



Argonne
NATIONAL
LABORATORY

... for a brighter future



U.S. Department
of Energy

UChicago ►
Argonne_{LLC}

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

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

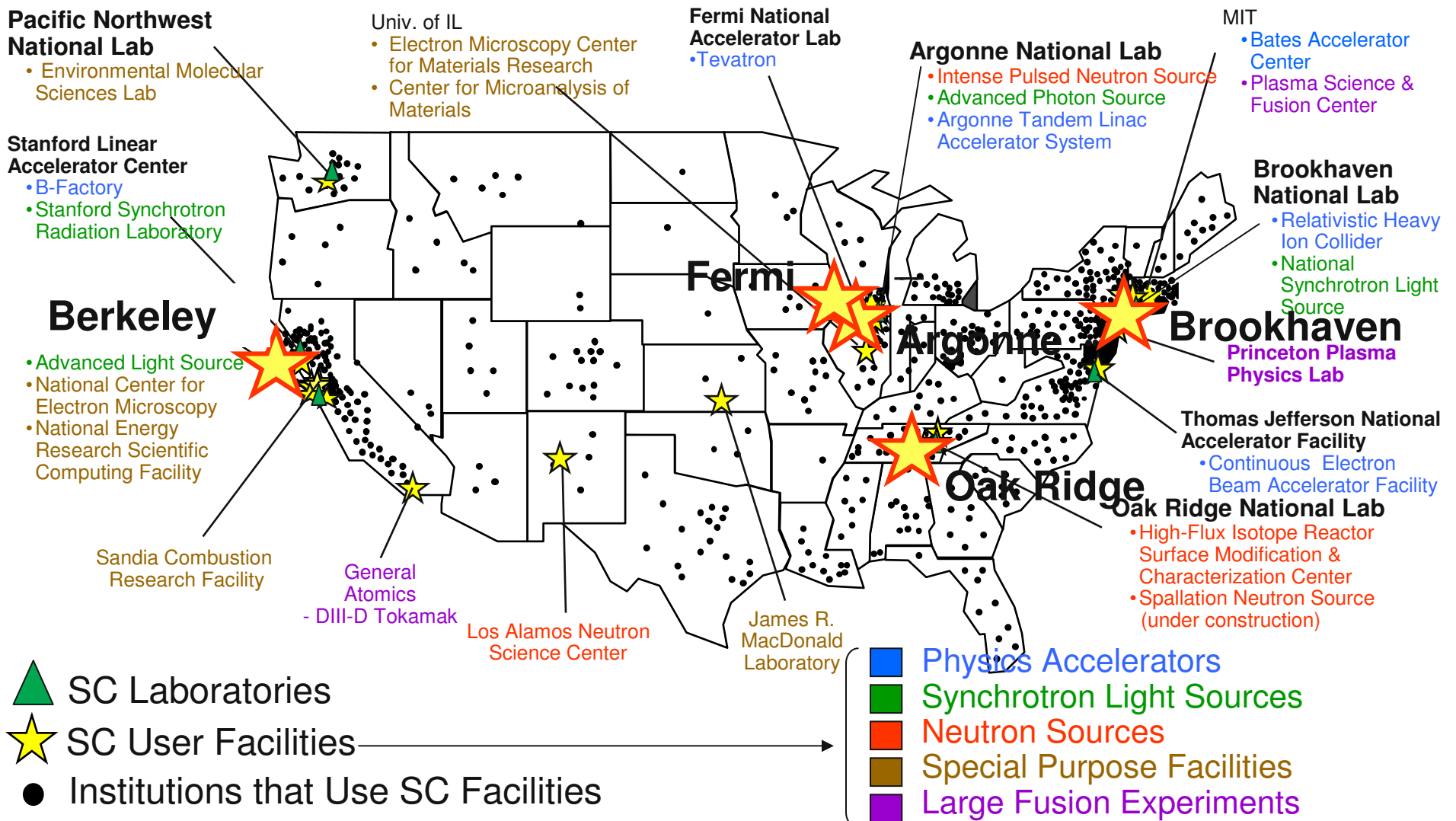
A photograph of a heavily congested city street, likely in New York City, showing a dense traffic jam. The street is filled with cars, including a yellow school bus, a white sedan, and a dark sedan in the foreground. In the background, there are tall brick buildings and streetlights. A white speech bubble with a black border is overlaid on the right side of the image, containing the text: "Yes, but it would still take us two hours to get downtown!!!".

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

Science is an End-to-End Problem

Geneva 

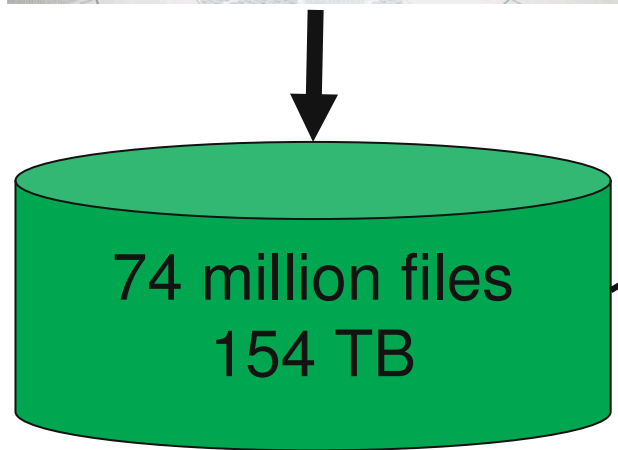
Cadarache 



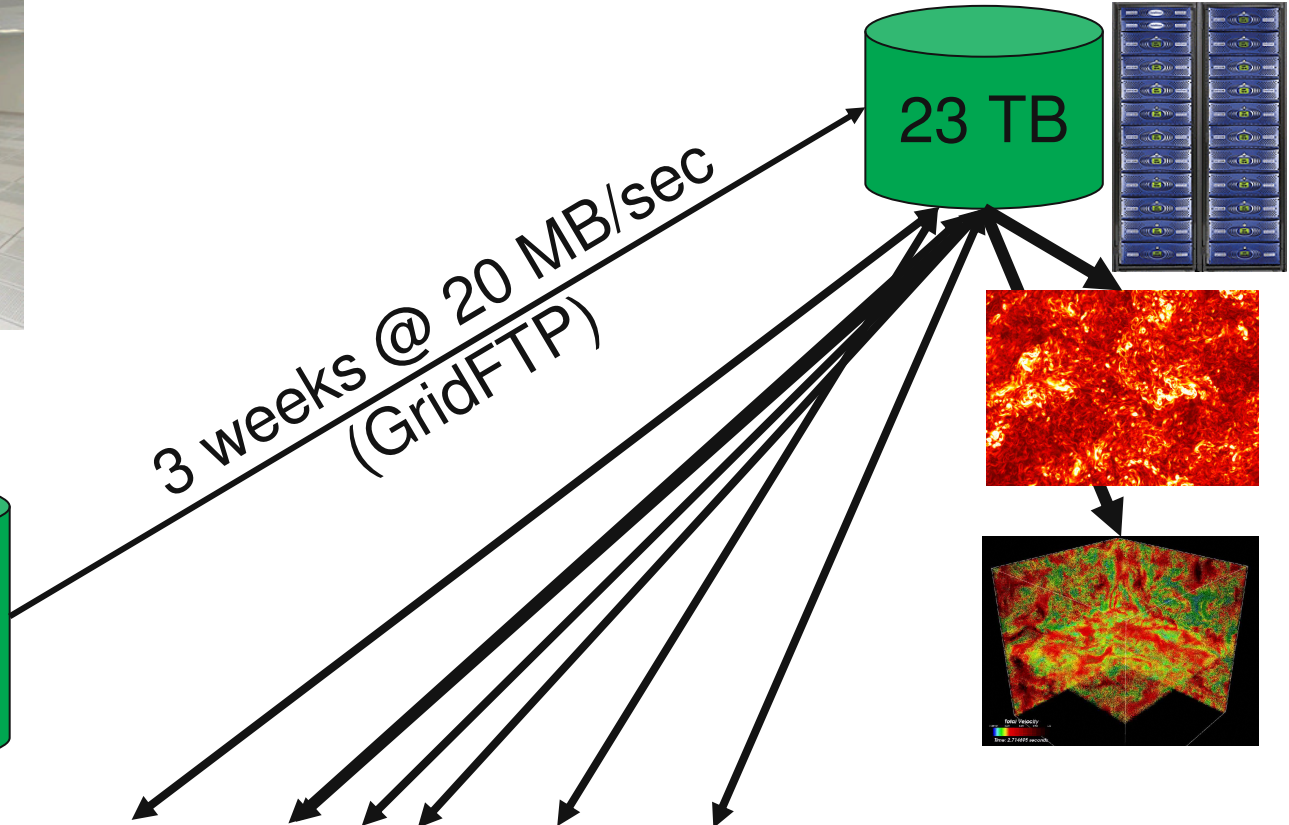
An Example: FLASH Turbulence Simulation (Robert Fisher, Don Lamb, et al.)



THE UNIVERSITY OF CHICAGO



Largest compressible homogeneous isotropic turbulence simulation



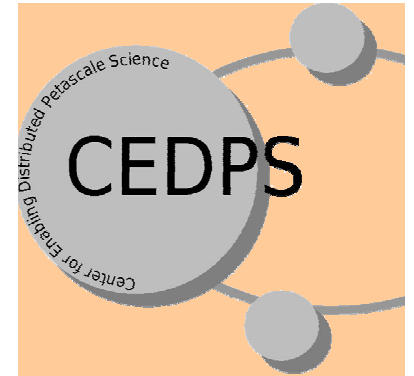
External users access turbulence dataset

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



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

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

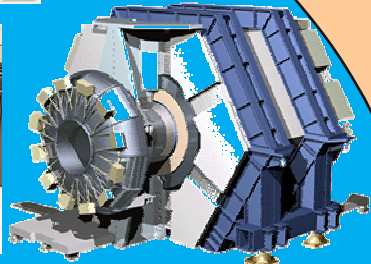
■ **Fast:** >10,000x faster than usual Internet

■ **Reliable:** recover from many failures

■ **Predictable:** data arrives when scheduled

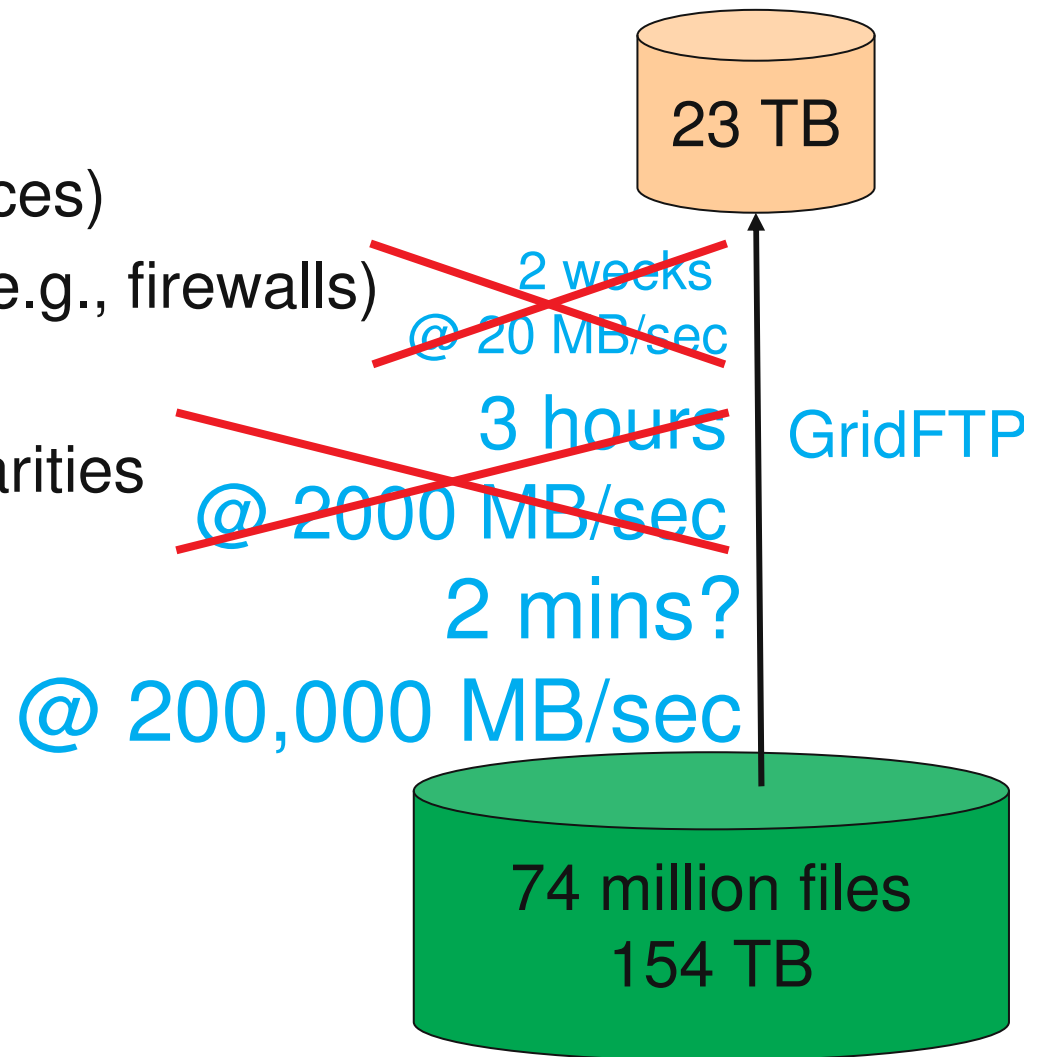
■ **Secure:** protect expensive resources & data

■ **Scalable:** deal with many users & much data



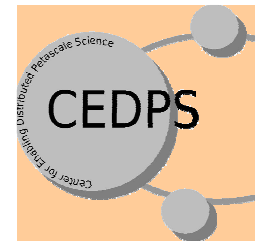
Data Delivery Challenges

- 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

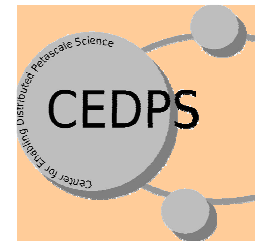
- Data complexity
- Parallelism (in diverse places)
- Network heterogeneities (e.g., firewalls)
- Space (or the lack of it)
- Protocols and their peculiarities
- Failures at many levels
- Deadlines
- Resource contention
- Multiple participants



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

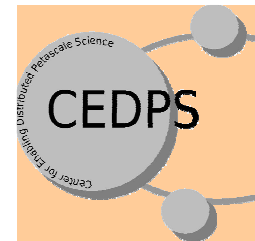


Data Delivery Challenges

- 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

GridFTP

MOPS



Data Delivery Challenges

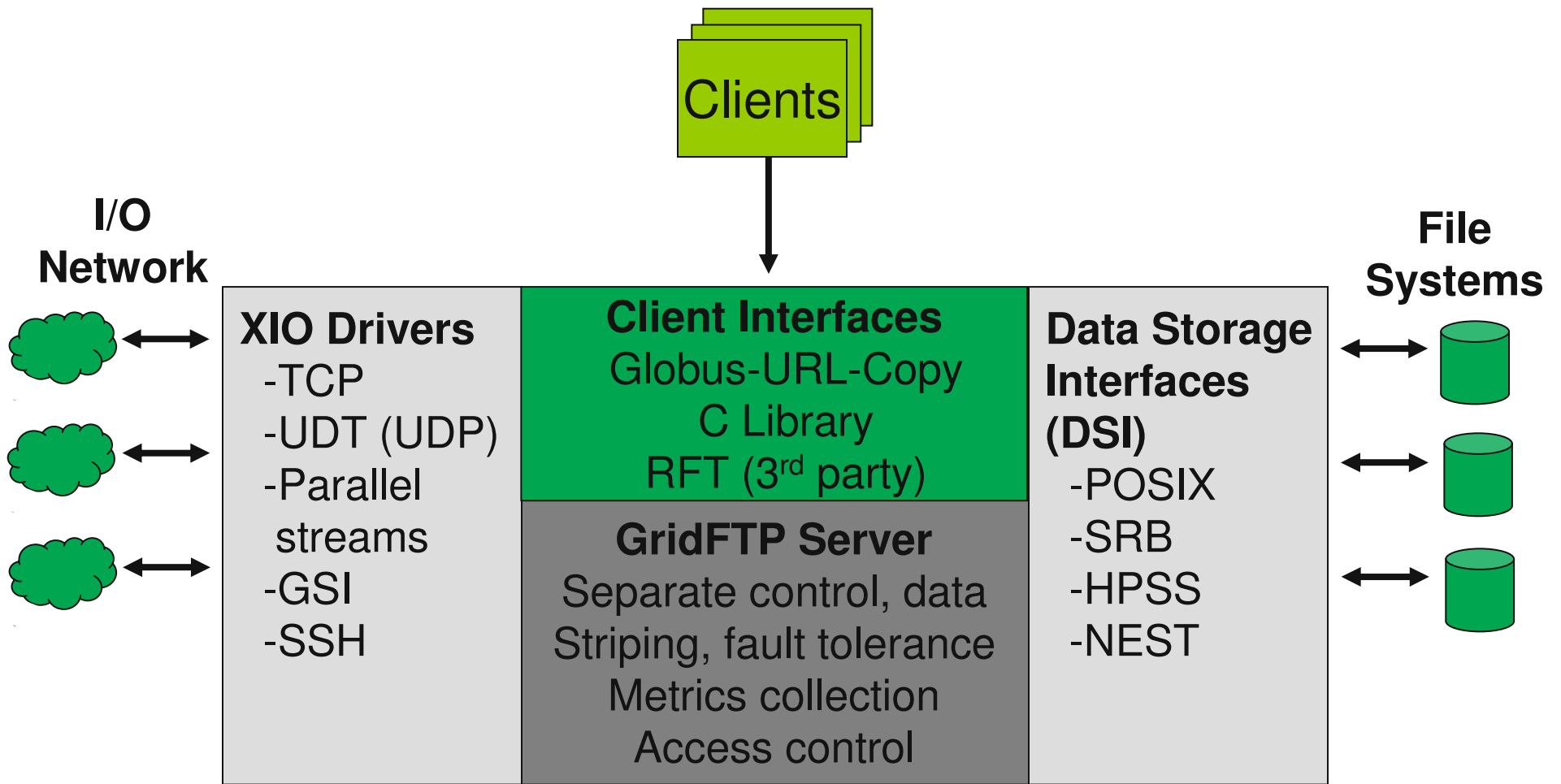
- Data complexity
- **Parallelism (in diverse places)**
- Network heterogeneities (e.g., firewalls)
- **Space (or the lack of it)**
- **Protocols and their peculiarities**
- **Failures at many levels**
- **Deadlines**
- **Resource contention**
- **Multiple participants**

GridFTP

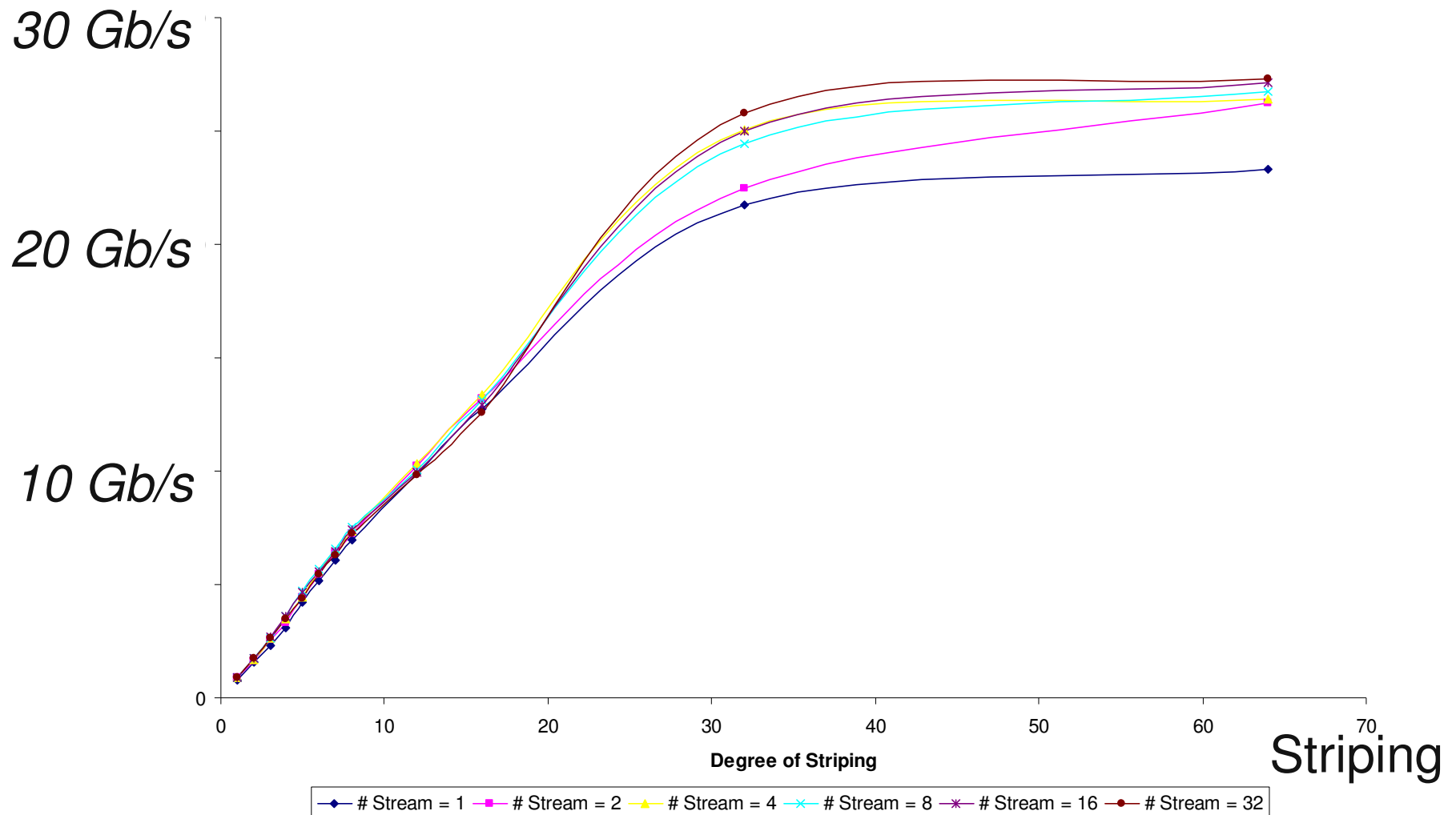
MOPS

DRS

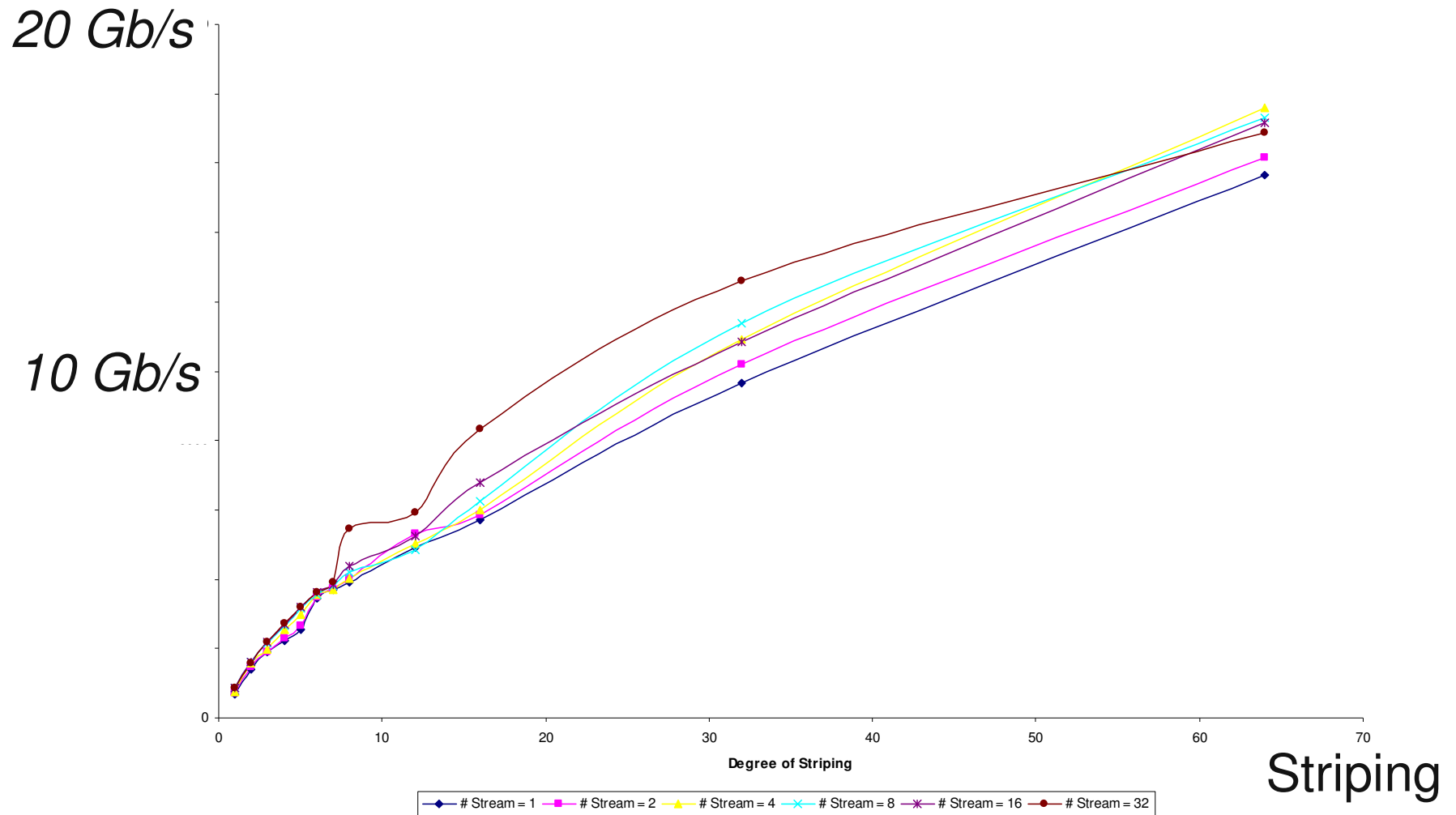
Current State of the Art: GridFTP



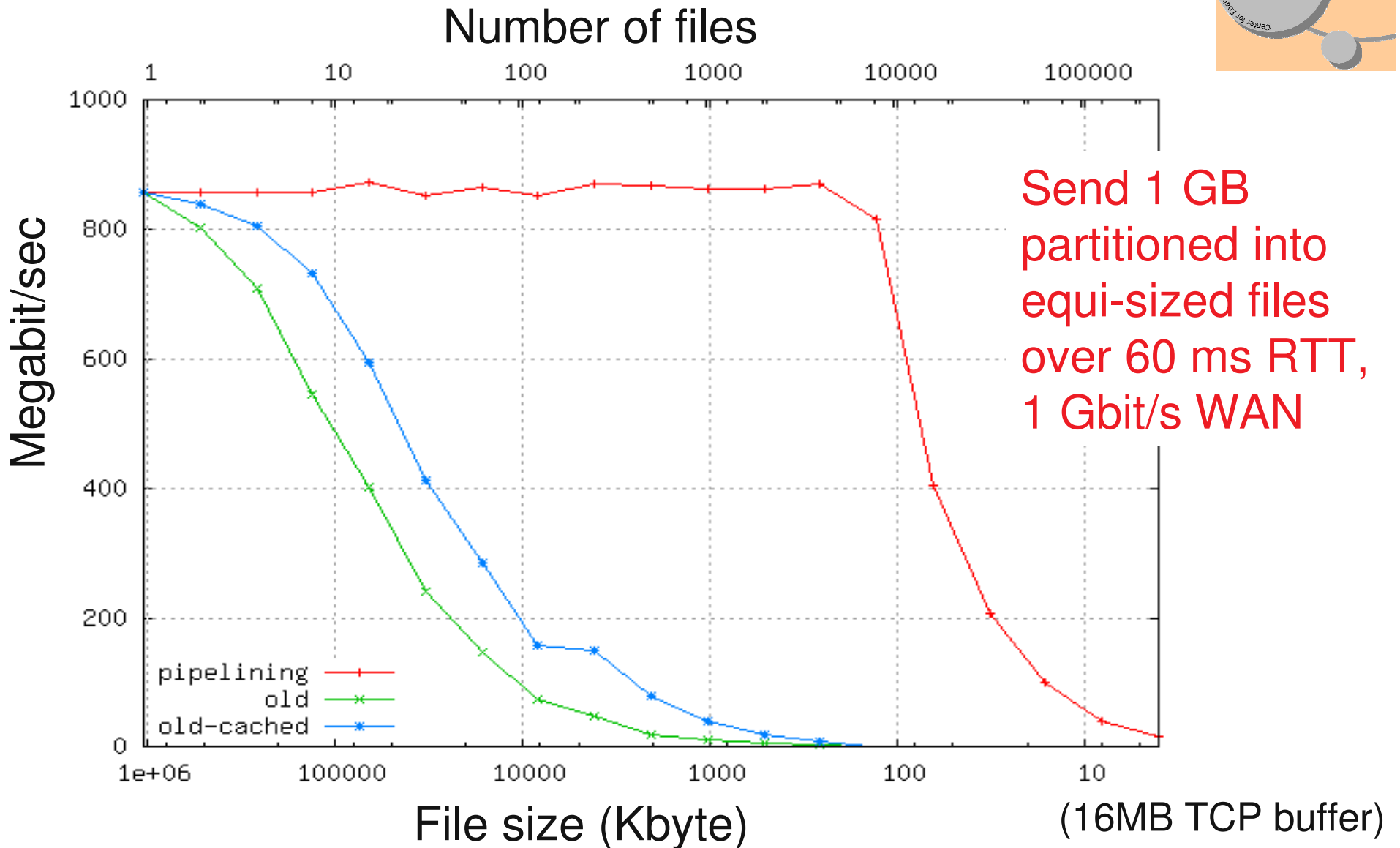
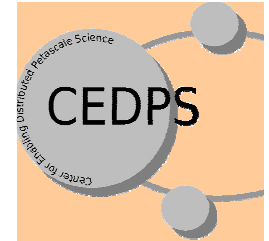
Memory to Memory over 30 Gigabit/s Network (San Diego — Urbana)



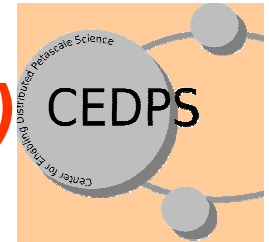
Disk to Disk over 30 Gigabit/s Network (San Diego — Urbana)



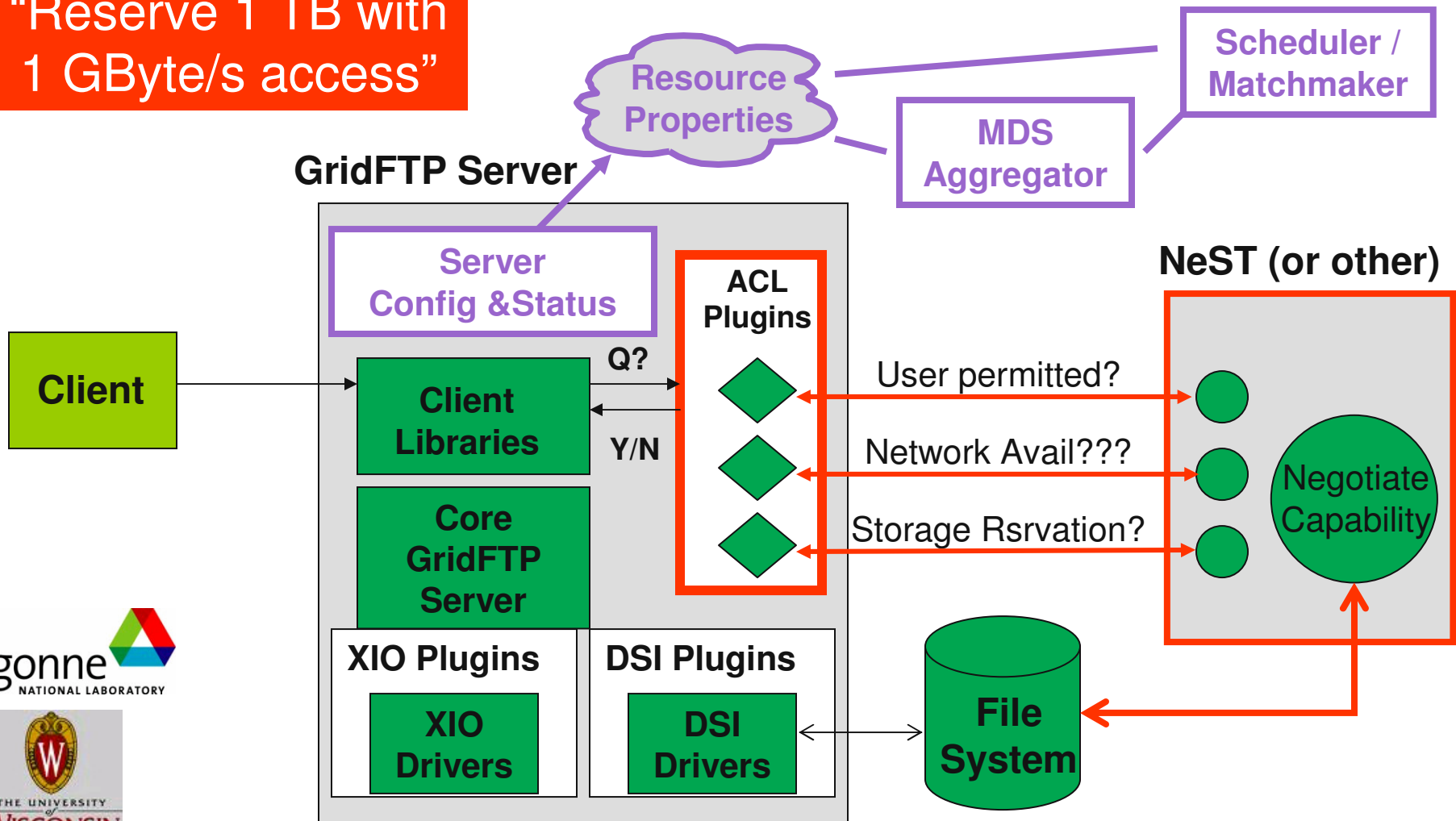
“Lots of Small Files” (LOSF) Optimization



Managed Object Placement Service (MOPS)



“Reserve 1 TB with 1 GByte/s access”





Data Replication Service

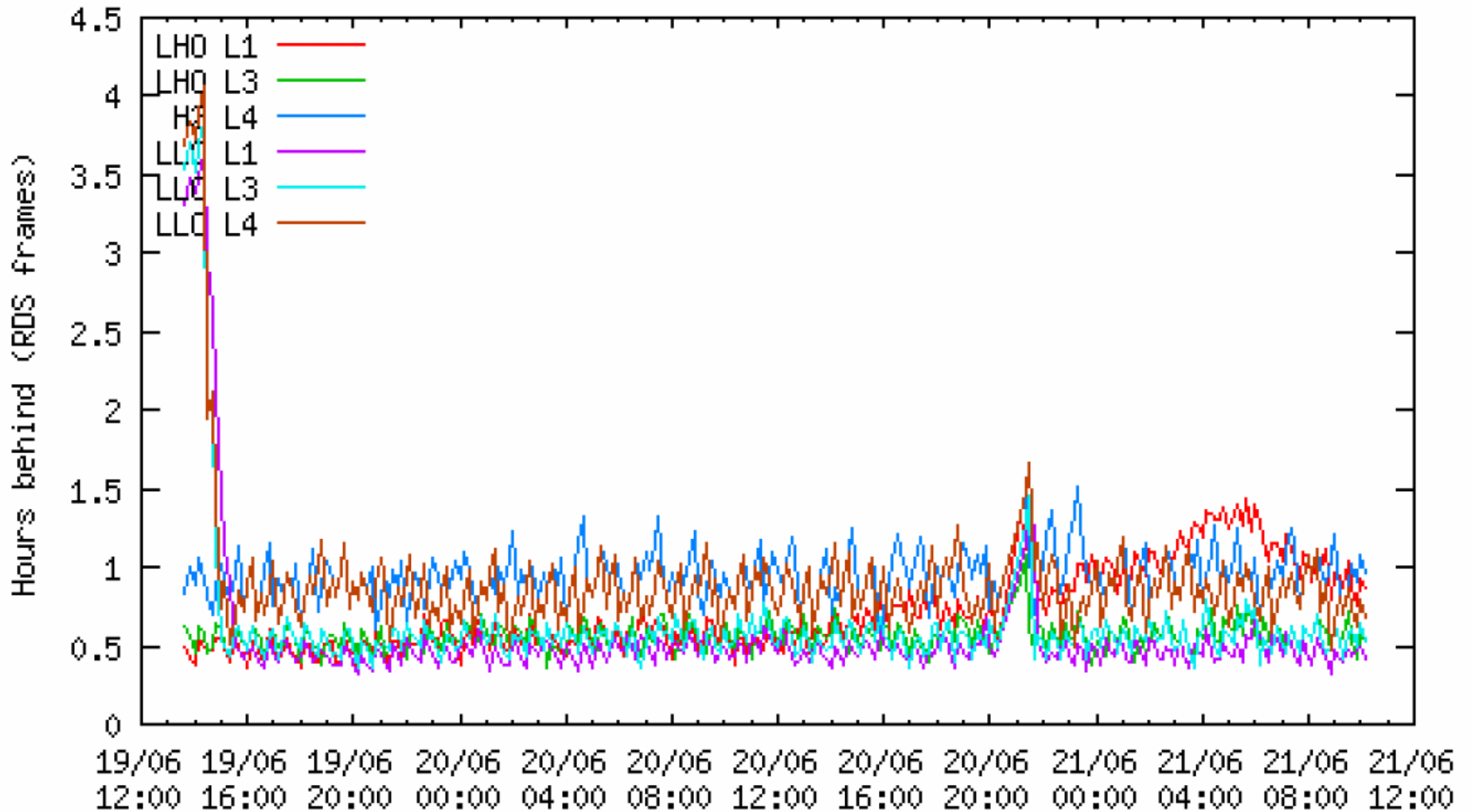


LIGO Gravitational Wave Observatory



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

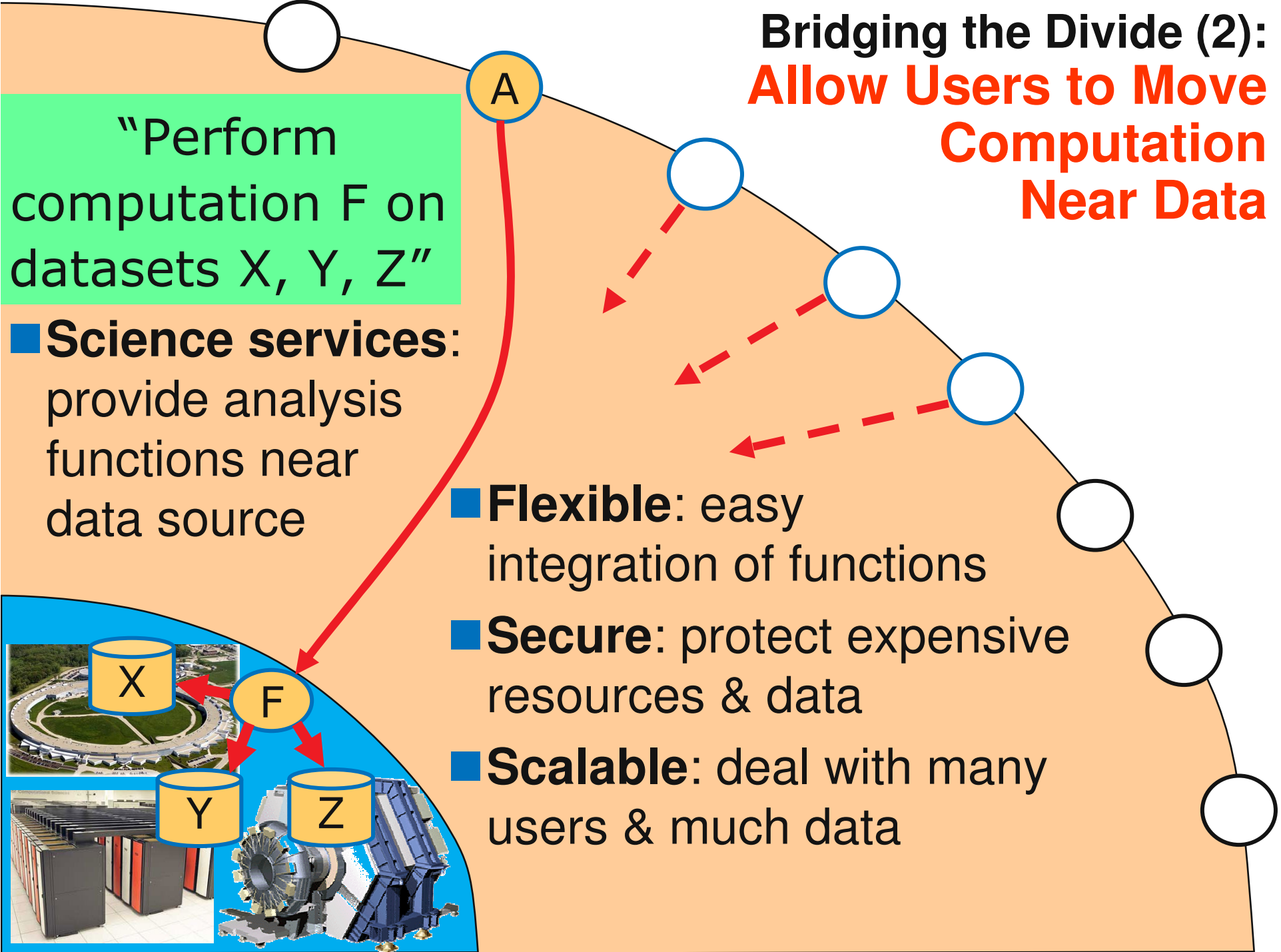
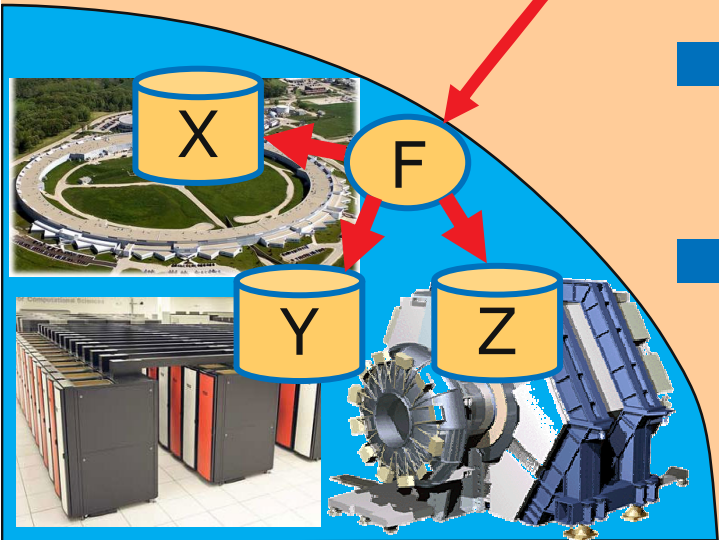


Bridging the Divide (2): Allow Users to Move Computation Near Data

“Perform computation F on datasets X, Y, Z”

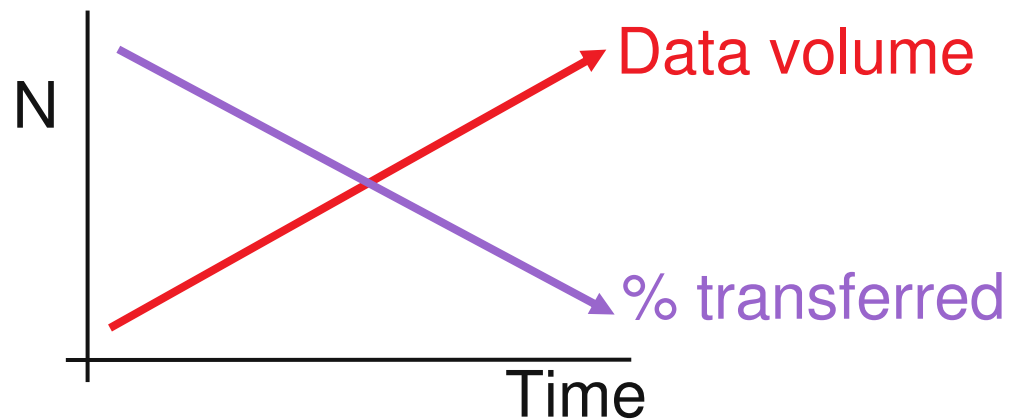
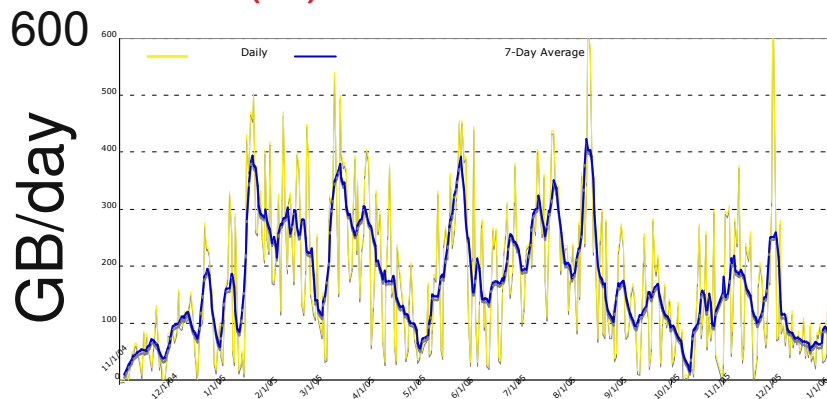
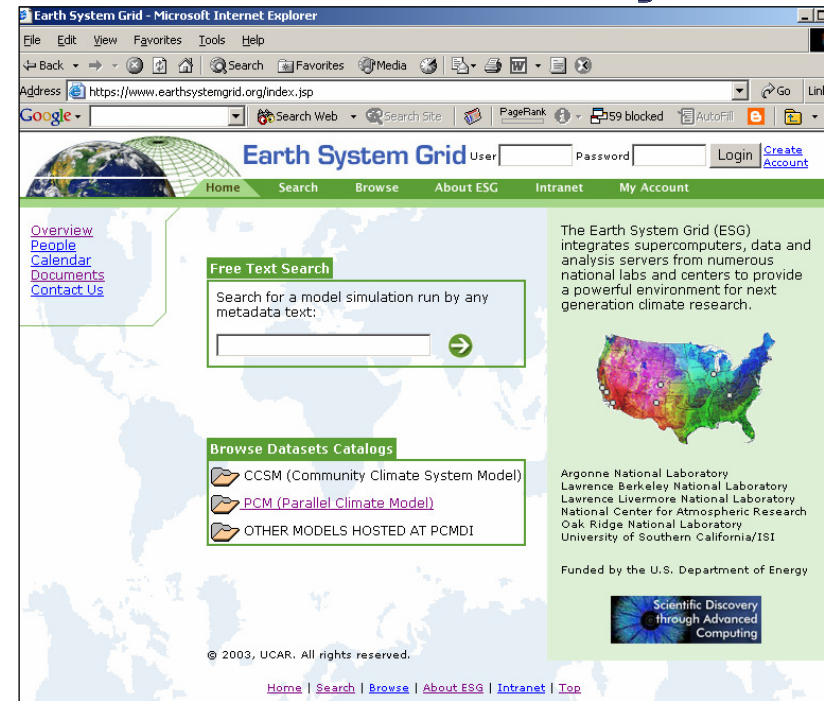
■ **Science services:** provide analysis functions near data source

- **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

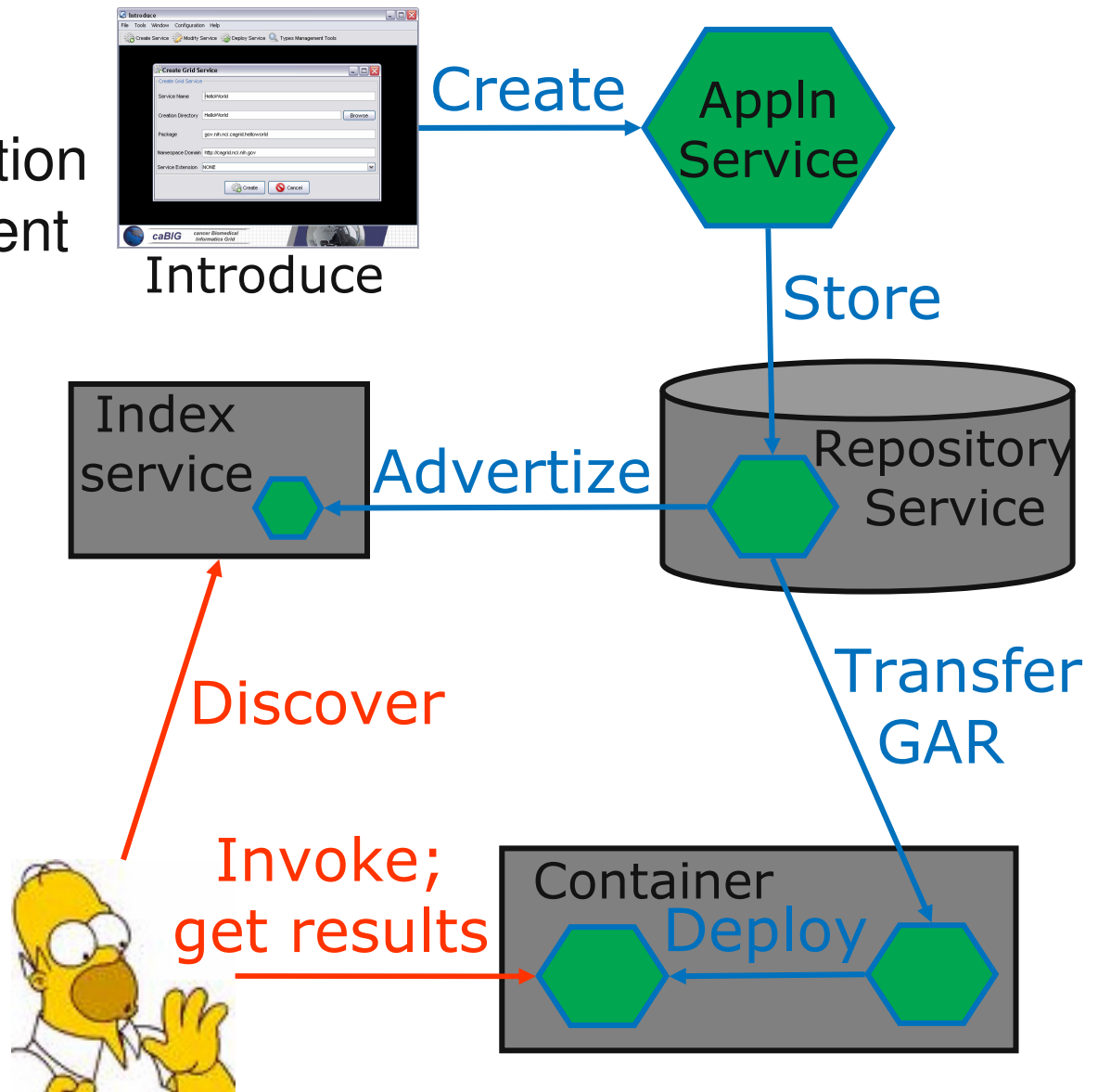
- Service authoring
 - Easy creation of “services” encapsulating data and/or computation
- Provisioning
 - Allocate resources to services and to other computations as demand changes
- Code portability and security
 - Encapsulation and portability of application code

Automated Service Creation Tools

- RAVE: Remote Application Virtualization Environment (Ravi Madduri et al.)

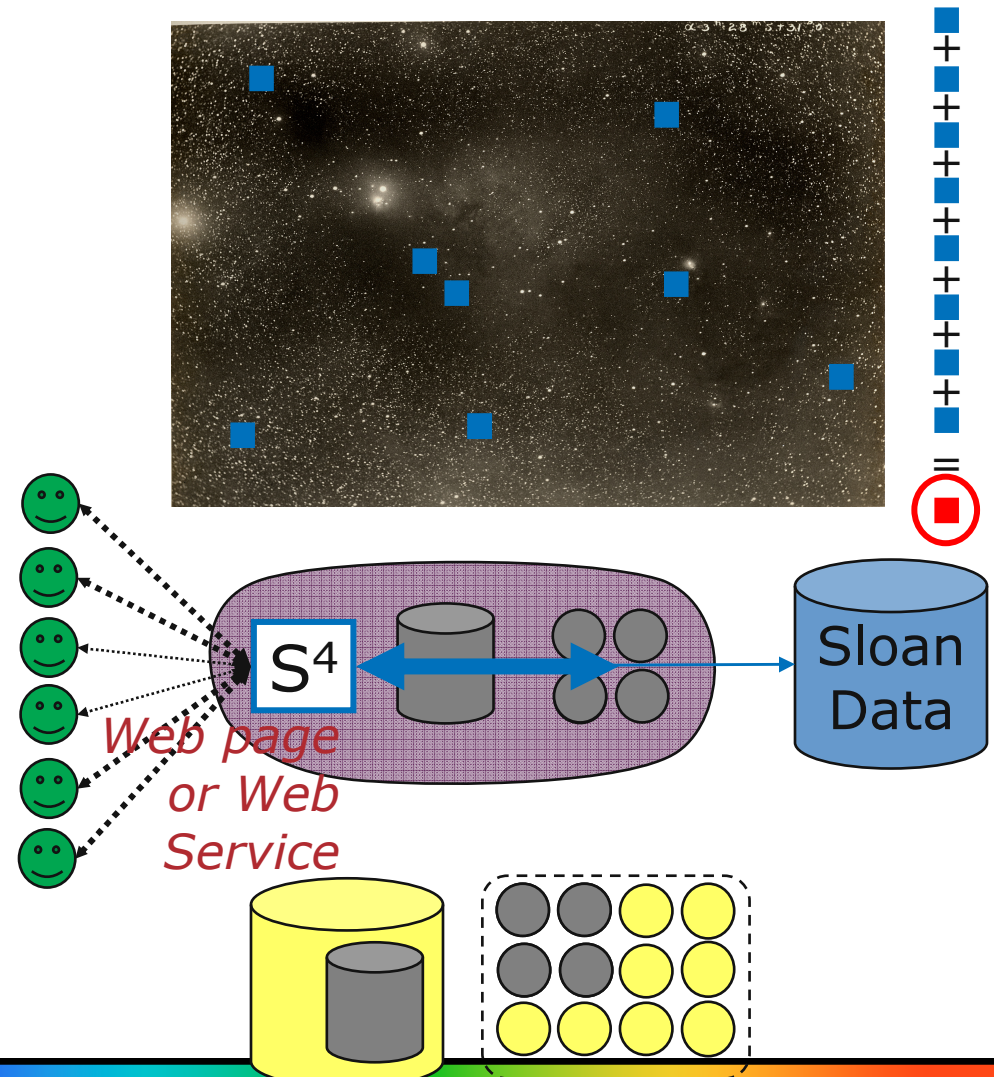
- Builds on Introduce
- Define service
- Create skeleton
- Discover types
- Add operations
- Configure security
- Wrap arbitrary executables

- pyGlobus tools (Keith Jackson et al., LBNL)

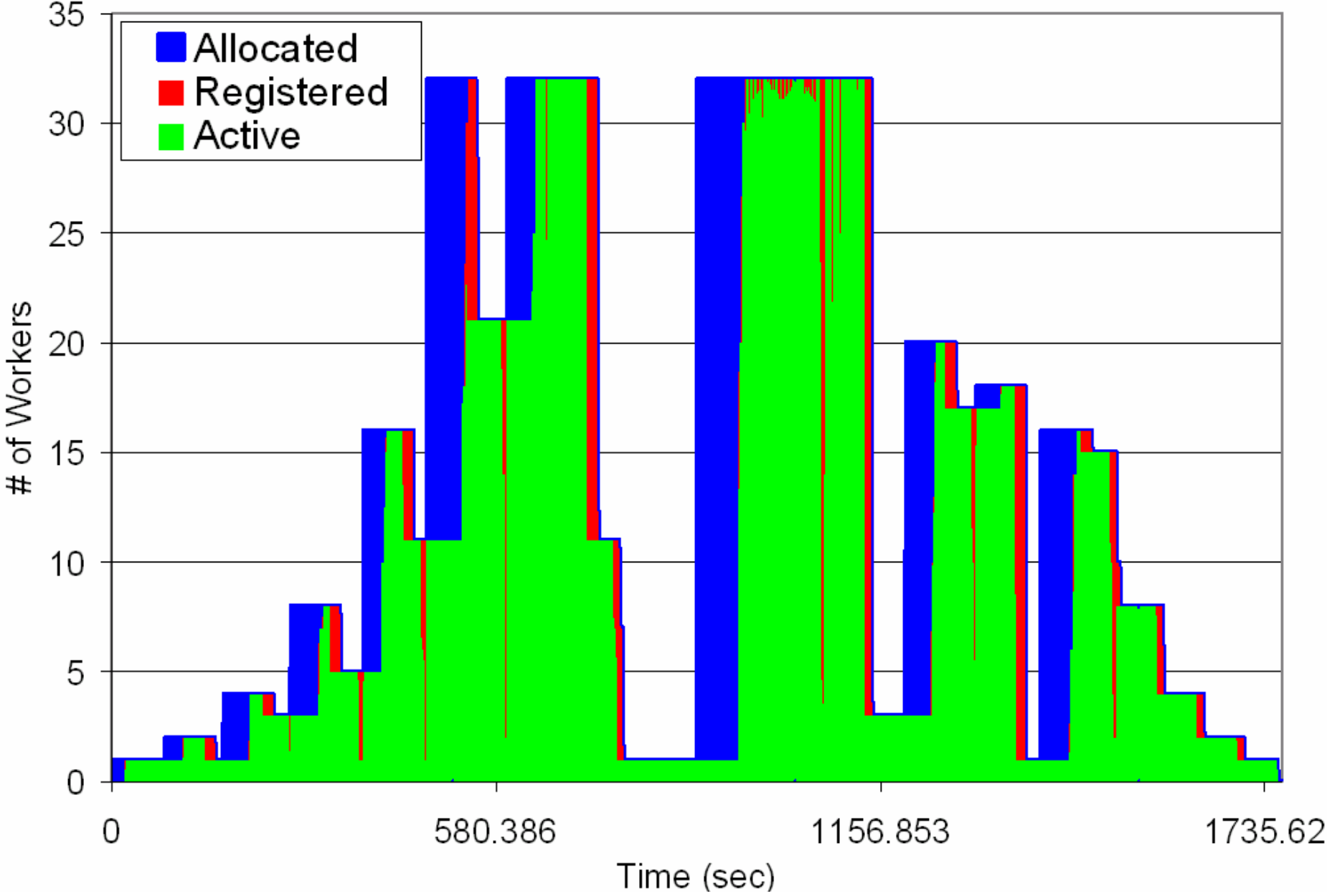


Provisioning: Stacking Service

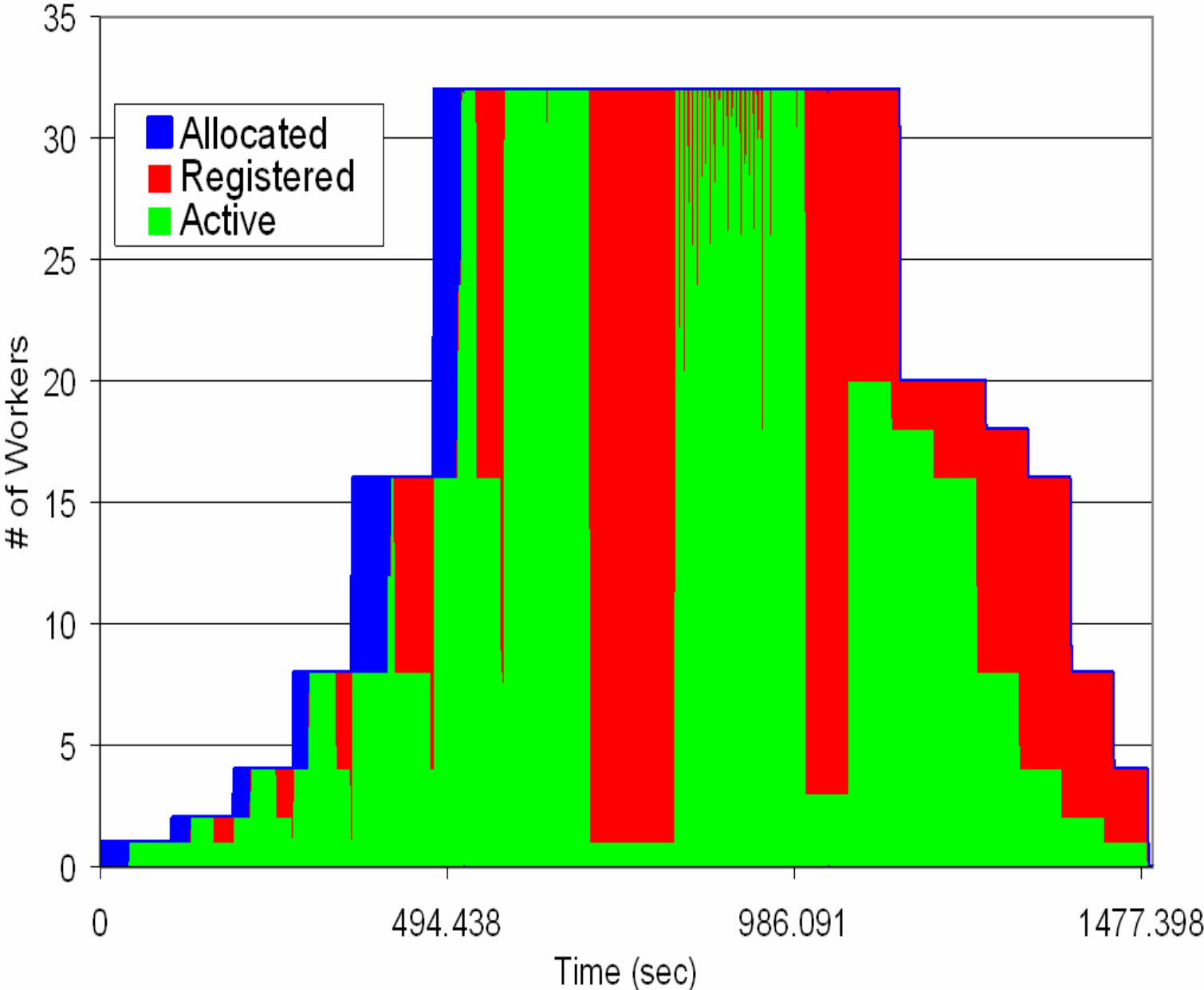
- Purpose
 - On-demand “stacks” of random locations within ~10TB dataset
- Challenge
 - Rapid access to 10-10K “random” files
 - Time-varying load
- Solution
 - Dynamic acquisition of compute, storage



Release after 15 Seconds Idle

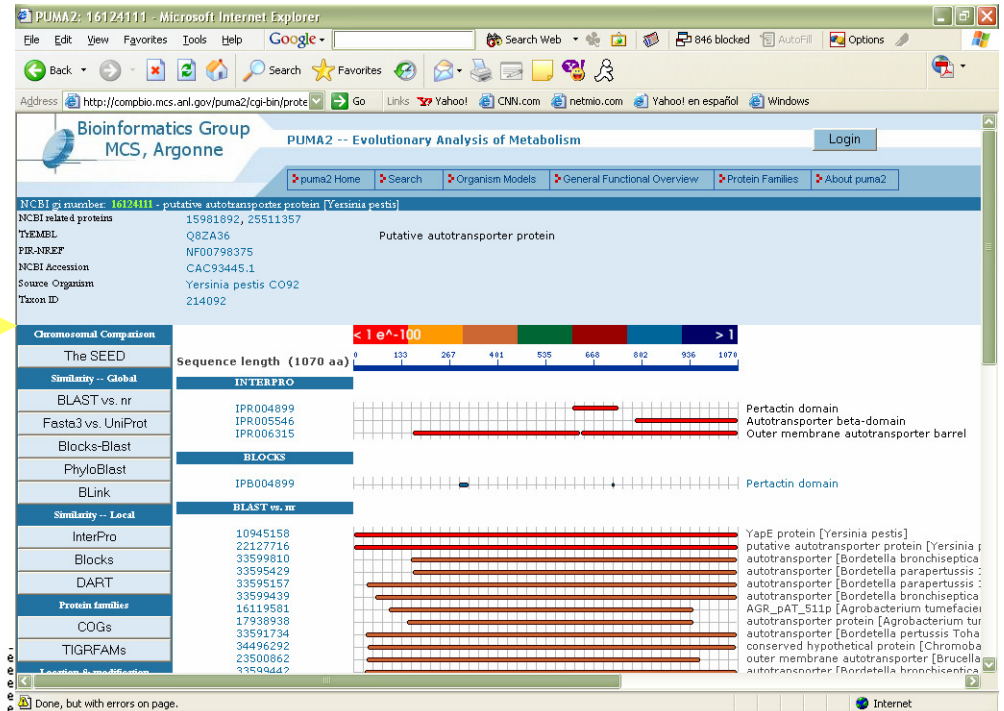


Release after 180 Seconds Idle



On-Demand Access to Computing in Biology

Public PUMA Knowledge Base
 Information about proteins analyzed against ~2 million gene sequences



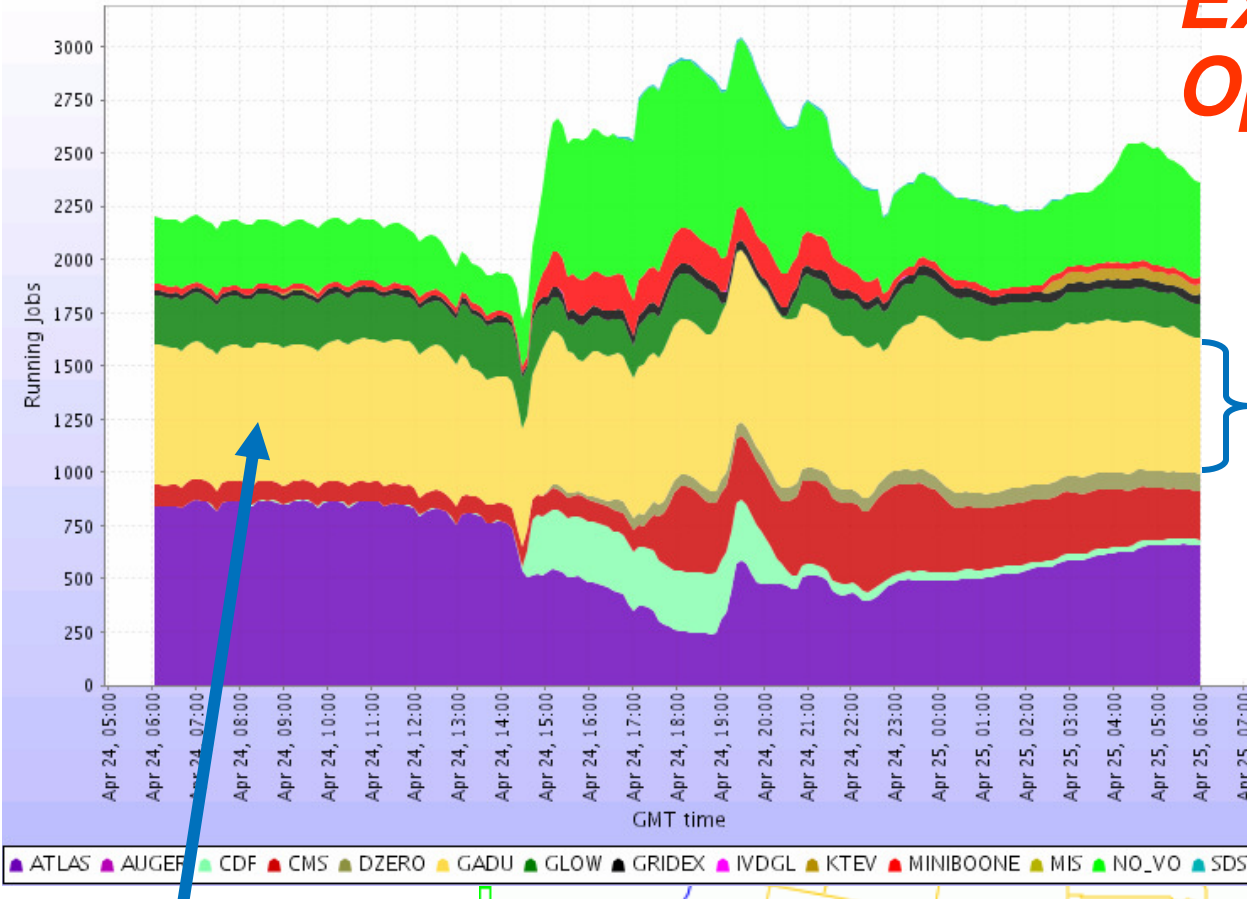
gi 23499780 gn REF_tigr BRA0013 gi 16080253 ref NP_391080.1 44.27 253 131 1 15 257 8 2603.7 e-30 134.4	gi 23499780 gn REF_tigr BRA0013 gi 123098409 ref NP_691875.1 43.48 253 133 2 16 258 5 2573.8 e-30 134.4
gi 23499780 gn REF_tigr BRA0013 gi 48637187 ref ZP_00294182.1 44.92 256 126 2 14 256 7 2591.1 e-30 134.4	gi 23499780 gn REF_tigr BRA0013 gi 52008400 gb IAWN25342.1 44.75 257 126 2 15 258 3 2561.9 e-30 134.4
gi 23499780 gn REF_tigr BRA0013 gi 48664015 ref ZP_00317908.1 44.49 245 134 1 13 257 5 2476.1 e-30 134.4	gi 23499780 gn REF_tigr BRA0013 gi 30348891 gb IAWN28934.1 39.53 253 138 3 18 257 5 2552.0 e-43 177.6
gi 23499780 gn REF_tigr BRA0013 gi 19655222 gb IAF93939.1 40.64 251 138 1 17 256 10 2602.7 e-43 177.6	gi 23499780 gn REF_tigr BRA0013 gi 12758806 gb IAA007757.1 43.03 251 130 4 18 256 11 2602.5 e-41 170.6
gi 23499780 gn REF_tigr BRA0013 gi 112597924 gb IAA185899.2 46.70 162 96 1 62 243 5 1856.8 e-39 162.5	gi 23499780 gn REF_tigr BRA0013 gi 46363318 ref ZP_0026079.1 39.58 240 136 2 14 253 6 2361.8 e-36 154.5
REF_tigr BRA0013 gi 39933731 ref NP_946007.1 34.90 255 134 1 13 257 5 2403.4 e-30 133.7	REF_tigr BRA0013 gi 48782600 ref ZP_00279106.1 35.92 245 134 1 13 256 3 2404.4 e-30 133.3
REF_tigr BRA0013 gi 41407534 ref NP_960370.1 36.09 266 137 1 14 256 6 2485.7 e-30 132.9	REF_tigr BRA0013 gi 48851585 ref ZP_00305793.1 32.39 247 136 1 12 255 5 2545.7 e-30 132.9
REF_tigr BRA0013 gi 15966306 ref NP_386659.1 36.50 263 134 1 12 256 3 2469.8 e-30 132.1	REF_tigr BRA0013 gi 17548526 ref NP_521866.1 36.36 264 134 1 12 256 3 2439.8 e-30 132.1
gi 23499780 gn REF_tigr BRA0013 gi 51891730 ref WP_074421.1 38.87 247 136 7 18 256 1 2403.4 e-30 133.7	gi 23499780 gn REF_tigr BRA0013 gi 1145881 gb IAA23739.1 33.87 246 147 3 13 253 3 2404.4 e-30 133.3
gi 23499780 gn REF_tigr BRA0013 gi 25029334 ref NP_739388.1 35.20 250 147 4 15 256 6 2485.7 e-30 132.9	gi 23499780 gn REF_tigr BRA0013 gi 21220953 ref NP_636732.1 36.52 257 138 6 12 255 5 2545.7 e-30 132.9
gi 23499780 gn REF_tigr BRA0013 gi 46314029 ref ZP_00214635.1 33.86 254 153 2 12 258 3 2465.7 e-30 132.9	gi 23499780 gn REF_tigr BRA0013 gi 41406852 ref NP_959683.1 35.61 238 149 2 16 253 2 2309.8 e-30 132.1
gi 23499780 gn REF_tigr BRA0013 gi 115644471 ref NP_229523.1 35.69 255 144 5 12 256 2 2469.8 e-30 132.1	gi 23499780 gn REF_tigr BRA0013 gi 123470090 ref ZP_00125423.1 35.20 250 145 4 12 253 3 2439.8 e-30 132.1
gi 23499780 gn REF_tigr BRA0013 gi 24935279 gb IAA64237.1 34.63 257 146 4 12 257 4 2499.8 e-30 132.1	gi 23499780 gn REF_tigr BRA0013 gi 48647655 ref ZP_00301815.1 36.05 258 145 9 12 257 4 2531.3 e-29 131.7
gi 23499780 gn REF_tigr BRA0013 gi 28851510 gb IAA054587.1 36.40 250 142 4 12 253 3 2431.3 e-29 131.7	gi 23499780 gn REF_tigr BRA0013 gi 127378783 ref NP_770312.1 36.25 251 143 3 14 255 7 2491.3 e-29 131.7
gi 23499780 gn REF_tigr BRA0013 gi 11708836 sp I50198 [LINC_PSEPA 34.23 260 143 4 12 257 4 2491.7 e-29 131.3	gi 23499780 gn REF_tigr BRA0013 gi 33594148 ref NP_381792.1 34.17 240 148 5 18 256 6 2363.7 e-29 130.2
gi 23499780 gn REF_tigr BRA0013 gi 33595116 ref NP_381750.1 34.17 240 148 5 18 256 6 2363.7 e-29 130.2	gi 23499780 gn REF_tigr BRA0013 gi 33283206 ref NP_333237.1 34.23 260 143 4 12 257 4 2491.7 e-29 131.3

Back Office Analysis on Grid
 Millions of BLAST, BLOCKS, etc., on OSG and TeraGrid

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

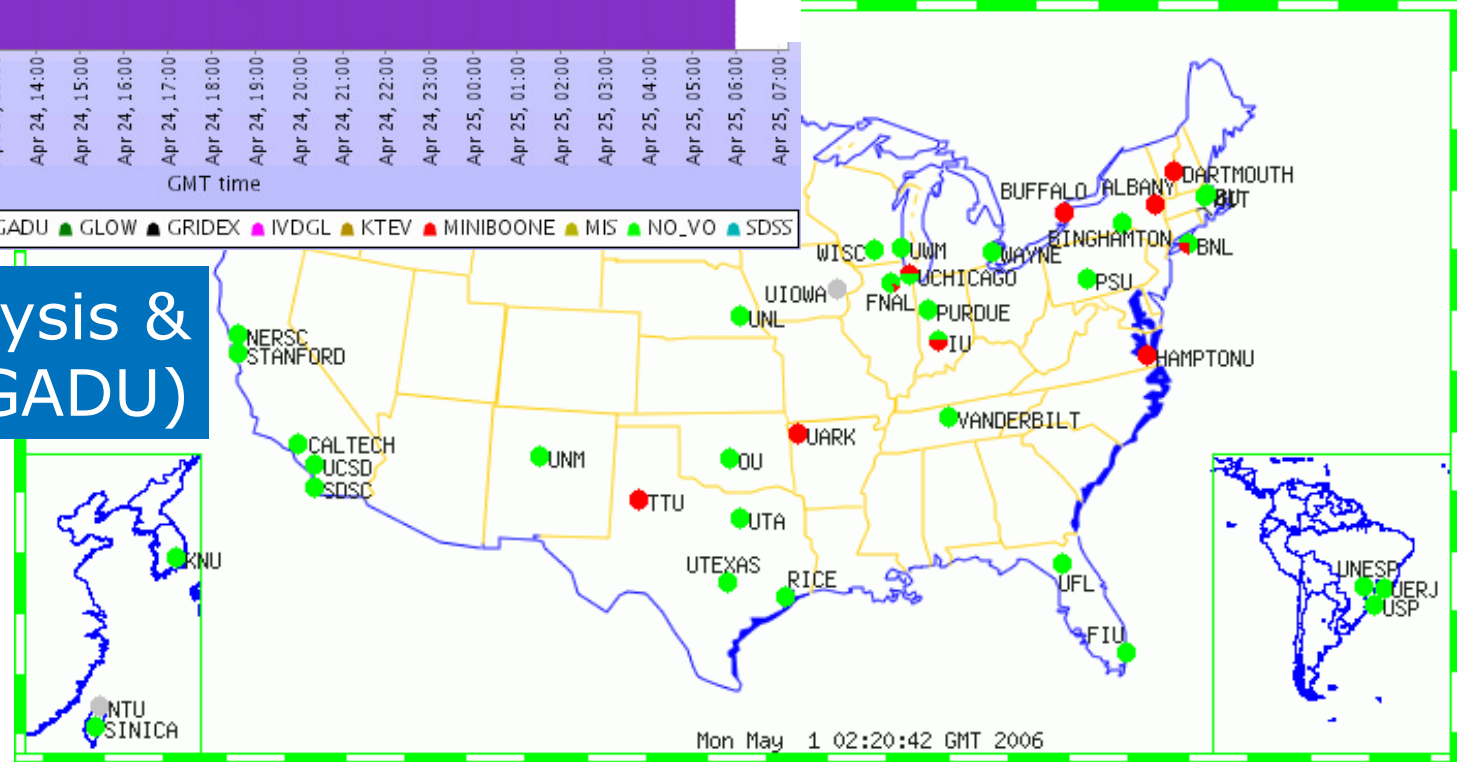
Execution on Open Science Grid

Running Jobs



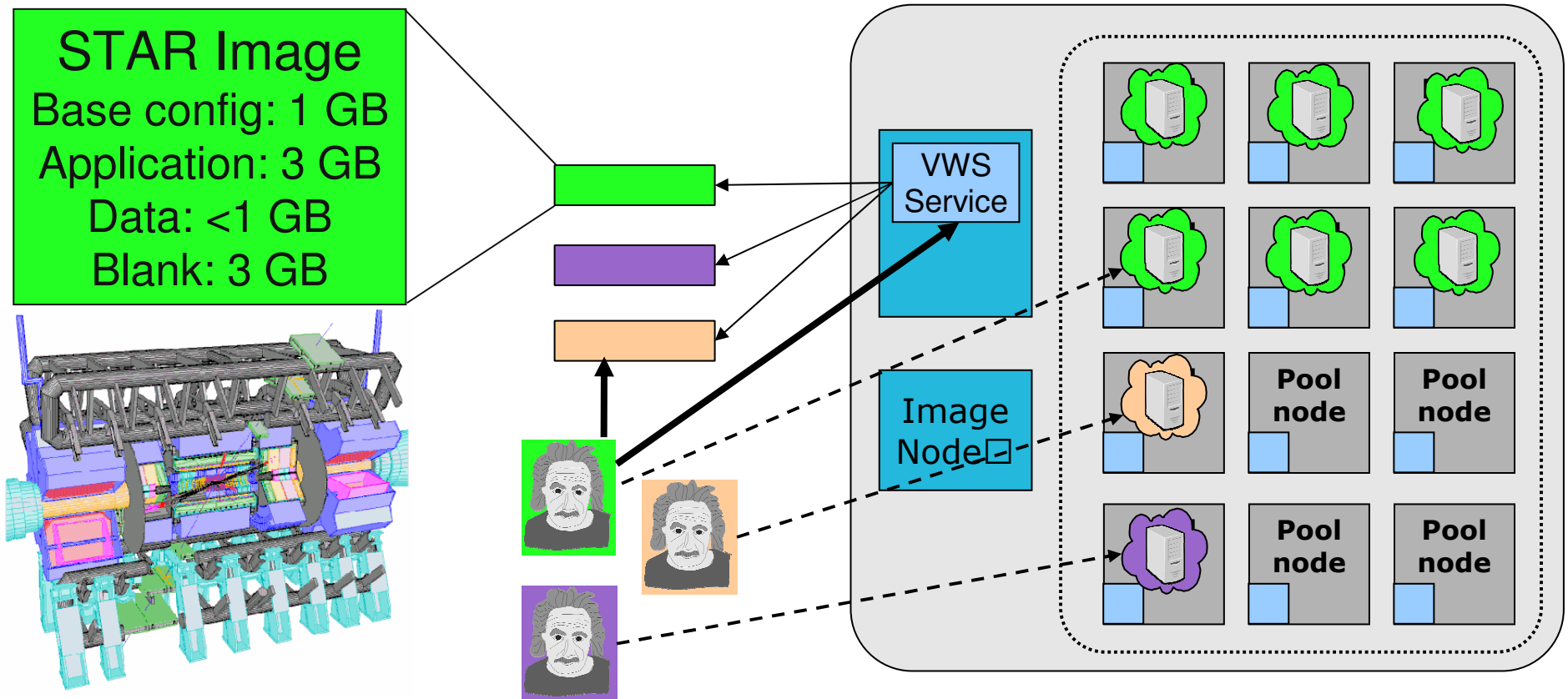
600-1000+ CPUs

Genome Analysis & DB Update (GADU)



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

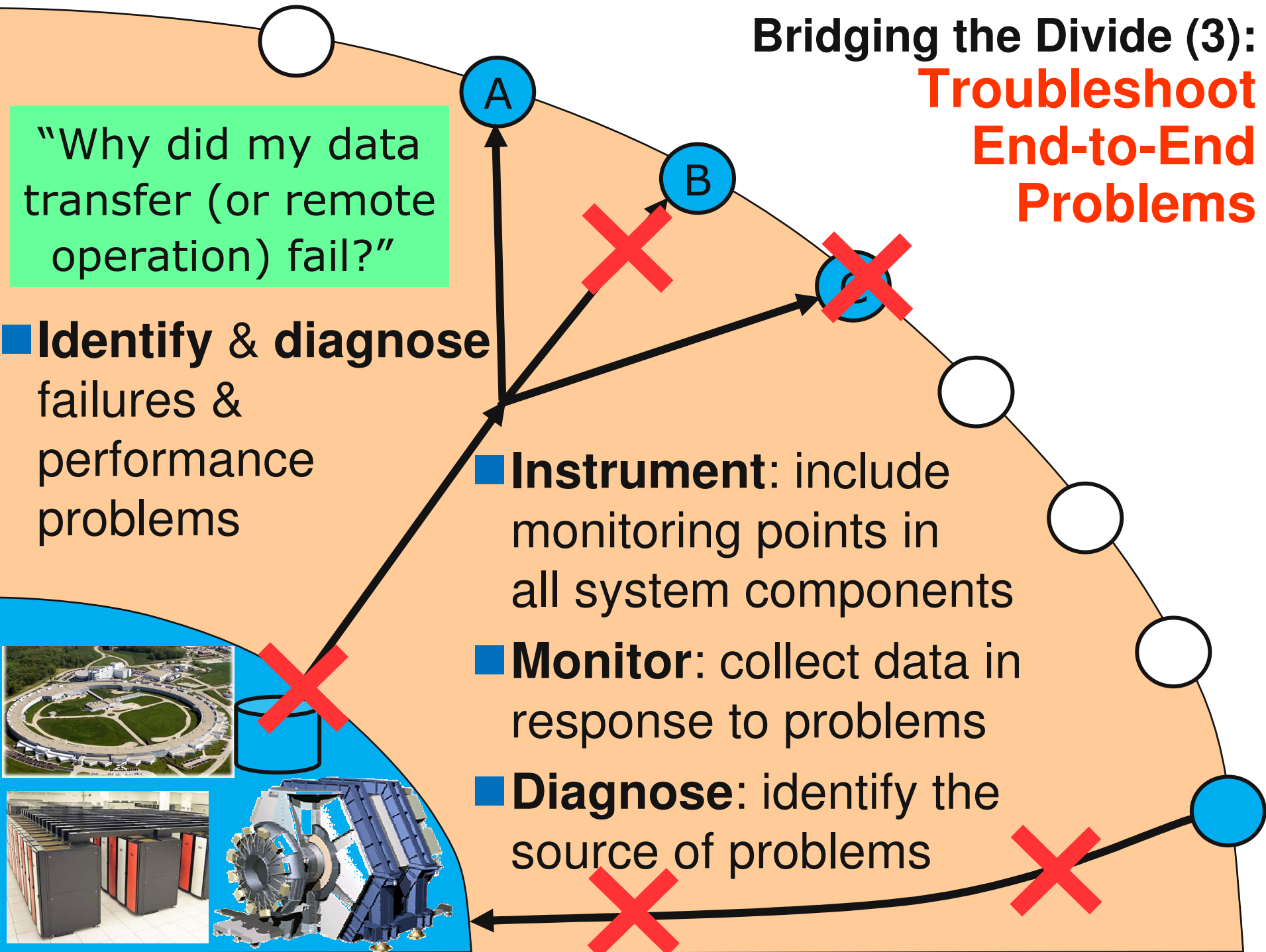
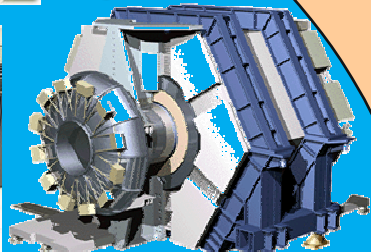
“Why did my data transfer (or remote operation) fail?”

■ **Identify & diagnose** failures & performance problems

■ **Instrument:** include monitoring points in all system components

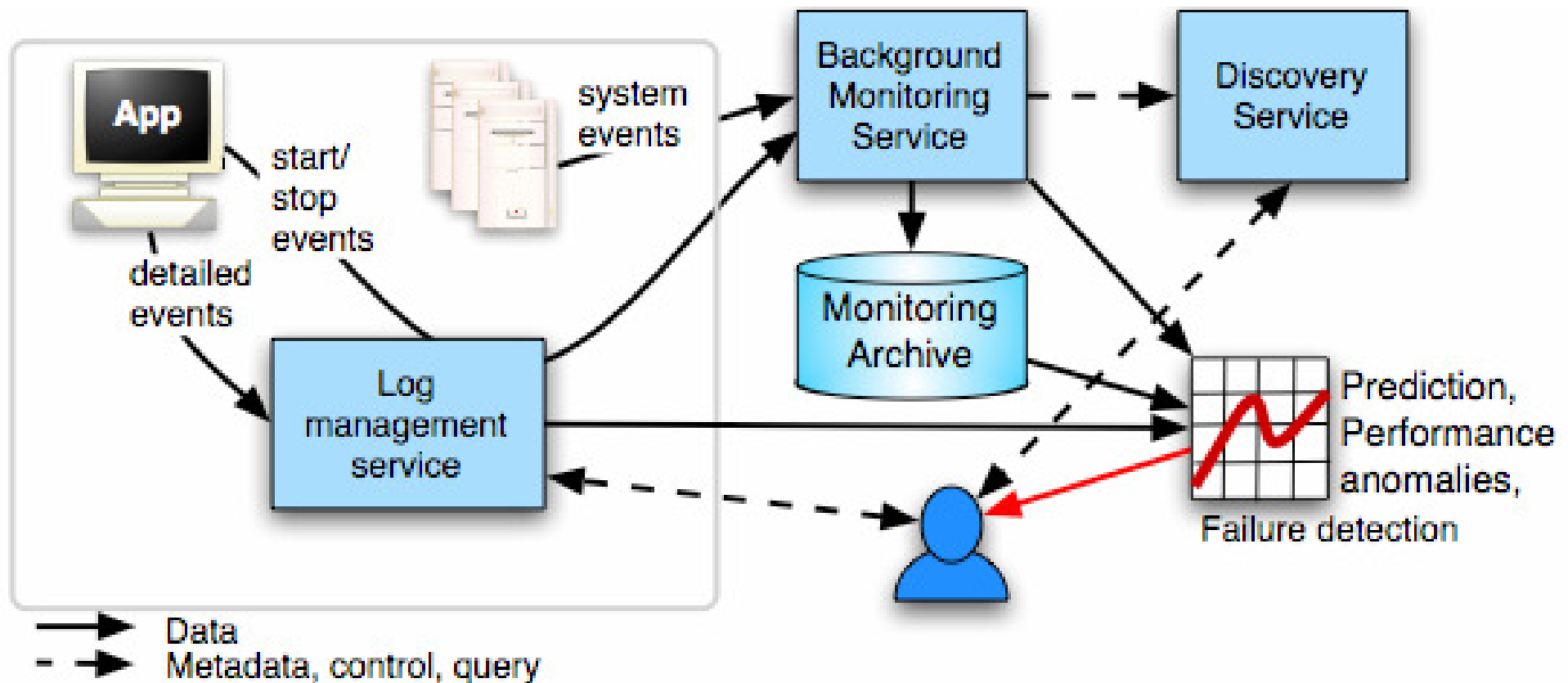
■ **Monitor:** collect data in response to problems

■ **Diagnose:** identify the source of problems

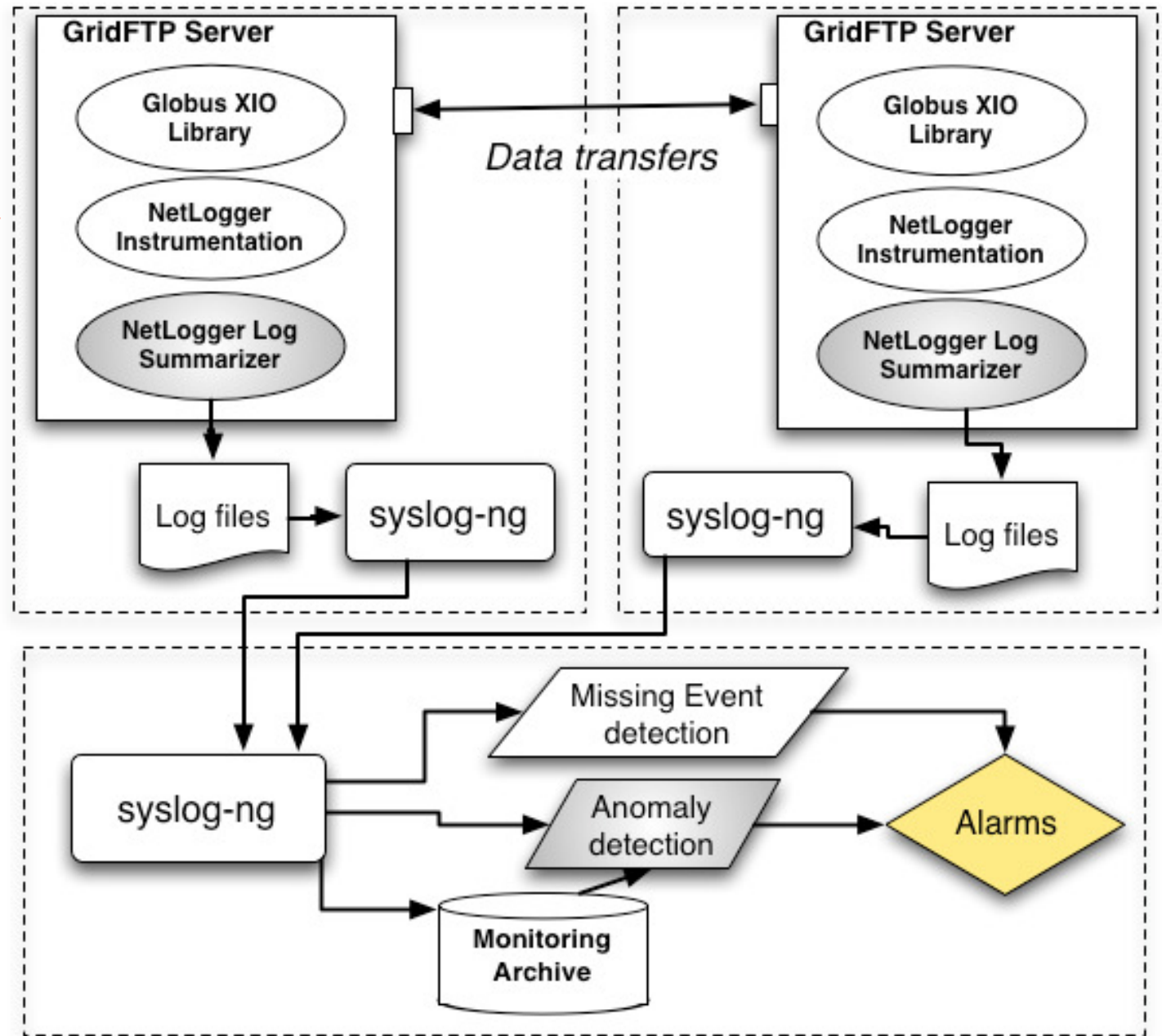


Troubleshooting Challenges and Approach

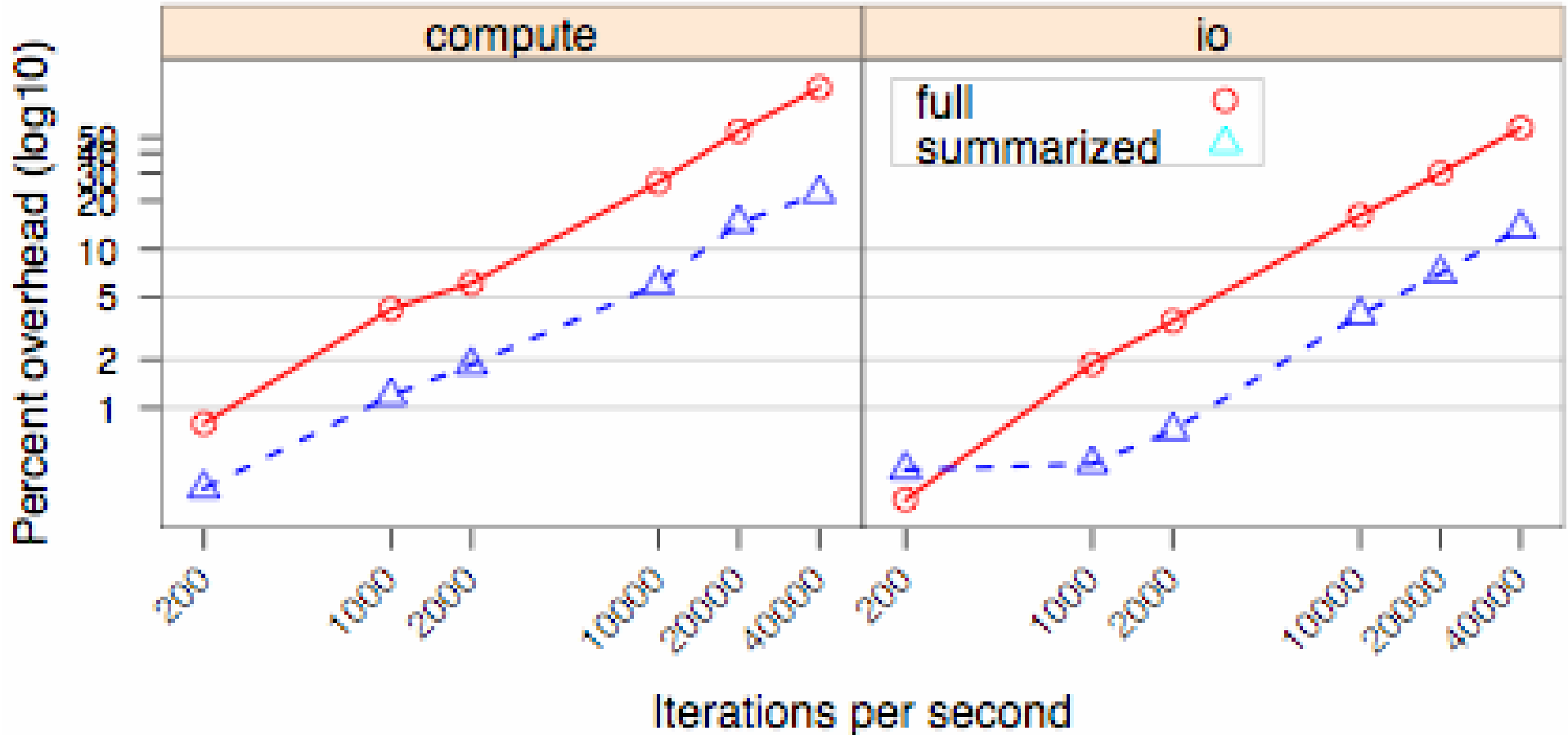
- Many devices
- Many failure cases
- Distributed responsibility
- Interactions between components



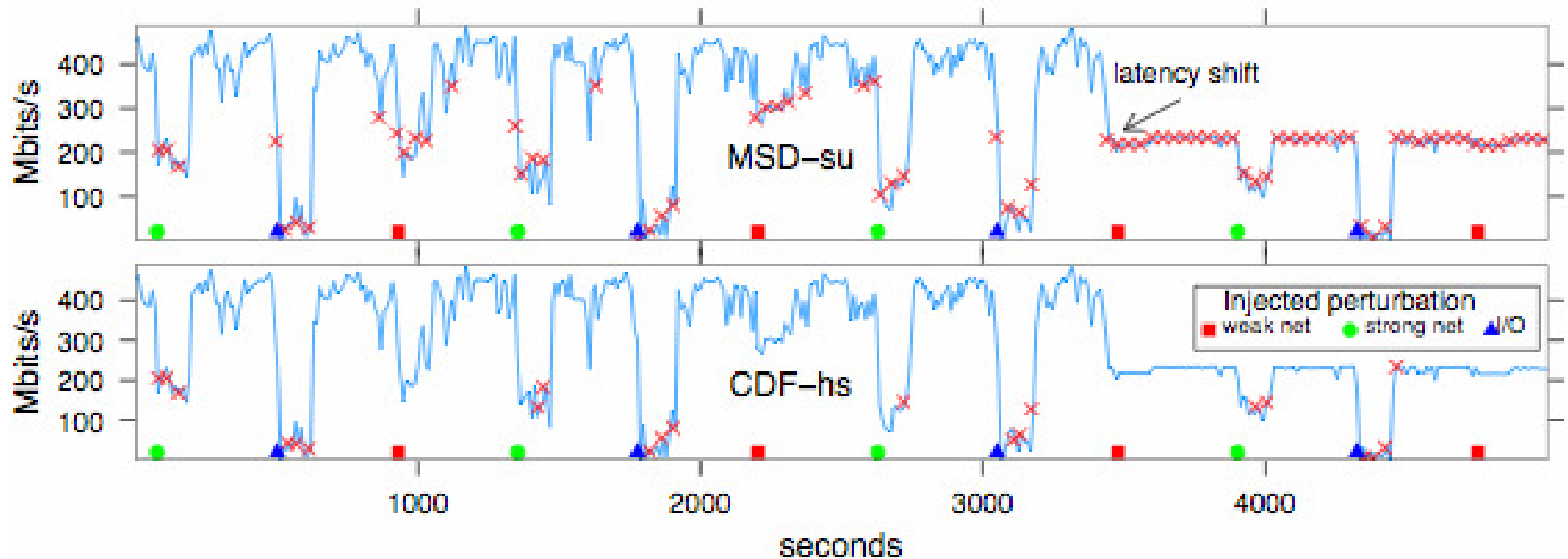
CEDPS Log Generation & Collection: GridFTP

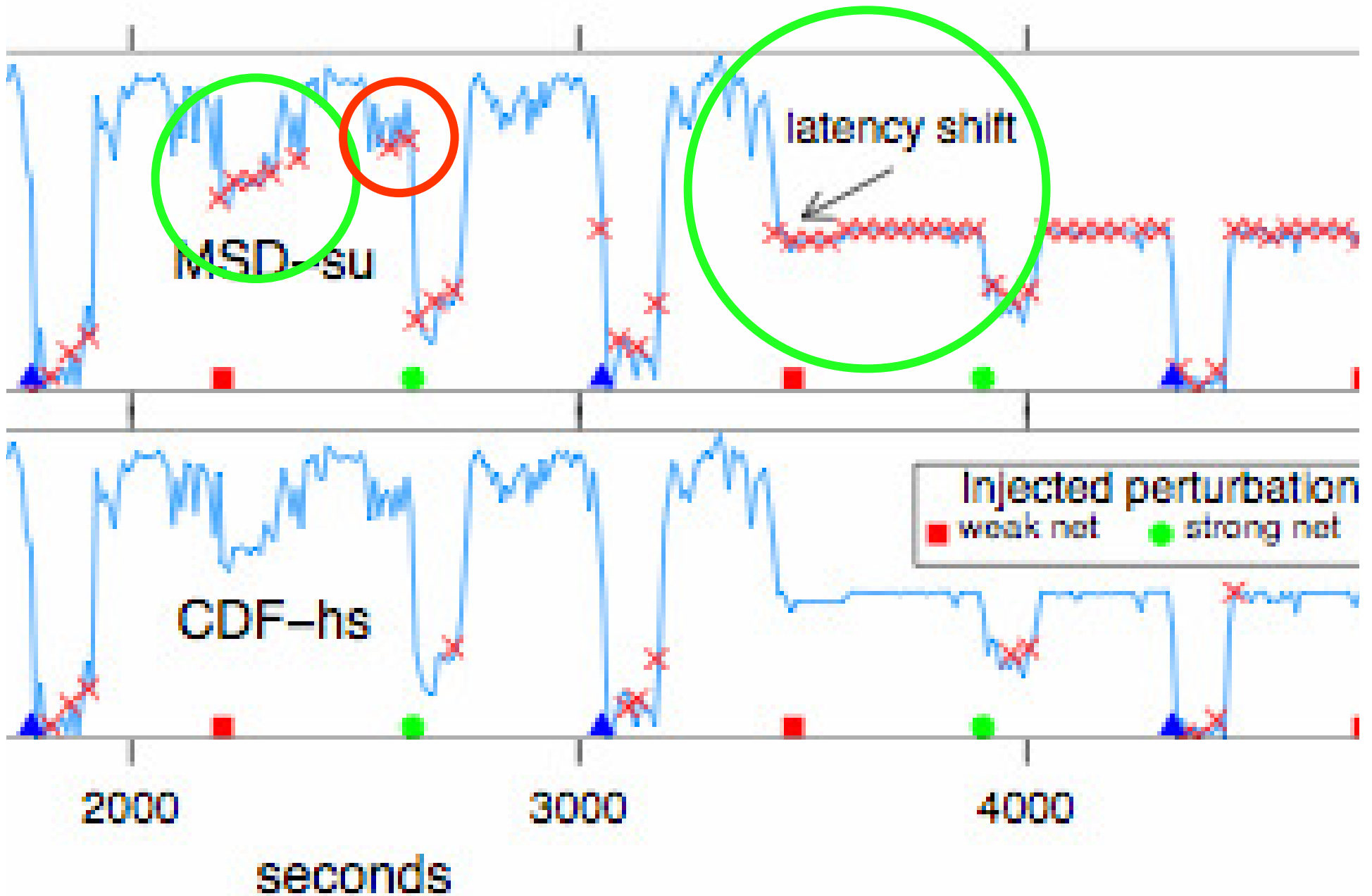


Perturbation Caused by Trace Generation



GridFTP with Injected Performance Perturbations





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

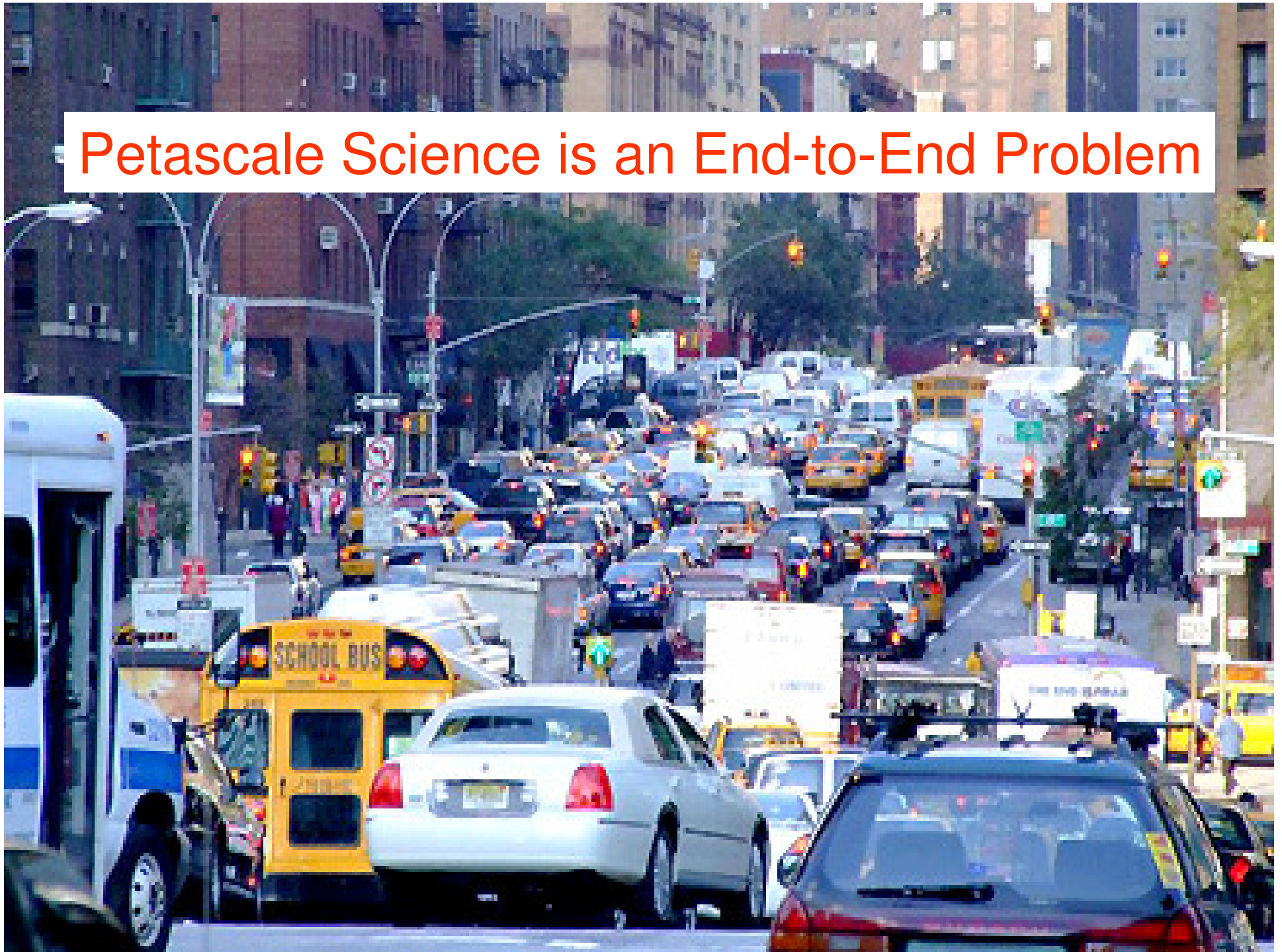
■ The last mile

- 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



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!