BLUE WATERS

SUSTAINED PETASCALE COMPUTING

May 4, 2010

The Crisis in Massive Storage

Dr. William Kramer

Blue Waters Deputy Director

NCSA NSF IBM

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Data

A Beautiful Disaster?

© AP

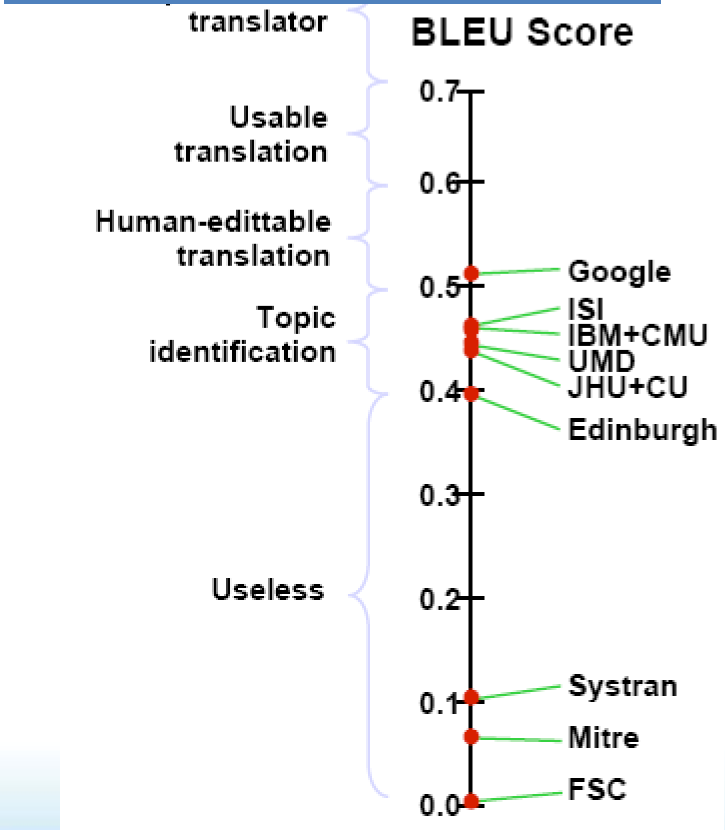
A Frightening Crisis?

The “Data Deluge” is like the “Boy Who Cried Wolf”

Easy Access to Data Give Advantages

- Being able to access and use data gives advantage
  - Science
  - Business
  - Competitiveness
- Simple contest resulted in Google Translate –
  - “Google's free online language *translation* service instantly translates text and web pages”
- The web/cloud generation expect immediate access to any information desired

Arabic translation: Google with more data beats others with more specialists

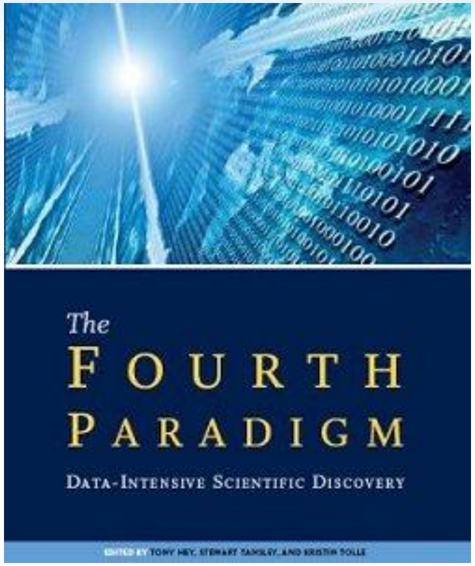


Is the Crisis Really Here This Time - Yes

- Exponential increase in sensors
  - Big Sensors
    - Limited in number – produce impressive amounts of organized data
    - Examples
      - Accelerators – O(10) world wide
      - Telescopes – O(100) world wide
      - Satellites – O(1000) world wide
  - Small Sensors
    - Relatively small amounts – but very large numbers
    - Examples
      - Environment Sensing
      - Building and Structure Sensing
      - Ships and Planes
      - Medical and life science lab instruments and testing
      - Video Surveillance

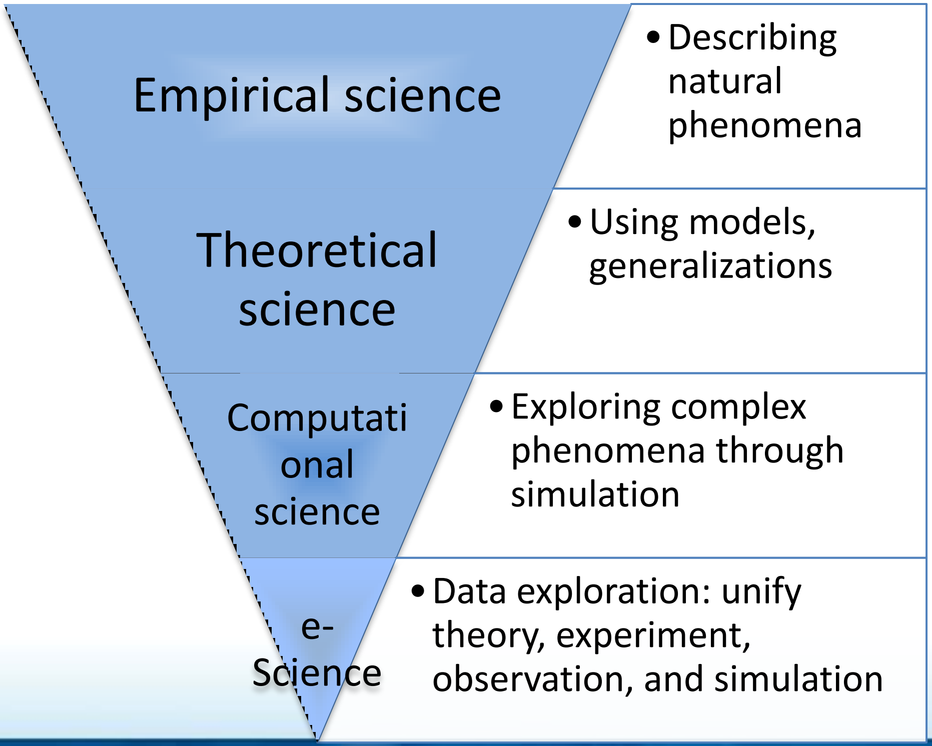
Is the Crisis Really Here This Time - Yes

- Increased Data Assimilation
- Increased coupling with simulation
- The “e-clouders”


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Jim Gray concluded economic necessity mandates putting the data near the application, since the cost of wide-area networking has fallen more slowly (and remains relatively higher) than all other IT hardware costs - Distributed Computing Economics. Queue 6, 3 (2008), 63–68

- Management and integration of data early into the science life cycle


Empirical science

- Describing natural phenomena

Theoretical science

- Using models, generalizations

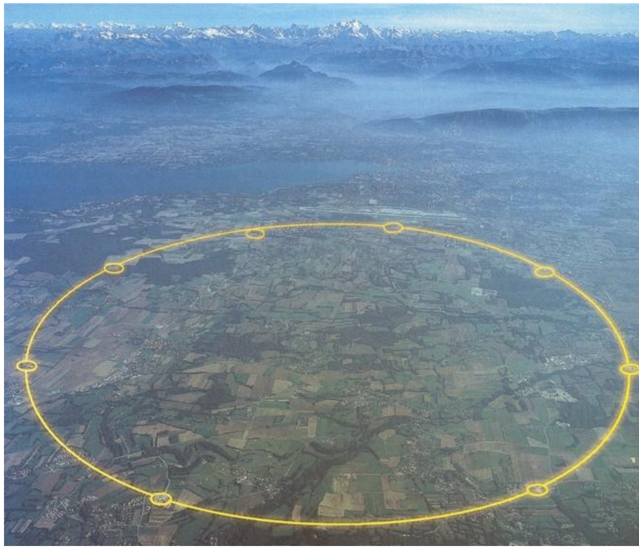
Computational science

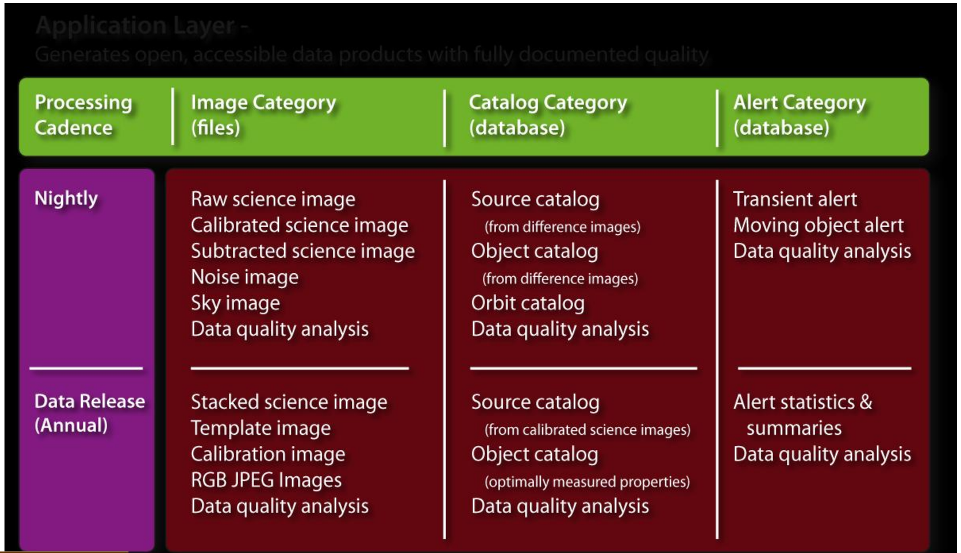
- Exploring complex phenomena through simulation

e-Science

- Data exploration: unify theory, experiment, observation, and simulation

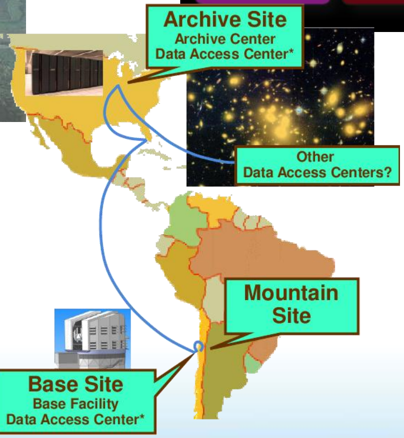
Big Sensors – LHC, LSST and SKA





Application Layer -
Generates open, accessible data products with fully documented quality

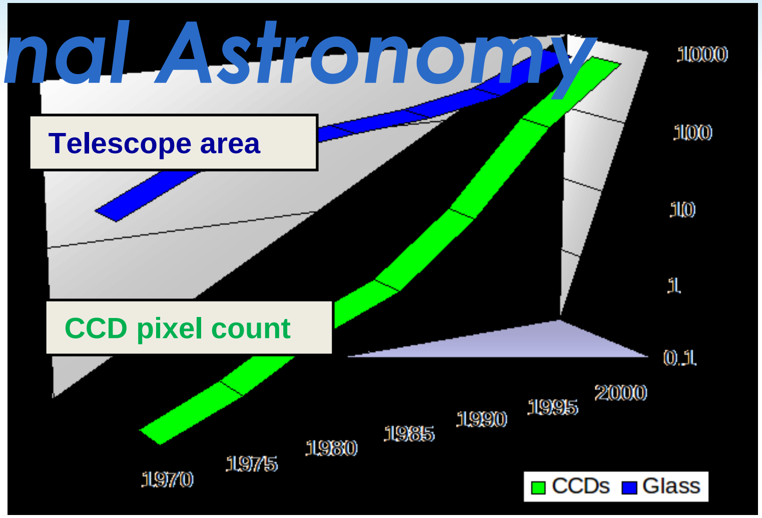
Processing Cadence	Image Category (files)	Catalog Category (database)	Alert Category (database)
Nightly	Raw science image Calibrated science image Subtracted science image Noise image Sky image Data quality analysis	Source catalog (from difference images) Object catalog (from difference images) Orbit catalog Data quality analysis	Transient alert Moving object alert Data quality analysis
Data Release (Annual)	Stacked science image Template image Calibration image RGB JPEG Images Data quality analysis	Source catalog (from calibrated science images) Object catalog (optimally measured properties) Data quality analysis	Alert statistics & summaries Data quality analysis



Data rates in observational Astronomy

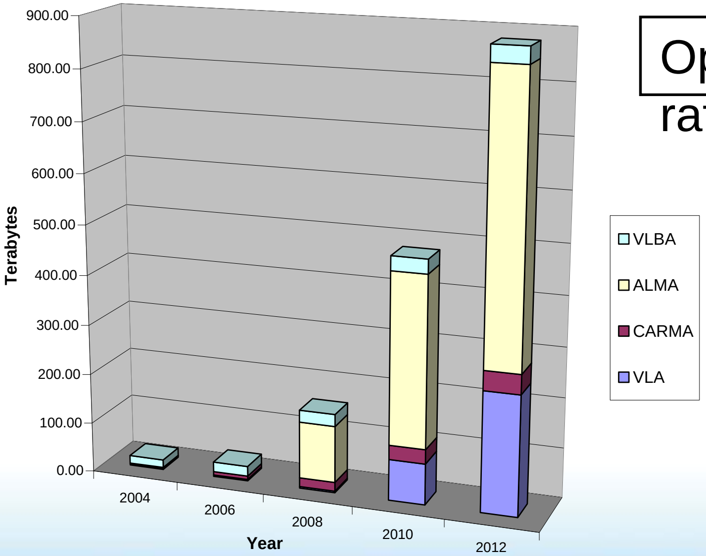
- Data rates are driven by:
  - Contemporary astrophysics questions require surveys of large cosmic volumes
  - Moore's Law advances in detector counts and data output
  - Increasingly sophisticated data processing needed.

- Data rates are **exponential** and require fundamentally new approaches to data management and processing.



(Szalay & Gray)

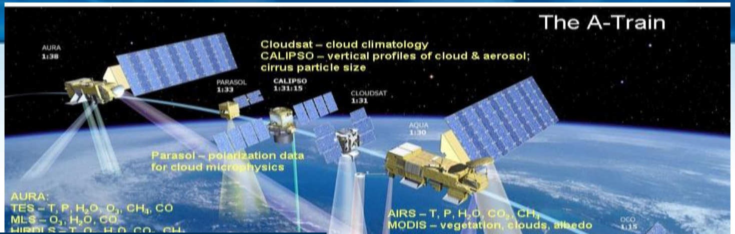
Optical data rates



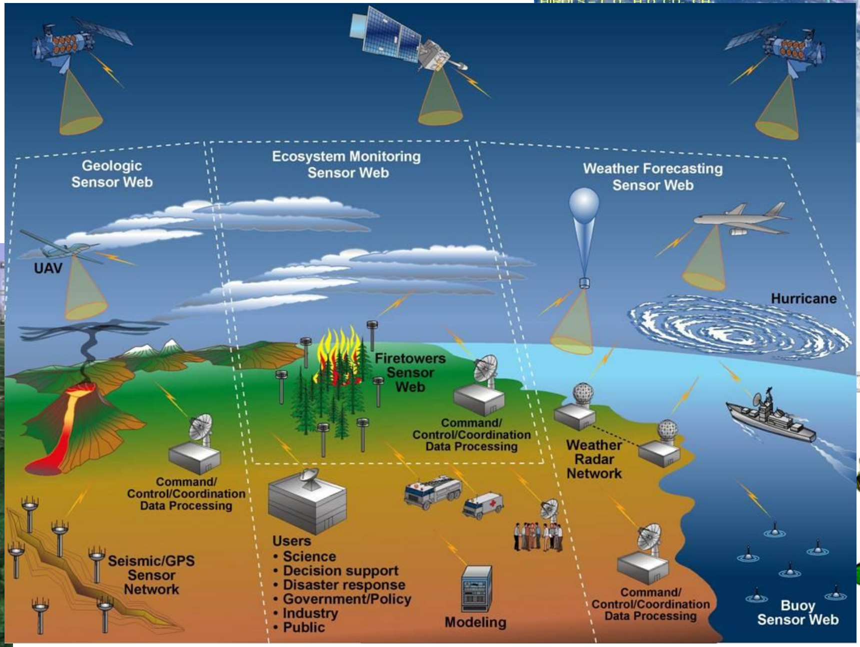
Radio data rates

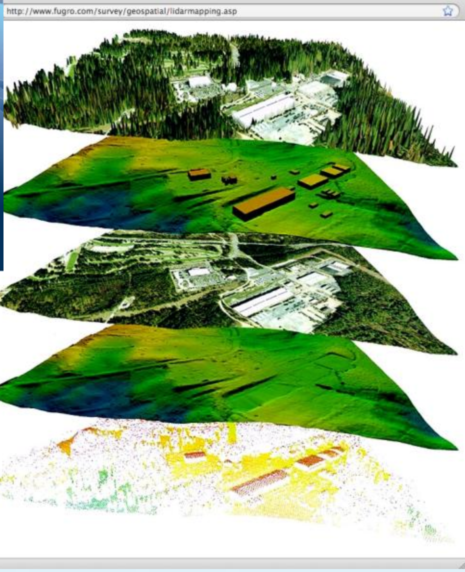
Slide courtesy of

Little Sensors – Geophysics Data Drivers


The A-Train
AURA 1:38
Cloudsat – cloud climatology
CALIPSO – vertical profiles of cloud & aerosol; cirrus particle size
PARASOL 1:33
CALIPSO 1:31:15
CLOUDSAT 1:31
AQUA 1:30
Parasol – polarization data for cloud microphysics
AURA:
TES – T, P, H2O, O3, CH4, CO
MLS – O3, H2O, CO
AIRS – T, P, H2O, CO2, CH4
MODIS – vegetation, clouds, albedo
OCO 1:15
OCO-XCO2 P(surface) T, H2O, cloud, aerosol

- Satellite
- Lidar

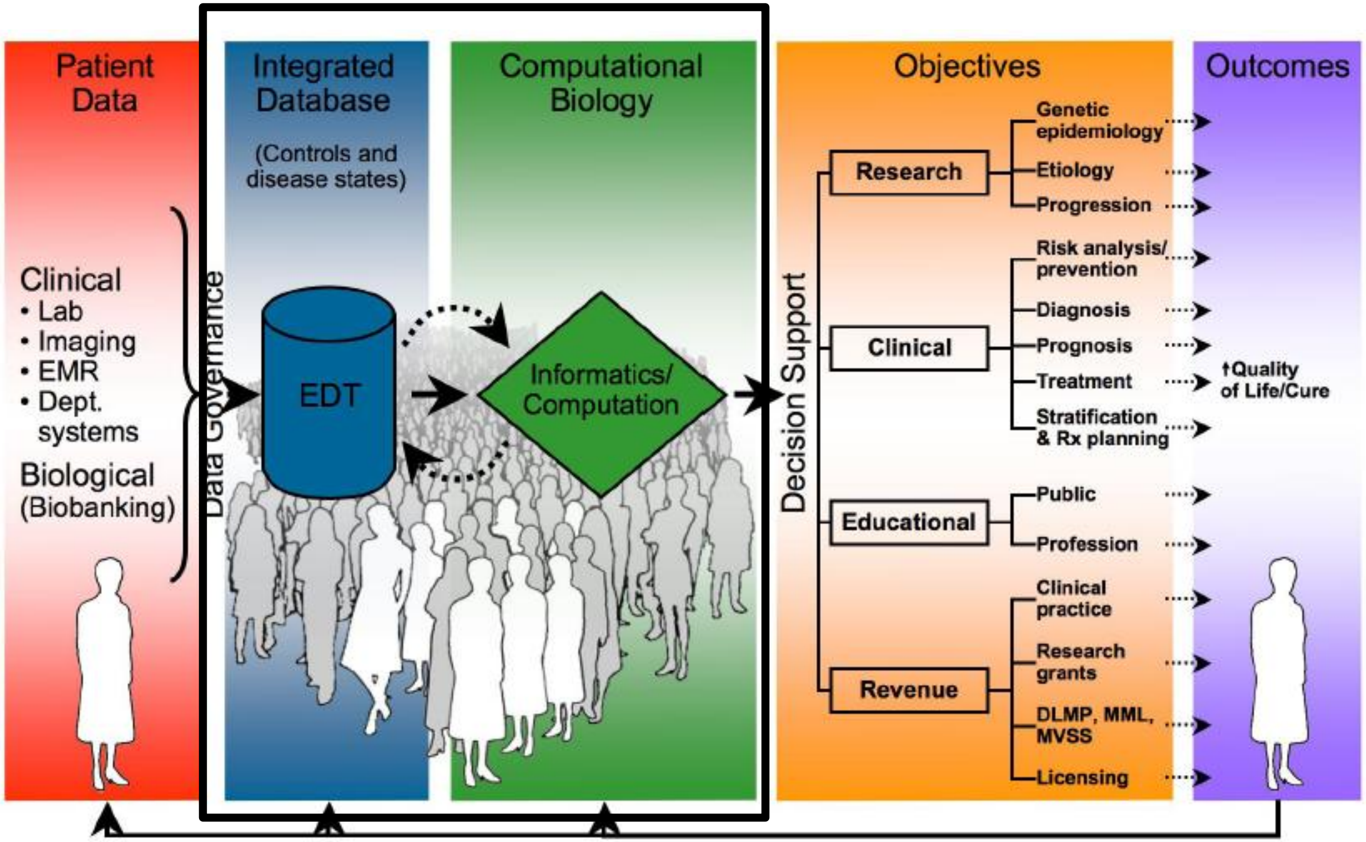

Geologic Sensor Web
Ecosystem Monitoring Sensor Web
Weather Forecasting Sensor Web
UAV
Hurricane
Firetowers Sensor Web
Command/ Control/Coordination Data Processing
Weather Radar Network
Command/ Control/Coordination Data Processing
Seismic/GPS Sensor Network
Users
• Science
• Decision support
• Disaster response
• Government/Policy
• Industry
• Public
Modeling
Command/ Control/Coordination Data Processing
Buoy Sensor Web



- Sensor webs
- GIS

Images Courtesy of Prof Praveen Kumar, Dept of Civil and Environmental Engineering, University of Illinois

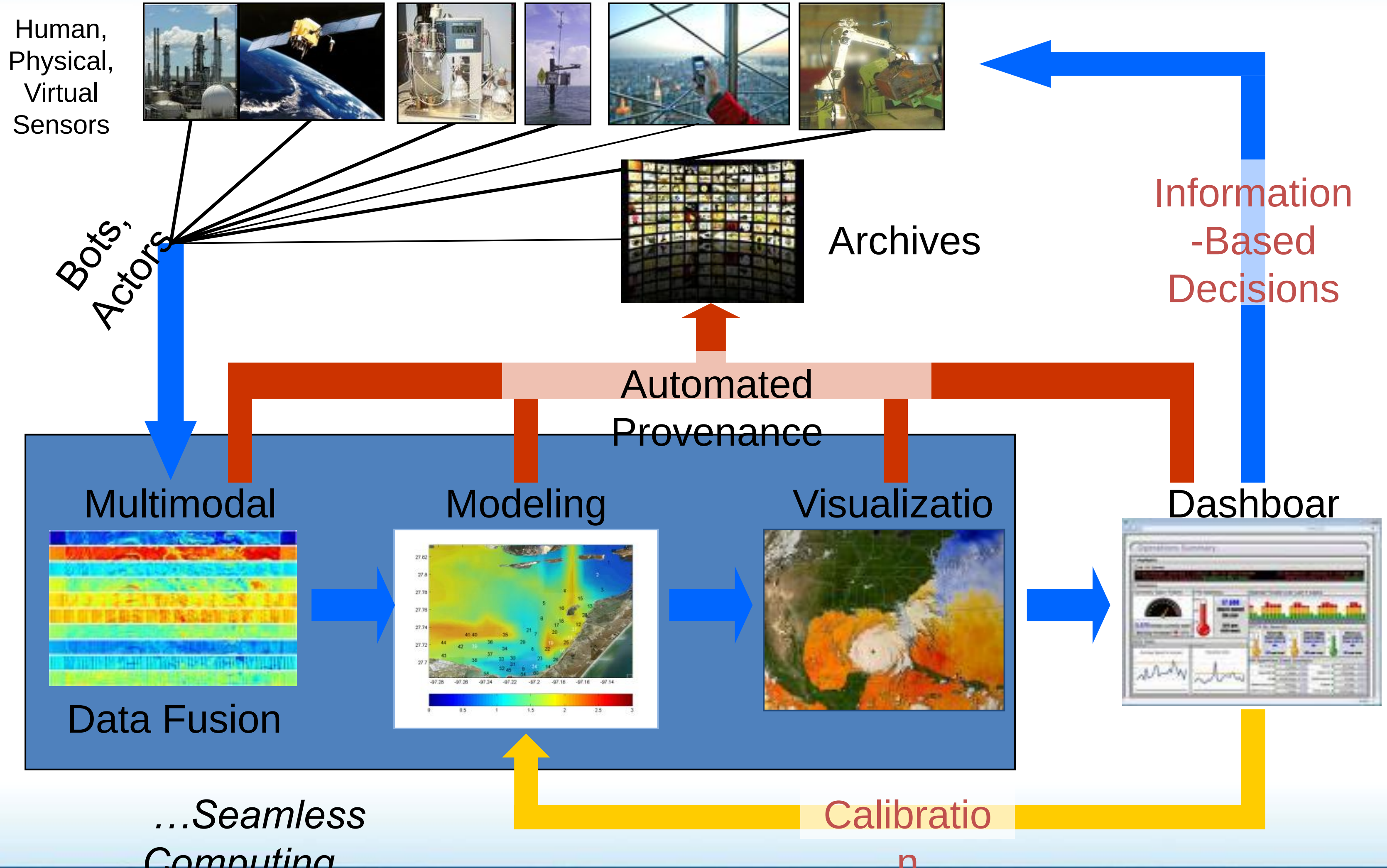
Little Sensors - Information-based medicine



Life Science – Medical “Opportunities”

- Data management solutions for high-throughput biology
  - “Next generation” sequencing currently generates tens of TB raw data per experiment, steep increases likely due to technical improvements in instruments 2-4x Moore’s Law
  - Other technologies are also rapidly increasing output: proteomics with prior spatial and/or chemical separation, high-throughput high-resolution imaging, epigenomics…
- Understanding the relationship between genotype and phenotype
  - Rapidly increasing production of full genome sequences from individuals within one species (mostly human) and from different species; millions of differences observed, thousands of genomes being sequenced
  - Identifying genomic determinants of phenotypic differences is a major data mining / statistical problem

Integrated, Information-Based Decision Making



Human, Physical, Virtual Sensors

Bots, Actors

Archives

Information -Based Decisions

Automated Provenance

Multimodal

Data Fusion

Modeling

Visualizatio

Dashboar

Calibratio
n

…Seamless Computing…

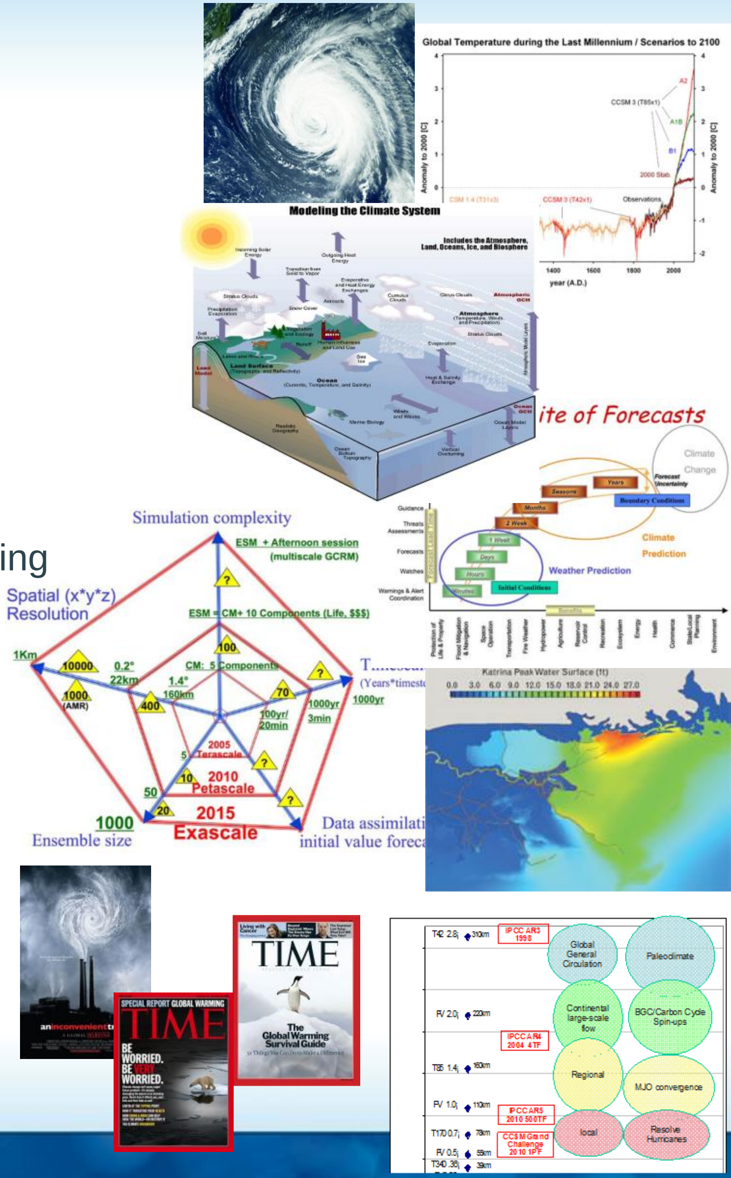
Sources include: Heidelberg Collaboratory, Minsker, GMES

Science Examples

Climate Science

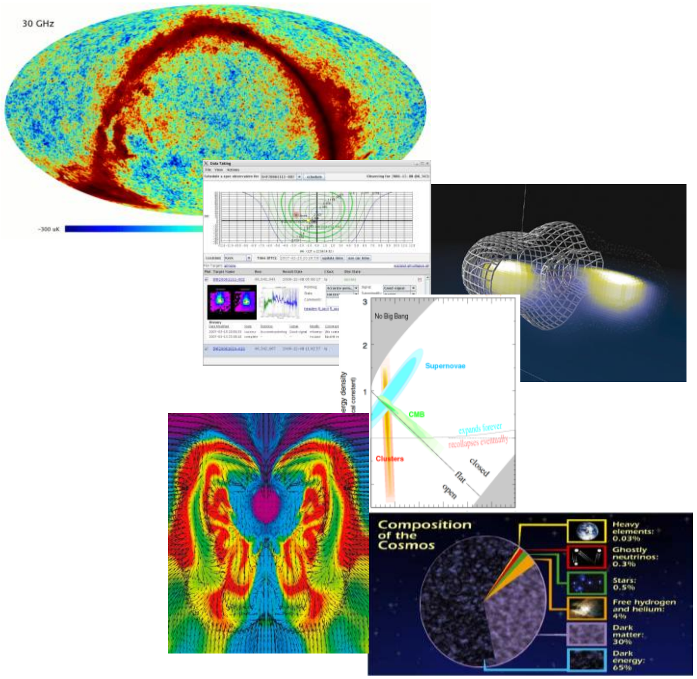
- Today: IPCC 4th Assessment Report - US Effort
  - Open Source - NCAR Community Climate System Model (CCSM-3)
    - Development effort: ~1 person-century
    - ~1 quadrillion operations/simulated year
    - 11,000 model years simulated with “T85” resolution
    - Rate of simulation: 3.5 simulated years/day
    - Data volume for IPCC: ~110 TB
  - Impact: Wide public acknowledgement of the global warming and policy change
- Near Term
  - Ensemble calculations - many runs to study mitigation
  - Regional impacts
  - Severe weather studies and mitigation
  - 1 PB on-line storage for a year
- **CCSM estimated storage for just IPCC assessments**
  - AR4 2004-2005 100 TB
  - AR5 2010-2011 1000 TB
  - AR6 2016-2017 10000 TB

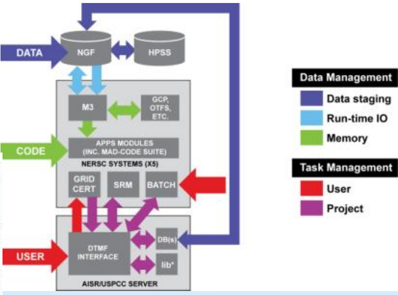
  (Above assume 1 degree ocean and 1 month output)

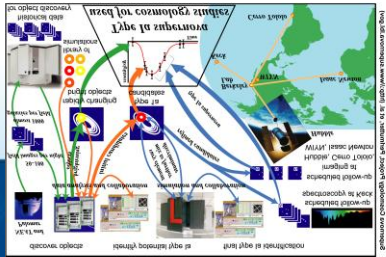


Science Examples *The Fate of the Universe*

- Today:
  - CMB Analysis (2006 Nobel Prize)
  - Supernova Factory
    - Find and examine in detail up to 300 nearby Type Ia supernovae - Discovered 34 supernovae during first year (more than entire pas) and now discovering 8-9 per month
      - First year: processed 250,000 images
      - Archived 6 TB of compressed data
  - Gravity Wave Search (Ligo, Cactus)
  - Supernova Simulation
  - Impact: The realization we can see only 5% of the universe = Dark Energy
- Upcoming
  - Analysis JDEM satellites and experiments
    - PLANCK
      - 1.5 billion pixels (12 GB) per sample - 2.5TB
        - >10,000 cores for computation

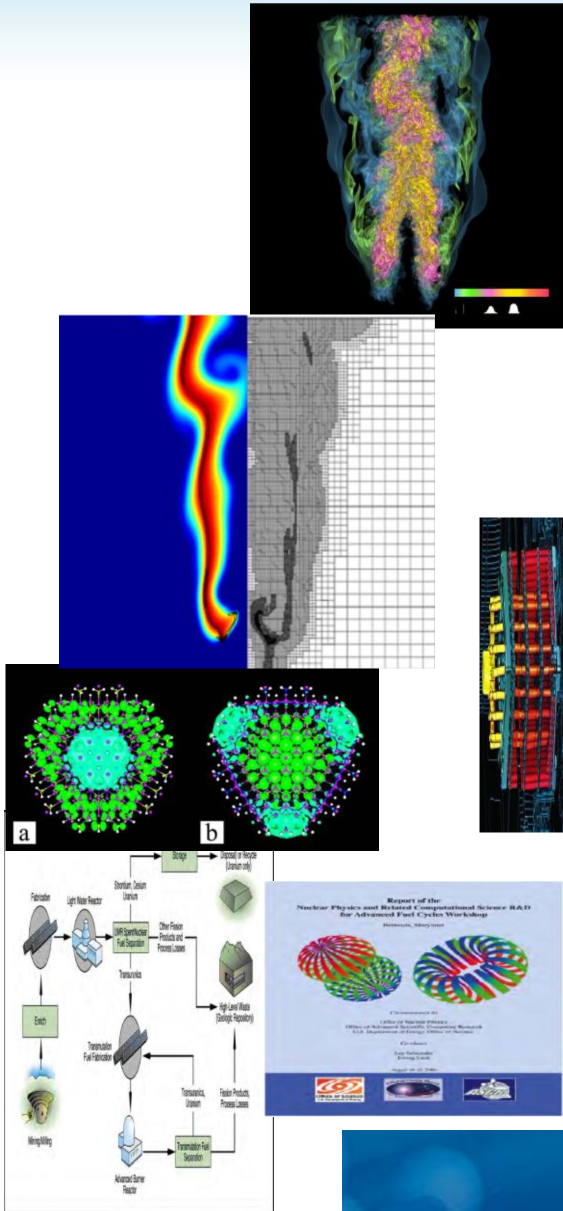






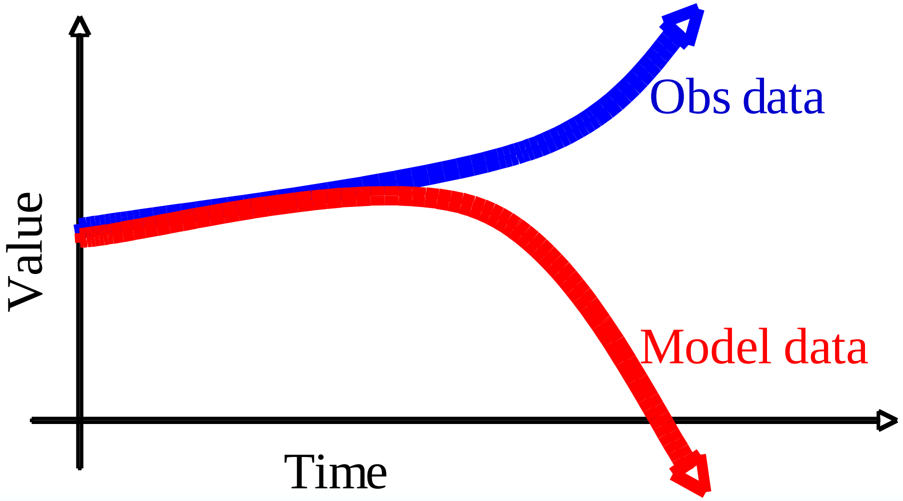
Science Example *Energy*

- Today
  - Combustion - 85% of US Energy usage
    - Today’s simulations need tremendous data resources
      - 0.3% of NERSC computational resources used 400-500 TB
      - 2% of needs 26 TB on-line for parallel access at all times
    - Coarse 3-D or high resolution 2-D
  - Nuclear Fusion
      - Design for ITER
- Future Goals
  - Combustion
    - High Fidelity 3-D, many species DNS and LES simulations using AMR
  - Nuclear Fusion
    - Operational ITER - Seven International Partners - 30 sites
    - ~2,000 test “shots” per year 1 TB raw data per test shot
    - Remote control rooms
  - Energy Efficiency
  - Solar and Alternative Energy
  - Nuclear Fission


a
b
Storage
Disposal or Recycle (Uranium only)
Fabrication
Light Water Reactor
Strontium, Cesium Uranium
LWR Spent Nuclear Fuel Separation
Other Fission Products and Process Losses
Transuranics
Enrich
High-Level Waste (Geologic Repository)
Transmutation Fuel Fabrication
Mining/Milling
Transuranics, Uranium
Fission Products, Process Losses
Advanced Burner Reactor
Transmutation Fuel Separation
Report of the
Nuclear Physics and Related Computational Science R&D
for Advanced Fuel Cycles Workshop
Bethesda, Maryland

Data Value Proposition

- The volume and complexity of observational data is overshadowing data from simulation
  - LHC
  - ITER
  - JDEM/SNAP
  - PLANCK
  - SciDAC
  - Genomic Program
  - Earth Systems Grid

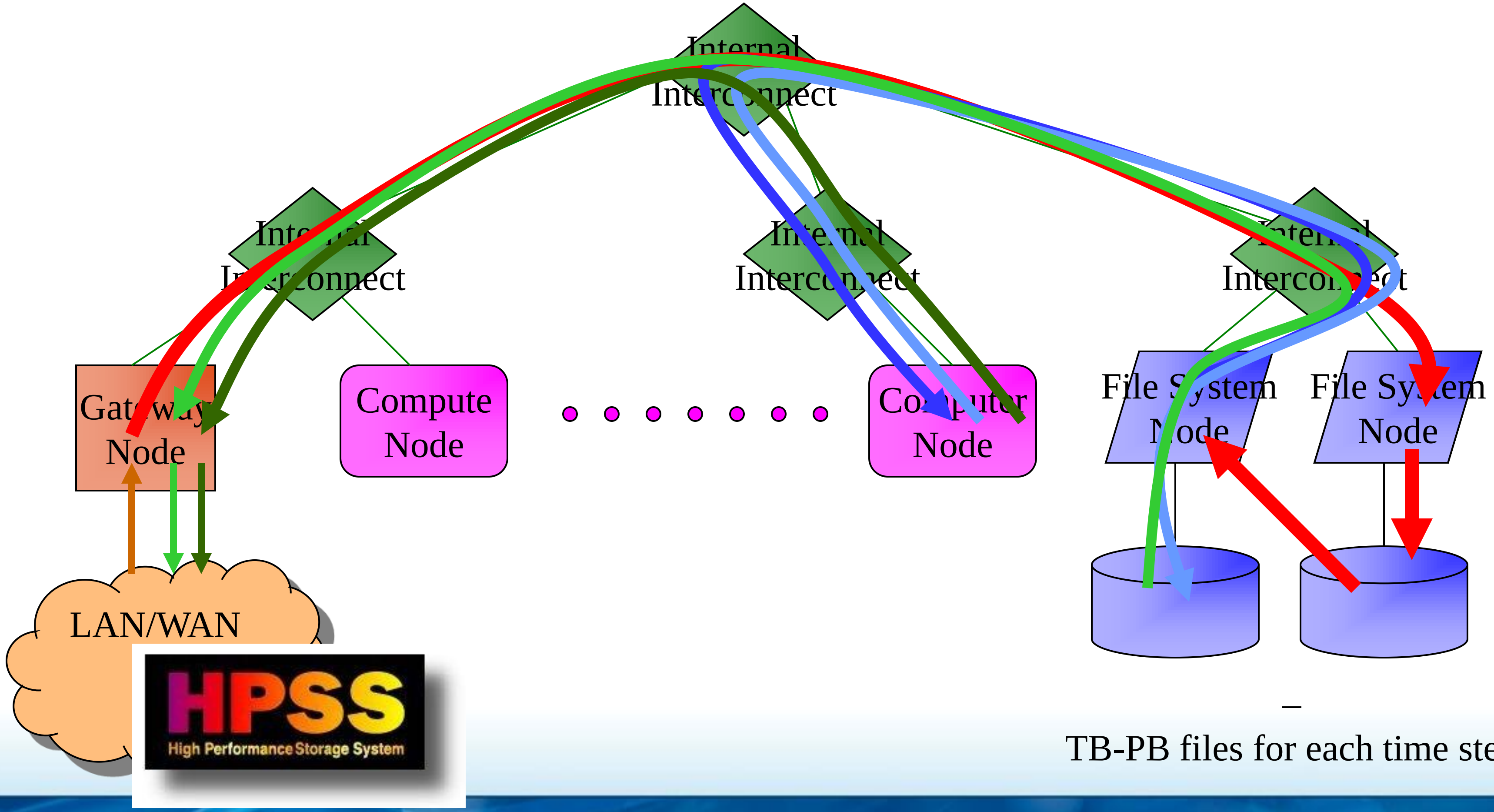


Courtesy L. Buja

Current State of the Practice Has Not Changed Much

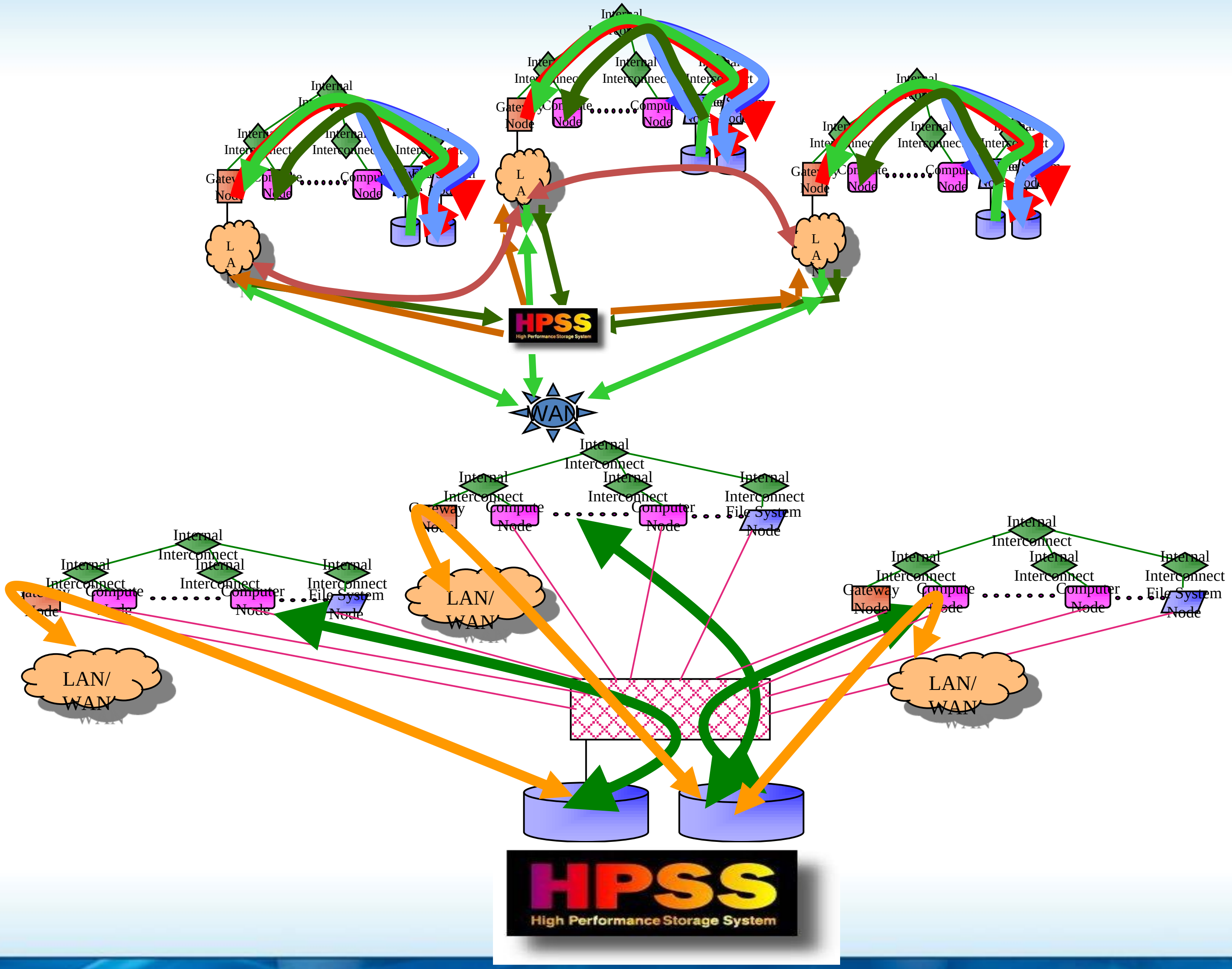
- Storage and I/O Strategies are a significant limitation to systems and science
- Locally attached disk is decreasing and generally not user accessible
- SANS systems for most parallel storage systems
  - FC, IB at the high end
  - Separate Disk Controllers
  - Range of disk types
- NFS still very common for low end commodity clusters – limitation of performance
  - pNFS not yet settled
  - Ethernet interconnect is common in data intensive computing farms – also a major limitation
- Parallel File systems
  - Production
    - GPFS, Lustre (supported), Pansas, cxfs
  - Experimental
    - PVFS, others
- Linear (tape storage)
  - “Today, storage silos and tape farms of various sorts are keeping up with the1.7-1.9 CAGR” for several more years – Kogge DARPA report

Data Divergence Problem



–

TB-PB files for each time step


Internal Interconnect
Gateway Node
Compute Node
File System Node
L A N
HPSS
High Performance Storage System
WAN
Internal Interconnect
Gateway Node
Compute Node
Computer Node
File System Node
LAN/ WAN
LAN/ WAN
LAN/ WAN
HPSS
High Performance Storage System

Current State of the Practice

- Application I/O strategies are primitive for the most part and inefficient
- Many systems are fundamentally limited for data access
- Science Teams spend O(FTEs) managing files and data
  - The philosophy of many storage systems is to cause users sufficient *pain so* they won’t store all that darn data
  - Or – 99% of the data is never used so why store it
  - Science teams focus on porting, Core performance and science – who has time for better I/O
- More science is becoming fundamentally limited by data access
  - HEP/NP and Genomics are the poster childs – but many other areas are impacted
  - Synergy of simulation and data assimilation is contributing to the issues
- Exascale planning is not really concentrating on I/O

A major motivation for the current *concentration movement* (aka *the cloud*) is the concentration of resources around data to reduce large scale data transfer

Disk Capacity - Transfer Rates Diverge

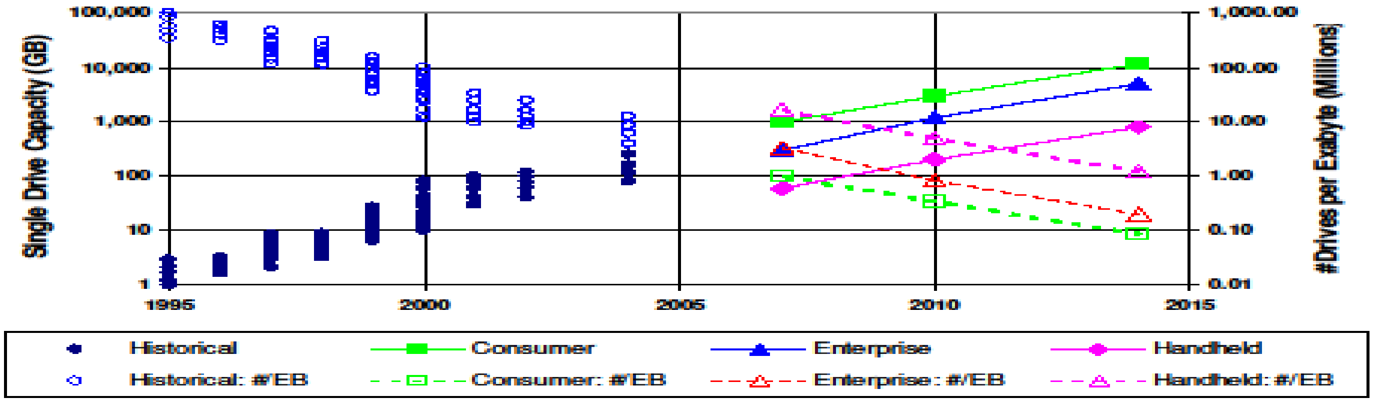


Figure 6.28: Disk capacity properties.

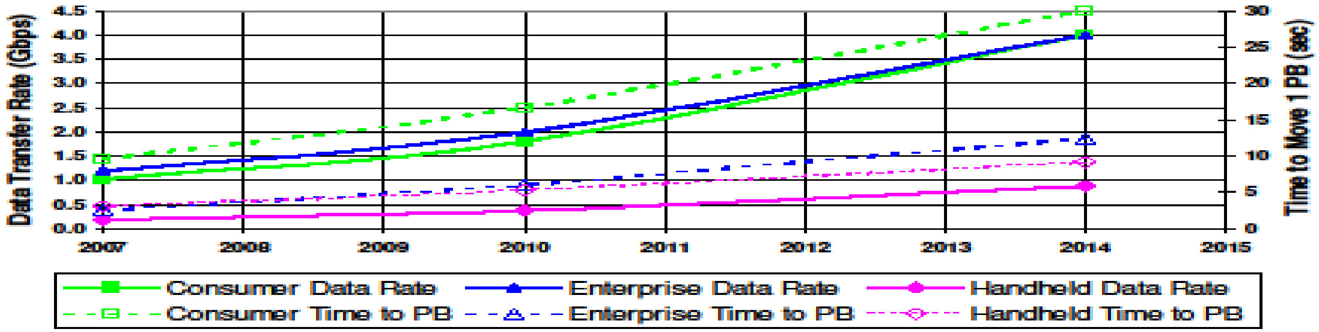


Figure 6.30: Disk transfer rate properties.

- Disk Capacity Grows at 10x / 6 years
- Disk transfer rate grows 3x-4.5x / 6 years
- Seek Rates (latency) are flat

User Desired Attributes Of Science Driven Data Subsystems

- Exa-scale Data for Petaflops/s systems
  - Zeta Scale Data for Exa-scale computation
  - Data at PB/s per system
- Within an HPC facility (a single administrative domain)
  - Automated data summarization, subsetting and feature recognition
  - Automated data movement controlled by facility policy and transparent to clients
    - Multiple storage device layers
    - Multiple storage fabrics
  - Uniform Namespace
    - Tightly coupled at each resource data layer
    - Coupled between layers
  - Highly parallel
    - Many processes/threads to many files – not what we want
    - Many processes/threads to one file
    - One process/thread to many files
  - Near equipment bandwidth and latency
    - Provisioned by Hardware - SW should have minimal impact on performance
  - Scalable metadata services
    - Not just creates/second - also stats per second
  - Sophisticated Analysis and Data management tools
  - Common and tuned Middle ware I/O libraries
- External access to the HPC facility

Example PS++ Use Case – *The Scientists' Data Cart*

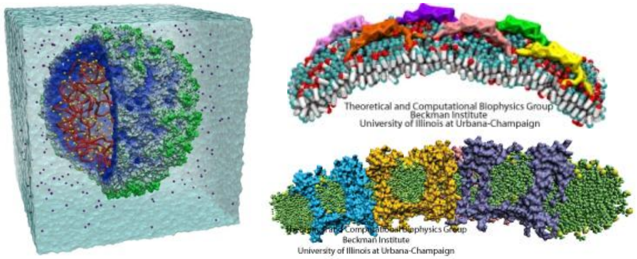
- Data Search and Retrieval
  - Web-based "Data Shopping"
  - Search "wizard" for typical and fast searches
  - Free form search and query for complex searches such as:
    - "Simulations using CCSM 3.0 with a resolution of T85 or greater run between January and May 2007"
    - "Supernova simulations for Type Ia US between January and May 2007"
  - Results determined by client roles
  - Data in different classes of annotations and persistence.
  - Put data "objects" (files) into their "data cart".
    - Carts have suites of actions such as download, compare, move for computation, visualize…
    - A "data cart" API defined so clients can plug in their own actions for analyzing and manipulating the data.
  - Similar functions will exist on the computational platforms in the form of shell scripts and tools.

Example PS++ Use Case - The Metadata Wizard

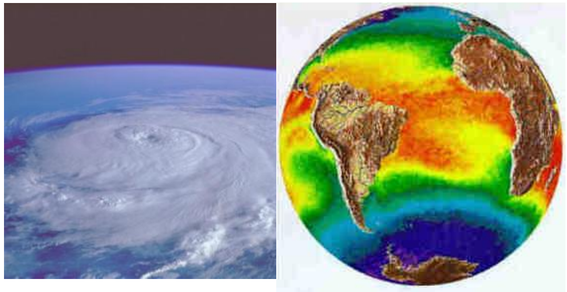
- Data Creation, Importation and Categorization
  - Default meta data defined
    - Example attributes (really in XML)
      - Originator, Source of data {simulation code, experiment, observation…), Sharing role {public, collaborative scientist, system-wide, project only, creator defined access control list, none), Formatting {flat, HDF5, netCDF…),Dates of creation and categorization, Data life time {temporary, N months (N=12?), infinite), Data integrity (i.e. {alternate copy location = [none, COS-dual, remote site…]}, {number of copies = [1,2, … N]})
  - Simulation data could be automatically annotated with creation time environment parameters
    - Standard methods used so data created at one site can be used at other sites
  - The Metadata wizard could annotate other data whenever metadata is not already associated.
    - The Wizard would be both web and command line based.
    - The Wizard extracts metadata automatically, and/or informs clients that metadata needs to be supplied when an automated process is insufficient. Site specific context and application specific translation may be feasible.
  - Data may flow from simulations done on other systems as well as from observations (satellites, ground sensing, etc.)
  - Must be infrastructure - not problem specific

Sustained Petascale computing will enable advances in a broad range of science and engineering disciplines:

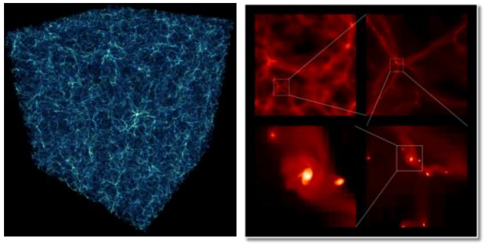
Molecular Science


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign
Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

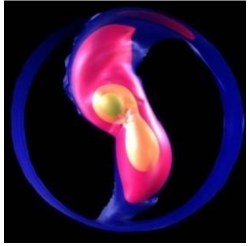
Weather & Climate Forecasting



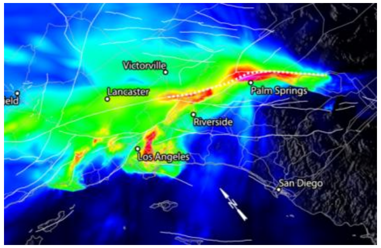
Astrophysics



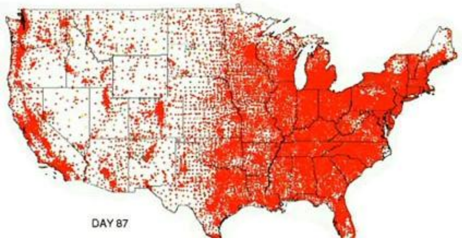
Astronomy



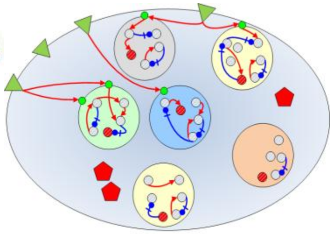
Earth Science



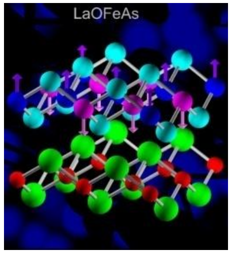
Health



Life Science



Materials



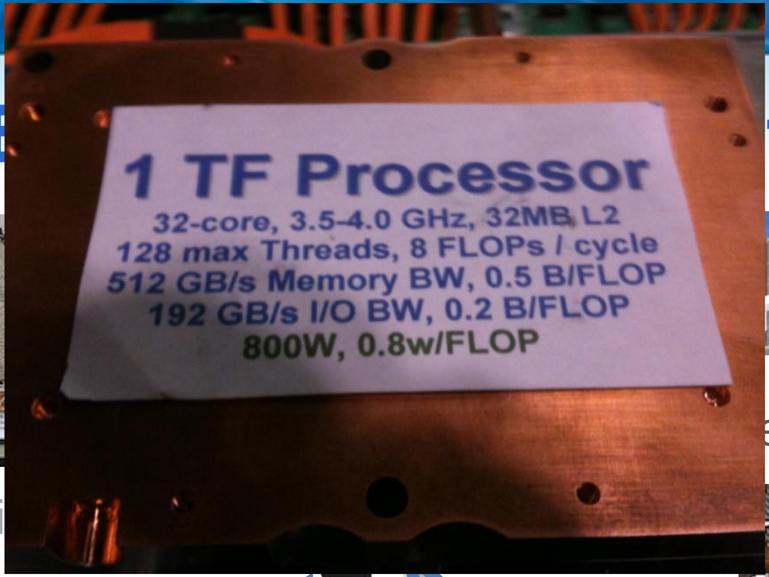
NSF Petascale Computing Resource Allocation (PRAC) Awardees

PIs	Field	Institutions
Schulten	Bio-molecular Dynamics	Illinois
Sugar	Quantum Chromodynamics	UC-Santa Barbara
O’Shea	Early galaxy formation	MSU
Nagamine	Cosmology	UNLV
Bartlett	Parallel language, Chemistry	U. FL
Bisset, Brown, Roberts	Social networks, Contagion	VA Tech, CMU, Research Triangle Inst.
Yeung	Turbulent flows	GA Tech.
Zhang	Materials science	Wm. & Mary
Wilhelmson	Tornadoes	Illinois

NSF Petascale Computing Resource Allocation (PRAC) Awardees(Cont'd)

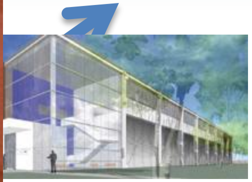
PIs	Field	Institutions
Jordan	Geophysics	U. So. CA
Lamm	Chemistry	IA St. U.
Woodward	Stellar hydrodynamics	U. of MN
Campanelli	General relativity, compact binaries	Rochester Inst. Tech.
Stan, Kirtman, Large, Randall	Climate	COLA (MD), U. Miami, UCAR, CO St. U.
Savrasov, Haule	Materials science	UC-Davis, Rutgers
Schnetter	Gamma-ray bursts	LSU
Tagkopoulos	Evolution	Princeton
Wang	Geophysics	U. of WY

From Chip to E tem



1 TF Processor

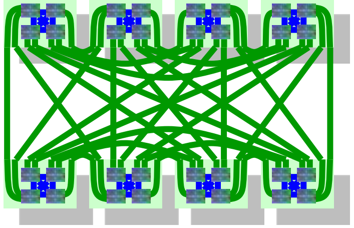
32-core, 3.5-4.0 GHz, 32MB L2
128 max Threads, 8 FLOPs / cycle
512 GB/s Memory BW, 0.5 B/FLOP
192 GB/s I/O BW, 0.2 B/FLOP
800W, 0.8w/FLOP



PCF

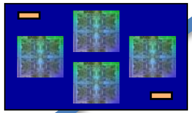
e Waters System

Rack/Bui

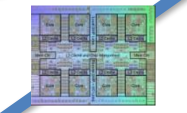


multiple MCMs

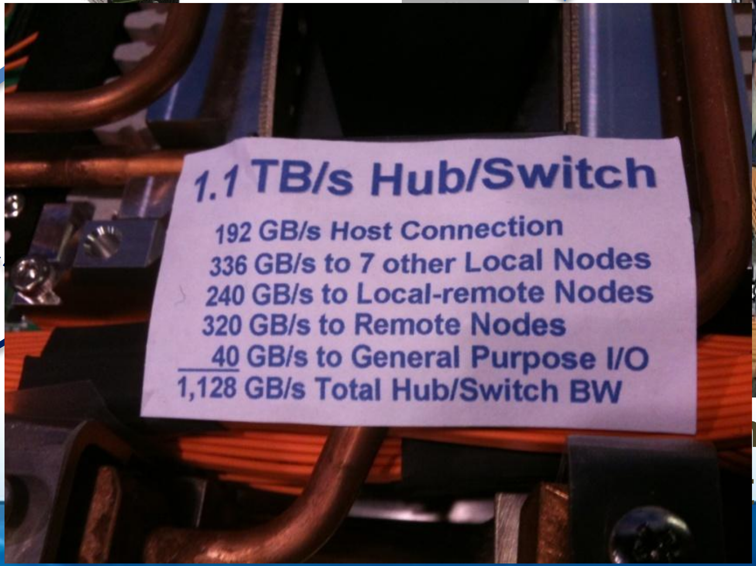
Near-li



Quad Chip MCM



Chip



1.1 TB/s Hub/Switch

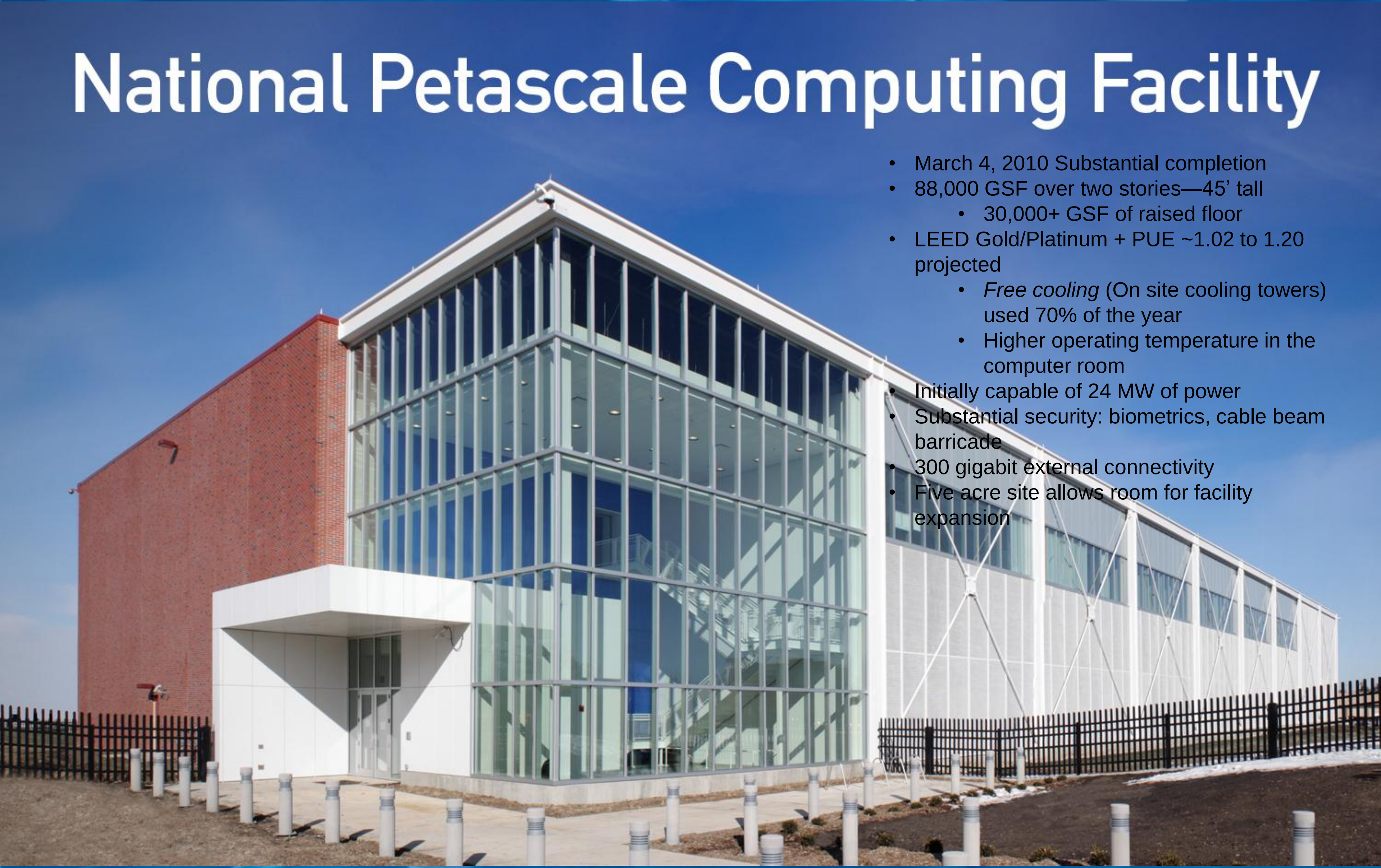
192 GB/s Host Connection
336 GB/s to 7 other Local Nodes
240 GB/s to Local-remote Nodes
320 GB/s to Remote Nodes
40 GB/s to General Purpose I/O
1,128 GB/s Total Hub/Switch BW

ndicates relative
t of public information

Blue Waters Computing System

System Attribute	Ranger		Blue Waters
Vendor	Sun		IBM
Processor	AMD Barcelona		IBM Power7
Peak Performance (PF)	0.579	17	>10
Sustained Performance (PF)	<0.05	>20	>1
Number of Cores/Chip	4	2	8
Number of Processor Cores	62,976	~3.5	>300,000
Amount of Memory (TB)	123	>8	>1
Interconnect Bisection BW (TB/s)	~4	>>10	
Amount of Disk Storage (PB)	1.73	>10	18
I/O Aggregate BW (TB/s)	?		1.5
Amount of Archival Storage (PB)	2.5 (20)	>200	>500
External Bandwidth (Gbps)	10	>10	100-400

National Petascale Computing Facility



- March 4, 2010 Substantial completion
- 88,000 GSF over two stories—45’ tall
    - 30,000+ GSF of raised floor
- LEED Gold/Platinum + PUE ~1.02 to 1.20 projected
    - *Free cooling* (On site cooling towers) used 70% of the year
    - Higher operating temperature in the computer room
- Initially capable of 24 MW of power
- Substantial security: biometrics, cable beam barricade
- 300 gigabit external connectivity
- Five acre site allows room for facility expansion

On-line File System is GPFS

- IBM is implementing scaling changes in GPFS for the HPCS/DARPA project.
- Blue Waters will implement those changes in a persistent manner
- GPFS configured to accommodate other local systems in a single namespace
- Performance requirements are appropriately scaled to BW characteristics

GHI

Near Line Storage is HPSS

- HPSS Hardware consists of three tape robots and appropriate numbers of tape drives
  - Expect to expand this thru the lifetime of BW

HPSS integrated with BW

  - GPFS-HPSS Interface
  - Import-Export Portal
    - Traditional HPSS commands

- NCSA is contributing RAIT implementation to the HPSS community as part of BW

NPCF Facility Wide File Systems

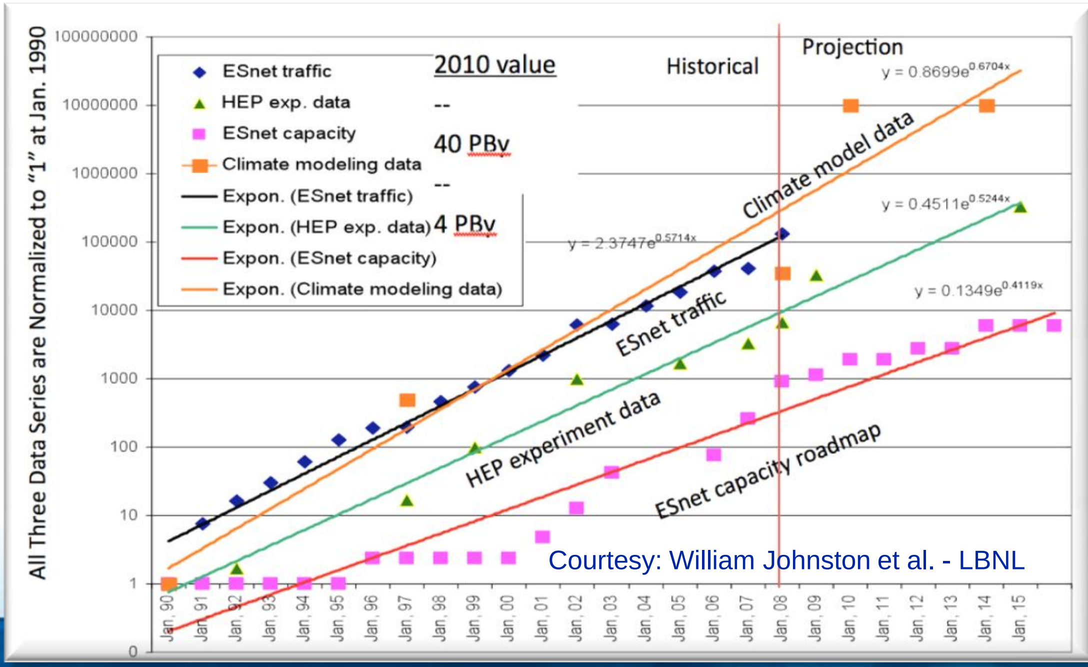
- Moving to a single global name space for all systems in a facility
- What is a FWFS? – a working definition
  - A production, facility-wide, high performance, parallel, shared file system
    - Makes scientific teams using systems more efficient and productive
    - Simplifies science team data management by providing a shared disk file system and single name space in production environment
    - Enables new paradigms
  - Global/Unified
    - A file system shared by all major systems without replication - N systems - M vendors
    - Uses consolidated storage and provides unified name space
    - Integration with Mass Storage
    - Integration with Grid is desired
  - Parallel
    - Provides performance that is scalable as the number of clients and storage devices increases
    - Performance very close to local parallel file systems

– Examples - NERSC- NGF (GPFS), ORNL- Spider (Lustre), DOD-(samfs)

CHALLENGES FOR THE MASS STORAGE COMMUNITY

Data Integration Challenges Facing Science

- **Models will generate more data in the near future than exist today**
- **How best to collect, distribute, and find data on a much larger scale?**
  - At each stage tools must be developed to improve efficiency
  - Substantially more ambitious community modeling projects (Petabyte (PB 10^{15}) and Exabyte (EB 10^{18})) will require a distributed database
- **Metadata describing extended modeling simulations (e.g., atmospheric aerosols and chemistry, carbon cycle, etc.)**
- **How to make information understandable to end-users so that they can interpret the data correctly**
- **Integration of multiple analysis tools, formats, data from unknown sources**
- **Trust and security on a global scale**



Courtesy: William Johnston et al. - LBNL

Slide info from Dean Willimas, LLNL

Data is Changing

- Much, much more data
- Finer grained/relatively smaller chunks
  - Many more files
  - Much more meta data
- More integration of different data formats

Usage is Changing

- The “e-cloud” generation expect immediate access to all data
  - Will not tolerate “feeling pain”
- More interdisciplinary merges of data
- More ad-hoc queries and combinations of data
  - Correlations
  - Re-analysis
- Tighter coupling of data analysis and simulation

Systems are Changing

- More layers of devices
  - Solid state storage devices
  - Different service levels of on-line storage devices
  - Near-line media continues to evolve
- More layers of SW
  - Parallelization of component layers
  - Open source versions
  - More layers interact directly

Resulting Challenges

- Continued device innovation
- Complexity
  - For systems and for users
- Resiliency
  - SW is fails at least as much as HW
  - Of the SW components in a large system, data services software failure rates are near the top
- Visualization and automatic feature recognition
- Need to serious consider complete re-engineering the software stack
- Need to engage new, non-traditional methods and communities

Resulting Challenges

- Modeling
  - For application use
  - What about for our systems?
- Improved assists for users to make good choices
  - Will our current clients use any we make
- Data Movement, Data Movement, Data Movement
  - From source to concentrated repository
  - Across layers of systems
  - Between repositories
- **The circumstances are right for Zeta/Yotta-byte Initiatives – but is the Mass Storage community**

Data

A Beautiful Disaster?

A Frightening Crisis?

Questions?

Dr. William Kramer
NCSA/University of Illinois
Blue Waters Deputy Director
wkramer@ncsa.uiuc.edu/ - http://www.ncsa.uiuc.edu/BlueWaters
(217) 333-6260

Storage Common Wisdom

- Old
  1. Users have a small number of large files
  2. Files are the lowest level unit of storage
  3. We need to cause users pain to move their files from place to place
  4. Users have all the files they need in each place they compute
  5. One system is sufficient for all the steps a workflow

- New
  1. Large numbers of small files dominate performance
  2. Objects are the lowest unit of storage
  3. It is more productive to systems and users to let systems to manage the placement of files
  4. User’s have data in many places and need to move the data frequently - even within a facility
  5. Job steps are best run on systems with the most appropriate balance