Expand Topical Areas Covered by UC-HiPACC

Mike Norman

San Diego Supercomputer Center

UC San Diego

Big Data Hijacks Exa-Scale Computing

BIG DATA *AND* EXTREME-SCALE COMPUTING

Introduction

In the past three years, the United States, the European Union, and Japan have each moved aggressively to develop their own plans for achieving exascale computing in the next decade. Such concerted planning by the traditional leaders of HPC speaks eloquently about both the substantial rewards that await the success of such efforts, and about the unprecedented technical obstacles that apparently block the path upward to get there. But while these exascale initiatives have understandably focused on the big challenges of exascale for hardware and software architecture, the relatively recent emergence of the phenomena of Big Data in a wide variety of scientific fields represents a tectonic shift that is transforming the entire research landscape on which all plans for exascale computing must play out. The workshop on **Big Data and Extreme-scale Computing** (**BDEC**) is premised on the idea that we must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change the national (and international) plans that are now being laid for achieving exascale computing.

Department of Energy Exascale Strategy

Report to Congress
June 2013

United States Department of Energy
Washington, DC 20585

leading in high-performance computing and large-scale data analysis for the long term will have a competitive advantage in a wide array of sectors, including basic science, national defense, advanced manufacturing, energy, health care, space, transportation, education, and information technology. Exascale (i.e., order of 10^{18} operations per second and order of 10^{18} bytes of storage) is the next stage of development in supercomputing, extending capability beyond today's petascale computers, to address the next generation of scientific, engineering, and large-data problems. This will include increases in computational capacity, memory capacity, and data storage. Measured against today's most advance multi-petascale computers, which deliver 10-20 PetaFlops (PF), exascale represents a 50- to 100-fold increase in computational power.

The plan presented in this report covers the research and development required to achieve exascale computing by early in the next decade, approximately 2022. This includes support for design activities in support of technologies needed for exascale computing, as well as engineering associated for integration of those technologies in exascale computers. Acquisition of computers beyond present-day petascale systems leading to exascale would be funded separately from what is presented in this plan, using the current planning and budgeting approach of acquiring systems at intermediate stages of performance to meet increasing demands culminating in exascale (e.g., 100 PF, 250 PF, 500 PF intermediate stages). The plan also calls for some investment in state-of-the-art next-generation technologies for which the payoff would be high but the risk is too great to be borne by industry.

The Relationship between Exascale and Large Data

As noted in the Advanced Scientific Computing Research Advisory Committee (ASCAC) report entitled "Synergistic Challenges in Data-Intensive Science and Exascale Computing" presented in March 2013, large data and exascale computing are inextricably linked. Exascale processing is expected to be essential for handling datasets envisioned in the future. DOE scientific experiments, for example the particle-physics experiments DOE conducts at the Large Hadron Collider, produce some of the largest datasets in use today. Similarly, large-scale climate and engineering simulations produce many petabytes of output today and are projected to grow to exabyte scale as computational resolution is increased in the future. Manipulation of these datasets will require exascale processing capacity. As currently planned, DOE exascale computers will be designed to process the very large datasets derived from anticipated computational, experimental, and observational sources. Although DOE research focuses on numerical simulations and analyses based on floating-point numbers, DOE is aware of and works with other U.S. Government organizations whose computational foci are based on non-floating-point computations.

Make **Data Science** and Equal Partner to **Computational Science**

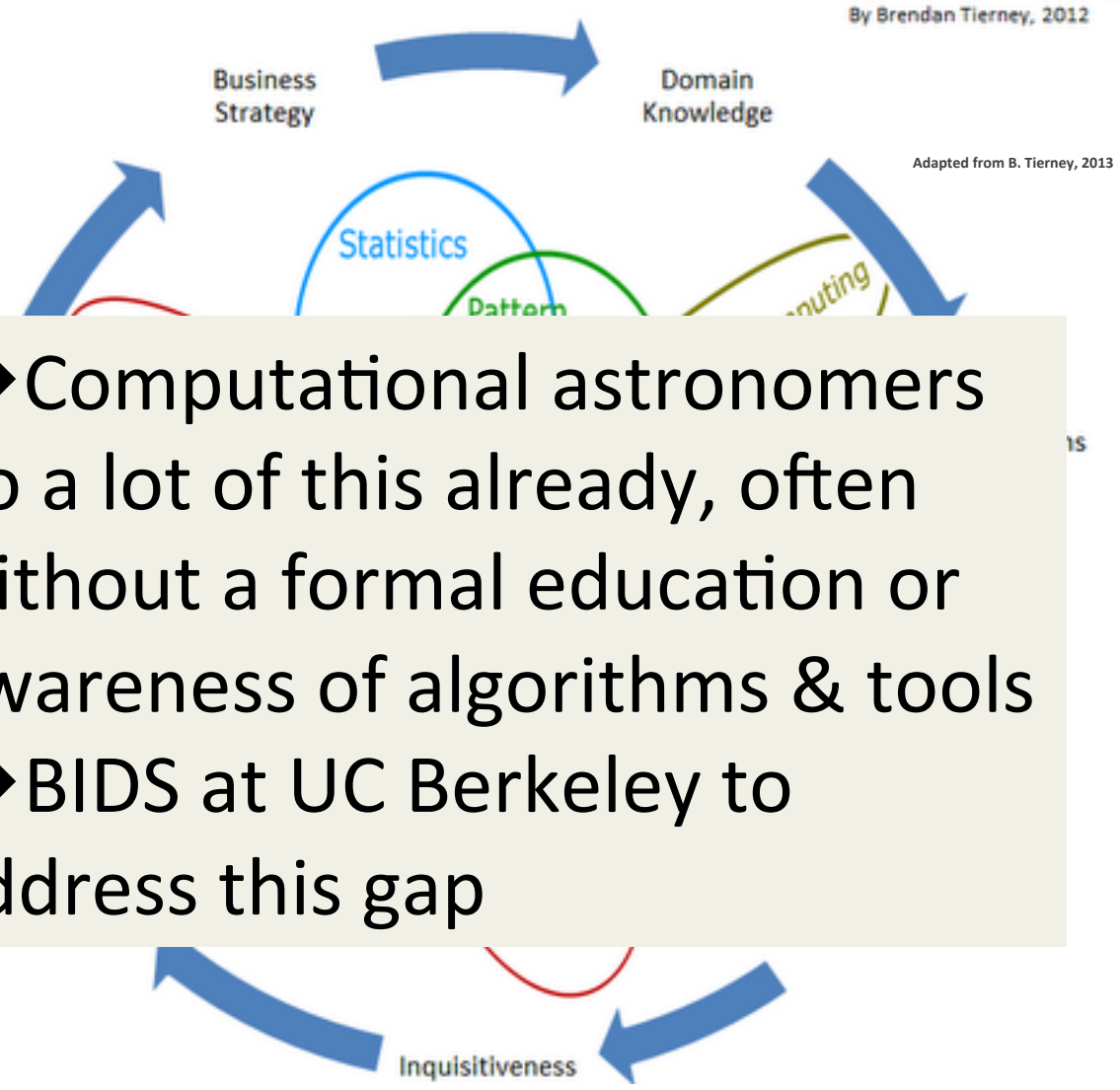
Era	Buzzword	Academic Discipline
ca. 1980-present	Supercomputing	Computational Science
ca. 2010-present	Big Data	Data Science

- Computational Science (and Engineering) programs sprung up by the hundreds in Universities in the 1990s
- Data Science (and Engineering) programs are just now springing up (e.g., Berkeley, UCSD)

Make **Data Science** and Equal Partner to **Computational Science**

- Observations
  - Supercomputers just another data source
  - Difficulty is in data analysis and data management
  - Big observing/experiment projects plan/budget-in DA/DM
  - NASA, DOE have a long legacy of practice; NSF just getting started

Data Science Is Multidisciplinary



➔Computational astronomers do a lot of this already, often without a formal education or awareness of algorithms & tools

➔BIDS at UC Berkeley to address this gap

What others are doing

- UCSD/SDSC
  - Institute for Data Science and Engineering (IDSE)
  - Collection of short courses focusing on big data systems, algorithms, tools, and applications
- LBNL & LLNL
  - Strengthening efforts in high-performance data science algorithms and architectures

Institute for Data Science and Engineering

Chaitan Baru

Associate Director, Data Initiatives
Director, Center for Large-Scale Data Systems Research
San Diego Supercomputer Center
University of California San Diego

SDSC UCSD
SAN DIEGO SUPERCOMPUTER CENTER

CLDS
Center for Large-scale Data Systems Research

Institute for Data Science and Engineering Principles

- Access to broad Data Science and Engineering curriculum
- Provide practical skills through hands-on exercises
- Offer flexible formats
  - Short, focused modules
  - At UCSD or at your location
  - Customizable content
- Objective, vendor- and technology-agnostic approach

Data Science and Engineering: Key Elements

- Understanding the Data Lifecycle
- Concepts and software for managing data
- Computing with data
- Extracting information from data
- Making decisions with data
- Technical architectures

Some Curriculum Themes

- Data management
  - SQL/noSQL, Structured/unstructured, Graph data, Distributed vs parallel
- Processing Models and Workflows
  - Parallel, distributed computing models
  - Dynamic scripts vs stable workflows
- Data Mining, Machine Learning, and Statistics
- Parallel Computing
- Visualization
  - Data vs Info Viz; Viz for Big Data
- Data Integration
  - Data fusion, ontologies, resolving terms and entities

Computation Overview and Update

Computation External Review Committee

February 11−13, 2014

Dona L. Crawford, Associate Director

Lawrence Livermore National Laboratory

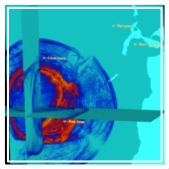
LLNL-PRES-649572

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Computation has leadership capabilities and frontier research in three areas of excellence

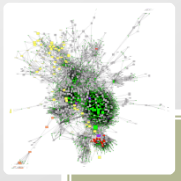
High Performance Computing

- HPC is the ecosystem associated with computer systems with massive numbers of multicore processors connected via high-speed interconnects running specialized system software to enable parallelism.



Computational Mathematics

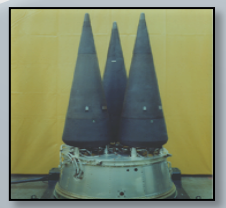
- Computational mathematics is the development and analysis of new models, algorithms, and software for predictive simulation of physical phenomenon.



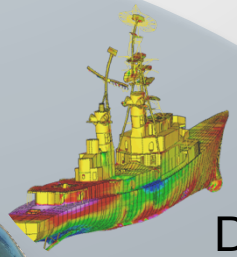
Data Science

- Data science is the development of new models and tools for data-driven discrete-event simulations, new architectures for data intensive computing, and pattern discovery through machine learning and graph analytics.

These areas are inextricably linked as building blocks for LLNL's modeling, simulation, and analytics capabilities

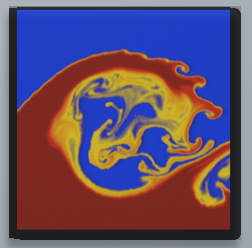


Stockpile Stewardship

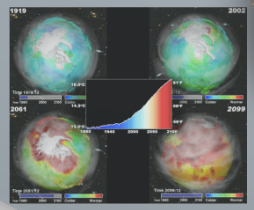


Defense

Cyber Security



Basic Science



Climate

Energy

Recommendation

- HiPACC leverage and contribute to burgeoning efforts in data science and engineering at
  - BIDS @ UC Berkeley
  - LBNL, LLNL, LANL
  - NASA Ames
  - SDSC/UCSD
  - others in UC?
- Strong relevance to Research and Education