Increasing Coherence Between Simulation and Data Analytics

Chesapeake Large Scale Data Analytics Conference
Annapolis, MD
October 25, 2016

Rob Leland
Vice President, Science & Technology
Chief Technology Officer
Sandia National Laboratories
Outline

- A tale of two visions
- Some background
- A charge from the National Strategic Computing Initiative
- Answers to three key questions
  - Why is an increasing coherence between simulation and analytics important?
  - What is really meant by “increasing coherence” between the two?
  - How might coherence be furthered in practice?
- A unifying vision
Vision 1: From a scientific perspective

Data analysis complements theory, experiment, and computation

From *The Fourth Paradigm: Data-Intensive Scientific Discovery* by Jim Gray
Vision 2: From a national security perspective

Graph matching example of data analytics

A key analytic primitive -- used to find a specific instance of an abstract pattern of interest

Some background

- **Simulation**
  - Computations to understand physical phenomena or conduct engineering

- **Large Scale Data Analytics (LSDA)**
  - Data Analytics = Discovering meaningful patterns in data
  - Large Scale = Requiring leading-edge processing and storage capabilities

- **LSDA is increasing in importance**
  - Pervasive
    - Commerce, finance, health care, science, engineering, national security, ...
  - Lasting societal significance
    - Internet search, genomics, climate modeling, Higgs particle, ...

- **LSDA is getting “harder”**
  - Captured data growing exponentially with time
  - Individual analysis becoming more sophisticated
  - More people examining more data more frequently
  - Aggregate work growing much faster than Moore’s Law

**The Economist:**
Executive Order 13782 of July 29, 2015
Creating a National Strategic Computing Initiative

On the authority vested in me as President by the Constitution and the laws of the United States of America, and as Commander in Chief of the Armed Forces of the United States, and by virtue of the authority vested in me, I hereby order:

Section 1. Policy. In order to maximize the benefits of HPC for the scientific community and to enable scientific and policy discoveries that are critical to national security and economic prosperity, the National Strategic Computing Initiative (NSCI) is created. NSCI is a collaborative effort among Department of Energy (DOE) laboratories, the National Science Foundation, and other agencies to advance the use of high-performance computing and data science technologies to support scientific discovery.

Section 2. Objectives. NSCI will:

(a) Accelerate the development and adoption of next-generation HPC systems and technologies;
(b) Support a diverse set of scientific disciplines and applications;
(c) Enhance collaboration among scientists, engineers, and technical leaders;
(d) Foster the development of new software tools and applications;
(e) Improve the training and education of the scientific workforce;
(f) Strengthen international cooperation in HPC.

Section 3. Implementation. The Director of the Office of Science (OAS) shall:

(a) Establish a National Strategic Computing Initiative Director to lead the initiative;
(b) Develop a multi-year plan to implement NSCI;
(c) Establish partnerships with industry, academia, and other federal agencies;
(d) Provide funding for research and development.

This order shall be implemented in accordance with applicable laws and regulations.

Barack Obama, President
July 29, 2015
NSCI Strategic Objectives

(1) Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.

(2) Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.

(3) Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").

(4) Increasing the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.

(5) Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.
Q1: Why is increasing coherence between simulation and analytics important?

- **For simulation**
  - HPC simulation must ride on some commodity curve
  - Larger market forces behind analytics
  - Can exploit commodity component technology from analytics

- **For analytics**
  - Large Scale Data Analytics problems becoming ever more sophisticated
  - Requiring more coupled methods
  - Can exploit architectural lessons from HPC simulation

- **For both: Integration of simulation and analytics in the same workflow**
  - Automation of analysis of data from simulation
  - Creation of synthetic data via simulation to augment analysis
  - Automated generation and testing of hypothesis
  - Exploration of new scientific and technical scenarios
  - …

**Mutual inspiration, technical synergy, and economies of scale in the creation, deployment, and use of HPC resources**
A challenge because simulation and analytics differ in many respects ...
Data structures describing simulation and analytics differ.

*Graphs from simulations may be irregular, but have more locality than those derived from analytics.*

**Computational Simulation of physical phenomena:**

- Climate modeling
- Car crash

**Large Scale Data Analytics:**

- Internet connectivity
- Yeast protein interactions

*Figures from Leland et. al. courtesy of Yelick, LBNL.*
Computation and communication patterns differ

The U.S. roadmap, which has spatial locality and is thus most similar of the three in structure to computational patterns that would arise in typical physical simulations.

The *Erdős-Rényi* graph, a well-studied example in graph theory work.

A scale-free graph, an example more reflective of real-world networks.

*Figure from Leland et. al. courtesy of Johnson, PNNL.*
Memory performance demands differ
A key differentiator in the performance of simulation and analytics

Standard benchmarks include:
• LINPACK (smallest data intensiveness; barely visible on graph)
• STREAM
• SPEC FP
• Spec Int

Area of the circle = relative data intensiveness (i.e. total amount of unique data accessed over a fixed interval of instructions)

Figure from Murphy & Kogge with adjustment to double radius of Linpack data point to make it visible.
Application code characteristics differ

Contrasting properties:

<table>
<thead>
<tr>
<th>Application code property</th>
<th>Simulation</th>
<th>Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial locality</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Temporal locality</td>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td>Memory footprint</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Computation type</td>
<td>May be floating-point dominated*</td>
<td>Integer intensive</td>
</tr>
<tr>
<td>Input-output orientation</td>
<td>Output dominated</td>
<td>Input dominated</td>
</tr>
</tbody>
</table>

* Increasingly, simulation work has become less floating-point dominated
Q2: So what do we really mean by “increasing coherence” between simulation and analytics?

- NOT one system ostensibly optimized for both simulation and analytics
- Greater commonality in underlying componentry and design principles
- Greater interoperability, allowing interleaving of both types of computations

... A more common hardware and software roadmap between simulation and analytics
And yet, there is hope ...
Simulation and analytics are evolving to become more similar in their architectural needs

- **Current challenges for the LSDA community**
  - Data movement
  - Power consumption
  - Memory/interconnect bandwidth
  - Scaling efficiency

- **Instruction mix for Sandia’s HPC engineering codes**
  - Memory operations 40%
  - Integer operations 40%
  - Floating point 10%
  - Other 10%

- **Common design impacts of energy cost trends**
  - Increased concurrency (processing threads, cores, memory depth)
  - Increased complexity and burden on
    - system software, languages, tools, runtime support, codes
Energy cost of moving data is becoming dominant

Energy cost for various common operations

Cost estimates for technology year
- 2012
- 2020

Emerging architectural and system software synergies

**Similar needs:**

<table>
<thead>
<tr>
<th>Architectural Characteristic</th>
<th>Simulation</th>
<th>Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation</td>
<td>Memory address generation dominated</td>
<td>Same</td>
</tr>
<tr>
<td>Primary memory</td>
<td>Low power, high bandwidth, semi-random access</td>
<td>Same</td>
</tr>
<tr>
<td>Secondary memory</td>
<td>Emerging technologies may offset cost, allowing much more memory</td>
<td>... require extremely large memory spaces</td>
</tr>
<tr>
<td>Storage</td>
<td>Integration of another layer of memory hierarchy to support checkpoint/restart</td>
<td>... to support out-of-core data set access</td>
</tr>
<tr>
<td>Interconnect technology</td>
<td>High bisection bandwidth, (for relatively coarse-grained access)</td>
<td>... (for fine-grained access)</td>
</tr>
<tr>
<td>System software (node-level)</td>
<td>Low dependence on system services, increasingly adaptive, resource management for structured parallelism</td>
<td>... highly adaptive, resource management for unstructured parallelism</td>
</tr>
<tr>
<td>System software (system-level)</td>
<td>Increasingly irregular workflows</td>
<td>Irregular workflows</td>
</tr>
</tbody>
</table>
Q3: How might coherence be furthered in practice?

- Making it an element of national strategy
  - Check via the NSCI
- Building this in to exascale computing efforts
  - Also a component of the NSCI
- Communicating with and enlisting the technical communities concerned
  - This forum and similar events
- Further developing the vision
  - Today’s dialogue session!
Acknowledgements
Additional references


- Jim Gray, *The Fourth Paradigm: Data-Intensive Scientific Discovery*