INDIANA UNIVERSITY

The Data Capacitor

Digital Library Brown Bag
February 22, 2006

Stephen Simms
Data Capacitor Project Manager
Research and Academic Computing
University Information Technology Services
ssimms@indiana.edu

INDIANA UNIVERSITY

Imagine

- You have been given hours of rare video
  - Jackson Pollack singing Jackson Browne
  - Jack Sprat eating fat
  - Ethel Merman singing Black Sabbath
- You feel compelled to:
  - Digitize
  - Examine
  - Edit
  - Catalog
  - Archive

INDIANA UNIVERSITY

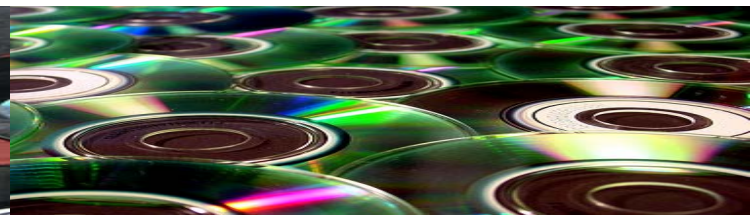
Wouldn’t It Be Nice...

- To have a temporary filespace fast enough to accept multiple streams of uncompressed video.
- To have a temporary filespace large enough to let you edit several 250 GB files at once.
- To have a temporary filespace which has a fast path to archival storage
- To have a system that could help you catalog and keep track of your extraordinary gift.

INDIANA UNIVERSITY

Data

- In the 21st Century everything is data
  - Nutritional data
  - Musical data
  - Patient data
- Raw material for
  - Research in the Arts and Humanities
  - Scientific advancement
  - Technological development



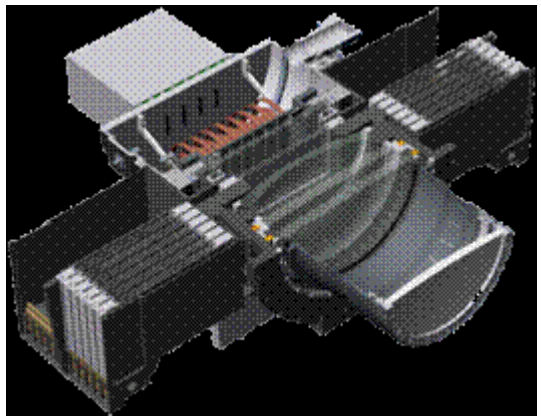
INDIANA UNIVERSITY

Better Technology = More Data

INDIANA UNIVERSITY

Better Telescopes

- Dr. Caty Pilachowski, Astronomer at IU
  - Investigating the evolution of stars
  - Project scientist for design of WIYN telescope
  - ODI promises 1 billion pixels/image by 2009



One Degree Imager

32k x 32k CCD

- LSST – Large Synoptic Survey Telescope
  - In the planning stages
  - Promise of 3 billion pixels/image by 2012

Better Televisions

- Ultra High Definition Television (UHDTV)
  - Japanese demonstration November 2005
  - 16 times more pixels than HDTV
  - Puts HD-DVD vs Blu-Ray debate in perspective

INDIANA UNIVERSITY

Data Challenges

- Where do I put it?
  - Storing the “data firehose”
    - Instrument
    - Simulation
    - Workflow
- How can I move it?
  - Pushing and pulling data
    - archive storage
    - computational and visualization resources
    - some other location across the network
- Can I find it again?
  - Insure data’s utility
    - through clear organization
    - documented provenance
    - provide a means for data discovery and make the data available

INDIANA UNIVERSITY

The Data Capacitor

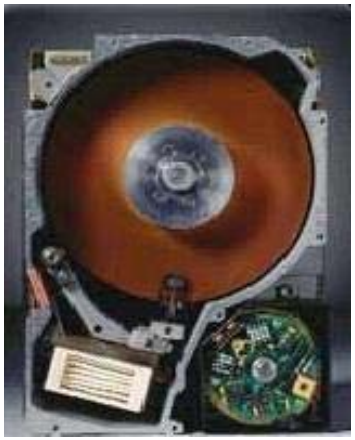
Last October, IU received $1.7 Million from the NSF to address data challenges in the form of a Major Research Instrumentation grant.

The Data Capacitor Project will provide:

- Hundreds of Terabytes of fast short term storage
- Servers for rapid data transfer
- Servers to provide web services

Why Capacitor?

- Capacitor
  - Provides transient storage of electrons
  - Absorbs and evens out peaks in electron flow
  - Provides fast discharge of electrons



- Data Capacitor
  - Provides transient storage of data
  - Absorbs and evens out peaks in data flow
  - Provides fast discharge of data

The Data Firehose

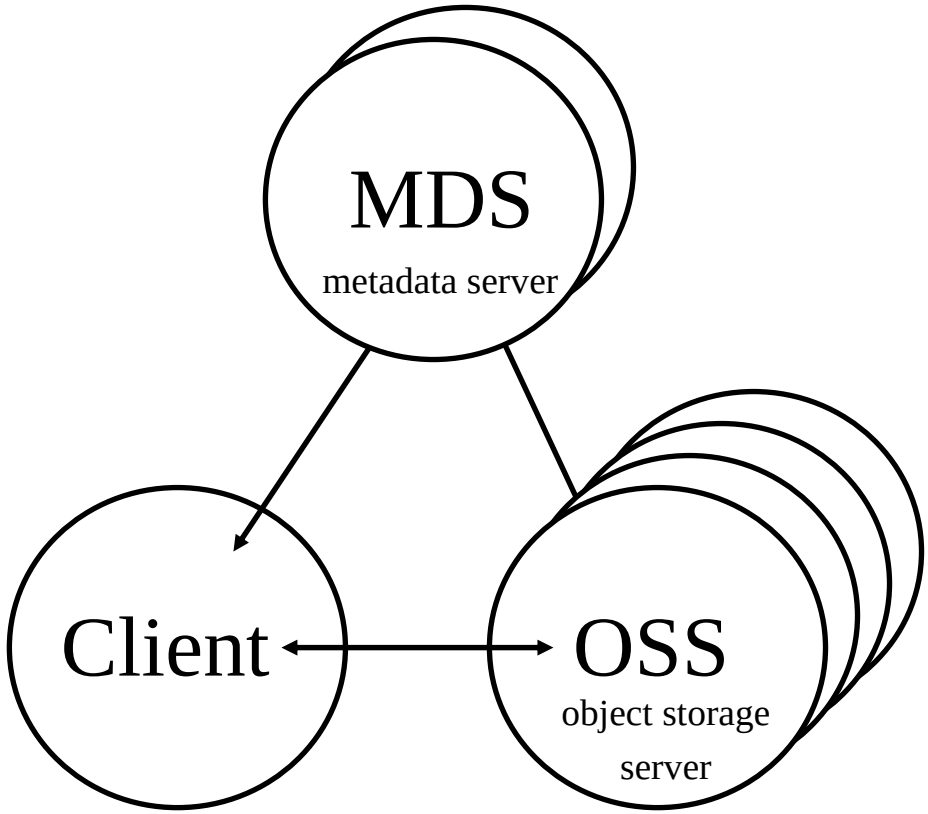
- Data in the network as uncompressible fluid
  - Network saturated?
  - Workflow clogged?

- Lots of fast storage as temporary reservoir
  - Quickly capture instrument or simulation data
  - Stage workflow data waiting for resources
  - Compress data before transfer to archive storage

INDIANA UNIVERSITY

Where you put it: Lustre

Scalable Object Storage

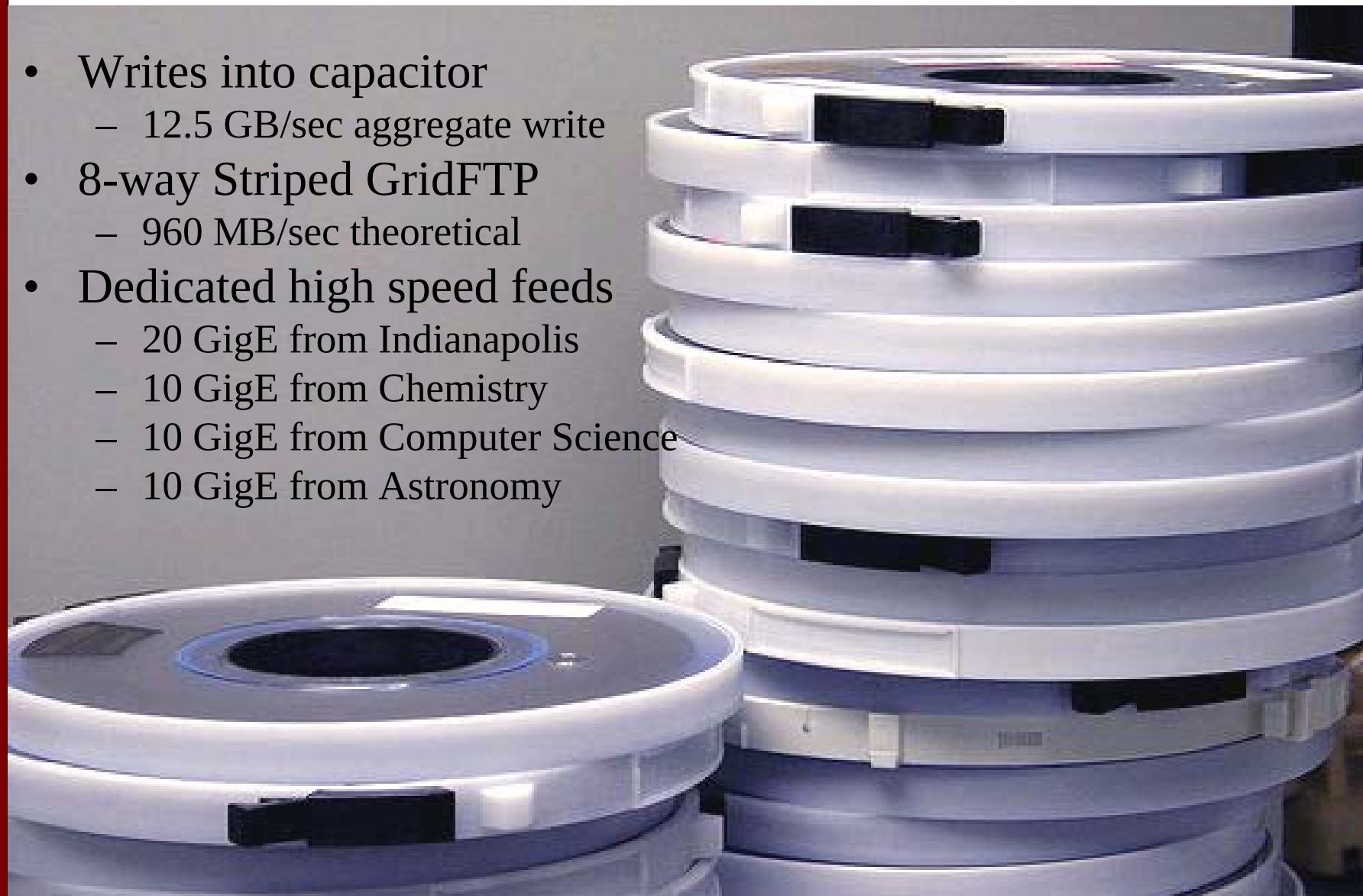


How you move it: Data Transfer Servers

- Not everyone will mount the Data Capacitor natively.
- Dedicated data servers will be available to local and remote users and will help ease the data transfer burden usually shouldered by:
  - Cluster head nodes
  - compute nodes
  - Gatekeepers
- We intend to support
  - NFS
  - CIFS
  - GridFTP
  - Striped GridFTP
  - Possibly pCIFS

INDIANA UNIVERSITY

Now? How about now?



- Writes into capacitor
  - 12.5 GB/sec aggregate write
- 8-way Striped GridFTP
  - 960 MB/sec theoretical
- Dedicated high speed feeds
  - 20 GigE from Indianapolis
  - 10 GigE from Chemistry
  - 10 GigE from Computer Science
  - 10 GigE from Astronomy


INDIANA UNIVERSITY
More Than Storage
I *know* that file is in here somewhere...

INDIANA UNIVERSITY

Data and Metadata Management

- Redundant servers running web services
  - -Data pre-processing
  - -Metadata acquisition
  - -Data discovery

INDIANA UNIVERSITY

Current Status

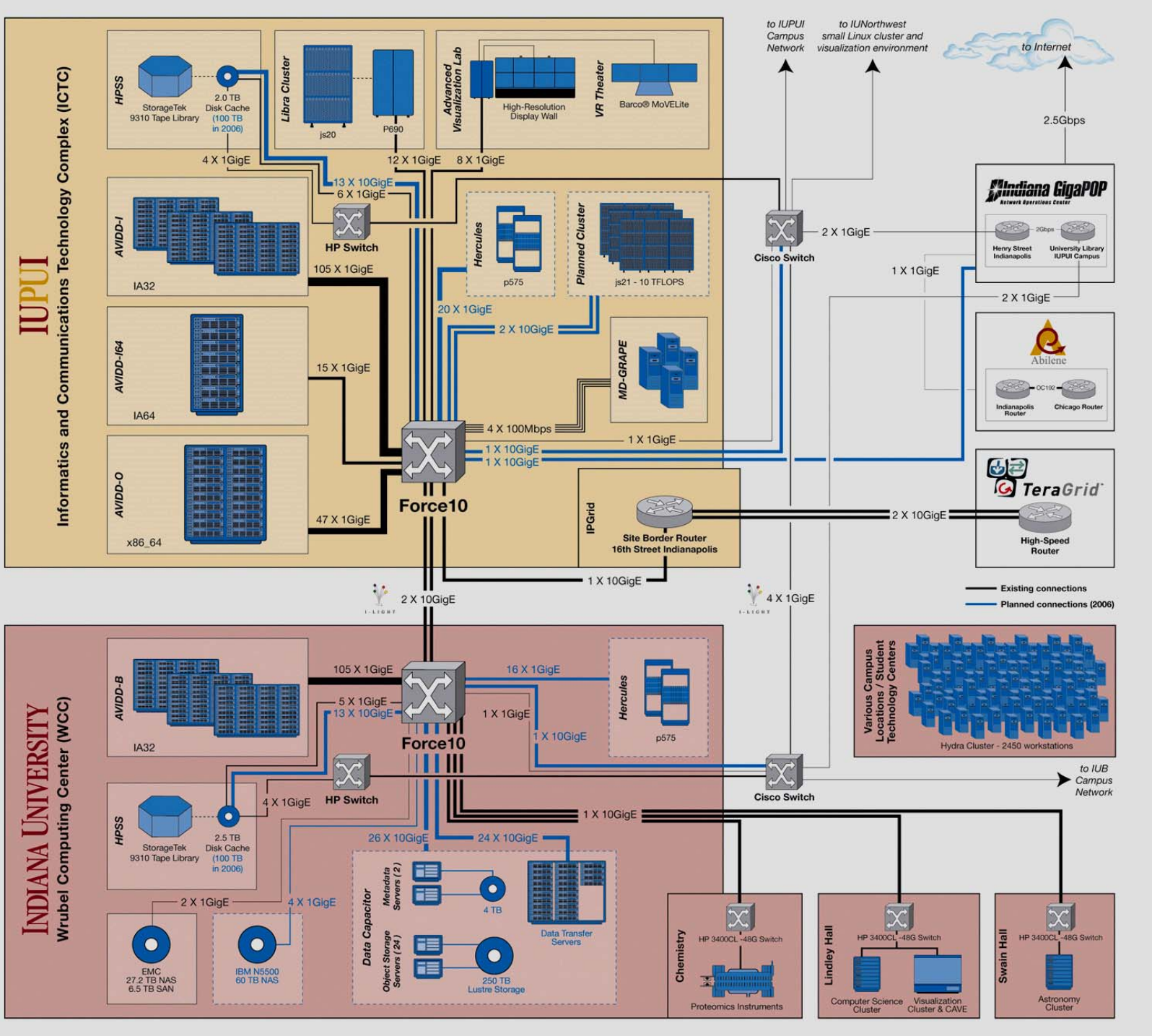
- Data Capacitor Prototype installed and deployed
- Request for vendor proposals issued.
- Vendor responses received and evaluated.
- Vendor interviews start February 24th.
- System to be installed and operational this Spring
- Work has already begun on local research projects:
  - Dr. David Clemmer - Proteomics
  - Dr. Chuck Horowitz - Plasma Pasta
  - Dr. John Huffman - X-ray Crystallography
- Work has already begun on the following national projects
  - Open Science Grid - Storage Element
  - TeraGrid - Lustre over WAN project

INDIANA UNIVERSITY

Maximize Usefulness Through One Stop Shopping

- Provide researchers with workflow opportunities that were previously difficult or impossible.
- Fast path to archive storage
- Shared filesystem for
  - Computation
  - Visualization
  - Instruments
  - Possibly desktops

INDIANA UNIVERSITY



INDIANA UNIVERSITY

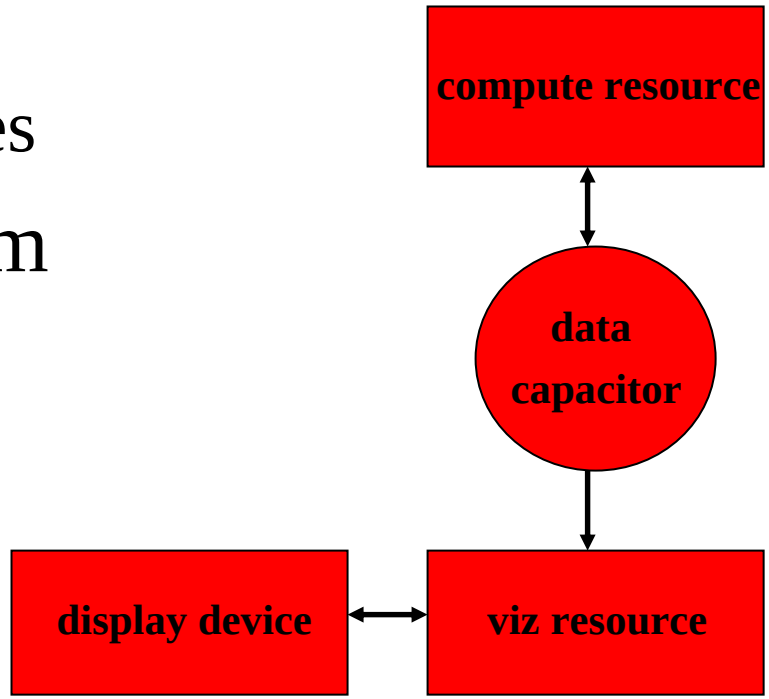
Dr. Chuck Horowitz

- IU Physicist
  - Interested in the behavior of neutron stars
  - Studying the behavior of nucleons under extreme pressure
  - Collaborating with UITS programmer Don Berry
  - Utilizing unique MDGRAPE-2 hardware

INDIANA UNIVERSITY

Plasma Pasta Workflow

- Particle interaction is simulated
  - Using specialized MDGRAPE-2 hardware
- Post processing
  - Creates VTK frames
- Visualization system
  - Ingests frames
  - Displays as movie



Kitware ParaView

File Edit View Select Source Filter Window Help

Selection Window

proton-000010_f.raw.vtk
Clip3
Clip4

Parameters Display Information

Statistics

Type: Polygonal
Number of cells: 347704
Number of points: 175272
Memory: 10.474 MBytes

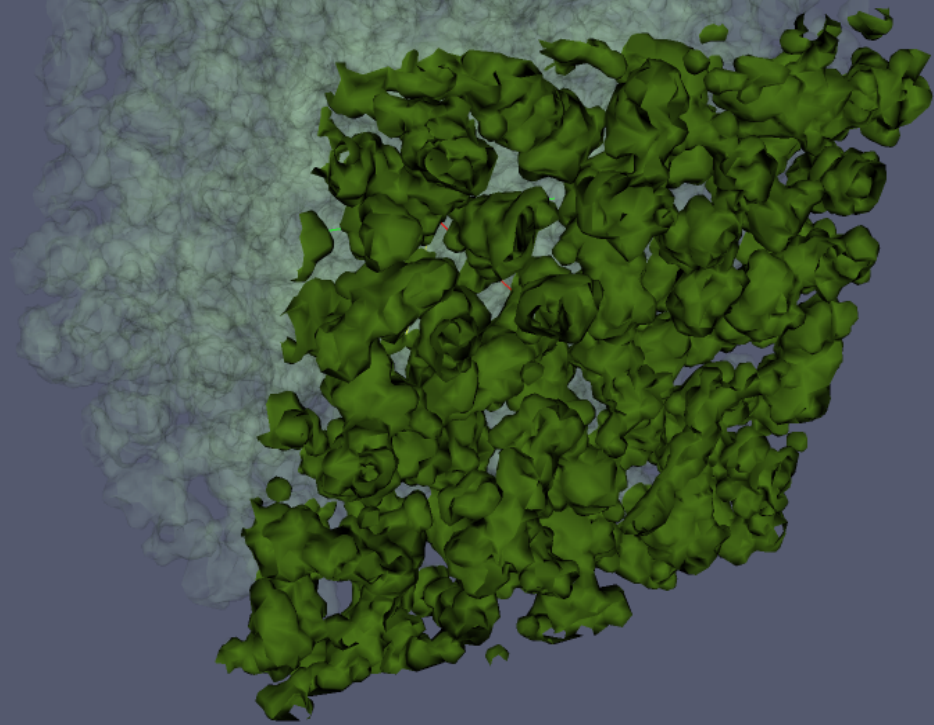
Data Arrays

Name	Type	Data Type	Data Range
scalars	point	1 - float	0.500000 - 0.500000
normals	point	3 - float	-0.999999 - 0.999998, -0.99

Bounds

X range: 0.000 to 127.000 (delta: 127.000)
Y range: 0.000 to 127.000 (delta: 127.000)
Z range: 0.000 to 127.000 (delta: 127.000)

3D View

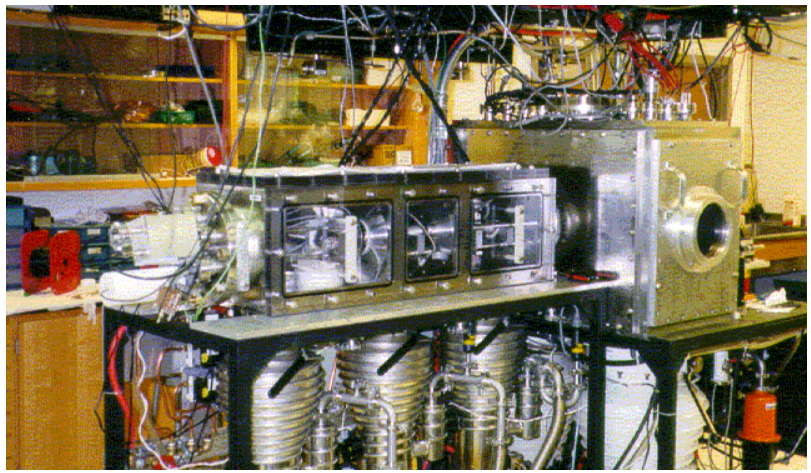


ParaView

INDIANA UNIVERSITY

Dr. David Clemmer

- Analytical Chemist at IU
  - Seeking to sequence biomolocules
  - Builds his own instruments
  - In 10 months “one run” has more than doubled in the amount of data produced

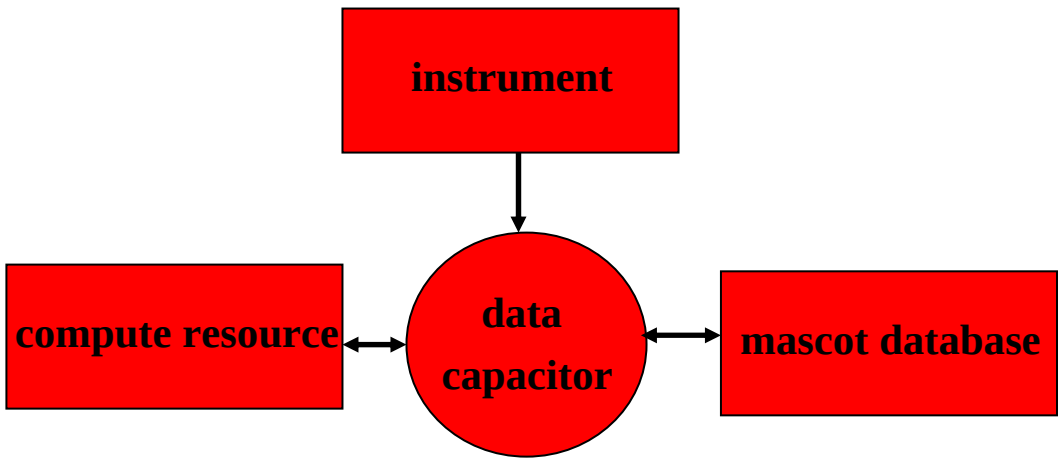


Ion mobility/time-of-flight instrument

INDIANA UNIVERSITY

Proteomics Workflow

- Data comes off of Ion Mobility / TOF instrument
- Data undergoes processing
- Database search using mascot

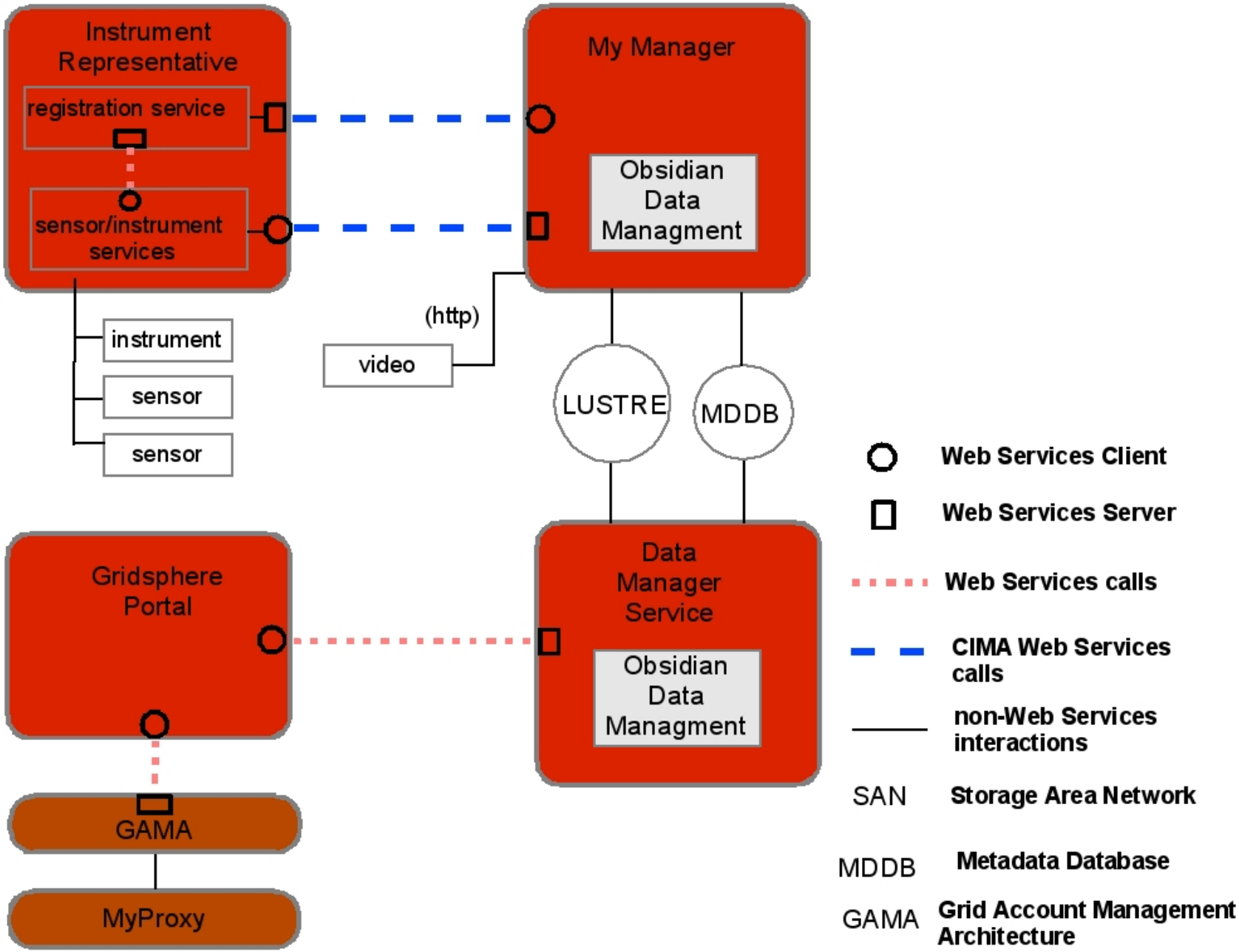


INDIANA UNIVERSITY

CIMA and X-ray Crystallography

- Dr. John Huffman
  - IU Chemist
  - Interested in creating large shared data repository
  - Interested in keeping metadata to “replay” an experiment
- Using Obsidian Architecture and Data Manager
  - Developed by Dr. Randy Bramley and Yu “Marie” Ma
  - Obsidian data architecture also in use:
    - astronomy
    - clinical radiation therapy
    - bioinformatics

INDIANA UNIVERSITY



INDIANA UNIVERSITY

Demonstration

- http://tinyurl.com/pccxf

INDIANA UNIVERSITY

Acknowledgements

Many Thanks to

Rick McMullen, John C. Huffman,

John N. Huffman, Kia Huffman,

Manny Plasencia

Research and Technical Services

Distributed Storage Support Group

INDIANA UNIVERSITY

Photo Acknowledgements

Many thanks to the following Creative Commons licensors at http://www.flickr.com:

Avlxyz
davesag
GFR
homemade
Jeffrey Gelens
Joseph Robertson
PPDIGITAL
Toni V

INDIANA UNIVERSITY

Thank you

Questions?