Multi-scale Atmospheric Modeling in a Cyberinfrastructure Era

John Helly
San Diego Supercomputer Center
University of California, San Diego

Outline

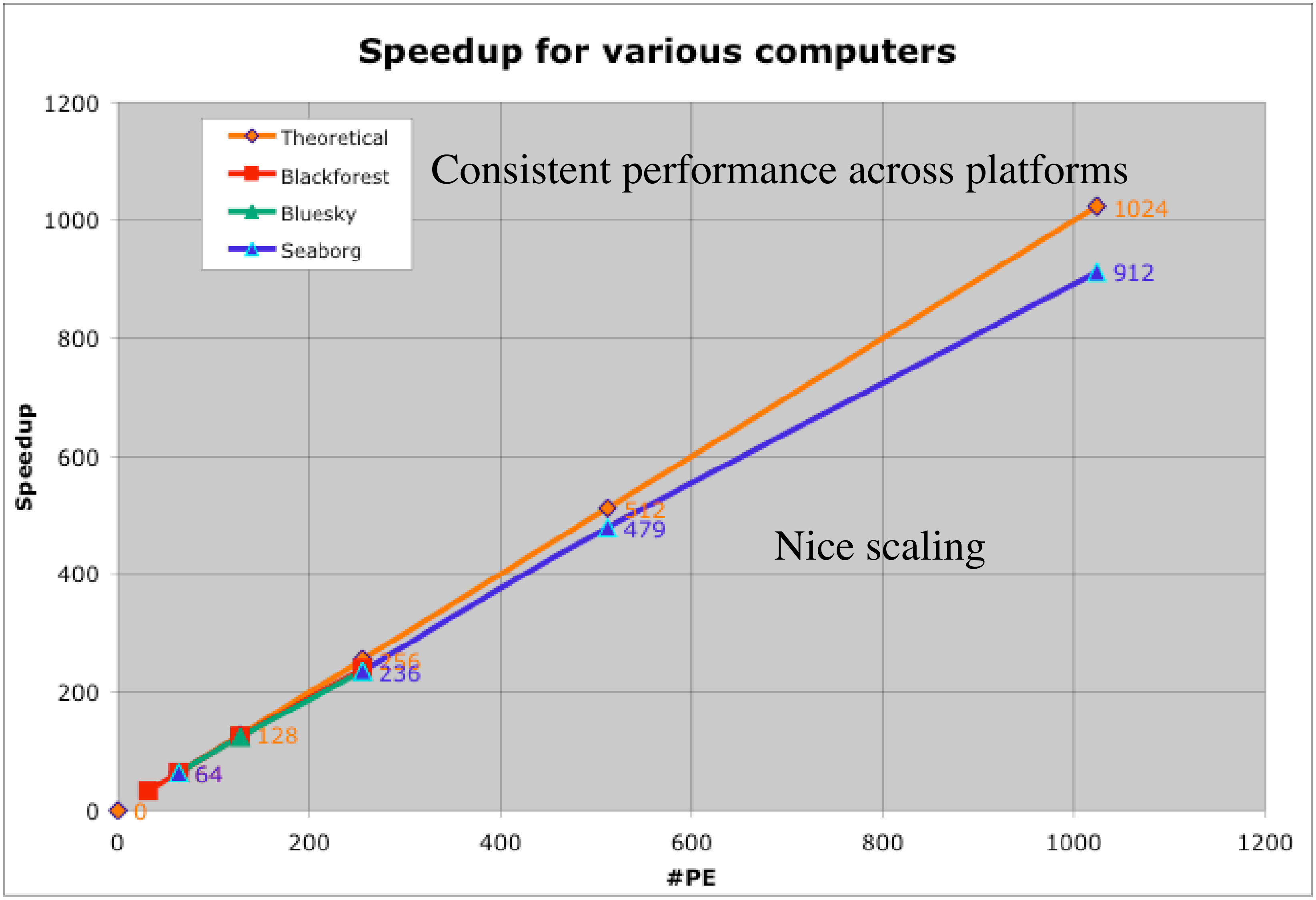
- MMF Requirements
  - Computation
  - Data
  - Networking
- Cyberinfrastructure Implications

MMF Computational Requirements

	Year(s)				
	2001	2002	2003	2005-06	2009-10
Simulated Duration (years)	1	1	See Plot For Current Results	100s	1000s
Performance (wall-clock days per simulated year)	20	1.5		Configuration dependent	
Number of Processors	64	512			

MMF Computational Requirements (cont.)

- Extrapolating from today's technology
  - 100 simulated years on NCAR's Bluesky today
  - 150 days on 512 PE (of 1216 PE available)
  - 24/7 availability, no errors, no restarts, no problems
- Things are a bit better with some new results …


Speedup for various computers
Theoretical
Blackforest
Bluesky
Seaborg
Consistent performance across platforms
Nice scaling
Speedup
#PE
0
64
128
236
256
479
512
912
1024
0
200
400
600
800
1000
1200

MMF Computational Requirements (cont.)

- Still requires better-than-linear scaling
  - 2 runs per year at that rate but we also need
    - developmental versus productions runs (ratio?)
    - ensembles (i. e., N)
    - intermediate-scale development and testing
  - Cannot commit an arbitrary number of processors sufficiently often using existing (or near) resources and approaches

Cyberinfrastructure implications

- distributed experiments (i.e., $||^N$)
- centralized vs distributed?
- distribution of software and data
- aggregation of results

$||^N$: Higher-order Parallelism

- Grid computing
  - Divide and conquer (old standby)
  - Maybe it's still the answer but does it scale?
    - Staging
    - Queuing
    - Synchronization
  - Many system engineering challenges
    - Data management
    - Network capacity
    - Configuration management

- How do you address the human and organizational aspects?

Cyberinfrastructure Implications

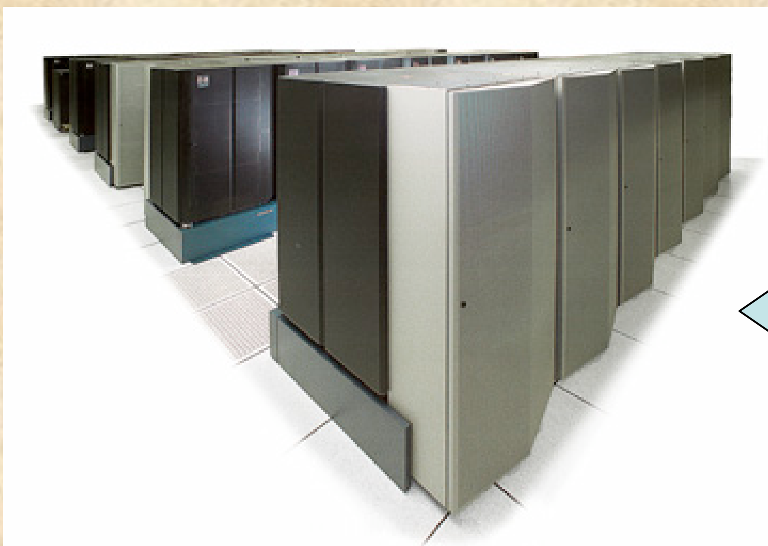
- What's the Right Mix?

Workstation

- How should they be connected?

SW/Data

SW/Data



SW/Data

Mainframes?

Clusters?

New SDSC Resources

- DataStar
  - 128 x 8 Power4, 1.5GHz, 16GB processors (01/2004)
  - 10 x 32 Power4, 1.7GHz, 128/256GB processors (now)
- Teragrid/Itanium
  - Cluster of 256 x 2, Madison 1.5GHz, 4GB processors
  - Cluster of 128 x 2, Madison 1.3GHz, 4GB processors
  - Both in house (operational 01/2004)
- Parallel File Systems
  - 5GBps (sustained)
  - 10GBps (peak)
  - To both of these machines

New SDSC Resources

- Fast Storage
  - 600 TB SAN (backbone)
  - Fiber channels to every node of both machines
  - And to Teragrid network (viz, external network)
- Tape Subsystem
  - ~2PB fast tape
  - ~2GBps transfer rate
  - Automatic migration from disk to tape according to disk quota management (and back)
  - Long-time parking of output data for post-processing
- Network
  - Phase 1: 40Gbps (to NCSA, lower to Argonne, CalTech)
  - Phase 2: 6 other institutions

Lessons-learned from Collaborative Code Development?

- Joint experiments with community codes require inter-organizational and inter-disciplinary collaboration
  - Scientists knowing the physics and phenomena
  - Programmers with architectural expertise and comprehension of the domain language
  - Sympathetic system administrators
  - Principal Investigators who are good managers

Regional Spectral Model

Example

- Joint code development and experiments
  - SDSC
  - Scripps Institution of Oceanography Climate Research Division
  - NOAA National Center for Environmental Prediction
- Goals
  - 10-12 km resolution of atmospheric processes (esp. ppt)
  - distributed access to common resources for joint experiments
  - preparing code to run on the Earth Simulator

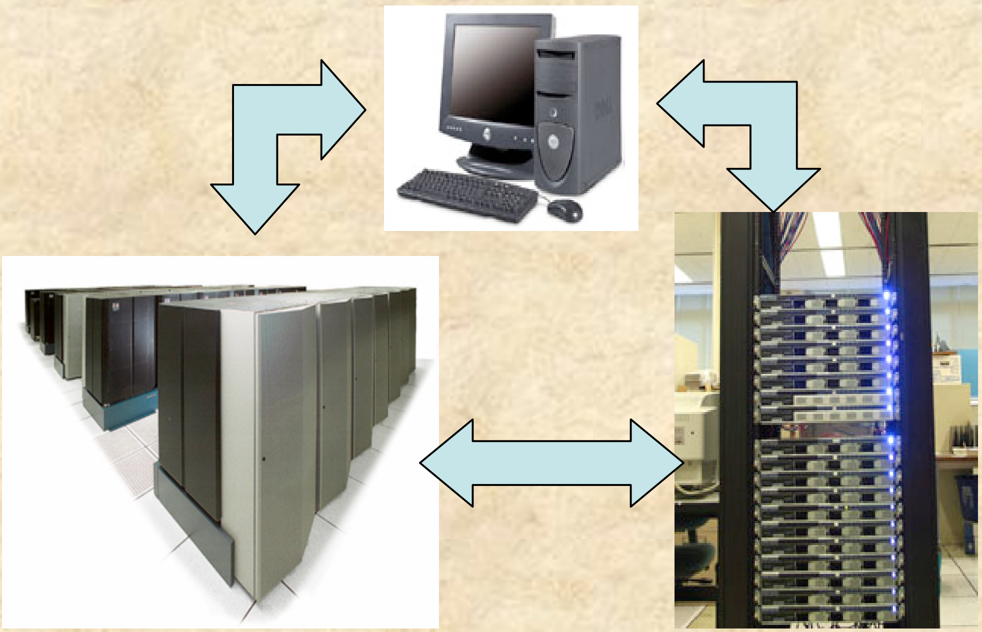
Regional Spectral Model (cont.)

- New community code resource
  - Supported platforms (i.e., build environments)
    - Power 3 / Power 4
    - Intel Architecture 32 / 64 (clusters)
    - SX-6 (Earth Simulator)
- Configuration management
  - Open source development model
    - Multiple developers (NOAA, SIO, SDSC)
    - Train domain-scientists in these techniques
    - Code changes versus CVS baseline
    - Test suites executed before ‘commit-to-baseline’

Biggest Issue (for now)

- Race between 'stable baseline' and new physics (i.e., new code)
  - Multiple platforms
  - Multiple build environments
  - Multi-processor (small n) testing
  - Multi-processor (big N) production
- Keeping a community code relevant
  - There will always be a lag
  - How much is too much?

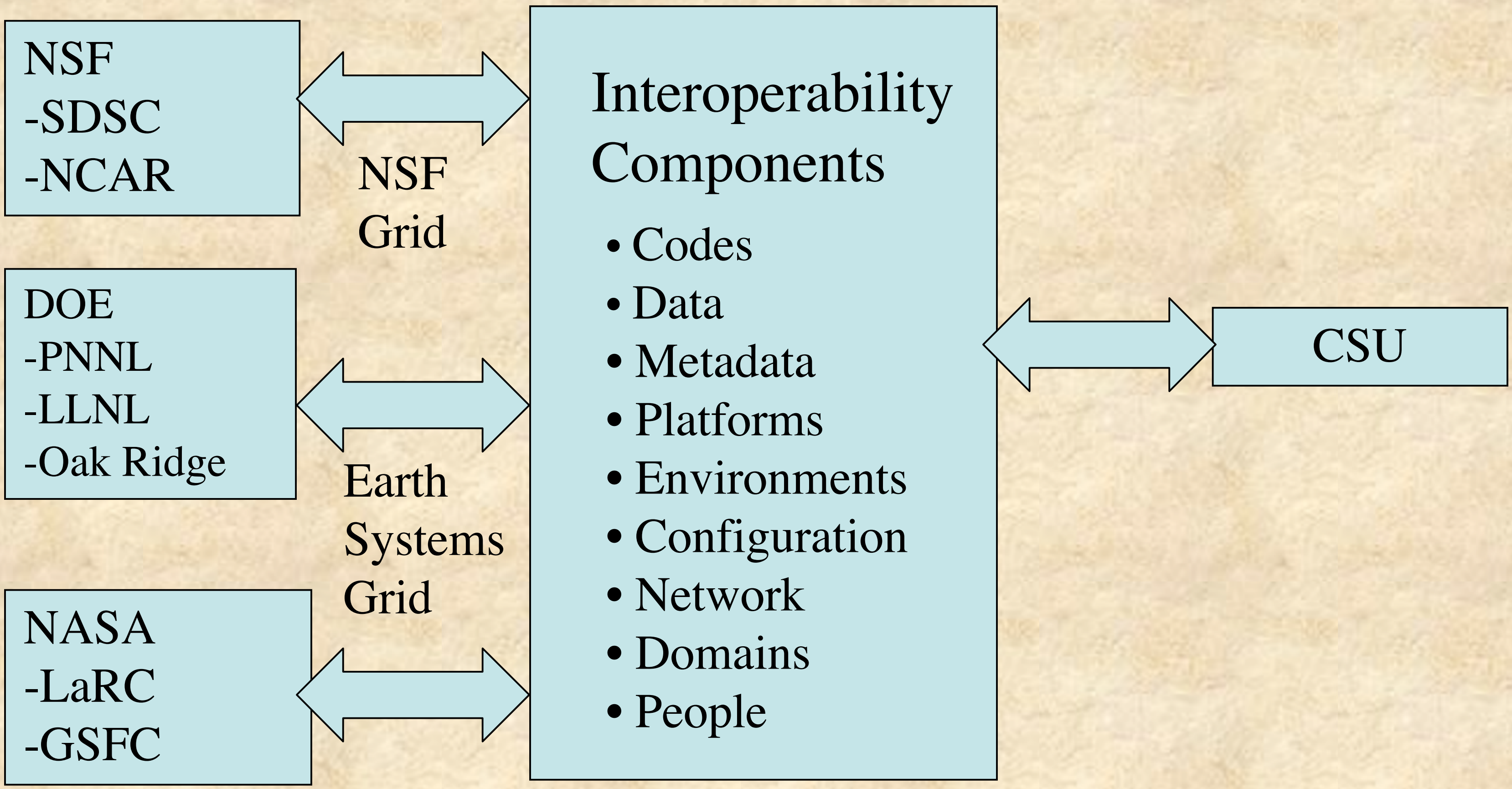
Cyberinfrastructure Implications



‘Intangible’ Tradeoffs

- Interoperability
- Logistics

Cyberinfrastructure for MMF

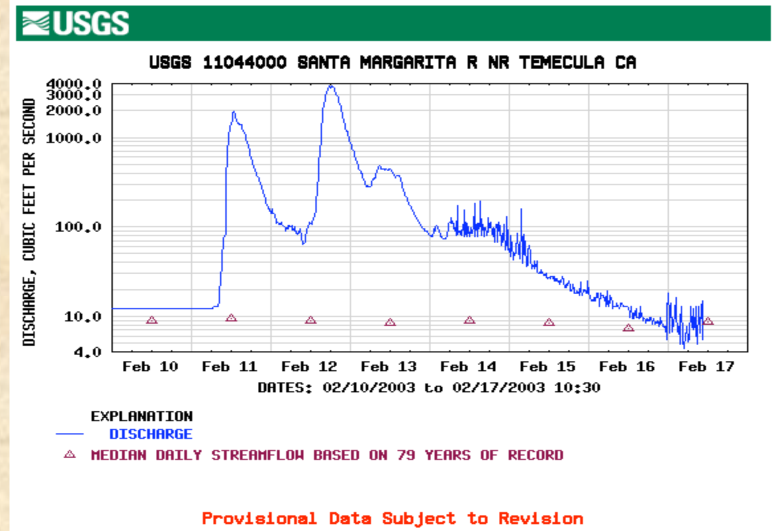


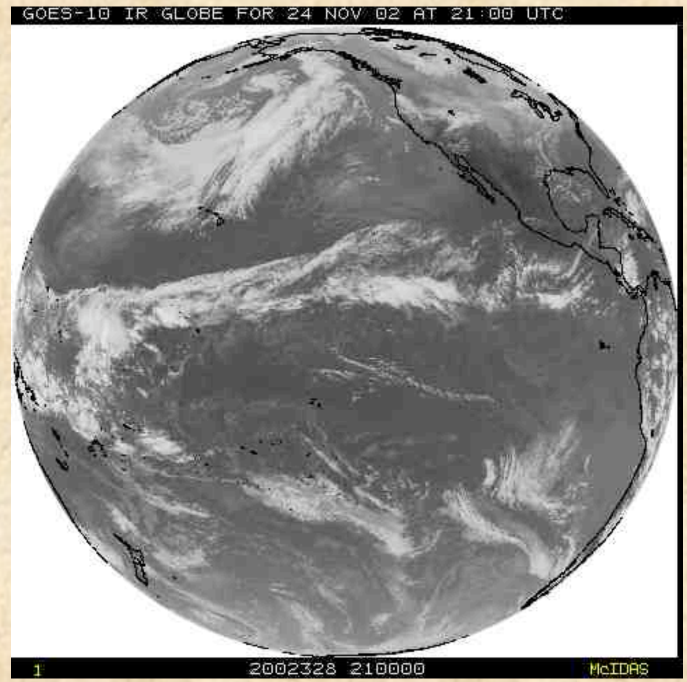
Cyberinfrastructure Implications

- Wide range of resource requirements
  - Development vs. production
    - Painless migration between them
  - Empirical data
    - verification and
    - validation
- Convenient access to resources for ‘many’ users
  - Dedicated resources or small queuing delays
  - Interactive atmosphere of trust and tradition
- High bandwidth, convenient, collegial interaction

Cyberinfrastructure Implications (cont.)

- Better-than-linear scaling requires organizational parallelism as well as architectural parallelism
  - A higher-degree of standardization is required as the number of participants grows
  - Configuration management becomes a critical, joint, cooperative activity





Data Loading

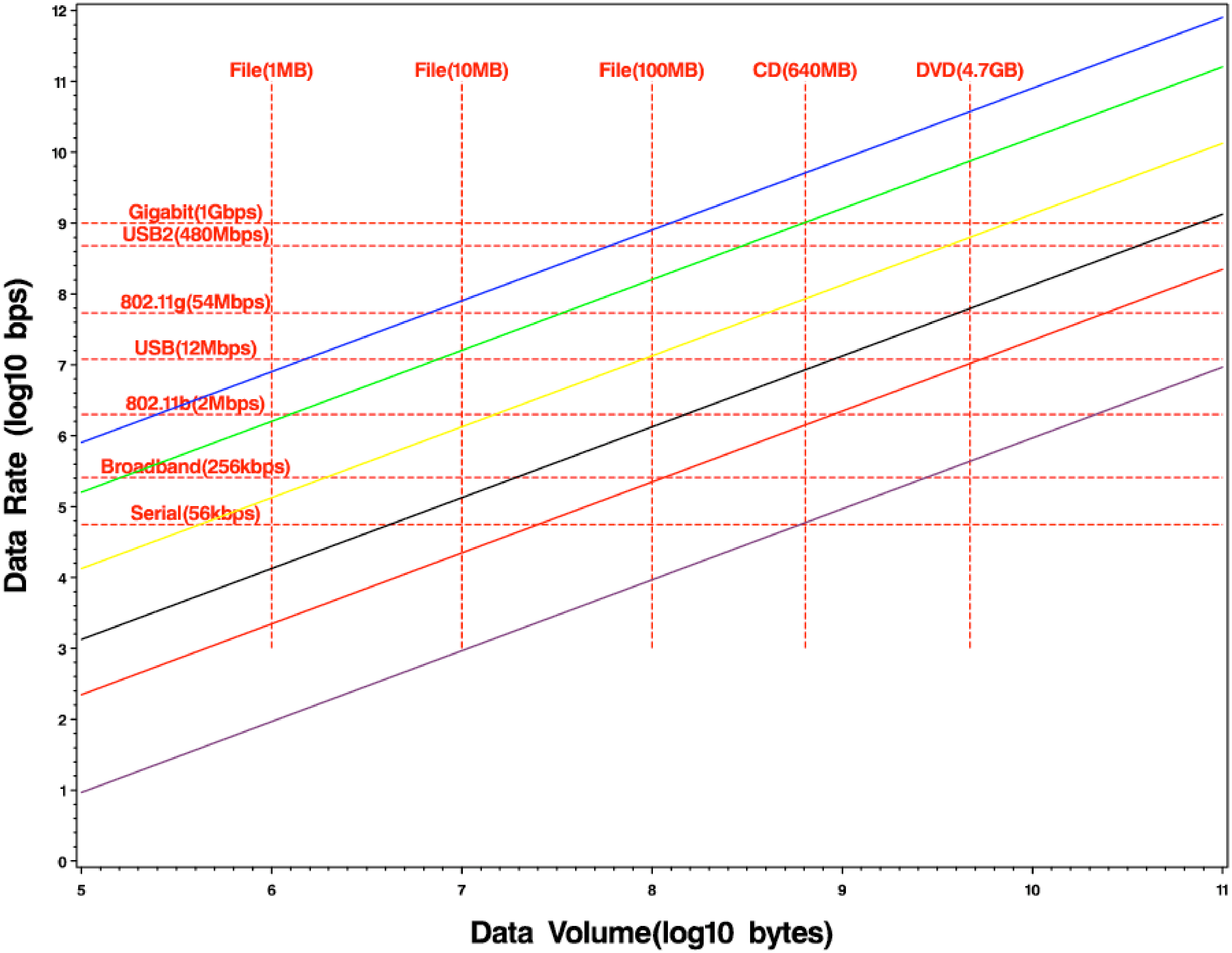
Staging, Execution, Analysis

- Staging (outbound)
  - Initialization : 0.1 Gbytes to each cluster then disseminated to each node
- Execution (held locally)
  - Data assimilation : 0.2 GBytes x nodes/cluster x clusters
  - Model output: Cloud Resolving Model @ 8192 grid cells
    - 1 year x (5 space-time, 6 prognostic, 10 diagnostic variables)
    - 10 Terabytes roughly per cluster (say 256 nodes)
- Analysis (inbound)
  - 10 Terabytes/run x number of clusters

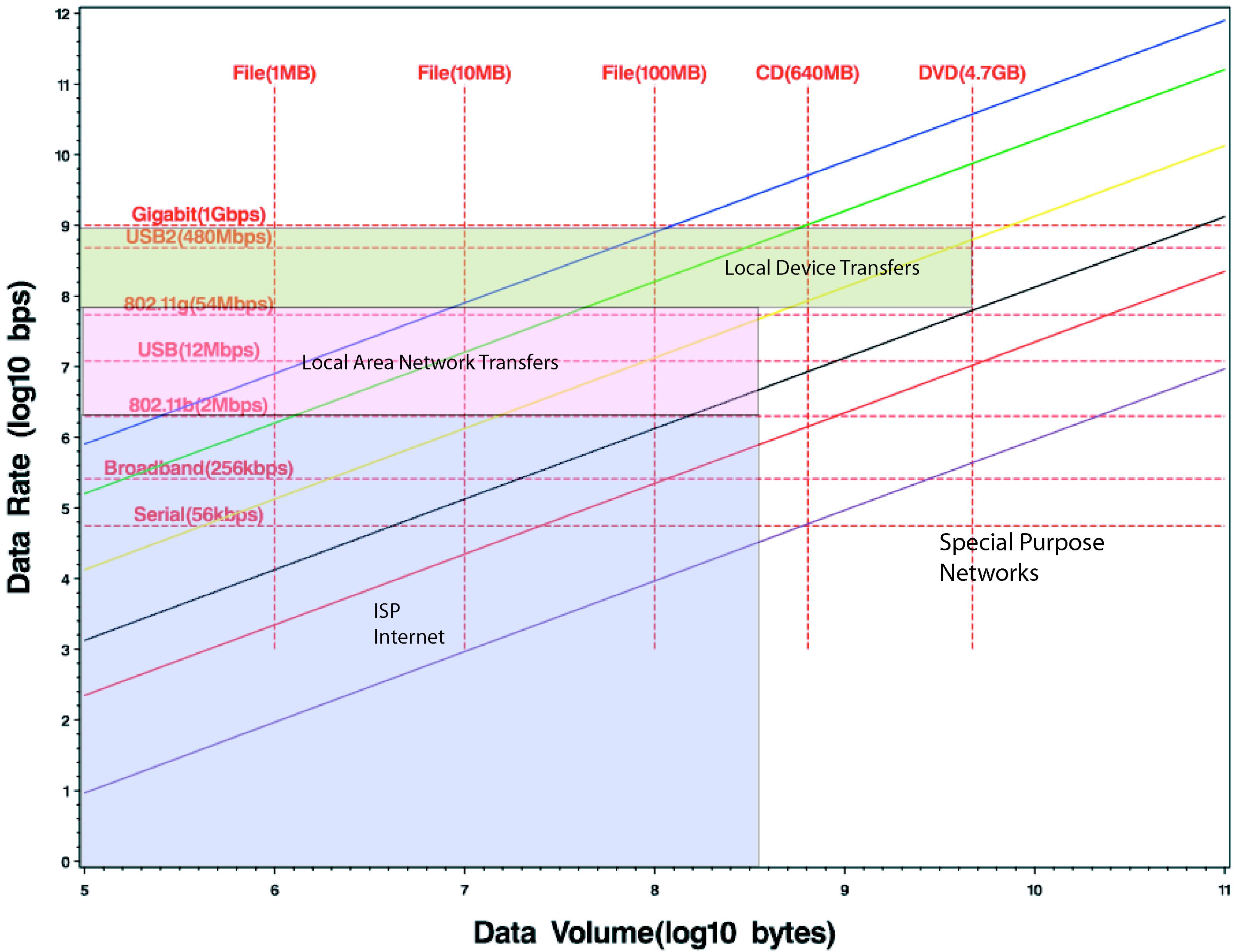
Cyberinfrastructure Implications

- Reality check
  - This implies that big fractions of a petabyte must be moved per run
    - Potentially 10's of petabytes per experiment (i.e., ensembles)
    - Analysis, Re-analysis
    - Archival
- Maybe not, but there's a catch to move less
  - Some form of data extraction on a space-time basis must otherwise exist everywhere

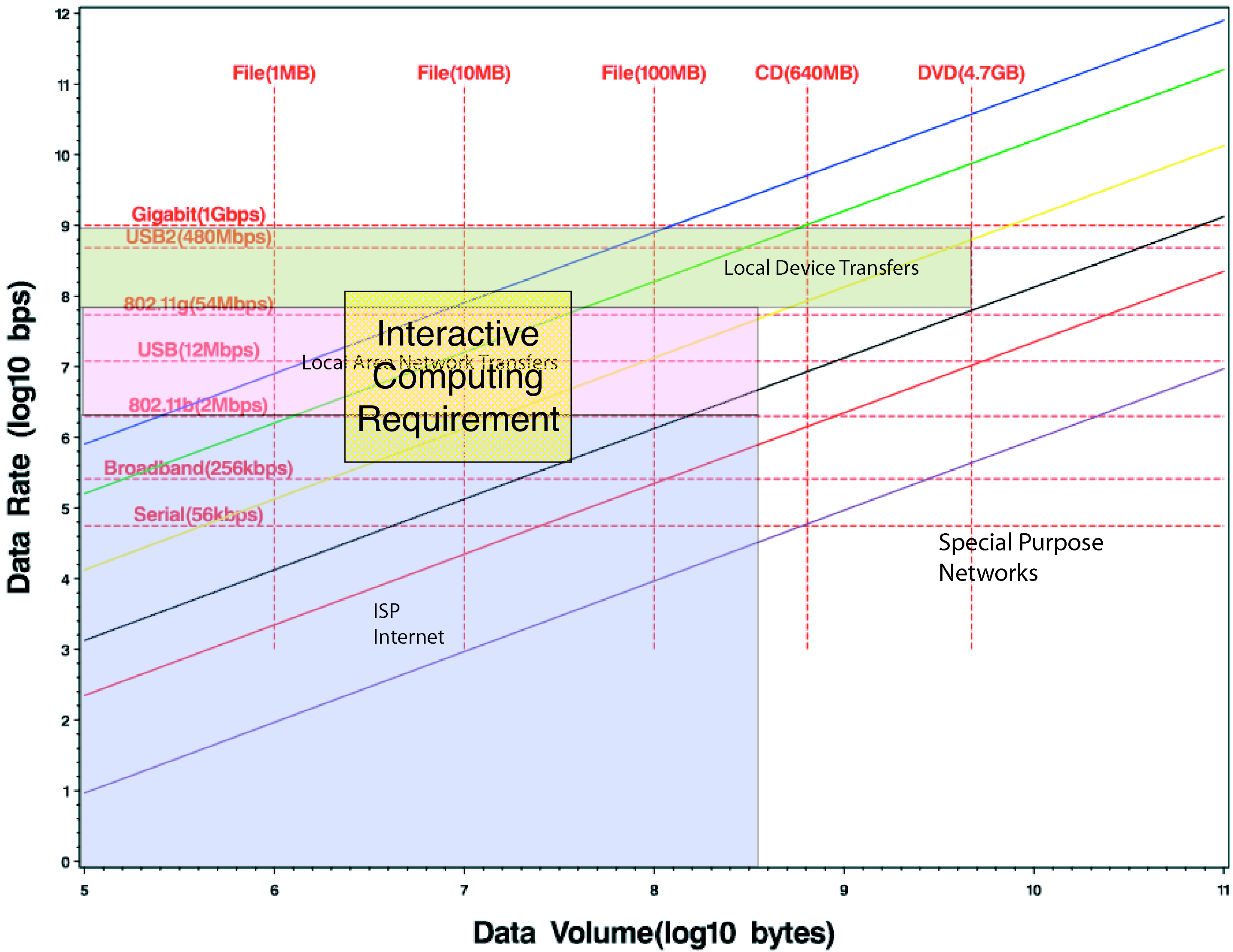
Network Requirements



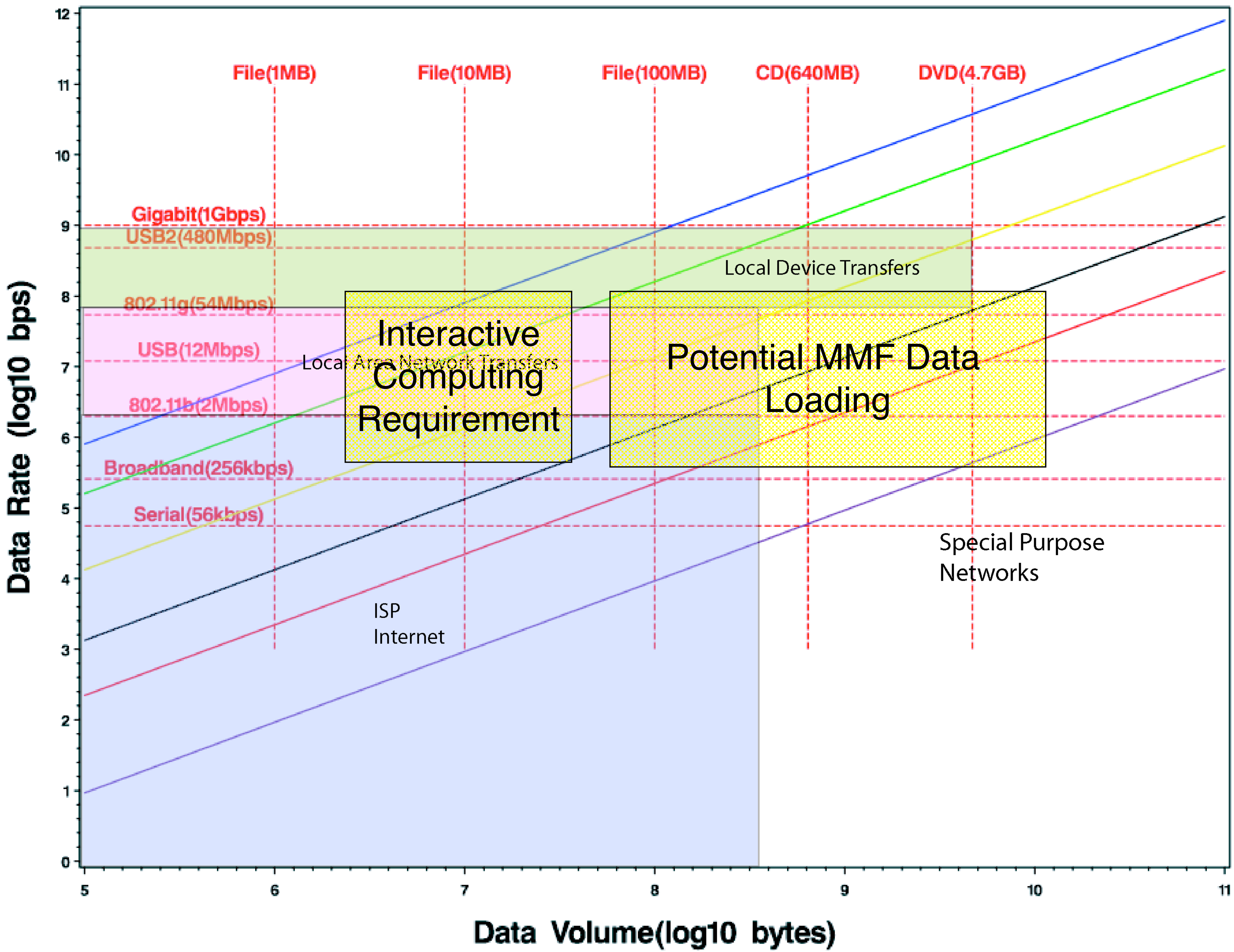
Data Rate (log10 bps)
Data Volume(log10 bytes)
File(1MB)
File(10MB)
File(100MB)
CD(640MB)
DVD(4.7GB)
Gigabit(1Gbps)
USB2(480Mbps)
802.11g(54Mbps)
USB(12Mbps)
802.11b(2Mbps)
Broadband(256kbps)
Serial(56kbps)
0
1
2
3
4
5
6
7
8
9
10
11
12
5
6
7
8
9
10
11



Data Rate (log10 bps)
Data Volume(log10 bytes)
File(1MB)
File(10MB)
File(100MB)
CD(640MB)
DVD(4.7GB)
Gigabit(1Gbps)
USB2(480Mbps)
802.11g(54Mbps)
USB(12Mbps)
802.11b(2Mbps)
Broadband(256kbps)
Serial(56kbps)
Local Device Transfers
Local Area Network Transfers
ISP Internet
Special Purpose Networks
0
1
2
3
4
5
6
7
8
9
10
11
12
5
6
7
8
9
10
11



File(1MB)
File(10MB)
File(100MB)
CD(640MB)
DVD(4.7GB)
Gigabit(1Gbps)
USB2(480Mbps)
802.11g(54Mbps)
USB(12Mbps)
802.11b(2Mbps)
Broadband(256kbps)
Serial(56kbps)
Local Device Transfers
Local Area Network Transfers
Interactive Computing Requirement
ISP Internet
Special Purpose Networks
Data Rate (log10 bps)
Data Volume(log10 bytes)



File(1MB)
File(10MB)
File(100MB)
CD(640MB)
DVD(4.7GB)
Gigabit(1Gbps)
USB2(480Mbps)
802.11g(54Mbps)
USB(12Mbps)
802.11b(2Mbps)
Broadband(256kbps)
Serial(56kbps)
Local Device Transfers
Local Area Network Transfers
Interactive Computing Requirement
Potential MMF Data Loading
ISP Internet
Special Purpose Networks
Data Rate (log10 bps)
Data Volume(log10 bytes)

How are we preparing for the Cyberinfrastructure Era?

Road to Cyberinfrastructure

Interoperability Components

- Codes
- Data
- Metadata
- Platforms
- Environments
- Configuration
- Network
- Domains
- People

Computing

- Teragrid
- NPACI SAC

Data

- GEON
- CUAHSI
- SIOExplorer

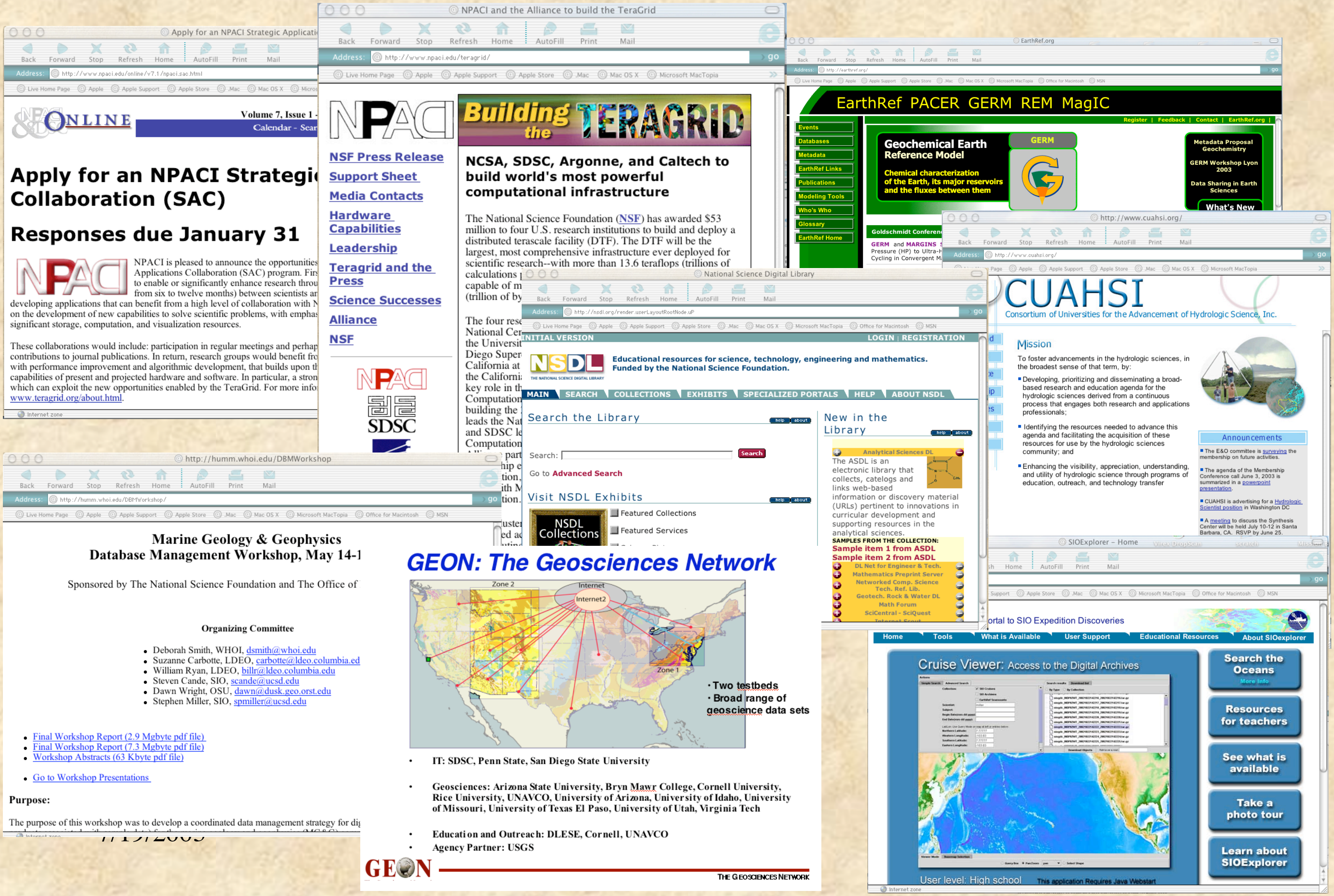
Community

- MG&G
- CI/Boulder
- MMF Workshop 2003
- GERM 2003
- Interop Workshop

MMF NPACI SAC

- NPACI Strategic Application Collaboration (SAC) with CSU
  - Porting to Blue Horizon and Teragrid environments
  - Provide performance data for MMF scaling in Grid environment
  - Infrastructure requirements for distributed, collaborative code development
- Establishing
  - joint working relationships
  - data management environment
  - RDTE procedures

NSF Technology Pathfinders



Cyberinfrastructure is Not Just 'Bits and Boxes'

- Human infrastructure
  - People
  - Interdisciplinary, multi-lingual, multi-cultural
- Broad-base of expertise
  - Across institutions
  - Across disciplines
  - Across systems
- Leveraging existing investments provides foundation for future CI

Summary

- MMF is a well-behaved, challenging computational task
  - New physics and phenomena at better resolution
  - Existing code scales well and performs consistently
  - Climate-scale, distributed experiments pose exciting cyberinfrastructure challenges
  - We are learning today to handle these challenges
- MMF provides an outstanding application at the right stage of code development (i.e., early)

Backup

Network Requirements

- Collaborative, interdisciplinary research requires joint access to common information and resources
  - Data includes empirical and derived data and model results
  - Analysis tools
  - RDTE environment
    - Research, Development, Testing and Evaluation
    - Compilers, hardware, build-environments
    - Especially important in a transient user/developer population (i.e., academia) due to spin-up time
- Standard trade-off between computing and communications
  - ‘Compute it there, transfer it here’
  - Cycles vs. bandwidth

Definitions of Interoperability

- IEEE
  - Ability of two or more systems or components to
    - exchange information and to
    - use the information that has been exchanged
- ISO
  - Attributes of software that bear on its ability to interact with specified systems