National Energy Research Scientific Computing Center (NERSC)

Jonathan Carter

SOS 11
June 12, 2007

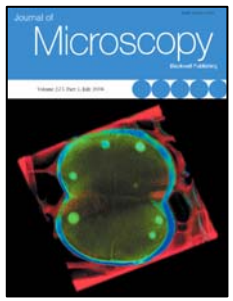
NERSC Mission

The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate the pace of scientific discovery by providing high performance computing, information, data, and communications services for research sponsored by the DOE Office of Science (SC).

Science-Driven Computing Strategy 2006 -2010

Office of Science
U.S. DEPARTMENT OF ENERGY

Science-Driven Computing



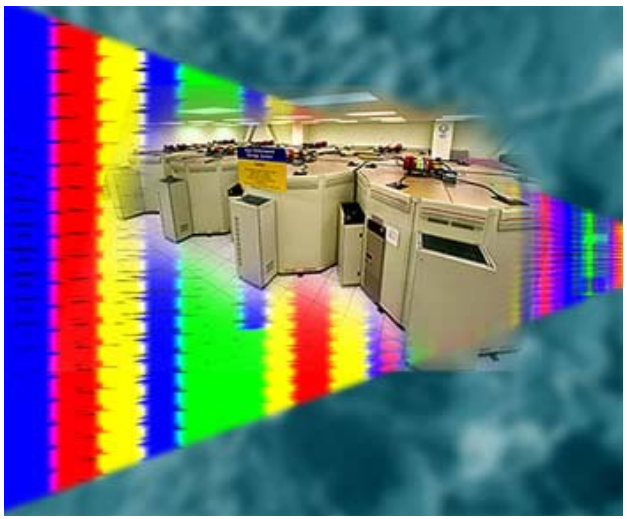
National Energy Research Scientific Computer Center (NERSC)

- computational facility for open science
- international user community
- 2500 users
- 300 projects

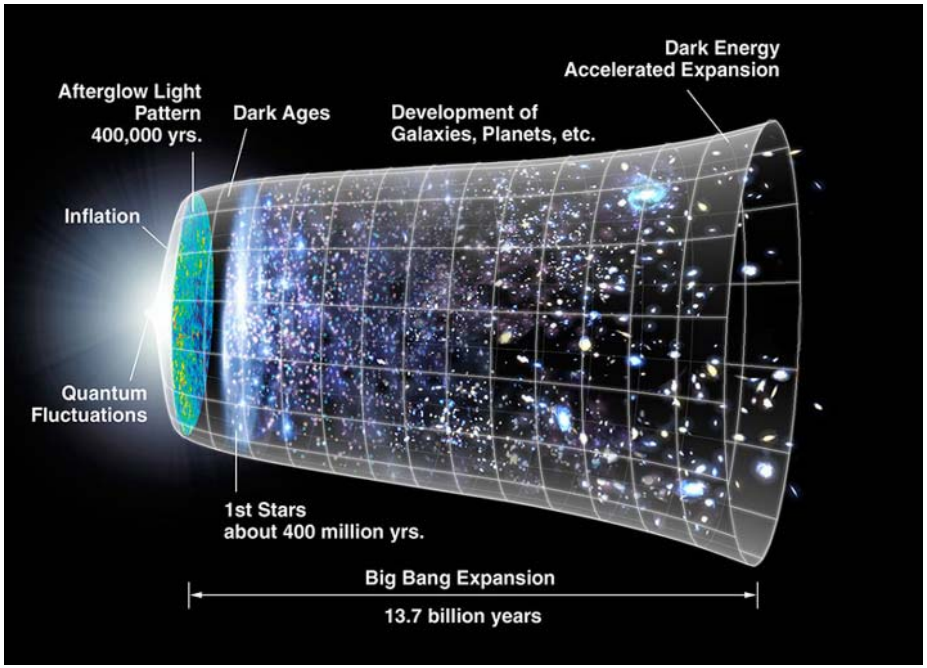
NERSC is enabling new science

Archiving for Genomics Research

- Production Genome Facility (PGF) at Joint Genome Institute (JGI) is producing sequence data at increasing rate
  - 2 million files per month of trace data (25 to 100 KB each)
  - 100 assembled projects per month (50 MB to 250 MB)
  - several very large assembled projects per year (~50 GB).
  - total about 2 TB per month on average
- NERSC and PGF staff collaborated to set up data pipeline using ESnet's new Bay Area MAN
- NERSC provides scientific data archive capability

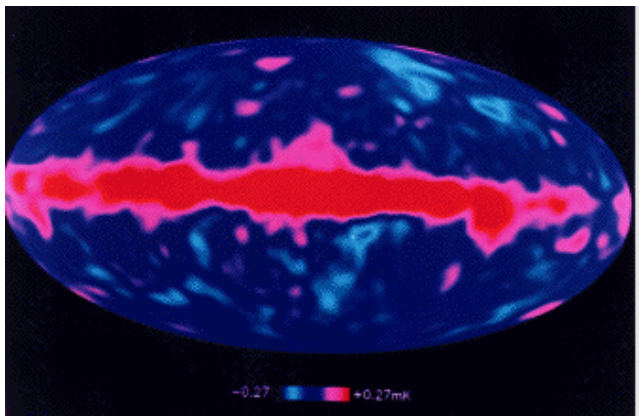


NERSC User George Smoot wins 2006 Nobel Prize in Physics



Mather and Smoot 1992

COBE Experiment showed anisotropy of CMB



Cosmic Microwave Background Radiation (CMB): an image of the universe at 400,000 years

Office of Science
U.S. DEPARTMENT OF ENERGY

CMB Computing at NERSC

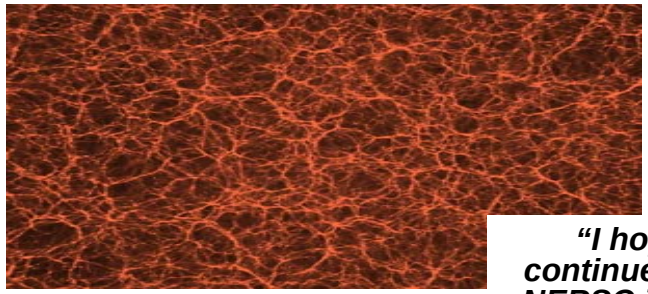
- CMB data analysis presents a significant and growing computational challenge, requiring
  - well-controlled approximate algorithms
  - efficient massively parallel implementations
  - long-term access to the best HPC resources
- DOE/NERSC has become the leading HPC facility in the world for CMB data analysis
  - about 1,000,000 CPU-hours/year for the last 4 years
  - 6.25 TB project disk space
  - 300 TB HPSS data
  - about 10 experiments and 100 users

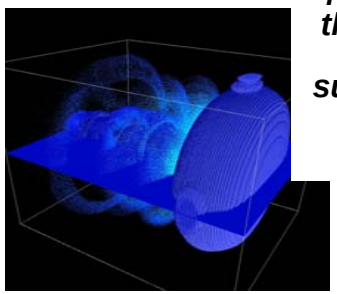
CMB is Characteristic for Large-Scale Projects at NERSC

- Petaflop/s and beyond computing requirements
- Algorithm and software requirements
- Use of new technology, e.g. NGF
- Service to a large international community
- Exciting science

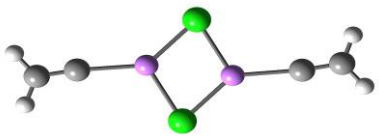
INCITE Program at NERSC 2006

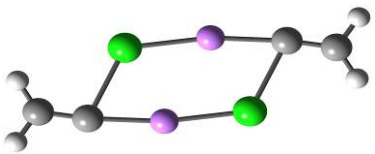
- Precision Cosmology Using the Lyman Alpha Forest – Mike Norman, SDSC
  - Increase understanding of the dark energy and dark matter
  - Large memory requirements needed 256 seaborg nodes, 10TB disk space
  - Generated > 60 TB data
  - Systems group and consulting staff shepherd through large debug jobs to track down memory consumption of code
- Particle-in-Cell Simulation of Laser Wakefield Particle Acceleration – Cameron Geddes, LBNL
  - Produce detailed 3D models of laser-driven wakefield particle accelerators
  - 4.6 million hours devoted to full 3D higher resolution simulations
  - Clarify mechanisms for beam formation and evolution
- Reactions of lithium carbenoids, lithium enolates, and mixed aggregates – Larry Pratt, Fisk University
  - Investigate the structure and reactions of organolithium compounds
  - Simulations involving electron correlation change equilibrium geometries significantly
  - Fine-grained parallelism requires powerful large-memory SMP nodes.





"I hope we can continue to work with NERSC in the future. I think your center did a fabulous job in supporting us. " Robert Harkness





Office of Science
U.S. DEPARTMENT OF ENERGY

Impact on Science Mission

Acknowledgments

A.A.G. wishes to thank Roland Assaraf for validating Zori against QMCMOL. A.A.G. also thanks Anthony Scemama for his contribution of electron pair localization function routines. Computer time was provided by the Department of Energy's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. D.D. was supported by the CREST Program of the National Science Foundation under Grant No. HRD-0318519.

P. N. and D. K. acknowledge support from a NASA LTSA and ATP grant. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract DE-AC03-76SF00098. We thank them for a generous allocation of computing time under the "Big Splash" award, without which this research would have been impossible.

We thank the RHIC Operations Group and RCF at BNL, and the NERSC Center at LBNL for their support. This work was supported in part by the HENP Divisions of the Office of Science of the U.S. DOE; the

Acknowledgements

H.W. thanks the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC03-76SF00098, for the allocation of computer time. M.T. thanks W. Domcke, M. Gelin, A. Pisliakov, and G. Stock for numerous helpful discussions. This work has been supported in part by a

The authors are grateful to Prof. R. J. Bartlett and Dr. M. Musial, who very kindly computed the CISDTQ and CCSDTQ numbers reported in this work. G. K-L. Chan would also like to thank Prof. N. C. Handy, who, as always, pointed him in the right direction. Most of the computations were carried out at the NERSC supercomputer centre, via DOE grant 12345, and the NERSC staff (in particular D. Skinner) are thanked for their assistance in many technical matters.

Acknowledgements This work was supported by the US Department of Energy and the National Science Foundation and used resources of the National Energy Research Scientific Computing Center at LBNL; C.G. was also supported by the Hertz Foundation. C.G. acknowledges his faculty advisor J. Wurtele. We appreciate contributions from G. Dugan, J. Faure, G. Fubiani, B. Nagler, K. Nakamura, N. Saleh, B. Shadwick, L. Archambault, M. Dickinson, S. Dimaggio, D. Syversrud, J. Wallig and N. Ybarrolaza.

Office of Science
U.S. DEPARTMENT OF ENERGY

Impact on Science Mission

- Majority of great science in SC is done with medium- to large-scale resources
- In 2006, NERSC users reported the publication of 1437 papers that were based wholly or partly on work done at NERSC (http://www.nersc.gov/news/reports/ERCAPpubs06.php)

Acknowledgments

A.A.G. wishes to thank Roland Assaraf for validating Zori against QMCMOL. A.A.G. also thanks Anthony Scemama for his contri-

time was provided by the Department of Energy's Innovative and Novel

program. D.D. was supported by the CREST Program of the National Science Foundation under Grant No. HRD-0318519.

and ATP grant. This research used resources of the National

ported by the Office of Science of the US Department of

for a generous allocation of computing time under the "Big

impossible

We thank the RHIC Operations Group and RCF at BNL, and the NERSC Center at LBNL for their support. This work was supported in part by the HENP Divisions of the Office of Science of the U.S. DOE; the

H.W. thanks the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC03-76SF00098, for the allocation

M. Gelin, A. Pisliakov, and G. Stock for numerous helpful dis-

in part by a

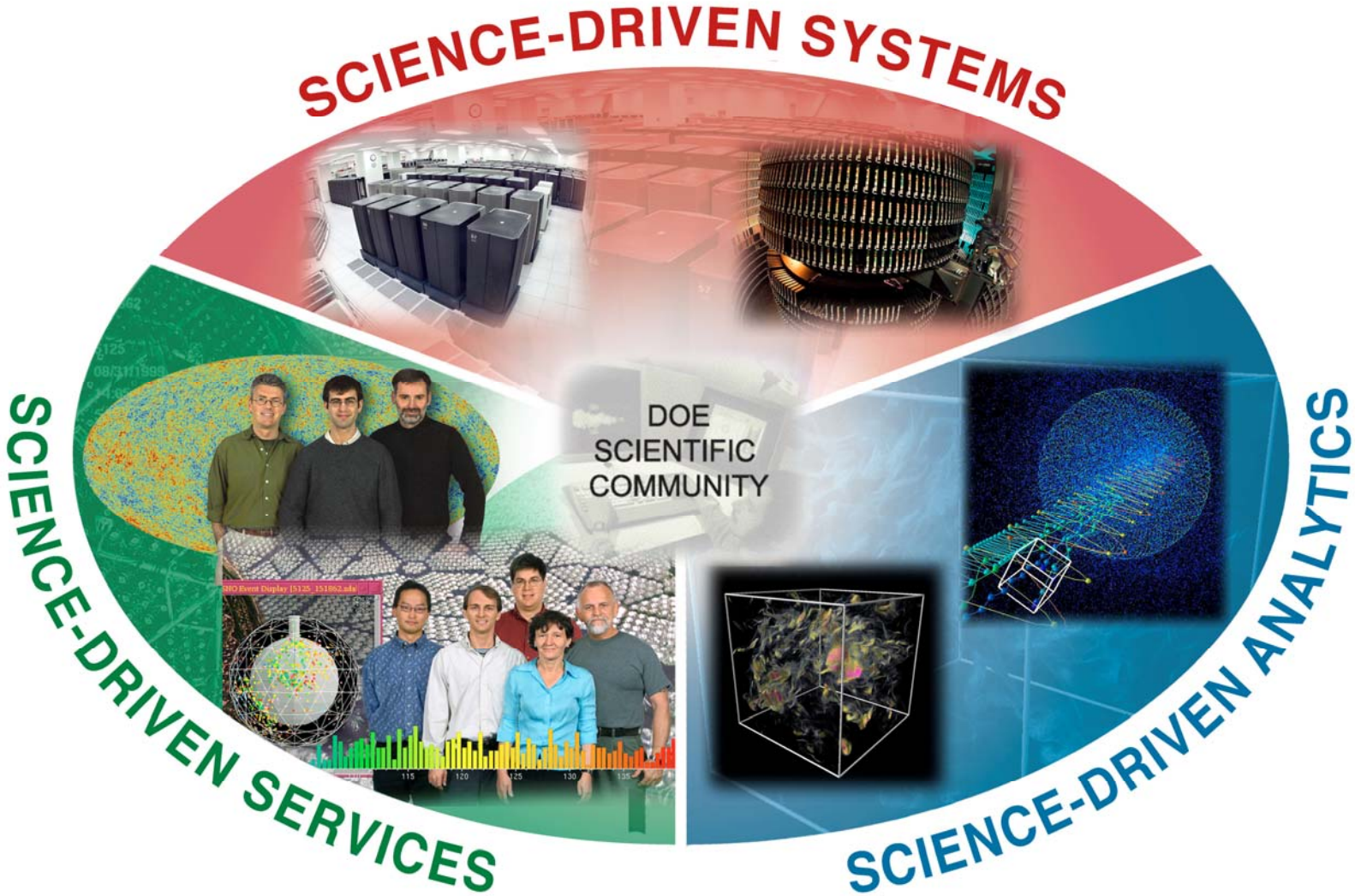
The authors are grateful to Prof. R. J. Bartlett and Dr. M. Musial, who very kindly computed the CISDTQ

would also like to thank Prof. N. C. Handy, who, as always, pointed him in the right direction. Most of the computations were carried out at the NERSC supercomputer centre, via DOE grant 12345, and the NERSC staff (in particular D. Skinner) are thanked for their assistance in many technical matters.

Acknowledgements This work was supported by the US Department of Energy and the National Science Foundation and used resources of the National Energy Research Scientific Computing Center at LBNL; C.G. was also supported by the Hertz Foundation. C.G. acknowledges his faculty advisor J. Wurtele. We appreciate contributions from G. Dugan, J. Faure, G. Fubiani, B. Nagler, K. Nakamura, N. Saleh, B. Shadwick, L. Archambault, M. Dickinson, S. Dimaggio, D. Syversrud, J. Wallig and N. Ybarrolaza.

Office of Science
U.S. DEPARTMENT OF ENERGY

Science-Driven Services



Science-Driven Services

- Provide the entire range of services from high-quality operations to direct scientific support
- Enable a broad range of scientists to effectively use NERSC in their research
- Concentrate on resources for scaling to large numbers of processors/cores, and for supporting multidisciplinary computational science teams

Science-Driven Services

- Consulting
  - One-on-one code tuning
  - Debugging
  - Software installation
  - Data manipulation advice: bbcp, gridftp, NGF, etc.
- Systems
  - Queue priority and time limits
  - Increased disk quotas
- Analytics
  - Large (INCITE/SciDAC) projects tend to produce the most output and often require large input datasets
- Networking
  - Network tuning, software installs (bbcp) to boost transfer bandwidth between NERSC and ORNL from 6.4 MB/s to 24 MB/s for combustion project
  - >60 TB transferred to SDSC using gridFTP at 25-55 MB/s for astrophysics project
  - Tuned network connections and replaced scp with hsi: transfer rate increased from 0.5 to 70 MB/s for astrophysics project

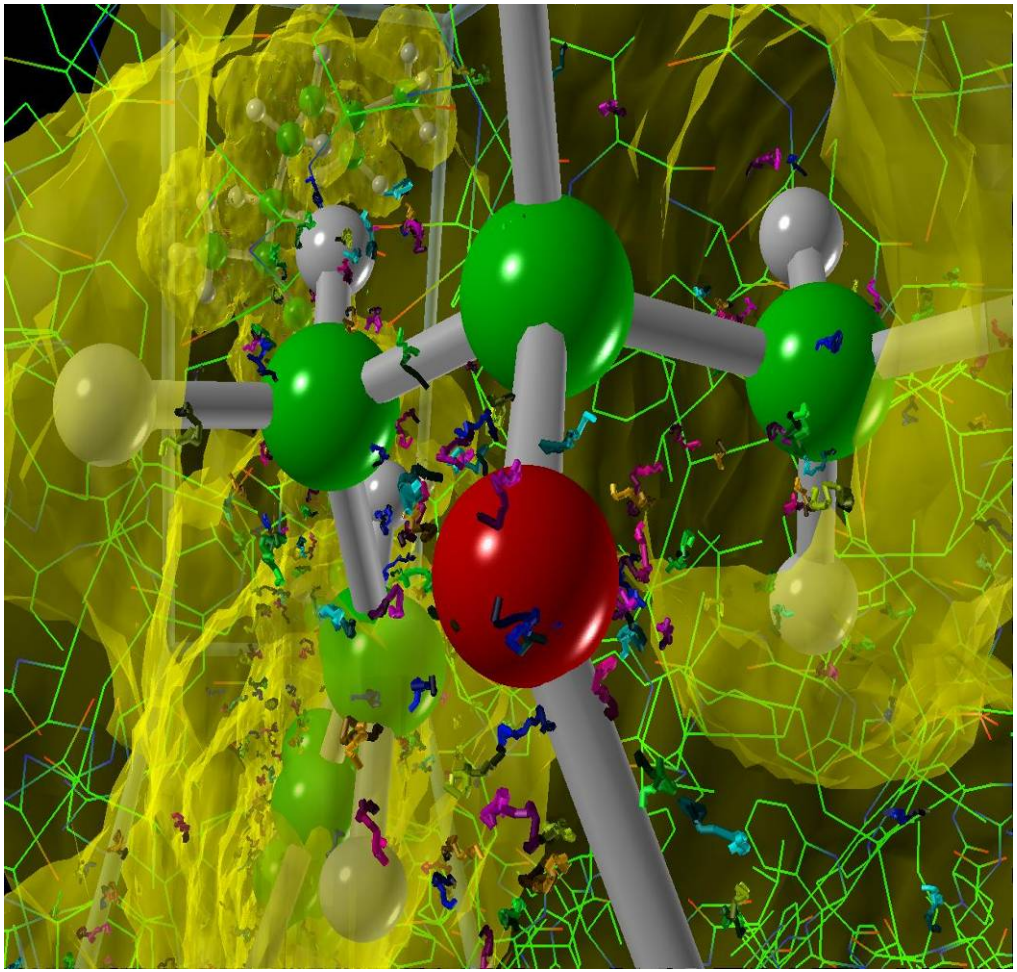
Example of Special Assistance

Photosynthesis Project

PI: William Lester, UC Berkeley

- **MPI tuning: 15-40% less MPI time**
- **Load balancing: scaling from 256 to 4,096 procs**
- **More efficient algorithm for random walk procedure**
- **Wrote parallel HDF5 I/O layer**

"We have benefited enormously from the support of NERSC staff."



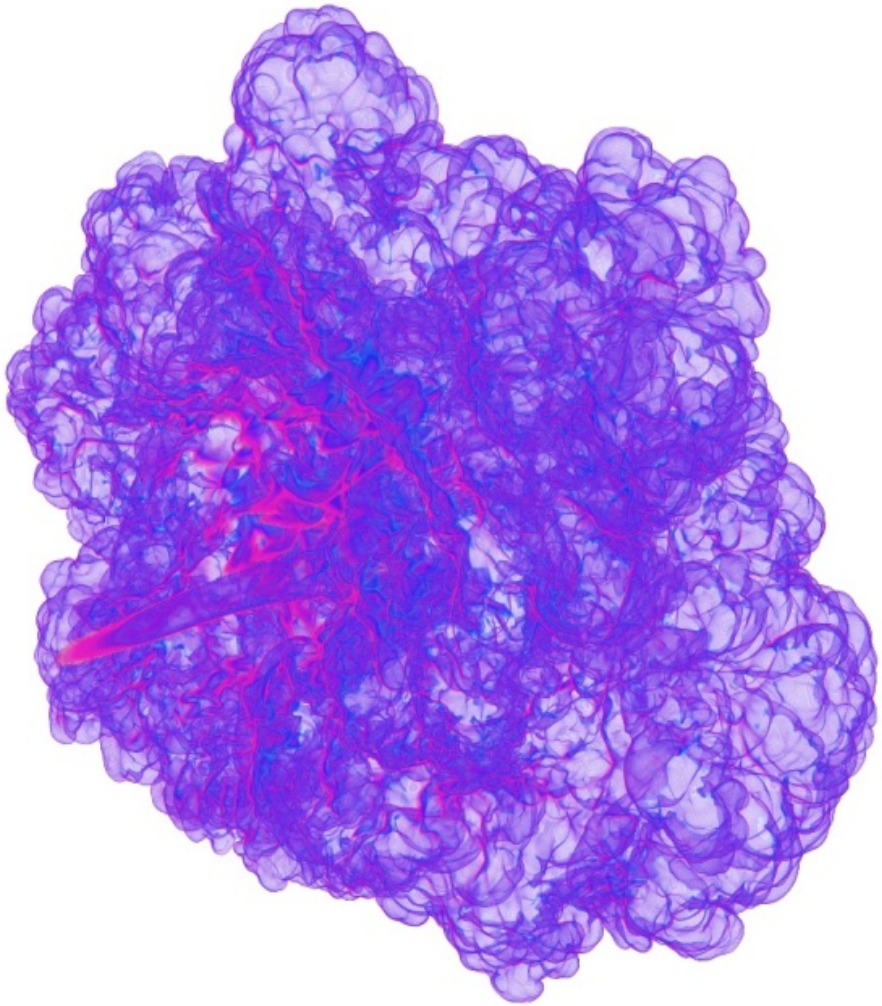
Example of Special Assistance

Thermonuclear Supernovae Project

PI: Tomasz Plewa, U. Chicago

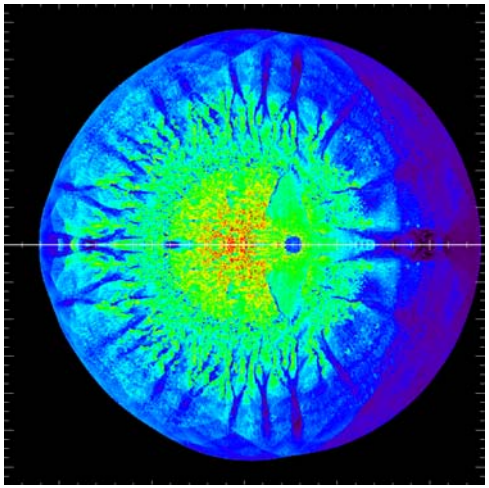
- **Resolved problems with large I/O by switching to a 64-bit environment**
- **Created automatic procedure for code check pointing**

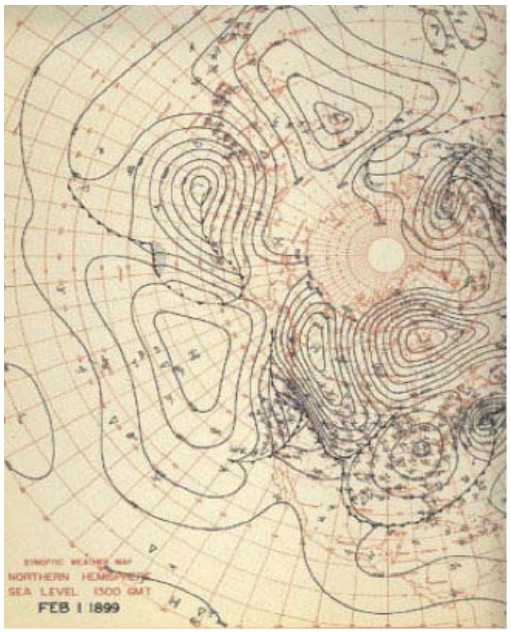
"We have found NERSC staff extremely helpful in setting up the computational environment, conducting calculations, and also improving our software."



Early Successes for INCITE 2007

- **Tuning the FLASH code for memory use to correct an error condition**
  - ***"We could not have asked for better or more support than we got from the folks at NERSC, in helping us to get on the NERSC machines quickly, in giving the job special status, and in helping us meet the challenges of running a large job on Bassi."***
    —Don Lamb
- **Tuning and debugging the global tropospheric circulation analysis code**
  - **Adding multi-level parallelism to bundle several associated parallel jobs**





Science-Driven Systems

Office of Science
U.S. DEPARTMENT OF ENERGY

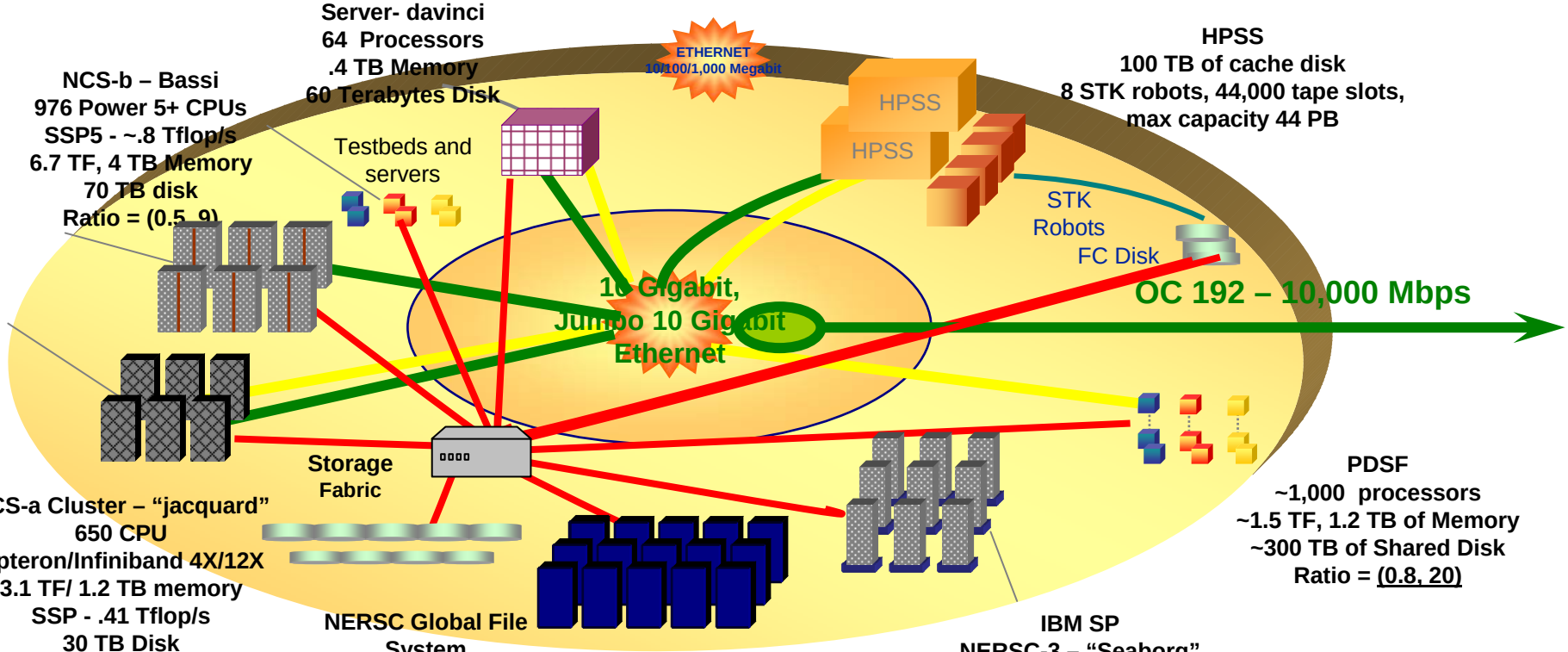
Science-Driven Systems

- Balanced and timely introduction of best new technology for complete computational systems (computing, storage, networking, analytics)
- Engage and work directly with vendors in addressing the SC requirements in their roadmaps
- Collaborate with DOE labs and other sites in technology evaluation and introduction

Computational System Strategy

- Balanced and timely introduction of the best new technologies for complete systems
  - Often the first and/or largest of its kind
- Computational systems address the widest breadth of DOE capability science
  - NERSC has two major computational systems in place at a time – called NERSC-n
    - NERSC's major computational systems arrive every three to four years
      - The oldest-generation system is replaced with the latest-generation system.
  - Modest-sized systems (NCSy) arrive between the major systems as funding and technology allows
    - Typically focus on a subset of the workload

NERSC Systems 2007



Visualization and Post Processing Server- davinci
64 Processors
.4 TB Memory
60 Terabytes Disk

Testbeds and servers

NCS-b – Bassi
976 Power 5+ CPUs
SSP5 - ~.8 Tflop/s
6.7 TF, 4 TB Memory
70 TB disk
Ratio = (0.5, 9)

HPSS
100 TB of cache disk
8 STK robots, 44,000 tape slots, max capacity 44 PB

PDSF
~1,000 processors
~1.5 TF, 1.2 TB of Memory
~300 TB of Shared Disk
Ratio = (0.8, 20)

NCS-a Cluster – "jacquard"
650 CPU
Opteron/Infiniband 4X/12X
3.1 TF/ 1.2 TB memory
SSP - .41 Tflop/s
30 TB Disk
Ratio = (.4,10)

NERSC Global File System
300 TB shared usable disk

Cray XT4
NERSC-5 - "Franklin"
SSP ~16 Tflop/s

IBM SP
NERSC-3 – "Seaborg"
6,656 Processors (Peak 10 TFlop/s)
SSP5 – .9 Tflop/s
7.8 Terabyte Memory
55 Terabytes of Shared Disk
Ratio = (0.8,4.8)

Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Office of Science
U.S. DEPARTMENT OF ENERGY

Franklin arrives, January 16, 2007

“Franklin”

Benjamin Franklin, one of America’s first scientists, performed ground breaking work in energy efficiency, electricity, materials, climate, ocean currents, transportation, health, medicine, acoustics and heat transfer.

Largest XT-4

9,740 nodes with 19,480 CPUs (cores)

102 Node Cabinets, 16 KWs per cabinet

39.5 TBs Aggregate Memory

16.1+ Tflop/s Sustained System Performance

Seaborg - ~.89

101.5 Tflop/s Theoretical System Peak Performance

Cray SeaStar2/3D Torus Interconnect (17x24x24)

6.3 TB/s Bi-Section Bandwidth

7.6 GB/s peak bi-directional bandwidth per link

50 Nanosecond per link latency

345 TBs of Usable Shared Disk

Sixty 4 Gbps Fibre Channel Data Connections

Four 10 Gbps Ethernet Network Connections

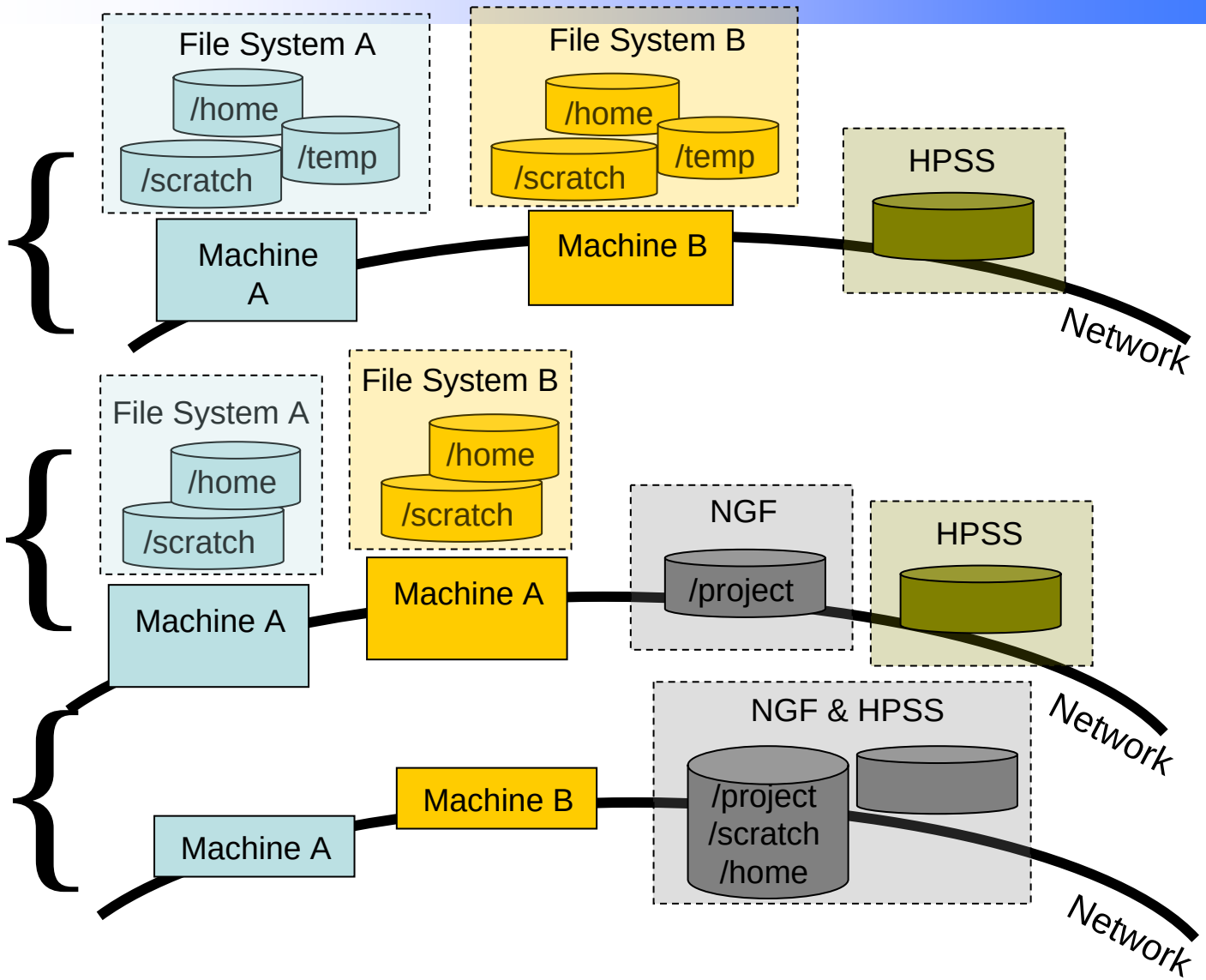
Sixteen 1 Gbps Ethernet Network Connections

2006: NERSC Global Filesystem (NGF) in Full Production

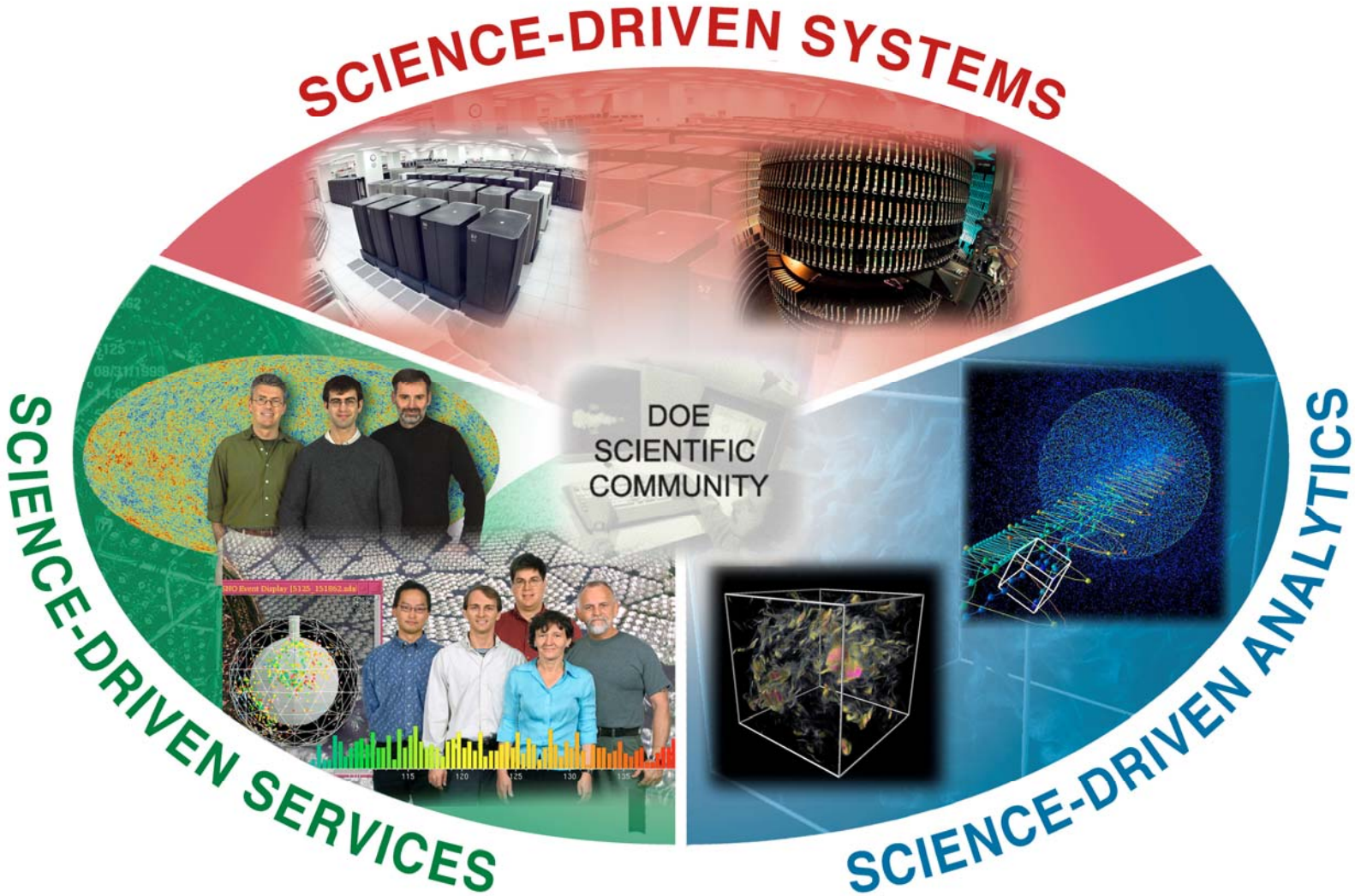
- After thorough evaluation and testing phase in production
- Based on IBM GPFS
- Seamless data access from **all** of NERSC's computational and analysis resources
- Single unified namespace makes it easier for users to manage their data across multiple system
- First production global filesystem spanning five platforms, three architectures, and four different vendors

NERSC Storage Roadmap

- **Past**
  - **Local Disk**
    - **/scratch**
    - **/home**
    - **/tmp**
  - **HPSS**
- **Now**
  - **Local Disk**
    - **/scratch**
    - **/home**
  - **NGF**
    - **/project**
  - **HPSS**
- **Future**
  - **Local Disk**
    - **/home**
  - **NGF-HPSS**
    - **/scratch**
    - **/project**



Science-Driven Analytics



Office of Science
U.S. DEPARTMENT OF ENERGY

What is the Analytics Program at NERSC?

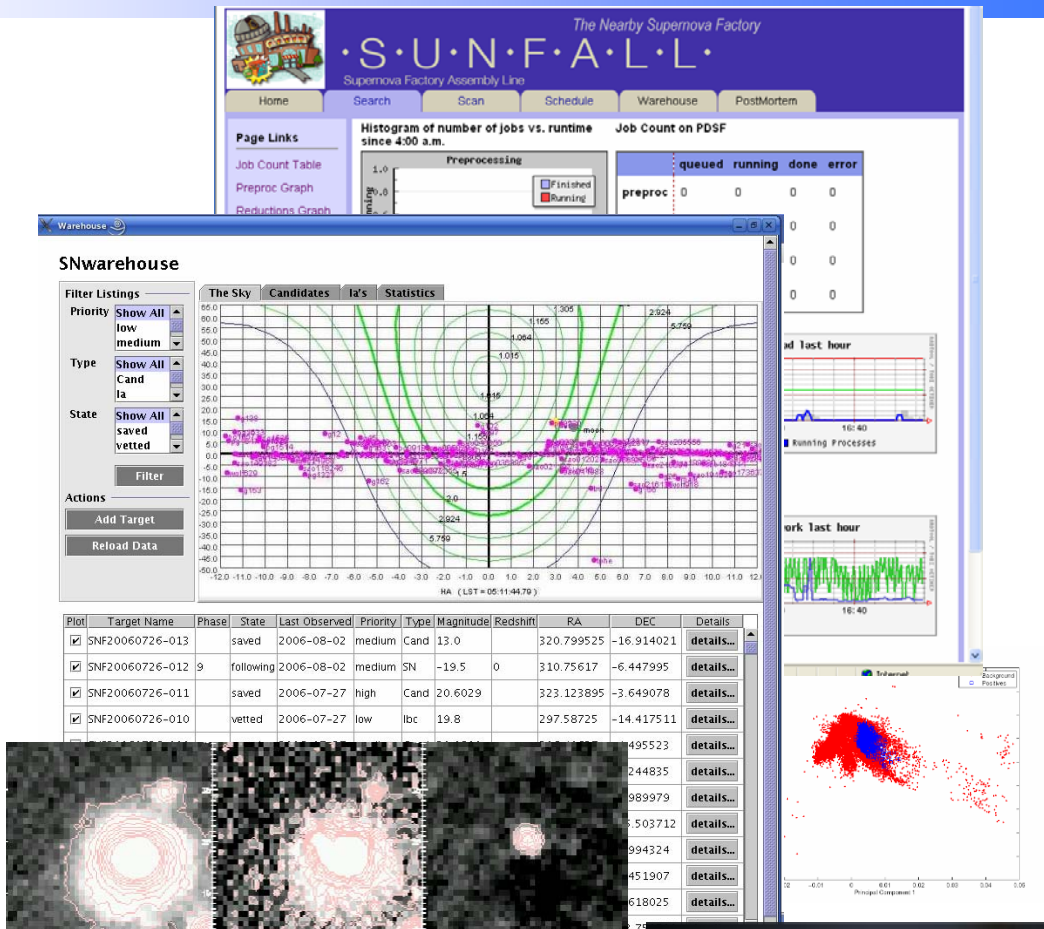
- Analytics is the *"science of analysis"*.
- At NERSC, the Analytics Program is the confluence of several key technologies:
  - Data management
    - Data storage/retrieval/sharing/movement, data indexing/querying
  - Data analysis and data mining
    - Compare datasets, find features within a dataset.
  - Data visualization of data
    - Primary method of data exploration.
  - Workflow management
    - Systematic approach to “data processing pipelines,” especially those that use multiple distributed resources and automate scientific data processing activities

NERSC Analytics Services

- A dedicated interactive analysis platform that is well integrated with the rest of the Center
- Software applications, libraries, etc. that are the building blocks for Analytics solutions
- In-depth, collaborative help to science stakeholders to implement effective analytics solutions
  - There is no “one-size-fits-all” solution due to the diversity of stakeholder problems
- Serving shared licenses to NERSC users
- Technology path-finding to stay abreast of latest developments in the field, including interactions with the CS research community

Production Analytics – SNFactory

- New supernova data analysis and workflow visualization tools (Sunfall and SNwarehouse) have improved usability and situational awareness, and enabled faster and easier access to data for supernova scientists worldwide
- Advanced image processing (Fourier contour analysis) and machine learning techniques running on NERSC platforms have achieved a >40% decrease in human workload in nightly supernova search (>75% FTE)



Summary

- NERSC supports a diverse range of requirements and science
- NERSC systems, services and analytics are highly regarded and valuable to the DOE science community
- NERSC is helping to create the highly successful systems of the future