BIG DATA AND EXTREME-SCALE COMPUTING

Exec Committee
Pete Beckman
Jean-Yves Berthou
Jack Dongarra
Yutaka Ishikawa
Satoshi Matsuoka
Philippe Ricoux

Following the International Exascale Software Initiative (IESP 2008-2012 => **Big Data and Extreme Computing** workshops series (BDEC)

http://www.exascale.org/bdec/

Overarching goal:

1. Create an international collaborative process focused on the co-design of software infrastructure for extreme scale science, addressing the challenges of both extreme scale computing and big data, and supporting a broad spectrum of major research domains,
2. Describe funding structures and strategies of public bodies with Exascale R&D goals worldwide
3. Establishing and maintaining a global network of expertise and funding bodies in the area of Exascale computing

1 – BDEC Workshop, Charleston, SC, USA, April 29-May 1, 2013

2 – BDEC Workshop, Fukuoka, Japan, February 26-28, 2014

3 – BDEC Workshop, Barcelona, Spain, January 28-30, 2015

1 - BDEC Workshop, Charleston, SC, USA, April 29-May 1, 2013

Big Data and Extreme-scale Computing (BDEC) Workshop, Charleston, SC, USA, April 29-May 1, 2013

- **Workflow Issues**
- **Architecture Challenges**
- **Higher Level Data Challenges : Data provenance, Policy based data management, Environments that support new types of data-driven research, Shared software infrastructure for intermediate processing**
- **Software Challenges: Tools to support real-time monitoring and observation of workflows, Coordination between data movement and compute services, Mechanisms to support fault tolerant workflows in data analysis, Mini-apps to support infrastructure co-design, Integration of widely used BD-capable data libraries into standard packages, Common tools for managing and exploring data, Interoperability Challenges**

BDEC Workshop Report (November 29, 2013)

Report on the Big Data and Extreme-scale Computing (BDEC) Workshop, Charleston, SC, USA, April 29-May1, 2013

1 Introduction

This report on the Big Data and Extreme-scale Computing (BDEC) workshop offers an initial account of the effort to develop a plan for sustained international cooperation in the design and development of a new generation software infrastructure for extreme scale science. The meeting, the first of a planned series, derived much of its impetus from the earlier work of the International Exascale Software Project (IESP) and the European Exascale Software Initiative (EESI, http://www.eesi-project.eu). The goal of the IESP was two-fold: 1) to produce a plan for a common, high quality computational environment for the peta/exascale systems that are expected to arrive over the next decade; and 2) to mobilize and coordinate the work of the international open source software community to create that environment. BDEC retains those goals but changes the point of view. The EESI coordinated the European contribution to IESP.

The IESP, working through a series of eight international meetings held from 2009 to 2012, built on a range of important earlier studies, including [1-4], to produce a widely read and cited "roadmap" document. The IESP Roadmap [5] presented a multidimensional analysis of the major challenges to be overcome in order to create a software infrastructure capable of supporting exaflop performance on next generation systems, and made a cogent case for the urgency of starting that work as soon as possible. Spurred in some degree by the work of the IESP and its Roadmap, the United States, the European Union, and Japan have, in the past three years, moved aggressively to develop their own plans for achieving exascale computing in the next decade. The EESI produced a European roadmap along with a set of recommendations to address the Petascal/Exascale challenge [10].

The first BDEC workshop marks a beginning of a distinct new phase of this community movement. The motivation for this second stage is based on the recognition that the "digital data deluge," which was sighted on the horizon well over a decade ago [6], has finally made landfall with impressive force. It is apparent that in the era of "Big Data," when every major field of science and engineering is producing, and needs to (repeatedly) process, truly extraordinary amounts of data, the many unsolved problems surrounding ***wide-area, multistage workflows****—the diverse patterns of when, where, and how all that data is to be produced, transformed, shared, and analyzed*—have to take center stage. Although the IESP Roadmap shows a clear awareness that extreme scale science inevitably means extreme scale data as well as extreme scale computing, IESP working groups, for the most part, adopted a traditional HPC (i.e., supercomputer centric) perspective. They were largely (and understandably) preoccupied by the impending software crisis caused by the move to the new paradigm in hardware and systems architecture, a paradigm that demands orders of magnitude more parallelism, places unprecedented constraints on energy consumption, and requires resilience to faults occurring at far higher frequencies than ever before. Data-driven workflow issues received some collateral discussion in the Roadmap, but the focus of attention for the IESP was on the revolutionary innovations in the system software stack that would be needed to address the steep challenges of emerging peta/exascale systems.

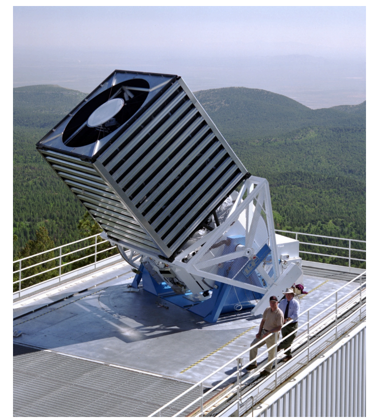
How Did We Get Here

- Previous BDEC meetings: US & JP
- Application Drivers: Astronomy, Medical, Genomics, Climate, Human Brain, Satellite images (GIS), Social Networks, etc.
- Good discussions on converged / shared problems:
  - Architecture, operations, software stack, algorithms

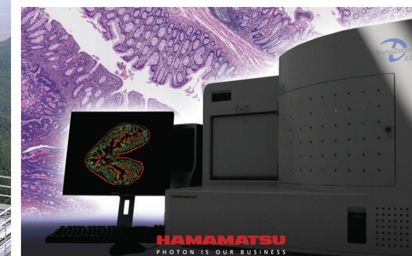
Instruments & Facilities

- "HPC Instrument" (Tsubame, Mira)
- SDSS, LSST, SKA, LOFAR , …
- APS(20x), SNS, …
- DNA Sequencers
- LHC / Atlas
- ARM

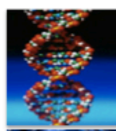
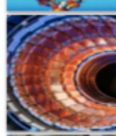
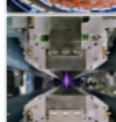
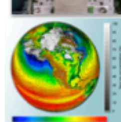
HPC



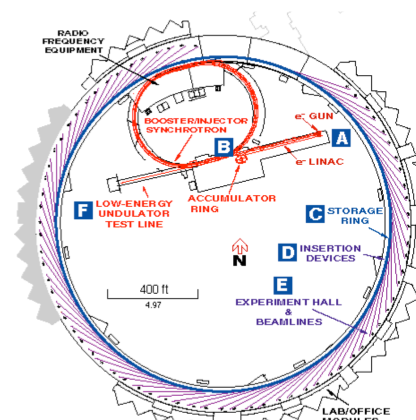
Sloan DSS



Tissue Samples

	**Genomics** Data Volume increases to 10 PB in FY21
	**High Energy Physics** (Large Hadron Collider) 15 PB of data/year
	**Light Sources** Approximately 300 TB/day
	**Climate** Data expected to be hundreds of 100 EB

ARM



APS

Comparing Architecture

Big Data	Extreme Computing
? Cost in memory and interconnect bandwidth	***Significant Cost*** in memory and interconnect bandwidth
Little Cost for resilient hardware in data storage	***Significant Cost*** in resilient hardware in shared file system
Little Cost for hardware to support system-wide resilience	***Significant Cost*** in resilience hardware to reduce whole-system MTTI
Significant Cost: ***increased aggregate IOPs***	Significant Cost: ***cutting-edge CPU performance features***
Often trades performance for capacity	***Often trades capacity for performance***

Comparing Operations

Big Data	Extreme Computing
Continuous access to long-lived "services" created by science community	***Periodic access*** to compute resources via job submitted to scheduler and queue
Time-shared access to elastic resources	***Space-shared*** compute resources for exclusive access during jobs
New hardware capacity ***purchased incrementally***	New tightly integrated system ***purchased every 4 years***
Users charged for all resources (storage, cpu, networking)	***Users charged for CPU hours, storage and networking is free***

Comparing Software

Big Data	Extreme Computing
Software responds to elastic resource demands	After allocation, ***resources static until termination***
Data access often ***fine-grained***	Data access is ***large bulk*** (aggregated) requests
Services are resilient to fault	***Applications restart after fault***
Often ***customized*** programming models	Widely ***standardized*** programming models
Libraries help ***move computation to storage***	Libraries help ***move data to CPUs***
Users routinely deploy their own services	***Users almost never deploy customized services***

Comparing Data

Scientific Big Data	Extreme Computing
Inputs ***arrive continuously***, streaming workflows	Inputs ***arrive infrequently***, buffering carefully managed
Data is ***unrepeatable*** snapshot in time	Data often ***reproducible*** (repeat simulation)
Data generated by sensors (***error: from measurement***)	Data generated from simulation (***error: from simulation***)
Data rate ***limited by sensors***	Data rate ***limited by platform***
Data often ***shared and curated*** by community	Data ***often private***
Often ***unstructured***	***Semi-structured***

What can we use from EC for BD?

HPC Software A Good Base

- MPI-IO, HDF5, pnetCDF, HPSS, other ad hoc solutions provide good building blocks
- Needed: Better abstract models, for both high and low level abstractions
  - "DSL" for data manipulation at scale
  - Such systems are data structure + methods (operators)
- Implementations that fully exploit good and clean semantics of access

Interoperability

- HDF5 provides strong support for many aspects of data provenance. Mechanisms exist in pnetCDF.
  - Should a base set be "automatic", much as file creation/ modify time is today?
  - Can we evolve to better interoperability, or are radically new models needed?
- Mathematical representation for continuous data
  - How should the information about the mapping of discrete ➔ continuous be stored *in the file?*
  - How should this be generalized to other representations?
- Accuracy of data values
  - How should accuracy be *efficiently* stored with file?
- Data formats impact performance and scalability
  - Optimizing for interoperability or performance *alone* may impede application
  - You **cannot** pick the format and then (successfully) say "make it fast"

BLUE WATERS
SUSTAINED PETASCALE COMPUTING
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Architecture

- Architecture:
  - What architectural changes are needed for extreme computing storage systems to make them better suited for BD?
    - Better small scale atomic I/O – Solid State Storage?
    - A new storage repository – non POSIX?
    - Seamless storage hierarchies
  - What operational changes are needed to support new storage architectures?
    - Yes – critical resource is bandwidth not CPU
  - Looking at future technologies, what future architectures are possible?
    - Interconnect is the most essential. Processor technology can be whatever it is.
    - Energy efficient memory

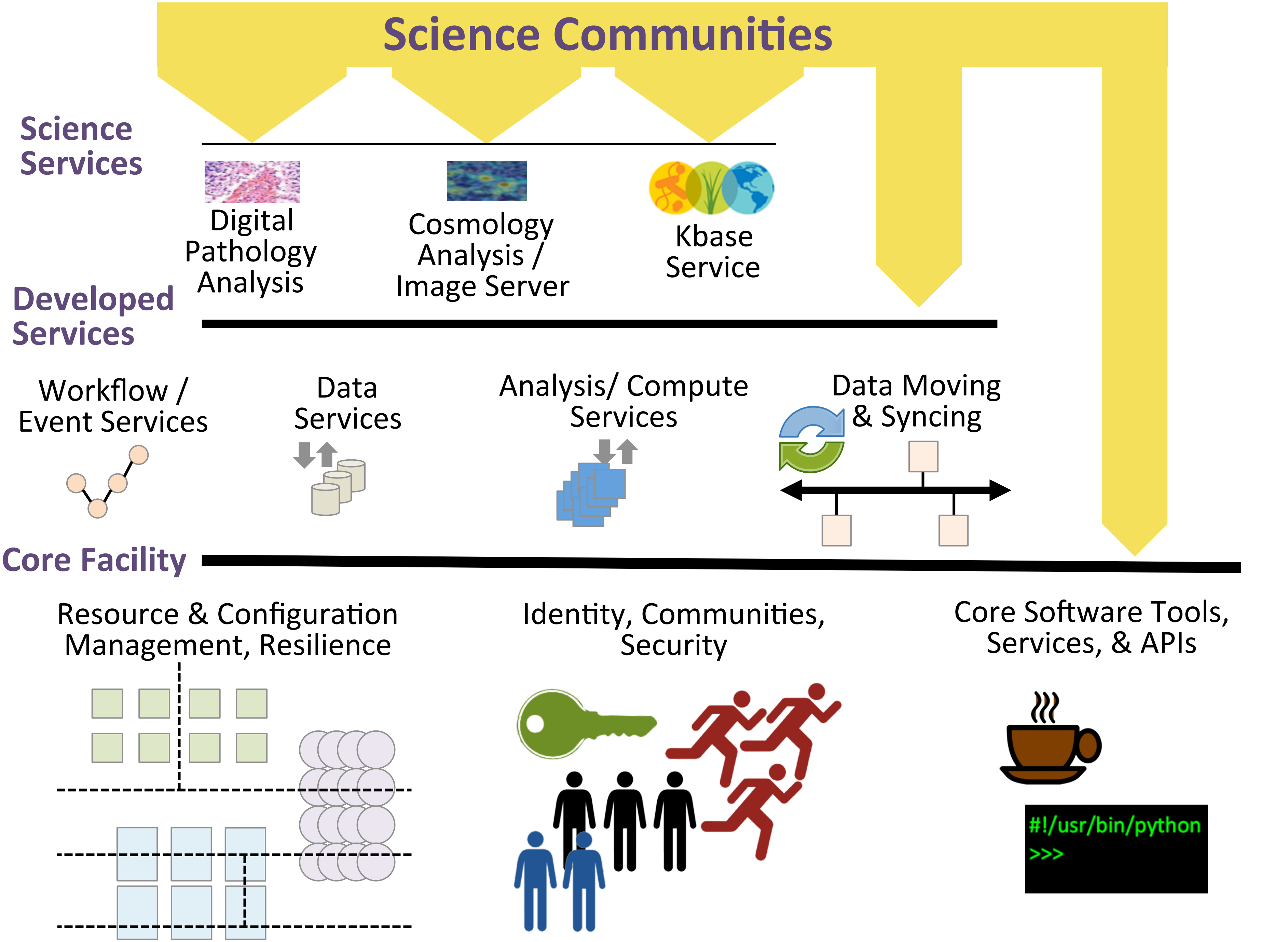
Big Data and Extreme Scale Computation Workshop - April 30 2013 - Charleston SC 9

Define Consistency Models for Access and Update

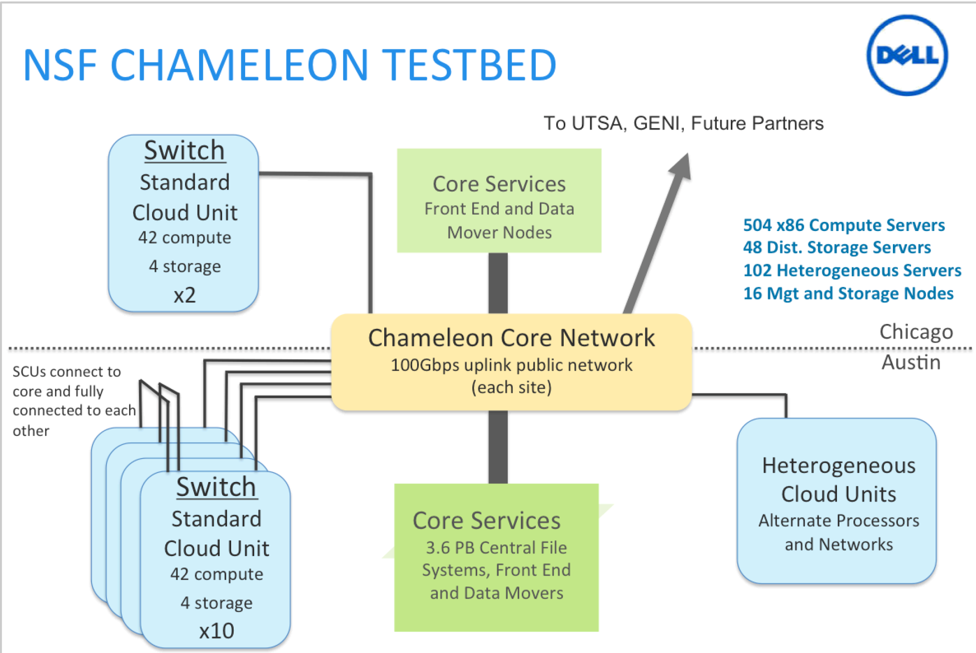
- Need consistency models that match use in applications
  - Or trade accuracy for speed
  - Already happened in search, e-commerce, even when solution is to trade accuracy for speed
    - Witness Amazon's pseudo cart implementation – items aren't really under your control ("in your cart") until you complete the purchase. But greatly simplifies data model.
    - Even though it angers customers on popular deals
- POSIX consistency model is stronger than sequential consistency and almost never what applications require
  - Even when strong consistency is needed, it is almost always on the granularity of a data object, not bytes in a file
  - Long history of file systems falsely claiming to be POSIX
- A bad alternative is the "do what is fast" consistency model – usually but not always works
  - Some systems have taken this route – both I/O and RDMA

18 PARALLEL@ILLINOIS

William Gropp
UIUC
William Kramer
NCSA



Science Communities
Science Services
Digital Pathology Analysis
Cosmology Analysis / Image Server
Kbase Service
Developed Services
Workflow / Event Services
Data Services
Analysis/ Compute Services
Data Moving & Syncing
Core Facility
Resource & Configuration Management, Resilience
Identity, Communities, Security
Core Software Tools, Services, & APIs
#!/usr/bin/python
>>>



New NSF Project to Support This Kind of Research

(see Kate Keahey)

CHAMELEON: A POWERFUL AND FLEXIBLE EXPERIMENTAL INSTRUMENT

- Large-scale instrument
  - Targeting Big Data, Big Compute, Big Instrument research
  - ~650 nodes (~14,500 cores), 2 sites (IU and TACC) connected with 100G network, 5 PB of storage over two sites,
- Reconfigurable instrument
  - Bare metal reconfiguration, operated as single instrument
  - Infrastructure based on OpenStack+Ironic and Grid'5000 technology
- Connected instrument
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users: wrap up your framework as an experimental environment that others can use
- Complementary instrument
  - Complementing GENI, Grid'5000, and potentially other testbeds

Look at Results from Previous Breakout Summaries (exascale.org)

Original Charge: Work toward ***identifying the research questions and promising directions***, not the answers, or how to spend other people's money

- Gaps in the Core Facilities / software arch
  - No elasticity, VMs not well supported, Identity and sharing difficult, etc…
- Missing workflows and BDEC mini-apps for key communities, linking to instruments, etc.
- Apps: Data integration, analysis, classification
  - Cleaning, filtering, transforming, classification, mapping/registration, event detection, prediction
- Data
  - Shared analysis difficult, composition and workflow poorly supported, no benchmarks, etc.

White Papers

- Presenters:
  - White paper presenters will have 6 minutes for their presentations and allowed to have only 4 slides.
  - Please send a pdf file with your presentation your session chair before the beginning of your session.

Goals for This Meeting: How do we Converge?

- Prepare for an initial draft report
  - Barcelona: Material needed to write draft
  - Present initial draft at BOF at SC15
- Report (2 parts)
  - A) What are the current plans / strategies in Asia, Europe, and the US for handling Big Data?
  - B) How do we get BD/EC Convergence?
    - What is missing, specific to each of these respective regional plans? What could be a coordinated, converged plan for BDEC, at international level

Breakout Sessions Introduction

- Breakout Questions to Help:
  – What are the main differences and commonalities between the HPC and BDA requirements/technologies/working-assumptions in this area?
  – Are there common needs/problems/interfaces could serve as the basis (or as stepping stones) along a path to (some reasonable level of) infrastructure and application convergence?
  – Are there interdomain testbeds that combine BDA and HPC workflows in ways that could help uncover pathways toward convergence?
  – What is/are the technology or new research that may be a game changer?
  – What action would be your number one priority to be taken rapidly to ensure success of the converge of Extreme computing and Big Data infrastructures?
  – What action would be your number one priority to be taken rapidly to ensure the emergence of efficient Extreme computing and Big Data applications?
  – How would you measure the success of the BDEC initiative?

Goals For This Meeting: How Do We Converge?

Breakout groups

Applications and Science
Chairs: David Keyes, Rosa Badia, Jean-Claude Andre

Architecture and Operation/Comprehensive Production Services
Chairs: Bill Kramer, Ewa Deelman, Francois Bodin

Algorithm and Applied Mathematics
Chairs: Hiroshi Nakashima, Philippe Ricoux, Alison Kennedy

Software Stack
Chairs: Franck Cappello, Kate Keahey, Satoshi Matsuoka