U.S. DEPARTMENT OF ENERGY | Office of Science

Network Research
Problems and Challenges for DOE Scientists

DOENET2025

February 1-2, 2016

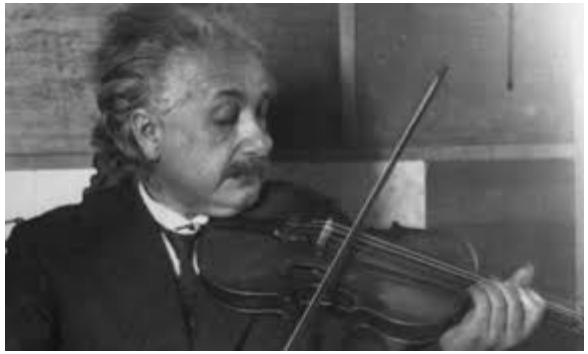
Introduction to DOE Networking

Richard Carlson
Program Manager
richard.carlson@science.doe.gov

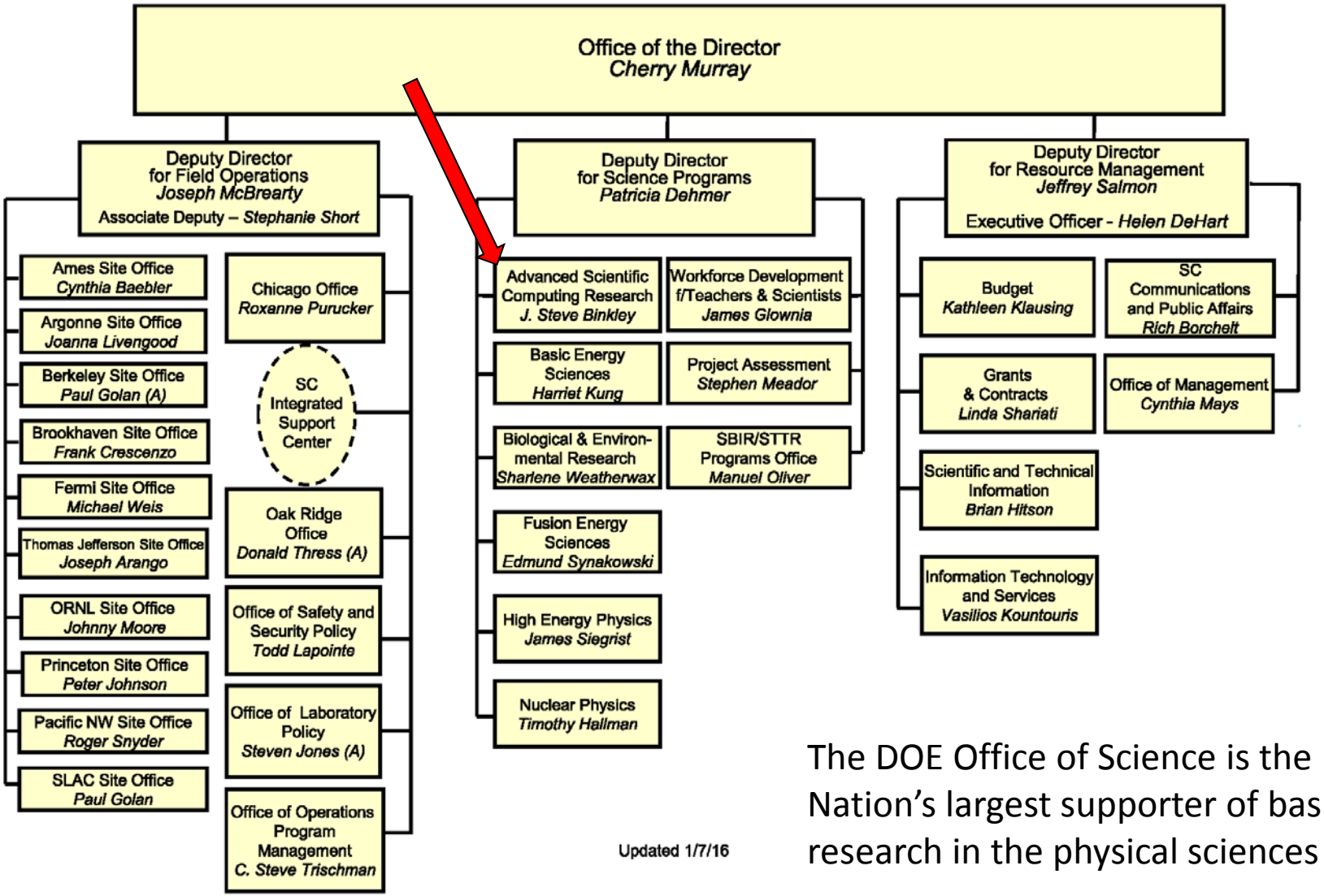
Talk Summary

- **DOE Scientists rely heavily on a robust, reliable, and performant network**
  - Science drivers HEP, BES, BER, …
- **DOE supports long-term fundamental research that may take years before investment returns are realized**
  - Globus, Fastbit, ADIOS, Adaptive Mesh Refinement
- **DOE will partner with Research and Educations networks to deploy advanced technologies without waiting for Vendors by-in**
  - Software Defined Networking and Exchange Points
  - Previous network research activities include
    - TCP Congestion Control
    - OSCARS/Terapath/Lambda Station
- **Workshop goals**
  - Identify problems and challenges
  - Avoid talking about potential solutions
  - Think outside the box

Spoilers!



DOE/SC - ASCR



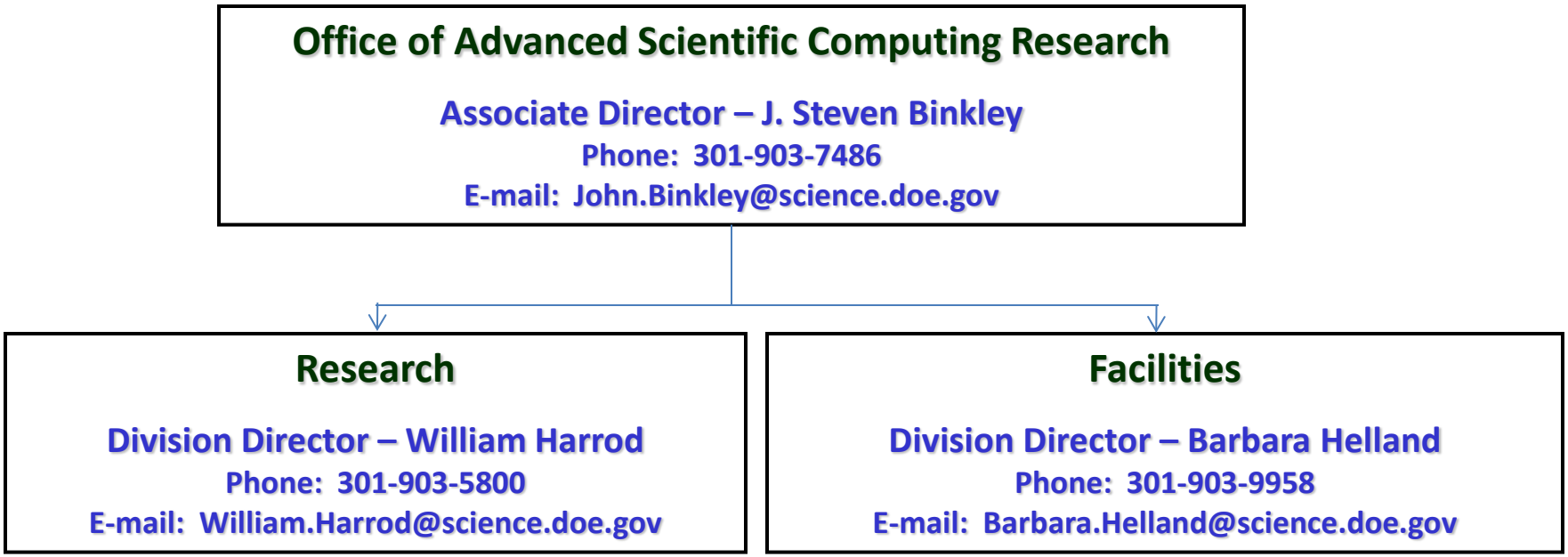
The DOE Office of Science is the Nation's largest supporter of basic research in the physical sciences

DOE and Office of Science Budgets

Department/Office/Division	FY14 Enacted	FY15 Enacted	FY16 President's Request	FY16 Enacted	Change between FY15 & FY16
Department of Energy	27,182.0	27,402.4	29,923.8	29,717.3	+8.4%
Office of Science	5,066.4	5,067.7	5,339.8	5,350.2	+5.6%
ASCR	478.1	541.0	621.0	621.0	+14.8%
BES	1,711.9	1,733.2	1,849.3	1,848.7	+6.7%
BER	609.7	592.0	612.4	609.0	+2.9%
FES	504.7	467.5	420.0	438.0	-6.3%
HEP	796.5	766.0	788.0	785.0	+3.8%
NP	569.1	595.5	624.6	617.1	+3.6%
ARPA-E	280.0	280.0	325.0	291.0	+3.9%

All figures in millions of U.S. Dollars

ASCR at a Glance

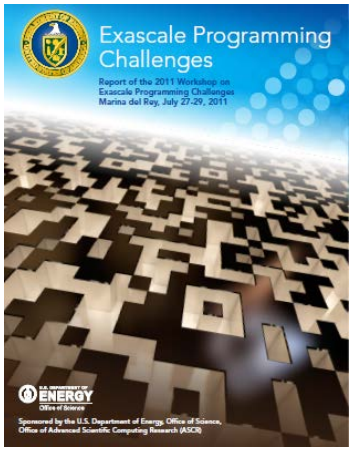


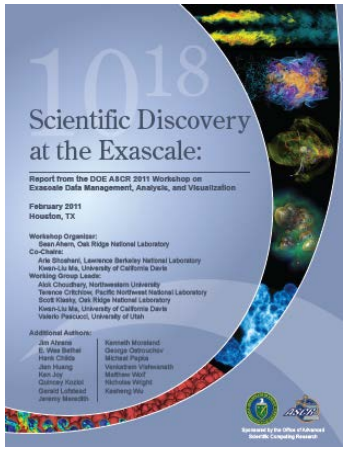
Relevant Websites

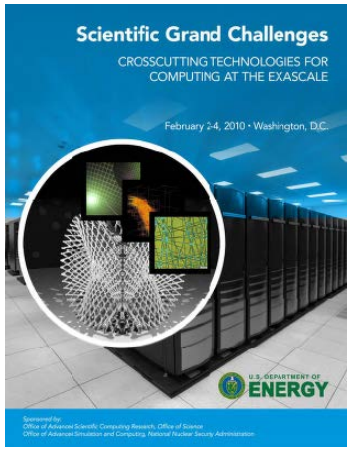
ASCR: science.energy.gov/ascr/
ASCR Workshops and Conferences:
science.energy.gov/ascr/news-and-resources/workshops-and-conferences/
SciDAC: www.scidac.gov
INCITE: science.energy.gov/ascr/facilities/incite/
Exascale Software: www.exascale.org
DOE Grants and Contracts info: science.doe.gov/grants/

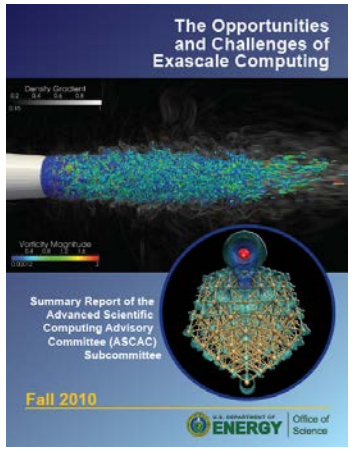
Fundamental Scientific Research

- **Applied Mathematics:** Algorithms and software to solve complex science problems;
- **Computer Science:** Advanced Operating Systems, runtime architectures, and analysis methods to achieve exascale based science;
- **Computational Partnerships:** CoDesign to pioneer the future of scientific applications;
- **Next Generation Networks for Science:** Enabling the future of collaborative and distributed science



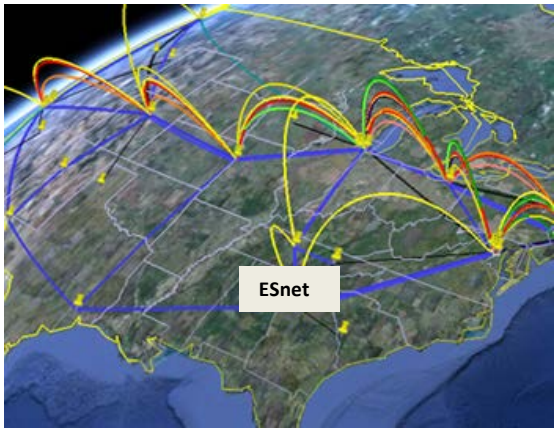






World Class Facilities

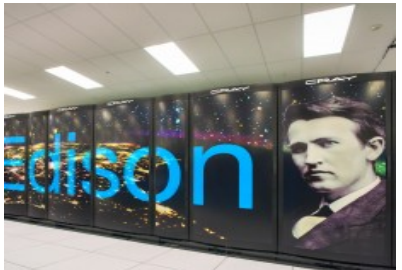
- **High Performance Production Computing for the Office of Science**
  - Characterized by a large number of projects (over 400) and users (over 4800)
- **Leadership Computing for Open Science**
  - Characterized by a small number of projects (about 50) and users (about 800) with computationally intensive projects
  - Cori, Summit, and Theta deployments in 2016/2017
- **International Networking– ESnet**
  - 44 x 100 Gbps terrestrial links, 340 Gbps transatlantic
  - 400 Gbps Terrestrial links in 2017/2018
- **Investing in the future – R&E Prototypes**



TOP500

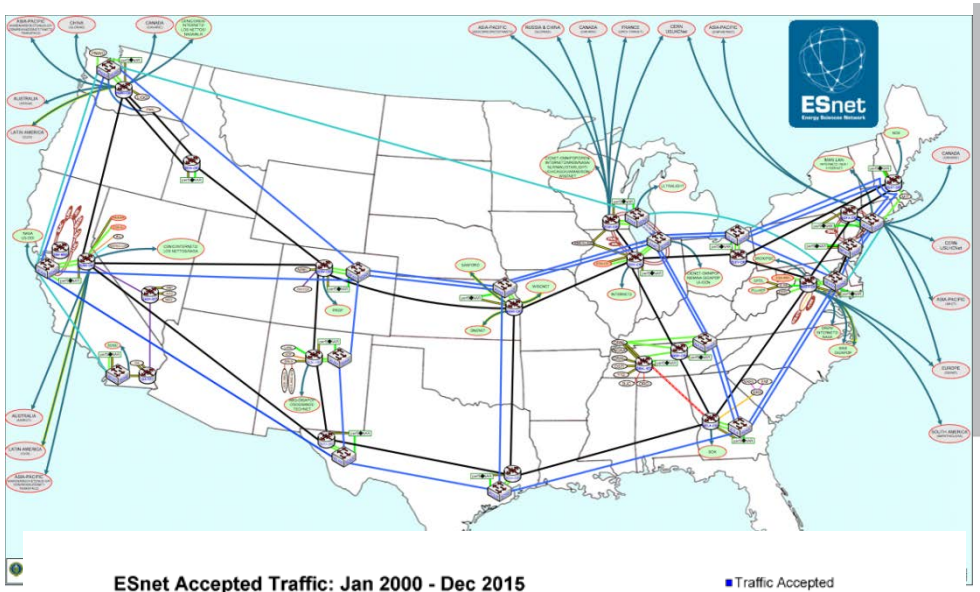
Titan at ORNL

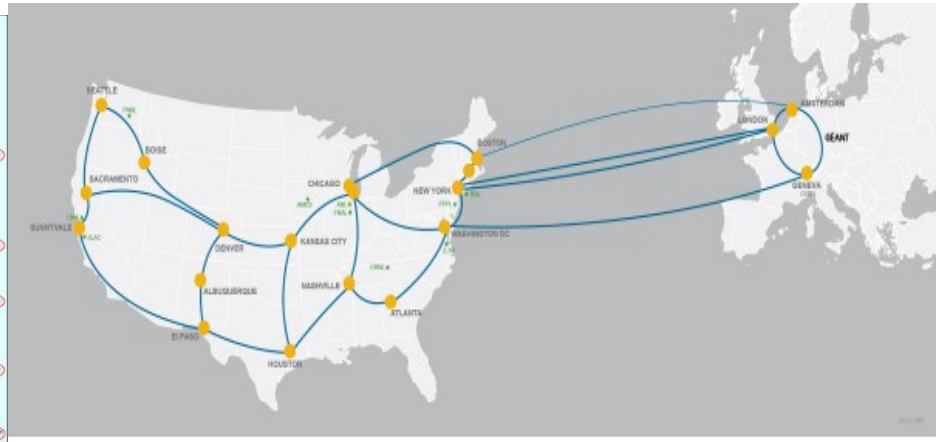
Mira at ANL



Edison at LBNL

ESnet Footprint and Traffic

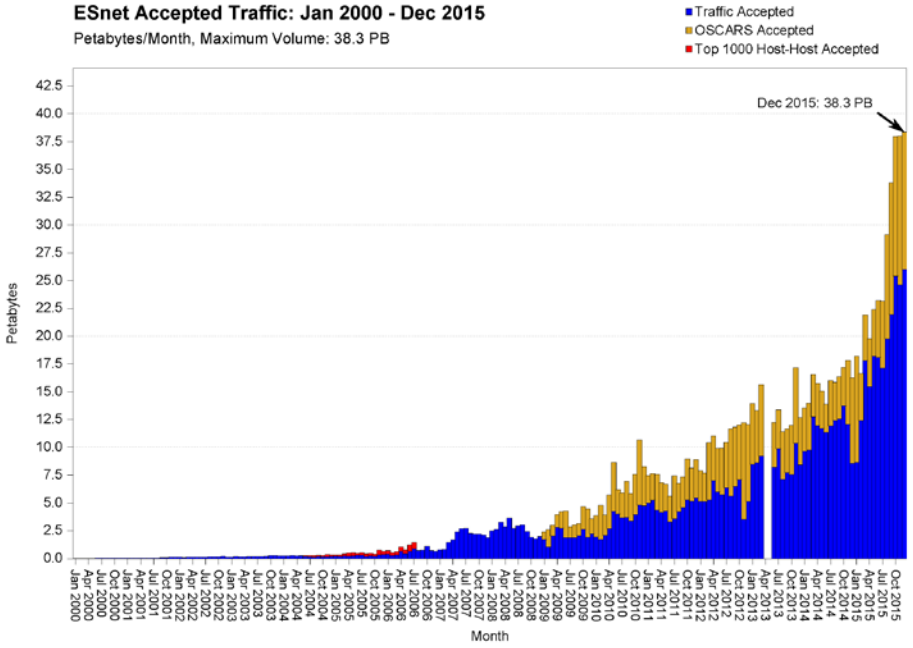


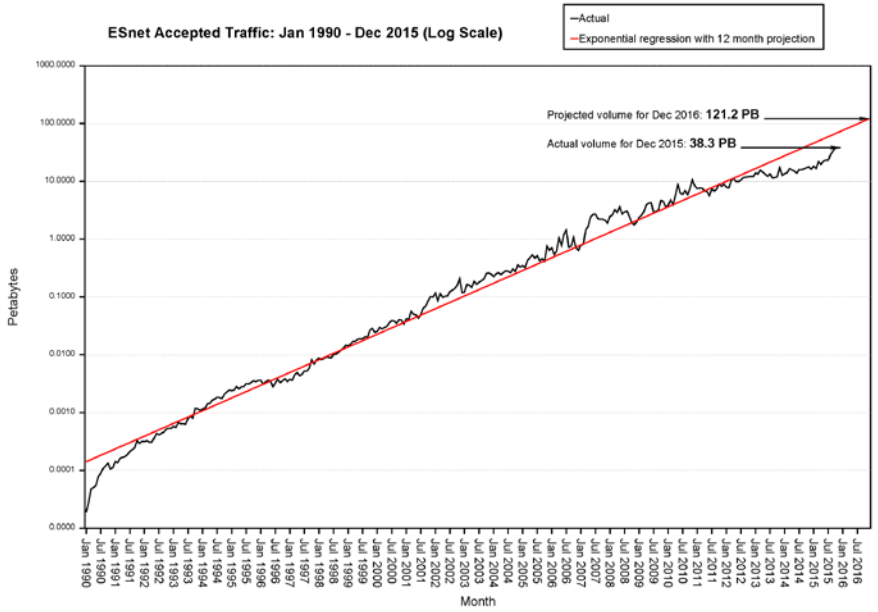


ESnet
ENERGY SCIENCES NETWORK

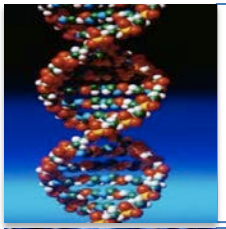
★ Department of Energy Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)	LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ANL	Argonne National Laboratory (Argonne, IL)	ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
BNL	Brookhaven National Laboratory (Upton, NY)	PNNL	Pacific Northwest National Laboratory (Richland, WA)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)	PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)	SLAC	SLAC National Accelerator Laboratory (Menlo Park, CA)



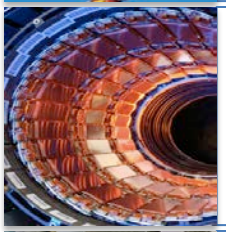


Extreme Scale Science is Causing a Data Explosion



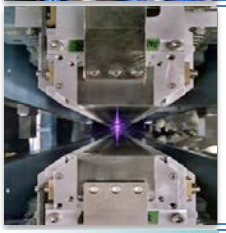
Genomics

Data Volume increases to 10 PB in FY21



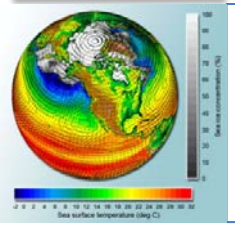
High Energy Physics (Large Hadron Collider)

15 PB of data/year



Light Sources

Approximately 300 TB/day



Climate

Data expected to be hundreds of 100 EB

Driven by exponential technology advances

Data sources

- Scientific Instruments
- Scientific Computing Facilities
- Simulation Results

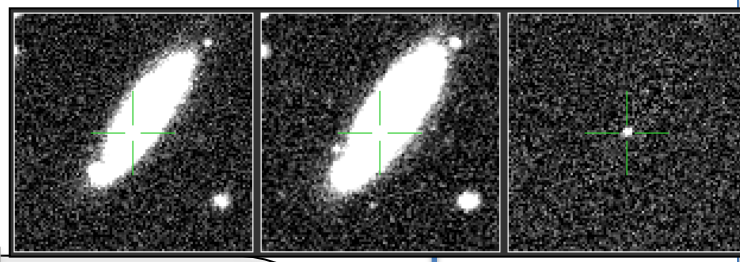
Big Data is part of Big Compute

- Using Big Data requires processing (e.g., search, transform, analyze, …)
- Exascale computing will enable timely and more complex processing of increasingly large Big Data sets

"Very few large scale applications of practical importance are NOT data intensive." – Alok Choudhary, IESP, Kobe, Japan, April 2012

Palomar Transit Factory

100 TBs of Reference Imaging



Palomar 48" Telescope

HPWREN Microwave Relay

SDSC to ESNET

NERSC Data Transfer Node

Networking Data Transfer

500 GB/night

Computing – I/O

Astrometric Solution

Reference Image Creation

Image Processing / Detrending

Image Subtraction

Nightly Image Stacking

Star/Asteroid Rejection

Transient Candidate

Real-Bogus ML Screening

Heavy DB Access

Real-Time Trigger

40 Minutes

1.5B objects in DB

Scanning Page

Wake Me Up – Real Time Trigger

Publish to Web

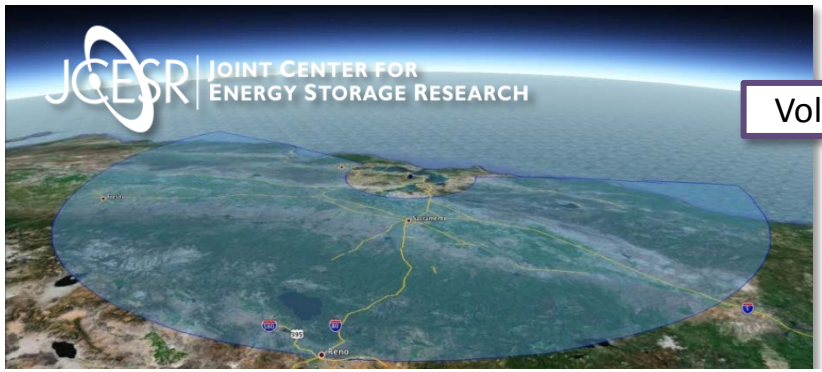
Outside Database for Triggers

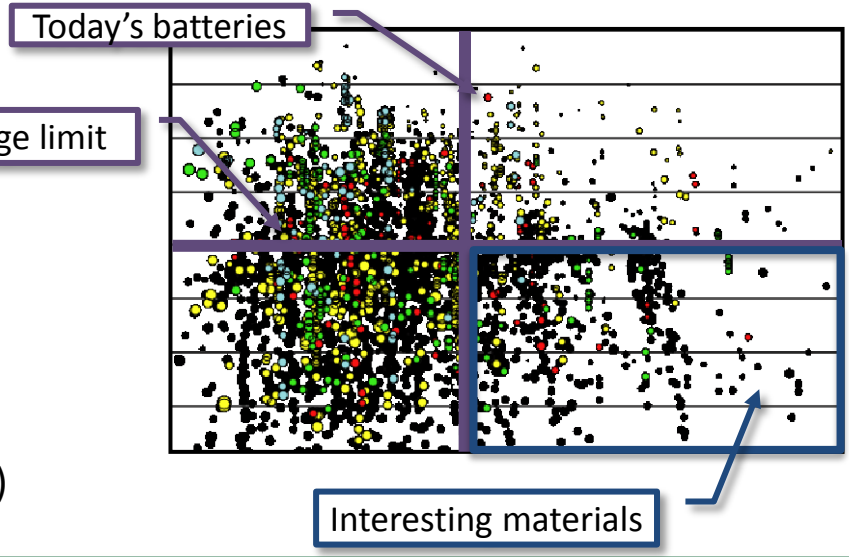
Web UI Marshal

Outside Telescope Follow-up

Computationally Intensive - Materials Genome

Computing 1000× today	Data services for industry and science
• Key to DOE's Energy Storage Hub • Tens of thousands of simulations used to screen potential materials • Need more simulations and fidelity for new classes of materials, studies in extreme environments, etc.	• Results from tens of thousands of simulations web-searchable • Materials Project launched in October 2012, now has >3,000 registered users • Increase U.S. competitiveness; cut in half 18 year time from discovery to market





By 2018:

- Increase energy density (70 miles → 350 miles)
- Reduce battery cost per mile ($150 → $30)

Collaboratively Intensive – Material Structures

Contribute to knowledge base

Knowledge base
Past experiments; simulations; literature; expert knowledge

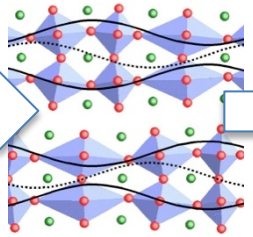
Select experiments (mins—hours)

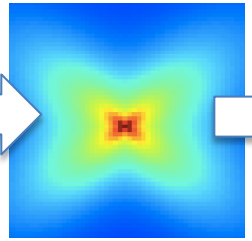
Simulations driven by experiments (mins—days)

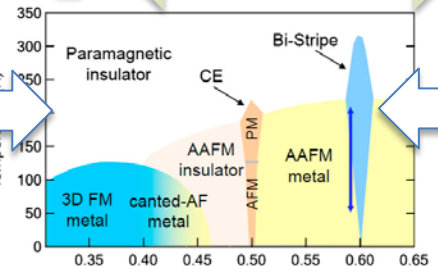
Knowledge-driven decision making

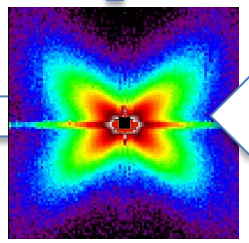
Detect errors (secs—mins)

La 60%
Sr 40%









Material composition

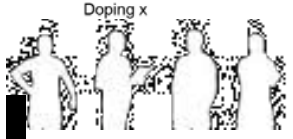
Simulated structure

Simulated scattering

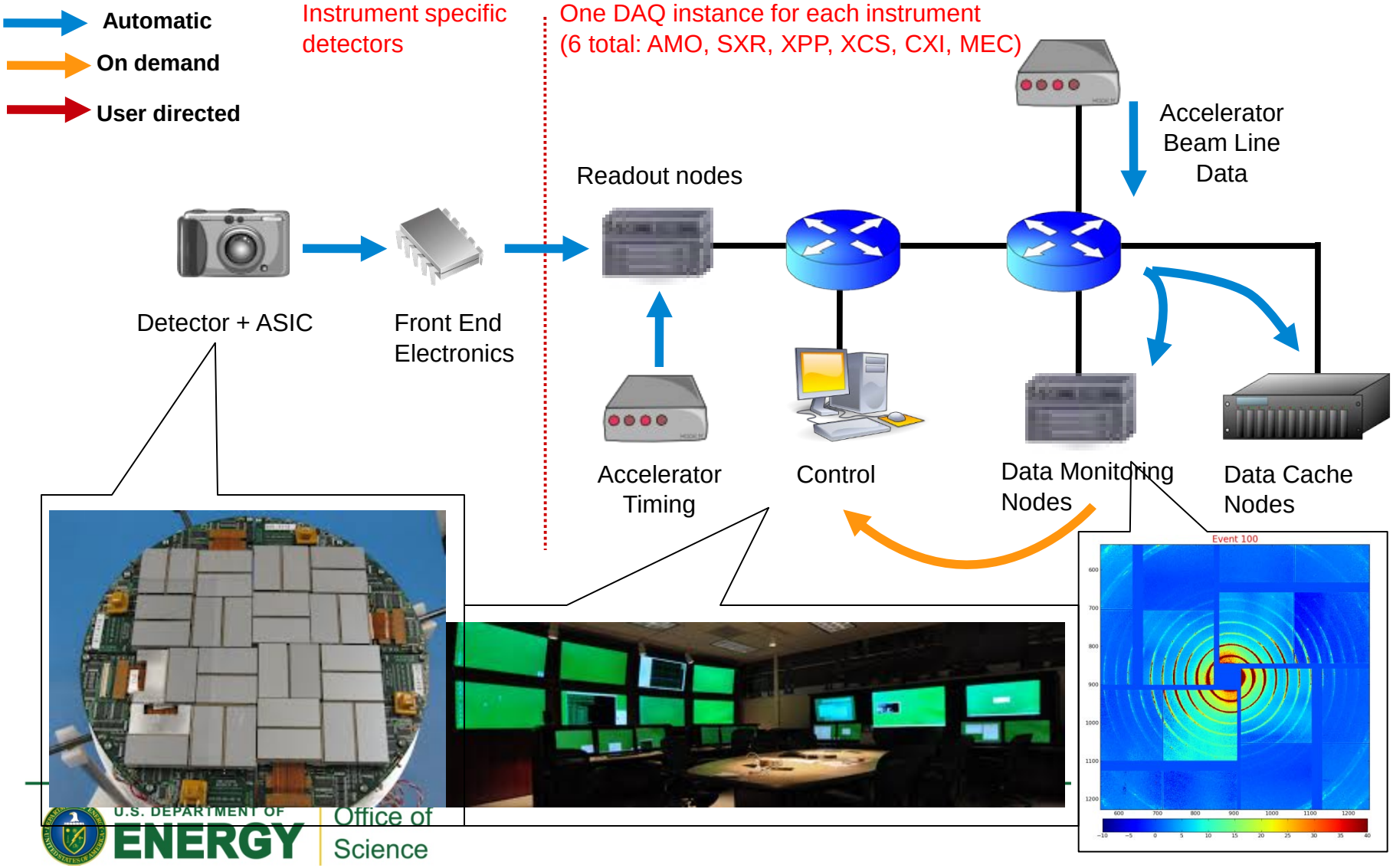
Experimental scattering

Sample

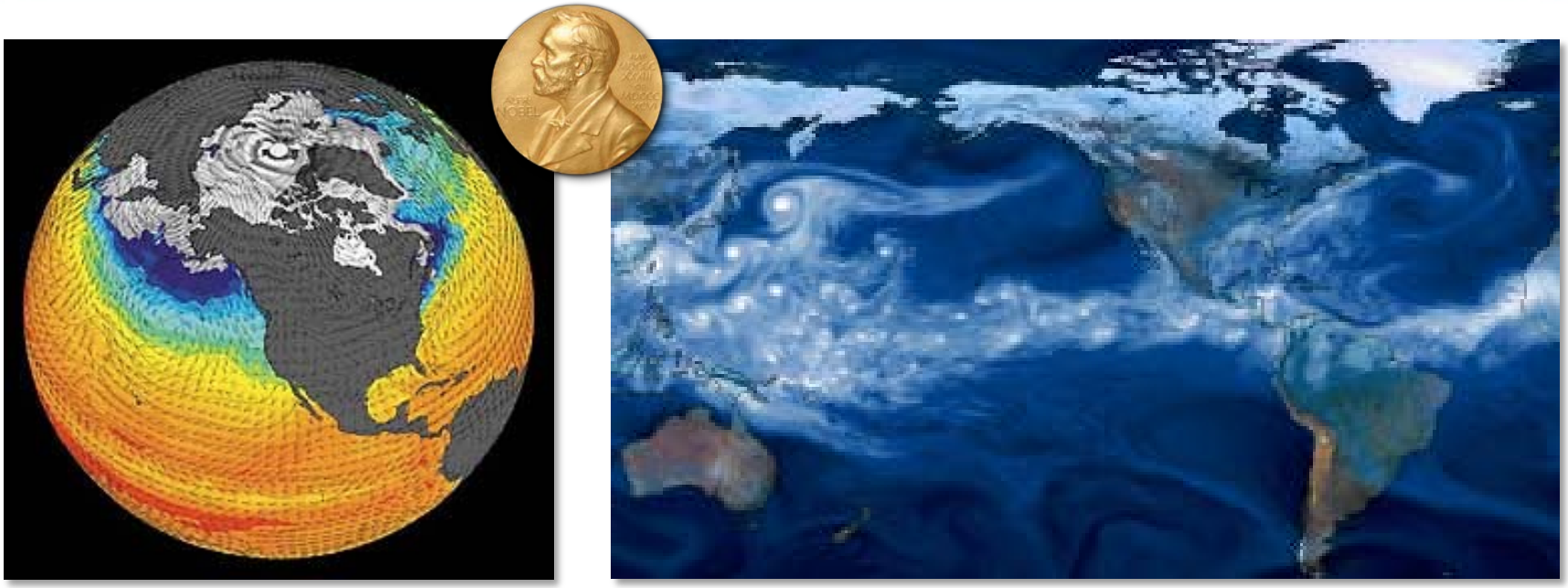
Evolutionary optimization



Data Flow: LCLS Data Acquisition



Computationally Intensive - Climate change analysis



Simulations

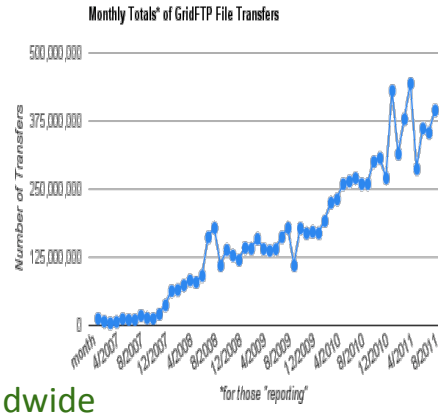
- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

Extreme data

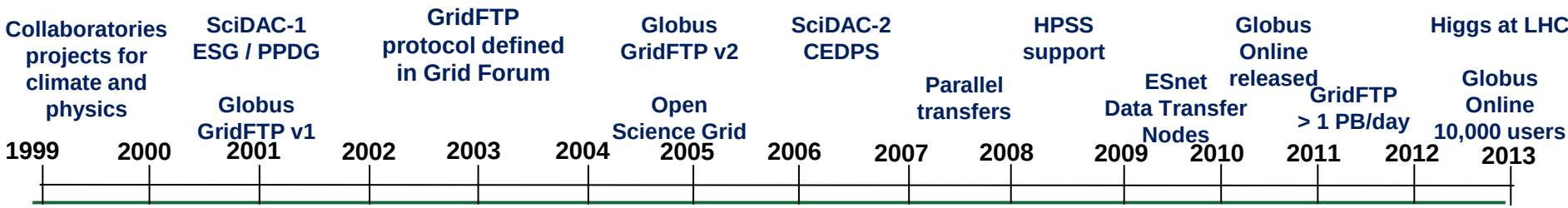
- "Reanalysis" projects need 100× more computing to analyze observations
- Machine learning and other analytics are needed today for petabyte data sets
- Combined simulation/observation will empower policy makers and scientists

High-Speed File Transfer, Synchronization, and Sharing with GridFTP and Globus Online

- **Problem**
  - High-speed collaborative science and modern DOE facilities producing big data need to share large numbers of files rapidly, reliably, and securely over long distances
  - Examples: High-energy physics must distribute 10+ PB worldwide, climate science produces 100 TB now, 10 EB soon; light sources can produce 500 TB/day
- **Solution**
  - (a) GridFTP protocol, high-performance Globus implementation; 10-100x speedup vs. existing methods; also provide reliability and security
  - (b) Globus Online, powerful cloud service for research data management, slashing expertise needs for file movement while enhancing reliability
  - Efficient software: GridFTP from globus.org (>1 PB moved per day); Globus Online at globusonline.org; two R&D 100 Awards



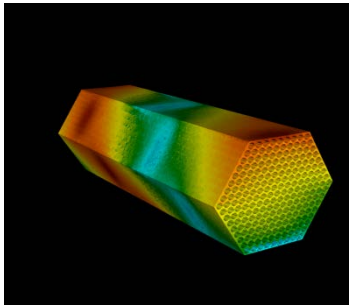
- **Impact**
  - LHC Higgs discovery: Globus GridFTP moves much of the data among 200 sites worldwide
  - Globus Online adopted by major DOE and NSF facilities: NERSC, ALCF, OLCF, APS, ALS, …
  - Testimonial: "I moved 100 7.3 GB files tonight in about 1.5 hours. I am very impressed – Globus Online is the most beneficial grid technology I have ever seen." – Steven Gotllieb, Indiana

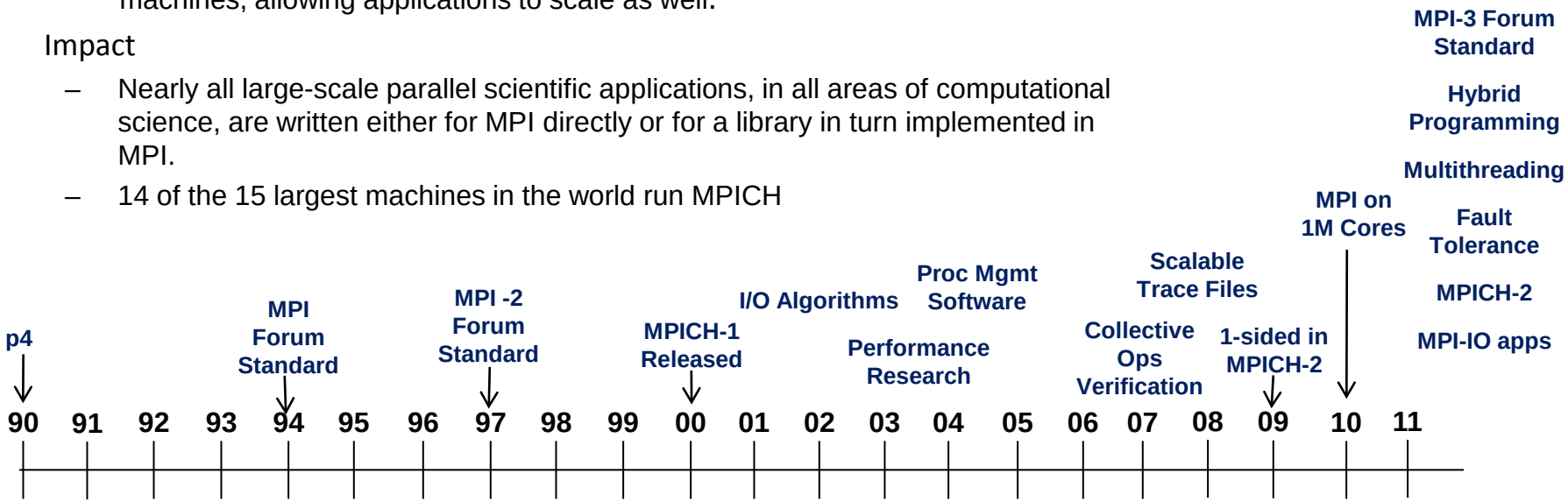


U.S. DEPARTMENT OF ENERGY | Office of Science

Portable Programming With MPI and MPICH

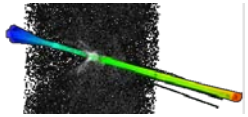
- Problem
  - Before MPI, development of parallel programs was stalled; application writers could not commit to a moving target approach to programming.
- Solution
  - Computer scientists worked with parallel computer vendors and application developers defined a standard programming interface: MPI (Message Passing Interface).
  - Argonne computer scientists developed the first complete implementation, MPICH, helping to promote adoption of the standard.
  - DOE support over the last 15 years has enabled MPICH to scale to larger and larger machines, allowing applications to scale as well.
- Impact
  - Nearly all large-scale parallel scientific applications, in all areas of computational science, are written either for MPI directly or for a library in turn implemented in MPI.
  - 14 of the 15 largest machines in the world run MPICH

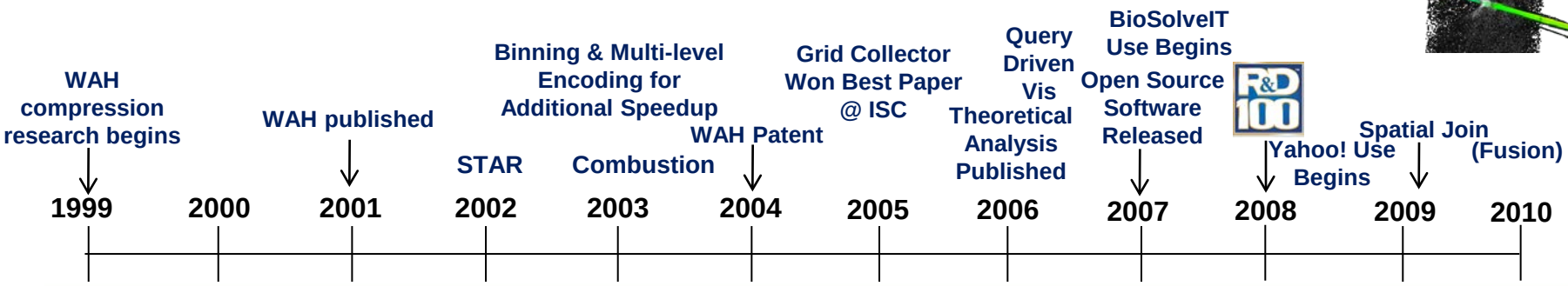




FastBit - Efficient Search Technology for Data Driven Science

- **Problem**
  - Quickly find records satisfying a set of user-specified conditions in a large, complex data set
  - Example: High-energy physics data –find a few thousand events based on conditions on energy level and number of particles in billions of collision events, with hundreds of variables,
- **Solution**
  - Developed new indexing techniques and a new compression method for the indexes, achieved 10-100 fold speedup compared with existing methods
  - Efficient software implementation: available open source from http://sdm.lbl.gov/fastbit/ (1000s of downloads), received a R&D 100 Award
- **Impact**
  - Laser Wakefield Particle Accelerator data analysis: FastBit acts as an efficient back-end for a visual analytics system, providing information for identifying and tracking particles
  - Combustion data analysis: FastBit identifies ignition kernels based on user specified conditions and tracks evolution of the regions
  - Testimonial "FastBit is at least 10x, in many situations 100x, faster than current commercial database technologies" – Senior Software Engineer, Yahoo! Inc





Looming Network Protocol Issues

- **Scientific Communities Demanding Robust and Reliable Network Infrastructure**
  - All Labs are multi-homed
    - Redundant paths via ESnet backbone
    - Separate connections to commercial and REN networks
  - Increased Demand for advanced services
    - Mix of Packet and Circuit Switched network
    - Mix of Optical and Electrical network
    - OSCARS on-demand circuits in daily use
- **Diverse mix of traffic generated by different Science Communities**
  - End-to-End bulk data transfers dominate
  - Complex/Interactive Supercomputing workflows on the horizon
  - Increase in Streaming Experimental/Observational data

Looming Transport Protocol Issues

- **Performance requires lossless network infrastructure**
  - Today ESnet engineers require 0% loss over the entire E2E path for acceptable performance
  - Current transport protocols have non-linear response to loss
- **Data Reliability and Integrity at 100 Gbps and beyond**
  - Data corruption and bit flips must be detected and/or corrected
  - Maintaining throughput over highly Parallel and multi-path network links

Additional Considerations

- **Security and integrity of sites, hosts, and nodes must be built-in instead of added on as an afterthought**
- **Measurement, Monitoring, Troubleshooting, and Operating the network must also be designed in from the start**

Multi-Domain Realities

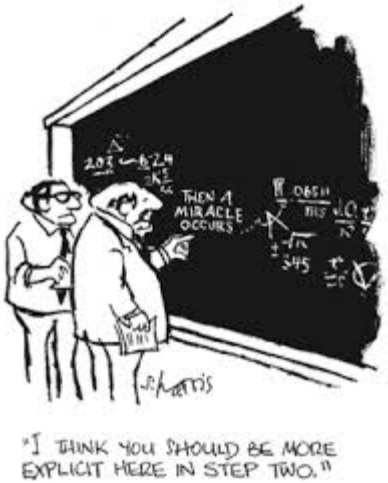
- **ESnet serves the DOE science community**
  - Peers with other REN's to reach individual scientists
  - 80% of the traffic enters or leaves the ESnet infrastructure
- **Campus and Regional networks play a major role in the U.S.**
- **National networks play a major role in the global science community**

Growing Community of Domain Scientists

- **The HEP community was experienced and sophisticated enough to create in-house networking expertise**
- **Other science communities do not have this cohesion or the knowledge needed to duplicate this activity**

Next Generation Networking for Science

- **The ASCR mission is to conduct the research needed to develop new knowledge in Applied Math, Computer Science and Networking**
- **The NGNS research activities include:**
  - Accelerate the development and deployment of technologies, protocols, tools, and high level services needed to support globally distributed science communities
  - Develop high-fidelity models and simulations that accurately describe and predict the observed behavior of scientific workflows, applications, computers and networks



Workshop Agenda

Monday, February 1, 2016

7:30 - 8:30	Continental Breakfast and Registration
8:30 - 9:00	Welcome and Introduction Rich Carlson, U.S. Department of Energy
9:00 - 10:00	Panel Presentations: Network Frontiers for DOE
10:00 - 10:30	Break
10:30 - 11:45	Panel Q&A Session
11:45 – 12:00	Break-out Session Change and Process
12:00 - 1:00	Lunch
1:00 - 2:30	Break-out Session 1: Discussions - Short Term [Terabyte/hour single application bulk data xfer]
2:30 - 3:00	Break-out Session 1: Report Out
3:00 - 3:30	Break
3:30 - 5:00	Break-out Session 2: Discussion - Medium Term [Petabyte/hour single application bulk data xfer]
5:00	Adjourn

Workshop Agenda

Tuesday, February 2, 2016

7:30 - 8:30	Continental Breakfast
8:30 - 9:00	Break-out Session 2: Report Out
9:00 - 10:30	Break-out Session 3: Discussion - Long Term [Exabyte single application bulk data xfer]
10:30 - 11:00	Break
11:00 - 11:30	Break-out Session 3: Report Out
11:30 – 12:00	Conclusions and Next Steps
12:00 - 1:00	Lunch
1:00 - 4:30	Report Writing

Workshop Goals

- **Identify the basic network/transport protocol research issues that inhibit or block scientists from effectively using the network**
  - Terabyte/hour to Exabyte/hour bulk data transfers on a routine basis while supporting a broad mix of other traffic
  - Interact with supercomputer simulation and experimental data analysis in real-time
  - Report faults and/or errors in a manner suitable for scientists and network operators
- **Avoid discussions about current/proposed solutions**
  - Clearly define the problem, not the solution
  - Think out-side the box!

Conclusions

- **DOE needs a robust and active Network Research program to meet the emerging needs of multiple Science Communities**
- **Projects will range from short term (1-3 years) to long term (10+ years)**
- **Basic research into network and transport protocols is required**
- **Managing and providing understandable information to scientists and engineers is also essential**