A Superfacility Model for Science

Kathy Yelick

Associate Laboratory Director for Computing Sciences
Lawrence Berkeley National Laboratory
Professor of Electrical Engineering and Computer Sciences
University of California at Berkeley
Science is poised for transformation
Old School Scientists: The Lone Scientist
Team Science

Lawrence introduces big team science 1931

LBNL the first National Lab
17-year-old Brittany Wegner creates breast cancer detection tool that is 99% accurate on a minimally invasive, previously inaccurate test.

Machine Learning + Online Data + Cloud Computing
Old School Scientific Workflow
Computing, experiments, networking and expertise in a “Superfacility” for Science

Slot-die printing of Organic photovoltaics

Data Growth is Outpacing Computing Growth

Graph based on average growth

Projected Data Rates Relative to 2010
HPC: It’s not just for simulation

Experimentation

Theory

Data Analysis

Simulation

Computing
HPC: It’s not just for simulation

Experimentation

Theory

Growth in Sequencers, CCDs, sensors, etc.

Data Analysis

Simulation

Computing
Integration of Simulation and Observational Science


Simulations aid in interpreting data

Intermediate Palomar Transient Factory with DESI, CMB-S4 and LSST coming

Image subtraction, machine learning in minutes
Real-Time MRI Challenge

3 min goal (1 sec/iteration)
Michael Driscoll HPC optimization

Compressed Sensing Approach by Mike Lustig et al
MRI results Wenwen Jiang
Old School Scientific Data Search
Automated Search, Meta-Data Analysis, and On-Demand Simulation

Challenge 2: Search needs to account for scale and lineage of data and the I/O challenges of future systems.

Challenge 3: The complexities and intricacies of scientific data, as well as, machine and deep learning algorithms require a careful consideration of the human factors. Machine learning techniques can help with learning about the data and generating metadata. However, this is not sufficient for scientific data, since the complexity of the data often requires specialized domain knowledge and understanding. Automated metadata generated from machine learning algorithms will likely need to be curated by humans to ensure accuracy. Additionally, the machine learning model needs to understand the terms or signals that might arise from a user’s query. Thus, it is important to understand how people interact and want to interact with scientific data search and machine-generated metadata labels.

Project Objectives: Designing a data integration ecosystem.

Our proposed techniques bring together a unique blend of skills that includes machine learning, human-computer interaction, and experience with scientific domains and users at facilities. Our goal is to make data a first-class discoverable resource at supercomputing centers through the powerful concept of search.

Figure 1 shows the conceptual system architecture that will be enabled by the research proposed in this proposal. The ScienceSearch framework has three key components: a) metadata generation, b) the Ground metadata storage framework, and c) an interface layer. The metadata generation framework uses a variety of machine learning techniques to generate the context of the data from both application data, as well as system level information. Ground is a data context service that provides the metadata storage layer. The interface layer allows the users to interact with the system to verify and validate automated metadata generated.

We envision the ScienceSearch framework will be available at supercomputing centers and users can make their data available to the system. The ScienceSearch framework will use the data sets and, ecosystem artifacts associated with the data (e.g., proposals, workflow and system logs, publications) to learn and generate metadata labels. The ScienceSearch framework will use active learning to surface the metadata labels to users for feedback. The users can validate, add, delete or edit labels. Similarly, we anticipate that...

Jobs submitted by “bots” based on queries; algorithms extract informatics for design
Computing and Networking
Facilities need to adapt
ESnet: Exponential data growth drives capacity

Science DMZ to deliver bandwidth to the end users
OSCARS for bandwidth reservation

100 Exabytes/year by 2024!
Bringing the Computer to the Experiment

LCLS/NERSC/Esnet Superfacility demo for Photosystem II

3x increase in ESnet load
Instruments and facilities require high-speed data network architectures like ScienceDMZ.
End-to-end, multi-domain network orchestration

Network is a ‘first class’ resource for a Superfacility workflow

Authen.ca on and identity management across domains

SDN for end-to-end Network @ Exascale (SENSE) project led by Monga @ESnet with ANL, Caltech, FNAL, NERSC, MAX/UMD
ESnet6 plans for superfacility support

“Hollow” Core optimized for performance
- **Programmable** to allocate bandwidth and monitor status
- **Scalable** – Leverage latest technology (e.g. FlexGrid spectral partitioning, tunable wave modulation)
- **Resilient** – Protection and restoration functions using next generation Traffic Engineering (TE) protocols

Services Edge optimized for flexibility
- **Programmable** to manage edge router/switch and retrieve telemetry information
- **Flexible** programmable switches (e.g. FPGA, NPU)
- **Dynamic** instantiation of services driven by SDN paradigms (e.g., virtualization, service chaining).
NERSC Cori has data partition (Haswell) and pre-exascale (KNL) NVRAM file system with close to 2 PB at 2 TB/sec
WAN-to-Cori optimized for streaming data: 100x faster from LCLS to Cori and Globus to CERN
Containers for HPC Systems

- Data analysis pipelines are often large, complex software stacks
- NERSC Shifter (with Cray), supports containers for HPC systems
- Used in HEP and NP projects (ATLAS, ALICE, STAR, LSST, DESI)

![Startup Time Graph](image-url)
Real-time queue prototyped at NERSC

- In 1998 dedicated hardware; now prototype queue on Cori
- <1% of NERSC allocation
- Cryo-Em, Mass spec, Telescopes, Accelerator, Light sources

Cryo-EM: Image classification
Nogales Lab

PTF: Image subtraction pipeline

ALS: 3D Reconstruction, rendered on SPOT web portal
Part 3

Research challenges are substantial
Software implementations at scale in pipeline

MicroCT imaging → Segmentation → Topological Analysis

Analysis ↪ Simulation ↪ Visualization
Interactive Analytics using Jupyter

Science notebooks through Jupyter (iPython)
- Widely used in science
- Interactive HPC LDRD

Deployed at NERSC:
- >100 users pre-production

Fernando Perez et al
Productive Programming

Spark

Speed
Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

• High failure rate
• Slow network
• Fast (local) disk

And Spark is still 10x+ slower than MPI
**SPARK Analytics on HPC**

**Weak scaling (fixed problem size per node): Power Iteration Benchmark**

- **Time (s)**
  - 120
  - 100
  - 80
  - 60
  - 40
  - 20
  - 0

- **Cluster**: 1, 2, 4, 8, 16
  - App
  - Fetch
  - JVM

- **Cluster with RDMA**: 1, 2, 4, 8, 16
  - App
  - Fetch
  - JVM

- **HPC**: 1, 2, 4, 8, 16
  - App
  - Fetch
  - JVM

**SPARK on HPC vs. clusters**

- Network, I/O, and virtualization all key to performance
- Increased scale from $O(100)$ to $O(10,000)$ cores

*Chaimov, Malony, Iancu, Ibrahim, Canon, Srinivasan*
CAMERA: Math for the Facilities

- Designing mathematical algorithms to allow real-time analysis next to the equipment
- New algorithms to transform manual into automatic analysis
- Inventing new math and models to match new acquisition technologies
- Robust and reliable codes and data flow: workflow environments
- Cultural and Sociological Challenges
- Multi-modal: Building the math that fuses information from multiple experiments
- Compare and integrate multiple analysis tools
- Real-time streaming ptychography—ALS, delivered to NSLS2, LANL, BESSY,
- Automatic image processing for the ALS/GE
- Fluctuation scattering and single particle imaging for the LCLS
- Workflow and access to remote supercomputers: XiCAM for ALS, SSRL, APS, NSLS2

CAMERA workshop on Tomography: Joint with APS, ESRF, DIAMOND, LNLS, LLNL, SSRL,...,
Automated detection and analysis of particle beams in laser-plasma accelerator simulations.

Fig. 5. Comparison of particle selection with/without MVEE: extracting the orientation and the axes of an enclosing ellipse from (a) produces (b), increasing the number of particles from 173 to 263. Colors indicate the density of particles, using only \((x, y)\)-coordinates, and black dots show potential particles to belong to the beam, according to the different methods.

During the searching for values that are approximately equal to \(\text{max}(f)\), we keep not only the maximum, but all bins where \(f \geq u \ast \text{max}(f)\), where \(u\) is an uncertainty or tolerance parameter, here empirically set to 0.85. While this value enables the detection of the main and the secondary beams (when present), lower values of \(u\) could be used to control the amount of particles to be selected at a lower accuracy of beam position. From this point, we refer to the subset of particles conditioned to \(u \ast \text{max}(f)\) and its adjacency, calculated for each time step, as "beam candidates".

Figure 4 (top) presents projections of Figure 3.b with their calculated beam candidates emphasized in red. These are the result of our first attempt to improve particle selection by using an algorithm known as minimum volume enclosing ellipsoid as in Khachiyan & Todd (1993), which is able to enclose previously selected particles and to include others based on a geometrically defined polytope. Figure 5 illustrates the algorithm when applied to LWFA data, showing the selected particles as black dots; these particles are not in the most dense region once the colors refers to \((x, y)\)-density calculation. When including compactness in \(\text{px}\), the most dense region happens further ahead. As distinct from calculating center of mass and forcing an \(\text{ad hoc}\) diameter or semi-major/minor axes, the minimum volume enclosing ellipsoid (MVEE) algorithm [Khachiyan & Todd (1993); Kumar & Yildirim (2005); Moshtagh (2009)] takes the subset of points and prescribes a polytope model to extrapolate a preliminary sub-selection to other particles likely to be in the bunch. The MVEE algorithm is a semidefinite programming problem and consists of a better approximation to the convexity of subsets of...
Architectures for Data vs. Simulation

Separate Jobs
Compute Intensive
Nearest Neighbor
All-to-All
Random Access

Different architectures for simulation? Can simulation use data architectures?
## Analytics vs. Simulation Kernels:

<table>
<thead>
<tr>
<th>7 Giants of Data</th>
<th>7 Dwarfs of Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic statistics</td>
<td>Monte Carlo methods</td>
</tr>
<tr>
<td>Generalized N-Body</td>
<td>Particle methods</td>
</tr>
<tr>
<td>Graph-theory</td>
<td>Unstructured meshes</td>
</tr>
<tr>
<td>Linear algebra</td>
<td>Dense Linear Algebra</td>
</tr>
<tr>
<td>Optimizations</td>
<td>Sparse Linear Algebra</td>
</tr>
<tr>
<td>Integrations</td>
<td>Spectral methods</td>
</tr>
<tr>
<td>Alignment</td>
<td>Structured Meshes</td>
</tr>
</tbody>
</table>
Machine Learning Mapping to Linear Algebra

- Logistic Regression, Support Vector Machines
- Dimensionality Reduction (e.g., NMF, CX/CUR, PCA)
- Clustering (e.g., MCL, Spectral Clustering)
- Graphical Model Structure Learning (e.g., CONCORD)
- Deep Learning (Convolutional Neural Nets)

- Sparse Matrix-Sparse Vector (SpMSpV)
- Sparse Matrix-Dense Vector (SpMV)
- Sparse Matrix Times Multiple Dense Vectors (SpMM)
- Sparse-Sparse Matrix Product (SpGEMM)
- Dense Matrix Vector (BLAS2)
- Sparse-Dense Matrix Product (SpDM$^3$)
- Dense Matrix Matrix (BLAS3)

Aydin Buluc
Random Access Analytics

- Genome assembly “needs shared memory”

**Global Address Space**
- Low overhead communication
- Remote atomics
- Partitions for any structure

Scales to 15K+ cores
Under 10 minutes for human
First ever solution

---

Specialization: End Game for Moore’s Law

NVIDIA builds deep learning appliance with P100 Tesla’s

Intel buys deep learning startup, Nervana

Google designs its own Tensor Processing Unit (TPU)
Data processing with special purpose hardware

- General trend towards specialization for performance
- Data processing (on raw data) will be first in DOE

Particle Tracking with Neuromorphic chips

Computing in Detectors

Deep learning processors for image analysis

FPGAS for genome analysis
Filtering, De-Noise and Compressing Data

AmeriFlux & FLUXNET: 750 users access carbon sensor data from 960 carbon flux data years

Arno Penzias and Robert Wilson discover Cosmic Microwave Background in 1965
Superfacility Vision

Programmable Networks
Distributed facility provides resilience, accessibility

Data platforms
High performance memory and storage systems

Archive, sharing, and reuse
Data is archived and retrieval through automated metadata

Terabit detectors
Detectors send terabit data streams.

In situ analysis
Advanced algorithms and specialized hardware near detectors

Automation and control
Real-time feedback to adjust telescopes or shot setup for experiments

Offline analysis
Combine with simulation to interpret data, improve models
The scientific process is poised to undergo a radical transformation based on the ability to access, analyze, simulate and combine large and complex data sets.
Superfacility: Integrated network of experimental and computational facilities and expertise

A single interconnected “facility” where data is acquired, stored, analyzed and served

Execution plan: one science area at a time