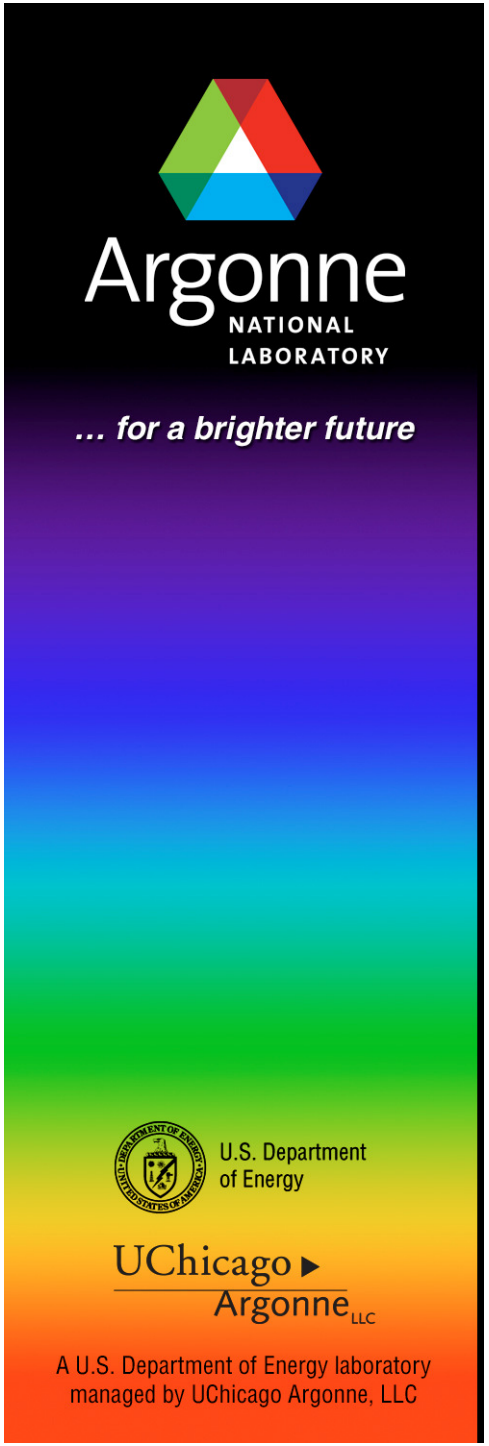


Enabling Distributed Petascale Science

Ian Foster

Argonne National Laboratory
University of Chicago

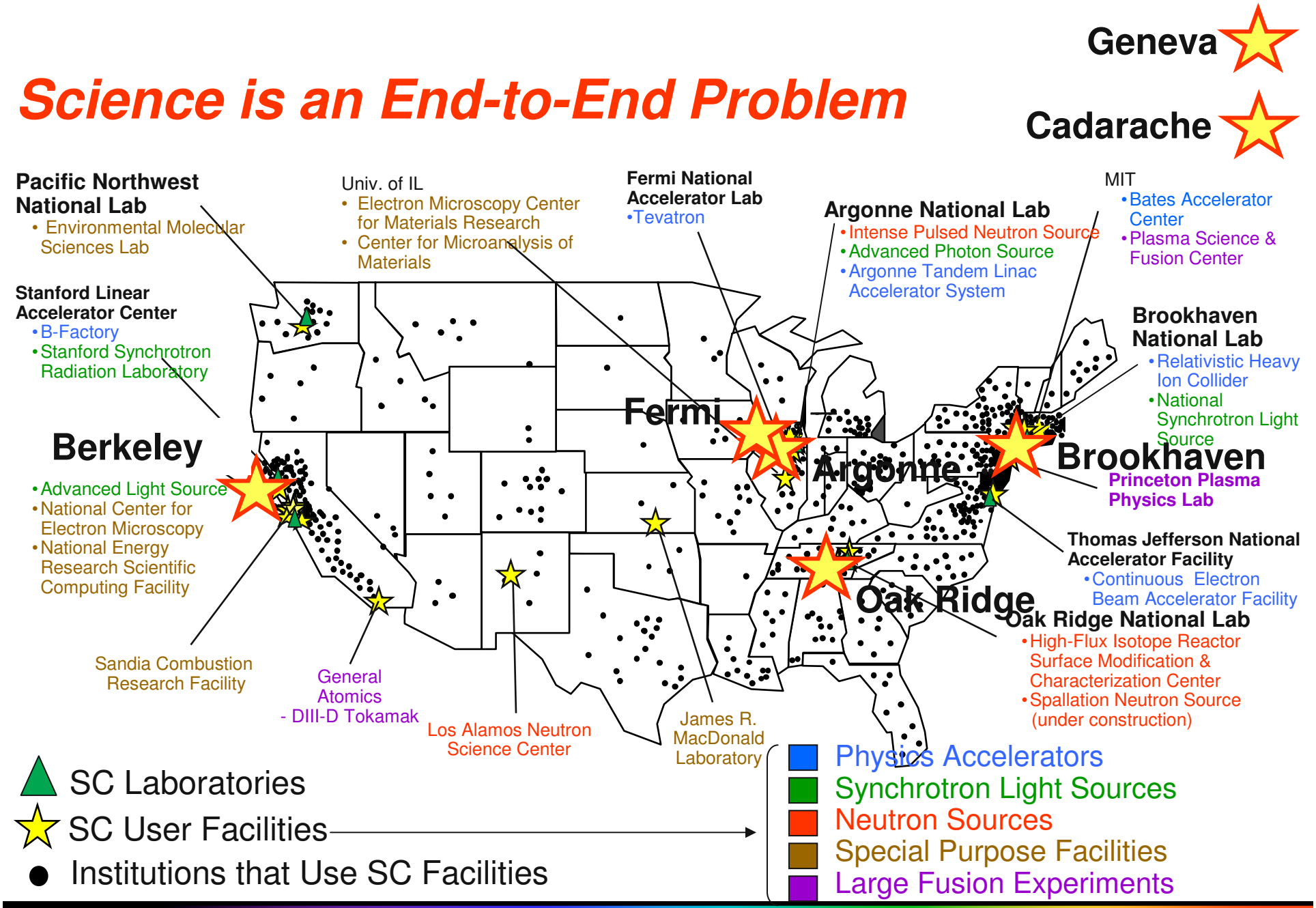
CEDPS Project Participants:

Jennifer **Schopf**, Kate **Keahey**, Dan **Fraser**,
John **Bresnahan**, Tim **Freeman**, Argonne
Keith **Jackson**, Brian **Tierney**, Dan **Gunter**, LBNL
Ann **Chervenak**, Carl **Kesselman**, USC/ISI
Miron **Livny**, Nick **LeRoy**, Wisconsin
Donald **Petravick**, Fermi


If planes had sped up by the same factor as computers over the past 50 years, we would cross the country in a tenth of a second


Yes, but it would still take us two hours to get downtown!!!
SCHOOL BUS

Science is an End-to-End Problem



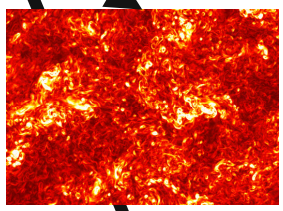
An Example: FLASH Turbulence Simulation (Robert Fisher, Don Lamb, et al.)

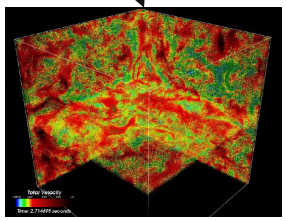
THE UNIVERSITY OF CHICAGO

74 million files
154 TB

3 weeks @ 20 MB/sec
(GridFTP)

23 TB





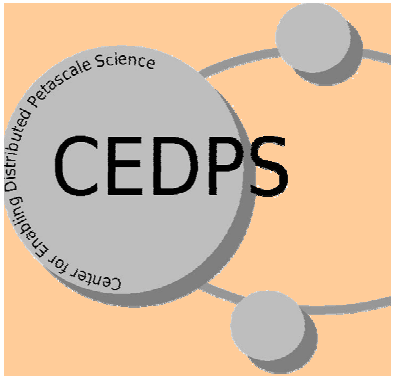
Largest compressible homogeneous isotropic turbulence simulation

External users access turbulence dataset

Elements of the End-to-End Problem Include …

- Massively parallel petascale simulation
- High-performance parallel I/O
- Remote visualization
- High-speed reliable data movement
- Terascale local analysis
- Data access and analysis by external users
- Troubleshooting problems in end-to-end system
- Security
- Orchestration of these various activities

Center for Enabling Distributed Petascale Science (CEDPS)



- Massively parallel petascale simulation
- High-performance parallel I/O
- Remote visualization
- **High-speed reliable data movement**
- Terascale local analysis
- **Data access and analysis by external users**
- **Troubleshooting problems in end-to-end system**
- Security
- Orchestration of these various activities

Argonne NATIONAL LABORATORY

BERKELEY LAB

Fermilab

ISI Information Sciences Institute USC School of Engineering

THE UNIVERSITY of WISCONSIN MADISON

Open Science Grid

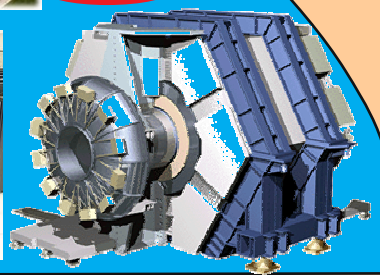
Earth System Grid

Bridging the Divide (1): Move Data to Users When & Where Needed

"Deliver this 100 Terabytes to locations A, B, C by 9am tomorrow"

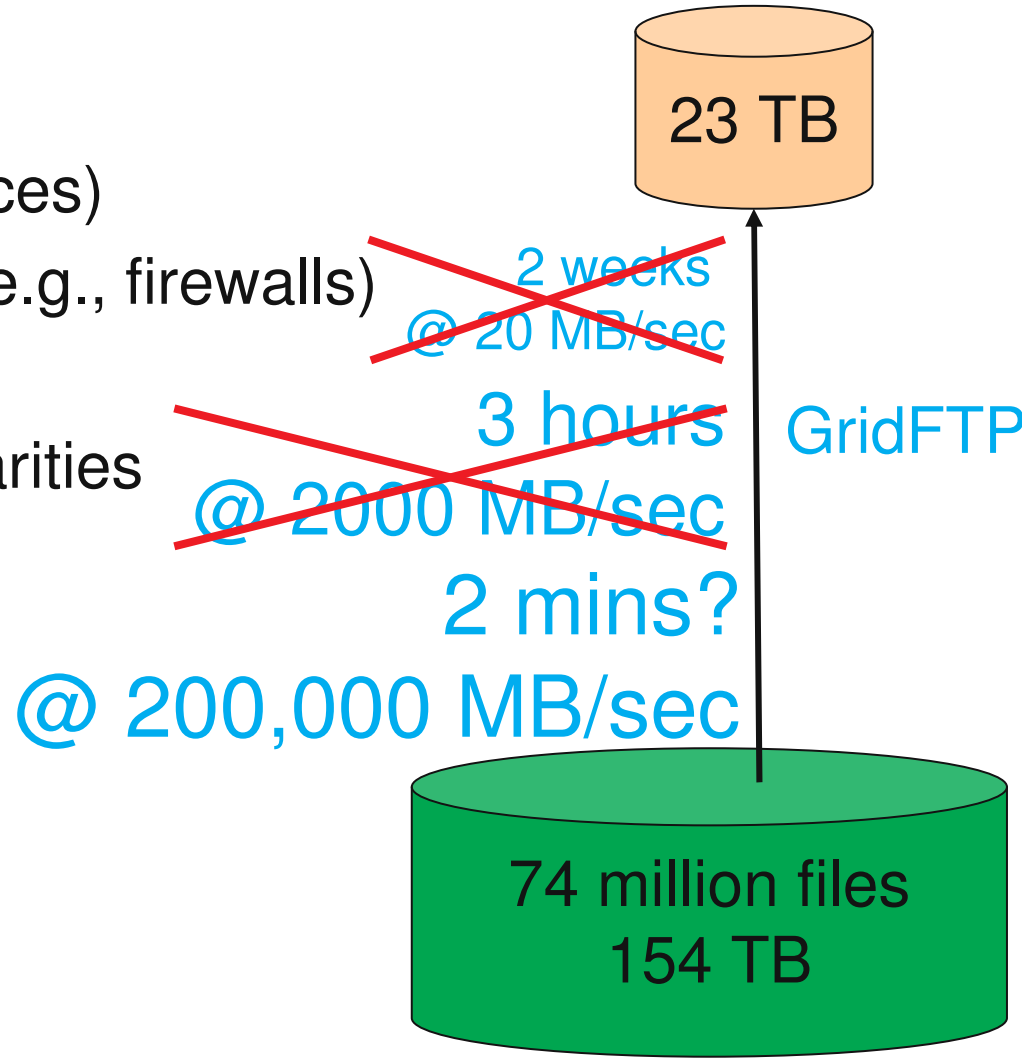
- **Fast**: >10,000x faster than usual Internet
- **Reliable**: recover from many failures
- **Predictable**: data arrives when scheduled
- **Secure**: protect expensive resources & data
- **Scalable**: deal with many users & much data





Data Delivery Challenges

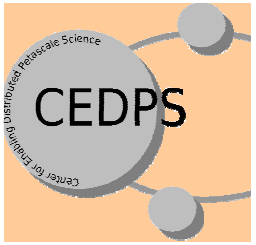
- Data complexity
- Parallelism (in diverse places)
- Network heterogeneities (e.g., firewalls)
- Space (or the lack of it)
- Protocols and their peculiarities
- Failures at many levels
- Deadlines
- Resource contention
- Multiple participants



Data Delivery Challenges

- Data complexity
- Parallelism (in diverse places)
- Network heterogeneities (e.g., firewalls)
- Space (or the lack of it)
- Protocols and their peculiarities
- Failures at many levels
- Deadlines
- Resource contention
- Multiple participants

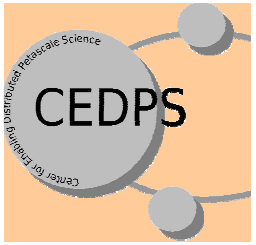
Data Delivery Challenges



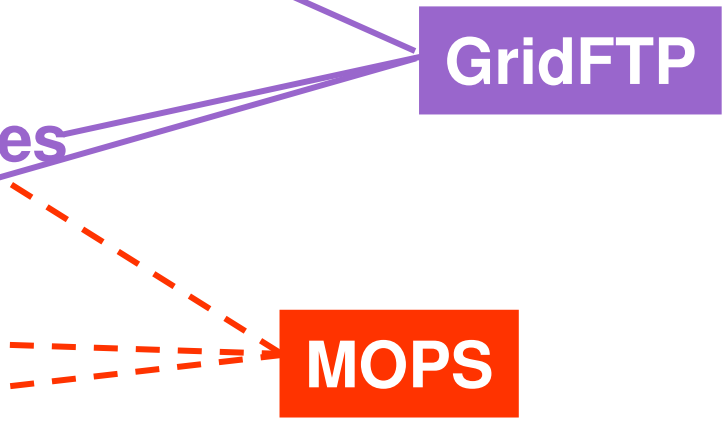
- Data complexity
- **Parallelism (in diverse places)**
- Network heterogeneities (e.g., firewalls)
- Space (or the lack of it)
- **Protocols and their peculiarities**
- **Failures at many levels**
- Deadlines
- Resource contention
- Multiple participants

GridFTP

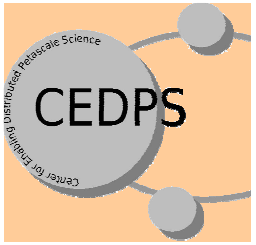
Data Delivery Challenges



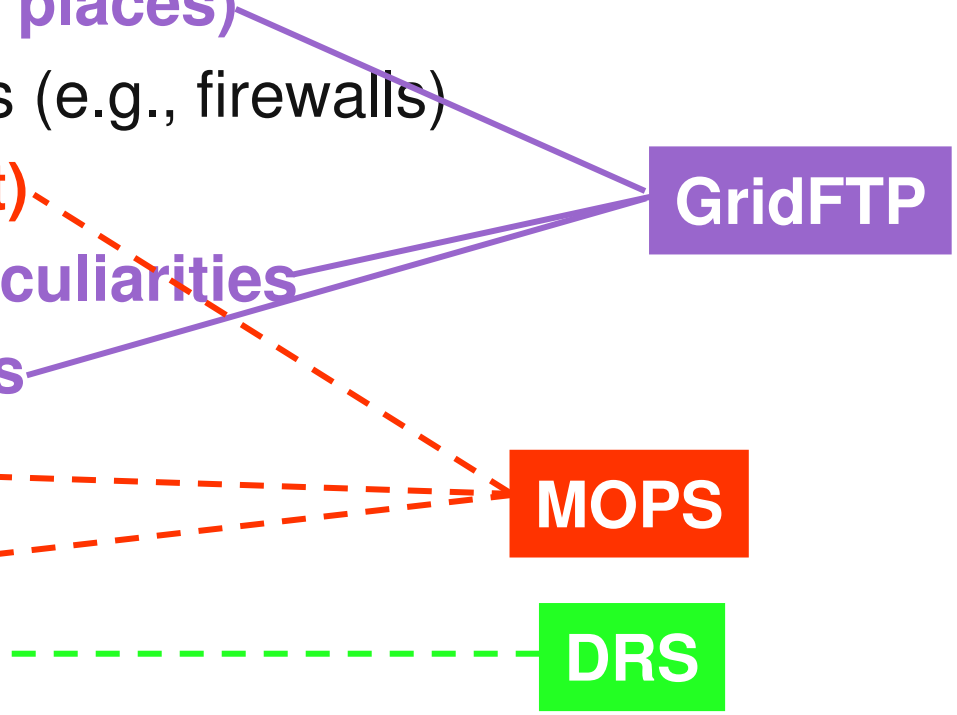
- Data complexity
- **Parallelism (in diverse places)**
- Network heterogeneities (e.g., firewalls)
- **Space (or the lack of it)**
- **Protocols and their peculiarities**
- **Failures at many levels**
- **Deadlines**
- **Resource contention**
- Multiple participants



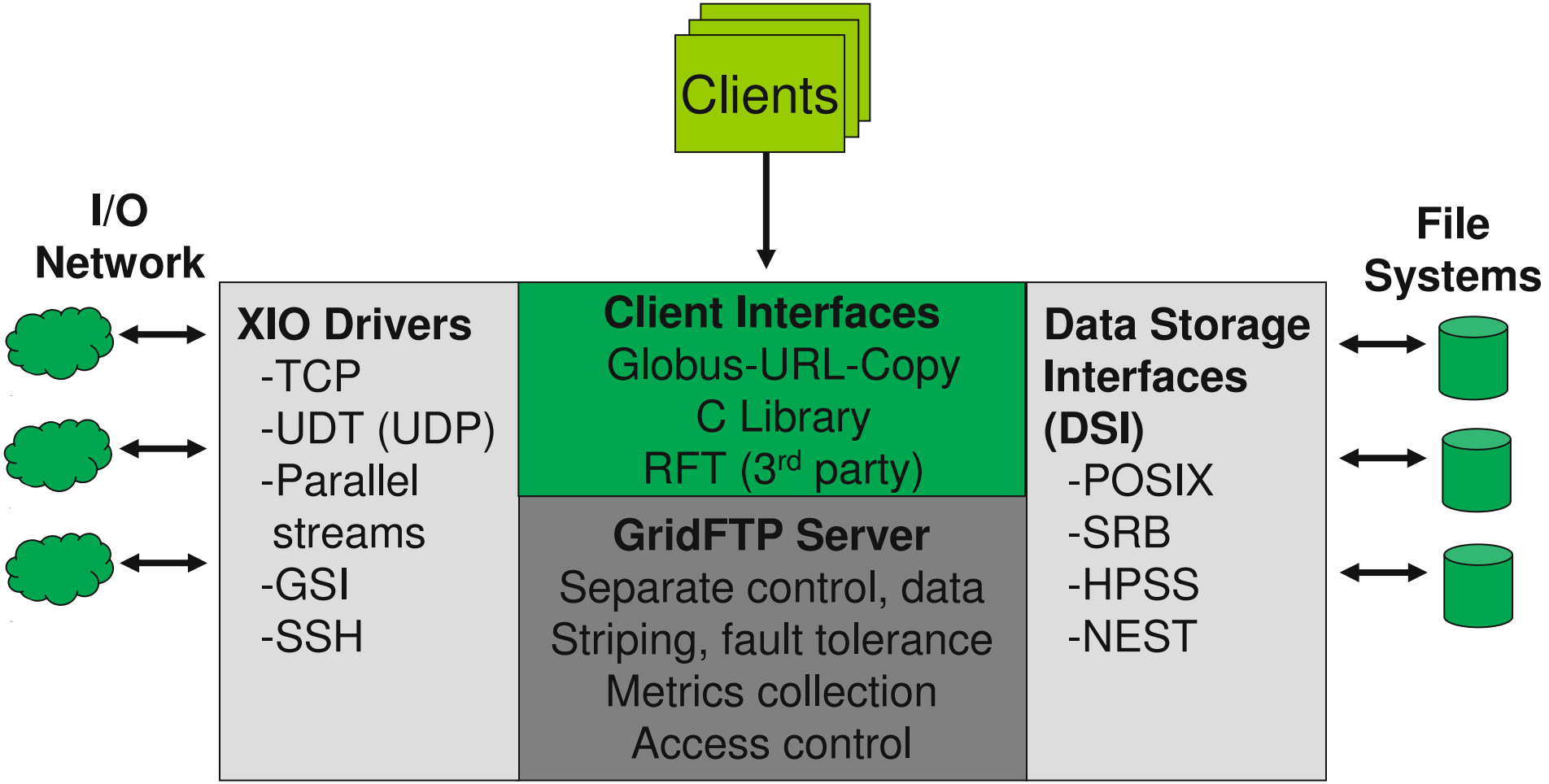
Data Delivery Challenges



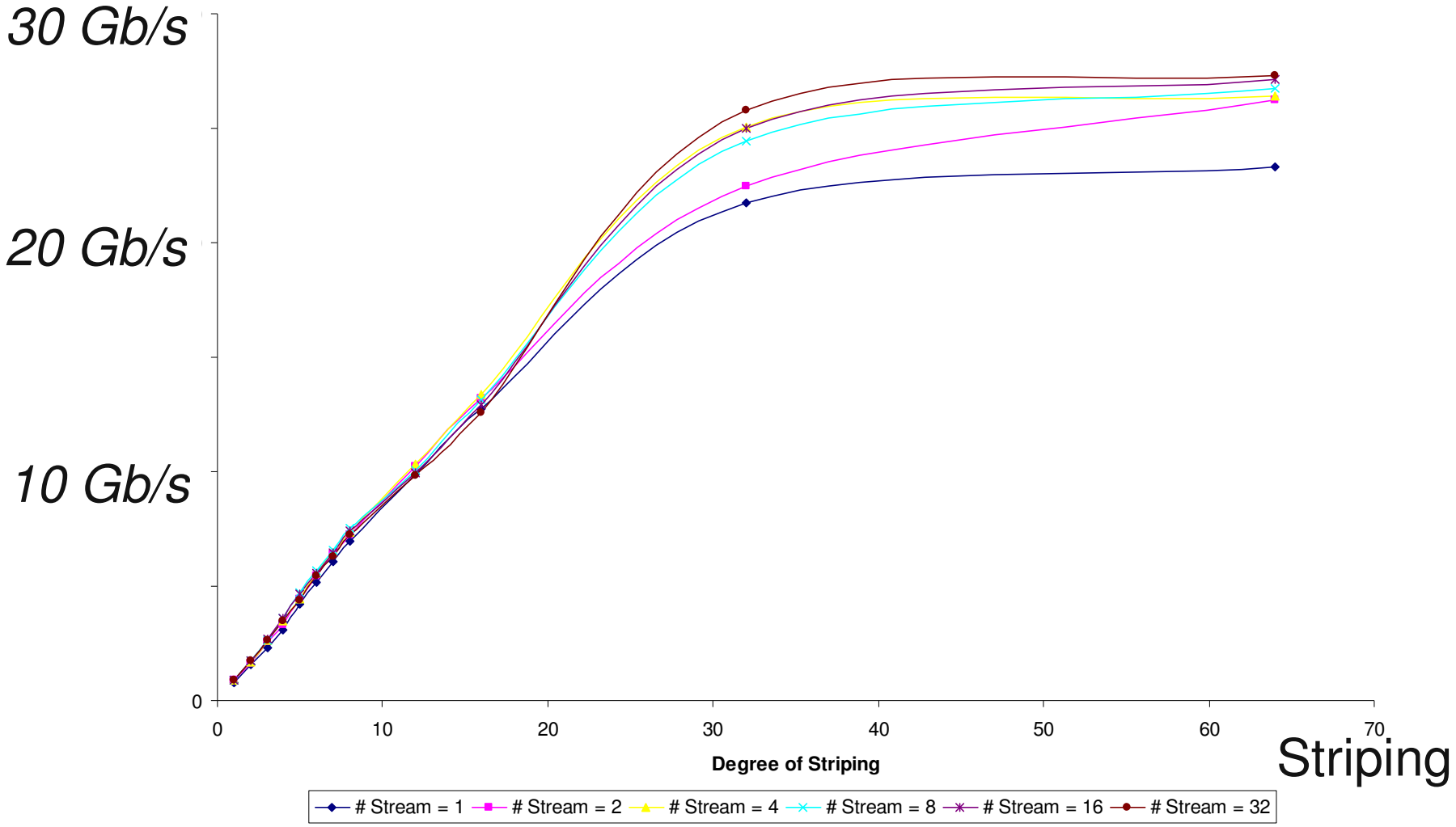
- Data complexity
- **Parallelism (in diverse places)**
- Network heterogeneities (e.g., firewalls)
- **Space (or the lack of it)**
- **Protocols and their peculiarities**
- **Failures at many levels**
- **Deadlines**
- **Resource contention**
- **Multiple** participants



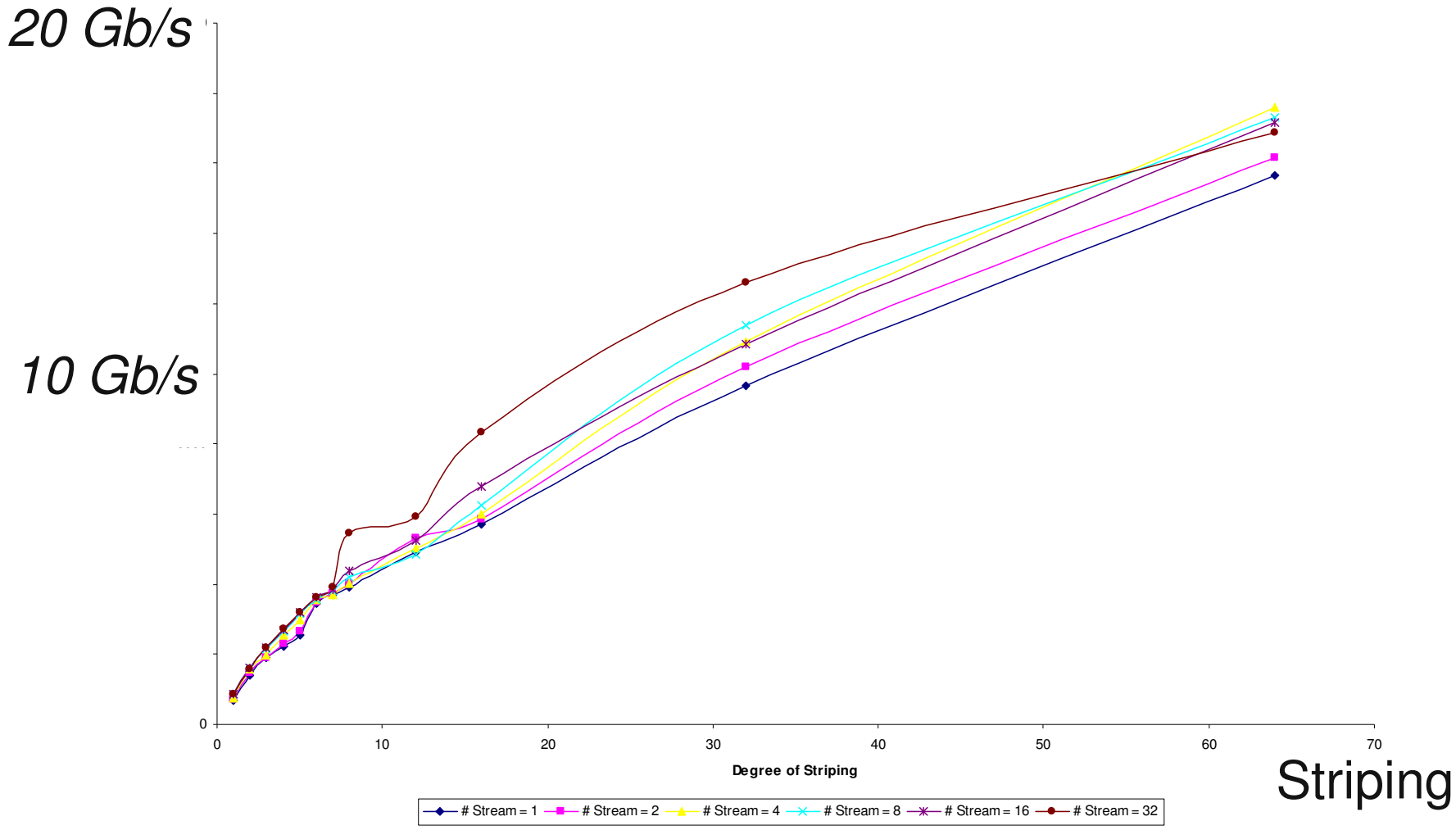
Current State of the Art: GridFTP



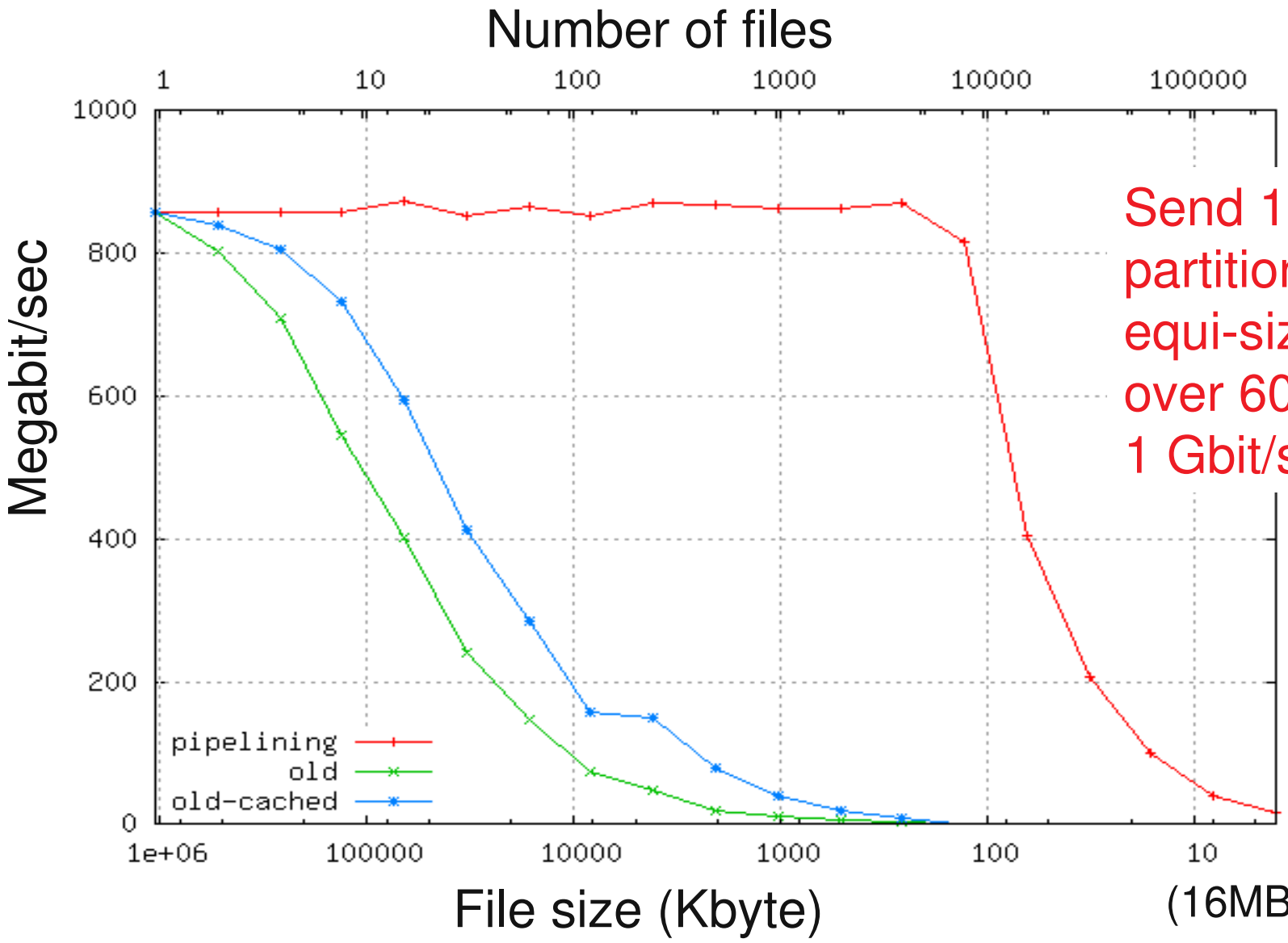
Memory to Memory over 30 Gigabit/s Network (San Diego — Urbana)



Disk to Disk over 30 Gigabit/s Network (San Diego — Urbana)



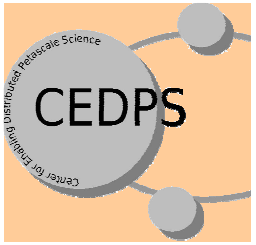
"Lots of Small Files" (LOSF) Optimization



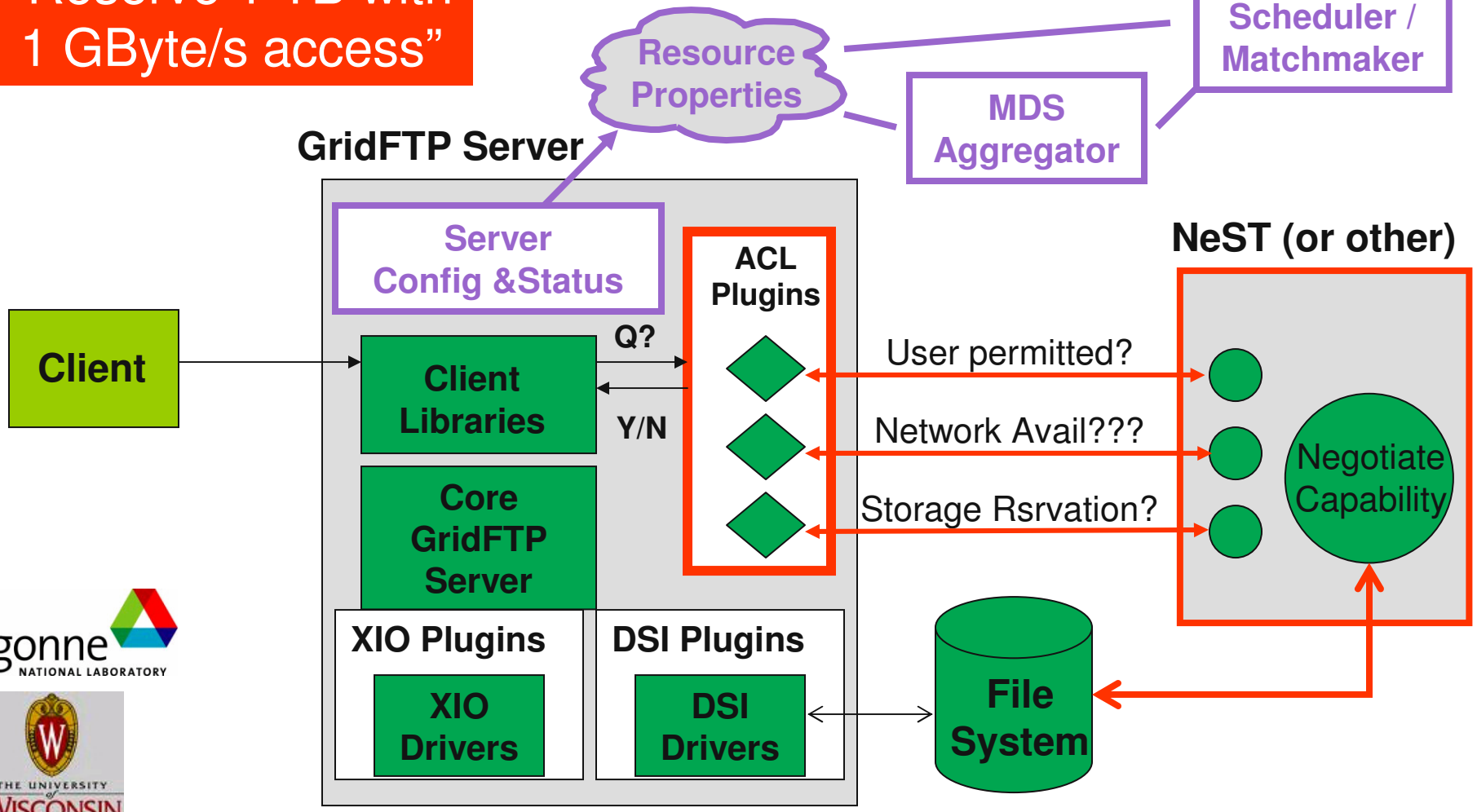
Send 1 GB partitioned into equi-sized files over 60 ms RTT, 1 Gbit/s WAN

(16MB TCP buffer)

Managed Object Placement Service (MOPS)



"Reserve 1 TB with 1 GByte/s access"

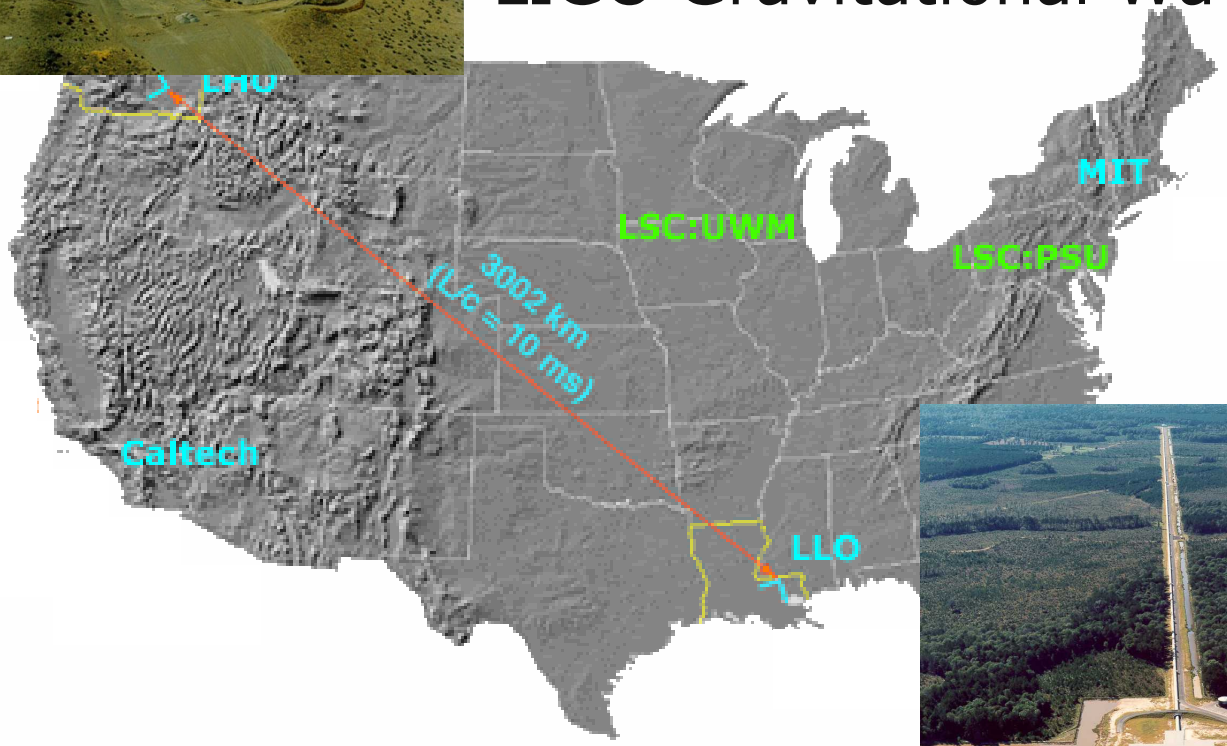


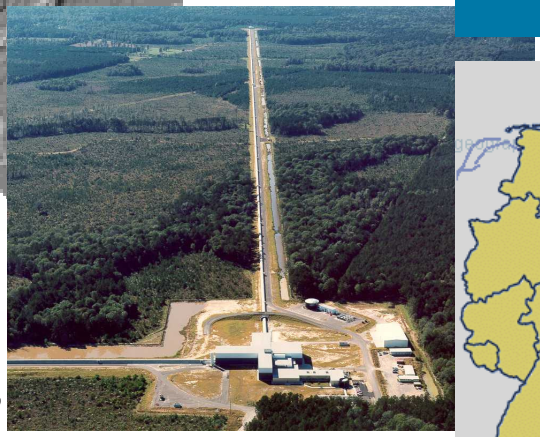
Argonne NATIONAL LABORATORY

THE UNIVERSITY of WISCONSIN MADISON

Data Replication Service

LIGO Gravitational Wave Observatory



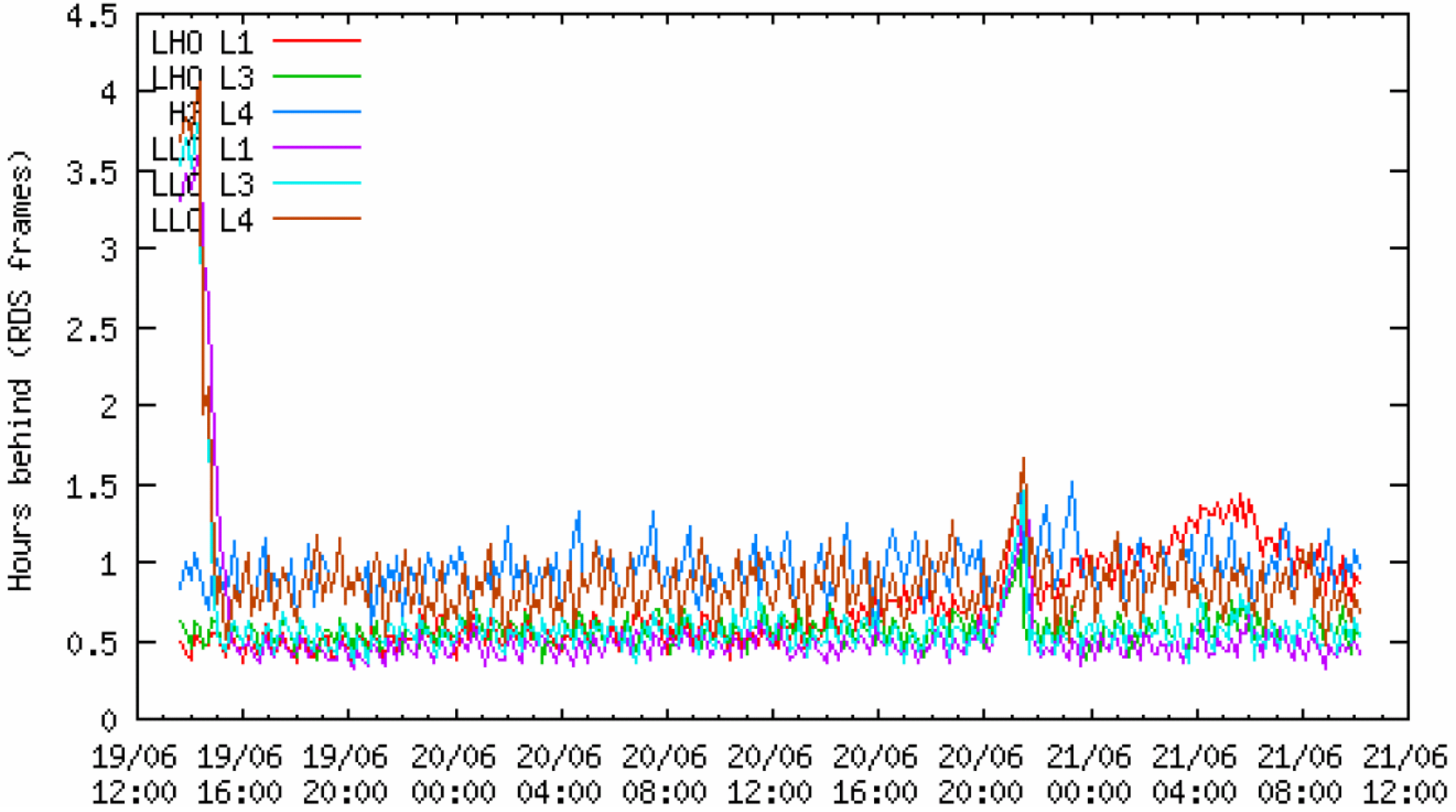


Replicating >1 Terabyte/day to 8 sites

770 TB replicated to date: >120 million replicas

MTBF = 1 month

Lag Plot for Data Transfers to Caltech

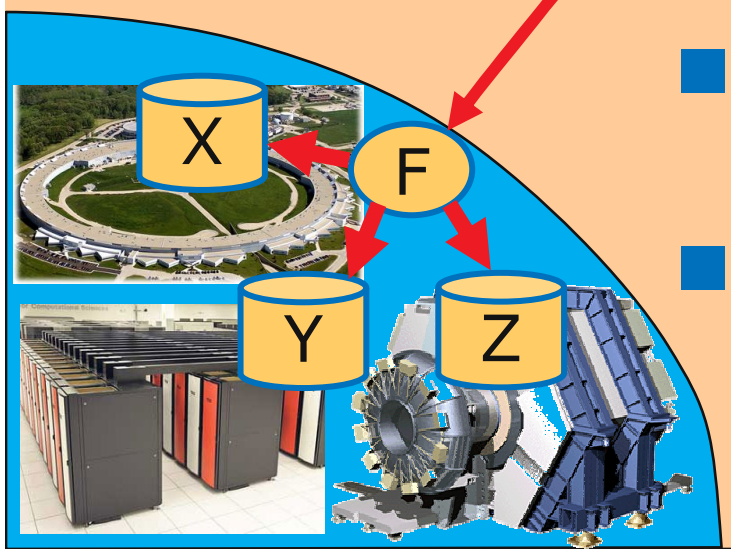


Credit: Kevin Flasch, LIGO

Bridging the Divide (2): Allow Users to Move Computation Near Data

"Perform computation F on datasets X, Y, Z"

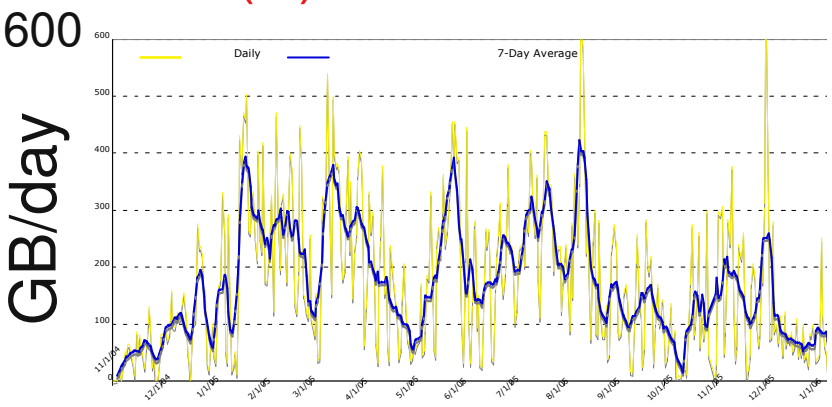
- **Science services**: provide analysis functions near data source

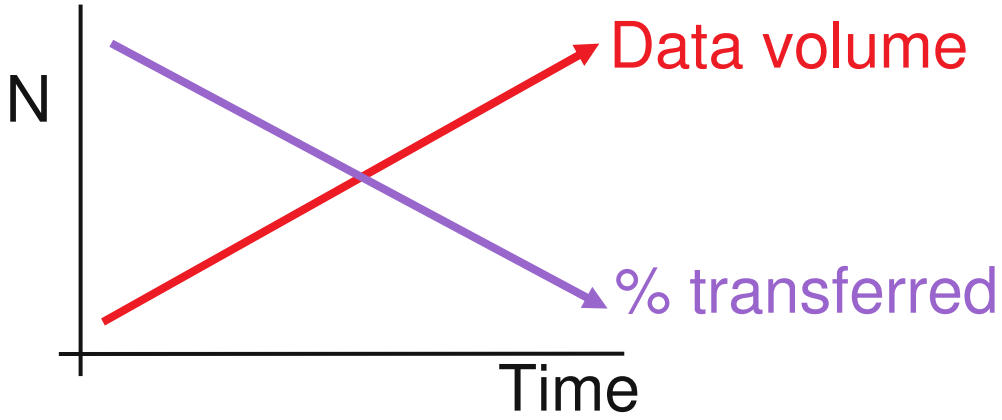


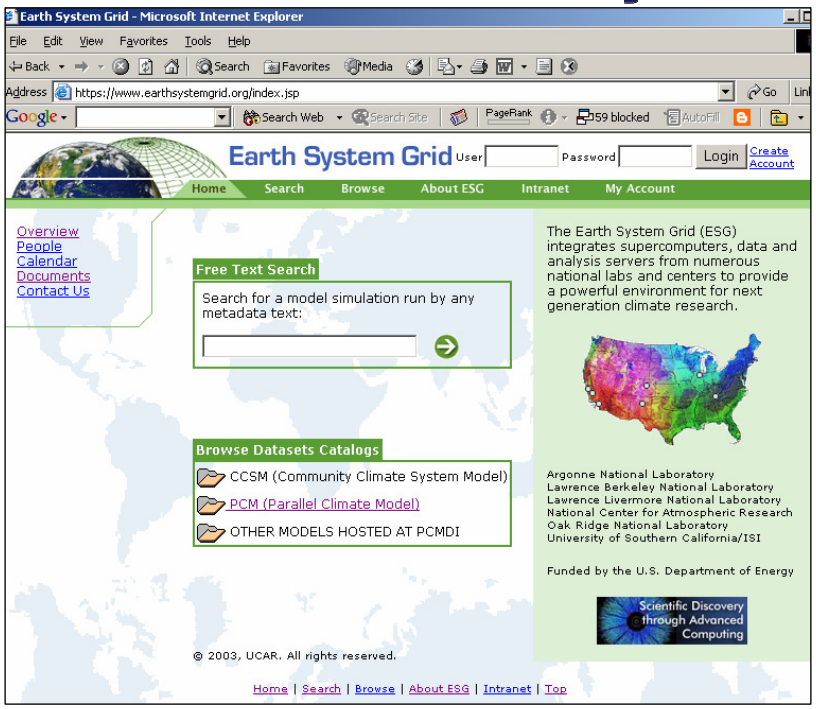
- **Flexible**: easy integration of functions
- **Secure**: protect expensive resources & data
- **Scalable**: deal with many users & much data

For Example …

- Entire datasets
  - X
- Data subsets
  - X[1:10, 1:50:2, 6]
- Predefined operations
  - ZonalMean(X)
- User-defined operations
  - f(X)





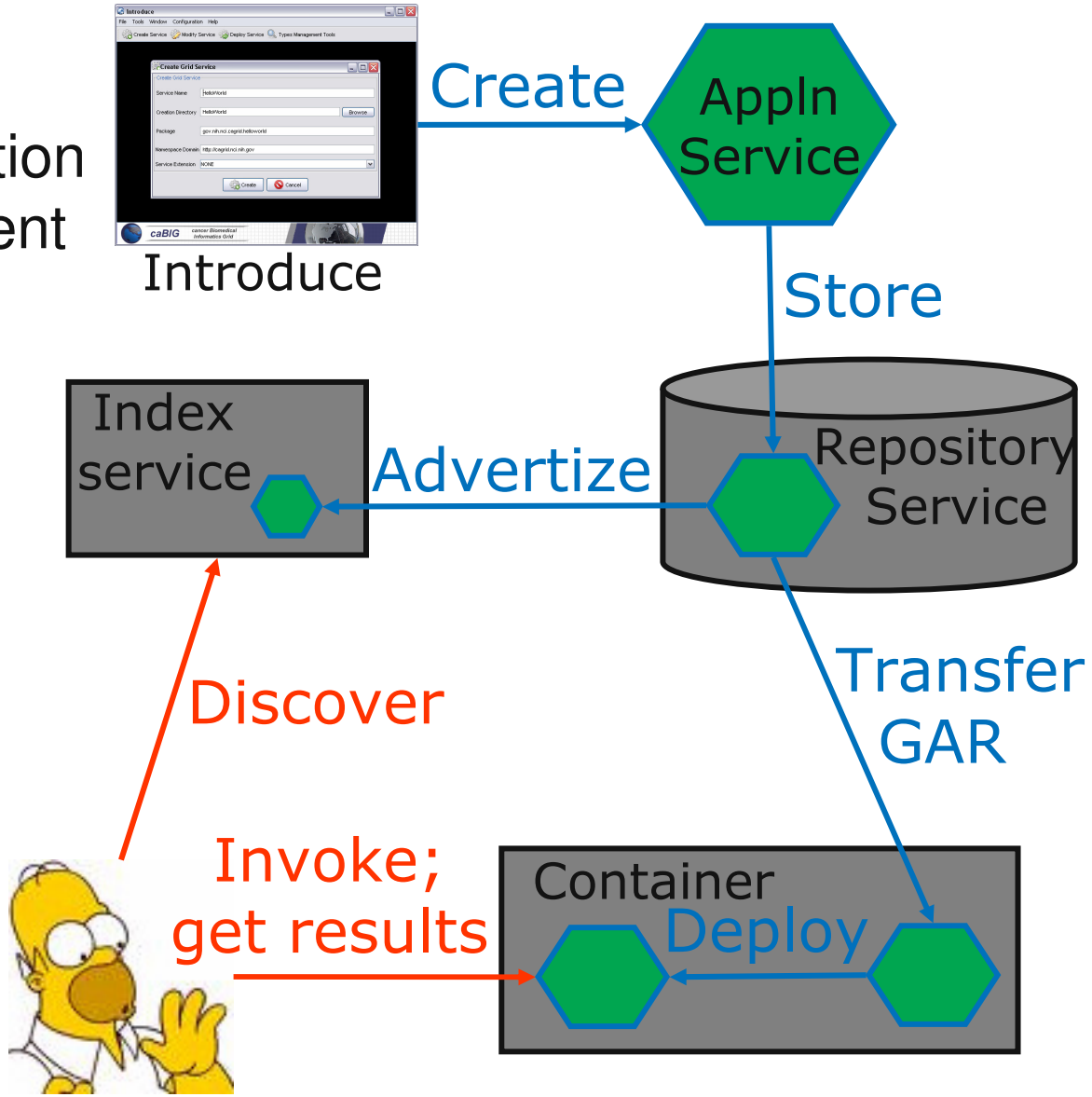


Server-Side Processing: Challenges

- Service authoring
  - Easy creation of “services” encapsulating data and/or computation
- Provisioning
  - Allocate resources to services and to other computations as demand changes
- Code portability and security
  - Encapsulation and portability of application code

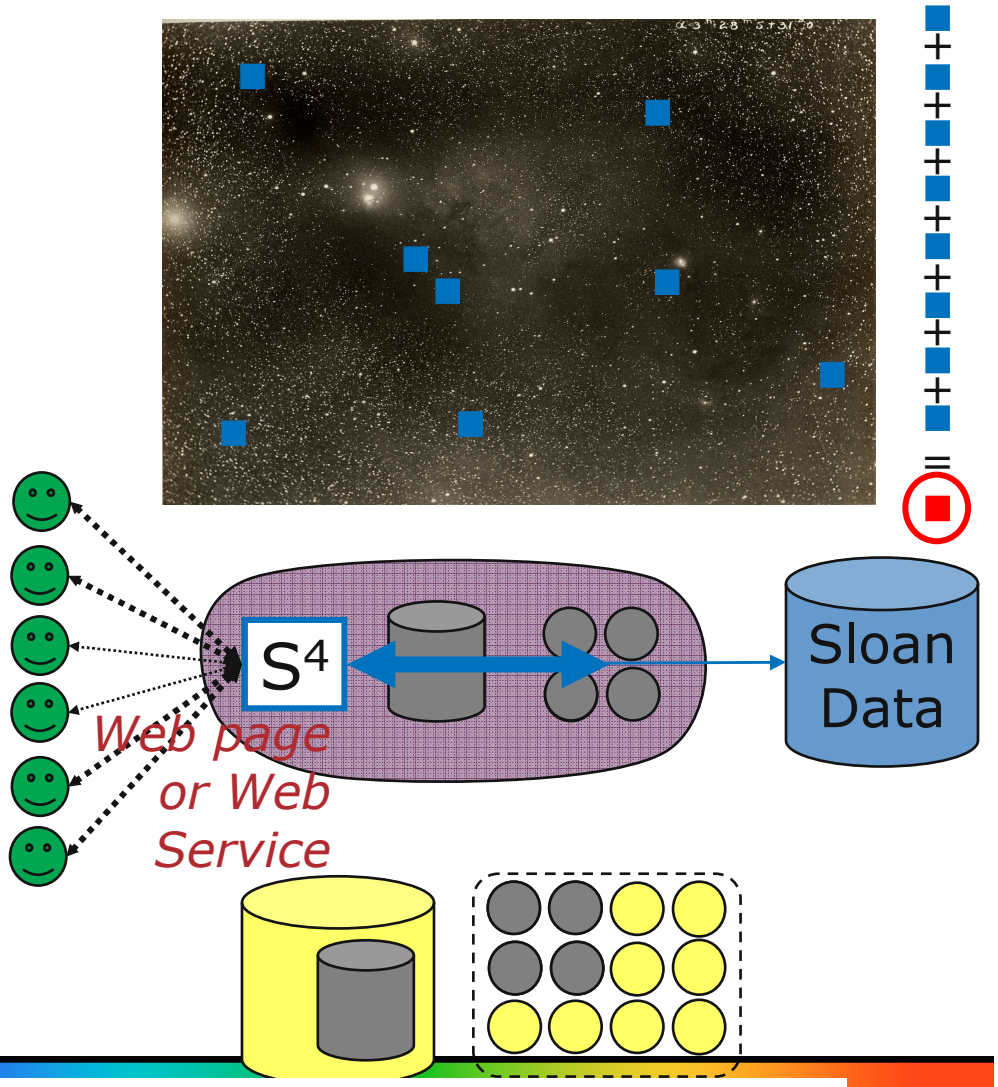
Automated Service Creation Tools

- RAVE: Remote Application Virtualization Environment (Ravi Madduri et al.)
  - Builds on Introduce
  - Define service
  - Create skeleton
  - Discover types
  - Add operations
  - Configure security
  - Wrap arbitrary executables
- pyGlobus tools (Keith Jackson et al., LBNL)

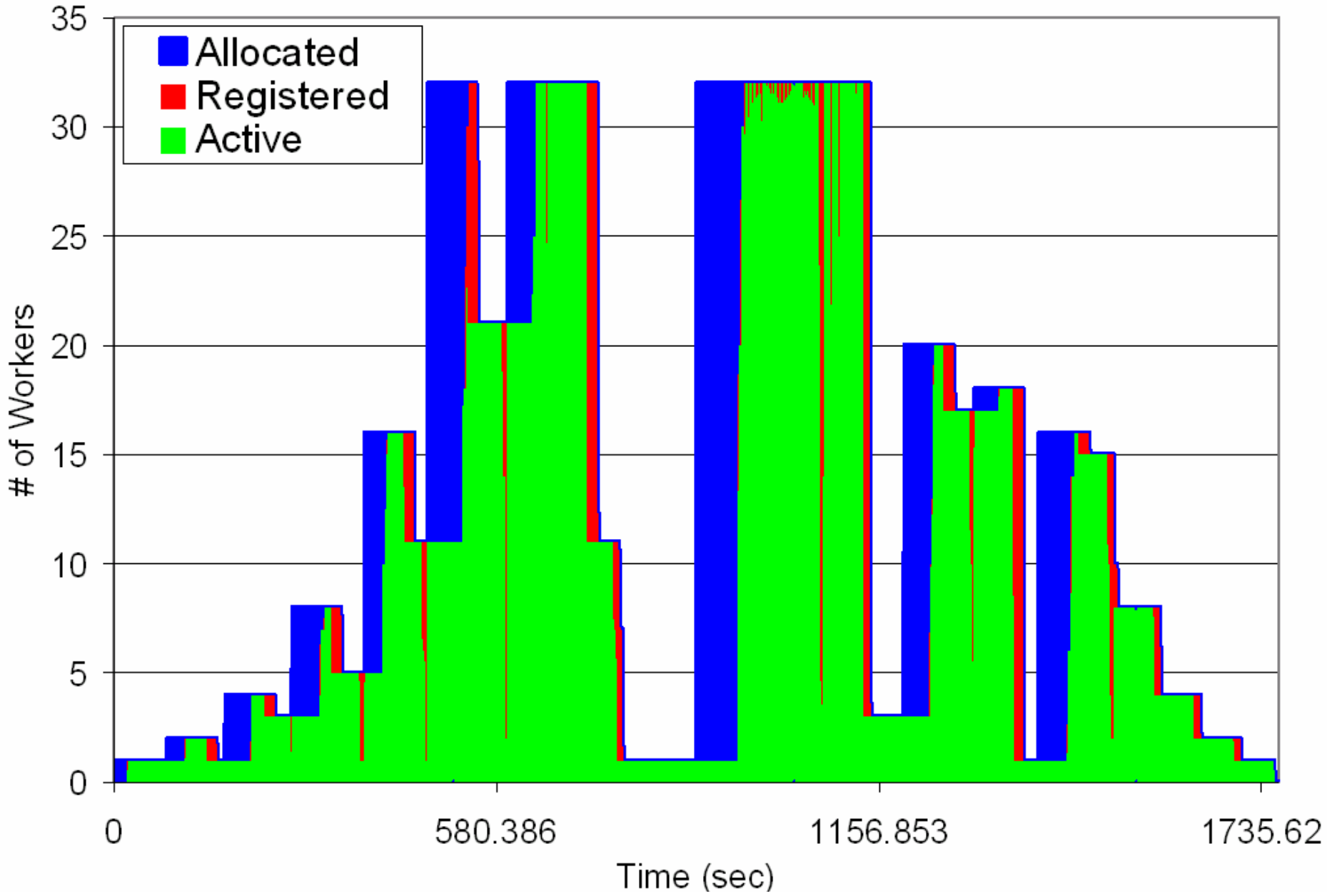


Provisioning: Stacking Service

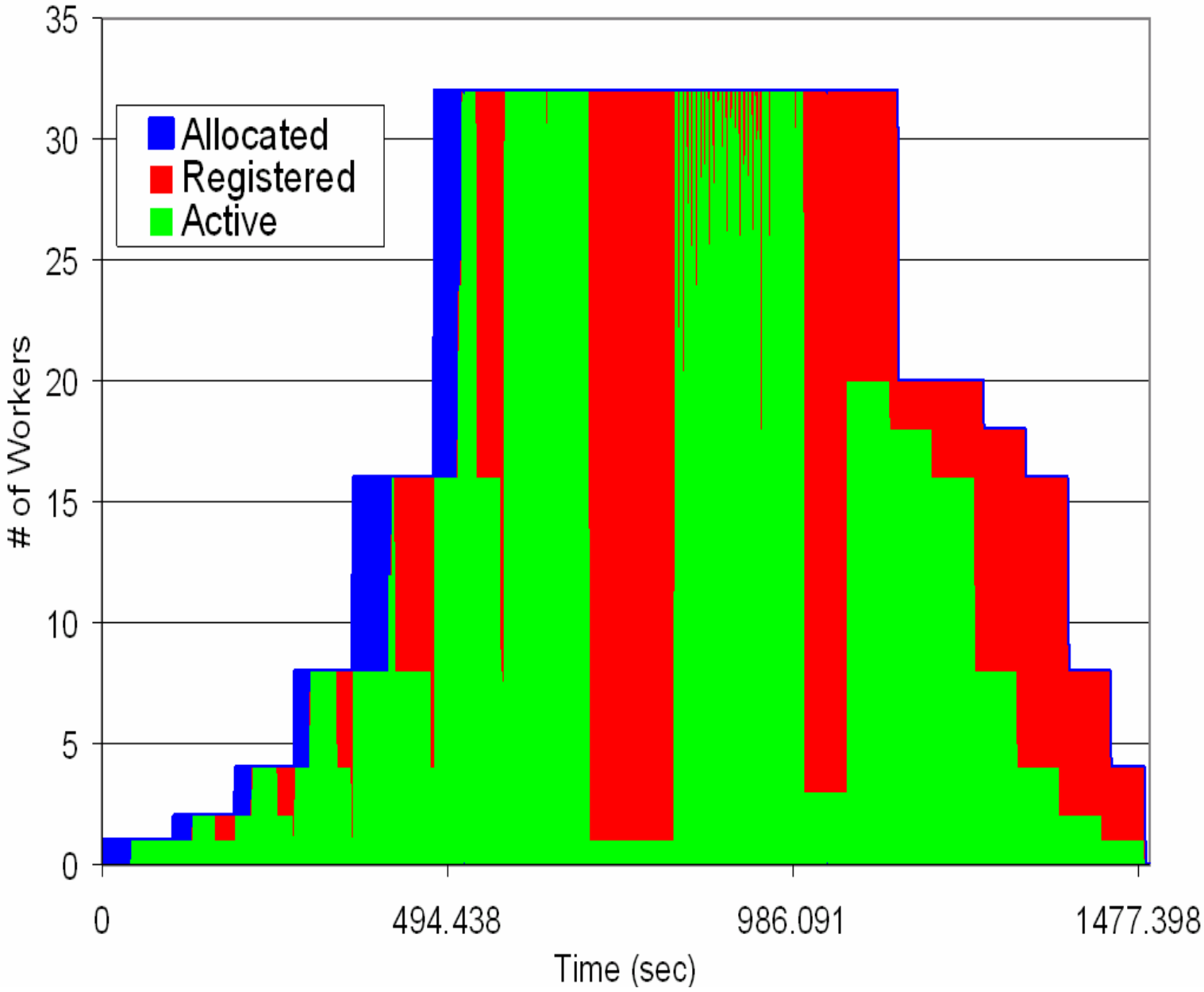
- Purpose
  - On-demand "stacks" of random locations within ~10TB dataset
- Challenge
  - Rapid access to 10-10K "random" files
  - Time-varying load
- Solution
  - Dynamic acquisition of compute, storage



Release after 15 Seconds Idle



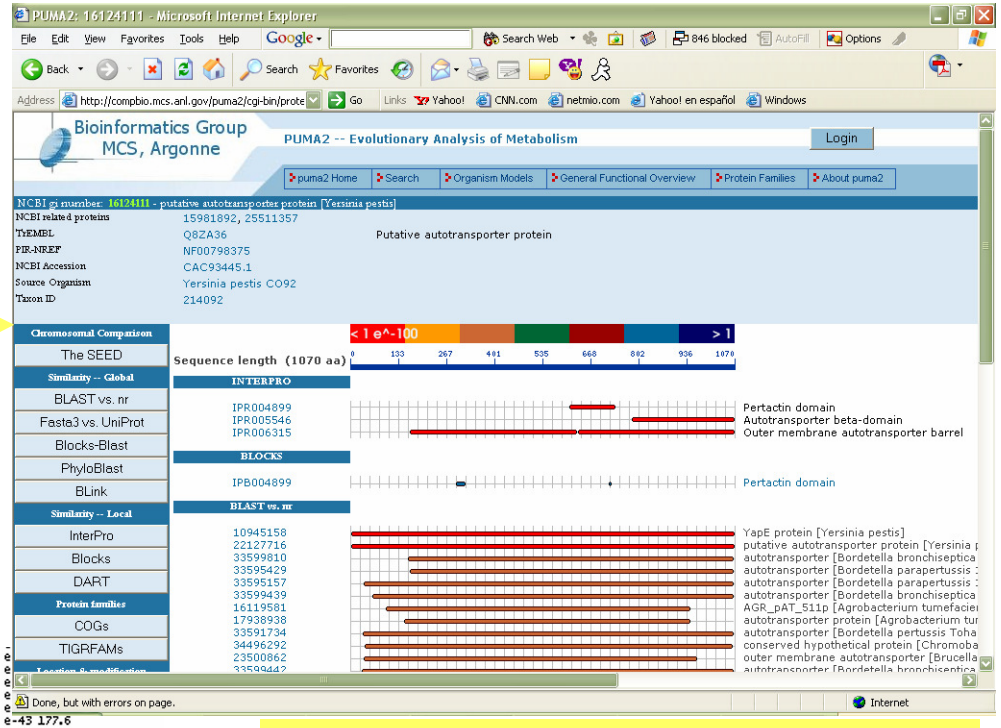
Release after 180 Seconds Idle

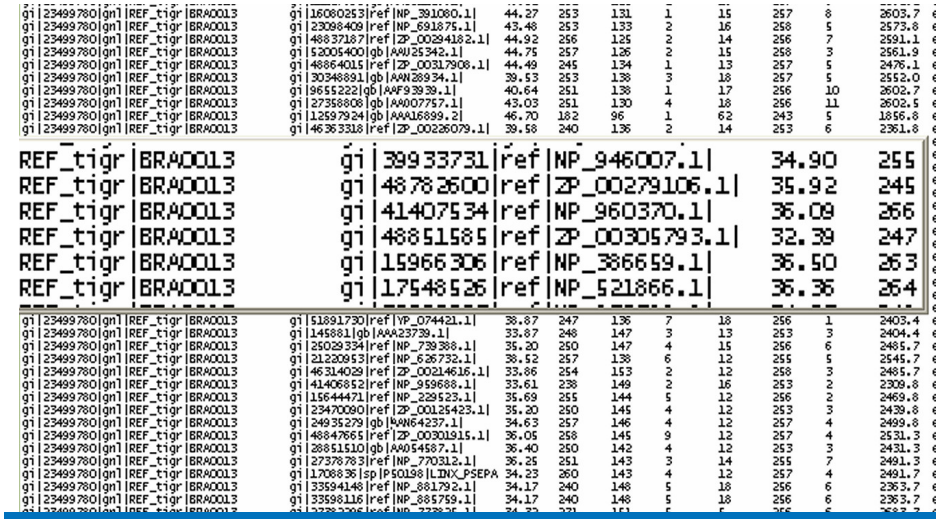


On-Demand Access to Computing in Biology

Public PUMA Knowledge Base

Information about proteins analyzed against ~2 million gene sequences



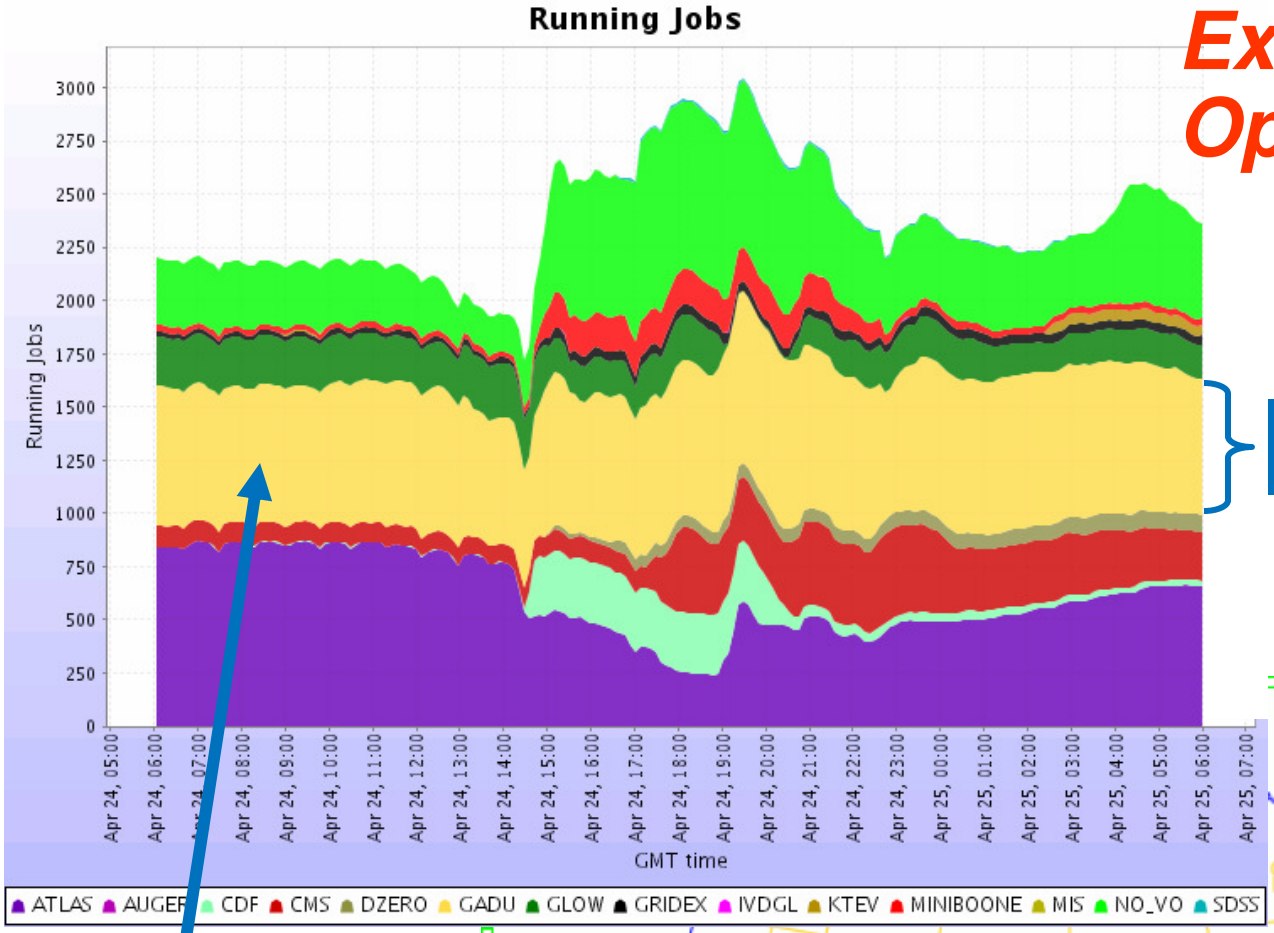


Back Office Analysis on Grid

Millions of BLAST, BLOCKS, etc., on OSG and TeraGrid

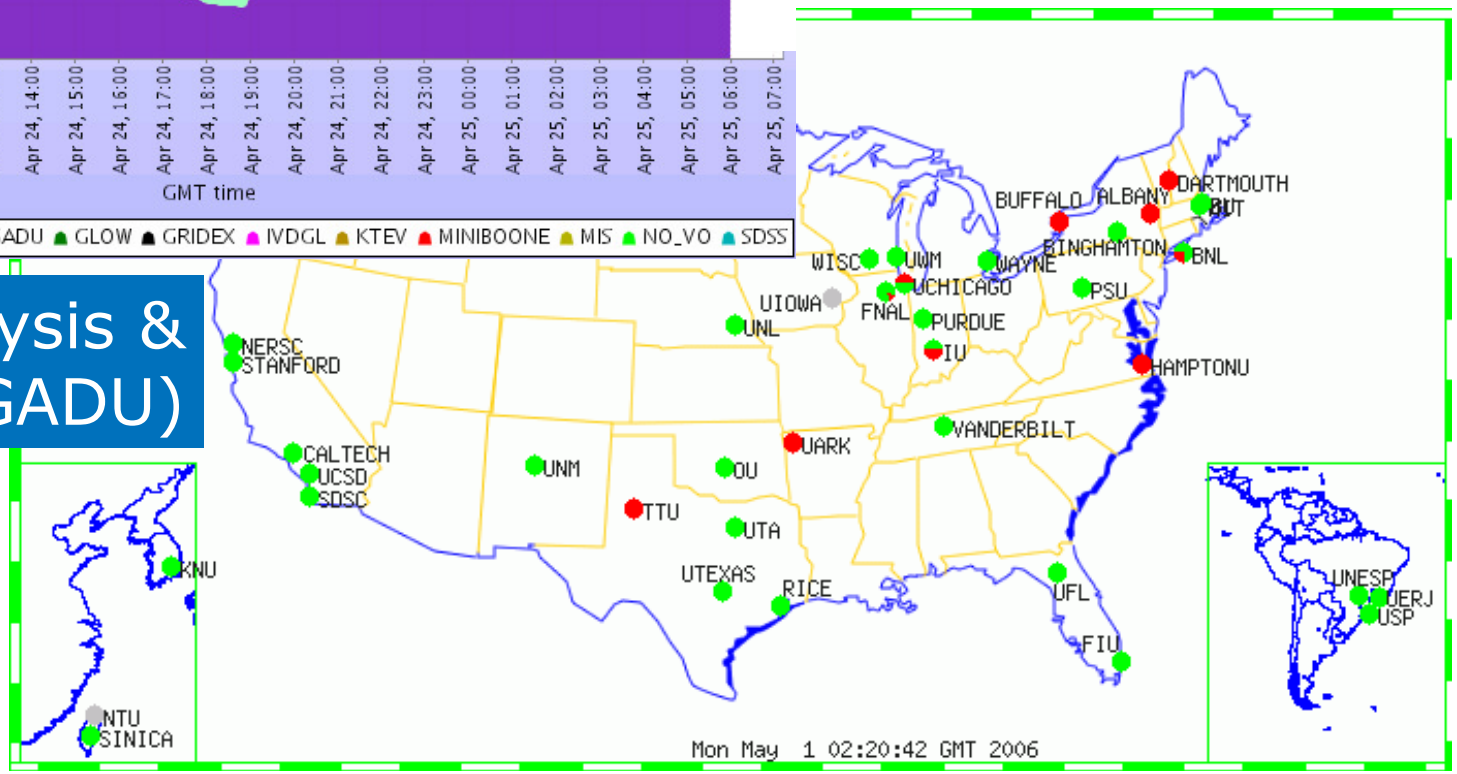
Natalia Maltsev et al., http://compbio.mcs.anl.gov/puma2

Execution on Open Science Grid



600-1000+ CPUs

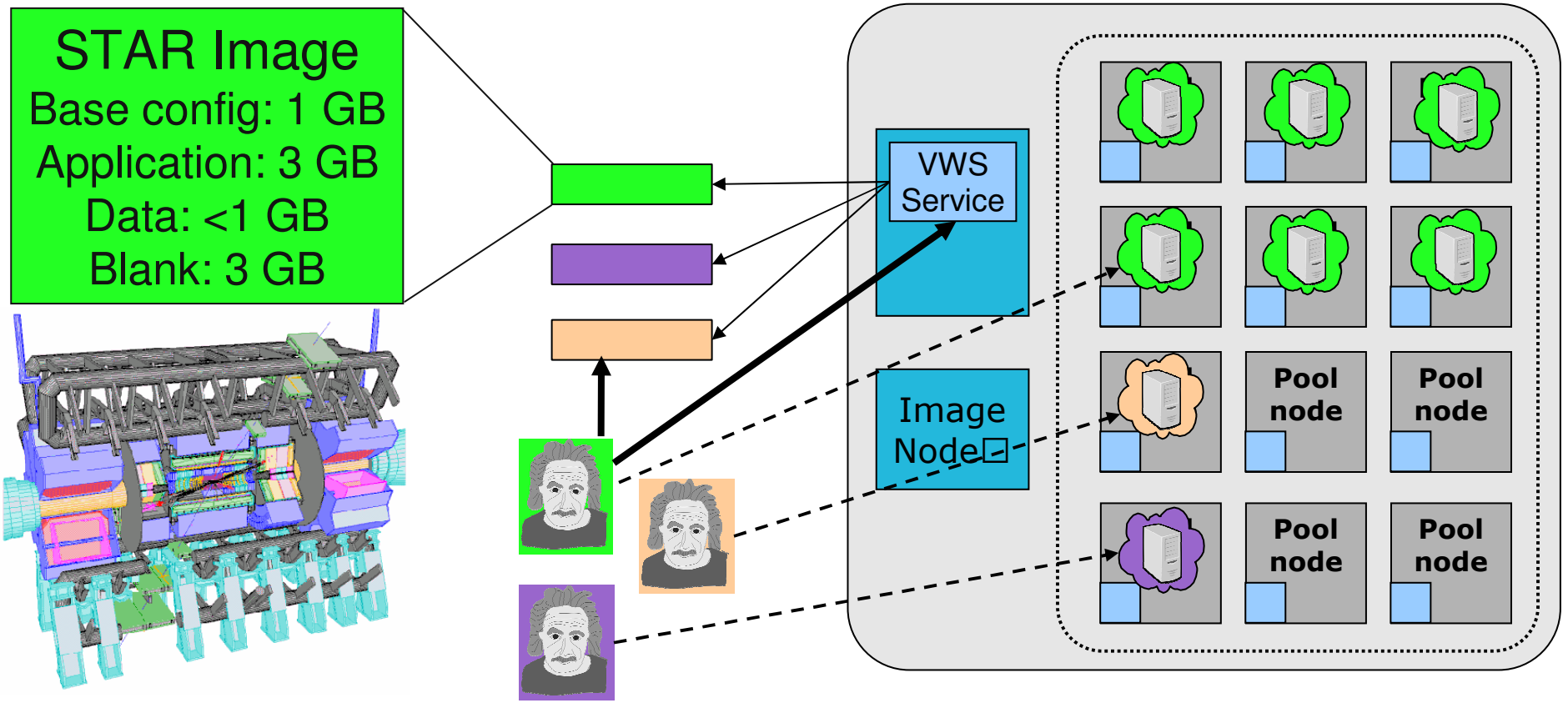
Genome Analysis & DB Update (GADU)



Open Science Grid

Configuration, Portability, and Encapsulation

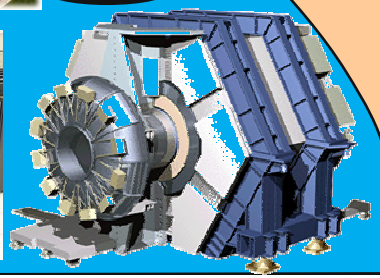
Virtual workspace service: use virtual machine (VM) technology to enable rapid deployment of complex codes on new computers



Bridging the Divide (3): Troubleshoot End-to-End Problems

"Why did my data transfer (or remote operation) fail?"

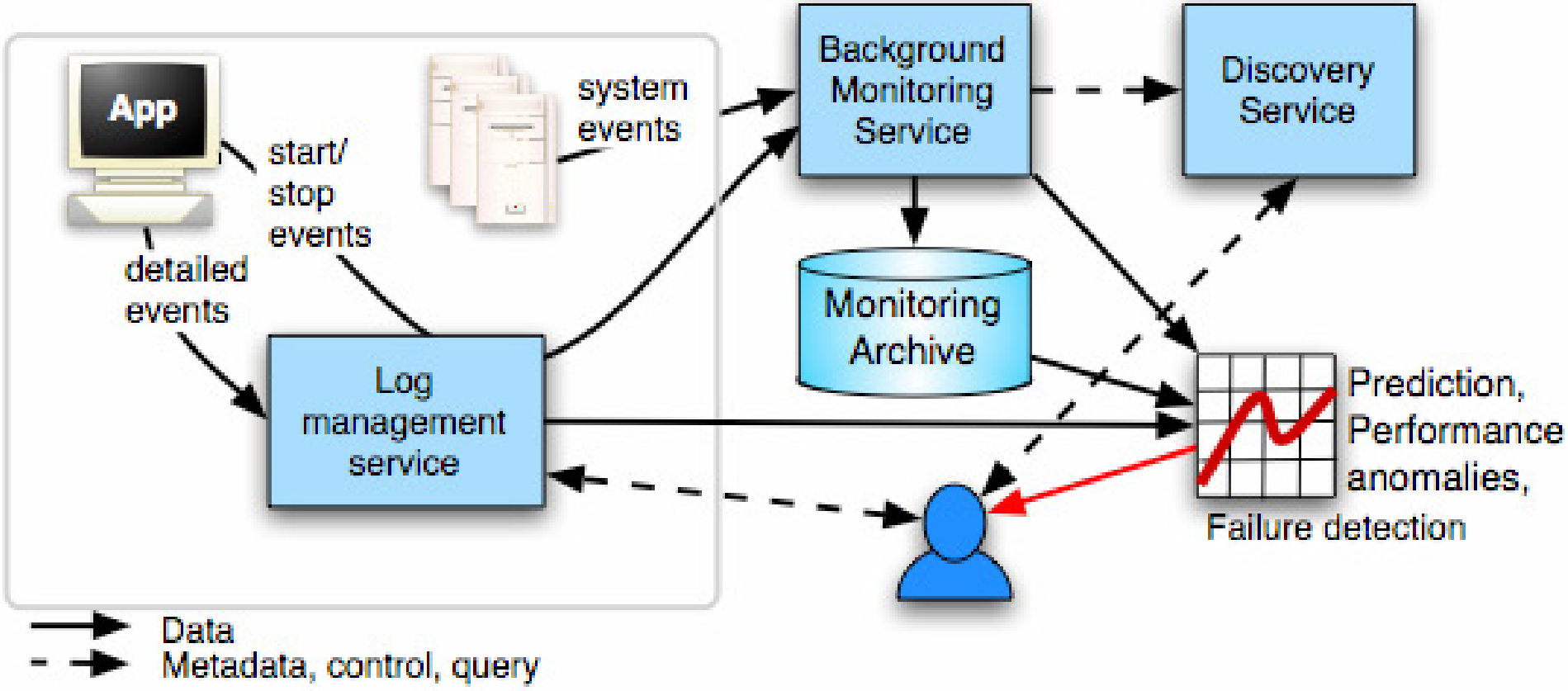
- **Identify** & **diagnose** failures & performance problems



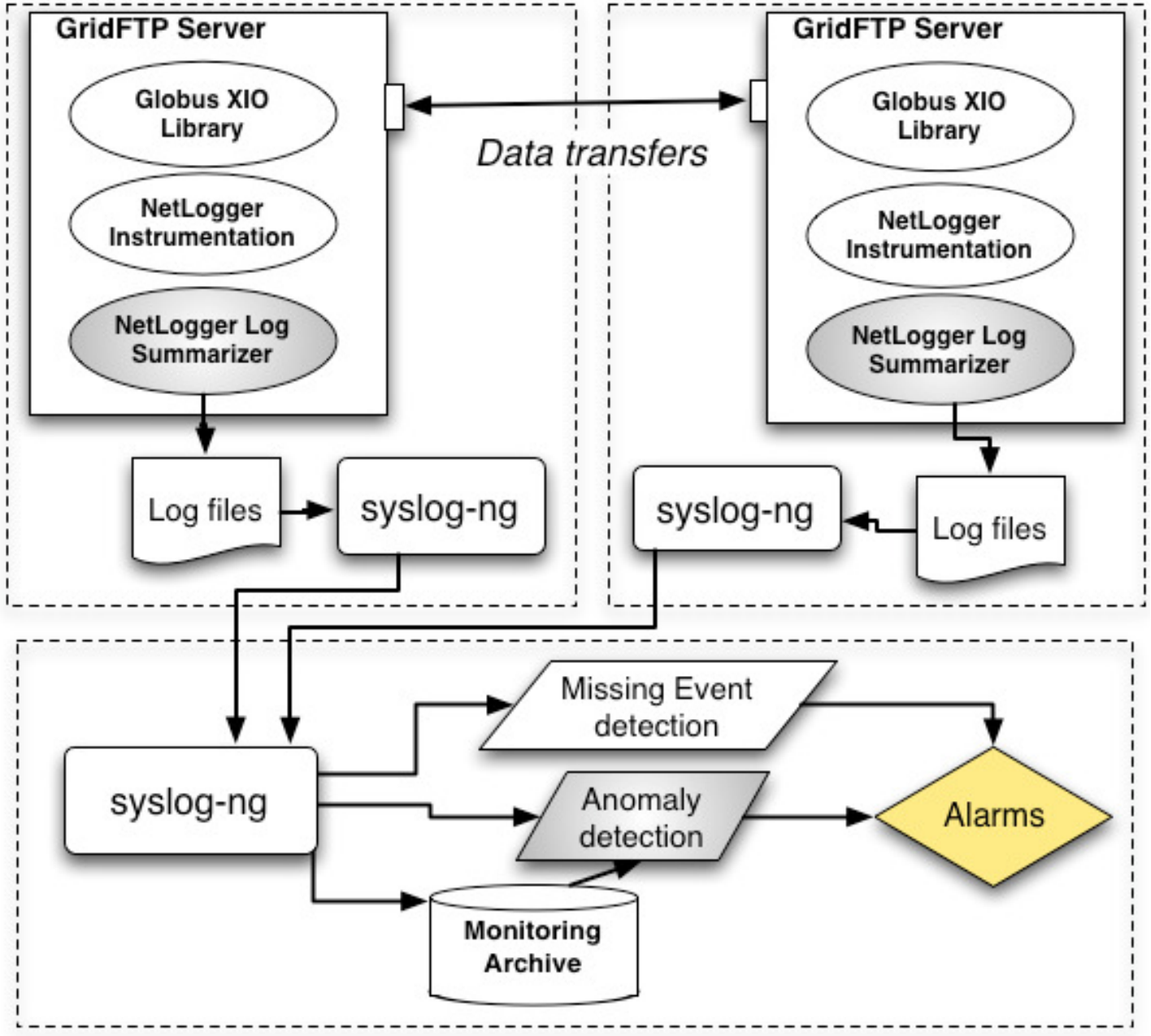
- **Instrument**: include monitoring points in all system components
- **Monitor**: collect data in response to problems
- **Diagnose**: identify the source of problems

Troubleshooting Challenges and Approach

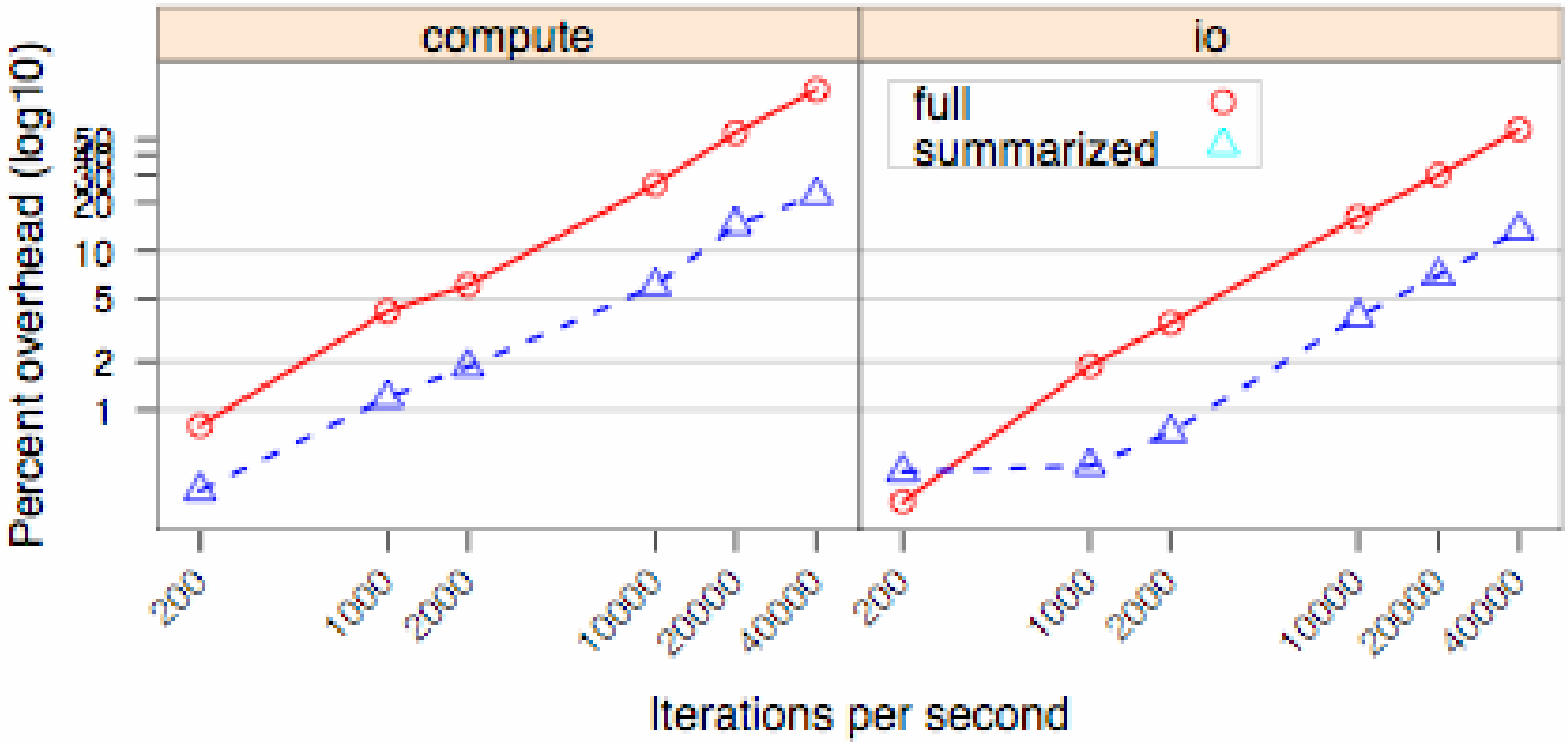
- Many devices
- Many failure cases
- Distributed responsibility
- Interactions between components



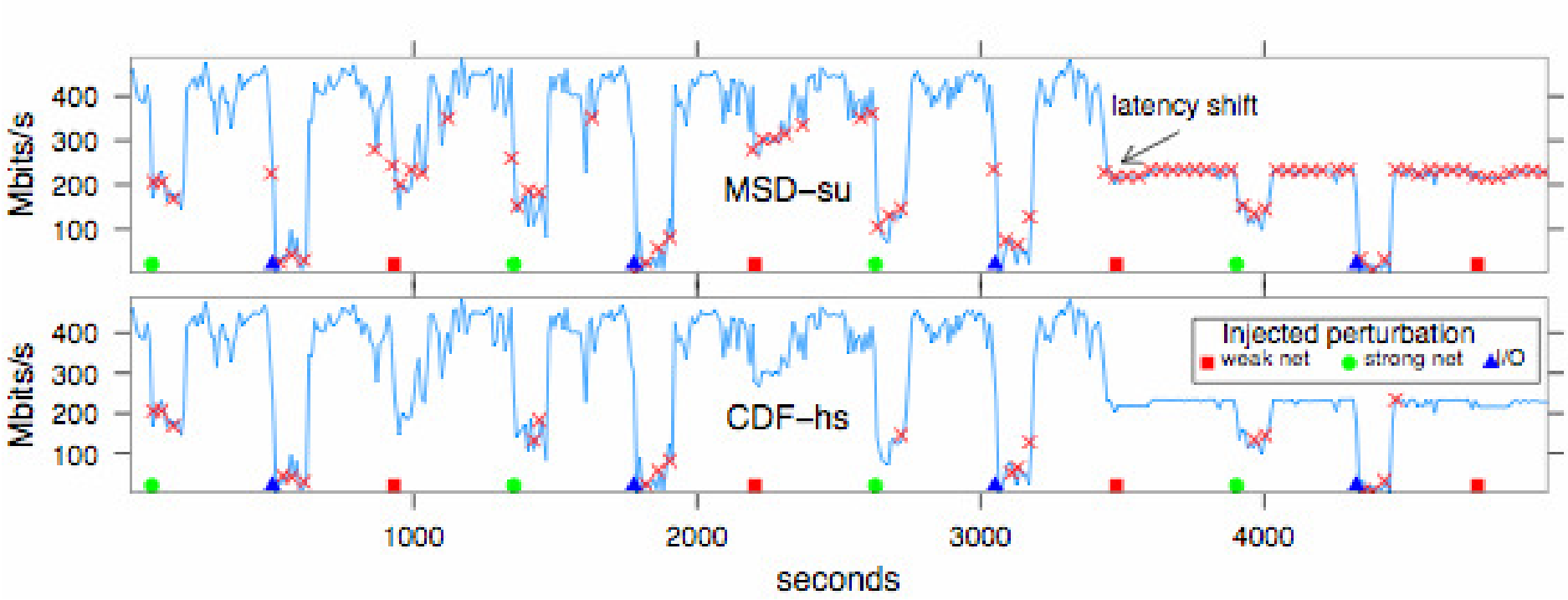
CEDPS Log Generation & Collection: GridFTP

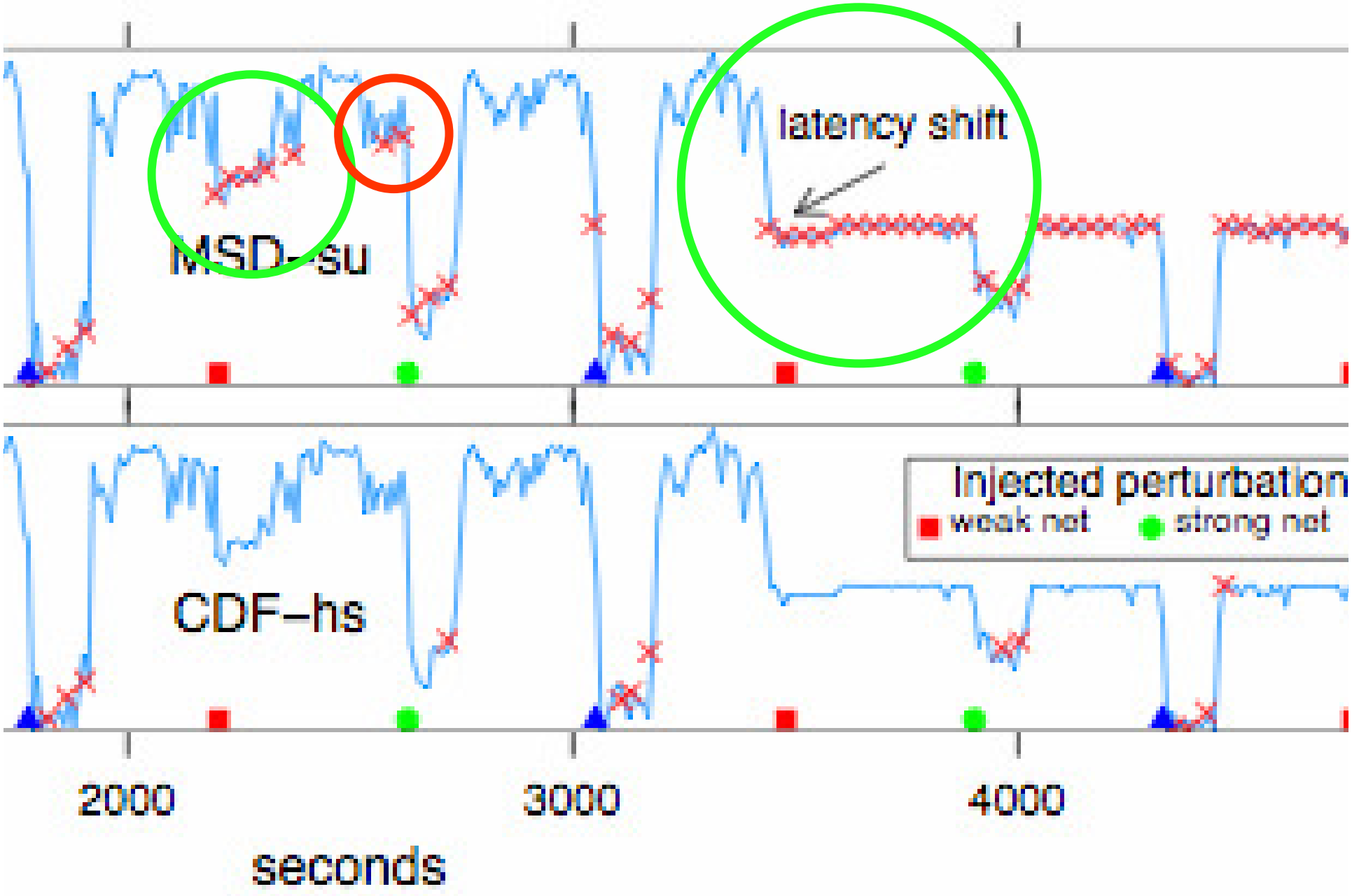


Perturbation Caused by Trace Generation



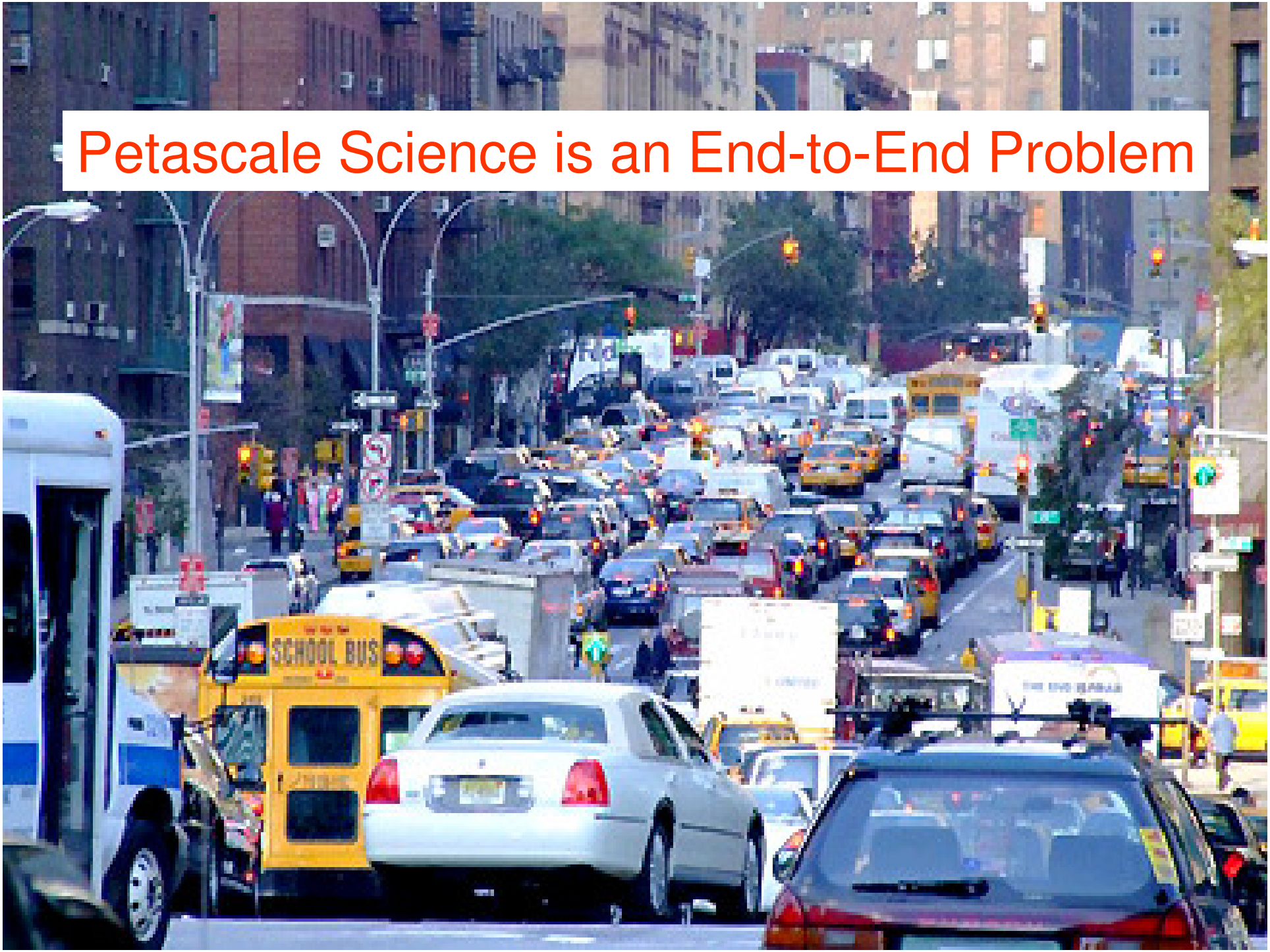
GridFTP with Injected Performance Perturbations




latency shift
MSD-su
CDF-hs
Injected perturbation
weak net
strong net
2000
3000
4000
seconds

Major Issues that are Currently Neglected

- Security
  - Managing who within dynamic "virtual organizations" can use what resources, access what data, perform what computations
  - Protecting end-to-end systems against attack
- The last mile
  - Need to beef up campus infrastructures to enable effective engagement with petascale science
- Connecting the ends
  - Enabling sharing of data and services among users of petascale (and terascale) resources


Petascale Science is an End-to-End Problem
SCHOOL BUS

Let Us Turbocharge your Science …

- Tools exist today, e.g.:
  - GridFTP for data transfer
  - DRS for data replication
  - RAVE and pyGlobus for service authoring
  - OSG for on-demand access to computing
- Yet better tools are on the way:
  - MOPS for storage and bandwidth management
  - Virtual machines for portability and encapsulation
  - End-to-end troubleshooting

Come to the CEDPS tutorial on Friday morning!