U.S. DEPARTMENT OF ENERGY | Office of Science

Science at Extreme Scale: Architectural Challenges and Opportunities

DOE Computer Graphics Forum
Argonne National Lab
April 22, 2014

Lucy Nowell, PhD
Computer Scientist and Program Manager
Advanced Scientific Computing Research
Lucy.Nowell@science.doe.gov

Today's Talk

- **Where we expected in 2010**
- **What we learned in 2011**
- **Where we are now**
- **What lies ahead?**

Quick-Facts about the DOE Office of Science

Advanced Scientific Computing Research (ASCR)

Basic Energy Sciences

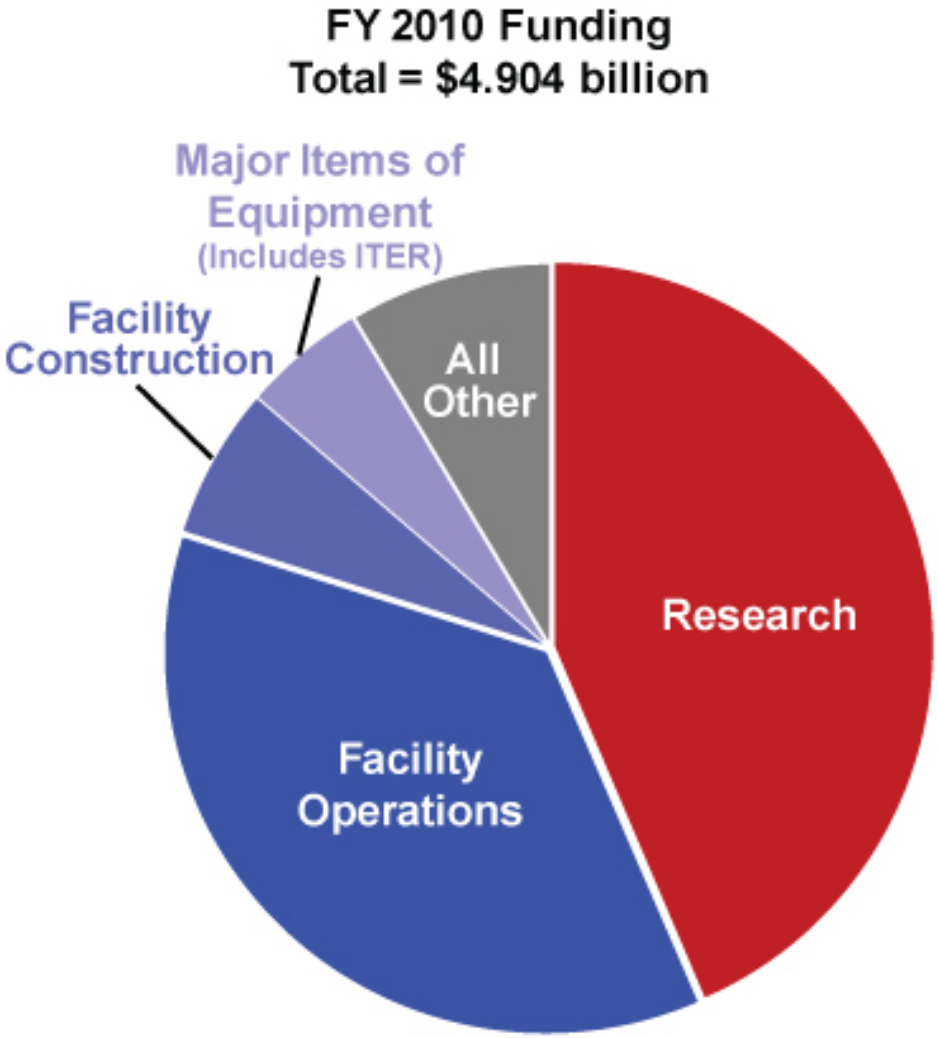
Biological and Environmental Research

Fusion Energy Sciences

High Energy Physics

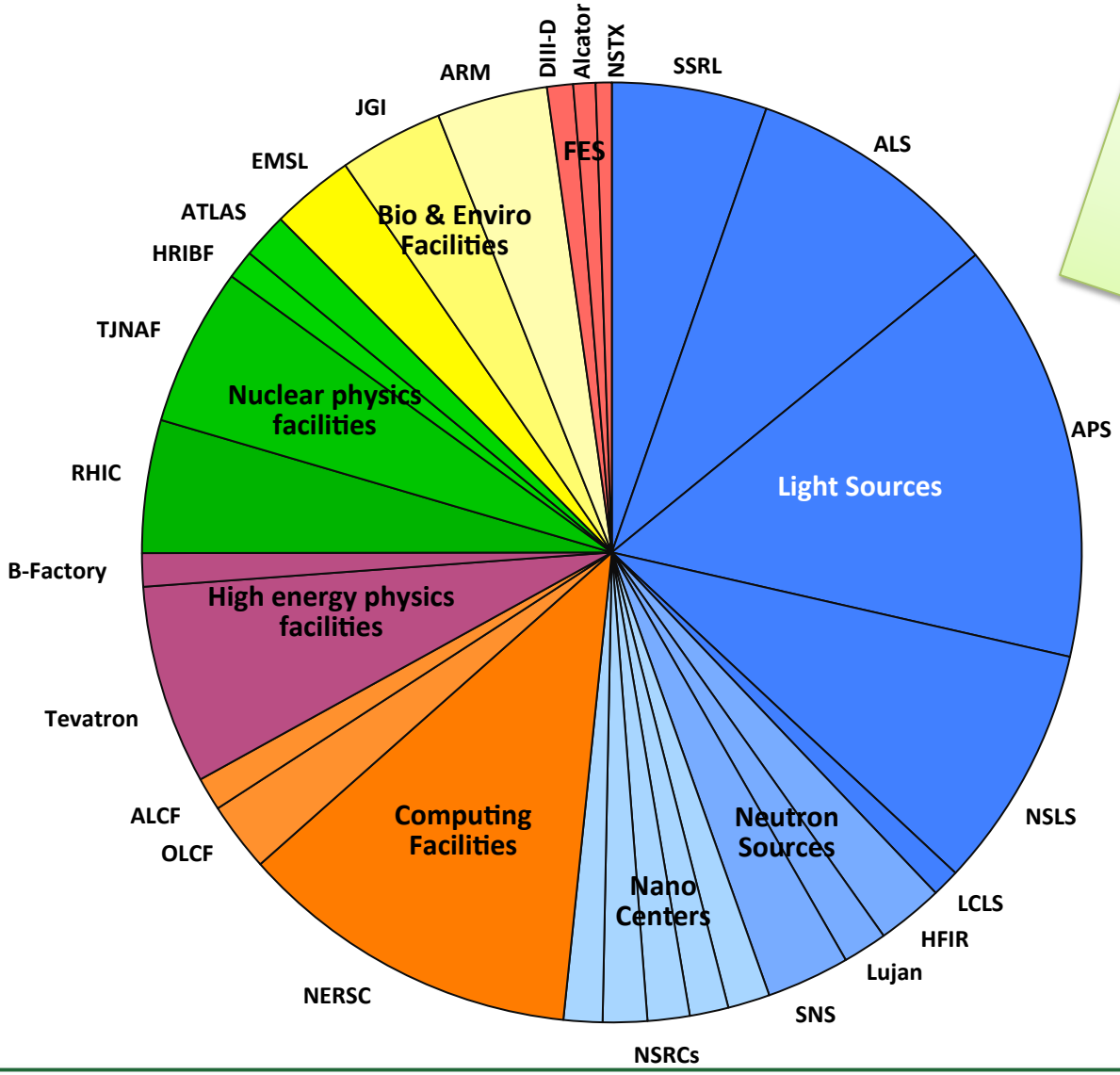
Nuclear Physics

DOE Office of Science User Facility Emphasis



Source: http://science.energy.gov/about/

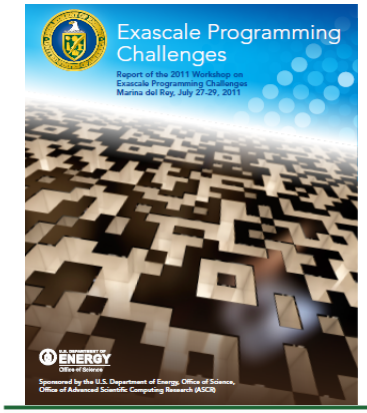
Users Come from all 50 States and D.C.

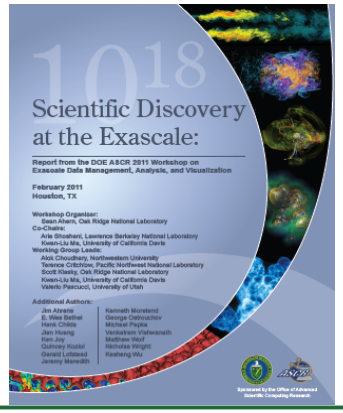


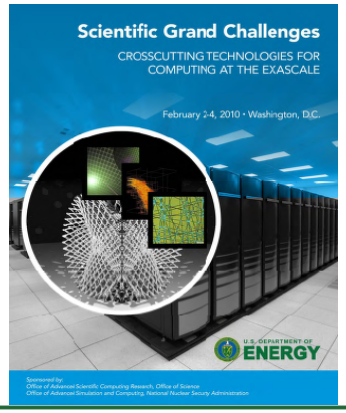
26,000 users/year at 32 national scientific user facilities

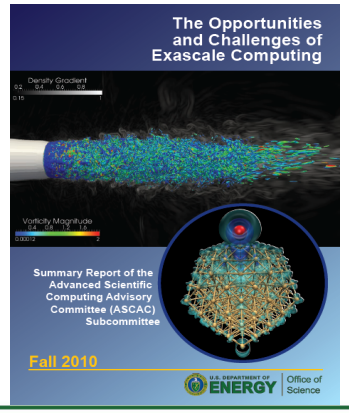
ASCR's Research

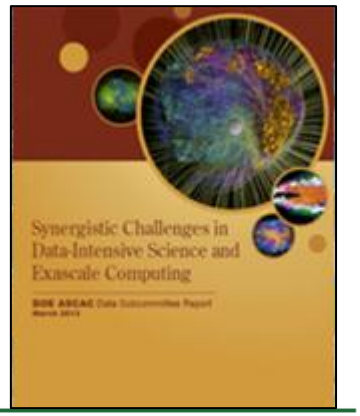
- Applied Mathematics
  - Emphasizes complex systems, uncertainty quantification, large data and exascale algorithms
- Computer Science
  - Exascale computing (architecture, many-core, power aware, fault tolerance), operating systems, compilers, performance tools; scientific data management, integration, analysis and visualization for petabyte to exabyte data sets
- Next Generation Networking
  - Networking, middleware, and collaboration technologies
- Partnerships
  - Co-Design and partnerships to pioneer the future of scientific applications;
- Research and Evaluation Prototypes
  - Fast Forward and Design Forward partnerships with Industry and Non-Recurring Engineering for the planned facility upgrades











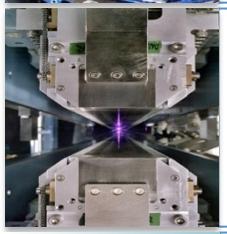
Extreme Scale Science Data Explosion

Genomics

Data Volume increases to 10 PB in FY21

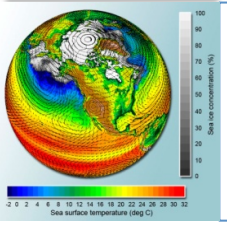
High Energy Physics (Large Hadron Collider)

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
  - Observational data
- **Big Data and Big Compute**
  - Analyzing Big Data requires processing (e.g., search, transform, analyze, …)
  - Extreme scale computing will enable timely and more complex processing of increasingly large Big Data sets

1 EB = 10^{18} bytes of storage

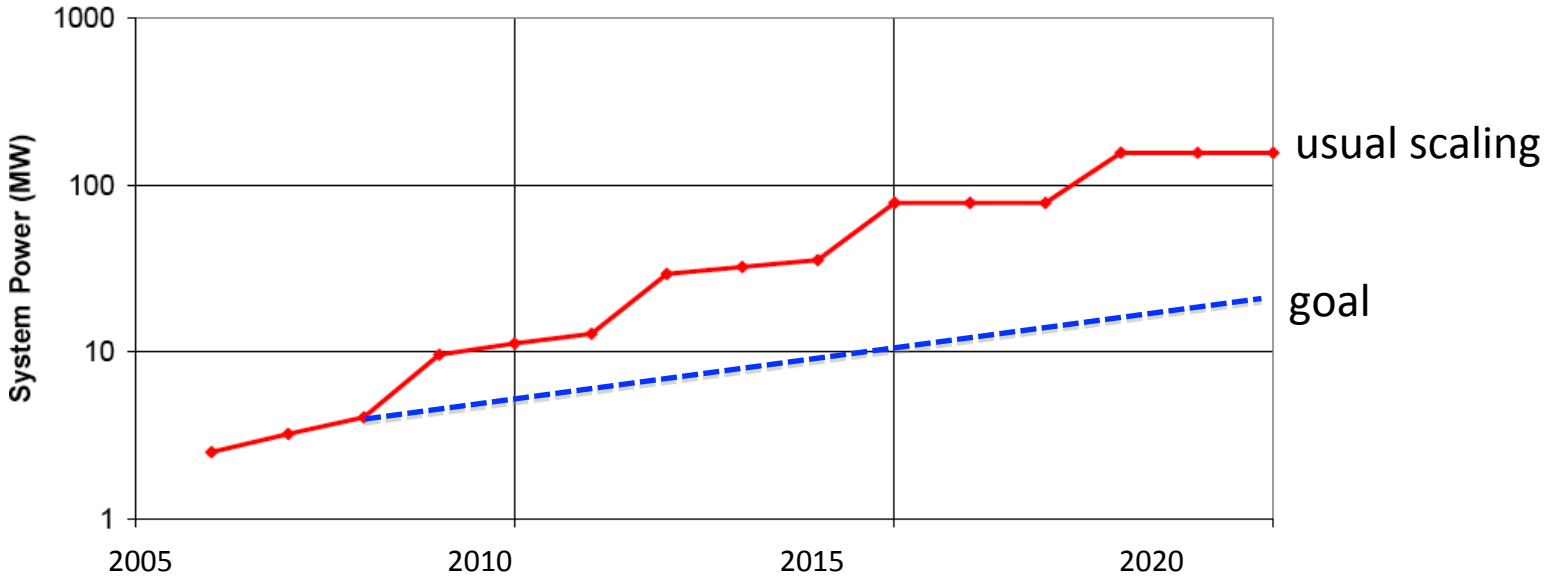
1 PB = 10^{15} bytes of storage

1 TB = 10^{12} bytes of storage

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

The Future is about Energy Efficient Computing

- At $1M per MW, energy costs are substantial
- 1 petaflop in 2010 used 3 MW
- 1 exaflop in 2018 at 200 MW with “usual” scaling
- 1 exaflop in 2018 at 20 MW is target



(Exa)Scale Changes Everything (Circa 2009)

	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Gf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1,000 cpus	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100

DOE Exascale Initiative Roadmap, Architecture and Technology Workshop, San Diego, December, 2009.

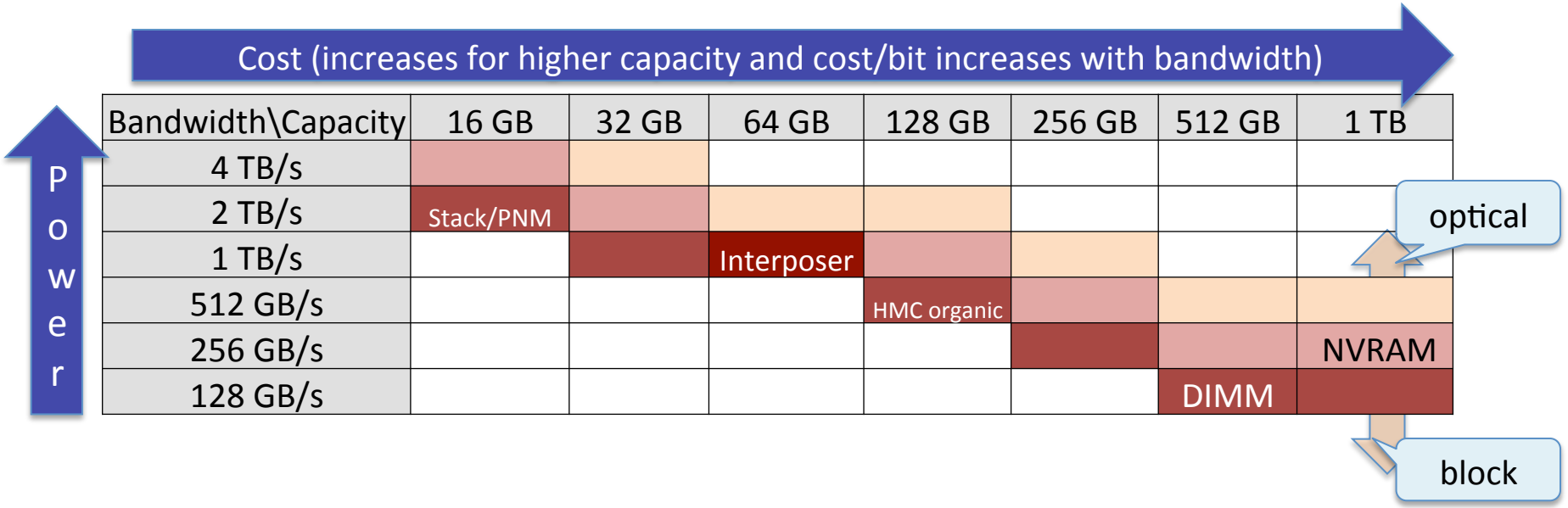
Potential System Architectures… *what did we get wrong?*

Systems	2009	2015	2018 2024
System peak	**2 Peta**	**100-200 Peta**	**1 Exa**
Power	**6 MW**	**~10 MW *15MW***	~~**~20 MW**~~
System memory	0.3 PB	~5 PB ***yes!***	10 PB
Node performance	125 GF	~~400 GF~~ ***3TF***	1-10TF 10-12TF
Node memory BW	25 GB/s	200 GB/s (2-level!!) ~~***100GB/s@100GB + 500GB/s@16GB***~~	>400 GB/s (2-level) ~~250GB/s@200GB + 4TB/s~~ @ 32-64GB
Node concurrency	12	O(100) *yes*	O(1000) *yes*
Interconnect BW (node)	1.5 GB/s	~~25~~ GB/s ***10-15GB/s***	~~50~~ GB/s ***100+ GB/s***
System size (nodes)	18,700	~~250,000 500,000~~ 30,000 – 60,000	O(million) ***yes***
Total concurrency	225,000	O(million)	O(billion)
Storage	15 PB	150 PB	500PB
IO	0.2 TB	10 TB/s ***+ burst buffer 100 TB***	50 TB/s ***+ burst buffer***
MTTI	days	days	O(1 day)

Potential System Architectures (2014 estimates)

Systems	2009	2015	2024
System peak	**2 Peta**	**100-200 Peta**	**1 Exa**
Power	**6 MW**	**10-15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	10 PB
Node performance	125 GF	3TF	10+TF
Node memory BW	25 GB/s	100GB @ 100GB/s 16GB @ 500GB/s	200GB @ 200GB/s 32GB @ 4TB/s
Node concurrency	12	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10-15 GB/s	100-400 GB/s
System size (nodes)	18,700	30k-60k	O(million)
Total concurrency	225,000	O(million)	O(billion)
Storage	15 PB	150 PB + 15 PB burst buffer	500 PB + 50 PB burst buffer
IO	0.2 TB	10 TB/s global PFS + 100 TB/s burst buffer	20 TB/s global PFS + 500 TB/s burst buf
MTTI	days	days	O(1 day)

Memory Speed vs. Capacity Conundrum



Bandwidth\Capacity	16 GB	32 GB	64 GB	128 GB	256 GB	512 GB	1 TB
4 TB/s							
2 TB/s	Stack/PNM						
1 TB/s			Interposer				
512 GB/s				HMC organic			
256 GB/s							NVRAM
128 GB/s						DIMM	

- Because of cost and power issues, we cannot have both high memory bandwidth and large memory capacity
- We evaluate the colored region which is feasible in 2017

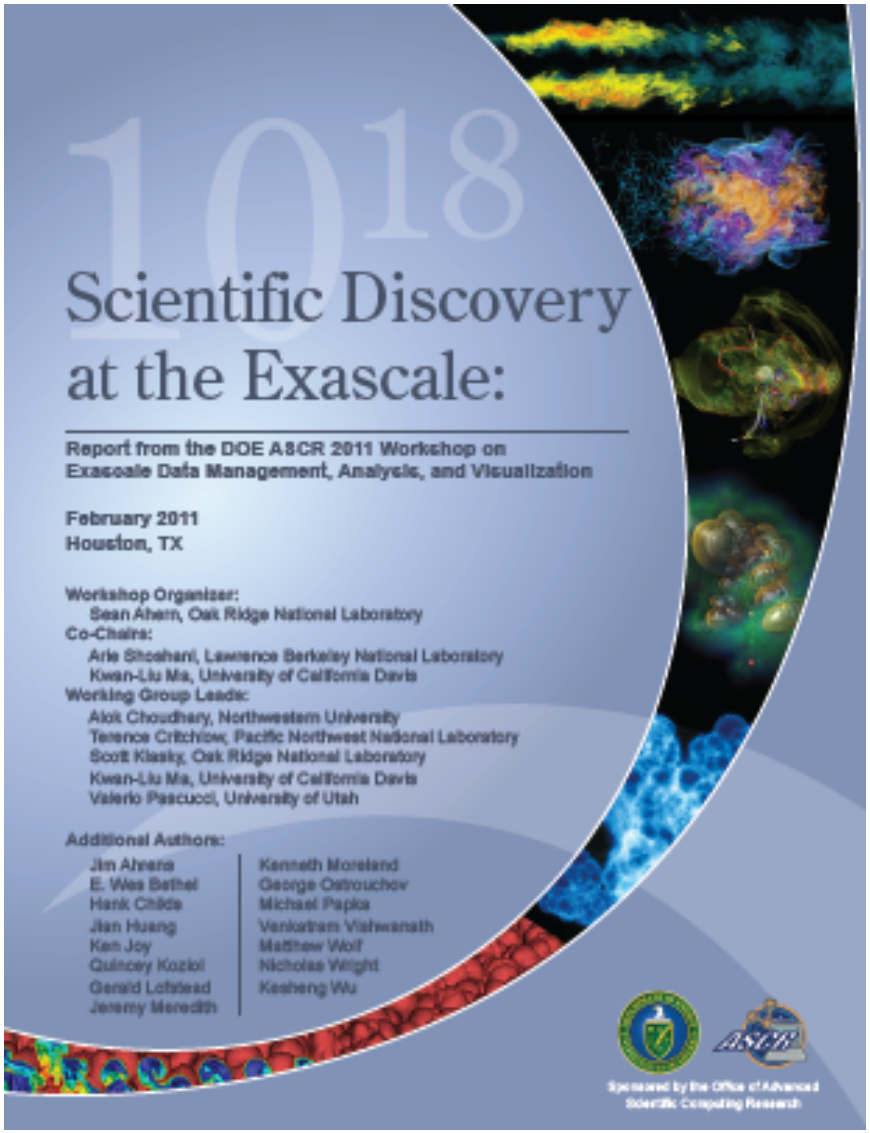
Compute intensive architecture concentrates power and $'s on upper-left

Data Intensive architecture concentrates more power and $'s on lower right

Future of Data Driven Science

- Hardware trends impact data driven science (in many cases more than compute intensive);
  - Data movement dominates power cost and thus must be minimized
  - Power cost drives down the memory footprint;
  - Disk read and write rates will fall further behind processors and memory;
- Data from instruments still on 18-24 month doubling;
- Significant hardware infrastructure needed to support data intensive science that probably will not be replicated at users' home institution (i.e., launching a petabyte file transfer at a users laptop is not friendly).

Scientific Discovery at the Exascale Workshop



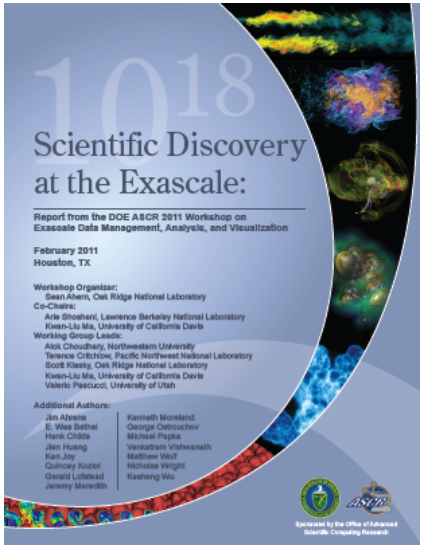
Scientific Discovery at the Exascale: Report from the DOE ASCR 2011 Workshop on Exascale Data Management, Analysis and Visualization, February 2011, Houston, TX

http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Exascale-ASCR-Analysis.pdf

Organizer: Sean Ahern, ORNL; Co-Chairs: Arie Shoshani, LBNL, and Kwan-Liu Ma, UC Davis

Workshop Report

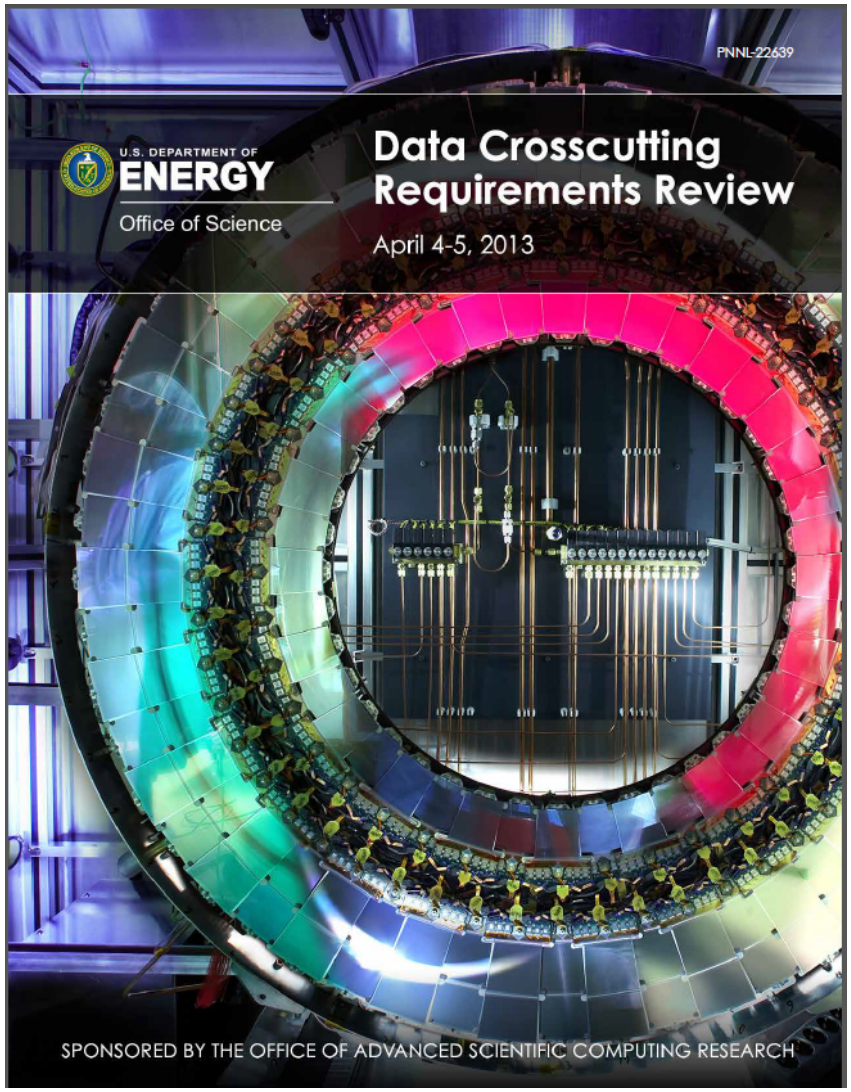
- **Principal Finding: "The disruptive changes posed by a progressive movement towards the exascale in HPC threaten to derail the scientific discovery process. Today's success in extracting knowledge from large HPC simulation output are not generally applicable to the exascale era, and simply scaling existing techniques to higher concurrency is not sufficient to meet the challenge." – p. 1**



Solicitation 14-1043

- **Funding Opportunity Announcement DE-FOA-0001043**
  - http://science.energy.gov/~/media/grants/pdf/foas/2014/SC_FOA_0001043.pdf
- **Five research themes:**
  1. **Usability and user interface design;**
  2. In situ methods for data management, analysis and **visualization**;
  3. Design of in situ workflows to support data management, processing, analysis and **visualization**;
  4. **New approaches to scalable interactive visual analytic environments**; and/or
  5. Proxy applications or workflows and/or simulations for data management, analysis and **visualization** software to support cod-design of extreme scale systems.

ASCR and BES, BER, HEP



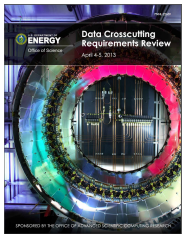
Data Crosscutting Requirements Review

In April 2013, a diverse group of researchers from the U.S. Department of Energy (DOE) scientific community assembled in Germantown, Maryland to assess data requirements associated with DOE-sponsored scientific facilities and large-scale experiments.

http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASCR_DataCrosscutting2_8_28_13.pdf

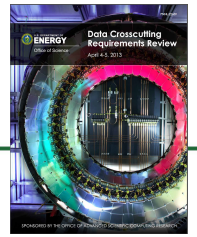
Research Challenges

- How can data be represented in the system so as to maximize its analytic value while also minimizing the power and memory cost of the analytic process?
- How can data provenance, which is essential for validation and later reuse/repurposing, be captured and stored without overburdening a system that is input/output (I/O) bound?
- For complex scientific problems that require integrated analysis of data from multiple simulations, observatories, and/or disciplines, how can the expected IO and memory constraints be overcome to support re-use and repurposing of data?



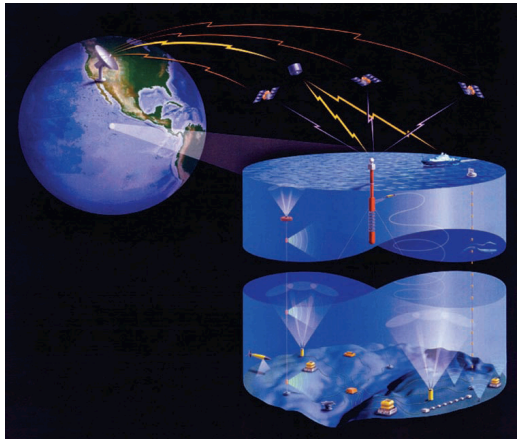
Research Challenges (cont.)

- In the context of these memory and I/O constrained systems, how can simulation data be compared to or integrated with observational/experimental data, both to validate the simulations and to support new types of analysis?
- What new abstractions are needed for long-term data storage that move beyond the concept of files to more richly represent the scientific semantics of experiments, simulations, and data points?
- How can data analysis contribute to generating the ten to one hundred billion way concurrency that future machines will support and need to mask latency?
- How can data management and analysis applications help to mitigate the impact of frequent hardware failures and silent faults?



Wisdom from NSF Visualization Challenge 2007

- **You can *do* science without graphics. But it's very difficult to *communicate* it in the absence of pictures. Indeed, some insights can only be made widely comprehensible as images. How many people would have heard of fractal geometry or the double helix or solar flares or synaptic morphology or the cosmic microwave background, if they had been described solely in words?**
- **To the general public, whose support sustains the global research enterprise, these and scores of other indispensable concepts exist chiefly as images.**



NSF Science and Visualization Challenge 2007, Special Report
http://www.nsf.gov/news/special_report/scivis/index.jsp?id=challenge

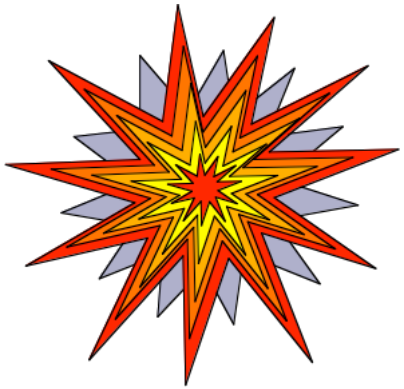
(Exa)Scale Changes Everything (Circa 2009)

	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Gf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1,000 cpus	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100

DOE Exascale Initiative Roadmap, Architecture and Technology Workshop, San Diego, December, 2009.

What Can We Do with all Those FLOPS?!

- **If FLOPS are "free," why can't we have more of them for users and usability?!**
  - Humans learn best by doing, not by looking…
  - Real-time interaction with visualizations generated in situ?
  - Computational steering?
  - Ways to enrich analysis with users' knowledge?
- **Making sure the methods we have now will work for the future is important, but what lies beyond that… at the edges of our imagination and creativity?**



Thank you!

Lucy Nowell

lucy.nowell@science.doe.gov

http://science.energy.gov/ascr/