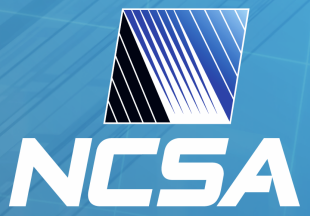
Big Data and Extreme Computing Workshop
Architecture and Operation

William Kramer
University of Illinois Urbana Champaign
National Center for Supercomputing Applications
Department of Computer Science



Architecture and Operations Working Group Participants

Bill Kramer, Ewa Deelman, Francois Bodin, Piyush Mehrotra, Marie-Christine Sawley, Giovanni Erbacci, Yutaka Ishikawa, Toshio Endo, Jean-Francois Lavignon, Pete Beckman, Osman Unsal, Jamie Kinney, Bronis de Supinski, Masaaki Kondo, Marek Michalewicz, Malcolm Muggeridge

Outline

- Current State
- Motivation
- General
- Production Services
- Architecture
- Executive Summary

- We will not be talking about architectures in detail since
  - We just had three days of great talks
  - To first order – the BD and EC architecture choices are the same in this period

Current State

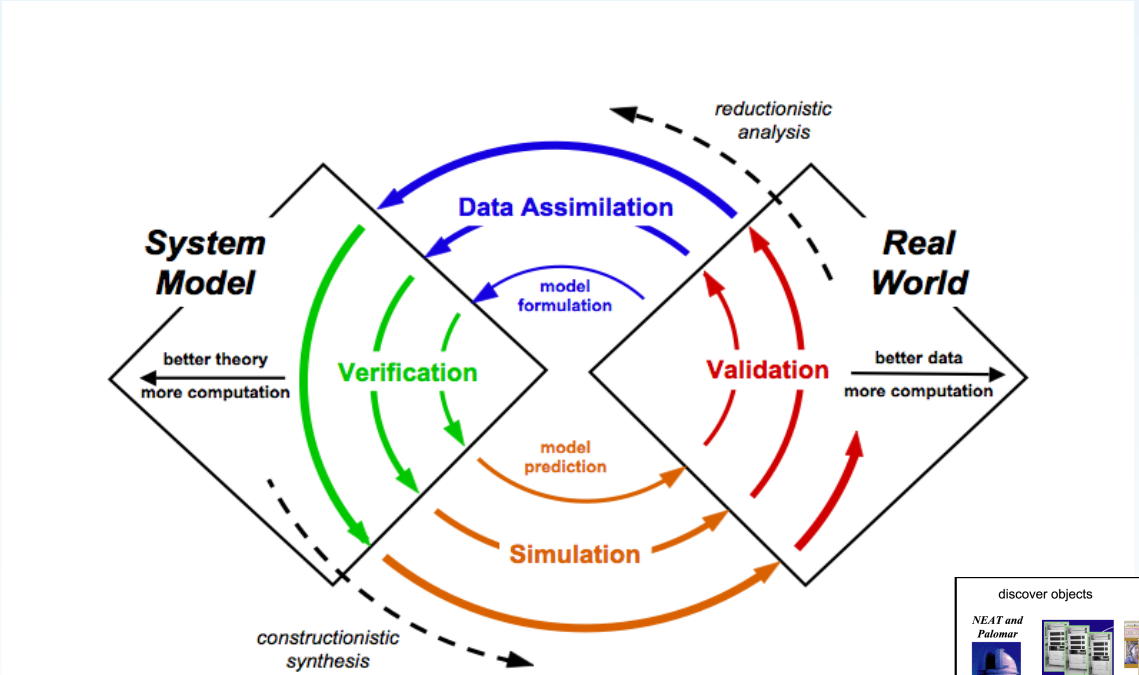
- In many ways, BD and EC have co-existed in the same facilities if not the same systems for many decades
- The perfect BD system shares many architectural attributes of the EC system
- Current Petascale EC resources have many examples of processing BD applications well
  - K machine is #1 of the Graph500 list
  - LSST analysis pipeline has been shown to run well and efficiently on Blue Waters
  - Examples of BD frameworks on EC file systems
    - *Benchmarking and Performance Studies of Mapreduce*, Hadoop Framework on Blue Waters Supercomputer, Manisha Gajbe, Kalyana Chadalavada, Gregory Bauer, William Kramer, International Conference on Advances in Big Data Analytics, July 27-30, 2015, Las Vegas, USA
  - Integrated EC and BD workflows exist
    - Ex – SCEC – Specfem3d + Cybershake – very large scale parallel phase and 100,000,000 job phase – co-exist on BW - published
  - The initial “user” member of Open Power Consortium was BD Google, and not that technology is what ORNL and LLNL will use for EC
    - Much EC technology now being used behind the scenes in cloud systems
  - EC resources in OSG
- Yet there remain significant differences between BD and EC
  - Actual and Perceived

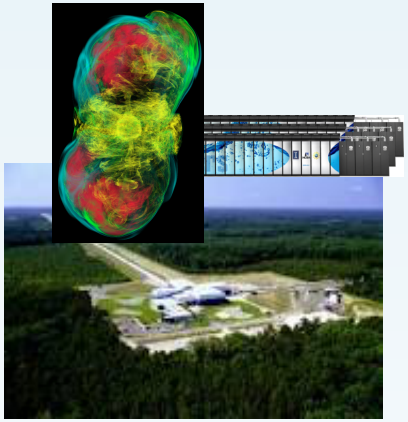
Motivation

- Systems are expensive and not integrating misses opportunities
  - Leveraging investments and purchasing power
- Integration of Computation and Observation cycles implicitly requires convergence
- Expanded cross disciplinary teams of researchers are needed to explore the most challenging problems for society
- Data Consolidation trends span BD and EC
- Understand what benefits from Convergence and what does not
  - Categorization of Data
  - Structured, Semi-structured and Unstructured Data
  - Computer Generated and Observed Data

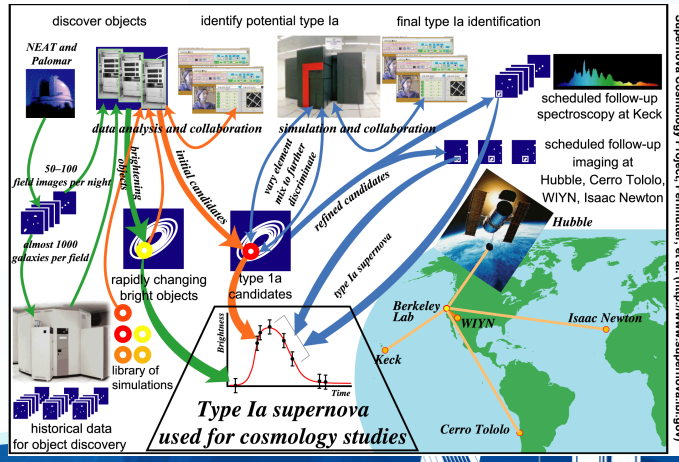
Real Motivation – Research is Changing

- Inference Spiral of System Science





- As models become more complex and new data bring in more information, we require ever increasing computational resources



Differences and commonalities between EC and BD

Difference

- **Cost and value models**
- **Project Approaches**
- Commercial software like Oracle
- Use VMs and cloud/grid methods
- Do not like waiting for resources
- Issues of data transport
- Use of shared data machines
- Aggregate I/O more important that individual IOPS
- Longer term data storage because of original data creation
- BD data sometimes unstructured
- Programming models and languages (Fortran/C vs Java)
- Parallelism and synchronization models
- Integer vs floating point performance
- Streaming data or data streaming over time Consistency and correctness models
- Data replication rather than protection
- Workflow complexity and optimization
- …..

Commonalities

- Interactivity is needed for BD as well as EC
- **Many BD methods are computationally intensive**
- **BD has much pre-computing done in the style of EC**
- **Energy Efficiency focus**
- Cost conscious focus
- **Time to solution focus**
- Use the basic underlying hardware features
- Different approaches in commercial and research
- **Both EC and BD deal with millions and billions of files per system**
  - EC is not large files
- SW Stacks
- Large amounts of data transfer
- High throughput use
- ….

NCSA

Technology game changer?

- Research that benefit both BD and EC, without need for convergence :
  - Processors that make use of 3D memory
  - high-capacity, high-bandwidth, cheap memory (what users want to see is a flat memory)
  - Accelerated computation (currently only FP)
  - high-speed interconnect (LAN and wide area)
  - high-speed to storage (bottleneck now is disk)
  - high-performance file systems
- Research to promote convergence:
  - speeds in/off the chip
  - data-aware algorithms
  - minimizing data movement, active storage
  - better use of SDN, virtualization
  - ability of software defined provisioning and management of resources
  - co-existing of VMs/containers in traditional EC systems/ Software and application stack control in EC
  - ease of validating applications in different environments
  - methods for co-location of computation and data
  - automated, optimal and easy use of deep memory hierarchies
  - on-the-fly data processing (percipient storage)
  - high-level abstractions for computation and data
  - benchmarks, application models, execution traces
  - efficient graph processing libraries

Effective Convergence of EC and BD Priorities

- New APIs and integrated execution models
- Ease of movement between cloud and extreme scale EC/BD resources
- More flexible, High-performance file/storage systems/repositories
  - Many files – not just a few large files
  - Multiple personalities/APIs?
  - Neither EC nor BD will be using strict POSIX
  - Virtualized I/O capability
  - design for a common storage capability
- Fine grained resource management
  - EC resource managers can today integrate long large jobs with modest length (O(10 minutes) small jobs – this works for thinks like structured data analysis pipelines (HEP, Genomics, …)
  - BD also has high throughput sub minute jobs that need new integration steps
  - Affinity for job steps
  - Need a way to specify the mix of resources that you need and the system would allocate them
  - Need for a malleable schedulers
  - Fast, lightweight launch
  - User Experience focus (scale vs time to complete, planned (batch) vs asynchronous immediate, …
  - Container support with low jitter
- Approaches to what to optimize

Effective Convergence of EC and BD Priorities

- Software Defined Resources
  - dynamic integration of memory and storage resources and associated API
  - runtime system that automate data movement between memory and storage
  - Network and topology configurations
  - Virtualized memory and storage systems with open APIs for transparent data movement and on-the fly processing
- Energy efficient resource management
- More collaboration between EC and BD researchers
  - benchmarks for EC and BD workloads
  - Technology discussions between research centers and cloud service providers
  - Work with a selected group of application, understand them and make them work in heterogeneous environments
    - Machine learning, Graph Analysis, etc.
  - Establish partner relationship with major “data creator” projects (LSST, SKA, next gen LIGO..)
- Understand cost models of BD and EC and use appropriately
- Allocation, Authorization and Authentication
  - Not really a firewall, etc.

Executive Summary Statements

- General:
  - We need to identify different levels of BD-EC convergence and evaluate their benefits.
    - Not all BD and EC has to, nor should converge – maybe driven by types of BD
  - We need to conduct a cost benefit analysis to determine where and how convergence would benefit the user communities and how to best prioritize the activities in a way that reflects the needs of the user community and the priorities of the funding organizations.
  - Address trans-national policies to encourage collaborations and flexible, efficient resources allocations. Exploit new synergies between countries and organizations.
- Architecture:
  - Both BD and EC will use the same architectural components (processors, memory technologies, interconnects, …) So the key issue are cost effective system balances and system software architectures.
  - More robust and dynamic methods to move the data where they are needed
  - Ensure I/O and storage technology research and productization targeting need of convergent system should receive sufficient focus and funding
- Operations:
  - We have a clear need for convergence of resource allocation and management mechanisms and services (that accommodate both styles of applications).
  - Set up a repository of reference components and workflow systems useful for EC
  - and BD
    applications.
  - Encourage resource providers to adopt a user-centric model that includes support for convergent BD/HPC applications.

NCSA

Bill's Additional Thoughts

- The basic technology building blocks will be the same for BD and EC systems
  - That does not imply the cost of the building blocks should be the same for BD and EC
  - Hence, the major challenges will be
    - System balances
    - System Software
    - How resources are managed
- It should not be a goal that all BD and all EC needs to use the same systems and services
- Many BD and EC uses are on same resources
- Really good low hanging fruit – workflow optimizations
  - Many inefficient components that are connected (often via files)
- Inconsistent funding models for large scale BD creators and large scale EC resources
- Reward metrics need to be similar for BD and EC
  - Methods developers
- Cultural and work methods pose at least as big a challenge as technical architecture challenges
- Need new resource management benchmarks – not just discrete application benchmarks

Bill's Additional Thoughts

- Architectural decisions are always mutli-variate optimization choices
  - Often expressed via benchmarks and Best Value
- Need to impedance match scale and complexity
- Both BD and EC need not data repository solutions
  - Lustre and GPFS – 15 years and still being made to work
  - Google changes their data repository every 2 years
- Convergence of EC and BD could take several paths
  1. Opportunistic use of resources and services
  2. An overall optimization across all both communities
  3. Sub-select areas of BD and/or EC that can efficiently leverage co-optimized architectures.
  - Examples:
    - BD/EC resources can exist with structured and semi-structure BD that need computational intensive analysis
    - May be reasonable to EC/BD cover HTC for O(10 minutes), but not O(10 second) sessions can easily co-exist
- Technology risks are across both EC and BD