terms of performance, but also in terms of its tuning complexity and its adaptability over different workload types. In fact, from the obtained results, we expect that the hybrid architecture is well suited to simplify the self configuration of complex application servers.

Balancing HPC Applications Through Smart Allocation of Resources in MT Processors

Carlos Boneti[2] Roberto Gioiosa[1], Francisco J. Cazorla[1], Julita Corbalan[1,2],
Jesus Labarta[1,2] and Mateo Valero[1,2]
[1]Barcelona Supercomputing Center, Spain
[2]Universitat Politecnica de Catalunya, Spain
{roberto.gioiosa, francisco.cazorla}@bsc.es, {cboneti, julita, jesus, mateo}@ac.upc.es

Many studies have shown that load imbalancing causes significant performance degradation in High Performance Computing (HPC) applications. Nowadays, Multi-Threaded (MT1) processors are widely used in HPC for their good performance/energy consumption and performance/cost ratios achieved sharing internal resources, like the instruction window or the physical register. Some of these processors provide the software hardware mechanisms for controlling the allocation of processor's internal resources. In this paper, we show, for the first time, that by appropriately using these mechanisms, we are able to control the tasks speed, reducing the imbalance in parallel applications transparently to the user and, hence, reducing the total execution time. Our results show that our proposal leads to a performance improvement up to 18% for one of the NAS benchmark. For a real HPC application (much more dynamic than the benchmark) the performance improvement is 8.1%. Our results also show that, if resource allocation is not used properly, the imbalance of applications is worsened causing performance loss.

Result Reuse in Design Space Exploration: A Study in System Support for Interactive Parallel Computing

Siu-Man Yau[1], Kostadin Damevski[2],
Vijay Karamcheti[1], Steven G. Parker[2] and Denis Zorin[1]
[1] Courant Institute of Mathematical Sciences, New York University
[2] Department of Computer Science, University of Utah

This paper presents a system supporting reuse of simulation results in multi-experiment computational studies involving independent simulations and explores the benefits of such reuse. Using a SCIRun-based defibrillator device simulation code (DefibSim) and the SimX system for computational studies, this paper demonstrates how aggressive reuse between and within computational studies can enable interactive rates for such studies on a moderate-sized 128-node processor cluster; a brute-force approach to the problem would require two thousand nodes or more on a massively parallel machine for similar performance. Key to realizing these performance improvements is exploiting optimization opportunities that present themselves at the level of the overall workflow of the study as opposed to focusing on individual simulations. Such global optimization approaches are likely to become increasingly important with the shift towards interactive and universal parallel computing.

SESSION 16

Algorithms - Theory

Approximating Max-min Linear Programs with Local Algorithms

Patrik Floréen, Petteri Kaski, Topi Musto and Jukka Suomela
Helsinki Institute for Information Technology HIIT, Department of Computer Science,
University of Helsinki, P.O. Box 68, FI-00014 University of Helsinki, Finland
E-mail: {firstname.lastname}@cs.helsinki.fi

A local algorithm is a distributed algorithm where each node must operate solely based on the information that was available at system startup within a constant-size neighbourhood of the node. We study the applicability of local algorithms to max-min LPs where the objective is to maximise $\min_k \sum_v c_{kv} x_v$ subject to $\sum_v a_{iv} x_v \leq 1$ for each i and $x_v \geq 0$ for each v. Here $c_{kv} \geq 0$, $a_{iv} \geq 0$, and the support sets $V_i = \{v : a_{iv} > 0\}$, $V_k = \{v : c_{kv} > 0\}$, $I_v = \{i : a_{iv} > 0\}$ and $K_v = \{k : c_{kv} > 0\}$ have bounded size. In the distributed setting, each agent v is responsible for choosing the value of x_v, and the communication network is a hypergraph $\mathcal{H}$ where the sets V_k and V_i constitute the hyperedges. We present inapproximability results for a wide range of structural assumptions; for example, even if $|V_i|$ and $|V_k|$ are bounded by some constants larger than 2, there is no local approximation scheme. To contrast the negative results, we present a local approximation algorithm which achieves good approximation ratios if we can bound the relative growth of the vertex neighbourhoods in $\mathcal{H}$.

Parallel Mining of Closed Quasi-Cliques

Yuzhou Zhang, Jianyong Wang, Zhiping Zeng and Lizhu Zhou
Tsinghua University
Beijing, 100084, China
zhangyz04@mails.tsinghua.edu.cn
{jianyong,dcszlz}@tsinghua.edu.cn
clipse.zeng@gmail.com

Graph structure can model the relationships among a set of objects. Mining quasi-clique patterns from large dense graph data makes sense with respect to both statistic and applications. The applications of frequent quasi-cliques include stock price correlation discovery, gene function prediction and protein molecular analysis. Although the graph mining community has devised many skills to accelerate the discovery process, mining time is always unacceptable, especially on large dense graph data with low support threshold. Therefore, parallel algorithms are desirable on mining quasi-clique patterns. Message passing is one of the most widely used parallel framework. In this paper, we parallelize the state-of-the-art closed quasi-clique mining algorithm called Cocain using message passing. The parallelized version of Cocain can achieve 30+ fold speedup on 32 processors in a cluster of SMPs. The techniques proposed in this work can be applied to parallelize other pattern-growth based frequent pattern mining algorithms.

Optimal Replication Transition Strategy in Distributed Hierarchical Systems

Chun-Chen Hsu[1,2], Chien-Min Wang[2], and Pangfeng Liu[1,3]
[1] Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
d95006,pangfeng@csie.ntu.edu.tw
[2]Institute of Information Science, Academia Sinica, Nankang, Taiwan
{tk,cmwang}@iis.sinica.edu.tw
[3]Graduated Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan

We study the replication transition problem in distributed hierarchical systems. Most distributed systems replicate data to increase data access efficiency. A replication strategy dictates where the replicas are stored in respond to the data access pattern, therefore a good strategy can effectively improve data access efficiency. However, the access pattern in a distributed system is constantly changing. As a result, a good replication strategy must evolve accordingly. The replication transition problem is to seek an efficient transition from one replication strategy to another, in order to cope with the dynamic access pattern. This paper focuses on solving the replication transition problem on tree topology, which is one of the most important models in Data Grid systems and web proxy systems from the literature. To the best of our knowledge, our work is the first that proposes an optimal algorithm for the replication transition problem on tree topology. The algorithm has a time complexity of $O(n \log \Delta log(n\Lambda))$, where n is the number of sites, Δ is the maximum degree in the tree and Λ is the largest communication delay in the network.

DVS Based Energy Minimization Algorithm for Parallel Machines

Jaeyeon Kang and Sanjay Ranka
Department of Computer and Information Science and Engineering, University of Florida
jkang, ranka@cise.ufl.edu

Energy consumption is a critical issue in parallel and distributed embedded systems. We present a novel algorithm for energy efficient scheduling of Directed Acyclic Graph (DAG) based applications on Dynamic Voltage Scaling (DVS) enabled systems. Experimental results show that our algorithm provides near optimal solutions for energy minimization with considerably smaller computational time and memory requirements compared to an existing algorithm that provides near optimal solutions.

SESSION 17

Applications - P2P Reliability and Trust

An Effective Pointer Replication Algorithm in P2P Networks

Jian Zhou, Laxmi N. Bhuyan and Anirban Banerjee
University of California, Riverside
{jianz, bhuyan, anirban}@cs.ucr.edu

Peer-to-Peer (P2P) networks have proven to be an efficient and successful mechanism for file sharing over the Internet. However, current P2P protocols have long worst case query latencies which prevents them from being employed for real time applications. Popularity of objects in these networks can change rapidly and augurs the need for a rapid and lightweight content replication strategy to reduce search and data-access latencies. In this paper, we propose an On-line Pointer Replication (OPR) algorithm in structured P2P networks which yields a significantly low worst case query latency. Also, the degree of replication achieved by OPR is dynamically adaptable to the instantaneous query arrival rate and churn characteristics of the system in order to reduce total control traffic. We evaluate and compare different replica placement strategies on the PlanetLab network as well as with simulations. Experimental results show that OPR outperforms the existing replica placement algorithms by at least 30%in average latency and around 40% in terms of maximum query latency.

Towards Reliable and Efficient Data Dissemination in Heterogeneous Peer-to-Peer Systems

Zhenyu Li
Institute of Computing Technology
Chinese Academy of Sciences (CAS)
Graduate School of CAS
Beijing, China
zyli@ict.ac.cn

Gaogang Xie and Zhongcheng Li
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
{xie, zcli}@ict.ac.cn

More and more emerging P2P applications require support for multi-source data dissemination. However, existing schemes on data dissemination are either only suitable for single source systems or inefficient in terms of delivery delay or message redundancy. This paper presents REM, a reliable and efficient multi-source data dissemination scheme. REM organizes member nodes in a ring-based structure which is reliable, locality-aware and heterogeneity-aware. An undirected delivery tree is built on top of the overlay and used by all nodes. Data messages are first flooded on the overlay and then delivered on the tree. Thus, REM combines the advantages of flooding-based scheme and tree-based scheme. We analyze the reliability, average hop complexity and redundancy rate theoretically and evaluate the performance by simulations. Specially, in large scale systems, compared with a randomized scheme, REM reduces the average delivery delay by about 25% while the redundancy rate is reduced from about 16% to 1%.

Distributed Asymmetric Verification in Computational Grids

Michael Kuhn, Stefan Schmid and Roger Wattenhofer
Computer Engineering and
Networks Laboratory (TIK)
ETH Zurich
CH-8092 Zurich, Switzerland
{kuhnmi,schmiste,wattenhofer}@tik.ee.ethz.ch

Lucrative incentives in grid computing do not only attract honest participants, but also cheaters. To prevent selfish behavior, verification mechanisms are required. Today's solutions mostly base on redundancy and inherently exhibit a considerable overhead. Often, however, the verification of a result takes much less time than its computation. In this paper we propose a distributed checking scheme that exploits this asymmetry. Our mechanism detects wrong results and excludes cheaters in a distributed manner and hence disburdens the central of the grid server. We show how the verification scheme is used in an application which aims at breaking the discrete logarithm problem by a parallel implementation of the Pollard-ρ algorithm. Our implementation extends the BOINC server software and is robust to various rational attacks even in the presence of colluders.

A Unified Model of Pollution in P2P Networks

Cong Shi, Dingyi Han, Xinyao Hu and Yong Yu
Apex Data and Knowledge Management Lab,
Department of Computer Science and Engineering,
Shanghai Jiaotong University, Shanghai, 200240, China.
{cshi, handy, montaguehu, yyu}@apex.sjtu.edu.cn

Nowadays many popular Peer-to-Peer (P2P) systems suffer from the simultaneous attacks of various pollution, including file-targeted attack and index-targeted attack. However, to our knowledge, there is no model that takes both of them into consideration. In fact, the two attacks impact the effect of each other. It makes the models considering either kind of pollution only fail to accurately illustrate the actual pollution. In this paper, we develop a unified model to remedy the defect. Through the analysis from the perspective of user behavior, the two attacks are integrated into the unified model as two factors impacting users' choice of the files to download. The modeled file proliferation processes are consistent to those measured in real P2P systems. Moreover, the co-effect of the two attacks is also analyzed. The extremum point of co-effect is found, which corresponds to the most efficient attack of pollution.

Further analysis of the model's accuracy requires the quantitative comparison between the modeled effects of pollution and the measured ones. Nonetheless, no such metric has ever been proposed, which also causes a lot of problems in evaluating the effect of pollution and antipollution techniques. To fix the deficiency, we propose several metrics to assess the effect of pollution, including abort ratio, average download time of unpolluted files, etc. These metrics estimate the effect of pollution from different aspects. They are applied to the analysis of pollution emulated by our unified model. The co-effect of pollution is captured by these metrics. Furthermore, the difference between our model and the previously developed ones is also reflected by them.

SESSION 18

Architecture - Networks

Fault Tolerance with Shortest Paths in Regular and Irregular Networks

Frank Olaf Sem-Jacobsen[1,2] and Olav Lysne[2]
frankose@ifi.uio.no, olavly@simula.no
[1]Department of Informatics [2]Networks and Distributed Systems
University of Oslo Simula Research Laboratory
Oslo, Norway Lysaker, Norway

Fault tolerance has become an important part of cur- rent supercomputers. Local dynamic fault tolerance is the most expedient way of tolerating faults by preconfiguring the network with multiple paths from every node/switch to every destination. In this paper we present a local shortest path dynamic fault-tolerance mechanism inspired by a solution developed for the Internet, that can be applied to any shortest path routing algorithm such as Dimension Ordered Routing, Fat Tree Routing, Layered Shortest Path, etc., and provide a solution for achieving deadlock freedom in the presence of faults. Simulation results show that 1) for fat trees this yields the to this day highest throughput and lowest requirements on virtual layers for dynamic one-fault tolerance, 2) we require in general few layers to achieve dead- lock freedom, and 3) for irregular topologies it gives at most a 10 times performance increase compared to FRoots.

Performance Adaptive UDP for High-Speed Bulk Data Transfer Over Dedicated Links

Ben Eckart and Xubin He
Tennessee Technological University
Electrical and Computer Engineering
Cookeville, TN 38505, USA
{bdeckart21, hexb}@tntech.edu

Qishi Wu
University of Memphis
Department of Computer Science
Memphis, TN 38152, USA
qishiwu@memphis.edu

New types of networks are emerging for the purpose of transmitting large amounts of scientific data among research institutions quickly and reliably. These exotic networks are characterized by being high-bandwidth, high-latency, and free from congestion. In this environment, TCP ceases to be an appropriate protocol for reliable bulk data transfer because it fails to saturate link throughput. Of the new protocols designed to take advantage of these networks, a subclass has emerged using UDP for data transfer and TCP for control. These high-speed variants of reliable UDP, however, tend to underperform on all but high-end systems due to constraints of the CPU, network, and hard disk. It is therefore necessary to build a high-speed protocol adaptive to the performance of each system. This paper develops such a protocol, Performance Adaptive UDP (henceforth PA-UDP), which aims to dynamically and autonomously maximize performance under different systems. A mathematical model and related algorithms are proposed to describe the theoretical basis behind effective buffer and CPU management. Based on this model, we implemented a prototype under Linux and the experimental results demonstrate that PA-UDP outperforms an existing high-speed protocol on commodity hardware in terms of throughput and packet loss. PA-UDP is efficient not only for high-speed research networks but also for reliable high-performance bulk data transfer over dedicated local area networks where congestion and fairness are typically not a concern.

Epoch-Based Reconfiguration: Fast, Simple, and Effective Dynamic Network Reconfiguration

J. M. Montañana, J. Flich and J. Duato
Dept. of Computer Engineering (DISCA)
Universidad Politécnica de Valencia (UPV)
Camino de Vera, 14, 46021–Valencia, Spain
E-mail: jmontana@gap.upv.es

Dynamic network reconfiguration is defined as the process of changing from one routing function to another while the network remains up and running. The main challenge is to avoid deadlocks and reduce packet dropping rate while keeping network service.

Current approaches either require the existence of extra network resources like e.g. virtual channels, their complexity is so high that their practical applicability is limited, or they affect to the performance of the network during the reconfiguration process.

In this paper we present EBR, a simple and fast method for dynamic network reconfiguration. EBR guarantees a fast and deadlock-free reconfiguration, but instead of avoiding deadlocks our mechanism is based on regressive deadlock recoveries. Thus, EBR allows cycles to be formed, and in the situation of a deadlock some packets may be dropped. However, as demonstrated, no packets need to be dropped in the working zone of the system. Also, the mechanism works in an asynchronous manner, does not require additional resources and works on any topology. In order to minimize the number of dropped packets, EBR uses an epoch marking system that guarantees that only packets potentially leading to a deadlock will be removed.

Evaluation results show that EBR works efficiently in different topologies and with different routing algorithms. When compared with current proposals, EBR always gets the best numbers in all the analyzed parameters (dropped packets, latency, throughput, reconfiguration time and resources required), thus achieving the good properties of all mechanisms.

Achieving 100% Throughput in Input-Buffered WDM Optical Packet Interconnects

Lin Liu and Yuanyuan Yang
Dept. of Electrical & Computer Engineering,
State University of New York, Stony Brook, NY 11794, USA

All-optical wavelength-division-multiplexing (WDM) interconnects are a promising candidate for future ultra high-speed interconnections due to the huge capacity of optics. Packet scheduling algorithms that can guarantee 100% throughput under various types of traffic enable an interconnect to achieve its full capacity. However, although such algorithms have been proposed for electronic interconnects, they cannot be directly applied to WDM optical interconnects due to the following reasons. First, almost all of these algorithms depend on the Virtual Output Queue (VOQ) technique which is currently difficult to implement in WDM optical interconnects due to lack of optical RAM; Second, a packet arriving at the input of a WDM interconnect now have more than one output wavelength channels to choose from due to wavelength conversion capability. The former motivates us to search for a new input buffering fabric that is more practical under the current optical technology and can achieve satisfactory performance, and the latter indicates that a new definition of "admissible traffic" may be needed for WDM optical interconnects. In this paper, we first introduce a new fiber-delay- line (FDL) based input buffering fabric that is able to provide flexible buffering delay in WDM optical interconnects. We then give a new definition of "admissible traffic" for a WDM optical interconnect, and propose the Most- Packet Wavelength-Fiber Pair First (MPWFPF) scheduling algorithm for WDM interconnects using such buffering fabric. We theoretically prove that with the new buffering fabric, MPWFPF can deliver 100% throughput for input-buffered WDM interconnects with no speedup required. Finally, we further propose a faster scheduling algorithm, WDM-iSLIP, that can efficiently determine an approximate optimal schedule with much lower time complexity. Extensive simulations have been conducted to verify the theoretical results, and test the performance of the proposed scheduling algorithms in input-buffered WDM interconnects with the new buffering fabric.

SESSION 19

Software - Language Features and Implementation

A Transparent Non-Invasive File Data Model for Algorithmic Skeletons

Denis Caromel and Mario Leyton
INRIA Sophia-Antipolis, Université de Nice Sophia-Antipolis, CNRS - I3S
2004, Route des Lucioles, BP 93, F-06902 Sophia-Antipolis Cedex, France
First.Last@sophia.inria.fr

A file data model for algorithmic skeletons is proposed, focusing on transparency and efficiency. Algorithmic skeletons correspond to a high-level programming model that takes advantage of nestable programming patterns to hide the complexity of parallel/distributed applications.

Transparency is achieved using a workspace factory abstraction and the proxy pattern to intercept calls on File type objects. Thus allowing programmers to continue using their accustomed programming libraries, without having the burden of explicitly introducing non-functional code to deal with the distribution aspects of their data.

A hybrid file fetching strategy is proposed (instead of lazy or eager), that takes advantage of annotated functions and pipelined multithreaded interpreters to transfer files in-advance or on-demand. Experimentally, using a BLAST skeleton application, it is shown that the hybrid strategy provides a good tradeoff between bandwidth usage and CPU idle time.

Parallelizing Irregular C Codes Assisted by Interprocedural Shape Analysis

R. Asenjo, R. Castillo, F. Corbera, A. Navarro, A. Tineo and E.L. Zapata
Dept. of Computer Architecture, University of Málaga,
Complejo Tecnologico, Campus de Teatinos, E-29071. Málaga, Spain.
{asenjo,rosa,corbera,angeles,tineo,ezapata}@ac.uma.es

In the new multicore architecture arena, the problem of improving the performance of a code is more in the software side than in the hardware one. However, optimizing irregular dynamic data structure based codes for such architectures is not easy, either by hand or compiler assisted. Regarding this last approach, shape analysis is a static technique that achieves abstraction of dynamic memory and can help to disambiguate, quite accurately, memory references in programs that create and traverse recursive data structures. This kind of analysis has promising applicability for accurate data dependence tests in loops or recursive functions that traverse dynamic data structures. However, support for interprocedural programs in shape analysis is still a challenge, especially in the presence of recursive functions. In this work we present a novel fully context-sensitive interprocedural shape analysis algorithm that supports recursion and can be used to uncover parallelism. Our approach is based on three key ideas: i) intraprocedural support based on “Coexistent Links Sets” to precisely describe the memory configurations during the abstract interpretation of the C code; ii) interprocedural support based on “Recursive Flow Links” to trace the state of pointers in previous calls; and iii) annotations of the read/written heap locations during the program analysis. We present preliminary experiments that reveal that our technique compares favorably with related work, and obtains precise memory abstractions in a variety of recursive programs that create and manipulate dynamic data structures. We have also implemented a data dependence test over our interprocedural shape analysis. With this test we have obtained promising results, automatically detecting parallelism in three C codes, which have been successfully parallelized.

Efficient Automated Marshaling of C++ Data Structures for MPI Applications

Wesley Tansey and Eli Tilevich
Center for High-End Computing Systems
Department of Computer Science
Virginia Tech
Email: {tansey,tilevich}@cs.vt.edu

We present an automated approach for marshaling C++ data structures in High Performance Computing (HPC) applications. Our approach utilizes a graphical editor through which the user can express a subset of an object's state to be marshaled and sent across a network. Our tool, MPI Serializer, then automatically generates efficient marshaling and unmarshaling code for use with the Message Passing Interface (MPI), the predominant communication middleware for HPC systems.

Our approach provides a more comprehensive level of support for C++ language features than the existing state of the art, and does so in full compliance with the C++ Language Standard. Specifically, we can marshal effectively and efficiently non-trivial language constructs such as polymorphic pointers, dynamically allocated arrays, non-public member fields, inherited members, and STL container classes. Additionally, our marshaling approach is also applicable to third party libraries, as it does not require any modifications to the existing C++ source code.

We validate our approach through two case studies of applying our tool to automatically generate the marshaling functionality of two realistic HPC applications. The case studies demonstrate that the automatically generated code matches the performance of typical hand-written implementations and surpasses current state-of-the-art C++ marshaling libraries, in some cases by more than an order of magnitude. The results of our case studies indicate that our approach can be beneficial for both the initial construction of HPC applications as well as for the refactoring of sequential applications for parallel execution.

A Deterministic Multi-Way Rendezvous Library for Haskell

Nalini Vasudevan	Satnam Singh	Stephen A. Edwards
Department of Computer Science	Microsoft Research Cambridge	Department of Computer Science
Columbia University	Cambridge CB3 0FB	Columbia University
New York, USA	United Kingdom	New York, USA
naliniv@cs.columbia.edu	satnams@microsoft.com	sedwards@cs.columbia.edu

The advent of multicore processors requires mainstream concurrent programming languages with high level concurrency constructs and effective debugging techniques. Unfortunately, many concurrent programming languages are non-deterministic and allow data races.

We present a deterministic concurrent communication library for an existing multi-threaded language. We implemented the SHIM communication model in the Haskell functional language, which supports asynchronous communication and transactional memory. The SHIM model uses multi-way rendezvous to guarantee determinism. We describe two implementations of the model in Haskell, demonstrating the ease of writing such a library.

We illustrate our library with examples and experimentally compare two implementations. We also compare our new model with equivalent sequential programs and parallel versions using Haskell's existing concurrency mechanisms.

SESSION 20

Algorithms - Fault Tolerance

An Optimal Checkpoint/Restart Model for a Large Scale High Performance Computing System

Yudan Liu[1], Raja Nassar[1], Chokchai (Box) Leangsuksun[1], Nichamon Naksinehaboon[1],
Mihaela Paun[1] and Stephen L. Scott[2]
[1]College of Engineering & Science
Louisiana Tech University, Ruston, LA 71270, USA
[2]Computer Science and Mathematics Division, Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
{yli010, nassar, box, nna003, mpaun}@latech.edu, scottsl@ornl.gov

The increase in the physical size of High Performance Computing (HPC) platform makes system reliability more challenging. In order to minimize the performance loss (rollback and checkpoint overheads) due to unexpected failures or unnecessary overhead of fault tolerant mechanisms, we present a reliability-aware method for an optimal checkpoint/restart strategy. Our scheme aims at addressing fault tolerance challenge, especially in a large-scale HPC system, by providing optimal checkpoint placement techniques that are derived from the actual system reliability. Unlike existing checkpoint models, which can only handle Poisson failure and a constant checkpoint interval, our model can deal with a varying checkpoint interval and with different failure distributions. In addition, the approach considers optimality for both checkpoint overhead and rollback time. Our validation results suggest a significant improvement over existing techniques.

Picking up the Pieces: Self-Healing in Reconfigurable Networks

Jared Saia and Amitabh Trehan
Department of Computer Science,
University of New Mexico, Albuquerque, NM 87131-1386
{saia, amitabh}@cs.unm.edu

We consider the problem of self-healing in networks that are reconfigurable in the sense that they can change their topology during an attack. Our goal is to maintain connectivity in these networks, even in the presence of repeated adversarial node deletion, by carefully adding edges after each attack. We present a new algorithm, DASH, that provably ensures that: 1) the network stays connected even if an adversary deletes up to all nodes in the network; and 2) no node ever increases its degree by more than 2 log n, where n is the number of nodes initially in the network. DASH is fully distributed; adds new edges only among neighbors of deleted nodes; and has average latency and bandwidth costs that are at most logarithmic in n. DASH has these properties irrespective of the topology of the initial network, and is thus orthogonal and complementary to traditional topology-based approaches to defending against attack.

We also prove lower-bounds showing that DASH is asymptotically optimal in terms of minimizing maximum degree increase over multiple attacks. Finally, we present empirical results on power-law graphs that show that DASH performs well in practice, and that it significantly outperforms naive algorithms in reducing maximum degree increase. We also present empirical results regarding stretch and discuss a heuristic to keep stretch low.

Wait-free Programming for General Purpose Computations on Graphics Processors

Phuong Hoai Ha
University of Tromsø
Department of Computer Science
Faculty of Science, N-9037 TromsøNorway
phuong@cs.uit.no

Philippas Tsigas
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg, Sweden
tsigas@cs.chalmers.se

Otto J. Anshus
University of Tromsø
Department of Computer Science
Faculty of Science, N-9037 TromsøNorway
otto@cs.uit.no

The fact that graphics processors (GPUs) are today's most powerful computational hardware for the dollar has motivated researchers to utilize the ubiquitous and powerful GPUs for general-purpose computing. Recent GPUs feature the single-program multiple-data (SPMD) multicore architecture instead of the single-instruction multiple-data (SIMD). However, unlike CPUs, GPUs devote their transistors mainly to data processing rather than data caching and flow control, and consequently most of the powerful GPUs with many cores do not support any synchronization mechanisms between their cores. This prevents GPUs from being deployed more widely for general-purpose computing.

This paper aims at bridging the gap between the lack of synchronization mechanisms in recent GPU architectures and the need of synchronization mechanisms in parallel applications. Based on the intrinsic features of recent GPU architectures, we construct strong synchronization objects like wait-free and t-resilient read-modify-write objects for a general model of recent GPU architectures without strong hardware synchronization primitives like test-and-set and compare-and-swap. Accesses to the wait-free objects have time complexity O(N), whether N is the number of processes. Our result demonstrates that it is possible to construct wait-free synchronization mechanisms for GPUs without the need of strong synchronization primitives in hardware and that wait-free programming is possible for GPUs.

Self-Stabilizing Algorithms for Sorting and Heapification

Doina Bein
Department of Computer Science
University of Texas at Dallas, TX 75083
siona@utdallas.edu

Ajoy K. Datta, Lawrence L. Larmore
School of Computer Science
University of Nevada, Las Vegas, NV 89154
{datta,larmore}@cs.unlv.edu

We present two space and time efficient asynchronous distributed self-stabilizing algorithms. The first sorts an oriented chain network and the second heapifies a rooted tree network. The time complexity of both solutions is linear — in terms of the nodes (for the chain) and height (for the tree). The chain sorting algorithm uses $O(m)$ bits per process where m represents the number of bits required to store any value in the network. The heapify algorithm needs $O(m \cdot D)$ bits per process where D is the degree of the tree.

SESSION 21

Applications - Sensors

Sweep Coverage with Mobile Sensors

Weifang Cheng[1], Mo Li[2], Kebin Liu[3], Yunhao Liu[2], Xiangyang Li[4] and Xiangke Liao[1] 973 WSN Joint Lab
[1]National University of Defense Technology, China
[2]Hong Kong University of Science and Technology, Hong Kong
[3]Shanghai Jiao Tong University, China
[4]Illinois Institute of Technology, Chicago, USA
Email: wfangch@nudt.edu.cn, limo@cse.ust.hk, liu@cse.ust.hk,
captainlkb2003@sjtu.edu.cn, xli@cs.iit.edu, xkliao@nudt.edu.cn

Many efforts have been made for addressing coverage problems in sensor networks. They fall into two categories, full coverage and barrier coverage, featured as static coverage. In this work, we study a new coverage scenario, sweep coverage, which differs with the previous static coverage. In sweep coverage, we only need to monitor certain points of interest (POIs) periodically so the coverage at each POI is time-variant, and thus we are able to utilize a small number of mobile sensors to achieve sweep coverage among a much larger number of POIs. We investigate the definitions and model for sweep coverage. Given a set of POIs and their sweep period requirements, we prove that determining the minimum number of required sensors (min-sensor sweep-coverage problem) is NP-hard, and it cannot be approximated within a factor of 2. We propose a centralized algorithm with constant approximation ratio $2 + \epsilon$ for the simplified problem where all sweep periods are identical. We further characterize the non-locality of the problem and design a distributed sweep algorithm, DSWEEP, cooperating sensors to provide required sweep requirements with the best effort. We conduct extensive simulations to study the performance of the proposed algorithms. Our simulations show that DSWEEP outperforms the randomized scheme in both effectiveness and efficiency.

SenCast: Scalable Multicast in Wireless Sensor Networks

Shaoliang Peng[1], Shanshan Li[1], Lei Chen[2], Nong Xiao[1] and Yuxing Peng[1]
[1]Department of Computer Science, National University of Defense Technology
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
E-mail: {pengshaoliang, shanshanli}@nudt.edu.cn, leichen@cse.ust.hk, {nxiao, pyx}@nudt.edu.cn

Multicast is essential for wireless sensor network (WSN) applications. Existing multicast protocols in WSNs are often designed in a P2P pattern, assuming small number of destination nodes and frequent changes on network topologies. In order to truly adopt multicast in WSNs, we propose a base-station modelbased multicast, SenCast, to meet the general requirements of applications. SenCast is scalable and energy-efficient for large group communications in WSNs. Theoretical analysis shows that SenCast is able to approximate the Minimum Nonleaf Nodes (MNN) problem to a ratio of $ln|R|$ (R is the set of all destinations), best known lowest bound. We evaluate our design through comprehensive simulations. Experimental results demonstrate that SenCast outperforms previous multicast protocols including the most recent work uCast.

Continuous Answering Holistic Queries Over Sensor Networks

Kebin Liu[1], Lei Chen[2], Minglu Li[1] and Yunhao Liu[2]
[1]Shanghai Jiao Tong University [2]Hong Kong University of Science and Technology
kebin@cse.ust.hk, leichen@cs.ust.hk, li-ml@cs.sjtu.edu.cn, liu@cs.ust.hk

Wireless sensor networks (WSNs) are widely used for various monitoring applications. Users issue queries to sensors and collect sensing data. Due to the low quality sensing devices or random link failures, sensor data are often noisy. In order to increase the reliability of the query results, continuous queries are often employed. In this work we focus on continuous holistic queries like Median. Existing approaches are mainly designed for non-holistic queries like Average. However, it is not trivial to answer holistic ones due to their non-decomposable property. We propose two schemes for answering queries under different data changing conditions. While sensor data changes slowly, based on the data correlation between different rounds, we propose one algorithm for getting the exact answers. When the data changing speed is high, we propose another approach to derive the approximate results. We evaluate both designs through extensive simulations. The results demonstrate that our approach significantly reduces the traffic cost compared with previous works while maintaining the same accuracy.

SESSION 22

Applications - Web Applications

CoSL: A Coordinated Statistical Learning Approach to Measuring the Capacity of Multi-tier Websites

Jia Rao and Cheng-Zhong Xu
Department of Electrical & Computer Engineering
Wayne State University, Detroit, Michigan 48202
{jrao, czxu}@wayne.edu

Website capacity determination is crucial to measurement-based access control, because it determines when to turn away excessive client requests to guarantee consistent service quality under overloaded conditions. Conventional capacity measurement approaches based on high-level performance metrics like response time and throughput may result in either resource over-provisioning or lack of responsiveness. It is because a website may have different capacities in terms of the maximum concurrent level when the characteristic of workload changes. Moreover, bottleneck in a multi-tier website may shift among tiers as client access pattern changes. In this paper, we present an online robust measurement approach based on statistical machine learning techniques. It uses a Bayesian network to correlate low level instrumentation data like system and user cpu time, available memory size, and I/O status that are collected at run-time to high level system states in each tier. A decision tree is induced over a group of coordinated Bayesian models in different tiers to identify the bottleneck dynamically when the system is overloaded. Experimental results demonstrate its accuracy and robustness in different traffic loads.

Simultaneous Transducers for Data-Parallel XML Parsing

Yinfei Pan, Ying Zhang and Kenneth Chiu
Department of Computer Science
State University of New York at Binghamton
Binghamton, NY 13902
ypan3@binghamton.edu, yzhang8@binghamton.edu, kchiu@cs.binghamton.edu

Though XML has gained significant acceptance in a number of application domains, XML parsing can still be a vexing performance bottleneck. With the growing prevalence of multicore CPUs, parallel XML parsing could be one option for addressing this bottleneck. Achieving data parallelism by dividing the XML document into chunks and then independently processing all chunks in parallel is difficult, however, because the state of an XML parser at the first character of a chunk depends potentially on the characters in all preceding chunks. In previous work, we have used a sequential preparser implementing a preparsing pass to determine the document structure, followed by a parallel full parse. The preparsing is sequential, however, and thus limits speedup. In this work, we parallelize the preparsing pass itself by using a simultaneous finite transducer (SFT), which implicitly maintains multiple preparser results. Each result corresponds to starting the preparser in a different state at the beginning of the chunk. This addresses the challenge of determining the correct initial state at beginning of a chunk by simply considering all possible initial states simultaneously. Since the SFT is finite, the simultaneity can be implemented efficiently simply by enumerating the states, which limits the overhead. To demonstrate effectiveness, we use an SFT to build a parallel XML parsing implementation on an unmodified version of libxml2, and obtained good scalability on both a 30 CPU Sun E6500 machine running Solaris and a Linux machine with two Intel Xeon L5320 CPUs for a total of 8 physical cores.

Computational Monitoring and Steering Using Network-Optimized Visualization and AjaxWeb Server

Mengxia Zhu
Dept of Computer Science
Southern Illinois University
Carbondale, IL 62901
mzhu@cs.siu.edu

Qishi Wu
Dept of Computer Science
University of Memphis
Memphis, TN 38152
qishiwu@memphis.edu

Nageswara S.V. Rao
Computer Science & Math Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830
raons@ornl.gov

We describe a system for computational monitoring and steering of an on-going computation or visualization on a remote host such as workstation or supercomputer. Unlike the conventional "launch-and-leave" batch computations, this system enables: (i) continuous monitoring of variables of an on-going remote computation using visualization tools, and (ii) interactive specification of chosen computational parameters to steer the computation. The visualization and control streams are supported over widearea networks using transport protocols based on stochastic approximation methods to provide stable throughput. Using performance models for transport channels and visualization modules, we develop a visualization pipeline configuration solution that minimizes end-to-end delay over widearea connections. The user interface utilizes Asynchronous JavaScript and XML (Ajax) technologies to provide an interactive environment that can be accessed by multiple remote users using web browsers. We present experimental results on a geographically distributed deployment to illustrate the effectiveness of the proposed system.

Scalable Data Dissemination Using Hybrid Methods

Wenhui Zhang and Vincenzo Liberatore
Division of Computer Science, EECS
Case Western Reserve University, Cleveland, Ohio 44106
E-mail: {wxz24, vxl11}@case.edu
Jonathan Beaver, Panos K. Chrysanthis and Kirk Pruhs
Dept. of Computer Science
University of Pittsburgh, Pittsburgh, PA 15260
E-mail: {beaver, panos, kirk}@cs.pitt.edu

Web server scalability can be greatly enhanced via hybrid data dissemination methods that use both unicast and multicast. Hybrid data dissemination is particularly promising due to the development of effective end-to-end multicast methods and tools. Hybrid data dissemination critically relies on document selection which determines the data transfer method that is most appropriate for each data item. In this paper, we study document selection with a special focus on actual end-point implementations and Internet network conditions. We individuate special challenges such as scalable and robust popularity estimation, appropriate classification of hot and cold documents, and unpopular large documents. We propose solutions to these problems, integrate them in MBDD (middleware support multicast-based data dissemination) and evaluate them on PlanetLab with collected traces. Results show that the multicast server can effectively adapt to dynamic environments and is substantially more scalable than traditional Web servers. Our work is a significant contribution to building practical hybrid data dissemination services.

SESSION 23

Software - Resource Management and Scheduling

An Adaptive Parallel Pipeline Pattern for Grids

Horacio González-Vélez and Murray Cole
Institute for Computing Systems
Architecture, School of Informatics,
University of Edinburgh, UK
e-mail: h.gv@ed.ac.uk, mic@inf.ed.ac.uk

This paper introduces an adaptive parallel pipeline pattern which follows the GRASP (Grid-Adaptive Structured Parallelism) methodology. GRASP is a generic methodology to incorporate structural information at compile time into a parallel program that enables it to adapt automatically to dynamic variations in resource performance. GRASP instruments the pipeline with a series of pragmatic rules, which depend on particular performance thresholds based on the computation/communication patterns of the program and the availability of resources in the grid. Our parallel pipeline pattern is implemented as a parameterisable C/MPI API using a variable-size input data vector and a stage function array. We have evaluated its efficiency using a numerical benchmark stage function in a non-dedicated computational grid environment.

Data Throttling for Data-Intensive Workflows

Sang-Min Park and Marty Humphrey
Department of Computer Science, University of Virginia, Charlottesville, VA 22904
{ sp2kn, humphrey }@cs.virginia.edu

Existing workflow systems attempt to achieve high performance by intelligently scheduling tasks on resources, sometimes even attempting to move the largest data files on the highest-capacity links. However, such approaches are inherently limited, in that there is only minimal control available regarding the arrival time and rate of data transfer between nodes, resulting in unbalanced workflows in which one task is idle while waiting for data to arrive. This paper describes a data throttling framework that can be exploited by workflow systems to uniquely regulate the rate of data transfers between the workflow tasks via a specially-created QoS-enabled GridFTP server. Our workflow planner constructs a schedule that both specifies when/where individual tasks are to be executed, as well as when and at what rate data is to be transferred. Simulation results involving a simple workflow indicate that our system can achieve a 30% speedup when nodes show a computation/communication ratio of approximately 0.5. We reinforce and confirm these results via the actual implementation of the Montage workflow in the wide area, obtaining a maximum speedup of 31% and an average speedup with 16%. Overall, we believe that our data throttling Grid workflow system both executes workflows more efficiently (by better establishing balanced workflow graphs) and operates more cooperatively with unrelated concurrent Grid activities by consuming less overall network bandwidth, allowing such unrelated activities to execute more efficiently as well.

A Dynamic Scheduling Approach for Coordinated Wide-Area Data Transfers Using GridFTP

Gaurav Khanna [1], Umit Catalyurek [2], Tahsin Kurc [2], Rajkumar Kettimuthu [3],
P. Sadayappan [1] and Joel Saltz [2]
[1]Department of Computer Science and Engineering, The Ohio State University
[2]Department of Biomedical Informatics, The Ohio State University
[3]Mathematics and Computer Science Division, Argonne National Laboratory

Many scientific applications need to stage large volumes of files from one set of machines to another set of machines in a wide-area network. Efficient execution of such data transfers needs to take into account the heterogeneous nature of the environment and dynamic availability of shared resources. This paper proposes an algorithm that dynamically schedules a batch of data transfer requests with the goal of minimizing the overall transfer time. The proposed algorithm performs simultaneous transfer of chunks of files from multiple file replicas, if the replicas exist. Adaptive replica selection is employed to transfer different chunks of the same file by taking dynamically changing network bandwidths into account. We utilize GridFTP as the underlying mechanism for data transfers. The algorithm makes use of information from past GridFTP transfers to estimate network bandwidths and resource availability. The efficiency of the algorithm is evaluated on a wide-area testbed.

Lightweight Process Migration and Memory Prefetching in OpenMosix

Roy S.C. Ho, Cho-Li Wang and Francis C.M. Lau
Department of Computer Science
The University of Hong Kong
{scho,clwang,fcmlau}@cs.hku.hk

We propose a lightweight process migration mechanism and an adaptive memory prefetching scheme called AMPoM (Adaptive Memory Prefetching in openMosix), whose goal is to reduce the migration freeze time in openMosix while ensuring the execution efficiency of migrants. To minimize the freeze time, our system transfers only a few pages to the destination node during process migration. After the migration, AMPoM analyzes the spatial locality of memory access and iteratively prefetches memory pages from remote to hide the latency of inter-node page faults. AMPoM adopts a unique algorithm to decide which and how many pages to prefetch. It tends to prefetch more aggressively when a sequential access pattern is developed, when the paging rate of the process is high or when the network is busy. This advanced strategy makes AMPoM highly adaptive to different application behaviors and system dynamics. The HPC Challenge benchmark results show that AMPoM can avoid 98% of migration freeze time while preventing 85-99% of page fault requests after the migration. Compared to openMosix which does not have remote page fault, AMPoM induces a modest overhead of 0-5% additional runtime. When the working set of a migrant is small, AMPoM outperforms openMosix considerably due to the reduced amount of data transfer. These results indicate that by exploiting memory access locality and prefetching, process migration can be a lightweight operation with little software overhead in remote paging.

SESSION 24

Algorithms - Sensor Networks

Efficient and Robust Sensor Data Aggregation Using Linear Counting Sketches

Yao-Chung Fan
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan, R. O. C.
dr938318@cs.nthu.edu.tw

Arbee L.P. Chen
Department of Computer Science
National Chengchi University
Taipei, Taiwan, R. O. C.
alpchen@cs.nccu.edu.tw

Sensor networks have received considerable attention in recent years, and are often employed in the applications where data are difficult or expensive to collect. In these applications, in addition to individual sensor readings, statistical aggregates such as Min and Count over the readings of a group of sensor nodes are often needed. To conserve resources for sensor nodes, in-network strategies are adopted to process the aggregates. One primitive in-network aggregation strategy is the tree-based aggregation, where the aggregates are computed along a spanning tree over a sensor network. However, a shortcoming with the tree-based aggregation is that it is not robust against communication failures, which are common in sensor networks. One of the solutions to overcome this shortcoming is to enable multi-path routing, by which each node broadcasts its reading or a partial aggregate to multiple neighbors. However, multi-path routing based aggregation typically suffers from the problem of overcounting sensor readings. In this study, we propose using the linear counting sketches for multi-path routing based in-network aggregation. We claim that the use of the linear counting sketches makes our approach considerably more accurate than previous approaches using the same sketch space. Our approach also enjoys low variances in term of the aggregate accuracy, and low overheads either in computations or sketch space. Through extensive experiments with real-world and synthetic data, we demonstrate the efficiency and effectiveness of using the linear counting sketches as a solution for the in-network aggregation.

Data Gathering in Wireless Sensor Networks with Mobile Collectors

Ming Ma and Yuanyuan Yang
Department of Electrical and Computer Engineering,
State University of New York, Stony Brook, NY 11794, USA

In this paper, we propose a new data gathering mechanism for large scale wireless sensor networks by introducing mobility into the network. A mobile data collector, or for convenience called M-collector in this paper, could be a mobile robot or a vehicle equipped with a powerful transceiver and battery, works like a mobile base station and gathers data while moving through the field. An M-collector starts the data gathering tour periodically from the static data sink, polls each sensor while traversing the transmission range of the sensor, then collects data directly from the sensor without relay (i.e., in a single hop), finally returns and uploads data to the data sink. Since data packets are gathered directly without relay and collision, the lifetime of sensors is expected to be prolonged, and sensors can be made very simple and inexpensive. We focus on the problem of minimizing the length of each data gathering tour and refer to this problem as the single-hop data gathering problem, or SHDGP for short. We first formalize the SHDGP problem into a mixed integer program and prove its NP-hardness. We then present a heuristic tour-planning algorithm for a single M-collector. For the applications with a strict distance/time constraint for each data gathering tour, we also utilize multiple M-collectors to traverse through several shorter sub-tours concurrently to satisfy the distance/time constraint. Our single-hop data gathering scheme will improve the scalability and solve intrinsic problems of large scale homogeneous networks, and can be used in both connected networks and disconnected networks. The simulation results demonstrate that the new data gathering algorithm can greatly reduce the moving distance of the collectors compared to the covering line approximation algorithm, and is close to the optimal algorithm in small networks. In addition, the proposed data gathering scheme can prolong the network lifetime significantly compared to a network which has only a static data collector, or a network in which the mobile collector can only move along straight lines.

Energy Efficient Sleep Scheduling Based on Moving Directions in Target Tracking Sensor Network

Bo Jiang, Kai Han and Binoy Ravindran
Virginia Polytechnic Institute & State University
ECE Dept. Virginia Tech
Blacksburg, VA 24061, USA
{bjiang,khan05,binoy}@vt.edu

Hyeonjoong Cho
ETRI
Daejoen, South Korea
raycho@etri.re.kr

This paper presents a target direction-based sleep scheduling algorithm (TDSS) for target tracking surveillance sensor networks. TDSS reduces the number of the proactively awakened sensor nodes and schedules their sleep pattern to enhance energy efficiency but suffer little performance loss. Both approaches are based on two probabilistic distribution models of target moving directions, normal distribution and linear distribution. We compare TDSS with the two models against the legacy circle-based proactively waking up scheme (Circle) and a working node reducing algorithm - MCTA. The evaluation result shows that TDSS achieves better energy efficiency but with less performance loss in terms of detection probability and detection delay.

SESSION 25

Applications - Additional Applications

All-Pairs: An Abstraction for Data-Intensive Cloud Computing

Christopher Moretti, Jared Bulosan, Douglas Thain and Patrick J. Flynn
Department of Computer Science and Engineering, University of Notre Dame

Although modern parallel and distributed computing systems provide easy access to large amounts of computing power, it is not always easy for non-expert users to harness these large systems effectively. A large workload composed in what seems to be the obvious way by a naive user may accidentally abuse shared resources and achieve very poor performance. To address this problem, we propose that production systems should provide end users with high-level abstractions that allow for the easy expression and efficient execution of data intensive workloads. We present one example of an abstraction – All-Pairs – that fits the needs of several data-intensive scientific applications. We demonstrate that an optimized All-Pairs abstraction is both easier to use than the underlying system, and achieves performance orders of magnitude better than the obvious but naive approach, and twice as fast as a hand-optimized conventional approach.

Junction Tree Decomposition for Parallel Exact Inference

Yinglong Xia[1] and Viktor K. Prasanna[2,1]
Computer Science Department[1]
Ming Hsieh Department of Electrical Engineering[2]
University of Southern California
Los Angeles, CA 90089, U.S.A.
{yinglonx, prasanna}@usc.edu

We present a junction tree decomposition based algorithm for parallel exact inference. This is a novel parallel exact inference method for evidence propagation in an arbitrary junction tree. If multiple cliques contain evidence, the performance of any state-of-the-art parallel inference algorithm achieving logarithmic time performance is adversely affected. In this paper, we propose a new approach to overcome this problem. We decompose a junction tree into a set of chains. Cliques in each chain are partially updated after the evidence propagation. These partially updated cliques are then merged in parallel to obtain fully updated cliques. We derive the formula for merging partially updated cliques and estimate the computation workload of each step. Experiments conducted using MPI on state-of-the-art clusters showed that the proposed algorithm exhibits linear scalability and superior performance compared with other parallel inference methods.

Accelerating Reed-Solomon Coding in RAID Systems with GPUs

Matthew L. Curry and Anthony Skjellum
University of Alabama at Birmingham
Computer and Information Sciences
115A Campbell Hall
1300 University Boulevard
Birmingham, Alabama 35294-1170
{curryml,tony}@cis.uab.edu

H. Lee Ward and Ron Brightwell
Sandia National Laboratories
Computation, Computers, Information and Mathematics
PO Box 5800
Albuquerque, New Mexico 87185-1319
{lee,rbbrigh}@sandia.gov

Graphical Processing Units (GPUs) have been applied to more types of computations than just graphics processing for several years. Until recently, however, GPU hardware has not been capable of efficiently performing general data processing tasks. With the advent of more general-purpose extensions to GPUs, many more types of computations are now possible. One such computation that we have identified as being suitable for the GPU's unique architecture is Reed-Solomon coding in a manner appropriate for RAID-type systems. In this paper, we motivate the need for RAID with triple-disk parity and describe a pipelined architecture for using a GPU for this purpose. Performance results show that the GPU can outperform a modern CPU on this problem by an order of magnitude and also confirm that a GPU can be used to support a system with at least three parity disks with no performance penalty.

SESSION 26

Applications - Applications and the Cell Processor

High Performance MPEG-2 Software Decoder on the Cell Broadband Engine

David A. Bader
College of Computing
Georgia Institute of Technology
bader@cc.gatech.edu

Sulabh Patel
Electronic Arts, Inc.
sulabh@gmail.com

The Sony-Toshiba-IBM Cell Broadband Engine is a heterogeneous multicore architecture that consists of a traditional microprocessor (PPE) with eight SIMD co-processing units (SPEs) integrated on-chip. While the Cell/B.E. processor is designed with multimedia applications in mind, there are currently no open-source, optimized implementations of such applications available. In this paper, we present the design and implementation behind the creation of an optimized MPEG-2 software decoder for this unique parallel architecture, and demonstrate its performance through an experimental study.

This is the first parallelization of an MPEG-2 decoder for a commodity heterogeneous multicore processor such as the IBM Cell/B.E. While Drake et al. have recently parallelized MPEG-2 using StreamIt for a streaming architecture, our algorithm is quite different and is the first to address the new challenges related to the optimization and tuning of a multicore algorithm with DMA transfers and local store memory. Our design and efficient implementation target the architectural features provided by the heterogeneous multicore processor. We give an experimental study on Sony PlayStation 3 and IBM QS20 dual-Cell Blade platforms. For instance, using 16 SPEs on the IBM QS20, our decoder runs 3.088 times faster than a 3.2 GHz Intel Xeon and achieves a speedup of over 10.545 compared with a PPE-only implementation. Our source code is freely-available through SourceForge under the CellBuzz project.

High-Speed String Searching Against Large Dictionaries on the Cell/B.E. Processor

Daniele Paolo Scarpazza[1], Oreste Villa[2] and Fabrizio Petrini[1]

[1] IBM TJ Watson Research Center
Cell Solution Department
Yorktown Heights, NY 10598 USA
dpscarpazza,fpetrin@us.ibm.com

[2] Politecnico di Milano
Dipartimento di Elettronica e Informazione
Milano I-20133, Italy
ovilla@elet.polimi.it

Our digital universe is growing, creating exploding amounts of data which need to be searched, filtered and protected. String searching is at the core of the tools we use to curb this explosion, such as search engines, network intrusion detection systems, spam filters, and anti-virus programs. But as communication speed grows, our capability to perform string searching in real-time seems to fall behind. Multi-core architectures promise enough computational power to cope with the incoming challenge, but it is still unclear which algorithms and programming models to use to unleash this power.

We have parallelized a popular string searching algorithm, Aho-Corasick, on the IBM Cell/B.E. processor, with the goal of performing exact string matching against large dictionaries. In this article we propose a novel approach to fully exploit the DMA-based communication mechanisms of the Cell/B.E. to provide an unprecedented level of aggregate performance with irregular access patterns.

We have discovered that memory congestion plays a crucial role in determining the performance of this algorithm. We discuss three aspects of congestion: memory pressure, layout issues and hot spots, and we present a collection of algorithmic solutions to alleviate these problems and achieve quasi-optimal performance.

The implementation of our algorithm provides a worst-case throughput of 2.5 Gbps, and a typical throughput between 3.3 and 4.4 Gbps, measured on realistic scenarios with a two-processor Cell/B.E. system.

Financial Modeling on the Cell Broadband Engine

Virat Agarwal
virat@cc.gatech.edu
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332

Lurng-Kuo Liu
lkliu@us.ibm.com
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

David A. Bader
bader@cc.gatech.edu
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332

High performance computing is critical for financial markets where analysts seek to accelerate complex optimizations such as pricing engines to maintain a competitive edge. In this paper we investigate the performance of financial workloads on the Sony-Toshiba-IBM Cell Broadband Engine, a heterogeneous multicore chip architected for intensive gaming applications and high performance computing. We analyze the use of Monte Carlo techniques for financial workloads and design efficient parallel implementations of different high performance pseudo and quasi random number generators as well as normalization techniques. Our implementation of the Mersenne Twister pseudo random number generator outperforms current Intel and AMD architectures by over an order of magnitude. Using these new routines, we optimize European Option (EO) and Collateralized Debt Obligation (CDO) pricing algorithms. Our Cell-optimized EO pricing achieves a speedup of over 2 in comparison with using RapidMind SDK for Cell, and comparing with GPU, a speedup of 1.26 as compared with using RapidMind SDK for GPU (NVIDIA GeForce 8800), and a speedup of 1.51 over NVIDIA GeForce 8800 (using CUDA). Our detailed analyses and performance results demonstrate that the Cell/B.E. processor is well suited for financial workloads and Monte Carlo simulation.

Parallel Biological Sequence Alignments on the Cell Broadband Engine

Abhinav Sarje and Srinivas Aluru
Department of Electrical and Computer Engineering
Iowa State University
{asarje, aluru}@iastate.edu

Sequence alignment and its many variants are a fundamental tool in computational biology. There is considerable recent interest in using the Cell Broadband Engine, a heterogenous multi-core chip that provides high performance, for biological applications. However, work so far has been limited to computing optimal alignment scores using quadratic space under the basic global/local alignment algorithm. In this paper, we present a comprehensive study of developing sequence alignment algorithms on the Cell exploiting its thread and data level parallelism features. First, we develop a Cell implementation that computes optimal alignments and adopts Hirschberg's linear space technique. The former is essential as merely computing optimal alignment scores is not useful while the latter is needed to permit alignments of longer sequences. We then present Cell implementations of two advanced alignment techniques – spliced alignments and syntenic alignments. In a spliced alignment, consecutive non-overlapping portions of a sequence align with ordered non-overlapping, but non-consecutive portions of another sequence. Spliced alignments are useful in aligning mRNA sequences with corresponding genomic sequences to uncover gene structure. Syntenic alignments are used to discover conserved exons and other sequences between long genomic sequences from different organisms. We present experimental results for these three types of alignments on the Cell BE and report speedups of about 4 on six SPUs on the Playstation 3, when compared to the respective best serial algorithms on the Cell BE and the Pentium 4 processor.

Workshop 1
Heterogeneity in Computing Workshop
HCW 2008

Divisible Load Scheduling with Result Collection on Heterogeneous Systems

Abhay Ghatpande, Hidenori Nakazato and Hiroshi Watanabe
School of Global Information and Telecommunication Studies
Waseda University, Tokyo
Email: abhay@toki.waseda.jp

Olivier Beaumont
INRIA Futurs - LaBRI
Bordeaux, France
Email: olivier.beaumont@labri.fr

Divisible Load Theory (DLT) is an established mathematical framework to study Divisible Load Scheduling (DLS). However, traditional DLT does not comprehensively deal with the scheduling of results back to source (i.e., result collection) on heterogeneous systems. In this paper, the DLSRCHETS (DLS with Result Collection on HETerogeneous Systems) problem is addressed. The few papers to date that have dealt with DLSRCHETS, proposed simplistic LIFO (Last In, First Out) and FIFO (First In, First Out) type of schedules as solutions to DLSRCHETS. In this paper, a new heuristic algorithm, ITERLP, is proposed as a solution to the DLSRCHETS problem. With the help of simulations, it is proved that the performance of ITERLP is significantly better than existing algorithms.

Scheduling DivisibleWorkloads on Heterogeneous Platforms under Bounded Multi-Port Model

Olivier Beaumont, Nicolas Bonichon and Lionel Eyraud-Dubois
Université de Bordeaux,
INRIA Bordeaux - Sud-Ouest,
France
{Olivier.Beaumont,Nicolas.Bonichon,Lionel.Eyraud-Dubois}@labri.fr

In this paper, we discuss complexity issues for scheduling divisible workloads on heterogeneous systems under the bounded multi-port model. To our best knowledge, this paper is the first attempt to consider divisible load scheduling under a realistic communication model, where the master node can communicate simultaneously to several slaves, provided that bandwidth constraints are not exceeded. In this paper, we concentrate on one round distribution schemes, where a given node starts its processing only once all data has been received. Our main contributions are (i) the proof that processors start working immediately after receiving their work (ii) the study of the optimal schedule in the case of 2 processors and (iii) the proof that scheduling divisible load under the bounded multi-port model is NP-complete. This last result strongly differs from divisible load literature and represents the first NP-completeness result when latencies are not taken into account.

Process Reassignment with Reduced Migration Cost in Grid Load Rebalancing

Lin Chen, Cho-Li Wang and Francis. C.M. Lau
Department of Computer Science
The University of Hong Kong
Pokfulam Road, Hong Kong
{lchen2, clwang, fcmlau}@cs.hku.hk

We study the load rebalancing problem in a heterogeneous grid environment that supports process migration. Given an initial assignment of tasks to machines, the problem consists of finding a process reassignment that achieves a desired better level of load balance with minimum reassignment (process migration) cost. Most previous algorithms for related problems aim mainly at improving the balance level (or makespan) with no explicit concern for the reassignment cost. We propose a heuristic which is based on local search and several optimizing techniques which include the guided local search strategy and the multi-level local search. The searching integrates both the change of workload and the migration cost introduced by a process movement into the movement selection, and enables a good tradeoff between low-cost movements and the improving balance level. Evaluations show that the proposed heuristic can find a solution with much lower migration cost for achieving the same balance level than previous greedy or local search algorithms for a range of problem cases.

Resource Allocation in a Client/Server Hybrid Network for Virtual World Environments

Luis Diego Briceño[1], Howard Jay Siegel[1,2], Anthony A. Maciejewski[1],
Ye Hong[1], Brad Lock[1], Mohammad Nayeem Teli[2], Fadi Wedyan[2],
Charles Panaccione[2], and Chen Zhang[2]

Colorado State University
[1]Department of Electrical & Computer Engineering
[2]Department of Computer Science
Fort Collins, CO 80523
{LDBricen, HJ, AAM, YHong, Bradley.Lock, Mohammad.Teli,
Fadi.Wedyan, Charles.Panaccione, Chen.Zhang}@colostate.edu

The creation of a virtual world environment (VWE) has significant costs, such as maintenance of server rooms, server administration, and customer service. The initial development cost is not the only factor that needs to be considered; factors such as the popularity of a VWE and unexpected technical problems during and after the launch can affect the final cost and success of a VWE. The capacity of servers in a client/server VWE is hard to scale and cannot adjust quickly to peaks in demand while maintaining the required response time. To handle these peaks in demand, we propose to employ users' computers as secondary servers. The introduction of users' computers as secondary servers allows the performance of the VWE to support an increase in users. In this study, we develop and implement five static heuristics to implement a secondary server scheme that reduces the time taken to compute the state of the VWE. The number of heterogeneous secondary servers, conversion of a player to a secondary server, and assignment of players to secondary servers are determined by the heuristics implemented in this study. A lower bound of the performance is derived to evaluate the results of the heuristics.

Optimizing Latency and Reliability of Pipeline Workflow Applications

Anne Benoit, Veronika Rehn-Sonigo and Yves Robert
ENS Lyon   Université de Lyon
LIP laboratory, UMR 5668, CNRS - ENS Lyon - INRIA - UCB Lyon 5668
46 allée d'Italie, 69364 Lyon Cedex 07, France
E-mail: {Anne.Benoit, Veronika.Sonigo, Yves.Robert}@ens-lyon.fr

Mapping applications onto heterogeneous platforms is a difficult challenge, even for simple application patterns such as pipeline graphs. The problem is even more complex when processors are subject to failure during the execution of the application. In this paper, we study the complexity of a bi-criteria mapping which aims at optimizing the latency (i.e., the response time) and the reliability (i.e., the probability that the computation will be successful) of the application. Latency is minimized by using faster processors, while reliability is increased by replicating computations on a set of processors. However, replication increases latency (additional communications, slower processors). The application fails to be executed only if all the processors fail during execution. While simple polynomial algorithms can be found for fully homogeneous platforms, the problem becomes NPhard when tackling heterogeneous platforms. This is yet another illustration of the additional complexity added by heterogeneity.

An Efficient, Model-Based CPU-GPU Heterogeneous FFT Library

Yasuhito Ogata[1,3], Toshio Endo[1,3], Naoya Maruyama[1,3] and Satoshi Matsuoka[1,2,3]
[1] Tokyo Institute of Technology
[2] National Institute of Informatics
[3] JST, CREST
ogata, endo@matsulab.is.titech.ac.jp
naoya.maruyama, matsu@is.titech.ac.jp

General-Purpose computing on Graphics Processing Units (GPGPU) is becoming popular in HPC because of its high peak performance. However, in spite of the potential performance improvements as well as recent promising results in scientific computing applications, its real performance is not necessarily higher than that of the current high-performance CPUs, especially with recent trends towards increasing the number of cores on a single die. This is because the GPU performance can be severely limited by such restrictions as memory size and bandwidth and programming using graphics-specific APIs. To overcome this problem, we propose a model-based, adaptive library for 2D FFT that automatically achieves optimal performance using available heterogeneous CPU-GPU computing resources. To find optimal load distribution ratios between CPUs and GPUs, we construct a performance model that captures the respective contributions of CPU vs. GPU, and predicts the total execution time of 2D-FFT for arbitrary problem sizes and load distribution. The performance model divides the FFT computation into several small sub steps, and predicts the execution time of each step using profiling results. Preliminary evaluation with our prototype shows that the performance model can predict the execution time of problem sizes that are 16 times as large as the profile runs with less than 20% error, and that the predicted optimal load distribution ratios have less than 1% error. We show that the resulting performance improvement using both CPUs and GPUs can be as high as 50% compared to using either a CPU core or a GPU.

On the Design, Control, and Use of a Reconfigurable Heterogeneous Multi-Core System-on-a-Chip

Tyrone Tai-On Kwok[1] and Yu-Kwong Kwok[1,2]
[1]Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
[2]Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373, USA

With the continued progress in VLSI technologies, we can integrate numerous cores in a single billion-transistor chip to build a multi-core system-on-a-chip (SoC). This also brings great challenges to traditional parallel programming as to how we can increase the performance of applications with increased number of cores. In this paper, we meet the challenges using a novel approach. Specifically, we propose a reconfigurable heterogeneous multi-core system. Under our proposed system, in addition to conventional processor cores, we introduce dynamically reconfigurable accelerator cores to boost the performance of applications. We have built a prototype of the system using FPGAs. Experimental evaluation demonstrates significant system efficiency of the proposed heterogeneous multi-core system in terms of computation and power consumption. Keywords: FPGA, heterogeneous multi-core, network-on-chip (NoC), parallel processing, reconfigurable computing, system-on-a-chip (SoC).

A Reputation Algorithm for a Self-Organizing System Based upon Resource Virtualization

Dan C. Marinescu, Chen Yu and Gabriela M. Marinescu
School of Electrical Engineering and Computer Science
University of Central Florida, Orlando, Fl, 32816
Email: (dcm, yuchen, magda)@cs.ucf.edu

John P. Morrison and Christoffer Norvik
Computer Science Department
University College Cork, Cork, Ireland
Email: (j.morrison, c.norvik)@cs.ucc.ie

In the service-oriented architecture introduced in [10] the reputation affects the prices producers could ask for their resources, the prices consumers of resources are willing to pay, and the com mission charged for every contract. In this paper we introduce the GRB algorithm to quantify the reputations of the participants. The system is self-organizing and this affects the space requirements and the communication complexity of the algorithm.

Automatic Middleware Deployment Planning on Heterogeneous Platforms

Pushpinder Kaur Chouhan [#1], Eddy Caron [*2], Frédéric Desprez [*3]
[#]Department of Computer Science,
University College Cork, Ireland.
[1]p.chouhan@cs.ucc.ie
[*]Université de Lyon
LIP, UMR ENS Lyon - CNRS - INRIA - UCBL, Lyon
46, Allée d'Italie. F-69364 Lyon, France
[2]ENS Lyon
Eddy.Caron@ens-lyon.fr
[3]INRIA
Frederic.Desprez@inria.fr

The use of many distributed, heterogeneous resources as a large collective platform offers great potential. A key issue for these grid platforms is middleware scalability and how middleware services can be mapped on the available resources. Optimizing deployment is a difficult problem with no existing general solutions. In this paper, we address the following problem: how to perform out an adapted deployment for a hierarchy of servers and resource brokers on a heterogeneous system?

Our objective is to generate a best platform from the available nodes so as to fulfill the clients demands. However, finding the best deployment among heterogeneous resources is a hard problem since it is close to find the best broadcast tree in a general graph, which is known to be NP-complete.

Thus, in this paper, we present a heuristic for middleware deployment on heterogeneous resources. We apply our heuristic to automatically deploy a distributed Problem Solving Environment on a large scale grid.We present experiments comparing the automatically generated deployment against a number of other reasonable deployments.

Asynchronous Genetic Search for Scientific Modeling on Large-Scale Heterogeneous Environments

Travis Desell, Boleslaw Szymanski and Carlos Varela
Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, U.S.A.
{deselt, szymansk, cvarela}@cs.rpi.edu

Use of large-scale heterogeneous computing environments such as computational grids and the Internet has become of high interest to scientific researchers. This is because the increasing complexity of their scientific models and data sets is drastically outpacing the increases in processor speed while the cost of supercomputing environments remains relatively high. However, the heterogeneity and unreliability of these environments, especially the Internet, make scalable and fault tolerant search methods indispensable to effective scientific model verification. The paper introduces two versions of asynchronous master-worker genetic search and evaluates their convergence and performance rates in comparison to traditional synchronous genetic search on both a IBM BlueGene supercomputer and using the MilkyWay@HOME BOINC Internet computing project (see http://milkyway.cs.rpi.edu). The asynchronous searches not only perform faster on heterogeneous grid environments as compared to synchronous search, but also achieve better convergence rates for the astronomy model used as the driving application, providing a strong argument for their use on grid computing environments and by the MilkyWay@Home BOINC internet computing project.

Enhancing Build-Portability for Scientific Applications Across Heterogeneous Platforms

Magdalena Sławińska, Jarosław Sławiński and Vaidy Sunderam
Dept. of Math and Computer Science, Emory University
400 Dowman Drive
Atlanta, GA 30322, USA
fmagg,jaross,vssg@mathcs.emory.edu

The paper addresses the build problem in the HPC arena that results from heterogeneity in hardware architectures, system software, and application build systems. We propose a new approach that generalizes current build systems, and enhances their portability across heterogeneous high-end platforms. Although the original application build system needs to be modified it is a one-time activity that allows us to extract target platform specific information and store it into profiles. Our toolkit exploits profiles to retrieve appropriate target-specific information during the actual build process. Our developed mechanism termed late binding enables dynamic concretization of platform-specific variables from profiles. This approach simplifies the build process for heterogeneous environments, promotes profile reuse, and improves its portability. In order to verify our approach in practice we have applied our methodology to a production molecular dynamics code (the CPMD application).

Enabling Personal Clusters on Demand for Batch Resources Using Commodity Software

Yang-Suk Kee and Carl Kesselman
Information Sciences Institute, University of Southern California
{yskee,carl}@isi.edu
Daniel Nurmi and Rich Wolski
University of California, Santa Barbara
{dan,rich}@cs.ucsb.edu

Providing QoS (Quality of Service) in batch resources against the uncertainty of resource availability due to the space-sharing nature of scheduling policies is a critical capability required for high-performance computing. This paper introduces a technique called personal cluster which reserves a partition of batch resources on user's demand in a best-effort manner. A personal cluster provides a private cluster dedicated to the user during a user-specified time period by installing a user-level resource manager on the resource partition. This technique not only enables cost-effective resource utilization and efficient task management but also provides the user a uniform interface to heterogeneous resources regardless of local resource management software. A prototype implementation using a PBS batch resource manager and Globus Toolkits based on Web Services shows that the overhead of instantiating a personal cluster of medium size is small, which is just about 1 minute for a personal cluster having 32 processors

Workshop 2

Workshop on Parallel and Distributed Real-Time Systems

WPDRTS 2008

Hybrid-Priority Real-Time Scheduling

Sanjoy Baruah — The University of North Carolina
Nathan Fisher — Wayne State University

A hybrid scheduling algorithm is proposed, which integrates features of the Fixed Priority (FP) and Earliest Deadline First (EDF) scheduling policies. It is shown that this hybrid scheduling algorithm is a generalization of both FP and EDF, and tends to retain most of the desirable properties and features of both individual policies. Two exact (i.e., necessary and sufficient) tests are derived for sporadic task systems scheduled by the hybrid scheduling algorithm.

Schedulability Analysis of Generalized Multiframe Traffic on Multihop-Networks Comprising Software-Implemented Ethernet-Switches

Björn Andersson
IPP Hurray Research Group
Polytechnic Institute of Porto, Portugal
bandersson@dei.isep.ipp.pt

Performance Assessment of OMG Compliant Data Distribution Middleware

Christian Esposito and Stefano Russo
Dipartimento di Informatica e Sistemistica,
Universita' di Napoli Federico II
Via Claudio 21, 80125 - Napoli, Italy
{christian.esposito, stefano.russo}@unina.it

Dario Di Crescenzo
Consorzio SESM Soluzioni Innovative,
Circum. Est. Napoli, 80014
Giugliano in Campania (NA), Italy
ddicrescenzo@sesm.it

Event-Driven Architectures (EDAs) are widely used to make distributed mission critical software systems more-efficient and scalable. In the context of EDAs, Data Distribution Service (DDS) is a recent standard by the Object Management Group that offers a rich support for Quality-of-Service and balances predictable behavior and implementation efficiency. The DDS specification does not outline how messages are delivered, so several architectures are nowadays available. This paper focuses on performance assessment of OMG DDS-compliant middleware technologies. It provides three contributions to the study of evaluating the performance of DDS implementations: 1) describe the challenges to be addressed; 2) propose possible solutions; 3) define a representative workload scenario for evaluating the performance and scalability of DDS platforms. At the end of the paper, a case study of DDS performance assessment, performed with the proposed benchmark, is presented.

Addressing the Challenges of Mission-Critical Information Management in Next-Generation Net-Centric Pub/Sub Systems with OpenSplice DDS

Douglas C. Schmidt and Hans van't Hag
PrismTech Corporation,
6 Lincoln Knoll Lane, Suite 100, Burlington, MA, 01803, USA
{doug.schmidt, hans.vanthag}@prismtech.com

An important class of new and existing applications requires cooperating entities to share and exchange information seamlessly in real-time, while meeting stringent requirements for quality of service (QoS) and heterogeneity. This paper describes how the recently adopted OMG DDS standard, and the OpenSplice DDS implementation of this standard, support the data distribution and management challenges posed by next-generation distributed systems, and systems of systems.

Shared Resources and Precedence Constraints with Capacity Sharing and Stealing

Luís Nogueira, Luś Miguel Pinho
IPP Hurray Research Group
School of Engineering (ISEP), Polytechnic Institute of Porto (IPP), Portugal
{luis,lpinho}@dei.isep.ipp.pt

This paper proposes a new strategy to integrate shared resources and precedence constraints among real-time tasks, assuming no precise information on critical sections and computation times is available. The concept of bandwidth inheritance is combined with a greedy capacity sharing and stealing policy to efficiently exchange bandwidth among tasks, minimising the degree of deviation from the ideal system's behaviour caused by inter-application blocking. The proposed Capacity Exchange Protocol (CXP) focus on exchanging extra capacities as early, and not necessarily as fairly, as possible. This loss of optimality is worth the reduced complexity as the protocol's behaviour nevertheless tends to be fair in the long run and outperforms other solutions in highly dynamic scenarios, as demonstrated by extensive simulations.

Automated Generation of Explicit Connectors for Component Based Hardware/Software Interaction in Embedded Real-Time Systems

Wolfgang Forster[1], Christof Kutschera[2], Andreas Steininger[1] and Karl M. Göschka[1]
[1]Vienna University of Technology
Karlsplatz 13, A-1040 Vienna, Austria
wolfgang.forster,andreas.steininger,karl.goeschka@tuwien.ac.at
[2]University of Applied Sciences Technikum Vienna
Department of Embedded Systems
Höchstädtplatz 5, A-1200 Vienna, Austria
kutschera@technikum-wien.at

The complexity of today's embedded real-time systems is continuously growing with high demands on dependability, resource-efficiency, and reusability. Two solution approaches address these needs: First, in the component based software engineering (CBSE) paradigm, software is decomposed into self-contained components with explicit interactions and context dependencies. Connectors represent the abstraction of interactions between these components. Second, components can be shifted from software to reconfigurable hardware, typically field programmable gate arrays (FPGAs), in order to meet real-time constraints. This paper proposes a component-based concept to support efficient hardware/software co-design: A hardware component together with the hardware/software connector can seamlessly replace a software component with the same functionality, while the particularities of the alternative interaction are encapsulated in the component connector. Our approach provides for tools that can generate all necessary interaction mechanisms between hardware and software components. A proof-of-concept application demonstrates the advantages of our concept: Rapid change and comparison of different partitioning decisions due to automated and faultless generation of the hardware/software connectors.

Real-Time Dynamic Relinking

Mathias Ekman
Bombardier Transportation
721 73 Västerås, Sweden
mathias.ekman@se.transport.bombardier.com

Henrik Thane
Mälardalen University
Department of Computer Science and Electronics
henrik.thane@mdh.se

In this paper we present a method for automatically, on demand, linking entire functions into statically linked running embedded multi-tasking real-time applications. The purpose is to facilitate dynamic instrumentation of deployed systems. The method makes it possible to dynamically instrument the target in run-time, without preparing the source code. Code segments that are modified are substituted on the function level by the introduction of a dynamic relink method. The actual modification of the execution binary is performed in a safe and controlled manner by a low interference task. An algorithm is introduced for reusing memory from obsolete functions.

SelfS – A Real-Time Protocol for Virtual Ring Topologies

B. Griese, A. Brinkmann and M. Porrmann
Heinz Nixdorf Institute, Paderborn, Germany
{bgriese, brinkman, porrmann}@hni.upb.de

Real-time automation systems have evolved from centrally controlled sensor-actor systems to complex distributed computing systems. Therefore, the communication system becomes a crucial component that strongly influences performance. In this paper we present a simple distributed communication protocol that meets hard real-time constraints without requiring complex synchronization mechanisms. An advantage of the distributed protocol is that network planning can be reduced to a minimum. The protocol is based on virtual rings and can be easily embedded into arbitrary network topologies. Besides a detailed evaluation and analysis of the protocol, the paper includes lower bounds on jitter and performance for arbitrary communication patterns.

Design and Optimization of a Distributed, Embedded Speech Recognition System

Chung-Ching Shen, William Plishker, and Shuvra S. Bhattacharyya
Dept. of Electrical and Computer Engineering, and Institute for Advanced Computer Studies
University of Maryland at College Park, USA
{ccshen, plishker, ssb}@umd.edu

In this paper, we present the design and implementation of a distributed sensor network application for embedded, isolated-word, real-time speech recognition. In our system design, we adopt a parameterized-dataflow-based modeling approach to model the functionalities associated with sensing and processing of acoustic data, and we implement the associated embedded software on an off-the-shelf sensor node platform that is equipped with an acoustic sensor. The topology of the sensor network deployed in this work involves a clustered network hierarchy. A customized time division multiple access protocol is developed to manage the wireless channel. We analyze the distribution of the overall computation workload across the network to improve energy efficiency. In our experiments, we demonstrate the recognition accuracy for our speech recognition system to verify its functionality and utility. We also evaluate improvements in network lifetime to demonstrate the effectiveness of our energy-aware optimization techniques.

Model-driven Specification of Component-based Distributed Real-time and Embedded Systems for Verification of Systemic QoS Properties

James H. Hill
Vanderbilt University
Nashville, TN, USA
j.hill@vanderbilt.edu

Aniruddha Gokhale
Vanderbilt University
Nashville, TN, USA
a.gokhale@vanderbilt.edu

The adage “the whole is not equal to the sum of its parts” is very appropriate in the context of verifying a range of systemic properties, such as deadlocks, correctness, and conformance to quality of service (QoS) requirements, for component-based distributed real-time and embedded (DRE) systems. For example, end-to-end worst case response time (WCRT) in component-based DRE systems is not as simple as accumulating WCRT for each individual component in the system because of inherent complexities introduced by the large solution space of possible deployment and configurations. This paper describes a novel process and tool-based artifacts that simplify the formal specification of component-based DRE systems for verification of systemic QoS properties. Our approach is based on the mathematical formalism of Timed Input/Output Automata and uses generative programming techniques for automating the verification of systemic QoS properties for component-based DRE systems.

Formal Modeling and Analysis of Real-Time Resource-Sharing Protocols in Real-Time Maude

Peter Csaba Ölveczky[1], Pavithra Prabhakar[2] and Xue Liu[3]
[1] Department of Informatics, University of Oslo
[2] Department of Computer Science, University of Illinois at Urbana-Champaign
[3] School of Computer Science, McGill University

This paper presents general techniques for formally modeling, simulating, and model checking real-time resource-sharing protocols in Real-Time Maude. The "scheduling subset" of our techniques has been used to find a previously unknown subtle bug in a state-of-the-art scheduling algorithm. This paper also shows how our general techniques can be instantiated to model and analyze the well known priority inheritance protocol.

Verification of Networks of Timed Automata Using mCRL2

Jan Friso Groote, Michel A. Reniers and Yaroslav S. Usenko
Laboratory for Quality Software, Department of Mathematics and Computer Science,
Technical University of Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

It has been our long time wish to combine the best parts of the real-time verification methods based on timed automata (TA) (the use of regions and zones), and of the process-algebraic approach of languages like LOTOS and timed muCRL. This could provide us with additional verification possibilities for Real-Time systems, not available in existing timed-automata-based tools like UPPAAL. The language mCRL2 offers a uniform framework for the specification of data and processes. Data are specified by equational specifications: one can declare sorts and functions working upon these sorts, and describe the meaning of these functions by equational axioms. Processes are described in process algebraic style, where the particular process syntax stems from ACP, extended with data-parametric ingredients: there are constructs for conditional composition, and for data-parametric choice and communication. As is common in process algebra, infinite processes are specified by means of (finite systems of) recursive equations. In mCRL2 such equations can also be data-parametric. In our previous work, we use the existing results to translate a timed automaton to a timed muCRL processes in the form of a Timed Linear Process Specification (TLPS). This translation uses a very simple sort Time to represent the real-time clock values. As a result we obtain a semantically equivalent specification in timed muCRL. The resulting timed muCRL process only uses a very restricted subset of the full syntax of the language: it has the form of a TLPS. In the present paper, we extend this approach to networks of communicating timed automata with shared variables In this paper, we apply the steps described before to the example of Fischer's protocol. After translating the timed automaton describing Fischer's protocol to an mCRL2 process, this process is linearized, i.e., transformed into an equivalent TLPS. The resulting TLPS is in the input format of the discretization and time abstraction transformations. Applying the results of our previous work results in a finite representation of the original system which can be used for establishing correctness easily.

Workshop 3
Reconfigurable Architectures Workshop
RAW 2008

Sustainable (Re-) Configurable Solutions for the High Volume SoC Market

Fabio Campi, Luca Ciccarelli
FTM, STMicroelectronics
Viale Olivetti 2, Agrate Brianza (MI)

Claudio Mucci
ARCES - University of Bologna
Viale Pepoli 3/2, Bologna, Italy

The application of embedded run-time configurable architectures to System-on-chip design has long been considered a possible major enabling factor, especially in the direction of lowering time-to-market of new products as well as mitigating NRE costs related to verification, bug-fixes and product upgrades. In fact, while achieving significant success in specific application fields, reconfigurable computing has so far mostly failed to reach the high-volume application specific standard products (ASSP) that both in terms of volumes and revenues represent the largest share of today's SoC market. This is essentially due to the area overhead induced by these solutions with respect to standard ASIC design styles, which is unaffordable for the low margins that characterize this specific product class. In this paper, the exploitation of mask-programmable hardware technologies for deploying high volume ASSP is evaluated as a possible mitigation factor to the above discussed issues. The paper provides an introduction to mask-programmable technologies as well as an overview and a classification of most significant available trends and solutions in the field. In particular, the application of mask-level programmability in the context of the most significant trends in reconfigurable architectures is thoroughly discussed. In the authors' opinion it is both useful and necessary to capitalize on and exploit the valuable legacy created by 10 years of exploration of reconfigurable architectures in the context of the new possibilities offered by the emergence of mask-programmable options as a significant factor in SoC design.

Design Methodology for Partial Dynamic Reconfiguration: A New Degree of Freedom in the HW/SW Codesign

Marco D. Santambrogio and Donatella Sciuto
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5
20133 Milano, Italy
{santambr,sciuto}@elet.polimi.it

Many emerging products in communication, computing and consumer electronics demand that their functionality remains flexible also after the system has been manufactured and that is why the reconfiguration is starting to be considered into the design flow as a new relevant degree of freedom, in which the designer can have the system autonomously modify its functionalities according to the application's changing needs. Therefore, reconfigurable devices, such as FPGAs, introduce yet another degree of freedom in the design workflow: the designer can have the system autonomously modify the functionality carried out by the IP core according to the application's changing needs while it runs. Research in this field is, indeed, being driven towards a more thorough exploitation of the reconfiguration capabilities of such devices, so as to take advantage of them not only at compile-time, i.e. at the time when the system is first deployed, but also at run-time, which allows the reconfigurable device to be reprogrammed without the rest of the system having to stop running. This paper presents emerging methodologies to design reconfigurable applications, providing, as an example the workflow defined at the Politecnico di Milano.

On the Efficiency and Accuracy of Hybrid Pseudo-Random Number Generators for FPGA-Based Simulations

Amirhossein Alimohammad, Saeed Fouladi Fard, Bruce F. Cockburn and Christian Schlegel
Dept. of Electrical and Computer Engineering, Univ. of Alberta, Edmonton, Canada
E-mail: {amir, saeed, cockburn, schlegel}@ece.ualberta.ca

Most commonly-used pseudo-random number generators (PNGs) in computer systems are based on linear recurrence. These deterministic PNGs have fast and compact implementations, and can ensure very long periods. However, the points generated by linear PNGs in fact have a regular lattice structure and are thus not suitable for applications that rely on the assumption of uniformly distributed pseudo-random numbers (PNs). In this paper we propose and evaluate several fast and compact linear, non-linear, and hybrid PNGs for a field-programmable gate array (FPGA). The PNGs have excellent equidistribution properties and very small autocorrelations, and have very long repetition periods. The distribution and long-range correlation properties of the new generators are efficiently, and much more rapidly, estimated at hardware speeds using designed modules within the FPGA. The results of these statistical tests confirm that the combination of several linear PNGs or the combination of even one small non-linear PNG with a linear PNG significantly improves the statistical properties of the generated PNs.

Rotationally Invariant Sparse Patch Matching on GPU and FPGA

Zachary K. Baker and Reid Porter
Los Alamos National Laboratory
Los Alamos, NM 87545
Email: {zbaker, rporter}@lanl.gov

Vector and data-flow processors are particularly strong at dense, regular computation. Sparse, irregular data layouts cause problems because their unpredictable data access patterns prevent computational pipelines from filling effectively. A number of algorithms in image processing have been proposed which are not dense, and instead apply local neighborhood operations to a sparse, irregular set of points.

Sparse and irregular data transfer is difficult for modern processors because they have more processing power than memory bandwidth. However, if the computation can be expanded while not increasing the bandwidth, modern processors can be made more efficient.

The application targeted in this paper is patch matching over large scenes. Given two sequential frames of video data, corresponding points between the two frames are found. Correspondences are determined by comparing small image patches around each point. By rotating and comparing patches of the image over a range of angles, it is possible to more accurately match them through the scene. Rotation and interpolation are required to produce an appropriate image to compare against.

Results for CPU, FPGA, and GPU are presented, with FPGA far outperforming the GPU or CPU due to its potential for high levels of hardware parallelism as the total volume of computation increases.

A Bandwidth Optimized SDRAM Controller for the MORPHEUS Reconfigurable Architecture

Sean Whitty and Rolf Ernst
Institute of Computer and Communication Network Engineering
Technical University of Braunschweig, Germany
{whitty, ernst}@ida.ing.tu-bs.de

High-end applications designed for the MORPHEUS computing platform require a massive amount of memory and memory throughput to fully demonstrate MORPHEUS's potential as a high-performance reconfigurable architecture. For example, a proposed film grain noise reduction application for high definition video, which is composed of multiple image processing tasks, requires enormous data rates due to its large input image size and real-time processing constraints. To meet these requirements and to eliminate external memory bottlenecks, a bandwidth-optimized DDR-SDRAM memory controller has been designed for use with the MORPHEUS platform and its Network On Chip interconnect. This paper describes the controller's design requirements and architecture, including the interface to the Network On Chip and the two-stage memory access scheduler, and presents relevant experiments and performance figures.

Reducing Energy by Exploring Heterogeneity in a Coarse-grain Fabric

Gayatri Mehta, Colin J. Ihrig and Alex K. Jones
University of Pittsburgh
{gmehta,cihrig,akjones}@ece.pitt.edu

This paper explores the impact of heterogeneity on energy consumption in a stripe-based coarse-grain fabric architecture. We examine the benefit of replacing 25-50% of functional blocks with dedicated vertical routes in the fabric. Additionally, we reduce the number of operations supported by the functional units from 23 to 16, 10 and 8. To assist in testing and examining the impact of these different architectures on energy consumption, an automation process was created to automatically generate fabric instances based on a Fabric Instance Model (FIM) written in XML. The FIM is also used as an input parameter to our heuristic mapper in order to program a particular fabric instance. Upon testing these instances, we found that the fabric with ALUs supporting 10-operations and using an 8:1 interconnect with 33% of the functional units replaced with dedicated pass gates provided the best energy versus mappability tradeoff, resulting in a 32% energy improvement and a 47% area savings over the baseline fabric with ALUs supporting 23-operations and using an 8:1 interconnect without dedicated vertical routes.

Reusable Context Pipelining for Low Power Coarse-Grained Reconfigurable Architecture

Yoonjin Kim and Rabi N. Mahapatra
Embedded Systems and Co-Design Group, Dept. of Computer Science
Texas A&M University, College Station, TX 77843
{ykim, rabi}@cs.tamu.edu

Coarse-grained reconfigurable architectures (CGRA) require many processing elements and a configuration memory unit (configuration cache) for reconfiguration of the ALU array elements. This structure consumes significant amount of power. Power reduction during reconfiguration is necessary for the reconfigurable architecture to be used as a competitive IP core in embedded systems. In this paper, we propose a power-conscious reusable context pipelining architecture for CGRA that efficiently reduces power consumption in configuration cache without performance degradation. Experimental results show that the proposed approach saves up to 57.97% of the total power consumed in the configuration cache with reduced configuration cache size compared to the previous approach.

Design of Steering Vectors for Dynamically Reconfigurable Architectures

Nick A. Mould[1], Brian F. Veale[2], John K. Antonio[3], Monte P. Tull[1] and John R. Junger[1]

[1]School of Electrical and
Computer Engineering
University of Oklahoma
Norman, OK 73019 USA
{nick_mould, tull, jjunger}
@ou.edu

[2]Austin, TX 78729 USA
veale@acm.org

[3]School of Computer Science
University of Oklahoma
Norman, OK 73019 USA
antonio@ou.edu

An architectural framework is studied that can perform dynamic reconfiguration. A basic objective is to dynamically reconfigure the architecture so that its configuration is well matched with the current computational requirements. The reconfigurable resources of the architecture are partitioned into N slots. The configuration bits for each slot are provided through a connection to one of N independent busses, where each bus can select from among K configurations for each slot. Increasing the value of K can increase the number of configurations that the architecture can reach, but at the expense of more hardware complexity to construct the busses. Our study reveals that it is often possible for the architecture to closely track ideal desired configurations even when K is relatively small (e.g., two or four). The input configurations to the collection of busses are defined as steering vectors; thus, there are K steering vectors, each having N equal sized partitions of configuration bits. A combinatorial approach is introduced for designing steering vectors that enables the designer to evaluate trade-offs between performance and hardware complexity associated with the busses.

Heuristics for Scheduling Reconfigurable Devices with Consideration of Reconfiguration Overheads

Josef Angermeier and Jürgen Teich
Department of Computer Science 12
University of Erlangen-Nuremberg, Germany
Am Weichselgarten 3, 91058 Erlangen
{angermeier, teich}@informatik.uni-erlangen.de

When using dynamically and partially reconfigurable FPGAs in embedded systems, the scheduler needs to fulfill area and time requirements for each task. While those demands are already well studied in literature, another characteristic peculiarity of reconfigurable systems has been rather neglected: the reconfiguration overhead. However, scheduling algorithms considering the exclusive access to the reconfiguration port can improve the latency of obtained schedules considerably. In this paper, we present new scheduling heuristics and a methodology to compare approaches which take into consideration the reconfiguration overheads with those which disregard them. Furthermore, our experimental results give insight into possible performance increases and present problem instances for which the reconfiguration latency is negligible.

Scheduling Reconfiguration at Runtime on the TRIPS Processor

Justin Teller
Department of Electrical and
Computer Engineering
The Ohio State University
Columbus, Ohio 43210C1272, USA
teller.9@osu.edu

Füsun Özgüner
Department of Electrical and
Computer Engineering
The Ohio State University
Columbus, Ohio 43210C1272, USA
ozguner@ece.osu.edu

Robert Ewing
Air Force Research Laboratory
Wright-Patterson Air Force Base
Dayton, Ohio 45433, USA
Robert.Ewing@wpafb.af.mil

We address the problem of scheduling parallel applications onto Heterogeneous Chip Multi-Processors (HCMPs) containing reconfigurable processing cores. To model reconfiguration, we introduce the novel Mutually Exclusive Processor Groups reconfiguration model, which captures many different modes of reconfiguration. The paper continues by proposing the Heterogeneous Earliest Finish Time with Mutually Exclusive Processor Groups (HEFTMEG) scheduling heuristic that uses our new reconfiguration model; at compile-time, HEFT-MEG schedules reconfigurations to occur at runtime, with the goal of choosing the most efficient configuration for different application phases. Scheduling reconfiguration to occur at runtime with HEFT-MEG improves the performance of GPS Acquisition, a software radio application, by about 23%, compared to the best single-configuration schedule on the same hardware.

Perimeter Quadrature-based Metric for Estimating FPGA Fragmentation in 2D HW Multitasking

J. Septién, D. Mozos, H. Mecha, J. Tabero and M. A. García de Dios
Universidad Complutense de Madrid
jseptien@dacya.ucm.es

This paper explains a new technique to estimate free area fragmentation, when hardware multitasking is being considered on a 2D FPGA. The importance of a good fragmentation metric is stated, as well its use as allocation heuristic and as defragmentation alarm. We present a new fragmentation metric based on the relative quadrature of the free area perimeter, showing examples of how it behaves with one or several holes and also with islands. Finally, we show how it can be used as cost function in a location selection heuristic, each time a task is loaded in the FPGA. Experimental results show that though it maintains a low complexity, this metric behaves better than most of the previous ones, discarding a lower amount of computing volume when the FPGA supports a heavy task load.

A 937.5ns Multi-context Holographic Configuration with A 30.75μs Retention Time

Mao Nakajima, Daisaku Seto and Minoru Watanabe
Electrical and Electronic Engineering
Shizuoka University
3-5-1 Jyohoku, Hamamatsu, Shizuoka 432-8561, Japan
Email: tmwatan@ipc.shizuoka.ac.jp

Optically reconfigurable gate arrays (ORGAs) have been developed to realize a large virtual gate count by adding a holographic memory onto a programmable gate array VLSI. However, in ORGAs, although a large virtual gate count can be realized by exploiting the large capacity storage capability of a holographic memory, the actual gate count, which is the gate count of a programmable gate array VLSI, is still important to increase the instantaneous performance. Nevertheless, in previously proposed ORGAVLSIs, the static configuration memory to store a single configuration context consumed a large implementation area of the ORGA-VLSIs and prevented the realization of large-gate-count ORGA-VLSIs. Therefore, to increase the gate density, a dynamic optically reconfigurable gate array (DORGA) architecture has been proposed. It uses the junction capacitance of photodiodes as dynamic memory, thereby obviating the static configuration memory. However, to date, although only a 1.83–.89 ms single-context holographic configuration and a retention time of 3.49– 5.61 s of DORGA architecture have been confirmed, the performance at nanosecond-scale reconfiguration with a multi-context DORGA architecture has never been analyzed. Therefore, this paper presents the experimental result of a 937.5 ns multi-context holographic configuration and a 30.75 μs retention time of DORGA architecture. The advantages of this architecture are discussed in relation to the results of this study.

An Adaptive and Scalable Multiprocessor System for Xilinx FPGAs Using Minimal Sized Processor Cores

Alexander Klimm, Lars Braun and Jürgen Becker
Universität Karlsruhe (TH)
Engesserstr. 5, 76131 Karlsruhe, Germany
klimm,braun,becker@itiv.uni-karlsruhe.de

In embedded systems, especially in multi-purpose platforms, the need for computational power can vary greatly. This depends mainly on the type of applications running on the platform as well as on limitations such as real-time constraints or the amount of data being processed. These factors can change during lifetime of the system and even during runtime. A highly adaptive computing platform is desirable, providing exactly the computational power and speed needed for any particular task at any particular time using minimal resources. This contribution discusses a highly scalable multiprocessor system composed of minimal sized processor cores. The proposed architecture is intended for computational intensive applications such as on-line routing and placement on FPGAs or applications in the cryptographic domain. While offering a lot of leeway in computational power it uses minimal resources on an FPGA, thus freeing hardware resources for applications running in parallel to the multiprocessor system.

A Design Flow Tailored for Self Dynamic Reconfigurable Architecture

Fabio Cancare', Marco D. Santambrogio and Donatella Sciuto
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5
20133 Milano, Italy
{santambr,rana,sciuto}@elet.polimi.it

Dynamic reconfigurable embedded systems are gathering, day after day, an increasing interest from both the scientific and the industrial world. The need of a comprehensive tool which can guide designers through the whole implementation process is becoming stronger. In this paper the authors introduce a new design framework which amends this lack. In particular the paper describes the entire low level design flow onto which the framework is based.

Run-Time Reconfigurable Adaptive Multilayer Network-on-Chip for FPGA-Based Systems

Michael Hübner[1], Lars Braun[1], Diana Göhringer[2] and Jürgen Becker[1]
[1] ITIV, Universität Karlsruhe (TH), Germany
{huebner, braun, becker} @itiv.uni-karlsruhe.de
[2] FGAN-FOM, Ettlingen Karlsruhe, Germany
goehringer@fom.fgan.de

Since the 1990s reusable functional blocks, well known as IP-Cores, were integrated on one silicon die. These Systems-on-Chip (SoC) used a bus-based system for intermodule communication. Technology and flexibility issues forced to introduce a novel communication system called Network-on-Chip (NoC). Around 1999 this method was introduced and until then it is investigated by several research groups with the aim to connect different IP-Blocks through an effective, flexible and scalable communication network. Exploiting the flexibility of FPGAs, the run-time adaptivity through run-time reconfiguration, opens a new area of research by considering dynamic and partial reconfiguration. This paper presents an approach for exploiting dynamic and partial reconfiguration with Xilinx Virtex-II FPGAs for a multi-layer Network-on-Chip and the related techniques for adapting the network while run-time to the requirements of an application.

Flexible Parallel Pipeline Network-on-Chip Based on Dynamic Packet Identity Management

Faizal A. Samman, Thomas Hollstein and Manfred Glesner,
Technische Universität Darmstadt
Institute of Microelectronic Systems
Karlstr. 15. Darmstadt, Hessen D-64283
faizal.samman, thomas.hollstein, glesner@mes.tu-darmstadt.de

This paper presents a network-on-chip (NoC) with flexible infrastructure based on dynamic wormhole packet identitity mapping management. The NoCs are developed based on VHDL-modular approach to support the design flexibility. The on-chip router switches packets using wormhole packet switching method with synchronous parallel pipeline technique. Contention-free routing algorithms and dynamic packet identity management are proposed to support wire-through-share methodology and identity-slot division multiple access technique. The on-chip routers are also equipped with packet flow control and automatic injection rate control mechanism to avoid packet-loss, when the NoC is congested. Some static and adaptive routing algorithms are implemented in the NoC to observe the performance of the NoC over some selected network traffic patterns and the logic consumption (using CMOS standard-cell library). Area overheads to implement several adaptive routings over static routing algorithm are less than 9%. Our NoC guarantees in-order and lossless message flits delivery.

Implementation of a Reconfigurable Data Protection Module for NoC-based MPSoCs

Leandro Fiorin, Slobodan Lukovic
ALaRI, Faculty of Informatics
University of Lugano
Lugano, Switzerland
{fiorin, lukovics}@alari.ch

Gianluca Palermo
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Milano, Italy
gpalermo@elet.polimi.it

Security issues are emerging to be a basic concern in modern SoC development. Since in the field of on-chip interconnections the security problem continues to remain mostly an unexplored topic, this paper proposes a novel technique for data protection that uses the communication subsystem as basis. The proposed architecture works as a firewall managing the memory accesses on the basis of a lookup table containing the access rights. This module, called Data Protection Unit (DPU), has been designed for MPSoC architectures and integrated in the Network Interfaces near the shared memory. We implement the DPU inside an MPSoC architecture on FPGA and we add features to the module to be aware of dynamic reconfiguration of the system software. Starting from a general overview of our design down to components' structure, we introduce the place and the role of the DPU module inside the system for a reconfigurable secure implementation of a MPSoC on FPGA. The description of the DPU concept, its implementation, and integration into the system are described in detail. Finally, the architecture is fully implemented on FPGA and tested on a Xilinx Virtex-II Pro board.

HARPE: A Harvard-based Processing Element Tailored for Partial Dynamic Reconfigurable Architectures

Alessio Montone, Vincenzo Rana, Marco D. Santambrogio and Donatella Sciuto
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5
20133 Milano, Italy
{santambr,rana,sciuto}@elet.polimi.it, alessio.montone@dresd.org

Aim of this paper is to propose a Reconfigurable Processing Element based on a Harvard Architecture, called HARPE. HARPE's architecture includes a MicroBlaze soft-processor in order to make HARPEs deployable also on devices not having processors on silicon die. In such a context, this work also introduces a novel approach for the management of processor data memory. The proposed approach allows the individual management of data and the dynamic update of the memory, thus making it possible to define Partially Dynamical Reconfigurable Multi Processing Element Systems, that consist of several master (e.g., soft-processors, hard-processors or HARPE cores) and slave components. Finally, the proposed methodology enables the possibility of creating a system in which both HARPEs and their memories (data and code) can be separately configured at run time with a partial configuration bitstream, in order to make the whole system more flexible with respect to changes occurring in the external environment.

Runtime Adaptive Multi-Processor System-on-Chip: RAMPSoC

Diana Göhringer[1], Michael Hübner[2], Volker Schatz[1] and Jürgen Becker[2]
FGAN-FOM, Germany[1]
ITIV, Universität Karlsruhe (TH), Germany[2]
{dgoehringer, schatz}@fom.fgan.de
{huebner, becker}@itiv.uni-karlsruhe.de

Current trends in high performance computing show, that the usage of multiprocessor systems on chip are one approach for the requirements of computing intensive applications. The multiprocessor system on chip (MPSoC) approaches often provide a static and homogeneous infrastructure of networked microprocessor on the chip die. A novel idea in this research area is to introduce the dynamic adaptivity of reconfigurable hardware in order to provide a flexible heterogeneous set of processing elements during run-time. This extension of the MPSoC idea by introducing run-time reconfiguration delivers a new degree of freedom for system design as well as for the optimized distribution of computing tasks to the adapted processing cells on the architecture related to the changing application requirements. The "computing in time and space" paradigm and the extension with the new degree of freedom for MPSoCs will be presented with the RAMPSoC approach described in this paper.

Algorithmic Skeletons for the Design of Partially Reconfigurable Systems

Florian Dittmann, Stefan Frank and Simon Oberthür
Heinz Nixdorf Institute, University Paderborn
Fürstenallee 11, 33102 Paderborn, Germany
email: {roichen,sfrank,oberthuer}@upb.de

Designing reconfigurable systems that beneficially exploit the spatial and temporal domain is a cumbersome task hardly supported by current design methods. In particular, if we aim to bridge the gap between application and reconfigurable substrate, we require concrete concepts that allow for utilizing the inherent parallelism and adaptiveness of reconfigurable devices. We propose algorithmic skeletons as sophisticated technique therefore. Algorithmic skeletons are programming templates for the parallel computing domain and therefore separate the structure of a computation from the computation itself. Hence, they offer a seminal means to extract temporal and spatial characteristics of an application, which can be used to make reconfigurability explicit. In this work, we show the conceptual background as well as a concrete implementation means of the method.

Balancing Reconfigurable Data Path Resources According to Application Requirements

Mateus Beck Rutzig, Antonio Carlos S. Beck and Luigi Carro
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre/RS - Brazil
{mbrutzig, caco, carro}@inf.ufrgs.br

Processor architectures are changing mainly due to the excessive power dissipation and the future break of Moore's law. Thus, new alternatives are necessary to sustain the performance increase of the processors, while still allowing low energy computations. Reconfigurable systems are strongly emerging as one of these solutions. However, because they are very area consuming and deal with a large number of applications with diverse behaviors, new tools must be developed to automatically handle this new problem. This way, in this work we present a tool aimed to balance the reconfigurable area occupied with the performance required by a given application, calculating the exact size and shape of a reconfigurable data path. Using as case study a tightly coupled reconfigurable array and the Mibench Benchmark set, we show that the solution found by the proposed tool saves four times area in comparison with the non-optimized version of the reconfigurable logic, with a decrease of only 5.8% on average of its original performance. This way, we open new applications for reconfigurable devices as low cost accelerators.

A Hybrid Processing Element based Reconfigurable Architecture for Hashing Algorithms

Deepak Sreedharan
Dept. of Electrical and Computer Engineering
Reconfigurable Computing Lab
University of Arizona
Tucson, AZ – 85721, USA
deepaks@email.arizona.edu

Ali Akoglu
Dept. of Electrical and Computer Engineering
Reconfigurable Computing Lab
University of Arizona
Tucson, AZ – 85721, USA
akoglu@ece.arizona.edu

Given the high computation demand for cryptography and hashing algorithms there is a need to develop flexible and high performance architectures. This paper proposes a methodology to derive processing elements as a starting point for the state-of-the-art reconfigurable computing and presents a case-study to show that application-specific reconfigurable computing has performance benefits close to fully-custom designs in addition to the intended reconfigurablity. We use hashing algorithms as a case study to propose a novel application-specific reconfigurable architecture based on a balanced mixture of coarse and fine grained processing elements with a tuned interconnect structure. For that purpose we introduce a methodology to derive hybrid grained processing elements and expose both fine and coarse grain parallelism based on a new common and recurring computation pattern extraction tool. After extracting the recurring patterns between SHA-1 and MD5 algorithms, we derive the unified interconnect architecture tailored to the control data dependencies of both the algorithms. That way the amount of reconfiguration on the proposed architecture when switching between the two algorithms is minimized. The proposed reconfigurable architecture is synthesized using the Synopsys design compiler targeted at TSMC 250 nm libraries. We compare its performance with ASIC technology on SHA-1 and MD5 algorithms. Results show that the proposed architecture which is reconfigurable between the two hashing algorithms has frequency of operation close to ASIC implementation of the individual algorithms for iterative and pipelined versions and results with 35% savings in area.

A Self-adaptive On-line Task Placement Algorithm for Partially Reconfigurable Systems

Yi Lu, Thomas Marconi, Georgi Gaydadjiev, Koen Bertels and Roel Meeuws
Computer Engineering Lab., TU Delft, The Netherlands
{yilu, thomas, georgi, rmeeuws}@ce.et.tudelft.nl, k.l.m.bertels@tudelft.nl

With the arrival of partial reconfiguration technology, modern FPGAs support swapping tasks in or out individually at run-time without interrupting other tasks running on the same FPGA. Although, implementing this feature achieves much better flexibility and device utilization, the challenge remains to quickly and efficiently place tasks arriving at run-time on such partially reconfigurable systems. In this paper, we propose an algorithm to handle this on-line, run-time task placement problem. By adding logical constraints on the FPGA and introducing our resources management solution, the simulation results show our algorithm has better overall performances compared with previous reported methods in terms of task rejection number, placement quality and execution time.

Configuration Scheduling Using Temporal Locality and Kernel Correlation

Santheeban Kandasamy, Andrew Morton and Wayne M. Loucks
Electrical and Computer Engineering
University of Waterloo
Waterloo, Ontario, Canada
fs2kandas,arrmorton,wmloucksg@uwaterloo.ca

This paper examines runtime decisions to configure hardware accelerators or execute in software. Traditionally, reconfigurable FPGAs are reconfigured on-demand with the hardware accelerator, as it is needed by the application. If the software kernel which the hardware kernel replaces is available too, then more sophisticated decision making on reconfigurations may lead to improved execution time and reduced power consumption. The temporal locality algorithm is proposed for applications where individual kernels dominate during differing execution modes. The kernel correlation algorithm is proposed for applications where sequences of kernels are invoked in regular patterns. SystemC simulation is used to compare these two scheduling algorithms against the on-demand policy. Both timing and power consumption results are presented. They indicate that a fairly large reconfiguration time is required for configuration scheduling to be beneficial.

Implementation of a Floating-point Matrix-vector Multiplication on a Reconfigurable Architecture

Fabio Garzia Department of Computer Systems Tampere University of Technology Tampere, Finland fabio.garzia@tut.fi	Claudio Brunelli Department of Computer Systems Tampere University of Technology Tampere, Finland claudio.brunelli@tut.fi
Davide Rossi DEIS Università degli studi di Bologna Bologna, Italy davide.rossi@studio.unibo.it	Jari Nurmi Department of Computer Systems Tampere University of Technology Tampere, Finland jari.nurmi@tut.fi

This paper describes the implementation of a floating-point 4x4 matrix-vector multiplication on a reconfigurable system. The 4x4 matrix-vector multiplication is meant to be used to perform two steps (transformation and perspective projection) of a 3D graphics application. The target system is based on a bus architecture with a general purpose core as master and the reconfigurable array as main accelerator. The system has been prototyped on a FPGA device. The matrix-vector multiplication has been successfully implemented on the reconfigurable block. Compared to the general purpose implementation it is convenient if the number of vectors to process is higher than seven. If hundreds of vectors are processed, the speed-up achievable reaches 3X.

Designing Coprocessors for Hybrid Compute Systems

Volker Hampel, Peter Sobe and Erik Maehle
University of Lübeck
Institute of Computer Engineering
{hampel, sobe, maehle}@iti.uni-luebeck.de

A Hybrid Compute System (HCS) combines standard CPUs and reconfigurable devices, usually FPGAs, in one system. Recently, these systems have become more attractive again, due to a closer and hence faster coupling of both computational components. From our work with several designs for the same application, we have found the communication between a CPU and a FPGA-based coprocessor to relate either to pipelining or to a bulk-wise transfer with buffered data processing. We identify conditions which determine whether the pipelined or the buffered style should be used in a design. A Reed/Solomon encoding coprocessor has been implemented for each of the communication architectures to serve as an example of how these conditions materialize and how they influence the performance.

A Rapid Prototyping Environment for High-speed Reconfigurable Analog Signal Processing

Joachim Becker, Fabian Henrici, Stanis Trendelenburg and Yiannos Manoli
Department of Microsystems Engineering (IMTEK), University of Freiburg, Germany
Email: jmbeck@imtek.de

This paper reports on a rapid-prototyping platform for high-frequency continuous-time analog filters to be used in communication front-ends. A Field Programmable Analog Array (FPAA) is presented, which implements a unique hexagonal topology of 55 tunable OTAs for reconfigurable instantiation of G_m-C filters in a 0.13 μm CMOS technology. With a maximum unity-gain bandwidth of 186 MHz, it is the first analog array to achieve a bandwidth, which allows processing of intermediate frequencies used in communication systems. In addition to the intuitive manual mapping of analog filters to the chip structure, a genetic algorithm with hardware in the loop is used for automated synthesis of transfer functions.

A Framework for Dynamic 2D Placement on FPGAs

Christian Schuck, Matthias Kühnle, Michael Hübner, Jürgen Becker
ITIV - Universität Karlsruhe (TH) – Germany
{schuck, kuehnle, huebner, becker}@itiv.uni-karlsruhe.de

The presented paper describes an approach of dynamic positioning of functional building blocks on Virtex (Xilinx) FPGAs. The modules can be of a variable rectangular shape. Further, the on-chip location of the area to be reconfigured can be freely chosen, so that any module can be placed anywhere within the defined dynamic region of the FPGA. Thus the utilization of the chip area can be optimized, which in turn reduces e.g. costly area and power consumption. This paper describes a runtime system and the necessary framework, which is able to manage the reconfigurable area. Further it shows how a NoC approach can be applied to shorten wire lengths for communication. This will in turn save routing resources and potentially increases clock speed.

Efficient and Accurate FPGA-based Simulator for Molecular Dynamics

Eunjung Cho and Anu G. Bourgeois
Dept. of Computer Science, Georgia State University, Atlanta, GA 30303
echo@student.gsu.edu, anu@cs.gsu.edu
and
José A. Fernández-Zepeda
Dept. of Computer Science, CICESE, Ensenada, B.C. 22860, MEXICO
fernan@cicese.mx

A Molecular Dynamics (MD) system is defined by the position and momentum of particles and their interactions. Solving the dynamics numerically and evaluating the interaction is computationally expensive even for a small number of particles in the system. We are focusing on long-ranged interactions, since the calculation time is $O(N^2)$ for an N particle system. There are many existing algorithms aimed at reducing the calculation time of MD simulations. Among the existing algorithms, Multigrid (MG) method reduces $O(N^2)$ calculation time to $O(N)$ time while still achieving reasonable accuracy. Another movement to achieve much faster calculation time is running MD simulation on special purpose processors and customized hardware with ASICs or an FPGAs. In this paper, we design and implement FPGA-based MD simulator with an efficient MG method.

Partial Run-time Reconfiguration of FPGA for Computer Vision Applications

Manish Birla[1] and K. N. Vikram[2]
[1]Siemens Corporate Technology, [1]Genesis Microchip,
Bangalore, India Bangalore, India
Manish.Birla@siemens.com vikramkn@ieee.org

FPGAs are a popular platform for implementation of computer vision applications, due to the inherent parallelism present in the programmable fabric. In addition to hardware acceleration through parallelization, modern FPGAs are also dynamically reconfigurable, thereby adding an additional dimension to the mapping of algorithms to hardware. Among the various uses for run-time reconfiguration, one application is the time multiplexing of limited hardware resources to carry out a considerably complex computation. This paper presents the use of partial reconfiguration for time multiplexing computations in the implementation of a computer vision application - human detection. The results obtained from the implementation of a proof-of-concept prototype on a Xilinx Virtex-4 FPGA are also presented.

Physical Layer Design Automation for RFID Systems

Shenchih Tung and Alex K. Jones
University of Pittsburgh
shst8@pitt.edu, akjones@ece.pitt.edu

While RFID is starting to become a ubiquitous technology, the variation between different RFID systems still remains high. This paper describes a design automation flow for fast implementation of the physical layer component of new RFID systems. Physical layer features are described using waveform features, which are used to automatically generate physical layer encoding and decoding hardware blocks. We present automated implementations of five protocols related to RFID including Manchester encoding for ultra high frequency (UHF) active tags, Pulse Interval encoding (PIE) for UHF passive tags, and Modified Miller encoding for lower frequency RFID tags. We have targeted reconfigurable devices to allow changes in the design and compared these implementations with a standard cell ASIC target.

A Self Adaptive Interfacing Concept for Consumer Device Integration into Automotive Entities

Benjamin Glas, Alexander Klimm, Oliver Sander, Klaus D. Müller-Glaser and Jürgen Becker
Institut für Technik der Informationsverarbeitung, Universität Karlsruhe (TH)
Email: {glas,klimm,sander,kmg,becker}@itiv.uni-karlsruhe.de

Innovation in today's automobiles is mainly driven by electronic development. Standard infotainment devices delivered by car manufacturers offer services such as navigation, television, DVD, telephony and more. Especially in the last few years a trend can be recognized extending from these OEM specific solutions towards integration of consumer electronic devices such as iPods, cell phones or the like. An approach solely based on software adjustments might not be sufficient to solve the problem of car to mobile device communication. Reconfigurable hardware offers a unique possibility of adapting to a given problem in hardware as well as software structures. The feature of partial dynamic reconfiguration in particular can contribute significantly to the flexibility of the overall system. Any adaptation of the running system should be automatic and going unnoticed by the user. This contribution depicts an approach that utilizes reconfigurable hardware for connecting consumer electronics to an automobile. Detection of protocols, loading of appropriate hardware interfaces, as well as integration of offered or requested services are the essential steps presented in this paper.

Modelling and Exploration of a Reconfigurable Array Using SystemC TLM

Henrik Svensson, Thomas Lenart and Viktor Öwall
Department of Electrical and Information Technology
Lund University
Box 118, SE-221 00 Lund, Sweden
{henrik.svensson,thomas.lenart,viktor.owall}@eit.lth.se

This paper presents a coarse-grained reconfigurable architecture based on an array of processing and memory cells. Memory cells are distributed and placed close to processing cells to reduce memory bottlenecks. Processing cells are instruction set processors with enhanced performance for communication-intensive inner loops. Processor communication is performed using a self-synchronizing protocol that simplifies algorithm mapping and manages unpredictable time variations. The reconfigurable architecture is described as a scalable and parameterizable SystemC transaction level model, which allows rapid architectural exploration. Our exploration environment SCENIC is used to setup scenarios, control the simulation models and to extract performance data during simulation. A case study demonstrates different implementations of a filter algorithm, and how exploration is used to tune and optimize for execution time, latency, or used resources.

Accelerating Matrix Decomposition with Replications

Yi-Gang Tai, Chia-Tien Dan Lo and Kleanthis Psarris
Department of Computer Science
University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249, USA
{ytai,danlo,psarris}@cs.utsa.edu

Matrix decomposition applications that involve large matrix operations can take advantage of the flexibility and adaptability of reconfigurable computing systems to improve performance. The benefits come from replication, which includes vertical replication and horizontal replication. If viewed on a space-time chart, vertical replication allows multiple computations executed in parallel, and horizontal replication renders multiple functions on the same piece of hardware. In this paper, the reconfigurable architecture that supports replications for matrix decomposition applications on reconfigurable computing systems is described, and issues including the comparison of algorithms on the system and data movement between the internal computation cores and the external memory subsystem are addressed. A prototype of such a system is implemented to prove the concept. It is expected to improve the performance and scalability of matrix decomposition involving large matrices.

A Coarse Grained and Hybrid Reconfigurable Architecture with Flexible Noc Router for Variable Block Size Motion Estimation

Ruchika Verma	Ali Akoglu
Department of ECE	Department of ECE
University of Arizona	University of Arizona
Tucson, Az-85721	Tucson, Az-85721
15206265149	15206265149
ruchikav@ece.arizona.edu	akoglu@ece.arizona.edu

This paper proposes a novel application-specific hybrid coarse-grained reconfigurable architecture with a flexible network on chip (NoC) mechanism. Architecture supports variable block size motion estimation (VBSME) with much less resources than ASIC based and coarse grained reconfigurable architectures. The intelligent NoC router supports full search motion estimation algorithm as well as other fast search algorithms like diamond, hexagon, big hexagon and spiral. Our model is a hierarchical hybrid processing element based 2D architecture which supports reuse of reference frame blocks between the processing elements through NoC routers. This reduces the transactions from/to the main memory. Proposed architecture is designed with Verilog-HDL description and synthesized by 90 nm CMOS standard cell library. Results show that our architecture reduces the gate count by 7x compared to its ASIC counterpart that only supports full search method. Moreover, the proposed architecture operates at a frequency comparable to ASIC based implementation to sustain 30fps. Our approach is based on a simple design which utilizes a high-level of parallelism with an intensive data reuse. Therefore, proposed architecture supports run-time reconfiguration for any block size and for any search pattern depending on the application requirement.

A CAM-based Intrusion Detection System for Single-packet Attack Detection

Ying Yu	Raymond R. Hoare	Alex K. Jones
Marvell Semiconductor	Concurrent EDA, LLC	University of Pittsburgh
yingyu@marvell.com	rayhoare@concurrenteda.com	akjones@ece.pitt.edu

Many telecommunications devices such as network switches contain content addressable memories (CAMs) for uses such as routing tables. CAMs, a class of associative memories, contain considerable logic for various forms of content matching and can be considered a class of reconfigurable logic engines. This paper demonstrates how a commercial ternary CAM and traditional RAM can be used with minimal additional logic to implement over 90% of the Snort 2.0 intrusion detection system (IDS) at line speeds of or exceeding 1 Gbs. In addition to simple matching techniques, sophisticated matching operations required by Snort can be implemented by levering an iterative approach that leverages a post processing action RAM. Additionally, a novel range encoding algorithm allows range matching required in the CAM for which other encodings either exceed the width provided by a CAM entry, or require excessive number of CAM entries to be scalable. The system was implemented for verification and performance evaluation in cycle accurate simulation using SystemC.

A Neocortex Model Implementation on Reconfigurable Logic with Streaming Memory

Christopher N. Vutsinas, Tarek M. Taha and Kenneth L. Rice
Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
{cvutsin, tarek, krice}@clemson.edu

In this paper we study the acceleration of a new class of cognitive processing applications based on the structure of the neocortex. Our focus is on a model of the visual cortex used for image recognition developed by George and Hawkins. We propose techniques to accelerate the algorithm using reconfigurable logic, specifically a streaming memory architecture utilizing available off-chip memory. We discuss the design of a streaming memory access unit enabling a large number of processing elements to be placed on a single FPGA thus increasing throughput. We present an implementation of our approach on a Cray XD1 and discuss possible extensions to further increase throughput. Our results indicate that using a two FPGA design with streaming memory gives a speedup of 71.9 times over a purely software implementation.

A Customizable FPGA IP Core Implementation of a General Purpose Genetic Algorithm Engine

Pradeep Fernando, Hariharan Sankaran, Srinivas Katkoori
CSE Department,University of South Florida
4202 E Fowler Avenue ENB 118
Tampa FL 33620 USA
katkoori@cse.usf.edu

Didier Keymeulen, Adrian Stoica, Ricardo Zebulum, Ramesham Rajeshuni
Jet Propulsion Laboratories
4800 Oak Grove Drive
Pasadena, CA, 91109 USA
didier.keymeulen@jpl.nasa.gov

Hardware implementation of Genetic Algorithms (GA) is gaining importance as genetic algorithms can be effectively used as an optimization engine for real-time applications (for e.g., evolvable hardware). In this work, we report the design of an IP core that implements a general purpose GA engine which has been successfully synthesized and verified on a Xilinx Virtex II Pro FPGA Device (XC2VP30). The placed and routed IP core has an area utilization of only 16% and clock period of 2.2ns (~450MHz). The GA core can be customized in terms of the population size, number of generations, cross-over and mutation rates, and the random number generator seed. The GA engine can be tailored to a given application by interfacing with the application specific fitness evaluation module as well as the required storage memory (to store the current and new populations). The core is soft in nature i.e., a gate-level netlist is provided which can be readily integrated with the user's system.

FPGA Implementation of a Vehicle Detection Algorithm Using Three-Dimensional Information

Masanori Hariyama, Kensaku Yamashita and Michitaka Kameyama
Graduate School of Information Sciences, Tohoku University
Aoba 6-6-05, Aramaki, Aoba, Sendai, Miyagi,980-8579, Japan
Email: {hariyama@, yamashita@kameyama., kameyama}@ecei.tohoku.ac.jp

This paper presents a vehicle detection algorithm using 3-dimensional(3-D) information and its FPGA implementation. For high-speed acquisition of 3-D information, feature-based stereo matching is employed to reduce search area. Our algorithm consists of some tasks with high degree of column-level parallelism. Based on the parallelism, we propose area-efficient VLSI architecture with local data transfer between memory modules and processing elements. Images are equally divided into blocks with some columns, and a block is allocated to a PE. Each PE performs the processing in parallel. The proposed architecture is implemented on FPGA (Altera Stratix EP1S40F1020C7). For specifications of image size 640 × 480, 100 frames/sec, and operating frequency 100MHz, only 11,000 logic elements ($< 30\%$) are required for 30PEs.

Defect Tolerance of Holographic Configurations in ORGAs

Kouji Shinohara and Minoru Watanabe
Electrical and Electronic Engineering
Shizuoka University
3-5-1 Jyohoku, Hamamatsu, Shizuoka 432-8561, Japan
Email: tmwatan@ipc.shizuoka.ac.jp

Optically reconfigurable gate arrays (ORGAs) have been developed as a type of multi-context field programmable gate array to realize fast reconfiguration and numerous reconfiguration contexts. Along with such advantages, ORGAs have high defect tolerance. They consist simply of a holographic memory, a laser diode array, and a gate array VLSI. Even if a gate array VLSI includes defective areas, the perfectly parallel programmable capability of ORGAs enables perfect avoidance of those defective areas through alternative use of other non-defective areas. Moreover, holographic memories to store contexts are known to have high defect tolerance because each bit of a reconfiguration context can be generated from the entire holographic memory. Moreover, the damage of some fraction rarely affects its diffraction pattern or a reconfiguration context. Therefore, ORGAs are very robust against component defects in devices such as laser arrays, gate arrays, and holographic memory, and are particularly useful for space applications, which require high reliability. However, to date, the degree to which defects in a holographic memory affects holographic configurations has never been analyzed. Therefore, this paper describes analysis results of defect limitations of holographic configurations.

Workshop 4

Workshop on High-Level Parallel Programming Models & Supportive Environments

HIPS 2008

A Framework for Automated Performance Bottleneck Detection

I-Hsin Chung, Guojing Cong, David Klepacki,
Simone Sbaraglia, Seetharami Seelam and Hui-Fang Wen
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598 USA
{ihchung,gcong,klepacki,ssbarag,sseelam,hfwen}@us.ibm.com

In this paper, we present the architecture design and implementation of a framework for automated performance bottleneck detection. The framework analyzes the time-spent distribution in the application and discovers the performance bottlenecks by using given bottleneck definitions. The user can query the application execution performance to identify performance problems. The design of the framework is flexible and extensible so it can be tailored based on the actual application execution environment and performance tuning requirement. To demonstrate the usefulness of the framework, we apply the framework on a practical DARPA application and show how it helps to identify performance bottlenecks. The framework helps to automate the performance tuning process and improve the user's productivity.

NOISEMINER: An Algorithm for Scalable Automatic Computational Noise and Software Interference Detection

Isaac Dooley, Chao Mei, Laxmikant Kale
University of Illinois at Urbana-Champaign
Department of Computer Science
Urbana, IL USA
{idooley2,chaomei2,kale}@uiuc.edu

This paper describes a new scalable stream mining algorithm called NOISEMINER that analyzes parallel application traces to detect computational noise, operating system interference, software interference, or other irregularities in a parallel application's performance. The algorithm detects these occurrences of noise during real application runs, whereas standard techniques for detecting noise use carefully crafted test programs to detect the problems. This paper concludes by showing the output of NOISEMINER for a real-world case in which 6 ms delays, caused by a bug in an MPI implementation, significantly limited the performance of a molecular dynamics code on a new supercomputer.

Towards Scalable Performance Analysis and Visualization Through Data Reduction

Chee Wai Lee, Celso Mendes and Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
cheelee@uiuc.edu, cmendes@uiuc.edu, kale@cs.uiuc.edu

Performance analysis tools based on event tracing are important for understanding the complex computational activities and communication patterns in high performance applications. The purpose of these tools is to help applications scale well to large numbers of processors. However, the tools themselves have to be scalable. As application problem sizes grow larger to exploit larger machines, the volume of performance trace data generated becomes unmanagable especially as we scale to tens of thousands of processors. Simultaneously, at analysis time, the amount of information that has to be presented to a human analyst can also become overwhelming.

This paper investigates the effectiveness of employing heuristics and clustering techniques in a scalability framework to determine a subset of processors whose detailed event traces should be retained. It is a form of compression where we retain information from processors with high signal content.

We quantify the reduction in the volume of performance trace data generated by NAMD, a molecular dynamics simulation application implemented using CHARM++. We show that, for the known performance problem of poor application grainsize, the quality of the trace data preserved by this approach is sufficient to highlight the problem.

Optimization via Reflection on Work Stealing in TBB

Arch Robison, Michael Voss and Alexey Kukanov
Intel Corporation
{Arch.Robison,MichaelJ.Voss,Alexey.Kukanov}@intel.com

Intel® Threading Building Blocks (Intel® TBB) is a C++ library for parallel programming. Its templates for generic parallel loops are built upon nested parallelism and a work-stealing scheduler. This paper discusses optimizations where the high-level algorithm inspects or biases stealing. Two optimizations are discussed in detail. The first dynamically optimizes grain size based on observed stealing. The second improves prior work that exploits cache locality by biased stealing. This paper shows that in a task stealing environment, deferring task spawning can improve performance in some contexts. Performance results for simple kernels are presented.

Model-Guided Performance Tuning of Parameter Values: A Case Study with Molecular Dynamics Visualization

Yoonju Lee Nelson[1], Bhupesh Bansal[2], Mary Hall[1], Aiichiro Nakano[2] and Kristina Lerman[1]
[1]USC/Information Sciences Institute; Marina del Rey, CA 90292
[2]University of Southern California; Los Angeles, CA 90089

In this paper, we consider the interaction between application programmers and tools that automatically search a space of application-level parameters that are believed to impact the performance of an application significantly. We study performance tuning of a large scientific application, the visualization component of a molecular dynamics simulation. The key contribution of the approach is the use of high-level programmer-specified models of the expected performance behavior of individual parameters. We use these models to reduce the search space associated with the range of parameter values and achieve results that perform close to that of a more exhaustive search of the parameter space. With this case study, we show the importance of appropriate parameter selection, with the difference between best-case and worst-case performance with a particular input data set and processor configuration of up to a factor of 17. We show that through the use of models, we can drastically reduce search time, examining only 0.3% to 5% of the search space, and usually select an implementation that is close to the best performance, within 0.84% to 15%, even though the models are not completely accurate.

Extended Characterization of DMA Transfers on the Cell BE Processor

Farshad Khunjush and Nikitas J. Dimopoulos
Department of Electrical and Computer Engineering
University of Victoria, Victoria, B.C., Canada,
{fkhunjus,nikitas}@ece.uvic.ca

The main contributors to message delivery latency in message passing environments are the copying operations needed to transfer and bind a received message to the consuming process/thread. A significant portion of the software communication overhead is attributed to message copying. Recently, a set of factors has been leading high-performance processor architectures toward designs that feature multiple processing cores on a single chip (a.k.a. CMP). The Cell Broadband Engine (BE) shows potential to provide high-performance to parallel applications (e.g., MPI applications). The Cell's non-homogeneous architecture along with small local storage in SPEs impose restrictions and challenges for parallel applications. In this work, we first characterize various data delivery mechanisms in the Cell BE processor; then, we propose techniques to facilitate the delivery of a message in MPI environments implemented in the Cell BE processor. We envision a cluster system comprising several cell processors each supporting several computation threads.

Programmability of the HPCS Languages: A Case Study with a Quantum Chemistry Kernel

Aniruddha G. Shet, Wael R. Elwasif, Robert J. Harrison, and David E. Bernholdt
Oak Ridge National Laboratory
PO Box 2008, Oak Ridge, TN 37831 USA
{shetag,elwasifwr,harrisonrj,bernholdtde}@ornl.gov

As high-end computer systems present users with rapidly increasing numbers of processors, possibly also incorporating attached co-processors, programmers are increasingly challenged to express the necessary levels of concurrency with the dominant parallel programming model, Fortran+MPI+OpenMP (or minor variations). In this paper, we examine the languages developed under the DARPA High-Productivity Computing Systems (HPCS) program (Chapel, Fortress, and X10) as representatives of a different parallel programming model which might be more effective on emerging high-performance systems. The application used in this study is the Hartree-Fock method from quantum chemistry, which combines access to distributed data with a task-parallel algorithm and is characterized by significant irregularity in the computational tasks. We present several different implementation strategies for load balancing of the task-parallel computation, as well as distributed array operations, in each of the three languages. We conclude that the HPCS languages provide a wide variety of mechanisms for expressing parallelism, which can be combined at multiple levels, making them quite expressive for this problem.

An Empirical Study of the Performance and Productivity of Two Parallel Programming Models

Imran Patel and John R. Gilbert
University of California Santa Barbara
Department of Computer Science
Santa Barbara, CA 93106 USA
{imran, gilbert}@cs.ucsb.edu

The choice of parallel programming models and languages is a major factor in program performance and programmer productivity in HPC. However, evaluation of their relative merits is usually done based on conventional wisdom and subjective beliefs. We present a quantitative approach to evaluate such hypotheses statistically and validate them with empirical data. We apply this approach to compare two languages representing the message passing (MPI) and shared memory programming (UPC) paradigms. We formulate hypothesis tests for comparing the performance and productivity of these two models and evaluate them with data from observational studies of HPC programmers. We present and analyze several results, some of which are statistically significant, that demonstrate the promise of empirical evaluation in HPC development.

Compiler-Controlled Extraction of Computation-Communication Overlap in MPI Applications

Dibyendu Das
IBM India Systems
Technology Lab
dibyendu.das@in.ibm.com

Manish Gupta
IBM India Systems
Technology Lab
mgupta7@in.ibm.com

Rajan Ravindran
IBM India Systems
Technology Lab
rrajan@in.ibm.com

W Shivani
NIT Warangal
shivani_w2002@yahoo.co.in

P Sivakeshava
NIT Warangal
sivakeshava@gmail.com

Rishabh Uppal
IIT Kanpur
rishabh@cse.iitk.ac.in

Exploiting Computation-Communication Overlap is a well-known requirement to speed up distributed applications. However, efforts till now use programmer expertise, rather than any automatic tool to do this. In our work we propose the use of an aggressive optimizing compiler (IBM's xl series) to automatically extract opportunities for computation communication overlap. We depend on aggressive inlining, dominator trees and SSA based use-def analyses provided by the compiler framework for exploiting such overlap. Our target is MPI applications. In such applications, we try to automatically move mpi_waits as well as split blocking mpi_send/recv to create more opportunities for overlap. Our objective is two-fold: firstly, our tool should relieve the programmer from the burden of hunting for overlap manually as much as possible, and secondly, it should aid in converging on parallel applications which benefit from such overlap quickly. These are necessary as MPI applications are quickly becoming complex and huge and manual overlap extraction is becoming cumbersome. Our early experience shows that it is not necessary that exploiting an overlap always leads to performance improvement. This corroborates with the fact that if we have an automatic tool, then, we can quickly discard such applications (or certain configurations of such applications) without spending person-hours to manually rewrite MPI applications for introducing non-blocking calls. Our initial experiments with the industry-standard NAS Parallel benchmarks show that we can get small-to-moderate improvements by utilizing overlap even in such highly tuned benchmarks. This augurs well for real-world applications that do not exploit overlap optimally.

Build to Order Linear Algebra Kernels

Jeremy G. Siek[1], Ian Karlin[2] and E. R. Jessup[2]
[1] Department of Electrical and Computer Engineering, University of Colorado
Email: jeremy.siek@colorado.edu
[2] Department of Computer Science, University of Colorado
Email: ian.karlin@colorado.edu, jessup@cs.colorado.edu

The performance bottleneck for many scientific applications is the cost of memory access inside linear algebra kernels. Tuning such kernels for memory efficiency is a complex task that reduces the productivity of computational scientists. Software libraries such as the Basic Linear Algebra Subprograms (BLAS) ameliorate this problem by providing a standard interface for which computer scientists and hardware vendors have created highly-tuned implementations. Scientific applications often require a sequence of BLAS operations, which presents further opportunities for memory optimization. However, because BLAS are tuned in isolation they do not take advantage of these opportunities. This phenomenon motivated the recent addition to the BLAS of several routines that perform sequences of operations. Unfortunately, the exact sequence of operations needed in a given situation is highly application dependent, so many more routines are needed. In this paper we present preliminary work on a domain-specific compiler that generates implementations for arbitrary sequences of basic linear algebra operations and tunes them for memory efficiency. We report experimental results for dense kernels and show speedups of 25% to 120% relative to sequences of calls to GotoBLAS and vendor-tuned BLAS on Intel Xeon and IBM PowerPC platforms.

Cache Optimization for Mixed Regular and Irregular Computations

Gudula Rünger and Michael Schwind
Department of Computer Science,
Technical University Chemnitz, Germany
{ruenger, schwi}@informatik.tu-chemnitz.de

Many large application programs suffer from a bad utilization of cache and memory hierarchy. Known transformation techniques change loop structures and/or data layout in order to improve the performance. However, those techniques are usually adapted to either regular or irregular computations. In this paper, we investigate a combination of transformations suitable for algorithms with both irregular and regular features. The transformations are composed of computational loop reordering and data storage reorderings. We show how the different transformations interact in terms of cache utilization. Experiments on different clusters show performance gains for the sequential as well as for the parallel version.

Array Optimizations for Parallel Implementations of High Productivity Languages

Mackale Joyner, Zoran Budimlić, Vivek Sarkar and Rui Zhang
Department of Computer Science, Rice University
{mjoyner, zoran, vsarkar, ruizhang}@cs.rice.edu

This paper presents an interprocedural rank analysis algorithm to automatically infer ranks of arrays in X10, a language that supports rank-independent specification of loop and array computations using regions and points. We use the rank analysis information to enable storage transformations on arrays. We evaluate a transformation that converts high-level multidimensional X10 arrays into lower-level multidimensional Java arrays, when legal to do so. Preliminary performance results for a set of parallel computational benchmarks on a 64-way AIX Power5+ SMP machine show that our optimizations deliver performance that rivals the performance of lower-level, hand-tuned code with explicit loops and array accesses, and up to two orders of magnitude faster than unoptimized, high-level X10 programs. The results show that our optimizations also help improve the scalability of X10 programs by demonstrating that relative performance improvements over the unoptimized versions increase as we scale the parallelism from 1 CPU to 64 CPUs.

Workshop 5

Workshop on Java and Components for Parallelism, Distribution and Concurrency

JAVAPDC 2008

A Survey of Concurrent Priority Queue Algorithms

Kristijan Dragičević and Daniel Bauer
IBM Research Laboratory Zurich

Algorithms for concurrent data structures have gained attention in recent years as multi-core processors have become ubiquitous. Using the example of a concurrent priority queue, this paper investigates different synchronization methods and concurrent algorithms. It covers traditional lock-based approaches, non-blocking algorithms as well as a method based on software transactional memory. Besides discussing correctness criteria for the various approaches, we also present performance results for all algorithms for various scenarios. Somewhat surprisingly, we find that a simple lock-based approach performs reasonable well, even though it does not scale with the number of threads. Better scalability is achieved by non-blocking approaches.

Investigating Software Transactional Memory on Clusters

Christos Kotselidis, Mohammad Ansari, Kimberly Jarvis, Mikel Luján, Chris Kirkham and Ian Watson
School of Computer Science
The University of Manchester
Oxford Road, M13 9PL, Manchester
{kotselidis, ansari, jarvis, mikel, chris, watson}@cs.manchester.ac.uk

Traditional parallel programming models achieve synchronization with error-prone and complex-to-debug constructs such as locks and barriers. Transactional Memory (TM) is a promising new parallel programming abstraction that replaces conventional locks with critical sections expressed as transactions. Most TM research has focused on single address space parallel machines, leaving the area of distributed systems unexplored. In this paper we introduce a flexible Java Software TM (STM) to enable evaluation and prototyping of TM protocols on clusters. Our STM builds on top of the ProActive framework and has as an underlying transactional engine the state-of-the-art DSTM2. It does not rely on software or hardware distributed shared memory for the execution. This follows the transactional semantics at object granularity level and its feasibility is evaluated with non-trivial TM-specific benchmarks.

Smart Proxies in Java RMI with Dynamic Aspect-Oriented Programming

Andrew Stevenson and Steve MacDonald
David R. Cheriton School of Computer Science
University of Waterloo, Waterloo, Ontario, Canada
{aastevenson, stevem}@uwaterloo.ca

Java RMI extends Java with distributed objects whose methods can be called from remote clients. This abstraction is supported using statically-generated proxy objects on the client to hide network communication. One limitation of Java RMI is that these proxies can only forward method calls from client to server. Many applications can benefit from smart proxies that shift some application responsibilities to the client. Normally, developers must manually implement such proxies. This paper describes an aspect-oriented approach to creating smart proxies. This approach allows a server to extend an existing proxy with new capabilities at runtime. This approach is demonstrated with two examples.

A Parallel Implementation of the Finite-Domain Time-Difference Algorithm using MPJ Express

Aamir Shafi, Aftab Hussain and Jamil Raza
NUST Institute of Information Technology
Pakistan
{aamir.shafi,aftab.husain,jamil.raza}@niit.edu.pk

This paper presents and evaluates a parallel Java implementation of the Finite-Difference Time-Domain (FDTD) method, which is a widely used numerical technique in computational electrodynamics. The Java version is parallelized using MPJ Express–a thread-safe messaging library. MPJ Express provides a full implementation of the mpiJava 1.2 API specification. This specification defines a MPI-like binding for the Java language. This paper describes our experiences of implementing the Java version of the FDTD method. Towards the end of this paper, we evaluate and compare the performance of the Java version against its C counterpart on a 32 processing core Linux cluster of eight compute nodes.

Comparison of the Conjugate Gradient of NAS Benchmark and of the Multisplitting Algorithm with the Jace Environment

Jacques M. Bahi, Raphaël Couturier and David Laiymani
University of Franche-Comte (LIFC)
IUT de Belfort-Montbéliard
Rue Engel Gros 90016 Belfort CEDEX - France
{jacques.bahi, raphael.couturier, david.laiymani}@univ-fcomte.fr

The aim of this paper is to study the behaviors of the well known Conjugate Gradient (CG) algorithm and the multisplitting algorithm in a grid context. We focus on the CG implementation used in the NAS benchmark and on the multisplitting approach which produces similar results (from a numerical point of view). By grid context we mean an architecture composed of several heterogeneous clusters geographically distributed and the use of a software environment able to tackle the heterogeneity of the nodes. Under these hypothesis, we performed a set of experiments on the Grid'5000 platform using the pure Java Jace V2 environment. We show that, by drastically reducing global synchronizations, the asynchronous multisplitting method outperforms the NAS CG implementation, with distant sites, whatever the communication protocol used.

Integrated Framework for Development and Execution of Component-based Grid Applications

Vladimir Getov
Harrow School of Computer Science, University of Westminster,
Watford Rd, Northwick Park, Harrow, London, HA1 3TP, U.K.
V.S.Getov@westminster.ac.uk

Component-based software technologies have emerged as a modern approach to software development for distributed and Grid applications. However, the lack of longer-term experience and the complexity of the target systems demand more research results in the field. This paper provides an overview of three different approaches to developing component-based Grid applications. In order to re-use legacy codes, the wrapper software approach can be adopted in its two flavours C hand-written or automatically generated wrapper code. Another approach applicable to existing object-oriented software is to componentise the code by introducing appropriate modifications. The third approach is component-oriented development from scratch. We compare and contrast the three approaches and highlight their advantages and weaknesses.

Scheduling ProActive Activities with an XPDL-based Workflow Engine

Nadia Ranaldo, Giancarlo Tretola and Eugenio Zimeo
Research Centre on Software Technology, Department of Engineering
University of Sannio 82100 Benevento C Italy
{ranaldo, tretola, zimeo}@unisannio.it

Composition represents today one of the most challenging approach to design complex software systems, especially in distributed environments. While two different views (in time and in space) are considered by researchers to compose applications, we aim at applying these two views in an integrated approach. In particular, we believe that large-scale composition in an open world is simplified by using composition in time, whereas, in the closed environment of a cluster, composition in space is more effective. In the paper, we propose the integration of a workflow engine with ProActive objects to support together the two kinds of composition at different scale-level. The paper shows the first results of this integration and highlights the plan for achieving a stronger integration with the use of GCM components.

Java ProActive vs. Fortran MPI: Looking at the Future of Parallel Java

Brian Amedro, Denis Caromel and Fabrice Huet
INRIA Sophia Antipolis C CNRS C I3S C UNSA
first.last@inria.fr

Vladimir Bodnartchouk
ActiveEon
vladimir.bodnartchouk@activeeon.com

About ten years after the Java Grande effort, this paper aims at providing a snapshot of the comparison of Fortran MPI, with Java performance for parallel computing, using the ProActive library. We first analyze some performance details about ProActive behaviors, and then compare its global performance from the MPI library. This comparative is based on the five kernels of the NAS Parallel Benchmarks. From those experiments we identify benchmarks where parallel Java performs as well as Fortran MPI, and lack of performance on others, together with clues for improvement.

A Practical Example of Convergence of P2P and Grid Computing: An Evaluation of JXTA's Communication Performance on Grid Networking Infrastructures

Gabriel Antoniu
INRIA - Rennes Research Centre
Campus de Beaulieu, 35042 Rennes cedex, France

Mathieu Jan
CEA, LIST
CEA Saclay, 91191 Gif-sur-Yvette, France

David Noblet
California Institute of Technology
1200 E California Blvd, MC 256-80, Pasadena, CA 91125, USA

As the size of today's grid computing platforms increases, the need for self-organization and dynamic reconfiguration becomes more and more important. In this context, the convergence of grid computing and peer-to-peer (P2P) computing seems natural. However, grid infrastructures are generally available as a hierarchical federation of SAN-based clusters interconnected by high-bandwidth WANs. In contrast, P2P systems usually run on the Internet, on top of random, generally flat network topologies. This difference may lead to the legitimate question of how adequate are the P2P communication mechanisms on hierarchical grid infrastructures. Answering this question is important, since it is essential to efficiently exploit the particular features of grid networking topologies in order to meet the constraints of scientific applications. This paper evaluates the communication performance of the JXTA P2P platform over high-performance SANs and WANs, for both J2SE and C bindings. We discuss these results, then we propose and evaluate several techniques able to improve the JXTA's performance on such grid networking infrastructures.

Middleware Support for Delay-Tolerant Service Provision in Disconnected Mobile Ad Hoc Networks

Yves Mahéo, Romeo Said and Frédéric Guidec
Valoria Laboratory, European University of Brittany
{Yves.Maheo,Romeo.Said,Frederic.Guidec}@univ-ubs.fr

The service-oriented approach is known to be adequate for building distributed applications for mobile ad hoc networks, and a number of successful research works using this approach in routed MANETs have already been conducted. Yet, providing support for service-oriented applications poses specific problems in disconnected MANETs where end-to-end connectivity is not guaranteed. This paper presents a prototype Java middleware support for delay-tolerant service provision in disconnected MANETs. It first describes the lower layer of this middleware, which implements a protocol for document dissemination that leverages on delay-tolerant, opportunistic and content-based communication. It then explains how service discovery and service invocation mechanisms can be implemented on top of this protocol. Finally, simulation results are presented, which show that this approach gives good results compared with an alternative approach that consists in exploiting single-hop interactions between service providers and service clients.

Workshop 6

Workshop on Nature Inspired Distributed Computing

NIDISC 2008

Distributed, Heterogeneous Resource Management Using Artificial Immune Systems

Lucas A. Wilson
Texas Advanced Computing Center (TACC)
The University of Texas at Austin
Austin, Texas 78758-4497
lwilson@tacc.utexas.edu

As high performance and distributed computing become more important tools for enabling scientists and engineers to solve large computational problems, the need for methods to fairly and efficiently schedule tasks across multiple, possibly geographically distributed, computing resources becomes more crucial. Given the nature of distributed systems and the immense numbers of resources to be managed in distributed and large-scale cluster environments, traditional centralized schedulers will not be extremely effective at providing timely scheduling information. In order to manage large numbers of resources quickly, less computationally intensive methods for scheduling tasks must be explored.

This paper proposes a novel resource management system based on the immune system metaphor, making use of the concepts in Immune Network Theory and Danger Theory. By emulating various elements in the immune system, the proposed manager could efficiently execute tasks on very large systems of heterogeneous resources across geographic and/or administrative domains. The distributed nature of the immune system is also exploited in order to allow efficient scheduling of tasks, even in extremely large environments, without the use of a centralized or hierarchical scheduler.

Resource-Centric Task Allocation in Grids with Artificial Danger Model Support

Young Choon Lee and Albert Y. Zomaya
Advanced Networks Research Group, School of Information Technologies,
The University of Sydney, NSW 2006, Australia
{yclee,zomaya}@it.usyd.edu.au

This paper addresses the problem of scheduling bagof- tasks (BoT) applications in grids and presents a novel heuristic, called the most suitable match with danger model support algorithm (MSMD) for these applications. Unlike previous approaches, MSMD is capable of efficiently dealing with BoT applications regardless of whether they are computationally or data intensive, or a mixture of both; this strength of MSMD is achieved by making scheduling decisions based on the suitability of resourceCtask matches, instead of completion time. MSMD incorporates an artificial danger model–based on the danger model in immunology–which selectively responds to unexpected behaviors of resources and applications, in order to increase fault-tolerance. The results from our thorough and extensive evaluation study confirm the superior performance of MSMD, and its generic applicability compared with previous approaches that only consider one or the other of the task requirements.

A Complex System Inspired Theoretical Formalism for Data Management in Peer-to-Peer Grids

Vijay Srinivas Agneeswaran
Distributed Information Systems Lab (LSIR),
School of Information and Communication Sciences
Swiss Federal Institute of Technology, Lausanne(EPFL), Switzerland.
D Janakiram
Distributed & Object Systems Lab,
Department of Computer Science & Engg.,
Indian Institute of Technology, Madras, India.

To our knowledge, theoretical foundations of large scale distributed systems have not been explored much in the literature. This paper provides a step in this direction. It explores two theories from complex systems in physics: Self-Organized Criticality (SOC) and Highly Optimized Tolerance (HOT) and applies the theories to data management middleware in grids. In particular, it explores replica management solutions over OpenDHT, the state of the art DHT deployment and compares the same with replica management solution of Virat, our P2P middleware. OpenDHT is based on SOC and achieves robustness but not optimality. The two layered P2P architecture of Virat combines elements of SOC and HOT to achieve robustness *and* optimal performance. We validate the theoretical conjecture with extensive performance studies over Intranet and Internet testbeds.

Adaptive and Dynamic Intrusion Detection by Means of Idiotypic Networks Paradigm

Marek Ostaszewski[1], Pascal Bouvry[1] and Franciszek Seredynski[2]
[1]Faculty of Sciences, Technology and Communication, University of Luxembourg
[2]Institute of Computer Science, Polish Academy of Sciences
[2]Institute of Computer Science, University of Podlasie

In this paper we present a novel intrusion detection architecture based on Idiotypic Network Theory (INIDS), that aims at dealing with large scale network attacks featuring variable properties, like Denial of Service (DoS). The proposed architecture performs dynamic and adaptive clustering of the network traffic for taking fast and effective countermeasures against such high-volume attacks. INIDS is evaluated on the MIT'99 dataset and outperforms previous approaches for DoS detection applied to this set.

Application of Cellular Automata to Create S-Box Functions

Miroslaw Szaban
Institute of Computer Science
University of Podlasie
Sienkiewicza 51, 08-110 Siedlce, Poland
mszaban@ap.siedlce.pl

Franciszek Seredynski
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
sered@pjwstk.edu.pl

Substitution boxes (S-boxes) play very important role in cryptography. In this paper we propose a new method based on applying Cellular Automata (CA) to create S-boxes. We present preliminary results suggesting that CA are able to realize a function of classical S-boxes. It was shown that proposed CA fulfills a number of creation conditions necessary to play a role of S-box table. Proposed CA offer a solution which produces result comparable with classical S-boxes.

A Genetic Programming Approach to Solve Scheduling Problems with Parallel Simulation

Andreas Beham[1], Stephan Winkler[1,2], Stefan Wagner[2] and Michael Affenzeller[2]
[1] Research Center Hagenberg, [2] Department of Software Engineering
Upper Austrian University of Applied Sciences, Campus Hagenberg
Softwarepark 11, A-4232 Hagenberg, Austria
{andreas.beham,stephan.winkler,stefan.wagner,michael.affenzeller}@heuristiclab.com

Scheduling and dispatching are two ways of solving production planning problems. In this work, based on preceding works, it is explained how these two approaches can be combined by the means of an automated rule generation procedure and simulation. Genetic programming is applied as the creator and optimizer of the rules. A simulator is used for the fitness evaluation and distributed over a number of machines. Some example results suggest that the approach could be successfully applied in the real world as the results are more than human competitive.

Design and Evaluation of Tabu Search Method for Job Scheduling in Distributed Environments

Fatos Xhafa and Javier Carretero
Dept. of Languages and Informatics Systems
Polytechnic University of Catalonia, Spain
fatos@lsi.upc.edu

Enrique Alba
Dept. of Languages and Computer Science
University of Málaga, Spain
eat@lcc.uma.es

Bernabé Dorronsoro
Faculty of Science, Technology and Communication
University of Luxembourg, Luxembourg
bernabe.dorronsoro@uni.lu

The efficient allocation of jobs to grid resources is indispensable for high performance grid-based applications. The scheduling problem is computationally hard even when there are no dependencies among jobs. Thus, we present in this paper a new tabu search (TS) algorithm for the problem of batch job scheduling on computational grids. We consider the job scheduling as a bi-objective optimization problem consisting of the minimization of the makespan and flowtime. The bi-objectivity is tackled through a hierarchic approach in which makespan is considered a primary objective and flowtime a secondary one. An extensive experimental study has been first conducted in order to fine-tune the parameters of our TS algorithm. Then, our tuned TS is compared versus two well known TS algorithms in the literature (one of them is hybridized with an ant colony optimization algorithm) for the problem. The computational results show that our TS implementation clearly outperforms the compared algorithms. Finally, we evaluated the performance of our TS algorithm on a new set of instances that better fits with the concept of computational grid. These instances are composed of a higher number of CheterogeneousC machines (up to 256) and emulate the dynamic behavior of these systems.

A Task Allocation Framework for Biological Sequence Comparison Applications in Heterogeneous Environments

Azzedine Boukerche[1], Marcelo Nardelli P. Santana[2] and Alba Cristina Magalhaes Alves de Melo[2]
[1]School of Information Technology, University of Ottawa, Canada, boukerch@site.uottawa.ca
[2]Department of Computer Science, University of Brasilia (UnB), Brazil,{nardelli, alves}@unb.br

Biological Sequence Comparison is a very important operation in computational biology since it is used to relate organisms and understand evolutionary processes. This article presents the design and evaluation of an allocation framework for biological sequence comparison applications that use dynamic programming and run in heterogeneous environments. Its goal is to determine which processors will execute the application, considering some characteristics of the heterogeneous environment, such as observed processor power and network bandwidth. The results obtained with four different task allocation policies in a 10-machine heterogeneous environment show that, for some sequence sizes, we were able to reduce the execution time of the parallel application in more than a half, when the appropriate number of processors is used.

Scheduling Strategies for the Bicriteria Optimization of the Robustness and Makespan

Louis-Claude Canon and Emmanuel Jeannot
LORIA, INRIA, Nancy University, CNRS, France
Email: {louis-claude.canon,emmanuel.jeannot}@loria.fr

In this paper we study the problem of scheduling a stochastic task graph with the objective of minimizing the makespan and maximizing the robustness. As these two metrics are not equivalent, we need a bicriteria approach to solve this problem. Moreover, as computing these two criteria is very time consuming we propose different approaches: from an evolutionary metaheuristic (best solutions but longer computation time) to more simple heuristics making approximations (bad quality solutions but fast computation time). We compare these different strategies experimentally and show that we are able to find different approximations of the Pareto front of this bicriteria problem.

A Parallel Insular Model for Location Areas Planning in Mobile Networks

Laidi Foughali
LIRE Laboratory
Unversité Mentouri - Constantine
Route Ain El Bey 25017 Constantine Algeria
laidi.foughali@lifl.fr

El-Ghazali Talbi
LIFL Laboratory
UMR USTL/CNRS 8022 - Bâtiment M3
59655 Villeneuve d'Ascq Cédex - FRANCE
el-ghazali.talbi@lifl.fr

Mohamed Batouche
Computer Science Department
CCIS King Saud University
Riyadh Kingdom of Saudi Arabia
batouche@ccis.edu.sa

The main interest of this paper is the optimization of the location areas planning in cellular radio networks. It is well known that the quality of service in mobile networks depends on many parameters, among them an optimal location area planning. Furthermore, it is more interesting to provide a logical organization for the already deployed networks.

In this paper, we propose the use of heuristics strategies and hybrid metaheuristics strategies to solve the location areas planning problem. The latter is formulated as a constrained planar graph partitioning problem by using a mathematical model which is based on a very realistic specification. Heuristics strategies are based on greedy algorithms while hybrid metaheuristics are based on genetic algorithms. New genetic operators have been designed to this specific problem. Moreover, parallel approaches have been proposed to improve the quality of solutions and speedup the search. Results obtained on real-life benchmarks show the effectiveness of the developed optimization algorithms.

Designing Hybrid Integrative Evolutionary Approaches to the Car Sequencing Problem

Arnaud Zinflou, Caroline Gagné and Marc Gravel
Universit du Québec à Chicoutimi, Québec, Canada
arnaud_zinflou@uqac.ca, caroline_gagne@uqac.ca, marc_gravel@uqac.ca

In this paper, we present three new integrative approaches for solving the classical car sequencing problem. These approaches are essentially based on a genetic algorithm which incorporates two crossover operators using an integer linear programming model. The two proposed hybrid crossover are combined efficiently in a genetic algorithm and we show that the hybrid approach outperforms a genetic algorithm with local search on the CSPLib benchmarks. Although that the computations time are long when integrative hybridization is used, this study well illustrates the interest of designing hybrid approaches exploiting the strengths of different methods.

Statistical Measures of a Fitness Landscape for the Vehicle Routing Problem

Zbigniew J. Czech
Silesia University of Technology, Gliwice,
and Silesia University, Sosnowiec, Poland
e-mail: zczech@polsl.pl

The work concerns the statistical measures of a fitness landscape in the context of the vehicle routing problem with time windows (VRPTW). The measures are determined by using a parallel simulated annealing algorithm as a tool for exploring a solution space. The landscape properties which are discovered allow us to evaluate the difficulty of the VRPTW benchmarking instances and to establish some parameters of the parallel algorithm.

Inference of Large-Scale Structural Features of Gene Regulation Networks Using Genetic Algorithms

Viet Anh Nguyen and Albert Y. Zomaya
School of Information Technologies, University of Sydney
NSW 2006, Australia
a.zomaya@usyd.edu.au

Considerable attempts have been made to develop models and learning strategies to infer gene networks starting from single connections. However, due to noise and other difficulties that arise from making measurements at the meso and nano levels, these so called bottom-up approaches have not been of much success. The need for methods that use a top-down approach to extract global statistics from expression data has emerged to deal with such difficulties. This paper presents a theoretical framework that employs global statistics learnt from gene expression data to infer different network structural properties of large-scale gene regulatory networks. The framework is inspired by genetic algorithms and designed with the aim to address the different weaknesses in existing approaches. Experimental results show that the developed system is more superior to previously published results.

Is a Non-uniform System of Creatures More Efficient Than a Uniform One?

Patrick Ediger, Rolf Hoffmann and Mathias Halbach
Technische Universität Darmstadt
FB Informatik, FG Rechnerarchitektur
Hochschulstraße 10, 64289 Darmstadt, Germany
{ediger, hoffmann, halbach}@ra.informatik.tu-darmstadt.de

We have analyzed the creatures exploration problem with non-uniform creatures. The creatures' task is to visit all empty cells in an environment containing obstacles with a minimum number of steps. Ten different algorithms with good performance from former investigations were used on 16 environments. New metrics were defined for such a multi agent system, like the absolute and relative efficiency. The efficiency relates the work of an agent system to the work of a reference system. A reference system is such a system that can solve the problem with the lowest number of creatures with potentially different algorithms. It turned out that the system CG-32a (32 creatures, algorithms C and G, alternating placed initially) is 31% respectively 33% more efficient than the reference systems. The relative efficiency was even higher when systems using the same algorithms were compared. Mixing different types of creatures resulted in efficiencies higher than one in 6% of all non-uniform systems.

Single Machine Scheduling with Job Ready and Delivery Times Subject to Resource Constraints

Władysław Adam Janiak Faculty of Computer Science and Management
Wrocław University of Technology
Łukasiewicza 5, 50-371 Wrocław, Poland
wladyslaw.janiak@interia.pl
Adam Janiak
Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology
Janiszewskiego 11/17, 50-372 Wrocław, Poland
adam.janiak@pwr.wroc.pl
Marie-Claude Portmann
Ecole desMines de Nancy
Parc de Saurupt, 54 042 Nancy Cedex, France
marie-claude.portmann@loria.fr

The paper deals with a single machine scheduling problem with job processing times dependent on continuously divisible resource, e.g. gas, power, energy, raw materials, catalyzer, financial outlay. Ready and delivery times are also given for each job. The problem is to find a schedule of jobs and resource allocation that minimize the time by which all jobs are delivered. Genetic approach, using some proved problem properties is constructed to solve the problem being strongly NPChard. Analysis of some computational experiment and conclusion remarks are also given.

Computer Experiments with a Parallel Clonal Selection Algorithm for the Graph Coloring Problem

Jacek Dąbrowski and Marek Kubale
Gdansk University of Technology,
ul. Gabriela Narutowicza 11/12,
80-952 Gdansk, Poland
{jacek.dabrowski, kubale}@eti.pg.gda.pl

Artificial immune systems (AIS) are algorithms that are based on the structure and mechanisms of the vertebrate immune system. Clonal selection is a process that allows lymphocytes to launch a quick response to known pathogens and to adapt to new, previously unencountered ones. This paper presents a parallel island model algorithm based on the clonal selection principles for solving the Graph Coloring Problem. The performance of the algorithm over a set of well-established benchmark graphs and random graphs is compared with a parallel Tabu Search algorithm.

A Study of Master-Slave Approaches to Parallelize NSGA-II

Juan J. Durillo, Antonio J. Nebro, Francisco Luna and Enrique Alba
Departamento de Lenguajes y Ciencias de la Computación
E.T.S.I. Informática
Universidad de Málaga (Spain)
{durillo,antonio, v,eat}@lcc.uma.es

Many of the optimization problems from the real world are multiobjective in nature, and the reference algorithm for multiobjective optimization is NSGA-II. Frequently, these problems present a high complexity, so classical metaheuristic algorithms fail to solve them in a reasonable amount of time; in this context, parallelism is a choice to overcome this fact to some extent. In this paper we study three parallel approaches (a synchronous and two asynchronous strategies) for the NSGA-II algorithm based on the master-worker paradigm. The asynchronous schemes are designed to be used in grid systems, so they can make use of hundreds of machines. We have applied them to solve a real world problem which lies in optimizing a broadcasting protocol using a network simulator. Our experiences reveal that significant time reductions can be achieved with the distributed approaches by using a grid system of more than 300 processors.

Workshop 7

Workshop on High Performance Computational Biology

HiCOMB 2008

Ordered Index Seed Algorithm for Intensive DNA Sequence Comparison

Dominique Lavenier
IRISA / CNRS
Campus de Beaulieu
35042 Rennes, France
lavenier@irisa.fr

This paper presents a seed-based algorithm for intensive DNA sequence comparison. The novelty comes from the way seeds are used to efficiently generate small ungapped alignments C or HSPs (High Scoring Pairs) C in the first stage of the search. W-nt words are first indexed and all the 4W possible seeds are enumerated following a strict order ensuring fast generation of unique HSPs. A prototype – written in C – has been realized and tested on large DNA banks. Speed-up compared to BLASTN range from 5 to 28 with comparable sensitivity.

Parallel, Scalable, Memory-Efficient Backtracking for Combinatorial Modeling of Large-Scale Biological Systems

Byung-Hoon Park[1], Matthew Schmidt[1,2], Kevin Thomas[3], Tatiana Karpinets[1] and Nagiza F. Samatova[1,2,4]
[1]Computer Science and Mathematics Division,
Oak Ridge National Laboratory, Oak Ridge, TN 37831
[2]Computer Science Department, North Carolina State University, Raleigh, NC 27695
[3]Cray Inc.
[4]Corresponding author: samatovan@ornl.gov

Data-driven modeling of biological systems such as protein-protein interaction networks is data-intensive and combinatorially challenging. Backtracking can constrain a combinatorial search space. Yet, its recursive nature, exacerbated by data-intensity, limits its applicability for large-scale systems. Parallel, scalable, and memory-efficient backtracking is a promising approach. Parallel backtracking suffers from unbalanced loads. Load rebalancing via synchronization and data movement is prohibitively expensive. Balancing these discrepancies, while minimizing end-to-end execution time and memory requirements, is desirable. This paper introduces such a framework. Its scalability and efficiency, demonstrated on the maximal clique enumeration problem, are attributed to the proposed: (a) representation of search tree decomposition to enable parallelization; (b) depth-first parallel search to minimize memory requirement; (c) least stringent synchronization to minimize data movement; and (d) on-demand work stealing with stack splitting to minimize processors' idle time. The applications of this framework to real biological problems related to bioethanol production are discussed.

Parallelized Preprocessing Algorithms for High-density Oligonucleotide Arrays

Markus Schmidberger and Ulrich Mansmann
Chair of Biometrics and Bioinformatics, IBE
University of Munich
81377 Munich, Germany
{schmidb,mansmann}@ibe.med.uni-muenchen.de

Studies of gene expression using high-density oligonucleotide microarrays have become standard in a variety of biological contexts. The data recorded using the microarray technique are characterized by high levels of noise and bias. These failures have to be removed, therefore preprocessing of raw data has been a research topic of high priority over the past few years.

Actual research and computations are limited by the available computer hardware. Furthermore most of the existing preprocessing methods are very time consuming. To solve these problems, the potential of parallel computing should be used. For parallelization on multicomputers, the communication protocol MPI (Message Passing Interface) and the R language will be used.

This paper proposes the new R language package affyPara for parallelized preprocessing of high-density oligonucleotide microarray data. Partition of data could be done on arrays and therefore parallelization of algorithms gets intuitive possible. The partition of data and distribution to several nodes solves the main memory problems and accelerates the methods by up to the factor ten.

On the Effectiveness of Rebuilding RNA Secondary Structures from Sequence Chunks

Michela Taufer[1], Thamar Solorio[2], Abel Licon[1,3], David Mireles[3] and Ming-Ying Leung[4]

[1] Dept. of Computer & Inf. Sciences
University of Delaware

[2] Dept. of Computer of Science
The University of Texas at Dallas

[3] Dept. of Computer Science
The University of Texas at El Paso

[4] Dept. of Math. Sciences and Bioinformatics Prog.
The University of Texas at El Paso

mtaufer@acm.org, tsolorio@hlt.utdallas.edu,
{alicon2, dvmireles}@miners.utep.edu, mleung@utep.edu

Despite the computing power of emerging technologies, predicting long RNA secondary structures with thermodynamics-based methods is still infeasible, especially if the structures include complex motifs such as pseudoknots. This paper presents preliminary results on rebuilding RNA secondary structures by an extensive and systematic sampling of nucleotide chunks. The rebuilding approach merges the significant motifs found in the secondary structures of the single chunks. The extensive sampling and prediction of nucleotide chunks are supported by grid technology as part of the RNAVLab functionality. Significant motifs are identified in the chunk secondary structures and merged in a single structure based on their recurrences and other statistical insights. A critical analysis of the strengths, weaknesses, and future developments of our method is presented.

A Parallel Architecture for Regulatory Motif Algorithm Assessment

Daniel Quest[1,2], Kathryn Dempsey[1], Mohammad Shafiullah[1], Dhundy Bastola[1], and Hesham Ali[1,2]
[1] College of Information Technology, University of Nebraska at Omaha
[2] Department of Pathology and Microbiology, University of Nebraska Medical Center
Omaha, NE 68182-0694 USA
E-mail: djquest@unmc.edu

Computational discovery of cis-regulatory motifs has become one of the more challenging problems in bioinformatics. In recent years, over 150 methods have been proposed as solutions, however, it remains difficult to characterize the advantages and disadvantages of these approaches because of the wide variability of approaches and datasets. Although biologists desire a set of parameters and a program most appropriate for cis-regulatory discovery in their domain of interest, compiling such a list is a great computational challenge. First, a discovery pipeline for 150+ methods must be automated and then each dataset of interest must used to grade the methods. Automation is challenging because these programs are intended to be used over a small set of sites and consequently have many manual steps intended to help the user in fine-tuning the program to specific problems or organisms. If a program is fine-tuned to parameters other than those used in the original paper, it is not guaranteed to have the same sensitivity and specificity. Consequently, there are few methods that rank motif discovery tools. This paper proposes a parallel framework for the automation and evaluation of cis-regulatory motif discovery tools. This evaluation platform can both run and benchmark motif discovery tools over a wide range of parameters and is the first method to consider both multiple binding locations within a regulatory region and regulatory regions of orthologous genes. Because of the large amount of tests required, we implemented this platform on a computing cluster to increase performance.

Design and Development of a FPGA-based Cascade Markov Model for Recognition of Steroid Hormone Response Elements

Maria Stepanova
Bioinformatics Research Centre, Nanyang Technological University,
50 Nanyang Drive,
Singapore 637553
mari0004@ntu.edu.sg

Feng Lin
School of Computer Engineering, Nanyang Technological University,
Block N4, Nanyang Avenue,
Singapore 639798
asflin@ntu.edu.sg

Valerie Lin
School of Biological Sciences, Nanyang Technological University,
60 Nanyang Drive,
Singapore 637551
CLLin@ntu.edu.sg

Steroid hormones are necessary for vital functions of most of vertebrates. Hormone molecules act within cells via interaction with their receptor proteins which are transcription factors. Identification of Hormone Response Elements (HREs) in DNA is essential for understanding the mechanism of hormone-mediated gene expression regulation. We present a systematic approach for recognition of HREs within promoters of vertebrate genes. The proposed approach is based on an experimentally validated dataset and a specifically reconstructed cascade Markov model for HRE recognition with reference to its complex composition. The approach provides a reliable accuracy for HRE prediction, and may be extended to other sequence motifs with repeated or multi-component structure. The developed FPGA implementation of the Markov model is tested using a Virtex-4 board. The model is trained for prediction of HREs in promoters of hormone-responsive genes, and for further study on direct targets for androgen, progesterone and glucocorticoid hormones.

Supporting High Performance Bioinformatics Flat-File Data Processing Using Indices

Xuan Zhang and Gagan Agrawal
Department of Computer Science and Engineering
Ohio State University
Columbus, OH, 43220
{zhangx,agrawal}@cse.ohio-state.edu

As an essential part of in vitro analysis, biological database query has become more and more important in the research process. A few challenges that are specific to bioinformatics applications are data heterogeneity, large data volume and exponential data growth, constant appearance of new data types and data formats. We have developed an integration system that processes data in their flat file formats. Its advantages include the reduction of overhead and programming efforts. In the paper, we discuss the usage of indicing techniques on top of this flat file query system. Besides the advantage of processing flat files directly, the system also improves its performance and functionality by using indexes. Experiments based on real life queries are used to test the integration system.

Sample-Align-D: A High Performance Multiple Sequence Alignment System Using Phylogenetic Sampling and Domain Decomposition

Fahad Saeed and Ashfaq Khokhar
Department of Electrical and Computer Engineering
University Of Illinois at Chicago
Chicago, IL 60607
Email: {fsaeed2, ashfaq}@uic.edu

Multiple Sequence Alignment (MSA) is one of the most computationally intensive tasks in Computational Biology. Existing best known solutions for multiple sequence alignment take several hours (in some cases days) of computation time to align, for example, 2000 homologous sequences of average length 300. Inspired by the Sample Sort approach in parallel processing, in this paper we propose a highly scalable multiprocessor solution for the MSA problem in phylogenetically diverse sequences. Our method employs an intelligent scheme to partition the set of sequences into smaller subsets using k-mer count based similarity index, referred to as k-mer rank. Each subset is then independently aligned in parallel using any sequential approach. Further fine tuning of the local alignments is achieved using constraints derived from a global ancestor of the entire set. The proposed Sample-Align-D Algorithm has been implemented on a cluster of workstations using MPI message passing library. The accuracy of the proposed solution has been tested on standard benchmarks such as PREFAB. The accuracy of the alignment produced by our methods is comparable to that of well known sequential MSA techniques. We were able to align 2000 randomly selected sequences from the Methanosarcina acetivorans genome in less than 10 minutes using Sample-Align-D on a 16 node cluster, compared to over 23 hours on sequential MUSCLE system running on a single cluster node.

Adaptive Locality-Effective Kernel Machine for Protein Phosphorylation Site Prediction

Paul D. Yoo, Yung Shwen Ho, Bing Bing Zhou and Albert Y. Zomaya
Advanced Networks Research Group
School of Information Technologies (J12)
University of Sydney, NSW 2006, Australia
dyoo4334@it.usyd.edu.au

In this study, we propose a new machine learning model namely, Adaptive Locality-Effective Kernel Machine (Adaptive-LEKM) for protein phosphorylation site prediction. Adaptive-LEKM proves to be more accurate and exhibits a much stable predictive performance over the existing machine learning models. Adaptive-LEKM is trained using Position Specific Scoring Matrix (PSSM) to detect possible protein phosphorylation sites for a target sequence. The performance of the proposed model was compared to seven existing different machine learning models on newly proposed PS-Benchmark_1 dataset in terms of accuracy, sensitivity, specificity and correlation coefficient. Adaptive-LEKM showed better predictive performance with 82.3% accuracy, 80.1% sensitivity, 84.5% specificity and 0.65 correlation-coefficient than contemporary machine learning models.

Impact of Multicores on Large-scale Molecular Dynamics Simulations

Sadaf R. Alam, Pratul K. Agarwal, Scott S. Hampton, Hong Ong and Jeffrey S. Vetter
Computer Science and Mathematics Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831
{alamsr,agarwalpk,hamptonss,hongong,vetter}@ornl.gov

Processing nodes of the Cray XT and IBM Blue Gene Massively Parallel Processing (MPP) systems are composed of multiple execution units, sharing memory and network subsystems. These multicore processors offer greater computational power, but may be hindered by resource contention. In order to understand and avoid such situations, we investigate the impact of resource contention on three scalable molecular dynamics suites: AMBER (PMEMD module), LAMMPS, and NAMD. The results reveal the factors that can inhibit scaling and performance efficiency on emerging multicore processors.

Workshop 8

Advances in Parallel and Distributed Computing Models

APDCM 2008

The Space Complexity of the Leader Election in Anonymous Networks

Ei Ando, Hirotaka Ono, Kunihiko Sadakane and Masafumi
Yamashita Department of Computer Science and Communication Engineering
Graduate School of Information Science and Electrical Engineering, Kyushu University

It is known that the leader election in anonymous networks is not always solvable, and solvable/unsolvable cases are characterized bythe network topologies. Therefore a distributed leader election algorithm is required to elect a leader when it is possible, otherwise recognize the impossibility and stop.Although former studies proposed several leader election algorithms, the space complexity of the problem is not considered well. This paper focuses on the space complexity, that is, the necessary or sufficient number of bits on processors to execute a leader election algorithm. First we show that only one bit memory is sufficient for a leader election algorithm which is specific to a fixed n. We then show that a general algorithm can solve the leader election for arbitrary n if each processor has $O(n \log d)$ bits memory where d is the maximum degree of a processor. Finally, we give a lower bound $\Omega(\log n)$ on the space complexity, that is, we show that it is impossible to construct a leader election algorithm if only $\log n$ bits are available for a processor.

Fault Tolerant Scheduling of Precedence Task Graphs on Heterogeneous Platforms

Anne Benoit, Mourad Hakem and Yves Robert
ENS Lyon, Universite de Lyon, LIP laboratory
UMR 5668, ENS Lyon C CNRS C INRIA C UCBL, Lyon, France
{Anne.Benoit,Mourad.Hakem,Yves.Robert}@ens-lyon.fr

Fault tolerance and latency are important requirements in several applications which are time critical in nature: such applications require guaranties in terms of latency, even when processors are subject to failures. In this paper, we propose a fault tolerant scheduling heuristic for mapping precedence task graphs on heterogeneous systems. Our approach is based on an active replication scheme, capable of supporting ε arbitrary fail-silent (fail-stop) processor failures, hence valid results will be provided even if ε processors fail. We focus on a bi-criteria approach, where we aim at minimizing the latency given a fixed number of failures supported in the system, or the other way round. Major achievements include a low complexity, and a drastic reduction of the number of additional communications induced by the replication mechanism. Experimental results demonstrate that our heuristics, despite their lower complexity, outperform their direct competitor, the FTBAR scheduling algorithm [3].

Offline and Online Master-Worker Scheduling of Concurrent Bags-of-Tasks on Heterogeneous Platforms

Anne Benoit [2,4,5] Loris Marchal [1,4,5] Jean-François Pineau [2,4,5]
Yves Robert [2,4,5] Frédéric Vivien [3,4,5]
[1] CNRS [2] ENS Lyon [3] INRIA [4] Université de Lyon
[5] LIP laboratory, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, France
{anne.benoit, loris.marchal, jean-francois.pineau, yves.robert, frederic.vivien}@ens-lyon.fr

Scheduling problems are already difficult on traditional parallel machines. They become extremely challenging on heterogeneous clusters, even when embarrassingly parallel applications are considered. In this paper we deal with the problem of scheduling multiple applications, made of collections of independent and identical tasks, on a heterogeneous master-worker platform. The applications are submitted online, which means that there is no a priori (static) knowledge of the workload distribution at the beginning of the execution. The objective is to minimize the maximum stretch, i.e. the maximum ratio between the actual time an application has spent in the system and the time this application would have spent if executed alone.

On the theoretical side, we design an optimal algorithm for the offline version of the problem (when all release dates and application characteristics are known beforehand). We also introduce several heuristics for the general case of online applications.

On the practical side, we have conducted extensive simulations and MPI experiments, showing that we are able to deal with very large problem instances in a few seconds. Also, the solution that we compute totally outperforms classical heuristics from the literature, thereby fully assessing the usefulness of our approach.

Self-Stabilizing Wavelets and ϱ-Hops Coordination

Christian Boulinier and Franck Petit
MIS / Université de Picardie Jules Verne
Amiens, France

In this paper, we first introduce a simple tool called the wavelet or ϱ-wavelet scheme. Wavelets deal with coordination among processes which are at most ϱ hops away of each other. We propose a self-stabilizing solution for this scheme. Our solution requires no underlying structure and works in arbitrary anonymous settings, i.e., where process identifiers are not required. We show that our solution provides a simple and generic self-stabilizing ϱ-infimum computation.

Next, we present a self-stabilizing ϱ-barrier synchronization protocol based on the wavelet scheme. We show that our protocol provides an efficient device in the design of local coordination problems at distance ϱ, such as the ϱ-Local Resource Allocation (LRA). In particular, we propose a solution for the popular ϱ-Local Mutual Exclusion (LME) problem. The solution to ϱ-LME also provides a transformer to transform algorithms written under ϱ-central daemon into algorithms working with any distributed daemon.

Why Shared Memory Matters to VLSI Design: The BSR Is as Powerful as Reconfiguration

Stefan D. Bruda and Yuanqiao Zhang
Department of Computer Science
Bishop's University
Sherbrooke, Quebec J1M 1Z7, Canada
{bruda|zhang}@cs.ubishops.ca

We investigate the relative computational power of parallel models with directed reconfigurable buses and the most powerful shared memory model, the broadcast with selective reduction (BSR). We show that these two models have the same computational power in a strong sense. We also show that the Collision write conflict resolution rule is universal on models with directed reconfigurable buses (meaning that complex conflict resolution rules such as Priority or even Combining can be simulated by Collision with constant-time overhead). Consequences of these finding, including some real time consideration, are discussed.

Mobility Control Schemes with Quick Convergence in Wireless Sensor Networks

Xiao Chen
Dept. of Comp. Sci.
Texas State Univ.
San Marcos, TX 78666
xc10@txstate.edu

Zhen Jiang
Dept. of Comp. Sci.
West Chester Univ.
West Chester, PA 19383
zjiang@wcupa.edu

Jie Wu
Dept. of Comp. Sci. and Eng.
Florida Atlantic Univ.
Boca Raton, FL 33431
jie@cse.fau.edu

In the near future, wireless sensor networks (WSN) performing sensing and communication tasks will be widely deployed as technology rapidly advances. Communication is one of the essential functionalities of these networks while power and computation resources in each sensor are limited. Recently, attention has been drawn to using mobility control to minimize energy consumption in wireless sensor networks. In this paper, we are going to provide quickly converging mobility control schemes to achieve optimal configuration in a single data flow. The key idea of our schemes is to use the optimal location information of each relay node as a guide for node movement while maintaining the connectivity of relay nodes along the data flow. Experimental results show that our schemes can speed up the convergence process to nearly the optimal and reduce the cost of it almost to the minimum, compared with the best results known to the date.

A Distributed Algorithmic Framework for Coverage Problems in Wireless Sensor Networks

Akshaye Dhawan and Sushil K. Prasad
Department of Computer Science
Georgia State University
Atlanta, Ga 30303
akshaye@cs.gsu.edu, sprasad@gsu.edu

One of the key challenges in Wireless Sensor Networks (WSNs) is that of extending the lifetime of the network while meeting some coverage requirements. In this paper we present a distributed algorithmic framework to enable sensors to determine their sleep-sense cycles based on specific coverage goals. The framework is based on our earlier work on the target coverage problem. We give a general version of the framework that can be used to solve network/graph problems for which melding compatible neighboring local solutions directly yields globally feasible solutions. We also apply this framework to several variations of the coverage problem, namely, target coverage, area coverage and k-coverage problems, to demonstrate its general applicability. Each sensor constructs minimal cover sets for its local coverage objective. The framework entails each sensor prioritizing these local cover sets and then negotiating with its neighbors for satisfying mutual constraints. We introduce a dependency graph model that can capture the interdependencies among the cover sets. Detailed simulations are carried out to further demonstrate the resulting performance improvements and effectiveness of the framework.

Maximal Strips Data Structure to Represent Free Space on Partially Reconfigurable FPGAs

Mostafa Elbidweihy and Jerry L. Trahan
Dept. of Electrical and Computer Engineering, Louisiana State University
Baton Rouge, LA USA
{melbid1@lsu.edu, trahan@ece.lsu.edu}

Partially reconfigurable devices allow the execution of multiple tasks simultaneously on the same chip. To schedule a set of tasks in a small amount of time, the scheduling algorithm will need to represent the free space efficiently. A data structure to represent free space should allow the scheduler to identify free space in which to place a new task and admit efficient updates after placing or removing a task. In this paper, we review some existing data structures and analyze their time complexity. We propose a new structure using maximal horizontal and vertical strips to represent the free space. These strips start and stop at task boundaries. Simulation and time analysis showed that this method has better time complexity than many other free space data structures and at the same time has a very reasonable rejection ratio on real-time tasks compared to other methods.

On the Performance and Cost of Some PRAM Models on CMP Hardware

Martti Forsell
VTT Technical Research Centre of Finland
Platform Architectures Team
Box 1100, FI-90571 Oulu, Finland
Martti.Forsell@VTT.Fi

The Parallel Random Access Machine is a very strong model of parallel computing that has resisted cost-effective implementation attempts for decades. Recently, the development of VLSI technology has provided means for indirect on-chip implementation, but there are different variants of the PRAM model that provide different performance, area and power figures and it is not known how their implementations compare to each others. In this paper we measure the performance and estimate the cost of practical implementations of four PRAM models including EREW, Limited Arbitrary CRCW, Full Arbitrary CRCW, Full Arbitrary Multioperation CRCW on our Eclipse chip multiprocessor framework. Interestingly, the most powerful model shows the lowest relative cost and highest performance/area and performance/power figures.

Component Labeling for *k*-Concave Binary Images Using an FPGA

Yasuaki Ito
Department of Information Engineering
Hiroshima University
1-4-1 Kagamiyama, Higashi Hiroshima
Hiroshima, 739C8527 Japan
yasuaki@cs.hiroshima-u.ac.jp

Koji Nakano
Department of Information Engineering
Hiroshima University
1-4-1 Kagamiyama, Higashi Hiroshima
Hiroshima, 739C8527 Japan
nakano@cs.hiroshima-u.ac.jp

Connected component labeling is a task that assigns unique IDs to the connected components of a binary image. The main contribution of this paper is to present a hardware connected component labeling algorithm for k-concave binary images designed and implemented in FPGA. Pixels of a binary image are given to the FPGA in raster order, and the resulting labels are also output in the same order. The advantage of our labeling algorithm is small latency and to use a small internal storage of the FPGA. We have implemented our hardware labeling algorithm in an Altera Stratix Family FPGA, and evaluated the performance. The implementation result shows that for a 10-concave binary image of 2048 × 2048, our connected component labeling algorithm runs in approximately 70ms and its latency is approximately 750ns.

A Scalable Configurable Architecture for the Massively Parallel GCA Model

J. Jendrsczok, P. Ediger and R. Hoffmann
TU Darmstadt
FB Informatik, FG Rechnerarchitektur
Hochschulstraße 10, 64289 Darmstadt, Germany
{jendrsczok, ediger, hoffmann}@ra.informatik.tu-darmstadt.de

The Global Cellular Automata model (GCA) is a massively parallel computation model which extends the classical Cellular Automata model (CA) with dynamic global neighbors. We present for that model a data parallel architecture which is scalable in the number of parallel pipelines and which uses application specific operators (adapted operators). The instruction set consists of control and RULE instructions. A RULE computes the next cell contents for each cell in the destination object. The machine consists of P pipelines. Each pipeline has an associated primary memory bank and has access to the global memory (real or emulated multi port memory).

The diffusion of particles was used as an example in order to demonstrate the adaptive operators, the machine programming and its performance. Particles which point to each other within a defined neighborhood search space are interchanged. The pointers are modified in each generation by a pseudo random function. The machine with up to 32 pipelines was synthesized for an Altera FPGA for that application.

Analytic Performance Models for Bounded Queueing Systems

Praveen Krishnamurthy and Roger D. Chamberlain
Dept. of Computer Science and Engineering, Washington University in St. Louis
{praveen,roger}@wustl.edu

Pipelined computing applications often have their performance modeled using queueing techniques. While networks with infinite capacity queues have well understood properties, networks with finite capacity queues and blocking between servers have resisted closed-form solutions and are typically analyzed with approximate solutions. It is this latter case that more closely represents the circumstances present for pipelined computation. In this paper, we extend an existing approximate solution technique and, more importantly, provide guidance as to when the approximate solutions work well and when they fail.

Manageable Peer-to-Peer Architecture for Video-on-Demand

Xuguang Lan, Nanning Zheng, Jianru Xue, Weike Chen, Bin Wang and Wen Ma
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
{xglan, nnzheng, jrxue, wkchen, bwang, wman}@aiar.xjtu.edu.cn

An efficiently manageable overlay network architecture, called AIRVoD, is proposed for video-ondemand, based on a distributed-centralized P2P network. The system consists of distributed servers, peers, SuperNodes and sub-SuperNodes. The distributed central servers act as the centralized index server to store the peer list, program list, and buffer information of peers. Each newly joined peer in AIRVoD periodically exchanges data availability information with the central server. Some powerful peers are selected to be sub-SuperNodes which store a larger part of the demanded program. The demanding peers can retrieve correspondingly unavailable data from partners selected from the central server, sub- SuperNodes and SuperNodes that have the original programs to supply the available data, by using the proposed parallel scheduling algorithm. There are four characteristics of this architecture: 1) easy to globally balance and manage: central server can identify, and cluster each newly joined peer, and allocate the load in the whole peer network; 2) efficient for dynamic networks: data transmission is dynamically determined according to data availability which can be derived from central server; 3) resilient, as the partnerships can adapt to quick switching among multi-suppliers under global balance; and 4) highly cost-effective: the powerful peers are taken full advantage to be larger suppliers. AIRVoD has been demonstrated based on CERNET OF CHINA.

BSP2OMP: A Compiler for Translating BSP Programs to OpenMP

Ami Marowka
Shenkar College of Engineering and Design
12 Anna Frank, Ramat-Gan, 52526, Israel
amimar2@yahoo.com

The convergence of the two widely used parallel programming paradigms, shared-memory and distributed-shared-memory parallel programming models, into a unified parallel programming model is crucial for parallel computing to become the next mainstream programming paradigm. We study the design differences and the performance issues of two parallel programming models: a sharedmemory programming model (OpenMP) and a distributedshared programming model (BSP). The study was carried out by designing a compiler for translating BSP parallel programs to an OpenMP programming model called BSP2OMP. Analysis of the compiler outcome, and of the performance of the compiled programs, show that the two models are based on very similar underlying principles and mechanisms.

On the Construction of Paired Many-to-Many Disjoint Path Covers in Hypercube-Like Interconnection Networks with Faulty Elements

Jung-Heum Park
School of Computer Science and Information Engineering
The Catholic University of Korea, Korea
j.h.park@catholic.ac.kr

Hee-Chul Kim
School of Computer Science and Information Communications Engineering
Hankuk University of Foreign Studies, Korea
hckim@hufs.ac.kr

Hyeong-Seok Lim
School of Electronics and Computer Engineering
Chonnam National University, Korea
hslim@chonnam.ac.kr

A paired many-to-many k-disjoint path cover (k-DPC) of a graph G is a set of k disjoint paths joining k distinct source-sink pairs in which each vertex of G is covered by a path. This paper is concerned with paired many-to-many disjoint path coverability of hypercube-like interconnection networks, called restricted HL-graphs. The class includes twisted cubes, crossed cubes, multiply twisted cubes, Möbius cubes, Mcubes, and generalized twisted cubes. We show that every restricted HL-graph of degree m with f or less faulty elements has a paired many-to-many k-DPC for any f and $k \geq 2$ with $f + 2k \leq m$. The result improves the known bound of $f + 2k \leq m - 1$ by one.

Design and Implementation of a Tool for Modeling and Programming Deadlock Free Meta-pipeline Applications

Shinichi Yamagiwa	Leonel Sousa
INESC-ID/IST	INESC-ID/IST
Rua Alves Redol, 9	Rua Alves Redol, 9
1000-029 Lisboa Portugal	1000-029 Lisboa Portugal
yama@inesc-id.pt	las@inesc-id.pt

The Caravela platform has been designed to develop a parallel and distributed stream-based computing paradigm, namely supported on the pipeline processing approach herein designated by meta-pipeline. This paper is focused on the design and implementation of a modeling tool for the meta-pipeline, namely to tackle the deadlock problem due to uninitialized input data stream in a pipeline-model. A new efficient algorithm is proposed to prevent deadlock situations by detecting uninitialized edges in a pipeline graph. The algorithm identifies the cyclic paths in a pipeline-graph and builds a reduced list with only the true cyclic paths that have to be really initialized. Further optimization techniques are also proposed to reduce the computation time and the required amount of memory. Moreover, this paper also presents a Graphical User Interface (GUI) for easy programming meta-pipeline applications, which provides an automatic validation procedure based on the proposed algorithm. Experimental results presented in this paper show the effectiveness of both the proposed algorithm and the developed GUI.

SIFT Implementation and Optimization for Multi-Core Systems

Qi Zhang [1,2], Yurong Chen [2], Yimin Zhang [2] and Yinlong Xu [1]
[1] Dept. of Computer Science, Univ. of Science and Technology of China
[2] Intel China Research Center, Intel Corporation
{yurong.chen}@intel.com

Scale Invariant Feature Transform (SIFT) is an approach for extracting distinctive invariant features from images, and it has been successfully applied to many computer vision problems (e.g. face recognition and object detection). However, the SIFT feature extraction is compute-intensive, and a real-time or even super-real-time processing capability is required in many emerging scenarios. Nowadays, with the multicore processor becoming mainstream, SIFT can be accelerated by fully utilizing the computing power of available multi-core processors. In this paper, we propose two parallel SIFT algorithms and present some optimization techniques to improve the implementation's performance on multi-core systems. The result shows our improved parallel SIFT implementation can process general video images in super-real-time on a dual-socket, quad-core system, and the speed is much faster than the implementation on GPUs. We also conduct a detailed scalability and memory performance analysis on the 8-core system and on a 32-core Chip Multiprocessor (CMP) simulator. The analysis helps us identify possible causes of bottlenecks, and we suggest avenues for scalability improvement to make this application more powerful on future large-scale multicore systems.

Self-organized Forensic Support in MANETs

Xiwei Zhao, Vinayak Ganapathy, Niki Pissinou and Kia Makki
Telecommunications & Information Technology Institute
College of Engineering & Computing, Florida International University
{xzhao001, vgana002, pissinou, makkik}@fiu.edu

A distributed hash table (DHT) based approach for supporting forensic capability in Mobile Ad Hoc Networks (MANETs) is presented. The DHT-based approach has been modified to inhibit recursive increase in bandwidth consumption due to forensic activity – the process of logging is associated with that of packet delivery via customizable decreasing functions. Simulation has revealed that this approach limits the bandwidth requirement for forensic activities, although it requires a trade-off between bandwidth consumption and effective logging. The focus is to design a self-organized logging system over networks with dynamic topology.

Workshop 9
Communication Architecture for Clusters
CAC 2008

Making an SCI Fabric Dynamically Fault Tolerant

Håkon Kvale Stensland[1,2], Olav Lysne[1,2], Roy Nordstrøm[3] and Hugo Kohmann[3]
[1] Simula Research Laboratory, Lysaker, Norway
[2] University of Oslo, Department of Informatics, Oslo, Norway
[3] Dolphin Interconnect Solutions, Oslo, Norway

In this paper we present a method for dynamic fault tolerant routing for SCI networks implemented on Dolphin Interconnect Solutions hardware. By dynamic fault tolerance, we mean that the interconnection network reroutes affected packets around a fault, while the rest of the network is fully functional. To the best of our knowledge this is the first reported case of dynamic fault tolerant routing available on commercial off the shelf interconnection network technology without duplicating hardware resources. The development is focused around a 2-D torus topology, and is compatible with the existing hardware, and software stack. We look into the existing mechanisms for routing in SCI. We describe how to make the nodes that detect the faulty component do routing decisions, and what changes are needed in the existing routing to enable support for local rerouting. The new routing algorithm is tested on clusters with real hardware. Our tests show that distributed databases like MySQL can run uninterruptedly while the network reacts to faults. The solution is now part of Dolphin Interconnect Solutions SCI driver, and hardware development to further decrease the reaction time is underway.

A Dynamic Routing Control System for High-Performance PC Cluster with Multi-path Ethernet Connection

Shin'ichi Miura[1], Taisuke Boku[1,2], Takayuki Okamoto[1] and Toshihiro Hanawa[2]
[1] Graduate School of Systems and Information Engineerings, University of Tsukuba
[2] Center for Computational Sciences, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
{miura,taisuke,okamoto,hanawa}@hpcs.cs.tsukuba.ac.jp

VLAN-based Flexible, Reliable and Expandable Commodity Network (VFREC-Net) is a network construction technology for PC clusters that allows multi-path network routing to be configured using inexpensive Layer-2 Ethernet switches based on tagged-VLAN technology. Current VFREC-Net system encounters problems with traffic balancing when the communication pattern of the application does not fit the network topology, due to its static routing scheme.

We have developed a framework to solve this problem by allowing dynamic rewriting of the routing table from the application level. The basic function of modifying the VLANID assignment for the communicating node pair from user application is supported. Using this function, it is possible to dynamically select the path to the destination from multiple paths.

We also provide an API library for MPI programming to use the above framework, and confirmed its effectiveness through communication optimization on the NPB Kernel-CG benchmark.

Exploiting Spatial Parallelism in Ethernet-based Cluster Interconnects

Stavros Passas, George Kotsis, Sven Karlsson and Angelos Bilas
Institute of Computer Science (ICS)
Foundation for Research and Technology - Hellas (FORTH)
P.O. Box 1385, Heraklion, GR-71110, Greece
{stabat,kotsis,svenka,bilas}@ics.forth.gr

In this work we examine the implications of building a single logical link out of multiple physical links. We use MultiEdge to examine the throughput-CPU utilization tradeoffs and examine how overheads and performance scale with the number and speed of links. We use low-level instrumentation to understand associated overheads, we experiment with setups between 1 and 8 1-GBit/s links, and we contrast our results with a single 10-GBit/s link. We find that: (a) Our base protocol achieves up-to 65% of the nominal aggregate throughput. (b) Replacing the interrupts with polling significantly impacts only the multiple link configurations, reaching 80% of nominal throughput. (c) The impact of copying on CPU overhead is significant, and removing copying results in up-to 66% improvement in maximum throughput, reaching almost 100% of the nominal throughput. (d) Scheduling packets over heterogeneous links requires simple but dynamic scheduling to account for different link speeds and varying load.

Synchronized Send Operations for Efficient Streaming Block I/O Over Myrinet

Evangelos Koukis, Anastassios Nanos and Nectarios Koziris
National Technical University of Athens
School of Electrical and Computer Engineering
Computing Systems Laboratory
{vkoukis, ananos, nkoziris}@cslab.ece.ntua.gr

Providing scalable clustered storage in a cost-effective way depends on the availability of an efficient network block device (nbd) layer. We study the performance of gmblock, an nbd server over Myrinet utilizing a direct disk-to-NIC data path which bypasses the CPU and main memory bus. To overcome the architectural limitation of a low number of outstanding requests, we focus on overlapping read and network I/O for a single request, in order to improve throughput. To this end, we introduce the concept of synchronized send operations and present an implementation on Myrinet/GM, based on custom modifications to the NIC firmware and associated userspace library. Compared to a network block sharing system over standard GM and the base version of gmblock, our enhanced implementation supporting synchronized sends delivers 81% and 44% higher throughput for streaming block I/O, respectively.

HPPNET: A Novel Network for HPC and Its Implication for Communication Software

Panyong Zhang[1,2,3], Can Ma[1,2,3], Jie Ma[1,2], Qiang Li[1,2,3] and Dan Meng[1,2]
[1]National Research Center for Intelligent Computing Systems,
Institute of Computing Technology, Chinese Academy of Sciences
[2]Key Laboratory of Computer System and Architecture, Chinese Academy of Sciences
[3]Graduate University of Chinese Academy of Sciences
{zhangpanyong, macan, majie, liqiang, md}@ncic.ac.cn

With the widespread adoption of multicore processors in high performance computing (HPC) environment, the balance between computation and communication moves towards computation. It has been becoming more important to design a high efficient network for HPC system, which commonly has two challengers associated: 1) To provide a communication environment with low latency, high bandwidth, and high small message processing rate; 2) and to efficiently support the partitioned global address space (PGAS) programming model. With respects to these needs this paper proposes a novel network named HPPNET. By adopting HyperTransport interface, separate channel design, on-load part of processing work to host, and transparent direct load/store in global physical address space, HPPNET can sufficiently support both needs for HPC. Meanwhile, we have adopted several key technologies to minimize the implication of new network for communication software. Evaluation shows that HPPNET hardware design can achieve high performance and bring no barrier to high efficiency software design. Our results also show that the 8 bytes remote store cost 0.4μs in HPPNET prototype.

Optimizing Non-blocking Collective Operations for InfiniBand

Torsten Hoefler
Open Systems Lab
Indiana University
Bloomington IN 47405
Email: htor@cs.indiana.edu

Andrew Lumsdaine
Open Systems Lab
Indiana University
Bloomington IN 47405
Email: lums@cs.indiana.edu

Non-blocking collective operations have recently been shown to be a promising complementary approach for overlapping communication and computation in parallel applications. However, in order to maximize the performance and usability of these operations it is important that they progress concurrently with the application without introducing CPU overhead and without requiring explicit user intervention. While studying nonblocking collective operations in the context of our portable library (libNBC), we found that most MPI implementations do not sufficiently support overlap over the InfiniBand network. To address this issue, we developed a low-level communication layer for libNBC based on the Open Fabrics InfiniBand verbs API. With this layer we are able to achieve high degrees of overlap without the need to explicitly progress the communication operations. We show that the communication overhead of parallel application kernels can be reduced up to 92% while not requiring user intervention to make progress.

Scaling Alltoall Collective on Multi-core Systems

Rahul Kumar, Amith Mamidala and D. K. Panda
Department of Computer Science and Engineering
The Ohio State University
{kumarra, mamidala, panda}@cse.ohio-state.edu

MPI_Alltoall is one of the most communication intense collective operation used in many parallel applications. Recently, the supercomputing arena has witnessed phenomenal growth of commodity clusters built using InfiniBand and multi-core systems. In this context, it is important to optimize this operation for these emerging clusters to allow for good application scaling. However, optimizing MPI_Alltoall on these emerging systems is not a trivial task.

InfiniBand architecture allows for varying implementations of the network protocol stack. For example, the protocol can be totally on-loaded to a host processing core or it can be off-loaded onto the NIC or can use any combination of the two. Understanding the characteristics of these different implementations is critical in optimizing a communication intense operation such as MPI_Alltoall. In this paper, we systematically study these different architectures and propose new schemes for MPI_Alltoall tailored to these architectures. Specifically, we demonstrate that we cannot use one common scheme which performs optimally on each of these varying architectures. For example, on-loaded implementations can exploit multiple cores to achieve better network utilization, and in offload interfaces aggregation can be used to avoid congestion on multi-core systems. We employ shared memory aggregation techniques in these schemes and elucidate the impact of these schemes on multi-core systems. The proposed design achieves a reduction in MPI Alltoall time by 55% for 512Byte messages and speeds up the CPMD application by 33%.

A Hybrid MPI Design Using SCTP and iWARP

Mike Tsai, Brad Penoff and Alan Wagner
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Email: {myct,penoff,wagner}@cs.ubc.ca

Remote Direct Memory Access (RDMA) and pointto- point network fabrics both have their own advantages. MPI middleware implementations typically use one or the other, however, the appearance of the Internet Wide Area RDMA Protocol (iWARP), RDMA over IP, and protocol off-load devices introduces the opportunity to use a hybrid design for MPI middleware that uses both iWARP and a transport protocol directly. We explore the design of a new MPICH2 channel device based on iWARP and the Stream Control Transmission Protocol (SCTP) that uses SCTP for all point-to-point MPI routines and iWARP for all remote memory access routines (i.e., one-sided communication).

The design extends the Ohio Supercomputer Center software-based iWARP stack and our MPICH2 SCTP-based channel device. The hybrid channel device aligns the semantics of the MPI routine with the underlying protocol that best supports the routine and also allows the MPI API to exploit the potential performance benefits of the underlying hardware more directly. We describe the design and issues related to the progress engine design and connection setup. We demonstrate how to implement iWARP over SCTP rather than TCP and discuss its advantages and disadvantages. We are not aware of any other software implementations of iWARP over SCTP, nor MPI middleware that uses both iWARP verbs and the SCTP API.

Design and Implementation of Open-MX: High-Performance Message Passing Over Generic Ethernet Hardware

Brice Goglin
Laboratoire Bordelais de Recherche en Informatique
INRIA Bordeaux - Sud-Ouest C France
Brice.Goglin@inria.fr

Open-MX is a new message passing layer implemented on top of the generic Ethernet stack of the Linux kernel. It provides high-performance communication on top of any Ethernet hardware while exhibiting the Myrinet Express application interface. Open-MX also enables wire-interoperability with Myricom's MXoE hosts. This article presents the design of the Open-MX stack which reproduces the MX firmware in a Linux driver. MPICH-MX and PVFS2 layers are already able to work flawlessly on Open-MX. The first performance evaluation shows interesting latency and bandwidth results on 1 and 10 gigabit hardware.

A Multithreaded Communication Engine for Multicore Architectures

François TRAHAY, Elisabeth BRUNET, Alexandre DENIS and Raymond NAMYST
INRIA, LABRI, Université Bordeaux 1
351 cours de la Libération
F-33405 TALENCE, FRANCE
{trahay,brunet,denis,namyst}@labri.fr

The current trend in clusters leads towards an increase of the number of cores per node. As a result, an increasing number of parallel applications is mixing message passing and multithreading as an attempt to better match the underlying architecture's structure. This naturally raises the problem of designing efficient, multithreaded implementations of MPI. In this paper, we present the design of a multithreaded communication engine able to exploit idle cores to speed up communications in two ways: it can move CPU-intensive operations out of the critical path (e.g. PIO transfers offload), and is able to let rendezvous transfers progress asynchronously. We have implemented these methods in the PM2 software suite, evaluated their behavior in typical cases, and we have observed good performance results in overlapping communication and computation.

Workshop 10

NSF Next Generation Software Program

NSFNGS 2008

Understanding the Performance of Streaming Applications Deployed on Hybrid Systems

Joseph Lancaster, Ron Cytron, and Roger D. Chamberlain
Dept. of Computer Science and Engineering
Washington University in St. Louis
lancaster@wustl.edu, cytron@wustl.edu, roger@wustl.edu

Significant performance gains have been reported by exploiting the specialized characteristics of hybrid computing architectures for a number of streaming applications. While it is straightforward to physically construct these hybrid systems, application development is often quite difficult. We have built an application development environment, Auto-Pipe, that targets streaming applications deployed on hybrid architectures. Here, we describe some of the current and future characteristics of the Auto-Pipe environment that facilitate an understanding of the performance of an application that is deployed on a hybrid system.

The Fresh Breeze Project: A Multi-Core Chip Supporting Composable Parallel Programming

Jack B. Dennis
MIT Computer Science and Artificial Intelligence Laboratory
dennis@csail.mit.edu

The Fresh Breeze Project concerns the architecture and design of a multicore chip that can achieve superior performance while supporting composability of parallel programs. The requirements of composability imply that the management of processor allocation and memory management must be sufficiently flexible to permit reassignment of resources according to the current needs of computations. The Fresh Breeze Programming model combines the spawn/join threading model of Cilk with a write-once memory model based on fixed-size chunks that are allocated and freed by efficient hardware mechanisms. This model supports computing jobs by many users, each consisting of a hierarchy of function activations. The model satisfies all six principles for supporting modular program construction. Within this programming model, it is possible for any parallel program to be used, without change, as a component in building larger parallel programs.

The Potential of Computation Reuse in High-Level Optimization of a Signal Recognition System

Melina Demertzi[1], Pedro C. Diniz[2], Mary W. Hall[1], Anna C. Gilbert[3] and Yi Wang[3]

[1]USC / Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, Calif. 90292
USA
{melina,mhall}@isi.edu

[2]Instituto Superior Técnico
Technical University of Lisbon
Tagus park, 2780 Porto Salvo
Portugal
pedro.diniz@tagus.ist.utl.pt

[3]University of Michigan
Department of Mathematics
Ann Arbor, Michigan 48109
USA
annacg,yiws@umich.edu

This paper evaluates the potential of exploiting computation reuse in a signal recognition system that is jointly optimized from mathematical representation, algorithm design and final implementation. Walsh wavelet packets in conjunction with a BestBasis algorithm are used to derive transforms that discriminate between signals. The FPGA implementation of this computation exploits the structure of the resulting transform matrices in several ways to derive a highly optimized hardware representation of this signal recognition problem. Specifically, we observe in the transform matrices a significant amount of reuse of subrows, thus indicating redundant computation. Through analysis of this reuse, we discover the potential for a 3X reduction in the amount of computation of combining a transform matrix and signal. In this paper, we focus on how the implementation might exploit this reuse in a profitable way. By exploiting a subset of this computation reuse, the system can navigate the tradeoff space of reducing computation and the extra storage required.

Enhancing the Effectiveness of Utilizing an Instruction Register File

David Whalley and Gary Tyson
Computer Science Dept., Florida State Univ., Tallahassee, FL 32306-4530
e-mail: {whalley,tyson}@cs.fsu.edu

This paper describes the outcomes of the NSF Grant CNS-0615085: CSR-EHS: Enhancing the Effectiveness of Utilizing an Instruction Register File. We improved promoting instructions to reside in the IRF and adapted compiler optimizations to better utilize an IRF. We show that an IRF can decrease the execution time penalty of using an L0/filter cache while further reducing energy consumption. Finally, we introduce a tagless hit instruction cache that significantly reduces energy consumption without increasing execution time.

Embedded Multi-Processor System-on-chip (MPSoC) Design Considering Process Variations

Feng Wang and Yuan Xie
Pennsylvania State University, University Park, PA, 16802
Email:fenwang, yuanxie @cse.psu.edu

As technology scales, the delay uncertainty caused by process variations has become increasingly pronounced in deep submicron designs. As a result, a paradigm shift from deterministic to statistical design methodology at all levels of the design hierarchy is inevitable. As part of two NSF projects (NSF CNS CAREER 0643902 and NSF CNS 0720659), a variation-aware task allocation and scheduling method for Multiprocessor System-on-Chip (MPSoC) architectures is proposed, to mitigate the impact of parameter variations. A new design metric, called performance yield and defined as the probability of the assigned schedule meeting the predefined performance constraints, is used to guide the task allocation and scheduling procedure. An efficient yield computation method for task scheduling complements and significantly improves the effectiveness of the proposed variation-aware scheduling algorithm. Experimental results show that our variation-aware scheduler achieves significant yield improvements.

Improving Code Caching Performance for Java Applications

Yu Sun and Wei Zhang
Dept of ECE, Southern Illinois University Carbondale, Carbondale IL 62901
{sunyu,zhang}@engr.siu.edu

Java applications rely on Just-In-Time (JIT) compilers or adaptive compilers to generate and optimize binary code at runtime to boost performance. In conventional Java Virtual Machines (JVM), however, the binary code is typically written into the data cache, and then is loaded into the instruction cache through the shared L2 cache or memory, which is not efficient. In this paper, we study different efficient code caching strategies to write and read the dynamically generated code faster. Our experimental results indicate that writing code directly into the instruction cache can improve the performance of Java application by up to 42.9%, with 9.6% on average.

Efficient Software Checking for Fault Tolerance

Jing Yu, María Jesús Garzarán and Marc Snir
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N.Goodwin ave, Urbana, IL 61820
{jingyu, garzaran, snir}@cs.uiuc.edu

Dramatic increases in the number of transistors that can be integrated on a chip make processors more susceptible to radiation-induced transient errors. For commodity chips which are cost- and energy-constrained, software approaches can play a major role for fault detection because they can be tailored to fit different requirements of reliability and performance. However, software approaches add a significant performance overhead because they replicate the instructions and add checking instructions to compare the results. In order to make software checking approaches more attractive, we use compiler techniqes to identify the "unnecessary" replicas and checking instructions. In this paper, we present three techniques. The first technique uses boolean logic to identify code patterns that correspond to outcome tolerant branches. The second technique identifies address checks before loads and stores that can be removed with different degrees of fault coverage. The third technique identifies the checking instructions and shadow registers that are unnecessary when the register file is protected in hardware. By combining the three techniques, the overheads of software approaches can be reduced by an average 50%.

Software Reliability with Architectural Uncertainties

Lance Fiondella and Swapna S. Gokhale
Dept. of Computer Science and Engineering
Univ. of Connecticut, Storrs, CT 06269
{lfiondella,ssg}@engr.uconn.edu

Architecture-based software reliability analysis can provide early identification of critical components which can then be targeted for cost-effective reliability improvement of the application. However, an important challenge in conducting this analysis early in the life cycle is that it is nearly impossible to estimate the architectural and component parameters with certainty. The issue of estimating software application reliability in the presence of uncertain component reliabilities has been addressed in the previous research. In this paper we consider the estimation of software reliability in the presence of architectural uncertainties. We present a methodology to estimate the confidence levels in the architectural parameters using limited testing or simulation data based on the theory of confidence intervals of the multinomial distribution. The sensitivity of the system reliability to uncertain architectural parameters can then be quantified by varying these parameters within their confidence intervals. The illustration of the methodology using a case study indicates that the impact of the uncertainty in a given architectural parameter on the overall application reliability is determined by the inherent branching behavior of the application and the component reliabilities.

Runtime Verification Methods for MPI

Ganesh L. Gopalakrishnan and Robert M. Kirby
School of Computing, University of Utah,
Salt Lake City, UT 84112
fganesh, kirbyg@cs.utah.edu
http://www.cs.utah.edu/formal verification

The Gauss group at the University of Utah has researched and developed runtime verification tools for MPI programs. Our tool, in-situ partial order (ISP), is being applied to several MPI benchmarks. At the same time, we are embarked on research that ensures the completeness of ISP. Our work on specifying the formal semantics of MPI has also encompassed MPI 2.0. These developments and our plans for our final (fourth) year are elaborated in this paper.

Scalable Dynamic Information Flow Tracking and its Applications

Rajiv Gupta, Neelam Gupta, Xiangyu Zhang
Dennis Jeffrey, Vijay Nagarajan, Sriraman Tallam and Chen Tian
UC Riverside & Univ. of Arizona & Purdue University

We are designing scalable dynamic information flow tracking techniques and employing them to carry out tasks related to debugging (bug location and fault avoidance), security (software attack detection), and data validation (lineage tracing of scientific data). The focus of our ongoing work is on developing online dynamic analysis techniques for long running multithreaded programs that may be executed on a single core or on multiple cores to exploit thread level parallelism.

Enhancing Application Robustness through Adaptive Fault Tolerance

Zhiling Lan, Yawei Li, Ziming Zheng and Prashasta Gujrati
Illinois Institute of Technology, Department of Computer Science
{lan,liyawei,zzheng11,gujrpra}@iit.edu

As the scale of high performance computing (HPC) continues to grow, application fault resilience becomes crucial. To address this problem, we are working on the design of an adaptive fault tolerance system for HPC applications. It aims to enable parallel applications to avoid anticipated failures via preventive migration, and in the case of unforeseeable failures, to minimize their impact through selective checkpointing. Both prior and ongoing work are summarized in this paper.

Improving Software Reliability and Productivity via Mining Program Source Code

Tao Xie, Mithun Acharya, Suresh Thummalapenta and Kunal Taneja
Department of Computer Science
North Carolina State University
Raleigh NC USA 27695
{xie, acharya}@csc.ncsu.edu, {sthumma, ktaneja}@ncsu.edu

A software system interacts with third-party libraries through various APIs. Insufficient documentation and constant refactorings of third-party libraries make API library reuse difficult and error prone. Using these library APIs often needs to follow certain usage patterns. These patterns aid developers in addressing commonly faced programming problems such as what checks should precede or follow API calls, how to use a given set of APIs for a given task, or what API method sequence should be used to obtain one object from another. Ordering rules (specifications) also exist between APIs, and these rules govern the secure and robust operation of the system using these APIs. These patterns and rules may not be well documented by the API developers. Furthermore, usage patterns and specifications might change with library refactorings, requiring changes in the software that reuse the library. To address these issues, we develop novel techniques (and their supporting tools) based on mining source code, assisting developers in productively reusing third party libraries to build reliable and secure software.

STAMP: A Universal Algorithmic Model for Next-Generation Multithreaded Machines and Systems

Michel Dubois
University of Southern California
Dept. of Electrical Engineering
Los Angeles, CA 90089-2562, USA
dubois@paris.usc.edu

Hyunyoung Lee
University of Denver
Dept. of Computer Science
Denver, CO 80208, USA
hlee@cs.du.edu

We propose a generic algorithmic model called STAMP (Synchronous, Transactional, and Asynchronous Multi- Processing) as a universal performance and power complexity model for multithreaded algorithms and systems. We provide examples to illustrate how to design and analyze algorithms using STAMP and how to apply the complexity estimates to better utilize CMP(Chip MultiProcessor)-based machines within given constraints such as power

Methodologies for Optimization of Distributed Algorithms and Middleware

Gurdip Singh, Valeriy Kolesnikov and Sanghamitra Das
234 Nichols Hall, Computing and Information Sciences
Kansas State University
Manhattan, KS 66506

Many frameworks have been proposed which provide distributed algorithms encapsulated as middleware services to simplify application design. The developers of such algorithms are faced with two opposing forces. One is to design generic algorithms that are reusable in a large number of applications. Efficiency considerations, on the other hand, force the algorithms to be customized to specific operational contexts. This problem is often attacked by simply re-implementing all or large portions of an algorithm. The goals of this project are to (a) Develop techniques to design distributed algorithms amenable to customization, (b) Develop infrastructure to analyze applications and the target platform specifications to determine when customization can take place, (c) Develop tools to perform the necessary code transformation and compositions to carry out the identified optimizations in the algorithms. We are pursuing a number of different approaches to achieve these goals. We have developed two frameworks, FRAMES and InDiGO, which provides tools and techniques to perform optimization following two different approaches. We have performed experimental studies to demonstrate the advantages of our infrastructure. Another approach based on proof-based optimization is also being pursued. This paper highlights the results from these efforts and discuss our future work plans to apply these techniques to distributed sensor network middleware.

Versatile Models of Systems Using MAP Queueing Networks

Giuliano Casale, Ningfang Mi and Evgenia Smirni
College of William and Mary
Department of Computer Science
Williamsburg, VA
{casale, ningfang, esmirni}@cs.wm.edu

Analyzing the performance impact of temporal dependent workloads on hardware and software systems is a challenging task that yet must be addressed to enhance performance of real applications. For instance, existing matrix-analytic queueing models can capture temporal dependence only in systems that can be described by one or two queues, but the capacity planning of real multi-tier architectures requires larger models with arbitrary topology. To address the lack of a proper modeling technique for systems subject to temporal dependent workloads, we introduce a class of closed queueing networks where service times can have non-exponential distribution and accurately approximate temporal dependent features such as short or long range dependence. We describe these service processes using Markovian Arrival Processes (MAPs), which include the popular Markov-Modulated Poisson Processes (MMPPs) as special cases. Using a linear programming approach, we obtain for MAP closed networks tight upper and lower bounds for arbitrary performance indexes (e.g., throughput, response time, utilization). Numerical experiments indicate that our bounds achieve a mean accuracy error of 2% and promote our modeling approach for the accurate performance analysis of real multi-tier architectures.

Matchmaking and Implementation Issues for a P2P Desktop Grid

Michael Marsh[1], Jik-Soo Kim[1], Beomseok Nam[1], Jaehwan Lee[1], San Ratanasanya[1],
Bobby Bhattacharjee[1], Peter Keleher[1], Derek Richardson[2], Dennis Wellnitz[2] and Alan Sussman[1]
[1]UMIACS and Department of Computer Science
[2]Department of Astronomy
University of Maryland, College Park, MD 20742
[1]{mmarsh,jiksoo,bsnam,jhlee,san,bobby,keleher,als}@cs.umd.edu
[2]{dcr,wellnitz}@astro.umd.edu

We present some recent and ongoing work in our decentralized desktop computing grid project. Specifically, we discuss matching jobs with compute nodes in a peer-to-peer grid of heterogeneous platforms, and the implementation of our algorithms in a concrete system.

Improving I/O Performance through Compiler-Directed Code Restructuring and Adaptive Prefetching

Seung Woo Son, Mahmut Kandemir
Dept. of Computer Science and Engineering
The Pennsylvania State University
{sson,kandemir}@cse.psu.edu

Mustafa Karakoy
Department of Computing
Imperial College
mk22@doc.ic.ac.uk

Due to the huge and continuously increasing disparity between CPU speeds and disk access latencies, large high-performance applications that exercise disks tend to waste a disproportionate percentage of their execution times waiting for disk requests to complete. This paper presents two compiler-directed approaches to improving the I/O performance of data-intensive applications: code restructuring for disk reuse maximization and adaptive I/O prefetching. The compiler-directed code restructuring improves I/O performance by reducing the number of disk accesses through increasing disk reuse. That is, the data in a given set of disks are reused as much as possible before moving onto other disks. Adaptive I/O prefetching, on the other hand, is motivated by the observation that the effectiveness of compiler-directed I/O prefetching reduces significantly due to harmful prefetches when multiple CPUs share the same set of disks. To reduce intra- and inter-CPU harmful prefetches, our adaptive I/O prefetching scheme obtains inter-thread data sharing patterns through profiling and, based on the extracted sharing patterns, divides the threads into clusters and assigns a dedicated I/O prefetcher thread to each cluster. Our experimental results clearly show that both these approaches improve the I/O performance dramatically over the conventional data locality oriented schemes and compiler-directed I/O prefetching schemes.

VT-ASOS: Holistic System Software Customization for Many Cores

Dimitrios S. Nikolopoulos, Godmar Back, Jyotirmaya Tripathi and Matthew Curtis-Maury
Department of Computer Science, Virginia Tech
dsn@cs.vt.edu, gback@cs.vt.edu, tripathi@cs.vt.edu, mfcurt@cs.vt.edu

VT-ASOS is a framework for holistic and continuous customization of system software on HPC systems. The framework leverages paravirtualization technology. VT-ASOS extends the Xen hypervisor with interfaces, mechanisms, and policies for supporting application-specific resource management schemes on many-core systems, while retaining the advantages of virtualization, including protection, performance isolation, and fault tolerance. We outline the VTASOS framework and present results from a preliminary prototype, which enables static customization of scheduler parameters and runtime adaptation of parallel virtual machines.

Experiences in Building a Multi-University Testbed for Research in Multimedia Communications

Ram Dantu
Network Security Laboratory
Department of Computer Science & Engineering
University of North Texas
Denton, Texas 76203 USA
rdantu@unt.edu

Henning Schulzrinne
Department of Computer Science
Columbia University
New York, New York 10027, USA
hgs@cs.columbia.edu

Paul Sroufe
Network Security Laboratory
Department of Computer Science & Engineering
University of North Texas
Denton, Texas 76203 USA
prs0010@unt.edu

The next generation Internet needs to support multimedia services such as Voice/Video over IP (VoIP) and IP-based television (IPTV) and media distribution. The goal of the project is to develop a testbed for understanding and analysis of vulnerabilities of IP-based multimedia networks. This is a collaborative project between University of North Texas, Columbia University, Purdue University, and University of California at Davis. This project was awarded in 2006 and in one year, we have developed a multi-university testbed with variety of PBX solutions including Asterisk, a multitude of voice and video IP phones, and multiple universities were connected through a secure VPN and Internet2. As with any endeavor, we have many unique experiences and issues, which sometimes cause setbacks. We have dealt with issues concerning interoperability between phones and servers, Network Address Translation (NAT) connectivity issues, and many other collaboration issues concerning people and technology. An interworking of students from multiple universities, faculty, system administrators and support personnel has brought the testbed resources together for a working environment. In this paper we described our progress and experiences in more detail and how to fulfill our mission statement while supporting the various collaborative efforts involved with this project. In addition, we described the current research activities based on the testbed. Finally we described the next steps in the testbed development.

Multidimensional Information Retrieval in Peer-to-Peer Networks

Duc A. Tran K. Nguyen
Department of Computer Science
University of Massachusetts, Boston, MA 02125
Email: duc@cs.umb.edu

This paper summarizes a key finding in our NSF project CNSC0615055/07530661. The project is aimed at development and implementation of a complete architectural design for information retrieval in decentralized networks. In this paper, we particularly focus on two main components: the communication architecture and the indexing architecture. The communication component serves as the foundation for the nodes to maintain connectivity in the system, while the indexing component prepares the information in a way that enables fast and efficient future search operations. A challenge is to overcome the curse of dimensionality. From both network and database perspectives, decentralized networks are complex; not only we have to accommodate the rapid growth of the network size, but also the multi-dimensionality of the data space. We describe our solution in the context of P2P networks, but it is also applicable to any Internet-based decentralized networks.

Project Status: RIVER: Resource Management Infrastructure for Consolidated Hosting in Virtualized Data Centers

Bhuvan Urgaonkar and Anand Sivasubramaniam
Department of Computer Science and Engineering
The Pennsylvania State University, University Park, PA 16802
bhuvan@cse.psu.edu and anand@cse.psu.edu

This paper describes the status of the NSFfunded project titled “River: Resource Management Infrastructure for Consolidated Hosting in Virtualized Data Centers” (CNS-0720456.)

Quality Assurance and Adaptation of Internet Services: Early Experience

Cheng-Zhong Xu
Department of Electrical and Computer Engineering
Wayne State University, Detroit, MI 48202
czxu@wayne.edu

End-to-end quality of web service is determined by both network delays and server response time with respect to web pages. A web page often contains an HTML base and multiple static or dynamic objects that are embedded in the base. This paper presents an overview of our recent research experience with server resource management for client-perceived pageview response time assurance and adaptation. The main thrusts are two feedback control approaches that regulate server resource allocation at run-time in response to the change of input traffic and server load conditions. One is a model predictive feedback controller and the other is model-free self-tuning fuzzy controller. Their effectiveness and sub-second control accuracy are demonstrated on PlanetLab.

Towards Energy-Efficient Code Dissemination in Wireless Sensor Networks

Youtao Zhang [1], Jun Yang [2] and Weijia Li [1]

[1] Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260

[2] Electrical and Computer Engineering Department
University of Pittsburgh
Pittsburgh, PA 15261

Post-deployment code dissemination has become an important design issue for many applications in wireless sensor networks (WSNs). While several dissemination protocols have been proposed, challenges still exist due to the high energy consumption of transmitting wireless signals. In this paper, we present update-conscious compilation (UCC) techniques for energy-efficient code dissemination in WSNs. An update-conscious compiler, when compiling the modified code, includes the compilation decisions that were made when generating the old binary. In most cases, matching the previous decisions improves the binary code similarity, reduces the amount of data to be transmitted to remote sensors, and thus, consumes less energy. In this paper, we focus on the development of update-conscious register allocation (UCCRA) algorithms. Our experimental results show that UCCRA can achieve great improvements over the traditional, update-oblivious approaches.

Failure Prediction in IBM BlueGene/L Event Logs

Yanyong Zhang
ECE Department, Rutgers University
{yyzhang}@ece.rutgers.edu

Anand Sivasubramaniam
CSE Department, Penn State University
anand@cse.psu.edu

In this paper, we present our effort in developing a failure prediction model based on event logs collected from IBM BlueGene/L. We first show how the event records can be converted into a data set that is appropriate for running classification techniques. Then we apply classifiers on the data, including RIPPER (a rule-based classifier), Support Vector Machines (SVMs), a traditional Nearest Neighbor method, and a customized Nearest Neighbor method. We show that the customized nearest neighbor approach can outperform RIPPER and SVMs in terms of both coverage and precision. The results suggest that the customized nearest neighbor approach can be used to alleviate the impact of failures.

Domain-Specific Library Generation for Parallel Software and Hardware Platforms

Franz Franchetti[1], Yevgen Voronenko[1], Peter A. Milder[1], Srinivas Chellappa[1],
Marek R. Telgarsky[1], Hao Shen[2], Paolo D'Alberto[3], Frédéric de Mesmay[1]
James C. Hoe[1], José M. F. Moura[1] and Markus Püschel[1]
[1]Carnegie Mellon University, [2]Technical University of Denmark, [3]Yahoo Inc.

We overview a library generation framework called Spiral. For the domain of linear transforms, Spiral automatically generates implementations for parallel platforms including SIMD vector extensions, multicore processors, field-programmable gate arrays (FPGAs) and FPGA accelerated processors. The performance of the generated code is competitive with the best available hand-written libraries.

Adaptive Tuning in a Dynamically Changing Resource Environment

Seyong Lee and Rudolf Eigenmann
School of ECE, Purdue University
West Lafayette, IN, 47907
{lee222,eigenman}@purdue.edu

We present preliminary results of a project to create a tuning system that adaptively optimizes programs to the underlying execution platform. We will show initial results from two related efforts. (i) Our tuning system can efficiently select the best combination of compiler options, when translating programs to a target system. (ii) By tuning irregular applications that operate on sparse matrices, our system is able to achieve substantial performance improvements on cluster platforms. This project is part of a larger effort that aims at creating a global information sharing system, where resources, such as software applications, computer platforms, and information can be shared, discovered, and adapted to local needs.

Automatic Generation of a Parallel Sorting Algorithm

Brian A. Garber, Dan Hoeflinger, Xiaoming Li, María Jesús Garzarán and David Padua
Department of C.S.
University of Illinois at Urbana-Champaign
{garber,hoefling, xli15, garzaran, padua}@cs.uiuc.edu

In this paper, we discuss a library generator for parallel sorting routines that examines the input characteristics (and the parameters they affect) to select the best performing algorithm. Our preliminary experimental results show that the automatic generation of a distributed memory parallel sorting routine provides up to a four fold improvement over standard parallel algorithms with typical parameters. With the recent importance of multicore processors, we are extending this work to shared memory. This provides new challenges specific to multicore systems. However, with their increasing popularity, this extension becomes very valuable.

Towards Effective Automatic Parallelization for Multicore Systems

Uday Bondhugula[1], Muthu Baskaran[1], Albert Hartono[1], Sriram Krishnamoorthy[1],
J. Ramanujam[2], Atanas Rountev[1] and P. Sadayappan[1]
[1]Dept. of Computer Science and Engineering
The Ohio State University
2015 Neil Ave. Columbus, OH, USA
{bondhugu,baskaran,hartonoa,krishnsr,rountev,saday}@cse.ohio-state.edu
[2]Dept. of Electrical & Computer Engg. and
Center for Computation & Technology
Louisiana State University
jxr@ece.lsu.edu

The ubiquity of multicore processors in commodity computing systems has raised a significant programming challenge for their effective use. An attractive but challenging approach is automatic parallelization of sequential codes. Although virtually all production C compilers have automatic shared-memory parallelization capability, it is rarely used in practice by application developers because of limited effectiveness. In this paper we describe our recent efforts towards developing an effective automatic parallelization system that uses a polyhedral model for data dependences and program transformations.

Adaptive Speculation in Behavior-Oriented Parallelization

Yunlian Jiang and Xipeng Shen
Computer Science Department
The College of William and Mary, Williamsburg, VA, USA
{jiang,xshen}@cs.wm.edu

Behavior-oriented parallelization is a technique for parallelizing complex sequential programs that have dynamic parallelism. Although the technique shows promising results, the software speculation mechanism it uses is not cost-efficient. Failed speculations may waste computing resource and severely degrade system efficiency. In this work, we propose adaptive speculation to predict the profitability of a speculation and dynamically enable or disable the speculation of a region. Experimental results demonstrate the effectiveness of the scheme in improving the efficiency of software speculation. In addition, the adaptive speculation can also enhance the usability of behavior-oriented parallelization by allowing users to label potential parallel regions more flexibly.

Towards Universal Code Generator Generation

Timothy Richards, Edward K. Walters II, J. Eliot B. Moss, Trek Palmer and Charles C. Weems
Department of Computer Science, University of Massachusetts Amherst

One of the most difficult tasks a compiler writer faces is the construction of the code generator. The code generator is that part of the compiler that translates compiler intermediate representation (IR) into instructions for a target machine. Unfortunately, implementing a code generator "by hand" is a difficult, time consuming, and error prone task. The details of both the IR and target instruction set must be carefully considered in order to generate correct and efficient code. This, in turn, requires an expert in both the compiler internals as well as the target machine. Even an expert, however, can produce a code generator that is difficult to verify and debug.

In this paper we present a universal approach for automating the construction of correct code generators. In particular, we show that both the compiler IR and target instruction set semantics can be described by a machine description language and leveraged by a heuristic search procedure to derive code generator patterns. We then utilize formal methods to determine if the IR and target sequence pairs that make up these patterns are semantically equivalent.

Automatic Construction of Coordinated Performance Skeletons

Jaspal Subhlok and Qiang Xu
University of Houston
Department of Computer Science
Houston, TX 77204

Performance prediction is particularly challenging for dynamic and unpredictable environments that cannot be modeled well, such as execution with sharing of CPU and bandwidth resources. Our approach to performance estimation in such scenarios is based on actual execution of short running customized performance skeletons for target applications. This work focuses on automatic construction of performance skeletons for parallel MPI programs. Logicalization of a family of traces to a single trace is presented as a key technique for skeleton construction. Compression of communication traces is achieved by identifying the loop structure from traces. Results are presented that demonstrate that logicalization and compression are accurate and efficient. Automatically constructed performance skeletons were able to effectively predict application performance in a variety of scenarios involving resource sharing and changes in the execution environment.

Immersive Real-Time Large-Scale Network Simulation: A Research Summary

Jason Liu
School of Computing and Information Sciences
Florida International University
Miami, Florida 33199
liux@cis.fiu.edu

Immersive real-time large-scale network simulation is a technique that supports simulation of large-scale networks to interact with real implementations of network protocols, network services, and distributed applications. Traffic generated by real network applications is carried by the virtual network simulated in real time where delays and losses are calculated according to the simulated network conditions. We emphasize network immersion so that the virtual network is indistinguishable from a physical testbed in terms of network behavior, yet allows the flexibility of simulation. In this paper we provide a summary of current research in immersive real-time large-scale network simulation, particularly in areas of hybrid network traffic modeling and scalable emulation infrastructure design.

Parallelizing Computer System Simulators

Derek Chiou, Dam Sunwoo, Hari Angepat, Joonsoo Kim, Nikhil A. Patil,
William Reinhart and D. Eric Johnson
The University of Texas at Austin
{derek,sunwoo,angepat,turo,npatil,wreinhar,dejohnso}@ece.utexas.edu

This paper describes NSF-supported work in parallelized computer system simulators being done in the Electrical and Computer Engineering Department at the University of Texas at Austin. Our work is currently following two paths: (i) the FAST simulation methodology that is capable of simulating complex systems accurately and quickly (currently about 1.2MIPS executing the x86 ISA, modeling an out-of-order superscalar processor and booting Windows XP and Linux) and (ii) the RAMP-White (White) platform that will soon be capable of simulating very large systems of around 1000 cores. We plan to combine the projects to provide fast and accurate simulation of multicore systems.

We Have It Easy, But Do We Have It Right?

Todd Mytkowicz and Amer Diwan
University of Colorado
{mytkowicz,diwan}@colorado.edu

Matthias Hauswirth
University of Lugano
Matthias.Hauswirth@gmail.com

Peter Sweeney
IBM Research
pfs@us.ibm.com

We show two severe problems with the state of the art in empirical computer system performance evaluation, observer effect and measurement context bias, and we outline the path toward a solution.

A System-Level Perspective for Efficient NoC Design

Amit Kumar, Niket Agarwal, Li-Shiuan Peh and Niraj K. Jha
Dept. of Electrical Engineering, Princeton University, Princeton, NJ 08544
famitk, niketa, peh, jhag@princeton.edu

With the advent of chip multiprocessors (CMPs) in mainstream systems, the on-chip network that connects different processing cores becomes a critical part of the design. There has been significant work in the recent past on designing these networks for efficiency and scalability. However, most network design evaluations use a stand-alone network simulator which fails to capture the system-level implications of the design. New design innovations, which might yield promising results when evaluated using such stand-alone models, may not look that attractive when evaluated in a full-system simulation framework.

In this work, we present GARNET, a detailed network model incorporated inside a full-system simulator which enables system-level performance and power modeling of network-level techniques. GARNET also facilitates accurate evaluation of techniques that simultaneously leverage the memory hierarchy as well as the interconnection network. We also discuss express virtual channels, a novel flow control technique which improves network energy/delay by creating virtual lanes in the network along which packets can bypass intermediate routers.

Improving Predictions forWater Spills Using DDDAS

Craig C. Douglas
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
craig.douglas@uky.edu

Paul Dostert
Department of Mathematics
University of Arizona
Tucson, AZ 85721-0089
dostert@math.arizona.edu

Yalchin Efendiev and Richard E. Ewing
Department of Mathematics
Texas A & M University
College Station, TX 77843-3368
efendiev@math.tamu.edu

Deng Li
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
li.deng@uky.edu

In dynamic data driven application systems, the predictions are improved based on measurements obtained in time. Predicted quantity often satisfies differential equation models with unknown initial conditions and source terms. A physical example of the problem we are attempting to solve is a major waste spill near a body of water. This can be, for example, near an aquifer, or possibly in a river or bay. Sensors can be used to measure where the contaminant was spilled, where it is, and where it will go. In this paper, we propose techniques for improving predictions by estimating initial conditions and source terms. We show how well we can solve the problem for a variety of data-driven models.

Designing and Parameterizing a Workflow for Optimization: A Case Study in Biomedical Imaging

Vijay S. Kumar[2], Mary Hall[1], Jihie Kim[1], Yolanda Gil[1],
Tahsin Kurc[2], Ewa Deelman[1], Varun Ratnakar[1] and Joel Saltz[2]

[1] University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

[2]Department of Biomedical Informatics
The Ohio State University
333 West 10th Avenue
Columbus, OH 43210

This paper describes our experience to date employing the systematic mapping and optimization of large-scale scientific application workflows to current and future parallel platforms. The overall goal of the project is to integrate a set of system layers – application program, compiler, run-time environment, knowledge representation, optimization framework, and workflow manager – and through a systematic strategy for workflow mapping, our approach will exploit the vast machine resources available in such parallel platforms to dramatically increase the productivity of application programmers. In this paper, we describe the representation of a biomedical imaging application as a workflow, our early experiences in integrating the set of tools brought together for this project, and implications for future applications.

A Wireless Sensor Network Based Closed-Loop System for Subsurface Contaminant Plume Monitoring

Qi Han
Dept. of Math and Computer Sciences
Colorado School of Mines
qhan@mines.edu

Anura P. Jayasumana
Dept. of Electrical and Computer Engineering
Colorado State University, Fort Collins, CO 80523
Anura.Jayasumana@Colostate.edu

Tissa Illangaskare and Toshihiro Sakaki
Division of Environmental Science and Engineering
Colorado School of Mines, Golden, CO 80401
{tissa,tsakaki}@mines.edu

A closed-loop contaminant plume monitoring system is being developed that integrates wireless sensor network based monitoring with numerical models for subsurface plumes. The system is based on a novel virtual sensor network architecture that supports the formation, usage, adaptation, and maintenance of dynamic subsets of nodes. This automated monitoring system is intended to capture transient plumes to assess the source, track plumes in real-time, and predict future plumes behavior using numerical models that can be continuously re-calibrated by sensor data. This paper presents recent progress made in (1) developing a proof-of-concept study using a porous media test bed with sensors deployed; and (2) developing distributed algorithms for virtual sensor networking.

Towards a Real-Time Data DrivenWildland Fire Model

Jan Mandel
Department of Mathematical Sciences
University of Colorado Denver
Denver, CO 80217-3364
Jan.Mandel@gmail.com

Jonathan D. Beezley
Department of Mathematical Sciences
University of Colorado Denver
Denver, CO 80217-3364
jon.beezley.math@gmail.com

Soham Chakraborty
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
sohaminator@gmail.com

Janice L. Coen
Mesoscale and Microscale Meteorology Division
National Center for Atmospheric Research
Boulder, CO 80307-3000
janicec@ucar.edu

Craig C. Douglas
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
craig.c.douglas@gmail.com

Anthony Vodacek
Center for Imaging Science
Rochester Institute of Technology
Rochester, NY 14623
vodacek@cis.rit.edu

Zhen Wang
Center for Imaging Science
Rochester Institute of Technology
Rochester, NY 14623
zxw7546@cis.rit.edu

A wildland fire model based on semi-empirical relations for the spread rate of a surface fire and post-frontal heat release is coupled with the Weather Research and Forecasting atmospheric model (WRF). The propagation of the fire front is implemented by a level set method. Data is assimilated by a morphing ensemble Kalman filter, which provides amplitude as well as position corrections. Thermal images of a fire will provide the observations and will be compared to a synthetic image from the model state.

Programming Support for Sensor-based Scientific Applications

Nanyan Jiang and Manish Parashar
The Applied Software Systems Laboratory
Department of Electrical and Computer Engineering
Rutgers University, Piscataway NJ 08855, USA
{nanyanj, parashar}@caip.rutgers.edu

Technical advances are enabling a pervasive computational ecosystem that integrates computing infrastructures with embedded sensors and actuators, and are giving rise to a new paradigm for monitoring, understanding, and managing natural and engineered systems – one that is information/data-driven. This research investigates programming systems for sensor-driven applications. It addresses abstractions and runtime mechanisms for integrating sensor systems with computational models for scientific processes, as well as for innetwork data processing, e.g., aggregation, adaptive interpolation and assimilation. The current status of this research, as well as initial results are presented.

Development of Laboratory and Computational Techniques for Optimal and Quantitative Understanding of Cellular Metabolic Networks

Xiaojiang Feng[1], Joshua Rabinowitz[1,2] and Herschel Rabitz[1]
Department of Chemistry[1] and Lewis-Sigler Institute for Integrative Genomics[2], Princeton University, Princeton, New Jersey 08544
xfeng@princeton.edu, joshr@genomics.princeton.edu, hrabitz@princeton.edu

This paper summarizes the development of laboratory and computational techniques for systematic and reliable understanding of metabolic networks. By combining a filter-based cell culture system and an optimized metabolite extraction protocol, a broad array of cellular metabolites can be reliably quantified following nutrient and other environment perturbations. A nonlinear closed-loop procedure was also developed for optimal bionetwork model identification. Computational illustrations and laboratory applications clearly demonstrate the capabilities of these techniques in understanding cellular metabolism, especially when they are integrated in an optimal fashion.

ECG Segmentation in a Body Sensor Network Using Hidden Markov Models

Huaming Li and Jindong Tan
Department of Electrical and Computer Engineering
Michigan Technological University
Houghton, MI 49931
{lhuaming, jitan}@mtu.edu

A novel approach for segmenting ECG signal in a body sensor network employing Hidden Markov Modeling (HMM) technique is presented. The parameter adaptation in traditional HMM methods is conservative and slow to respond to these beat interval changes. Inadequate and slow parameter adaptation is largely responsible for the low positive predictivity rate. To solve the problem, we introduce an active HMM parameter adaptation and ECG segmentation algorithm. Body sensor networks are used to pre-segment the raw ECG data by performing QRS detection. Instead of one single generic HMM, multiple individualized HMMs are used. Each HMM is only responsible for extracting the characteristic waveforms of the ECG signals with similar temporal features from the same group, so that the temporal parameter adaptation can be naturally achieved.

Autonomic Power and Performance Management of High-performance Servers

Bithika Khargharia[1], Salim Hariri[1],
Wael Kdouh[2], Manal Houri[2], Hesham El-Rewini[2], Mazin Yousif[3]
[1]NSF Center for Autonomic Computing, University of Arizona, Tucson, AZ
[2]Southern Methodist University, Dallas, TX, [3]Intel Corp., Hillsboro, OR
E-mail: bithikak @email.arizona.edu

With the increased complexity of platforms coupled with data centers' servers sprawl, power consumption is reaching unsustainable limits. Researchers have addressed data centers' power & performance management at different hierarchies going from server clusters to servers to individual components within the server. This paper presents a novel technique for autonomic power & performance management of a high-performance server platform that consists of multi-core processor and multi-rank memory subsystems. Both the processor and/or the memory subsystem are dynamically reconfigured (expanded or contracted) to suit the application resource requirements. The reconfigured platform creates the opportunity for power savings by transitioning any unused platform capacity (processor/memory) into low-power states for as long as the platform performance remains within given acceptable thresholds. The platform power expenditure is minimized subject to platform performance parameters, which is formulated as an optimization problem. Our experimental results show around 58.33% savings in power as compared to static power management techniques.

Energy-efficient High-performance Storage System

Jun Wang, (juwang@mail.ucf.edu)
Office 407-823-0449; Fax 407-823-5835
School of Electrical Engineering and Computer Science
University of Central Florida, Orlando, Florida, 32816

The energy consumed by the high-performance storage subsystems increases the cost of operating computing systems for I/O intensive applications, both directly and indirectly. The development of energy-efficient storage systems can greatly reduce the cost of acquiring and operating large computing platforms, thereby making them cost-effective for a broad range of computational problems. This paper develops energy-efficient, high-performance, conventional disk based arrays called EEAID including RAID-1 and RAID-5 by exploiting existing redundancy at both cache and disk levels. In addition, we give a complete solution of dynamic disk power management and performance control for EERAID.

Using Game Theory for Scheduling Tasks on Multi-Core Processors for Simultaneous Optimization of Performance and Energy

Ishfaq Ahmad
Department of Computer Science and Engineering
University of Texas at Arlington
iahmad@cse.uta.edu

Sanjay Ranka
Department of Computer Science
University of Florida
ranka@cis.ufl.edu

Samee Ullah Khan
Department of Electrical and Computer Engineering, Colorado State University
samee.khan@colostate.edu

Multi-core processors are beginning to revolutionize the landscape of high-performance computing. In this paper, we address the problem of power-aware scheduling/mapping of tasks onto heterogeneous and homogeneous multi-core processor architectures. The objective of scheduling is to minimize the energy consumption as well as the makespan of computationally intensive problems. The multi-objective optimization problem is not properly handled by conventional approaches that try to maximize a single objective. Our proposed solution is based on game theory. We formulate the problem as a cooperate game. Although we can guarantee the existence of a Bargaining Point in this problem, the classical cooperative game theoretical techniques such as the Nash axiomatic technique cannot be used to identify the Bargaining Point due to low convergence rates and high complexity. Hence, we transform the problem to a max-max-min problem such that it can generate solutions with fast turnaround time.

A Prefetching Scheme for Energy Conservation in Parallel Disk Systems

Adam Manzanares, Kiranmai Bellam and Xiao Qin
Department of Computer Science and Software Engineering
Auburn University
Auburn, Alabama 36830, USA
{acm0008, kzb0008, xqin}@auburn.edu

Large-scale parallel disk systems are frequently used to meet the demands of information systems requiring high storage capacities. A critical problem with these large-scale parallel disk systems is the fact that disks consume a significant amount of energy. To design economically attractive and environmentally friendly parallel disk systems, we developed two energy-aware prefetching strategies for parallel disk systems with disk buffers. First, we introduce a new buffer disk architecture that can provide significant energy savings for parallel disk systems while achieving high performance. Second, we design a prefetching approach to utilize an extra disk to accommodate prefetched data sets that are frequently accessed. Third, we develop a second prefetching strategy that makes use of an existing disk in the parallel disk system as a buffer disk. Compared with the first prefetching scheme, the second approach lowers the capacity of the parallel disk system. However, the second approach is more cost-effective and energy-efficient than the first prefetching technique. Finally, we quantitatively compare both of our prefetching approaches against two conventional strategies including a dynamic power management technique and a non-energy-aware scheme. Using empirical results we show that our novel prefetching approaches are able to reduce energy dissipation in parallel disk systems by 44% and 50% when compared against a non-energy aware approach. Similarly, our strategies are capable of conserving 22% and 30% of the energy when compared to the dynamic power management technique.

Managing Power, Performance and Reliability Trade-offs

Padma Raghavan, Mahmut Kandemir, Mary Jane Irwin and Konrad Malkowski
{raghavan, kandemir, mji, malkowsk}@cse.psu.edu
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802

We present recent research on utilizing power, performance and reliability trade-offs in meeting the demands of scientific applications. In particular we summarize results of our recent publications on (i) phase-aware adaptive hardware selection for power-efficient scientific computations, (ii) adapting application execution to reduced CPU availability, and (iii) a helper thread based EDP reduction scheme for adapting application execution in CMPs.

Measuring Power and Temperature from Real Processors

Francisco J. Mesa-Martinez, Michael Brown, Joseph Nayfach-Battilana and Jose Renau
Dept. of Computer Engineering, University of California Santa Cruz
http://masc.soe.ucsc.edu

The modeling of power and thermal behavior of modern processors requires challenging validation approaches, which may be complex and in some cases unreliable. In order to address some of the difficulties associated with the validation of power and thermal models, this document describes an infrared measurement setup that simultaneously captures run-time power consumption and thermal characteristics of a processor. We use infrared cameras with high spatial resolution ($10 \times 10\mu m$) and high frame rate (125Hz) to capture thermal maps. Power measurements are obtained with a multimeter at a sampling rate of 1KHz. The synchronized traces can then be used in the validation process of possible thermal and power processor activity models.

Adaptive Power Control for Server Clusters

Xiaorui Wang and Ming Chen
University of Tennessee, Knoxville, TN 37996
{xwang, mchen11}@eecs.utk.edu

Power control is becoming a key challenge for effectively operating a modern data center. In addition to reducing operation costs, precisely controlling power consumption is an essential way to avoid system failures caused by power capacity overload or overheating due to increasing high-density. Control-theoretic techniques have recently shown a lot of promise on power management thanks to their better control performance and theoretical guarantees on control accuracy and system stability. However, existing work oversimplifies the problem by controlling a single server independently from others. As a result, at the cluster level where multiple servers are correlated by common workloads and share common power supplies, power cannot be shared to improve application performance. In this paper, we propose a cluster-level power controller that shifts power among servers based on their performance needs, while controlling the total power of the cluster to be lower than a constraint. Our controller features a rigorous design based on an optimal multi-input-multi-output control theory. Empirical results demonstrate that our controller outperforms a state-of-the-art controller, by having better application performance and more accurate power control.

DOSA: Design Optimizer for Scientific Applications

David A. Bader[1] and Viktor K. Prasanna[2]
[1]College of Computing
Georgia Institute of Technology
Altanta, GA 30332
bader@cc.gatech.edu
[2]Department of EE-Systems
University of Southern California
Los Angeles, CA 90089
prasanna@ganges.usc.edu

In this paper we briefly introduce our new framework, called "Design Optimizer for Scientific Applications" (DOSA) which allows the programmer or compiler writer to explore alternative designs and optimize for speed (or power) at design-time and use a run-time optimizer. The run-time system is a portable interface that enables dynamic application optimization by interfacing with the output of DOSA. As an illustration we demonstrate speed up for two applications: Parallel Exact Inference and Community Identification in large-scale networks.

Data Mining on the Grid for the Grid

Nitesh V. Chawla, Douglas Thain, Ryan Lichtenwalter and David A. Cieslak
Department of Computer Science and Engineering
University of Notre Dame, USA
{nchawla, dthain, rlichten, dcieslak}@cse.nd.edu

Both users and administrators of computing grids are presented with enormous challenges in debugging and troubleshooting. Diagnosing a problem with one application on one machine is hard enough, but diagnosing problems in workloads of millions of jobs running on thousands of machines is a problem of a new order of magnitude. Suppose that a user submits one million jobs to a grid, only to discover some time later that half of them have failed, Users of large scale systems need tools that describe the overall situation, indicating what problems are commonplace versus occasional, and which are deterministic versus random. Machine learning techniques can be used to debug these kinds of problems in large scale systems. We present a comprehensive framework from data to knowledge discovery as an important step towards achieving this vision.

Toward Improved Tumor Targeting for Image Guided Neurosurgery with Intra-operative Parametric Search using Distributed and Grid Computing

Andriy Fedorov and Nikos Chrisochoides
Department of Computer Science
College of William and Mary
Williamsburg, VA 23185
{fedorov,nikos}@cs.wm.edu

We describe a high-performance distributed software environment for real-time nonrigid registration during Image-Guided Neurosurgery (IGNS). The implementation allows to perform volumetric non-rigid registration of MRI data within two minutes of computation time, and can enable large-scale parametric studies of the registration algorithm. We explore some of the parameters of the non-rigid registration, and evaluate their impact on registration accuracy using ground truth. The results of the evaluation motivate running the registration process within the Grid environment, particularly when searching for optimal parameters of intraoperative registration. Based on the results of our evaluation, distributed parametric searching of optimal registration settings can significantly improve the registration accuracy in some regions of the brain, as compared to the accuracy achieved using the default parameters.

Introducing Gravel: An MPI Companion Library

Anthony Danalis, Aaron Brown, Lori Pollock and Martin Swany
Department of Computer and Information Sciences
University of Delaware, Newark, DE 19716
{danalis, brown, pollock, swany}@cis.udel.edu

A non-trivial challenge in high performance, cluster computing is the communication overhead introduced by the cluster interconnect. A common strategy for addressing this challenge is the use of communication-computation overlapping. In this paper, we introduce Gravel, a portable communication library designed to inter-operate with MPI for improving communication-computation overlapping. Selected MPI calls are semi-automatically replaced by Gravel calls only in key locations in an application where performance is critical. Gravel separates the data transfers from the handshake messages, enabling application developers to utilize Remote Data Memory Access (RDMA) directly, whether the data exchange scheme of their application is pure one-sided, or a more traditional two-sided. The Gravel API is much simpler than existing low level libraries and details (i.e., pointers) are hidden from the application layer so Gravel is usable in FORTRAN applications. This paper presents an overview of Gravel.

Verifying Linked Data Structure Implementations

Karen Zee
MIT CSAIL
Cambridge, MA
kkz@csail.mit.edu

Viktor Kuncak
EPFL I&C
Lausanne, Switzerland
viktor.kuncak@epfl.ch

Martin Rinard
MIT CSAIL
Cambridge, MA
rinard@csail.mit.edu

The Jahob program verification system leverages state of the art automated theorem provers, shape analysis, and decision procedures to check that programs conform to their specifications. By combining a rich specification language with a diverse collection of verification technologies, Jahob makes it possible to verify complex properties of programs that manipulate linked data structures. We present our results using Jahob to achieve full functional verification of a collection of linked data structures.

A Stochastic Model for Robust Resource Allocation in Heterogeneous Parallel and Distributed Computing Systems

Jay Smith[1,2], Howard Jay Siegel[2,3] and Anthony A. Maciejewski[2]

[1]DigitalGlobe
Longmont, CO 80503 USA
Email: jtsmith@digitalglobe.com

Colorado State University
[2]Dept. of Electrical and Computer Engineering
[3]Dept. of Computer Science
Fort Collins, CO 80503-1373 USA
Email: {hj, aam}@engr.colostate.edu

This paper summarizes some of our research in the area of robust static resource allocation for distributed computing systems operating under imposed Quality of Service (QoS) constraints. Often, these systems are expected to function in a physical environment replete with uncertainty, which causes the amount of processing required over time to fluctuate substantially. Determining a resource allocation that accounts for this uncertainty in a way that can provide a probabilistic guarantee that a given level of QoS is achieved is an important research problem. The stochastic robustness metric described in this research is based on a mathematical model where the relationship between uncertainty in system parameters and its impact on system performance are described stochastically.

Steps Toward Activity-Oriented Computing

João Pedro Sousa[2], Vahe Poladian[1], David Garlan[1], Bradley Schmerl[1] and Peter Steenkiste[1]

[1] Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
garlan,schmerl,steenkiste,poladian@cs.cmu.edu

[2]George Mason University
Fairfax, VA 22030
jpsousa@gmu.edu

Most pervasive computing technologies focus on helping users with computer-oriented tasks. In this NSF-funded project, we instead focus on using computers to support user-centered “activities” that normally do not involve the use of computers. Examples may include everyday tasks around such as answering the doorbell or doing laundry. A focus on activity-based computing brings to the foreground a number of unique challenges. These include activity definition and representation, system design, interfaces for managing activities, and ensuring robust operation. Our project focuses on the first two challenges.

Software Monitoring with Bounded Overhead

Sean Callanan, Daniel J. Dean, Michael Gorbovitski, Radu Grosu, Justin Seyster,
Scott A. Smolka, Scott D. Stoller and Erez Zadok
Stony Brook University

In this paper, we introduce the new technique of High- Confidence Software Monitoring (HCSM), which allows one to perform software monitoring with bounded overhead and concomitantly achieve high confidence in the observed error rates. HCSM is formally grounded in the theory of supervisory control of finite-state automata: overhead is controlled, while maximizing confidence, by disabling interrupts generated by the events being monitored– and hence avoiding the overhead associated with processing these interrupts!for as short a time as possible under the constraint of a user-supplied target overhead Otarget. HCSM is a general technique for software monitoring in that HCSM-based instrumentation can be attached at any system interface or API. A generic controller implements the optimal control strategy described above. As a proof of concept, and as a practical framework for software monitoring, we have implemented HCSM-based monitoring for both bounds checking and memory leak detection. We have further conducted an extensive evaluation of HCSM’s performance on several real-world applications, including the Lighttpd Web server, and a number of special-purpose micro-benchmarks. Our results demonstrate how confidence grows in a monotonically increasing fashion with the target overhead, and that tight confidence intervals can be obtained for each target-overhead level.

Workshop 11

High-Performance, Power-Aware Computing

HPPAC 2008

Saving 200kW and $200K/year by Power-aware Job/Machine Scheduling

Junichi Hikita, Akio Hirano and Hiroshi Nakashima
Kyoto University
ACCMS North Bldg., Yoshida Hommachi, Kyoto
606-8501, JAPAN
{hikita, hirano, h.nakashima}@media.kyoto-u.ac.jp

This paper reports our 3.75-year empirical study on power-aware operations of Kyoto University's supercomputer system. The supercomputer system of 10 TFlops had required about 540 kW on average in its first fiscal year 2004. After that and one-year try-and-error of power efficient operation, we implemented a simple but effective scheduler of jobs and machine powering to improve the perload power efficiency by up to 39% and to save 200 kW and $200,000 electric charge in the fiscal year 2006. The power-aware scheduler tries to minimize the number of active nodes in the system of eleven nodes keeping sufficient computational power for given loads. Thus the poweraware scheduler has not degraded, but has significantly improved, the service quality in terms of the average job-waiting time.

Reducing Wasted Resources to Help Achieve Green Data Centers

Jordi Torres, David Carrera, Kevin Hogan, Ricard Gavaldà Vicenç Beltran, Nicolás Poggi
Barcelona Supercomputing Center (BSC) - Technical University of Catalonia (UPC)
torres@ac.upc.edu

In this paper we introduce a new approach to the consolidation strategy of a data center that allows an important reduction in the amount of active nodes required to process a heterogeneous workload without degrading the offered service level. This article reflects and demonstrates that consolidation of dynamic workloads does not end with virtualization. If energy-efficiency is pursued, the workloads can be consolidated even more using two techniques, memory compression and request discrimination, which were separately studied and validated in previous work and are now to be combined in a joint effort. We evaluate the approach using a representative workload scenario composed of numerical applications and a real workload obtained from a top national travel website. Our results indicate that an important improvement can be achieved using 20% less servers to do the same work. We believe that this serves as an illustrative example of a new way of management: tailoring the resources to meet high level energy efficiency goals.

Power Management of Distributed Web Servers by Controlling Server Power State and Traffic Prediction for QoS

Takayuki Imada, Mitsuhisa Sato, Yoshihiko Hotta and Hideaki Kimura
Graduate School of Systems and Information Engineering, University of Tsukuba
{imada, msato, hotta, kimura}@hpcs.cs.tsukuba.ac.jp

In this paper, we propose a scheme based on server node state control, including stand-by/wake-up and processor power control, to achieve aggressive power eduction while satisfying Quality of Service (QoS). Decreasing power consumption on Web servers is currently a challenging new problem to be solved in a data center or warehouse. Although Web servers are configured to have maximum performance, the actual access rate to the servers can be small in a specific period, such as midnight, so it may be possible to reduce the power consumption of the servers while satisfying QoS with lower server performance. In order to reduce power consumption on the server nodes, we now have to consider the power consumption of the entire node rather than only processor power by Dynamic Voltage and Frequency Scaling (DVFS). We implemented the proposed scheme to the distributed Web server system using the power-profile of server nodes and considering load increment based on traffic prediction method and evaluated the proposed scheme with a Web server benchmark workload based on SPECWeb99. The result reveals that the proposed scheme achieved an energy saving of approximately 17% with sufficient QoS performance on the distributed Web server system.

A Simple Power-Aware Scheduling for Multicore Systems When Running Real-Time Applications

Diana Bautista, Julio Sahuquillo, Houcine Hassan, Salvador Petit and José Duato
Department of Computer Engineering (DISCA)
Universidad Politécnica de Valencia, Spain
diabaura@fiv.upv.es, {jsahuqui,husein,spetit}@disca.upv.es

High-performance microprocessors, e.g., multithreaded and multicore processors, are being implemented in embedded real-time systems because of the increasing computational requirements. These complex microprocessors have two major drawbacks when they are used for real-time purposes. First, their complexity difficults the calculation of the WCET (Worst Case Execution Time). Second, power consumption requirements are much larger, which is a major concern in these systems.

In this paper we propose a novel soft power-aware real-time scheduler for a state-of-the-art multicore multithreaded processor, which implements dynamic voltage scaling techniques. The proposed scheduler reduces the energy consumption while satisfying the constraints of soft real-time applications. Different scheduling alternatives have been evaluated, and experimental results show that using a fair scheduling policy, the proposed algorithm provides, on average, energy savings ranging from 34% to 74%.

System-level, Thermal-aware, Fully-loaded Process Scheduling

Dong Li, Hung-Ching Chang, Hari K. Pyla, Kirk W. Cameron
Department of Computer Science, Virginia Tech
{lid, hcchang , harip, cameron} @ vt.edu

Processor power consumption produces significant heat and can result in higher average operating temperatures. High operating temperatures can lead to reduced reliability and at times thermal emergencies. Previous thermal-aware techniques use Dynamic Voltage and Frequency Scaling (DVFS) or multithreaded or multicore process migration to reduce thermals. However, these methods do not gracefully handle scenarios where processors are fully loaded, i.e. there are no free threads or cores for process scheduling. We propose techniques to reduce processor temperature when processors are fully loaded. We use system-level compiler support and dynamic runtime instrumentation to identify the relative thermal intensity of processes. We implement a thermal-aware process scheduling algorithm that reduces processor thermals while maintaining application throughput. We favor "cool" processes by reducing time slice allocations for "hot" processes. Results indicate that our thermal-aware scheduling can reduce processor thermals by up to 3 degrees Celsius with little to no loss in application throughput.

Energy Efficient Packet Classification Hardware Accelerator

Alan Kennedy and Xiaojun Wang
HDL Lab, School of Electronic Engineering,
Dublin City University, Dublin 9, Ireland
alan.kennedy@eeng.dcu.ie

Bin Liu
Department of Computer Science and Technology
Tsinghua University, Beijing P.R.China
liub@tsinghua.edu.cn

Packet classification is an important function in a router's line-card. Although many excellent solutions have been proposed in the past, implementing high speed packet classification reaching up to OC-192 and even OC-768 with reduced cost and low power consumption remains a challenge. In this paper, the HiCut and HyperCut algorithms are modified making them more energy efficient and better suited for hardware acceleration. The hardware accelerator has been tested on large rulesets containing up to 25,000 rules, classifying up to 77 Million packets per second (Mpps) on a Virtex5SX95T FPGA and 226 Mpps using 65nm ASIC technology. Simulation results show that our hardware accelerator consumes up to 7,773 times less energy compared with the unmodified algorithms running on a StrongARM SA-1100 processor when classifying packets. Simulation results also indicate ASIC implementation of our hardware accelerator can reach OC-768 throughput with less power consumption than TCAM solutions.

Performance Evaluation of Parallel Applications on Next Generation Memory Architecture with Power-Aware Paging Method

Yuto Hosogaya[1,3], Toshio Endo[1,3] and Satoshi Matsuoka[1,2,3]
[1] Tokyo Institute of Technology
[2] National Institute of Informatics
[3] JST, CREST
{hosogaya, endo}@matsulab.is.titech.ac.jp, matsu@is.titech.ac.jp

With increasing demand for low power high performance computing,reducing power of not only CPUs but also memory is becoming important. In typical general-purpose HPC environments, DRAM is installed in an over-provisioned fashion to avoid swapping, although in most cases not all such memory is used, leading to unnecessary and excessive power consumption, even in a standby state. We propose a next generation low power memory system that reduces required DRAM capacity while minimizing application performance degradation. In this system, both DRAM and MRAM, fast non-volatile memory, are used as main memory, while flash memory is used as a swap device. Our profile-based paging algorithm optimizes memory accesses by using faster memory as much as possible, reducing accesses to slower memory. Simulated results of our architecture show that the overall energy consumption of the memory system can be reduced to 25% by in the best case by reducing DRAM capacity, with only 17% performance loss in application benchmarks.

Modeling and Analysis of Power in Multicore Network Processors

S. Huang[1], Y. Luo[2] and W. Feng[1]

[1] Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
{huangs,feng}@cs.vt.edu

[2] Dept. of Electrical and Computer Engineering
University of Massachusetts Lowell
Lowell, MA 01854
yan_luo@uml.edu

With the emergence of multicore network processors in support of high-performance computing and networking applications, power consumption has become a problem of increasing significance. Lower-power multicore processors, such as the Intel IXP network processors, have been employed in network devices to try to address this problem; however, effective tools are still needed to assist in the exploration of the "performance versus power" design space.

In this paper, we present a composite power model that simultaneously leverages three existing power-modeling tools (Cacti, Wattch, and Orion) to model the power consumption of Intel IXP2400. This model is then integrated with an open-source Intel IXP2400 NP simulator called NePSim2 and validated against its datasheet to within 5% of actual power consumption. We utilize the simulator and power model to explore how architectural parameters affect the performance and power consumption of a set of highperformance security applications, thus delivering valuable insights to chip architects for designing energy-efficient multicore network processors.

Power Efficiency in High Performance Computing

Shoaib Kamil	John Shalf	Erich Strohmaier
LBNL/UC Berkeley	LBNL/NERSC	LBNL/CRD
sakamil@lbl.gov	jshalf@lbl.gov	estrohmaier@lbl.gov

After 15 years of exponential improvement in microprocessor clock rates, the physical principles allowing for Dennard scaling, which enabled performance improvements without a commensurate increase in power consumption, have all but ended. Until now, most HPC systems have not focused on power efficiency. However, as the cost of power reaches parity with capital costs, it is increasingly important to compare systems with metrics based on the sustained performance per watt. Therefore we need to establish practical methods to measure power consumption of such systems in-situ in order to support such metrics. Our study provides power measurements for various computational loads on the largest scale HPC systems ever involved in such an assessment. This study demonstrates clearly that, contrary to conventional wisdom, the power consumed while running the High Performance Linpack (HPL) benchmark is very close to the power consumed by any subset of a typical compute-intensive scientific workload. Therefore, HPL, which in most cases cannot serve as a suitable workload for performance measurements, can be used for the purposes of power measurement. Furthermore, we show through measurements on a large scale system that the power consumed by smaller subsets of the system can be projected straightforwardly and accurately to estimate the power consumption of the full system. This allows a less invasive approach for determining the power consumption of large-scale systems.

Towards Energy Efficient Scaling of Scientific Codes

Yang Ding, Konrad Malkowski, Padma Raghavan and Mahmut Kandemir
Department of Computer Science & Engineering
Pennsylvania State University, University Park, PA 16802, USA
{yding, malkowsk, raghavan, kandemir}@cse.psu.edu

Energy consumption is becoming a crucial concern within the high performance computing community as computers expand to the peta-scale and beyond. Although the peak execution rates on tuned dense matrix operations in supercomputers have consistently increased to approach the peta-scale regime, the linear scaling of peak execution rates has been achieved at the expense of cubic growth in power with systems already appearing in the megawatt range. In this paper, we extend the ideas of algorithm scalability and performance iso-efficiency to characterize the system-wide energy consumption. The latter includes dynamic and leakage energy for CPUs, memories and network interconnects. We propose analytical models for evaluating energy scalability and energy efficiency. These models are important for understanding the power consumption trends of data intensive applications executing on a large number of processors. We apply the models to two scientific applications to explore opportunities when using voltage/frequency scaling for energy savings without degrading performance. Our results indicate that such models are critical for energy-aware high-performance computing in the tera- to peta-scale regime.

Workshop 12

High Performance Grid Computing

HPGC 2008

Using Bandwidth Data to Make Computation Offloading Decisions

Rich Wolski, Selim Gurun, Chandra Krintz and Dan Nurmi
Computer Science Dept., Univ. of California, Santa Barbara

We present a framework for making computation offloading decisions in computational grid settings in which schedulers determine when to move parts of a computation to more capable resources to improve performance. Such schedulers must predict when an offloaded computation will outperform one that is local by forecasting the local cost (execution time for computing locally) and remote cost (execution time for computing remotely and transmission time for the input/output of the computation to/from the remote system). Typically, this decision amounts to predicting the bandwidth between the local and remote systems to estimate these costs. Our framework unifies such decision models by formulating the problem as a statistical decision problem that can either be treated "classically" or using a Bayesian approach. Using an implementation of this framework, we evaluate the efficacy of a number of different decision strategies (several of which have been employed by previous systems). Our results indicate that a Bayesian approach employing automatic change-point detection when estimating the prior distribution is the best-performing approach.

NETEMBED: A Network Resource Mapping Service for Distributed Applications

Jorge Londoño and Azer Bestavros
Computer Science Department Boston University
{jmlon, best}@cs.bu.edu

Emerging configurable infrastructures (large-scale overlays, grids, distributed testbeds, and sensor networks among others) comprise diverse sets of computing resources and network conditions. The distributed applications to be deployed on these infrastructures exhibit increasingly complex constraints and requirements on the resources they require. Thus, a common problem facing the efficient deployment of distributed applications on these infrastructures is that of mapping application-level requirements onto the network in such a manner that the requirements of the application are realized. We present two new techniques to tackle this combinatorially-hard problem that thanks to a number of heuristics, are able to find feasible solutions or determine the non-existence of a solution in most cases, where otherwise the problem would be intractable. These techniques are also false negative free, a common problem among other heuristics currently in use.

Large-Scale Experiment of Co-allocation Strategies for Peer-to-Peer SuperComputing in P2P-MPI

Stéphane Genaud
AlGorille Team - LORIA
Campus Scientifique - BP 239,
F-54506 Vandoeuvre-lès-Nancy, France
stephane.genaud@loria.fr

Choopan Rattanapoka
LSIIT-ICPS, UMR 7005 CNRS-ULP
Pôle API, Boulevard Sébastien Brant,
67412 Illkirch, France
choopan@icps.u-strasbg.fr

High Performance computing generally involves some parallel applications to be deployed on the multiples resources used for the computation. The problem of scheduling the application across distributed resources is termed as co-allocation. In a grid context, co-allocation is difficult since the grid middleware must face a dynamic environment. Middleware architecture on a Peer-to-Peer (P2P) basis have been proposed to tackle most limitations of centralized systems. Some of the issues addressed by P2P systems are fault tolerance, ease of maintenance, and scalability in resource discovery. However, the lack of global knowledge makes scheduling difficult in P2P systems.

In this paper, we present the new developments concerning locality awareness as well as co-allocation strategies available in the latest release of P2P-MPI. i) The spread strategy tries to map processes on hosts so as to maximize the total amount of available memory while maintaining locality of processes as a secondary objective. ii) The concentrate strategy tries to maximize locality between processes by using as many cores as hosts offer. The co-allocation scheme has been devised to be simple for the user and meets the main high performance computing requirement which is locality. Extensive experiments have been conducted on Grid5000 with up to 600 processes on 6 sites throughout France. Results show that we achieved the targeted goals in these real conditions.

Service Replication in Grids: Ensuring Consistency in a Dynamic, Failure-Prone Environment

André Luckow
Institute of Computer Science
University of Potsdam
luckow@cs.uni-potsdam.de

Bettina Schnor
Institute of Computer Science
University of Potsdam
schnor@cs.uni-potsdam.de

A major challenge in a service-oriented environment as a Grid is fault tolerance. The more resources and services involved, the more complicated and error-prone becomes the system. Migol [14] is a Grid middleware, which addresses the fault tolerance of Grid applications and services. Migol's core component is its registry service called Application Information Service (AIS). To achieve fault tolerance and high availability the AIS is replicated on different sites. Since a registry is a stateful Web service, the replication of the AIS is no trivial task.

In this paper, we present our concept for active replication of Grid services. Migol's Replication Service uses a token-based algorithm and certificate-based security to provide secure group communication. Further, we show in different experiments that active replication in a real Grid environment is feasible.

Experimental Validation of Grid Algorithms: a Comparison of Methodologies

Emmanuel Jeannot
INRIA Nancy Grand-Est
LORIA, Nancy University, CNRS
Emmanuel.Jeannot@loria.fr

The increasing complexity of available infrastructures with specific features (caches, hyperthreading, dual core, etc.) or with complex architectures (hierarchical, parallel, distributed, etc.) makes models either extremely difficult to build or intractable. Hence, it raises the question: how to validate algorithms if a realistic analytic analysis is not possible any longer? As for some other sciences (physics, chemistry, biology, etc.), the answer partly falls in experimental validation. Nevertheless, experiment in computer science is a difficult subject that opens many questions: what an experiment is able to validate? What is a “good experiments”? How to build an experimental environment that allows for “good experiments”? etc. In this paper we will provide some hints on this subject and show how some tools can help in performing “good experiments”. More precisely we will focus on three main experimental methodologies, namely real-scale experiments (with an emphasis on PlanetLab and Grid’5000), Emulation (with an emphasis on Wrekavoc: http://wrekavoc.gforge.inria.fr) and simulation (with an emphasis on SimGRID and Grid-Sim). We will provide a comparison of these tools and methodologies from a quantitative but also qualitative point of view.

A Modeling Approach for Estimating Execution Time of Long-Running Scientific Applications

Seyed Masoud Sadjadi[1], Shu Shimizu[2], Javier Figueroa[1,3], Raju Rangaswami[1], Javier Delgado[1], Hector Duran[4], Xabriel J. Collazo-Mojica[5]
[1]Florida International University (FIU), Miami, Florida, USA; [2]IBM Tokyo Research Laboratory, Tokyo, Japan; [3]University of Miami, Coral Gables, Florida, USA; [4]University of Guadalajara, CUCEA, Mexico; [5]University of Puerto Rico, Mayaguez Campus, Puerto Rico;

In a Grid computing environment, resources are shared among a large number of applications. Brokers and schedulers find matching resources and schedule the execution of the applications by monitoring dynamic resource availability and employing policies such as first-come-first-served and back-filling. To support applications with timeliness requirements in such an environment, brokering and scheduling algorithms must address an additional problem - they must be able to estimate the execution time of the application on the currently available resources. In this paper, we present a modeling approach to estimating the execution time of long-running scientific applications. The modeling approach we propose is generic; models can be constructed by merely observing the application execution “externally” without using intrusive techniques such as code inspection or instrumentation. The model is cross-platform; it enables prediction without the need for the application to be profiled first on the target hardware. To show the feasibility and effectiveness of this approach, we developed a resource usage model that estimates the execution time of a weather forecasting application in a multi-cluster Grid computing environment. We validated the model through extensive benchmarking and profiling experiments and observed prediction errors that were within 10% of the measured values. Based on our initial experience, we believe that our approach can be used to model the execution time of other time-sensitive scientific applications; thereby, enabling the development of more intelligent brokering and scheduling algorithms.

Supporting a Visualization Application on a Self-Adapting Grid Middleware

Liang Chen, HanWei Shen and Gagan Agrawal
Department of Computer Science and Engineering
Ohio State University, Columbus OH 43210
{chenlia,hwshen,agrawal}@cse.ohiostate.edu

This paper describes how we have used a self-adapting middleware to implement a distributed and adaptive volume rendering application. The middleware we have used is GATES (Grid-based Adap-Tive Execution on Streams), which allows processing of streaming data in a distributed environment. A challenge in supporting such an application on streaming data is to balance the visualization quality and the speed of processing, which can be automatically done by the GATES middleware. We describe how we divide the application into a number of processing stages, and what adaptation parameters we use. Our experimental studies have focused on evaluating the self-adaptation enabled by the middleware, and measuring the overhead associated with the use of middleware.

Assuring Information Quality in E-Science

Ying Su[1], Zhanming Jin[2] and Peng Jie[1]
[1] Institute of Scientific and Technical Information of China, Beijing, CHN
[2] School of Economics and Management, Tsinghua University, Beijing, CHN
E-mail: {suy4, jinzhm}@sem.tsinghua.edu.cn, pengj@istic.ac.cn

In this paper we outline a framework for assuring information quality (IQ) in an e-Science context, using semiotics theory, semantic explanation of resources, and data couplings. Scientists can define the quality characteristics that are of importance in their particular domain by applying an IQ semiotics, which classifies and organizes these domain-specific quality characteristics within this quality assurance framework. Resource Description Framework (RDF) is used to explain data resources, with reference to IQ indicators defined in the semiotics. Data couplings - again defined in RDF - are used to represent mappings between data elements and the IQ semiotics. As a practical illustration of our approach, we present a case study from the domain of natural science.

Oncogrid: A Proposal of Grid Infrastructure for the Establishment of a National Health Information System on Childhood Cancer

Higor A. V. Alves, Moacir A. Campos. Jr., Francisco J. A. Fernandes,
Marcia N. S. Kondo, André N. de Mello, Adilson Y. Hira, Marcelo K. Zuffo
The Laboratory of Integrated Systems - Polytechnic School of the University of São Paulo
Av. Prof. Luciano Gualberto, Trav. 3, 158
São Paulo, São Paulo, Brazil
{higor, macampos, fjose, mkondo, andrenm, ayhira, mkzuffo}@lsi.usp.br
Paola R. G. Accioly, Luiz Guimarães, Magdala A. Novaes
Telehealth Group of the Federal University of Pernambuco
Av. Prof. Moraes Rego, s/n, Cidade Universitária
Recife, Pernambuco, Brazil
{paola.accioly, luiz.guimaraes, magdala.novaes}@nutes.ufpe.br

In Brazil, a country of continental proportions, law requires that health institutions directly involved in the cancer treatment in the country have to register and consolidate cancer patients' data. This information is important for the management and evaluation of the cancer treatment, and for the definition of public policies. In this work we propose the use of a high-performance computational infrastructure, based on grid computing, to ease and accelerate the consolidation of the remotely distributed patients' records in these institutions. In particular, we present a case study of implementation of a tool for patients' survival estimation based on the Kaplan-Meier Method. We also discuss the issues related to architecture, implementation and integration, as well as the benefits and prospects for the use of this infrastructure on the establishment of a national health information system.

Grid-Based Research, Development, and Deployment in New York State

Russ Miller[1,2], Jonathan J. Bednasz[3], Kenneth Chiu[4], Steven M. Gallo[3], Madhu Govindaraju[4],
Michael Lewis[4], Catherine L. Ruby[1], and Charles M. Weeks[2]
[1]Department of Computer Science & Engineering, SUNY-Buffalo, Buffalo, NY 14260
[2]Hauptman-Woodward Medical Research Institute, 700 Ellicott Street, Buffalo, NY 14203
[3]Center for Computational Research, SUNY-Buffalo, Buffalo, NY 14203
[4]Department of Computer Science, Binghamton University, Binghamton, NY 13902
miller@buffalo.edu

In this paper, we present cyberinfrastructure and grid computing efforts in New York State. In particular, we focus on fundamental efforts in Binghamton and Buffalo, including the design, development, and deployment of the New York State Grid, as well as a grass-roots New York State Initiative.

Monitoring for Multi-Middleware Grid

Gevorg S. POGHOSYAN and Marcel KUNZE
Forschungszentrum Karlsruhe
Steinbuch Centre for Computing
Karlsruhe Institute of Technologies
{Gevorg.Poghosyan, Marcel.Kunze}@kit.edu

Within the framework of the German Grid Computing Initiative (D-Grid), we study the monitoring systems and software suites that are used to collect the information from computational grids working with single or multiple middleware systems. Based on these investigations we build the prototypes of monitoring systems and implement it in the D-Grid infrastructure. A concept of Site Check Center (SCC) suggested to providing a unified interface for access to data from different test-benchmark systems working with more than one middleware software. A Vertical hierarchal architecture for exchange of information and building the network of monitoring systems is suggested and employed. A concept for separation between consumer and resource/service provider related monitoring information is proposed. Furthermore, we study the integration of monitoring components into general computational multi-middleware grid infrastructure developed according to specific community needs.

Workshop 13

Workshop on System Management Techniques, Processes, and Services

SMTPS 2008

SonD : A Fast Service Deployment System Based on IP SAN

Yang Yin[1,2], Zhenjun Liu[1], Haiying Tang[1,2], Shuo Feng[1],
Yajun Jia[3], Xiaoming Han[1] and Lu Xu[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences,Beijing, 100080, China
[2]Graduate School of the Chinese Academy of Sciences, Beijing, 100080, China
[3]Geophysical Research Institute, Exploration and Development Institute of Xinjiang
Oilfield Company, PretroChina, Wulumuqi, 830013, China
{yinyang, liuzhenjun, tanghaiying, fengshuo, hanxiaoming, xulu}@nrchpc.ac.cn
jyjun@petrochina.com.cn

Deploying operating systems and services in an enterprise environment can be a time-consuming and cumbersome task. With the dramatic increase of the number of computers managed by the system administrator, the work of management is more complicated than before. Moreover, services provided by servers in data center change frequently, which requires a flexible service deployment system to achieve on demand deployment. To maximize return on investment (ROI) and stay competitive in today's fast-paced global economy, administrators must be able to deploy systems or services quickly and effectively with minimal downtime.

This paper presents SonD, a system for providing dynamic, flexible, and quick service and operating system deployment in a LAN environment. The techniques used by SonD include an iterative snapshot mechanism based on an IP SAN system, which provides writable snapshots to achieve fast creation of the service, a dynamic service mapping method, and a network boot technology achieved by both software and hardware solutions. This design results in a system which can rapidly, flexibly, dynamically and reliably distribute OSes and services to many clients simultaneously. For example, SonD can deploy 50 windows XP systems in a few minutes on commodity PCs. We describe SonD's design and implementation, and review the key technologies for our system. We evaluate its performance and apply it to oil exploration. The results show that SonD fits well with the service deployment usage in cluster computing environment.

An MPI Tool for Automatically Discovering the Switch Level Topologies of Ethernet Clusters

Joshua Lawrence and Xin Yuan
Department of Computer Science, Florida State University, Tallahassee, FL 32306
{lawrence, xyuan}@cs.fsu.edu

We present an MPI topology discovery tool for homogeneous Ethernet switched clusters. Unlike existing Ethernet topology discovery methods that rely on Simple Network Management Protocol (SNMP) queries to obtain the topology information, our tool infers the topology from end-to-end measurements. The tool works on clusters connected by managed and/or unmanaged Ethernet switches, and does not require any special privilege. We discuss the theoretical foundation of the tool, present the algorithms used, and report our evaluation of the tool.

Scalable Server Provisioning with HOP-SCOTCH

David Daly, Marcio A. Silva, José E. Moreira
IBM T.J. Watson Research Center
Yorktown Heights, NY, 10958, U.S.A.
{dmdaly, masilva, jmoreira}@us.ibm.com

The problem of provisioning servers in a cluster infrastructure includes the issues of coordinating access and sharing of physical resources, loading servers with the appropriate software images, supporting storage access to users and applications, and providing basic monitoring and control services for those servers. We have developed a system called HOP-SCOTCH that automates the provisioning process for large clusters. Our solution relies on directory services to implement access control. It uses network boot and managed root disks to control the image of each server. We leverage IBM's Global Storage Architecture to provide storage to users and applications. Finally, interfaces are provided to access the services both programmatically and interactively. We demonstrate the scalable behavior of HOP-SCOTCH by experimenting with a cluster of 40 blade servers. We can provision all 40 servers with a brand new image in under 15 minutes.

Agent Multiplication: An Economical Large-scale Testing Environment for System Management Solutions

Kyung Dong Ryu[1], David Daly[1], Mary Seminara[1], Sukhyun Song[2] and Paul G. Crumley[1]

[1]IBM T.J. Watson Research Center
Yorktown Heights, NY
{kryu, dmdaly, seminara, pgc}@us.ibm.com

[2]Department of Computer Science
University of Maryland, College Park, MD
shsong@cs.umd.edu

System management solutions are designed to scale to thousands or more machines and networked devices. However, it is challenging to test and verify the proper operation and scalability of management software given the limited resources of a testing lab. We have developed a method called agent multiplication, in which one physical testing machine is used to represent hundreds of client machines. This provides the necessary client load to test the performance and scalability of the management software and server within limited resources. In addition, our approach guarantees that the test environment remains consistent between test runs, ensuring that test results can be meaningfully compared. We used agent multiplication to test and verify the operation of a server managing 4,000 systems. We exercised the server functions with only 8 test machines. Applying this test environment to an early version of a real enterprise system management solution we were able to uncover critical bugs, resolve race conditions, and examine and adjust thread prioritization levels for improved performance.

Ovis-2: A Robust Distributed Architecture for Scalable RAS

J. M. Brandt, B. J. Debusschere, A. C. Gentile, J. R. Mayo,
P. P. Pébay, D. Thompson and M. H. Wong
Sandia National Laboratories, Livermore CA 94550 U.S.A.
{brandt,bjdebus,gentile,jmayo,pppebay,dcthomp,mhwong}@sandia.gov

Resource utilization in High Performance Compute clusters can be improved by increased awareness of system state information. Sophisticated run-time characterization of system state in increasingly large clusters requires a scalable fault-tolerant RAS framework. In this paper we describe the architecture of Ovis-2 and how it meets these requirements. We describe some of the sophisticated statistical analysis, 3-D visualization, and use cases for these. Using this framework and associated tools allows the engineer to explore the behaviors and complex interactions of low level system elements while simultaneously giving the system administrator their desired level of detail with respect to ongoing system and component health.

Early Experiences in Application Level I/O Tracing on Blue Gene Systems

Seetharami Seelam[1], I-Hsin Chung[1], Ding-Yong Hong[2], Hui-Fang Wen[1] and Hao Yu[1]
[1]IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598 USA
{sseelam,ihchung,hfwen,yuh}@us.ibm.com
[2]Department of Computer Science
National Tsing Hua University
HsinChu 300, Taiwan
dyhong@sslab.cs.nthu.edu.tw

On todays massively parallel processing (MPP) supercomputers, it is increasingly important to understand I/O performance of an application both to guide scalable application development and to tune its performance. These two critical steps are often enabled by performance analysis tools to obtain performance data on thousands of processors in an MPP system. To this end, we present the design, implementation, and early experiences of an application level I/O tracing library and the corresponding tool for analyzing and optimizing I/O performance on Blue Gene (BG) MPP systems. This effort was a part of IBM HPC Toolkit for BG systems. To our knowledge, this is the first comprehensive application-level I/O monitoring, playback, and optimizing tool available on BG systems. The preliminary experiments on popular NPB BTIO benchmark show that the tool is much useful on facilitating detailed I/O performance analysis.

Efficient Algorithms for Periodic Real-Time Tasks to Optimal Discrete Voltage Schedules

Da-Ren Chen
Dept. of Information Management,
Hwa Hsia Institute of Technology,
Taiwan, R.O.C
danny@cc.hwh.edu.tw

Shu-Ming Hsieh
Dep. of Electronic Engineering
Hwa Hsia Institute of Technology,
Taiwan, R.O.C
stanley@cc.hwh.edu.tw

Ming-Fong Lai
Science & Technology Research
and Information Center National
Applied Research Laboraries,
Taiwan, R.O.C
mflai@mail.stpi.org.tw

In this paper, we pay attention to the periodic task scheduling on a variable voltage processor with d discrete voltage/speed levels. We propose an intra-task DVS algorithm which constructs a minimum energy schedule for k tasks in $O(d + k \log k)$ time. We also give an inter-task DVS algorithm with $O(d + n \log n)$ time, where n denotes the number of jobs. Previous approaches solve this problem by first generating a canonical schedule and then adjusting the speed in $O(dn \log n)$ or $O(n^3)$ time. However, they do not consider that the length of a canonical schedule depends on the LCM of those of task periods and is of exponential length in general. In our approach, the tasks with arbitrary periods are first transformed into harmonic periods and then profile their key features. Afterward, the (optimal) discrete solution can be computed directly from these features.

Adaptive B-Greedy (ABG): A Simple yet Efficient Scheduling Algorithm

Hongyang Sun and Wen-Jing Hsu
School of Computer Engineering,
Nanyang Technological University, Singapore
{sunh0007, hsu}@ntu.edu.sg

In order to improve processor utilizations on parallel systems, adaptive scheduling with parallelism feedback was recently proposed. A-Greedy, an existing adaptive scheduler, offers provably-good job execution time and processor utilization. Unfortunately, it suffers from unstable feedback and hence unnecessary processor reallocations even when the job has constant parallelism. This problem may cause difficulties in the management of system resources. We propose a new adaptive scheduler called ABG (for Adaptive B-Greedy), which ensures both performance and stability. In a direct comparison with A-Greedy using simulated data parallel jobs, ABG shows an average 50% reduction in wasted processor cycles and an average 20% improvement in running time. For a set of jobs, ABG also outperforms A-Greedy by 10% to 15% on average in terms of both makespan and mean response time, provided the system is not heavily loaded. Our detailed analysis shows that ABG indeed offers improved transient and steady-state behaviors in terms of control-theoretic metrics. Using trim analysis, we show that ABG provides nearly linear speedup for individual jobs and good processor utilizations. Using competitive analysis, we also show that ABG offers good makespan and mean response time bounds.

A Hash-TLB Approach for MMU Virtualization in Xen/IA64

Xiantao Zhang [1,2], Anthony X. F. Xu [2], Qi Li [2,3], David K. Y. Yau [4], Sihan Qing [5], Huanguo Zhang [1]
[1]School of Computer, Wuhan University, Wuhan Hubei, 430079, China
[2]Intel OTC, SSG, Intel China Software Center, Shanghai, 200241, China
[3]Department of Computer Science, Tsinghua University, Beijing, 100084, China
[4]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
[5]Institute of Software, Chinese Academy of Sciences, Beijing, 100080, China
xiantao.zhang,anthony.xu@intel.com liqi@csnet1.cs.tsinghua.edu.cn yau@cs.purdue.edu

With advances in hardware-assisted full virtualization technologies, system virtualization based on the virtual machine monitor (VMM) has received much recent attention. Using the Xen/IA64 hardware virtual machine implemented on Intel® Virtualization Technology for Itanium® (VT-i), we investigate the design of a virtual software hash translation lookaside buffer (TLB) based on the virtual hash page table (VHPT). Experimental results show that the proposed design can significantly improve the performance of the hardware virtual machine of Xen/IA64. Our contributions are the following. First, we design and implement in the VMM a virtual hash TLB algorithm to optimize the system performance of VT-i guest virtual machines. Second, we quantify experimentally the performance benefits of the hash TLB for VT-i guest virtual machines and analyze the performance impact of the software VHPT walker with the hash TLB algorithm. Lastly, we present experiments to verify, in an SMP virtual machine system environment, the superior scalability of the hash TLB approach.

A Theoretical Analysis of Wireless Localization Using RF-based Fingerprint Matching

Jie Yang, Yingying Chen
{jyang, yingying.chen}@stevens.edu
Deparment of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030

Accurately obtaining the position of mobile devices is critical to high-level applications. In indoor environments, localization approaches employing RF-based fingerprint matching is an active research area because it can reuse the existing communication infrastructure, as well as reduce the signal uncertainty to achieve better location accuracy. In this paper, we provide a theoretical analysis of the localization performance when using fingerprint matching schemes. Specifically, we derived an analytic expression for the Cumulative Distribution Function (CDF) of the location error and investigated the mathematical relationship between the location error and the sampling points. Particularly, we studied the effects of the number of sampling points and the distance between adjacent sampling points. We further conducted experiments using an 802.11 network in a real office building environment. Our performance evaluation provides strong evidence that theoretical analysis is effective as the experimental results match our analytic study very well.

Mining Joules and Bits: Towards a Long-Life Pervasive System

Shweta Medhekar, Richard Howard, Wade Trappe and Yanyong Zhang
WINLAB, Rutgers University, North Brunswick, NJ 08902
{shwetam,reh,trappe,yyzhang}@winlab.rutgers.edu
Peter Wolniansky
Radio Technology Systems, 82 Mt Tabor Way, Ocean Grove, NJ 07756
wolniansky@optonline.net

In this paper, we investigate one of the major challenges in pervasive systems: energy efficiency, by exploring the design of an RFID system intended to support the simultaneous and real time monitoring of thousands of entities. These entities, which may be individuals or inventory items, each carry a low-power transmit-only tag and are monitored by a collection of networked base-stations reporting to a central database. We have built a customized transmit-only tag with a small form-factor, and have implemented a real-time monitoring application intended to verify the presence of each tag in order to detect potential disappearance of a tag (perhaps due to item theft). Throughout the construction of our pervasive system, we have carefully engineered it for extended tag lifetime and reliable monitoring capabilities in the presence of packet collisions, while keeping the tags small and inexpensive. The major challenge in this architecture (called Roll-Call) is to supply the energy needed for long range continuous tracking for a year or more while keeping the tags (called PIPs) small and inexpensive. We have used this as a model problem for optimizing cost, size and lifetime across the entire pervasive, persistent system from firmware to protocol.

Workshop 14

Workshop on Parallel and Distributed Scientific and Engineering Computing

PDSEC 2008

Efficient Parallelization Method for Large Scale Beam Dynamics Simulations in Linear Accelerators

J. Xu, B. Mustapha, V.N. Aseev, P.N. Ostroumov and J. Nolen
Physics Division, Argonne National Laboratory,
9700 S. Cass Ave., Argonne, IL 60439, USA

Large scale beam dynamics simulations are important to support the design and operations of an accelerator. The beam dynamics code TRACK, originally developed at Argonne National Laboratory (ANL) has become a general beam dynamics code for hadron linacs and has found wide applications worldwide. In order to combine the unique features of TRACK with large scale and fast parallel computing we have recently developed parallel version of the code: PTRACK. It has been successfully benchmarked on different platforms: BG/L and Jazz at ANL, Iceberg at ARSC and Seaborg at NERSC. Large scale end-to-end simulations of the FNAL proton driver has been performed, where 100 millions particles were tracked. The actual parallel version has the potential of simulating 10^9 particles on 10 racks with 20,480 processors of BG/L at ANL, which will be available in 2008. This paper focus on efficient parallelization method been used in PTRACK.

Parallel Implementations of Three Scientific Applications Using LB_Migrate

Rohit Chaube[1,3], Ioana Banicescu[2,3] and Ricolindo L. Cari?no[4]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science and Engineering
[3]Center for Computational Sciences – HPC[2]
[4]Center for Advanced Vehicular Systems
Mississippi State University
{rohit@hpc,ioana@cse,rlc@cavs}.msstate.edu

In this paper we focus on the implementation of large scientific applications with LB Migrate, a dynamic load balancing library. The library employs dynamic loop scheduling techniques to address performance degradation factors due to load imbalance, provides a flexible interface with the native data structure of the application, and performs data migration. The library is reusable and it is not application specific. For initial testing, the library was employed in three applications: the profiling of an automatic quadrature routine, the simulation of a hybrid model for image denoising, and N-body simulations. We discuss the original applications without the library, the changes made to the applications to be able to interface with the library, and we present experimental results. Performance results indicate that the library adds minimal overhead, up to 6%, and it varies from application to application. However the benefits gained from the use of the library are substantial.

Fast and Efficient Formation Flocking for a Group of Autonomous Mobile Robots

Naixue Xiong[1,2], Yingshu Li[1]
[1]Depa. of Computer Science
Georgia State University, Atlanta, USA
fnxiong, ylig@cs.gsu.edu

Laurence T. Yang
Depa. of Computer Science
St. Francis Xavier University, Canada
lyang@stfx.ca

Jong Hyuk Park
Depa. of Computer Science and Engineering
Kyungnam University
449 Wolyoung-dong, Masan
Kyungnam, 631-701, Korea
parkjonghyuk1@hotmail.com

Yan Yang[2], Sun Tao[3]
[2]School of Information Science
Japan Advanced Insti. of Scie. and Tech.
[3]Wuhan Second Ship Design
Research Institute,Wuhan, 430064, China
y.yang@jaist.ac.jp

The control and coordination of mobile robots in groups that can freely cooperate and move on a plane is a widely studied topic in distributed robotics. In this paper, we focus on the flocking problem: there are two kinds of robots: the leader robot and the follower robots. The follower robots are required to follow the leader robot wherever it goes (following), while keeping a formation they are given in input (flocking). A novel scheme is proposed based on the relative motion theory. Extensive theoretical analysis and simulation results demonstrate that this scheme provides the follower robots an efficient method to follow the leader as soon as possible with the shortest path. Furthermore, this scheme is scalable, and the processing load for every robot is not increased with the addition of more robots. Keywords. Distributed system, Mobile robots, Flocking algorithm, Motion theory, Control and coordination.

Evaluation of Dynamic Scheduling Methods in Simulations of Storm-time Ion Acceleration

Ioannis Riakiotakis, Georgios Goumas and Nectarios Koziris
National Technical University of Athens
School of Electrical and Computer Engineering
Computing Systems Laboratory
{iriak, goumas, nkoziris}@cslab.ece.ntua.gr
Fiori-Anastasia Metallinou and Ioannis A. Daglis
National Observatory of Athens
Institute for Space Applications and Remote Sensing
{ametal, daglis}@space.noa.gr

In this paper we investigate the applicability of classic dynamic loop scheduling methods on a numerical simulation code that calculates the trajectories of charged particles in the earth's magnetosphere. The numerical application under consideration investigates the influence of substorm-induced electric fields that cause magnetospheric disturbances, responsible for severe hazards in human activities and technology infrastructures in the near-earth space environment. The computational time to simulate the motion of each particle is dependent on the inital conditions applied and may greatly vary between different particles. This fact leads to great load imbalances in parallel execution scenarios and thus to degraded overall performance. For this reason we apply dynamic scheduling techniques to load-balance the tasks in homogeneous, heterogeneous and loaded distributed-memory parallel platforms and select the most appropriate among the available strategies.

Large Scale Execution of a Bioinformatic Application on a Volunteer Grid

Viktors Bertis[6], Raphaël Bolze [1,3,4], Frédéric Desprez [2,3,4] and Kevin Reed [5]
[1] CNRS [2] INRIA [3] Université de Lyon
[4] LIP laboratory, UMR 5668, ENS Lyon C CNRS C INRIA C UCBL, Lyon, France
[5] IBM Centers for Solution Innovation, Chicago, United States
[6] IBM Systems & Technology Group, Austin, United States
{raphael.bolze,frederic.desprez}@ens-lyon.fr
{bertis,knreed}@us.ibm.com

Large volunteer desktop platforms are now available for several applications. This paper presents the work that we did to prepare the first phase of the Help Cure Muscular Dystrophy project to run on World Community Grid. The project was launched on December 19, 2006, and took 26 weeks to complete. We present performance evaluation of the overall execution and compare a volunteer grid with a dedicated one.

Evaluation and Tuning of the Level 3 CUBLAS for Graphics Processors

Sergio Barrachina, Maribel Castillo, Francisco D. Igual and Rafael Mayo
Enrique S. Quintana-Ortí
Depto. de Ingeniería y Ciencia de Computadores
Universidad Jaume I
12.071CCastellón, Spain
{barrachi,castillo,figual,mayo,quintana}@icc.uji.es

The increase in performance of the last generations of graphics processors (GPUs) has made this class of platform a coprocessing tool with remarkable success in certain types of operations. In this paper we evaluate the performance of the Level 3 operations in CUBLAS, the implementation of BLAS for NVIDIA R GPUs with unified architecture. From this study, we gain insights on the quality of the kernels in the library and we propose several alternative implementations that are competitive with those in CUBLAS. Experimental results on a GeForce 8800 Ultra compare the performance of CUBLAS and the new variants.

A Lightweight Scalable I/O Utility for Optimizing High-End Computing Applications

Shujia Zhou, Bruce H. Van Aartsen and Thomas L. Clune
NASA Goddard Space Flight Center, Greenbelt, MD USA 20771
Shujia.Zhou@nasa.gov

Filesystem I/O continues to be a major performance bottleneck for many High-End Computing (HEC) applications and in particular for Earth science models, which often generate a relatively large volume of data for a given amount of computational work. The severity of this I/O bottleneck rapidly increases with the number of processors utilized. Consequently, considerable computing resources are wasted, and the sustained performance of HEC applications such as climate and weather models is highly constrained. To alleviate much of this bottleneck, we have developed a lightweight software utility designed to improve performance of typical scientific applications by circumventing bandwidth limitations of typical HEC filesystems. The approach is to exploit the faster interprocessor bandwidth to move output data from compute nodes to designated I/O nodes as quickly as possible, thereby minimizing the I/O wait time. This utility has successfully demonstrated a significant performance improvement within a major NASA weather application.

Outlier Detection in Performance Data of Parallel Applications

Katharina Benkert [1,2], Edgar Gabriel [1] and Michael M. Resch [2]
[1] Parallel Software Technologies Laboratory, Department of Computer Science,
University of Houston, Houston, TX 77204-3010, USA.
Email: gabriel@cs.uh.edu
[2] High Performance Computing Center Stuttgart (HLRS),
Nobelstr. 19, 70569 Stuttgart, Germany.
Email: {benkert, resch}@hlrs.de

When an adaptive software component is employed to select the best-performing implementation for a communication operation at runtime, the correctness of the decision taken strongly depends on detecting and removing outliers in the data used for the comparison. This automatic decision is greatly complicated by the fact that the types and quantities of outliers depend on the network interconnect and the nodes assigned to the job by the batch scheduler. This paper evaluates four different statistical methods used for handling outliers, namely a standard interquartile range method, a heuristic derived from the trimmed mean value, cluster analysis and a method using robust statistics. Using performance data from the Abstract Data and Communication Library (ADCL) we evaluate the correctness of the decisions made with each statistical approach over three fundamentally different network interconnects, namely a highly reliable InfiniBand network, a Gigabit Ethernet network having a larger variance in the performance, and a hierarchical Gigabit Ethernet network.

MUPHY: A Parallel High Performance MUlti PHYsics/Scale Code

Massimo Bernaschi and Sauro Succi
Istituto Applicazioni Calcolo, CNR,
Viale del Policlinico, 137
00161 Rome, Italy
Maria Fyta, Efthimios Kaxiras, Simone Melchionna and Jayanta K. Sircar
Department of Physics and School of Engineering and Applied Sciences,
Harvard University,
Cambridge MA 02138, USA

We present a parallel high-performance program that combines microscopic Molecular Dynamics (MD) with a mesoscopic Lattice Boltzmann (LB) method to simulate the translocation of biopolymers through nanometer size pores explicitly taking into account the interactions of the molecule with the surrounding fluid. The parallel implementation exhibits excellent scalability on the BlueGene platform and includes techniques which may improve the flexibility and efficiency of other complex multi-physics applications.

Optimizing Network Performance of Computing Pipelines in Distributed Environments

Qishi Wu, Yi Gu
Dept of Computer Science
University of Memphis
Memphis, TN 38152
{qishiwu,yigu}@memphis.edu

Mengxia Zhu
Dept of Computer Science
Southern Illinois University
Carbondale, IL 62901
mzhu@cs.siu.edu

Nageswara S.V. Rao
Computer Science & Math Div
Oak Ridge National Laboratory
Oak Ridge, TN 37831
raons@ornl.gov

Supporting high performance computing pipelines over wide-area networks is critical to enabling large-scale distributed scientific applications that require fast responses for interactive operations or smooth flows for data streaming. We construct analytical cost models for computing modules, network nodes, and communication links to estimate the computing times on nodes and the data transport times over connections. Based on these time estimates, we present the Efficient Linear Pipeline Configuration method based on dynamic programming that partitions the pipeline modules into groups and strategically maps them onto a set of selected computing nodes in a network to achieve minimum end-to-end delay or maximum frame rate. We implemented this method and evaluated its effectiveness with experiments on a large set of simulated application pipelines and computing networks. The experimental results show that the proposed method outperforms the Streamline and Greedy algorithms. These results, together with polynomial computational complexity, make our method a potential scalable solution for large practical deployments.

A Parallel Sewing Method for Solving Tridiagonal Toeplitz Strictly Diagonally Dominant Systems

M. Majedi, R. E. Shaw, L. E. Garey
University of New Brunswick
Dept. of Computer Science and Applied Statistics
Saint John, NB, Canada E2L 4L5
maryam.majedi@unb.ca, reshaw@unbsj.ca, legarey@unbsj.ca

The large scale of linear systems of equations results in costly solving time. These systems usually have specific properties that can be used for designing fast algorithms. In addition, using parallel programming on distributed memory clusters enables us to get the results even faster. This work introduces a new fast parallel algorithm for solving systems with a strictly diagonally dominant three-band Toeplitz coefficient matrix. We call this new method the sewing method because the boundaries sew the adjacent subsystems together.

A Linear Solver for Benchmarking Partitioners

Kambiz Ghazinour, Ruth E. Shaw, Eric Aubanel[1] and Larry E. Garey
University of New Brunswick
Dept. of Computer Science & Applied Statistics
Saint John, NB, Canada E2L 4L5
[1]University of New Brunswick
Dept. of Computer Science
Fredericton, NB, Canada E3B 5A3
Kambiz.gh@unb.ca, reshaw@unbsj.ca, aubanel@unb.ca, legarey@unbsj.ca

A number of graph partitioners are currently available for solving linear systems on parallel computers. Partitioning algorithms divide the graph that arises from the linear system into a specified number of partitions such that the workload per processor is balanced and the communication between the processors is minimized. The measure of partition quality is often taken to be the number of edges cut by the partition. Ultimately the quality of a partition will be reflected in the execution time of the parallel application. In this paper, we introduce a linear solver benchmark that enables comparison of partition quality. This work also serves to motivate further work on developing benchmarks for graph partitioners.

SWAMP: Smith-Waterman using Associative Massive Parallelism

Shannon Steinfadt and Dr. Johnnie W. Baker
Department of Computer Science, Kent State University, Kent, Ohio 44242 USA
ssteinfa@cs.kent.edu, jbaker@cs.kent.edu

One of the most commonly used tools by computational biologists is some form of sequence alignment. Heuristic alignment algorithms developed for speed and their multiple results such as BLAST [1] and FASTA [2] are not a total replacement for the more rigorous but slower algorithms like Smith-Waterman [3]. The different techniques complement one another. A heuristic can filter dissimilar sequences from a large database such as GenBank [4] and the Smith-Waterman algorithm performs more detailed, in-depth alignment in a way not adequately handled by heuristic methods. An associative parallel Smith-Waterman algorithm has been improved and further parallelized. Analysis between different algorithms, different types of file input, and different input sizes have been performed and are reported here. The newly developed associative algorithm reduces the running time for rigorous pairwise local sequence alignment.

A Parallel Implementation of Fault Simulation on a Cluster ofWorkstations

Kyunghwan Han and Soo-Young Lee
Department of Electrical and Computer Engineering
Auburn University, Auburn, AL 36849
leesooy@eng.auburn.edu

A cluster of workstations may be employed for reducing fault simulation time greatly. Fault simulation can be parallelized by partitioning fault list, the test vector or both. In this study, parallel fault simulation algorithm called PAUSIM has been developed by parallelizing AUSUM which consists of logic simulation and two steps of fault simulation for sequential logic circuits. Compared to other algorithms, PAUSIM-CY avoids redundant work by a judicious task decomposition. Also, it adopts a cyclic fault partitioning method based on the LOG partitioning and local redistribution, resulting in a well-balanced load distribution. The results from the parallel implementation using MPI show a significant speed-up by PAUSIM-CY over other existing parallel algorithms.

Implementation of NAMD Molecular Dynamics Non-bonded Force-field on the Cell Broadband Engine Processor

Guochun Shi, Volodymyr Kindratenko
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
{gshi,kindr}@ncsa.uiuc.edu

We present results of porting an important kernel of a production molecular dynamics simulation program, NAMD, to the Cell/B.E. processor. The non-bonded force-field kernel, as implemented in the NAMD SPEC 2006 CPU benchmark, has been implemented. Both single-precision and double-precision floating-point kernel variations are considered, and performance results obtained on the Cell/B.E., as well as several other platforms, are reported. Our results obtained on a 3.2 GHz Cell/B.E. blade show linear speedups when using multiple synergistic processing elements.

A Distributed Simulation Framework for Conformal Radiotherapy

G. Yaikhom, J. P. Giddy, D. W. Walker and P. Downes
School of Computer Science, Cardiff University, Cardiff, UK
{g.yaikhom, j.p.giddy, d.w.walker, p.downes}@cs.cardiff.ac.uk
E. Spezi and D. G. Lewis
Department of Medical Physics, Velindre Cancer Centre, Cardiff, UK
{emiliano.spezi, geraint.lewis}@velindre-tr.wales.nhs.uk

This paper describes the RTGrid distributed simulation framework for conformal radiotherapy. We introduce novel approaches through which several distributed computing technologies are made accessible to Monte Carlo simulations for radiotherapy dose calculations. Currently, radiotherapy treatment planning is typically performed on PCs and workstations which lack the computational power to run Monte Carlo simulations quickly enough to be useful to clinicians. Therefore, although Monte Carlo simulation techniques offer highly accurate doses, they are seldom chosen for clinical deployment. The RTGrid project is investigating both the capability and capacity modes of exploiting grid computing for radiotherapy treatment planning using Monte Carlo simulations.

Trace-based Automatic Padding for Locality Improvement with Correlative Data Visualization Interface

MARCO HÖ BBEL
Fachgruppe Informatik
Universität Bayreuth
Hoebbel@Uni-Bayreuth.de

THOMAS RAUBER
Fachgruppe Informatik
Universität Bayreuth
Rauber@Uni-Bayreuth.de

CARSTEN SCHOLTES
Fachgruppe Informatik
Universität Bayreuth
Carsten.Scholtes@Uni-Bayreuth.de

The efficient use of the cache hierarchy of an execution platform often has a major impact on the performance of an application. It is often difficult for the application programmer or the compiler to determine a suitable memory layout for the application data, since the interactions between the memory accesses cannot be fully anticipated before the program execution.

This paper introduces an approach to improve the cache efficiency by dynamically padding memory allocations by a post-compilation tool. Data structures like histograms are used to evaluate cache simulations of memory traces of the application considered and to compute optimized pad sets. These can then be used for later runs of the application with different data sets. The accumulated representation of the references' memory accesses additionally offers a visualization interface to the algorithm-specific memory access pattern of each reference captured. As implied above, the advantage of the method is that it also allows for an improvement of the cache usage of binary only applications for which no source code is available. Post optimization cache behavior analyses as well as run-time measurements show that the cache hit rates of the runtime-modified applications are considerably increased by applying the generated pad set.

On Synthesizing Workloads Emulating MPI Applications

Javier Navaridas, Jose Miguel-Alonso, Francisco Javier Ridruejo.
Department of Computer Architecture and Technology,
The University of the Basque Country P.O. Box 649, 20080 San Sebastian, SPAIN
Tel. +34 943018019 Fax +34 943015590
{javier.navaridas, j.miguel, franciscojavier.ridruejo}@ehu.es

Evaluation of high performance parallel systems is a delicate issue, due to the difficulty of generating workloads that represent, with fidelity, those that will run on actual systems. In this paper we make an overview of the most usual methodologies used to generate workloads for performance evaluation purposes, focusing on the network: random traffic, patterns based on permutations, traces, execution-driven, etc. In order to fill the gap between purely synthetic and application-driven workloads, we present a set of pseudo-synthetic workloads that mimic applications behavior, emulating some widely-used implementations of MPI collectives and some communication patterns commonly used in scientific applications. This mimicry is done in terms of spatial distribution of messages as well as in the causal relationship among them. As an example of the proposed methodology, we use a subset of these workloads to confront tori and fat-trees.

Performance Effects of Gram-Schmidt Orthogonalization on Multi-Core Infiniband Clusters

Gudula Rünger and Michael Schwind
Department of Computer Science,
Technical University Chemnitz, Germany
{ruenger, schwi}@informatik.tu-chemnitz.de

In this article, we investigate how the distribution of dense matrices in a cluster of multi-core systems affects the performance of linear algebra codes. These codes have the property of communicating only within rows or columns of a processor grid. Especially, we consider orthogonalization methods like the Gram- Schmidt orthogonalization. Experiments show the performance for different topology aware mappings on two different multi-core clusters connected with Infiniband. We show how the performance is influenced by these mappings.

Finite-Difference Time-Domain on the Cell/B.E. Processor

Meilian Xu and Parimala Thulasiraman
University of Manitoba
Dept. of Computer Science
Winnipeg, Manitoba, Canada
{maryx,thulasir}@cs.umanitoba.ca

Finite-Difference Time-Domain (FDTD) is a kernel used to solve problems in electromagnetics applications such as microwave tomography. It is a data-intensive and computation-intensive problem. However, its computation scheme indicates that an architecture with SIMD support has the potential to bring performance improvement over traditional architectures without SIMD support. The Cell Broadband Engine (Cell/B.E.) processor is an implementation of a heterogeneous multicore architecture. It consists of one conventional microprocessor, PowerPC Processor Element (PPE), and eight SIMD co-processor elements, Synergistic Processor Elements (SPEs). One unique feature of an SPE is that it has 128-entry 128-bit uniform registers which support SIMD. Therefore, FDTD may be mapped well on Cell/B.E. processor. However, each SPE can directly access only 256KB local store (LS) both for instructions and data. The size of LS is much less than what is needed for an accurate simulation of FDTD which requires large number of fine-grained Yee cells. In this paper, we design the algorithm on Cell/B.E. by efficiently using the asynchronous DMA (direct memory access) mechanism available on an SPE transferring data between its LS and the main memory via the high bandwidth bus on-chip EIB (Element Interconnect Bus). The new algorithm was run on an IBM Blade QS20 blades running at 3.2GHz. For a computation domain of 600×600 Yee cells, we achieve an overall speedup of 14.14 over AMD Athlon and 7.05 over AMD Opteron at the processor level.

Parallel Performance Wizard: A Performance Analysis Tool for Partitioned Global-Address-Space Programming

Hung-Hsun Su Max Billingsley III and Alan D. George
High-performance Computing and Simulation (HCS) Research Lab
Dept. of Electrical and Computer Engineering
University of Florida, Gainesville, Florida 32611-6200
{su,billingsley,george}@hcs.ufl.edu

Given the complexity of parallel programs, developers often must rely on performance analysis tools to help them improve the performance of their code. While many tools support the analysis of message-passing programs, no tool exists that fully supports programs written in programming models that present a partitioned global address space (PGAS) to the programmer, such as UPC and SHMEM. Existing tools with support for message-passing models cannot be easily extended to support PGAS programming models, due to the differences between these paradigms. Furthermore, the inclusion of implicit and one-sided communication in PGAS models renders many of the analyses performed by existing tools irrelevant. For these reasons, there exists a need for a new performance tool capable of handling the challenges associated with PGAS models. In this paper, we first present background research and the framework for Parallel Performance Wizard (PPW), a modularized, event-based performance analysis tool for PGAS programming models. We then discuss features of PPW and how they are used in the analysis of PGAS applications. Finally, we illustrate how one would use PPW in the analysis and optimization of PGAS applications by presenting a small case study using the PPW version 1.0 implementation.

Towards an Adaptive Task Pool Implementation

M. Hofmann and G. Rünger
Department of Computer Science
Chemnitz University of Technology, Germany
E-mail: {mhofma,ruenger}@informatik.tu-chemnitz.de

Task pools can be used to achieve the dynamic load balancing that is required for an efficient parallel implementation of irregular applications. However, the performance strongly depends on a task pool implementation that is well suited for the specific application. This paper introduces an adaptive task pool implementation that enables a stepwise transition between the common strategies of central and distributed task pools. The influence of the task size on the parallel performance is investigated and it is shown that the adaptive implementation provides the flexibility to adapt to different situations. Performance results from benchmark programs and from an irregular application for anomalous diffusion simulation are presented to demonstrate the need for an adaptive strategy. It is shown that profiling information about the overhead of the task pool implementation can be used to determine an optimal task pool strategy.

Performance of MPI Broadcast Algorithms

Daniel M. Wadsworth and Zizhong Chen
Jacksonville State University
MCIS Department
Jacksonville, AL 36265, USA
dan@wadsworthsystems.com and zchen@jsu.edu

As cluster computing has gotten cheaper and more powerful, ever larger clusters are being built and ever larger problems are being tackled. Many, if not most, of these problems require the broadcast of data not only at the beginning of the computation but also as it progresses. Given the comparative slowness of communication as opposed to computation, any improvement in communication time can have a significant impact on the time required to complete a computation. This paper looks at the broadcast function of MPI and explores some alternative implementations for the common cluster architecture in which some number of computing nodes are connected via an Ethernet switch. It then compares these implementations to the built-in implementation of broadcast in MPICH2 and suggests a new implementation.

DLBEM: Dynamic Load Balancing Using Expectation-Maximization

Han Zhao, Xinxin Liu and Xiaolin Li
Scalable Software Systems Laboratory
Department of Computer Science
Oklahoma State University
Stillwater, OK 74048, USA
Email: {haz, xinxin, xiaolin}@cs.okstate.edu

This paper proposes a dynamic load balancing strategy called DLBEM based on maximum likelihood estimation methods for parallel and distributed applications. A mixture Gaussian model is employed to characterize workload in data-intensive applications. Using a small subset of workload information in systems, the DLBEM strategy reduces considerable communication overheads caused by workload information exchange and job migration. In the meantime, based on the Expectation-Maximization algorithm, DLBEM achieves near accurate estimation of the global system state with significantly less communication overheads and results in efficient workload balancing. Simulation results for some representative cases on a two-dimensional 16*16 grid demonstrate that DLBEM approach achieves even resource utilization and over 90% accuracy in the estimation of the global system state information with over 70% reduction on communication overheads compared to a baseline strategy.

Workshop 15

Performance Modeling, Evaluation, and Optimisation of Ubiquitous Computing and Networked Systems

PMEO 2008

Heuristic Algorithms for Routing and Wavelength Assignment in WDM Optical Networks

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
lik@newpaltz.edu

We evaluate the average-case performance of eight offline heuristic algorithms to solve the routing and wavelength assignment problem and the related throughput maximization problem in wavelength division multiplexing optical networks. Our experimental performance evaluation is conducted by extensive simulations on a wide range of WDM optical networks. We find offline routing and wavelength assignment algorithms and throughput maximization algorithms which perform better than previously studied online algorithms.

A Robust and Energy Efficient Protocol for RandomWalk in Ad hoc Networks with IEEE 802.11

Adnan Noor Mian, Roberto Beraldi and Roberto Baldoni
DIS, Università di Roma "La Sapienza",
Via Ariosto, 25
Roma, 00185, Italy
{adnan, beraldi, baldoni}@dis.uniroma1.it

This paper is about energy efficient and robust implementation of random walks in mobile wireless networks. While random walk based algorithm are often proposed to solve many problems in wireless networks, their implementation is usually done at the application layer so that many characteristics of the wireless transmissions are not exploited. In this paper we show that we can greatly reduce the energy requirements to perform a walk by better exploiting the broadcast nature of the transmissions. We propose a robust, energy efficient distributed next hop selection algorithm. To evaluate the algorithm we present a simulation study performed with ns-2. We found that in the proposed algorithm energy is reduced to more than 4 times and the selection delay is reduced to more than 8 times as compared to a standard next hop selection implementation.

PLP: Towards a Realistic and Accurate Model for Communication Performances on Hierarchical Cluster-based Systems

Wahid Nasri, Olfa Tarhouni and Nadia Slimi
ESSTT, Département d'Informatique,
5, Avenue Taha Hussein - B.P. 56, Bab Menara - 1008 Tunis, Tunisia
Wahid.Nasri@ensi.rnu.tn

Today, due to many reasons, such as the inherent heterogeneity, the diversity, and the continuous evolving of actual computational supports, writing efficient parallel applications on such systems represents a great challenge. One way to answer this problem is to optimize communications of such applications. Our objective within this work is to design a realistic model able to accurately predict the cost of communication operations on execution environments characterized by both heterogeneity and hierarchical structure. We principally aim to guarantee a good quality of prediction with a neglected additional overhead. The proposed model was applied on point-to-point and collective communication operations and showed by achieving experiments on a hierarchical cluster-based system with heterogeneous resources that the predicted performances are close to measured ones.

DIBS: Dual Interval Bandwidth Scheduling for Short-term Differentiation

Humzah Jaffar and Xiaobo Zhou
Department of Computer Science
University of Colorado at Colorado Springs
Colorado Springs, CO 80918

Liqiang Zhang
Dept. of Computer and Information Sciences
Indiana University South Bend
South Bend, IN 46615

Packet delay and bandwidth are two important metrics for measuring quality of service (QoS) of Internet services. While proportional delay differentiation (PDD) has been studied intensively in the context of Differentiated Services, few studies were conducted for per-class bandwidth differentiation. In this paper, we design and evaluate an efficient bandwidth differentiation approach. The DIBS (dual interval bandwidth scheduling) approach focuses on the short-term bandwidth differentiation of multiple classes because many Internet transactions take place in a small time frame. It does so based on the normalized instantaneous bandwidth, measured by the use of packet size and packet delay. It also proposes to use a look-back interval and a look-ahead interval to trade off differentiation accuracy and scheduling overhead. We implemented DIBS in the Click Modular Software Router. Extensive experiments have demonstrated its feasibility and effectiveness in achieving short-term bandwidth differentiation. Compared with the representative PDD algorithm WTP, DIBS can achieve better bandwidth differentiation when the interclass packet size distributions are different. Compared with the representative weighted fair queueing algorithm PGPS, DIBS can achieve more accurate or comparable bandwidth differentiation at various workload situations, with better delay differentiation and lower cost.

The Impact of Routing Schemes on Group Communication Throughput in Scalable Networks

Ahmed Yassin Al-Dubai
School of Computing, Napier University, 10 Colinton Road
Edinburgh, EH10 5DT,UK
Email: a.aldubai@napier.ac.uk

Multicast communication has been a major avenue for many studies in interconnection networks. However, such a communication pattern has been studied under either limited operating conditions or within the context of deterministic routing algorithms. This paper investigates the impact of routing algorithms, both deterministic and adaptive routing on the multicast communication over interconnection network. In particular, we focus on the provision of efficient multicast communication algorithms in interconnection networks. Using detailed simulation experiments, different multicast algorithms have been compared for a range of system sizes, traffic loads, and destination nodes. For the case of multicast latency, our proposed algorithms exhibit the best performance when the traffic load is high and the start-up overhead does not dominates the propagation overhead. The results also highlight the impact of adaptive routing on both latency and throughput when designing efficient multicast algorithms. Thus, these results demonstrate significant potential to be applied to current and future generation interconnection networks.

The Effect of Real Workloads and Stochastic Workloads on the Performance of Allocation and Scheduling Algorithms in 2D Mesh Multicomputers

S. Bani-Mohammad, M. Ould-Khaoua,
and Lewis M. Mackenzie
Glasgow University, Computing
Science,
Glasgow G12 8RZ, UK.
{saad, mohamed, lewis}@dcs.gla.ac.uk

I. Ababneh
University of Science
and Technology,
Computing Science,
Irbid, Jordan.
ismael@just.edu.jo

J. D. Ferguson
Strathclyde University,
Computer and Information
Sciences,
Glasgow G1 1XH, UK.
john.ferguson@cis.strath.ac.uk

The performance of the existing non-contiguous processor allocation strategies has been traditionally carried out by means of simulation based on a stochastic workload model to generate a stream of incoming jobs. To validate the performance of the existing algorithms, there has been a need to evaluate the algorithms' performance based on a real workload trace. In this paper, we evaluate the performance of several well-known processor allocation and job scheduling strategies based on a real workload trace and compare the results against those obtained from using a stochastic workload. Our results reveal that the conclusions reached on the relative performance merits of the allocation strategies when a real workload trace is used are in general compatible with those obtained when a stochastic workload is used.

A Static Data Placement Strategy towards Perfect Load-Balancing for Distributed Storage Clusters

Deepthi K.Madathil, Rajani B. Thota, Paulina Paul and Tao Xie
Department of Computer Science, San Diego State University
5500 Campanile Drive, San Diego, CA 92182, USA
Deepthi-madathil@rohan.sdsu.edu
Rajani-thota@rohan.sdsu.edu
ppaul@rohan.sdsu.edu
xie@cs.sdsu.edu

Applications like cluster-based Video-On-Demand (VOD) systems are inherently data-intensive because clients frequently retrieve data stored on a distributed storage subsystem interconnected by a high-speed local network. To meet the Quality-of-Service (QoS) imposed by the clients, quick responses to access requests are fundamental for these applications. Among the numerous ways to reduce response times, data placement, has attracted much attention from researchers due to its effectiveness and low cost. In this paper, we propose a novel load-balancing and performance oriented static data placement strategy, called perfect balancing (PB), which can be applied to distributed storage subsystems in clusters to noticeably improve system responsiveness. The basic idea of PB is to balance the load across local disks and to minimize the discrepancy of service times of data on each disk simultaneously. A comprehensive experimental study shows that PB reduces mean response time up to 19.04% and 8.67% over the two well known data placement algorithms Greedy and SP respectively.

A Locally-Optimizing Approach for Multichannel Assignment and Routing

Lei Tang, Xiaoyan Hong and Bo Gu
Department of Computer Science,
University of Alabama, Tuscaloosa, AL 35487
fltang, hxy, bgug@cs.ua.edu

In this paper, we propose a distributed online multichannel assignment and routing algorithm called LOAM. Using LOAM algorithm, a node in the network assigns channels for the incoming traffic flows and forwards them based on the channel usage information of its nearby nodes. We also present a model to analyze the throughput of multichannel networks and prove LOAM can achieve more than a fraction of the optimal throughput. Furthermore, we compare the achievable throughput of LOAM with a centralized offline channel assignment algorithm in average cases considering different channel rate and traffic load scenarios, the results of which demonstrate that LOAM can achieve more than 90% of theoretical maximum throughput and has a more balanced load distribution. Finally, we evaluate LOAM in a realistically simulated 802.11a wireless multihop network, compared with a random channel assignment algorithm. The simulation results show that LOAM can deliver up to 25 percent more messages with a shorter delivery latency.

Performance Comparison of SGI Altix 4700 and SGI Altix 3700 Bx2

Subhash Saini, Dennis C. Jespersen, Dale Talcott,
Jahed Djomehri and Timothy Sandstrom
NASA Advanced Supercomputing Division, NASA Ames Research
Center, Moffett Field, California, 94035-1000, USA
{Subhash.Saini, Dennis.C.Jespersen, Dale.R.Talcott,
Mohammad.J.Djomehri, Timothy.A.Sandstrom}@nasa.gov

Suitability of the next generation of high-performance computing systems for petascale simulations will depend on a balance between factors such as processor performance, memory performance, local and global network performance, and Input/Output (I/O) performance. As the supercomputing industry develops new technologies for these subsystems, achieving system balance becomes challenging. In this paper, we evaluate the performance of a newly introduced dual-core-based SGI Altix 4700 system and we compare its performance with that of a single-core-based SGI Altix 3700 Bx2 system. We used the High-Performance Computing Challenge (HPCC) benchmarks and five real-world applications, three from computational fluid dynamics, one from climate modeling and one from nanotechnology. Our study shows that the SGI Altix 4700 performs slightly better than the SGI Altix 3700 Bx2 up to 128 processors, while the performance of the systems is almost the same beyond 128 processors, when the communication time dominates the compute time.

Accurately Measuring Collective Operations at Massive Scale

Torsten Hoefler[1,2], Timo Schneider[2] and Andrew Lumsdaine[1]

[1]Open Systems Laboratory
Indiana University
501 N. Morton Street
Bloomington, IN 47404 USA
{htor,lums}@cs.indiana.edu

[2]Dept. of Computer Science
Technical University of Chemnitz
Strasse der Nationen 62
Chemnitz, 09107 GERMANY
{htor,timos}@cs.tu-chemnitz.de

Accurate, reproducible and comparable measurement of collective operations is a complicated task. Although Different measurement schemes are implemented in well-known benchmarks, many of these schemes introduce different systematic errors in their measurements. We characterize these errors and select a window-based approach as the most accurate method. However, this approach complicates measurements significantly and introduces a clock synchronization as a new source of systematic errors. We analyze approaches to avoid or correct those errors and develop a scalable synchronization scheme to conduct benchmarks on massively parallel systems. Our results are compared to the window-based scheme implemented in the SKaMPI benchmarks and show a reduction of the synchronization overhead by a factor of 16 on 128 processes.

A Mesh Hybrid Adaptive Service Discovery Protocol (MesHASeDiP): Protocol Design and Proof of Correctness

Kaouther Abrougui and Azzedine Boukerche
SITE university of Ottawa
Email:{abrougui,boukerch}@site.uottawa.ca

The characteristics of Wireless Mesh Networks (WMNs) have motivated us in the design of an efficient service discovery protocol that considers the capabilities of such networks. In this paper we propose a novel service discovery technique for WMNs. Our approach reduces the discovery overhead by integrating the discovery information in the routing layer. Based on an adaptively adjusted advertisement zone of service providers, we have combined proactive and reactive service discovery strategies to come up with an efficient hybrid adaptive service discovery protocol for WMNs. Our protocol optimizes the network overhead. We will show that our proposed protocol is scalable and that it outperforms existing service discovery protocols in terms of message overhead.

A Novel Distribution Signaling Protocol Based on Mobile Agents and OBS for Intelligent Optical Networks

Wei Li[1], Yuan Li[1,2], Hong Liu[1], Nan Chi[1] and Dexiu Huang[1]
[1] Wuhan National Lab of Optoelectronics,
Huazhong University of Science and Technology, Wuhan,430074, P.R.China
[2] Department of Computer Science,
Central China Normal University, Wuhan,430079,P.R.China.
E-mail:hongliu28@gmail.com; weilee@hust.edu.cn

A novel optical signaling protocol is proposed in which the mobile agent MANet is used to transport the distribution network control and management information among the nodes. The 1310nm band is designed to be a control optical wavelength channel to transmit BHPs and 1550nm band is used to be burst packets optical wavelength channels to transmit BPs. Referring to RSVP-TE in DCM ASON, significations and formats of signaling messages in the protocol are defined. The key design of scheme is to utilize the structure of active packets in the application of mobile agent MANet to encapsulate signaling messages and traffic together into one BHP to shorten the set-up time of an optical path. Such structure is also used to process distributed disposals to alleviate the workload to promote cooperating calculation ability of each node. Good efficiencies are shown by simulations of establishment time and blocking probability in a 40-nodes optical mesh network.

Ad-hoc Routing Metrics and Applied Weighting for QoS support

Migas, Nikos and Buchanan W. J.
School of Computing, Napier University, EH10 5DT, Scotland, UK
{n.migas, w.buchanan}@napier.ac.uk

In the vast majority of ad-hoc routing protocols, the hop-counting mechanisms for identifying the optimal route are dominant. However, this approach oversimplifies such a complex decision by ignoring the fact that participating devices may have considerably unequal performance characteristics and current utilisation status. Accordingly, it is possible for an optimal route to be composed of devices with high utilisation status, or, low battery reserves, which results in an overall unreliable route. This research work tackles this by identifying the best metrics that can describe any route within a graph, in terms of overall throughput, reliability, and minimum energy consumption. Simulations were carried out by varying critical factors of mobile devices such as battery reserves, memory and CPU utilisation, and results recorded the effect that this has on the device's overall routing metric. This paper also presents the threshold values, which turn the device from routing-capable to routing-incapable state.

Proportionally-Fair Best Effort Flow Control in Network-on-Chip Architectures

Mohammad S. Talebi[1], Fahimeh Jafari[1,2], Ahmad Khonsari[3,1] and Mohammad H. Yaghmaee[2]
[1] IPM, School of Computer, Tehran, Iran
[2] Ferdowsi University of Mashhad, Mashahhad, Iran
[3] ECE Department, University of Tehran,Tehran, Iran
mstalebi@ipm.ir, jafari@ipm.ir, ak@ipm.ir, hyaghmae@ferdowsi.um.ac.ir

The research community has recently witnessed the emergence of Multi-Processor System on Chip (MPSoC) platforms consisting of a large set of embedded processors. Particularly, Interconnect networks methodology based on Network-on-Chip (NoC) in MP-SoC design is imminent to achieve high performance potential. More importantly, many well established schemes of networking and distributed systems inspire NoC design methodologies. Employing end-to-end congestion control is becoming more imminent in the design process of NoCs. This paper presents a centralized congestion scheme in the presence of both elastic and streaming flow traffic mixture. In this paper, we model the desired Best Effort (BE) source rates as the solution to a utility maximization problem which is constrained with link capacities while preserving Guaranteed Service (GS) traffics services requirements at the desired level. We proposed an iterative algorithm as the solution to the maximization problem which has the benefit of low complexity and fast convergence. The proposed algorithm may be implemented by a centralized controller with low computation and communication overhead.

Web-Based E-Learning in 3D Large Scale Distributed Interactive Simulations Using HLA/RTI

L. Ahmad, A. Boukerche, A. Al Hamidi, A. Shadid and R. Pazzi
Paradise Research Lab, School of Information Technology and Engineering (SITE)
University of Ottawa, Ottawa, Canada
{lahma016, boukerch, aalha033, ashad099, rwerner}@site.uottawa.ca

In this paper, we investigate the integration of HLA/RTI with VRML/X3D in order to provide a training facility that can be accessed through a traditional web browser. We apply our proposed architecture on a test bed that consists of a cancer treatment scenario within a 3D virtual Radiology Department. We have incorporated the High Level Architecture HLA standard as an important corner stone in allowing the e-learning environment to be distributed and interactive. Combining Virtual Reality concepts with the real pivots of HLA/RTI can create a very strong foundation for such a valuable and required class of applications by utilizing all reliable technologies to make it as flexible as possible. We have ensured that any foreseen future optimizations could be incorporated easily without any overhead.

Performance Evaluation of NSF Application Benchmarks on Parallel Systems

Rod Fatoohi
San Jose State University
San Jose, California, 95192, USA
rfatoohi@sjsu.edu

The National Science Foundation (NSF) recently released a set of application benchmarks that would be a key factor in selecting the next-generation high performance computing environment. These benchmarks are designed to capture the salient attributes of those science and engineering applications placing the most stringent demands on the system to be provisioned. The application benchmarks consist of six codes that require large amount of memory and work with large data sets. In this work, we study the complexity, performance, and scalability of these codes on four machines: a 512-processor SGI Altix 3700, a 512-processor SGI Altix 3700/BX2, a 512-processor dual-core based SGI Altix 4700, and a 128-processor Cray Opteron cluster interconnected by the Myrinet network. We evaluated these codes for two different problem sizes using different numbers of processors. Our results show that per processor the SGI machines, using the Intel Itanium-2 processor, are faster than the Cray cluster, using the AMD Opteron processor, by a factor of up to three. Also, we found out that some of these codes scale up very well as we increase the number of processors while others scaled up poorly. In addition, one of the codes achieved about 2/3 of the peak rate of an SGI Altix processor. Moreover, the dual-core based system achieved comparable performance results to the single-core based system. Finally, we provide some limitations and concluding remarks.

A Novel High-Performance and Low-Power Mesh-Based NoC

Reza Sabbaghi-Nadooshan[1,2], Mehdi Modarressi[3,4] and Hamid Sarbazi-Azad[3,4]
[1]Islamic Azad University Science and Research Branch, Tehran, Iran
[2]Islamic Azad University Central Tehran Branch, Tehran, Iran
[3]IPM School of computer science, Tehran, Iran
[4]Sharif University of Technology, Tehran, Iran
r_sabbaghi@iauctb.ac.ir, modarressi@ce.sharif.edu, azad@sharif.edu

In this paper, a 2D shuffle-exchange based mesh topology, or 2D SEM (Shuffle-exchange Mesh) for short, is presented for network-on-chips. The proposed two-dimensional topology applies the conventional well-known shuffle-exchange structure in each row and each column of the network. Compared to an equal sized mesh which is the most common topology in onchip networks, the proposed shuffle-exchange based mesh network has smaller diameter but for an equal cost. Simulation results show that the 2D SEM effectively reduces the power consumption and improves performance metrics of the on-chip networks with regard to the conventional mesh topology.

Workshop 16

Dependable Parallel, Distributed and Network-Centric Systems

DPDNS 2008

Achieving and Assuring High Availability

K. Trivedi	G. Ciardo	B. Dasarathy	M. Grottke	A. Rindos	B. Varshaw
Duke University	UC Riverside	Telcordia Univ.	Erlangen-Nbg.	IBM	IBM

kst@ee.duke.edu, ciardo@cs.ucr.edu, das@research.telcordia.com
michael.grottke@wiso.uni-erlangen.de, rindos@us.ibm.com, vashaw@us.ibm.com

We discuss availability aspects of large software-based systems. We classify faults into Bohrbugs, Mandelbugs and aging-related bugs, then examine mitigation methods for the last two bug types. We also consider quantitative approaches to availability assurance.

An Overlay Protection Layer against Denial-of-Service Attacks

Hakem Beitollahi Geert Deconinck
Hakem.Beitollahi@esat.kuleuven.be Geert.Deconinck@esat.kuleuven.be

Katholieke Universiteit Leuven
Electrical Engineering
Kasteelpark Arenberg 10, Leuven, Belgium

Today Internet is becoming an emerging technology for remote control of industrial applications, where one site needs to control another site remotely (e.g. power plants controllers). Denial-of-Service (DoS) attacks may cause significant disruptions to the Internet which will threaten the operation of such network based control systems. Over lay networks have been proposed to protect Internet appli cation sites by location-hiding technique. This paper an alyzes a large domain of previous approaches against this problem. This paper addresses how an interface to an overlay net work can be designed such that communication services among geographically distributed application sites are se cured against DoS attacks. This paper presents a novel ar chitecture called overlay protection layer (OPL) that proac tively protect application sites from DoS attacks. Through simulation this paper shows DoS attacks have a negligible chance to disrupt communications services via the OPL architecture. Even if attackers attack 50% of over lay nodes via a Distributed DoS attack still 75% of commu nication channels are available.

Extending Algorithm-Based Fault Tolerance to Tolerate Fail-Stop Failures in High Performance Distributed Environments

Zizhong Chen
Jacksonville State University
MCIS Department
Jacksonville, AL 36265, USA
zchen@jsu.edu

It has been proved in previous algorithm-based fault tolerance that, for matrix matrix multiplication, the checksum relationship in the input checksum matrices is preserved at the end of the computation no mater which algorithm is used. However, whether this checksum relationship can be maintained in the middle of the computation or not remains open. In this paper, we first demonstrate that this checksum relationship is not maintained in the middle of the computation for most algorithms for matrix matrix multiplication. We then prove that, however, for the outer product version matrix matrix multiplication algorithm, this checksum relationship can be maintained in the middle of the computation. Based on this checksum relationship maintained in the middle of the computation, we demonstrate that fail-stop process failures (which are often tolerated by checkpointing or message logging) in the outer product version matrix-matrix multiplication can be tolerated without checkpointing or message logging.

An Integrated Approach for Reliability and Dependability of Wireless Mesh Networks

Georg Lukas, André Herms, Svilen Ivanov and Edgar Nett
University of Magdeburg
Institute for Distributed Systems
Universitätsplatz 2, 39106 Magdeburg, Germany
{glukas,aherms,svilen,nett}@ivs.cs.uni-magdeburg.de

Wireless Mesh Networks (WMNs) are gaining popularity in many application areas because of their low cost and high exibility. Technically, a WMN forms a distributed network-centric system which aims to provide communication services to the application layer, but has no built-in dependability provisions. To make the network dependable, many problems have to be solved on different layers. With this paper, we describe our ongoing work to provide an integrated solution to increase the dependability of WMNs. Our approach combines network coverage planning on the physical layer, bandwidth management on the link layer and live network monitoring to improve the reliability, availability and maintainability of a WMN. We provide fine-grained means to improve the predictability of the network components, thus making the WMN more dependable. In this paper, we present first results of our work, and describe how they are interleaved.

Rate Avalanche: The Performance Degradation in Multi-rate 802.11 WLANs

Liqiang Zhang and Yu-Jen Cheng
Dept. of Computer & Information Sciences
Indiana University South Bend
South Bend, IN 46615, USA

Xiaobo Zhou
Dept. of Computer Science
University of Colorado at Colorado Springs
Colorado Springs, CO 80918, USA

The Request-to-Send/Clear-to-Send (RTS/CTS) exchange was defined as an optional mechanism in DCF (Distributed Coordination Function) access method in IEEE 802.11 standard to deal with the hidden node problem. However, in most infrastructure-based WLANs, it is turned off with the belief that the benefit it brings might not even be able to pay off the transmission overhead it introduces. While this is often true for networks using fixed transmission rate, our investigation leads to the opposite conclusion when multiple transmission rates are exploited in WLANs. In particular, through extensive simulations using realistic channel propagation and reception models, we found out that in a heavily loaded multi-rate WLAN, a situation that we call rate avalanche often happens if RTS/CTS is turned off. The rate avalanche effect could significantly degrade the network performance even if no hidden node presents. Our investigation also reveals that, in the absence of effective and practical loss-differentiation mechanisms, simply turning on the RTS/CTS could dramatically enhance the network performance in most cases. Various scenarios/conditions are extensively examined to study their impact on the network performance for RTS/CTS on and off respectively. Our study provides some important insights about using the RTS/CTS exchange in mutlirate 802.11 WLANs.

Efficient and Dependable Overlay Networks

Jawwad Shamsi and Monica Brockmeyer
Wayne State University
{jshamsi, mbrockmeyer}@wayne.edu

We describe QoSMap, a novel approach for mapping overlay networks with stringent QoS requirements. QoSMap computes high quality overlay paths by utilizing direct underlay paths that fulfill the application's topology and link-specific QoS requirements. It also improves application dependability by increasing resiliency against QoS failures under dynamic network conditions. The increased resiliency is achieved through the use of supplemental indirect paths which have a single intermediate node and which also meet application QoS requirements. For evaluation, we compared QoSMap with a simple QoS approach for overlay mapping. We compared the two approaches for fulfilling varying application requests with strict QoS requirements of latency and packet loss on PlanetLab. We observed that QoSMap yielded high quality overlays. Further, the use of indirect paths upon the QoS failure of the direct path avoided overlay reconfiguration and resulted in increased dependability and longer lifetime for the application overlays.

Fault Tolerant Practices on Network Processors for Dependable Network Processing

Yan Luo and Jie Fan
Dept. of Electrical and Computer Engineering
University of Massachusetts Lowell
yan_luo@uml.edu, jie_fan@student.uml.edu

In this paper, we study how to provide dependable network processing through multi-core based network processors (NPs). We present the performance analysis results of an NP based network system and motivate our research. We propose to use the redundant cores available in an NP to handle faults. We outline a fault-tolerant (FT) task model specifically for NP based applications and describe our implementation taking advantage of hardware features of an Intel IXP2xxx NP. A set of experiments are conducted to evaluate the performance and effectiveness of a FT-enabled NP system. The experiment results show that our faulttolerant design can effectively improve the schedulability of the system.

State Management for Distributed Python Applications

Gregory D. Benson
Department of Computer Science
University of San Francisco
benson@cs.usfca.edu

We present a novel state management mechanism that can be used to capture the complete execution state of distributed Python applications. This mechanism can serve as the foundation for a variety of dependability strategies including checkpointing, replication, and migration. Python is increasingly used for rapid prototyping parallel programs and, in some cases, used for high-performance application development using libraries such as NumPy. Building on Stackless Python and the River parallel and distributed programming environment, we have developed mechanisms for state capture at the language level. Our approach allows for migration and checkpointing of applications in heterogeneous environments. In addition, we allow for preemptive state capture so that programmers need not introduce explicit snapshot requests. Our mechanism can be extended to support application or domain-specific state capture. To our knowledge, this is the first general checkpointing scheme for Python. We describe our system, the implementation, and give some initial performance figures.

A Fault-tolerant System for Java/CORBA Objects

Atul Kumar
Accenture Technology Labs
IBC Knowledge Park, Bannerghatta Road, Bangalore - 560029, India
atulk@ieee.org

Frameworks like CORBA facilitate the development of distributed applications through the use of off-the-shelf components. Though the use of distributed components allows faster building of applications, it also reduces the application availability as failure of any component can make the application unavailable. In this paper we present the design and implementation of a fault-tolerant system for CORBA objects implemented in Java. The proposed fault tolerant system employs object replication. We use a three tier architecture in which the middle tier manages replication and acts as a proxy for replicated objects. The proxy ensures consistency and transparency. In the current implementation, the proxy uses the primary-site approach to ensure strong consistency. Saving and restoring of objects' state is done transparently and it does not require object implementation to have special functions implemented for this purpose.

Accuracy vs Efficiency of Hyper-exponential Approximations of the Response Time Distribution of MMPP/M/1 Queues

Paolo Romano, Bruno Ciciani, Andrea Santoro and Francesco Quaglia
Dipartimento di Informatica e Sistemistica
Sapienza Università di Roma

The Markov Modulated Poisson Process (MMPP) has been shown to well describe the flow of incoming traffic in networked systems, such as the GRID and the WWW. This makes the MMPP/M/1 queue a valuable instrument to evaluate and predict the service level of networked servers. In a recent work we have provided an approximate solution for the response time distribution of the MMPP/M/1 queue, which is based on a weighted superposition of M/M/1 queues (i.e. a hyper-exponential process). In this article we address the tradeoff between the accuracy of this approximation and its computational cost. By jointly considering both accuracy and cost, we identify the scenarios where such approximate solution could be effectively used in support of network servers (dynamic) configuration and evaluation strategies, aimed at ensuring the agreed dependability levels in case of, e.g., request redirection due to faults.

Improving Data Availability for a Cluster File System through Replication

Jin Xiong, Jianyu Li and Rongfeng Tang
Key Laboratory of Computer System and Architecture
Institute of Computing Technology
Chinese Academy of Sciences
{xj, lijianyu, rf_tang}@ncic.ac.cn

Yiming Hu
Department of Electrical & Computer
Engineering and Computer Science
University of Cincinnati
yhu@ececs.uc.edu

Data availability is a challenging issue for largescale cluster file systems built upon thousands of individual storage devices. Replication is a well-known solution used to improve data availability. However, how to efficiently guarantee replicas consistency under concurrent conflict mutations remains a challenge. Moreover, how to quickly recover replica consistency from a storage server crash or storage device failure is also a tough problem. In this paper, we present a replication-based data availability mechanism designed for a large-scale cluster file system prototype named LionFS. Unlike other replicated storage systems that serialize replica updates, LionFS introduces a relaxed consistency model to enable concurrent updating all replicas for a mutation operation, greatly reducing the latency of operations. LionFS ensures replica consistency if applications use file locks to synchronize the concurrent conflict mutations. Another novelty of this mechanism is its light-weight log, which only records failed mutations and imposes no overhead on failure-free execution and low overhead when some storage devices are unavailable. Furthermore, recovery of replica consistency needs not stop the file system services and running applications. Performance evaluation shows that our solution achieves 50-70% higher write performance than serial replica updates. The logging overhead is shown to be low, and the recovery time is proportional to the amount of data written during the failure.

Investigating Fau75 Tolerant Computing Systems Reliability

Salvatore Distefano
University of Messina, Engineering Faculty
Contrada di Dio, S. Agata, 98166 Messina, Italy.
Tel: +39 090 3977318, fax: +39 090 3977471
Email: sdistefano@unime.it

Nowadays, computers and networks represent the heart of a great part of modern technologies. Computing systems are widely used in many application areas, and they are desired to achieve various complex and safety-critical missions. As consequence, greater attention is lavished on performance and dependability evaluation of computing systems. This brings to the specification of precise techniques and models, that consider and evaluate aspects before (consciously or unconsciously) approximated or ignored at all. On the other hand, the increasing importance assumed by such systems is translated in terms of tighter and tighter constraints, requirements and/or policies (QoS, fault tolerance, maintenance, redundancy, etc.) according to the systems' criticism. The evaluation must therefore take into great account such dynamic behaviors, carefully identifying and quantifying dependencies among devices. In this paper we face the problem of individuating and evaluating the most common dynamic behaviors and dependencies affecting fault tolerant computing systems. We propose some models to represent such aspects in terms of reliability/availability, basing on dynamic reliability block diagrams (DRBD), a new formalism derived from RBD we developed. In this way we want to provide the guidelines for adequately evaluating fault tolerant computing system reliability/availability.

Fault-Tolerant Static Scheduling for Grids

Bernhard Fechner, Udo Hönig, Jörg Keller and Wolfram Schiffmann
FernUniversität in Hagen
Department of Mathematics and Computer Science
58084 Hagen, Germany
{Bernhard.Fechner,Udo.Hoenig,Joerg.Keller,Wolfram.Schiffmann}@fernuni-hagen.de

While fault-tolerance is desirable for grid applications because of the distributed and dynamic nature of grid resources, it has seldom been considered in static scheduling. We present a fault-tolerant static scheduler for grid applications that uses task duplication and combines the advantages of static scheduling, namely no overhead for the faultfree case, and of dynamic scheduling, namely low overhead in case of a fault. We also give preliminary experimental results on our scheme.

Workshop 17

International Workshop on Security in Systems and Networks

SSN 2008

Countering Phishing Threats with Trust-by-Wire in Packet-switched IP Networks – A Conceptual Framework

Stephan Kubisch, Harald Widiger,
Peter Danielis, Jens Schulz and Dirk Timmermann
University of Rostock
Institute of Applied Microelectronics
and Computer Engineering
18051 Rostock, Germany
Tel./Fax: +49 (381) 498-7276 / -1187251
{stephan.kubisch;harald.widiger}@uni-rostock.de

Thomas Bahls and Daniel Duchow
Nokia Siemens Networks
Broadband Access Division
17489 Greifswald, Germany
Tel./Fax: +49 (3834) 555-642 / -602
{thomas.bahls;daniel.duchow}@nsn.com

During the last years, the Internet has grown into a mass-medium for communication and information exchange. Millions of people are using the Internet for business and in social life. Users can be reached easily and cost-effectively. Unfortunately the Internet's open structure is the reason for its frequent misuse for illegal and criminal actions such as dissembling phishing attacks. Thus, anti-phishing techniques are needed to recognize potential phishing threats. But mostly these techniques are only of reactive nature, are soon circumvented by expert frauds, or are not efficient enough. This paper describes an anti-phishing framework. A concept for trust management and a mechanism called IPclip are presented. The main idea of IPclip is to guarantee trust-by-wire in packet-switched networks by providing trustworthy location information along with every IP packet. This information is used as supplementary and trustworthy trigger to identify potential phishing threats. Besides, the proposed framework allows for tracing the threat's origin by using a set of location information.

The Development of Novel Switching Devices by Using Embedded Microprocessing System Running Linux

F. Yan and K. H. Yeung
City University of Hong Kong

This paper discusses the development of novel switching devices using embedded microporcessing systems. It first reviews on basic Ethernet Switch operations and the Spanning tree protocol (STP). It then gives a brief analysis on Linux STP implementation and bridge configuration tools. The aim of the analysis is to lay a foundation for latter discussion on how novel switching devices can be developed based on the bridging codes. To facilitate the discussion, a new kind of Ethernet switches (with modified STP operations) proposed by the authors of this paper is used as an example. This practical example demonstrates how kernel programming on Linux and some modifications on configuration tools can be made to develop new switching devices. Experiments on the newly developed switches are also reported. The results show that these new switches can provide better security for STP networks. At the end of the paper, discussion on how to port the design to embedded miscroprocessing systems is also given.

What is the Best Way to Prove a Cryptographic Protocol Correct?

Sreekanth Malladi
College of Business and Information Systems
Dakota State University
Madison, SD - 57042
Sreekanth.Malladi@dsu.edu

Gurdeep S. Hura
Dept of Math and Computer Science
University of Maryland-Eastern Shore
Princess Anne, MD - 21853
gshura@umes.edu

In this paper, we identify that protocol verification using invariants have significant limitations such as inapplicability to some protocols, non-standard attacker inferences and non-free term algebras. We argue that constraint solving for bounded process analysis can be used in conjunction with decidability of context-explicit protocols as a verification tool and can overcome those limitations. However, this is possible only when new decidability results are obtained for protocol security, especially in presence of non-standard inferences and non-free term algebras. Those results are under progress and will be soon submitted for publication.

Internal Forensic Acquisition for Mobile Equipments

Gianluigi Me
Dipartimento di Informatica, Sistemi e Produzione, Università di Roma "Tor Vergata"
me@disp.uniroma2.it
Maurizio Rossi

Several technologies and products have emerged in the market lately for Personal Digital Assistants and smartphone data forensic acquisition. The smartphone market is very huge and provides a great variety of manufacturers and models causing a strong heterogeneity of the tools adopted to retrieve smartphone contents in a forensically sound way: in fact, in most cases, the mobile devices manufacturers implement their own (proprietary) protocols on the proprietary cable-jack and the proprietary OSs, causing the forensic operators to be overwhelmed by the one-on-one tools for every single mobile device. This paper aims to propose a new methodology and a tool to acquire the data by using the removable memory cards (e.g. SD, mini SD, MMC etc). This approach could represent a breakthrough in the mobile forensics, since the developed tool could replace the plethora of the hardware tools currently used. In this paper, firstly, we will summarize the current seizure methodology and its related problems when applied to the mobile device scenario. Then, we will introduce an alternative methodology to seize and examine the data from internal memory, overcoming some problems of the traditional methodology. At the end, we will show some experimental results of this methodology, obtained from a real device.

Network Traffic Behavior Analysis by Decomposition into Control and Data Planes

Basil AsSadhan, Hyong Kim, José M. F. Moura and Xiaohui Wang
Carnegie Mellon University
Electrical and Computer Engineering Department
5000 Forbes Ave, Pittsburgh, PA, USA
bsadhan, kim, moura, xiaohuiw@ece.cmu.edu

In this paper, we analyze network traffic behavior by decomposing header traffic into control and data planes to study the relationship between the two planes. By computing the cross-correlation between the control and data traffics, we observe a general 'similar' behavior between the two planes during normal behavior, and that this similarity is affected during abnormal behaviors. This allows us to focus on abnormal changes in network traffic behavior. We test our approach on the Network Intrusion Dataset provided by the Information Exploration Shootout (IES) project and the 1999 DARPA Intrusion detection Evaluation Dataset from the MIT Lincoln Lab. We find that TCP control and data traffic have high correlation levels during benign normal applications. This correlation is reduced when attacks that affect the aggregate traffic are present in the two datasets.

Decentralized Mediation Security

Richard Whittaker, Gonzalo Argote-Garcia and Peter J. Clarke
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
email:{rwhitt01, gargo001, clarkep}@cis.fiu.edu

Raimund K. Ege
Department of Computer Science
Northern Illinois University
DeKalb, IL 60115 USA
email:ege@niu.edu

Information systems that use the mediation architecture are well suited to accessing heterogeneous data in a distributed environment. These mediation systems utilize a central security mediator that is primarily concerned with securing the internal structure of the mediation system. The current problem with this approach is that clients are unable to have authority and administer the security of their data during the collaboration with the mediation system. This creates a black box scenario from the perspective of the client towards the mediation system, which can have great consequences in transactions that deal with sensitive data within an untrusted environment. In this paper, we present a solution in the form of a decentralized mediator that provides all parties involved in a transaction within the mediation system autonomy over their security that is security model independent and dynamic at runtime. Our approach allows the client to manage the access to data during a transaction with the mediation system.

Defining a Simple Metric for Real-time Security Level Evaluation of Multi-sites Networks

A. K. Ganame and J. Bourgeois
LIFC, University of Franche-Comte
Montbeliard, France
{Julien.Bourgeois,Abdoul.Ganame}@univ-fcomte.fr
http://rma.pu-pm.univ-fcomte.fr

In previous research work, we have developed a centralized Security Operation Center (SOC) and a distributed SOC. These environments are very useful to react to intrusions or to analyze security problem because they provide a global view of the network without adding any kinds of software on network components. They therefore lack the possibility to have a real-time metric which measures the security health of the different sites. The idea is to have, in one look, an indication of the security level of all the sites of the network. In this article, we propose to define such a metric which gives the user 3 states for a given network.

Overfort: Combating DDoS with Peer-to-Peer DDoS Puzzle

Soon Hin Khor
Interdisciplinary Information Studies
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033 Japan
khor.soon.hin@iii.u-tokyo.ac.jp

Akihiro Nakao
Interdisciplinary Information Studies
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033 Japan
nakao@iii.u-tokyo.ac.jp

The Internet community has been long convinced that Distributed Denial-of-Service (DDoS) attacks are difficult to combat since IP spoofing prevents traceback to the sources of attacks. Even if traceback is possible, the sheer number of sources that must be shutdown renders traceback, by itself, ineffective. Due to this belief, much effort has been focused on winning the "arms race" against DDoS by over-provisioning resources. This paper shows how Overfort can possibly withstand DDoS onslaughts without being drawn into an arms race by using higher-level traceback to DDoS agents' local DNSes (LDNSes) and dealing with those LDNSes instead. Overfort constructs an on-demand overlay using multiple overlay-ingress gateways with their links partitioned into many virtual linksłeach with different bandwidth and IPłleading to the server to project the illusion of multiple server IPs. An attacker will be faced with the daunting puzzle of finding all the IPs and thereafter the confusion of how much traffic to clog each IP with. Furthermore, Overfort has a mechanism to segregate LDNSes that are serving DDoS agents and restrict them to a limited number of IPs thus saving the other available IPs for productive use. Both proliferation of access channels to the server and LDNS segregation mechanism are the key components in Overfort to defend against DDoS with significantly less resources.

The Deterministic Protocol for Rational Secret Sharing

Maleka S
Indian Institute of Technology Madras
maleka.smile@gmail.com

Amjed Shareef
Indian Institute of Technology Madras
amjedshareef@gmail.com

C. Pandu Rangan
Indian Institute of Technology Madras
rangan@cs.iitm.ernet.in

We consider the rational secret sharing problem introduced by Halpern and Teague, where players prefer to get the secret than not to get the secret and with lower preference, prefer that as few of the other players get the secret. The impossibility of a deterministic protocol for rational secret sharing is proved by Halpern and Teague. The impossibility result is based on the fact that a rational player always chooses a dominating strategy and so there is no incentive for a player to send his secret share. This rational behavior makes secret sharing impossible, but there is an interesting way by which we can force rational players to cooperate for achieving successful secret sharing. A rational player may be deterred from exploiting his short term advantage by the threat of punishment that reduces his long term payoff. This can be captured by the repeated interaction of players. Hence, we study rational secret sharing in a scenario, where players interact repeatedly in several rounds which enables the possibility of secret sharing among rational players. In our model, the dealer, instead of sending shares, forms polynomials of the secret shares and sends points on that polynomial (say subshares) to the players. The dealer constructs polynomials in a manner that the degrees of polynomials used differ by at most one and each player is not aware of the degree of polynomial employed for others. The players distribute shares in terms of subshares. We show a surprising result on the deterministic protocol for rational secret sharing problem in synchronous model. This is the first protocol that achieves rational secret sharing in a reasonable model to the best of our knowledge.

A Remote Anonymous Attestation Protocol in Trusted Computing

Jiqiang Liu, Jia Zhao and Zhen Han
Computer and Information Technology Department, Beijing Jiaotong University
Beijing 100044, P R China
Email: {jqliu,04112070,zhan}@bjtu.edu.cn

Remote attestation is an important attribute in trusted computing. One of the purpose of remote attestation is to attest the remote platform is trusty but not revealing the actual identity of the platform. Direct Anonymous Attestation(DAA) is a kind of scheme which is adopted by Trusted Computing Group in the specification 1.2 to hide the privacy of the platform. But DAA involves various of zero-knowledge proofs and is not efficient to implement. To guarantee the trustworthiness and privacy, we propose a remote anonymous attestation protocol based on ring signature in this paper. We also show that our protocol is secure under the RSA assumption in random oracle model. Furthermore, the attestation protocol does not need the third party and extra zero-knowledge proof, which makes it very efficient in realization.

Workshop 18

International Workshop on Hot Topics in Peer-to-Peer Systems

HOTP2P 2008

The Design of a Versatile, Secure P2PSIP Communications Architecture for the Public Internet

David A. Bryan and Bruce B. Lowekamp
College of William and Mary / SIPeerior Technologies, Inc.
Williamsburg, VA, USA
{bryan, lowekamp}@cs.wm.edu

Marcia Zangrilli
SIPeerior Technologies, Inc.
Williamsburg, VA, USA
marcia@sipeerior.com

Communications systems, encompassing VoIP, IM, and other personal media, present different challenges for P2P environments than other P2P applications. In particular, reliable communication implies that each resource (person) is unique and must be reliably located, without false negatives. Because of their prevalence in real deployments, the overlay must use endpoints behind NATs as peers and must be resilient against DoS attacks that attempt to disrupt the system's routing properties or DoS a particular person. We have designed and implemented a P2P communications system that addresses these issues, now deployed as both a commercial and academic project, which has resulted in a leading proposal for a P2PSIP standard in the IETF. We present the design tradeoffs necessary to meet the requirements of a reliable communications system and provide guidance on appropriate choices for designers of other similar systems in the future. In particular, the practical issues of non-transitive routing, NAT traversal required by our endpoints, and the prevention of DoS attacks have proven to be more critical than strict performance metrics in selecting DHT identifiers, topology, and routing algorithms. Where a central authority exists, certificates can be stored in the overlay and allow more efficient DHT algorithms to be used. We explain how security and routing schemes can help preserve the integrity, scalability, and performance of P2PSIP communication Systems.

Multi-Source Video Multicast in Peer-to-Peer Networks

Francisco de Asís Lóez-Fuentes and Eckehard Steinbach
Technische Universität München
Institute of Communication Networks, Media Technology Group
80333 München, Germany
{fcoasis, Eckehard.Steinbach}@tum.de

We propose a novel framework for multi-source video streaming in peer-to-peer (P2P) networks. Multiple videos are distributed to all requesting peers exploiting full collaboration between the sources, the requesting peers and helper peers. Each source distributes its own video sequence while additionally forwarding video data received from other sources. A single peer is selected to redistribute a particular video block to the peers which would like to receive the videos. Our goal is to maximize the overall throughput or alternatively the aggregate video quality of multiple concurrent streaming sessions. We also consider the special cases of "same throughput" or "same video quality" for all streams. We formulate the rate allocation and redistribution as an optimization problem and evaluate our framework for three different scenarios. In the first scenario, the rate allocation is jointly decided for all participating peers. In the second scenario, the rate allocation is also decided jointly, but additionally either same rate or same video quality streams are enforced. Our third scenario assumes separate distribution for every source. In this case, the peers divide their upload capacity equally among the different video sequences. Our results show the superior performance of joint rate allocation compared to independent allocation and the effectiveness of our framework.

Towards Feasibility and Scalability of Text Search in Peer-to-Peer Systems

Akshay Lal, Varun Gupta, Khaled Harfoush and Injong Rhee
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
Email: [alal, vgupta, harfoush, rhee]@cs.ncsu.edu

In this paper, we introduce a search engine, Dgoogle, designed for large scale P2P systems. Dgoogle is purely textbased, does not organize documents based on pre-defined keywords or based on their semantics. It is simple to implement and can tolerate variations in the wording of text queries. Compared to existing proposals, such as Inverted Indices, Dgoogle does not stress network bandwidth and offers an order of magnitude of savings in storage overhead and in the response time to user queries. Furthermore, Dgoogle's performance is not affected by long queries or by processing popular query words. Simulation results validate the efficacy of our proposal.

Replication in Bandwidth-Symmetric Bittorrent Networks

M. Meulpolder, D.H.J. Epema, H.J. Sips
Parallel and Distributed Systems Group
Department of Computer Science, Delft University of Technology, the Netherlands
{M.Meulpolder,D.H.J.Epema,H.J.Sips}@tudelft.nl

The popular and well-known BitTorrent peer-to-peer protocol offers fast file distribution in a highly scalable way. Several studies have investigated the properties of this protocol, mostly focusing on heterogeneous end-user environments such as the Internet, with asymmetric connections. In this paper however, we focus on the usage of the BitTorrent protocol in homogeneous local environments with symmetric bandwidth properties. Compared with a traditional client-server setup, the use of BitTorrent in such settings can offer huge benefits in performance and scalability, allowing bandwidth sharing and high speed file distribution. We aim to improve the performance of such networks with a novel mechanism for replication using so-called replicators, which replicate a subset of the files in the system. A mathematical model of the resulting Replicated BitTorrent is presented and validated by emulation. Furthermore, we present simulation results that provide insight in the performance of Replicated BitTorrent networks with dynamic peer arrivals and departures. The results show that Replicated BitTorrent significantly improves download times in local bandwidth-symmetric BitTorrent networks.

Game Theoretical Analysis of Incentives for Large-scale, Fully Decentralized Collaboration Networks

Thomas Bocek[1], Michael Shann[1], David Hausheer[1] and Burkhard Stiller[2]
[1]Communication Systems Group
Department of Informatics IFI, University of Zurich
Email:{bocek, hausheer, stiller}@ifi.uzh.ch, mikeshann@access.uzh.ch
[2]Computer Engineering and Networks Laboratory
ETH Zurich
Email: stiller@tik.ee.ethz.ch

One of the key challenges in peer-to-peer networks is the design of incentives to encourage peers to share their resources. Incentives are necessary in order to exploit the full potential of these systems. The tit-for-tat incentive scheme, as used in BitTorrent for example, has proven to be a successful approach in P2P file sharing systems, where peers have direct relations and share the same kind of resources. However, in P2P systems where different kind of resources are shared between peers with non-direct relations, the design of incentives remains a challenge. In this paper, a large-scale, fully decentralized P2P collaboration network is shown, where peers share not only bandwidth and storage space, but also contribute by editing articles and voting for or against changes. A new incentive scheme is proposed which supports non-direct relations and provides incentives for sharing different kind of resources. The incentive scheme is based on a reputation system that assigns a reputation value to every peer reflecting its previous behavior in the network. Based on this value, the service level is differentiated, i.e. the higher a peer's reputation the better the quality of service it can get from the network. The service differentiation has been analyzed and simulated with rational, irrational and altruistic peers based on game theory concepts.

Neighbor-of-neighbor Routing Over Deterministically Modulated Chord-like DHTs

Giovanni Chiola and Marina Ribaudo
DISI, Università di Genova
chiolag@acm.org, ribaudo@disi.unige.it

We define a DHT system whose aim is to combine the routing efficiency typical of randomized networks, i.e. average path length $O(\log n/\log\log n)$ with the ease of coding and the start-up efficiency of an optimized uniform system. Our proposed system is non-uniform, even though it is not explicitly randomized. In particular, the difference in terms of finger tables between two adjacent peers in a large system is very moderate, and predictable based on an approximation of the peers' keys. These characteristics allow us to adopt a (multi-hop) Neighbor of Neighbor routing algorithm, which asymptotically achieves $O(\log n/\log\log n)$ average path length, and that in practice may even slightly outperform randomized networks containing more than 100,000 peers.

On the Need for Query-centric Unstructured Peer-to-peer Overlays

William Acosta and Surendar Chandra
University of Notre Dame, Notre Dame, IN 46556, USA
{wacosta,surendar}@nd.edu

Hybrid P2P systems rely on the assumption that sufficient objects exist nearby in order to make the unstructured search component efficient. This availability depends on the object annotations as well as on the terms in the queries. Earlier work assumed that the object annotations and query terms follow Zipf-like long-tail distribution. We show that the queries in real systems exhibit more complex temporal behavior. To support our position, first we analyzed the names and annotations of objects that were stored in two popular P2P sharing systems; Gnutella and Apple iTunes. We showed that the names and annotations exhibited a Zipf like long tail distribution. The long tail meant that over 98% of the objects were insufficiently replicated (less than 0.1% of the peers). We also analyzed a query trace of the Gnutella network and identified the popularity distribution of the terms used in the queries. We showed that the set of popular query terms remained stable over time and exhibited a similarity of over 90%. We also showed that despite the Zipf popularity distributions of both query terms and file annotation terms, there was little similarity over time (< 20%) between popular file annotation terms and popular file terms. Prior P2P search performance analysis did not take this mismatch between the query terms and object annotations into account and thus overestimated the system performance. There is a need to develop unstructured P2P systems that are aware of the temporal mismatch of the object and query popularity distributions.

LiteLoad: Content Unaware Routing for Localizing P2P protocols

Shay Horovitz and Danny Dolev
School of Computer Science and Engineering
Hebrew University of Jerusalem Jerusalem, Israel
{horovitz,dolev}@cs.huji.ac.il

In today's extensive worldwide Internet traffic, some 60% of network congestion is caused by Peer to Peer sessions. Consequently ISPs are facing many challenges like: paying for the added traffic requirement, poor customer satisfaction due to degraded broadband experience, purchasing costly backbone links and upstream bandwidth and having difficulty to effectively control P2P traffic with conventional devices.

Existing solutions such as caching and indexing of P2P content are controversial as their legality is uncertain due to copyright violation, and therefore hardly being installed by ISPs. In addition these solutions are not capable to handle existing encrypted protocols that are on the rise in popular P2P networks.

Other solutions that employ traffic shaping and blocking degrade the downloading throughput and cause end users to switch ISPs for a better service.

LiteLoad discerns patterns of user communications in Peer to Peer file sharing networks without identifying the content being requested or transferred and uses least-cost routing rules to push peer-to-peer transfers into confined network segments. This approach maintains the performance of file transfer as opposed to traffic shaping solutions and precludes internet provider involvement in caching, cataloguing or indexing of the shared content. Simulation results expresses the potential of the solution and a proof of concept of the key technology is demonstrated on popular protocols, including encrypted ones.

Absolute Slicing in Peer-to-peer Systems

Alberto Montresor and Roberto Zandonati
University of Trento, Italy
alberto.montresor@unitn.it, r.zandonati@studenti.unitn.it

Peer-to-peer (P2P) systems are slowly moving from application-specific architectures to a generic service-oriented design framework. The idea is to allow a dynamic collection of P2P applications to cohabit into a single system, with applications starting and terminating at will, or even changing their requirements at run-time. This raises an interesting problem in connection with managing resource assignment in a large-scale, heterogeneous and unreliable environment. Recently, the distributed slicing service has been proposed to allow for an automatic partitioning of P2P networks into groups (slices) that represent a controllable amount of some resource. A particular instantiation of such service has been described, called ordered slicing, in which nodes are ranked based on some metrics and then assigned to a slice based on their position in the ranking. In this paper, we present an alternative version of the problem called absolute slicing. Here, the goal is to assign a specified number of nodes to a slice and maintain such assignment in spite of churn. We propose a simple algorithm that solves the problem by combining well-known protocols such as peer sampling and aggregation, and we experimentally evaluate its performance.

Efficiency of Tree-structured Peer-to-peer Service Discovery Systems

Eddy Caron[1,2,4], Frédéric Desprez[1,3,4], Cédric Tedeschi[1,2,4]
[1] University of Lyon [2] ENS Lyon [3] INRIA
[4] LIP UMR 5668 CNRS ENS Lyon INRIA UCBL, France.

The efficiency of service discovery is a crucial point in the development of fully decentralized middlewares intended to manage large scale computational grids. The work conducted on this issue led to the design of many peerto- peer fashioned approaches. More specifically, the need for flexibility and complexity in the service discovery has seen the emergence of a new kind of overlays, based on tries, also known as lexicographic trees.

Although these overlays are efficient and well designed, they require a costly maintenance and do not accurately take into account the heterogeneity of nodes and the changing popularity of the services requested by users.

In this paper, we focus on reducing the cost of the maintenance of a particular architecture, based on a dynamic prefix tree, while enhancing it with some load balancing techniques that dynamically adapt the load of the nodes in order to maximize the throughput of the system. The algorithms developed couple a self-organizing prefix tree overlay with load balancing techniques inspired by similar previous works undertaken for distributed hash tables.

After some simulation results showing how our load balancing heuristics perform in such an overlay and compare to other heuristics, we provide a fair comparison of this architecture and similar overlays recently proposed.

A Framework for Peer-to-Peer Service Interaction

Jun Li
Department of Computer Science
ETH Zurich
Zurich, Switzerland
jli@ethz.ch

Although structured/unstructured peer-to-peer (P2P) algorithms proliferate in recent years, building P2P application is still complex and time-consuming due to inadequate infrastructure support. On the other hand, service-oriented architecture significantly reduces developing complexity by further decoupling dependent components. Based on the idea of using services as fundamental elements for developing P2P applications, we propose a P2P service framework to facilitate P2P application development. The framework chooses eQuus as the DHT implementation for the reason of robustness and integrates with OSGi to maximize its usability. Group communication helps the framework to achieve replication with almost free cost. Pull and push based service discovery mechanism enables applications unaware of the underlying remote distribution. Through relaying mechanism, our framework is able to traverse NATs. Based on our framework, we also implemented QChat - a pure P2P chat application.

Workshop 19

Workshop on Large-Scale, Volatile Desktop Grids

PCGRID 2008

Optimizing the Data Distribution Layer of BOINC with BitTorrent

Fernando Costa and Luis Silva
Departamento Engenharia Informatica,
Universidade de Coimbra,
Polo II, 3030-Coimbra, Portugal
{flcosta,luis}@dei.uc.pt

Gilles Fedak
Laboratoire de Recherche en
Informatique/INRIA Futurs,
France
fedak@lri.fr

Ian Kelley
School of Computer Science,
Cardiff University,
United Kingdom
I.R.Kelley@cs.cardiff.ac.uk

In this paper we show how we applied BitTorrent data distribution techniques to the BOINC middleware. Our goal was to decentralize BOINC's data model to take advantage of client network capabilities. To achieve this, we developed a prototype that adds BitTorrent functionality for task distribution and conducted smallscale tests of the environment. Additionally, we measured the impact of the BitTorrent components in both the BOINC client and server, and compared it with the original implementation. Our preliminary results indicate that the BitTorrent client had a negligible influence on the BOINC client's computation time, even in the case where it was seeding extensively. The BOINC server, on the contrary, showed an unexpectedly low bandwidth output when seeding the file, as well as spikes on CPU usage. Current results show the BitTorrent scenario allows clients to share the burden of data distribution on BOINC with almost no negative influence on compute time. This paper will discuss the tests that were performed, how they were evaluated, as well as some improvements that could be made to future tests to enhance server-side efficiency.

Efficient Content Search in iShare, a P2P based Internet-Sharing System

Seyong Lee, Xiaojuan Ren and Rudolf Eigenmann
School of ECE, Purdue University
West Lafayette, IN, 47907
{lee222,renx,eigenman}@purdue.edu

This paper presents an efficient content search system, which is applied to iShare, a distributed peer-to-peer(P2P) Internet-sharing system. iShare facilitates the sharing of diverse resources located in different administrative domains over the Internet. For efficient resource management, iShare organizes resources into a hierarchical name space, which is distributed over the underlying structured P2P network. However, iShare's search capability has a fundamental limit inherited from the underlying structured P2P system's search capability. Most existing structured P2P systems do not support content searches. There exists some research that provides content search functionality, but the approaches do not scale well and incur substantial overheads on data updates. To address these issues, we propose an efficient hierarchical-summary system, which enables an efficient content search and semantic ranking capability over traditional structured P2P systems. Our system uses a hierarchical name space to implement a summary hierarchy on top of existing structured P2P overlay networks, and uses a Bloom Filter as a summary structure to reduce space and maintenance overhead. We implemented the proposed system in iShare, and the results show that our search system finds all relevant results regardless of summary scale and the search latency increases very slowly as the network grows.

Probabilistic Allocation of Tasks on Desktop Grids

Joshua Wingstrom and Henri Casanova
Information and Computer Sciences Dept.
University of Hawai'i at Manoa, Honolulu, U.S.A.

While desktop grids are attractive platforms for executing parallel applications, their volatile nature has often limited their use to so-called "high-throughput" applications. Checkpointing techniques can enable a broader class of applications. Unfortunately, a volatile host can delay the entire execution for a long period of time. Allocating redundant copies of each task to hosts can alleviate this problem by increasing the likelihood that at least one instance of each application task completes successfully. In this paper we demonstrate that it is possible to use statistical characterizations of host availability to make sound task replication decisions. We find that strategies that exploit such statistical characterizations are effective when compared to alternate approaches. We show that this result holds for real-world host availability data, in spite of only imperfect statistical characterizations.

Computing Low Latency Batches with UnreliableWorkers in Volunteer Computing Environments

Eric M. Heien, Noriyuki Fujimoto, Kenichi Hagihara
Graduate School of Information Science and Technology, Osaka University
Toyonaka, Osaka 560-8531, Japan
{e-heien, fujimoto, hagihara}@ist.osaka-u.ac.jp

Internet based volunteer computing projects such as SETI@home are currently restricted to performing coarse grained, embarrassingly parallel tasks. This is partly due to the "pull" nature of task distribution in volunteer computing environments, where workers request tasks from the master rather than the master assigning tasks to arbitrary workers. In this paper we develop algorithms for computing batches of medium grained tasks with soft deadlines in pullstyle volunteer computing environments. Using assumptions about worker availability intervals based on previous studies, we develop models of unreliable workers in volunteer computing environments. These models are used to develop algorithms for task distribution in volunteer computing systems with a high probability of meeting batch deadlines. We develop algorithms for perfectly reliable workers, computation-reliable workers and unreliable workers. The effectiveness of the algorithms is demonstrated by using traces from actual execution environments.

Scheduling Algorithms for Multiple Bag-of-Task Applications on Desktop Grids: A Knowledge-Free Approach

Cosimo Anglano, Massimo Canonico
Dipartimento di Informatica, Università del Piemonte Orientale (Italy),
email:{cosimo.anglano,massimo.canonico}@unipmn.it

Desktop Grids are being increasingly used as the execution platform for a variety of applications that can be structured as Bag-of-Tasks (BoT). Scheduling BoT applications on Desktop Grids has thus attracted the attention of the scientific community, and various schedulers tailored towards them have been proposed in the literature. However, previous work has focused on scheduling a single BoT application at a time, thus neglecting other scenarios in which several users submit multiple BoT applications at the same time. This paper aims at filling this gap by proposing a set of scheduling algorithms able to deal with multiple BoT applications. The performance of these algorithm has been evaluated, by means of simulation, for a large set of operational scenarios obtained by varying both the workload submitted to the Desktop Grid and the characteristics of the involved resources. Our results show that, although there is no a clear winner among the proposed solutions, knowledge-free strategies (that is, strategies that do not require any information concerning the applications or the resources) can provide good performance.

Increasing the Amount of Work Completed by Volunteer Computing Projects with Task Distribution Policies

David Toth and David Finkel
Worcester Polytechnic Institute
Computer Science Department
Worcester, MA 01609
USA
toth@cs.wpi.edu, dfinkel@cs.wpi.edu

Volunteer computing projects rely on volunteers running clients on their computers that contribute to projects when the computers' owners allow them to. These projects allow people to solve problems that were previously too computationally intensive to solve. However, due to the relatively small fraction of the population that participates in volunteer computing projects, it's very important to use the donated CPU cycles as efficiently as possible. Volunteer computing clients use two different methods to retrieve tasks: retrieving one task at a time when the client has no more work to do or retrieving multiple tasks at once and storing them in a buffer. We simulate these different task retrieval policies to compare the number of tasks completed by clients using the different policies. Our simulations showed that clients that retrieve one task at a time complete more tasks than clients that retrieve multiple tasks at once and buffer them. Our simulations also showed that there was not a significant gain in the amount of work that could be completed by devising a more complicated adaptive policy.

Defeating Colluding Nodes in Desktop Grid Computing Platforms

Gheorghe Cosmin Silaghi
Dept. of Business Information Systems
Babes.-Bolyai University, Cluj, Romania
gsilaghi@econ.ubbcluj.ro

Filipe Araujo and Luis Moura Silva
CISUC, Dept. of Informatics Engineering
University of Coimbra, Portugal
{filipius, luis}@dei.uc.pt

Patricio Domingues
School of Technology and Management
Polytechnic Institute of Leiria, Portugal
patricio@estg.ipleiria.pt

Alvaro E. Arenas
Rutherford Appleton Laboratory
STFC, Didcot, UK
a.e.arenas@rl.ac.uk

Desktop Grid systems reached a preeminent place among the most powerful computing platforms in the planet. Unfortunately, they are extremely vulnerable to mischief, because volunteers can output bad results, for reasons ranging from faulty hardware (like over-clocked CPUs) to intentional sabotage. To mitigate this problem, Desktop Grid projects replicate work units and apply majority voting, typically on 2 or 3 results. In this paper, we observe that this form of replication is powerless against malicious volunteers that have the intention and the (simple) means to ruin the project using some form of collusion. We argue that each work unit needs at least 3 voters and that voting pools with conflicts enable the master to spot colluding malicious nodes. Hence, we postprocess the voting pools in two steps: i) we use a statistical approach to identify nodes that were not colluding, but submitted bad results; ii) we use a rather simple principle to go after malicious nodes which acted together: they might have won conflicting voting pools against nodes that were not identified in step i. We use simulation to show that our heuristic can be quite effective against colluding nodes, in scenarios where honest nodes form a majority.

Optimal Spot-checking to Minimize the Computation Time in Volunteer Computing

Kan Watanabe, Masaru Fukushi and Susumu Horiguchi
Graduate School of Information Sciences, Tohoku University
{can, mfukushi, susumu}@ecei.tohoku.ac.jp

This paper proposes an optimization technique for spot-checking to minimize the computation time of volunteer computing (VC) systems with malicious participants who return erroneous results. Credibility based voting with spot-checking is one of the most promising approaches to guaranteeing the computational correctness of VC systems. There is an optimal value for the spot-check rate; that is, the increase of the spot-check rate reduces the necessary redundancy for voting, which in turn increases the computation time due to the spot-checking itself. From this tradeoff, the estimation of the optimal spot-check rate is therefore the major concern to minimize the computation time required for the credibility based voting. The key idea is to represent the mathematical expectation of the computation time as a function of the spot-check rate to obtain an estimate of the optimal spot-check rate before the computation. It is shown by Monte Carlo simulations that the proposed technique can always obtain an approximate estimate of the optimal spot-check rate and minimize the computation time with an uncertainty of 1%.

Providing Security to the Desktop Data Grid

Jesus Luna, Michail Flouris, Manolis Marazakis and Angelos Bilas
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH)
PO Box 1385. GR-71110. Heraklion, Greece.
{jluna, flouris, maraz, bilas}@ics.forth.gr

Volunteer Computing is becoming a new paradigm not only for the Computational Grid, but also for institutions using production-level Data Grids because of the enormous storage potential that may be achieved at a low cost by using commodity hardware within their own computing premises. However, this novel "Desktop Data Grid" depends on a set of widely distributed and untrusted storage nodes, therefore offering no guarantees about neither availability nor protection to the stored data. These security challenges must be carefully managed before fully deploying Desktop Data Grids in sensitive environments (such as eHealth) to cope with a broad range of storage needs, including backup and caching.

In this paper we propose a cryptographic protocol able to fulfil the storage security requirements related with a generic Desktop Data Grid scenario, which were identified after applying an analysis framework extended from our previous research on the Data Grid's storage services. The proposed protocol uses three basic mechanisms to accomplish its goal: (a) symmetric cryptography and hashing, (b) an Information Dispersal Algorithm and the novel (c) "Quality of Security" (QoSec) quantitative metric. Although the focus of this work is the associated protocol, we also present an early evaluation using an analytical model. Our results show a strong relationship between the assurance of the data at rest, the QoSec of the Volunteer Storage Client and the number of fragments required to rebuild the original file.

Implementing an Industrial-Strength Academic Cyberinfrastructure at Purdue University

Preston M. Smith [1,2], Thomas J. Hacker [2,1,3] and Carol X. Song [1]
[1]Rosen Center for Advanced Computing,
[2]Computer & Information Technology, College of Technology
[3]Discovery Park Cyber Center
Purdue University, West Lafayette, IN

Purdue University operates one of the largest cycle recovery systems in existence in academia based on the Condor workload management system. This system represents a valuable and useful cyberinfrastructure (CI) resource supporting research and education for campus and national users. During the construction and operation of this CI, we encountered many unforeseen challenges and benefits unique to an actively used infrastructure of this size. The most significant problems were integrating Condor with existing campus HPC resources, managing resource and user growth, coping with the distributed ownership of compute resources around campus, and integrating this CI with the Tera-Grid and Open Science Grid. In this paper, we describe some of our experiences and establish some best practices, which we believe will be valuable and useful to other academic institutions seeking to operate a production campus cyberinfrastrucure of a similar scale and utility.

Simplifying Resource Sharing in Voluntary Grid Computing with the Grid Appliance

David Isaac Wolinsky and Renato J. Figueiredo
Advanced Computing and Information Systems Laboratory
University of Florida
davidiw@ufl.edu, renato@acis.ufl.edu

Research projects in many fields are increasingly reliant on the use of computer-based simulation and computing grids. Many projects have successfully leveraged voluntary computing infrastructures by developing and distributing "@home" applications using the BOINC framework. Through generous contributions from the general public, these systems now have a computing backbone on which to have their data processed or simulations run. A shortcoming of such systems is that most users are often limited to contributing resources and few users are capable of developing or porting their own applications in order to use these resources. While many users are satisfied with receiving points (an intangible good) in return for their contribution, the need to port applications presents a barrier to entry to many other users who can potentially benefit from using the voluntary resources. In this paper, we describe enhancements made to the "Grid Appliance", a virtual machine based system which enables an execution environment in which users are given the opportunity to voluntarily share (providing and using) resources and run unmodified x86/Linux applications. Voluntary grids introduce a host of issues to tackle, most importantly getting users involved quickly. With that in mind, the Grid Appliance provides many tools for making a user-friendly environment for users, developers, and administrators. This paper summarizes the challenges of getting users involved, reducing the overhead for administrators, and describes the solutions used in the Grid Appliance. Index TermsłVoluntary Grid Computing, Virtual Computing, Virtual Machines, Virtual Networking.

Provisioning of Virtual Environments forWide Area Desktop Grids through Redirect-on-write Distributed File System

Vineet Chadha, David Wolinsky and Renato J. Figueiredo
Advanced Computing and information Systems Laboratory
University of Florida, Gainesville, FL
chadha,davidiw,renato@acis.ufl.edu

We describe and evaluate a thin client solution for desktop grid computing based on virtual machine appliances whose images are fetched on-demand and on a per-block basis over wide-area networks. The approach uses a distributed file system redirection mechanism which enables the use of unmodified NFS clients/servers and local buffering of file system modifications during the appliance's lifetime. The file system redirection technique is achieved through user-level proxies, and can be integrated with virtual private network overlays to provide transparent access to image servers even if they are behind firewalls. We have implemented and evaluated a prototype system which allows thin client diskless appliances to boot over a proxy VM bringing on-demand only a small fraction of the appliance image (16MB out of 900MB) and showing low runtime overhead for CPU-intensive applications. The paper also presents decentralized mechanisms to support seamless image version upgrades.

Centralized BOINC Resources Manager for Institutional Networks

Daniel Lombraña González
University of Extremadura, Spain
daniellg@unex.es

Francisco Fernández de Vega
University of Extremadura, Spain
fcofdez@unex.es

Germán Galeano Gil
Junta de Extremadura, Spain
ggaleano@edu.juntaextremadura.net

Ben Segal
CERN, Switzerland
B.Segal@cern.ch

This paper describes a new BOINC based approach to manage desktop grid environments belonging to institutions such as Universities or Companies. While the BOINC model delegates to users the decision of selecting the projects to which their own resources will collaborate with, we consider here the case of institutions which owns the computers provided to their staff. Our model removes users from the projects selection step and transfers this responsibility to computers' owner: the institution itself. Our solution describes a new model for handling BOINC enabled computers in a centralized way, while also provides a new software management tool. This tool allows an institution to remotely manage all its BOINC enabled resources from a centralized point of view. We have set up and tested the model in a modern University, with distributed computer labs on three different cities. We show how the resources were remotely assigned to different projects selected by the institution. Finally, We describe results supporting the effectiveness of the approach.

Workshop 20

Workshop on Multi-Threaded Architectures and Applications

MTAAP 2008

Architecture Optimizations for Synchronization and Communication on Chip Multiprocessors

Sevin Fide and Stephen Jenks
Department of Electrical Engineering and Computer Science
University of California, Irvine, USA
fsevin.fide, stephen.jenksg@uci.edu

Chip multiprocessors (CMPs) enable concurrent execution of multiple threads using several cores on a die. Current CMPs behave much like symmetric multiprocessors and do not take advantage of the proximity between cores to improve synchronization and communication between concurrent threads. Thread synchronization and communication instead use memory/cache interactions. We propose two architectural enhancements to support fine grain synchronization and communication between threads that reduce overhead and memory/cache contention. Register-Based Synchronization exploits the proximity between cores to provide low-latency shared registers for synchronization. This approach can save significant power over spin waiting when blocking events that suspend the core are used. Prepushing provides software controlled data forwarding between caches to reduce coherence traffic and improve cache latency and hit rates. We explore the behavior of these approaches, and evaluate their effectiveness at improving synchronization and communication performance on CMPs with private caches. Our simulation results show significant reduction in inter-core traffic, latencies, and miss rates.

Effect of Context Aware Scheduler on TLB

Satoshi Yamada and Shigeru Kusakabe
Grad. School of Information Sci. and Electrical Eng.
Kyushu University
744, Motooka, Nishi-ku, Fukuoka city, 819-0395, Japan
satoshi@ale.csce.kyushu-u.ac.jp
kusakabe@csce.kyushu-u.ac.jp

Multiprocessing and multithreading increase frequencies of thread schedulings and context switches. Context switches can decrease the system throughput because they disturb locality of references and cause cache misses such as Translation Lookaside Buffer (TLB), L1, and L2 misses. To reduce the overhead due to context switches, we focus on the advantage of threads which share the same memory address space (sibling threads). In switching sibling threads, the Operating System (OS) does not have to switch their memory address spaces. Thus, we can omit a part of switching procedure and expect an efficient use of memory hierarchy. To utilize this advantage, we have proposed a context aware scheduler, which aggregates sibling threads. By using our scheduler, we can expect to reduce the switches of memory address spaces in executing multithreaded applications. Because switching memory address spaces results in TLB flushes on some kinds of processors, our scheduler can be effective in reducing TLB misses on such platforms. In this paper, we evaluate the effect of our context aware scheduler on reducing TLB misses. For this purpose, we run several multithreaded applications simultaneously with our context aware scheduler and measure the number of TLB misses. From these experiments, we conclude that our context aware scheduler is effective in reducing TLB misses and contributes to enhance the system throughput without imposing the degradation of throughput on any application.

A Dissipative Resource Distribution Policy for SMT Processors

Hongzhou Chen, Lingdi Ping, Xuezeng Pan, Kuijun Lu and Xiaoping Chen
College of Computer Science and Technology, Zhejiang University, China
honjoychan@gmail.com

The performance in Simultaneous Multithreading (SMT) processors is significantly influenced by how optimally the common resources are shared among all threads. However, the programs running on threads often take on changing complicated behavior at runtime. Existing resource distribution methods either rely mainly on front-end fetching policies, or make the resource distribution decision according to the restricted information from the pipeline instead of the direct in-flight performance. It's difficult for them to catch various resource requirements efficiently and obtain potential gain. This paper presents a dissipative resource distribution policy (DRD) for SMT processors, in which the resource allocations evolve targeting the direct runtime performance. The self-organization mechanism established for DRD and the introduction of additional chaos cooperate together to form the fluctuation in the distribution evolvement procedure, and maintain persistent evolvement in changing program behavior. Simulation results show that DRD benefits much from its dissipative structure, and obtains potential fairness gains as well as pleasing throughput IPC increments. The meaning and setting of important DRD parameters are also discussed.

Early Experience with Out-of-Core Applications on the Cray XMT

Daniel Chavarría-Miranda[1], Andres Márquez[1], Jarek Nieplocha[2], Kristyn Maschhoff[4] and Chad Scherrer[3]
[1]High-Performance Computing
[2]Computational Sciences and Mathematics Division
[3]Computational Mathematics
Pacific Northwest National Laboratory
{daniel.chavarria, andres.marquez, jarek.nieplocha, chad.scherrer}@pnl.gov
[4]Cray, Inc. kristyn@cray.com

This paper describes our early experiences with a preproduction Cray XMT system that implements a scalable shared memory architecture with hardware support for multithreading. Unlike its predecessor, the Cray MTA-2 that had very limited I/O capability, the Cray XMT offers Lustre, a scalable high-performance parallel filesystem. Therefore it enables development of out-of-core applications that can deal with very large data sets that otherwise would not fit in the system main memory. Our application performs statistically-based anomaly detection for categorical data that can be used for analysis of Internet traffic data. Experimental results indicate that the preproduction version of the machine is able to achieve good performance and scalability for the in- and out-of-core versions of the application.

A Lock-free Multi-threaded Algorithm for the Maximum Flow Problem

Bo Hong
Drexel University, Philadelphia, PA 19104
bohong@coe.drexel.edu

The maximum flow problem is an important graph problem with a wide range of applications. In this paper, we present a lock-free multi-threaded algorithm for this problem. The algorithm is based on the push-relabel algorithm proposed by Goldberg. By using re-designed push and relabel operations, we derive our algorithm that finds the maximum flow with $O(|V|^2|E|)$ operations. We demonstrate that as long as a multi-processor architecture supports atomic 'read-update-write' operations, it will be able to execute the multi-threaded algorithm free of any lock usages. The proposed algorithm is expected to significantly improve the efficiency of solving maximum flow problem on parallel/multicore architectures.

Design of Scalable Dense Linear Algebra Libraries for Multithreaded Architectures: the LU Factorization

Gregorio Quintana-Ortí, Enrique S. Quintana-Ortí
Departamento de Ingeniería y Ciencia de Computadores
Universidad Jaume I
12.071–Castellón, Spain
{gquintan,quintana}@icc.uji.es
Ernie Chan, Robert A. van de Geijn, Field G. Van Zee
Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712
{echan,field,rvdg}@cs.utexas.edu

The scalable parallel implementation, targeting SMP and/or multicore architectures, of dense linear algebra libraries is analyzed. Using the LU factorization as a case study, it is shown that an algorithm-by-blocks exposes a higher degree of parallelism than traditional implementations based on multithreaded BLAS. The implementation of this algorithm using the SuperMatrix runtime system is discussed and the scalability of the solution is demonstrated on two different platforms with 16 processors.

Reducing the Run-time of MCMC Programs by Multithreading on SMP Architectures

Jonathan M. R. Byrd, Stephen A. Jarvis and Abhir H. Bhalerao
University of Warwick
Department of Computer Science
Coventry, CV4 7AL, UK
fJ.M.R.Byrd, Stephen.Jarvis, Abhir.Bhaleraog@dcs.warwick.ac.uk

The increasing availability of multi-core and multiprocessor architectures provides new opportunities for improving the performance of many computer simulations. Markov Chain Monte Carlo (MCMC) simulations are widely used for approximate counting problems, Bayesian inference and as a means for estimating very high-dimensional integrals. As such MCMC has found a wide variety of applications in fields including computational biology and physics, financial econometrics, machine learning and image processing. This paper presents a new method for reducing the runtime of Markov Chain Monte Carlo simulations by using SMP machines to speculatively perform iterations in parallel, reducing the runtime of MCMC programs whilst producing statistically identical results to conventional sequential implementations. We calculate the theoretical reduction in runtime that may be achieved using our technique under perfect conditions, and test and compare the method on a selection of multi-core and multi-processor architectures. Experiments are presented that show reductions in runtime of 35% using two cores and 55% using four cores.

Memory Model Effects on Application Performance for a Lightweight Multithreaded Architecture

Sheng Li[1], Shannon Kuntz[2], Peter Kogge[2] and Jay Brockman[2]
[1]Department of Electrical Engineering, [2]Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
{sli2, skuntz, kogge, jbb}@nd.edu

In this paper, we evaluate the effects of a partitioned global address space (PGAS) versus a flat, randomized distributed global address space (DGAS) in the context of a lightweight multithreaded parallel architecture. We also execute the benchmarks on the Cray MTA-2, a multithreaded architecture with a DGAS mapping. Key results demonstrate that distributing data under the PGAS mapping increases locality, effectively reducing the memory latency and the number of threads needed to achieve a given level of performance. In contrast, the DGAS mapping provides a simpler programming model by eliminating the need to distribute data and, assuming sufficient application parallelism, can achieve similar performance by leveraging large numbers of threads to hide the longer latencies.

Facilitating Efficient Synchronization of Asymmetric Threads on Hyper-Threaded Processors

Nikos Anastopoulos and Nectarios Koziris
National Technical University of Athens
School of Electrical and Computer Engineering
Computing Systems Laboratory
{anastop,nkoziris}@cslab.ece.ntua.gr

So far, the privileged instructions MONITOR and MWAIT introduced with Intel Prescott core, have been used mostly for inter-thread synchronization in operating systems code. In a hyper-threaded processor, these instructions offer a "performance-optimized" way for threads involved in synchronization events to wait on a condition. In this work, we explore the potential of using these instructions for synchronizing application threads that execute on hyper-threaded processors, and are characterized by workload asymmetry. Initially, we propose a framework through which one can use MONITOR/MWAIT to build condition wait and notification primitives, with minimal kernel involvement. Then, we evaluate the efficiency of these primitives in a bottom-up manner: at first, we quantify certain performance aspects of the primitives that reflect the execution model under consideration, such as resource consumption and responsiveness, and we compare them against other commonly used implementations. As a further step, we use our primitives to build synchronization barriers. Again, we examine the same performance issues as before, and using a pseudo-benchmark we evaluate the efficiency of our implementation for fine-grained inter-thread synchronization. In terms of throughput, our barriers yielded 12% better performance on average compared to Pthreads, and 26% compared to a spin-loops-based implementation, for varying levels of threads asymmetry. Finally, we test our barriers in a realworld scenario, and specifically, in applying thread-level Speculative Precomputation on four applications. For this multithreaded execution scheme, our implementation provided up to 7% better performance compared to Pthreads, and up to 40% compared to spin-loops-based barriers.

Qthreads: An API for Programming with Millions of Lightweight Threads

Kyle B. Wheeler	Richard C. Murphy	Douglas Thain
University of Notre Dame	Sandia National Laboratories	University of Notre Dame
South Bend, Indiana, USA	Albuquerque, New Mexico, USA	South Bend, Indiana, USA
kwheeler@cse.nd.edu	rcmurphy@sandia.gov	dthain@cse.nd.edu

Large scale hardware-supported multithreading, an attractive means of increasing computational power, benefits significantly from low per-thread costs. Hardware support for lightweight threads is a developing area of research. Each architecture with such support provides a unique interface, hindering development for them and comparisons between them. A portable abstraction that provides basic lightweight thread control and synchronization primitives is needed. Such an abstraction would assist in exploring both the architectural needs of large scale threading and the semantic power of existing languages. Managing thread resources is a problem that must be addressed if massive parallelism is to be popularized. The qthread abstraction enables development of large-scale multithreading applications on commodity architectures. This paper introduces the qthread API and its Unix implementation, discusses resource management, and presents performance results from the HPCCG benchmark.

Workshop 21

Workshop on Parallel and Distributed Computing in Finance

PDCoF 2008

Intrusion of Agent-Based Social Simulation in Economic Theory: Will Heterogeneous Agent Bring More Clarity in the Context of IT-Outsourcing?

Bogdan Werth and Scott Moss
Manchester Metropolitan University Business School, Centre for Policy Modelling, Aytoun Street,
Aytoun Building, M1 3GH Manchester, United Kingdom
{bogdan, scott}@cfpm.org

Past research on IT outsourcing has mainly focused on the transaction itself with the use of the Transaction Cost Theory as a primary investigation framework. The notion of asset specificity was blindly accepted in its primary definition, which embodies a bone of contention for many academics due to its looseness of definition. This paper introduces an alternative metric of asset specificity in a qualitative way. It goes down the line of Agent-Based Social Simulation and represents actors of the outsourcing process as heterogeneous agents. This approach is contrary to a widely held perception of economic standard literature with its assumption of homogeneous actors.

Parallel Option Pricing with Fourier Space Time-stepping Method on Graphics Processing Units

Vladimir Surkov
Department of Computer Science
University of Toronto, Toronto, Canada
vsurkov@cs.toronto.edu

With the evolution of Graphics Processing Units (GPUs) into powerful and cost-efficient computing architectures, their range of application has expanded tremendously, especially in the area of computational finance. Current research in the area, however, is limited in terms of options priced and complexity of stock price models. This paper presents algorithms, based on the Fourier Space Time-stepping (FST) method, for pricing single and multi-asset European and American options with Lévy underliers on a GPU. Furthermore, the single-asset pricing algorithm is parallelized to attain greater efficiency.

Multi-threaded Data Mining of EDGAR CIKs (Central Index Keys) from Ticker Symbols

Douglas A. Lyon
Chairman, Computer Engineering Department
Fairfield University
1073 North Benson Rd.
Fairfield, CT 06824
lyon@docjava.com

This paper describes how use the Java Swing HTMLEditorKit to perform multi-threaded web data mining on the EDGAR system (Electronic Data- Gathering, Analysis, and Retrieval system). EDGAR is the SEC's (U.S. Securities and Exchange Commission) means of automating the collection, validation, indexing, acceptance, and forwarding of submissions. Some entities are regulated by the SEC (e.g. publicly traded firms) and are required, by law, to file with the SEC.

Our focus is on making use of EDGAR to get information about company filings. These offers are filed with companies, using their Central Index Key (CIK). The CIK is used on the SEC's computer system to identify entities that filed a disclosure with the SEC. We show how to map a stock ticker symbol into a CIK.

The methodology for converting the web data source into internal data structures is based on using HTML as the input into a context-sensitive parser-callback facility. Screen scraping is a popular means of data mining, but the unstructured nature of HTML pages makes this a challenge.

The stop-and-wait nature of HTTP queries, as well as the non-deterministic nature of the response time, adversely impacts performance. We show that a combination of caching and multi-threading can improve performance by several orders of magnitude.

Large Scale Distribution of Stochastic Control Algorithms for Gas Storage Valuation

Constantinos Makassikis, Stéphane Vialle
SUPELEC, IMS research group, 2 rue Edouard Belin, 57070 Metz, France
LORIA, ALGORILLE project team, BP 239, 54506 Vandoeuvre-lès-Nancy, France
Constantinos.Makassikis@supelec.fr, Stephane.Vialle@supelec.fr
Xavier Warin
EDF - R&D, OSIRIS group, 1 Avenue Charles de Gaulle, 92141 Clamart, France
Xavier.Warin@edf.fr

This paper introduces the distribution of a stochastic control algorithm which is applied to gas storage valuation, and presents its experimental performances on two PC clusters and an IBM Blue Gene/L supercomputer. This research is part of a French national project which gathers people from the academic world (computer scientists, mathematicians, ...) as well as people from the industry of energy and finance in order to provide concrete answers on the use of computational clusters, grids and supercomputers applied to problems of financial mathematics.

The designed distribution allows to run gas storage valuation models which require considerable amounts of computational power and memory space while achieving both speedup and size-up: it has been successfully implemented and experimented on PC clusters (up to 144 processors) and on a Blue Gene supercomputer (up to 1024 processors). Finally, our distributed algorithm allows to use more computing resources in order to maintain constant the execution time while increasing the calculation accuracy.

Asynchronous Task Dispatch for High Throughput Computing for the eServer IBM Blue Gene® Supercomputer

Amanda Peters[1], Alan King[2], Tom Budnik[1], Pat McCarthy[1],
Paul Michaud[3], Mike Mundy[1], James Sexton[2], Greg Stewart[1]
[1]IBM Systems and Technology Group
Rochester, MN 55901
{apeters, tbudnik, pjmccart, mmundy, gregstew}@us.ibm.com
[2]IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
kingaj, sextonjc @us.ibm.com
3IBM Software Group
Houston, TX 77056
pkmichaud @us.ibm.com

High Throughput Computing (HTC) environments strive “to provide large amounts of processing capacity to customers over long periods of time by exploiting existing resources on the network” according to Basney and Livny. A single Blue Gene/L rack can provide thousands of CPU resources into HTC environments. This paper discusses the implementation of an asynchronous task dispatch system that exploits a recently released feature of the Blue Gene/L control system – called HTC mode – and presents data on experimental runs consisting of the asynchronous submission of multiple batches of thousands of tasks for financial workloads. The methodology developed here demonstrates how systems with very large processor counts and light-weight kernels can be configured to deliver capacity computing at the individual processor level in future petascale computing systems.

Parallel Numerical Simulation of Strategic Bankruptcy

Yu Cai and Howard Qi
Michigan Technological University
{cai,howardqi}@mtu.edu

In this paper, we extend the strategic default bankruptcy model to predict risky premium on defaultable bonds in a more realistic economic environment. By considering interest rates, taxes and the voltality of business operations, the model becomes considerably complicated, which imposes significant challenges on the mathematical framework as well as the computation power required in simulating the stochastic process. Since it is hard to obtain a closed form analytical solution for the framework, numerical simulation is an alternative. We present a dynamic block allocation algorithm for parallel Quasi-Monte Carlo(QMC) simulation. The convergence speed of the model is also studied. Simulation results show that our model can be used to estimate risk and risk premium in financial economics.

Parameterization Based on Randomized Quasi-Monte Carlo Methods

Giray Ökten and Matthew Willyard
Department of Mathematics
Florida State University
Tallahassee, FL 32306-4510
okten@math.fsu.edu
mwillyar@math.fsu.edu

We present a theoretical framework where any randomized quasi-Monte Carlo method can be viewed and analyzed as a parameterization method for parallel quasi- Monte Carlo. We present deterministic and stochastic error bounds when different processors of the computing environment run at different speeds. We implement two parameterization methods, both based on randomized quasi-Monte Carlo, and apply them to pricing digital options and collateralized mortgage obligations. Numerical results are used to compare the parameterization methods by their parallel performance as well as their Monte Carlo efficiency.

Option Pricing, Maturity Randomization and Grid Computing

Marina Marena
Department De Castro
Università di Torino
Torino, Italy
marena@econ.unito.it

Daniele Marazzina
Department SEMeQ
Università A.Avogadro
Novara, Italy
marazzina@eco.unipmn.it

Gianluca Fusai
Department SEMeQ
Università A.Avogadro
Novara, Italy
fusai@eco.unipmn.it

By geometric randomization of the option maturity, we transform the n-steps backward recursion that arises in option pricing into an integral equation. The option price is then obtained solving n independent integral equations. This is accomplished by a quadrature procedure that transforms each integral equation in a linear system. Since the solution of each linear system is independent one of the other, we can exploit the potentiality of the grid architecture AGA. We compare different quadrature methods of the integral equation with Monte Carlo simulation. Therefore we price options (such as plain vanilla, single and double barrier options) when the underlying evolves according to different exponential Lévy processes.

Do Price Limits Inhibit Futures Prices?

Latha Shanker — Concordia University
Narayanaswamy Balakrishnan — McMaster University

We investigate the effect of daily price limits, which trigger a trading halt if the limit is hit, in futures markets. Empirically, it has been observed that futures price limits are rarely hit. This could be because traders avoid putting in bid-ask quotes which could trigger a trading halt. If this is true, futures prices would cluster in a narrow region close to the limits without hitting them. We test this for the British pound futures contract for a period in which limits are imposed, by comparing the number of daily observations of futures prices which fall in the narrow region, with that predicted at the 99% confidence level by the distribution of the 'true' daily futures price. Our tests require that we calculate all possible combinations of a long time series, which we find we can do relatively efficiently with a parallel program.

Financial Evaluation of Participating Life Insurance Policies in Distributed Environments

Stefania Corsaro, Pasquale Luigi De Angelis, Zelda Marino, Francesca Perla and Paolo Zanetti
University of Naples "Parthenope"
Via Medina 40, 80133 Naples, Italy
{corsaro,deangelis,marino,perla,zanetti}@uniparthenope.it

In this paper we discuss the development of a parallel software for the numerical simulation of Participating Life Insurance Policies on distributed architectures. The use of stochastic pricing models, together with the request of solutions in a "useful" time, which have to be sufficiently accurate for the financial problem as well, make the financial problem a very computationally intensive one; as a consequence, advanced architectures are mandatory for effective decisional processes. The mathematical models, which describe the considered problems, usually require the evaluation of multidimensional integrals and the solution of Stochastic Differential Equations. The involved integrals are solved by means of Monte Carlo method in conjunction with the Antithetic Variates variance reduction technique, while Differential Equations are approximated via a fully implicit Euler scheme. The parallelization strategy we adopt relies on the parallelization of Monte Carlo algorithm. We implement and test the software on cluster architectures.

Towards the Development of a Decentralized Market Information System: Requirements and Architecture

René Brunner, Felix Freitag and Leandro Navarro
Computer Architecture Department, Polytechnic University of Catalonia
08014, Barcelona
Spain
{rbrunner, felix, leandro}@ac.upc.edu

In a market, information about its specifications and the behavior of its participants is essential for sophisticated and efficient negotiation strategies. However, there is currently no completely researched system to provide and consult an overall knowledge of economic information in distributed markets. These markets are implemented for example by Grid applications and gained importance over the last few years. This paper presents the economic information requirements and a high-level architecture overview for a Decentralized Market Information System (DMIS). The proposed system acquires economic data in a distributed environment for providing it to individual traders or other participants in a decentralized manner. First, we outline the economic information requirements which the system needs to achieve. Therefore their properties and a privacy model has to be considered. Then we propose an architecture for the system which combines technologies of distributed information aggregation system and distributed publish-subscribe models, based on a structured overlay network. The architecture has been designed to meet both the economic information requirements and that of scalability and robustness of a large-scale distributed environment. Initial measurements confirm the proof-of-concept implementation of the existing prototype.

Workshop 22

Workshop on Large-Scale Parallel Processing

LSPP 2008

Data Access Optimizations for Highly Threaded Multi-Core CPUs with Multiple Memory Controllers

Georg Hager, Thomas Zeiser and Gerhard Wellein
Regionales Rechenzentrum Erlangen
91058 Erlangen, Germany
{georg.hager,thomas.zeiser,gerhard.wellein}@rrze.uni-erlangen.de

Processor and system architectures that feature multiple memory controllers are prone to show bottlenecks and erratic performance numbers on codes with regular access patterns. Although such effects are well known in the form of cache thrashing and aliasing conflicts, they become more severe when memory access is involved. Using the new Sun UltraSPARC T2 processor as a prototypical multi-core design, we analyze performance patterns in low-level and application benchmarks and show ways to circumvent bottlenecks by careful data layout and padding.

Experiences in Scaling Scientific Applications on Current-generation Quad-core Processors

Kevin Barker, Kei Davis, Adolfy Hoisie, Darren Kerbyson, Mike Lang,
Scott Pakin and José Carlos Sancho
Performance and Architecture Lab (PAL), Los Alamos National Laboratory, USA
{kjbarker,kei,hoisie,djk,mlang,pakin,jcsancho}@lanl.gov

In this work we present an initial performance evaluation of AMD and Intel's first quad-core processor offerings: the AMD Barcelona and the Intel Xeon X7350. We examine the suitability of these processors in quad-socket compute nodes as building blocks for large-scale scientific computing clusters. Our analysis of intra-processor and intra-node scalability of microbenchmarks and a range of large-scale scientific applications indicates that quad-core processors can deliver an improvement in performance of up to 4x per processor but is heavily dependent on the workload being processed. While the Intel processor has a higher clock rate and peak performance, the AMD processor has higher memory bandwidth and intra-node scalability. The scientific applications we analyzed exhibit a range of performance improvements from only 3x up to the full 16x speed-up over a single core. Also, we note that the maximum node performance is not necessarily achieved by using all 16 cores.

Locality Aware MPI Communication on a Commodity Opto-Electronic Hybrid Network

Shin'ichiro Takizawa and Toshio Endo
Tokyo Institute of Technology
{takizawa,endo}@matsulab.is.titech.ac.jp

Satoshi Matsuoka
Tokyo Institute of Technology
National Institute of Informatics
matsu@is.titech.ac.jp

Future supercomputers with millions of processors would pose significant challenges in their interconnection networks due to difficulty in design constraints such as space, cable length, cost, power consumption, etc. Instead of huge switches or bisection bandwidth restricted topologies such as a torus, we propose a network which utilizes both fully-connected lower-bandwidth electronic packet switching (EPS) network and low-power optical circuit switching (OCS) network. Optical circuits, connected sparingly to only a limited set of nodes to conserve power and cost, are used in a supplemental fashion as "shortcut" routes only when a node communicates substantially across EPS switches, while short latency communication is handled by EPS only. Our MPI inter-node communication algorithm accommodates for such a network by appropriate scheduling of nodes according to application communication patterns, in particular utilizing relatively high EPS local switch bandwidth to forward messages to nodes with optical connections for shortcutting in order to maximize overall throughput. Simulation studies confirm that our proposal effectively avoids contentions in the network in high-bandwidth applications with nominal additions of optical circuitry to existing machines.

Symbolic Expression Analysis for Compiled Communication

Shuyi Shao[1], Yu Zhang[2], Alex K. Jones[3] and Rami Melhem[4]

[1,4]University of Pittsburgh
Department of CS
Pittsburgh, PA 15260 USA
{[1]syshao, [4]melhem}@cs.pitt.edu

[2,3]University of Pittsburgh
Department of ECE
Pittsburgh, PA 15260 USA
[2]yuz28@pitt.edu,[3]akjones@ece.pitt.edu

Enabling circuit switching in multiprocessor systems has the potential to achieve more efficient communication with lower cost compared to packet/wormhole switching. However, in order to accomplish this efficiently, assistance from the compiler is required to reveal the communication pattern in the parallel application. In this paper we present symbolic expression analysis techniques in a MPI parallel compiler. Symbolic expression analysis allows the identification and representation of the communication pattern and also assists in the determination of communication phases in MPI parallel applications at compile-time. We demonstrate that using the compiler analysis based on symbolic expression analysis to determine the communication pattern and phases provides an average of 2.6 times improvement in message delay over a threshold-based runtime system for our benchmarks with a maximum improvement of 9.7 times.

Optimization of Infiniband for Scientific Applications

Gregory Johnson, Darren J. Kerbyson and Mike Lang
Performance and Architecture Lab (PAL), Los Alamos National Laboratory, USA
{djk,mlang}@lanl.gov

The achievable performance on Infiniband networks is governed by the latencies and bandwidths of communication channels as well as by contention within the network. Currently Infiniband statically routes messages and thus do not take into account dynamic loading of the channels. By interrogating the network routing tables we quantify the contention that occurs for a number of communication patterns using a large-scale (1024 processor) system. Empirical data confirms our contention calculation almost exactly. Custom routing tables are defined that provide both optimum and worst-case performance for a largerange of communication patterns. Performance differences can be as large as 12× (from optimum to worst-case). Two large-scale applications show a runtime improvement of between 10-20% and up to 40% improvement in just their communication time when using optimized routing tables. The approach taken is applicable to many Infiniband systems, and we expect the performance improvements to be even greater on larger-scale systems.

A Scalable, Asynchronous Spanning Tree Algorithm on a Cluster of SMPs

Guojing Cong	Hanhong Xue
IBM Research	IBM System & Technology Group
Yorktown Heights, NY	Poughkeepsie, NY
gcong@us.ibm.com	hanhong@us.ibm.com

Large-scale data science applications require manipulating large graphs distributed across multiple processors. In this paper we present our experimental study of an asynchronous, distributed spanning tree algorithm that handles the challenging random, sparse graphs with billions of vertices. With a constant number of barriers, our implementation scales to 1024 processors on a cluster of SMPs. Our algorithm sheds new light on the design and implementation of graph algorithms on distributed-memory machines.

On the Concept of Simultaneous Execution of Multiple Applications on Hierarchically Based Cluster and the Silicon Operating System

N.Venkateswaran, Vinoth Krishnan Elangovan, Karthik Ganesan
TP Ramnath Sai Sagar, Sriram Aananthakrishanan, Shreyas Ramalingam
Shyamsundar Gopalakrishnan, Madhavan Manivannan, Deepak Srinivasan
Viswanath Krishnamurthy, Karthik Chandrasekar, Vishwanath Venkatesan
Balaji Subramaniam, Vidya Sangkar L, Aravind Vasudevan
Shrikanth Ganapathy, Sriram Murali and Murali Thyagarajan
Waran Research Foundation (WARFT), Chennai, India.

In this paper we present a novel cluster paradigm and silicon operating system. Our approach in developing the competent cluster design revolves around an execution model to aid the execution of multiple independent applications simultaneously on the cluster, leading to cost sharing across applications. The execution model should envisage simultaneous execution of multiple applications (running traces of multiple independent applications in the same node at an instant, without time sharing) and on all the partitions(nodes) of a single cluster, without sacrificing the performance of individual application, unlike in the current cluster models. Performance scalability is achieved as we increase the number of nodes, the problem size of the individual independent applications, due to non-dependency across applications and hence increase in the number of non-dependent operations(as the problem sizes of the applications get increased) and this leads to better utilization of the unused resources within the node. This execution model is very much dependent on the node architecture for performance scalability. This would be a major initiative towards achieving performance Cost-Effective Supercomputing.

Application-specific Topology-aware Mapping for Three Dimensional Topologies

Abhinav Bhatelé and Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
bhatele2@uiuc.edu, kale@uiuc.edu

The fastest supercomputers today such as Blue Gene/L and XT3 are connected by a 3-dimensional torus/mesh interconnect. Applications running on these machines can benefit from topology-awareness while mapping tasks to processors at runtime. By co-locating communicating tasks on nearby processors, the distance traveled by messages and hence the communication traffic can be minimized, thereby reducing communication latency and contention on the network. This paper describes preliminary work utilizing this technique and performance improvements resulting from it in the context of a n-dimensional k-point stencil program. It shows that for a fine-grained application with a high communication to computation ratio, topology-aware mapping has a significant impact on performance. Automated topology-aware mapping by the runtime using similar ideas can relieve the application writer from this burden and result in better performance. Preliminary work towards achieving this for a molecular dynamics application, NAMD, is also presented. Results on up to 32,768 processors of IBM's Blue Gene/L and 2,048 processors of Cray's XT3 support the ideas discussed in the paper.

Using Hardware Multithreading to Overcome Broadcast/Reduction Latency in an Associative SIMD Processor

Kevin Schaffer and Robert A. Walker
Department of Computer Science
Kent State University
Kent, Ohio 44242
{kschaffe, walker}@cs.kent.edu

The latency of broadcast/reduction operations has a significant impact on the performance of SIMD processors. This is especially true for associative programs, which make extensive use of global search operations. Previously, we developed a prototype associative SIMD processor that uses hardware multithreading to overcome the broadcast/reduction latency. In this paper we show, through simulations of the processor running an associative program, that hardware multithreading is able to improve performance by increasing system utilization, even for processors with hundreds or thousands of processing elements. However, the choice of thread scheduling policy used by the hardware is critical in determining the actual utilization achieved. We consider three thread scheduling policies and show that a thread scheduler that avoids issuing threads that will stall due to pipeline dependencies or thread synchronization operations is able to maintain system utilization independent of the number of threads.

Faster Matrix-Vector Multiplication on GeForce 8800GTX

Noriyuki Fujimoto
Graduate School of Information Science and Technology, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan
fujimoto@ist.osaka-u.ac.jp

Recently a GPU has acquired programmability to perform general purpose computation fast by running ten thousands of threads concurrently. This paper presents a new algorithm for dense matrix-vector multiplication on NVIDIA CUDA architecture. The experimental results on GeForce 8800GTX show that the proposed algorithm runs maximum 15.69 (resp., 32.88) times faster than the sgemv routine in NVIDIA's BLAS library CUBLAS 1.1 (resp., IntelMath Kernel Library 9.1 on one-core of 2.0 GHz Intel Xeon E5335 CPU with SSE3 SIMD instructions) for matrices with order 16 to 12800. The performance, including the data transfer between CPU and GPU, of Jacobi's iterative method for solving linear equations shows that the proposed algorithm is practical for some real applications.

GPU Acceleration of Numerical Weather Prediction

John Michalakes
National Center for Atmospheric Research
Boulder, CO
michalak@ucar.edu

Manish Vachharajani
University of Colorado
Boulder, CO
manishv@colorado.edu

Weather and climate prediction software has enjoyed the benefits of exponentially increasing processor power for almost 50 years. Even with the advent of large-scale parallelism in weather models, much of the performance increase has come from increasing processor speed rather than increased parallelism. This free ride is nearly over. Recent results also indicate that simply increasing the use of large-scale parallelism will prove ineffective for many scenarios. We present an alternative method of scaling model performance by exploiting emerging architectures using the fine-grain parallelism once used in vector machines. The paper shows the promise of this approach by demonstrating a 20× speedup for a computationally intensive portion of the Weather Research and Forecast (WRF) model on an NVIDIA 8800 GTX Graphics Processing Unit (GPU). We expect an overall 1.3× speedup from this change alone.