BDEC Architecture Breakout 1

Rick Stevens (ANL)
Satoshi Matsuoka (Tokyo Tech)

Requirements “Big Data Arch” vs. EC

- Streaming in/out large volume of data continuously (BD) vs. burst batch staging (EC)
- Machine balance towards higher I/O bandwidth and IOPs (High Amdahl number), integer operations, …, (BD) vs. floating point (EC) and high BW (BD&EC)
- BW intensive High-performance storage for intermediate data in the workflow (Q: capacity? Performance? Retention time lengh?), (BD) vs. in-situ memory processing (EC)
- Data publication (Q: nature of publication? Whole data sets vs. summary, …) (BD)
- Exabyte capacity requirements along with bandwidth for extreme scale (BD&EC?)

Architectural Trends

- Flops: cheap (heavily SIMDized)
- Integer Ops: less capable than flops (heavy multithreading, branchy)
- Memory volume: manageable (petabytes), subject to power constraints (LPDDR3)
- Memory Bandwidth expensive (c.f. HBM) to remote chips
  - Data movement expensive
  - Deeper memory hierarchy (cache alll the way to 3rd tier)
  - DRAM:Flash:HDD – Cost&Bandwidth&IOPS 100:10:1 (or greater)
- Network : integration of network onto the processor
  - latency 1 microsec, > 100Gbps, at worst will match local I/O BW
  - Full bisection decreasing, however photonics to rescue?
  - SDN capabilities (scatter/gather and other intelligent routing)
- No Perturbation and Bit reproducibility---NOT!

Points of discussion

- What are the new architectural axis? different, same?
- Do we retain the traditional memory/storage hierarchy or not?
  - What would be the shape of the hierarchy? Pyramid? Bell? (depends on use case, cost, technologies…)
  - How would the hierarchy look? In terms of the mode of access? (DB, streaming, memory-mapped, …), & architectural support
  - Distinguish mutability (over time), read/write, …
- What are the architectural "walls" in the current architecture that need to be overcome? (like parallel file system with current I/O nodes arch.)
- What is the main mode of access – object storage vs. file system
- Homogeneous or heterogeneous
  - In the light of memory capacity requirements
  - Specialization to certain types of data analysis
  - Acceleration achieved through network attachment (instead of PCI-e) or other means, but subject to system BW limitations?
- How much total system BW (e.g. bisection & local BW) do we need in BD?
- Long-term storage retention, curation, provenance – Tape, HDD, or even NVM ? – "the buck stops here!"

What is the “right” programming model, and their architectural implications

- What are the primitives of abstractions beyond POSIX? (HDF5, NetCDF, …)
- What are the architectural support for the primitives? Lessons from Classic DB (OODB in particular), map-reduce, …
- Declarative (SQL, Map-reduce, Scala/Erlang, …), Framework-level stuff, …
- Interactive (Map-reduce,
- BD-DSLs (or BD-DSL builders)

Example BDEC Activities

- What are the apps that highlight Big Data?
  - “Big Data” mini(proto)-apps
  - Customizing the architecture to the BD requirement
  - BD&EC merger-object store for semi-persistent data
  - Quantifiable metrics and categorization of “big data” apps
- What existing machines and apps can we use and how?
  - Real machines, emulator(e.g., RAMDISK), simulator, modeling…
  - Open source tools?

notes

BDEC Architectural Questions (1)

- Physical Changes
  - Orders of magnitude scalability with massive data
  - Faster storage, more bandwidth, Amdahl number
  - Short vs. long term, varying data access patterns (streaming vs. random)
  - Change in locality, integration of compute & storage
  - Application driven networking (SDN)
  - Different tradeoff of HW with compute sacrifice, or coexist? Interger vs. FP, SPMD vs. Task Parallel
  - Unstructured sparse problems vs. big data algorithms
  - Power aware, etc. reduction of precision, nearline persistence

- Operational Changes
  - Bottom up queries, interactivity, in-situ data processing & visualization …
  - Workflow, provenance, long-term archive

BDEC Architectural Questions (2)

- Quantitative Measures, Trends, …
  - How much capacity, BW do we need and when?
  - Where in the system?
  - Reliablity and resilience metrics?
- Future Technologies
  - Architectural Trends
    - DRAM:Flash:HDD – Cost&BE 100:10:1
    - NV-DIMM and NV technologies
  - Disruptive technologies
    - Devices, Networks, Architectures

Software Breakout

V. Sarkar, W. Kramer, W. Nagel

BDEC 2013

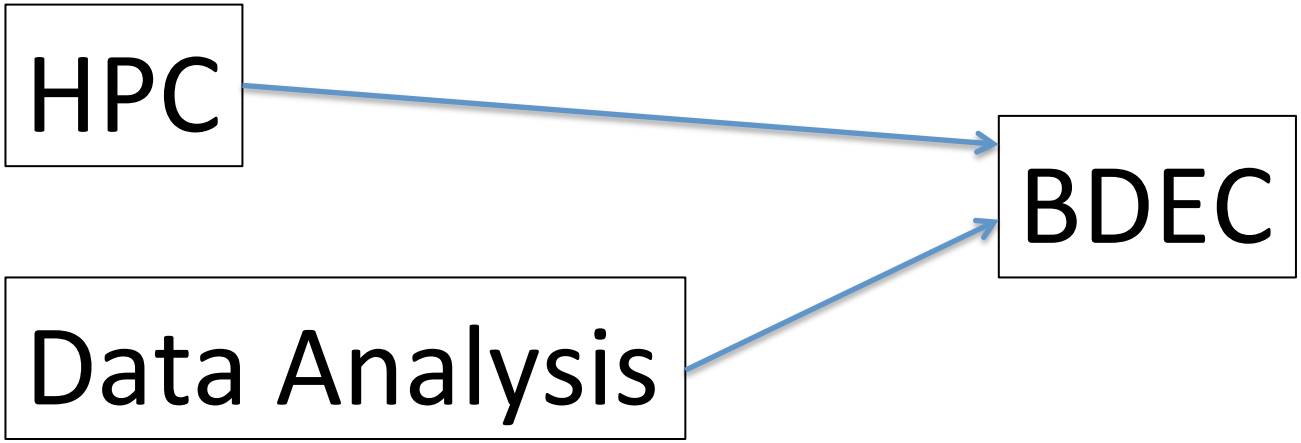
Participants

* Alok Choudhary	Northwestern University
* Geoffrey Fox	Indiana University
* Kelly Gaither	Texas Advanced Computing Center
* Michael Heroux	Sandia National Laboratories
* William Kramer	University of Illinois
* Andrew Lumsdaine	Indiana University
* Bernd Mohr	Jülich Supercomputing Centre
* Shinichi Morishita	University of Tokyo
* Wolfgang E. Nagel	TU Dresden, ZIH
* Kenji Ono	RIKEN Advanced Institute for Computational Science
* Philippe Ricoux	Total SA
* Joel Saltz	Emory University
* Vivek Sarkar	Rice University
* Arie Shoshani	Lawrence Berkeley National Laboratory
* Toyotaro Suzumura	Tokyo Institute of Technology
* Ranga Vatsavai	Oak Ridge National Laboratory
* Jean-Pierre Vilotte	Institut de Physique du Globe de Paris

Context: a collision course

Software:

- What software are you currently using to manage and explore your data?
- What algorithms and software libraries/tools need development and improvement to address your big data needs?
- As you look to the future, what are the holes/gaps that have no planned solution?



Software technology/trends	Current practice	Future BDEC requirements
Storage	Parallel file systems for HPC, databases for data analysis	Integrated approach: scalable object-based storage, middleware, …
Scientist's Workflow	Static data location, move data to compute ahead of time or perform compute where data is located	Interactive visualization, distributed dynamic data location and movement, agility
Data exploration	Low-level queries	Goal-oriented questions at domain level, automatic index generation
Programming model	Compute-centric, legacy HPC languages or inefficient/ immature "big data" languages	DSLs for data-driven computing: persistent data, first-class support for data layouts, semantics of continuous data
Data sharing	Copy and share	Automatic/transparent data access, self-organizing, move compute to data
Data management		Content curation, security (at scale)
Data model	Databases (relational, array database), File formats (HDF5, NetCDF)	Domain-specific models

Software technology/ trends	Current practice	Future BDEC requirements
Statistical analysis, machine learning	R, Mahout, …	Scalable ML runtimes
Data analysis (volume, veracity, velocity, variety), data mining	Static, predictable data sets	Velocity and variety of data (dynamic, unpredictable, load balancing challenges)
Resource management	Batch use, interactive use, reservations, CPU-centric	Peak provisioning, on-demand, adaptive jobs, integrated, data-oriented resource management (BW, …)
Monitoring, introspection	Split responsibility --- between users, sysadmins	Holistic view
Metadata, annotations	Manual	Automatic
Adaptation, auto-tuning		

Software technology/trends	Current practice	Future BDEC requirements
Industry trends	Lowest common denominator --- parallel file systems, tiered storage	Piggyback on industry trends for databases, data analysis (where appropriate)
Education	Parallel computing (OpenMP, MPI)	Integrate parallel computing and data science courses?
Massive numbers of data sources/sensors	Batch, store and move	Flexible, low-latency, high-bandwidth tradeoffs
Software development environment	Classic IDEs are programmer-centric	Scientist workbench with unified view of compute and data
Domain-specific libraries/ frameworks	Compute-centric frameworks	Integrated frameworks for HPC and data access
Operating systems	Minimal use of abstractions, performance tuning, devirtualization	Distributed access control, resource hierarchies, virtualization (OpenStack?)
Runtime systems	Separate runtimes for compute (tasks, threads), minimal runtime support for persistent data (parallel I/O)	Integrated runtime for parallel computation and data access
Resilience	Coarse-grained checkpoint-restart	Fine-grained checkpoint, task idempotence, data integrity

BACKUP SLIDES START HERE

Software technology/trends	Current practice	Future BDEC requirement
Domain-specific requirements for data resilience		Many domains can recover from missing data
Data compression		Much larger compression factors needed in the future

Gaps/Challenges

- HPC Viz
  - Interactive exploration of data
    - E.g., Turbulence, cosmology
  - Need mechanisms for distributed data sets
- Information visualization
- Differences in Data Generation & Data Consumption
  - E.g., data slice over multiple time steps (transpose)
  - Data reduction, filtering
  - Real-time monitoring
- Need taxonomy of workflows

International Collaboration, Frameworks, Funding, and Co-design

BDEC Breakout 3: Day 1

Participants

- Jean-Yves Berthou, French National Research Agency (ANR)
- Bronis de Supinski, Lawrence Livermore National Laboratory
- Yutaka Ishikawa, University of Tokyo
- Daniel Katz, National Science Foundation
- Terry Moore, University of Tennessee
- William Tang, Princeton University
- Pier Luigi Vidale, University of Reading
- Vladimir Voevodin, Research Computing Center -- Moscow State University

International Collaborations

- Generally accepted that international collaborations are essential to accelerate progess
- Current and Budding Funding Efforts
  - Reviewed G8 project
  - Japan-US collaboration is underway
    - Targeting open source production software and interfaces
    - Small workshop at Tsukuba last month
    - Informal steering committee meeting on Sunday
      - Looking to invite other countries interested in project
    - Workshop Tuesday at ISC
- Data collaborations are user driven, may lead to stronger relationships

Co-design

- Efforts in progress in Japan
  - One team designing mini-apps
  - Three teams for new architecture; will use mini-apps to evaluate designs
- US
  - Office of Science
    - Current projects: Extreme Materials (Tim German), Nuclear Energy (Andrew Siegel), Combustion (Jackie Chen)
    - Call coming out in area of big data
  - NNSA has established a tri-lab co-design project (leads are Sriram Swaminarayan, Mike Heroux and Rob Neely)
- European Commission
  - Three projects: numerical weather modeling; CFD; Engineering application;
  - A second call has been launched
- European countries have call that will be open to international collaborations (Germany 2012-14; France 2011-13, 2014-18 to be confirmed)
- Belmont Forum call on E-Infrastructure & Data Management for Natural Sciences
  - Phase 1: 2013-14; Phase 2: 2015
  - Three goals related data accessibility, integration, analysis
- Call for exascale simulations from RFBR (Russian agency)

Issues to Consider Going Forward

- What are our models for possible models for frameworks for collaboration?
- How do we bring big data and extreme scale computing together?
  - What are the common assumptions and goals?
  - What are the sociological barriers? People working in big data and extreme scale computing often come from significantly different viewpoints
  - Big data issues may be more pervasive in their domains
  - Driver is to get the science output in a useful amount of time

Report on draft plan of international collaboration

International Collaboration on
System Software
2013/3/19

Goals and Framework for Collaboration

- Goal
  - Build upon JP-US cooperative agreements to tackle the challenge of exascale
- Focus
  - Exascale open source software that can be transitioned into production and supported by vendors
  - Software below the applications (system software)
- Framework
  - Share open source software where it is helpful
  - Collaborate on research where beneficial
  - Create and share interfaces and later standards to facilitate building exascale ecosystem

Organization

- Steering committee is organized
  - Today: USA and Japan
  - The 1st meeting at ISC, and the 2nd meeting will be in January of 2014
  - We will plan a BOF at SC13
  - Discussing collaboration with other countries
- Technical committee to be organized
  - Initial technical projects will lead selection
- Progress reports every 6 months

Early Outcomes

- Sharing testbeds of interesting hardware where possible
- Sharing Mini-apps
- Collaborate on stretch goals / capabilities that could be given to vendors in future procurements.
- Shared interfaces for low-level software components

Initial Technical Areas of Collaboration Proposed

Members are candidates

- Low-level communication layer
  - JP: Yutaka Ishikawa (U. of Tokyo/Riken), Mitsuhisa Sato (U. of Tsukuba), Atsushi Hori(Riken), Masayuki Hatanaka (Riken), Hiromichi Takagi (NEC), Shinji Sumimoto (Fujitsu)
  - US: Rajeev Thakur (ANL), Pavan Balaji (ANL), Paul Hargrove (LBNL)
- Kernel System Programming Interface
  - JP: Yutaka Ishikawa (U. of Tokyo/Riken), Balazs Gerofi‘(U. of Tokyo), Atsushi Hori (Riken), Hiromichi Takagi (NEC), Masaaki Shimizu (Hitachi), Shinji Sumimoto (Fujitsu), Masaaki Shimizu (Hitachi), Hiromichi Takagi (NEC), Shinji Sumimoto (Fujitsu)
  - US: Pete Beckman (ANL), Ron Brightwell (SNL),
- Process/Thread model for PGAS
  - JP: Atsushi Hori (Riken), Akio Shimada (Riken), Mitsuhisa Sato(U. of Tsukuba), Masaaki Shimizu (Hitachi), (NEC), Shinji Sumimoto (Fujitsu)
  - US: Pavan Balaji (ANL), Pete Beckman (ANL), Kathy Yelick (LBNL)

Initial Technical Areas of Collaboration Proposed

- File System / I/O Forwarding / Hierarchical File System
  - JP: Osamu Tatebe (U. of Tsukuba), Yutaka Ishikawa (U. of Tokyo), Atsushi Hori (Riken), Masaaki Shimizu (Hitachi), Shinji Sumimoto (Fujitsu), (NEC)
  - US: Rob Ross (ANL), Lee Ward (SNL)
- Full System Power Management / Live data / Fine-grained controls (from kernel to control system)
  - JP: Masaaki Kondo (UEC Tokyo), Hiroshi Nakamura (U. of Tokyo), (Fujitsu)
  - US: Pete Beckman (ANL), Martin Shultz (LLNL)
- Hint interface for communication libraries
  - JP: Takeshi Nanri (Kyushu U.), Shinji Sumimoto (Fujitsu), Hidetomo Shibamura (ISIT), Toshiya Takami (Kyushu Univ.)
  - US: (possibly, some person in ARMCI group @ PNNL)

Technical Areas of Collaboration Under Negotiation

- Hint interface for communication libraries
  - JP: Takeshi Nanri (Kyushu U.), Shinji Sumimoto (Fujitsu), Hidetomo Shibamura (ISIT), Toshiya Takami (Kyushu Univ.)
  - US: (possibly, some person in ARMCI group @ PNNL)

Future Plan

- Workshops at ISC'13 9:00AM – 1:00PM, 18th of June
  - 1st part: introduction of our activities
  - 2nd part: how other countries join us
- (Unofficial) Steering committee is organized at ISC'13 before BOF
  - Members: US and Japan
  - Other countries are invited at this meeting
  - Unofficial technical committee meeting is organized depending on who can come to ISC'13
- When is the official technical committee meeting ?
  - It will be decided later

Preparing HPC Codes and Software for Exascale Computing: Early results of the G8 Exascale Projects

Charleston – April, 2013

G8 Research Councils Initiative on Multilateral Research Funding

Jean-Yves Berthou
ANR – Director of the ICT dept.

G8 Research Councils Initiative on Multilateral Research Funding

- G8 Heads of Research Councils (HORCs) meeting in Kyoto, Japan, in May 2008: Proposal for a **multilateral funding activity**:
  - Multilateral research projects can address global challenges in ways that are beyond the capacity of national or bilateral activities.
  - The G8 HORCs framework provides the unique opportunity to pilot a new modality for conducting international research.
- Goals: supporting **excellent** and **interdisciplinary** research on topics of **global relevance** best tackled through a multinational approach.
- Research topics defined separately for each call.
- 7 Funding Agencies: NSERC (Canada), ANR (France), DFG (Germany), JSPS (Japan), RFBR (Russia), RCUK (UK), NSF (USA).
- Principles:
  - A common call text with selection criteria predefined together
  - A **multinational peer review** process in 2 stages (pre-proposals and full proposals)
  - Consortia consisting of partners from at least 3 of the participating countries.
  - **National funding** according to normal terms and conditions for national project funding.
  - Funding meant for **collaborative research**, not merely for networking, mobility or communication.

G8HORCS calls planned

- 2011 : *Interdisciplinary Program on Application Software towards Exascale Computing for Global Scale Issues*

- 2012 : *Interdisciplinary Programme on Material Efficiency – A First Step towards Sustainable Manufacturing*

- 2013 : *Interdisciplinary Programme on Coastal vulnerability & Fresh water security* – Belmont Forum (IGFA - International Group of Funding Agencies for Global Change Research)

What next?

- Continuation of the initiative?
- In which format?

Preparing HPC Codes and Software for Exascale Computing: Early results of the G8 Exascale Projects

G8 Exascale workshop – November 2012

1. Share the early results of these 6 projects and their impact on science
2. Explore common lessons learned in terms of Exascale research
3. Present early feedback on this innovative multinational collaborative pilot program
4. Discuss between funding agencies, PIs and the Exascale community related to the G8 tool: call, award management, cooperation between partners, consortium management

Funding Organizations

- The Natural Sciences and Engineering Research Council of Canada (NSERC)
- The French National Research Agency (ANR)
- The German Research Foundation (DFG)
- The Japan Society for the Promotion of Science (JSPS)
- The Russian Foundation for Basic Research (RFBR)
- The Research Councils of the United Kingdom (RCUK)
- The U.S. National Science Foundation (NSF)

First feedbacks : the good

Unique opportunity to enable collaborations between 3 to 6 different countries (4 projects with 5 or 6 countries).

- Makes improvement available in all these countries , usually bilateral collaborations or regional collaborations (EC, US, Asia).
- Doing international comparison is of great value, lead to some most important recent advances
- Transforming the frontiers , foster int. collaboration, speed-up the research

Mix of top world researchers

Involvement of student (10 per projects), key contributions from young scientists (unique opportunity for them), training of new generation of fusion scientists, training sessions for young scientists

Access to top world facilities (HPC computers and Data). As an example, the ECS project: access to top world HPC computers, BLueWaters (NCSA), Intrepid (ANL), K computer (Riken), Tsubame2(Titech), Jugene (PRACE), Marenostrum2 (BSC)

Interest of:

- non G8 countries: Spain joined ECS project, China joined Nu-Fuse project
- Companies: Initiation collaboration with Intel and IBM (SEISMIC IMAGING project)

First feedbacks, the bad (or the less good)

Projects : research engaged need to continue beyond the 3 years project

All : One call does not make sense, need 5 years visibility at least

All : very little amount of money

NFOs: management burden not sustainable for some agencies (Canada)

All : Need to find way to make sustainable, long term, multi-lateral collaborations

BELMONT FORUM

- World's **major and emerging funders** of global environmental change research, and international science councils
- Acting as Council of Principals for **IGFA**, a larger group of funding agencies

- **Australia/CSIRO**
- **Austria/BMWF**
- **Brazil/FAPESP**
- **Canada/NSERC**
- **China/NSFC**
- **European Commission/DG R&I**
- **France/CNRS&ANR**
- **Germany/DFG&BMBF**
- **India/MoES**
- **Japan/MEXT&JST**
- **Norway/RCN**
- **South Africa/NRF**
- **United Kingdom/NERC**
- **United States/NSF**
- ***International Council for Science (ICSU)***
- ***International Social Sciences Council (ISSC)***

BELMONT FORUM

- Initiated in 2009, by NSF (US) and NERC (UK)
- **Belmont Challenge**:

to accelerate delivery of the environmental research needed
to remove critical barriers to sustainability
by aligning and mobilizing international resources

- Convergence with other processes of collective thinking within the Global Environmental Change community

⇒ **S&T Alliance for Global Sustainability** with **Future Earth** as first major action

⇒ The BF's **International Opportunity Fund**

The International Opportunity Fund

- A yearly call with 2/3 thematic **Collaborative Research Actions** (CRAs)
- Main goals:
  - **Address the Belmont Challenge priorities** - Deliver knowledge needed for action
  - **Support Future Earth** by promoting innovative types of research
  - **Lever IGFA/Belmont Forum member's existing investments** through international added value
  - **Bring together new partnerships** of natural, geo scientists, humanities and social scientists, and stakeholders
- A flexible tool
  - ***A la carte*** for a given CRA
  - Suitable for **various kinds of incentives**: networking, clustering, integration, capacity building…
  - **Open to any funder** (BF, IGFA or not)
  - **Everything is common** (scoping, call, selection, scientific follow-up) **but the money**
  - Possibility of **joint call** with other international initiatives

IOF 2012

- CRAs on **Freshwater Security** and **Coastal Vulnerability**, joint with **G8HORCs**, aligned with FP7 and NSERC
- **130 pre-proposals,** involving more than **1000 partners** from ~ **50 countries** (high number of partners from non-BF member countries, coming on board with their own funding)
- **53 full proposals**
- **13 funded projects; 2-3-year; ~ 1-2 M€** projects

BELMONT FORUM

IOF 2013

- E-infrastructure and Data Management
- Food Security and Land Use Change

2014 CRA proposals under **scoping process**

- **Arctic** : Data Observing System and Sustainability Science
- **Biodiversity and Ecosystem Services** : Improving our Capacity to Predict Global Changes
- **Climate Services** : Seasonal to Decadal Predictability of Regional Climate (incl. monsoons and polar areas)

Others CRAs proposals to come,

including CRAs with special emphasis on HSS

Facilitating International Collaboration through New Funding Opportunities

Maria Uhle
Program Director, International Activities
GEO Directorate, NSF

e-Infrastructures & Data Management

Objective

Develop transformative concepts and approaches to create integrated data management infrastructures that can help meet the Belmont Challenge and Future Earth

Ultimate Goals

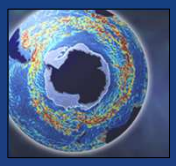
BELMONT FORUM

- Making environmental data/information more visible and accessible to a wide range of potential users
- Tools to facilitate the integrated analysis of data/information from both natural and social sciences, visualization of results
- Development of new, knowledge from public and private sector data/information assets

Initial Ideas

BELMONT FORUM

Results of a scoping workshop:



- Data standards;
- **Interoperability;**
- Data provenance;
- Semantics;
- **Architectures;**
- Security models;
- **Governance and legal issues** (e.g., intellectual property rights, privacy)

Mechanisms

- Phase I (2013 - 2014)
  - Building Communities
  - Developing a Community Strategy and Implementation Plan
  - Inform Phase II
    - 18-month **"Knowledge Hub"**
- Phase II (2015)
  - Deliver on Implementation Plan through medium to large projects

27 Feb 2013 5

Timeline

March - June 2013:

- Identify national researchers to participate in Steering Committee and program

July 2013:

- First Meetings of Steering Committee and 'All Community' meeting to design 18-month Community Building and Strategy Development Program

July 2013 – September 2014:

- Program of international community building and strategy development activities, including mapping exercises, workshops, exchanges, summer-schools

27 Feb 2013 6

Timeline

September 2014:

- Community Strategy and Implementation Plan produced, with recommendations for Phase II Belmont Forum Call to develop 'Exemplars'

January 2015:

- Belmont Forum decides on Phase II call

Where should we go from here?