Trends in Computational and Data-enabled Science and Need for Coordinated e-Infrastructure

Edward Seidel
Director, National Center for Supercomputing Applications
Founder Professor of Physics, Professor of Astronomy
University of Illinois at Urbana-Champaign
Collaborations Grow as Problems Become More Complex; Systems and Data Sizes, too!

TRENDS IN SCIENCE
Precisely 100 years ago, Einstein predicted gravitational waves. Feb 11, 2016 their historic discovery was announced!

Three tech trends made this possible:
• Instrumentation
• HPC
• Data Science
They are intricately connected!
2 Black Hole Collision: Changing Culture of Science

1600-1972: Galileo - Hawking.
1 person, no computer, 50 KB

1994: 10 people, NCSA Cray Y-MP, 50MB

1998: 15 people, NCSA Origin, 50GB
Community Einstein Toolkit

“Einstein Toolkit: open software for astrophysics to enable new science, facilitate interdisciplinary research and use emerging petascale computers and advanced CI.”

• Consortium: 94 members, 47 sites, 15 countries

Simulation: Luciano Rezzolla, Max Planck Institut für Gravitationsphysik (AEI)
LIGO Scientific Collaboration

- 1004 Authors on the breakthrough paper...
- How do they collaborate?
- They have a data sharing agreement and services to support them
Gravitational Wave Astrophysics

Interplay between Simulation and Experiment

Models & Simulation

Theory

Scientific Discovery!

\[ G_{\mu\nu} = 8\pi T_{\mu\nu} \]

Compact binaries, supernovae collapse, gamma-ray bursts, oscillating NSs, gravitational waves, …

News Flash!

Sept 14, 2016,
Announced Last Month:
Gravitational waves and black holes discovered!
“...that simple chirp, which rose to the note of middle C before abruptly stopping, seems destined to take its place among the great sound bites of science, ranking with Alexander Graham Bell’s “Mr. Watson — come here” and Sputnik’s first beeps from orbit.” NY Times
But HPC is only part of greater ecosystem!

BREAKTHROUGH HPC SYSTEMS SUPPORT BREAKTHROUGH SCIENCE
Blue Waters Computing System

Cray XE6/XK7 - 276 Cabinets

XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core

XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores - 11,354,112 cuda cores
– 4,224 K20X GPUs, 4 GB per FP core

Blue Waters Computing System

Aggregate Memory – 1.6 PB
13.2 Peak PF

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

Spectra Logic: 300 usable PB

100-300 Gbps WAN

Soneion: 26 usable PB

Cray XE6/XK7 - 276 Cabinets

XE6 Compute Nodes - 5,688 Blades – 22,640 Nodes – 362,240 FP (bulldozer) Cores – 724,480 Integer Cores
4 GB per FP core

XK7 GPU Nodes
1,056 Blades – 4,224 Nodes
33,792 FP Cores - 11,354,112 cuda cores
– 4,224 K20X GPUs, 4 GB per FP core

Blue Waters Computing System

Aggregate Memory – 1.6 PB
13.2 Peak PF

10/40/100 Gb Ethernet Switch

External Servers

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

Spectra Logic: 300 usable PB

100-300 Gbps WAN

Soneion: 26 usable PB
100M atoms! AIDS Virus, Chromatophore
Schulten, UIUC
“When I die and go to heaven there are two matters on which I hope for enlightenment. One is quantum electrodynamics, and the other is the turbulent motion of fluids. And about the former I am rather optimistic.” Horace Lamb, 1932…

- Turbulent flow via direct numerical simulation (DNS) just possible on modern Petascale systems
  - 100PF+ would allow complete engine simulation with turbulence
  - 40 Exaflop would allow complete airplane simulation!
Real World Science and Engineering Require Integration of Compute and Data

FUTURE DIRECTIONS
The Growth of Data

“I’m still here...”

But I’m your new baby big brother...

Data Tsunami

With millions of processors...
Data-enabled Transformation of Science

How can I publish, discover, verify data in this new world?

Astronomy 1500-2000:
- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility

Sloan Digital Sky Survey 2000+
- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”

DES (now)
- 200GB/night
- PB in decade

LSST (6 years)
- Record data for decade
- SDSS/night!
- 200 PB/decade
Integrating it all: Multi-Messenger Astronomy…

Astronomy c. 2020!

- New era: seeing events as they occur
- Here now
  - ALMA, EVLA
  - Ice Cube neutrinos
- On horizon
  - 20-30m optical?
  - LSST
  - aLIGO, Indigo
- SKA = exabytes
- Simulations integrate physics

Communities communicate by sharing data, software… knowledge…
Scenarios like this in all fields...
National Academy Reports…

“Convergence [is] a culture shift for academic organizations that are traditionally organized around discipline-based departments….”

“…most difficult problems do not respect disciplinary boundaries…convergent science, [integrates] insights and approaches from many fields…barriers to convergent science…

We need e-Infrastructures that better support convergence…. 
Modern Research & Education Ecosystem

Blue Waters

Key challenge will be to create a deeply integrated compute and data environment that supports complex problem solving for academia and industry...

Education: I need all of this to start to solve my problem!
Fundamental Problems with this

- It is piecemeal…
- It is not integrated…
- It’s chaotic: Pieces come and go…
- Big instrument projects have
  - Can’t depend on anyone
  - Therefore they “roll their own”:
    - $100s of millions each
- It’s expensive: need about ~$1B…
- But with current model
  - We are (presumably) wasting money…
  - It is harder to support the way science is moving, towards convergence…

What is needed are long term planning horizons before construction! Can this be done?
Participants

Steering Committee

The efforts of developing this concept are guided by a steering committee consisting of:

**Dan Atkins**
W.K. Kellogg Professor of Community Information
Professor of Electrical Engineering and Computer Science
College of Engineering, University of Michigan

**Alan Blatecky**
Interdisciplinary Research and Education

DATA AS ENABLER FOR COLLABORATIVE WORK
Materials Innovation
Grand Challenge Communities

• Combining approaches in a digital world
  • Theory and computation
  • Instrumentation
  • Data and informatics

• Cyberinfrastructure
  • Software centers
  • Data services + Instruments
  • Computing

• Policy
  • Open data will accelerate discovery, enhance interdisciplinarity, speed innovation, commercialization
“Long Tail” vs Big Data: Advanced Photon Source Upgrade

Highly integrated computing/data services at ANL

Curves for APS, ESRF and SP8 upgrades based on present designs, assuming identical undulators

Different from LHC: serve many disciplines, require highly integrated computing facility “nearby”…

Thanks to Ian Foster
We aim to make it simple for materials datasets and resources to be ...

Published
- Identified
- Described
- Curated
- Verifiable
- Accessible
- Preserved

and

Discovered
- Searched
- Browsed
- Shared
- Recommended
- Accessed

Source: Ian Foster
Building Grand Challenge Communities around Data

NSF BIG DATA HUBS
Alaska & Hawaii are part of the West region. US Territories can participate in any region.

**MIDWEST**
- 106 Personnel
- 79 Organizations
- 12 states
- UIUC/NCSA (PI)
- Indiana U (co-PI)
- Iowa State (co-PI)
- U of M (co-PI)
- UND (co-PI)

**NORTHEAST**
- 193 Personnel
- 99 Institutions
- 9 States
- UCSD/SDSC (PI)
- Berkeley (PI)
- UW (PI)
- Georgia Tech (PI)
- UNC/RENCI (PI)
- Columbia (PI)

**WEST**
- 86 Personnel
- 47 Organizations
- 13 States
- UCSD/SDSC (PI)
- UCSD/SDSC (PI)
- UCSD/SDSC (PI)

**SOUTH***
- 116 Personnel
- 95 Organizations
- 15 States + DC
- UCSD/SDSC (PI)
- UCSD/SDSC (PI)
- UCSD/SDSC (PI)

BD Hubs

Points indicate affiliations of individuals named as steering council members and/or task leads.

*South points indicate Senior Personnel.
Researchers, educators, students communicate by sharing data…this is central to enabling everything above! Services needed to make it work!
Summary

• Major paradigm and culture shifts: research is changing dramatically
  • Complex problems require collaborations at new scales
  • Computing and data capabilities growing at an unprecedented rate
• Major infrastructure projects are highly computing & data intensive
  • Data services are needed to support
  • A comprehensive, integrated approach would better serve science, reduce costs
• Data sharing supports int’l cooperation