Inria-UIUC/NCSA-ANL-BSC-JSC-Riken Joint Laboratory on Extreme Scale Computing (JLESC)

Franck Cappello,
Argonne National Laboratory
University of Illinois at Urbana Champaign

Why? And What is it?

JLESC Purpose

Making the bridge between Petascale and Extreme computing

Funding institutions

JLESC International, virtual organization

Researchers, Students, Engineers, From CS, Maths And also other discisplines

Full Partners

JLESC Objectives

Context:
Computational and data focused simulation and analytics at scale.

JLESC time line

- Created in 2009 with Inria and UIUC as the partners
- Argonne joined in 2013
- Barcelona Supercomputing Center joined in 2014
- Julich Supercomputing center joined in 2015
- Riken/AICS joined in 2015

→ Running full speed (may enlarge the number of partners later)

The JLESC agreement was signed in 2014 for 4 years.
Renewed if the majority of full partners agree

Challenges from Petascale to Extreme scale

- Address Heterogeneity
  - Cores: (slim) compute cores, accelerators, (fat) system cores
  - Deeper memory hierarchy: Caches, on socket high bandwidth memory, DRAM, 3D Xpoint (fast non volatile), NVM (NAND based)
- Load balancing
  - Static reasons: process variation-gate length, wire width, etc.,
  - Dynamic reasons: power management, Turbo modes, etc.
  - Heterogeneous Applications (multi-physics, workflows)
- Resilience
  - not only for process crash, DUE and radiation induced SDC
  - But also for Bugs that could lead to systematic errors
- Plateauing of the HDD storage bandwidth
  - Memory keeps increasing (X10 PB), storage I/O is not (1TB/s).
  - In situ data analytics, Burst Buffers, Compression (error bound lossy compression)

→ A lot of pressure on Applications, Numerical algorithms, Runtimes, System and I/O developers

JLESC Research topics

- Computational sciences, application, mini-apps
- Parallel programming models and libraries,
- Numerical algorithms and libraries,
- Data analytics, graph algorithms,
- Parallel I/O systems and libraries,
- Storage infrastructure design and efficiency
- System and application performance analysis and modeling (Tools),
- Resilience (checkpoint/restart, SDC detection)

JLESC Match Making/Reporting

- A 2 to 3-days workshop every six months (next ones Lyon and Kobe)
- A summer school every year
- Researcher/student exchanges. (from days to 1 year or more)

inside HPC

News | HPC Hardware | HPC Software | Industry Segments | White Papers | Resources | Special Repo

Sign up for our newsletter and get the latest HPC news and analysis. Email Address

Home » Industry Segments » Collaboration » Top HPC Centers Meet in Barcelona at JLESC

Top HPC Centers Meet in Barcelona at JLESC

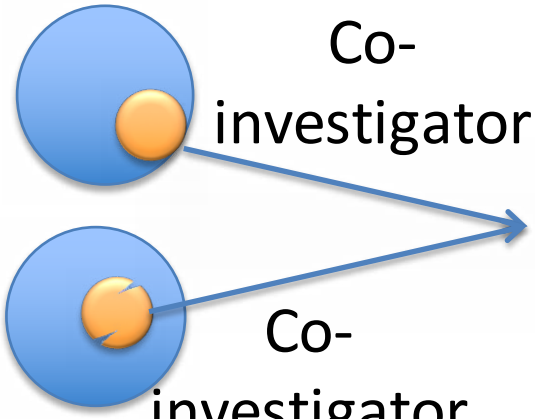
July 6, 2015 by Rich Brueckner

researchers from six of the
est supercomputing centers got
ther in Barcelona at the
nning of this month for the Joint
oratory for Extreme-Scale
puting (JLESC) to discuss the
lenges for future
ercomputers. Marc Snir
lighted US Government
grams to promote new
ercomputing systems; Thomas
ert discussed the EU Flagship
nan Brain Project'; and Akira
wa explained the point of view
n Riken, Japan's largest research institution.

JLESC 2015: 3rd Joint Laboratory for Extreme-Scale Computing

JLESC Research Project

Partner institutions



Co-investigator

Co-investigator

JLESC Joint Projects

Research

New ideas
Publications,
Software

Research results
Presented at
JLESC workshops

Evaluated by the JLESC committees

Reporting

A project proposal:
A statement of goals for the project,
a program of work,
and a schedule to accomplish the goals.

A project report presented in the JLESC activity report

Evaluated by the JLESC committees

JLESC Facilities

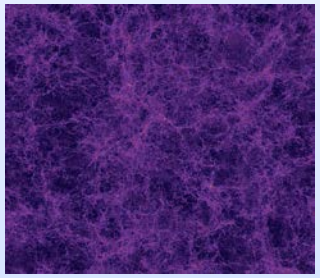
Joint Laboratory for Extreme-Scale Computing

An institution involved in JLESC will provide to JLESC visitors involved in collaborations (if available):

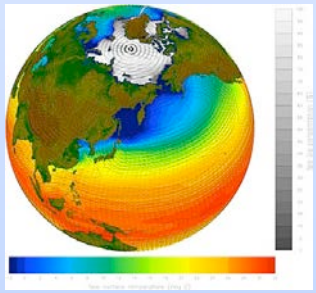
- Office space, internet access, administrative support, and
- Access to its local HPC resources during their visit.

What Collaborations?

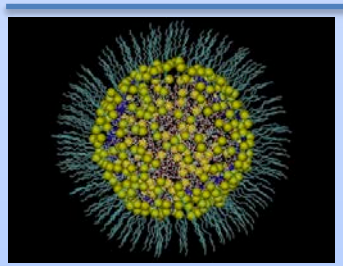
Applications are the drivers of JLPC research



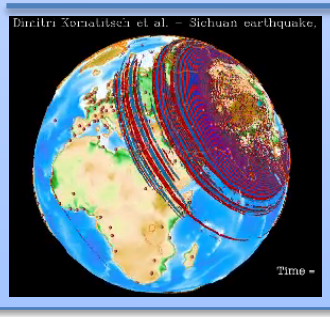
HACC: (Hardware/Hybrid Accelerated Cosmology Code) Framework. New Cosmological N-Body Framework, Designed for extreme performance AND portability, including heterogeneous systems, Supports multiple programming models, Memory efficient, In situ analysis framework
Largest Executions on Mira



The Community Earth System Model (**CESM/ACME**) is a fully-coupled, global climate model that provides state-of-the-art computer simulations of the Earth's past, present, and future climate states. The CESM project will address important areas of climate system research. In particular, it is aimed at understanding and predicting the climate system. **Multiple coupled modules**



NAMD, recipient of a 2002 Gordon Bell Award, is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. Based on Charm++ parallel objects, NAMD scales to hundreds of processors on high-end parallel platforms. **Large scale distributed objects program**



SPECFEM3D simulates seismic wave propagation in sedimentary basins or any other regional geological model. SPECFEM3D version 2.0, named "Sesame", uses the continuous Galerkin spectral-element method, to simulate forward and adjoint coupled acoustic-(an)elastic seismic wave propagation on arbitrary unstructured hexahedral meshes. **Hybrid code mixing CPUs and GPUs**

Active Projects (Apps&Numerics)

- Comparison of Meshing and CFD Methods for Accurate Flow Simulations on HPC system

  Members: M. Tsubokura (RIKEN), A. Lintermann (JSC)

- Strong Scalability Enhancements to FMM for Molecular Dynamics Simulations

  Members: A. Amer, P. Balaji (ANL), I. Kabadshow (JSC), D. Haensel (JSC)

  Fast Integrators for Scalable Quantum Molecular Dynamics

  Members: A. Schleife (UIUC), E. Constantinescu (ANL)

  HPC libraries for solving dense symmetric eigenvalue problems

  Members: T. Imamura (RIKEN), I. Gutheil (JSC)

  Shared Infrastructure for Source Transformation Automatic Differentiation

  Members: L. Hascoët (INRIA), S. Hari Krishna Narayanan, Paul Hovland (ANL)

- Reducing Communication in Sparse Iterative and Direct Solvers

  Members: A. Bienz, L. Olson, B. Gropp (UIUC), L. Grigori, (INRIA)

  Iterative and direct solvers in a hybrid MPI/OpenMP computational mechanics code (Alya)

  Members: G. Houzeaux (BSC), Luc Giraud, Pierre Ramet (Inria)

  Optimizing ChASE eigensolver for Bethe-Salpeter computations on multi-GPUs

  Members: A. Schleife (UIUC), E. Di Naopli (JSC)

Active Projects (Prog. Models, Runtime & tools)

Enhancing Asynchronous Parallelism in OmpSs with Argobots

Members: P. Balaji, A. Amer, S. Seo (ANL), J. Labarta, R. M. Badia, X. Teruel, V. Beltran Querol (BSC)

- ## Load rebalancing of a particle code

  Members: C. Lachat, E. Jeannot, A. Denis (Inria), G. Houzeaux, M. Vazquez (BSC)

- ## Developer tools for porting & tuning parallel applications on extreme-scale parallel systems

  Members: B. J. N. Wylie C. Feld (JSC), M. Tsuji H. Muraj (RIKEN), J. Gimenez (BSC)

Study the use of the Folding hardware-based profiler to assist on data distribution for heterogeneous memory systems in HPC

Members: A. J. Pena, H. Servat, J. Labarta (BSC), P. Balaji (ANL)

Active Projects (I/O, Visualization, Storage)

Joint Laboratory for Extreme-Scale Computing

Argonne NATIONAL LABORATORY · Barcelona Supercomputing Center · JÜLICH FORSCHUNGSZENTRUM · RIKEN

Resource management and scheduling for data-intensive HPC workflows

Members: N.l Cheriere, Gabriel Antoniu (Inria), M. Dorier, R. Ross (ANL)

Elastic provisioning for data streams

Members: L. Pineda, O. Marcu, Alexandru Costan, Gabriel Antoniu (Inria), B. Subramaniam, Kate Keahey (ANL),

Toward taming large and complex data flows in datacentric supercomputing

Members: E. Jeannot (Inria), François Tessier (Inria, now ANL), V. Vishwanath (ANL)

- Extreme-Scale Workflow Tools: Swift, Decaf, Damaris, and FlowVR

  Members: J. M. Wozniak, T. Peterka, M. Dreher, M. Dorier, M. Dreher, R. Ross (ANL), S. Ibrahim, O. Yildiz, G. Antoniu, B. Raffin (INRIA)

- Smart in situ visualization

  Members: L. Rahmani, L. Bougé, G. Antoniu(Inria), M. Dorier (Inria, now ANL), T. Peterka (ANL)

Exploiting Omnisc'IO to imporve scheduling, prefetching and caching in Exascale systems

Members: S. Ibrahim, G. Antoniu (Inria), M. Dorier (Inria now ANL), R. Ross (ANL)

- Towards Interference-aware scheduling in HPC systems

  Members: O. Yildiz, S. Ibrahim, G. Antoniu (Inria), M. Dorier (Inria, now ANL), R. Ross (ANL)

Modeling and avoiding execution interferences

Members: F. Wagner, R. Bleuse, D. Trystram (Inria), G. Bauer (UIUC), F. Cappello, (ANL)

Active Projects (Resilience)

- **Evaluating Lossy Compression for HPC checkpointing**

  Members: J. Calhoun L. Olson M. Snir (UIUC), Sheng Di, F. Cappello (ANL)

- **New Techniques to Design Silent Data Corruption Detectors**

  Members: O. Subasi, L. Bautista Gomez (ANL, now BSC), O. Unsal, J. Labarta (BSC), S. Di, P.-L. Guhur, F. Cappello (ANL), A. Benoit, Y. Robert, A. Cavelan, H. Sun (Inria)

- **Hybrid resilience for MPI + tasks codes**

  Members: T. Martsinkevich (Inria, now Riken), F. Cappello (ANL), O.Subasi, O. Unsal (BSC)

- **Optimization of fault tolerance strategy for workflow**

  Members: S. Di, T. Peterka , F. Cappello (ANL), A. Cavelan , A. Benoit, Yves Robert (Inria)

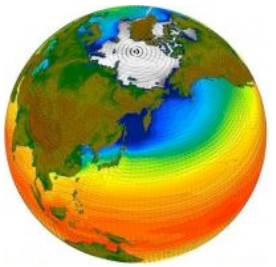
Success stories

The G8 (multi-disciplinary) ECS project for Enabling Climate Simulation at Extreme Scale

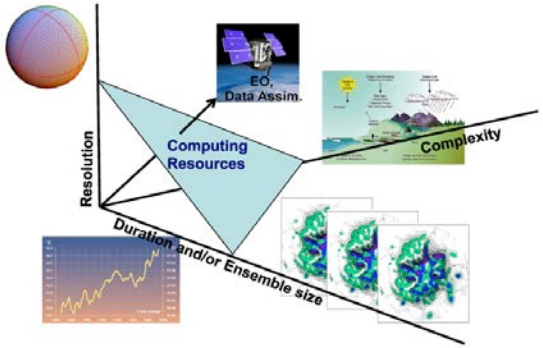
F. Cappello (Main PI, ANL-UIUC, INRIA), M. Snir (ANL, UIUC), G. Bosilca (UTK), L. Giraud (INRIA), J. Labarta (BSC), R. Loft (NCAR), S. Matsuoka (Titech), M. Sato (U. Tsukuba), F. Wolf (GRS), O. Unsal, A. Weaver (U. Victoria), J. White (NCAR), D. Wuebbles (UIUC)

- Objectives:
  - Understand what will be the major obstacles that the climate community will face at Exascale
  - Propose and evaluate possible ways to overcome these obstacles
- Target Codes and Systems:
  - CESM (North America) and NICAM (Japan)
  - K Computer, BlueGene P/Q, Blue Waters, Tsubame 2
- Results:
  - Identified scenarios of climate simulations at extreme scale
  - Modeling and auto-tuning/scheduling for intra-node heterogen
  - Identify likely scalability bottlenecks & define possible solutions
  - New hybrid fault-tolerant protocols

NICAM CESM



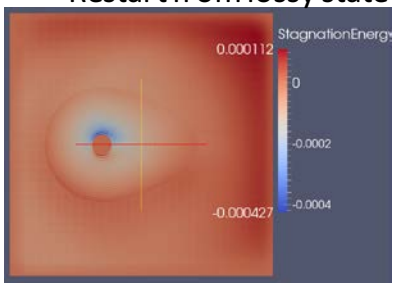
- 58 participants
- 40 publications, 25 talks
- 7 software prototypes
- 11 advised students
- 24 visits



Cracking the Fault Tolerance Challenge of Exascale Systems (since 2009)

Team: Osman Unsal, Omer Subasi (BSC), Tatiana Martsinlevich, Tomas Ropars, Amina Guermouche, Yves Robert (Inria), Leonardo Bautista Gomez, Sheng Di, Franck Cappello, Marc Snir (ANL and UIUC)

Restart from lossy state



First challenge: Reduce Checkpoint/restart time

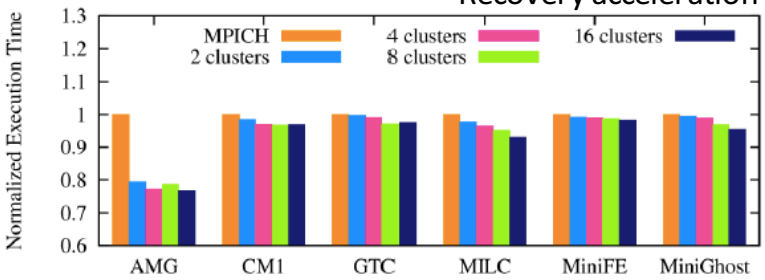
→ FTI (fault tolerance Interface) for multi-level checkpointing

→ Lossy compression of checkpoints and restart from lossy states

Second challenge: Reduce as much as possible checkpoint/restart overhead

→ Formulation of optimal checkpoint intervals for multi-level checkpointing

Recovery acceleration



Third Challenge: Accelerating restart

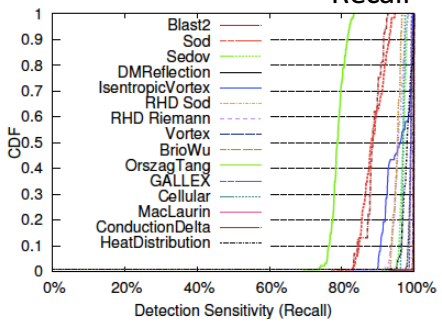
→ Identified the send-determinism and developed several low overhead message logging protocol.

→ Solved memory occupation issue of message logging.

Fourth Challenge: Detection of Systematic and non-systematic Silent Data Corruptions

→ Developed a new class of low overhead SDC detectors based on surrogate functions (using prediction, machine learning and exploiting numerical properties)

Recall



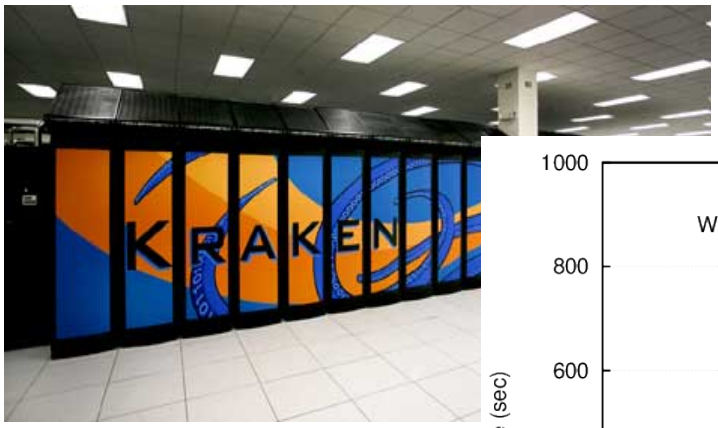
> 10 visits, > 10 publications, 5 software prototypes

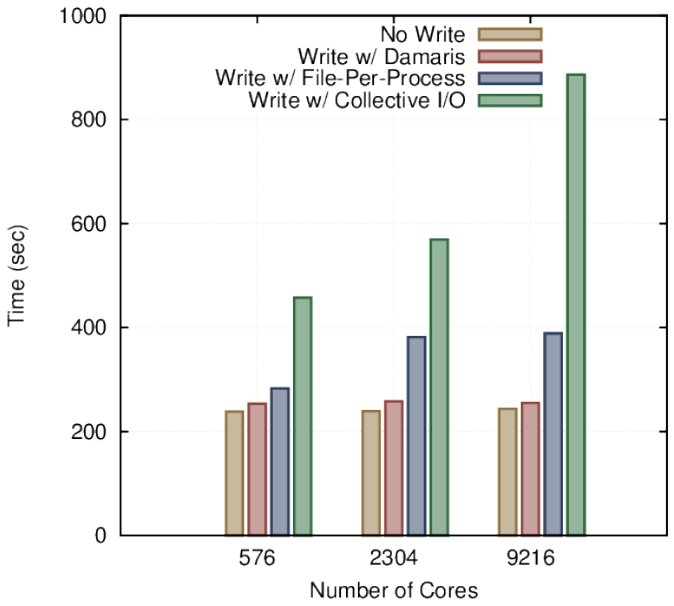
Addressing I/O Bottleneck (A critical problem at Exascale)

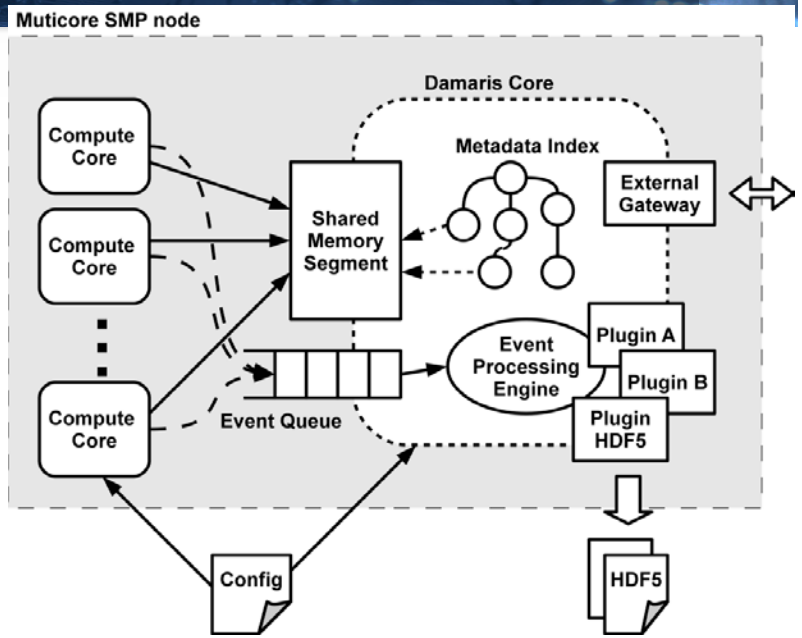
Damaris: A Middleware for I/O on Multicore

Team: Gabriel Antoniu (Inria), Matthieu Dorier (Inria now ANL), Rob Ross (ANL), Marc Snir (UIUC)

Idea : one dedicated I/O core per multicore node
Originality : shared memory, asynchronous processing
Implementation: software library
Application: tornado simulation, fluid dynamics
Transferred to the Blue Waters supercomputer (2011)

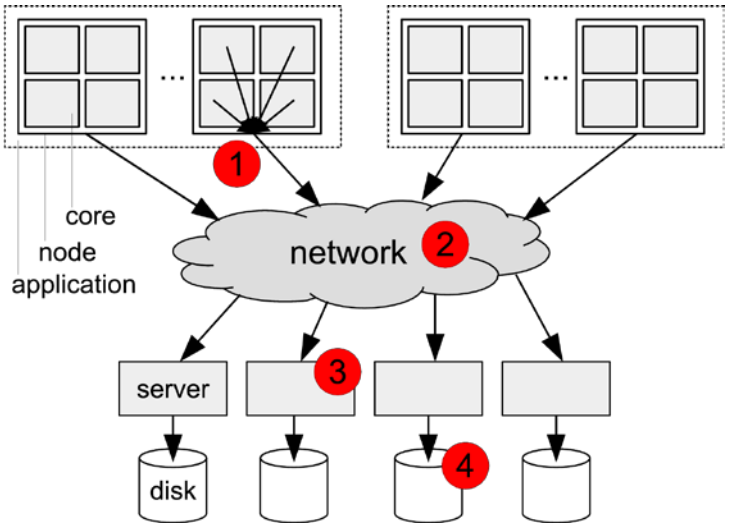






- Scales on 10,000+ cores on Kraken (11th of Top500 in 2010)
- Scales on 16,000+ cores on Titan (1st of Top500 in 2012)
- x12 less files
- x15 write throughput
- Jitter fully hidden
- Predictable performance
- Damaris ADT project: (2015-2017)

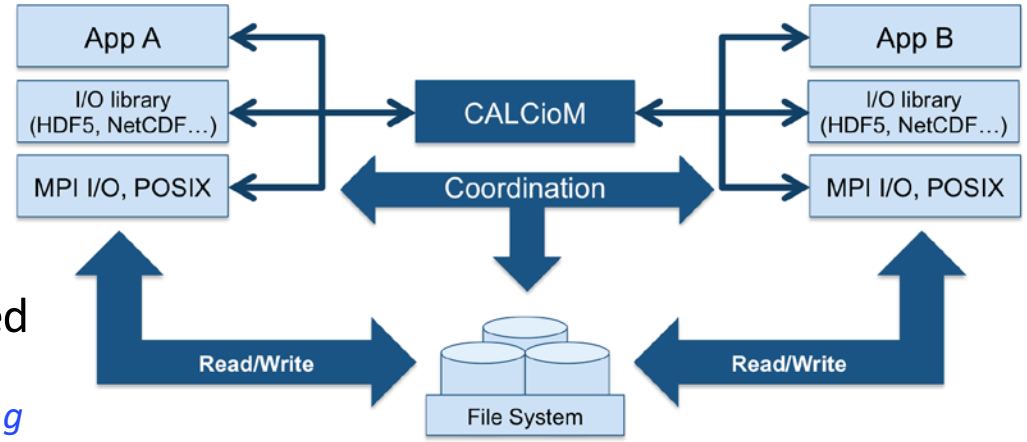
Addressing I/O Interference



- Identifying potential root causes of interference
  - Network interface and protocols
  - Network components
  - Storage servers, schedulers
  - Storage devices (disks, SSDs)
- Interplay between the above sources

O. Yildiz et al. IPDPS 2016, *On the Root Causes of Cross-Application I/O Interference in HPC Storage Systems*

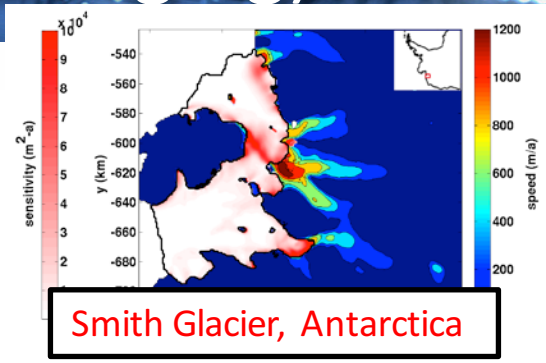
- Mitigating I/O interference: the CALCioM approach
- Main ideas:
  - applications communicate with one another
  - different coordination strategies can be implemented



M. Dorier et al. IPDPS 2014, *CALCioM: Mitigating I/O Interference in HPC Systems through Cross-Application Coordination*

Paving the road of Algorithmic Differentiation (Fundamental and technically challenging)

Laurent Hascoët, Valeri Pascual (Inria)
Paul Hovland, Sri Hari Krishna Narayanan, Michel Schanen,
Jean Utke* (ANL).



Smith Glacier, Antarctica

- AD is a technique for computing derivatives of programs
  - Example of Derivatives use: **sensitivity analysis**
    - Study the sensitivity of the model outputs WRT uncertain input parameters
    - If R is a function of X and Y, hen the uncertainty in R is obtained by:
    - Where åX and åY are the uncertainty on X and Y

$$R = R(X, Y, \ldots)$$
$$\delta R = \sqrt{\left(\frac{\partial R}{\partial X} \cdot \delta X\right)^2 + \left(\frac{\partial R}{\partial Y} \cdot \delta Y\right)^2 + \ldots}$$

  - Challenges: Efficiency of derivative code, complexity of AD tool development, dynamic allocation of memory.

- Efficiency: Developed an implementation of the Christianson method for differentiating fixed point iterations within OpenAD (with Dan Goldberg from Edinburgh). Application: MITgcm package halfpipe_streamice
- Complexity of AD Tools: Developing TapenadeXAIF -- An interface between Inria's Tapenade and OpenAD from Argonne. Leverages Tapenade's source analysis and OpenAD's differentiation algorithms. Application: OpenAD regression suite
- Memory Allocation: Developed ADMM – the first library to support AD of codes containing dynamic memory allocation/deallocation. Application: Chemical engineering process model: simulated moving bed (SMB)

Promising Stories

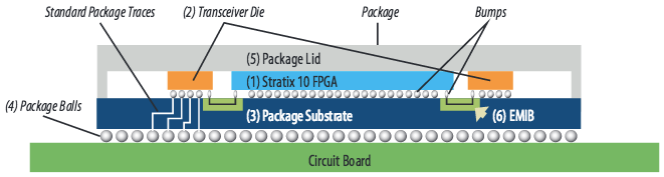
Beyond Moore's law: Reconfigurable Computing

How to increase the performance when Frequency, Power, and Integration are bounded?

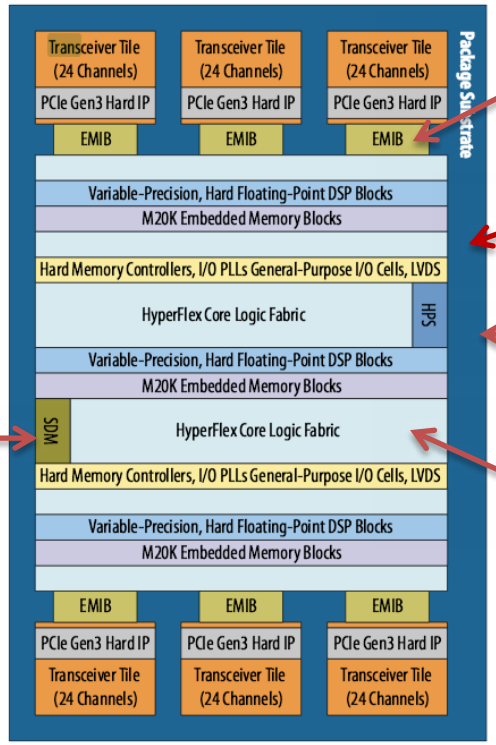
1) Dedicated architectures (Anton), 2) Reconfigurable

Example: Stratix 10 FPGA SoC:

- 10 TFLOPs (single-precision IEEE 754 DSPs)
- ~100 GFlops/Watt (SP)
- 64 bit quad-core ARM (1.5 GHz)
- 1 Tbps bandwidth (Hybrid Memory Cube)
- Up to 144 30 Gbps transceivers
- 5.5M logic elements (FPGA)
- 1 GHz fabric clocking (2X previous gen.)



Multi-die on substrate



Secure Device Manager

Embedded Multi-Die Interconnect Bridge

Intel 14 nm tri-gate technology (*Knights Landing*)

Quad 64 bits ARM Cortex-A53 Hard Processor (1.5 GHz)

5.5M logic elements
1 GHz fabric clocking
(2X previous gen.)

-Discussion initiated December 2015 JLESC workshop on FPGA for HPC and data analytics
-Specific workshop at ANL on FPGA (Riken, BSC, UIUC and ANL) with Xilinx and Altera
-Session on advanced architecture in the June 2016 JLESC workshop
-Specific workshop at UIUC on FPGA

What Impact?

Evaluation Criteria for the Scientific/Advisory Committee

- Impact on extreme scale platforms
- Impact on the broad HPC community
- Impact on HPC research in the different institutions

- Quality and number of publications
- Software prototypes – open-source or not, number of downloads
- Integration of the software prototypes into Extreme Scale Platform software stack
- Research grants
- Visibility: *Keynotes, press releases, interviews*
- Education of young researchers: advising and teaching
- Transfer, involvement of industry, patents, standardization
- Research activities management (including workshop organization, meeting organization)
- Dissemination (to a large public audience)

Quantitative evaluation (not counting 2015-2016)

Number of active projects: 10 to 23

Total number of Person Months: 829

Number of visits: 63

Number of accepted publications: 61

Number of software : 6

Student advising: 29

Number of additional funding: 5

Number of student awards: 4

JLESC as a Research Object

JLESC is the most successful international collaborations of UIUC

Why?

Scoot Poole (Human Science) uses JLESC as a research object:

- Understand what make collaborations work
- Formalize
- Help other collaborations

Thank you