# Joint-lab workshop Nov. 19-21 2012

### The workshop will take place at Argonne National Laboratory.

### This event is supported by INRIA, ANL, UIUC and NCSA, French Ministry of Foreign Affairs as well as by EDF

## Schedule under construction

| Main Topics | Schedule | Speakers | Types of presentation | Topic | Download |
|---|---|---|---|---|---|
| | **Sunday 18th:**<br><br>**Bus to restaurant** | | **From Argonne Guest House** | The Bus will leave at **6:30PM**. Since a bus is available, no taxi will be covered. | |
| | Sunday Nov. 18th 19:00 | Dinner | Giordano's<br>641 PLAINFIELD RD<br>WILLOWBROOK, IL 60521<br>(630) 325-6710 | http://www.giordanos.com/<br>http://maps.google.com/maps?f=q&hl=en&q=641%20PLAINFIELD%20RD.,+WILLOWBROOK,+IL+60527+US&ie=UTF8&z=15&om=1&iwloc=A | |
| **Workshop Day 1** (Room 1416, TCS conference center) | **Monday Nov. 19th** | | | | |
| | 07:30-8:30 | Transportation: Guest House to TCS (building 240) | | (Entrance of the conference center) | |
| | 08:00 | Contiental Breakfast and Registration | | *Food available in Room 1416, Lunch seating in room 1416 (second half)* | |
| **Welcome and Introduction** | 08:30 | Franck Cappello, INRIA & UIUC, Marc Snir ANL | Opening | Welcome, formal opening and workshop details | Opening-8th-Workhsop.pdf |
| | 08:40 | Marc Snir | Opening | ANL presentation and vision of the collaboration | snir JLPC 11-19-12.pdf |
| | 08:50 | Bill Gropp | Opening | UIUC/NCSA update and vision of the collaboation | Overview-gropp.pdf |
| | 09:00 | Frederic Desprez | Opening | INRIA update on HPC strategy and vision of the collaboration | |
| **Big Apps, Big DATA - Big I/O**<br>chair: Rajeev Thakur | 09:15 | Robert Jacob | Trends in HPC | **Climate modeling and challenges of exascale** | climateINRIA.pdf |
| | 09:45 | Rob Ross, ANL | Trends in HPC | **Trends in HPC I/O and File systems** | ross_hpc-io-trends-2012-11.pdf |
| | 10:15 | Break | | | |
| | 10:45 | Rob Pennington, NCSA | Trends in HPC | **Scientific Data Generators – Plague or a Panacea?** | Pennington INRIA Illinois workshop Nov 19 2012.pdf |
| | 11:15 | Andrew Chien, ANL | Potential collaboration | **Presto/Blockus: Towards a Scalable R Programming System** | Presto-Blockus-11-19-2012-Final.pdf |
| | 11:45 | Matthieu Dorier, INRIA | **Joint Results** | **I/O and in-situ visualization: recent results with the Damaris approach** | DORIER-JLPC-November2012.pdf |
| | 12:15 | Lunch | | | |
| **Programming Models /Runtime** chair: Sanjay Kale | 13:30 | Wen-Mei Hwu, UIUC | Potential collaboration | **Scalability, Performance, and Numerical Stability of Many-core GPU Algorithms - A Case Study of Tri-diagonal Solvers** | Blue-Waters-joint-lab-scalability-11-19-2012-v2.pdf |
| | 14:00 | Pavan Balaji, ANL | Potential collaboration | **MPI-ACC: A Unified Data Movement Infrastructure for MPI and Accelerators** | 2012-11-19-inria-mpiacc.pdf |
| | 14:30 | Christian Klein, INRIA | Potential collaboration | **Cooperative Resource Management for Parallel and Distributed Systems** | CristianKleinCooRM.pdf |
| | 15:00 | E. Francesquini, INRIA | Potential collaboration | **The Actor Model and Multi-core Architectures** | Francesquini_19Nov.pdf |
| | 15:30 | Break | | | |
| **Numerical algorithms and Methods**<br>Chair: Paul Hovland | 16:00 | Stefan Wild, ANL | Potential collaboration | **Numerical optimization for "automatic" tuning of codes** | wild.pdf |
| | 16:30 | Laura Grigori | Joint Results | **Iterative methods, preconditioning, and their application to CMB data analysis** | LGrigori_JLPC_Nov2012.pdf |
| | 17:00 | Laurent Hascoet, INRIA | Early Results | **The Data-Dependence graph of Adjoint Codes** | slidesAdjDDArgonne.pdf |
| | 17:30 | Adjourn | | | |
| | 17:30-18:30 | Transportation: TCS (building 240) to Guest House | | | |
| | 18:30 | Transportation to restaurant: **B US leaves guest house at 6:30PM** | | | |
| | 19:00 | Dinner | Jameson's<br>Woodridge 1001 W. 75th Street Woodridge, IL 60517<br>630.910.9700 | http://www.jamesons-charhouse.com/index.html<br>MAP | |
| | | | | | |
| **Workshop Day 2** (Main room) | **Tuesday Nov. 20th** | | | | |
| | 07:30-8:30 | Transportation: Guest House to TCS (building 240) | | (Entrance of the conference center) | |
| | 08:00 | Contiental Breakfast and Registration | | *Food available in Room 1416, Lunch seating in room 1416 (second half)* | |

| | | | | | |
|---|---|---|---|---|---|
| **Big Systems**<br>Chair: Jean François Mehaut | 08:30 | Pete Beckman, ANL | Trends | **New Directions in Extreme-Scale Operating Systems and Runtime Software** | Beckman-INRIA-NCSA-Thinking.pdf |
| | 09:00 | Bill Kramer, UIUC/NCSA | Trends | **Is There a Life After the Top500? (Or What to Do About the Top Problems with the TOP500 List)** | INRIA-NCSA-ANL-112012a-sent.pdf |
| **Cloud**<br>Chair: Gabriel Antoniu | 09:30 | Ian Foster, ANL | Potential collaboration | **Big Process for Big Data** | Big Process Big Data Foster November 2012.pdf |
| | 10:00 | Christine Morin, INRIA | Potential collaboration | **Work in Progress on Cloud Computing in Myriads Team and Contrail European Project** | 2012-11-20-Contrail-Myriads.pdf |
| | 10:30 | Break | | | |
| | 11:00 | Frederic Desprez, INRIA | Potential collaboration | **Workflow Allocations and Scheduling on IaaS Platforms, from Theory to Practice** | desprez-workflow.pdf |
| **Resilience:**<br>Chair: Christine Morin | 11:30 | Yves Robert | Early Result | **Performance modeling of checkpointing under failure prediction** | yrobert-ckpt-prediction.pdf |
| | 12:00 | Rinku Gupta, ANL | Potential collaboration | **CIFTS: An infrastructure for coordinated and comprehensive system-wide fault tolerance.** | anl-inria-cifts.pdf |
| | 12:30 | Ana Gainaru, UIUC | Early Results | **Coupling failure prediction, proactive and preventive checkpoint for current production HPC systems.** | againaru.pdf |
| | 13:00 | Lunch | | *Food buffet in Room 1407, Lunch seating in room 1416 (second half)* | |
| | | | | **Parallel Session** | |
| **Mini workshop on Numerical libraries**<br>Chair: Paul Hovland (room 1406, TCS conference center) | 8:30 | Tim Tautges | Potential collaboration | **Mesh-based Data and Algorithms across the Simulation Process: anecdotes, activities, and opportunities** | meshdata.pdf |
| | 09:00 | Bill Gropp, UIUC | Potential collaboration | TBA | |
| | 09:30 | Laura Grigori, INRIA | Potential collaboration | TBA | |
| | 10:00 | Break | | | |
| | 10:30 | Anshu Dubey, ANL | Potential collaboration | **Optimizing Scientific Codes While Retaining Portability** | Anshu_JS.pdf |
| | 11:00 | Discussion | | | |
| | 12:00 | Adjourn | | | |
| | 13:00 | Lunch | | | |
| | | | | **Parallel Sessions** | |
| **Mini workshop on Performance Modeling and simulation**<br>Chair: Marc Snir | 14:30 | Sanjay Kale, UIUC | Potential collaboration | **A perspective on the BigSim approach to performance prediction** | |
| | 15:00 | Arnaud Legrand, INRIA | Potential collaboration | **SimGrid for HPC** | legrand_Argonne.pdf |
| | 15:30 | Torsten Hoefler, ETH | Potential collaboration | **Performance Modeling for Parallel Software Development and Tuning** | hoefler-perfmodeling.pdf |
| | 16:00 | Break | | | |
| | 16:30 | Laercio Pila | Potential collaboration | **A Performance Measurement Approach for Modeling Latency and Bandwidth of Large Scale Multicore Machines** | llpilla-joint-lab-2012-2.pdf |
| | 17:00 | Discussion | | | |
| | 18:00 | Adjourn | | | |
| | 18:00-18:30 | Transportation: TCS (building 240) to Guest House | | | |
| | 19:00 | Dinner | Meggaiano's<br>240 Oakbrook Center Oak Brook, IL 60523 | [http://www.maggianos.com/EN/Oak-Brook_Oak-Brook_IL/Pages/LocationLanding.aspx?AspxAutoDetectCookieSupport=1 ] MAP | |
| | | | | | |
| **Mini workshop on Cloud**<br>Chair: Kate Keahey | 14:30 | Kate Keahey, ANL | Potential collaboration | **Infrastructure Outsourcing in Multi-Cloud Environment** | jointlab-anl.pdf |
| | 15:00 | Narayan Deai, ANL | Potential collaboration | **Building Clouds for Technical Computing** | magellan - joint ws 11-12.pdf |
| | 15:30 | Jonathan Rouzaud, INRIA | Potential collaboration | **Provisioning Virtual Machines in Federated Clouds** | cloud_allocation.pdf |
| | 16:00 | Break | | | |
| | 16:30 | Michael Wilde | Potential collaboration | **Swift: simpler parallel programming for cloud and HPC domains**<br>http://www.ci.uchicago.edu/swift (Swift for clouds and clusters)<br>http://www.mcs.anl.gov/exm (Swift for extreme-scale domains) | SwiftforCloudandHPC.Wilde.JointLab.2012.1120.pdf |
| | 17:00 | Francieli Zanon Boito | Potential collaboration | **Application Aware I/O Scheduling in the Parallel File System Server Side** | francieli_8jlpc_workshop.pdf |
| | 17:30 | Discussion | | | |
| | 18:00 | Adjourn | | | |
| | 18:00-18:30 | Transportation: TCS (building 240) to Guest House | | | |
| | 18:30 | Transportation to restaurant: **BUS leaves guest house at 6:30PM** | | | |
| | 19:00 | Dinner | Maggiano's<br>240 Oakbrook Center Oak Brook, IL 60523 | [http://www.maggianos.com/EN/Oak-Brook_Oak-Brook_IL/Pages/LocationLanding.aspx?AspxAutoDetectCookieSupport=1 ] MAP | |
| | | | | | |
| **Workshop Day 3** (Main room) | **Wednesday Nov 21st** | | | | |
| | 07:30-8:30 | Transportation: Guest House to TCS (building 240) | | (Entrance of the conference center) | |
| | 08:00 | Contiental Breakfast and Registration | | *Food available in Room 1416, Lunch seating in room 1416 (second half)* | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | **Parallel Sessions** | |
| **Mini workshop on Programming models /runtime**<br>Chair: Pavan Balaji | 08:30 | Emmanuel Jeannot, INRIA | Results | **Process placement with unbalanced architecture** | Jeannot-Argonne-2012.pdf |
| | 09:00 | Sanjay Kale, UIUC | Trend | Charm++ update | 2012_11_INRIA_Argonne_CharmPDF.pdf |
| | 09:30 | Andra Hugo, Raymond Namyst, INRIA | Potential collaboration | **Composing multiple StarPU applications over heterogeneous machines: a supervised approach** | Composability_StarPU.pdf |
| | 10:00 | Break | | | |
| | 10:30 | Jim Dinan | Potential collaboration | **A One-Sided View of HPC: Global-View Models and Portable Runtime Systems** | dinan_aiu12_slides.pdf |
| | 11:00 | Sebastien Fourestier | Potential collaboration | **Parallel repartitioning and re-mapping in Scotch** | Fourestier_Pellegrini_-_Parallel_repartitioning_and_remapping_in_Scotch.pdf |
| | 11:30 | Timo Schneider, ETH | Potential collaboration | **Optimization Principles for Collective Neighborhood Communications** | schneider-neighbor-colls.pdf |
| | 12:00 | Discussion | | | |
| | 12:30 | Closing | | | |
| | 13:00 | Lunch | | | |
| | 13:15-14:15 | Transportation: TCS (building 240) to Guest House | | | |
| | | | | | |
| **Mini workshop on Resilience**<br>Chair: Franck Cappello | 08:30 | Mohamed Slim Bouguerra | Result | TBA | inria_uiuc_11_2012_workshop_new.pdf |
| | 09:00 | Amina Guermouche, INRIA | Result | **Unified Model for Assessing Checkpointing Protocols at Extreme-Scale** | 2012-11-21_Jointlab.pdf |
| | 09:30 | Bogdan Nicolae, IBM | Result | **I-Ckpt: Leveraging memory access patterns and inline collective deduplication to improve scalability of CR** | 8th_jlpc.pdf |
| | 10:00 | Break | | | |
| | 10:30 | Tatiana Martsinkevich, INRIA | Result | **Fully distributed recovery for send-determinism applications** | distributed recovery.pdf |
| | 11:00 | Peter Brune, ANL | Trends | **Multilevel Resiliency for PDE Simulations** | resilientfas.pdf |
| | 11:30 | Xiang Ni, Estaban Menese | Results | **Scalable in-memory checkpoint with automatic restart on failure** | inria_xiang_ft.pdf |
| | 12:00 | Discussion | | | |
| | 12:30 | Closing | | | |
| | 13:00 | Lunch | | Box Lunches | |
| | 13:15-14:15 | Transportation: TCS (building 240) to Guest House | | | |

## Abstracts

### Robert Jacob

Climate modeling and challenges of exascale

Climate models have been called one of the most complex computer applications ever developed. The combination of geophysical fluid dynamics and the many important non-fluid phenomena in climate is the main source of this complexity. The current construction of climate models will be briefly reviewed and the challenges exascale poses for business-as-usual climate modeling will be examined.

### Robert Pennington

Scientific Data Generators – Plague or a Panacea?

Large scientific instruments are typically also large data generators and HPC systems are no exception. The resulting data from scientific instruments are the basis for new discoveries as well as for comparison to other works. The potential for these two aspects to become more effective is getting an increased level of attention by researchers and by research sponsors. Opportunities to extend the availability of datasets generated within HPC environments are being taken up by a number of often highly collaborative domains and the results have significant effect and long term impact on the respective domains. The resources available, particularly with large scale HPC systems, provide an environment for generating or extending unique opportunities for science that is data driven. The question is how to accelerate this process beyond a limited number of domains.

### Wen-mei Hwu, University of Illinois

Scalability, Performance, and Numerical Stability of Many-core GPU Algorithms - A Case Study of Tri-diagonal Solvers

The IMPACT group at the University of Illinois has been working on the co-design of scalable algorithmsand programming tools for massively threaded computing devices. A major challenge that we areaddressing is to simultaneously achieve scalability, performance, and numerical stability for tri-diagonalsolvers. In this talk, I will go over the major building blocks involved: memory layout and dynamictiling. I will show experimental results to demonstrate how these building blocks jointly enable the firstscalable, numerically stable tri-diagonal solver that matches the numerical stability of MKL and surpassthe performance of CUSPARSE.

### Robert Ross, ANL

Trends in HPC I/O and File systems

All aspects of HPC systems are undergoing change as we move into petascale and towards exascale computing. The traditional "I/O software stack" is no exception: the layers, capabilities, and abstractions in the stack are all in flux as we consider how to best support future HPC applications. This talk will discuss these developmental trends, using ongoing work at Argonne as examples of some directions of study.

### Ian Foster

Big Process for Big Data

Large and diverse data result in challenging data management problems that researchers and facilities are often ill-equipped to handle. I propose a new approach to these problems based on the outsourcing of research data management tasks to software-as-a-service providers. I argue that this approach can both achieve significant economies of scale and accelerate discovery by allowing researchers to focus on research rather than mundane information technology tasks. I present early results with the approach in the context of Globus Online.

### Andrew Chien

Presto/Blockus: Towards a Scalable R Programming System

We are studying simple extensions of the R programming system to allow R programmers to have simple, scalable access to multi-core and cluster scale-out parallelism, enabling access to larger memories and higher computation speeds. Subsequent objectives include scale-vertical to secondary storage, which promises computing over "Big Data" in modest size systems. This effort is joint with HP and several other institutions.

### Andra Hugo, INRIA

Composing multiple StarPU applications over heterogeneous machines: a supervised approach

Enabling HPC applications to perform efficiently when invoking multiple parallel libraries simultaneously is a great challenge. Even if a single runtime system is used underneath, scheduling tasks or threads coming from different libraries over the same set of hardware resources introduces many issues, such as resource oversubscription, undesirable cache flushes or memory bus contention.
In this talk, I will present an extension to the StarPU runtime system that enables multiple StarPU kernels to simultaneously run over the same CPU+GPU architecture. Further on, I will present some experimental results showing the improvements our solution brings to the efficiency of parallel applications composing several parallel libraries (e.g.: libraries in the domain of dense linear algebra or fluid mechanics). Eventually, I will give some insights about the main challenges of the composability problem and I will present the main topics we are interested in for the future work.

### Pete Beckman, ANL

New Directions in Extreme-Scale Operating Systems and Runtime Software

For more than a decade, extreme-scale operating systems and runtime software have been evolving very slowly. Today's large-scale systems use slightly retooled "node" operating systems glued together with ad hoc local agents to handle I/O, job launch, and management. These extreme-scale systems are only slightly more tightly integrated than are generic Linux clusters with InfiniBand. As we look forward to a new era for large-scale HPC systems, we see that power and fault management will become key design issues. Software management of power and support for resilience must now be part of the whole-system design. Extreme-scale operating systems and runtime software will not be simply today's node code with a few control interfaces, but rather a tightly integrated "global OS" that spans the entire platform and works cooperatively across portions of the machine in order to manage power and provide resilience.

### Sebastien Fourestier, INRIA

Parallel repartitioning and re-mapping in Scotch

Scotch is a software package for sequential and parallel graph partitioning, static mapping, sparse matrix block ordering, clustering and sequential mesh and hypergraph ordering. As a research project, it is subject to continuous improvement, resulting from several on-going research tasks. Our talk will address several new features we have recently added to Scotch. We will present some threaded algorithms for shared-memory coarsening and refinement. We will also show early results regarding its parallel repartitioning and sequential remapping functionalities.

### Anshu Dubey, ANL

Optimizing Scientific Codes While Retaining Portability

Optimization of large scientific codes for production is a balancing act between portability and performance. In face of future hardware architecture challenges, retaining portability while obtaining acceptable performance is expected to more challenging than ever. The first part of my presentation will be about experiences with pragmatic optimizations of FLASH, a multiphysics simulation code with a wide user base. The second part will discuss ideas for addressing the future challenges.

### Michael Wilde, ANL

Swift: simpler parallel programming for cloud and HPC domains

### Ana Gainaru, UIUC

Coupling failure prediction, proactive and preventive checkpoint for current production HPC systems.

A large percentage of computing capacity in today's large high-performance computing systems is wasted due to failures and recoveries. A way of reducing the overhead induced by these strategies is by combining them with failure avoidance methods. Failure avoidance is based on a prediction model that detects fault occurrences ahead of time and allows preventive measures to be taken, such as task migration or checkpointing the application. This talk presents the implementation and results of a prototype implementation of proactive checkpointing based on the ELSA toolkit coupled with periodic multi-level checkpointing based on FTI. The proactive checkpointing is implemented as a level zero (L0) in a four-level scheme, providing the fastest checkpoint, which is necessary to act quickly between the failure prediction and the moment of the failure. We evaluate the proposed approach on the TSUBAME system and we show that the overhead in comparison with a preventive checkpoint execution only represents only 2% to 6%.

### Peter Brune

Multilevel Resiliency for PDE Simulations

Co-Authors: Mark Adams, Jed Brown, Peter Brune (speaking), Barry Smith

Multilevel methods for the solution of partial differential equations are the de-facto fast algorithms for large-scale computations. The utilization of these method necessitates progressively smaller approximations of the solution to the problem, potentially on a smaller subset of the machine. These algorithms present a tempting target for enabling efficient extreme-scale resiliency, as the multilevel structure may be used to efficiently compress the PDE solution and check for algorithmic correctness. We discuss the components of multilevel methods and their use for resilient computation. We speculate on possibilities for the integration of these methods into simulations.

### Stefan Wild

Numerical optimization for "automatic" tuning of codes

Heterogeneity and rapid evolution of modern architectures increasingly demand that scientific codes be tuned in order to achieve high performance on different machines. Empirical performance tuning seeks high-performing code variants based on their measured performance on a target machine, but several obstacles remain in making this procedure "automatic." In this talk we provide an overview of the search problem in performance tuning, as formulated through a derivative-free, mixed-integer optimization problem. We explore modeling formulations for the problem, local and global algorithms, and potential trade-offs between competing objectives such as run time and energy consumption.

### Rinku Gupta

CIFTS: An infrastructure for coordinated and comprehensive system-wide fault tolerance.

The need for leadership class fault-tolerance continues to increase as emerging high performance systems move towards offering exascale level performance. While most high-end systems do provide mechanisms for detection, notification and perhaps handling of hardware and software related faults, the individual components present in the system perform these actions separately. Knowledge about occurring faults is seldom shared between different software and almost never on a system-wide basis. A typical system contains numerous software that could benefit from such knowledge, include applications, middleware libraries, job schedulers, file systems, math libraries, monitoring software, operating systems, and check pointing software.

The Coordinated Infrastructure for Fault Tolerant Systems (CIFTS) initiative provides the foundation necessary to enable systems to adapt to faults in a holistic manner. CIFTS achieves this through the Fault Tolerance Backplane (FTB), providing a unified management and communication framework, which can be used by any system software to publish fault-related information. In this talk, I will present some of the work done by the CIFTS group towards the development of FTB and FTB-enabled components; and discuss the potential and challenges of such system-wide inter-layer fault tolerance frameworks.

### Bogdan Nicolae

I-Ckpt: Leveraging memory access patterns and inline collective deduplication to improve scalability of CR

With increasing scale and complexity of supercomputing and cloud computing architectures, faults are becoming a frequent occurrence. For a large class of applications that run for a long time and are tightly coupled, Checkpoint-Restart (CR) is the only feasible method to survive failures. However, exploding checkpoint sizes that need to be dumped to storage pose a major scalability challenge. To tackle with this challenge, this talk focuses on two techniques: (1) leveraging knowledge of memory access patterns to minimize overhead of asynchronous checkpointing; (2) an inline collective memory contents deduplication scheme that attempts to identify and eliminate duplicate memory pages across all processes before they are saved to storage. Several extensions and future work directions are also discussed.

### Jonathan Rouzaud-Cornabas

Provisioning Virtual Machines in Federated Clouds

With the increasing number of Cloud offers and their heterogeneity , it becomes harder and harder for the Cloud Users to select the proper cloud(s) and resources.
Moreover, the selection process is strongly related to the application itself and the users' requirements (deadline, cost, etc.).
In this talk, we will present our early work on selecting and provisioning Virtual Machines in Federated Clouds. Our current work focuses on running Bag Of Tasks.
We will show our Cloud Broker Simulator based on SimGrid and how it can be used to help selecting resources for a given application based on a set of requirements.
Finally, we will conclude on presenting the future challenges such as taking into account a large set of scientific computing applications such as workflows.
Rouzaud-Cornabas Jonathan

### Laurent Hascoet

The Data-Dependence graph of Adjoint Codes

Automatic Differentiation (AD) is the primary means of obtaining analytic derivatives from a numerical model given as a computer program. Therefore, it is an essential productivity tool in numerous computational science and engineering domains. Computing gradients with the adjoint mode of AD via source transformation is a particularly beneficial but also challenging use of AD. From another viewpoint, Data-Dependence Graphs are one of the key tools to study and improve the performance of programs, particularly in view of their parallel execution. Basic parallelizability properties are classically expressed as properties of the Data-Dependence Graph of a code. We explore the relation between the Data-Dependence graphs of a program and of its adjoint, thus explaining why many parallel properties of a code also apply to its adjoint.

### Jim Dinan

A One-Sided View of HPC: Global-View Models and Portable Runtime Systems

Global-view and one-sided parallel programming models provide a promising alternative to conventional approaches by enabling programmers to aggregate the memory of multiple nodes and allowing them to access any data, regardless of its physical location. This model for asynchronous data movement also decouples synchronization from communication, enabling a greater degree of asynchrony. These properties are of critical importance to scientific computing applications, which must cope with rapidly evolving system architectures, and where new simulation and analysis techniques have exposed greater sparsity and computational imbalance.

In this talk, I will present recent and ongoing work on portable one-sided communication interfaces and global-view parallel programming systems. This work focuses on the evolution of the MPI-2 remote memory access (RMA) communication interface into the new MPI-3 RMA interface, and on the utilization of these interfaces to support higher-level parallel programming interfaces. I will describe work, in which we have used the MPI RMA interface to provide the first portable, one-sided implementation of Global Arrays and its impact on the NWChem computational chemistry suite. In addition, I will describe current and ongoing work in the deployment, implementation, and performance tuning of MPI-3 RMA.

### Arnault Legrand

SimGrid for HPC

In this talk, I will briefly present the history and goals of the SimGrid simulation toolkit. Although SimGrid was primarily designed in 1999 to perform scheduling studies on heterogeneous systems such as Grid, recent developments have made it a very effective alternative for conducting simulation studies for P2P and HPC compared to many ad hoc (but often short lived) simulators. I will thus present the current status of research and developments in SimGrid as well as the future directions we intend to address.

### Matthieu Dorier

I/O and in-situ visualization: recent results with the Damaris approach

As dumping large amounts of data to parallel file systems starts to highly impact the performance of HPC simulations as well as the practicability of subsequent analysis tasks, new approaches to I/O and data analysis must be found. Damaris proposes to relocate I/O and analysis tasks in dedicated cores interacting with the simulation through shared memory.
In this talk, we will provide a quick recall on the Damaris approach to scalable, efficient, jitter-free I/O, along with past results. We will then move to more recent works and results using Damaris for in-situ visualization with the CM1 atmospheric model and the Nek5000 CFD simulation. This presentation will include a demo of Damaris providing in-situ visualization to a sample simulation through the VisIt visualization software.

### Amina Guermouche

Unified Model for Assessing Checkpointing Protocols at Extreme-Scale

In this talk, we present a unified model for several well-known checkpoint/restart protocols. The proposed model is generic enough to encompass both extremes of the checkpoint/restart space, from coordinated approaches to a variety of uncoordinated checkpoint strategies (with message logging). We identify a set of crucial pa- rameters, instantiate them and compare the expected efficiency of the fault tolerant protocols, for a given application/platform pair. We then propose a detailed analysis of several scenarios, including some of the most powerful currently available HPC platforms, as well as anticipated Exascale designs. The results of this analytical comparison are corroborated by a comprehensive set of simulations. Altogether, they outline comparative behaviors of checkpoint strategies at very large scale, thereby providing insight that is hardly accessible to direct experimentation.

### Yves Robert

Performance modeling of checkpointing under failure prediction

This talk deals with the impact of fault prediction techniques on checkpointing strategies. We extend the classical analysis of Young and Daly in the presence of a fault prediction system, which is characterized by its recall and its precision, and which provides either exact or window-based time predictions. We succeed in deriving the optimal value of the checkpointing period (thereby minimizing the waste of resource usage due to checkpoint overhead) in all scenarios. These results allow to analytically assess the key parameters that impact the performance of fault predictors at very large scale. In addition, the results of this analytical evaluation are nicely corroborated by a comprehensive set of simulations, thereby demonstrating the validity of the model and the accuracy of the results.

### Torsten Heofler

Performance Modeling for Parallel Software Development and Tuning

Scientific applications are commonly developed and used over the lifecycle of multiple parallel computing architectures. Despite all efforts to develop performance-portable parallel programming environments, several changes are often necessary to adapt the code to new architectures and systems. Performance modeling has been discussed as a viable tool to support all stages of the software development process of parallel applications and to support co-design of different layers. In this talk, we particularly focus on the last, and most expensive stage, the continuous porting and improvement phase. We show how to apply semi-analytic modeling techniques to understand the structure of large parallel applications and to pinpoint bottlenecks and viable targets for code improvements. To do this, we combine techniques for optimizing serial and parallel codes and we demonstrate several real-world application examples. We expect that our methodology can help to improve the effectivity of the software development process for parallel computing.

### Timo Schneider

Optimization Principles for Collective Neighborhood Communications

Abstract: Many scientific applications work in a bulk-synchronous mode of iterative communication and computation steps. Even though the communication steps happen at the same time, important patterns such as stencil computations cannot be expressed as collective communications in MPI. Neighborhood collective operations allow to specify arbitrary collective communication relations during runtime and enable optimizations similar to traditional collective calls. We show a number of optimization opportunities and algorithms for different communication scenarios. We also show how users can assert additional constraints that provide new optimization opportunities in a portable way. Our communication and protocol optimizations result in a performance i mprovement of up to a factor of two for stencil communications. We found that our optimization heuristics can automatically generate communication schedules that are comparable to hand-tuned collectives. With those optimizations, we are able to accelerate arbitrary collective communication patterns, such as regular and irregular stencils with optimization methods for collective communications

### **Narayan Desai**

Building Clouds for Technical Computing

Abstract: Virtualization is not an obvious match for technical computing workloads, particularly due to performance overhead and inefficiency. In the Magellan project, we have built a moderate scale private cloud, assessed the impact of running applications in this environment, and begun using the system for workloads well-suited to it. In this talk, I'll describe our current state, including application sweet spots, several performance engineering goals we have hit, as well as our near term plans for system improvements.

### **Tatiana V. Martsinkevich**

Distributed recovery for SPMD-deterministic applications

Abstract: Hybrid checkpointing and message logging protocols exhibit good scalability and potentially can be employed in future exa-scale systems. Having on hands the data provided by such a protocol it is still an open problem how to recover fast from a fault and moreover, how to guarantee that with the increase of the size of a system, recovery will not become a bottleneck. We discuss about our proposed distributed recovery protocol, it's implementation and show some preliminary results that we have got.

### Christine Morin

Work in Progress on Cloud Computing in Myriads Team and Contrail European Project

Myriads research team at Inria Rennes – Bretagne Atlantique carries out several research projects in the area of cloud computing with a particular interest in autonomous resource and elastic application management. In our talk, we provide an overview of recent results and on-going research activities on the design and implementation of IaaS and PaaS services. In particular, we focus on the work in progress in the framework of the Contrail European Project we coordinate. Contrail project aims at developing an open source cloud computing software stack for the execution of elastic services on top of federated IaaS clouds and for SLA management in cloud federations.

### Kate Keahey

Infrastructure Outsourcing in Multi-Cloud Environment

Infrastructure clouds created ideal conditions for users to outsource their infrastructure needs by offering on-demand, short-term access, pay-as-you-go business model, the use of virtualization technologies which provide a safe and cost-effective way for users to manage and customize their environments, and sheer convenience, as users and institutions no longer have to have specialized IT departments and can focus on their core mission instead. These key innovations however also bring challenges which include high levels of failure; lack of interoperability between cloud providers, and lack of tools that allow users to leverage the on-demand growing and shrinking of infrastructure. All these factors prevent users from capitalizing on the infrastructure cloud opportunity.

In this talk, I will describe a multi-cloud auto-scaling service that enables the user to leverage "computational power on tap" provided by infrastructure clouds, i.e., allows the user to easily deploy resources across multiple private, community, and commercial clouds; provides high availability in that it allows users to replace failed resources; and scales to demand. The policies governing scaling are customizable based on system and application-specific indicators. We will describe the service architecture and implementation and discuss results obtained in the sustained deployment and management of thousands of virtual machines on EC2.

### Frederic Desprez

Workflow Allocations and Scheduling on IaaS Platforms, from Theory to Practice

Many scientific applications are described through workflow structures. Due to the increasing level of parallelism offered by modern computing infrastructures, workflow applications now have to be composed not only of sequential programs, but also of parallel ones. Cloud platforms bring on-demand resource provisioning and pay-as-you-go billing model. Then the execution of a workflow corresponds to a certain budget. The current work addresses the problem of resource al- location for non-deterministic workflows under budget constraints. We present a way of transforming the initial problem into sub-problems that have been studied before. We propose two new allocation algorithms that are capable of determining resource allocations under budget constraints and we present ways of using them to address the problem at hand. Then we present a first implementation of a workflow management system based on DIET/Phantom/Nimbus of the FutureGrid platform for the Ramses workflow, a n-body simulations of dark matter interactions application.

### Cristian Klein*, Christian Perez, INRIA

Cooperative Resource Management for Parallel and Distributed Systems

HPC resources, such as Supercomputers, Clusters, Grids and HPC Clouds, are managed by RMS that multiplex resources among multiple users and decide how computing nodes are allocated to user applications. Optimizing resource allocation to applications is critical to ensure their efficient execution. However, current RMS, such as batch schedulers, only offer a limited interface. In most cases, the application has to blindly choose resources at submittal without being able to adapt its choice to the state of the target resources, neither before it started nor during execution.

The goal of this work is to improve resource management, so as to allow applications to efficiently allocate resources. We achieve this by proposing software architectures that promote collaboration between the applications and the RMS, thus, allowing applications to negotiate the resources they run on. The first contribution deals with moldable applications, for which resources are only negotiated before they start. We propose CooRMv1, a centralized RMS architecture, which delegates resource selection to the application launchers. Simulations show that the solution is both scalable and fair. The results are validated through a prototype implementation deployed on Grid'5000.

The second contribution deals with run-time negotiation of resources, so as to improve support for malleable and evolving applications. We propose CooRMv2, a centralized RMS architecture, that enables efficient scheduling of evolving applications, especially non-predictable ones. It allows applications to inform the RMS about their maximum expected resource usage, through pre-allocations. Resources which are pre-allocated but unused can be filled by malleable applications. Simulation results show that considerable gains can be achieved.

### Laura Grigori

Iterative methods, preconditioning, and their application to CMB data analysis

In this talk we will review recent advances in minizing communication for linear algebra operations.  Communication avoiding algorithms refer to a new class of algorithms that provably minimize communication, in terms of both volume of communication and number of messages transferred on the critical path of the algorithms.  After a brief introduction of communication avoiding algorithms for dense operations that have been introduced in the recent years, this talk will focus mainly on iterative methods, incomplete LU factorizations, two level preconditioners, and their impact on a challenging application in astrophysics, the CMB data analysis.

### E. Francesquin, INRIA

The  Actor Model and Multi-core Architectures
Advisors: A. Goldman (USP Sao Paulo), J-F. Méhaut (UJF-CEA, Grenoble)

The Actor model for parallel and concurrent programming has been in use for at least two decades [1]. However, it was not until recently that the interest in this model has been rekindled due, in part, to the emergence of multi and many-core architectures. In this model, there is no shared memory and the communication is entirely based on message passing. Newest multi-core machines have a hierarchical memory structure, meaning that the time need to send a message from one actor to the  other changes significantly depending on their location. This difference is specially noticeable if the machine in question is a NUMA machine. Applications developed using the actor model trust the runtime environment to do an efficient actor-to-core mapping. In this presentation we will show some of our ongoing work that aims at the creation of an efficient actor runtime system for these machine architectures. We will show the importance of taking into account not only the machine architecture but also the application characteristics during the scheduling decisions. In order to validate our findings we will use an actual virtual machine, specifically the Erlang VM, exercised using synthetic benchmarks and real applications such as Sim-Diasca, an open source discrete simulation engine developed in Erlang by EDF

### Laercio Pilla, INRIA

A Performance Measurement Approach for Modeling Latency and Bandwidth of Large Scale Multicore Machines
Advisors: P. Navaux (UFRGS/Brésil), J-F. Méhaut (UJF-CEA, Grenoble)

In this presentation, I will discuss our approach for modeling parallel machines by extending current hierarchical topology models with benchmarked information. We are working to measure the actual distance between hardware components using latency and bandwidth as measurement metrics. We report these metrics for the different levels of cache, memory, and network available in the parallel machine. This information can be provided to other algorithms interested in locality and communication costs, such as schedulers, load balancers, and memory policies. Our main case study applies our machine topology model in a load balancing algorithm developed in Charm++.

### Tim Tautges

Mesh-based Data and Algorithms across the Simulation Process: anecdotes, activities, and opportunities

### Emmanuel Jeannot

Process placement with unbalanced architecture
We describe how, within the TreeMatch tool, we handle the case where the architecture is unbalanced (e.g. where not all the core of a given architecture are usable).

### Sanjay Kale

A perspective on the BigSim approach to performance prediction

Over the past 10 years, in progress made as funding permitted, we have developed an approach to predicting performance of applications running on large scale computers using runs on substantially smaller computers. The emulation-followed-by-simulation model of BigSim allows for a spectrum of different resolution levels, both for computation and communication. It also makes it easy for the application developers to use the prediction machinery: one simply runs the application as if running on the future machine, in an emulation mode. This helps catch the code-scalability bugs in the first place. We developed techniques for handling the memry-overload problem or emulation. For simulation, both PDES and sequential simulation is used. One of the main utilities of the model is in predicting the network performance especially as it pertains to contention. I will describe the past experience and potential future plans for the BigSim approach.

### Pavan Balaji

MPI-ACC: A Unified Data Movement Infrastructure for MPI and Accelerators

Bill Kramer

Is There a Life After the Top500? (Or What to Do About the Top Problems with the TOP500 List)