U.S. DEPARTMENT OF ENERGY | Office of Science

ASCR Data Activities

Rich Carlson
Carolyn Lauzon

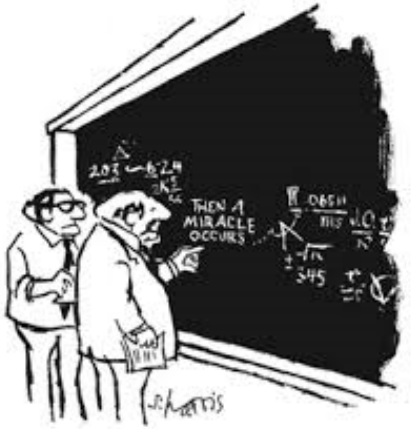
‘Spoilers’

- **For DOE: ‘Big Data’ is an old issue**
  - “Science drivers derive from the increasingly multidisciplinary nature of scientific research and from new classes of scientific problems that can be addressed when massive databases, powerful computers, and the like are accessible online.” Blueprint for Future Science Middleware and Grid Research and Infrastructure: Aug 26-28, 2002
  - “Such a capability [supporting large-science applications] demands revolutionary advances in network technologies for tasks such as petabyte data transfers at tera bit-per-second speeds, computational steering, interactive and collaborative visualization, and remote instrument control” Network Provisioning and Protocols for DOE Large-Science Applications: Apr 10-11, 2003
  - “Current facilities, like SLAC National Accelerator Laboratory’s Linac Coherent Light Source, can produce up to 18 terabytes (TB) per day, while upgraded detectors at Lawrence Berkeley National Laboratory’s Advanced Light Source will generate ~10TB per hour. The expectation is that these rates will increase by over an order of magnitude in the coming decade.” ASCR/BES Data Workshop 2011: Oct 24-25, 2011

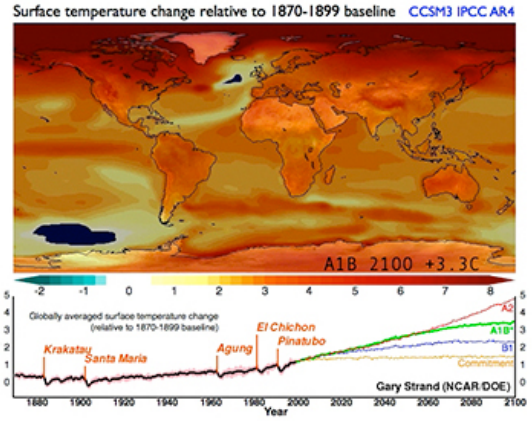
The 3 legs of Scientific Discovery

Unique instruments collect observational data

Theory and Math to create predictions and test hypothesis

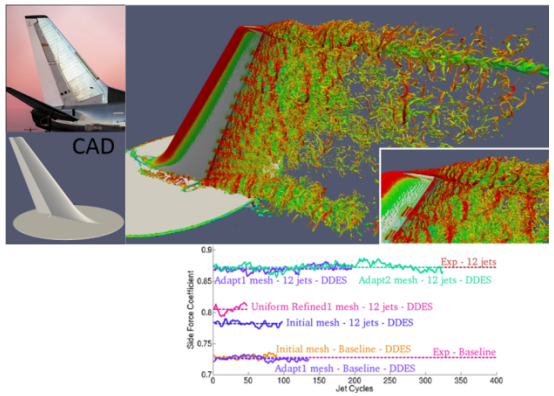


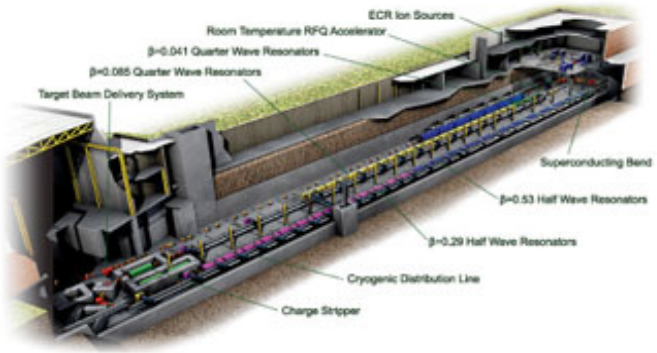
Computational models explore virtual worlds

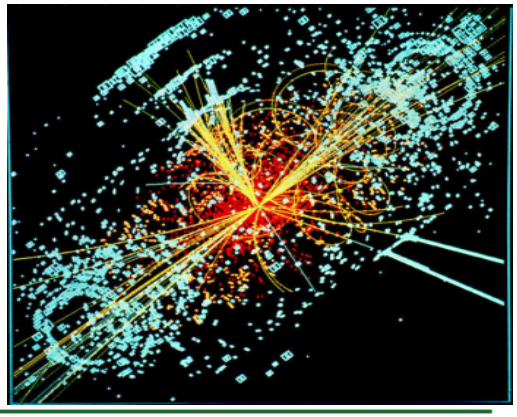


Big Data in Science

- **Reports from 1990's and 2000's all say that DOE science has a data problem**
  - Detector technology is increasing data volumes and rates
  - Remote access to unique instruments by individuals, small teams and large communities
  - High performance networks reducing the distance barrier
  - Computational science program leverages HPC advances



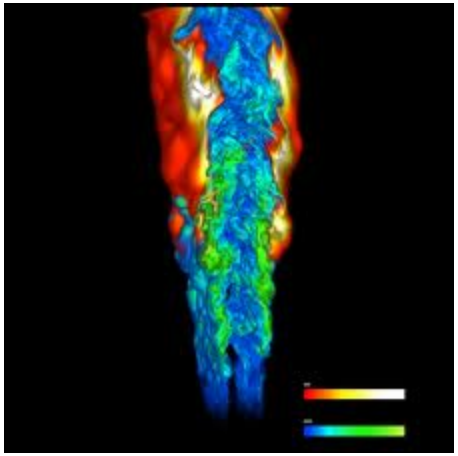




OASCR/MICS Programs

- **Scientific Collaboratories**
  - Develop middleware tools and services
    - Open Science Grid (OSG)
    - Earth Systems Grid (ESG)
    - Particle Physics Data Grid (PPDG)
    - Fusion Grid

- **Scientific Discovery through Advanced Computing (SciDAC)**
  - Bring computer scientists and domain scientists together to solve large complex problems
    - Combustion
    - Climate
    - Material Structures



OASCR – Office of Advanced Scientific Computing Research
MICS – Mathematical, Information, and Computational Science

Scientific Computing Uses

- **High Performance Computing**
  - Solve large computationally intensive problems
    - Climate
    - Combustion
    - Material Structures

- **High Throughput Computing**
  - Solve large numbers of computational problems
    - Particle collision analysis
    - Genome sequencing

Scientific Data Research

- **Distributed Data**
  - Developing new mechanisms that allow computers to process and analyze data streamed from unique instruments
  - Develop new mechanisms that allow multiple scientists to simultaneously examine, compare, and manipulate simulation and experimental data
- **Computational Data**
  - Developing applications and libraries that can take advantage of computers with millions of cores and billions of processing threads
  - Developing new mechanisms to deal with analyzing large amounts of simulation data
  - Developing new mechanisms to visualize and display complex data structures and features

NGNS Research Program Activities

- **Revive Scientific Collaboration program**
  - Link to Exascale Computing Initiative
    - Extreme scale computing spans desktop (teraflop) midrange (petaflop) and Leadership (Exaflop) machines
    - Research needed to make this a seamless computing ecosystem instead of separate computing classes
  - Accelerating Scientific Knowledge Discovery (ASKD)
    - Base research programs to needed to realize seamless computing ecosystem vision
    - Meet the emerging data needs of the experimental and observational science communities
  - Software sustainability (Thomas)
  - SciDAC SDAV (Scalable Data-Management, Analysis, and Visualization) Institute

OSTP memo on Open Data Policies

- **SC Response**
  - SC working group on Digital Data (SCWGDD) representing the 6 program offices
    - Data used to validate research results must be publicly available with paper describing results
    - require proposals to contain a Data Management Plan
    - ASCR considers SW a data artifact
- **NIST responds**
  - Public working group (NBD-WG) to develop architecture and definition documents
  - Interagency Technical Advisory Group (iTAG)
- **NAS responds**
  - BRDI board
- **NITRD responds**
  - Big Data WG
  - Large Scale Networking subcommittees
- **Industry responds**
  - Research Data Alliance https://rd-alliance.org

ASCR Workshops and Reviews

- **2011 Scientific Collaborations for Extreme-Scale Science (SCESS)**
  - https://indico.bnl.gov/materialDisplay.py?materialId=1&confId=403
- **2013 ASCR/HEP Data Requirements and Cross Cutting**
  - http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/HEP_ASCR_Data_Summit_Report_April_2013.pdf
  - http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASCR_DataCrosscutting2_8_28_13.pdf
- **2012 ASCR/BES**
  - http://science.energy.gov/~/media/ascr/pdf/research/scidac/ASCR_BES_Data_Report.pdf
- **2012 Geant4 (HEP)**
  - http://science.energy.gov/~/media/ascr/pdf/research/scidac/GEANT4-final.pdf
- **2013 Synergistic Challenges in Data-Intesnive Science and Exascale Computing**
  - http://science.energy.gov/~/media/ascr/ascac/pdf/reports/2013/ASCAC_Data_Intensive_Computing_report_final.pdf
- **2013 Accelerating Scientific Knowledge Discovery (ASKD)**
  - http://www.orau.gov/ASKD2013/ASKD_Report_V1_0.pdf
- **2014 Software Productivity for Extreme Scale Science**
  - http://www.orau.gov/swproductivity2014/
- **ESnet and NERSC requirements workshops**
  - http://www.es.net/about/science-requirements/reports/
  - http://www.nersc.gov/science/hpc-requirements-reviews/

Virtual Data Facility

- **Jan 2013: ALD's present concept at Facilities Strategic Planning meeting**
- **Summer 2013: NERSC presents updated plan**
- **Jan 2014: ASCR Facilities DD launches VDF planning project**
  - ASCR facilities center directors & ASCR PM
  - Target workshop in April/May
  - Generate FY16 budget justification text
    - Scope, cost, user community, structure
    - Hardware and Software split
    - Open/closed deployment
- **2013: BER begins its own VDF project, hires PM**

Report URLs

Geant 4 (dated 9/12)	**http://science.energy.gov/~/media/ascr/pdf/research/scidac/GEANT4-final.pdf**
BES/ASCR (dated 6/12)	http://science.energy.gov/~/media/ascr/pdf/research/scidac/ASCR_BES_Data_Report.pdf
HEP/ASCR data summit (4/13)	http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/HEP_ASCR_Data_Summit_Report_April_2013.pdf
Data X-cut (4/4 -4/5)	http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASCR_DataCrosscutting2_8_28_13.pdf
ASKD WG (8/14)	http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASKD_Report_V1_0.pdf
NERSC Req. Summary (posted 01/13)	http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/Data-Requirements-from-NERSC-v5.pdf

For NERSC Req. Reviews looking to 2017

ASCR	**TBD**
BER	http://www.nersc.gov/assets/HPC-Requirements-for-Science/BER2017/BER2017FinalJune7.pdf
BES	TBD
HEP	TBD
NP	TBD
FES	TBD

For ESnet (most recent final reports posted. I believe more up-to-date ones will be posted soon).

ASCR (2012)	**http://www.es.net/assets/pubs_presos/ASCR-Net-Req-Review-2012-Final-Report.pdf**
BER (2012 report)	http://www.es.net/assets/pubs_presos/BER-Net-Req-Review-2012-Final-Report.pdf
BES (2010)	http://www.es.net/assets/Uploads/BES-Net-Req-Workshop-2010-Final-Report.pdf
NP(2011)	http://www.es.net/assets/pubs_presos/NP-Net-Req-Workshop-2011-Final-Report.pdf
FES (2011)	http://www.es.net/assets/pubs_presos/FES-Net-Req-Workshop-2011-Final-Report.pdf
HEP (2009)	http://www.es.net/assets/Papers-and-Publications/HEP-Net-Req-Workshop-2009-Final-Report.pdf