CEDPS: Center for Enabling Distributed Petascale Science

Jennifer Schopf

Argonne National Laboratory
University of Chicago

www.cedps.net

Argonne NATIONAL LABORATORY

... for a brighter future

BERKELEY LAB

Fermilab

ISI Information Sciences Institute USC School of Engineering

THE UNIVERSITY of WISCONSIN MADISON

U.S. Department of Energy

UChicago Argonne LLC

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

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

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

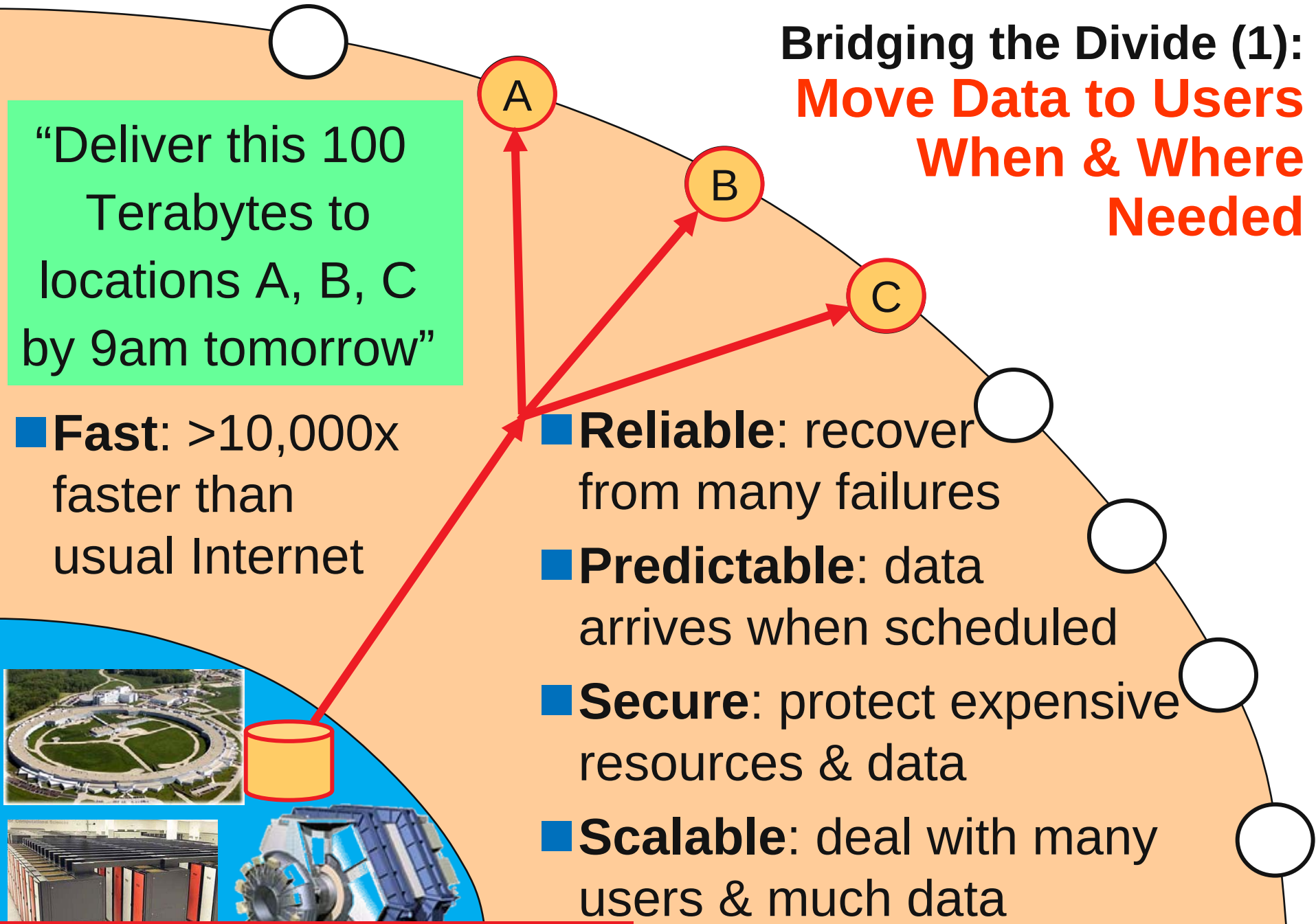
Center for Enabling Distributed Petascale Science (CEDPS)

- 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

CEDPS

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

Area Lead: Ann Chervenak, ISI

Data Delivery Challenges

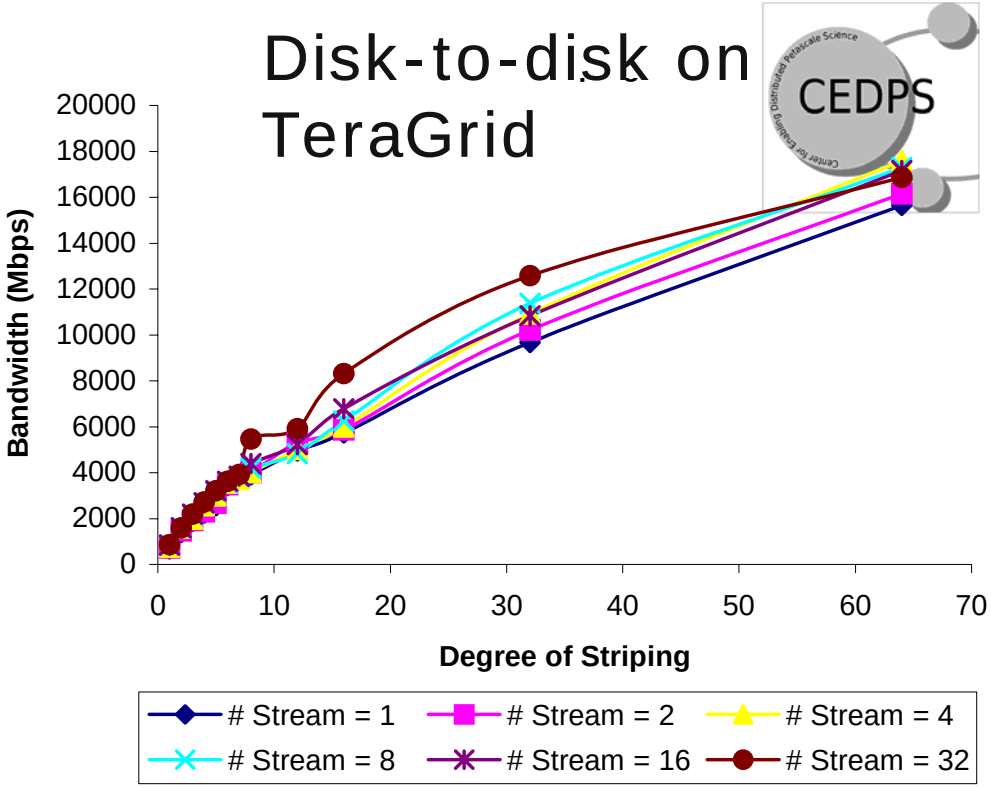
- Data complexity
- Parallelism (in diverse places)
- Network heterogeneities (e.g., firewalls)
- Space (or the lack of it)
- Protocols and their peculiarities
- Failures at many levels
- Deadlines
- Resource contention
- Multiple participants

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

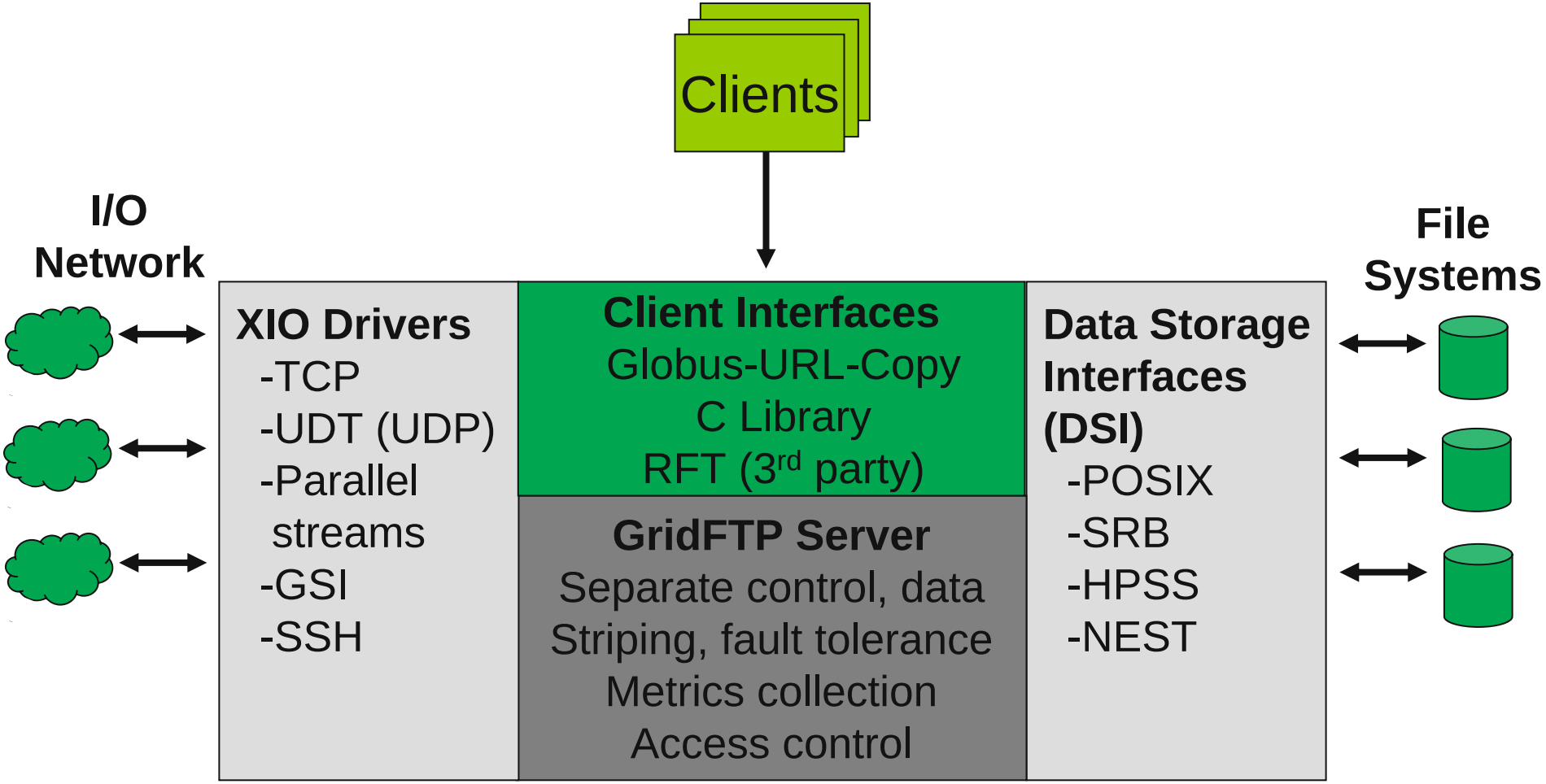
GridFTP

GridFTP in GT4



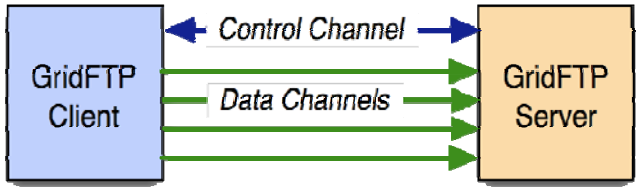
- A high-performance, secure, reliable data transfer protocol
- Optimized for high-bw wide-area networks
- GSI support for security
  - SSH tunneling now available
- 3rd party and partial file transfer support
- XIO for different transports
- Parallelism and striping → multi-Gb/sec wide area transport

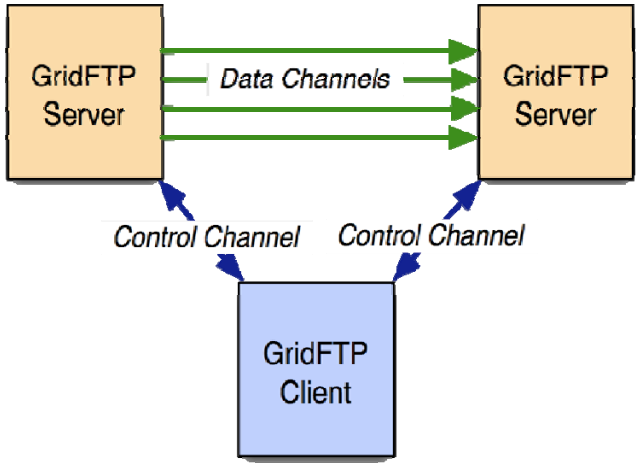
Current State of the Art: GridFTP



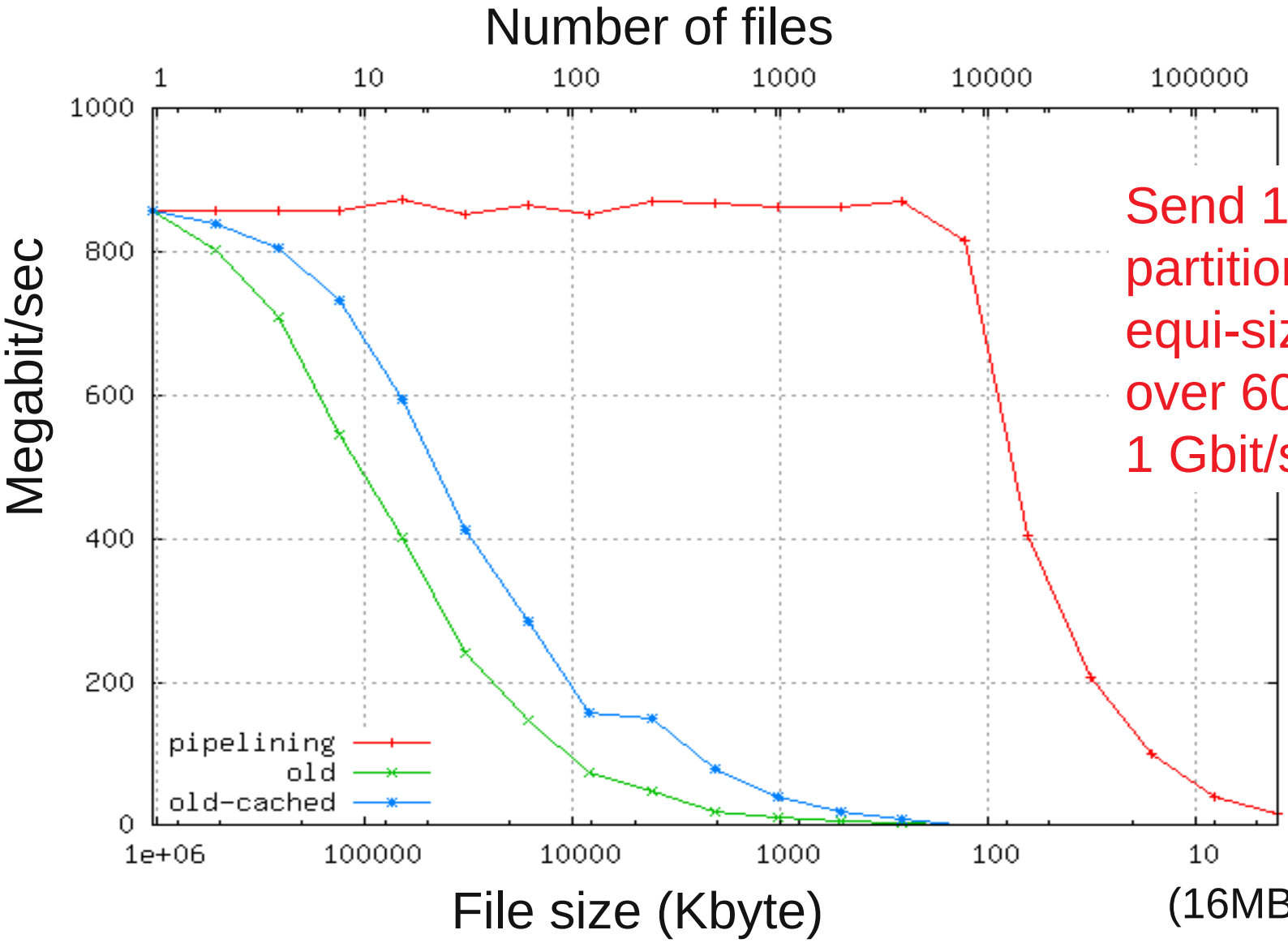
Recent Improvements: GridFTP over SSH

- The Problem
  - Not all users require GSI and the need for certificate infrastructure.
- The Solution
  - Use SSH for Control Channel
  - Data channel remains as is, so performance is still GridFTP
- Included in 4.1.2 development release





"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)

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

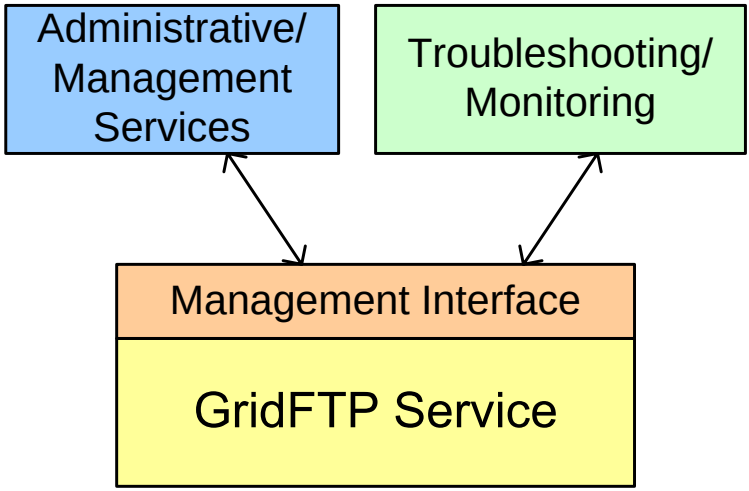
MOPS: Managed Object Placement Service

New Service: MOPS
Managed Object Placement Service

- Enhancement to today's GridFTP—that allows for management of:
  - Space, Bandwidth, Connections
  - Other resources needed to endpoints of data transfers
- Building blocks:
  - GridFTP server (Globus): Add resource management
  - NeST storage appliance (U Wisconsin): Provides storage and connection management
  - dCache storage management (Fermi): Improve scalability and fault tolerance

MOPS Features: Merge GridFTP with NeST

- Better internal resource management
- To overcome issues with GridFTP servers overwhelming resources
- Management interface
  - System admins will prescribe resource limits for GridFTP service (maximum CPU, memory usage, connections, bandwidth)
  - MOPS will report back on current state of its resources to administrative services, troubleshooting
- Release 1.0 available Oct 1, 2007
- http://www.cedps.net/wiki/index.php/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

GridFTP

MOPS: Managed Object Placement Service

DRS: Data Replication Service

New Service: Data Placement Service

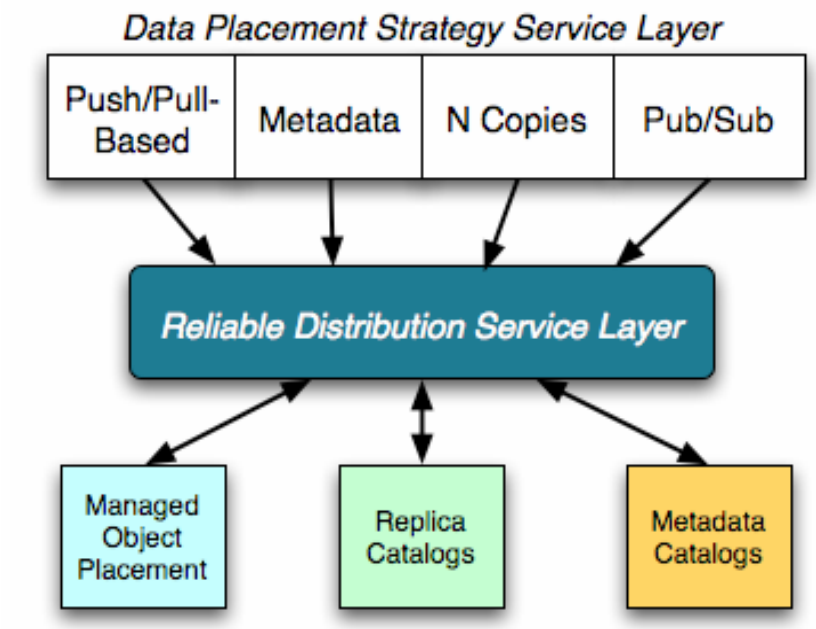
- Data placement and distribution services implement different data distribution and placement behaviors
  - Decide where to place objects and replicas
  - Policy-driven, based on needs of application and the VO
  - Effectively creates a placement workflow
- Currently designing the first-generation data placement service as part of CEDPS
  - DOE SciDAC Center for Enabling Distributed Petascale Science
- Seeking application input on the type of placement services they need

Data Placement Policies

- Place explicit list of data items
  - Similar to existing Globus Data Replication Service
- Metadata- or subscription-based placement
  - Place data where it is likely to be accessed by scientists and/or used in performing computations
  - Use results of metadata queries for data with certain attributes or subscriptions
- N-Copies: maintain N copies of data items
  - Placement service checks existing replicas, creates/delete replicas to maintain N copies of each
  - Keeps track of lifetime of allocated storage space, migrates data as necessary

Reliable Distribution Layer

- Responsible for carrying out the distribution or placement "plan" generated by higher-level service
- Provide feedback to higher level placement services on the outcome of the placement workflow
- Call on lower-level services to coordinate

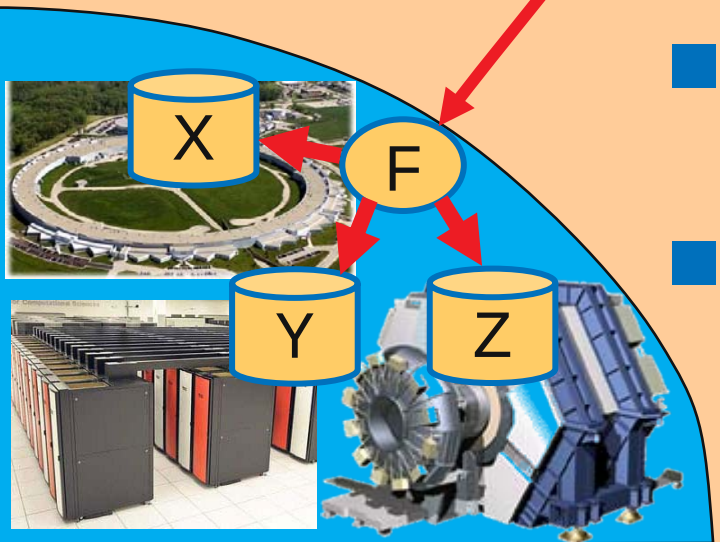


- Release 1.0 available Oct 1, 2007
- http://www.cedps.net/wiki/index.php/Data

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

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

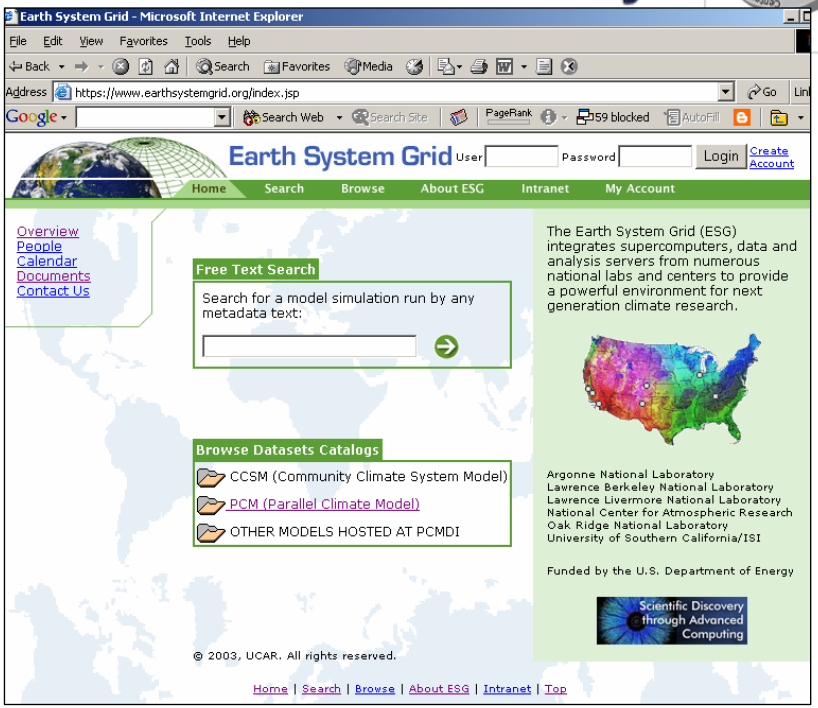
- **Science services**: provide analysis functions near data source

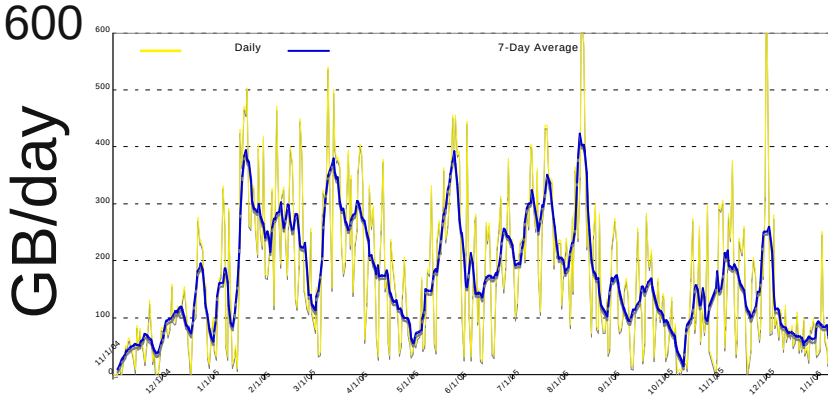


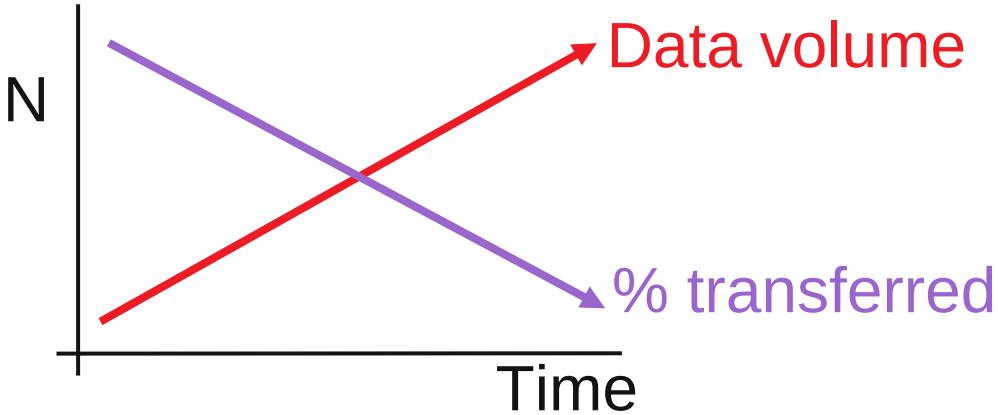
- **Flexible**: easy integration of functions
- **Secure**: protect expensive resources & data
- **Scalable**: deal with many users & much data

For Example …

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





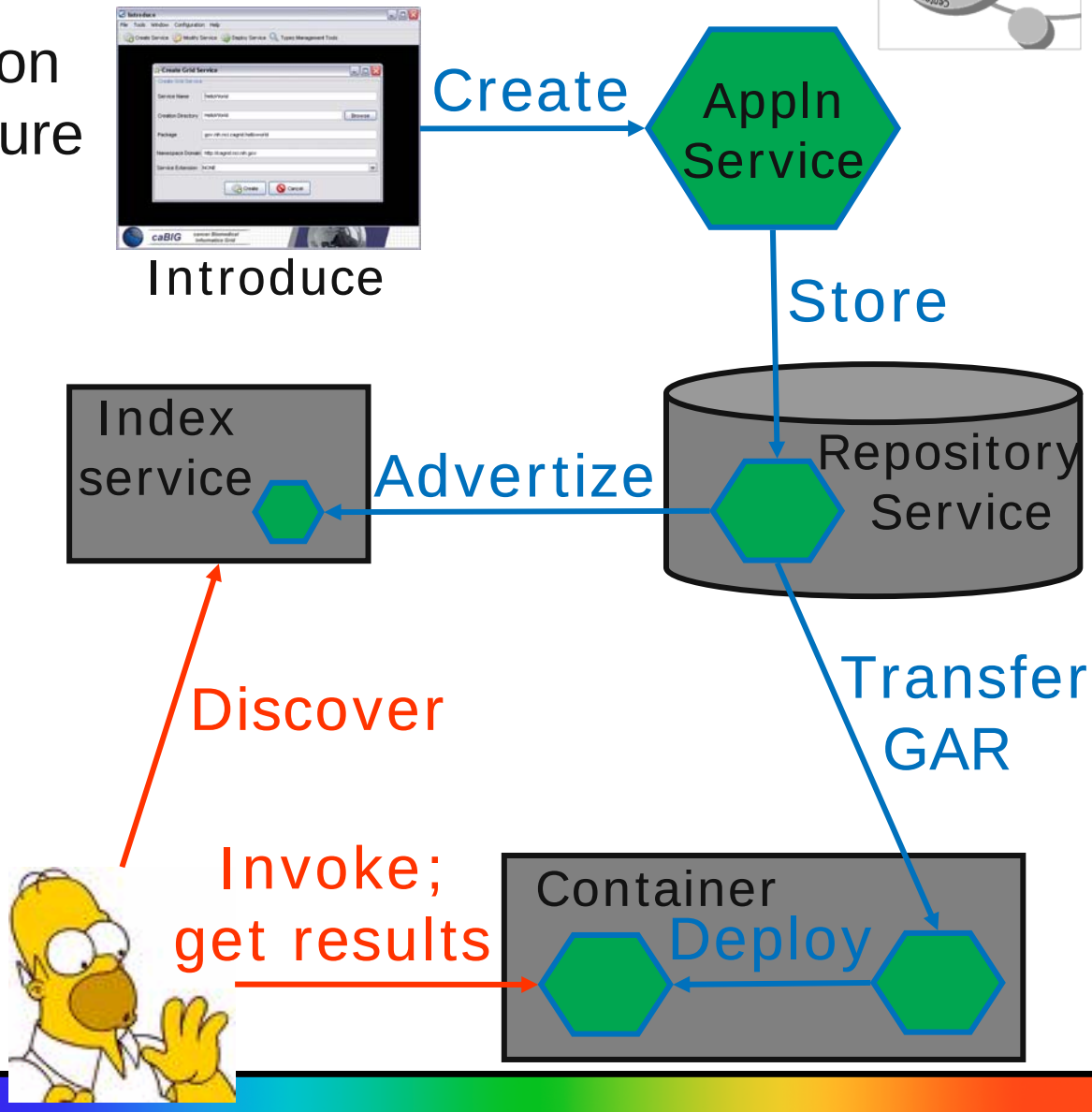


Server-Side Processing: Challenges

- 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

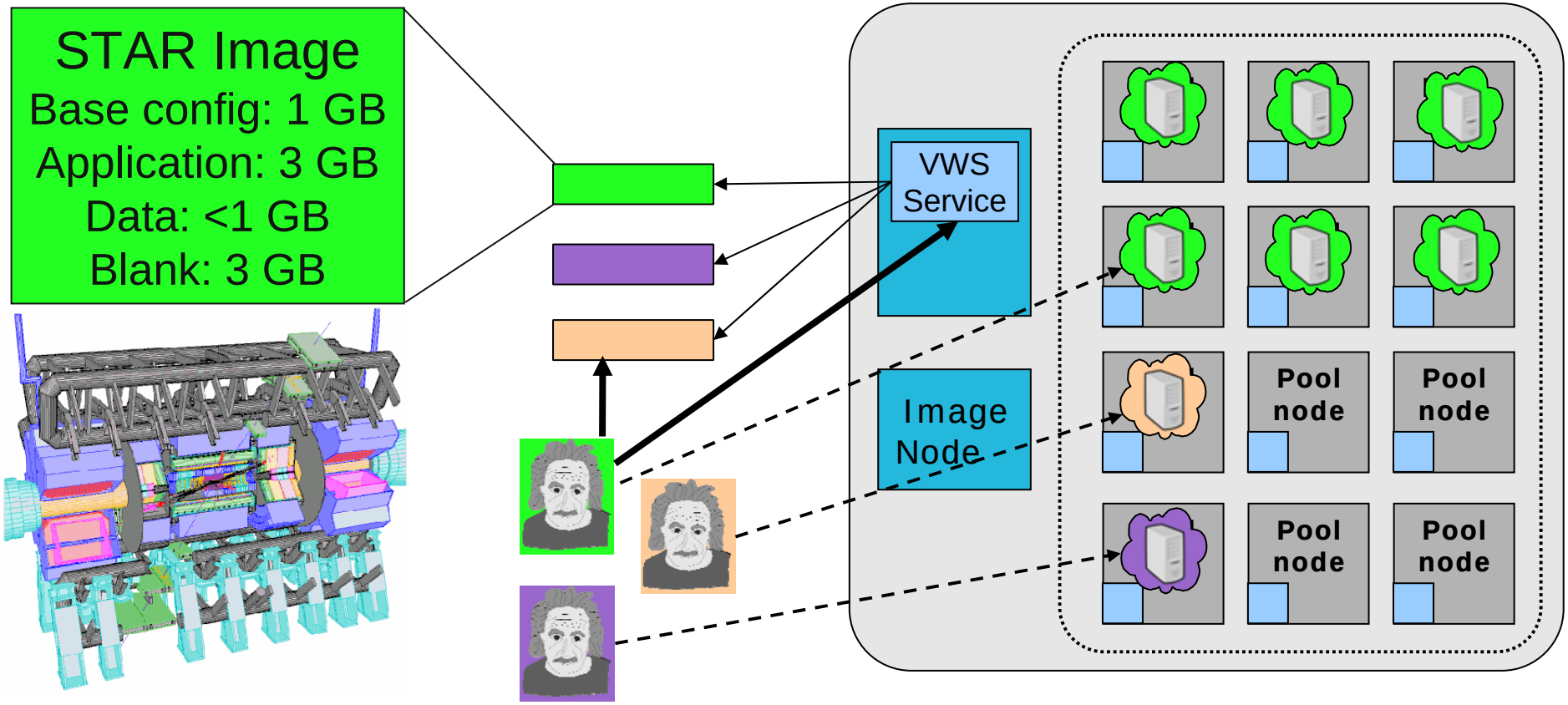
Automated Service Creation Tools

- RAVi: Remote Application Virtualization Infrastructure (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)



Configuration, Portability, and Encapsulation

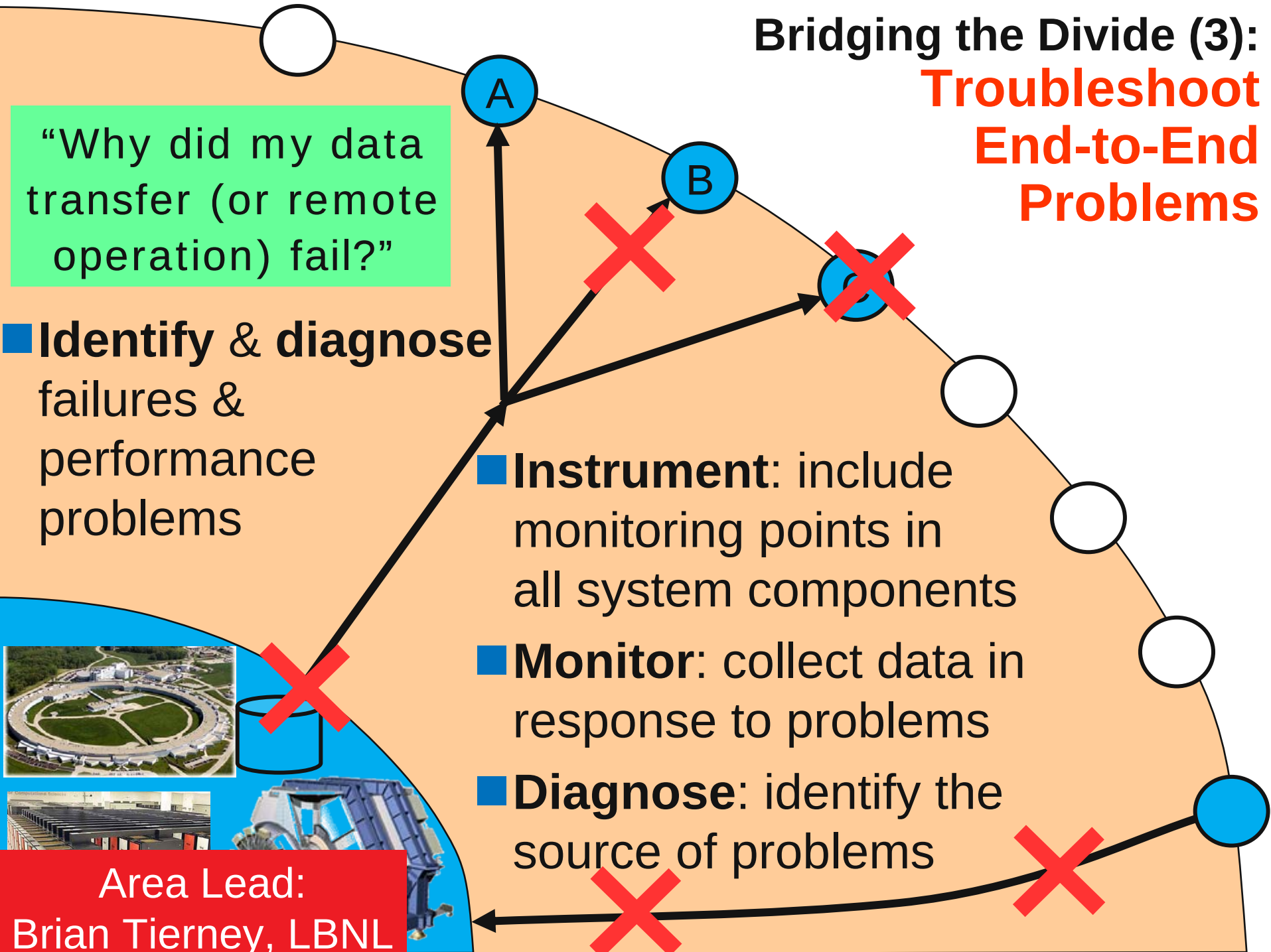
Virtual workspace service: use virtual machine (VM) technology to enable rapid deployment of complex codes on new computers



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

Area Lead:
Brian Tierney, LBNL

Common Logging: The Problem

- Assume your distributed compute job normally takes 30 minutes to complete. But…
  - 3 hours have passed and the job has not yet completed.
- What, if anything, is wrong?
  - Is the job still running or did one of the software components crash?
  - Is the network particularly congested?
  - Is the CPU particularly loaded?
  - Is there a disk problem?
  - Was a software library containing a bug installed somewhere?

Unified Logging

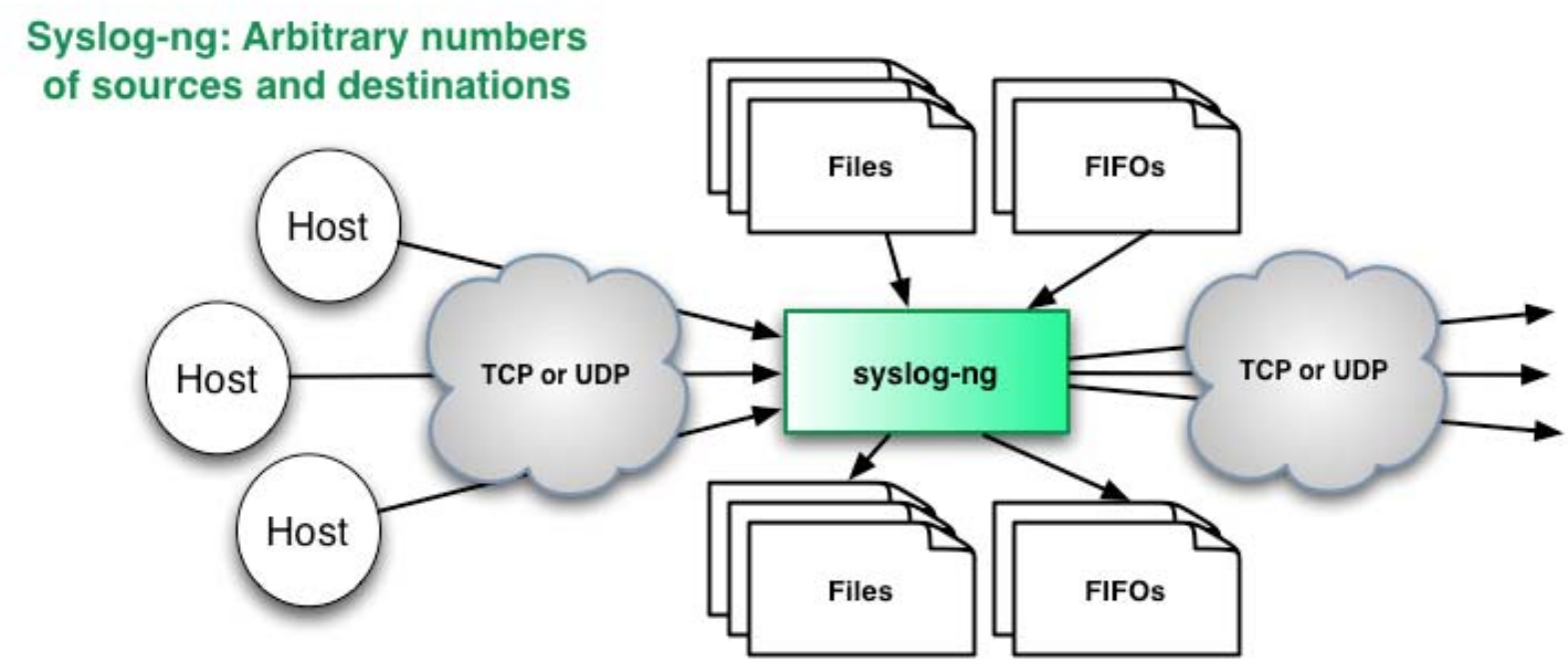
- "Standard" log format
  - CEDPS has defined a "Best Practice for Logging" document
  - Name-value pairs
  - Begins and ends for all actions
  - http://www.cedps.net/wiki/index.php/LoggingBestPractice
- Log file collection mechanism

Log Collection

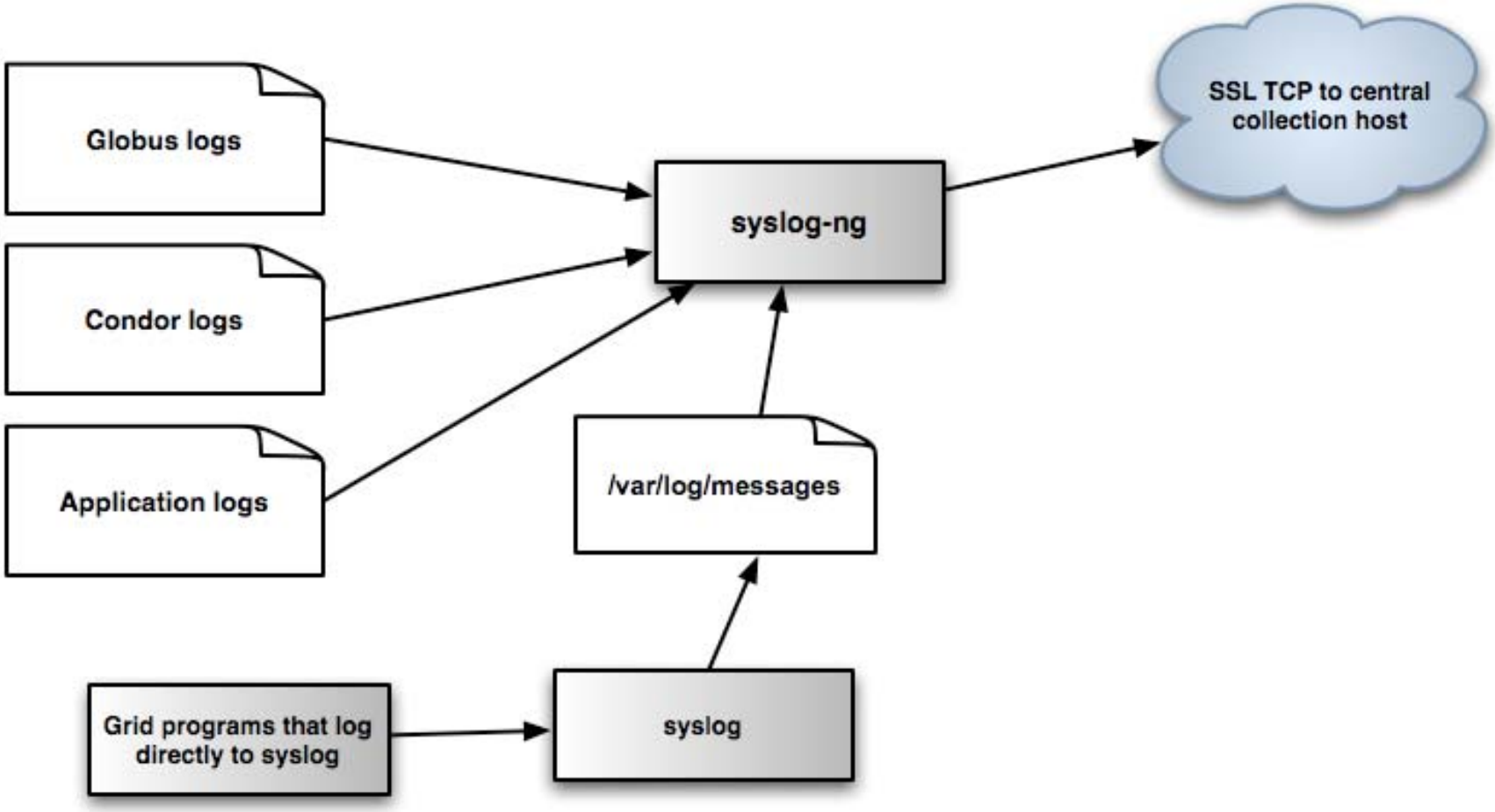
- No need to invent something new for this
  - syslog-ng fills all requirements
    - *Open source, runs on all major OSes*
    - *Fault tolerant, secure (via stunnel), scalable, easy to configure, etc.*
    - *Large user base*

http://www.balabit.com/products/syslog-ng/

Log Collecting With syslog-ng

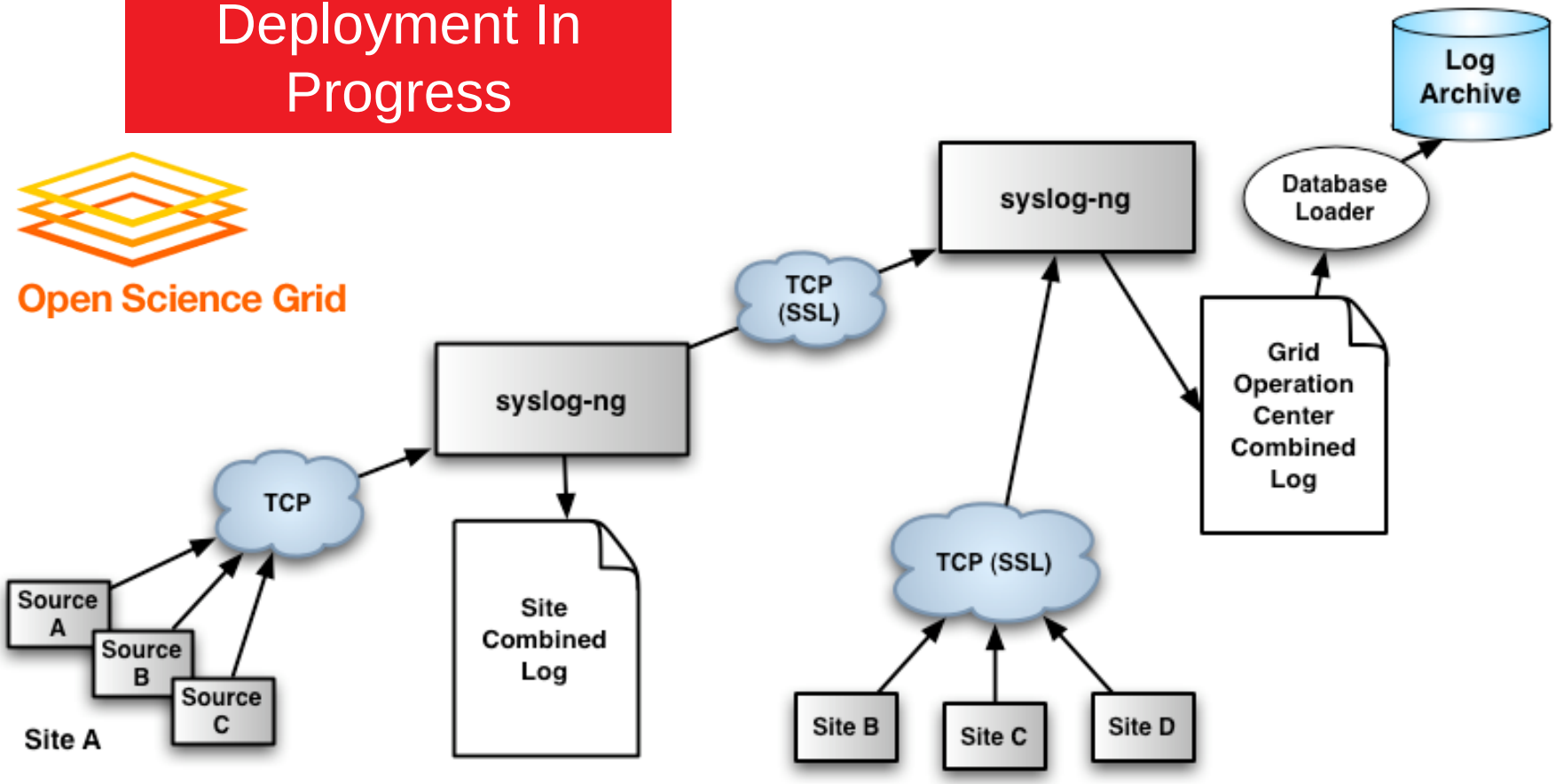


Sample Site Deployment



Syslog-ng Deployment for OSG

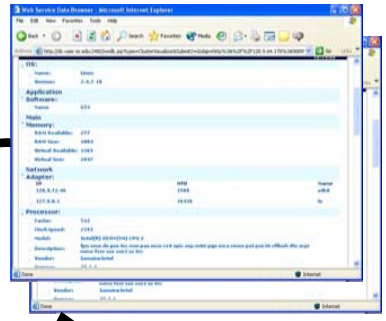
Deployment In Progress



http://www.cedps.net/wiki/index.php/Troubleshooting

Troubleshooting

- Many devices and failure cases
- Find the problems first:
  - Index service and portal for basic data
  - Trigger service for alarms on errors

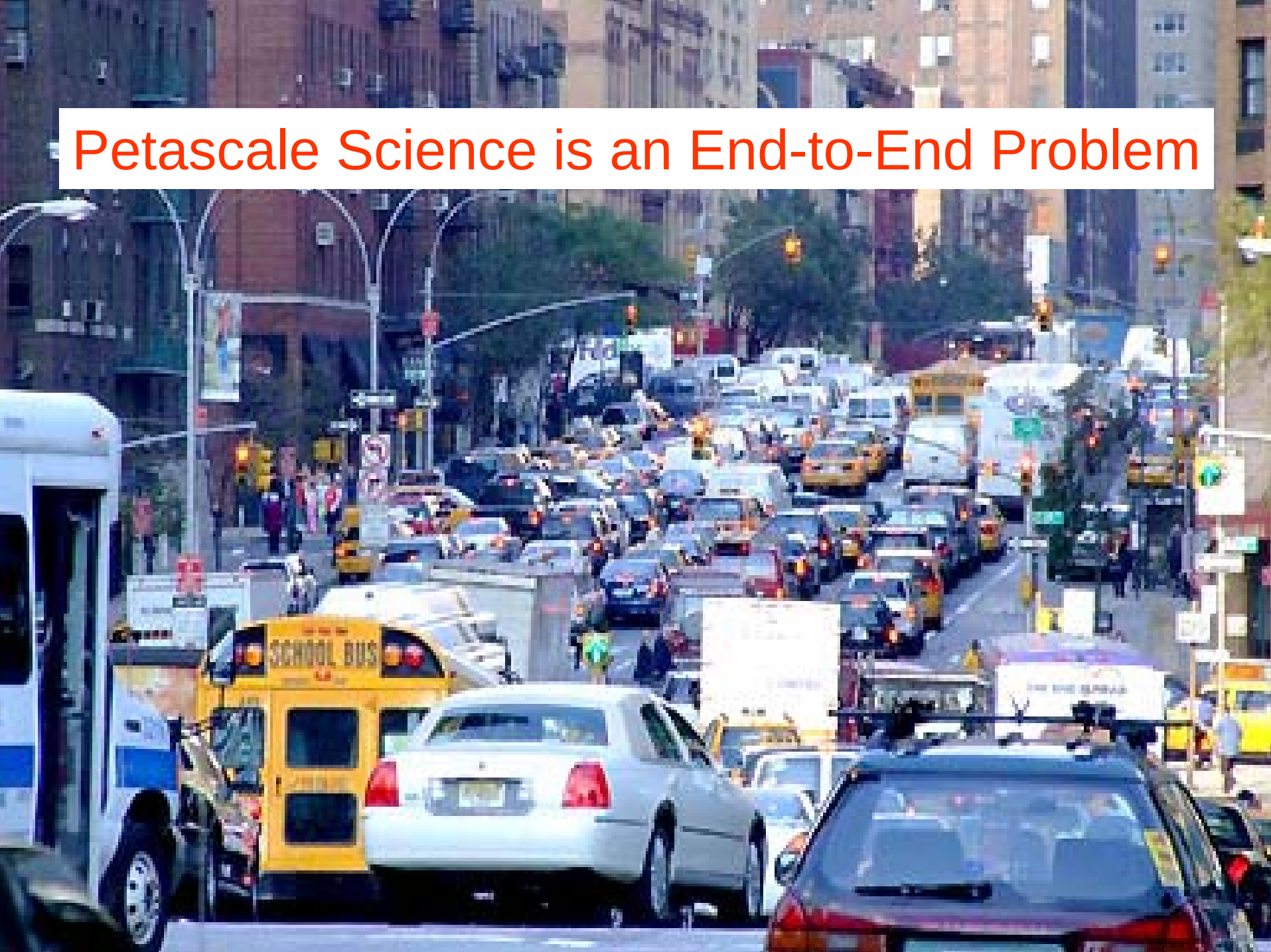


Trigger action

Globus MDS4 Data

In Production Use

Petascale Science is an End-to-End Problem



Let Us Turbocharge your Science …

- Tools exist today, e.g.:
  - GridFTP for data transfer
  - DRS for data replication
  - Common logging for easier troubleshooting
  - OSG central log collection and reliability
- Yet better tools are on the way:
  - MOPS for storage and bandwidth management
  - RAVi and Workspace services
  - End-to-end troubleshooting

Please tell us about your end-to-end needs!

Jennifer Schopf jms@mcs.anl.gov

http://www.cedps.net