Software-as-a-Service for Research Data Management

Raj Kettimuthu

Argonne National Laboratory and
University of Chicago

Exploding data volumes

Astronomy

MACHO et al.: 1 TB

Palomar: 3 TB

2MASS: 10 TB

GALEX: 30 TB

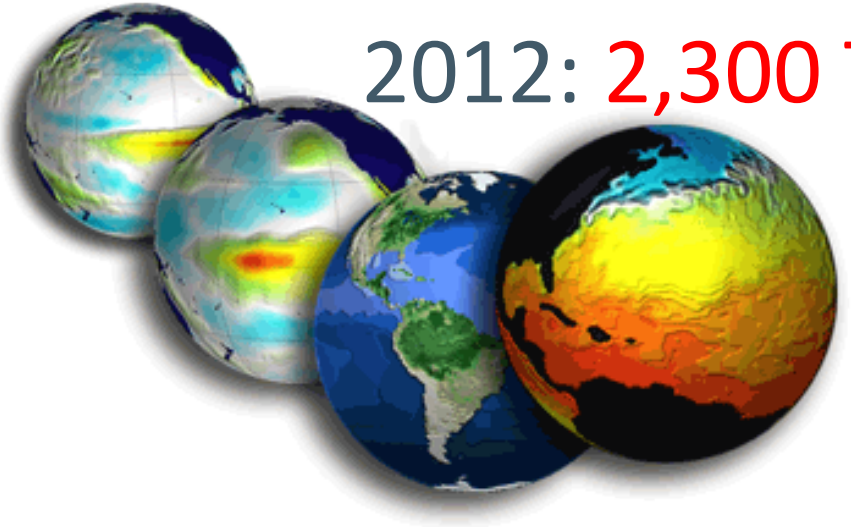
Sloan: 40 TB

Pan-STARRS: 40,000 TB

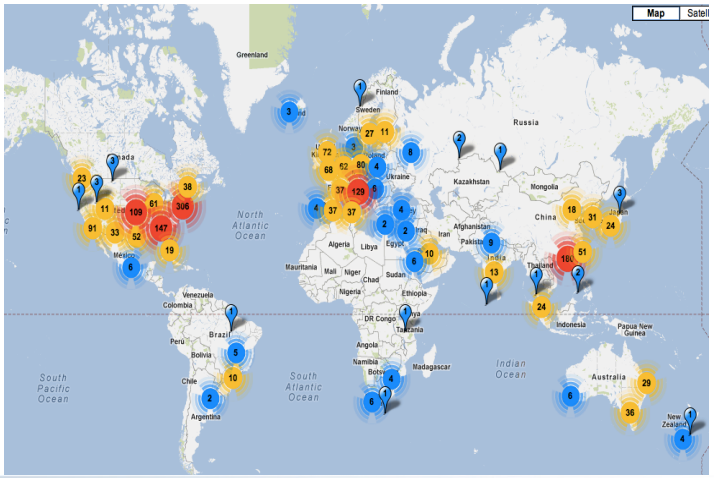
Climate

2004: 36 TB

2012: 2,300 TB

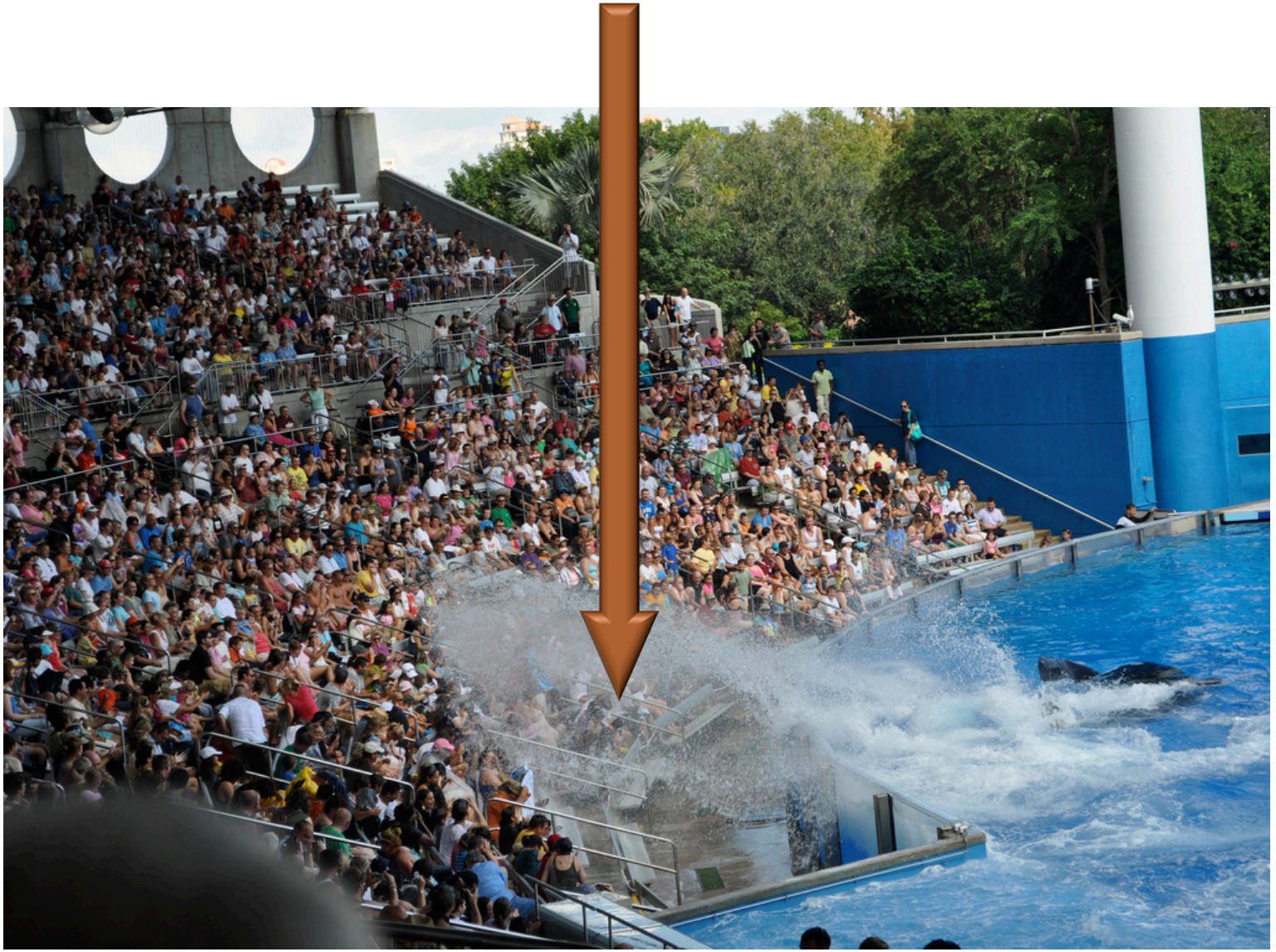


Genomics



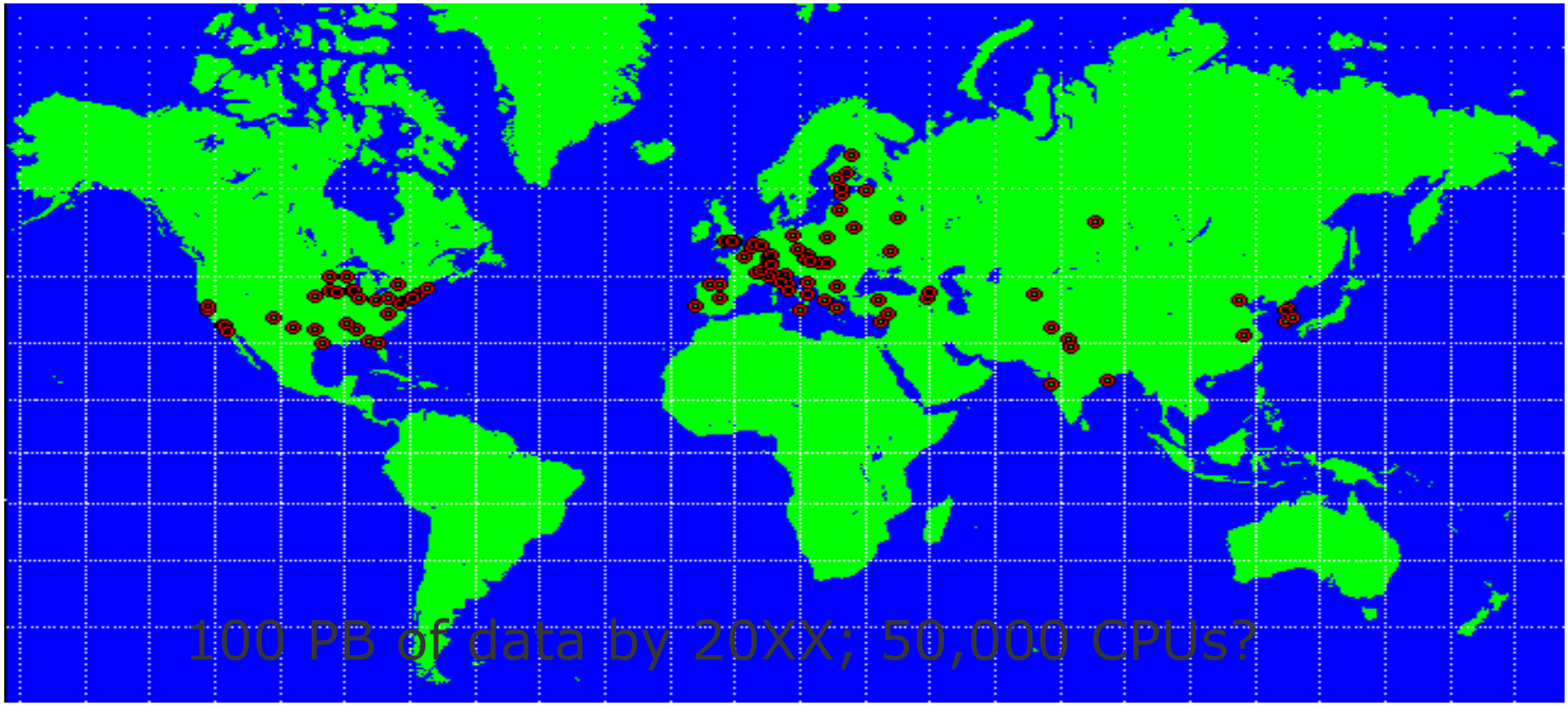
10^5 increase in data volumes in 6 years

Data Deluge

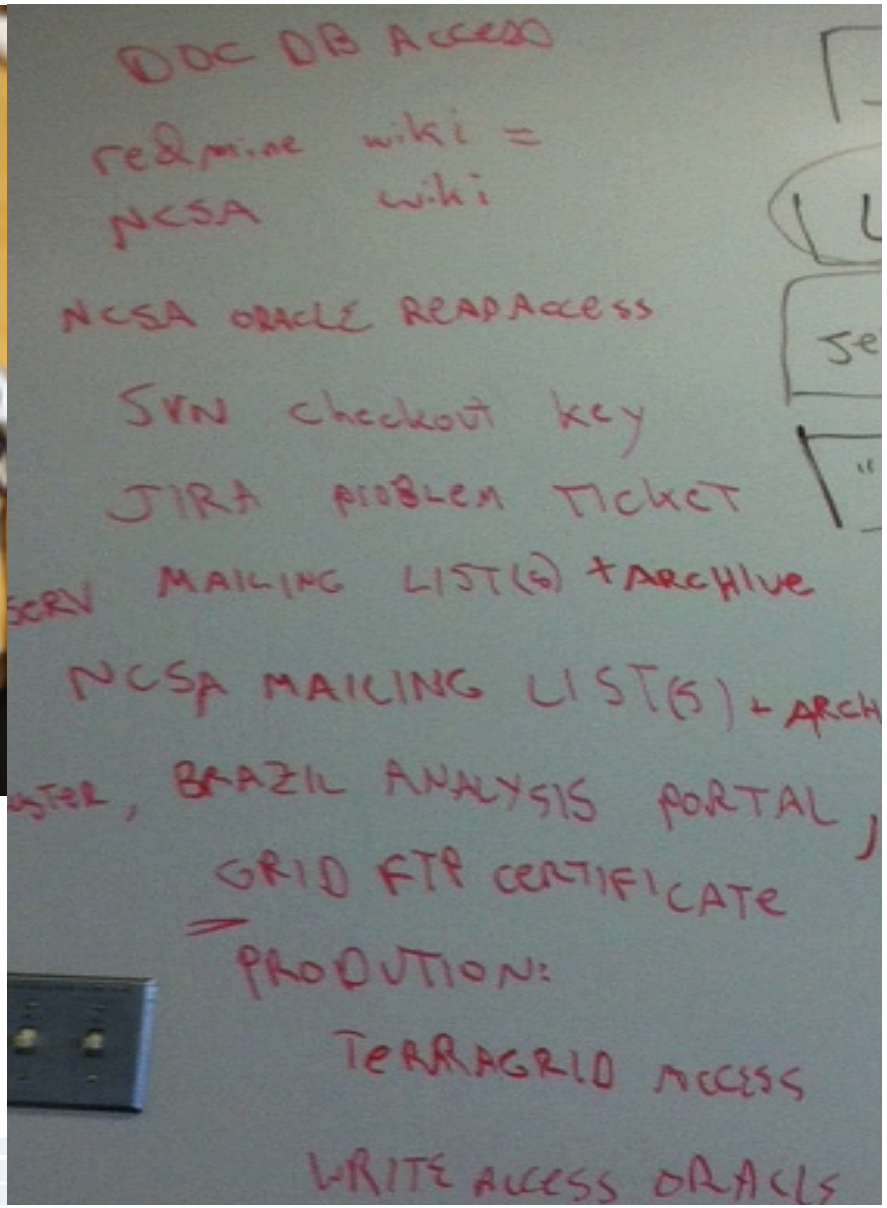


Large Hadron Collider

1800 Physicists, 150 Institutes, 32 Countries



But small and medium science is suffering



- Data deluge
- Ad-hoc solutions
- Inadequate software, hardware & IT staff

Research Data Management as a Service

Accelerate discovery and innovation worldwide by providing research data management as a service

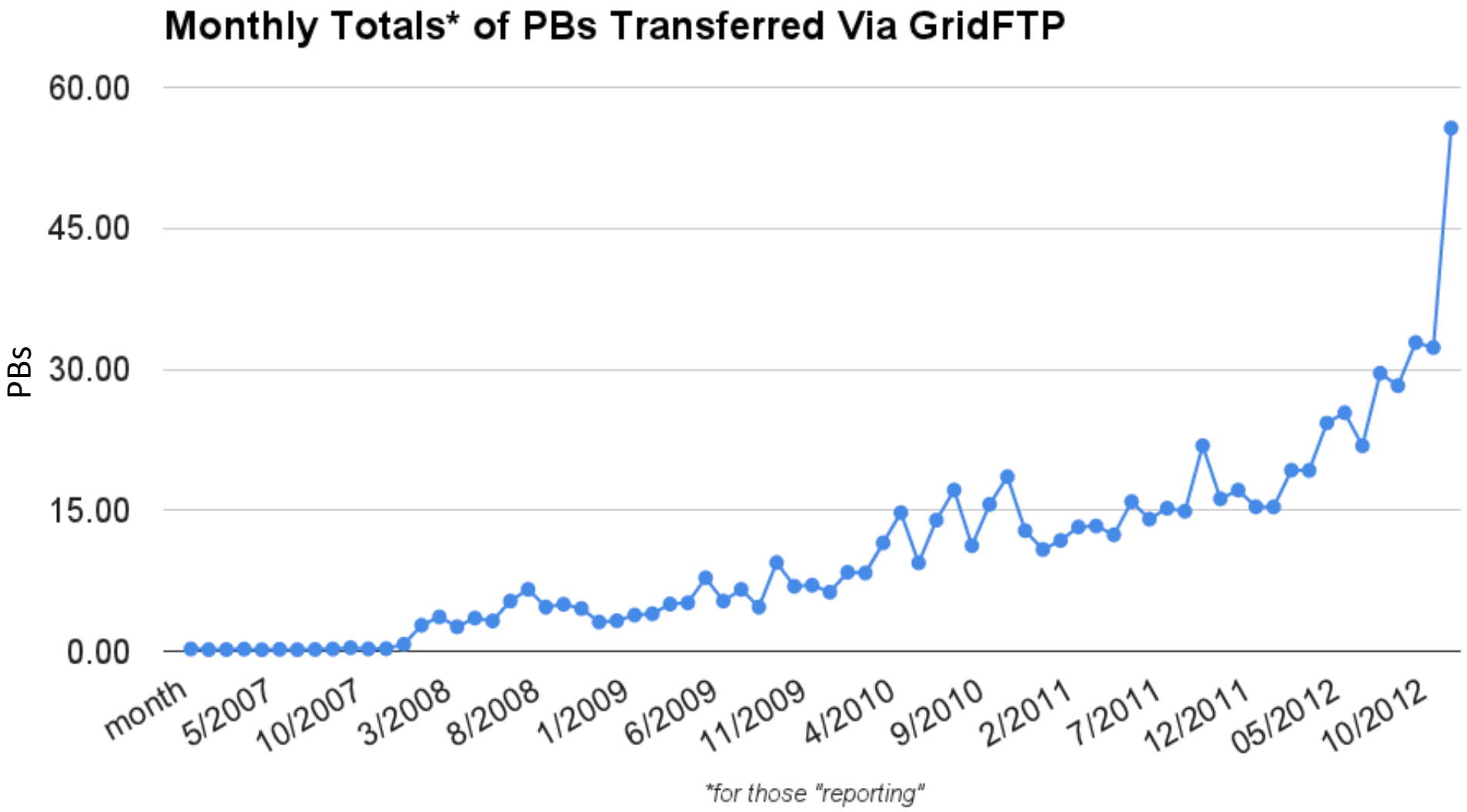
Leverage the cloud to

- provide millions of researchers with unprecedented access to powerful tools;
- enable a massive shortening of cycle times in time-consuming research processes; and
- reduce research IT costs dramatically via economies of scale


STARBUCKS COFFEE
STARBUCKS
COFFEE

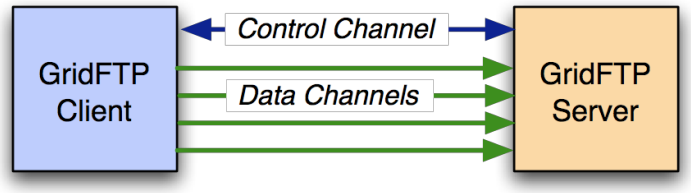
GridFTP servers around the world

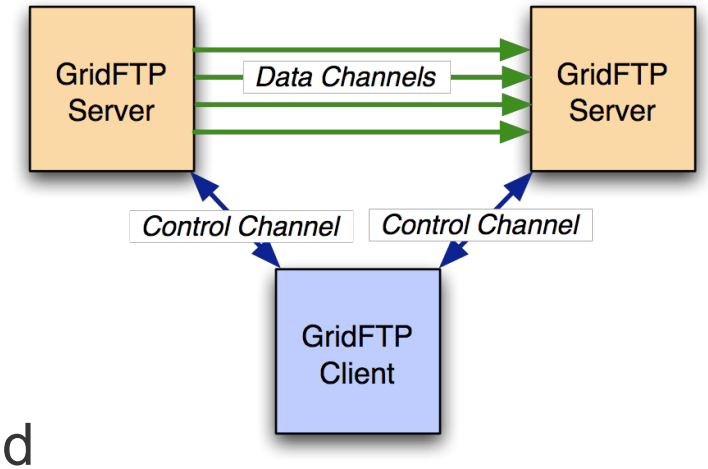
GridFTP usage



GridFTP

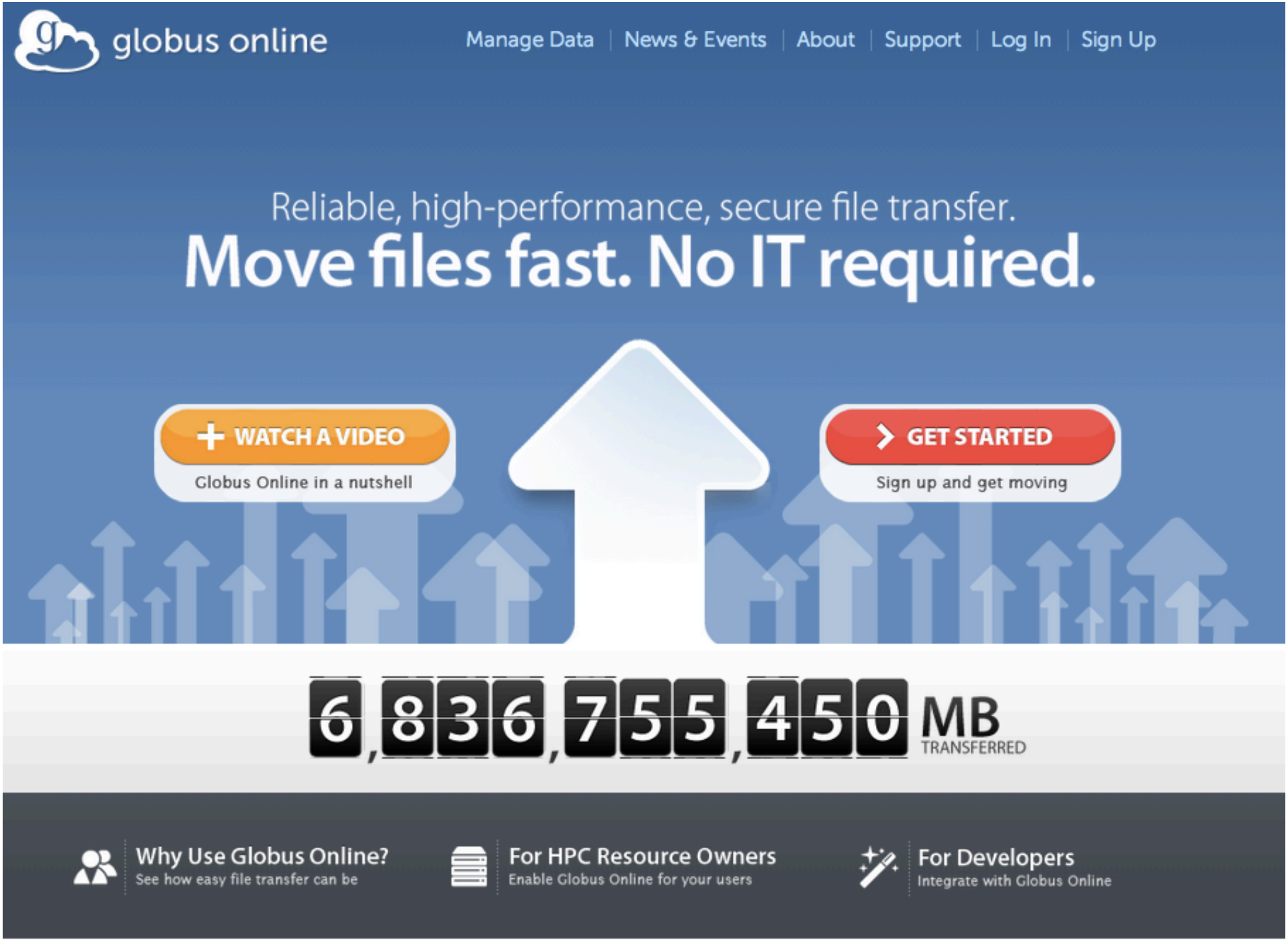
- Extension of the standard FTP
- Two channel protocol like FTP
- Control Channel
  - Command/Response
  - Used to establish data channels
  - Basic file system operations
    - eg. mkdir, delete etc
- Data channel
  - Pathway over which *file* is transferred
  - Many different underlying protocols can be used
    - MODE command determines the protocol





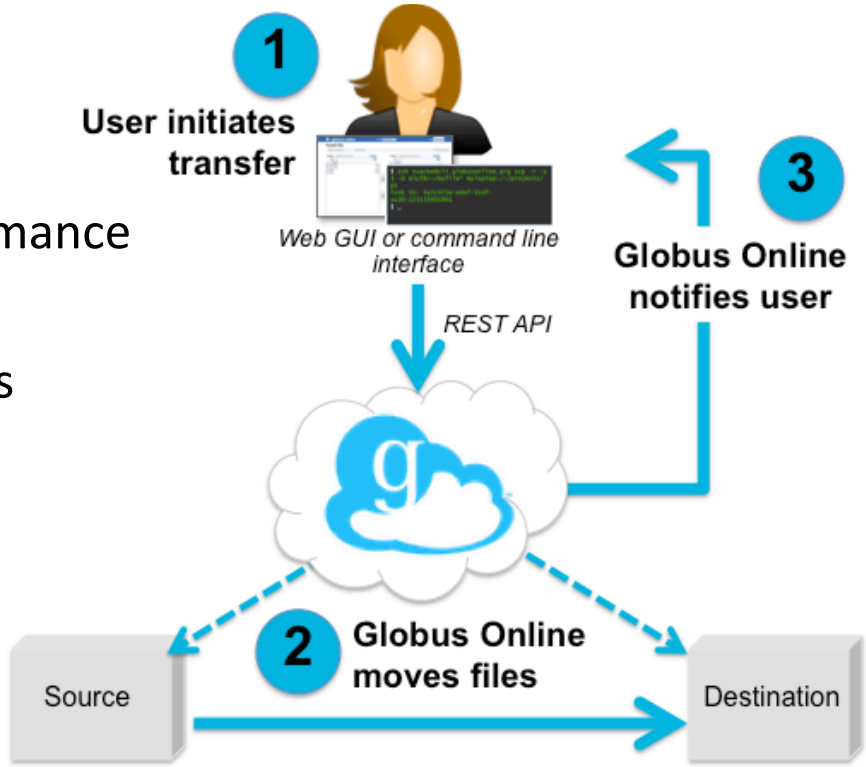
GridFTP Adoption

- GridFTP has been around for more than a decade now
- Until about a year or two ago, GridFTP was mostly used only by big science projects
  - LHC, ESG, LIGO etc
- Two key reasons
  - Security configuration was difficult both for end users and GridFTP server administrators
  - End users were not able to handoff the data movement task to some generic client and forget about it


globus online
Manage Data
News & Events
About
Support
Log In
Sign Up
Reliable, high-performance, secure file transfer.
Move files fast. No IT required.
WATCH A VIDEO
Globus Online in a nutshell
GET STARTED
Sign up and get moving
6,836,755,450 MB
TRANSFERRED
Why Use Globus Online?
See how easy file transfer can be
For HPC Resource Owners
Enable Globus Online for your users
For Developers
Integrate with Globus Online

What is Globus Online?

- Move, Sync, Share files
  - Easy “fire-and-forget” transfers
  - Share with any Globus user or group
  - Automatic fault recovery & High performance
  - Across multiple security domains
  - Web, command line and REST interfaces
- Minimize IT costs
  - Software as a Service (SaaS)
    - No client software installation
    - New features automatically available
  - Consolidated support & troubleshooting
  - Simple endpoint installation with Globus Connect and GridFTP
- Recommended by XSEDE, Blue Water, NERSC, ALCF, ESnet, many Universities



LIVE DEMO

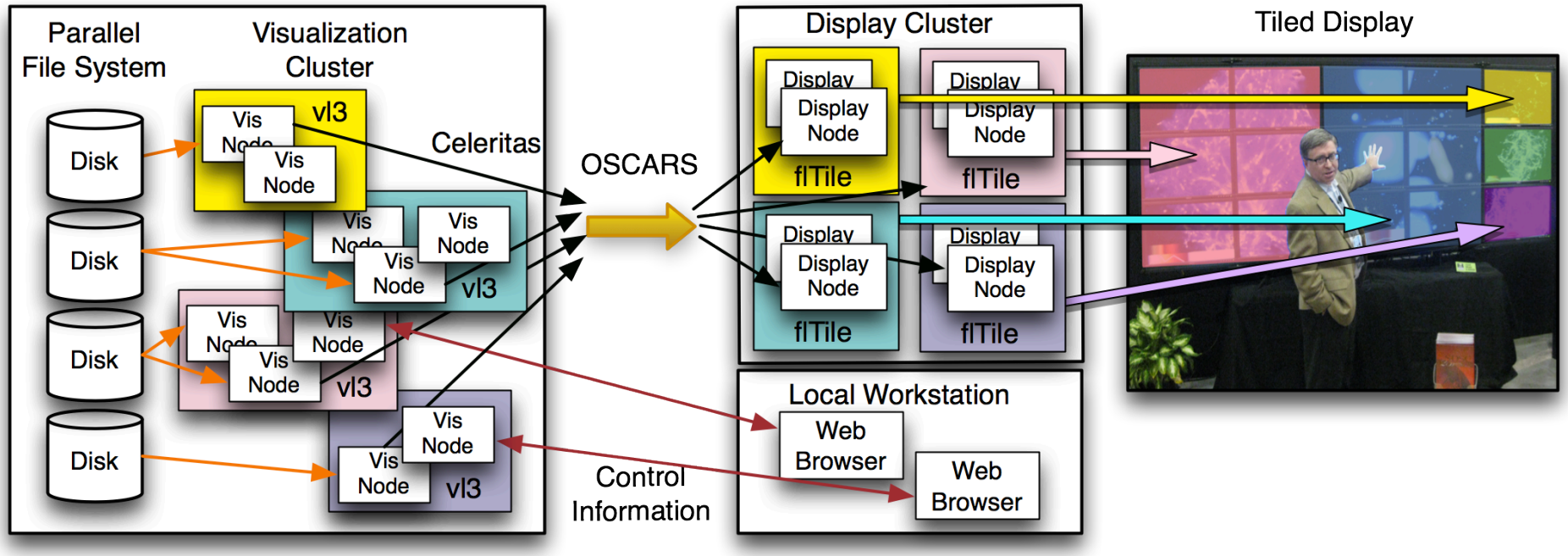
Early adoption is encouraging

Early adoption is encouraging

- 7,500 registered users; ~100 daily
- >10PB moved; >500MM files
- 10x (or better) performance vs. scp
- 99.9% availability
- Entirely hosted on AWS

Resource Aware Protocols

Interactive Remote Visualization of ENZO Cosmology



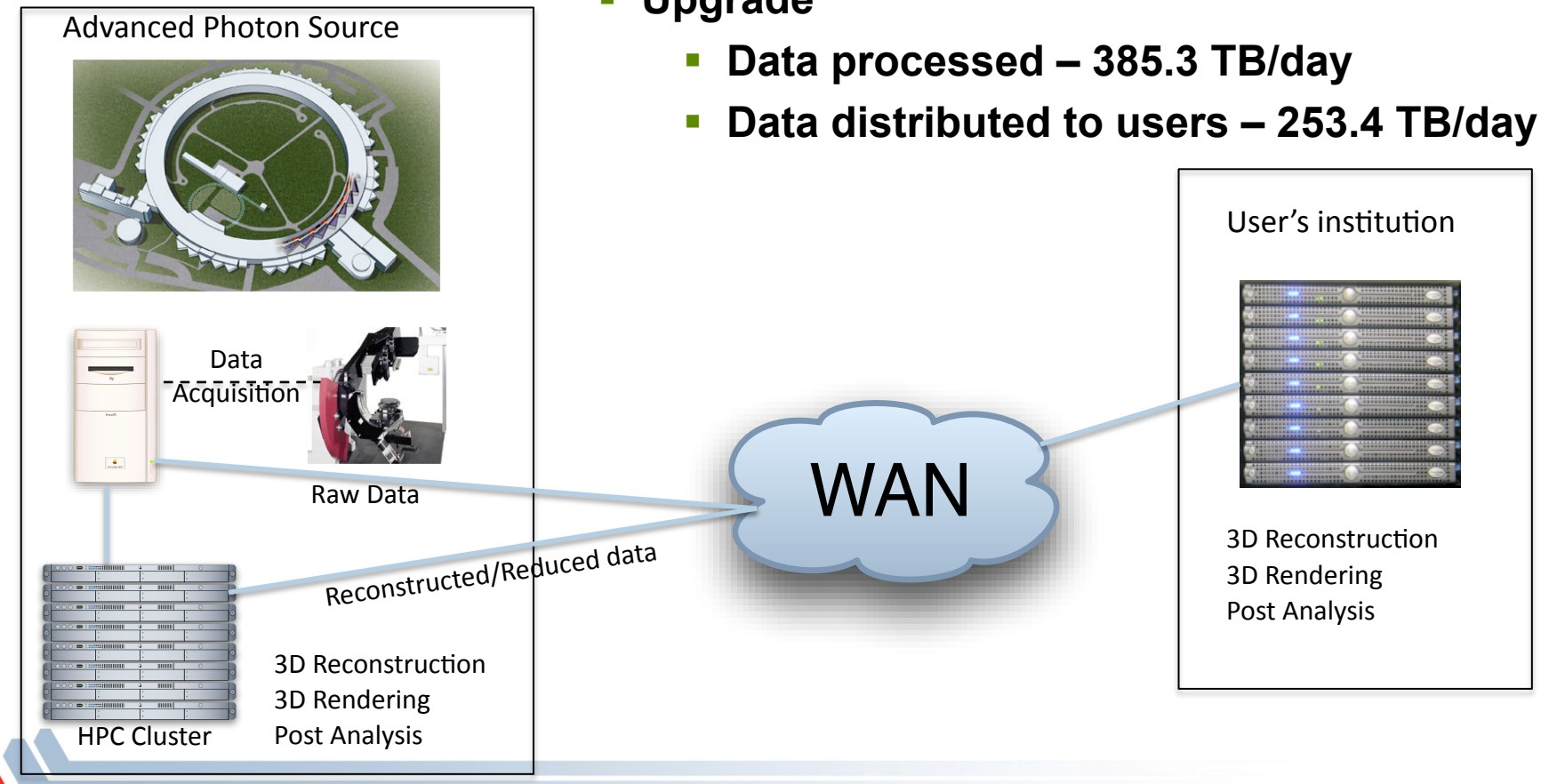
Argonne National Laboratory

San Diego

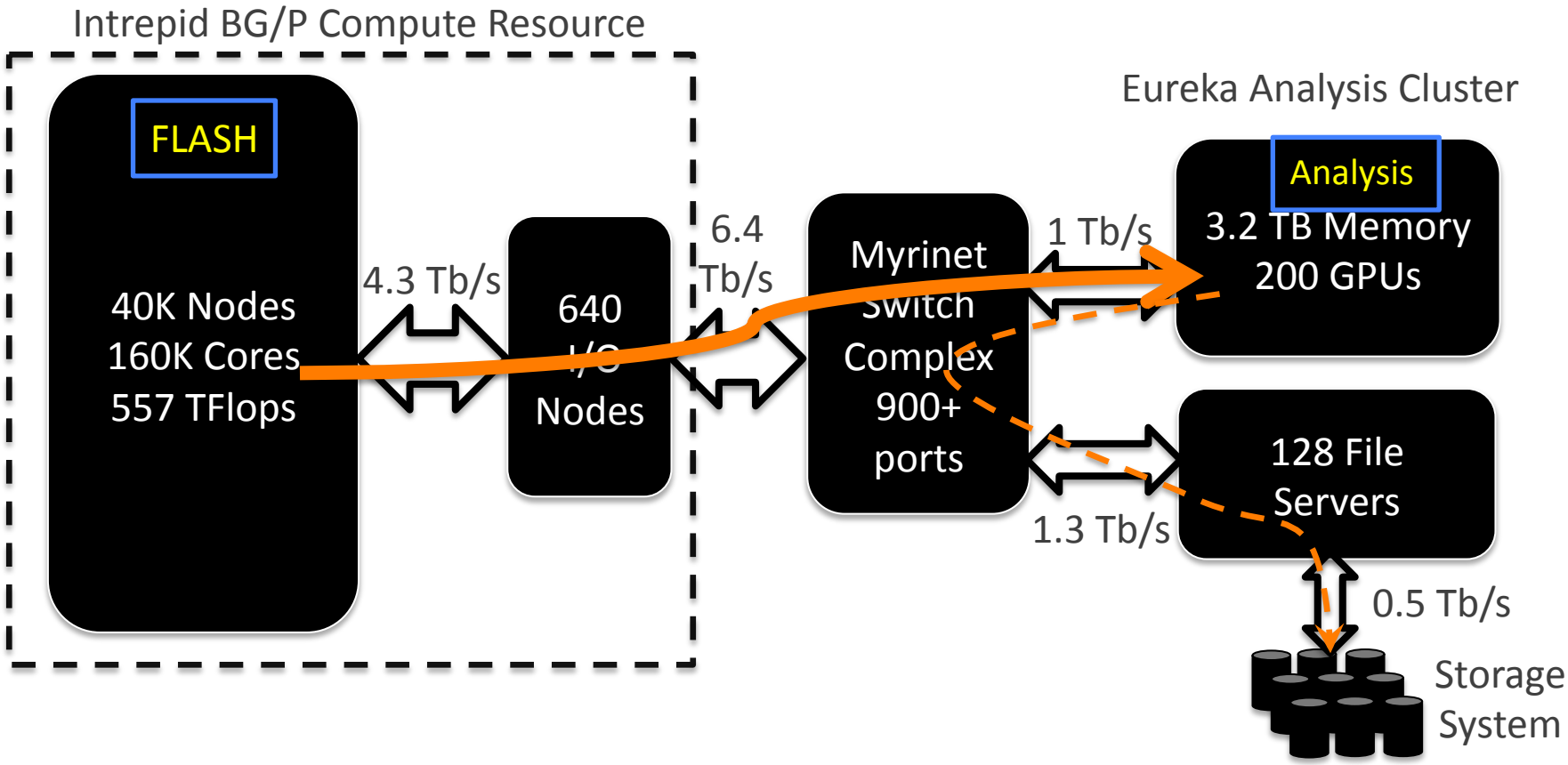
New Orleans - SC'10 Show floor

Tomography at APS

- **Current**
  - **Data processed – 5.6 TB/day**
  - **Data distributed to users – 3.3 TB/day**
- **Upgrade**
  - **Data processed – 385.3 TB/day**
  - **Data distributed to users – 253.4 TB/day**

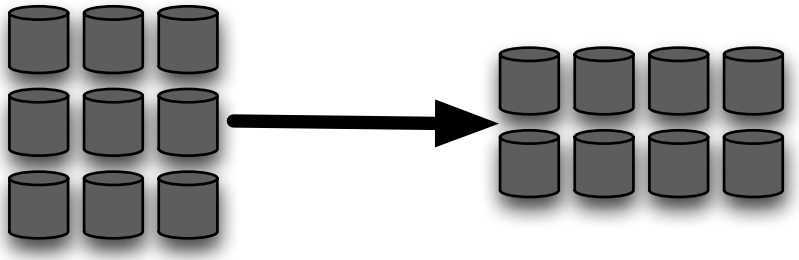


Simulation-time Data Analysis and Visualization of FLASH Astrophysics Simulation

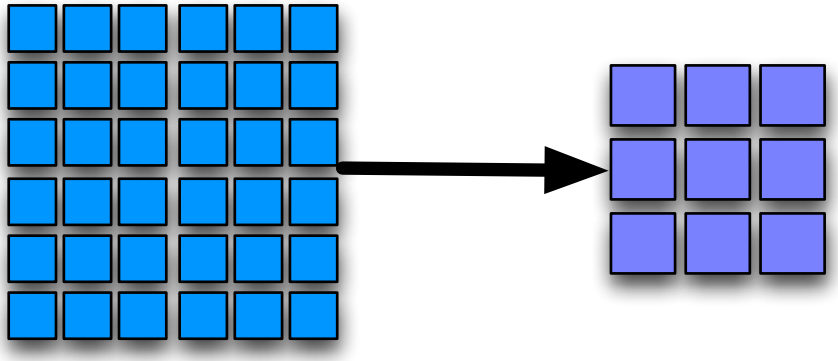


Simulation-time data analysis is critical to reduce the data written to storage and to generate faster insights

Data Movement Trends

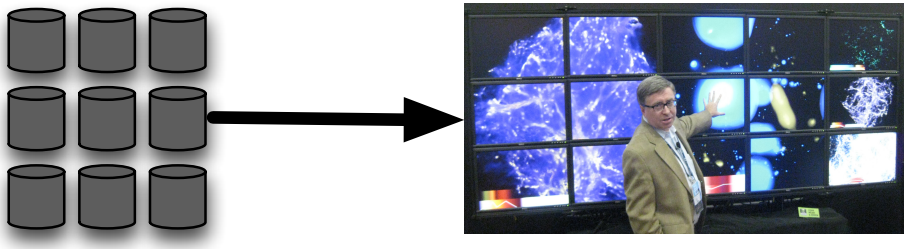


Disk-to-Disk Transfers

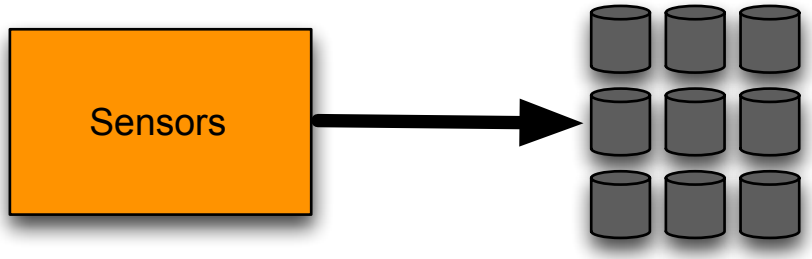


Memory-to-Memory Transfers

Parallel M-to-N Data Flows



Disk-to-Memory Transfers



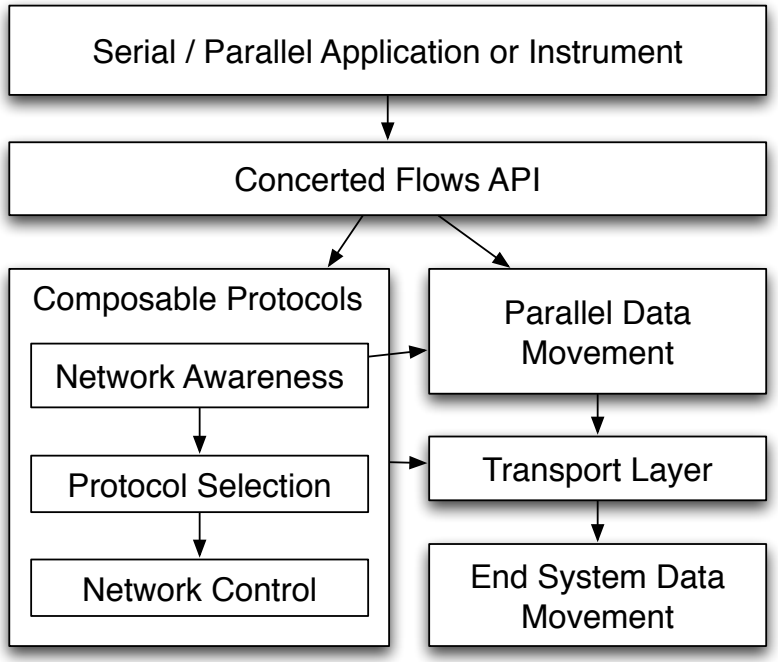
Memory-to-Disk Transfers

Network Characteristics

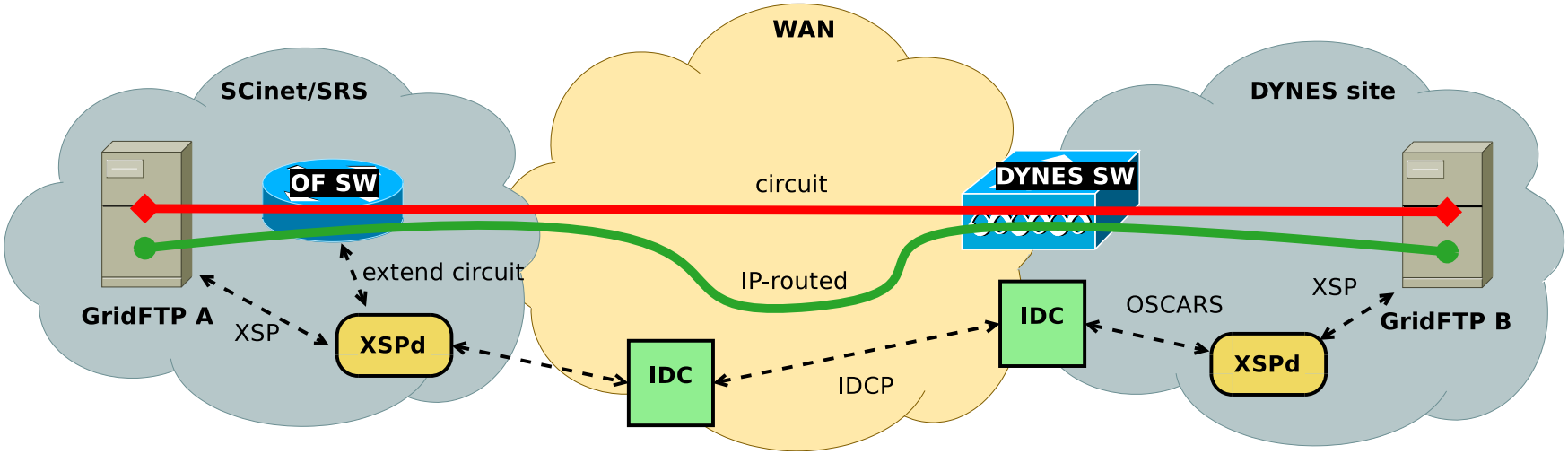
- Network Type
  - Shared or dedicated
  - Circuit or packet or hybrid
- Network activity
  - Over-utilized or under-utilized
- Network Topology
  - Parallel paths
  - Bandwidth, latency, loss rate
- LAN (within a leadership facility), MAN or WAN
- Network is no longer a blackbox
  - Software Defined Networking
  - Topology and link state information available
  - Guaranteed bandwidth

Concerted Flows

- Develop new parallel protocols that are
- Capture the diverse flow characteristics and needs
- Leverages feedback from network agents and exploits topology to design flow and congestion control for parallel data movement
- Build a knowledge base capturing the data transfer patterns of several DOE applications



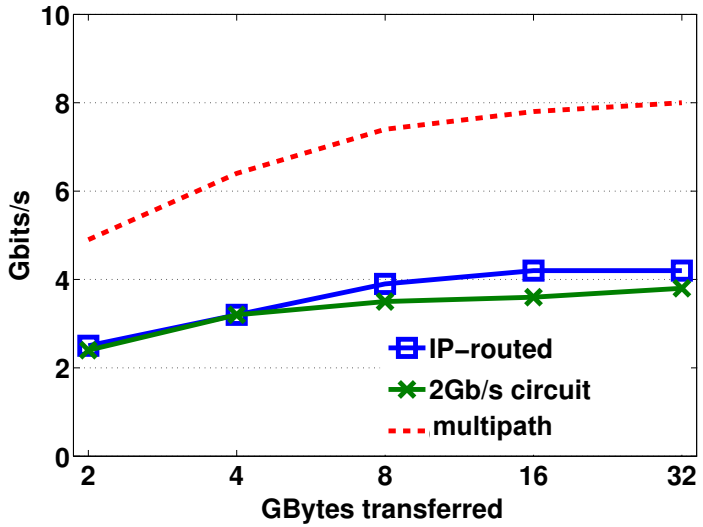
Exploiting multiple paths



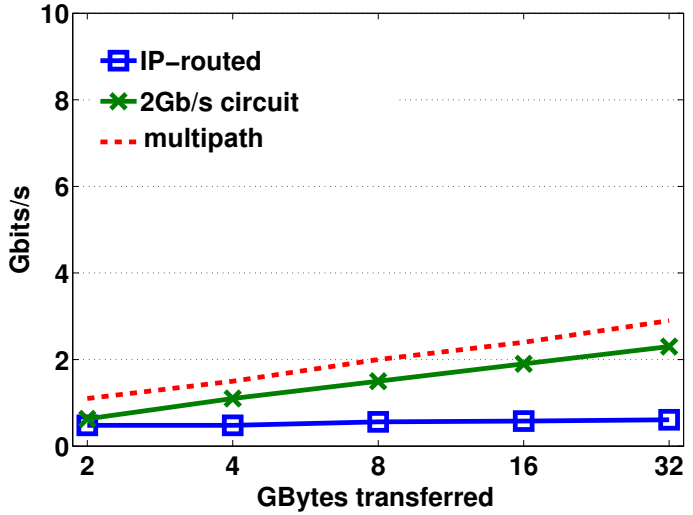
- Take advantage of the multiple interfaces in the multi-homed data transfer nodes
- Use circuit as well as production IP link
- Data will flow even while the circuit is being setup
- Once the circuit is setup, use both the paths to improve throughput

Exploiting multiple paths

Transfer between NERSC and ANL



Transfer between Umich and Caltech



Default, commodity IP routes

+ Dedicated circuits

= Significant performance gains

For More Information

- Visit www.globusonline.org/signup to:
  - Get a free account to start moving and sharing files
- Visit www.globusonline.org for:
  - Tutorials, FAQs, Pro Tips, Troubleshooting
  - Papers, Case Studies
- Visit support.globusonline.org or contact support@globusonline.org for:
  - Help
  - Forums
- Follow us at @globusonline on Twitter and Globus Online on Facebook

Questions?