# Proceedings of the 25th IEEE International Parallel & Distributed Processing Symposium, Workshops and Phd Forum

**IPDPS 2011 Advance Program Abstracts**

**Abstracts for 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.**

# Contents

# Workshop 1
# Heterogeneity in Computing Workshop
# HCW 2011

# Time Utility Functions for Modeling and Evaluating Resource Allocations in a Heterogeneous Computing System

Luis Diego Briceño[1], Bhavesh Khemka[1], Howard Jay Siegel[1,2], Anthony A. Maciejewski[1],
Christopher Groër[3], Gregory Koenig[3], Gene Okonski[4], and Steve Poole[3,4]

**Colorado State University**
**[1] Department of Electrical and Computer Engineering**
**[2] Department of Computer Science**
**Fort Collins, CO 80523**

**[3]Oak Ridge National Laboratory**
**Oak Ridge, TN 37830**

**[4]Department of Defense**
**Washington, DC 20001**

email: {LDBricen, Bhavesh.Khemka, HJ, AAM}@colostate.edu
{GroerCS, Koenig, SPoole}@ornl.gov
okonskitg@verizon.net

## Abstract

This study considers a heterogeneous computing system and corresponding workload being investigated by the Extreme Scale Systems Center (ESSC) at Oak Ridge National Laboratory (ORNL). The ESSC is part of a collaborative effort between the Department of Energy (DOE) and the Department of Defense (DoD) to deliver research, tools, software, and technologies that can be integrated, deployed, and used in both DOE and DoD environments. The heterogeneous system and workload described here are representative of a prototypical computing environment being studied as part of this collaboration. Each task can exhibit a time-varying *importance* or *utility* to the overall enterprise. In this system, an arriving task has an associated priority and precedence. The priority is used to describe the importance of a task, and precedence is used to describe how soon the task must be executed. These two metrics are combined to create a utility function curve that indicates how valuable it is for the system to complete a task at any given moment. This research focuses on using time-utility functions to generate a metric that can be used to compare the performance of different resource schedulers in a heterogeneous computing system. The contributions of this paper are: (a) a mathematical model of a heterogeneous computing system where tasks arrive dynamically and need to be assigned based on their priority, precedence, utility characteristic class, and task execution type, (b) the use of priority and precedence to generate time-utility functions that describe the value a task has at any given time, (c) the derivation of a metric based on the total utility gained from completing tasks to measure the performance of the computing environment, and (d) a comparison of the performance of resource allocation heuristics in this environment.

# Optimized Barriers for Heterogeneous Systems Using MPI

Jan C. Meyer and Anne C. Elster
Norwegian University of Science and Technology
Dept. of Computer and Information Science
Sem Sælands v. 7-9, NO-7491 Trondheim, Norway
{janchris,elster}@idi.ntnu.no

## Abstract

The heterogeneous communication characteristics of clustered SMP systems create great potential for optimizations which favor physical locality. This paper describes a novel technique for automating such optimizations, applied to barrier operations. Portability poses a challenge when optimizing for locality, as costs are bound to variations in platform topology. This challenge is addressed through representing both platform structure and barrier algorithms as input data, and altering the algorithm based on benchmark results which can be easily obtained from a given platform. Our resulting optimization technique is empirically tested on two modern clusters, up to eight dual quad-core nodes on one, and up to ten dual hex-core nodes on another. Included test results show that the method captures performance advantages on both systems without any explicit customization, and produces specialized barriers of superior performance to a topology-neutral implementation.

# Characterizing Task-Machine Affinity in Heterogeneous Computing Environments

Abdulla M. Al-Qawasmeh[1], Anthony M. Maciejewski[1], Rodney Roberts[3], and Howard J. Siegel[1,2]

[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
Colorado State University, Fort Collins, Colorado, USA
[3]Department of Electrical and Computer Engineering
Florida State University, Tallahassee, Florida, USA

## Abstract

Many computing environments are heterogeneous, i.e., they consist of a number of different machines that vary in their computational capabilities. These machines are used to execute task types that vary in their computational requirements. Characterizing heterogeneous computing environments and quantifying their heterogeneity is important for many applications. In previous research, we have proposed preliminary measures for machine performance homogeneity and task-machine affinity. In this paper, we build on our previous work by introducing a complementary measure called the task difficulty homogeneity. Furthermore, we refine our measure of task-machine affinity to be independent of the task type difficulty measure and the machine performance homogeneity measure. We also give examples of how the measures can be used to characterize heterogeneous computing environments that are based on real world task types and machines extracted from the SPEC benchmark data.

## Scheduling on unspecified heterogeneous distributed resources

Daniel Millot and Christian Parrot
Computer Science Dept.
TELECOM sudParis
Évry, France
{Daniel.Millot, Christian.Parrot}@it-sudparis.eu

### Abstract

In this paper, we present an adaptive method for scheduling parallel applications on unspecified distributed memory platforms. The presented method can be used to schedule parallel applications when the total workload and the execution parameters (communication speed, available computing power...) are unspecified. When used to schedule divisible load applications according to a masterworker model, this method delivers the workload through multiple rounds. In order to maximize the throughput of each worker, it can prevent both idleness in the use of workers and contentions in the use of the links between master and workers. Before focusing on the proposed scheduling method, the paper recalls the underlying methods on which its development relies. The paper then gives a theoretical analysis of the method before presenting results of simulations obtained with the SimGrid framework on a limited distributed memory platform.

## MO-Greedy: an extended beam-search approach for solving a multi-criteria scheduling problem on heterogeneous machines

Louis-Claude Canon
Ensimag, INRIA Rhône-Alpes

Emmanuel
INRIA Bordeaux Sud-Ouest, LaBRI

### Abstract

Optimization problems can often be tackled with respect to several objectives. In such cases, there can be several incomparable Pareto-optimal solutions. Computing or approximating such solutions is a major challenge in algorithm design. Here, we show how to use an extended beam-search technique to solve a multi-criteria scheduling problem for heterogeneous machines. This method, called MO-Greedy (for Multi-Objective greedy), allows the design of a multi-objective algorithm when a single-objective greedy one is known. We show that we can generate, in a single execution, a Pareto front optimized with respect to the preferences specified by the decision maker. We compare our approach to other heuristics and an approximation algorithm and show that the obtained front is, on average, better with our method.

## A Model-based Schedule Representation for HeterogeneousMapping of Dataflow Graphs

Hsiang-Huang Wu, Chung-Ching Shen, Nimish Sane, William Plishker, Shuvra S. Bhattacharyya
Department of Electrical & Computer Engineering, and
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland, USA
{hhwu, ccshen, nsane, plishker, ssb}@umd.edu

### Abstract

Dataflow-based application specifications are widely used in model-based design methodologies for signal processing systems. In this paper, we develop a new model called the dataflow schedule graph (DSG) for representing a broad class of dataflow graph schedules. The DSG provides a graphical representation of schedules based on dataflow semantics. In conventional approaches, applications are represented using dataflow graphs, whereas schedules for the graphs are represented using specialized notations, such as various kinds of sequences or looping constructs. In contrast, the DSG approach employs dataflow graphs for representing both application models and schedules that are derived from them. Our DSG approach provides a precise, formal framework for unambiguously representing, analyzing, manipulating, and interchanging schedules. We develop detailed formulations of the DSG representation, and present examples and experimental results that demonstrate the utility of DSGs in the context of heterogeneous signal processing system design.

## A Waterfall Model to Achieve Energy Efficient Task Mapping for Large Scale GPU Cluster

Wenjie Liu[1], Zhihui Du[1,*], Yu Xiao[2], David A. Bader[3], and Chen Xu[2]
[1] Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China
[*]Corresponding Author's Email: duzh@tsinghua.edu.cn
[2]Beijing University of Posts and Telecommunications, China
[3]College of Computing, Georgia Institute of Technology, Atlanta, GA, 30332, USA

### Abstract

High energy consumption has become a critical problem for supercomputer systems. GPU clusters are becoming an increasingly popular architecture for building supercomputers because of its great improvement in performance. In this paper, we first formulate the tasks mapping problem as a mini-mal energy consumption problem with deadline constraint. Its optimizing object is very different from the traditional mapping problem which often aims at minimizing makespan or minimiz-ing response time. Then a Waterfall Energy Consumption Model, which abstracts the energy consumption of one GPU cluster system into several levels from high to low, is proposed to achieve an energy efficient tasks mapping for large scale GPU clusters. Based on our Waterfall Model, a new task mapping algorithm is developed which tries to apply different energy saving strategies to keep the system remaining at lower energy levels. Our mapping algorithm adopts the Dynamic Voltage Scaling, Dynamic Resource Scaling and $\beta$-migration for GPU sub-task to significantly reduce the energy consumption and achieve a better load balance for GPU clusters. A task generator based on the real task traces is developed and the simulation results show that our mapping algorithm based on the Waterfall Model can reduce nearly 50% energy consumption compared with traditional approaches which can only run at a high energy level. Not only the task deadline can be satisfied, but also the task execution time of our mapping algorithm can be reduced.

# An Agent-Based Approach to Reconciling Data Heterogeneity in Cyber-Physical Systems

Jing Lin†, Sahra Sedigh†, and Ali R. Hurson‡
†Department of Electrical and Computer Engineering
‡Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri 65409–0040
Email: {jlpg2,sedighs,hurson}@mst.edu

## Abstract

Computing and communication devices in any cyber-physical system (CPS) of non-trivial scale exhibit significant heterogeneity. Critical infrastructure systems, which are prime examples of CPSs, are no exception. The extent of networking capability, decentralized control, and more generally, integration between the cyber and physical infrastructures can vary greatly within a large-scale CPS. Other manifestations of heterogeneity in CPSs are in the resolution, syntax, and semantics of data collected by sensors from the physical infrastructure. Similar challenges complicate the use of databases that maintain past sensor data, device settings, or information about the physical infrastructure. The work presented in this paper aims to address these challenges by using the summary schemas model (SSM), which enables heterogeneous data sources to be queried with an unrestricted view and/or terminology. This support for imprecise queries significantly broadens the scope of data that can be used for intelligent decision support and carries the promise of increased reliability and performance for the CPS. We seek to ensure that ambiguity and imprecision do not accompany this expanded scope. The ultimate goal of a CPS is to fortify and streamline the operation of its physical infrastructure. The success of this task is contingent upon correct and efficient interpretation of data describing the state of the physical components, and the constraints to which it is subject. To this end, we propose agent-based semantic interpretation services that extract meaningful and useful information from raw data from heterogeneous sources, aided by the SSM. The proposed approach is described in the context of intelligent water distribution networks, which are cyber-physical critical infrastructure systems responsible for reliable delivery of potable water. The methodology is general, and can be extended to a broad range of CPSs, including smart power grids and intelligent transportation systems.

# NVCR: A Transparent Checkpoint-Restart Library for NVIDIA CUDA

Akira Nukada*[†] , Hiroyuki Takizawa[‡] and Satoshi Matsuoka*[§][†]
Scientific Information and Computing Center, Tokyo Institute of Technology
Email: nukada@matsulab.is.titech.ac.jp,matsu@is.titech.ac.jp
[†] CREST, Japan Science and Technology Agency
[‡] Graduate School of Information Sciences, Tohoku University
Email: tacky@isc.tohoku.ac.jp
[§] National Institute of Informatics

## Abstract

Today, CUDA is the de facto standard programming framework to exploit the computational power of graphics processing units (GPUs) to accelerate various kinds of applications. For efficient use of a large GPU-accelerated system, one important mechanism is checkpoint-restart that can be used not only to improve fault tolerance but also to optimize node/slot allocation by suspending a job on one node and migrating the job to another node. Although several checkpoint-restart implementations have been developed so far, they do not support CUDA applications or have some severe limitations for CUDA support. Hence, we present a checkpoint-restart library for CUDA that first deletes all CUDA resources before checkpointing and then restores them right after checkpointing. It is necessary to restore each memory chunk at the same memory address. To this end, we propose a novel technique that replays memoryrelated API calls. The library supports both CUDA runtime API and CUDA driver API. Moreover, the library is transparent to applications; it is not necessary to recompile the applications for checkpointing. This paper demonstrates that the proposed library can achieve checkpoint-restart of various applications at acceptable overheads, and the library also works for MPI applications such as HPL.

# Use of Internet Embedding Tools for Heterogeneous Resources Aggregation

Olivier Beaumont
INRIA Bordeaux Sud-Ouest
Bordeaux, France
Email: olivier.beaumont@labri.fr

Nicolas Bonichon
Université de Bordeaux
Bordeaux, France
Email: nicolas.bonichon@labri.fr

Philippe Duchon
Université de Bordeaux
Bordeaux, France
Email: philippe.duchon@labri.fr

Hubert Larchevêque
Institut Polytechnique de Bordeaux
Bordeaux, France
Email: hubert.larcheveque@labri.fr

## Abstract

In this paper we are interested in large scale distributed platforms like BOINC, consisting of heterogeneous resources and using the Internet as underlying communication network. In this context, we study a resource clustering problem, where the goal is to build clusters having at least a given capacity and such that any two participants to the same cluster are not too far from each other. In this context, the distance between two participants corresponds to the latency of a communication between them. Our goal is to provide algorithms with provable approximation ratios. In such large scale networks, it is not realistic to assume that the whole latency matrix (that gives the latency between any two participants) is known, and we need to rely on embedding tools such as Vivaldi or Sequoia. These tools enable to work on compact descriptions and well described metric spaces in which the distance between two points can be obtained directly from a small amount of information available at each node. We present the Bin Covering under Distance Constraint problem (BCDC for short), and propose dedicated algorithms for this problem for each metric space induced by each of the embedding tools. Then, we propose a comparison of these algorithms based on actual latency measures, that enables to decide which algorithm/embedding tool pair offers in practice for realistic datasets the best balancing between distance prediction and approximation ratios for the resource clustering problem.

# Workshop 2
# Reconfigurable Architectures Workshop
# RAW 2011

# A High-Level Power Model for MPSoC on FPGA

Roberta Piscitelli and Andy D. Pimentel
Computer Systems Architecture group
Informatics Institute, University of Amsterdam, The Netherlands
Email: {r.piscitelli,a.d.pimentel}@uva.nl

## Abstract

This paper presents a framework for high-level power estimation of multiprocessor systems-on-chip (MPSoC) architectures on FPGA. The technique is based on abstract execution profiles, called event signatures, and it operates at a higher level of abstraction than, e.g., commonly-used instruction-set simulator (ISS) based power estimation methods and should thus be capable of achieving good evaluation performance. As a consequence, the technique can be very useful in the context of early system-level design space exploration. We integrated the power estimation technique in a system-level MPSoC synthesis framework. Subsequently, using this framework, we designed a range of different candidate architectures which contain different numbers of Microblaze processors and compared our power estimation results to those from real measurements on a Virtex-6 FPGA board.

# Modular Framework for Multi-level Multi-device MPSoC Simulation

Christoph Roth[†] , Gabriel Marchesan Almeida[*] , Oliver Sander[†] , Luciano Ost[*] ,
Nicolas Hébert[*] , Gilles Sassatelli[*] , Pascal Benoit[*] , Lionel Torres[*] , Jürgen Becker[†]
[†] Institute for Information Processing Technology (ITIV)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
{christoph.roth, sander, becker}@kit.edu
[*] Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
Department of Microelectronics
Montpellier, France
{marchesan, copelloost, hebert, sassatelli, pbenoit, torres}@lirmm.fr

## Abstract

Multi-Processor System-on-Chips (MPSoCs) as evolution of traditional embedded system architectures demand for a detailed exploration on architectures and system design. Simulation time and complexity are major issues, as such systems get more and more complex. This paper presents a novel simulation framework aiming the simulation of multiple MPSoC systems on different abstraction levels while using several simulation platforms concurrently in order to address aforementioned challenges. We introduce the concept of MultiX-Simulation which enables a tradeoff between performance and accuracy and allows for the flexible adaptation to various use cases. The backbone of the framework is formed by the High Level Architecture, a simulation middleware for Parallel Discrete Event Simulation. Also within this contribution different performance aspects of the framework are evaluated by using various multi-granular MPSoC system models.

# A heterogeneous Multicore System on Chip with run-time reconfigurable virtual FPGA Architecture

M. Hübner[1], P. Figuli[1], R. Girardey[1], D. Soudris[2], K. Siozios[2], J. Becker[1]
[1]Karlsruhe Institute of Technology - KIT, Germany
{michael.huebner, peter.figuli, becker}@kit.edu
[2]National Technical University of Athens, Greece, School of Electrical and Computer Engineering
{dsoudris, ksiop@microlab.ntua.gr}

## Abstract

System design, especially for low power embedded applications often profit from a heterogeneous target hardware platform. The application can be partitioned into modules with specific requirements e.g. parallelism or performance in relation to the provided hardware blocks on the multicore hardware. The result is an optimized application mapping and a parallel processing with lower power consumption on the different cores on the hardware. This paper presents a heterogeneous platform consisting of a microprocessor and a field programmable gate array (FPGA) connected via a standard AMBA bus. The novelty of this approach is that the FPGA is realized as virtual reconfigurable hardware upon a traditional off the shelf FPGA device. The advantage with this approach is that the specification of the virtual FPGA stays unchanged, independent to the underlying hardware and provides therefore features, which the exploited physical host FPGA cannot provide. A special feature of the presented virtual FPGA amongst others is the dynamic reconfigurability which is for example not available with all off the shelf FPGAs. Furthermore the concept of FPGA virtualization enables the re-use of hardware blocks on other physical FPGA devices. This paper presents the hardware platform and describes the tool chain for the heterogeneous system on chip.

# A Scalable Microarchitecture Design that Enables Dynamic Code Execution for Variable-Issue Clustered Processors

Ralf Koenig, Timo Stripf, Jan Heisswolf, Juergen Becker
Karlsruhe Institute of Technology, Karlsruhe, Germany
{ralf.koenig, stripf, heisswolf, becker}@kit.edu

## Abstract

The dynamic run-time complexity of embedded applications is steadily increasing. Currently, only specialized Multiprocessor System-on-Chip (MPSoC) architectures can deliver the required processing power as well as energy efficiency. Although today's MPSoCs incorporate different, potentially reconfigurable cores, their ability to dynamically balance exploitable instruction-, data-, and thread-level parallelism is still very limited. In this paper, we present a novel coarse-grained reconfigurable architecture that can be adapted to operate on different computation granularities and types of parallelism at run time, depending on the application's needs. Our contributions comprise different microarchitectural techniques realizing dynamic operation execution for Run-time Scalable Issue Width (RSIW) processor instances. These enable to adapt on demand the issue width of out-of-order RSIW processor instances. Our results show that significant performance improvements can be obtained by our dynamic operation execution technique compared to atomic instruction execution.

# CRM: Configurable Range Memory for Fast Reconfigurable Computing

Jongkyung Paek
School of EECS
Seoul National University
Seoul, South Korea
jongkpaek@dal.snu.ac.kr

Jongeun Lee*
School of ECE
UNIST
Ulsan, South Korea
jlee@unist.ac.kr

Kiyoung Choi
School of EECS
Seoul National University
Seoul, South Korea
kchoi@snu.ac.kr

## Abstract

Application-specific hardware and reconfigurable processors can dramatically speed up compute-intensive kernels of applications, offloading the burden of main processor. To minimize the communication overhead in such a coprocessor approach, the two processors can share an on-chip memory, which may be considered by each processor as a scratchpad memory. However, this setup poses a significant challenge to the main processor, which now must manage data on the scratchpad explicitly, often resulting in superfluous data copy. This paper presents an enhancement to scratchpad, called *Configurable Range Memory* (CRM), that can reduce the need for explicit management and thus reduce data copy and promote data reuse on the shared memory. Our experimental results using benchmarks from DSP and multimedia applications demonstrate that our CRM architecture can significantly reduce the communication overhead compared to the architecture without shared memory, while not requiring explicit data management.

# Tuple Spaces in Hardware for Accelerated Implicit Routing

Zachary K. Baker and Justin L. Tripp
Los Alamos National Laboratory
Los Alamos, NM 87545
E-mail: {zbaker, jtripp}@lanl.gov

## Abstract

Organizing and optimizing data objects on networks with support for data migration and failing nodes is a complicated problem to handle as systems expand to hundreds of thousands of nodes. The goal of this work is to demonstrate that high levels of speedup can be achieved by moving responsibility for finding, fetching, and staging data into an FPGA-based network interface. We present a system for implicit routing of data via FPGA-based network cards. In this system, data structures are requested by name, and the network cooperatively finds the data and returns the information to the requester. This is achieved through successive examination of hardware hash tables implemented in the individual FPGA network cards. By avoiding the complex network software stacks between nodes, the data is quickly transferred entirely through FPGA-FPGA interaction. The performance of this system is approximately 26x faster vs. the software network on a per-node basis. This is due to the improved speed of the hash tables, higher levels of network abstraction and lowered latency between the network nodes.

# High Speed Partial Run-Time Reconfiguration Using Enhanced ICAP Hard Macro

Simen Gimle Hansen, Dirk Koch and Jim Torresen
Department of Informatics, University of Oslo, Norway
{simenha, koch, jimtoer}@ifi.uio.no

## Abstract

Achieving high speed run-time reconfiguration is important for the adaptation of partial reconfiguration in many applications. The reconfiguration speed that is currently available today is somehow artificially limited by the FPGA vendors, while the fabrication process technologies used for building the latest devices today are capable of achieving much higher reconfiguration speed. In this paper we will present a new design and implementation method for achieving high speed partial run-time reconfiguration that exceeds the specified reconfiguration speed of today's FPGAs. By adding custom logic around the Internal Configuration Access Port (ICAP) to implement an enhanced ICAP hard macro, we will investigate the partial run-time reconfiguration speed and explore the limits of the ICAP interface. This is done by using overclocking of the ICAP. Compared with previously work on high-speed reconfiguration, using the enhanced ICAP hard macro will significantly increase the reconfiguration speed.

# Online Routing of FPGA Clock Networks for Module Relocation in Partial Reconfigurable Multi Clock Designs

Christian Schuck, Bastian Haetzer*, Michael Hübner, Jürgen Becker
Institut für Technik der Informationsverarbeitung - ITIV
Karlsruhe Institute of Technology - KIT
Vincenz-Prieβnitz-Strasse 1 76131 Karlsruhe
{schuck, huebner, becker}@kit.edu
Embedded Systems Engineering Group (ESE) - ITI
University Stuttgart
Pfaffenwaldring 47 70569 Stuttgart
haetzer@informatik.uni-stuttgart.de

## Abstract

Module-based partial reconfiguration of FPGAs offers great possibilities for runtime flexibility. It enables hardware tasks to swap in and out the design without interruption of the entire system. In this context the techniques of module relocation and the 2-dimensional reconfiguration have been successfully applied in order to reduce the storage requirement for partial bit-streams and to shorten the reconfiguration times significantly. Besides the adaptation on functional level, multiple clock domains and dynamic frequency scaling are key techniques to achieve an adaptation on power and performance level as well. However, current approaches of module relocation provide no support for designs with multiple clock domains. In this paper we present a new method of online clock network routing, which solves this problem. The concept is based on a "on the fly" manipulation of configuration bits which determine the clk-inputs of the single slices. The method is implemented in hardware in order to maximize configuration performance. Our results show that a memory saving of 66% for an example design can be achieved while the original reconfiguration speed could be maintained. Figures for hardware usage are also given.

# Securing Boot of an Embedded Linux on FPGA

Florian Devic
SAS NETHEOS &
LIRMM UMR - CNRS 5506
University of Montpellier 2
Montpellier, FRANCE
Email: f.devic@netheos.net

Lionel Torres
LIRMM UMR - CNRS 5506
University of Montpellier 2
Montpellier, FRANCE
Email: lionel.torres@lirmm.fr

Benoît Badrignans
SAS NETHEOS
Montpellier, FRANCE
Email: b.badrignans@netheos.net

## Abstract

The growing complexity of embedded systems makes reconfiguration and embedded OSs (Operating Systems) more and more interesting. FPGAs (Field-Programmable Gate Arrays) are able to perform such a feature with success.

With most of the FPGAs, the OS is stored into an external memory (usually Flash) and running on a processor embedded into the FPGA. We consider that FPGA embedded processor is able to process the OS update through, for instance, an insecure network.

However, these features may give rise to security flaws affecting the system integrity or freshness. Integrity can be altered by spoofing or modifying data in order to introduce malicious code. In the same way, freshness can be affected by replaying an old configuration in order to downgrade the system.

This work proposes a trusted computing mechanism taking into account the whole security chain from bitstream-to-kernel-boot ensuring, both hardware and software, integrity while preventing replay attacks.

This paper summarizes the current counter-measures ensuring integrity, confidentiality and freshness of the bitstream. Then we propose a solution to protect OS kernel against malicious modifications thanks to already trusted bitstream power-up.

We also evaluate the area and performance overhead of the proposed architecture and its improvement using asymmetric cryptography. Adding security and increasing performances, this solution generates between 0 and 40% of area overhead depending on the re-usability consideration.

# Hyperelliptic Curve Cryptoarchitecture for fast Execution of Schnorr and Okamoto Authentication Protocols

Alexander Klimm, Sebastian Vogel, and Jürgen Becker
Karlsruher Institut für Technologie (KIT)
Institut für Technik der Informationsverarbeitung
Vincenz-Prie*β*nitz-Str. 1, 76131 Karlsruhe, Germany
{klimm,vogel,becker}@kit.edu

## Abstract

A hyperelliptic curve cryptoarchitecture for execution of authentication protocols under hard realtime constraints is proposed in this contribution. Tailoring the architecture of the computation platform to the characteristics of the underlying cryptographic protocol enables for short authentication times, minimal hardware usage, while providing a higher level of security compared to today's standard cryptosystems. Due to the low hardware footprint it is ideally suited for integration into embedded systems with limited resources such as electronic control units in automobiles. All cryptographic functionality is strictly bound to an internal processing module thus hiding the complexity of the HECC operations. Automotive access control systems is am example of an application area that can be matched perfectly by the proposed architecture. Experiments show that a massive speed up can be achieved which can be exploited to save power or introduce additional security measures on algorithmic level.

# A Reconfigurable Multi-core Cryptoprocessor for Multi-channel Communication Systems

Michael Grand[1], Lilian Bossuet[2], Guy Gogniat[3], Bertrand Le Gal[1], Jean-Philippe Delahaye[4], Dominique Dallet[1]
[1]IMS Laboratory, University of Bordeaux, firstname.lastname@ims-bordeaux.fr
[2]Hubert Curien Laboratory, University of Lyon, lilian.bossuet@univ-st-etienne.fr
[3] Lab-STICC Laboratory, University of Bretagne Sud, guy.gogniat@univ-ubs.fr
[4] CELAR, DGA, jean-philippe.delahaye@dga.defense.gouv.fr

## Abstract

This paper presents a reconfigurable Multi-Core Crypto-Processor MCCP) especially designed to secure multi-channel and multi-standard communication systems. Such component meets many constraints like high throughput and flexibility. In contrast, a classical mono-core approach either provides limited throughput or does not allow simple management of multi-channel streams. Nevertheless, parallelism is not sufficient for a multi-standard radio. It is therefore essential to increase MCCP flexibility. To achieve these results, our work takes advantage of the Xilinx FPGA hardware reconfiguration. The proposed MCCP can reach a maximum throughput of 1.7 Gbps at 190 MHz with several AES cipher modes. It uses about 4000 slices on a Virtex 4 FPGA.

# Migrating Static Systems to Partially Reconfigurable Systems on Spartan-6 FPGAs

Christian Beckhoff
Email: christian@recobus.de

Dirk Koch and Jim Torresen
Department of Informatics, University of Oslo
P.O. Box 1080 Blindern, N-0316 Oslo, Norway
Email: {koch, jimtoer}@ifi.uio.no
Web: http://www.matnat.uio.no/forskning/prosjekter/crc/

## Abstract

In this paper we present a design flow for migrating a static only FPGA system into a system featuring partial runtime reconfiguration. The migration can lower device cost and benefits future extendibility. The migration flow with our tool GoAhead is described in detail. In contrast to present EDA tools, GoAhead is the first tool that allows to build systems with partial runtime reconfiguration on Spartan-6 FPGAs. A case study will demonstrate the migration of a static only design into a partially reconfigurable system.

# ReBit: A tool to manage and analyse FPGA-based reconfigurable systems

M. D. Santambrogio[1,2] , A. Cazzaniga[2] , A. Bonetto[2] , Donatella Sciuto[2]
[1] Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
santambr@mit.edu
[2] Dipartimento di Elettronica e Informazione
Politecnico di Milano
{santambr, sciuto}@elet.polimi.it, {andrea.cazzaniga, alessandra.bonetto}@dresd.org

## Abstract

Partial dynamic reconfiguration of FPGAs is a methodology that allows the efficient use of FPGAs resources and an improved degree of flexibility with respect to static hardware when designing an architecture on FPGA. Recently several tools, aiming at supporting the designer in the implementation and the validation processes involved in partial reconfiguration, have been released.

Within this scenario we introduce a framework, called ReBit, intended to be complementary to the most important of tool suite available today, e.g. Xilinx ISE suite, improving the existing features and adding new ones, such as partial bitstream scheduling policy testing and algorithmic bus macros placement, using different APIs, integrated in the framework. These features have been validated using different Xilinx FPGAs Spartan 3, Virtex II Pro and Virtex 4.

# OpenPR: An Open-Source Partial-Reconfiguration Toolkit for Xilinx FPGAs

Ali Asgar Sohanghpurwala[*] , Peter Athanas[*] , Tannous Frangieh[*] and Aaron Wood[†]
[*] Bradley Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA 24061
Email: asohangh,athanas,tannous@vt.edu
[†] Information Sciences Institute East
Email: awood@east.isi.edu

## Abstract

The Xilinx Partial Reconfiguration Early Access Software Tools for ISE 9.2i has been an instrumental package for performing a wide variety of research on Xilinx FPGAs, and is now superseded with the corresponding non-free add-on for ISE 12.3. The original package was developed and offered by Xilinx as a downloadable add-on package to the Xilinx ISE 9.2 tools. The 9.2i toolkit provided a methodology for creating rectangular partial reconfiguration modules that could be swapped in and out of a static baseline design with one or more PR slots. This paper presents a new PR toolkit called OpenPR that, for starters, provides similar functionality as the Xilinx PR Toolkit, yet is extendable to explore other modes of partial reconfiguration. The distinguishing feature of this toolkit is that it is being released as open source, and is intended to extend to the needs of individual researchers.

# Occam-pi as a High-level Language for Coarse-Grained Reconfigurable Architectures

Zain-ul-Abdin and Bertil Svensson
Centre for Research on Embedded Systems (CERES),
Halmstad University, Halmstad, Sweden.
Email: Zain-ul-Abdin@hh.se, Bertil.Svensson@hh.se

## Abstract

Recently we proposed occam-pi as a high-level language for programming coarse grained reconfigurable architectures. The constructs of occam-pi combine ideas from CSP and pi-calculus to facilitate expressing parallelism, communication, and reconfigurability. The feasability of this approach was illustrated by developing a compiler framework to compile occam-pi implementations to the Ambric architecture.

In this paper, we demonstrate the applicability of occam-pi for programing an array of functional units, eXtreme Processing Platform (XPP). This is made possible by extending the compiler framework to target the XPP architecture, including automatic floating to fixed-point conversion. Different implementations of a FIR filter and a DCT algorithm were developed and evaluated on the basis of performance and resource consumption. The reported results reveal that the approach of using occam-pi to program the category of coarse grained reconfigurable architectures appears to be promising. The resulting implementations are generally much superior to those programmed in C and comparable to those hand-coded in the low-level native language NML.

## Dynamic Reconfiguration for Irregular Code using FNC-PAE Processor Cores

Eberhard Schüler, Martin Vorbach, Frank May
PACT XPP Technologies AG
Munich, Germany
es | mv | fm@pactxpp.com

Markus Weinhardt
Faculty of Engineering and Computer Science
Hochschule Osnabrück
Osnabrck, Germany
mweinhardt@computer.org

### Abstract

This paper describes PACT XPP Technologies' Function-PAE (FNC-PAE) Processor Core which was designed for executing irregular, control-flow dominated code efficiently in embedded systems. It combines aspects of dynamically reconfigurable coarse-grain arrays and VLIW processors. The silicon-proven FNC-PAE Cores are tightly integrated with the XPP reconfigurable dataflow array. We present the FNC-PAE architecture, its development environment (assembler, C compiler, and simulator), application examples, and performance data collected from the fully working prototype chip.

## A Replacement Technique to Maximize Task Reuse in Reconfigurable Systems

Juan Antonio Clemente, Daniel Mozos
Departamento de Arquitectura de Computadores y
Automática (DACyA)
Universidad Complutense de Madrid (UCM)
Madrid, Spain
ja.clemente@fdi.ucm.es
Javier Resano
Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
Zaragoza, Spain
jresano@unizar.es

### Abstract

Dynamically reconfigurable hardware is a promising technology that combines in the same device both the high performance and the flexibility that many recent applications demand. However, one of its main drawbacks is the reconfiguration overhead, which involves important delays in the task execution, usually in the order of hundreds of milliseconds, as well as high energy consumption. One of the most powerful ways to tackle this problem is configuration reuse, since reusing a task does not involve any reconfiguration overhead. In this paper we propose a configuration replacement policy for reconfigurable systems that maximizes task reuse in highly dynamic environments. We have integrated this policy in an external task- graph execution manager that applies task prefetch by loading and executing the tasks as soon as possible (ASAP). However, we have also modified this ASAP technique in order to make the replacements more flexible, by taking into account the mobility of the tasks and delaying some of the reconfigurations. In addition, this replacement policy is a hybrid design-time/run-time approach, which performs the bulk of the computations at design time in order to save run-time computations. Our results illustrate that the proposed strategy outperforms other state-of- the-art replacement policies in terms of reuse rates and achieves near-optimal reconfiguration overhead reductions. In addition, by performing the bulk of the computations at design time, we reduce the execution time of the replacement technique by 10 times with respect to an equivalent purely run-time one.

# Integrated Temporal Planning, Module Selection and Placement of Tasks for Dynamic Networks-on-Chip

Philipp Mahr, Steffen Christgau, Christian Haubelt
Department of Computer Science
University of Potsdam
August-Bebel-Str. 89, 14482 Potsdam
Email: {*pmahr, hyperion, haubelt*}@cs.uni-potsdam.de

Christophe Bobda
College of Engineering
University of Arkansas
Fayettville, Arkansas 72701
Email: cbobda@uark.edu

## Abstract

In this work dynamic module selection is integrated in a scheduling and placement flow of tasks for a Dynamic Network-on-Chip. Several implementations (modules) of a task are considered, which differ in size and execution time. In contrast, most state-of-the-art flows consider one module per task, therefore having a static module selection during compile time. Tasks arrive and need to be scheduled and placed by finding a feasible start time and place, such that they meet their deadlines and area requirements. Tasks that do not meet these requirements are rejected. Heuristics for module selection are presented and integrated in an $O(n \log n)$ scheduling and placement flow using EDF-Next-Fit. Evaluation of the dynamic module selection heuristics is performed using synthetic benchmarks. The results show a lower rejection rate of tasks when compared to static module selection.

# Enhancing Resource Utilization with Design Alternatives in Runtime Reconfigurable Systems

Alexander Wold, Dirk Koch, Jim Torresen
Department of Informatics, University of Oslo, Norway
Email: {alexawo,koch,jimtoer}@ifi.uio.no

## Abstract

Average resource utilization in reconfigurable sys- tems is dependent on how effectively modules are placed. Re- sources may be left unused because modules usually have differ- ent resource requirements, and consequently different module bounding boxes. When fitting such different modules on an FPGA, the placement can be imperfect. In order to investigate the effect of module design alternatives on resource utilization, we model the placement problem as a constraint satisfaction problem. In particular we are interested in real world FPGAs. The constraint satisfaction problem is solved using a constraint solver. The constraint problem is modelled as a minimization problem in order to attain the optimal placement. We have in- vestigated module design alternatives which include representing the modules with different layout. We have found an increase in average resource utilization of 11alternatives. The computation time is short, so the method could be applied as a part of an interactive tool.

# Asymmetric Large Size Signed Multipliers Using Embedded Blocks in FPGAs

Shuli Gao
Department of ECE
Royal Military College of Canada
Kingston ON. Canada
Email: shuli.gao@rmc.ca

Dhamin Al-Khalili
Department of ECE
Royal Military College of Canada
Kingston ON. Canada

Noureddine Chabini
Department of ECE
Royal Military College of Canada
Kingston ON. Canada

## Abstract

In this paper, asymmetric non-pipelined large size signed multipliers are implemented using symmetric and asymmetric embedded multipliers in FPGAs. Decomposition of the operands, and consequently the multiplication process, are performed for the efficient use of the embedded blocks. Partial products are organized in various configurations, and the additions of the products are performed in an optimized manner. A heuristic method has been developed, which analyzes the timing and the area at each stage of the adder tree. The optimization algorithm, which is referred to as “Delay-Table” method has led to the minimization of the total critical path delay with reduced utilization of FPGA resources. The asymmetric signed multipliers are implemented in Xilinx FPGAs using 1818-bit and 2518-bit embedded signed multipliers. The implementation results have demonstrated an improvement in terms of speed and number of embedded blocks compared to the standard realization. The improvements are 27.1% in speed and 10.9% in the use of embedded multipliers when using the symmetric embedded blocks. The improvements increase further to 28.5% and 36.6%, respectively, when using asymmetric embedded multipliers.

# Just-in-time Instruction Set Extension - Feasibility and Limitations for an FPGA-based Reconfigurable ASIP Architecture

Mariusz Grad and Christian Plessl
Paderborn Center for Parallel Computing, University of Paderborn
{ mariusz.grad | christian.plessl } @uni-paderborn.de

## Abstract

In this paper, we study the feasibility of moving the instruction set customization process for reconfigurable ASIPs to runtime under the precondition that current FPGA devices and tools are used. To this end we propose a tool flow for just-in-time ASIP customization which identifies suitable custom instructions to accelerate arbitrary binary applications that execute on a virtual machine. The tool flow targets our previously introduced Woolcano reconfigurable ASIP architecture, which augments the PowerPC core in a Xilinx Virtex 4FX CPU with runtime reconfigurable instructions. We evaluate the tool flow with a comprehensive set of applications from the SPEC2006, SPEC2000, MiBench, and SciMark2 benchmark suites and compare the speedups that can be achieved with the overhead of the complete ASIP specialization process. We show that an average speedup of 5× can be achieved for benchmarks from the embedded computing domain. The overhead of custom instruction identification and hardware generation for these benchmarks is less than 50 minutes and will be compensated if the applications execute for more than 2 hours. Finally, we evaluate caching strategies to reduce the time until this break even point is reached.

# The SecretBlaze: A Configurable and Cost-Effective Open-Source Soft-Core Processor

Lyonel Barthe, Luís Vitório Cargnini, Pascal Benoit, Lionel Torres
LIRMM, UMR 5506, CNRS
University of Montpellier 2
161, rue Ada, 34392 Montpellier, France
e-mail: {firstname.lastname}@lirmm.fr

## Abstract

Owing to the flexibility they offer, soft-core processors are becoming an attractive alternative for embedded system designs. Through the use of reconfigurable computing devices, such flexibility provides designers the opportunity to quickly uncover best design trade-offs and to achieve specific application goals. The present work introduces the SecretBlaze, a highly configurable and open-source RISC soft-core processor. Details about its modular approach that aims to strike the balance between the design quality and the efficiency of the datapath are particularly discussed. Synthesis and performance results also demonstrates that the SecretBlaze is a very promising processor.

# Reconfigurable Instruction Decoding for a Wide-Control-Word Processor

Alen Bardizbanyan, Magnus Själander, and Per Larsson-Edefors
VLSI Research Group, Dept. of Computer Science and Engineering,
Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
{alenb,hms,perla}@chalmers.se

## Abstract

Fine-grained control through the use of a wide control word can lead to high instruction-level parallelism, but unless compressed the words require a large memory footprint. A reconfigurable fixed-length decoding scheme can be created by taking advantage of the fact that an application only uses a subset of the datapath for its execution. We present the first complete implementation of the FlexCore processor, integrating a widecontrol- word datapath with a run-time reconfigurable instruction decompressor. Our evaluation, using three different EEMBC benchmarks, shows that it is possible to reach up to 35% speedup compared to a five-stage pipelined MIPS processor, assuming the same datapath units. In addition, our VLSI implementations show that this FlexCore processor offers up to 24% higher energy efficiency than the MIPS reference processor.

# Native Double-Precision LINPACK Implementation on a Hybrid Reconfigurable CPU

Thang Viet Huynh
Graz University of Technology, Graz, Austria
Signal Processing and Speech Communication Laboratory
thang.huynhviet@tugraz.at
Manfred Mücke, Wilfried N. Gansterer
University of Vienna, Vienna, Austria
Research Lab Computational Technologies and Applications
{manfred.muecke|wilfried.gansterer}@univie.ac.at

## Abstract

Applications requiring double precision (DP) arithmetic executed on embedded CPUs without native DP support suffer from prohibitively low performance and power efficiency. Hybrid reconfigurable CPUs, allowing for reconfiguration of the instruction set at runtime, appear as a viable computing platform for applications requiring instructions not supported by existing fixed architectures. Our experiments on a Stretch S6 as prototypical platform show that limited reconfigurable resources on such architectures are sufficient for providing native support of DP arithmetic. Our design using a DP fused multiply-accumulate (FMA) extension instruction achieves a peak performance of 200 MFlop/s and a sustained performance of 22.7 MFlop/s at a clock frequency of 100 MHz. It outperforms LINPACK using software-emulated DP floating-point arithmetic on the S6 by a factor of 5.7 while achieving slightly higher numerical accuracy. In single precision, multiple floating-point operators can be implemented in parallel on the S6.

# Distributed Resource Management in Massively Parallel Processor Arrays

Vahid Lari, Frank Hannig, and Jürgen Teich
Hardware/Software Co-Design, Department of Computer Science
University of Erlangen-Nuremberg, Germany
Email: {vahid.lari, hannig, teich}@cs.fau.de

## Abstract

This paper proposes a methodology for applications to automatically claim linear arrays of processing elements within massively parallel processor arrays at run-time depending on the available degree of parallelism or dynamic computing requirements. Using this methodology, parallel programs running on individual processing elements gain the capability of autonomously managing the available processing resources in their neighborhood. We present different protocols and architectural support for gathering and transporting the result of a resource exploration for informing a configuration loader about the number and location of the claimed resources. Timing and data overhead cost of four different approaches are mathematically evaluated. In order to verify and compare these decentralized algorithms, a simulation platform has been developed to compare the data overhead and scalability of each approach for different sizes of processor arrays.

# Distributed security for communications and memories in a multiprocessor architecture

Pascal Cotret*, Jérémie Crenne*, Guy Gogniat*, Jean-Philippe Diguet*, Lubos Gaspar†, and Guillaume Duc‡
*Laboratoire Lab-STICC, Université de Bretagne-Sud, Lorient (France)
name.surname@univ-ubs.fr
†Laboratoire Hubert-Curien, Université Jean-Monnet, Saint-Etienne (France)
lubos.gaspar@univ-st-etienne.fr
‡Département COMELEC, Télécom ParisTech, Paris (France)
guillaume.duc@telecom-paristech.fr

## Abstract

The need for security in embedded systems has strongly increased since several years. Nowadays, it is possible to integrate several processors in a single chip. The design of such multiprocessor systems-on-chip (MPSoC) must be done with a lot of care as the execution of applications may lead to potential vulnerabilities such as revelation of critical data and private information. Thus it becomes mandatory to deal with security issues all along the design cycle of the MPSoC in order to guarantee a global protection. Among the critical points, the protection of the communications is very sensible as most of the data are exchanged through the communication architecture of the system. This paper targets this point and proposes a solution with distributed enhancements to secure data exchanges and to monitor communications within a MPSoC. In order to validate our contribution, a case study based on a generic multiprocessor architecture is considered.

# Fast hardware computation of x mod z

J. T. Butler
Department of Electrical and Computer Engineering
Naval Postgraduate School
Monterey, CA U.S.A.
T. Sasao
Department of Computer Science and Electronics
Kyushu Institute of Technology
Iizuka, Fukuoka, JAPAN

## Abstract

We show a high-speed hardware implementation of $x$ mod $z$ that can be pipelined in $O(n - m)$ stages, where $x$ is represented in $n$ bits and $z$ is represented in $m$ bits. It is suitable for large $x$. We offer two versions. In the first, the value of $z$ is fixed by the hardware. For example, using this circuit, we show a random number generator that produces more than 11 million random numbers per second on the SRC-6 reconfigurable computer. In the second, $z$ is an independent input. This is suitable for RNS number system applications, for example. The second version can be pipelined in $O(n)$ stages.

# Improving Reconfigurable Hardware Energy Efficiency and Robustness via DVFS-Scaled Homogeneous MP-SoC

Roberto Airoldi, Fabio Garzia and Jari Nurmi
Tampere University of Technology
P.O. Box 553, FIN-33101, Tampere, Finland
Email: name.surname@tut.fi

## Abstract

This paper presents the study of Dynamic Voltage and Frequency Scaling (DVFS) technique applied to an existing multi-core architecture composed of 9 computational nodes interconnected by a hierarchical Network-on-Chip. The architecture was synthesized and characterized in area/power utilizing 65nm standard cell technology. For the analysis of the achievable energy/power saving, a representative algorithm from wireless communications was utilized as test case. Energy and power reduction results achieved with DVFS were then compared to the ones obtainable via clock gating. The results show that DVFS guarantees higher energy savings than clock gating. Moreover, when considering power consumption DVFS improves the system performance by a factor of 3 when compared to clock gating, improving hardware robustness to soft errors related to power integrity phenomena.

# FPGA Implementation of the Greedy Perimeter Stateless Routing for WSNs

Georgios-Grigorios Mplemenos
Department of Electronic
and Computer Engineering
Technical University of Crete
Chania, Crete, Greece
Email: mplemenos@mhl.tuc.gr

Ioannis Papaefstathiou
Department of Electronic
and Computer Engineering
Technical University of Crete
Chania, Crete, Greece
Email: ygp@mhl.tuc.gr

## Abstract

One of the most rapidly expanding areas, nowadays, in networking systems is the Wireless Sensor Network (WSN) one. Typically WSNs rely on multi-hop routing protocols which must be able to establish communication among nodes and guarantee packet deliveries. In this paper, we present a novel approach for the implementation of the WSN routing protocols, which takes advantage of modern FPGAs in order to provide faster routing decisions while consuming significantly less energy than existing systems. Despite our focus on a particular routing protocol (GPSR), the platform developed has the additional advantage that due to the reconfigurability feature of the FPGA it can efficiently execute different routing protocols based on the requirements of the different WSN applications. As our real world experiments demonstrate, we accelerated the execution of the most widely used WSN routing protocol (GPSR) by at least 31 times when compared to the speed achieved when the exact same protocol is executed on a low power Intel Atom processor. More importantly by utilizing a high-end FPGA the overall energy consumption was reduced by more than 90%.

# A Multi-level Reconfigurable Architecture for a Wireless Sensor Node Coprocessing Unit

François Philipp and Manfred Glesner
Microelectronic Systems Research Group
Technische Universität Darmstadt
Merckstra$\beta$e 25, 64283 Darmstadt, Germany
{francoisp,glesner}(at)mes.tu-darmstadt.de

## Abstract

An architecture for a dynamically reconfigurable coprocessing unit for a wireless sensor node is presented in this paper. The hardware accelerator is tailored for the processing of sensor data streams that cannot be efficiently handled by low-power microcontrollers. High energy-efficiency is obtained by using reconfigurable computing mechanisms. The details of the architecture including multiple, specialized, reconfigurable processing stages are introduced. For a fast and efficient dynamic reconfiguration, the size of the reconfiguration data is minimized by using multiple reconfiguration levels. Reconfigurable function units are reused for the implementation of different algorithms sharing arithmetical properties. Thus, only a few resources are required for reconfiguration, limiting the power consumption of the node while its computation performance is increased with application-specific accelerators. An application example illustrates the utilization of the processing unit for the acquisition of accelerometer data.

# On an Hybrid and General Control Scheme for Algorithms Represented as a Polytope

Roberto Pérez-Andrade and César Torres-Huitzil
Information Technology Laboratory
Advanced Studies Center of the National
Polytechnic Institute, CINVESTAV
Ciudad Victoria, Tamaulipas, Mexico
email: {jrperez, ctorres}@tamps.cinvestav.mx

René Cumplido and Juan M. Campos
Department of Computer Science
National Institute for Astrophysics,
Optics and Electronics, INAOE
Santa Maria Tonantzintla, Puebla, Mexico
email: rcumplido, jcampos@ccc.inaoep.mx

## Abstract

This paper presents a general and hybrid (centralized and distributed) approach for the activation of processing elements (PEs) inside of a processor array using the polytope model. The proposed approach is suitable of being implemented on reconfigurable systems since by changing some mathematical expressions, the proposed control approach is able to provide activation patterns for different algorithms based on the polytope model. We have taken the Cholesky decomposition as example for developing our hybrid control towards a generalization of this scheme.

## Bitonic Sorting on Dynamically Reconfigurable Architectures

J. Angermeier, E. Sibirko, R. Wanka, and J. Teich
Hardware/Software Co-Design, Department of Computer Science
University of Erlangen-Nuremberg, Germany

### Abstract

Sorting is one of the most investigated tasks computers are used for. Up to now, not much research has been put into increasing the flexibility and performance of sorting applications by applying *reconfigurable* computer systems. There are parallel sorting algorithms (*sorting circuits*) which are highly suitable for VLSI hardware realization and which outperform sequential sorting methods applied on traditional software processors by far. But usually they require a large area that increases with the number of keys to be sorted. This drawback concerns ASIC and statically reconfigurable systems.

In this paper, we present a way to adopt the well-known Bitonic sorting method to dynamically reconfigurable systems such that this drawback is overcome. We present a detailed description of the design and actual implementation, and we present experimental results of our approach to show its benefits in performance and the trade-offs of our approach.

## A FPGA-based Accelerator to Speed-up Matrix Multiplication of Floating Point Operations

Holanda, B.;Pimentel, R.;Barbosa,J.; Camarotti, R.
Informatics Center - CIn
Federal University of Pernambuco - UFPE
Recife - Pernambuco - Brazil
{ bhtcd, rwpa, jpfb, rcfr }@cin.ufpe.br
Silva-Filho,A.;João,L.;Souza,V.;Ferraz,J.;Lima,M.
Informatics Center - CIn
Federal University of Pernambuco - UFPE
Recife - Pernambuco - Brazil
{ agsf, jcblc, vlss, jdf2, mel }@cin.ufpe.br

### Abstract

Field Programmable Gate Arrays (FPGAs) are able to provide a high computational parallelism that can be exploited to achieve high performance improvements in intensive data processing problems. In this paper our efforts were directed towards developing a PC cluster based on nodes that use FPGAs as co-processors. The target application is a floating-point large dense matrix multiplication. Experimental results for just one node of the cluster, consisting of a Xilinx Virtex 5 VLX50T with a PCI interface, showed performance improvements compared with the Intel Core2 Quad at 2.66 GHz, achieving a speed-up of 1.19 times. Other analyses in terms of frequency variation and power dissipation have been made by considering different matrix sizes running in one node of the cluster. Recently, the platform has been updated for a powerful Gidel plaftorm, the PROCe III 260E. This new platform consists of 1 FPGA Stratix III per board. In this board, it is possible to allocate up to 40 MACs per FPGA, reaching an overall speed-up of approximately 11.2 per node of the cluster when compared with the same general-purpose processor. A full example is presented in this paper.

# Compression Based Efficient and Agile Configuration Mechanism for Coarse Grained Reconfigurable Architectures

Syed. M. A. H. Jafri*†§ , Ahmed Hemani*, Kolin Paul*‡ , Juha Plosila†§ , and Hannu Tenhunen*†§
Email: jafri@kth.se, hemani@kth.se, kolin@gmail.com, juplos@utu.fi, hannu@kth.se
*Royal Institute of Technology, Sweden; †University of Turku, Finland;
‡ Indian Institute of Technology Delhi, India; § Turku Centre for Computer Science, Finland

## Abstract

This paper considers the possibility of speeding up the configuration by reducing the size of configware in coarsegrained reconfigurable architectures (CGRAs). Our goal was to reduce the number of cycles and increase the configuration bandwidth. The proposed technique relies on multicasting and bitstream compression. The multicasting reduces the cycles by configuring the components performing identical functions simultaneously, in a single cycle, while the bitstream compression increases the configuration bandwidth. We have chosen the dynamically reconfigurable resource array (DRRA) architecture as a vehicle to study the efficiency of this approach. In our proposed method, the configuration bitstream is compressed offline and stored in a memory. If reconfiguration is required, the compressed bitstream is decompressed using an online decompresser and sent to DRRA. Simulation results using practical applications showed upto 78% and 22% decrease in configuration cycles for completely parallel and completely serial implementations, respectively. Synthesis results have confirmed nigligible overhead in terms of area (1.2 %) and timing.

# FeatureVerilog: Extending Verilog to support Feature-Oriented Programming

YE Jun, TAN Qingping, LI Tun, CAO Guorong
School of Computer Science
National University of Defense Technology
Changsha, China
yejungolden82@gmail.com

## Abstract

Nowadays, systems are getting harder and harder to develop and maintain. For a long time, researchers have tried to solve these problems by increasing the level of abstraction, but RTL description languages (e.g. Verilog) are still being widely used. This paper proposes FeatureVerilog, a novel hardware description language that extends Verilog to support Feature-Oriented Programming (FOP). FeatureVerilog does not avoid the detailed description of RTL, but it can organize it in a more reasonable way. We have implemented a prototype pre-compiler for FeatureVerilog and used it to re-develop the OpenRISC 1200 project. The comparison of our implementation and the original one shows that FeatureVerilog can eliminate the duplicate code in the latter implementation effectively.

# Workshop 3

# Workshop on Nature Inspired Distributed Computing

# NIDISC 2011

# Adaptive Neighborhoods for Cellular Genetic Algorithms

Bernabé Dorronsoro
Interdisciplinary Centre for Security, Reliability, and Trust
University of Luxembourg
Luxembourg
Email: bernabe.dorronsoro@uni.lu

Pascal Bouvry
Faculty of Science, Technology and Communication
University of Luxembourg
Luxembourg
Email: pascal.bouvry@uni.lu

## Abstract

Cellular genetic algorithms (cGAs) are a kind of genetic algorithms (GAs) with decentralized population in which interactions among individuals are restricted to close ones. The use of decentralized populations in GAs allows to keep the population diversity for longer, usually resulting in a better exploration of the search space and, therefore in a better performance of the algorithm. However, the use of decentralized populations supposes the need of several new parameters that have a major impact on the behavior of the algorithm. In the case of cGAs, these parameters are the population and neighborhood shapes. Hence, we propose in this work two new adaptive techniques that allow removing the neighborhood to use from the algorithm's configuration. As a result, one of the new adaptive cGAs outperform the compared cGAs with fixed neighborhoods in the continuous and combinatorial domains.

# Hybrid MPI/OpenMP Strategy for Biological Multiple Sequence Alignment with DIALIGN-TX in Heterogeneous Multicore Clusters

Emerson de Araujo Macedo,
Alba Cristina Magalhaes Alves de Melo,
Gerson Henrique Pfitscher
Department of Computer Science
University of Brasilia (UnB)
Brasilia, Brazil
{emerson, albamm, gerson}@cic.unb.br

Azzedine Boukerche
School of Information and Technology Engineering
University of Ottawa
Ottawa, Canada
boukerch@site.uottawa.ca

## Abstract

Multiple Sequence Alignment (MSA) is a fundamental problem in Bioinformatics that aims to align more than two biological sequences in order to emphasize similarity regions. This problem is known to be NP-Complete, so heuristic methods are used to solve it. DIALIGN-TX is an iterative heuristic method for MSA that is based on dynamic programming and generates alignments by concatenating ungapped regions with high similarity. This paper proposes an MPI/OpenMP master/slave parallel strategy to run DIALIGN-TX in heterogeneous multicore clusters, with multiple allocation policies. The results obtained in a 28-core heterogeneous cluster with real sequence sets show that the execution time can be drastically reduced. Also, we show that an appropriate choice of the allocation policy and the master node has great impact on the overall system performance.

# Parallelization Strategies for Ant Colony Optimisation on GPUs

JoséM. Cecilia, José M. García
Computer Architecture
Department
University of Murcia
30100 Murcia, Spain
Email: chema, jmgarcia@ditec.um.es

Manuel Ujaldón
Computer Architecture
Department
University of Malaga
29071 Málaga, Spain
Email: ujaldon@uma.es

Andy Nisbet, Martyn Amos
Novel Computation Group
Division of Computing and IS
Manchester Metropolitan University
Manchester M1 5GD, UK
Email: a.nisbet,m.amos@mmu.ac.uk

## Abstract

Ant Colony Optimisation (ACO) is an effective population-based meta-heuristic for the solution of a wide variety of problems. As a population-based algorithm, its computation is intrinsically massively parallel, and it is therefore theoretically well-suited for implementation on Graphics Processing Units (GPUs). The ACO algorithm comprises two main stages: *Tour construction* and *Pheromone update*. The former has been previously implemented on the GPU, using a task-based parallelism approach. However, up until now, the latter has always been implemented on the CPU. In this paper, we discuss several parallelisation strategies for *both* stages of the ACO algorithm on the GPU. We propose an alternative *data-based* parallelism scheme for *Tour construction*, which fits better on the GPU architecture. We also describe novel GPU programming strategies for the *Pheromone update* stage. Our results show a total speed-up exceeding 28x for the *Tour construction* stage, and 20x for *Pheromone update*, and suggest that ACO is a potentially fruitful area for future research in the GPU domain.

# Multiple Particle Collision Algorithm applied to Radiative Transference and Pollutant Localization Inverse Problems

Eduardo Fávero Pacheco da Luz
Post-graduation program in Applied Computing
National Institute for Space Research
São José dos Campos/SP - Brazil
eduardo.luz@lac.inpe.br
José Carlos Becceneri, Haroldo Fraga de Campos Velho
Computing and Applied Mathematics Associated Laboratory
National Institute for Space Research
São José dos Campos/SP - Brazil
{becce, haroldo}@lac.inpe.br

## Abstract

The Multiple Particle Collision Algorithm (MPCA) is a nature-inspired stochastic optimization method developed specially for high performance computational environments. Its advantages resides in the intense use of computational power provided by multiple processors in the task of search the solution space for a near optimum solution. This work presents the application of MPCA in solving two inverse problems written as optimization problems, its advantages and disadvantages are also described, so are the obtained results.

# Multi-Environmental Cooperative Parallel Metaheuristics for Solving Dynamic Optimization Problems

Mostepha Redouane Khouadjia[*][†], Briseida Sarasola[‡], Enrique Alba[‡], Laetitia Jourdan[*] and El-Ghazali Talbi[*][†]
[*]National Institute for Research in Computer Science and Control INRIA Lille Nord-Europe, France
[†]LIFL, UMR CNRS 8022, Université de Lille-1, Villeneuve d'Ascq, France
Email: {mostepha-redouane.khouadjia, laetitia.jourdan}@inria.fr, talbi@lifl.fr
[‡]E.T.S.I. Informática, Universidad de Málaga, Málaga, Spain
Email: {briseida, eat}@lcc.uma.es

## Abstract

Dynamic optimization problems are problems in which changes occur over time. These changes could be related to the optimization objective, the problem instance, or involve problem constraints. In most cases, they are seen as an ordered sequence of sub-problems or environments, that must be solved during a certain time interval. The usual approaches tend to solve each sub-problem when a change happens, dealing always with one single environment at each time instant. In this paper, we propose a multi-environmental cooperative model for parallel metaheuristics to tackle Dynamic Optimization Problems. It consists in dealing with different environments at the same time, using different algorithms that exchange information coming from these environments. A parallel multi-swarm approach is presented for solving the Dynamic Vehicle Routing Problem. The effectiveness of the proposed approach is tested on a well-known set of benchmarks, and compared with other metaheuristics from the literature. Experimental results show that our multi-environmental approach outperforms conventional metaheuristics on this problem.

## Ant Colony Optimization for Route Planning Using Link Travel Time Predictions

Rutger Claes
DistriNet Labs
Katholieke Universiteit Leuven
Leuven, Belgium
rutger.claes@cs.kuleuven.be

Tom Holvoet
DistriNet Labs
Katholieke Universiteit Leuven
Leuven, Belgium
tom.holvoet@cs.kuleuven.be

### Abstract

Finding the shortest path in a road network is a well known problem. Various proven static algorithms such as Dijkstra and A* are extensively evaluated and implemented. When confronted with dynamic costs, such as link travel time predictions, alternative route planning algorithms have to be applied. This paper applies Ant Colony Optimization combined with link travel time predictions to find routes that reduce the time spend by travels by taking into account link travel time predictions. The presented algorithm is evaluated using simulations in real world traffic networks.

## Nature-Inspired Evaluation of Data Types for Trust Management in MANETs

Marcin Seredynski
Interdisciplinary Centre for Security, Reliability and Trust
University of Luxembourg
Luxembourg, Luxembourg
marcin.seredynski@uni.lu

Pascal Bouvry
Faculty of Sciences, Technology and Communication
University of Luxembourg
Luxembourg, Luxembourg
pascal.bouvry@uni.lu

### Abstract

In mobile ad hoc networks cooperation in packet forwarding is usually based upon the mixture of direct and indirect reciprocity. Nodes evaluate the level of cooperation of other network participants using a trust system. Its performance depends on classes of data used for description of forwarding behaviour of nodes. This paper studies the influence of two particular classes of data referred to as personal and general. A dynamic Nature-inspired model based on replicator equation is used to study the influence of these classes on the performance of the nodes. Computational experiments demonstrate that the distinction between personal and general data is not necessary if node's priority is the forwarding efficiency or battery conservation. However, if network user wants to maximise his throughput or to create the best possible protection against selfish nodes, than prioritising personal data improves his performance.

# Use of Meta-Heuristics for Design of Fuel Loading Pattern in light water Reactors comprising some Radial and Axial Heterogeneities

Do Jean-Michel, Lautard Jean-Jacques, Baudron
Anne-Marie, Douce Siegfried
DEN/DM2S/SERMA/
CEA Saclay
91191, France
e-mail: jean-michel.do@cea.fr

Arnaud Gilles
DEN/DM2S/SFME/
CEA Saclay
91191, France
e-mail: gilles.arnaud@cea.fr

## Abstract

The third generation nuclear core should be attractive from the nuclear fuel cycle management aspect. The main goal is to achieve a good conversion of fertile isotopes into fissile isotopes while respecting safety constraints. Preliminary studies have shown the interest of the core loaded with fertile and heterogeneous fissile assemblies. The purpose of this work is to find out an optimized loading pattern and axial repartition of fissile and fertile elements in regards to the following criteria: power peak and void coefficient minimization. Due to the huge number of possible combinations, we have replaced the global optimization by two simple forms: first optimization of the fertile / fissile repartitions on an assembly in 1D representation, and secondly optimization of the core's loading pattern in 3D representation. The evaluations of the criteria have been done using the neutron physics transport solver MINOS integrated in the new lattice-core APOLLO3 code. The research algorithms used in this study are of an evolutionary algorithm, a "Max Min ant system colony algorithm" and a particle swarm adapted to our multi criteria approach. These algorithms are distributed using the heterogeneous island's method. A second level of parallelism has been introduced inside each island. Therefore we could increase the number of evaluations while respecting the CPU time limitation of the batch queues available. There are several solutions that stand out from the simulations. Some of them are completely original and have proved to be relevant a posteriori. This method could then be used as a decision support tool for the exploitation reactors and also for the design of new reactors.

# A Shared-Memory ACO-Based Algorithm for Numerical Optimization

Peter Korošec, Jurij Šilc
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
{peter.korosec, jurij.silc}@ijs.si

Marian Vajteršic, Rade Kutil
Department of Scientific Computing
University of Salzburg
Salzburg, Austria
{marian, rkutil}@cosy.sbg.ac.at

## Abstract

Numerical optimization techniques are applied to a variety of engineering problems. The objective function evaluation is an important part of the numerical optimization and is usually realized as a black-box simulator. For efficient solving the numerical optimization problem, new shared-memory approach is proposed. The algorithm is based on an ACO meta-heuristics, where indirect coordination between ants drives the search procedure towards the optimal solution. Indirect coordination offers a high degree of parallelism and therefore relatively straightforward shared-memory implementation. For the communication between processors, the Intel-OpenMP library is used. It is shown that speed-up strongly depends on the simulation time. Therefore, algorithm's performance, according to simulator's time complexity, is experimentally evaluated and discussed.

# A Genetic Algorithm with a Penalty Function in the Selective Travelling Salesman Problem on a Road Network

Anna Piwonska
Bialystok University of Technology
Faculty of Computer Science
Wiejska 45A, 15-351 Biaystok, Poland
Email: a.piwonska@pb.edu.pl

Franciszek Seredynski
Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland
and
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
Email: sered@ipipan.waw.pl

## Abstract

The Selective Travelling Salesman Problem (STSP) is a version of the Travelling Salesman Problem (TSP) where it is not necessary to visit all vertices. Instead of it, with each vertex a number meaning a profit is associated. The problem is to find a cycle which maximizes collected profit but does not exceed a given cost constraint. A direct application of the STSP, e.g. in Intelligent Transportation Systems, is finding an optimal tour in road networks. However, while the classic STSP is defined on a complete graph, a road network is in general not complete and often has a rather sparse edge set. This paper presents the STSP defined on a road network (R-STSP). Since R-STSP is NP-hard and stands the problem with a constraint, the genetic algorithm (GA) with a penalty function is proposed. Computer experiments performed on the real road network in Poland have shown that this GA outperforms the GA searching only the feasible solution space.

# An energy-efficient kernel framework for large-scale data modeling and classification

Paul D. Yoo*, Jason W.P. Ng
Dept. Computer Engineering, Khalifa University of
Science, Technology and Research (KUSTAR)
Etisalat British Telecom (BT) Innovation Centre
Abu Dhabi, P.O Box 127788, U.A.E
{paul.yoo, jason.ng}@kustar.ac.ae
Albert Y. Zomaya
Centre for Distributed and High Performance
Computing, School of Information Technologies (J12),
University of Sydney, NSW 2006, Australia
albert.zomaya@sydney.edu.au

## Abstract

Energy-efficient computing has now become a key challenge not only for data-center operations, but also for many other energy-driven systems, with the focus on reducing of all energy-related costs, and operational expenses, as well as its corresponding and environmental impacts. Intelligent machine-learning systems are typically performance driven. For instance, most non-parametric model-free approaches are often known to require high computational cost in order to find the global optima. Designing more accurate machine-learning systems to satisfy the market needs will hence lead to a higher likelihood of energy waste due to the increased computational cost. This paper thus introduces an energy-efficient framework for large-scale data modeling and classification. It can achieve a test error comparable to or better than the state-of-the-art machine-learning models, while at the same time, maintaining a low computational cost when dealing with large-scale data. The effectiveness of the proposed approaches has been demonstrated by our experiments with two large-scale KDD datasets: Mtv-1 and Mtv-2.

# On the Resilience of [distributed] EAs against Cheaters in Global Computing Platforms

Sebastien Varrette* , Emilia Tantar* and Pascal Bouvry*
* Computer Science and Communication (CSC) Research Unit
University of Luxembourg
16, rue Richard Coudenhove-Kalergi
L-1359 Luxembourg, Luxembourg
Emails: {Firstname.Name@uni.lu}

## Abstract

GC platforms such as BOINC[BOINC04] are nowadays considered as the most powerful distributed computing systems worldwide. Based on volunteer computing and various forms of incentives, such architecture also attract cheaters who seek to obtain rewards with little or no contribution to the system. Cheating can be modelized as alteration of output values produced by some or all tasks of the program beeing executed. In complement of generic approaches typically based on task duplication, this article focus on ABFT[ABFT84] technics where the fault tolerance scheme is tailored to the algorithm performed to make it resilient to a limited number of faults i.e. falsified results produced by cheaters in this case. More precisely, the aspects of dEA against cheaters in environment is studied. Our main contribution consists in the formal analysis of the impact of cheating faults on this context, together with a theoretical proof of convergence towards valid solutions despite the presence of malicious acts. Whereas the resilience of against a simpler model of fault (i.e. crash faults) has been observed in the literature[Vega06, Hidalgo07, Morales08, Gonzales09], this study is the first to propose a theoretical proof of this behaviour, additionally against a more complex kind of fault. By the variety of problems addressed by EAs, this study will hopefully promote their usage in the future developments around GC platforms.

# Workshop 4

# Workshop on High Performance Computational Biology

# HiCOMB 2011

# Parallelizing BLAST and SOM algorithms with MapReduce-MPI library

Seung-Jin Sul, Andrey Tovchigrechko
J. Craig Venter Institute
Rockville, MD
atovtchi@jcvi.org

## Abstract

Most bioinformatics algorithms are developed in a serial form due to a fast pace of changes in the subject domain and the fact that many bioinformatics tasks can be parallelized as collections of serial jobs communicating at the file system level (High-Throughput Computing, HTC). Recently, a MapReduce-MPI library was made available by Sandia Lab to ease porting of a large class of serial ap-plications to the High Performance Computing (HPC) architectures dominating large federated resources such as NSF TeraGrid. Using this library, we have created two open-source bioinformatics applications. The first one addresses a problem of adapting existing com-plex and highly optimized serial bioinformatics algorithm to HPC architecture in a minimally invasive way. We built a parallel BLAST implementation that calls the high-level methods of unmodified NCBI C++ Toolkit. We demonstrated scaling for up to 1000 cores on TACC Ranger cluster when processing the sufficiently large input datasets. Using unmodified NCBI Toolkit ensures that the results are compatible across the multitude of settings in the original serial algorithm, and that future versions of the upstream code can be easily integrated. The second application is a Self-Organizing Map (SOM) machine-learning algorithm, popular in bioinformatics applications such as metagenomic binning. The nature of the SOM requires a global synchronization step with a frequency that necessitates the use of an HPC environment. Our implementation of the "batch SOM" uses a mix of MapReduce-MPI and direct MPI calls and scales to 1000 cores as well. This allows easy processing of datasets with a size that is out of range of the serial SOM implementations. Both im-plementations are available in the open source at http://andreyto.github.com/mgtaxa/.

# Algorithms for a parallel implementation of Hidden Markov Models with a small state space

Jesper Nielsen, Andreas Sand
Bioinformatics Research Centre
Aarhus University
Aarhus, Denmark
Email: {jn,asand}@birc.au.dk

## Abstract

Two of the most important algorithms for Hidden Markov Models are the forward and the Viterbi algorithms. We show how formulating these using linear algebra naturally lends itself to parallelization. Although the obtained algorithms are slow for Hidden Markov Models with large state spaces, they require very little communication between processors, and are fast in practice on models with a small state space.

We have tested our implementation against two other implementations on artificial data and observe a speed-up of roughly a factor of 5 for the forward algorithm and more than 6 for the Viterbi algorithm. We also tested our algorithm in the Coalescent Hidden Markov Model framework, where it gave a significant speed-up.

# Improving CUDASW++, a Parallelization of Smith-Waterman for CUDA Enabled Devices

Doug Hains, Zach Cashero, Mark Ottenberg,
Wim Bohm and Sanjay Rajopadhye
Colorado State University
Department of Computer Science

## Abstract

CUDASW++ is a parallelization of the Smith-Waterman algorithm for CUDA graphical processing units that computes the similarity scores of a query sequence paired with each sequence in a database. The algorithm uses one of two kernel functions to compute the score between a given pair of sequences: the inter-task kernel or the intra-task kernel. We have identified the intra-task kernel as a major bottleneck in the CUDASW++ algorithm. We have developed a new intra-task kernel that is faster than the original intra-task kernel used in CUDASW++. We describe the development of our kernel as a series of incremental changes that provide insight into a number of issues that must be considered when developing any algorithm for the CUDA architecture. We analyze the performance of our kernel compared to the original and show that the use of our intra-task kernel substantially improves the overall performance of CUDASW++ on the order of three to four giga-cell updates per second on various benchmark databases.

# An ultrafast scalable many-core motif discovery algorithm for multiple GPUs

Yongchao Liu, Bertil Schmidt, Douglas L. Maskell
School of Computer Engineering
Nanyang Technological University
Singapore
e-mail: {liuy0039, asbschmidt, asdouglas}@ntu.edu.sg

## Abstract

The identification of genome-wide transcription factor binding sites is a fundamental and crucial problem to fully understand the transcriptional regulatory processes. However, the high computational cost of many motif discovery algorithms heavily constraints their application for large-scale datasets. The rapid growth of genomic sequences and gene transcription data further deteriorates the situation and establishes a strong requirement for time-efficient scalable motif discovery algorithms. The emergence of many-core architectures, typically CUDA-enabled GPUs, provides an opportunity to reduce the execution time by an order of magnitude without the loss of accuracy. In this paper, we present mCUDA-MEME, an ultrafast scalable many-core motif discovery algorithm for multiple GPUs based on the MEME algorithm. Our algorithm is implemented using a hybrid combination of the CUDA, OpenMP and MPI parallel programming models in order to harness the powerful compute capability of modern GPU clusters. At present, our algorithm supports OOPS and ZOOPS models, which are sufficient for most motif discovery applications. mCUDAMEME achieves significant speedups for the starting point search stage (and the overall execution) when benchmarked, using real datasets, against parallel MEME running on 32 CPU cores. Speedups of up to 1.4 (1.1) on a single GPU of a Fermi-based Tesla S2050 quad-GPU computing system and up to 10.8 (8.3) on the eight GPUs of a two Tesla S2050 system were observed. Furthermore, our algorithm shows good scalability with respect to dataset size and the number of GPUs (availability:https://sites.google.com/site/yongchaosoftware/mc uda-meme).

# Parallel Mapping Approaches for GNUMAP

Nathan L. Clement*, Mark J. Clement†, Quinn Snell† and W. Evan Johnson*§
*Department of Computer Science
University of Texas at Austin, Austin, TX 78712
Email: nathanlclement@gmail.com
†Department of Computer Science
Brigham Young University, Provo, UT 84602
‡Department of Statistics
Brigham Young University, Provo, UT, 84602
§Department of Oncological Sciences, Huntsman Cancer Institute
University of Utah, Salt Lake City, UT, 84105

## Abstract

Mapping short next-generation reads to reference genomes is an important element in SNP calling and expression studies. A major limitation to large-scale whole-genome mapping is the large memory requirements for the algorithm and the long run-time necessary for accurate studies. Several parallel implementations have been performed to distribute memory on different processors and to equally share the processing requirements. These approaches are compared with respect to their memory footprint, load balancing, and accuracy. When using MPI with multi-threading, linear speedup can be achieved for up to 256 processors.

# Computing the Phylogenetic Likelihood Function Out-of-Core

Fernando Izquierdo-Carrasco and Alexandros Stamatakis
The Exelixis Lab, Scientific Computing Group
Heidelberg Institute for Theoretical Studies
D-69118 Heidelberg, Germany
{Fernando.Izquierdo-Alexandros.Stamatakis}@h-its.org

## Abstract

The computation of the phylogenetic likelihood function for reconstructing evolutionary trees from molecular sequence data is both memory- and compute-intensive. Based on our experience with the user community of RAxML, memory-shortages (as opposed to CPU time limitations) are currently the prevalent problem regarding resource availability, that is, lack of memory hinders large-scale biological analyses. To this end, we study the performance of an out-of-core execution of the phylogenetic likelihood function by means of a proof-of-concept implementation in RAxML. We find that RAM miss rates are below 10%, even if only 5% of the required data structures are held in RAM. Moreover, we show that our proof-of-concept implementation runs more than 5 times faster than the respective standard implementation when paging is used. The concepts presented here can be applied to all programs that rely on the phylogenetic likelihood function and can contribute significantly to enabling the computation of whole-genome phylogenies.

# Efficient Nonserial Polyadic Dynamic Programming on the Cell Processor

Li Liu, Mu Wang, Jinlei Jiang, Ruizhe Li, and Guangwen Yang
Computer Science and Technology Department, Tsinghua University, Beijing, 100084 China
liuli03@mails.tsinghua.edu.cn; {jjlei, ygw}@tsinghua.edu.cn

## Abstract

Dynamic programming (DP) is an effective technique for many search and optimization problems. However, the high arithmetic complexity limits its extensive use. Although modern processor architectures with multiple cores and SIMD (single instruction multiple data) instructions provide increasingly high computing power, even the state-of-the-art fully optimized algorithm still largely underutilizes modern multi-core processors. In this paper we propose to improve one family of DP, nonserial polyadic DP (NPDP), targeting a heterogeneous multi-core architecture, the Cell Broadband Engine. We first design a new data layout which efficiently utilizes the on-chip memory system of the Cell processor. Next we devise a CellNPDP algorithm with two tiers. The first tier is a SPE (a co-processor on the Cell processor) procedure which efficiently computes a block of data that can fit into one SPEs local store. The second tier is a parallel procedure which enables all SPEs to efficiently compute all blocks of data. To evaluate CellNPDP, we use both performance modeling and experiments. The performance model reveals that the processor utilization of NPDP can be independent of the problem size. To empirically evaluate CellNPDP, we use two platforms: the IBM QS20 dual-Cell blade and a CPU platform with two latest quad-core CPUs. On both platforms, the processor utilization of CellNPDP is larger than 60%, which demonstrates that our optimizations and CellNPDP can be architecture-independent. Compared to the state-of-the-art fully optimized algorithm on the CPU platform, CellNPDP is 44-fold faster for single-precision and 28-fold faster for double-precision, which is a significant improvement to NPDP.

# FENZI: GPU-enabled Molecular Dynamics Simulations of Large Membrane Regions based on the CHARMM force field and PME

Narayan Ganesan, Michela Taufer
Dept. of Computer & Inf. Sciences
University of Delaware
Email: {ganesan, taufer}@udel.edu

Brad Bauer, Sandeep Patel
Dept. of Chemistry and Biochemistry
University of Delaware
Email: {babauer, patel}@udel.edu

## Abstract

When studying membrane-bound protein receptors, it is necessary to move beyond the current state-of-the-art simulations that only consider small membrane patches and implicit solvent. Limits of traditional computer platforms negatively impact the model's level of realism and the computational scales achievable. On the other hand, multi-core platforms such as GPUs offer the possibility to span length scales in membrane simulations much larger and with higher resolutions than before.

To this end, this paper presents the design and implementation of an advanced GPU algorithm for Molecular Dynamics (MD) simulations of large membrane regions in the NVT, NVE, and NPT ensembles using explicit solvent and Particle Mesh Ewald (PME) method for treating the conditionally-convergent electrostatic component of the classical force field. A key component of our algorithm is the redesign of the traditional PME method to better fit on the multithreading GPU architecture. This has been considered a fundamentally hard problem in the molecular dynamics community working on massively multithreaded architecture. Our algorithm is integrated in the code FENZI (*yun dong de FEN ZI* in Mandarin or *moving molecules* in English). The paper analyzes both the performance and accuracy of large-scale GPU-enabled simulations of membranes using FENZI, showing how our code can enable multi-nanosecond MD simulations per day, even when using PME.

# Divide-and-conquer approach to the parallel computation of elementary flux modes in metabolic networks

Dimitrije Jevremovic, Daniel Boley
Department of Computer Science & Engineering
University of Minnesota
Minneapolis, MN, USA
Email: jevrem@cs.umn.edu, boley@cs.umn.edu
Carlos P Sosa
IBM, and
Biomed. Infor. & Comp'l. Biology, Univ. of Minnesota
Rochester, MN, USA
Email: cpsosa@us.ibm.com

## Abstract

Elementary flux modes are an important class of metabolic pathways used to characterize the functioning and behavior of metabolic networks of biochemical reactions in a biological cell. The computation of the elementary flux modes is accomplished by using the so-called Nullspace Algorithm whose high computational cost and memory requirements still limit the computation to relatively small metabolic networks. We combine a "combinatorial" parallelization with a novel divide-and-conquer paradigm into a new implementation of the Nullspace Algorithm with lower memory requirements. We discuss the disadvantages of the combinatorial parallelization and divide-and-conquer ideas and explain why their combination attains more computational power. The improved parallel Nullspace Algorithm is used to compute up to nearly 50 million elementary flux modes for a metabolic network for yeast, a task which was previously not possible using either of the two approaches individually.

# Sensitive and Specific Identification of Protein Complexes in "Perturbed" Protein Interaction Networks from Noisy Pull-Down Data

William Hendrix
North Carolina State University
Raleigh, NC 27695

Tatiana Karpinets
Oak Ridge National Laboratory
Oak Ridge, TN 37831

Byung-Hoon Park
Oak Ridge National Laboratory
Oak Ridge, TN 37831

Eric Schendel
North Carolina State University
Raleigh, NC 27695

Alok Choudhary
Northwestern University
Evanston, IL 60208

Nagiza F. Samatova
North Carolina State University, Raleigh, NC 27695
Oak Ridge National Laboratory, Oak Ridge, TN 37831
Corresponding author: samatovan@ornl.gov

## Abstract

High-throughput mass-spectrometry technology has enabled genome-scale discovery of protein-protein interactions. Yet, computational inference of protein interaction networks and their functional modules from large-scale pull-down data is challenging. Over-expressed or "sticky" bait is not specific; it generates numerous false positives. This "curse" of the technique is also its "blessing"—the sticky bait can pull-down interacting components of other complexes, thus increase sensitivity. Finding optimal trade-offs between coverage and accuracy requires tuning multiple "knobs," i.e., method parameters. Each selection leads to a putative network, where each network in the set of "perturbed" networks differs from the others by a few added or removed edges. Identification of functional modules in such networks is often based on graph-theoretical methods such as maximal clique enumeration. Due to the NP-hard nature of the latter, the number of tunings to explore is limited. This paper presents an efficient iterative framework for sensitive and specific detection of protein complexes from noisy protein interaction data.

# An Efficient and Scalable Implementation of SNP-pair Interaction Testing for Genetic Association Studies

Lars Koesterke*, Dan Stanzione, Matt Vaughn
Texas Advanced Computing Center
The University of Texas
Austin, Texas, USA
*e-mail: lars@tacc.utexas.edu

Stephen M. Welch*, Waclaw Kusnierczyk
Department of Agronomy
Kansas State University
Manhattan, Kansas, USA
*e-mail: welchsm@ksu.edu

Jinliang Yang, Cheng-Ting Yeh, Dan Nettleton, Patrick S. Schnable*
Iowa State University
Ames, Iowa, USA
*email: Schnable@iastate.edu

## Abstract

This paper describes a scalable approach to one of the most computationally intensive problems in molecular plant breeding, that of associating quantitative traits with genetic markers. The fundamental problem is to build statistical correlations between particular loci in the genome of an individual plant and the expressed characteristics of that individual. While applied to plants in this paper, the problem generalizes to mapping genotypes to phenotypes across all biology. In this work, a formulation of a statistical approach for identifying pairwise interactions is presented. The implementation, optimization and parallelization of this approach are then presented, with scalability results.

# Workshop 5

# Advances in Parallel and Distributed Computing Models

# APDCM 2011

# Efficient Work-Stealing Strategies for Fine-Grain Task Parallelism

Adnan
Graduate School of Systems And Information Engineering
University Of Tsukuba
Tsukuba, Japan
E-mail: adnantoo@hpcs.cs.tsukuba.ac.jp

Mitsuhisa Sato
Center for Computational Sciences
University of Tsukuba
Tsukuba, Japan
E-mail: msato@cs.tsukuba.ac.jp

## Abstract

Herein, we describe extended work-stealing strategies for StackThreads/MP, in which thieves steal from the bottom of a victim's logical stack not just the bottommost task but multiple chained tasks. These new strategies offer two advantages: reducing the total cost of stealing tasks and reducing the total idle time. In addition, these strategies attempt to preserve the sequential execution order of tasks in the chain. We evaluated these extended work-stealing strategies by using the unbalanced tree search (UTS) benchmark and could demonstrate its advantages over the original work-stealing strategy and other OpenMP task implementations and Cilk implementation as well. Extended work-stealing strategies exhibit significant improvement with respect to the UTS benchmark, even if the task is very fine-grain and non-uniform.

# A Packet Aggregation Mechanism For Real Time Applications Over Wireless Networks

Paulo H. Azevêdo[1] , Marcos F. Caetano[2] and Jacir L. Bordim[1]
[1]Department of Computer Science
[2]Department of Electrical Engineering
University of Brasilia (UnB), 70910-900, Brasilia, Brazil
E-mail: paulo.a.filho@aiesec.net, {caetano|bordim}@cic.unb.br

## Abstract

This work presents a packet aggregation mechanism tailored for real time applications over wireless network environments. In the core of the proposed mechanism lies an elaborated packet holding time estimation. This estimation is used to allow for packet aggregation along a multihop path. The proposed mechanism was evaluated and compared with another prominent packet aggregation scheme. The simulation results have shown that the proposed mechanism is capable to keep jitter and total delay within application bounds. Furthermore, the proposed scheme allows for substantial reduction on the number of packet transmissions as well as the overall packet overhead. The savings in terms of packet transmissions reached nearly 80% in the evaluated scenarios. These results have shown that the proposed scheme is able to cope with varying network link capacity and strict application timing requirements.

# An analysis of memory bandwidth limited scenarios for NUMA and GPU system

Jens Breitbart
Research Group Programming Languages / Methodologies
Universität Kassel, Germany
Email:jbreitbart@uni-kassel.de

## Abstract

The processing power and parallelism in hardware is expected to increase rapidly over the next years, whereas memory bandwidth per flop and the amount of main memory per flop will be falling behind. These trends will result in both more algorithms to become limited by memory bandwidth, and overall memory requirements to become an important factor for algorithm design. In this paper we study the Gau-Seidel stencil as an example of a memory bandwidth limited algorithm. We consider GPUs and NUMA systems, which are both designed to provide high memory bandwidth at the cost of making algorithm design more complex. The mapping of the non-linear memory access pattern of the Gau-Seidel stencil to the different hardware is important to achieve high performance. We show that there is a trade-off between overall performance and memory requirements when optimizing for optimal memory access pattern. Vectorizing on the NUMA system and optimizing to utilize all processors on the GPU does not pay off in terms of performance per memory used, which we consider an important measurement regarding the trends named before.

# Dynamic Monopolies in Colored Tori

Sara Brunetti$^*$ , Elena Lodi$^*$ , Walter Quattrociocchi$^*$
$^*$University of Siena, Italy,
{lodi, sara.brunetti,walter.quattrociocchi}@unisi.it

## Abstract

The *information diffusion* has been modeled as the spread of an information within a group through a process of social influence, where the diffusion is driven by the so called *influential network*. Such a process, which has been intensively studied under the name of *viral marketing*, has the goal to select an initial good set of individuals that will promote a new idea (or message) by spreading the ”rumor” within the entire social network through the word-of-mouth. Several studies used the *linear threshold model* where the group is represented by a graph, nodes have two possible states (active, non-active), and the threshold triggering the adoption (activation) of a new idea to a node is given by the number of the active neighbors.

The problem of detecting in a graph the presence of the minimal number of nodes that will be able to activate the entire network is called *target set selection* (TSS). In this paper we extend TSS by allowing nodes to have more than two colors. The multicolored version of the TSS can be described as follows: let $G$ be a torus where every node is assigned a color from a finite set of colors. At each local time step, each node can recolor itself, depending on the local configurations, with the color held by the majority of its neighbors.

We study the initial distributions of colors leading the system to a monochromatic configuration of color $k$, focusing on the minimum number of initial $k$-colored nodes. We conclude the paper by providing the time complexity to achieve the monochromatic configuration.

# Distributed Construction of Nested Clusters with Inter-cluster Routing

Alain BUI
PRiSM (UMR CNRS 8144)
Université de Versailles
St-Quentin-en-Yvelines
alain.bui@prism.uvsq.fr

Simon CLAVIÈRE
PRiSM (UMR CNRS 8144)
Université de Versailles
St-Quentin-en-Yvelines
simon.claviere@prism.uvsq.fr

Devan SOHIER
PRiSM (UMR CNRS 8144)
Université de Versailles
St-Quentin-en-Yvelines
devan.sohier@prism.uvsq.fr

## Abstract

This paper presents an algorithm to compute a binary hierarchy of nested clusters. A token browses the network and recruits nodes to its cluster. When the cluster reaches a maximal size defined by a parameter $K$ of the algorithm, it is divided (when possible) and tokens are created in both of the new clusters. The new clusters are then built and divided in the same fashion.

To allow the division of clusters, a spanning tree is built for each cluster. At each division, routing information is stored. The browsing scheme chosen is a random walk, in order to ensure (local) load balancing.

The naming process used for the clusters, along with the information stored during each division, allows routing to any cluster. This routing scheme is based on the spanning trees of the clusters.

# A Model for Coherent Distributed Memory For Race Condition Detection

Franck Butelle Camille Coti
Franck.Butelle@lipn.univ-paris13.fr Camille.Coti@lipn.univ-paris13.fr
LIPN, CNRS-UMR7030, Université Paris 13, F-93430 Villetaneuse, France

## Abstract

We present a new model for distributed shared memory systems, based on remote data accesses. Such features are offered by network interface cards that allow one-sided operations, remote direct memory access and OS bypass. This model leads to new interpretations of distributed algorithms allowing us to propose an innovative detection technique of race conditions only based on logical clocks. Indeed, the presence of (data) races in a parallel program makes it hard to reason about and is usually considered as a bug.

# Using Memory Access Traces to Map Threads and Data on Hierarchical Multi-core Platforms

Eduardo Henrique Molina da Cruz,
Marco Antonio Zanata Alves,
Alexandre Carissimi,
Philippe Olivier Alexandre Navaux
PPGC Graduate Program in Computer Science
Institute of Informatics
UFRGS Federal University of Rio Grande do Sul
Porto Alegre, RS, Brazil
Email: {ehmcruz, mazalves, asc, navaux}@inf.ufrgs.br
Christiane Pousa Ribeiro,
Jean-Francois Méhaut
INRIA Mescal Research Team
LIG Laboratory
Grenoble University, France
Email: {Christiane.Pousa, Jean-Francois.Mehaut}@imag.fr

## Abstract

In parallel programs, the tasks of a given application must cooperate in order to accomplish the required computation. However, the communication time between the tasks may be different depending on which core they are executing and how the memory hierarchy and interconnection are used. The problem is even more important in multi-core machines with NUMA characteristics, since the remote access imposes high overhead, making them more sensitive to thread and data mapping. In this context, process mapping is a technique that provides performance gains by improving the use of resources such as interconnections, main memory and cache memory. The problem of detecting the best mapping is considered NP-Hard. Furthermore, in shared memory environments, there is an additional difficulty of finding the communication pattern, which is implicit and occurs through memory accesses. This work aims to provide a method for static mapping for NUMA architectures which does not require any prior knowledge of the application. Different metrics were adopted and an heuristic method based on the Edmonds matching algorithm was used to obtain the mapping. In order to evaluate our proposal, we use the NAS Parallel Benchmarks (NPB) and two modern multi-core NUMA machines. Results show performance gains of up to 75% compared to the native scheduler and memory allocator of the operating system.

# A Matching-Based Automata for Distributed Graph Algorithms

J. Paul Daigle and Sushil K. Prasad
Department of Computer Science
Georgia State University
Atlanta, Georgia 30303, USA

## Abstract

A consistent problem with distributed algorithms for graph problems is scalability. Here we present a scalable automata that can easily be modified to solve a number of problems. We focus on the Minimum Weighted Vertex Cover problem, showing that our approach improves on the best known algorithm, reducing the number of communication rounds from $O(\log n)$ to $O(\log \Delta)$. We then show how our framework can be modified to solve for Edge Coloring and Independent Set.

# GALS-based LPSP: Implementation of a Novel Architecture for Low Power High Performance Security Processors

Hala A. Farouk,
Computer Engineering,
Arab Academy for Science
and Technology,
Alexandria, Egypt.
Email: mail@halafarouk.info

Mahmoud T. El-Hadidi,
Electronics and
Electrical Communication,
Cairo University,
Giza, Cairo, Egypt.
Email: hadidi@eun.eg

Ahmed Abou El Farag,
Computer Engineering,
Arab Academy for Science
and Technology,
Alexandria, Egypt.
Email: abouelfarag@aast.edu

## Abstract

Current architectures for processors that run security applications are optimized for either high-performance or low energy consumption. We propose an implementation for an architecture that not only provides high performance and low energy consumption but also mitigates security attacks on the cryptographic algorithms which are running on it. The security is taken as a new dimension in the design process of this new processor architecture; the Globally-Asynchronous Locally-Synchronous-based Low Power Security Processor (GALS-based LPSP). GALS-based LPSP inherits the scheduling freedom and high performance from the dataflow architectures and the low energy consumption and flexibility from the GALS systems. In this paper a prototype of the GALS-based LPSP is implemented as a soft core on the Virtex-5 (xc5-vlx155t) FPGA. The architectural features that allow the processor to mitigate Side-Channel attacks are explained in detail and tested on the current encryption standard, the AES. The performance analysis reveals that the GALS-based LPSP achieves two times higher throughput with one and a half times less energy consumption than the currently used embedded processors.

# Self-Stabilizing Master-Slave Token Circulation and Efficient Topology Computation in a Tree of Arbitrary Size

Wayne Goddard and Pradip K Srimani
School of Computing
Clemson University
Clemson, SC 29634-0974
{goddard, srimani}@cs.clemson.edu

## Abstract

Self-stabilizing algorithms represent an extension of distributed algorithms in which nodes of the network have neither coordination, synchronization, nor initialization. We consider the model introduced by Lee et al. where there is one designated master node and all other nodes are anonymous and as simple as possible. We provide here a master-slave algorithm that achieves token circulation in a tree, and using that, enables the master node to compute the topology of the tree. We show that the complexity of the algorithm is at most 2n rounds O($n^2$) steps, and the space complexity at the slave nodes is almost constant in the sense that storage needed collectively at all n slave nodes is 2n bits.

# From Simulation to Experiment: A Case Study on Multiprocessor Task Scheduling

Sascha Hunold
CNRS / LIG Laboratory
Grenoble, France
sascha.hunold@imag.fr

Henri Casanova
Dept. of Information and Computer Sciences
University of Hawai'i at Mānoa, U.S.A.
henric@hawaii.edu

Frédéric Suter
IN2P3 Computing Center, CNRS/IN2P3
Lyon-Villeurbanne, France
Frederic.Suter@cc.in2p3.fr

## Abstract

Simulation is a popular approach for empirically evaluating the performance of algorithms and applications in the parallel computing domain. Most published works present results without quantifying simulation error. In this work we investigate accuracy issues when simulating the execution of parallel applications. This is a broad question, and we focus on a relevant case study: the evaluation of scheduling algorithms for executing mixed-parallel applications on clusters. Most such scheduling algorithms have been evaluated in simulation only. We compare simulations to real-world experiments in a view to identify which features of a simulator are most critical for simulation accuracy. Our first finding is that simple yet popular analytical simulation models lead to simulation results that cannot be used for soundly comparing scheduling algorithms. We then show that, by contrast, simulation models instantiated based on brute-force measurements of the target execution environment lead to usable results. Finally, we develop empirical simulation models that provide a reasonable compromise between the two previous approaches.

# Execution and Time Models for Pervasive Sensor Networks

Ajay D. Kshemkalyani, Ashfaq A. Khokhar, Min Shen
University of Illinois at Chicago
Chicago, IL 60607, USA
{ajay, ashfaq, mshen6}@uic.edu

## Abstract

Sensor-actuator networks and interactive ubiquitous environments are distributed systems in which the sensor-actuators communicate with each other by message-passing. This paper makes three contributions. First, it gives a general system and execution model for such sensor-actuator networks in pervasive environments. Second, it examines the range of time models that are useful for specifying properties, and for implementation, in such distributed networks, and places approaches and limitations in perspective. Third, it shows that although the partial order time model has not been seen to be useful as a specification tool in real applications of sensornets, yet, it is useful for real applications in pervasive sensornets because (under certain conditions) it can serve as a viable alternative to physically synchronized clocks that provide the linear order time model.

# Reducing Download Times in Peer-to-Peer File Sharing Systems with Stochastic Service Capacities

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

## Abstract

The main problem for an individual user peer in a peer-to-peer network with heterogeneous source peers is the peer selection problem, namely, switching among source peers and finally settling on one, while keeping the total time of probing and downloading to a minimum. There has been little investigation on selecting source peers with stochastic service capacities. The main contribution of this paper is to address the problem of reducing download times in peer-to-peer file sharing systems with stochastic service capacities. A precise analysis of the expected download time is given when the service capacity of a source peer is a random variable. A chunk-based switching and peer selection algorithm using the method of probing high-capacity peers is proposed and the expected download time of the algorithm is analyzed. Two subproblems of the optimal choice of the threshold of high-capacity source peers and the optimal order of probing are also solved. The performance of the algorithm is compared with the random chunk-based switching method. It is shown that noticeable performance improvement can be obtained.

# Exploring Weak Dependencies in DAG Scheduling for Exact Inference

Nam Ma
Computer Science Department
University of Southern California
Los Angeles, CA 90089
Email: namma@usc.edu

Yinglong Xia
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
Email: yxia@us.ibm.com

Viktor K. Prasanna
Ming Hsieh Department of
Electrical Engineering
University of Southern California
Los Angeles, CA 90089
Email: prasanna@usc.edu

## Abstract

Many computational solutions can be expressed as directed acyclic graphs (DAGs) with weighted nodes. In parallel computing, a fundamental challenge is to efficiently map computing resources to the tasks, while preserving the precedence constraints among the tasks. Traditionally, such constraints are preserved by starting a task after all its preceding tasks are completed. However, for a class of DAG structured computations, a task can be partially executed with respect to each preceding task. We define such relationship between the tasks as weak dependency. %A typical example is exact inference in junction trees, where a clique can be partially updated with respect to any one of its preceding cliques. In this paper, we adapt a traditional DAG scheduling scheme to exploit weak dependencies in a DAG. We perform experiments to study the impact of weak dependency based scheduling method on the execution time using a representative set of task graphs for exact inference in junction trees. For a class of task graphs, on a state-of-the-art general-purpose multicore system, the weak dependency based scheduler runs 4x faster than a baseline scheduler that is based on the traditional scheduling method.

# The Impact of Non-coherent Buffers on Lazy Hardware Transactional Memory Systems

Anurag Negi* Rubén Titos-Gil† Manuel E. Acacio† Jośe M. García† Per Stenstrom *
Universidad de Murcia †
Chalmers University of Technology *
{rtitos,meacacio,jmgarcia}@ditec.um.es
{negi,per.stenstrom}@chalmers.se

## Abstract

When supported in silicon, transactional memory (TM) promises to become a fast, simple and scalable parallel programming paradigm for future shared memory multiprocessor systems. Among the multitude of hardware TM design points and policies that have been studied so far, lazy conflict resolution designs often extract the most concurrency, but their inherent need for lazy versioning requires careful management of speculative updates. In this paper we study how coherent buffering, in private caches for example, as has been proposed in several hardware TM proposals, can lead to inefficiencies. We then show how such inefficiencies can be substantially mitigated by using complete or partial non-coherent buffering of speculative writes in dedicated structures or suitably adapted standard per-core write-buffers. These benefits are particularly noticeable in scenarios involving large coarse grained transactions that may write a lot of non-contended data in addition to actively shared data. We believe our analysis provides important insights into some overlooked aspects of TM behaviour and would prove useful to designers wishing to implement lazy TM schemes in hardware.

# Alphabets: An Extended Polyhedral Equational Language

S. Rajopadhye
Computer Science Department
Colorado State University
Sanjay.Rajopadhye@colostate.edu

G. Gupta, D-G. Kim
Apolent Corporation
Fort Collins CO
{gautam, daegonk}@apolent.com

## Abstract

The polyhedral model is a formalism for reasoning about an important class of compute- and data-intensive kernels in many programs. We extend the model to include (i) while loops, and (ii) nonaffine dependence functions, together with additional reduction-like operators like argmin and k-argmin. We propose a equational language Alphabets, that extends an earlier language Alpha. We prove its closure properties under program transformations, discuss its denotational semantics, and provide operational semantics in the form of a demand-driven code generator. Our work focuses on expressibility, and complements most previous efforts to extend the polyhedral model that address legality of transformations, techniques to choose them to optimize particular criteria, and the dependence analysis needed to bring a program into an extended model.

# On the Performance of Window-Based Contention Managers for Transactional Memory

Gokarna Sharma and Costas Busch
Department of Computer Science
Louisiana State University
Baton Rouge, LA 70803, USA
{gokarna, busch}@csc.lsu.edu

## Abstract

This paper considers a family of greedy contention managers for transactional memory for executing *windows of transactions*, which aim to provide both good theoretical and practical performance guarantees at the same time. The main approach behind window-based contention managers is to use random delays at the beginning of the window, which have the property that the conflicting transactions are shifted inside the window and their execution times may not coincide. Thus, conflicting transactions can execute at different time slots and potentially many conflicts are avoided. In this paper, window-based contention managers are considered for eager conflict management software transactional memory systems and evaluated using sorted link list, red-black tree, skip list, and vacation benchmarks. The performance of window-based contention managers is compared through experiments with Polka, the published best contention manager, Greedy, the first contention manager with provable theoretical and practical performance properties, and Priority, a simple priority based contention manager. The results show that window-based contention managers have comparable performance with Polka, and outperform Greedy and Priority, sometimes by significant margins. The evaluation results confirm their benefits in practical performance throughput and other transactional metrics such as aborts per commit, execution time overhead, etc., along with their non-trivial provable properties. This is a significant step toward the design of scalable transactional memory schedulers.

# CRT-based Decryption using DSP blocks on the Xilinx Virtex-6 FPGA

Bo Song, Yasuaki Ito, and Koji Nakano
Department of Information Engineering
School of Engineering, Hiroshima University
1-4-1 Kagamiyama Higashi-Hiroshima, Hiroshima, 739-8527, JAPAN

## Abstract

The main contribution of this paper is to present an efficient hardware algorithm for Chinese Remainder Theorem (CRT) based RSA decryption using Montgomery multiplication algorithm. Our hardware algorithm supporting up-to 2048-bit RSA decryption is designed to be implemented using one DSP48E1 block, one Block RAM and few logic blocks in the Xilinx Virtex-6 FPGA. The implementation results show that our RSA core for 1024-bit RSA decryption runs in 11.263ms. Quite surprisingly, the multiplier in DSP block used to compute Montgomery multiplication works in more than 95% clock cycles during the processing. Hence, our implementation is close to optimal in the sense that it has only less than 5% overhead in multiplication and no further improvement is possible as long as CRT-based Montgomery multiplication based algorithm is applied. We have also succeeded in implementing 320 RSA cores in one Xilinx Virtex-6 FPGA XC6VLX240T-1 which work in parallel. The implemented parallel 320 RSA cores achieve 26.2 Mbit/s throughput for 1024-bit RSA decryption.

## Heuristics and Evaluations of Energy-Aware Task Mapping on Heterogeneous Multiprocessors

Wei Sun and Tomoyoshi Sugawara
A1231, System Platforms Research Labs, NEC Corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa, 211-8666 Japan
Email: w-sun, tom-sugawara@ap.jp.nec.com

### Abstract

On various heterogeneous multiprocessor platforms, it is necessary to optimize the throughput and the energy consumption. The problem of optimally mapping tasks onto a set of given heterogeneous processors for minimum overall completion time has been known, in general, to be NP-complete. The energy consumption of a task may be very different in heterogeneous processors. However, if the most energy-efficient processor for each task is always chosen in the task mapping, the completion time of a set of tasks may grow wildly in the worst case. DVS (Dynamic Voltage Scaling) technique is currently available in a larger number of processors to effectively reduce dynamic power dissipation and consequently to save a proportion of total energy consumption, but meanwhile the execution time of a task running in lower voltage definitely becomes longer. Hence, the task mapping problem in terms of time, energy and voltage turns more complicated and harder to solve along with the heterogeneity. Moreover, today most processors only support discrete DVS and thus the optimization problem tends to be Integer Linear Programming problems for which, as we know, there is no polynomial time algorithm unless P = NP. In this paper we formulate and study the optimization problem of reducing overall completion time and the total energy consumption, and then some heuristics, which are experimentally evaluated and compared.

## Towards an Effective Unified Programming Model for Many-Cores

Ana Lucia Varbanescu[*†] , Pieter Hijma[*] , Rob van Nieuwpoort[*‡] and Henri Bal[*]
[*] Computing Systems Group, Vrije Universiteit Amsterdam, The Netherlands
[†] Parallel and Distributed Systems Group, Delft University of Technology, The Netherlands
[‡] ASTRON, Dwingeloo, The Netherlands
e-mail: analucia,rob,hphijma,bal@cs.vu.nl

### Abstract

Building an effective programming model for many-core processors is challenging. On the one hand, the increasing variety of platforms and their specific programming models force users to take a hardware-centric approach not only for implementing parallel applications, but also for designing them. This approach diminishes portability and, eventually, limits performance. On the other hand, to effectively cope with the increased number of large-scale workloads that require parallelization, a portable, application-centric programming model is desirable. Such a model enables programmers to focus first on extracting and exploiting parallelism from their applications, as opposed to generating parallelism for specific hardware, and only second on platform-specific implementation and optimizations.

In this paper, we first present a survey of programming models designed for programming three families of many-cores: general purpose many-cores (GPMCs), graphics processing units (GPUs), and the Cell/B.E.. We analyze the usability of these models, their ability to improve platform programmability, and the specific features that contribute to this improvement.

Next, we also discuss two types of generic models: parallelism-centric and application-centric. We also analyze their features and impact on platform programmability. Based on this analysis, we recommend two application-centric models (OmpSs and OpenCL) as promising candidates for a unified programming model for many-cores and we discuss potential enhancements for them.

## Elimination Techniques of Redundant Data Transfers among GPUs and CPU on Recursive Stream-based Applications

Pablo Lamilla Álvarez and Shinichi Yamagiwa
School of Information,
Kochi University of Technology/JST PRESTO,
Kami, Kochi 782-8502 Japan

Masahiro Arai and Koichi Wada
Department of Computer Science
University of Tsukuba,
Tsukuba Ibaraki 305-8573 Japan

### Abstract

Applying the stream-based computing approach, the general purpose computing on graphics processing units has become to be considered as a breakthrough to overcome the performance bottleneck as seen in the recent CPU architecture. However, the program potentially includes the data transfer overhead if it has recursive I/Os. During the recursive operation in the GPU-based program, the output streams are copied to the input ones and this overhead degrades the performance. This paper proposes the best method to eliminate the transfer overheads and shows design and implementation of the method based on CUDA and OpenCL. The experimental evaluation using realistic applications shows the method eliminates the transfer overhead and the method exploits the potential performance of GPU.

## Performance Acceleration of Kernel Polynomial Method Applying Graphics Processing Units

Shixun Zhang and Shinichi Yamagiwa
School of Information,
Kochi University of Technology/
JST PRESTO,
Kami, Kochi 782-8502 Japan

Masahiko Okumura and Seiji Yunoki
Computational Condensed Matter Physics Laboratory,
RIKEN ASI,
Wako, Saitama, 351-0198 Japan,
JST CREST,
Kawaguchi, Saitama 332–0012, Japan, and
Computational Materials Science Research Team,
RIKEN AICS,
Kobe, Hyogo, 650-0047 Japan

### Abstract

The Kernel Polynomial Method (KPM) is one of the fast diagonalization methods used for simulations of quantum systems in research fields of condensed matter physics and chemistry. The algorithm has a difficulty to be parallelized on a cluster computer or a supercomputer due to the fine-gain recursive calculations. This paper proposes an implementation of the KPM on the recent graphics processing units (GPU) where the recursive calculations are able to be parallelized in the massively parallel environment. This paper also illustrates performance evaluations regarding the cases when the actual simulation parameters are applied, the one for increased intensive calculations and the one for increased amount of memory usage. Finally, it concludes that the performance on GPU promises very high performance compared to the one on CPU and reduces the overall simulation time.

# Workshop 6

# Communication Architecture for Scalable Systems

# CASS 2011

# Improving Routing Performance in InfiniBand Networks with 2-Level Generalized Fat-tree Topologies

Wickus Nienaber  Santosh Mahapatra  Xin Yuan
Department of Computer Science, Florida State University
{nienaber, mahapatr, xyuan}@cs.fsu.edu

## Abstract

This paper focuses on deterministic single-path routing schemes on 2-level generalized fat-trees. We develop a routing algorithm that is optimal in terms of worst-case permutation performance. In comparison to existing routing schemes for such topologies, our algorithm also improves the average performance of common communication patterns including bisect patterns, full permutation patterns, and dissemination (Bruck) patterns on various 2-level generalized fat-trees as demonstrated in our evaluation results.

# A Unified Algorithm for both Randomized Deterministic and Adaptive Routing in Torus Networks

Keith D. Underwood
Intel Corporation
Albuquerque, NM, USA
Email: keith.d.underwood@intel.com

Eric Borch
Intel Corporation
Fort Collins, CO, USA
Email: eric.borch@intel.com

## Abstract

Torus networks are popular in large scale, high performance computing installations due to their use of relatively short cables and their incremental expandability. There are basically two types of traditional torus routing approaches: deterministic dimension ordered routing and adaptive routing. Traditional approaches to deterministic routing have known shortcomings under some traffic patterns, but adaptive routing creates challenges at the network end-point for programming models that expect ordered messages (e.g. MPI and SHMEM). This paper presents a new approach that supports both adaptive routing and improved throughput for deterministically routed (and therefore ordered) messages. In addition, whereas most current approaches to adaptive routing are designed for either mesh networks or virtual cut-through torus networks, the new routing algorithm allows for adaptive routing of messages on wormhole routed torus networks. The result is a routing algorithm that achieves a substantial portion of the benefit of adaptive routing while maintaining message ordering.

# Profile Guided MPI Protocol Selection for Point-to-Point Communication Calls

Aniruddha Marathe and David K. Lowenthal
Department of Computer Science
The University of Arizona
Tucson, AZ
{amarathe,dkl}@cs.arizona.edu

Zheng Gu, Matthew Small, and Xin Yuan
Department of Computer Science
Florida State University
Tallahassee, FL
{zgu,small,xyuan}@cs.fsu.edu

## Abstract

Improving communication performance is critical to achieving high performance in message-passing programs. Designing new, efficient protocols to realize point-to-point and collective communication operations has therefore been an active area of research. However, the best protocol for a given communication routine is both application and architecture specific.

This paper contributes a new method of selection of the optimal protocol for a given point-to-point communication pair. Our technique analyzes the MPI communication call profile of an application and uses a computation and communication model we have developed to choose the proper protocol for each communication phase. We have applied our system to MPI applications such as CG, Sweep3D and Sparse Matrix multiplication, as well as synthetic applications. Our scheme yields an improvement in total execution time of up to 20% compared to MVAPICH2 and up to 3.2% compared to the best, highly optimized communication protocol for the real applications. Furthermore, experiments on the synthetic applications show that the savings can be much more pronounced.

# Dynamic Time-Variant Connection Management for PGAS Models on InfiniBand

Abhinav Vishnu,[*] Manoj Krishnan,[*] and Pavan Balaji[‡]
[*]Pacific Northwest National Laboratory, Richland, WA 99352
Email: {abhinav.vishnu, manoj}@pnl.gov

[‡]Mathematics and Computer Science Division
Argonne National Laboratory, Argonne, IL 60439
Email: {balaji}@mcs.anl.gov

## Abstract

InfiniBand (IB) has established itself as a promising network infrastructure for high-end cluster computing systems as evidenced by its usage in the Top500 supercomputers today. While the IB standard describes multiple communication models (including reliable-connection (RC), and unreliable datagram (UD)), most of its promising features such as remote direct memory access (RDMA), hardware atomics and network fault tolerance are only available for the RC model which requires connections between communicating process pairs. In the past, several researchers have proposed on-demand connection management techniques that establish connections when there is a need to communicate with another process. While such techniques work well for algorithms and applications that only communicate with a small set of processes in their life-time, there exists a broad set of applications that do not follow this trend. For example, applications that perform dynamic load balancing and adaptive work stealing have a small set of communicating neighbors at any given time, but over time the total number of neighbors can be very high; in some cases, equal to the entire system size.

In this paper, we present a dynamic time-variant connection management approach that establishes connections on-demand like previous approaches, but further intelligently tears down some of the unused connections as well. While connection tear-down itself is relevant for any programming model, different models have different complexities. In this paper, we study the Global Arrays (GA) PGAS model for two reasons: (1) the simple one-sided communication primitives provided by GA and other PGAS models ensure that connection requests are always initiated by the origin process without explicit synchronization with the target process—this makes connection tear-down simpler to handle; and (2) GA supports applications in several domains such as computational chemistry (NWChem) and computational biology (ScalaBLAST) that demonstrate this behavior making it an obvious first target for the proposed enhancements. We evaluate our proposed approach using NWChem computational chemistry application using up to 6144 processes, and show that our approach can significantly reduce the memory requirements of the communication library while maintaining its performance.

# Audit: New Synchronization for the GET/PUT Protocol

Atsushi Hori [#1] , Jinpil Lee [†2] , Mitsuhisa Sato [#3†]
[1]hori(a)riken.jp, [2]jinpil(a)hpcs.cs.tsukuba.ac.jp, [3]msato(a)cs.tsukuba.ac.jp
[#]Riken Advanced Institute for Computational Science
7-1-26 Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo, 650-0047 JAPAN
[†]Center for Computational Sciences, University of Tsukuba
1-1-1 Tennodai Tsukuba-city, Ibaraki, 305-8577 JAPAN

## Abstract

The GET/PUT protocol is considered as an effective communication API for parallel computing. However, the one-sided nature of the GET/PUT protocol lacks the synchronization functionality on the target process. So far, several techniques have been proposed to tackle this problem. The APIs for the synchronization proposed so far have failed to hide the implementation details of the synchronization. In this paper, a new synchronization API for the GET/PUT protocol is proposed. The idea here is to associate the synchronization flags with the GET/PUT memory regions. By doing this, the synchronization flags are hidden from users, and users are free from managing the associations between the memory regions and synchronization flags. The proposed API, named ”Audit,” does not incur additional programing and thus enables natural parallel programing. The evaluations show that *Audit* exhibits better performance than the Notify API proposed in ARMCI.

# Efficient and contention-free virtualisation of fat-trees

Frank Olaf Sem-Jacobsen[1,2] , Ashild Grønstad Solheim[1,2] , Olav Lysne[1,2] ,
Tor Skeie[1,2] , and Thomas Sødring[2]

[1]Department of Informatics
University of Oslo
Oslo, Norway

[2]Networks and Distributed Systems
Simula Research Laboratory
Lysaker, Norway

## Abstract

Maintaining high system utilisation is a key factor for data centres. However, strictly partitioning the datacentre resources to fully isolate the concurrent applications (contention freedom) leads to poor system utilisation because of fragmentation. We present an allocation algorithm for fat-trees (which are commonly found in large-scale data centres) capable of increasing system utilisation while maintaining application isolation. Results show at least a 10% increase in system utilization compared to regular contention free allocation mechanisms, at the cost of a slight reduction in network performance or application isolation.

# Fat-Trees Routing and Node Ordering Providing Non-Blocking MPI Global Collectives for Exascale Jobs

Eitan Zahavi
Mellanox Technologies LTD
Yokneam, Israel
eitan@mellanox.co.il

## Abstract

As the size of High Performance Computing clusters grows, the increasing probability of interconnect hot spots degrades the latency and effective bandwidth the network provides. This paper presents a solution to this scalability problem for real life constant bisectional-bandwidth fat-tree topologies. It is shown that maximal bandwidth and cut-through latency can be achieved for MPI global collective traffic. To form such congestion-free configuration, MPI programs should utilize collective communication, MPI-node- order should be topology aware,and the packets routing should match the MPI communication patterns. First, we show that MPI collectives can be classified into unidirectional and bidirectional shifts. Using this property, we propose a scheme for congestion-free routing of the global collectives in fully and partially populated fat trees running a single job. Simulation results of the proposed routing, MPI-node-order and communication patterns show a 40% throughput improvement over previously published results for all-to-all collectives.

# Optimizing MPI Collectives using Efficient Intra-node Communication Techniques over the Blue Gene/P Supercomputer

Amith R Mamidala[1] Daniel Faraj[2] Sameer Kumar[1] Douglas Miller[2]
Michael Blocksome[2] Thomas Gooding[2] Philip Heidelberger[1] Gabor Dozsa[3]
[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598
[2] IBM Systems and Technology Group, Rochester, MN, 55901
[3] Barcelona Supercomputing Center, Spain
{amithr, faraja, sameerk, dougmill, blocksom, tgooding, philiph, gdozsa}@us.ibm.com

## Abstract

The Blue Gene/P (BG/P) supercomputer consists of thousands of compute nodes interconnected by multiple networks. Out of these, a 3D torus equipped with direct memory access (DMA) engine is the primary network. BG/P also features a collective network which supports hardware accelerated collective operations such as broadcast and allreduce. One of the operating modes on BG/P is the virtual node mode where the four cores can be active MPI tasks, performing inter-node and intra-node communication.

This paper proposes software techniques to enhance MPI Collective communication primitives, MPI_Bcast and MPI_Allreduce in virtual node mode by using cache coherent memory subsystem as the communication method within the node. The paper describes techniques leveraging atomic operations to design concurrent data structures such as broadcast-FIFOs to enable efficient collectives. Such mechanisms are important as we expect the core counts to rise in the future and having such data structures makes programming easier and efficient. We also demonstrate the utility of shared address space techniques for MPI collectives, wherein a process can access the peer's memory by specialized system calls. Apart from cutting down the copy costs, such techniques allow for seamless integration of network protocols with intra-node communication methods. We propose intra-node extensions to multi-color network algorithms for collectives using light weight synchronizing structures and atomic operations. Further, we demonstrate that shared address techniques allow for good load balancing and are critical for efficiently using the hardware collective network on BG/P. When compared to current approaches on the 3D torus, our optimizations provide performance up to almost 3 folds for MPI_Bcast and a 33% performance gain for MPI_Allreduce(in virtual node mode). We also see improvements up to 44% for MPI_Bcast using the collective tree network.

# ConnectX-2 CORE-Direct Enabled Asynchronous Broadcast Collective Communications

Manjunath Gorentla Venkata and Richard L. Graham and
Joshua S. Ladd and Pavel Shamis
Oak Ridge National Laboratory
Computer Science and Mathematics Division
Oak Ridge, TN
manjujv, rlgraham, laddjs, shamisp@ornl.gov
Ishai Rabinovitz and Vasily Filipov and Gilad Shainer
Mellanox Technologies, Inc
Yokneam, Israel
ishai, vasily, shainer@mellanox.co.il

## Abstract

This paper describes the design and implementation of InfiniBand (IB) CORE-*Direct* based blocking and nonblocking broadcast operations within the Cheetah collective operation framework. It describes a novel approach that fully offloads collective operations and employs only user-supplied buffers. For a 64 rank communicator, the latency of CORE-*Direct* based hierarchical algorithm is better than production-grade Message Passing Interface (MPI) implementations, 150% better than the default Open MPI algorithm and 115% better than the shared memory optimized MVAPICH implementation for a one kilobyte (KB) message, and for eight mega-bytes (MB) it is 48% and 64% better, respectively. Flat-topology broadcast achieves 99.9% overlap in a polling based communication-computation test, and 95.1% overlap for a wait based test, compared with 92.4% and 17.0%, respectively, for a similar Central Processing Unit (CPU) based implementation.

# Dodging Non-Uniform I/O Access in Hierarchical Collective Operations for Multicore Clusters

Brice Goglin and Stéphanie Moreaud
INRIA, LaBRI, Université of Bordeaux
351, cours de la Libération
F-33405 Talence – France
Email: {Brice.Goglin,Stephanie.Moreaud}@inria.fr

## Abstract

The increasing number of cores led to scalability issues in modern servers that were addressed by using non- uniform memory interconnects such as HyperTransport and QPI. These technologies reintroduced Non-Uniform Memory Access (NUMA) architectures. They are also responsible for Non-Uniform Input/Output Access (NUIOA), as I/O devices may be directly connected to a single processor, thus getting faster access to some cores and memory banks than to the others. In this paper, we propose to adapt MPI collective operations to NUIOA constraints. These operations are now often based on the combination of multiple strategies depending on the underlying cluster topology, with local leader processes being used as intermediate. Our strategy focuses on electing these leaders according to the locality of processes and network interfaces so as to give them privileged network access. We validate our approach on a hierarchical Broadcast operation which brings up to 25% throughput improvement between 64 processes.

# Workshop 7
# High-Performance, Power-Aware Computing
# HPPAC 2011

## LAPP: A Low Power Array Accelerator with Binary Compatibility

Naveen Devisetti, Takuya Iwakami, Kazuhiro Yoshimura, Takashi Nakada, Jun Yao, and Yasuhiko Nakashima
Nara Institute of Science and Technology, Nara, Japan
Email:venkatarama-d, takuya-iw, kazuhiro-y, nakada, yaojun, nakashim@is.naist.jp

### Abstract

Recently, reconfigurable architectures are becoming popular to achieve good energy efficiency. In this paper we designed an energy efficient, highperformance accelerator, named Linear Array Pipeline Processor (LAPP). LAPP works to accelerate existing machine code executions to improve performance while maintaining the binary compatibility, instead of using special codes. With its highly reconfigurable feature, LAPP architecture is designed to effectively work with unit gating through a sufficiently long period to conceal the gating penalty, and thereby incurs minimum power consumption for a given workload. Specifically, codes are mapped fixedly onto Functional Unit (FU) array with minimized caches and registers, and they are pipeline executed with data stream. The synthesized results show that the area of a 36-stage LAPP is equal to 9.5 times that of a traditional processor core area. Compared to a Many-Core Processor (MCP) of the same area, an LAPP-simulator based estimation indicates that LAPP can achieve about 10 times the power efficiency for 9 image processing workloads.

## Design and Analysis of Heuristic Algorithms for Power-Aware Scheduling of Precedence Constrained Tasks

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

### Abstract

Energy-efficient scheduling of sequential tasks with precedence constraints on multiprocessor computers with dynamically variable voltage and speed is investigated as combinatorial optimization problems. In particular, the problem of minimizing schedule length with energy consumption constraint and the problem of minimizing energy consumption with schedule length constraint are considered. Our scheduling problems contain three nontrivial subproblems, namely, precedence constraining, task scheduling, and power supplying. Each subproblem should be solved efficiently so that heuristic algorithms with overall good performance can be developed. Such decomposition of our optimization problems into three subproblems makes design and analysis of heuristic algorithms tractable. Three types of heuristic power allocation and scheduling algorithms are proposed for precedence constrained sequential tasks with energy and time constraints, namely, pre-power-determination algorithms, post-power-determination algorithms, and hybrid algorithms. The performance of our algorithms are analyzed and compared with optimal schedules analytically. Such analysis has not been conducted in the literature for any algorithm. Therefore, our investigation in this paper makes initial contribution to analytical performance study of heuristic power allocation and scheduling algorithms for precedence constrained sequential tasks. Our extensive simulation data demonstrate that for wide task graphs, the performance ratios of all our heuristic algorithms approach one as the number of tasks increases.

# Performance, Power, and Thermal Analysis of Low-Power Processors for Scale-Out Systems

Phillip Stanley-Marbell
IBM Research–Zürich
CH-8803 Rschlikon, Switzerland
pst@zurich.ibm.com

Victoria Caparrós Cabezas
IBM Research–Zürich
CH-8803 Rschlikon, Switzerland
vca@zurich.ibm.com

## Abstract

There is increased interest, in high-performance computing as well as in commercial datacenters, in so-called scale-out systems, where large numbers of low-cost and low-power-dissipation servers are used for workloads which have available coarse-grained parallelism. One target class of devices for building scale-out systems is the class of low-power processors, such as those based on the ARM architecture, the Power Architecture, and the Intel Atom processor.

This article presents a detailed characterization of three contemporary low-power processors covering all the aforementioned ISAs, all implemented in state-of-the-art 45nm semiconductor processes. Processor performance, power dissipation, thermal load, and board-level power dissipation apportionment are presented, via a combination of hardware performance counters, OS-level timing measurements, current measurements, and thermal imaging via a microbolometer array.

It is demonstrated that while certain processors might provide low power dissipation, the most energy-efficient platform depends on the characteristics of the application, and the design of the entire platform (including integrated versus on-board peripherals, power supply regulators, etc.). The lowest-power platform showed a power-efficiency advantage of almost four times lower idle power dissipation, and almost five times lower active power dissipation for a single-threaded workload, versus the highest-power-dissipation platform studied. The latter however achieved a factor of two better energy-efficiency than its closest competitor, when executing a throughput-oriented workload, due to significantly better compute performance and available hardware concurrency.

# Power Consumption of Mixed Precision in the Iterative Solution of Sparse Linear Systems

Hartwig Anzt, Vincent Heuveline, Björn Rocker
Institute for Applied and Numerical Mathematics 4
Karlsruhe Institute of Technology, Germany
{hartwig.anzt,vincent.heuveline,bjorn.rocker}@kit.edu
Maribel Castillo, Juan C. Fernández,
Rafael Mayo, Enrique S. Quintana-Ortí
Depto. de Ingeniería y Ciencia de Computadores
Universidad Jaume I, Spain
{castillo,jfernand,mayo,quintana}@icc.uji.es

## Abstract

This paper presents a detailed analysis of a mixed precision iterative refinement solver applied to a linear system obtained from the 2D discretization of a fluid flow problem. The total execution time and energy need of different soft- and hardware implementations are measured and compared with those of a plain GMRES-based solver in double precision. The time and energy consumption of individual parts of the algorithm are monitored as well, enabling a deeper insight and the possibility of optimizing the energy consumption of the code on a general-purpose multi-core architecture and systems accelerated by a graphics processor.

# Rack Aware Scheduling in HPC data centers

Vikas Ashok Patil
Department of Computer Science
University at Buffalo,
State University of New York,
Buffalo, USA
vikasash@buffalo.edu

Dr. Vipin Chaudhary
Department of Computer Science
University at Buffalo,
State University of New York,
Buffalo, USA
vipin@buffalo.edu

## Abstract

Energy consumption in high performance computing data centers has become a long standing issue. With rising costs of operating the data center, various techniques need to be employed to reduce the overall energy consumption. Currently, among others there are techniques that guarantee reduced energy consumption by powering on/off the idle nodes. However, most of them do not consider the energy consumed by other components in a rack. Our study addresses this aspect of the data center. We show that we can gain considerable energy savings by reducing the energy consumed by these rack components. In this regard, we propose a scheduling technique that will help schedule jobs with the above mentioned goal. We claim that by our scheduling technique we can reduce the energy consumption considerably without affecting other performance metrics of a job. We implement this technique as an enhancement to the well known Maui scheduler and present our results. We compare our technique with various currently available Maui scheduler configurations. We simulate a wide variety of workloads from real cluster deployments using the simulation mode of Maui. Our results consistently show about 7 to 14% savings over the currently available Maui scheduler configurations. We shall also see that our technique can be applied in tandem with most of the existing energy aware scheduling techniques to achieve enhanced energy savings.

# Evaluation of the Energy Performance of Dense Linear Algebra Kernels on Multi-Core and Many-Core Processors

Maribel Castillo, Manel Dolz, Juan C. Fernández,
Rafael Mayo, Enrique S. Quintana-Ortí, Vicente Roca
Depto. de Ingeniería y Ciencia de Computadores
Universidad Jaume I
12.071 - Castellón, Spain
{castillo,dolzm,jfernand,mayo,quintana,vroca}@icc.uji.es

## Abstract

We evaluate the power consumption of three key kernels from several tuned dense linear algebra libraries on three general-purpose multi-core processors and a graphics processor, representative of the state-of-the-art in computing technology. The results of this study provide basic insights on the energy scalability of multi- and many-core designs and multi-threaded software as the building blocks of future EXAFLOPS systems.

# Emerging Trends on the Evolving Green500: Year Three

Tom Scogland, Balaji Subramaniam, and Wu-chun Feng
{tom.scogland,balaji,wfeng}@vt.edu
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060

## Abstract

It has been traditionally viewed that as the scale of a supercomputer increases, its energy efficiency decreases due to performance that scales sub-linearly and power consumption that scales at least linearly with size. However, based on the first three years of the Green500, this view does not hold true for the fastest supercomputers in the world. Many reasons for this counterintuitive trend have been proposed — with improvements in feature size, more efficient networks, and larger numbers of slower cores being amongst the most prevalent.

Consequently, this paper provides an analysis of emerging trends in the Green500 and delves more deeply into how larger-scale supercomputers compete with smaller-scale supercomputers with respect to energy efficiency. In addition, our analysis provides a compelling early indicator of the future of exascale computing. We then close with a discussion on the evolution of the Green500 based on community feedback.

# Dynamic Frequency Scaling and Energy Saving in Quantum Chemistry Applications

Vaibhav Sundriyal
Masha Sosonkina
Fang Liu Ames Laboratory/DOE
Iowa State University
Ames, IA 50011, USA
{vaibhavs,masha,fangliu}@scl.ameslab.gov

Michael W. Schmidt
Department of Chemistry
Iowa State University
Ames, IA 50011, USA
mike@si.msg.chem.iastate.edu

## Abstract

Modern high-performance computing system design is becoming increasingly aware of the energy proportional computing to lower the operational costs and raise reliability. At the same time, high-performance application developers are taking pro-active steps towards less energy consumption without a significant performance loss. One way to accomplish this is to change the processor frequency dynamically during application execution. In this paper, a representative computationally-intensive HPC application GAMESS is considered with the aim to investigate the energy saving potential of its various stages. GAMESS is a quantum chemistry software package used worldwide to perform *ab initio* electronic structure calculations.

This paper presents energy consumption characteristics of two Self-Consistent Field method implementations in GAMESS, which radically differ in their computer resource usages. The dynamic frequency scaling optimization is applied to these implementations and serves as verification for the proposed general energy savings model. The developed model provides the minimum of on the compute node energy consumption under a given performance loss tolerance for various processor frequencies.

# Temperature Aware Load Balancing for Parallel Applications: Preliminary Work

Osman Sarood, Abhishek Gupta, Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{sarood1, gupta59, kale}@illinois.edu

## Abstract

Increasing number of cores and clock speeds on a smaller chip area implies more heat dissipation and an ever increasing heat density. This increased heat, in turn, leads to higher cooling cost and occurrence of hot spots. Effective use of dynamic voltage and frequency scaling (DVFS) can help us alleviate this problem. But there is an associated execution time penalty which can get amplified in parallel applications. In high performance computing, applications are typically tightly coupled and even a single overloaded core can adversely affect the execution time of the entire application. This makes load balancing of utmost value. In this paper, we outline a temperature aware load balancing scheme, which uses DVFS to keep core temperatures below a user-defined threshold with minimum timing penalty. While doing so, it also reduces the possibility of hot spots. We apply our scheme to three parallel applications with different energy consumption profiles. Results from our technique show that we save up to 14% in execution time and 12% in machine energy consumption as compared to frequency scaling without using load balancing. We are also able to bound the average temperature of all the cores and reduce the temperature deviation amongst the cores by a factor of 3.

# Design and Evaluation of a Novel PCI Express Direct Link PEARL and Its Implementation PEACH

Toshihiro Hanawa[*†], Taisuke Boku[*†], Shin'ichi Miura[†], Mitsuhisa Sato[*†], and Kazutami Arimoto[‡]
[*] Graduate School of Systems and Information Engineering
[†] Center for Computational Sciences
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577 Japan
Email: {hanawa,taisuke,miura,msato}@hpcs.cs.tsukuba.ac.jp
[‡] System Core Development Division
Renesas Electronics Corporation
4-1 Mizuhara, Itami, Hyogo 664-0005 Japan
Email: kazutami.arimoto.jx@renesas.com

## Abstract

We have proposed PEARL, which is a power-aware, high-performance, dependable communication link using PCI Express as a direct communication device, for application in a wide range of parallel processing systems, from high-end embedded systems to small-scale high-performance clusters. The PEACH chip used to realize PEARL connects four ports of PCI Express Gen 2 with four lanes and uses an M32R processor with four cores and several DMACs. We also develop the PEACH board as a network interface card for implementing the PEACH chip. The preliminary evaluation results indicate that the PEACH board achieves a maximum performance of 1.1 Gbyte/s. In addition, through power-aware control, the power consumption can be reduced by up to 0.7 watts, and both the time required to reduce the number of lanes and the time required to change from Gen 2 to Gen 1 are 10 $\mu$s.

# Workshop 8

# High-Performance Grid and Cloud Computing Workshop

# HPGC 2011

## Cloud Management: Challenges and Opportunities

Tim Forell, Dejan Milojicic, Vanish Talwar
Hewlett-Packard
[first.lastname]@hp.com

### Abstract

Cloud computing offers on-demand access to shared resources and services, hosted in warehouse sized data centers at cloud providers. Effective management of these shared resources and services is one of the key requirements for the delivery of cloud computing. However, there are several challenges to achieve effective cloud management. These include scale, multiple levels of abstraction, federation, sustainability, and dynamism. In this paper, we outline these challenges, and then describe specific examples of new management architectures that address these challenges. We focus on driving principles for the new designs, and give an illustration of its deployment on OpenCirrus - a research cloud testbed. Overall, the paper provides open issues and challenges in cloud management for the research community to address.

## Energy-Aware Application-Centric VM Allocation for HPC

H. Viswanathan, E.K. Lee, I. Rodero, D. Pompili, M. Parashar
NSF Center for Autonomic Computing
Rutgers University, New Brunswick, NJ
{hari_viswanathan, eunkyung_lee, irodero, pompili, parashar}@cac.rutgers.edu
M. Gamell
Open University of Catalonia
Barcelona, Spain
marcgamell@uoc.edu

### Abstract

Virtualized datacenters and clouds are being increasingly considered for traditional High-Performance Computing (HPC) workloads that have typically targeted Grids and conventional HPC platforms. However, maximizing energy efficiency, cost-effectiveness, and utilization of datacenter resources while ensuring performance and other Quality of Service (QoS) guarantees for HPC applications requires careful consideration of important and extremely challenging tradeoffs.

An innovative application-centric energy-aware strategy for Virtual Machine (VM) allocation is presented. The proposed strategy ensures high resource utilization and energy efficiency through VM consolidation while satisfying application QoS. While existing VM allocation solutions are aimed at satisfying only the resource utilization requirements of applications along only one dimension (CPU utilization), the proposed approach is more generic as it employs knowledge obtained through application profiling along multiple dimensions. The results of our evaluation show that the proposed VM allocation strategy enables significant reduction either in energy consumption or in execution time, depending on the optimization goals.

# ENIGMA: Distributed Virtual Disks for Cloud Computing

Matteo Zola*, Valerio Bioglio†, Cosimo Anglano*, Rossano Gaeta†, Marco Grangetto†, Matteo Sereno†
* Dipartimento di Informatica, Università del Piemonte Orientale, Alessandria, Italy
{mzola, cosimo.anglano}@di.unipmn.it
† Dipartimento di Informatica, Università di Torino, Italy
{bioglio, rossano, grangetto, matteo}@di.unito.it

## Abstract

We propose ENIGMA, a distributed infrastructure that provides Cloud Computing infrastructures with *virtual disks* by abstracting the storage resources provided by a set of physical nodes and exposing to Cloud Computing users, applications, and Virtual Machines a set of virtual block storage devices, that can be used exactly as standard physical disks.

ENIGMA is designed to provide large storage capacity, high availability, strong confidentiality, and data access performance comparable to that of traditional storage virtualization solutions. To achieve all these design goals, ENIGMA exploits erasure-coding techniques, whereby each sector of a virtual disk is encoded as a set of *n fragments*, that are independently stored on a set of physical storage nodes, $k$ of which ($k \leq n$) are sufficient to reconstruct that sector.

We present the ENIGMA architecture and we show how the coding of sectors of a virtual disk ensures high availability in spite of failure of individual storage nodes as well as confidentiality in face of several types of attacks. We also briefly discuss performance results of ENIGMA.

# Managed GridFTP

John Bresnahan, Michael Link, Rajkumar Kettimuthu, Ian Foster
Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL
Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL

## Abstract

GridFTP extends the standard FTP protocol to provide a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks. The Globus GridFTP implementation has become the preeminent high-performance data transfer tool for the Grid community. Its modular architecture enables a standard GridFTP-compliant client access to any storage system that can implement its data storage interface, including the HPSS archival storage system, SRB, the GPFS parallel file system, and POSIX file systems. Its eXtensible I/O interface allows GridFTP to target high-performance wide-area communication protocols such as UDT, FAST TCP, and RBUDP. The Globus GridFTP server implementation already implements the concept of "striping," where multiple data movers are aggregated as one logical resource. However, there exists no mechanism in GridFTP to manage the use of server resources by the clients. When many transfer sessions occur simultaneously, the GridFTP server can overwhelm the transfer host and/or the underlying storage system. Moreover, there is no flexibility in the management of data movers in striped configuration. The data movers must be statically configured, and they cannot come and go dynamically. In this paper, we present a framework to manage the GridFTP resources efficiently so as to avoid overburdening host resources, prevent client starvation, and enable dynamic addition or removal of data movers.

# RBAC for High Performance Computing Systems Integration in Grid Computing and Cloud Computing

Anil L. Pereira
Southwestern Oklahoma State University
Weatherford, OK 73096, USA
e-mail: anil.pereira@swosu.edu

## Abstract

This paper describes a Role-based Access Control (RBAC) mechanism for distributed High Performance Computing (HPC) systems that will facilitate scalable evaluation, management and enforcement of access control policies. The RBAC mechanism forms an enhanced security framework for Grids and Clouds that will allow for interoperability between technologies in the two domains. The mechanisms being proposed here are important because the current lack of software tools and security standards in accessing distributed HPC systems and transporting Large Data Sets can add immensely to overheads in data processing or data integration times. RBAC models make policy management scalable and by virtue of being modular allow for more sophisticated access control models to be integrated with them. This paper shows how existing security standards can be leveraged for the specification and management of RBAC policies with the aim to allow disparate applications, systems and security domains to interoperate. The eXtensible Access Control Markup Language (XACML) can be used for policy specification and management across disparate organizations and the Security Assertion Markup Language (SAML) can be used for authentication and authorization assertions across the same. Both standards can be leveraged to facilitate policy management and enforcement, and delegation of rights. Authorization servers like Shibboleth can be leveraged for use as RBAC system components.

# e-Science Infrastructure Integration Invariants to Enable HTC and HPC Interoperability Applications

M. Riedel, M.S. Memon,
A.S. Memon, D. Mallmann, Th. Lippert
Jlich Supercomputing Centre
Forschungszentrum Jülich
Jlich, Germany
m.riedel@fz-juelich.de

D. Kranzlmüller
Ludwig Maximilians University Munich
Munich, Germany

A. Streit
Karlsruhe Institute of Technology
Karlsruhe, Germany

## Abstract

During the past decade, significant international and broader interdisciplinary research is increasingly carried out by global collaborations that often share resources within a single production e-science infrastructure. More recently, increasing complexity of e-science applications embrace multiple physical models (i.e. multi-physics) and consider longer and more detailed simulation runs as well as a larger range of scales (i.e. multi-scale). This increase in complexity is creating a steadily growing demand for cross-infrastructure operations that take the advantage of multiple e-science infrastructures with a more variety of resource types. Since interoperable e-science infrastructures are still not seamlessly provided today we proposed in earlier work the Infrastructure Interoperability Reference Model (IIRM) that represents a trimmed down version of the Open Grid Service Architecture (OGSA) in terms of functionality and complexity, while on the other hand being more specifically useful for production and thus easier to implement. This contribution focuses on several important reference model invariants that are often neglected when infrastructure integration activities are being performed thus hindering seamless interoperability in many aspects. In order to indicate the relevance of our invariant definitions, we provide insights into two accompanying cross-infrastructure use cases of the bio-informatics and fusion science domain.

# Cyberinfrastructure Usage Modalities on the TeraGrid

Daniel S. Katz
Computational Institute
University of Chicago & Argonne National Laboratory
Chicago, IL, USA
d.katz@ieee.org
David Hart
Computational & Information Systems Laboratory
National Center for Atmospheric Research
Boulder, CO, USA
dhart@ucar.edu

Chris Jordan
Texas Advanced Computing Center
University of Texas Austin
Austin, TX, USA
ctjordan@tacc.utexas.edu

Amit Majumdar
San Diego Supercomputer Center
University of California San Diego
San Diego, CA, USA
majumdar@sdsc.edu

J.P. Navarro
Computational Institute
Argonne National Laboratory & University of Chicago
Chicago, IL, USA
navarro@mcs.anl.gov

Warren Smith
Texas Advanced Computing Center
University of Texas Austin
Austin, TX, USA
wsmith@tacc.utexas.edu

John Towns
National Center for Supercomputing Applications
University of Illinois
Urbana, IL, USA
jtowns@ncsa.illinois.edu

Von Welch
Indiana University
Bloomington, IN, USA
vwelch@indiana.edu

Nancy Wilkins-Diehr
San Diego Supercomputer Center
U. of California San Diego
San Diego, CA, USA
wilkinsn@sdsc.edu

## Abstract

This paper is intended to explain how the TeraGrid would like to be able to measure ”usage modalities.” We would like to (and are beginning to) measure these modalities to understand what objectives our users are pursuing, how they go about achieving them, and why, so that we can make changes in the TeraGrid to better support them.

# A Framework for Elastic Execution of Existing MPI Programs

Aarthi Raveendran Tekin Bicer Gagan Agrawal
Department of Computer Science and Engineering, Ohio State University
{raveendr,bicer,agrawal}@cse.ohio-state.edu

## Abstract

There is a clear trend towards using cloud resources in the scientific or the HPC community, with a key attraction of cloud being the *elasticity* it offers. In executing HPC applications on a cloud environment, it will clearly be desirable to exploit elasticity of cloud environments, and increase or decrease the number of instances an application is executed on during the execution of the application, to meet time and/or cost constraints. Unfortunately, HPC applications have almost always been designed to use a fixed number of resources.

This paper describes our initial work towards the goal of making existing MPI applications elastic for a cloud framework. Considering the limitations of the MPI implementations currently available, we support adaptation by terminating one execution and restarting a new program on a different number of instances. The components of our envisioned system include a decision layer which considers time and cost constraints, a framework for modifying MPI programs, and a cloud-based runtime support that can enable redistributing of saved data, and support automated resource allocation and application restart on a different number of nodes.

Using two MPI applications, we demonstrate the feasibility of our approach, and show that outputting, redistributing, and reading back data can be a reasonable approach for making existing MPI applications elastic.

# Optimal Load Distribution for Multiple Heterogeneous Blade Servers in a Cloud Computing Environment

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
{*Email: lik@newpaltz.edu*}

## Abstract

*Given a group of heterogeneous blade servers in a cloud computing environment or a data center of a cloud computing provider, each having its own size and speed and its own amount of preloaded special tasks, we are facing the problem of optimal distribution of generic tasks over these blade servers, such that the average response time of generic tasks is minimized. Such performance optimization is important for a cloud computing provider to efficiently utilize all the available resources and to deliver the highest quality of service. We develop a queueing model for a group of heterogeneous blade servers, and formulate and solve the optimal load distribution problem of generic tasks for multiple heterogeneous blade servers in a cloud computing environment in two different situations, namely, special tasks with and without higher priority. Extensive numerical examples and data are demonstrated and some important observations are made. It is found that server sizes, server speeds, task execution requirement, and the arrival rates of special tasks all have significant impact on the average response time of generic tasks, especially when the total arrival rate of generic tasks is large. It is also found that the server size heterogeneity and the server speed heterogeneity do not have much impact on the average response time of generic tasks. Furthermore, larger (smaller, respectively) heterogeneity results in shorter (longer, respectively) average response time of generic tasks.*

# Using Hierarchical Dependency Data Flows to Enable Dynamic Scalability on Parallel Patterns

Jeremy Villalobos
Computer Science Department
University of North Carolina
Charlotte, North Carolina 28223
Email: jeremyvillalobos@gmail.com

Barry Wilkinson
Computer Science Department
University of North Carolina
Charlotte, North Carolina 28223
Email: abw@uncc.edu

## Abstract

Hierarchical dependencies are presented as an extension to data flow programming that allows parallel programs dynamically scale on a heterogeneous environment. The concept can help Grid parallel programs to cope with changes in processors, or Cloud and multi-core frameworks to manage energy use. A data stream with dependencies can be split, which in turn allows for a greater use of processors. The concept shows a 6% overhead when running with split dependencies on shared memory. The overhead on a cluster environment is masked by the network delay. Hierarchical dependencies show a 18.23% increase in non-functional code when the feature was added to a 5-point stencil implementation.

# Hybrid Heuristic for Scheduling Data Analytics Workflow Applications in Hybrid Cloud Environment

Mustafizur Rahman, Xiaorong Li and Henry Palit
Institute of High Performance Computing (IHPC)
Agency for Science Technology and Research (A*STAR)
Singapore
mmrahman@csse.unimelb.edu.au
lixr@ihpc.a-star.edu.sg
henry@ihpc.a-star.edu.sg

## Abstract

Effective scheduling is a key concern for the execution of performance driven applications, such as workflows in dynamic and cost driven environment including Cloud. The majority of existing scheduling techniques are based on meta-heuristics that produce good schedules with advance reservation given the current state of Cloud services or heuristics that are dynamic in nature, and map the workflow tasks to services on-the-fly, but lack the ability of generating schedules considering workflow-level optimization and user QoS constraints. In this paper, we propose an Adaptive Hybrid Heuristic for user constrained data-analytics workflow scheduling in hybrid Cloud environment by integrating the dynamic nature of heuristic based approaches as well as workflow-level optimization capability of meta-heuristic based approaches. The effectiveness of the proposed approach is illustrated by a comprehensive case study with comparison to existing techniques.

# Workshop 9

# Workshop on System Management Techniques, Processes, and Services

# SMTPS 2011

# Evaluating Load Generation in Virtualized Environments for Software Performance Testing

Marco A. S. Netto[1] , Suzane Menon[1] , Hugo V. Vieira[1] , Leandro T. Costa[1] ,
Flavio M. de Oliveira[1] , Rodrigo Saad[2] , Avelino Zorzo[1]
[1]Pontifical Catholic University of Rio Grande do Sul (PUCRS) – Faculty of Informatics
[2]Dell Inc.
Porto Alegre, Brazil

## Abstract

Before placing a software system into production, it is necessary to guarantee it provides users with a certain level of Quality-of-Service. Intensive performance testing is then necessary to achieve such a level and the tests require an isolated computing environment. Virtualization can therefore play an important role for saving energy costs by reducing the number of servers required to run performance tests and for allowing performance isolation when executing multiple tests in the same computing infrastructure. Load generation is an important component in performance testing as it simulates users interacting with the target application. This paper presents our experience in using a virtualized environment for load generation aimed at performance testing. We measured several performance metrics and varied system load, number of virtual machines per physical resource, and the CPU pinning schema for comparison of virtual and physical machines. The two main findings from our experiments are that physical machines produced steadier and faster response times under heavy load and that the pinning schema is an important aspect when setting up a virtualized environment for load generation.

# New Metrics for Scheduling Jobs on Clusters of Virtual Machines

Yanbin Liu, Norman Bobroff,
Liana Fong, Seetharami Seelam
IBM T. J. Watson Research Center
{ygliu,bobroff,llfong,sseelam}@us.ibm.com

Javier Delgado
Florida International University
{javier.delgado}@fiu.edu

## Abstract

As the virtualization of resources becomes popular, the scheduling problem of batch jobs on virtual machines requires new approaches. The dynamic and sharing aspects of virtual machines introduce unique challenges and complexity for the scheduling problems of batch jobs. In this paper, we propose a new set of metrics, called potential capacity (PC) and equilibrium capacity (EC), of resources that incorporate these dynamic, elastic, and sharing aspects of co-located virtual machines. We then show that we mesh this set of metrics smoothly into traditional scheduling algorithms. We evaluate the performance in using the metrics in a widely used greedy scheduling algorithm and show that the new scheduler improves job speedup for various configurations when compared to a similar algorithm using traditional physical machine metrics such as available CPU capacity.

## Secondary Job Scheduling in the Cloud with Deadlines

Shiyao Chen[†] Ting He[‡] Ho Yin Starsky Wong[‡] Kang-Won Lee[‡] Lang Tong[†]

### Abstract

The highly dynamic nature of the cloud environment leads to a time-varying resource utilization and the cloud provider can potentially accommodate secondary jobs with the remaining resource. To better implement the idea of resource reutilization in the cloud environment, the problem of secondary job scheduling with deadlines under time-varying resource capacity is considered in this paper. A transformation is proposed to reduce the offline problem with time-varying processor capacity to that with constant capacity. For online scheduling of underloaded system, it is shown that the earliest deadline first (EDF) scheduling algorithm achieves competitive ratio 1. For the overloaded system, an online scheduling algorithm V-Dover is proposed with asymptotically optimal competitive ratio when a certain admissibility condition holds. It is further shown that, in the absence of the admissibility condition, no online scheduling algorithm exists with a positive competitive ratio. Simulation results are presented to illustrate the performance advantage of the proposed V-Dover algorithm.

## Parallax - A New Operating System for Scalable, Distributed, and Parallel Computing

Dr. Rao Mikkilineni
Kawa Objects Inc., USA
and
Ian Seyler
Return Infinity, Canada
rao@kawaobjects.com, ian.seyler@returninfinity.com

### Abstract

Parallax, a new operating system, implements scalable, distributed, and parallel computing to take advantage of the new generation of 64-bit multi-core processors. Parallax uses the Distributed Intelligent Managed Element (DIME) network architecture, which incorporates a signaling network overlay and allows parallelism in resource configuration, monitoring, analysis and reconfiguration on-the-fly based on workload variations, business priorities and latency constraints of the distributed software components. A workflow is implemented as a set of tasks, arranged or organized in a directed acyclic graph (DAG) and executed by a managed network of DIMEs. These tasks, depending on user requirements are programmed and executed as loadable modules in each DIME. Parallax is implemented using the assembler language at the lowest level for efficiency and provides a C/C++ programming API for higher level programming.

# Reducing Shared Cache Contention by Scheduling Order Adjustment on Commodity Multi-Cores

Yingxin Wang Yan Cui Pin Tao Haining Fan Yu Chen Yuanchun Shi
Tsinghua University
{wyxmails, ccuiyyan}@gmail.com,{taopin, fhn, yuchen, shiyc}@tsinghua.edu.cn

## Abstract

Due to the limitation of power and processor complexity on traditional single core processors, multi-core processors have become the mainstream. One key feature on commodity multi-cores is that the last level cache (LLC) is usually shared. However, the shared cache contention can affect the performance of applications significantly. Several existing proposals demonstrate that task co-scheduling has the potential to alleviate the contention, but it is challenging to make co-scheduling practical in commodity operating systems.

In this paper, we propose two lightweight practical *cache-aware* co-scheduling methods, namely static SOA and dynamic SOA, to solve the cache contention problem on commodity multi-cores. The central idea of the two methods is that *the cache contention can be reduced by adjusting the scheduling order properly*. These two methods are different from each other mainly in the way of acquiring the process's cache requirement. The static SOA (static scheduling order adjustment) method acquires the cache requirement information statically by offline profiling, while the dynamic SOA (dynamic scheduling order adjustment) captures the cache requirement statistics by using performance counters. Experimental results using multi-programmed NAS workloads suggest that the proposed methods can greatly reduce the effect of cache contention on multi-core systems. Specifically, for the static SOA method, the execution time can be reduced by up to 15.7%, the number of cache misses can be reduced by up to 11.8%, and the performance improvement remains obvious across the cache size and the length of time slice. For the dynamic SOA method, the execution time reduction can achieve up to 7.09%.

# Characterization of the Communication Patterns of Scientific Applications on Blue Gene/P

Pier Giorgio Raponi[1,2]
[1]Scuola Superiore Sant'Anna
Via Moruzzi 1, 56124 Pisa, Italy
raponi@sssup.it

Fabrizio Petrini,[2] Robert Walkup,[2] Fabio Checconi[2]
[2]IBM TJ Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598
{fpetrin,walkup,fchecco}@us.ibm.com

## Abstract

This paper examines the communication characteristics of a collection of scientific applications selected from the LLNL's Sequoia suite of benchmarks and the ANL's workload. By using an instrumentation library built on top of MPI we collect and characterize the applications's messaging behavior: the type of communication patterns and primitives used, the amount of time spent for communication, the message sizes, the total amount of data exchanged, and the impact of collective primitives; through communication matrices we visualize the actual communication patterns to highlight symmetries and other relevant peculiarities. Our analysis exposes several similarities between the applications–namely the utilization of common low-dimensional stencils, and the use of a small set of collective primitives, in particular all-reduces with small vectors. Overall, our study provides a better understanding of the communication characteristics of several important scientific applications and benchmarks.

# Privacy Protection in Service Discovery for Large-scale Distributed Computing Systems

Jun Yeol Choi, Zhong Yuan Li, and
Hee Yong Youn
School of Information and Communication Engineering
Sungkyunkwan University, Suwon, Korea
trustcjy@gmail.com, lzy.source@gmail.com,
youn@ece.skku.ac.kr
Ohyoung Song
School of Electrical and Electronics Engineering
Chung-Ang University
Seoul 156-756, Republic of Korea
song@cau.ac.kr

## Abstract

Service discovery is one of the integral tasks of the system running in distributed computing environment. Even though numerous researches on service discovery have been conducted, the security and privacy issue have not yet been addressed adequately. Here letting only legitimate user and service provider for service discovery is a challenging problem. In this paper we present a scheme for secure service discovery by employing an incremental progressive exposure approach, where the user and service provider exchange only partial encrypted information at each step. They compare the information sent from the other party with their own information. The exchange operation continues until a mismatch occurs or the required service is found finally. Computer simulation reveals that the proposed approach significantly reduces the number of message exchanges compared to the existing progressive exposure scheme.

# Workshop 10

# International Workshop on Data-Intensive Computing in the Clouds

# DataCloud 2011

# Towards Scalable One-Pass Analytics Using MapReduce

Edward Mazur, Boduo Li, Yanlei Diao, Prashant Shenoy
Department of Computer Science
University of Massachusetts, Amherst
Email: {mazur, boduo, yanlei, shenoy}@cs.umass.edu

## Abstract

An integral part of many data-intensive applications is the need to collect and analyze enormous datasets efficiently. Concurrent with such application needs is the increasing adoption of MapReduce as a programming model for processing large datasets using a cluster of machines. Current MapReduce systems, however, require the data set to be loaded into the cluster before running analytical queries, and thereby incur high delays to start query processing. Furthermore, existing systems are geared towards batch processing.

In this paper, we seek to answer a fundamental question: what architectural changes are necessary to bring the benefits of the MapReduce computation model to incremental, one-pass analytics, i.e., to support stream processing and online aggregation? To answer this question, we first conduct a detailed empirical performance study of current MapReduce implementations including Hadoop and MapReduce Online using a variety of workloads. By doing so, we identify several drawbacks of existing systems for one-pass analytics. Based on the insights from our study, we conclude by listing key design requirements and arguing for architectural changes of MapReduce systems to overcome their current limitations and fully embrace incremental one-pass analytics.

# Transactions a la carte - Implementation and Performance Evaluation of Transactional Support on top of Amazon S3

Francis Gropengießer, Kai-Uwe Sattler
Department of Computer Science and Automation
Ilmenau University of Technology, Germany
first.last@tu-ilmenau.de

## Abstract

Most providers of Cloud-based database services favor availability and processing performance over consistency and hence provide only limited transaction support. Guaranteeing strict consistency is left to application developers. In this paper, we present a flexible transactional framework built on top of Amazon S3. It is intended for processing structured data under consideration of user-defined ACID guarantees ranging from non-transactional data querying and updating to full transactional data processing ensuring strict ACID properties. It supports the development of Cloud-based database applications without explicitly writing code for transactions. Instead, transactional semantics are weaved-in automatically using aspect-oriented programming techniques. In this way, application developers as well as end-users profit from transactional guarantees without having to care about them.

# XML Processing in the Cloud: Large-Scale Digital Preservation in Small Institutions

Peter Wittek[*], Thierry Jacquin, Hervé Déjean, Jean-Pierre Chanod[†], and Sándor Darányi[*]
[*]Swedish School of Library and Information Science
University of Bors
Bors, Sweden
Email: peterwittek@acm.org,sandor.daranyi@hb.se
[†]Xerox Research Centre Europe
Meylan, France
Email: {thierry.jacquin,herve.dejean,jean-pierre.chanod}@xrce.xerox.com

## Abstract

Digital preservation deals with the problem of retaining the meaning of digital information over time to ensure its accessibility. The process often involves a workflow which transforms the digital objects. The workflow defines document pipelines containing transformations and validation checkpoints, either to facilitate migration for persistent archival or to extract metadata. The transformations, nevertheless, are computationally expensive, and therefore digital preservation can be out of reach for an organization whose core operation is not in data conservation. The operations described the document workflow, however, do not frequently reoccur. This paper combines an implementation-independent workflow designer with cloud computing to support small institution in their ad-hoc peak computing needs that stem from their efforts in digital preservation.

# An adaptive framework for the execution of data-intensive MapReduce applications in the Cloud

Martin Koehler, Yuriy Kaniovskyi, Siegfried Benkner
University of Vienna
Faculty of Computer Science
Department of Scientific Computing
{koehler,yk,sigi}@par.univie.ac.at

## Abstract

Cloud computing technologies play an increasingly important role in realizing data-intensive applications by offering a virtualized compute and storage infrastructure that can scale on demand. A programming model that has gained a lot of interest in this context is MapReduce, which simplifies processing of large-scale distributed data volumes, usually on top of a distributed file system layer. In this paper we report on a self-configuring adaptive framework for developing and optimizing data-intensive scientific applications on top of Cloud and Grid computing technologies and the Hadoop framework. Our framework relies on a MAPE-K loop, known from autonomic computing, for optimizing the configuration of data-intensive applications at three abstraction layers: the application layer, the MapReduce layer, and the resource layer. By evaluating monitored resources, the framework configures the layers and allocates the resources on a per job basis. The evaluation of configurations relies on historic data and a utility function that ranks different configurations regarding to the arising costs. The optimization framework has been integrated in the Vienna Grid Environment (VGE), a service-oriented application development environment for providing applications on HPC systems, clusters and Clouds as services. An experimental evaluation of our framework has been undertaken with a data-analysis application from the field of molecular systems biology.

# The Gfarm File System on Compute Clouds

Kenji KOBAYASHI, Shunsuke MIKAMI, Hiroki KIMURA, Osamu TATEBE
Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, JAPAN
{kobayashi, mikami, hiroki, tatebe}@hpcs.cs.tsukuba.ac.jp

## Abstract

Due to its ability to provide requested resources on demand, the use of cloud computing for data-intensive computing is expected to skyrocket in the coming years. The present IaaS cloud infrastructure is designed such that the compute cloud and the storage cloud are separate components. However, the design has one major problem, namely a trade-off has to be made between the network cost and storage performance. To cope with these problems, we are proposing high performance file system service in compute cloud that transparently utilizes file access locality on dynamic configuration of cloud computing infrastructure. We evaluated the system using micro benchmarks, MapReduce applications, and MPI-IO applications. Our system shows the scalable file I/O performance.

# Evaluating Adaptive Compression to Mitigate the Effects of Shared I/O in Clouds

Matthias Hovestadt, Odej Kao, Andreas Kliem and Daniel Warneke
Technische Universität Berlin
Email: firstname.lastname@tu-berlin.de

## Abstract

IaaS clouds have become a promising platform for scalable distributed systems in recent years. However, while the virtualization techniques of such clouds are key to the cloud's elasticity, they also result in a reduced and less predictable I/O performance compared to traditional HPC setups. Besides the regular performance degradation of virtualized I/O itself, it is also the potential loss of I/O bandwidth through co-located virtual machines that imposes considerable obstacles for porting data-intensive applications to that platform. In this paper we examine adaptive compression schemes as a means to mitigate the negative effects of shared I/O in IaaS clouds. We discuss the decision models of existing schemes and analyze their applicability in virtualized environments. Based on an evaluation using XEN, KVM, and Amazon EC2, we found that most decision metrics (like CPU utilization and I/O bandwidth) are displayed inaccurately inside virtual machines and can lead to unreasonable levels of compression. As a remedy, we present a new adaptive compression scheme for virtualized environments which solely considers the application data rate. Without requiring any calibration or training phase our adaptive compression scheme can improve the I/O throughput of virtual machines significantly as shown through experimental evaluation.

# Performance Evaluation of Range Queries in Key Value Stores

Pouria Pirzadeh
Department of Computer Science
University of California, Irvine
Irvine, USA
Email: pouria@uci.edu

Junichi Tatemura, Hakan Hacigumus
NEC Laboratories America, Inc.
Cupertino, USA
Email: {tatemura, hakan}@sv.nec-labs.com

## Abstract

Recently there has been a considerable increase in the number of different Key-Value stores, for supporting data storage and applications on the cloud environment. While all these No-SQL solutions try to offer highly available and scalable services on the cloud, they are significantly different with each other in terms of the architecture and types of the applications, they try to support. Considering three widely-used such systems: Cassandra, HBase and Voldemort; in this paper we compare them in terms of their support for different types of query workloads. We are mainly focused on the range queries. Unlike HBase and Cassandra that have built-in support for range queries, Voldemort does not support this type of queries via its available API. For this matter, practical techniques are presented on top of Voldemort to support range queries. Our performance evaluation is based on mixed query workloads, in the sense that they contain a combination of short and long range queries, beside other types of typical queries on key-value stores such as lookup and update. We show that there are trade-offs in the performance of the selected system and scheme, and the types of the query workloads that can be processed efficiently.

# CloudClustering: Toward an iterative data processing pattern on the cloud

Ankur Dave
University of California, Berkeley
Berkeley, California, USA
ankurd@eecs.berkeley.edu

Wei Lu, Jared Jackson, Roger Barga
Microsoft Research
Redmond, Washington, USA
{weilu,jaredj,barga}@microsoft.com

## Abstract

As the emergence of cloud computing brings the potential for large-scale data analysis to a broader community, architectural patterns for data analysis on the cloud, especially those addressing iterative algorithms, are increasingly useful. MapReduce suffers performance limitations for this purpose as it is not inherently designed for iterative algorithms.

In this paper we describe our implementation of CloudClustering, a distributed k-means clustering algorithm on Microsoft's Windows Azure cloud. The k-means algorithm makes a good case study because its characteristics are representative of many iterative data analysis algorithms. CloudClustering adopts a novel architecture to improve performance without sacrificing fault tolerance. To achieve this goal, we introduce a distributed fault tolerance mechanism called the buddy system, and we make use of data affinity and checkpointing. Our goal is to generalize this architecture into a pattern for large-scale iterative data analysis on the cloud.

# Towards intelligent data placement for scientific workflows in collaborative cloud environment

Xin Liu
School of Computer Engineering
Nanyang Technological University
Singapore
Email: liu_xin@pmail.ntu.edu.sg

Anwitaman Datta
School of Computer Engineering
Nanyang Technological University
Singapore
Email: anwitaman@ntu.edu.sg

## Abstract

Recently emerged cloud computing offers a promising platform for executing scientific workflow applications due to its similar performance compared to the grid, lower cost, elasticity and so on. Collaborative cloud environments, which share resources of multiple geographically distributed data centers owned by different organizations enable researchers from all over the world to conduct their large scale data intensive research together through Internet. However, since scientific workflows consume and generate huge amount of data, it is thus essential to manage the data effectively for the purpose of high performance and cost effectiveness. In this paper, we propose intelligent data placement strategy to improve performance of workflows while minimizing data transfer among data centers. Specifically, at the startup stage, the whole dataset is divided into small data items which are then distributed among multiple data centers by considering these data centers' computation capability, storage budget, data item correlation, etc. During the runtime stage, when intermediate data is generated, it is placed on the suitable data centers using linear discriminant analysis by taking into account the same metrics as at the startup stage, as well as data centers' past behaviors (i.e., trustworthiness in terms of task delay). Simulation results demonstrate the promise of our data placement strategy by showing that compared to existing data placement strategies, our proposal effectively places the data to improve computation progress on the whole while minimizing the communication overheads incurred by data movement.

# iMapReduce: A Distributed Computing Framework for Iterative Computation

Yanfeng Zhang[*], Qinxin Gao[*], Lixin Gao[†], Cuirong Wang[*]
[*]Northeastern University, China
[†]University of Massachusetts Amherst, USA
Email: threewells14@gmail.com, gaoqx@mail.neuq.edu.cn, lgao@ecs.umass.edu, wangcr@mail.neuq.edu.cn

## Abstract

Relational data are pervasive in many applications such as data mining or social network analysis. These relational data are typically massive containing at least millions or hundreds of millions of relations. This poses demand for the design of distributed computing frameworks for processing these data on a large cluster. MapReduce is an example of such a framework. However, many relational data based applications typically require parsing the relational data iteratively and need to operate on these data through many iterations. MapReduce lacks built-in support for the iterative process. This paper presents iMapReduce, a framework that supports iterative processing. iMapReduce allows users to specify the iterative operations with map and reduce functions, while supporting the iterative processing automatically without the need of users' involvement. More importantly, iMapReduce significantly improves the performance of iterative algorithms by (1) reducing the overhead of creating a new task in every iteration, (2) eliminating the shuffling of the static data in the shuffle stage of MapReduce, and (3) allowing asynchronous execution of each iteration, *i.e.,* an iteration can start before all tasks of a previous iteration have finished. We implement iMapReduce based on Apache Hadoop, and show that iMapReduce can achieve a factor of 1.2 to 5 speedup over those implemented on MapReduce for well-known iterative algorithms.

# Designing the Cloud-based DOE Systems Biology Knowledgebase

Carina Lansing, Yan Liu, Jian Yin, Abigail Corrigan, Zoe Guillen, Kerstin Kleese van Dam, Ian Gorton
Fundamental and Computational Science Division
Pacific Northwest National Laboratory
Richland, WA, USA

## Abstract

Systems Biology research, even more than many other scientific domains, is becoming increasingly data-intensive. Not only have advances in experimental and computational technologies lead to an exponential increase in scientific data volumes and their complexity, but increasingly such databases are providing the basis for new scientific discoveries. To engage effectively with these community resources, integrated analyses, synthesis and simulation software is needed, supported by scientific workflows. In order to provide a more collaborative, community driven research environment for this heterogeneous setting, the Department of Energy (DOE) has decided to develop a federated, cloud based cyber infrastructure the Systems Biology Knowledgebase (Kbase). In this context the Pacific Northwest National Laboratory (PNNL) has been defining and testing the basic federated cloud-based system architecture and developed a prototype implementation. Community wide accessibility of biological data and the capability to integrate and analyze this data within its changing research context were seen as key technical functionalities the Kbase needs to enable. In this paper we describe the results of our investigations into the design of this cloud based federated infrastructure for:

- Semantics driven data discovery, access and integration
- Data annotation, publication and sharing
- Workflow enabled data analysis
- Project based collaborative working

We describe our approach, exemplary use cases and our prototype implementation that demonstrates the feasibility of this approach.

# Workshop 11

# Workshop on High-Level Parallel Programming Models & Supportive Environments

# HIPS 2011

# Towards a Methodology for Deliberate Sample-Based Statistical Performance Analysis

Geoff Stoker and Jeffrey K. Hollingsworth
University of Maryland
College Park, Maryland
gstoker, hollings@cs.umd.edu

## Abstract

Dynamic performance analysis of long-running programs in the high performance computing community increasingly relies on statistical profiling techniques to provide performance measurement results. Systematic sampling rates used to generate the statistical data are typically selected in an ad hoc manner with little formal regard for the context provided by the program being analyzed and the underlying system on which it is run. In an effort to provide a more effective statistical profiling process and additional rigor we argue in favor of the general principle of deliberate sampling rate selection. We present our idea for a methodology of systematic sample rate selection based on a performance measurement model incorporating the effect of sampling on both measurement precision and perturbation effects.

# Parallel Object Contracts for High Performance Computing

Kostadin Damevski
Virginia State University
Petersburg, VA 23806
Email: kdamevski@vsu.edu

Tamara Dahlgren
Lawrence Livermore National Laboratory
Livermore, CA 94550
Email: dahlgren1@llnl.gov

## Abstract

Modern high-performance computing applications are increasingly complex due to the nature of the problems being solved and the often unique, parallel platforms on which they are ultimately executed. Improving the quality of these applications through well-established software engineering techniques, such as design-by-contract, is of growing importance. This paper introduces preliminary work pursuing new interface contract clauses for parallel programming. Existing contract capabilities in the Babel compiler, which is a language interoperability tool tailored for high-performance computing, will be extended with parallel contract constructs for synchronization and method order sequencing. The goal of this work is to enhance existing contract mechanisms without imposing significant synchronization overhead. Preliminary results from an initial investigation of collective and threading assertions are presented.

# A (radical) proposal addressing the non-scalability of the irregular MPI collective interfaces

Jesper Larsson Träff
Department of Scientific Computing, Faculty of Computer Science
University of Vienna
Nordbergstrasse 15/3C, A-1090 Vienna, Austria
Email: traff@par.univie.ac.at

## Abstract

In both the regular and the irregular MPI collectives there is a correspondence between the argument lists and certain MPI derived datatypes. As a means to address and alleviate well-known memory and performance scalability problems in the *irregular* (or *vector*) collective interface definitions of MPI we propose to push this correspondence to its natural limit, and replace the interfaces of the MPI collectives with a new set of interfaces that specify all data sizes and displacements solely by means of derived datatypes. This reduces the number of collective (communication and reduction) interfaces from 16 to 10 (in the extreme to only 7), significantly generalizes the operations, unifies regular and irregular collective interfaces, makes it possible to decouple certain algorithmic decisions from the collective operation, and moves the interface scalability issue from the collective interfaces to the MPI derived datatypes. To complete the proposal we discuss the memory scalability of the derived datatypes and suggest a number of new datatypes for MPI, some of which may be of independent interest. Finally, we introduce additional, new functionality for scalable construction of datatypes as needed for the new collective interfaces. We give examples throughout of the benefits of this alternative set of collective interfaces. It is argued that an implementation can straightforwardly be undertaken within any reasonable MPI implementation.

# Auto-Generation of Parallel Finite-Differencing Code for MPI, TBB and CUDA

D. P. Playne
Computer Science, IIMS
Massey University
Auckland, New Zealand
d.p.playne@massey.ac.nz

K. A. Hawick
Computer Science, IIMS
Massey University
Auckland, New Zealand
k.a.hawick@massey.ac.nz

## Abstract

Finite-difference methods can be useful for solving certain partial differential equations (PDEs) in the time domain. Compiler technologies can be used to parse an application domain specific representation of these PDEs and build an abstract representation of both the equation and the desired solver. This abstract representation can be used to generate a language-specific implementation. We show how this framework can be used to generate software for several parallel platforms: Message Passing Interface (MPI), Threading Building Blocks(TBB) and Compute Unified Device Architecture(CUDA). We present performance data of the automatically-generated parallel code and discuss the implications of the generator in terms of code portability, development time and maintainability.

# Translating Chapel to Use FREERIDE: A Case Study in Using a HPC Language for Data-Intensive Computing

Bin Ren Gagan Agrawal Brad Chamberlain[†] Steve Deitz[†]
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210
{ren, agrawal}@cse.ohio-state.edu Cray Inc.
901, Fifth Avenue, Suite 1000
Seattle, WA 98164
bradc@cray.com, stdeitz@microsoft.com

## Abstract

In the last few years, the growing significance of data-intensive computing has been closely tied to the emergence and popularity of new programming paradigms for this class of applications, including Map-Reduce, and new high-level languages for data-intensive computing. The ultimate goal of these efforts in data-intensive computing has been to achieve parallelism with as little effort as possible, while supporting high efficiency and scalability. While these are also the goals that the parallel language/compiler community has tried meeting for the past several decades, the development of languages and programming systems for data-intensive computing has largely been in isolation to the developments in general parallel programming.

Such independent developments in the two areas, i.e., data-intensive computing and high productivity languages lead to the following questions: I) Are HPC languages suitable for expressing data-intensive computations? and if so, II.a) What are the issues in using them for effective parallel programming? or, if not, II.b) What characteristics of data-intensive computations force the need for separate language support?.

This paper takes a case study to address these questions. Particularly, we study the suitability of Chapel for expressing data-intensive computations. We also examine compilation techniques required for directly invoking a data-intensive middleware from Chapel's compilation system. The data-intensive middleware we use in this effort is FREERIDE that has been developed at Ohio State. We show how certain transformations can enable efficient invocation of the FREERIDE functions from the Chapel compiler. Our experiments show that after certain optimizations, the performance of the version of Chapel compiler that invokes FREERIDE functions is quite comparable to the performance of hand-written data-intensive applications.

# Using Shared Arrays in Message-Driven Parallel Programs

Phil Miller Aaron Becker Laxmikant Kalé
Parallel Programming Laboratory
Department of Computer Science
University of Illinois at Urbana-Champaign
Email: {mille121,abecker3,kale}@illinois.edu

## Abstract

This paper describes a safe and efficient combination of the object-based message-driven execution and shared array parallel programming models. In particular, we demonstrate how this combination engenders the composition of loosely coupled parallel modules safely accessing a common shared array. That loose coupling enables both better flexibility in parallel execution and greater ease of implementing multi-physics simulations. As a case study, we describe how the parallelization of a new method for molecular dynamics simulation benefits from both of these advantages. We also describe a system of typed handle objects that embed some of the determinacy constraints of the Multiphase Shared Array programming model in the C++ type system, to catch some violations at compile time. The combined programming model communicates in terms of these handles as a natural means of detecting and preventing errors.

# Monitoring and Debugging DryadLINQ Applications with Daphne

Vilas Jagannath, Zuoning Yin, Mihai Budiu
vbangal2@illinois.edu, zyin2@uiuc.edu, mbudiu@microsoft.com

## Abstract

Debugging and optimizing large-scale applications is still more art than engineering discipline. This document describes our experience in building a set of tools to help DryadLINQ application developers understand and debug their programs.

The core infrastructure for our tools is a portable library which provides a DryadLINQ job object model (i.e., a local representation of the distributed state of an executed application). Layered on the job object model we have built a variety of interactive and batch tools for: performance data collection and analysis, distributed state visualization, failure diagnostics, debugging, and profiling.

# SkelCL - A Portable Skeleton Library for High-Level GPU Programming

Michel Steuwer, Philipp Kegel, and Sergei Gorlatch
Department of Mathematics and Computer Science
University of Münster, Münster, Germany
Email: {michel.steuwer,philipp.kegel,gorlatch}@uni-muenster.de

## Abstract

While CUDA and OpenCL made general-purpose programming for Graphics Processing Units (GPU) popular, using these programming approaches remains complex and error-prone because they lack high-level abstractions. The especially challenging systems with multiple GPU are not addressed at all by these low-level programming models. We propose SkelCL – a library providing so-called algorithmic skeletons that capture recurring patterns of parallel computation and communication, together with an abstract vector data type and constructs for specifying data distribution. We demonstrate that SkelCL greatly simplifies programming GPU systems. We report the competitive performance results of SkelCL using both a simple Mandelbrot set computation and an industrial-strength medical imaging application. Because the library is implemented using OpenCL, it is portable across GPU hardware of different vendors.

# Toolchain for Programming, Simulating and Studying the XMT Many-Core Architecture

Fuat Keceli, Alexandros Tzannes, George C. Caragea, Rajeev Barua and Uzi Vishkin
University of Maryland, College Park
keceli,tzannes,gcaragea,barua,vishkin@umd.edu

## Abstract

The Explicit Multi-Threading (XMT) is a general-purpose many-core computing platform, with the vision of a 1000-core chip that is easy to program but does not compromise on performance. This paper presents a publicly available toolchain for XMT, complete with a highly configurable cycle-accurate simulator and an optimizing compiler.

The XMT toolchain has matured and has been validated to a point where its description merits publication. In particular, research and experimentation enabled by the toolchain played a central role in supporting the ease-of-programming and performance aspects of the XMT architecture. The compiler and the simulator are also important milestones for an efficient programmer's workflow from PRAM algorithms to programs that run on the shared memory XMT hardware. This workflow is a key component in accomplishing the dual goal of ease-of-programming and performance.

The applicability of our toolchain extends beyond specific XMT choices. It can be used to explore the much greater design space of shared memory many-cores by system researchers or by programmers. As the toolchain can practically run on any computer, it provides a supportive environment for teaching parallel algorithmic thinking with a programming component. Unobstructed by techniques such as decomposition-first and programming for locality, this environment may be useful in deferring the teaching of these techniques, when desired, to more advanced or platform-specific courses.

## An Optimized Reduction Design to Minimize Atomic Operations in Shared Memory Multiprocessors

Ettore Speziale [†] [1]
Dipartimento di Elettronica
ed Informazione
Politecnico di Milano
Milano, Italy
speziale@elet.polimi.it

Andrea di Biagio [†]
Dipartimento di Elettronica
ed Informazione
Politecnico di Milano
Milano, Italy
dibiagio@elet.polimi.it

Giovanni Agosta
Dipartimento di Elettronica
ed Informazione
Politecnico di Milano
Milano, Italy
agosta@elet.polimi.it

### Abstract

Reduction operations play a key role in modern massively data parallel computation. However, current implementations in shared memory programming APIs such as OpenMP are often computation bottlenecks due to the high number of atomic operations involved. We propose a reduction design that exploits the coupling with a barrier synchronization to optimize the execution of the reduction. Experimental results show how the number of atomic operations involved is dramatically reduced, which can lead to significant improvement in scaling properties on large numbers of processing elements. We report a speedup of 1.53x on the *312.swim_m* SPEC OMP2001 benchmark and a speedup of 4.02x on the *streamcluster* benchmark from the PARSEC suite over the baseline.

## Automatically Inserting Synchronization Statements in Divide-and-Conquer Programs

Pieter Hijma, Rob V. van Nieuwpoort, Ceriel J.H. Jacobs, and Henri E. Bal
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
Email: {pieter,rob,ceriel,bal}@cs.vu.nl

### Abstract

Divide-and-conquer is a well-known and important programming model that supports efficient execution of parallel applications on multi-cores, clusters, and grids. In a divide-and-conquer system such as Satin or Cilk, recursive calls are automatically transformed into jobs that execute asynchronously. Since the calls are non-blocking, consecutive calls are the source of parallelism. However, the programmer has to manually enforce synchronization with `sync` statements that indicate where the system has to wait for the result of the asynchronous jobs.

In this paper, we investigate the possibility to automatically insert `sync` statements to relieve the programmer of the burden of thinking about synchronization. We investigate whether correctness can be guaranteed and to what extent the amount of parallelism is reduced. We discuss the code analysis algorithms that are needed in detail.

To evaluate our approach, we have extended the Satin divide-and-conquer system, which targets efficient execution on grids, with a sync generator. The fact that Satin uses Java as a base language helps the sync generator to reason about control flow and aliasing of references to objects. Our experiments show that, with our analysis, we can automatically generate synchronization points in virtually all real-life cases: in 31 out of 35 real-world applications the `sync` statements are placed optimally. The automatic placement is correct in all cases, and in one case the sync generator corrected synchronization errors in an application (FFT).

# Exploiting Hierarchical Parallelism Using UPC

Lingyuan Wang, Saumil Merchant, and Tarek El-Ghazawi
Department of Electrical and Computer Engineering, The George Washington University
{lwanghpc, smerchan, tarek}@gwu.edu

## Abstract

High-Performance Computing (HPC) systems are increasingly moving towards an architecture that is deeply hierarchical. However, the execution model with single-level parallelism embodied in legacy parallel programming models falls short in exploiting the multi-level parallelism opportunities in both hardware architectures and applications. This makes the use of richer execution models imperative in order to fully exploit hierarchical parallelism.

Partitioned Global Address Space (PGAS) languages such as Unified Parallel C (UPC) are growing in popularity because of their ability to provide a globally shared address space with locality awareness. While UPC provides a welcome improvement over message passing libraries, users still program with a single level of parallelism in the context of SPMD. In this paper, we explore two explicit hierarchical programming approaches based on UPC to improve programmability and performance on hierarchical architectures. The first approach orchestrates computations on multiple sets of thread groups; the second approach extends UPC with nested, shared memory multi-threading. This paper presents a detailed description of proposed approaches and demonstrates their effectiveness in the context of the NAS Parallel Benchmarks and the Unbalanced Tree Search (UTS). Experimental results indicate that the hierarchical model not only provides greater expressive power but also enhances performance; all three benchmarks exceed the performance of the standard UPC implementations after being incrementally enhanced with hierarchical parallelism.

# Preserving Collective Performance Across Process Failure for a Fault Tolerant MPI

Joshua Hursey, Richard L. Graham
Oak Ridge National Laboratory, Oak Ridge, TN USA 37831
Email: {hurseyjj,rlgraham}@ornl.gov

## Abstract

Application developers are investigating Algorithm Based Fault Tolerance (ABFT) techniques to improve the efficiency of application recovery beyond what traditional techniques alone can provide. Applications will depend on libraries to sustain failure-free performance across process failure to continue to efficiently use High Performance Computing (HPC) systems even in the presence of process failure. Optimized Message Passing Interface (MPI) collective operations are a critical component of many scalable HPC applications. However, most of the collective algorithms are not able to handle process failure. Next generation MPI implementations must provide fault aware versions of such algorithms that can sustain performance across process failure. This paper discusses the design and implementation of fault aware collective algorithms for tree structured communication patterns. The three design approaches of rerouting, lookup avoiding and rebalancing are described, and analyzed for their performance impact relative to a similar fault unaware collective algorithm. The analysis shows that the rerouting approach causes up to a four times performance degradation while the rebalancing approach can bring the performance within 1% of the fault unaware performance. Additionally, this paper introduces the reader to a set of run-through stabilization semantics being developed by the MPI Forum's Fault Tolerance Working Group to support ABFT. This paper underscores the need for care to be taken when designing new fault aware collective algorithms for fault tolerant MPI implementations.

## Data Centric Techniques for Mapping Performance Measurements

Nick Rutar and Jeffrey K. Hollingsworth
Computer Science Department
University of Maryland
College Park, MD 20742, USA
{rutar,hollings}@cs.umd.edu

### Abstract

Traditional methods of performance analysis offer a code centric view, presenting performance data in terms of blocks of contiguous code (statement, basic block, loop, function). Data centric techniques, combined with hardware counter information, allow various program properties including cache misses and cycle count to be mapped directly to variables. We introduce mechanisms for efficiently collecting data centric performance numbers independent of hardware support. We create extended data centric mappings, which we call variable blame, that relates data centric information to high level data structures. Finally, we show performance data gathered from three parallel programs using our technique.

## Harnessing the Power of GPUs without Losing Abstractions in SAC and ArrayOL: A Comparative Study

Jing Guo[*] , Wendell Rodrigues[†] , Jeyarajan Thiyagalingam[‡] ,
Frédéric Guyomarc'h[†] ,
Pierre Boulet[†] , and Sven-Bodo Scholz[*]
[*] Department of Computer Science
University of Hertfordshire
Hatfield, United Kingdom
Email: j.guo@herts.ac.uk
[†] LIFL - USTL - INRIA Lille Nord Europe, France
[‡] Oxford e-Research Centre, University of Oxford
Oxford, United Kingdom

### Abstract

Over recent years, using Graphics Processing Units (GPUs) has become as an effective method for increasing the performance of many applications. However, these performance benefits from GPUs come at a price. Firstly extensive programming expertise and intimate knowledge of the underlying hardware are essential for gaining good speedups. Secondly, the expressibility of GPU-based programs are not powerful enough to retain the high-level abstractions of the solutions. Although the programming experience has been significantly improved by existing frameworks like CUDA and OpenCL, it is still a challenge to effectively utilise these devices while still retaining the programming abstractions. To this end, performing a source-to-source transformation, whereby a high-level language is mapped to CUDA or OpenCL, is an attractive option. In particular, it enables one to retain high-level abstractions and to harness the power of GPUs without any expertise on the GPGPU programming. In this paper, we compare and analyse two such schemes. One of them is a transformation mechanism for mapping a image/signal processing domain-specific language, ArrayOL, to OpenCL. The other one is a transformation route for mapping a high-level general purpose array processing language, Single Assignment C (SaC) to CUDA. Using a real-world image processing application as a running example, we demonstrate that albeit the fact of being general purpose, the array processing language be used to specify complex array access patterns generically. Performance of the generated CUDA code is comparable to the OpenCL code created from domain-specific language.

# The Essence of Synchronisation in Asynchronous Data Flow

Clemens Grelck
Institute of Informatics
University of Amsterdam
Amsterdam, Netherlands
Email: c.grelck@uva.nl

## Abstract

We discuss the aspect of synchronisation in the language design and implementation of the asynchronous data flow language S-Net. Synchronisation is a crucial aspect of any coordination approach. S-Net provides a particularly simple construct, the synchrocell. As a primitive S-Net language construct synchrocell implements a one-off synchronisation of two data items of different type on a stream of such data items. We believe this semantics captures the essence of synchronisation, and no simpler design is possible.

While the exact built-in behaviour as such is typically not what is required by S-Net application programmers, we show that in conjunction with other language features S-Net synchrocells meet typical demands for synchronisation in streaming networks quite well. Moreover, we argue that their simplistic design, in fact, is a necessary prerequisite to implement an even more interesting scenario: modelling state in streaming networks of stateless components. We finish with the outline of an efficient implementation by the S-Net runtime system.

# Large Volume Testing for Executable Formal Specification using Hadoop

Shigeru KUSAKABE
Grad. School of Info. Sci. & Electrical Engineering,
Kyushu University
Fukuoka, 819-0395, Japan
kusakabe@ait.kyushu-u.ac.jp

## Abstract

Formal methods are mathematically-based techniques for specifying, developing and verifying a component or system, in order to increase the confidence regarding to the reliability and robustness of the target. Formal methods can be used at different levels with different techniques, and one approach is to use model-oriented formal languages such as VDM languages in writing specifications. During model development, we can test executable specifications in VDM-SL and VDM++. In a lightweight formal approach, we test formal specifications to increase our confidence as we do in implementing software codes with conventional programming languages.

While the specific level of rigor depends on the aim of the project, millions of tests may be conducted in developing highly reliable mission-critical software in a light-weight formal approach.

In this paper, we introduce our approach to supporting large volume of testing for executable formal specifications using Hadoop, an implementation of MapReduce programming model. We are able to automatically distribute interpretation of specifications in VDM languages by using Hadoop. We also apply a property-based data-driven testing tool, QuickCheck, over MapReduce so that specification can be checked with thousands of tests by using formally described pre/post conditions and invariant functions. As an advanced property-based data-driven testing approach, we also discuss model-based testing. We observed scalable performance in testing large amount of data for executable specifications in our approaches.

## Address translation optimization for Unified Parallel C multi-dimensional arrays

Olivier Serres, Ahmad Anbar, Saumil G. Merchant, Abdullah Kayi and Tarek El-Ghazawi
NSF Center for High-Performance Reconfigurable Computing (CHREC),
Department of Electrical and Computer Engineering,
The George Washington University
{serres, anbar, apokayi}@gwmail.gwu.edu, {smerchan, tarek}@gwu.edu

### Abstract

Partitioned Global Address Space (PGAS) languages offer significant programmability advantages with its global memory view abstraction, one-sided communication constructs and data locality awareness. These attributes place PGAS languages at the forefront of possible solutions to the exploding programming complexity in the many-core architectures. To enable the shared address space abstraction, PGAS languages use an address translation mechanism while accessing shared memory to convert shared addresses to physical addresses. This mechanism is already expensive in terms of performance in distributed memory environments, but it becomes a major bottleneck in machines with shared memory support where the access latencies are significantly lower. Multi- and many-core processors exhibit even lower latencies for shared data due to on-chip cache space utilization. Thus, efficient handling of address translation becomes even more crucial as this overhead may easily become the dominant factor in the overall data access time for such architectures. To alleviate address translation overhead, this paper introduces a new mechanism targeting multi-dimensional arrays used in most scientific and image processing applications. Relative costs and the implementation details for UPC are evaluated with different workloads (matrix multiplication, Random Access benchmark and Sobel edge detection) on two different platforms: a many-core system, the TILE64 (a 64 core processor) and a dual-socket, quad-core Intel Nehalem system (up to 16 threads). Our optimization provides substantial performance improvements, up to 40x. In addition, the proposed mechanism can easily be integrated into compilers abstracting it from the programmers. Accordingly, this improves UPC productivity as it will reduce manual optimization efforts required to minimize the address translation overhead.

## Explicit Platform Descriptions for Heterogeneous Many-Core Architectures

Martin Sandrieser, Siegfried Benkner and Sabri Pllana
Department of Scientific Computing
Faculty of Computer Science - University of Vienna, Austria
Email: {ms, sigi, pllana}@par.univie.ac.at

### Abstract

Heterogeneous many-core architectures offer a way to cope with energy consumption limitations of various computing systems from small mobile devices to large data-centers. However, programmers typically must consider a large diversity of architectural information to develop efficient software. In this paper we present our ongoing work towards a Platform Description Language (PDL) that enables to capture key architectural patterns of commonly used heterogeneous computing systems. PDL architecture patterns support programmers and toolchains by providing platform information in a well-defined and explicit manner. We have developed a source-to-source compiler that utilizes PDL descriptors to transform sequential task-based programs to a form that is convenient for execution on heterogeneous many-core computing systems. We show various usage scenarios of our PDL and demonstrate our approach for a commonly used scientific kernel.

# DAGuE: A generic distributed DAG engine for high performance computing

George Bosilca[*], Aurelien Bouteiller[*], Anthony Danalis[*†], Thomas Herault[*], Pierre Lemarinier[*], Jack Dongarra[*
[*]University of Tennessee Innovative Computing Laboratory
[†]Oak Ridge National Laboratory
[‡]University Paris-XI

## Abstract

The frenetic development of the current architectures places a strain on the current state-of-the-art programming environments. Harnessing the full potential of such architectures has been a tremendous task for the whole scientific computing community.

We present DAGuE a generic framework for architecture aware scheduling and management of micro-tasks on distributed many-core heterogeneous architectures. Applications we consider can be represented as a Direct Acyclic Graph of tasks with labeled edges designating data dependencies. DAGs are represented in a compact, problem-size independent format that can be queried on-demand to discover data dependencies, in a totally distributed fashion. DAGuE assigns computation threads to the cores, overlaps communications and computations and uses a dynamic, fully-distributed scheduler based on cache awareness, data-locality and task priority. We demonstrate the efficiency of our approach, using several micro-benchmarks to analyze the performance of different components of the framework, and a Linear Algebra factorization as a use case.

# Workshop 12

# Workshop on Parallel and Distributed Scientific and Engineering Computing

# PDSEC 2011

# The Challenges of Programming for Exascale Class Machines (keynote)

Robert L. Clay
Scalable Modeling and Analysis Systems
Sandia National Laboratories
MS 9155, P.O. Box 969, Livermore California 94551
e-mail: rlclay@sandia.gov

## Abstract

As high-performance computing (HPC) advances toward exascale-class machines, there is increasing realization within the HPC community that major challenges lay ahead. These challenges span not only the system architecture and hardware, but also the entire software stack, including the applications. In particular, several key assumptions that are central to the current HPC application programming paradigm will not hold at this scale. In this talk we examine some of the key challenges from the perspective of the application programmer, and discuss ways to address or at least mitigate them. Hopefully, this will be somewhat of a 'call to arms' for application developers who wish to prepare for the coming wave of exascale-class HPC systems.

# Hyperspectral Data Processing in a High Performance Computing Environment: A Parallel Best Band Selection Algorithm

Stefan A. Robila, Gerald Busardo
Department of Computer Science
Montclair State University
Montclair, NJ, USA
robilas@mail.montclair.edu

## Abstract

Hyperspectral data are characterized by a richness of information unique among various visual representations of a scene by representing the information in a collection of grayscale images with each image corresponding to a narrow interval in the electromagnetic spectrum. Such detail allows for precise identification of materials in the scene and promises to support advances in imaging beyond the visible range. However, hyperspectral data are considerably large and cumbersome to process and efficient computing solutions based on high performance computing are needed. In this paper we first provide an overview of hyperspectral data and the current state of the art in the use of HPC for its processing. Next we discuss the concept of best band selection, a fundamental feature extraction problem in hyperspectral imagery that, besides exhaustive search has only non optimal solutions. We provide an elegant algorithm that performs an exhaustive search for the solution using a distributed, multicore environment and MPI in order to show how using such a solution provides significant improvement over traditional sequential platforms. Additional experiments on the robustness of the algorithm in terms of data and job sizes are also provided.

# Profiling Methodology and Performance Tuning of the Unified Model for Weather and Climate Simulations

Peter E. Strazdins[*] , Margaret Kahn[†] , Joerg Henrichs[‡] , Tim Pugh[§] , Mike Rezny[¶]
[*] School of Computer Science, The Australian National University,
Email: Peter.Strazdins@cs.anu.edu.au
[†] NCI National Facility, Australia,
Email: Margaret.Kahn@anu.edu.au
[‡] Oracle,
Email: joerg.henrichs@oracle.com
[§] Bureau of Meteorology, Australia
Email: T.Pugh@bom.gov.au
[¶] Monash Weather and Climate, Monash University,
Email: Michael.Rezny@monash.edu

## Abstract

Global weather and climate modelling is a compute-intensive task that is mission-critical to government departments concerned with meteorology and climate change. The dominant component of these models is a global atmosphere model. One such model, the Met Office Unified Model (MetUM), is widely used in both Europe and Australia for this purpose.

This paper describes our experiences in developing an efficient profiling methodology and scalability analysis of the MetUM version 7.5 at both low scale and high scale atmosphere grid resolutions. Variability within the execution of the MetUM and variability of the run-time of identical jobs on a highly shared cluster are taken into account. The methodology uses a lightweight profiler internal to the MetUM which we have enhanced to have minimal overhead and enables accurate profiling with only a relatively modest usage of processor time.

At high-scale resolution, the MetUM scaled to core counts of 2048, with load imbalance accounting a significant fraction the loss from ideal performance. Recent patches have removed two relatively small sources of inefficiency.

Internal segment size parameters gave a modest performance improvement at low-scale resolution (such as are used in climate simulation); this however was not significant a higher scales. Near-square process grid configurations tended to give the best performance. Byte-swapping optimizations vastly improved I/O performance, which has in turn a large impact on performance in operational runs.

# A Fast Heuristic for Scheduling Parallel Software with Respect to Energy and Timing Constraints

Margarete Sackmann, Peter Ebraert, Dirk Janssens
Ansymo (Antwerp Systems and Software Modeling)
Universiteit Antwerpen
Antwerp, Belgium
Email: {margarete.sackmann, peter.ebraert, dirk.janssens}@ua.ac.be

## Abstract

Embedded systems with several heterogeneous processors require scheduling techniques that are able to make use of the parallel processors while at the same time keeping resource consumption low. In this paper, we therefore introduce a scheduling algorithm that considers timing and energy requirements for computation and communication and allows deadlines on program parts. The software applications are represented by Synchronous Dataflow Graphs (SDF) as these allow one to represent software concurrency and enable effective scheduling on multi-processor platforms. We suggest a scheduling heuristic that generates a number of possible schedules for an SDF representation of a parallelizable program, ranging from fast schedules with high energy consumption to slow schedules with lower energy consumption. Besides being able to choose only schedules that satisfy a specific global deadline or do not exceed a certain energy consumption, we allow individual constraints on single nodes of the SDF graph. The problems arising from including such constraints into the scheduling and adapting the heuristic accordingly are explained. The performance of our scheduling heuristic is illustrated on randomly generated graphs.

# Deployment on GPUs of an application in computational atomic physics

P. Fortin[*], R. Habel[†], F. Jézéquel[*], J.-L. Lamotte[*] and N.S. Scott[‡]
[*]UPMC Univ Paris 06, UMR 7606, Laboratoire d'Informatique de Paris 6,
4 place Jussieu, 75252 Paris CEDEX 05, France
Email: {Pierre.Fortin,Fabienne.Jezequel,Jean-Luc.Lamotte}@lip6.fr
[†]Télécom SudParis, 9 rue Charles Fourier, 91011 Evry CEDEX, France
Email: Rachid.Habel@telecom-sudparis.eu
[‡]School of Electronics, Electrical Engineering & Computer Science,
The Queen's University of Belfast, Belfast BT7 1NN, UK
Email: ns.scott@qub.ac.uk

## Abstract

This paper describes the deployment on GPUs of PROP, a program of the 2DRMP suite which models electron collisions with H-like atoms and ions. Because performance on GPUs is better in single precision than in double precision, the numerical stability of the PROP program in single precision has been studied. The numerical quality of PROP results computed in single precision and their impact on the next program of the 2DRMP suite has been analyzed. Successive versions of the PROP program on GPUs have been developed in order to improve its performance. Particular attention has been paid to the optimization of data transfers and of linear algebra operations. Performance obtained on several architectures (including NVIDIA Fermi) are presented.

## The Large Scale Data Facility: Data Intensive Computing for scientific experiments

Ariel O. García, Serguei Bourov,
Ahmad Hammad, Jos van Wezel
Karlsruhe Institute of Technology (KIT)
Steinbuch Centre for Computing
Karlsruhe, Germany
Email: garcia@kit.edu

Volker Hartmann, Thomas Jejkal,
Patrick Neuberger, Rainer Stotzka
Karlsruhe Institute of Technology (KIT)
Institute for Data Processing and Electronics
Karlsruhe, Germany

### Abstract

The Large Scale Data Facility (LSDF) at the Karlsruhe Institute of Technology was started end of 2009 with the aim of supporting the growing requirements of data intensive experiments. In close cooperation with the involved scientific communities, the LSDF provides them not only with adequate storage space but with a directly attached analysis farm and –more importantly– with value added services for their big scientific data-sets. Analysis workflows are supported through the mixed Hadoop and OpenNebula Cloud environments directly attached to the storage, and enable the efficient processing of the experimental data. Metadata handling is a central part of the LSDF, where a metadata repository, community specific metadata schemes, graphical tools, and APIs were developed for accessing and efficiently organizing the stored data-sets.

## Efficiently Computing Tensor Eigenvalues on a GPU

Grey Ballard
UC Berkeley
Computer Science Department
Berkeley, CA
ballard@cs.berkeley.edu

Tamara Kolda and Todd Plantenga
Sandia National Laboratories
Livermore, CA
{tgkolda,tplante}@sandia.gov

### Abstract

The tensor eigenproblem has many important applications, generating both mathematical and application-specific interest in the properties of tensor eigenpairs and methods for computing them. A tensor is an $m$-way array, generalizing the concept of a matrix (a 2-way array). Kolda and Mayo have recently introduced a generalization of the matrix power method for computing real-valued tensor eigenpairs of symmetric tensors. In this work, we present an efficient implementation of their algorithm, exploiting symmetry in order to save storage, data movement, and computation. For an application involving repeatedly solving the tensor eigenproblem for many small tensors, we describe how a GPU can be used to accelerate the computations. On an NVIDIA Tesla C 2050 (Fermi) GPU, we achieve 318 Gflops/s (31% of theoretical peak performance in single precision) on our test data set.

# Dynamic adaptations in ab-initio nuclear physics calculations on multicore computer architectures

Avinash Srinivasa
Masha Sosonkina
Ames Laboratory/DOE
Iowa State University
Ames, IA 50011, USA
{avinashs,masha}@scl.ameslab.gov

Pieter Maris
James P. Vary
Physics Department
Iowa State University
Ames, IA 50011, USA
{pmaris,jvary}@iastate.edu

## Abstract

Computational resource availability often changes during the course of execution of an application. This is especially true in modern multi-user cluster environments where users can run many high-performance applications simultaneously which share resources such as Processing Elements (PEs), I/O, main memory, network. In such a scenario, it would be greatly advantageous to have applications augmented with adaptive capabilities, particularly during run-time. This involves targeting a computationally intensive part of the application and invoking appropriate adaptations so as to be able to adjust to the dynamically changing system conditions, to prevent drastic performance loss. In this paper, the parallel application MFDn (Many Fermion Dynamics for nuclear structure) used for ab-initio nuclear physics calculations is integrated with a middleware tool for invoking such adaptations. In particular, the multi-threaded Lanczos diagonalization procedure in MFDn is targeted to observe the effect on performance of dynamically changing the number of threads during the iterative process. Performance gains between two to seven times were observed in the presence of competing applications by incorporating these adaptation strategies.

# Comprehensive Performance Monitoring for GPU Cluster Systems

Karl Fuerlinger
Ludwig-Maximilians-University (LMU) Munich
Computer Science Department, MNM Team
Oettingenstr. 67, 80538 Munich, Germany
Email: fuerling@nm.ifi.lmu.de

Nicholas J. Wright and David Skinner
NERSC
Lawrence Berkeley National Laboratory
Berkeley, California 94720, USA
Email: {njwright, deskinner}@lbl.gov

## Abstract

Accelerating applications with GPUs has recently garnered a lot of interest from the scientific computing community. While tools for optimizing individual kernels are readily available, there is a lack of support for the specific needs of the HPC area. Most importantly, integration with existing parallel programming models (MPI and threading) and scalability to the full size of the machine are required. To address these issues we present our work on monitoring and performance evaluation of the CUDA runtime environment in the context of our scalable and efficient profiling tool IPM. We derive metrics for GPU utilization and identify missed opportunities for GPU-CPU overlap. We evaluate the monitoring accuracy and overheads of our approach and apply it to a full scientific application.

## Speculative Contention Avoidance in Software Transactional Memory

Ehsan Atoofian
Electrical Engineering Department
Lakehead University
Thunder Bay, Canada
atoofian@lakeheadu.ca

### Abstract

The well-known drawbacks of lock-based programming have forced researchers to explore new alternatives for parallel programming of which Transactional memory (TM) is a promising one. TM has been shown to scale well in multi-core processors when applications have data access patterns with a few conflicts. In contrast, if conflicts happen frequently in programs TM results in poor performance. In this paper, we introduce speculative contention avoidance (SCA) to prevent conflicts in TMs. SCA dynamically controls number of concurrently executing transactions and serializes those transactions that are likely to conflict. This technique is built on the concept of contention locality, which describes the likelihood that a previously aborted transaction fails again in the future. We find that contentions are highly predictable in TMs and exploit simple predictors which rely on history of transactions to speculate contentions. We have incorporated SCA into TL2 and compared the performance of the new implementation with the original STM using Stamp v0.9.7 benchmark suite. Our results show that SCA is highly effective at preventing contentions and significantly improves performance.

## Parallel Automatic Registration of Large Scale Microscopic Images on Multiprocessor CPUs and GPUs

Lee Cooper
Center for Comprehensive Informatics
Emory University
Atlanta, GA, USA
Email: lee.cooper@emory.edu

Kun Huang
Biomedical Informatics Department
Ohio State University
Columbus, OH, USA
Email: khuang@bmi.osu.edu

Manuel Ujaldon
Computer Architecture Department
University of Malaga
Malaga, Spain
Email: ujaldon@uma.es

### Abstract

During the present decade, emerging architectures like multicore CPUs and graphics processing units (GPUs) have steadily gained popularity for their ability to deploy high computational power at a low cost. In this paper, we combine parallelization techniques on a cooperative cluster of multicore CPUs and multisocket GPUs to apply their joint computational power to an automatic image registration algorithm intended for the analysis of high-resolution microscope images. Registration methods pose a computational challenge within the biomedical field due to the large size of microscope image data sets, which typically extend to the Terabyte scale. We analyze this application to identify those parts which are more favorable to the CPU and GPU execution models and decompose the process accordingly. Performance results are presented for two sets of images: mouse placenta (16K x 16K pixels) and mouse mammary tumor (23K x 62K pixels). Execution times are shown on different multi-node, multi-socket and multi-core configurations to provide performance insights about the most effective approach.

# Asynchronous peer-to-peer distributed computing for financial applications

Thierry GARCIA
Université de Toulouse
INP – ENSEEIHT – IRIT
BP 7122, 2 Rue Camichel
F-31071 Toulouse Cedex, France
e-mail: thierry.garcia@enseeiht.fr

Ming CHAU
Advanced Solutions Accelerator
199 Rue de l'Oppidum
F-34170 Castelnau le Lez, France
e-mail: mchau@advancedsolutionsaccelerator.com

The Tung Nguyen, Didier El-Baz
CNRS - LAAS 7 avenue du colonel Roche
F-31077 Toulouse, France
Université de Toulouse; UPS, INSA, INP, ISAE.
e-mail: ttnguyen@laas.fr elbaz@laas.fr

Pierre SPITERI
Université de Toulouse
INP – ENSEEIHT – IRIT
BP 7122, 2 Rue Camichel
F-31071 Toulouse Cedex, France
e-mail: pierre.spiteri@enseeiht.fr

## Abstract

This paper deals with the numerical solution of financial applications, more specifically the computation of American and European options derivatives modelled by boundary values problems. In such applications we have to solve large-scale algebraic linear systems. We concentrate on synchronous and asynchronous parallel iterative algorithms carried out on peer-to-peer networks. The properties of the operators arising in the discretized problem ensure the convergence of the parallel iterative synchronous and asynchronous algorithms. Computational experiments performed on peer-to-peer networks are presented and analyzed.

# Design and Performance of a Scalable, Parallel Statistics Toolkit

Philippe Pébay$^\star$, David Thompson$^\star$, Janine Bennett$^\circ$, and Ajith Mascarenhas$^\circ$ $^2$
Sandia National Laboratories
MS {$^\star$9159,$^\circ$9051}, P.O. Box 969
Livermore, CA 94551 U.S.A.
Email: {pppebay,dcthomp,jcbenne,aamasca}@sandia.gov

## Abstract

Most statistical software packages implement a broad range of techniques but do so in an ad hoc fashion, leaving users who do not have a broad knowledge of statistics at a disadvantage since they may not understand all the implications of a given analysis or how to test the validity of results. These packages are also largely serial in nature, or target multicore architectures instead of distributed-memory systems, or provide only a small number of statistics in parallel.

This paper surveys a collection of parallel implementations of statistics algorithm developed as part of a common framework over the last 3 years. The framework strategically groups modeling techniques with associated verification and validation techniques to make the underlying assumptions of the statistics more clear. Furthermore it employs a design pattern specifically targeted for distributed-memory parallelism, where architectural advances in large-scale high-performance computing have been focused. Moment-based statistics (which include descriptive, correlative, and multicorrelative statistics; principal component analysis (PCA); and k-means statistics) scale nearly linearly with the data set size and number of processes. Entropy-based statistics (which include order and contingency statistics) do not scale well when the data in question is continuous or quasi-diffuse but do scale well when the data is discrete and compact. We confirm and extend our earlier results by now establishing near-optimal scalability with up to 10,000 processes.

# Graph Model for Minimizing the Storage Overhead of Distributing Data for the Parallel Solution of Two-Phase Flows

Oliver Fortmeier, Alin A. Bastea, and H. Martin Bücker
RWTH Aachen University, Institute for Scientific Computing, Aachen, Germany
{fortmeier,bastea,buecker}@sc.rwth-aachen.de

## Abstract

We consider a finite element method for the parallel solution of two-phase flow problems using a level set approach. Here, two systems of equations result from the discretization of the governing partial differential equations. Rather than investigating the solution of these systems, we focus on finding a data distribution for their assembly. We formulate a new combinatorial problem that minimizes the overhead in storage requirement to represent the systems while, at the same time, balancing the computational effort to assemble these systems in parallel. We model this problem by introducing a weighted undirected graph. We then transform the problem to a (standard) graph partitioning problem in which a weighted sum of certain edges is minimized subject to balancing a weighted sum of all vertices. Numerical experiments are carried out illustrating the feasibility of the new approach for an application using up to 512 processes of a cluster of quad-core processors.

# A Comparative Benchmarking of the FFT on Fermi and Evergreen GPUs

Mohamed F. Ahmed
Computer Science and Engineering
The American University in Cairo
Email: MohamedAhmed@aucegypt.edu

Omar Haridy
Media Engineering and Technology
German University in Cairo
Email: Omar.Haridy@student.guc.edu.eg

## Abstract

GPGPUs are gaining importance as compute accelerators in HPC. Many studies provided insights of how NVIDIA GPUs outperform CPUs in many compute intensive algorithms. Also few studies demonstrated ATI GPUs compute capabilities. However, it is very important to measure and analyze the relative power of each architecture. In this paper, we analyze the architecture of NVIDIA's Fermi and AMD's Evergreen processors and demonstrate the best practices and techniques to best utilize the capabilities of each architecture. We implemented the FFT on both cards utilizing our findings using CUDA and OpenCL frameworks. We reached new performance ceilings on both GPUs. We also share our experience and findings on performance optimizations on both cards.

# GPU Accelerating for Rapid Multi-core Cache Simulation

Wan Han, Long Xiang, Gao Xiaopeng and Li Yi
State Key Laboratory of Virtual Reality Technology and System
School of Computer Science and Engineering, Beihang University
Beijing, China
wanhan@cse.buaa.edu.cn long@buaa.edu.cn gxp@buaa.edu.cn liyi@cse.buaa.edu.cn

## Abstract

To find the best memory system for emerging workloads, traces are obtained during application's execution, then caches with different configurations are simulated using these traces. Since program traces can be several gigabytes, simulation of cache performance is a time consuming process. Compute unified device architecture (CUDA) is a software development platform which enables programmers to accelerate the general-purpose applications on the graphics processing unit (GPU). This paper presents a real time multi-core cache simulator, which was built based on the Pin tool to get the memory reference; and fast method for multi-core cache simulation using the CUDA-enabled GPU. The proposed method is accelerated by the following techniques: execution parallelism exploration; memory latency hiding; a novel trace compression methodology. We describe how these techniques can be incorporated into CUDA code. Experimental results show that the hybrid parallel method of time-partitioning combines with set-partitioning presented here is 11.10x speedup compared to the CPU serial simulation algorithm. The present simulator can characterize cache performance of single-threaded or multi-threaded workloads at the speeds of 6-15 MIPS. It can simulates 6 cache configurations within one single pass at this speeds compared to CMP$im, which can only simulate one cache configuration each simulation pass at the speeds of 4-10 MIPS.

# Evaluation of likelihood functions for data analysis on Graphics Processing Units

Sverre Jarp, Alfio Lazzaro, Julien Leduc, Andrzej Nowak and Felice Pantaleo
CERN openlab
European Organization for Nuclear Research, CERN
Geneva, Switzerland
Emails: {sverre.jarp, alfio.lazzaro, julien.leduc, andrzej.nowak, felice.pantaleo}@cern.ch

## Abstract

Data analysis techniques based on likelihood function calculation play a crucial role in many High Energy Physics measurements. Depending on the complexity of the models used in the analyses, with several free parameters, many independent variables, large data samples, and complex functions, the calculation of the likelihood functions can require a long CPU execution time. In the past, the continuous gain in performance for each single CPU core kept pace with the increase on the complexity of the analyses, maintaining reasonable the execution time of the sequential software applications. Nowadays, the performance for single cores is not increasing as in the past, while the complexity of the analyses has grown significantly in the Large Hadron Collider era. In this context a breakthrough is represented by the increase of the number of computational cores per computational node. This allows to speed up the execution of the applications, redesigning them with parallelization paradigms. The likelihood function evaluation can be parallelized using data and task parallelism, which are suitable for CPUs and GPUs (Graphics Processing Units), respectively. In this paper we show how the likelihood function evaluation has been parallelized on GPUs. We describe the implemented algorithm and we give some performance results when running typical models used in High Energy Physics measurements. In our implementation we achieve a good scaling with respect to the number of events of the data samples.

# Exploiting multi-core architectures in clusters for enhancing the performance of the parallel Bootstrap simulation algorithm

César A. F. De Rose, Paulo Fernandes, Antonio M. Lima, Afonso Sales, and Thais Webber
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681 – Prédio 32 – 90619-900 – Porto Alegre, Brazil
{cesar.derose, paulo.fernandes, antonio.lima, afonso.sales, thais.webber}@pucrs.br

## Abstract

The solution of Markovian models is usually non-trivial to be performed using iterative methods, so it is well-fitted to simulation approaches and high performance implementations. The Bootstrap simulation method is a novel simulation technique of Markovian models that brings a considerable improvement in the results accuracy, notwithstanding its higher computation cost when compared to other simulation alternatives. In this paper, we present three parallel implementations of the Bootstrap simulation algorithm, exploiting a multi-core SMP cluster. We discuss some practical implementation issues about processing and communication demands, as well as present an analysis of speedup and efficiency considering different models' sizes and simulation trajectory lengths. Finally, future works point out some improvements to achieve even better results in terms of accuracy.

# Self-adaptive Evidence Propagation on Manycore Processors

Yinglong Xia
IBM T.J. Watson Research Center
1101 Kitchawan Rd.
Yorktown Heights, NY 10598, USA
Email: yxia@us.ibm.com

Viktor K. Prasanna
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089, USA
Email: prasanna@usc.edu

## Abstract

Evidence propagation is a major step in exact inference, a key problem in exploring probabilistic graphical models. Evidence propagation is essentially a series of computations between the potential tables in cliques and separators of a given junction tree. In real applications, the size of the potential tables varies dramatically. Thus, to achieve scalability over dozens of threads remains a fundamental challenge for evidence propagation on manycore processors. In this paper, we propose a self-adaptive method for evidence propagation on manycore processors. Given an arbitrary junction tree, we convert evidence propagation in the junction tree into a task dependency graph. The proposed self-adaptive scheduler dynamically adjusts the number of threads for scheduling or executing tasks according to the task dependency graph. Such a self-adaptability prevents the schedulers being too idle or too busy during the scheduling process. We implemented the proposed method on the Sun UltraSPARC T2 (Niagara 2) platform that supports up to 64 hardware threads. Through a set of experiments, we show that the proposed method scales well with respect to various input junction trees and exhibits superior performance when compared with several baseline methods for evidence propagation.

# Flexible Development of Dense Linear Algebra Algorithms on Massively Parallel Architectures with DPLASMA

George Bosilca*, Aurelien Bouteiller*, Anthony Danalis*†, Mathieu Faverge*, Azzam Haidar*,
Thomas Herault*‡, Jakub Kurzak*, Julien Langou§¶, Pierre Lemarinier*, Hatem Ltaief*,
Piotr Luszczek*, Asim YarKhan* and Jack Dongarra*†
*University of Tennessee Innovative Computing Laboratory
†Oak Ridge National Laboratory
‡University Paris-XI
§University of Colorado Denver

¶Research was supported by the National Science Foundation grant no. NSF CCF-811520

## Abstract

We present a method for developing dense linear algebra algorithms that seamlessly scales to thousands of cores. It can be done with our project called DPLASMA (Distributed PLASMA) that uses a novel generic distributed Direct Acyclic Graph Engine (DAGuE). The engine has been designed for high performance computing and thus it enables scaling of tile algorithms, originating in PLASMA, on large distributed memory systems. The underlying DAGuE framework has many appealing features when considering distributed-memory platforms with heterogeneous multicore nodes: DAG representation that is independent of the problem-size, automatic extraction of the communication from the dependencies, overlapping of communication and computation, task prioritization, and architecture-aware scheduling and management of tasks. The originality of this engine lies in its capacity to translate a sequential code with nested-loops into a concise and synthetic format which can then be interpreted and executed in a distributed environment. We present three common dense linear algebra algorithms from PLASMA (Parallel Linear Algebra for Scalable Multi-core Architectures), namely: Cholesky, LU, and QR factorizations, to investigate their data driven expression and execution in a distributed system. We demonstrate through experimental results on the Cray XT5 Kraken system that our DAG-based approach has the potential to achieve sizable fraction of peak performance which is characteristic of the state-of-the-art distributed numerical software on current and emerging architectures.

# An Efficient Search Algorithm without Memory for Peer-to-Peer Networks

Naixue Xiong
School of Information Technology,
Jiangxi University of Finance and Economics,
Nanchang, China 330013
e-mail: xiongnaixue@gmail.com

Yuhua Liu, Shishun Wu
Department of Computer Science
Huazhong Normal University
Wuhan, 430079, China
e-mail: yhliu@mail.ccnu.edu.cn

Laurence T. Yang
Department of Computer Science
St. Francis Xavier University
Antigonish, NS, B2G 2W5, Canada
e-mail: ltyang@stfx.ca

Kaihua Xu
College of Physical Science and Technology
Huazhong Normal University
Wuhan 430079, China,
e-mail: xukh@phy.ccnu.edu.cn

## Abstract

In recent years, as the rapid development of the technology about Peer-to-Peer (P2P) networks and the cloud computing technology, various applications of P2P technology become very widespread in most cloud computing distributed network applications. P2P cloud computing networks are unstructured and are an important component to implement next generation internet. How to quickly and efficiently search the resources in P2P networks has become one of the most critical issues, and it is one of the greatest concerns to users. This paper first describes the basic flooding P2P network search method, which is from the analyzing the pros and cons of several new search methods. After that, a cache-based search algorithm is proposed: When a node of the remaining load capacity is high, it will become the center node and form a joint topology area with the nearby nodes together, then the center node and ordinary nodes also need to store the index cache. At the local region the overheating resources will be copied to the local (that is, the contents cache). The simulation results show that the algorithm can effectively improve the hit rates of resources searching, and reduce the query delay in cloud computing networks.

# Workshop 13

# Dependable Parallel, Distributed and Network-Centric Systems

# DPDNS 2011

# Solving k-Set Agreement with Stable Skeleton Graphs

Martin Biely* , Peter Robinson‡ , and Ulrich Schmid† * EPFL, Switzerland, biely@ecs.tuwien.ac.at
ECS Group, Technische Universität Wien, Austria, s@ecs.tuwien.ac.at
‡ Division of Mathematical Sciences, Nanyang Technological University, Singapore, peter.robinson@ntu.edu.sg

## Abstract

In this paper we consider the k-set agreement problem in distributed message-passing systems using a round-based approach: Both synchrony of communication and failures are captured just by means of the messages that arrive within a round, resulting in round-by-round communication graphs that can be characterized by simple communication predicates. We introduce the weak communication predicate PSources(k) and show that it is tight for k-set agreement, in the following sense: We (i) prove that there is no algorithm for solving (k-1)-set agreement in systems characterized by PSources(k), and (ii) present a novel distributed algorithm that achieves k-set agreement in runs where PSources(k) holds. Our algorithm uses local approximations of the stable skeleton graph, which reflects the underlying perpetual synchrony of a run. We prove that this approximation is correct in all runs, regardless of the communication predicate, and show that graph-theoretic properties of the stable skeleton graph can be used to solve *k*-set agreement if PSources(k) holds.

# Compact Route Computation: Improving Parallel BGP Route Processing for Scalable Routers

Xuezhi Jiang*† , Mingwei Xu* , Qi Li*
* Tsinghua University, † Shijiazhuang Mechanized Infantry Institute
{jxz,xmw,liqi}@csnet1.cs.tsinghua.edu.cn

## Abstract

Nowadays Internet routers are overwhelmed by a large quantity of BGP (Border Gateway Protocol) updates triggered by route changes. The fast growth of the Internet size further aggravates router workloads and exacerbates routing convergence performance. Scalable routers, such as cluster routers and ForCES, are proposed to exploit distributed control plane (DCP) with multiple control elements (CEs) to scale route processing capacity. Previous studies show that most route updates are duplicated in the Internet. Traditional parallel computation schemes only consider calculating route in parallel but most routes are computed in vain since they will not be finally selected. This paper proposes a simple and novel idea of compact route computation (CRC) to reduce BGP route processing load and improve routing convergence performance. Our scheme partitions Adj-RIBs-in among multiple CEs in the granularity of prefixes, which makes non-consecutive updates for a prefix queued adjacently in distributed control plane. Route computations triggered by a prefix's consecutive updates are compacted into one. We evaluate our scheme by simulations with real BGP update data, and results show that our scheme is very effective to reduce route computation workloads. For example, for a scalable router with 4 CEs and the updates received from 24 neighbors, our scheme reduces 60% route computation load. It can distinctly reduce route computation load with more CEs.

# Towards Persistent Connections using Failure Detectors

Naohiro HAYASHIBARA
Dept. of Computer Science,
Faculty of Computer Science and Engineering,
Kyoto Sangyo University
Kyoto, Japan
Email: naohaya@cse.kyoto-su.ac.jp

## Abstract

In the application layer, communication between two nodes is usually done on a communication channel. On the other hand, connection and disconnection occur intermittently in lower layers. Persistent connection is often required by several applications (e.g., HTTP, large data transfer, etc.). It has to be available continuously till the application closes it explicitly. To realize such a persistent connection, transient link failure has to be masked from the application layer.

In this paper, we propose a way to provide persistent connections in TCP using accrual failure detector. Once the connection is established, the session layer can properly handle link failures, for example, it can suspend while link failure or unstable link condition, and resume after the recovery of the link. To realize persistent connections, we first implemented an interrogation-based accrual failure detector and evaluated its performance in several environments. Then, proposed the procedure for persistent connections using the failure detector.

# A Monitoring and Audit Logging Architecture for Data Location Compliance in Federated Cloud Infrastructures

Philippe Massonet, Syed Naqvi,
Christophe Ponsard and Joseph Latanicki
Centre d'Excellence en Technologies de l'Information
et de la Communication (CETIC), Belgium
(Philippe.Massonet, Syed.Naqvi, Christophe.Ponsard)@cetic.be
Thales Theresis, France
Joseph.Latanicki@thalesgroup.com
Benny Rochwerger
and Massimo Villari
IBM Haifa Research Lab, Israel
rochwer@il.ibm.com
University of Messina, Italy
mvillari@unime.it

## Abstract

Current cloud infrastructures have opaque service offerings where customers cannot monitor the underlying physical infrastructure. This situation raises concerns for meeting compliance obligations by critical business applications with data location constraints that are deployed in a Cloud. When federated cloud infrastructures span across different countries where data can migrate from one country to another, it should be possible for data owners to monitor the location of their data. This paper shows how an existing federated Cloud monitoring infrastructure can be used for data location monitoring without compromising Cloud isolation. In the proposed approach collaboration is required between the cloud infrastructure provider (IP) and the user of the cloud, the service provider (SP): the IP monitors the virtual machines (VM) on the SP's behalf and makes the infrastructure level monitoring information available to him. With the monitoring information the SP can create the audit logs required for compliance auditing. The proposed logging architecture is validated by an e-Government case study with legal data location constraints.

# Dependable Autonomic Cloud Computing with Information Proxies

D. Cenk Erdil
Computer Engineering Department
İstanbul Bilgi University
34060 İstanbul, Türkiye
cenk.erdil@bilgi.edu.tr

## Abstract

Autonomic computing systems promise to manage themselves on a set of basic rules specified to higher level objectives. One of the challenges in making this possible is *dependable collaboration* among peers in a large-scale network. Effective maintenance of next generation distributed systems, such as clouds and second generation grids, will be nearly impossible without autonomic computing, with ever increasing scale of such systems. In addition, due to the nature of autonomous clouds to form administrative boundaries, dependable collaboration becomes a much harder problem. Employing information proxies may help improve such collaboration in existence of administrative boundaries. Although a general proxy definition can refer to many contexts, we focus on such proxies for dependable collaboration for distributed resource scheduling. Our definition of information proxies, and the particular areas we make use of them mainly contribute to the *self-configuring* and *self-optimizing* fundamentals of the autonomic computing paradigm in general. By simulation, we show that information proxies help improve resource scheduling decisions that support large-scale autonomic computing systems.

# A Fault-tolerant High Performance Cloud Strategy for Scientific Computing

Ekpe Okorafor
Computer Science and Engineering
African University of Science and Technology (AUST)
Abuja, Nigeria
Email: eokorafor@aust.edu.ng

## Abstract

Scientific computing often requires the availability of a massive number of computers for performing large scale experiments. Traditionally, high-performance computing solutions and installed facilities such as clusters and super computers have been employed to address these needs. Cloud computing provides scientists with a completely new model of utilizing the computing infrastructure with the ability to perform parallel computations using large pools of virtual machines (VMs).

The infrastructure services (Infrastructure-as-a-service), provided by these cloud vendors, allow any user to provision a large number of compute instances. However, scientific computing is typically characterized by complex communication patterns and requires optimized runtimes. Today, VMs are manually instantiated, configured and maintained by cloud users. These coupled with the latency, crash and omission failures in service providers, results in an inefficient use of VMs, increased complexity in VM-management tasks, a reduction in the overall computation power and increased time for task completion.

In this paper, a high performance cloud computing strategy is proposed that combines the adaptation of a parallel processing framework, such as the Message Passing Interface (MPI) and an efficient checkpoint infrastructure for VMs, enabling its effective use for scientific computing. By developing such a mechanism, we can achieve optimized runtimes comparable to native clusters, improve checkpoints with low interference on task execution and provide efficient task recovery. In addition, checkpointing is used to minimize the cost and volatility of resource provisioning, while improving overall reliability. Analysis and simulations show that the proposed approach compares favorably with the native cluster MPI implementations.

# Evaluation of Simple Causal Message Logging for Large-Scale Fault Tolerant HPC Systems

Esteban Meneses
Department of Computer Science
University of Illinois
Urbana, Illinois, USA
emenese2@illinois.edu

Greg Bronevetsky
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, California, USA
greg@bronevetsky.com

Laxmikant V. Kalé
Department of Computer Science
University of Illinois
Urbana, Illinois, USA
kale@illinois.edu

## Abstract

The era of petascale computing brought machines with hundreds of thousands of processors. The next generation of exascale supercomputers will make available clusters with millions of processors. In those machines, mean time between failures will range from a few minutes to few tens of minutes, making the crash of a processor the common case, instead of a rarity. Parallel applications running on those large machines will need to simultaneously survive crashes and maintain high productivity. To achieve that, fault tolerance techniques will have to go beyond checkpoint/restart, which requires all processors to roll back in case of a failure. Incorporating some form of message logging will provide a framework where only a subset of processors are rolled back after a crash. In this paper, we discuss why a simple causal message logging protocol seems a promising alternative to provide fault tolerance in large supercomputers. As opposed to pessimistic message logging, it has low latency overhead, especially in collective communication operations. Besides, it saves messages when more than one thread is running per processor. Finally, we demonstrate that a simple causal message logging protocol has a faster recovery and a low performance penalty when compared to checkpoint/restart. Running NAS Parallel Benchmarks (CG, MG, BT and DT) on 1024 processors, simple causal message logging has a latency overhead below 5%.

## Building a Fault Tolerant MPI Application: A Ring Communication Example

Joshua Hursey, Richard L. Graham
Oak Ridge National Laboratory, Oak Ridge, TN USA 37831

### Abstract

Process failure is projected to become a normal event for many long running and scalable High Performance Computing (HPC) applications. As such many application developers are investigating Algorithm Based Fault Tolerance (ABFT) techniques to improve the efficiency of application recovery beyond what existing checkpoint/restart techniques alone can provide. Unfortunately for these application developers the libraries that their applications depend upon, like Message Passing Interface (MPI), do not have standardized fault tolerance semantics. This paper introduces the reader to a set of run-through stabilization semantics being developed by the MPI Forum's Fault Tolerance Working Group to support ABFT. Using a well-known ring communication program as the running example, this paper illustrates to application developers new to ABFT some of the issues that arise when designing a fault tolerant application. The ring program allows the paper to focus on the communication-level issues rather than the data preservation mechanisms covered by existing literature. This paper highlights a common set of issues that application developers must address in their design including program control management, duplicate message detection, termination detection, and testing. The discussion provides application developers new to ABFT with an introduction to both new interfaces becoming available, and a range of design issues that they will likely need to address regardless of their research domain.

## Algorithm-Based Recovery for Newton's Method without Checkpointing

Hui Liu, Teresa Davies, Chong Ding, Christer Karlsson, and Zizhong Chen
Department of Mathematical and Computer Sciences
Colorado School of Mines
Golden, CO 80401, USA
{huliu,tdavies,cding, ckarlsso, zchen}@mines.edu

### Abstract

Checkpointing is the most popular fault tolerance method used in high-performance computing (HPC) systems. However, increasing failure rates requires more frequent checkpoints, thus makes checkpointing more expensive. We present a checkpoint-free fault tolerance technique. It takes advantage of both data dependencies and communication-induced redundancies of parallel applications to tolerate fail-stop failures. Under the specified conditions, our technique introduces no additional overhead when there is no actual failure in the computation and recover the lost data with low overhead. We add fault-tolerant capacity to Newton's method by using our scheme and diskless checkpointing. Numerical simulations indicate that our scheme introduces much less overhead than diskless checkpointing does.

# Predicting Node Failure in High Performance Computing Systems from Failure and Usage Logs

Nithin Nakka[§], Ankit Agrawal[†], Alok Choudhary[†]
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign,
1308 W. Main St., Urbana, IL, USA, 61801.
nakka@illinois.edu
[†]Department of Electrical Engineering and Computer Science, Northwestern University,
2145 Sheridan Rd., Tech Inst. Bldg., Evanston, IL, USA, 60202.
{ankitag, choudhar}@eecs.northwestern.edu

## Abstract

In this paper, we apply data mining classification schemes to predict failures in a high performance computer system. Failure and Usage data logs collected on supercomputing clusters at Los Alamos National Laboratory (LANL) were used to extract instances of failure information. For each failure instance, past and future failure information is accumulated  time of usage, system idle time, time of unavailability, time since last failure, time to next failure. We performed two separate analyses, with and without classifying the failures based on their root cause. Based on this data, we applied some popular decision tree classifiers to predict if a failure would occur within 1 hour. Our experiments show that our prediction system predicts failures with a high-degree of precision up to 73% and recall of about 80%. We also observed that employing the usage data along with the failure data has improved the accuracy of prediction.

# Achieving Target MTTF by Duplicating Reliability-Critical Components in High Performance Computing Systems

Nithin Nakka[‡], Alok Choudhary[†], Gary Grider[§], John Bent[§], James Nunez[§] and Satsangat Khalsa[§]
[‡] Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA
nakka@crhc.illinois.edu
[†] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
choudhar@eecs.northwestern.edu
[§]Los Alamos National Laboratories, Albuquerque, New Mexico, U.S.A.
{ggrider, johnbent, jnunez, satsang}@lanl.gov

## Abstract

Mean Time To failure, MTTF, is a commonly accepted metric for reliability. In this paper we present a novel approach to achieve the desired MTTF with minimum redundancy. We analyze the failure behavior of large scale systems using failure logs collected by Los Alamos National Laboratory. We analyze the root cause of failures and present a choice of specific hardware and software components to be made fault-tolerant, through duplication, to achieve target MTTF at minimum expense. Not all components show similar failure behavior in the systems. Our objective, therefore, was to arrive at an ordering of components to be incrementally selected for protection to achieve a target MTTF. We propose a model for MTTF for tolerating failures in a specific component, system-wide, and order components according to the coverage provided. Systems grouped based on hardware configuration showed similar improvements in MTTF when different components in them were targeted for fault-tolerance.

# Workshop 14

# International Workshop on Hot Topics in Peer-to-Peer Systems

# HOTP2P 2011

# Controlling P2P Applications via Address Harvesting: The Skype Story

Anat Bremler-Barr[*], Omer Dekel[*], Ran Goldschmidt[†] and Hanoch Levy[‡]
[*]The Interdisciplinary Center, Herzliya, Israel. Email: bremler@idc.ac.il, omer.dekel@gmail.com
[†] University of Haifa, Israel, Email: ran.goldschmidt@gmail.com
[‡] Tel-Aviv University, Tel-Aviv, Israel Email: hanoch@cs.tau.ac.il

## Abstract

P2P applications have become a dominant force in the Internet, both as an economic factor and as a traffic contributor. A ”battle of power” is ongoing between the application providers and the Internet Service Providers (ISPs) on who will control this traffic. This is motivated by both economic incentives and QoS objectives.

Little is known to the ISPs about the architecture of such applications or about the identity of their sessions; these are hidden by the application providers (assisted by their distributed control structure) making the ISPs' life harder.

We are interested in Skype, as a very popular representative of distributed P2P applications. We explore the possibility of getting control/blocking Skype sessions by harvesting its Super Nodes (SNs), and blocking the network clients from connecting to them. Using experimental results and an analytical model we show that it is possible to collect a large enough number of SNs to block, with a probability higher than 95%. We further use the model to show that our approach is robust against possible strategies that can be adopted by Skype to maximize its resilience to blocking.

The results derived and the vulnerability to SN harvesting, though discussed in the context of Skype, are general and may hold true for other Super Node based P2P systems.

# SeederTrading: Trading Swarm Capacity for Improving Content Distribution

HyunYong Lee
National Institute of Information and
Communications Technology (NiCT)
Tokyo, Japan
ifjesus7@gmail.com

Masahiro Yoshida and Akihiro Nakao
The University of Tokyo
Tokyo, Japan
yoshida@nakao-lab.org, nakao@iii.u-tokyo.ac.jp

## Abstract

A surplus seeding capacity is often observed in large swarms with many seeders while small swarms with few seeders suffer from low download performance. In this paper, we propose a multi-swarm collaboration approach, called *SeederTrading* to improve overall content distribution performance by exploiting non-uniform distribution of seeders over swarms. In a nutshell, we enable the over-seeded swarms to share their surplus seeding capacity with the under-seeded swarms. To realize the SeederTrading, we resolve several issues including estimation of swarm capacity, selection of swarms to be matched, and seeder trading across the selected swarms. Our extensive measurement shows that 12.9% of under-seeded swarms can achieve their maximum download throughput by utilizing the surplus seeding capacity. Simulation results show that the under-seeded swarms can improve the download performance by over 20% while retaining the content distribution performance of over-seeded swarms. We also show that the trading seeding capacity across swarms can be done within 60 seconds.

# Virtual Direction Multicast for Overlay Networks

Suat Mercan
University of Nevada - Reno
Computer Science and Engineering
Reno, NV 89557
smercan@cse.unr.edu

Murat Yuksel
University of Nevada - Reno
Computer Science and Engineering
Reno, NV 89557
yuksem@cse.unr.edu

## Abstract

In this paper we propose Virtual Direction Multicast (VDM) which aims to minimize network usage and disconnection time for video multicast applications on peer-to-peer overlay networks. It locates the end hosts relative to each other based on a virtualized orientation scheme. It builds multicast tree by connecting the nodes which are estimated to be in the same virtual direction. By using the concept of directionality, we target to use minimal resources in the underlying network. We compare VDM against a Closest Node Multicast (CNM) protocol that connects nearby nodes to construct the multicast tree. Simulation results show that our proposed technique VDM consistently outperforms CNM under different churn rates.

# Parallel Processing Framework on a P2P System Using Map and Reduce Primitives

Kyungyong Lee[*] , Tae Woong Choi[†] , Arijit Ganguly[‡] , David I. Wolinsky[*] ,
P.Oscar Boykin[*] , Renato Figueiredo[*]
[*] ACIS Lab. Department of ECE. University of Florida
E-mail: {klee, davidiw, boykin, renato}@acis.ufl.edu
[†] Samsung SDS, Seoul, South Korea
E-mail: taewoong.choi@samsung.com
[‡] Amazon Web Service. Amazon.com
E-mail: arijit@amazon.com

## Abstract

This paper presents a parallel processing framework for structured Peer-To-Peer (P2P) networks. A parallel processing task is expressed using Map and Reduce primitives inspired by functional programming models. The Map and Reduce tasks are distributed to a subset of nodes within a P2P network for execution by using a self-organizing multicast tree. The distribution latency cost of multicast method is $O(log(N))$, where $N$ is a number of target nodes for task processing. Each node getting a task performs the Map task, and the task result is summarized and aggregated in a distributed fashion at each node of the multicast tree during the Reduce task. We have implemented this framework on the Brunet P2P system, and the system currently supports predefined Map and Reduce tasks or tasks inserted through Remote Procedure Call (RPC) invocations. A simulation result demonstrates the scalability and efficiency of our parallel processing framework. An experiment result on PlanetLab which performs a distributed K-Means clustering to gather statistics of connection latencies among P2P nodes shows the applicability of our system in applications such as monitoring overlay networks.

# SPUN: A P2P Probabilistic Search Algorithm based on Successful Paths in Unstructured Networks

D.M Rasanjalee Himali and Sushil K. Prasad
Department of Computer Science
Georgia State University
Atlanta, GA
dmrhimali@student.gsu.edu, sprasad@gsu.edu

## Abstract

Efficient searching for information is an important goal in peer-to-peer (P2P) networks. Searching in an unstructured P2P network is particularly challenging due to the random nature of the P2P overlay links. In this paper, we propose a novel probabilistic search mechanism called SPUN, that increases the success ratio of queries while keeping the bandwidth consumption considerably low. SPUN is an informed search mechanism that improves upon state-of-art probabilistic mechanism, namely, the Adaptive Probabilistic Search (APS). The core principle of our algorithm is to exploit the successful query paths that develop during the lifetime of P2P network and converge toward the target objects. Our work introduces a new neighbor selection criterion which allows a peer to evaluate its neighbors based on the strength of successful paths the neighbor leads to. We also introduce a peer profile exchange mechanism that supports the reduction in uncertainty in peer selection decision. Our extensive simulation results confirm that our path-based algorithm performs 25% better than APS and several of its variants and is capable of achieving higher success ratios with fewer walkers each with an average message size of 71 bytes.

# Betweenness Centrality Approximations for an Internet Deployed P2P Reputation System

Dimitra Gkorou, Johan Pouwelse, and Dick Epema
Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology
e-mail: {D.Gkorou,J.A.Pouwelse,D.H.J.Epema}@tudelft.nl

## Abstract

In the Bartercast reputation mechanism of the BitTorrent-based P2P client Tribler, peers compute local, subjective reputations of other peers by applying a flow-based algorithm to a locally maintained Bartercast graph with peers as nodes and bandwidth contributions as edges. We have previously shown that the computed reputations are more accurate when a peer takes the node with the highest Betweenness Centrality (BC) in its local Bartercast graph as the initial point in this algorithm rather than itself. BC is a powerful metric for identifying central nodes in complex network analysis, but its computation in large and dynamic networks is costly, and previously proposed approximation methods are only designed for static networks. In this paper, first we assess the stability of the nodes with the highest BC values in growing synthetic random and scale-free, and Bartercast graphs. Next, we evaluate three BC approximation methods proposed in the literature in terms of their ability to identify the top-most central nodes. We show that these approximations are efficient and highly accurate in scale-free and Bartercast graphs, but less so in random graphs. Finally, we integrate the three BC approximations into Bartercast, and we evaluate the quality of the reputations they yield.

# When KAD meets BitTorrent - Building a Stronger P2P Network

Juan Pablo Timpanaro, Thibault Cholez[**], Isabelle Chrisment[*], Olivier Festor
INRIA Nancy-Grand Est, France
[*]LORIA - ESIAL, Henri Poincaré University, Nancy 1, France
[**]Université Technologique de Troyes
Email: {juanpablo.timpanaro, thibault.cholez, olivier.festor}@inria.fr
Email: {isabelle.chrisment}@loria.fr

## Abstract

The current wave of evolution that leads BitTorrent towards a fully decentralized architecture is both promising and risky. Related work demonstrates that BitTorrent's Mainline DHT is exposed to several identified security issues. In parallel, the KAD DHT has been the core of intense research and was improved over years. In this paper, we present a study that motivates the integration of both worlds. We provide a performance comparison of both DHTs in terms of publishing efficiency. We investigate the security threats and show that the current BitTorrent Mainline DHT is more vulnerable to attacks than KAD while the download service of BitTorrent has much better performance. Given the strengths and weaknesses of both DHTs, we propose a design in which the two P2P networks can be merged to form a fully distributed, efficient and safe P2P ecosystem.

# An Adaptive Response Routing Mechanism to Improve DHT Performance in the Presence of NATs

David A. Bryan
Cogent Force, LLC and College of William and Mary
Williamsburg, VA, USA
dbryan@ethernot.org

## Abstract

Asymmetric network connectivity, frequently the result of NATs in the network environment, can significantly impact the selection of an appropriate DHT response routing technique. We show that neither response mechanism commonly used today (symmetric recursive or direct response) function well for all network NAT densities. We present an adaptive response routing algorithm for use in generic libraries or protocols where the deployment environment is not known in advance. Through worst-case analytical analysis and average-case simulation we show that our technique offers a significant performance advantage in environments (such as the Internet) where NAT densities are not known a priori.

# A New Protocol to Determine the NAT Characteristics of a Host

Sebastian Holzapfel, Matthäus Wander, Arno Wacker, Lorenz Schwittmann, and Torben Weis
University of Duisburg-Essen, Distributed Systems Group,
Bismarckstrae 90, 47057 Duisburg, Germany
{sebastian.holzapfel|matthaeus.wander|arno.wacker|lorenz.schwittmann|torben.weis}@vs.uni-due.de

## Abstract

The shortage of IPv4 addresses and the very slow transition to IPv6 leads to well-established pragmatic solutions in the Internet: today many hosts are still using IPv4 and are connected to the Internet over a Network Address Translation router. For many applications, which need inbound connections, like e.g. voice-over-IP or peer-to-peer-based systems it is necessary to determine the characteristics of the surrounding network environment, i.e. the behavior of the used router. In most cases this information is required to successfully establish inbound connections. Therefore, we present in this paper a new protocol to determine the characteristics of the used router. Our protocol is backward compatible to the well known STUN protocol while providing more detailed results. Furthermore, our protocol can be used in a fully decentralized way, i.e. without any centralized servers, making it suitable for pure peer-to-peer-based systems. We complete the presentation of our new protocol with an evaluation through a field experiment.

# Performance Prediction in a Decentralized Environment for Peer-to-Peer Computing

Bogdan Florin Cornea[*] , Julien Bourgeois[*] , The Tung Nguyen[†], Didier El-Baz[†]
[*] LIFC, University of Franche-Comté; 1 cours Leprince Ringuet; 25201 Montbéliard; France.
{bogdan.cornea, julien.bourgeois}@univ-fcomte.fr
[†] CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse, France.
University of Toulouse; UPS, INSA, INP, ISAE; LAAS; F-31077 Toulouse France.
{ttnguyen, elbaz}@laas.fr

## Abstract

P2PDC is an environment for high performance peer to peer computing that allows direct communication between peers. This environment is based on P2PSAP, a self adaptive communication protocol. P2PDC is suited to the solution of large scale numerical simulation problems via distributed iterative methods. dPerf is a performance prediction environment for parallel and distributed applications, with primary interest in programs written in C, C++, Fortran for P2PDC. The dPerf performance prediction tool makes use of static and dynamic analyses combined with trace-based simulation. In this paper, we present a decentralized version of P2PDC and show how dPerf predicts performance for the P2PDC environment. We present new features of P2PDC aimed at making it more scalable and robust. Through experiments with P2PDC and dPerf, we show how to properly choose a peer to peer computing system which can match the computing power of a cluster.

# Workshop 15

# Workshop on Multi-Threaded Architectures and Applications

# MTAAP 2011

# Hardware Transactional Memory for High-Performance Computing Applications

Karl Ljungkvist*, Martin Tillenius*, David Black-Schaffer*, Sverker Holmgren*,
Martin Karlsson*, Elisabeth Larsson*
*Department of Information Technology
Uppsala University
⟨firstname⟩.⟨lastname⟩@it.uu.se

## Abstract

This work investigates the benefits of transactional memory for high-performance and scientific computing by examining n-body and unstructured mesh applications on a prototype computer system with transactional memory support in hardware. Transactional memory systems have the potential to both improve performance, through reduced lock overhead, and ease development, by providing simplified synchronization semantics. To date, only a few early experiments have been presented on actual transactional memory hardware, with none of them investigating the benefits of transactional memory for scientific computing.

We investigate straight-forward implementations to see if replacing locks by transactions can yield better performance and compete with more complex algorithms. Our experiments show that using transactions is the fastest way to concurrently update shared floating-point variables, which is of interest in many scientific computing applications. However, if it is possible, avoiding concurrent updates altogether yields the best performance.

# A micro threading based concurrency model for parallel computing

Qiang YANG, C.R.JESSHOPE, Jian FU
Institute of informatics
University of Amsterdam
The Netherlands
{ q.yang, c.r.jesshope, j.fu }@uva.nl

## Abstract

The continuing launch of various multi-core processors popularizes parallel computing of gaining higher performance; however, it also exerts pressures on software developers concerning how to make full use of the multiplicity of resources to get the required bene-fits due to high coupling of parallel programs to specified hardware. In this paper, we propose SVP, a micro-threading based concurrency model, as an alternative to efficiently program on and manage multi-core, even many-core systems. SVP isolates programming from executing resources so that programs are compiled once and execute regardless of actual cores available, not only numbers, but also properties to achieve scalable speedup. Such lower coupling, better flexibility and portability in parallel applications are highly appreciated. To investigate SVP, tool chains and a fully functional software simulator of an SVP many-core chip have been developed for test and verification. Using this infra-structure, we are able to show how much we can gain from SVP and how speedup is scaled by running the same binary code of Game of Life on a scalable many-core platform.

# Comparison of Parallel Programming Models for Multicore Architectures

Deepak Shekhar T.C, Kiran Varaganti, Rahul Suresh, Rahul Garg and Ramalingam Ramamoorthy
Frontier Research Group, Samsung India Software Operations Pvt Ltd, Level 3, Block-B,
Bagmane Tridib, 65/2, Bagmane Tech Park, Byrasandra, CV Raman Nagar, Bengaluru - 560093, India
Email: deepak.tc@samsung.com, v.kiran@samsung.com, rahul.suresh@samsung.com,
rahul.garg@samsung.com, ramalingam.r@samsung.com

## Abstract

Multicore computing platforms have emerged as the most common computing platform to overcome challenges stemming from high power densities and thermal hot spots in conventional microprocessors. However, providing multiple cores does not directly translate into increased performance or better energy efficiency for most applications. The burden is placed on developers and tools to find and exploit parallelism and eventually utilize all of the available computing resources. Since multicore applications are more complex than single core applications, the software development tools play a crucial role to help programmers create high performance and correct software. In this paper we compare the most popular programming models OpenMP, GCD and Pthreads by applying these models to parallelize face detection and automatic speech recognition applications.

# A Work-stealing Framework for mixed-mode parallel Applications

Martin Wimmer, Jesper Larsson Träff
Department of Scientific Computing, Faculty of Computer Science
University of Vienna
Nordbergstrasse 15/3C
A-1090 Vienna, Austria
Email: {wimmer,traff }@par.univie.ac.at

## Abstract

Parallelizing complex applications even for well-behaved parallel systems often calls for different parallelization approaches within the same application. In this paper we discuss three applications from the literature that for both reasons of efficiency and expressive convenience benefit from a mixture of task and more tightly coupled data parallelism. These three applications, namely Quicksort, list ranking, and LU factorization with partial pivoting, are paradigms for recursive, mixed-mode parallel algorithms that can neither easily nor efficiently be expressed in either a purely data-parallel or a purely task-parallel fashion. As a solution we present a shared-memory programming framework that allows tasks to dynamically spawn subtasks with a given degree of parallelism for implementing tightly coupled parallel parts of the algorithm. All three paradigmatic applications can naturally be expressed in this framework, which in turn can be supported by an extended, non-conventional work-stealing scheduler, which we also briefly sketch. Using our new algorithm for work-stealing with deterministic team-building we are able to show, beyond the improved, more natural implementability, in many cases better scalability and sometimes absolute performance than with less natural implementations based on pure task-parallelism executed with conventional work-stealing. Detailed performance results using an Intel 32-core system substantiate our claims.

# Tracking Structure of Streaming Social Networks

David Ediger Jason Riedy David A. Bader Henning Meyerhenke
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

## Abstract

Current online social networks are massive and still growing. For example, Facebook has over 500 million active users sharing over 30 billion items per month. The scale within these data streams has outstripped traditional graph analysis methods. Real-time monitoring for anomalies may require dynamic analysis rather than repeated static analysis. The massive state behind multiple persistent queries requires shared data structures and flexible representations. We present a framework based on the STINGER data structure that can monitor a global property, connected components, on a graph of 16 million vertices at rates of up to 240 000 updates per second on 32 processors of a Cray XMT. For very large scale-free graphs, our implementation uses novel batching techniques that exploit the scale-free nature of the data and run over three times faster than prior methods. Our framework handles, for the first time, real-world data rates, opening the door to higher-level analytics such as community and anomaly detection.

# Toward Parallel Document Clustering

Jace A. Mogill, David J. Haglin
Pacific Northwest National Laboratory
Richland, WA, 99354 USA
{jace.mogill, david.haglin}@pnl.gov

## Abstract

A key challenge to automated clustering of documents in large text corpora is the high cost of comparing documents in a multi-million dimensional document space. The Anchors Hierarchy is a fast data structure and algorithm for localizing data based on a triangle inequality obeying distance metric, the algorithm strives to minimize the number of distance calculations needed to cluster the documents into "anchors" around reference documents called "pivots". We extend the original algorithm to increase the amount of available parallelism and consider two implementations: a complex data structure which affords efficient searching, and a simple data structure which requires repeated sorting. The sorting implementation is integrated with a text corpora "Bag of Words" program and initial performance results of end-to-end document processing workflow are reported.

# Evaluating In-Clique and Topological Parallelism Strategies for Junction Tree-Based Bayesian Inference Algorithm on the Cray XMT

George Chin Jr., Sutanay Choudhury, Lars Kangas, Sally McFarlane, and Andres Marquez
Pacific Northwest National Laboratory
Richland, Washington USA
{George.Chin, Sutanay.Choudhury, Lars.Kangas, Sally.McFarlane, Andres.Marquez}@pnl.gov

## Abstract

Long viewed as a strong statistical inference technique, Bayesian networks have emerged as an important class of applications for high-performance computing. We have applied an architecture-conscious approach to parallelizing the Lauritzen-Spiegelhalter Junction Tree algorithm for exact inferencing of Bayesian networks. In optimizing the Junction Tree algorithm, we have implemented both in-clique and topological parallelism strategies to best leverage the fine-grained synchronization and massive-scale multithreading of the Cray XMT architecture. Two topological techniques were developed to parallelize the evidence propagation process through the Bayesian network. The first technique involves performing intelligent scheduling of junction tree nodes based on the tree's topology and the relative sizes of nodes. The second technique involves decomposing the junction tree into a finer state graph representation to offer many more opportunities for parallelism. We evaluate these optimizations on five different Bayesian networks and report our findings and observations. From this development and evaluation, we demonstrate the application of massive-scale multithreading for load balancing and use of implicit parallelism-based compiler optimizations for designing scalable inferencing algorithms.

# High Performance Data Mining Using R on Heterogeneous Platforms

Prabhat Kumar, Berkin Ozisikyilmaz, Wei-Keng Liao, Gokhan Memik, Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL, USA
{pku649, boz283, wkliao, memik, choudhar}@ece.northwestern.edu

## Abstract

The exponential increase in the generation and collection of data has led us in a new era of data analysis and information extraction. Conventional systems based on general-purpose processors are unable to keep pace with the heavy computational requirements of data mining techniques. High performance co-processors like GPUs and FPGAs have the potential to handle large computational workloads. In this paper, we present a scalable framework aimed at providing a platform for developing and using high performance data mining applications on heterogeneous platforms. The framework incorporates a software infrastructure and a library of high performance kernels. Furthermore, it includes a variety of optimizations which increase the throughput of applications. The framework spans multiple technologies including R, GPUs, multi-core CPUs, MPI, and parallel-netCDF harnessing their capabilities for high-performance computations. This paper also introduces the concept of interleaving GPU kernels from multiple applications providing significant performance gain. Thus, in comparison to other tools available for data mining, our framework provides an easy-to-use and scalable environment both for application development and execution. The framework is available as a software package which can be easily integrated in the R programming environment.

# Bandwidth Reduction Through Multithreaded Compression of Seismic Images

Ahmed A. Aqrawi and Anne C. Elster
Norwegian University of Science and Technology
Department of Computer and Information Science
Trondheim, Norway
aqrawi@stud.ntnu.no, elster@idi.ntnu.no

## Abstract

One of the main challenges of modern computer systems is to overcome the ever more prominent limitations of disk I/O and memory bandwidth, which today are thousands-fold slower than computational speeds.

In this paper, we investigate reducing memory bandwidth and overall I/O and memory access times by using multithreaded compression and decompression of large datasets. Since the goal is to achieve a significant overall speedup of I/O, both level of compression achieved and efficiency of the compression and decompression algorithms, are of importance. Several compression methods for efficient disk access for large seismic datasets are implemented and empirically tested on on several modern CPUs and GPUs, including the Intel i7 and NVIDIA c2050 GPU. To reduce I/O time, both lossless and lossy symmetrical compression algorithms as well as hardware alternatives, are tested.

Results show that I/O speedup may double by using an SSD vs. HDD disk on larger seismic datasets. Lossy methods investigated include variations of DCT-based methods in several dimensions, and combining these with lossless compression methods such as RLE (Run-Length Encoding) and Huffman encoding. Our best compression rate (0.16%) and speedups (6 for HDD and 3.2 for SSD) are achieved by using DCT in 3D and combining this with a modified RLE for lossy methods. It has an average error of 0.46% which is very acceptable for seismic applications.

A simple predictive model for the execution time is also developed and shows an error of maximum 5% vs. our obtained results. It should thus be a good tool for predicting when to take advantage of multithreaded compression. This model and other techniques developed in this paper should also be applicable to several other data intensive applications.

# An Efficient k-means Algorithm on CUDA

Jiadong Wu and Bo Hong
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
jwu65, bohong@gatech.edu

## Abstract

The *k*-means algorithm is widely used for unsupervised clustering. This paper describes an efficient CUDA-based *k*-means algorithm. Different from existing GPU-based k-means algorithms, our algorithm achieves better efficiency by utilizing the triangle inequality. Our algorithm explores the trade-off between load balance and memory access coalescing through data layout management. Because the effectiveness of the triangle inequity depends on the input data, we further propose a hybrid algorithm that adaptively determines whether to apply the triangle inequality. The efficiency of our algorithm is validated through extensive experiments, which demonstrate improved performance over existing CPU-based and CUDA-based k-means algorithms, in terms of both speed and scalability.

# Workshop 16

# Workshop on Large-Scale Parallel Processing

# LSPP 2011

# Parallel sparse matrix-vector multiplication as a test case for hybrid MPI+OpenMP programming

Gerald Schubert[1] , Georg Hager[1] , Holger Fehske[2] , Gerhard Wellein[1]

[1]Erlangen Regional Computing Center
University of Erlangen-Nuremberg
91058 Erlangen, Germany
gerald.schubert@rrze.uni-erlangen.de

[2]Institute for Physics
University of Greifswald
17487 Greifswald, Germany
fehske@physik.uni-greifswald.de

## Abstract

We evaluate optimized parallel sparse matrix-vector operations for two representative application areas on widespread multicore-based cluster configurations. First the single-socket baseline performance is analyzed and modeled with respect to basic architectural properties of standard multicore chips. Going beyond the single node, parallel sparse matrix-vector operations often suffer from an unfavorable communication to computation ratio. Starting from the observation that nonblocking MPI is not able to hide communication cost using standard MPI implementations, we demonstrate that explicit overlap of communication and computation can be achieved by using a dedicated communication thread, which may run on a virtual core. We compare our approach to pure MPI and the widely used "vector-like" hybrid programming strategy.

# HIO: A Library for High Performance I/O and Data Management

William W. Dai
Computer, Computational, and Statistical Sciences Division
Los Alamos National Laboratory
E-mail: dai@lanl.gov

## Abstract

A library for parallel IO and data management has been developed for large-scale multi-physics simulations. The goal of the library is to provide sustainable, interoperable, efficient, scalable, and convenient tools for parallel IO and data management for high-level data structures in applications, and to provide tools for the connection between applications. The high-level data structures include one- and multi-dimensional arrays, structured meshes, unstructured meshes, and the meshes generated through adaptive mesh refinement. The IO mechanism can be collective and non-collective. The data objects suitable for the library could be either large or small data sets. Even for small data sets, the IO performance is close to one of MPI-IO performance.

# Parallel Processor Core for Semantic Search Engines

Suneil Mohan, Aalap Tripathy, Amitava Biswas, Rabi Mahapatra
Department of Computer Science and Engineering,
Texas A&M University,
College Station, Texas, USA
{suneil,aalap,amitabi,rabi}@cse.tamu.edu

## Abstract

Superior and fast semantic comparison improves the quality of web-search. Semantic comparison involves dot product computation of large sparse tensors which is time consuming and expensive. In this paper we present a low power parallel architecture that consumes only 15.41 Watts and demonstrates a speed-up in the order of $10^5$ compared to a contemporary hardware design, and in the order of $10^4$ compared to a purely software approach. Such high performance low power architecture can be used in semantic routers to elegantly implement energy efficient distributed search engines.

# MTASC: A Multithreaded Associative SIMD Processor

Kevin Schaffer and Robert A. Walker
Department of Computer Science
Kent State University
Kent, Ohio, USA
{kschaffe, walker}@cs.kent.edu

## Abstract

In this paper we describe the architecture of MTASC, a multithreaded associative SIMD processor, and a cycle-accurate instruction set simulator for that architecture. We show, through simulations of a set of five associative benchmarks, that this architecture is capable of significantly improving the performance of associative code over a single-threaded architecture, especially for processors with a large number of PEs. Furthermore, we show that the amount of improvement in performance is highly dependent on the frequency of reduction instructions in the code being executed.

# High Precision Integer Multiplication with a GPU

Niall Emmart and Charles Weems
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610, USA.
Email: nemmart@yrrid.com, weems@cs.umass.edu

## Abstract

We have improved our prior implementation of Strassen's algorithm for high performance multiplication of very large integers on a general purpose graphics processor (GPU). A combination of algorithmic and implementation optimizations result in a factor of 2.3 speed improvement over our previous work, running on an NVIDIA 295. We have also reoptimized the implementation for an NVIDIA 480, from which we obtain a factor of up to 10 speedup in comparison with a Core i7 processor of the same technology generation. This paper discusses how we adapted the algorithm to operate within the limitations of the GPU and how we dealt with other issues encountered in the implementation process, as well as reporting performance results for a multiplications ranging from 255K bits, to 24.512M bits in size.

# Distributed B&B: A Pure Peer-to-Peer Approach

Mathieu Djamaï, Bilel Derbel, Nouredine Melab
Laboratoire d'Informatique Fondamentale de Lille (LIFL)
Universitédes Sciences et Technologies de Lille.
Lille,France
Email : {mathieu.djamai,bilel.derbel,nouredine.melab}@lifl.fr

## Abstract

The state-of-the-art large scale approach for solving NP-hard permutation-based problems using parallel Branch-and-Bound (B&B) techniques is based on a Master-Slave model which is known to be limited in terms of scalability. In this paper, we present a new Peer-to-Peer (P2P) approach that can handle a huge amount of computational resources in a fully distributed way, that is without the need of any centralized coordinator. To achieve that, we propose simple and efficient fully distributed algorithms dealing with major parallel B&B issues such as work sharing, dynamic load balancing and termination detection. We argue that our P2P approach has a scalability which is exponentially better in theory compared to the Master-Slave technique while having a negligible communication overhead in a worst case-scenario, namely poly-logarithmic. The approach has been implemented and experimented using the Grid'5000 nation-wide French grid. Through extensive simulations involving up to 150 000 peers, we show that, compared to the state-of-the-art Master-Slave technique, our P2P approach enables (i) to improve the parallel efficiency up to a ratio of 6 to 1, (ii) to significantly speed up the B&B search process, namely by up to 7 factor in terms of number of solutions explored in the search space and, (iii) to keep the communication overhead relatively low, namely by a factor of at most 11 without penalizing the search process.

## An Adaptive Framework for Large-scale State Space Search

Yanhua Sun, Gengbin Zheng, Pritish Jetley, Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{sun51, gzheng, pjetley2, kale}@illinois.edu

### Abstract

State space search problems abound in the artificial intelligence, planning and optimization literature. Solving such problems is generally NP-hard. Therefore, a brute-force approach to state space search must be employed. It is instructive to solve them on large parallel machines with significant computational power. However, writing efficient and scalable parallel programs has traditionally been a challenging undertaking. In this paper, we analyze several performance characteristics common to all parallel state space search applications. In particular, we focus on the issues of grain size, the prioritized execution of tasks and the balancing of load among processors in the system. We demonstrate the techniques that are used to scale such applications to large scale. We have incorporated these techniques into a general search engine framework that is designed to solve a broad class of state space search problems. We demonstrate the efficiency and scalability of our design using three example applications, and present scaling results up to 16,384 processors.

## Fault-Tolerant Mechanism for Hierarchical Branch and Bound Algorithm

A. Bendjoudi
Centre de Recherche sur l'Information Scientifique
et Technique CERIST, DTISI
3 rue des frères Aissou, 16030 Ben-Aknoun, Algiers, Algeria
Université Abderrahmane Mira de Béjaia
Route de Targa Ouzemmour, 06000 Béjaia, Algeria
Email: ahcene.bendjoudi@gmail.com
N. Melab and E-G. Talbi
Université des Sciences et Technologies de Lille, France
LIFL/UMR CNRS 8022
59655 - Villeneuve d'Ascq cedex - France
Email: {nouredine.melab,el-ghazali.talbi}@lifl.fr

### Abstract

Solving exactly large instances of Combinatorial Optimization Problems *(COPs)* using Branch and Bound *(B&B)* algorithms requires a huge amount of computing resources. These resources can be offered by computational grids and the scalability can be achieved using Hierarchical Master/Worker-based B&B pushing the limits of the traditional Master/Worker paradigm. However, the resources offered by grids are most of the time unreliable, volatile, and heterogeneous. Therefore, they must take into account fault tolerance. In this paper, we present FTH-B&B, a fault tolerant hierarchical B&B, in order to deal with the fault tolerance issue. It is composed of several fault tolerant Master/Worker-based sub-B&Bs organized hierarchically into groups and perform independently fault tolerant mechanism. Beside, a fault recovery mechanism is introduced to recover and avoid redundant exploration of sub-problems in case of failures. In addition, we propose a mechanism to maintain the hierarchy safe and balanced during the lifetime of the algorithm. Our algorithm is applied to the Flow-Shop scheduling problem *(FSP)* and implemented on top of the ProActive grid middleware. It has been promisingly experimented on the Grid'5000 French nation-wide grid and shows its ability to remain efficient even in presence of failures.

# Hierarchical Mapping for HPC Applications

I-Hsin Chung
IBM T.J. Watson Research Center
Yorktown Heights, NY USA
ihchung@us.ibm.com

Che-Rung Lee Jiazheng Zhou Yeh-Ching Chung
National Tsing-Hua University
Hsin-Chu, Taiwan
{cherung,jzzhou,ychung}@cs.nthu.edu.tw

## Abstract

As the high performance computing systems scale up, mapping the tasks of a parallel application onto physical processors to allow efficient communication becomes one of the critical performance issues. Existing algorithms were usually designed to map applications with regular communication patterns. Their mapping criterion usually overlooks the size of communicated messages, which is the primary factor of communication time. In addition, most of their time complexities are too high to process large scale problems.

In this paper, we present a hierarchical mapping algorithm (HMA), which is capable of mapping applications with irregular communication patterns. It first partitions tasks according to their run-time communication information. The tasks that communicate with each others more frequently are regarded as strongly connected. Based on their connectivity strength, the tasks are partitioned into supernodes based on the algorithms in spectral graph theory. The hierarchical partitioning reduces the mapping algorithm complexity to achieve scalability. Finally, the run-time communication information will be used again in fine tuning to explore better mappings. With the experiments, we show how the mapping algorithm helps to reduce the point-to-point communication time for the PDGEMM, a ScaLAPACK matrix multiplication computation kernel, up to 20% and the AMG2006, a tier 1 application of the Sequoia benchmark, up to 7%.

# A Performance Model of Direct Numerical Simulation for Analyzing Large-Scale Systems

Darren J. Kerbyson and Kevin J. Barker
Performance and Architecture Lab
Pacific Northwest National Laboratory
Richland, WA, USA
e-mail: {Darren.Kerbyson,Kevin.Barker}@pnl.gov

## Abstract

In this work we develop, validate and use a performance model of the combustion code DNS3D that uses Direct Numerical Simulation to model turbulence in fluid motion. This is an important application area and is expected to use large fractions of future large-scale systems. From a thorough analysis of the code, its key computational characteristics are coupled with the performance characteristics of a parallel system using an analytical performance model that is parameterized in terms of key application and system aspects. The performance model is validated on two current systems, a multi-core AMD Opteron system with an Infiniband interconnect, and an IBM Power5+ system with an HPS interconnect. High prediction accuracies are demonstrated on these two systems. The model can be used to explore changes in the performance characteristics of either the application or of the systems, and in particular to examine the processing possibilities for DNS3D on future systems. We illustrate here how the model can be used to explore the impact of changes in the characteristics of either the system or of the application.

# Investigating the Impact of the Cielo Cray XE6 Architecture on Scientific Application Codes

Courtenay Vaughan, Mahesh Rajan, Richard Barrett, Doug Doerfler, and Kevin Pedretti
Center for Computing Research, Sandia National Laboratories
P.O. Box 5800, MS 1319, Albuquerque, NM, 87185-1319, USA
Email: ctvaugh, mrajan, rfbarre, dwdoerf, ktpedre@sandia.gov

## Abstract

Cielo, a Cray XE6, is the Department of Energy NNSA Advanced Simulation and Computing (ASC) campaign's newest capability machine. Rated at 1.37 PFLOPS, it consists of 8,944 dual-socket oct-core AMD Magny-Cours compute nodes, linked using Cray's Gemini interconnect. Its primary mission objective is to enable a suite of the ASC applications implemented using MPI to scale to tens of thousands of cores. Cielo is an evolutionary improvement to a successful architecture previously available to many of our codes, thus enabling a basis for understanding the capabilities of this new architecture. Using three codes strategically important to the ASC campaign, and supplemented with some micro-benchmarks that expose the fundamental capabilities of the XE6, we report on the performance characteristics and capabilities of Cielo.

# Workshop 17

# Workshop on Large-Scale, Volatile Desktop Grids

# PCGRID 2011

# Emulating Volunteer Computing Scheduling Policies

David P. Anderson
University of California, Berkeley
davea@ssl.berkeley.edu

## Abstract

Volunteer computing systems such as BOINC use several interacting scheduling policies, which must address multiple requirements across a large space of usage scenarios. In developing BOINC, we need to design and optimize these policies without direct access to the target nodes (volunteered PCs). To do this, we developed an emulation-based system that predicts the policies' behavior in specific scenarios. This system has been useful in the design and evaluation of policies, in software development, and in the resolution of problems occurring in the field.

# Computing the Tree of Life - Leveraging the Power of Desktop and Service Grids

Adam L. Bazinet and Michael P. Cummings
Center for Bioinformatics and Computational Biology
University of Maryland
College Park, MD 20742 USA
pknut777@umiacs.umd.edu, mike@umiacs.umd.edu

## Abstract

The trend in life sciences research, particularly in molecular evolutionary systematics, is toward larger data sets and ever-more detailed evolutionary models, which can generate substantial computational loads. Over the past several years we have developed a grid computing system aimed at providing researchers the computational power needed to complete such analyses in a timely manner. Our grid system, known as The Lattice Project, was the first to combine two models of grid computing - the service model, which mainly federates large institutional HPC resources, and the desktop model, which harnesses the power of PCs volunteered by the general public. Recently we have developed a ”science portal” style web interface that makes it easier than ever for phylogenetic analyses to be completed using GARLI, a popular program that uses a maximum likelihood method to infer the evolutionary history of organisms on the basis of genetic sequence data. This paper describes our approach to scheduling thousands of GARLI jobs with diverse requirements to heterogeneous grid resources, which include volunteer computers running BOINC software. A key component of this system provides a priori GARLI runtime estimates using machine learning with random forests.

# Volunteer Cloud Computing: MapReduce over the Internet

Fernando Costa, Luis Silva
CISUC, Dept. of Informatics Engineering
University of Coimbra, Portugal

Michael Dahlin
Laboratory for Advanced Systems Research
University of Texas Austin

## Abstract

Volunteer Computing harnesses computing resources of machines from around the world to perform distributed independent tasks, following a master/worker model. Despite the recent increase in popularity and power in middleware such as BOINC, there are still several limitations in existing systems. Current research is oriented towards optimizing existing applications, while the number of active users and projects has reached a plateau.

A programming paradigm that has been significantly popular and is used by several systems on the cloud is MapReduce. The main advantage of this paradigm is that it can be used to solve a vast amount of different problems, by breaking them into simple steps and taking advantage of distributed resources. Volunteer Computing provides these resources, and although it cannot match the conditions offered by a cluster, it has other advantages that can be leveraged. In this paper, we try to increase the computational power of Volunteer Computing systems by allowing more complex applications and paradigms such as MapReduce to be run, thus opening new avenues and possibilities for the use of computational devices scattered through the Internet.

We created a BOINC prototype that can run MapReduce jobs (BOINC-MR), using a pull-model in which communication is always initiated by the client. By running experiments on a small cluster, with multiple variables, we were able to evaluate a few initial scenarios with this paradigm. We used a simple MapReduce application, word count, as proof of concept, just to demonstrate a typical execution.

## Evolving N-Body Simulations to Determine the Origin and Structure of the Milky Way Galaxy's Halo using Volunteer Computing

Travis Desell, Malik Magdon-Ismail, Boleslaw Szymanski,
Carlos A. Varela
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY, USA
deselt,magdon,szymansk,cvarela@cs.rpi.edu
Benjamin A. Willett, Matthew Arsenault,
Heidi Newberg
Department of Physics, Applied Physics and Astronomy
Rensselaer Polytechnic Institute
Troy, NY, USA
willeb,arsenm2,heidi@rpi.edu

### Abstract

This work describes research done by the MilkyWay@Home project to use N-Body simulations to model the formation of the Milky Way Galaxy's halo. While there have been previous efforts to use N-Body simulations to perform astronomical modeling, to our knowledge this is the first to use evolutionary algorithms to discover the initial parameters to the N-Body simulations so that they accurately model astronomical data. Performing a single 32,000 body simulation can take up to 200 hours on a typical processor, with an average of 15 hours. As optimizing the input parameters to these N-Body simulations typically takes at least 30,000 or more simulations, this work is made possible by utilizing the computing power of the 35,000 volunteered hosts at the MilkyWay@Home project, which are currently providing around 800 teraFLOPS. This work also describes improvements to an open-source framework for generic distributed optimization (FGDO), which provide more efficient validation in performing these evolutionary algorithms in conjunction the Berkeley Open Infrastructure for Network Computing (BOINC).

## DGVCS Security from a Different Perspective: Challenges and Hurdles

Tobias Dussa
Computer Emergency Response Team (CERT)
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
Email: tobias.dussa@kit.edu

### Abstract

This paper discusses Desktop Grid and Volunteer Computing System (DGVCS) security issues from a different point of view, namely that of an organisation security team that is external to the unit running DGVCS. In an existing large-scale environment, it is likely that there are techniques and procedures in place that deal with all manners of security. Introducing DGVCS into such structures brings along unique challenges. Besides modifications of the techniques and procedures already deployed, highly-distributed systems like DGVCS require additional effort especially when handling and containing incidents: With DGVCS, cross-site and cross-organisation collaboration is necessary in order to be able to effectively counter security threats.

# Attic: A Case Study for Distributing Data in BOINC Projects

AbdelHamid Elwaer, Andrew Harrison, Ian Kelley and Ian Taylor
School Of computer Science and Informatics
Cardiff University
Cardiff, UK
a.elwaer@cs.cardiff.ac.uk

## Abstract

The BOINC toolkit is a very successful software infrastructure for the development of “volunteer computing” projects. For data distribution, the standard BOINC architecture provides uses a centralized server, or fixed set of pre-configured servers, each with a copy of a project’s input files. This mirrored configuration has provided a way for BOINC projects to increase the size of their user-base and manage their data as projects grow. However, such a centrally administered approach can be limiting in terms of performance metrics, such as scalability and reliability, and also for management, since such servers have to be manually installed and configured for use. In this paper, we study an alternative approach, which is capable of providing a more lightweight and dynamic environment for distributing data, called Attic. Attic makes use of a secure decentralized approach and BitTorrent-like file swarming techniques to serve data and manage load. To compare our approach to the current centralized infrastructure, we show three experiments where we have integrated Attic with BOINC; each compares how Attic performs against BOINC as fragments (chunks) of the original data file are distributed and downloaded concurrently from multiple data centers. Our results show that careful choice of chunk size, by analyzing a server’s uploading capabilities and file size, can result in large throughput gains for Attic when compared to the current BOINC data distribution paths.

# libboincexec: A generic virtualization approach for the BOINC middleware

Diogo Ferreira, Filipe Araujo
CISUC
Dept. of Informatics Engineering
University of Coimbra, Portugal
{defer,filipius}@dei.uc.pt

Patricio Domingues
Research Center for Informatics and Communiations
School of Technology and Management
Polytechnic Institute of Leiria, Leiria, Portugal
patricio@estg.ipleiria.pt

## Abstract

BOINC is a client-server desktop grid middleware that has grown to power very large computational projects. BOINC clients request computing jobs to a central server and run them alongside other regular applications. Unfortunately, this kind of execution causes two kinds of problems. Firstly, developers must port their application to every single operating system target, which usually means maintaining several different versions of the same application. Secondly, any application running natively on desktop grid hardware is a potential security threat to the volunteer client. During the course of this research we sought an efficient and generic method for alternative execution of jobs in BOINC clients. Our approach is strongly guided by the principles of non-intrusiveness and contains two main components. The first is a library, libboincexec, which is able to control several virtual machines monitors. The second is a modified BOINC wrapper that provides the glue between libboincexec and the middleware. Through the use of this solution we are able to effectively use virtual machines to perform computation on desktop grids. This computation is inherently safe because virtual machines provide sandboxing. Additionally, by targeting the same virtual operating system, the problem of maintaining different versions of an application does not exist, thereby solving the heterogeneity problem of desktop grid nodes.

## How to make BOINC-based desktop grids even more popular?

Peter KACSUK[a,b]
[a] MTA SZTAKI, 1518 Budapest P.O. Box 63., Hungary
[b] University of Westminster, London, UK
e-mail: kacsuk@sztaki.hu

### Abstract

Although BOINC-based volunteer desktop grids are quite popular there are new potential usage scenarios that could make BOINC systems even more popular both for volunteer computing and for establishing institutional desktop grids. In the paper we show three major new usage scenarios that will significantly increase the use of BOINC in the future. The extension of service grids with support of inexpensive BOINC resources will turn the attention of service grid users towards BOINC systems. The access of BOINC systems via workflow-oriented science gateways will enlarge the number of user communities that can use BOINC systems in their daily work. Finally, providing QoS requirements in BOINC systems by extending them with on-demand cloud resources will attract user communities whose applications have got some time constraints.

## Fedak,Distributed Results Checking for MapReduce in Volunteer Computing

Mircea Moca, Gheorghe Cosmin Silaghi
Babeș-Bolyai University of Cluj-Napoca, România
Email: {mircea.moca, gheorghe.silaghi}@econ.ubbcluj.ro

Gilles Fedak
INRIA/Université de Lyon, France
Email: Gilles.Fedak@inria.fr

### Abstract

MapReduce is a promising approach to support data-intensive applications on Volunteer Computing Systems. Existent middleware like BitDew allows running MapReduce applications in a Desktop Grid environment. If the Desk- top Grid is deployed in the Internet under the Volunteer Computing paradigm, it harnesses untrustable, volatile and heterogeneous resources and the results produced by MapRe- duce applications can be subject of sabotage. However, the implementation of large-scale MapReduce presents significant challenges with respect to the state of the art in Desktop Grid. A key issue is the design of the result certification, an operation needed to verify that malicious volunteers do not tamper with the results of computations. Because the volume of data produced and processed is so large that cannot be sent back to the server, the result certification cannot be centralized as it is currently implemented in Desktop Grid systems. In this paper we present a distributed result checker based on the Majority Voting method. We evaluate the efficiency of our approach using a model for characterizing errors and sabotage in the MapReduce paradigm. With this model, we can compute the aggregated probability with which a MapReduce implementation produces an erroneous result. The challenge is to capture the aggregated probability for the entire system, composed from probabilities resulted from the two phases of computation: Map and Reduce. We provide a detailed analysis on the performance of the result verification method and also discuss the generated overhead of managing security. We also give guidelines about how the result verification phase should be configured, given a MapReduce application.

## Building online computing service over volunteer grid resources

Mark Silberstein
Technion
marks@cs.technion.ac.il

### Abstract

Volunteer computing grids have traditionally been used for massively parallel workloads, such as processing data from large scientific experiments. We argue that the domain of volunteer grids can be extended well beyond this specific niche, by enhancing them with built-in mechanisms for integration with with standard clusters, grids and clouds, to compensate for unexpected fluctuations in resource availability and quality of service. The resulting capabilities for on-demand dynamic expansion of the resource pool, together with sophisticated scheduling mechanisms will turn volunteer grids into a powerful execution platform for on-line interactive computing services.

We will show our experience with the GridBoT system, which implements these ideas. GridBoT is part of a production high performance online service for genetic linkage analysis, called Superlink-online. The system enables anyone with the Internet access to submit genetic data, and easily and quickly analyze it as if using a supercomputer. The analyses are automatically parallelized and executed via GridBoT on over 45,000 non-dedicated machines from the Superlink@Technion volunteer grid, as well as on 9 other grids and clouds, including the Aamazon EC2. Since 2009 the system has served more than 300 geneticists from leading research institutions worldwide, and executed over 6500 different real analysis runs, with about 10 million tasks consumed over 420 CPU years.

## Increasing Participation in Volunteer Computing

David Toth
Computer Science Department
Merrimack College
North Andover, MA 01845
USA
david.toth@merrimack.edu

Russell Mayer
Political Science Department
Merrimack College
North Andover, MA 01845
USA
russell.mayer@merrimack.edu

Wendy Nichols
Communication Studies Department
Merrimack College
North Andover, MA 01845
USA
wendy.nichols@merrimack.edu

### Abstract

Volunteer computing is a distributed computing model where a large problem requiring the computational power of a supercomputer can be solved by thousands or millions of personal computers. In volunteer computing, people allow their personal computers to work together on a large problem when the computers are not otherwise in use. The owners of the computers are not compensated for the use of their computers. As volunteer computing has become an accepted way to solve large problems, the number of volunteer computing projects has increased significantly over the years. However, as the number of volunteer computing projects increases, the participation in volunteer computing must increase to allow volunteer computing to remain a viable way to solve large problems. We studied the reasons that people do not participate in order to determine how participation can be increased. We found that the likelihood of people participating in such projects was not affected by how these projects were described. We also collected information about factors that influenced the subjects' decisions.

# Workshop 18

# Parallel Computing and Optimization

# PCO 2011

# A Parallel Exact Solver for the Three-Index Quadratic Assignment Problem

François Galea
Embedded Real Time Systems Laboratory
CEA, LIST
Gif-sur-Yvette, France
Email: francois.galea@cea.fr

Bertrand Le Cun
Laboratoire PRiSM
University of Versailles-Saint-Quentin, UVSQ
Versailles, France
Email: Bertrand.Lecun@prism.uvsq.fr

## Abstract

Computers with multiple processor cores using shared memory are now ubiquitous. This paper reports an implementation of a branch-and-bound–based exact algorithm for the Three-Index Quadratic Assignment Problem (Q3AP) on multicore processors. Our parallel implementation has two levels of parallelism. The first, the most common parallelizes the tree search procedure using the Bob++ framework. The second one parallelizes the computation of the lower bound using the SIMD instruction set extensions of modern processors.

# Distributed Bees Foraging-Based Algorithm for Large-Scale Problems

Antonio Gómez-Iglesias, Francisco Castejón
National Fusion Laboratory
Euratom-CIEMAT Association
Madrid, Spain
{antonio.gomez,francisco.castejon}@ciemat.es
Miguel A. Vega-Rodríguez
Dept. Technologies of Computers and Communications
University of Extremadura
Cáceres, Spain
mavega@unex.es

## Abstract

The optimization process of large-scale computational problems presents various issues that need to be solved in order to achieve favorable results. One of the most significant challenges faced is the computational costs of these problems. These costs force researchers to use large-scale computational resources. Parallel implementations of different algorithms may be considered. However, these techniques present issues when dealing with problems in which a high variability in the execution time of the evaluation function takes place. This paper presents an algorithm designed to efficiently optimize these large-scale optimization problems with different execution times for the evaluation of the candidate solutions. Production grid computing infrastructures are used to carry out these optimizations. The own paradigm of the grid introduces new challenges in the process that must to be taken into account in the design of the algorithm.

## On-Line Optimization of Publish/Subscribe Overlays

Eddy Caron, Benjamin Depardon
University of Lyon - LIP Laboratory - UMR CNRS
ENS Lyon - INRIA - UCBL 5668, FRANCE
{Eddy.Caron,Benjamin.Depardon}@ens-lyon.fr

Ajoy K. Datta, Lawrence L. Larmore
University of Nevada
Las Vegas, USA
{Datta, Larmore}@cs.unlv.edu

### Abstract

Loosely coupled applications can take advantage of the publish/subscribe communication paradigm. In this latter, *subscribers* declare which events, or which range of events, they wish to monitor, and are asynchronously informed whenever a *publishers* throws an event. In such a system, when a publication occurs, all peers whose subscriptions contain the publication must be informed. In our approach, the subscriptions are represented by a DR-tree, which is an R-tree where each minimum bounding rectangle is supervised by a peer. Instead of attempting to statically optimize the DR-tree, we give an on-line algorithm, the *work function algorithm*, which continually changes the DR-tree in response to the sequence of publications, in attempt to dynamically optimize the structure. The competitiveness of this algorithm is computed to be at most 5 for any example where there are at most three subscriptions and the R-tree has height 2. The benefit of the on-line approach is that no prior knowledge of the distribution of publications in the attribute space is needed.

## Communication in Parallel Algorithms for Constraint-Based Local Search

Yves Caniou
JFLI, CNRS / NII
Tokyo, Japan
Yves.Caniou@ens-lyon.fr

Philippe Codognet
JFLI, CNRS / UPMC / University of Tokyo
Tokyo, Japan
Philippe.Codognet@lip6.fr

### Abstract

We address the issue of parallelizing constraint solvers based on local search methods for massively parallel architectures, involving several thousands of CPUs. We present a family of a constraint-based local search algorithms and investigate their performance results on hardwares with several hundreds of processors. The first method is a basic independent multiple-walk algorithm: each processor runs a local search starting from a distinct initial configuration and the first one which will reach a solution will notify the others and stop all computations. These simple methods have good performances, and good speedups can be achieved up to a few hundreds of processors. Then we consider 2 versions with communication between processors: 1) every $c$ iterations, each processor sends the current value (cost) of its configuration to others and a processor who received a better cost from another processor can decide to stop its current search with a probability $p$; 2) the number of iterations corresponding to the cost is also transfered. Both the received cost and the number of iterations have to be better for a processor to decide to draw a probability and restart.

Several experiments involving more than 100 processors have been conducted and different values of $p$ have been tried to consider more or less ”autistic” processors. However results show that it is very difficult to achieve better performance than the initial method without communication.

# Performance Optimization with Energy Constraint in Heterogeneous Multiple Computer Systems

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
*Email: lik@newpaltz.edu*

## Abstract

The problem of minimizing average task response time in heterogeneous multiple computer systems with energy constraint is considered. The average task response time in an entire system of multiple computers is formulated as a function of power allocations to the computers. The average task response time is minimized subjected to the constraint that the expected energy consumption of all the computers over certain period of time does not exceed a given energy budget. The minimization problem is solved by finding an optimal power allocation to the computers. An algorithm is developed to solve our optimization problem.

# Obtaining Simultaneous Equation Models through a unified shared-memory scheme of metaheuristics

Francisco Almeida,
Departamento de Estadística, Investigación Operativa y Computación,
University of La Laguna,
Tenerife Island, Spain,
falmeida@ull.es

Domingo Giménez,
Departamento de Informática y Sistemas,
University of Murcia, Spain,
domingo@um.es

Jose J. López-Espín,
Centro de Investigación Operativa,
University Miguel Hernández,
Elche, Spain,
jlopez@umh.es

## Abstract

A Simultaneous Equation Model represents simultaneous dependencies in a set of variables. These models are normally created by experts in the field, but in some cases it is difficult to obtain such a model, for example due to a large number of variables, to unclear dependencies, etc. Furthermore, sometimes it is necessary to evaluate models composed of different variables before to obtain the values of the variables in the model and subsequently a satisfactory model. It is possible to develop metaheuristics to help the expert in the automatic generation of satisfactory models. But it is necessary to experiment with several metaheuristics and tune them for the problem. Furthermore, inside a metaheuristic a large number of models are evaluated, and when the number of variables is large, the evaluation of the models is very time consuming. This paper presents some metaheuristics for obtaining Simultaneous Equation Models from a set of values of the variables. A unified shared-memory scheme for metaheuristics is used, which allows the easy application and tuning of different metaheuristics and combinations of them. Shared-memory versions of the metaheuristics are developed to reduce the execution time. To obtain parallel versions of the metaheuristics quickly, the unified metaheuristic scheme is used, so obtaining a unified parallel scheme for metaheuristics. The different functions in the scheme are parallelized independently, and each function is parameterized with a different number of threads, which allows us to select a different number of threads for each function and metaheuristic, so adapting the parallel scheme to the metaheuristic, the computational system and the problem. Experiments with GRASP, genetic algorithms, scatter search and combinations of them are shown.

# Distributed-Memory Parallel Algorithms for Matching and Coloring

Ümit V. Çatalyürek*, Florin Dobrian†, Assefaw Gebremedhin‡, Mahantesh Halappanavar§, Alex Pothen‡
* Depts. of Biomedial Informatics and Electrical & Computer Engineering, The Ohio State University
Email: umit@bmi.osu.edu
† Conviva
Email: dobrian@conviva.com
‡ Department of Computer Science, Purdue University
Email: {agebreme,apothen}@purdue.edu
§ Pacific Northwest National Laboratory
Email: Mahantesh.Halappanavar@pnl.gov

## Abstract

We discuss the design and implementation of new highly-scalable distributed-memory parallel algorithms for two prototypical graph problems, edge-weighted matching and distance-1 vertex coloring. Graph algorithms in general have low concurrency, poor data locality, and high ratio of data access to computation costs, making it challenging to achieve scalability on massively parallel machines. We overcome this challenge by employing a variety of techniques, including speculation and iteration, optimized communication, and randomization. We present preliminary results on weak and strong scalability studies conducted on an IBM Blue Gene/P machine employing up to tens of thousands of processors. The results show that the algorithms hold strong potential for computing at petascale.

## Automated, Parallel Optimization Algorithms for Stochastic Functions

Dheeraj Chahal* , Steven J. Stuart† , Sebastian Goasguen* and Colin J. Trout‡
* School of Computing
Clemson University, Clemson, SC 29634
Email: dchahal,sebgoa@clemson.edu
† Department of Chemistry
Clemson University, Clemson, SC 29634
Email: ss@clemson.edu
‡ Department of Chemistry
Saint Francis University, Loretto, PA 15940
Email: cjtstr@mail.francis.edu

### Abstract

We propose a hierarchical framework and new parallel algorithms for stochastic function optimization under conditions where the function to be optimized is subject to random noise, the variance of which decreases with sampling time. This is the situation expected for many real-world and simulation applications where results are obtained from sam- pling, and contain experimental error or random noise. Our new optimization algorithms are based on a downhill simplex algorithm, with extensions that alter the timing of simplex operations based on the level of noise in the function evaluations. Three proposed optimization methods, which we term maxnoise, point-to-point comparison, and a combination of these two, are tested on the Rosenbrock function and found to be better than previous stochastic optimization methods. The parallel framework implementing the optimization algorithms is also new, and is based on a master-worker architecture where each worker runs a massively parallel program. The parallel implementation allows the sampling to proceed independently on multiple processors, and is demonstrated to scale well up to over 100 vertices . It is highly suitable for clusters with an ever increasing number of cores per node. The new methods have been applied successfully to the reparameterization of the TIP4P water model, achieving thermodynamic and structural results for liquid water that are as good as or better than the original model, with the advantage of a fully automated parameterization process.

## Efficient implementation of the Simplex method on a CPU-GPU system

Mohamed Esseghir Lalami, Vincent Boyer, Didier El-Baz
CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse, France
Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse France
Email: mlalami@laas.fr vboyer@laas.fr elbaz@laas.fr

### Abstract

The Simplex algorithm is a well known method to solve linear programming (LP) problems. In this paper, we propose a parallel implementation of the Simplex on a CPU-GPU systems via CUDA. Double precision implementation is used in order to improve the quality of solutions. Computational tests have been carried out on randomly generated instances for non-sparse LP problems. The tests show a maximum speedup of 12:5 on a GTX 260 board.

# IPDPS 2011 PhD Forum

# An Algorithm-Based Recovery Scheme for Exascale Computing

Hui Liu
Department of Mathematical and Computer Sciences
Colorado School of Mines
Golden, CO 80401, USA
huliu@mines.edu

## Abstract

We present an algorithm-based recovery scheme for Exascale computing, which uses both data dependencies and communication-induced redundancies of parallel codes to tolerate fault with low overhead. For some applications, our scheme significantly reduces checkpoint size and introduces no overhead when there is no actual failure in the computation. Fault tolerance Newton's method by tailoring our scheme to the algorithm is performed. Numerical simulations indicate that our scheme introduces much less overhead than diskless checkpointing does.

# Lightweight Methods for Automated Design of Self-Stabilization

Aly Farahat
Department of Computer Science
Michigan Technological University
Houghton, MI 49931, USA

Ali Ebnenasir
Department of Computer Science
Michigan Technological University
Houghton, MI 49931, USA

## Abstract

Software systems are an integral part of almost every modern piece of technology. In complex inter-networked systems, software dependability issues are of a major concern, especially robustness to transient faults. The Self-Stabilization (SS) property guarantees that a system will regain its normal behavior in a finite amount of time when faults stop occurring, and will maintain its normal execution as long as there are no faults. SS is a desirable property of robust network protocols, however, it is intellectually challenging to manually design and verify SS.

In this paper, we propose a new vision for automated design and verification of SS. Specifically, our approach puts forward a paradigm of *synthesize-in-small-scale* and *generalize*. We automatically synthesize SS for small instances of non-stabilizing protocols and provide techniques that help designers in generalizing these small solutions for larger instances of the synthesized protocols. Our preliminary results are promising as we have automatically generated self-stabilizing protocols that are the same as their manually designed versions, and alternative solutions for well known problems. We also discuss our plans for future work.

## A Parallel Time-dependent Multimodal Shortest Path Algorithm Based on Geographical Partitioning

H. Ayed
LITA, University Paul Verlaine
CRP Henri Tudor Luxembourg
hedi.ayed@tudor.lu

### Abstract

This paper deals with the Time Dependent Multimodal Transport Problem (TDMTP), the combination of public and private transport modes to move passengers or goods from a source to a destination departing at a given time. To solve these problems, we previously proposed different serial algorithms. Although efficient on several benchmarks, they are not scalable since the limitation of either processing time or memory. To overcome these limits, we propose in this work a parallel algorithm for solving the TDMTP in very large networks. This new ad hoc parallel approach takes advantage of the possible decomposition of real transport networks according to transport modes and regions. We implemented our algorithm on a parallel shared memory machine, being tested in a super computer cluster. The first experiments on random theoretical graphs as well as on real networks are very promising.

## Communication Optimization Beyond MPI

Andrew Friedley
Indiana University
Bloomington, IN, USA

Andrew Lumsdaine
Indiana University
Bloomington, IN, USA

### Abstract

The Message Passing Interface (MPI) is the de-facto standard for parallel processing on high-performance computing systems. As a result, significant research effort has been spent on optimizing the performance of MPI implementations. However, MPI's specific semantics can limit performance when using modern networks with Remote DMA capabilities. We propose a compiler-assisted approach to optimization of MPI-based applications by transforming MPI calls to a one-sided (as opposed to message passing) communication model. In this paper we present a research plan for developing new optimizations using this approach, then show preliminary results with up to a 40% increase in bandwidth over MPI by using a simpler one-sided communication model.

## Efficient Agreement Protocols in Asynchronous Distributed Systems

Izabela Moise
Université de Rennes 1, IRISA / INRIA Rennes, France
Izabela.Moise@irisa.fr [3]

### Abstract

The Consensus problem is a central paradigm of fault-tolerant distributed computing. In a purely asynchronous system, Consensus is impossible to solve in deterministic manner. However, by enriching the system with some synchrony assumptions, several solutions have been proposed in order to circumvent the impossibility result, among which the Paxos approach introduced by Lamport. The repeated use of consensus protocols as building-blocks for higher-level applications, motivated research to optimize their performances. Two main optimizations have been proposed that reduce the latency of learning a decision. However, the second optimization succeeds only in favorable circumstances, otherwise it can be counterproductive and increase the total latency of the protocol. We proposed an efficient framework, called the Paxos-MIC protocol that combines two optimizations dynamically. It solves several consensus instances and guarantees the persistence of all decision values. The protocol is adaptive as it tries to obtain the best performance gain depending on the current context. Between two consecutive consensus instances, the leader determines if the second optimization has to be triggered or not. We analyzed its behavior when both optimizations (or just the first) are used. We studied favorable and unfavorable scenarios where the second optimization may lead to an additional cost.

## Efficient Verification Solutions for Message Passing Systems

Subodh Sharma and Ganesh Gopalakrishnan
School of Computing, University of Utah Salt Lake City, UT

### Abstract

We examine the problem of automatically and efficiently verifying the absence of communication related bugs in message passing systems, specifically in programs written using Message Passing Interface (MPI) API. A typical debugging or testing tool will fail to achieve this goal because they do not provide any guar- antee of coverage of non-deterministic communication matches in a message passing program. While dynamic verification tools do provide such a guarantee, they are quickly rendered useless when an interleaving explosion is witnessed. The general problem is difficult to solve, though we propose that specialized techniques can be developed that can work on top of dynamic verification schedulers thus making them more efficient. In this work, we provide point solutions to deal with the interleaving explosion. Specifically, we present algorithms that accomplish the following tasks: (i) identifying irrelevant message passing operations (Barriers) in MPI programs that add to the verification complexity and degrade application's performance and (ii) reducing sub- stantially the relevant set of interleavings using symmetry patterns; that needs to be explored for the detection of refusal deadlocks in MPI programs.

# Decentralized Network Bandwidth Prediction and Node Search

Sukhyun Song
Department of Computer Science
University of Maryland, College Park, MD 20742, U.S.A.
Email: shsong@cs.umd.edu

## Abstract

My PhD research addresses how to exploit network bandwidth information and increase the performance of data-intensive wide-area distributed applications. The goal is to solve four specific problems: i) design a decentralized algorithm for network bandwidth prediction, ii) design a decentralized algorithm to find bandwidth-constrained clusters, iii) design a decentralized algorithm to find bandwidth-constrained centroids, and iv) develop a wide-area MapReduce system with optimized data locality as an application of the three algorithms for bandwidth prediction and node search.

# Large-Scale Parallel Monte Carlo Tree Search on GPU

Kamil Rocki, Reiji Suda
Department of Computer Science
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, 113-8654, Tokyo, Japan
Email: kamil.rocki, reiji@is.s.u-tokyo.ac.jp

## Abstract

Monte Carlo Tree Search (MCTS) is a method for making optimal decisions in artificial intelligence (AI) problems, typically move planning in combinatorial games. It combines the generality of random simulation with the precision of tree search. The motivation behind this work is caused by the emerging GPU-based systems and their high computational potential combined with relatively low power usage compared to CPUs. As a problem to be solved I chose to develop an AI GPU(Graphics Processing Unit)-based agent in the game of Reversi (Othello) which provides a sufficiently complex problem for tree searching with non-uniform structure and an average branching factor of over 8. I present an efficient parallel GPU MCTS implementation based on the introduced 'block-parallelism' scheme which combines GPU SIMD thread groups and performs independent searches without any need of intra-GPU or inter-GPU communication. I compare it with a simple leaf parallel scheme which implies certain performance limitations. The obtained results show that using my GPU MCTS implementation on the TSUBAME 2.0 system one GPU can be compared to 100-200 CPU threads depending on factors such as the search time and other MCTS parameters in terms of obtained results. I propose and analyze simultaneous CPU/GPU execution which improves the overall result.

# Memory-aware algorithms and scheduling techniques: from multicore processors to petascale supercomputers

Mathias Jacquelin
LIP, École Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Email: Mathias.Jacquelin@ens-lyon.fr

## Abstract

This paper presents several memory-aware algorithms whose design is optimized for different target platforms. Complex memory architectures have spread in a wide range of systems, from multicore processors within cell phones to supercomputers. This trend enlightens the need to deal with heterogeneity and non uniform memory accesses.

As the memory wall is closing in, taking memory architecture into consideration has become fundamental for large-scale platforms. Designing algorithms and scheduling tasks on such heterogeneous platforms is a challenging task. We present several results in that area as well as future research plans.

# Memory Hierarchy Aware Parallel Priority Based Data Structures

Dinesh Agarwal, Abinashi Dhungel
Georgia State University
Department of Computer Science
Atlanta, Georgia 30303

## Abstract

Since the introduction of multicore architectures there have been a lot of speculations on how to utilize these computationally advanced machines efficiently. Our work explores the bilateral filtering kernel which is a fundamental low-cost non-iterative image processing kernel used to perform directionally dependent image smoothing. At a first glance it appears to be an embarrassingly parallel stencil computation - an easy target for implementation on multicore architectures with multiple levels of cache hierarchies. However, the set of contributing neighbors for the computation is not known apriori, thus the well studied methods for stencil computations are not applicable here. Moreover, the kernel is compute-intensive, it requires a careful observation of hotspots to devise efficient approaches to improve overall performance. Although, research has been conducted on bilateral filtering kernel implementations on multicore architectures, to the best of our knowledge no significant improvements have been reported. In this paper, we discuss why well studied methods do not improve performance, and propose the use of low level parallelism and reduction methods to improve the performance of bilateral filtering. We also present thorough results on a host of architectures including AMD Barcelona, AMD Shanghai, AMD Phenom, Intel Clovertown, Intel Nehalem quad core and Intel Nehalem 32 core machines.

# Parallel Algorithms for Bayesian Networks Structure Learning with Applications to Systems Biology

Olga Nikolova
Bioinformatics and Computational Biology, Department of Computer Engineering
Iowa State University, Ames, Iowa, USA

## Abstract

Bayesian networks (BN) are probabilistic graphical models which are widely utilized in modeling complex biological interactions in the cell. Learning the structure of a BN is an NP-hard problem and existing exact and heuristic solutions do not scale to large enough domains to allow for meaningful modeling of many biological processes. In this work, we present efficient parallel algorithms which push the scale of both exact and heuristic BN structure learning. We demonstrate the applicability of our methods by implementations on an IBM Blue Gene/L and an AMD Opteron cluster, and discuss their significance for future applications to systems biology.

# Fault tolerant data acquisition through dynamic load balancing

MichałSimon [1,2]
[1]Faculty of Automatic Control, Electronics and Computer Science
Silesian University of Technology
Gliwice, Polska
2CERN
Geneva, Switzerland
E-mail: michal.simon@cern.ch

## Abstract

Modern detectors used in high energy physics experiments are complex instruments designed to register collisions of particles at a rate in the MHz range. Data that correspond to a single collision of particles, referred to as an event, are acquired from millions of readout channels, and filtered, first by dedicated hardware, and then by computing farms running sophisticated filtering algorithms. In case of data acquisition systems with single-stage software filtration, due to the high input rate (the order of 100 kHz), the data are usually distributed in a static way between filtering nodes. However, the static distribution determines strongly the system, and results in decreased fault tolerance. The main objective of the presented studies is to increase the system's overall fault tolerance through dynamic load balancing. The proposed method aims to balance the workload inside heterogeneous systems, as well as, homogeneous systems, where the imbalance could be caused by faults. Moreover, our research includes developing a scalable load balancing protocol along with a distributed asynchronous load assignment policy. As a case study we consider the Data Acquisition system of the Compact Muon Solenoid experiment at CERN's new Large Hadron Collider.

# A Codesigned Fault Tolerance System for Heterogeneous Many-Core Processors

Keun Soo Yim and Ravishankar Iyer

## Abstract

This paper presents an efficient fault tolerance system for heterogeneous many-core processors. The efficiencies and coverage of the presented fault tolerance are optimized by customizing the techniques for different types of components in the highest layers of system abstractions and codesigning the techniques in a way that separates algorithms and mechanisms.

# Towards a storage backend optimized for atomic MPI-IO for parallel scientific applications

Viet-Trung Tran
ENS Cachan, Brittany
IRISA
Rennes, France
Email: viet-trung.tran@irisa.fr

## Abstract

We consider the challenges of building data management systems that are optimized for parallel I/O of scientific applications. One of those challenges focuses on the requirement of high I/O throughput for concurrent data accesses while obeying MPI atomicity semantics. In most existing implementations, the atomicity is often implemented through locking-based schemes, which have proven inefficient. In this proposal, we argue that a novel versioning-based scheme can avoid the need to perform expensive synchronization by using multiple snapshots of the same data, which is much more efficient. We describe our prototype of a versioning-based storage back-end and report on promising experimental results.

# Programming Heterogeneous Systems

David M. Kunzman and Laxmikant V. Kale
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
{kunzman2, kale}@illinois.edu

## Abstract

Various specialized hardware designs, such as Cell, GPGPUs, and MIC, have gained traction as alternative hardware designs capable of delivering higher flop rates than conventional designs. However, a drawback of these accelerators is that they simultaneously increase programmer burden in terms of code complexity and decrease portability by requiring hardware specific code to be interleaved throughout application code. The structure of the application code itself typically requires modification when targeting accelerators. Further, balancing the application workload across the cores becomes problematic, especially if a given computation must be split across a mixture of core types with variable performance characteristics.

Our research aims to address the complications that arise in heterogeneous systems by understanding how the application build process and underlying runtime system can assist the programmer in developing parallel programs that target such platforms. We are developing a unified programming model that can be used for all cores, host and accelerator alike. We discuss the modifications we have made to the runtime system, along with discussing future modifications. We have demonstrated a simple molecular dynamics (MD) program executing on a mixture of x86 and Cell processors without requiring hardware specific code within the application code.

# Data Parallel Programming Model for Many-Core Architectures

Yongpeng Zhang
North Carolina State University
yzhang25@ncsu.edu

## Abstract

Emerging accelerating architectures, such as GPUs, have proved successful in providing significant performance gains to various application domains. This is done by exploiting data parallelism in existing algorithms. However, programming in a data-parallel fashion imposes extra burdens to programmers, who are used to writing sequential programs. New programming models and frameworks are needed to reach a balance between programmability, portability and performance. We start from stream processing domain and propose GStream, a general-purpose, scalable data streaming framework on GPUs. The contributions of GStream are as follows: (1) We provide powerful, yet concise language abstractions suitable to describe conventional algorithms as streaming problems. (2) We project these abstractions onto GPUs to fully exploit their inherent massive data-parallelism. (3) We demonstrate the viability of streaming on accelerators. Experiments show that the proposed framework provides flexibility, programmability and performance gains for various benchmarks from a collection of domains, including but not limited to data streaming, data parallel problems, numerical codes and text search. This work lays a foundation to our future work to develop more general data parallel programming models for many-core architectures.

# Detection and Correction of Silent Data Corruption for Large-Scale High-Performance Computing

David Fiala
North Carolina State University
Department of Computer Science
Raleigh, NC, USA
dfiala@ncsu.edu

## Abstract

Faults have become the norm rather than the exception for high-end computing on clusters with 10s/100s of thousands of cores, and this situation will only become more dire as we reach exascale computing. Exacerbating this situation, some of these faults will not be detected, manifesting themselves as silent errors that will corrupt memory while applications continue to operate but report incorrect results. This paper introduces RedMPI, an MPI library residing in the profiling layer of any standards-compliant MPI implementation. RedMPI is capable of both online detection and correction of soft errors that occur in MPI applications without requiring code changes to application source code. By providing redundancy, RedMPI is capable of transparently detecting corrupt messages from MPI processes that become faulted during execution. Furthermore, with triple redundancy RedMPI "votes" out MPI messages of a faulted process by replacing corrupted results with corrected results from unfaulted processes. We present an evaluation of RedMPI on an assortment of applications to demonstrate the effectiveness and assess associated overheads.

Fault injection experiments establish that RedMPI is not only capable of successfully detecting injected faults, but can also correct these faults while carrying a corrupted application to successful completion without propagating invalid data.

# Improving Job Scheduling on Production Supercomputers

Wei Tang (Illinois Institute of Technology)
Zhiling Lan (Illinois Institute of Technology)
Narayan Desai (Argonne National Laboratory)

## Abstract

Job scheduling is a critical task on large-scale supercomputers, where small variety in scheduling policies can result in substantial differences in performance or resource utilization. Tremendous research has been focused on improving job scheduling theoretically. This work aims at addressing the job scheduling problem from practice. Driven by the practical motivating problems, we design and implement job scheduling schemes which can be easily deployed on production machines. All the schemes are evaluated by event-driven simulations using real workload from the production Blue Gene/P system at Argonne National Laboratory. Experimental results show our schemes can effectively improve job scheduling in terms of user satisfaction and system utilization.

# Towards a Self-Adaptive Data Management System for Cloud Environments

Alexandra Carpen-Amarie*
INRIA/IRISA, Rennes, France
Email: alexandra.carpen-amarie@inria.fr

## Abstract

Cloud computing is an increasingly popular paradigm that gained interest from both scientific community and industry. As data volumes processed by applications running on clouds increase, the need for efficient and secure data management emerges as a crucial requirement. More specifically, storage systems intended for very large scales have to address a series of challenges, such as a scalable architecture, data location transparency or high throughput under concurrent accesses, requirements that come with a major drawback: the complexity of configuring and tuning the system's behavior. Such challenges can be overcome if the system is outfitted with a set of self-management components that enable an autonomic behavior. They heavily relies on introspection mechanisms, which play the crucial role of exposing the system's behavior accurately and in real time.

This PhD research focuses on enhancing a distributed data-management system with self-management capabilities, so that it can meet the requirements of the Cloud storage services in terms of data availability, reliability and security. We focus on the case of BlobSeer, a system designed to store massive data, while leveraging a large-scale deployment and heavy data-access concurrency.

# An Integrated Scratch Management Service for HPC Centers

Henry M. Monti
Department of Computer Science
Virginia Tech.
hmonti@cs.vt.edu

## Abstract

To sustain emerging data-intensive scientific applications, High Performance Computing (HPC) centers invest a notable fraction of their operating budget on a specialized fast storage system, scratch space, which is designed for storing the data of currently running and soon-to-run HPC jobs. Instead, it is often used as a standard file system, wherein users arbitrarily store their data, without any consideration to the center's overall performance. To remedy this, centers periodically scan the scratch in an attempt to purge transient and stale data. This practice of supporting a cache workload using a file system and disjoint tools for staging and purging results in suboptimal use of the scratch space. This work addresses the above issues by proposing a new perspective, where the HPC scratch space is treated as a cache, and data population, retention, and eviction tools are integrated with scratch management. Using this approach, data is moved to the scratch space only when it is needed, and unneeded data is removed as soon as possible.

# Policy Based Data Placement in High Performance Scientific Computing

Muhammad Ali Amer
University of Southern California, Information Sciences Institute
Marina del Rey, CA, USA
amer@isi.edu

## Abstract

The recent increased use of workflow management systems by large scientific collaborations presents the challenge of highly dynamic data placement in distributed systems. Such data placement may be constrained by the enforcement of data placement policies. We present a Policy based Data Placement Service that enforces data placement policies, interfaces with workflow managers and improves workflow efficiency by offloading data staging jobs from workflows.

# Automatic Generation of Executable Communication Specifications from Parallel Applications

Xing Wu
Department of Computer Science
North Carolina State University

Frank Mueller
Department of Computer Science
North Carolina State University

Scott Pakin
Los Alamos National Laboratory

## Abstract

Portable parallel benchmarks are widely used and highly effective for (a) the evaluation, analysis and procurement of high-performance computing (HPC) systems and (b) quantifying the potential benefits of porting applications for new hardware platforms. Yet, past techniques to synthetically parametrized hand-coded HPC benchmarks prove insufficient for today's rapidly-evolving scientific codes particularly when subject to multi-scale science modeling or when utilizing domain-specific libraries.

To address these problems, this work contributes novel methods to automatically generate highly portable and customizable communication benchmarks from HPC applications. We utilize ScalaTrace, a lossless, yet scalable, parallel application tracing framework to collect selected aspects of the run-time behavior of HPC applications. We subsequently generate benchmarks with identical run-time behavior from the collected traces in the CONCEPTUAL language, a domain-specific language that enables the expression of sophisticated communication patterns using a rich and easily understandable grammar yet compiles to ordinary C+MPI. Experimental results demonstrate that the generated benchmarks are able to preserve the run-time behavior of the original applications. This ability to automatically generate performance-accurate benchmarks from parallel applications is novel and without any precedence, to our knowledge.

# Scout: High-Performance Heterogeneous Computing Made Simple

James A. Jablin
Brown University
jjablin@cs.brown.edu

Patrick McCormick
Los Alamos National Laboratory
pat@lanl.gov

Maurice Herlihy
Brown University
mph@cs.brown.edu

## Abstract

Researchers must often write their own simulation and analysis software. During this process they simultaneously confront both computational and scientific problems. Current strategies for aiding the generation of performance-oriented programs do not abstract the software development from the science. Furthermore, the problem is becoming increasingly complex and pressing with the continued development of many-core and heterogeneous (CPU-GPU) architectures. To achieve high performance, scientists must expertly navigate both software and hardware. Co-design between computer scientists and research scientists can alleviate but not solve this problem. The science community requires better tools for developing, optimizing, and future-proofing codes, allowing scientists to focus on their research while still achieving high computational performance.

Scout is a parallel programming language and extensible compiler framework targeting heterogeneous architectures. It provides the abstraction required to buffer scientists from the constantly-shifting details of hardware while still realizing high-performance by encapsulating software and hardware optimization within a compiler framework.

# Building Dynamic Computing Infrastructures over Distributed Clouds

Pierre Riteau
University of Rennes 1, IRISA
INRIA Rennes - Bretagne Atlantique
Rennes, France
Email: Pierre.Riteau@irisa.fr

## Abstract

The emergence of cloud computing infrastructures brings new ways to build and manage computing systems, with the flexibility offered by virtualization technologies. In this context, this PhD thesis focuses on two principal objectives. First, leveraging virtualization and cloud computing infrastructures to build distributed large scale computing platforms from multiple cloud providers, allowing to run software requiring large amounts of computation power. Second, developing mechanisms to make these infrastructures more dynamic. These mechanisms, providing inter-cloud live migration, offer new ways to exploit the inherent dynamic nature of distributed clouds.

# Error Correction and Clustering Algorithms for Next Generation Sequencing

Xiao Yang
Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA
xyang@iastate.edu

## Abstract

Next generation sequencing (NGS) revolutionized genomic data generation by enabling high-throughput parallel sequencing, making large scale genomic data analysis a crucial task. To improve NGS data quality, we developed an efficient algorithm that uses a flexible read decomposition method to improve accuracy of error correction. We further proposed a statistical framework to differentiate infrequently observed subreads from sequencing errors in the prevalence of genomic repeats. To enable the analysis of microbial organism composition in environmental samples, we developed a parallel solution for metagenomic sequence clustering integrating sketching, quasi-clique enumeration and MapReduce techniques.

# Performance Analysis of Long-running Applications

Zoltán Szebenyi[*‡] , Felix Wolf[*†‡] , Brian J. N. Wylie[*]
[*] Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany
[†] German Research School for Simulation Sciences, 52062 Aachen, Germany
[‡] RWTH Aachen University, 52056 Aachen, Germany
{z.szebenyi, f.wolf, b.wylie}@fz-juelich.de

## Abstract

With the growing complexity of supercomputing applications and systems, it is important to constantly develop existing performance measurement and analysis tools to provide new insights into application performance characteristics and thereby help scientists and engineers utilize computing resources more efficiently. We present the various new techniques developed, implemented and integrated into the Scalasca toolset specifically to enhance performance analysis of long-running applications. The first is a hybrid measurement system seamlessly integrating sampled and event-based measurements capable of low-overhead, highly detailed measurements and therefore particularly convenient for initial performance analyses. Then we apply iteration profiling to scientific codes, and present an algorithm for reducing the memory and space requirements of the collected data using iteration profile clustering. Finally, we evaluate the complete integration of all these techniques in a unified measurement system.

## p2MATLAB: Productive Parallel MATLAB for the Exascale

Vipin Sachdeva
College of Computing
Georgia Institute of Technology
Atlanta, GA
vipin@gatech.edu

### Abstract

MATLAB®and its open-source implementation Octave have proven to be one of the most productive environments for scientific computing in recent years. There have been multiple efforts to develop an efficient parallel implementation of MATLAB including by Mathworks®(Parallel Computing Toolbox), MIT Lincoln Labs (pMatlab) and several other organizations. However, most of these implementations seem to suffer from issues in performance or productivity or both. With the rapid scaling of high-end systems to hundreds of thousands of cores, and discussions of exascale systems in the near future, a scalable parallel Matlab would be of immense benefit to practitioners in the scientific computing industry. In this paper, we first describe our work to create an efficient pMatlab running on the IBM BlueGene/P architecture, and present our experiments with several important kernels used in scientific computing including from HPC Challenge Awards. We explain the bottlenecks with the current pMatlab implementation on BlueGene/P architecture, specially at high processor counts and then outline the steps required to develop a parallel MATLAB/Octave implementation, p$^2$Matlab, which is truly scalable to hundreds of thousands of processors.

## A Framework for Auto-tuning and Automatic Verification of Numerical Solver Codes on GPU Platforms

Allison S. Gehrke, Ilkyeun Ra, Daniel A. Connors
Dept. of Computer Science and Engineering and Dept. of Electrical Engineering
University of Colorado Denver
Denver, CO 80204 U.S.A
{Allison.Gehrke, Ilkyeun.Ra, Dan.Connors}@ucdenver.edu

### Abstract

Emerging multi-core processor designs create a computing paradigm capable of advancing numerous scientific areas, including medicine, data mining, biology, physics, and earth sciences. However, the trends in multi-core hardware technology have advanced far ahead of the advances in software technology and programmer productivity. For the most part, current scientists only leverage multi-core and GPU (Graphical Processing Unit) computing platforms after painstakingly uncovering the inherent task and data-level parallelism in their application. In many cases, the development does not realize the full potential of the parallel hardware.

There exists an opportunity to meet the challenges in optimally mapping scientific application domains to multi-core computer systems through the use of compile-time and link-time optimization strategies. We are exploring a code compilation framework that automatically generates and tunes numerical solver codes for optimal performance on graphical processing units. The framework advances computational simulation in kinetic modeling by significantly reducing the execution time of scientific simulations and enabling scientists to compare results to previous models and to extend, modify, and test new models without code changes.