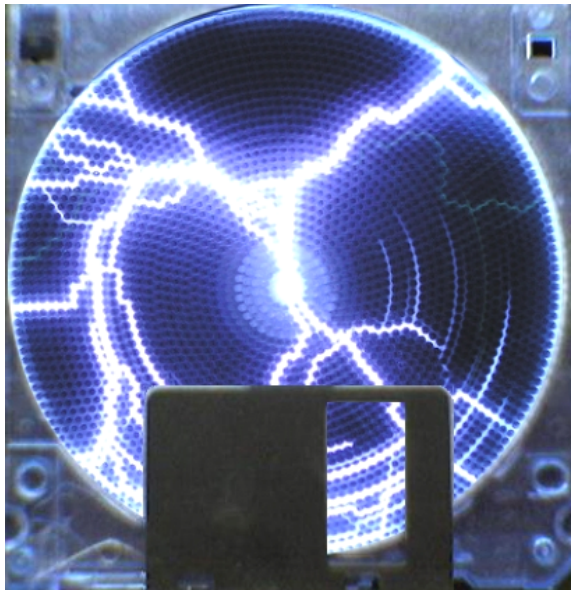
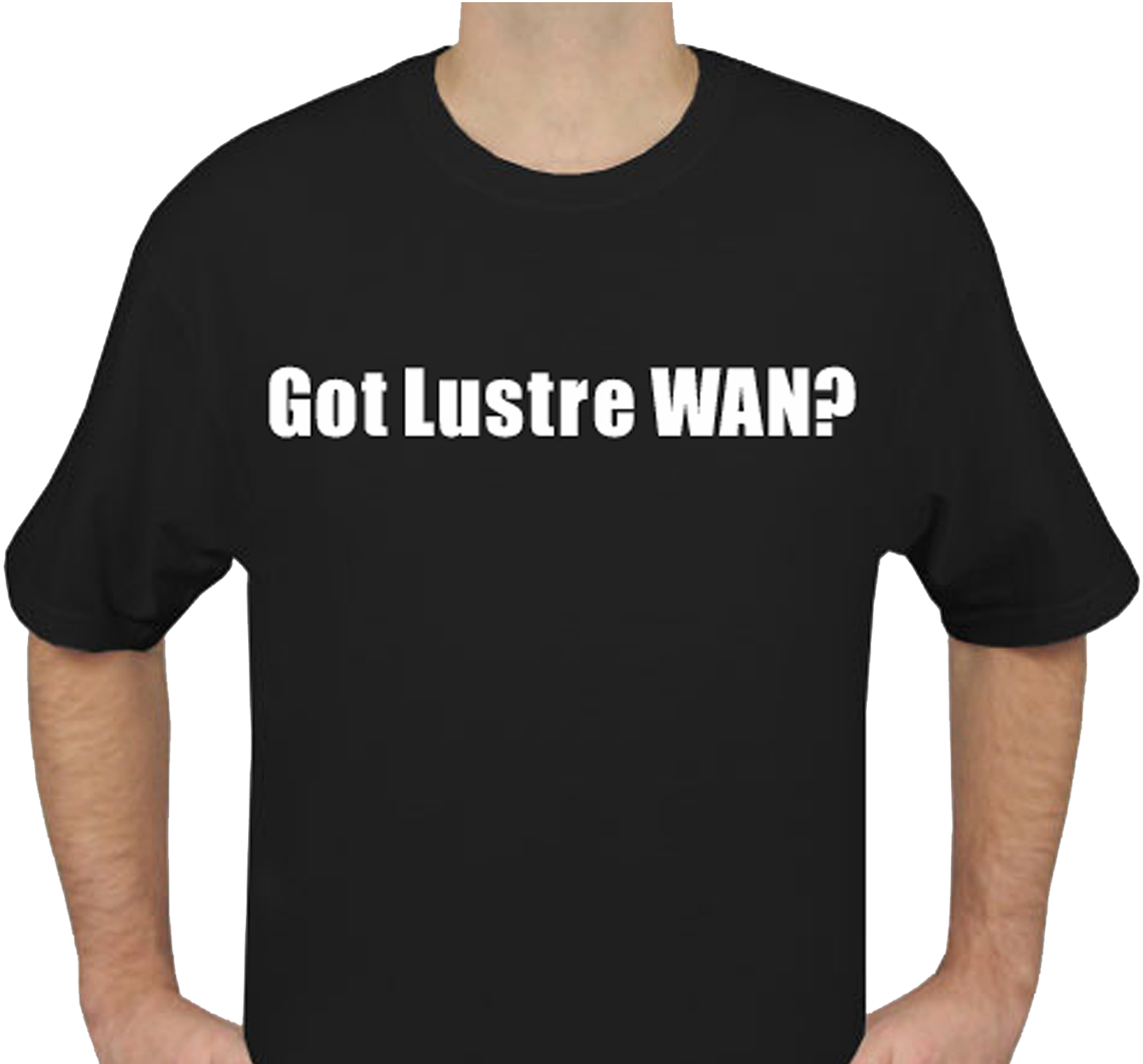
Indiana University's Lustre WAN:
Empowering Production Workflows on the TeraGrid



Stephen C. Simms
Manager, Data Capacitor Project
TeraGrid Site Lead, Indiana University
ssimms@indiana.edu


Got Lustre WAN?

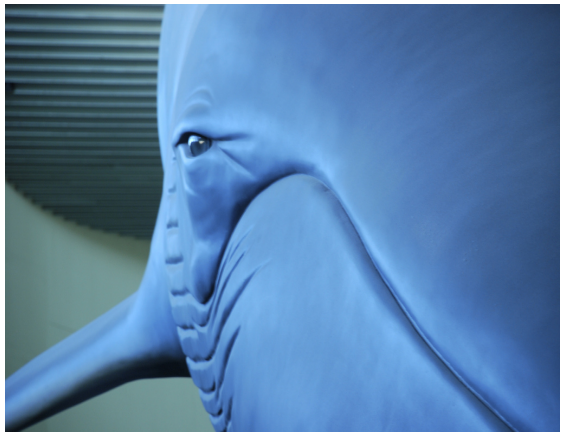
The Data Capacitor Project

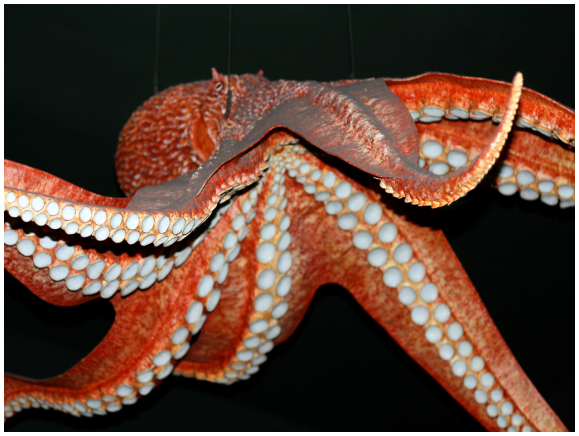
NSF Funded in 2005

535 Terabytes Lustre storage

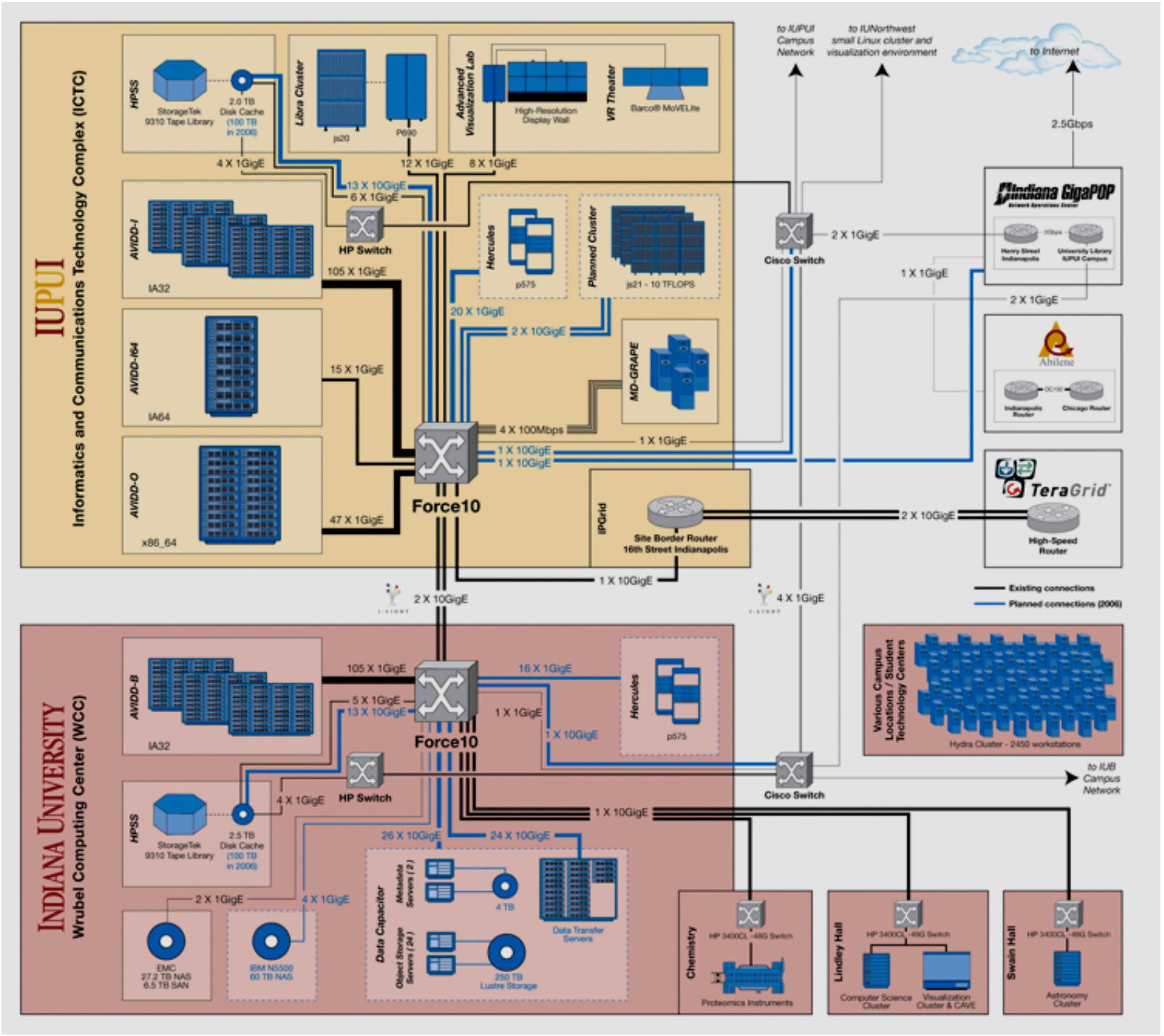
14.5 GB/s aggregate write

Short term storage





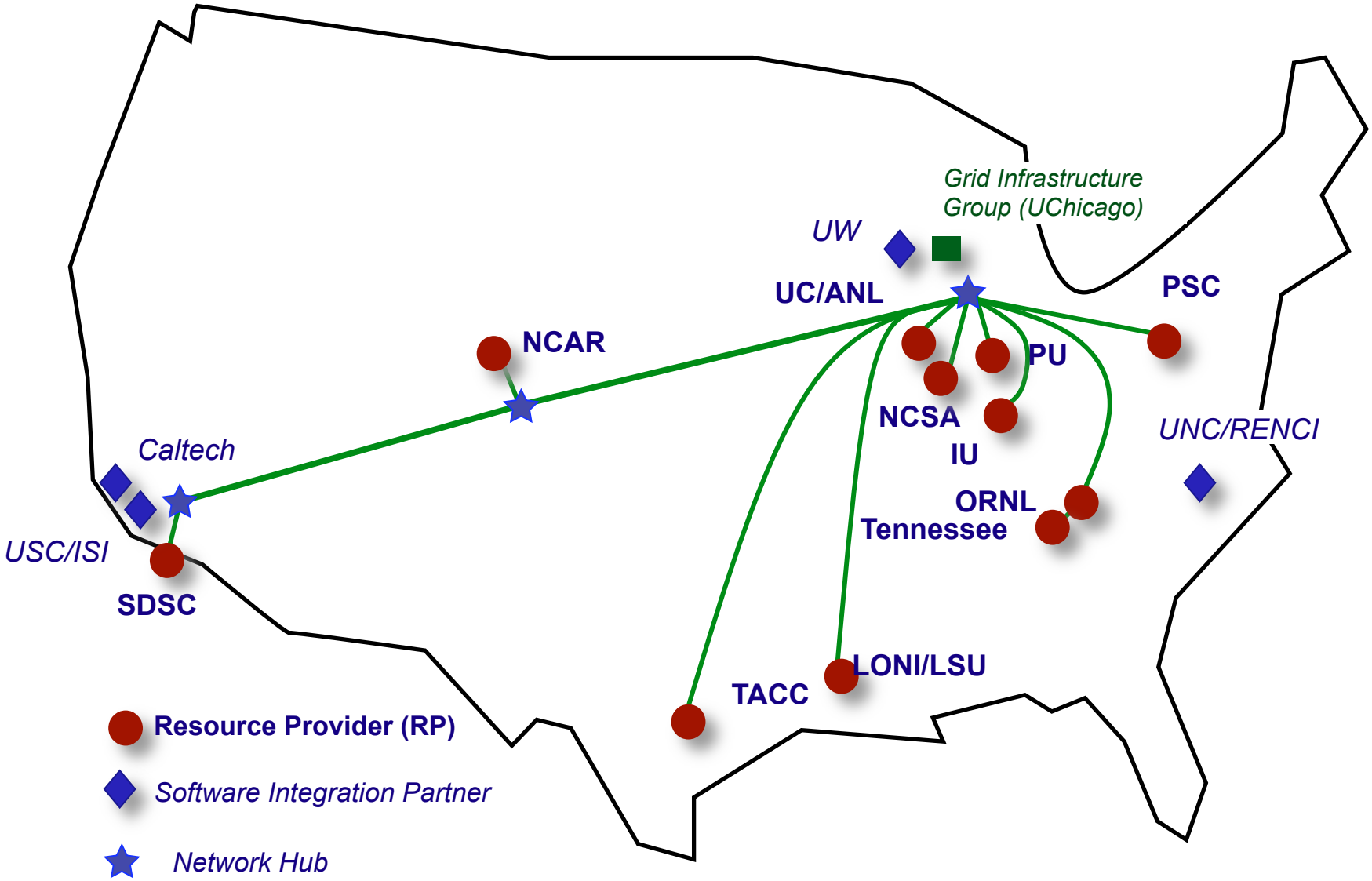
http://www.flickr.com/photos/shadowstorm/404158384/
http://www.flickr.com/photos/dvd5/163647219/
http://www.flickr.com/photos/vidiot/431357888/


IUPUI
Informatics and Communications Technology Complex (ICTC)
HPSS
StorageTek 9310 Tape Library
2.0 TB Disk Cache (100 TB in 2006)
Libra Cluster
js20
P690
Advanced Visualization Lab
High-Resolution Display Wall
VR Theater
Barco® MoVELite
4 X 1GigE
12 X 1GigE
8 X 1GigE
13 X 10GigE
6 X 1GigE
AVIDD-I
IA32
HP Switch
105 X 1GigE
Hercules
p575
Planned Cluster
js21 - 10 TFLOPS
20 X 1GigE
2 X 10GigE
AVIDD-I64
IA64
15 X 1GigE
MD-GRAPE
4 X 100Mbps
1 X 1GigE
1 X 10GigE
1 X 10GigE
AVIDD-O
x86_64
47 X 1GigE
Force10
IPGrid
Site Border Router
16th Street Indianapolis
2 X 10GigE
1 X 10GigE
to IUPUI Campus Network
to IUNorthwest
small Linux cluster and visualization environment
to Internet
2.5Gbps
Cisco Switch
2 X 1GigE
1 X 1GigE
Indiana GigaPOP
Henry Street Indianapolis
University Library IUPUI Campus
2 X 1GigE
Abilene
Indianapolis Router
Chicago Router
TeraGrid
High-Speed Router
2 X 10GigE
Existing connections
Planned connections (2006)
2 X 10GigE
4 X 1GigE
INDIANA UNIVERSITY
Wrubel Computing Center (WCC)
AVIDD-B
IA32
105 X 1GigE
5 X 1GigE
13 X 10GigE
Force10
16 X 1GigE
1 X 1GigE
1 X 10GigE
Hercules
p575
Various Campus Locations / Student Technology Centers
Hydra Cluster - 2450 workstations
HP Switch
Cisco Switch
to IUB Campus Network
HPSS
StorageTek 9310 Tape Library
2.5 TB Disk Cache (100 TB in 2006)
4 X 1GigE
1 X 10GigE
26 X 10GigE
24 X 10GigE
Data Capacitor
Metadata Servers (2)
Object Storage Servers (24)
4 TB
250 TB Lustre Storage
Data Transfer Servers
2 X 1GigE
4 X 1GigE
EMC 27.2 TB NAS 6.5 TB SAN
IBM N5500 60 TB NAS
Chemistry
HP 3400CL-48G Switch
Proteomics Instruments
Lindley Hall
HP 3400CL-48G Switch
Computer Science Cluster
Visualization Cluster & CAVE
Swain Hall
HP 3400CL-48G Switch
Astronomy Cluster

The TeraGrid

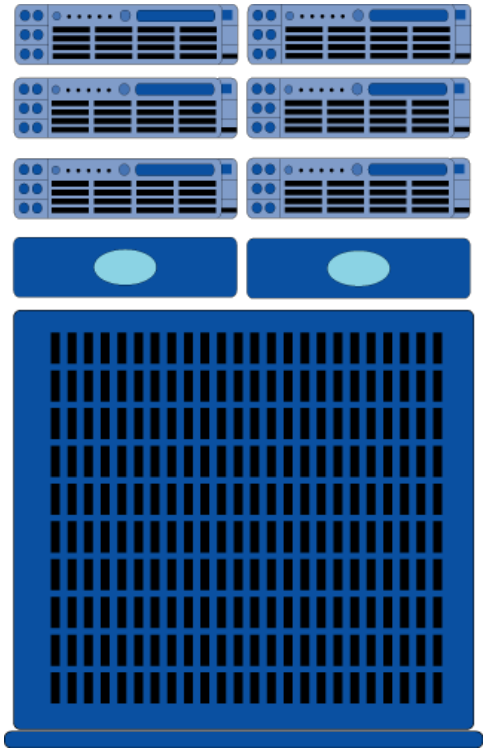
- TeraGrid is an open scientific discovery infrastructure combining leadership class resources at eleven partner sites to create an integrated, persistent computational resource.
- ANL, IU, Louisiana Optical Network Initiative (LONI), NCSA, NICS, ORNL, PSC, Purdue, SDSC,TACC, NCAR

The TeraGrid Map



IU's Data Capacitor WAN

- 1 pair Dell PowerEdge 2950 for MDS
- 2 pair Dell PowerEdge 2950 for OSS
  - 2 x 3.0 GHz Dual Core Xeon
  - Myrinet 10G Ethernet
  - Dual port Qlogic 2432 HBA (4 x FC)
  - 2.6 Kernel (RHEL 5)
- DDN S2A9550 Controller
  - Over 2.4 GB/sec measured throughput
  - 360 Terabytes of spinning SATA disk
- Currently running Lustre 1.6.7.2
  - Upgrading to 1.8.1.1 in May
- Announced production at LUG 2008
  - Allocated on Project by Project basis



Networking


CIC Omnipop


IP grid


manLan
Manhattan Landing
Exchange Point


NLR
light the future


INTERNET 2


TransPAC


AMPATH


I - LIGHT

IU UID Mapping

Lightweight

- Not everyone needs / wants kerberos
- Not everyone needs / wants encryption

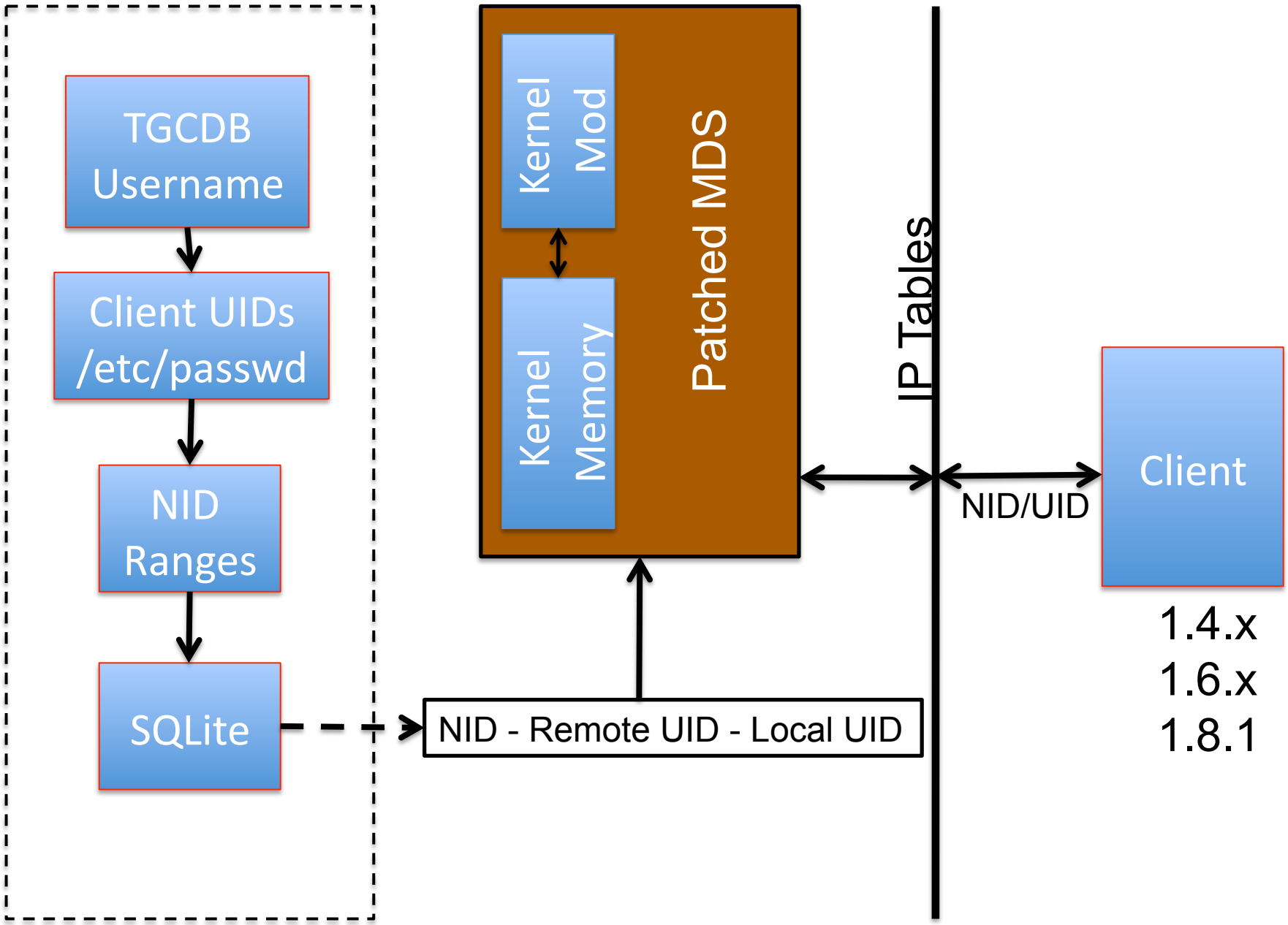
Only change MDS code

- Want to maximize clients we can serve

Simple enough to port the code forward

IU UID Mapping cont'd

- UID lookups on the MDS call a pluggable kernel module
  - Binary tree stored in memory
  - Based on NID or NID range
  - Remote UID mapped to Effective UID


TGCDB Username
Client UIDs /etc/passwd
NID Ranges
SQLite
NID - Remote UID - Local UID
Kernel Mod
Kernel Memory
Patched MDS
IP Tables
NID/UID
Client
1.4.x
1.6.x
1.8.1

UID Mapping

- Userspace – Kernel Space Barrier
  - Only crossed when we update the table
- Create a Forest of Binary Trees
  - Forward and Inverse Lookups for each UID
  - Time consumed for lookup is predictable
- Speed over Space
    - Consume memory rather than on the fly lookups
    - Every UID node consumes 6 Ints
    - 300 Users approximately 300KB

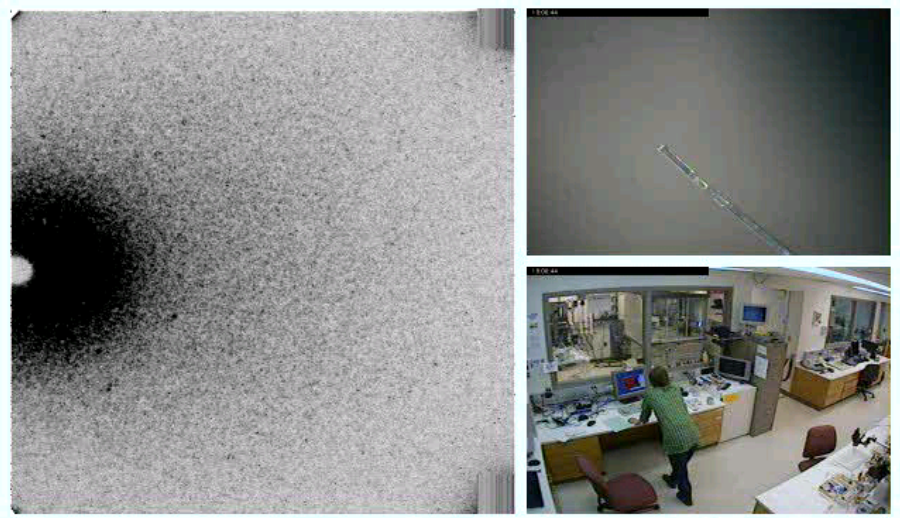
IU's Lustre WAN on the TeraGrid

- 8 Sites currently mounting IU DC-WAN
  - IU, LONI, NCSA, NICS, PSC, Purdue, SDSC, TACC
- 5 Sites mounting on compute resources
  - IU, LONI, NCSA, PSC, TACC
- Average of 93% capacity for the last quarter
- 2009 uptime of 96%
  - Filesystem availability to users

Data In and Data Out

Common Instrument Middleware Architecture

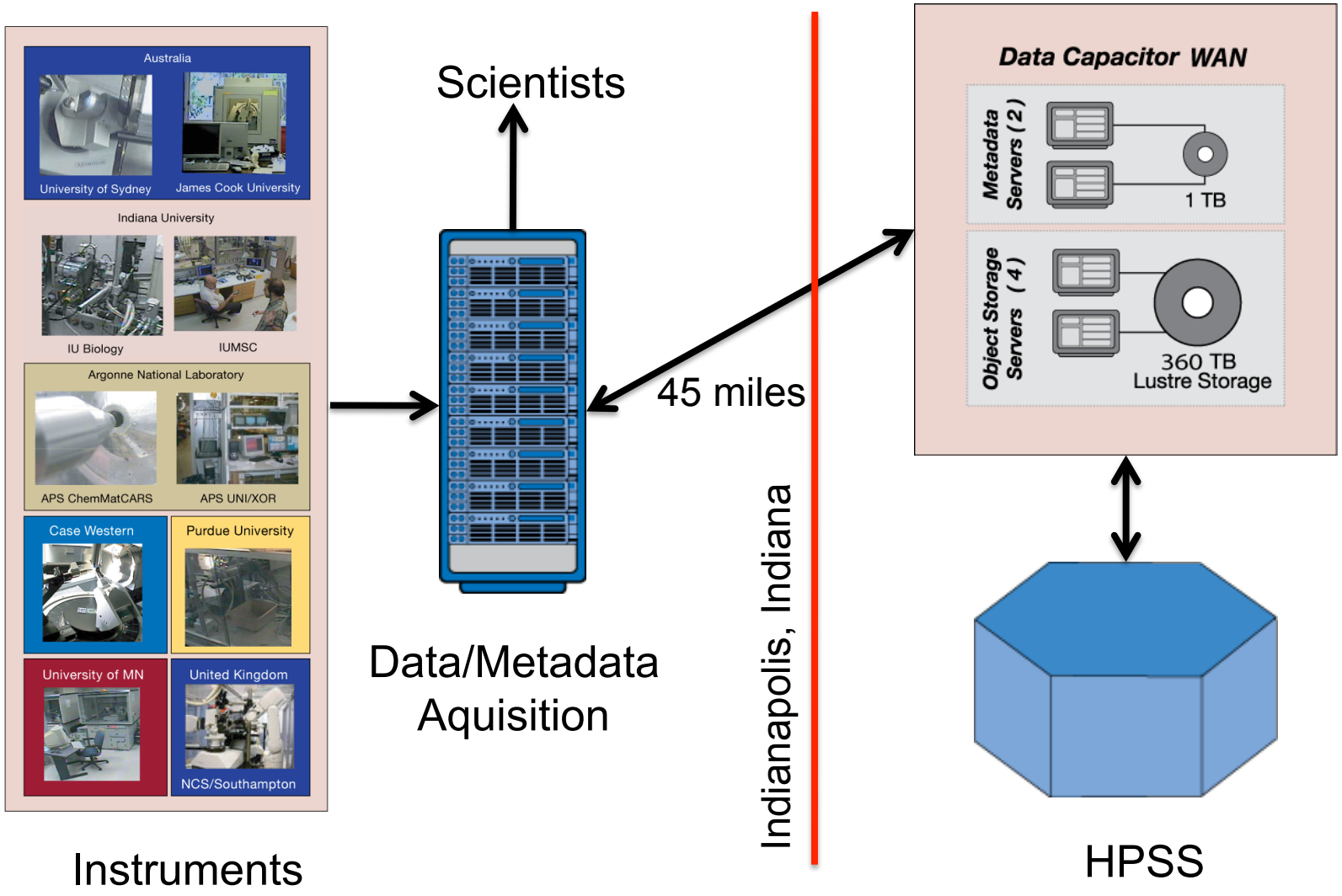
CIMA



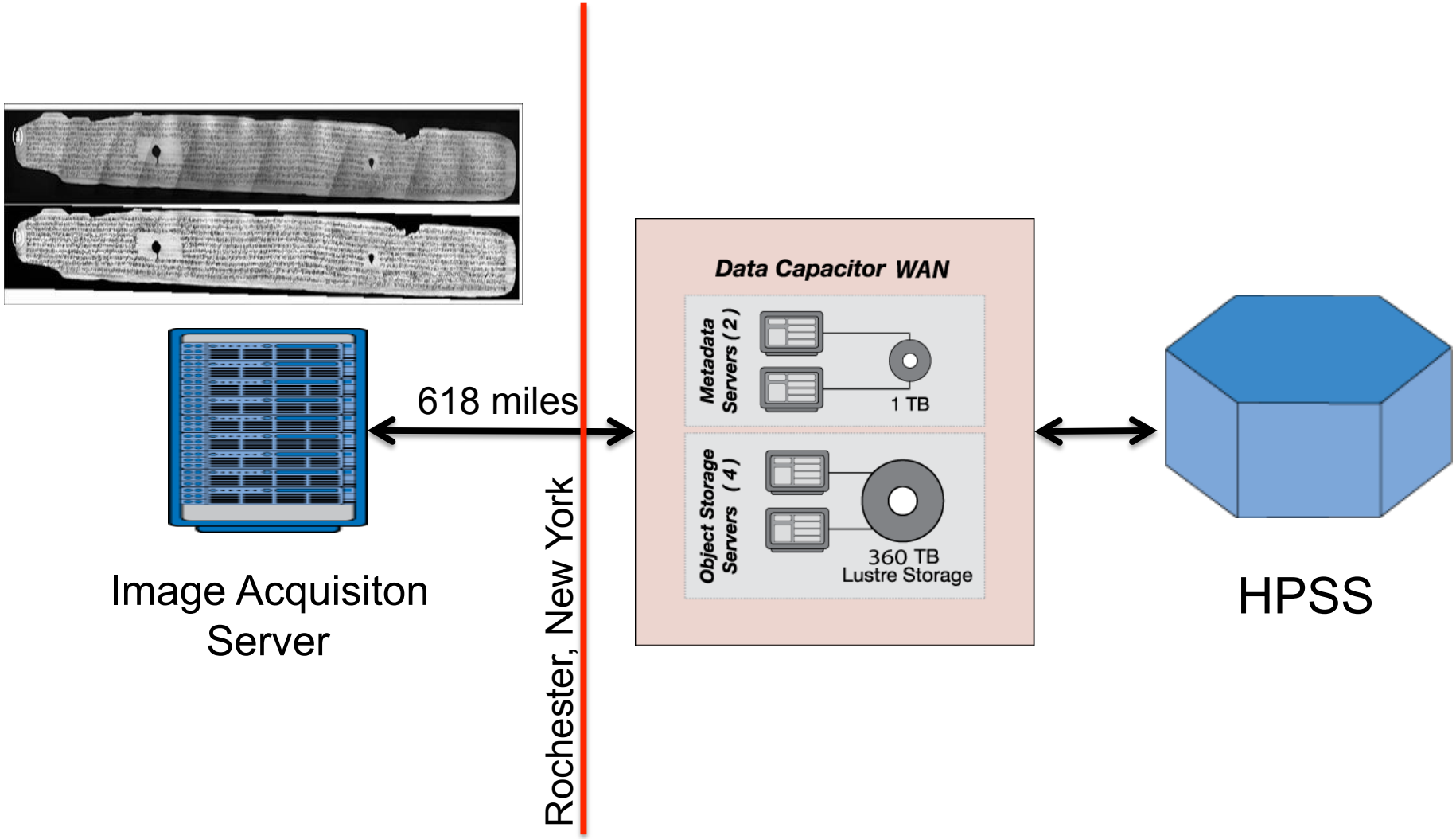
Time	2005-11-04 23:05:19 (UTC)	Instrument bay temperature (C)	19.50
Current data frame	053141.001	X-ray coolant water in (C)	16.40
Crystal temperature (C)	-148.20	X-ray coolant water out (C)	22.30
Instrument enclosure humidity (%)	45.60	CCD Chip Temperature (C)	-55.47
Instrument enclosure temperature (C)	22.70	Frame #	1
Instrument bay humidity	55.60		

X-ray diffractometer output

CIMA cont'd

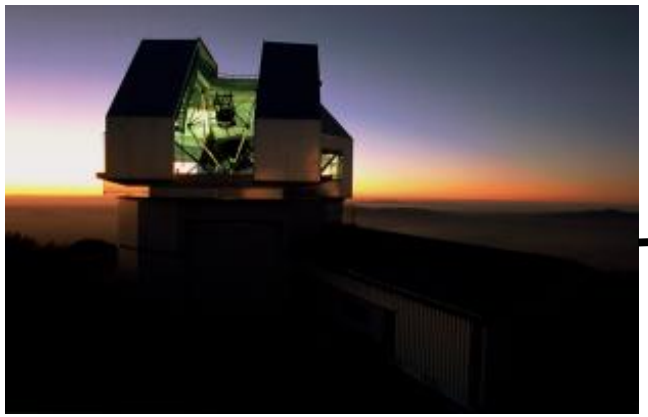

Australia
University of Sydney
James Cook University
Indiana University
IU Biology
IUMSC
Argonne National Laboratory
APS ChemMatCARS
APS UNI/XOR
Case Western
Purdue University
University of MN
United Kingdom
NCS/Southampton
Instruments
Scientists
Data/Metadata
Aquisition
45 miles
Indianapolis, Indiana
Data Capacitor WAN
Metadata Servers (2)
1 TB
Object Storage Servers (4)
360 TB
Lustre Storage
HPSS

Preserving Sarvamoola Granthas



One Degree Imager (ODI)

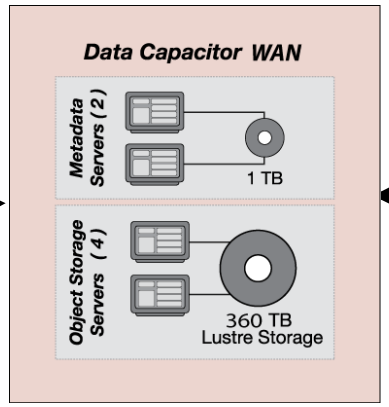
WIYN Telescope

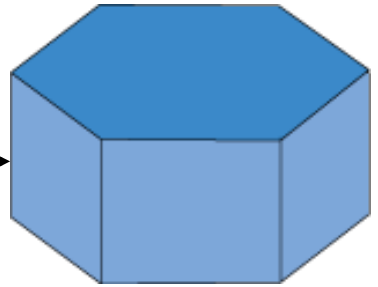


NOAO/AURA/NSF

1726 miles

Tucson, Arizona





HPSS

Ethnographic Video for Instruction and Analysis

EVIA

346 miles

Video Acquisiton Server

Ann Arbor, Michigan

Data Capacitor WAN

Metadata Servers (2)

1 TB

Object Storage Servers (4)

360 TB Lustre Storage

HPSS

Samba

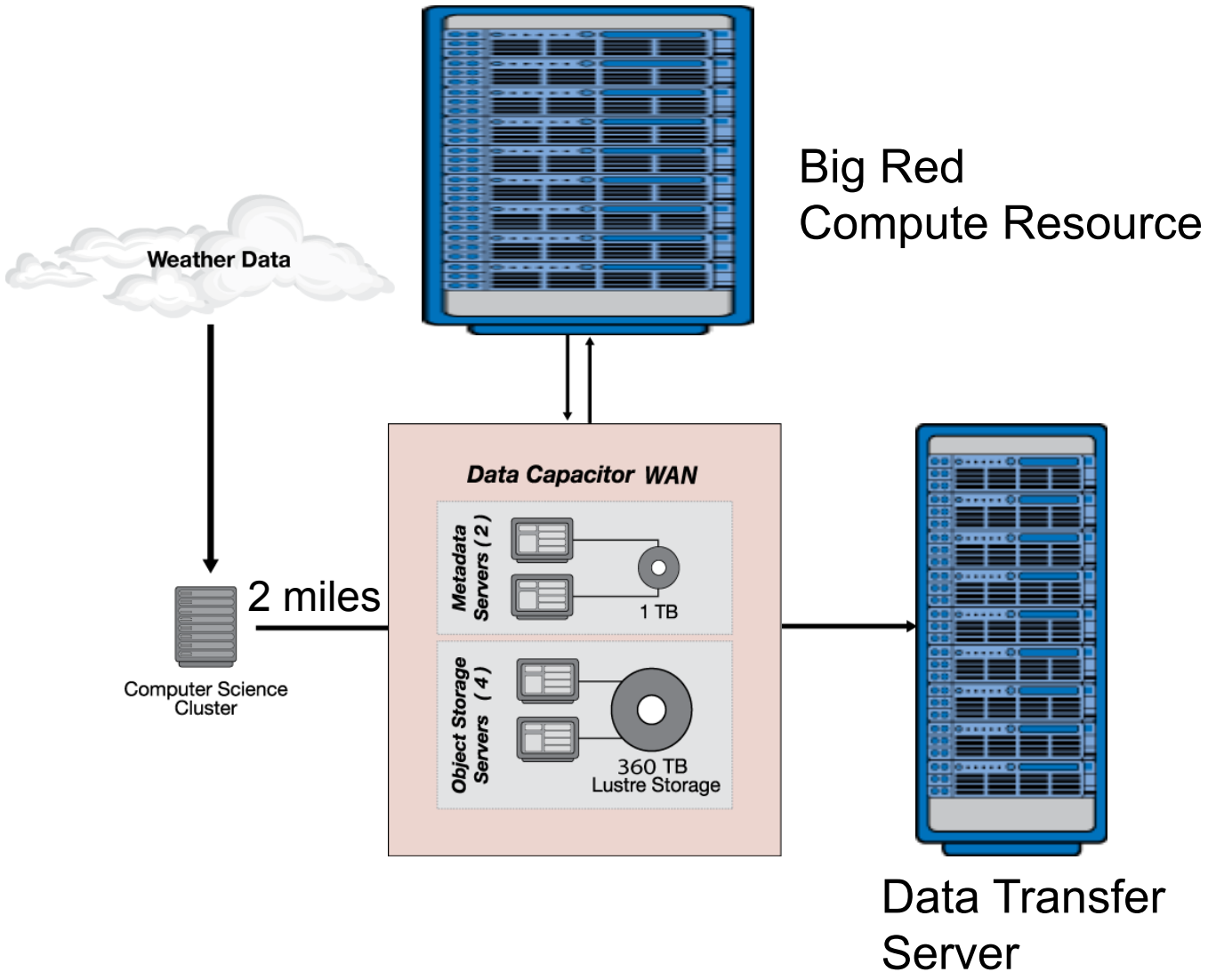
1 mile

Compression/Annotation Server

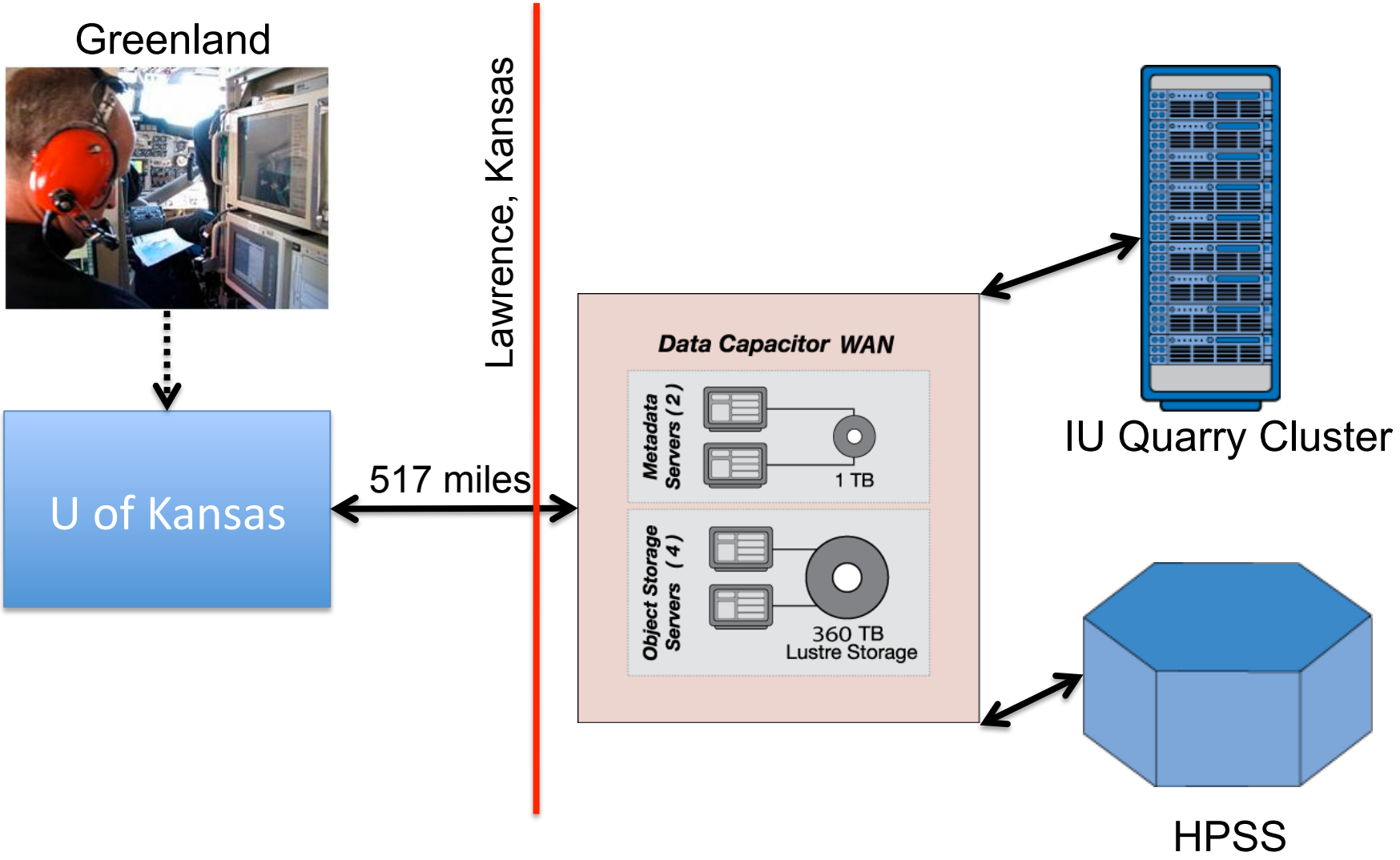
One Stop Resource Shopping

Linked Environments for Atmospheric Discovery

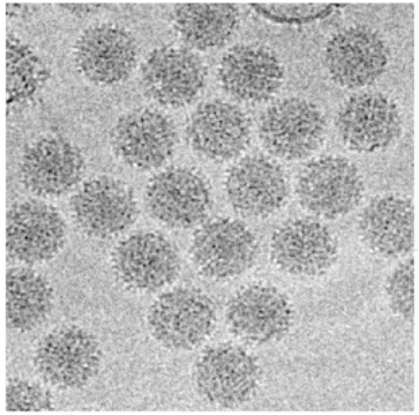
LEAD

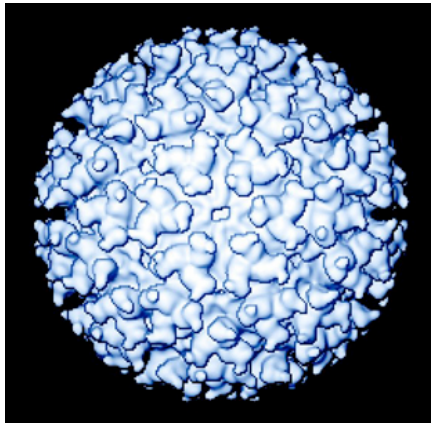


Center for the Remote Sensing of Ice Sheets (CReSIS) Workflow



CRYO Electron Microscopy





3 miles

Samba

Data Capacitor WAN

Metadata Servers (2)

1 TB

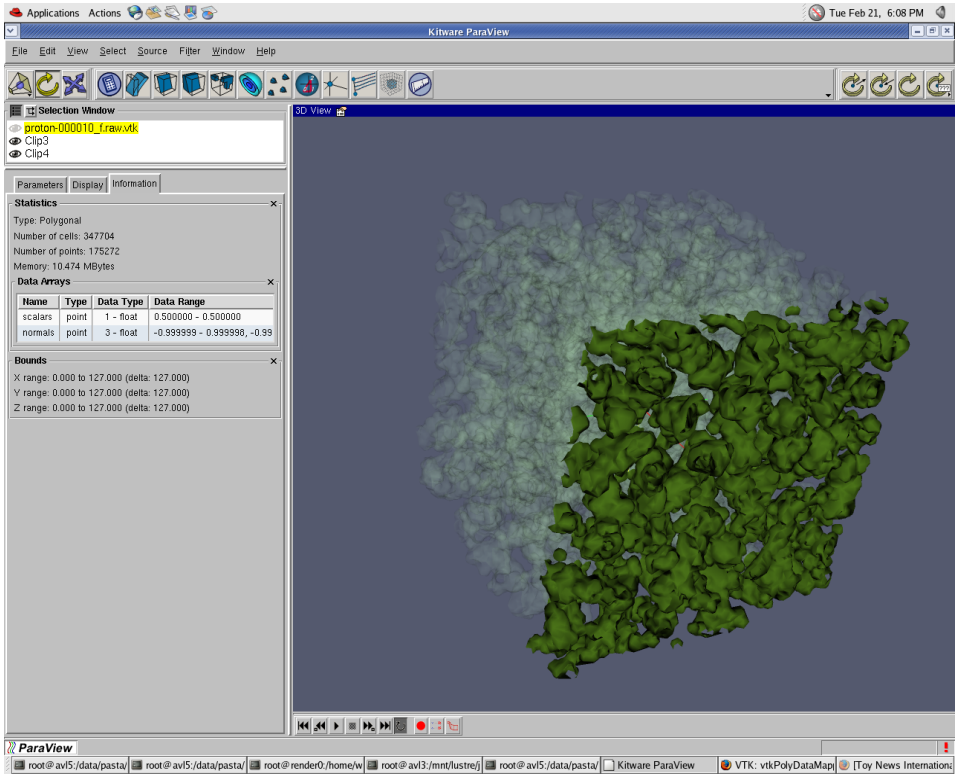
Object Storage Servers (4)

360 TB Lustre Storage

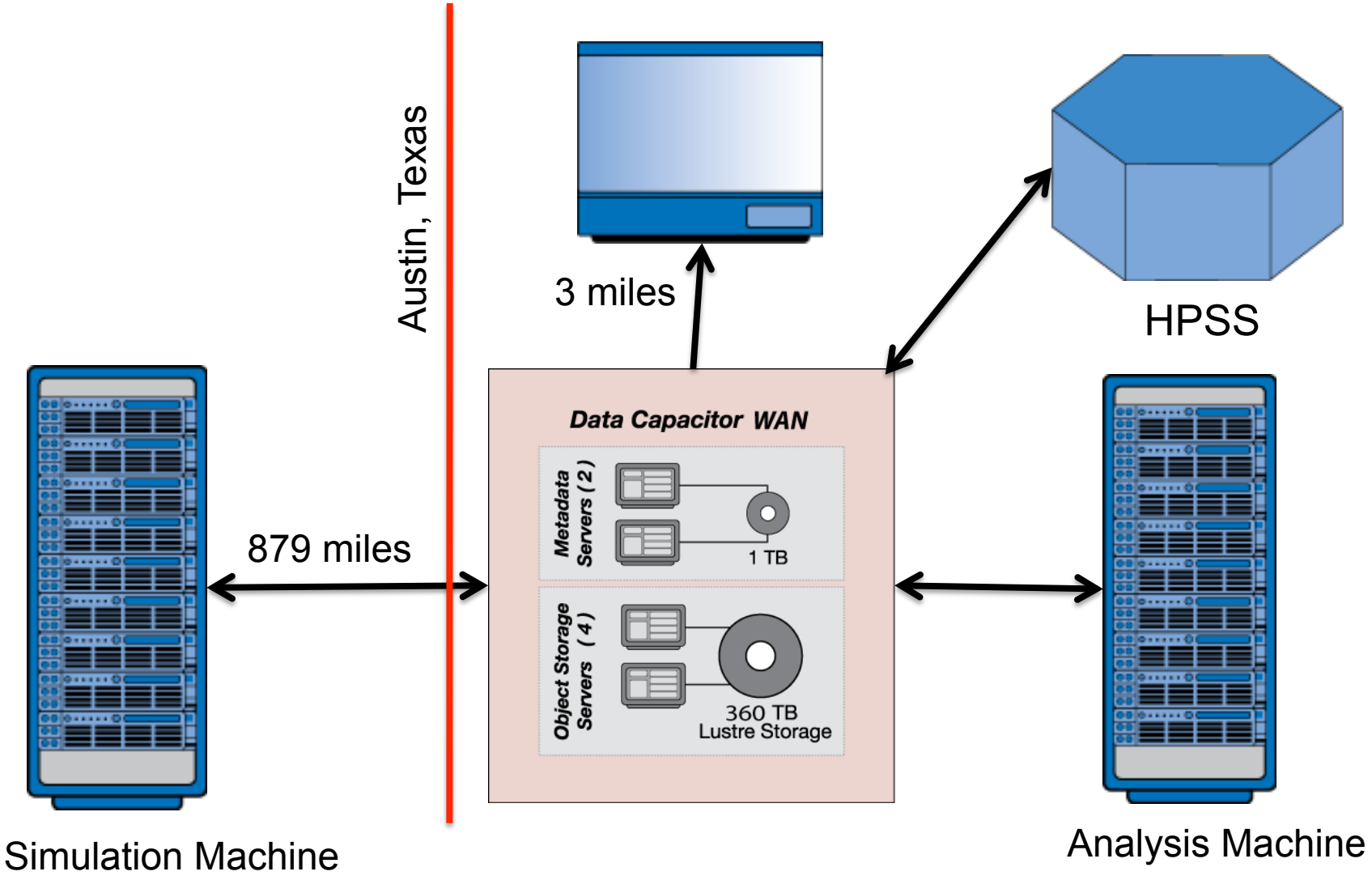
Big Red

HPSS

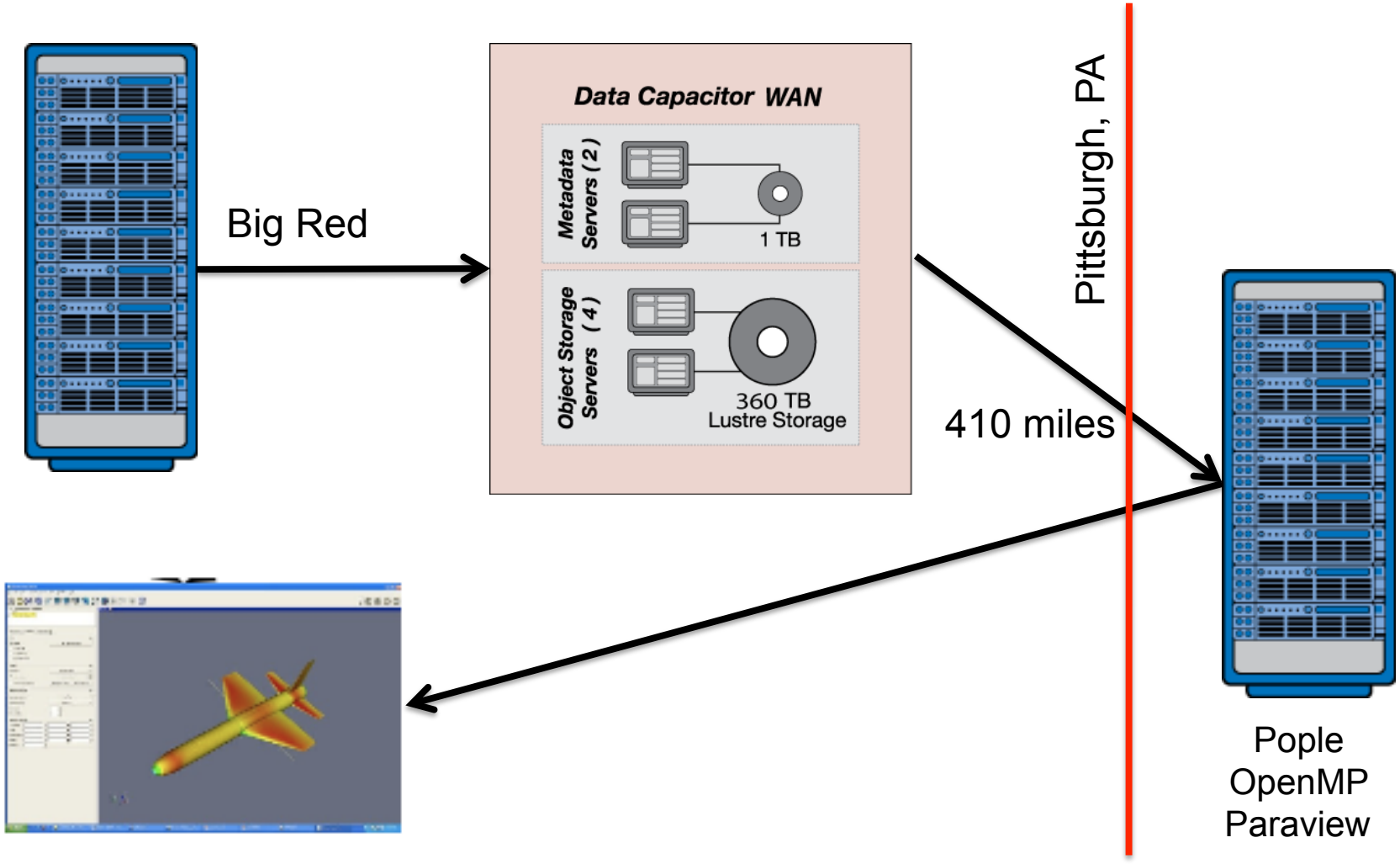
Equation of State Simulations and Plasma Pasta



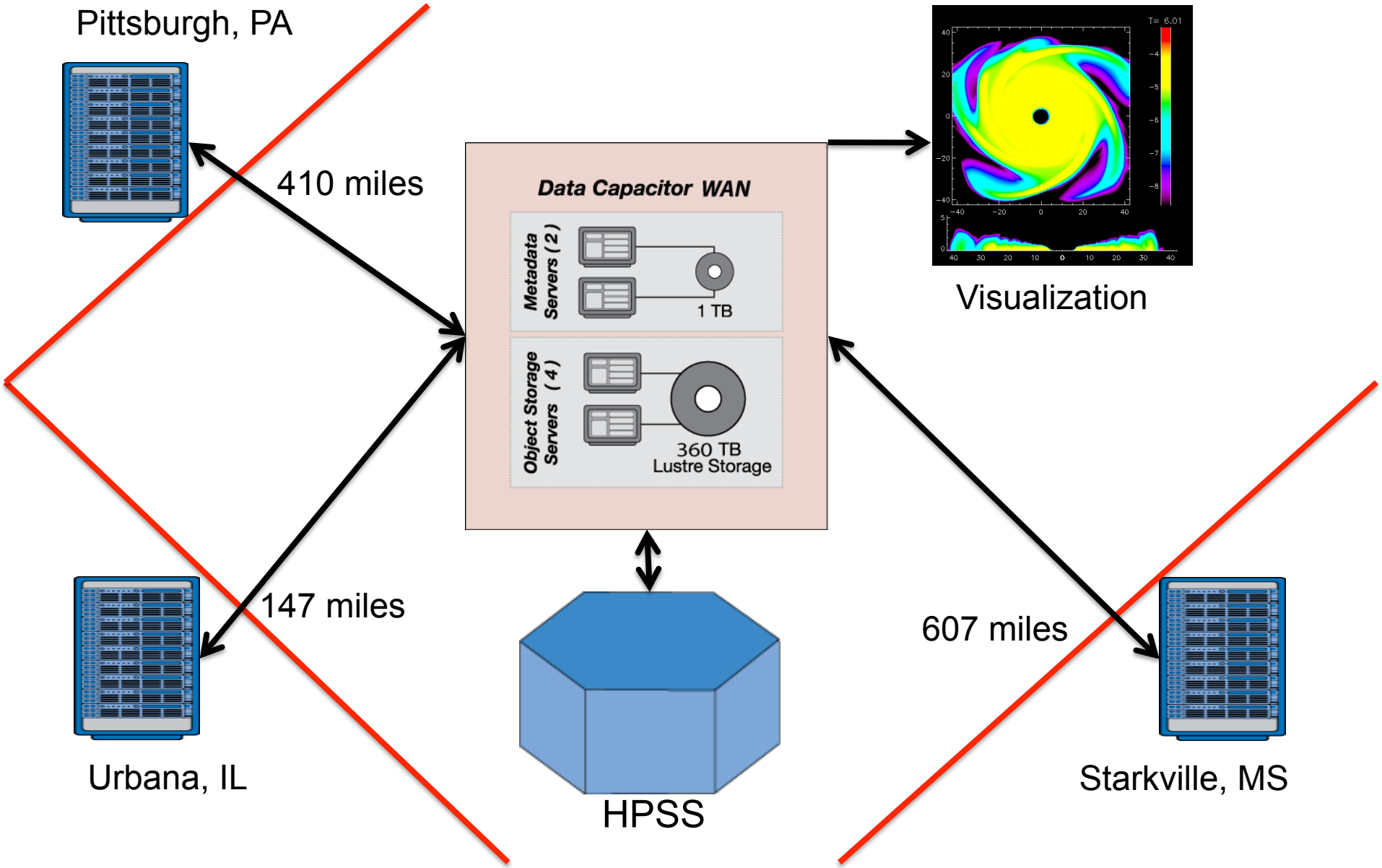
EOS and Plasma Pasta



Computational Fluid Dynamics



Gas Giant Planet Research



TeraGrid Future

- TeraGrid distributed Lustre WAN filesystem
  - 1.8.x
  - Distributed OSSs
    - NSF has funded servers to be deployed at five sites
    - IU, NCSA, NICS, PSC, TACC
  - IU's UID mapping code
  - PSC will run MDS
  - Will move to 2.0 code in the future

Non TeraGrid Exploration

- Dresden
  - ZIH (Technische Universitaet Dresden)
- Denmark
  - Risø – National Laboratory for Sustainable Energy
- Finland
  - Metsähovi Radio Observatory

Many Thanks

- Josh Walgenbach, Justin Miller, Nathan Heald, James McGookey, Resat Payli, Suresh Marru, Robert Henschel, Scott Michael, Tom Johnson, Chuck Horowitz, Don Berry, Scott, Teige, David Morgan, Matt Link (IU)
- Kit Westneat (DDN)
- Oracle support and engineering
- Michael Kluge, Guido Juckeland, Matthias Mueller (ZIH,Dresden)
- Thorbjorn Axellson (CReSIS)
- Greg Pike and ORNL
- Doug Balog, Josephine Palencia, and PSC
- Trey Breckenridge, Roger Smith, Joey Jones (Mississippi State University)

Support for this work provided by the National Science Foundation is gratefully acknowledged and appreciated (CNS-0521433)

Thank you!

Questions?

ssimms@indiana.edu

dc-team-l@indiana.edu

http://datacapacitor.iu.edu