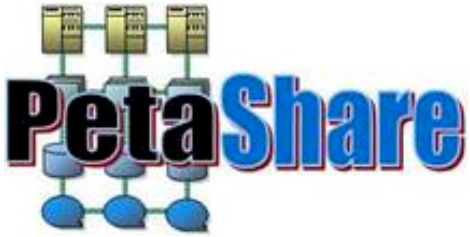


PetaShare:

Enabling Data Intensive Science

Tevfik Kosar

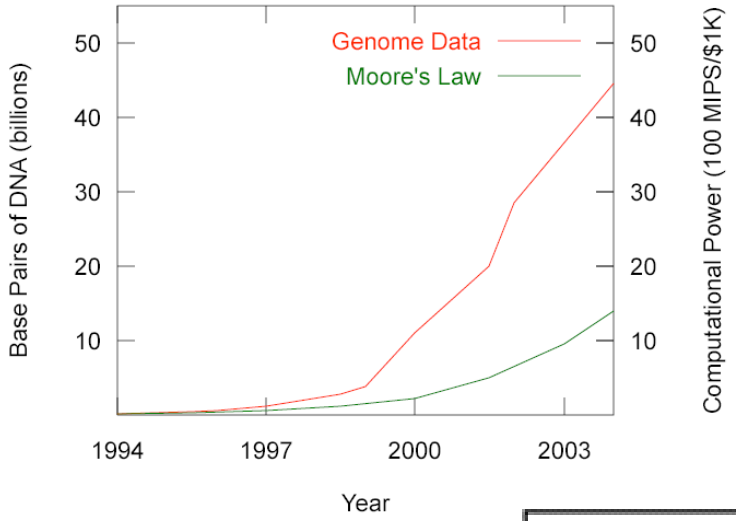
Center for Computation & Technology
Louisiana State University

June 25, 2007

AT LOUISIANA STATE UNIVERSITY

The Data Deluge

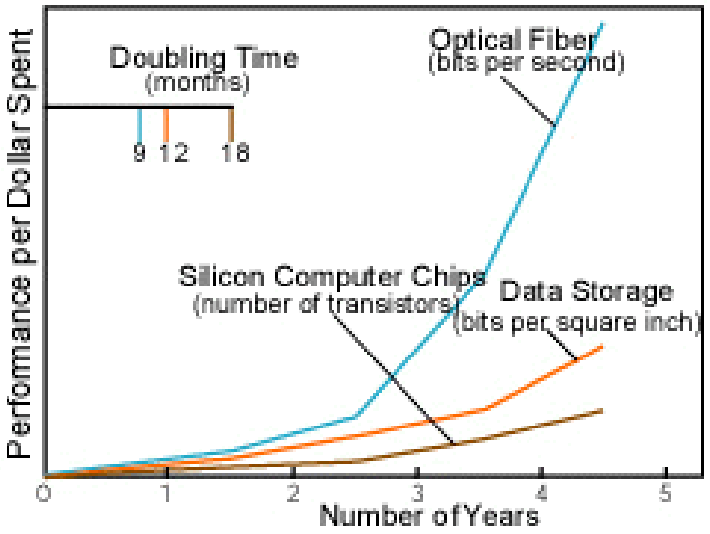
Scientific data outpaced Moore's Law!



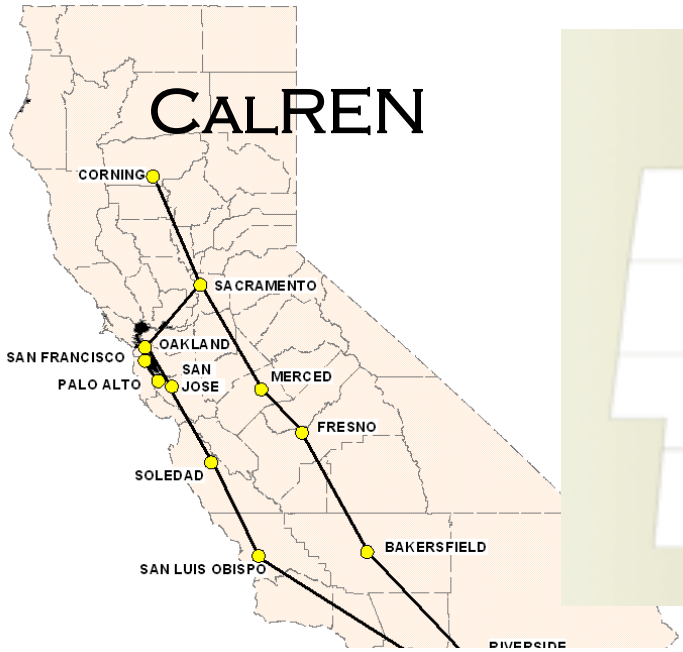
Demand for data in all areas of science!

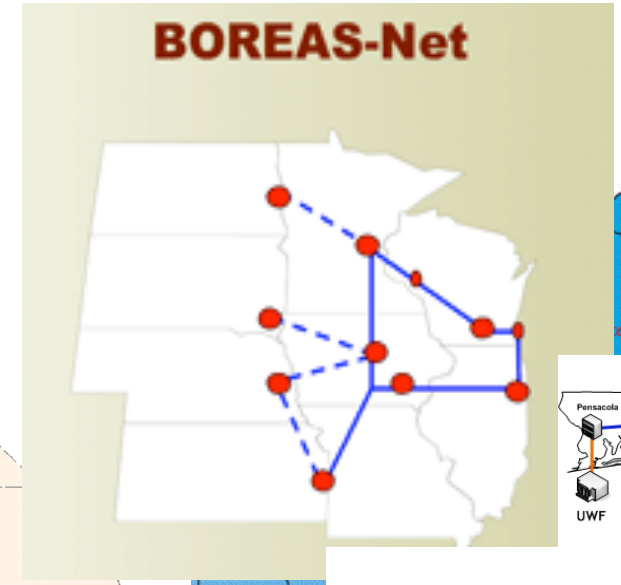
Application	Area	Data Volume
VISTA	Astronomy	100 TB/year
LIGO	Astrophysics	250 TB/year
WCER EVP	Educational Technology	500 TB/year
LSST	Astronomy	1000 TB/year
BLAST	Bioinformatics	1000 TB/year
ATLAS/CMS	High Energy Physics	5000 TB/year

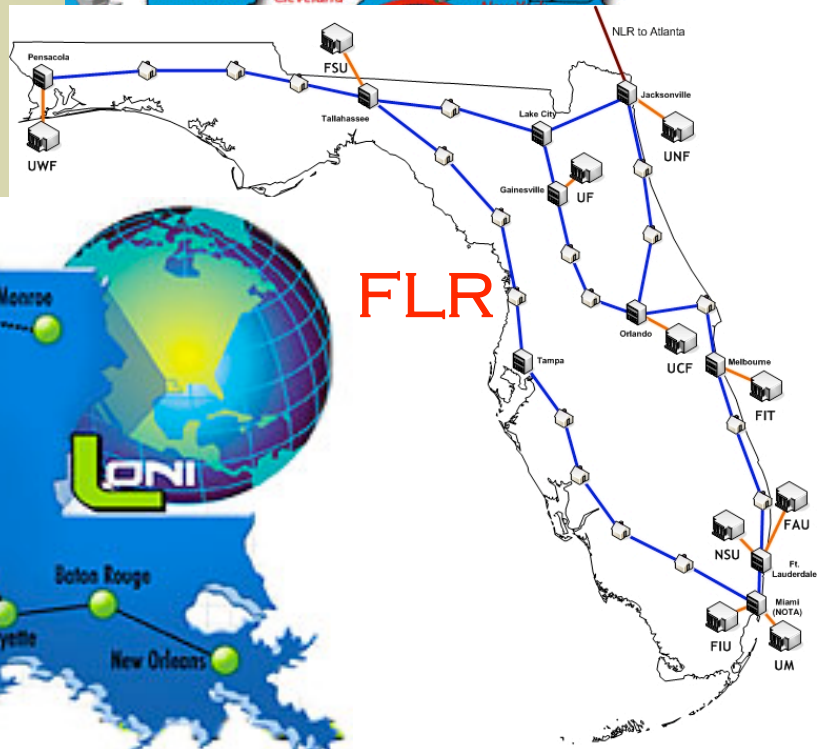
Computational Power (100 MIPS/$1K)



The Lambda Blast







DONE?..

- Each state is getting
  - Fast optical networks
  - Powerful computational grids

- But this solves only part of the problem!
- Researchers at these institutions still not be able to share and even process their own data

GRID today

The Leading Source for Global News and Information from the evolving Grid ecosystem,
including Grid, SOA, Virtualization, Storage, Networking and Service-Oriented IT

Applications:

NSF Funds LSU $1 Million for PetaShare Development

The National Science Foundation (NSF) recently funded Louisiana State University (LSU) $1 million for the development of PetaShare, which is seen as "a system might become an important testbed for future grids, and a leading site in next-generation petascale research."

The unbounded increase in the size of data generated by scientific applications necessitates collaboration and sharing among the nation's education and research institutions. Simply purchasing high-capacity, high-performance storage systems and adding them to the existing infrastructure of the collaborating institutions does not solve the underlying and highly challenging data handling problem. Scientists are compelled to spend a great deal of time and energy on solving basic data-handling issues, such as the physical location of data, how to access it, and/or how to move it to visualization and/or compute resources for further analysis.

LSU assistant professor Tevfik Kosar and his team aim to develop an innovative distributed data archival, analysis and visualization cyberinfrastructure for data intensive collaborative research, which they call PetaShare. PetaShare will enable transparent handling of underlying data sharing, archival and retrieval mechanisms, and will make data available to scientists for analysis and visualization on demand. PetaShare will enable scientists to focus on their primary research problem, assured that the underlying infrastructure will manage the low-level data handling issues.

Proposal Status | MAIN

Organization: Louisiana State University & Agricultural and Mechanical College

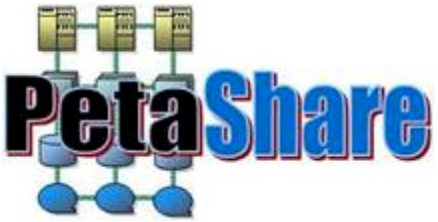
Review #1

Proposal Number: 0619843

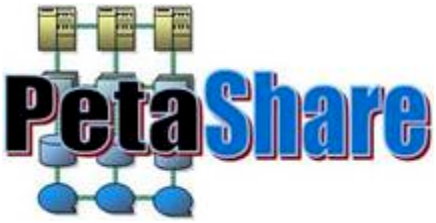
"..The PetaShare system might become an important testbed for future Grids, and a leading site in next generation Peta-scale research."

".. has a potential to serve as a catalyst for coalescing researchers who might otherwise not develop the incentive to collaborate. ."

This is the type of challenge that Grid-related researchers should be pursing, since Terabyte storage is already a commodity.

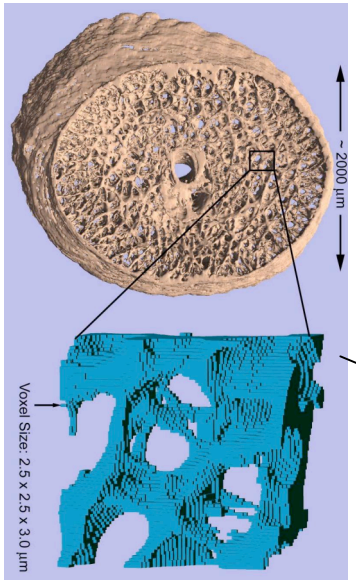


- Goal: enable domain scientists to focus on their primary research problem, assured that the underlying infrastructure will manage the low-level data handling issues.
- Novel approach: treat data storage resources and the tasks related to data access as first class entities just like computational resources and compute tasks.
- Key technologies being developed: data-aware storage systems, data-aware schedulers (i.e. Stork), and cross-domain meta-data scheme.
- Provides and additional 200TB disk, and 400TB tape storage

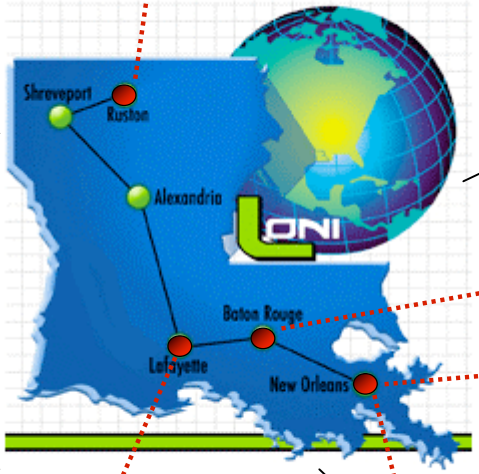


- PetaShare exploits 40 Gb/sec LONI connections between 5 LA institutions: LSU, LaTech, Tulane, ULL, and UNO.
- PetaShare links more than fifty senior researchers and two hundred graduate and undergraduate research students from ten different disciplines to perform multidisciplinary research.
- Application areas supported by PetaShare include coastal and environmental modeling, geospatial analysis, bioinformatics, medical imaging, fluid dynamics, petroleum engineering, numerical relativity, and high energy physics.

Participating institutions in the PetaShare project, connected through LONI. Sample research of the participating researchers pictured (i.e. biomechanics by Kodiyalam & Wischusen, tangible interaction by Ullmer, coastal studies by Walker, and molecular biology by Bishop).



LaTech
- High Energy Physics
- Biomedical Data Mining

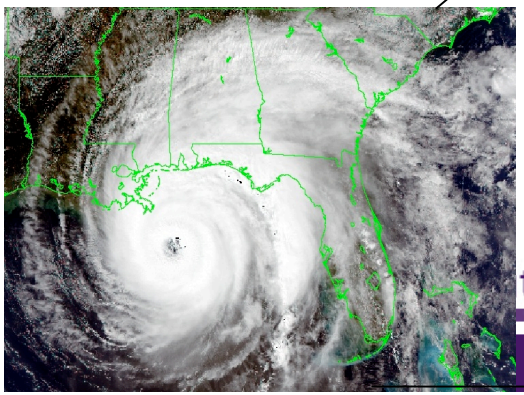


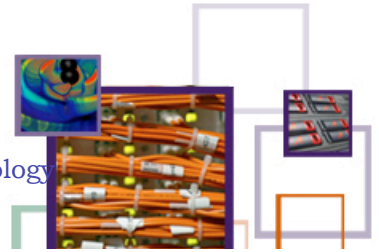
LSU
- Coastal Modeling
- Petroleum Engineering
- Computational Fluid Dynamics
- Synchrotron X-ray Microtomography

UNO
- Biophysics

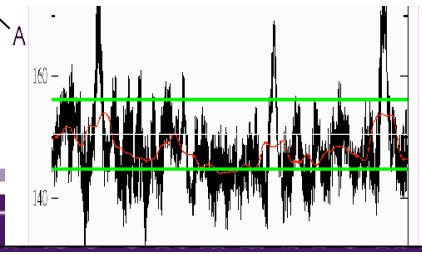
ULL
- Geology
- Petroleum Engineering

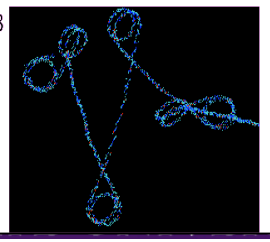
Tulane
- Molecular Biology
- Computational Cardiac Electrophysiology

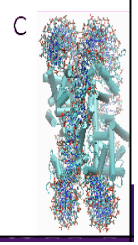




tation & Technology



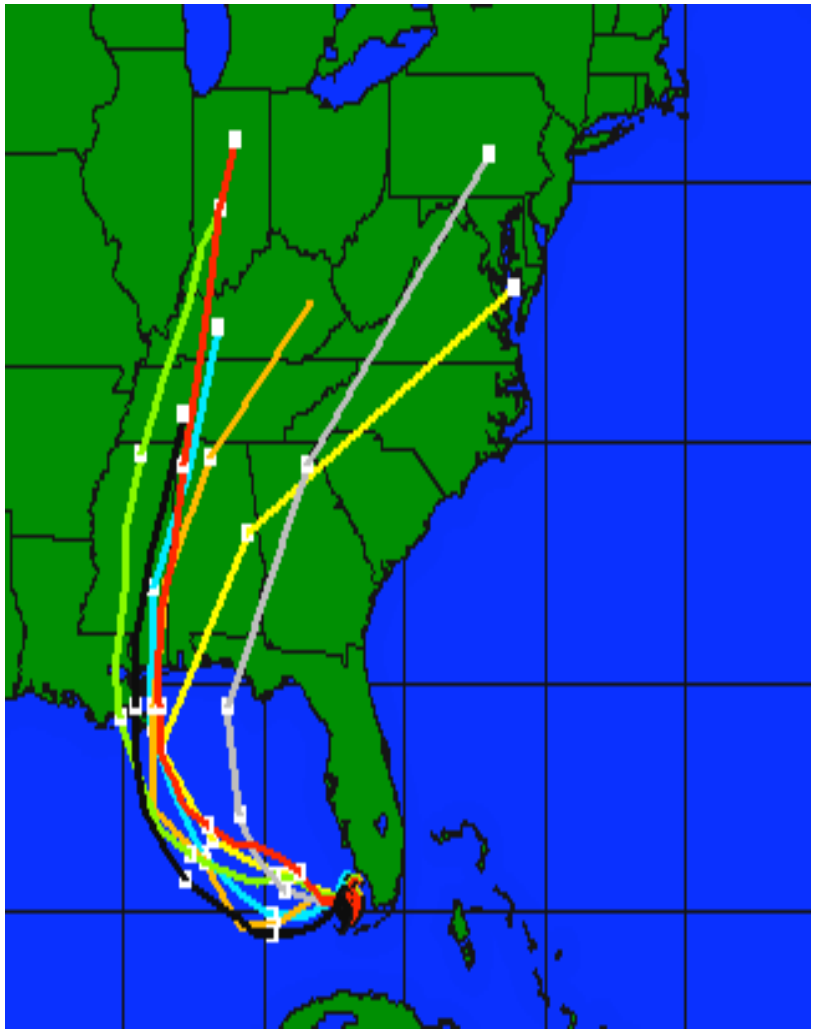




PetaShare Science Drivers

Coastal Studies

- Walker, Levitan, Mashriqui, Twilley (LSU)
- The Earth Scan Lab: with its three antennas, it captures 40GB of data from six satellites each day. (➔ 15 TB/year)
- Hurricane Center
  - Storm surge modeling, hurricane track prediction
- Wetland Biochemistry Institute
  - Coastal Ecosystem preservation
- SCOOP data archive



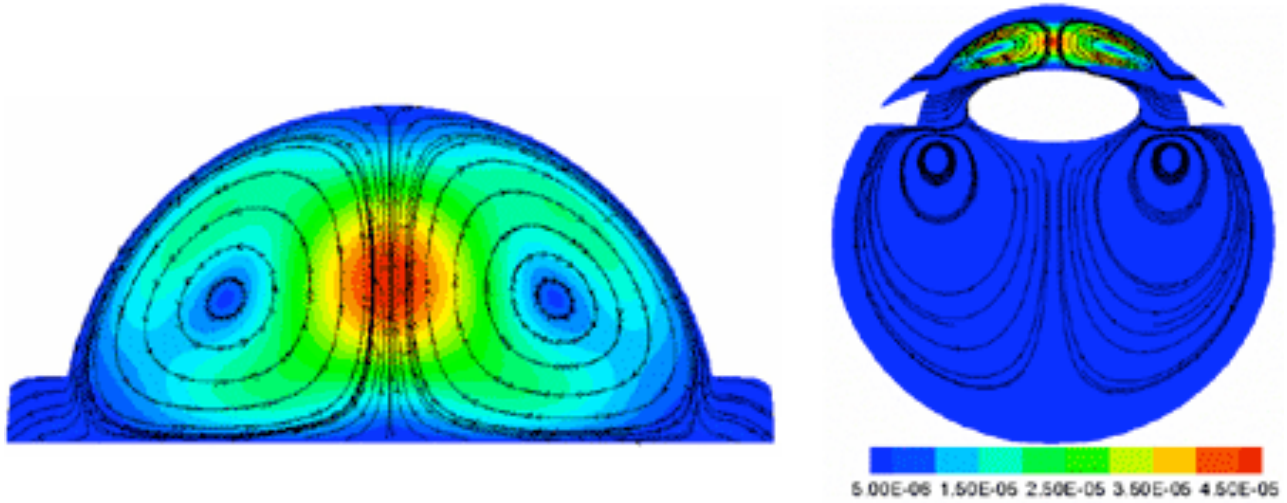
Petroleum Engineering

- White, Allen, Lei et al. (LSU, ULL, SUBR)
- UCoMS project – reservoir simulation and uncertainty analysis
- 26M simulations, each generating 50MB of data ➔ 1.3 PB of data total

- Drilling processing and real-time monitoring is data-intensive as well ➔ real-time visualization and analysis of TB's of streaming data

Computational Fluid Dynamics

- Acharya et al. (LSU)
- Focusing on simulation of turbulent flows including Direct Numerical Simulations (DNS), Large Eddy Simulations (LES), and Reynolds-Averaged Navier Stokes Simulations (RANS).
- In DNS, ~10,000 instances of flow field must be stored and analyzed, each instance may contain 150M discrete variables. Resulting data set ~ 10 TB.

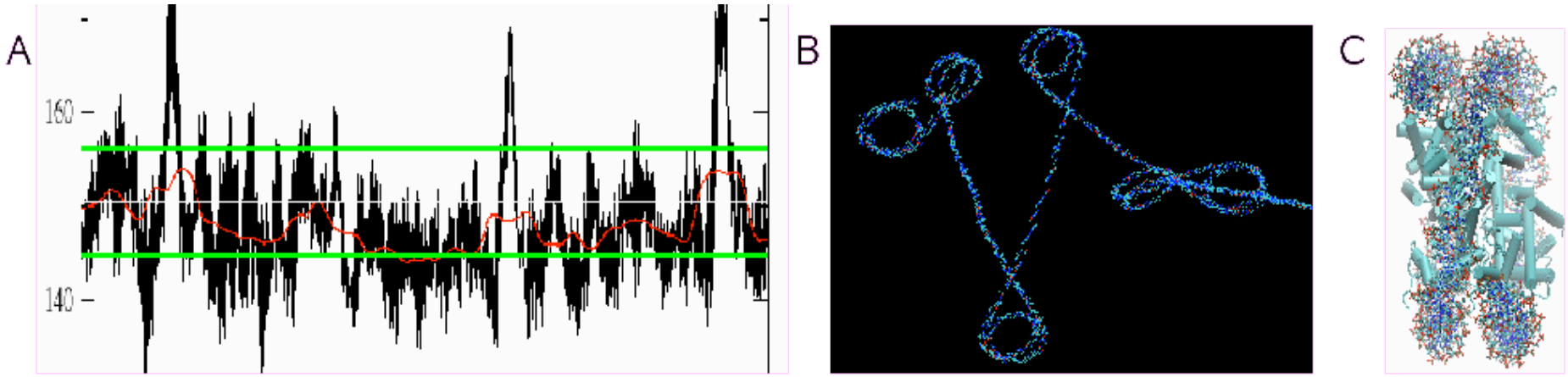


Molecular Biology

•Winters-Hilt (UNO)
•Biophysics and molecular biology – gene structure analysis
•Generates several terabytes of channel current measurements per month
•Generated data being sent to UC-Santa Cruz, Harvard and other groups

•Bishop (Tulane)
•Study the structure and dynamics of nucleosomes using all atom molecular dynamics simulations
•Each simulation requires 3 weeks of run time on a 24-node cluster, and 50-100 GB of storage ➔ 1-2 TB data per year

* Both access to the Genome database but separately!



And Others…

- Numerical Relativity
  - *Seidel et al (LSU)*
- High Energy Physics
  - *Greenwood, McNeil (LaTech, LSU)*
- Computational Cardiac Electrophysiology
  - *Trayanova (now at JHU)*
- Synchrotron X-ray Microtomography
  - *Wilson, Butler (LSU)*
- Bio Data Mining
  - *Dua (LaTech)*

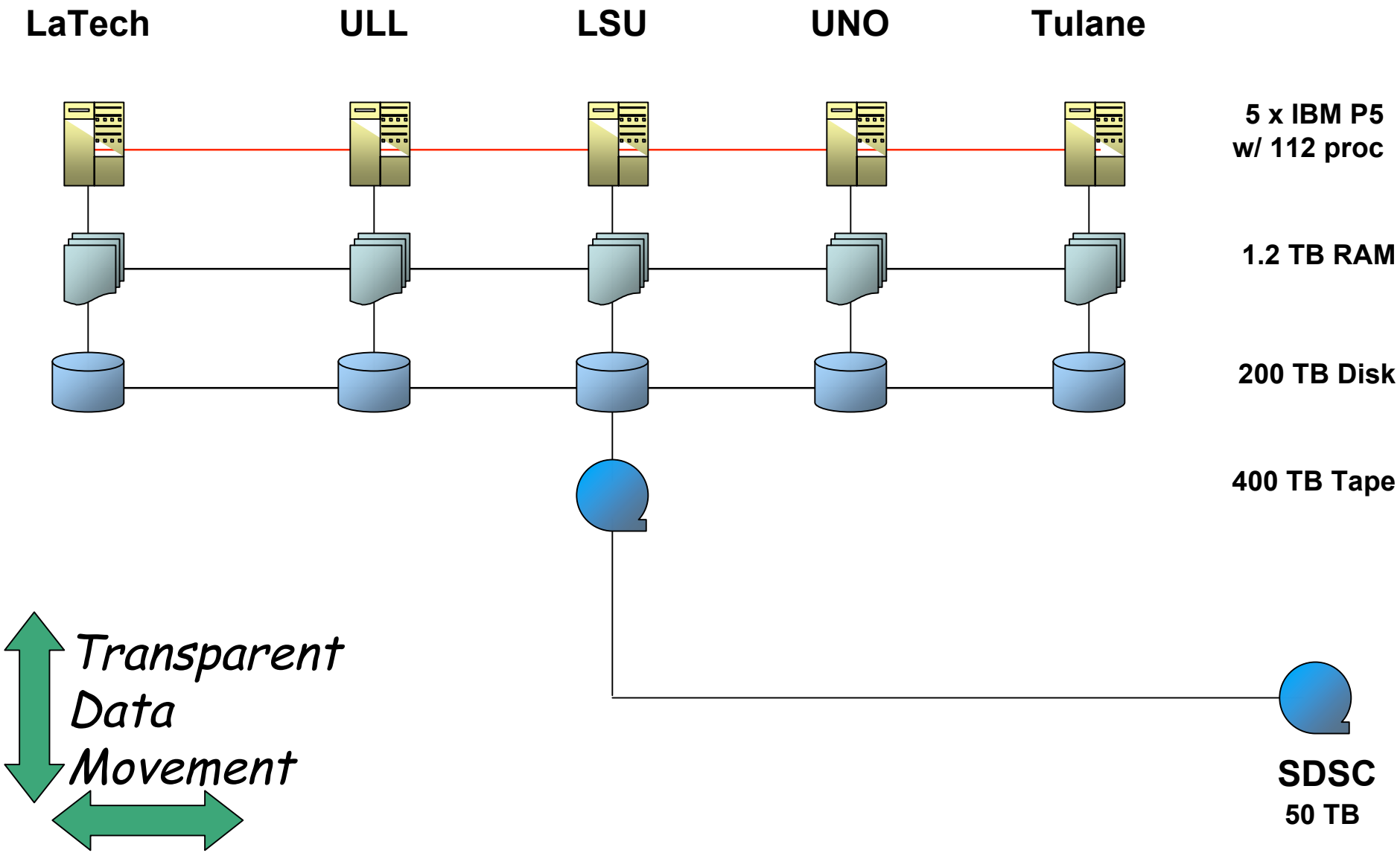
CS Research

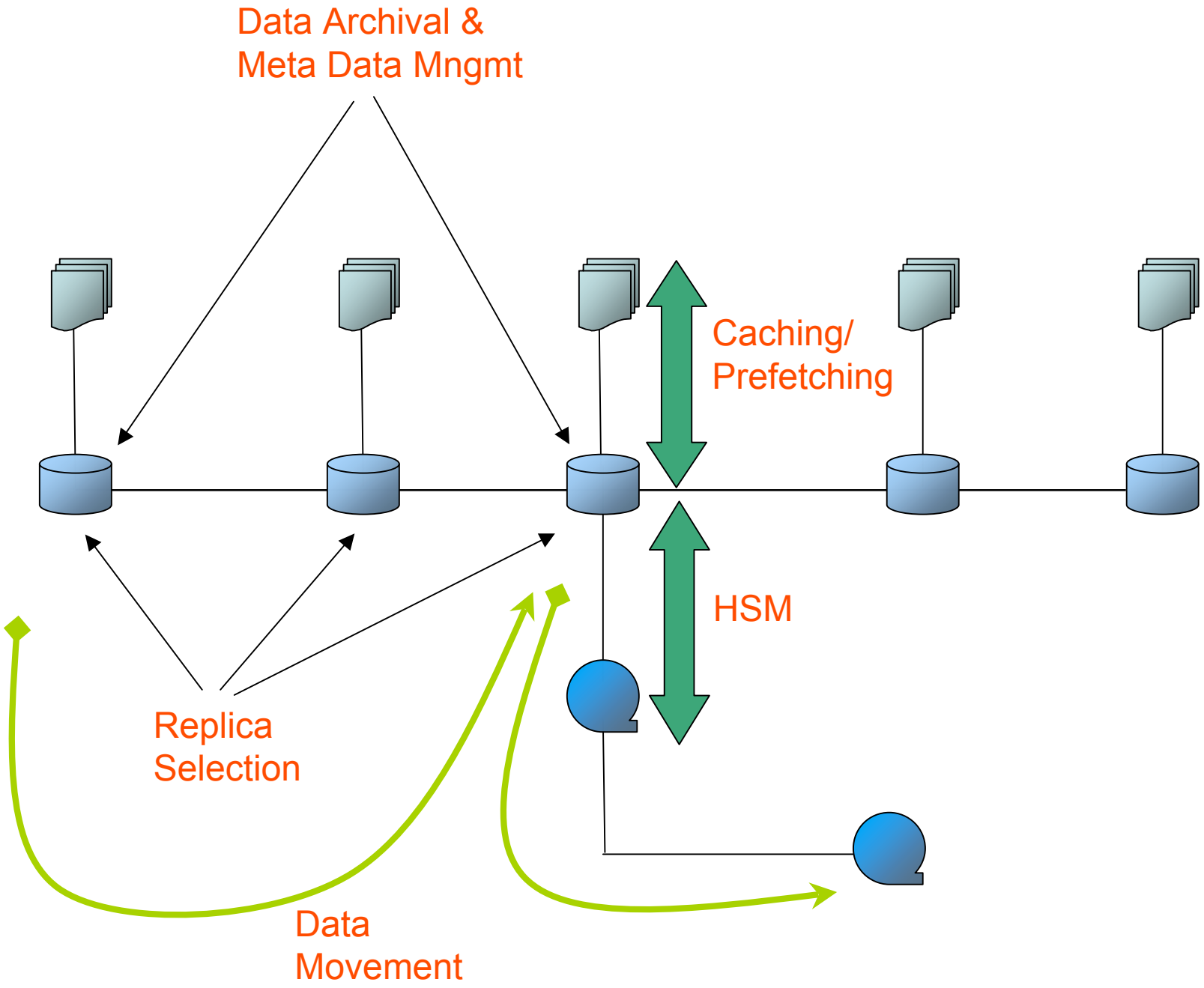
- Distributed Data Handling (Kosar)
- Grid Computing (Allen, Kosar)
- Visualization (Hutanu, Karki)
- Data Mining (Dua, Abdelguerfi)
- Database Systems (Triantaphyllou)

People involved with PetaShare Development and Usage

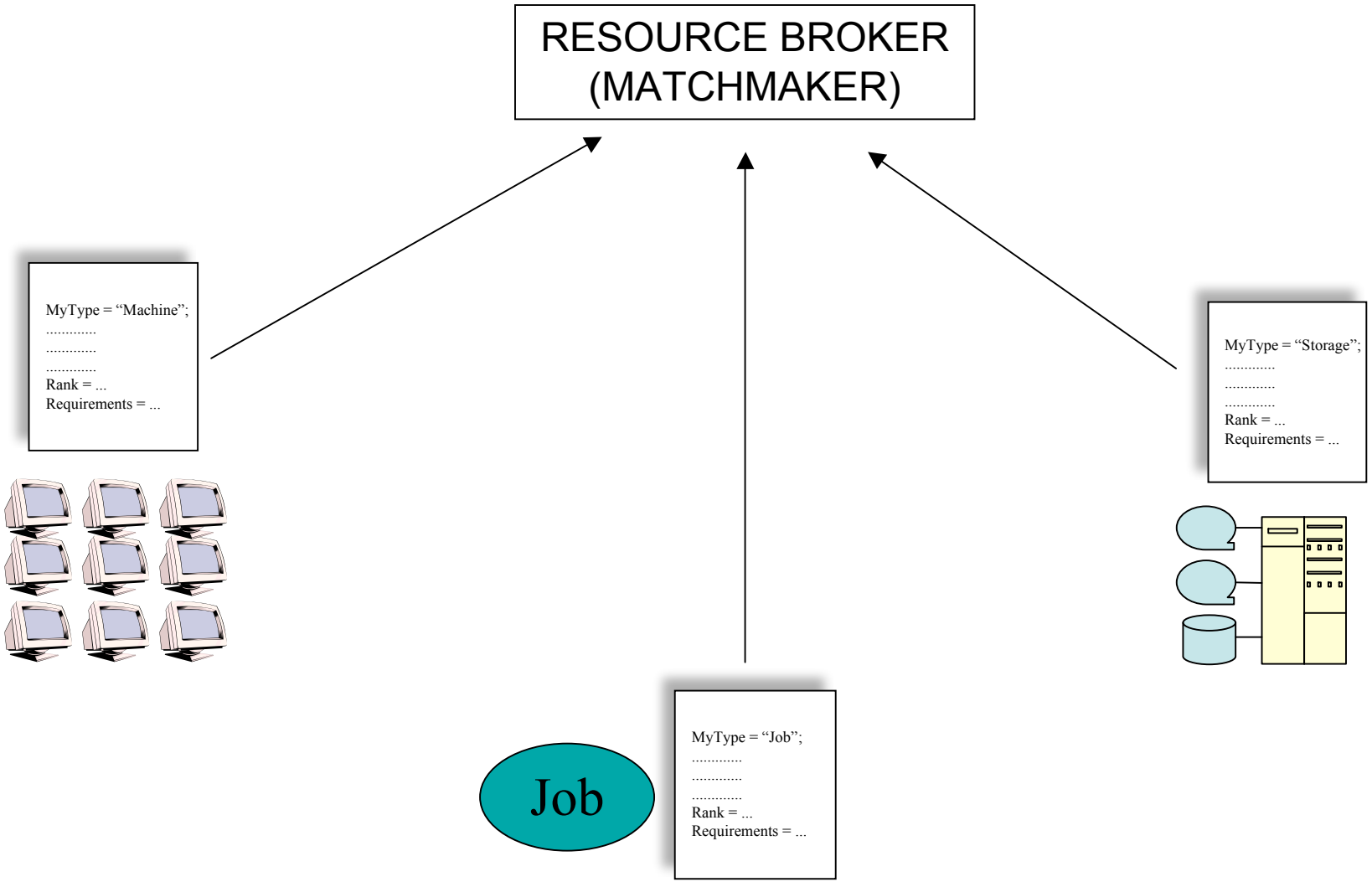
	Senior Personnel			Postdocs			Graduate Students			Undergraduates		
	Women	Minority	Total	Women	Minority	Total	Women	Minority	Total	Women	Minority	Total
LaTech			2						5			2
LSU	3		27			5	5	4	47	2	2	24
Tulane	1		3			2	3	5	11		1	3
ULL	7	4	18			8	13	5	31			4
UNO			7				5	4	32	3	8	17
Total	11	4	57			15	26	18	126	5	11	50

PetaShare Overview




Data Archival &
Meta Data Mngmt
Caching/
Prefetching
HSM
Replica
Selection
Data
Movement

Storage Systems as First Class Entities



Data-Aware Storage

- Storage server advertises:
  - Metadata information
  - Location information
  - Available and used storage space
  - Maximum connections available (eg. Max FTP conn, Max GridFTP conn, Max HTTP conn)
- Scheduler takes these into account
  - Allocates a connection before data placement
  - Allocates storage

Data-Aware Schedulers

- Traditional schedulers not aware of characteristics and semantics of data placement jobs

```
Executable = genome.exe
Arguments  = a b c d
```

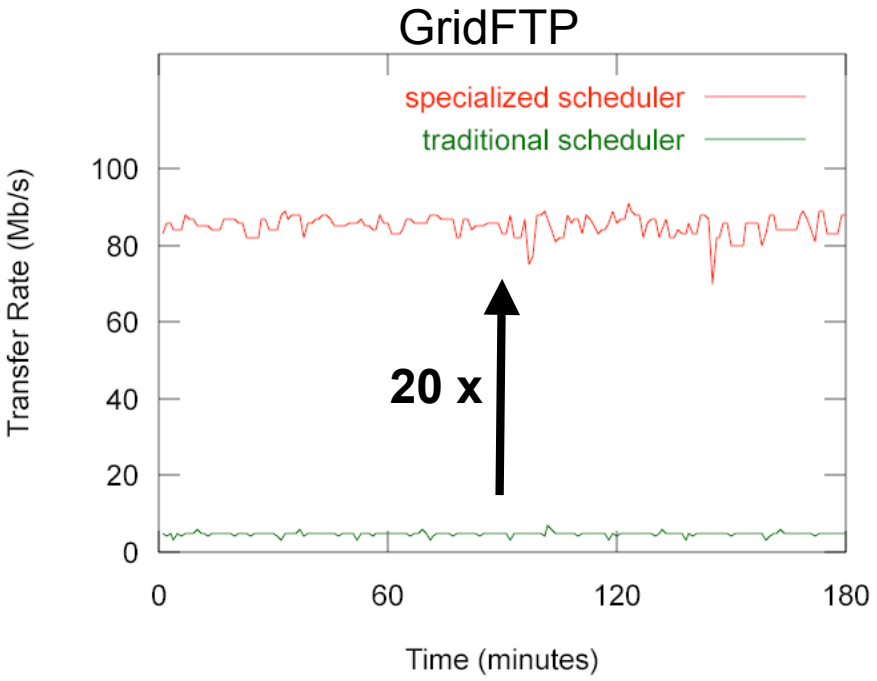
```
Executable = globus-url-copy
Arguments  = gsiftp://host1/f1
             gsiftp://host2/f2
             -p 4 -tcp-bs 1024
```

Any difference?

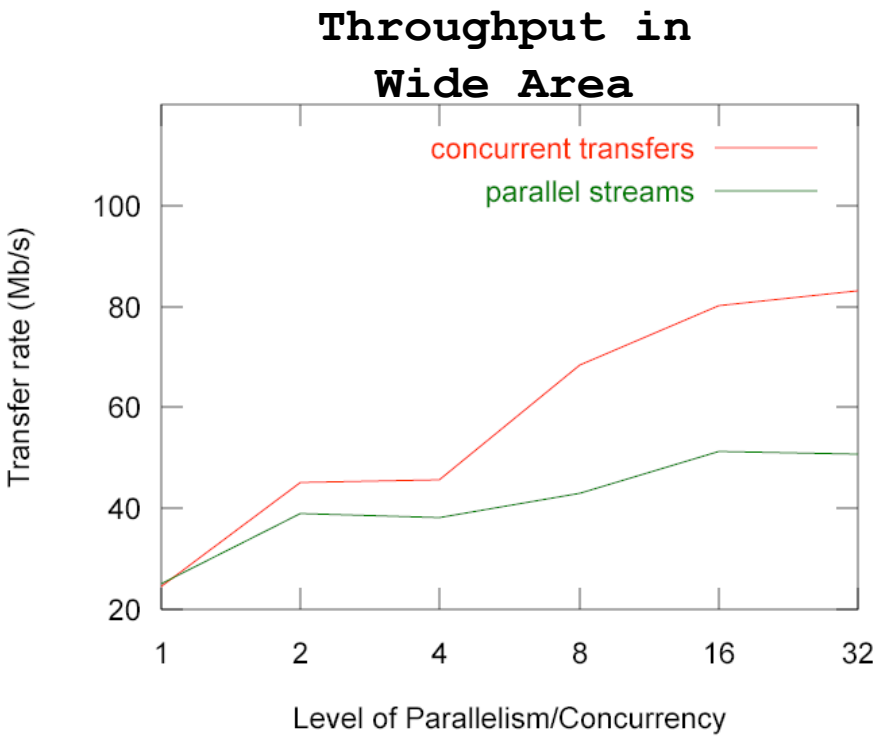
[ICDCS'04]

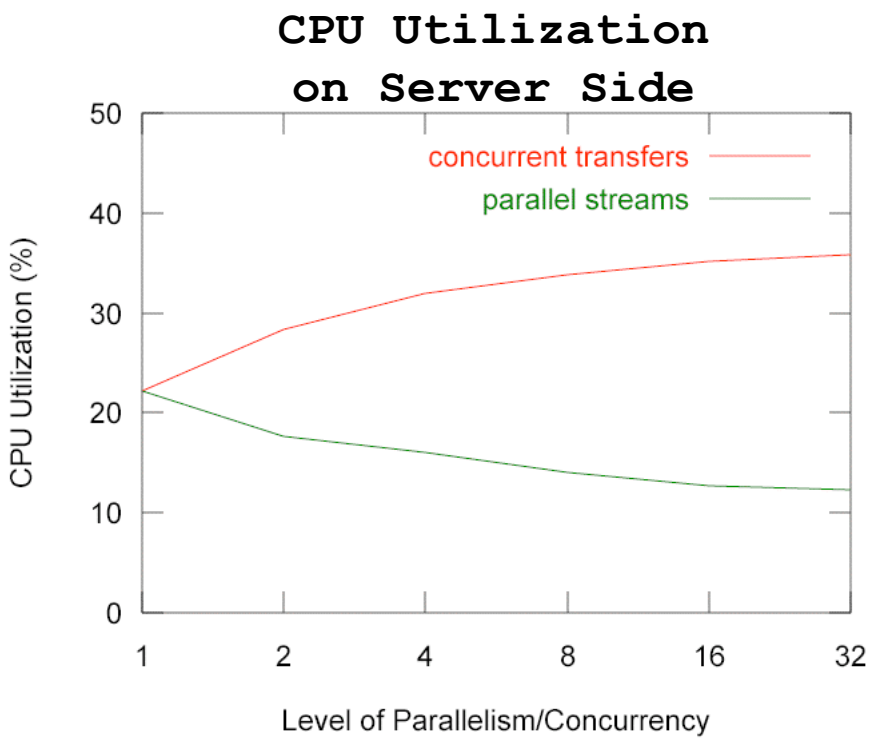
Data-Aware Schedulers

- What type of a job is it?
  - transfer, allocate, release, locate..
- What are the source and destination?
- Which protocols to use?
- What is available storage space?
- What is best concurrency level?
- What is the best route?
- What are the best network parameters?
  - tcp buffer size
  - I/O block size
  - # of parallel streams



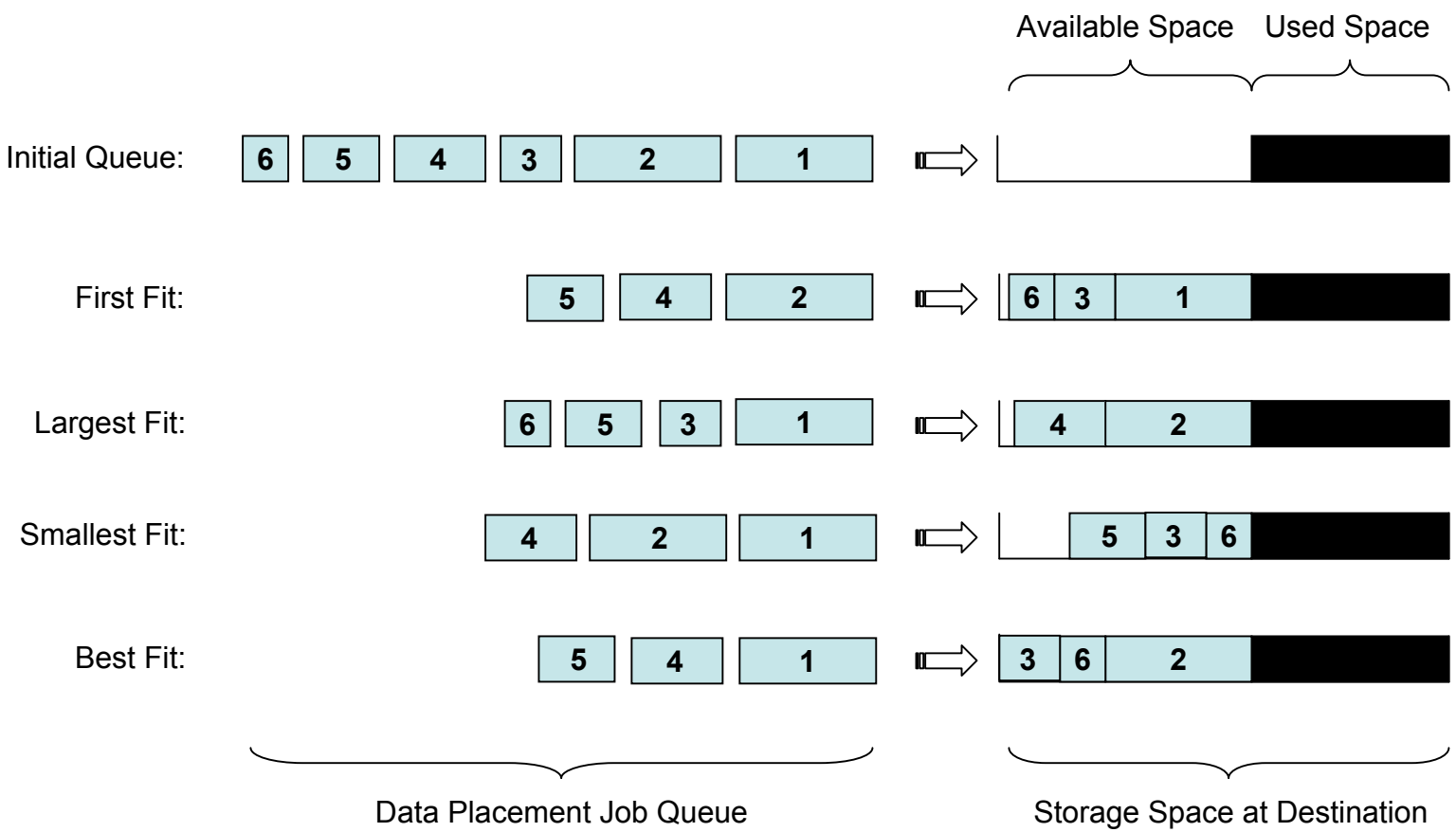
Optimizing Throughput and CPU Utilization at the same Time





- Definitions:
  - Concurrency: transfer n files at the same time
  - Parallelism: transfer 1 file using n parallel streams

Storage Space Management



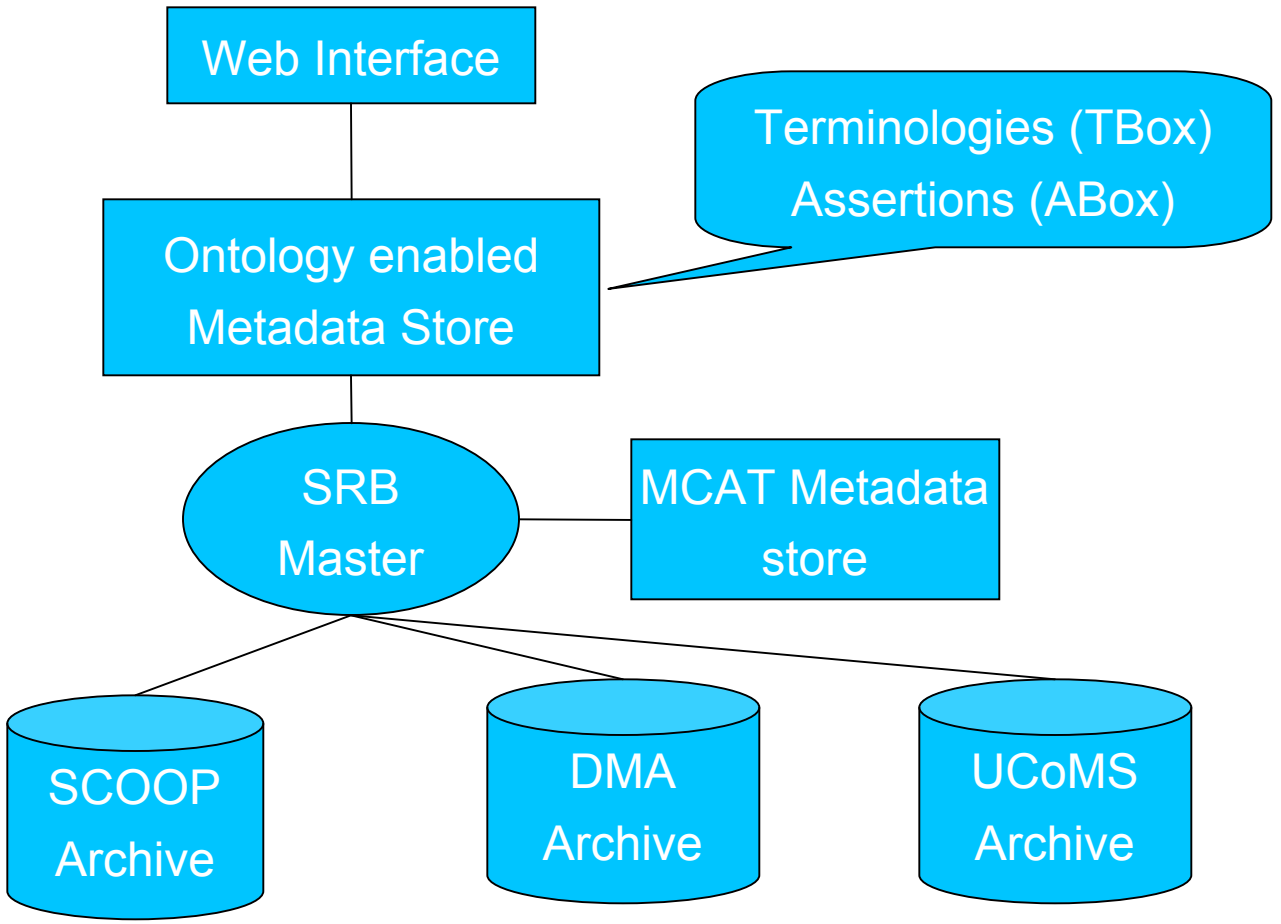
Cross-Domain Metadata

- SCOOP – Coastal Modeling
  - Model Type, Model Name, Institution Name, Model Init Time, Model Finish Type, File Type, Misc information …
- UCoMS – Petroleum Engineering
  - Simulator Name, Model Name , Number of realizations, Well number, output, scale, grid resolution …
- DMA – Scientific Visualization
  - Media Type, Media resolution, File Size, Media subject, Media Author, Intellectual property information, Camera Name …
- NumRel - Astrophysics
  - Run Name, Machine name, User Name, Parameter File Name, Thorn List, Thorn parameters …

Problem Definition

- Managing structured data over different knowledge entities
- Simulation metadata are tightly coupled with their specific knowledge domain
- Interoperability is the key, i.e., offer user coherent view over different data sets

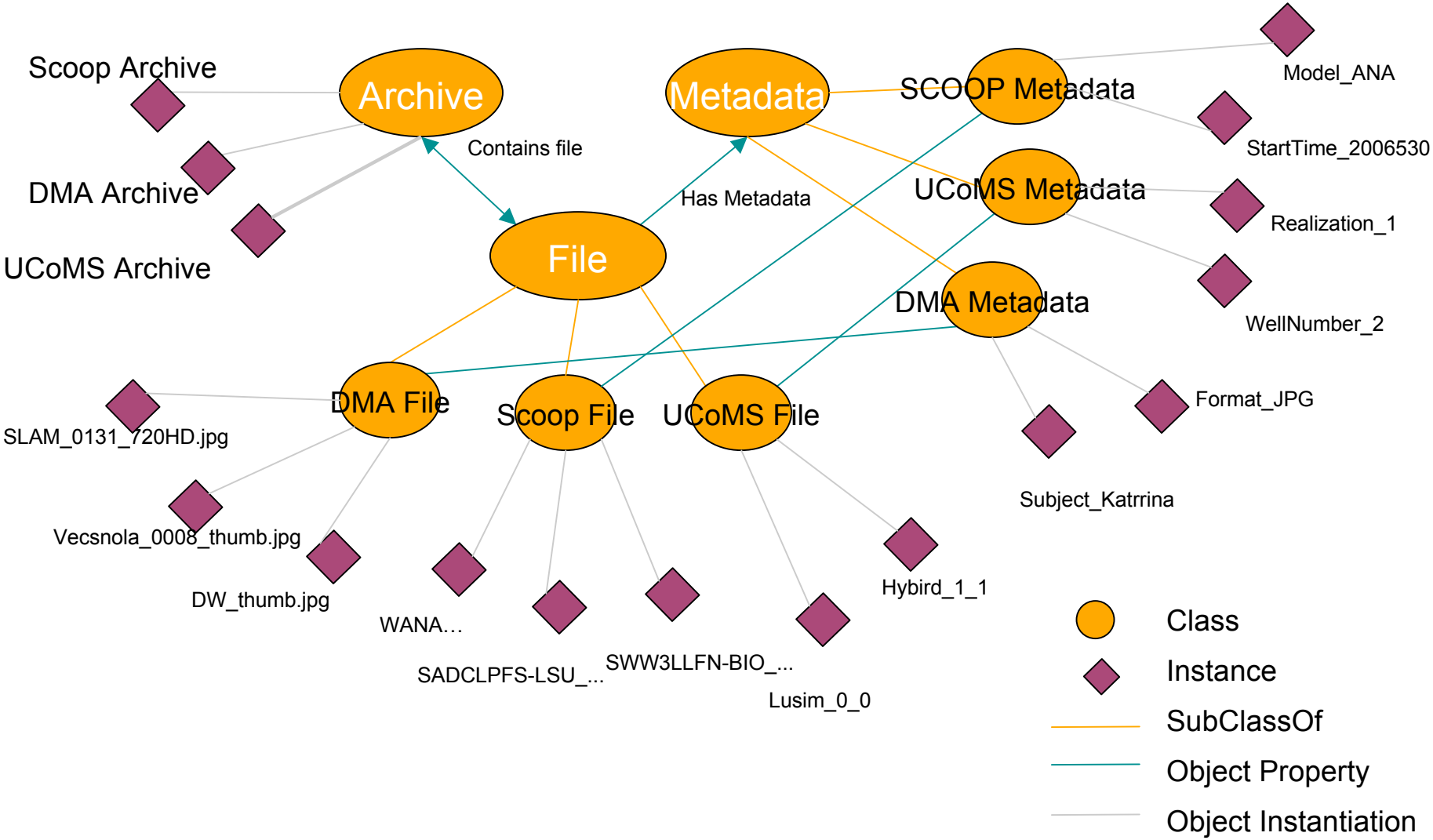
Architecture Graph



Implementation - Components

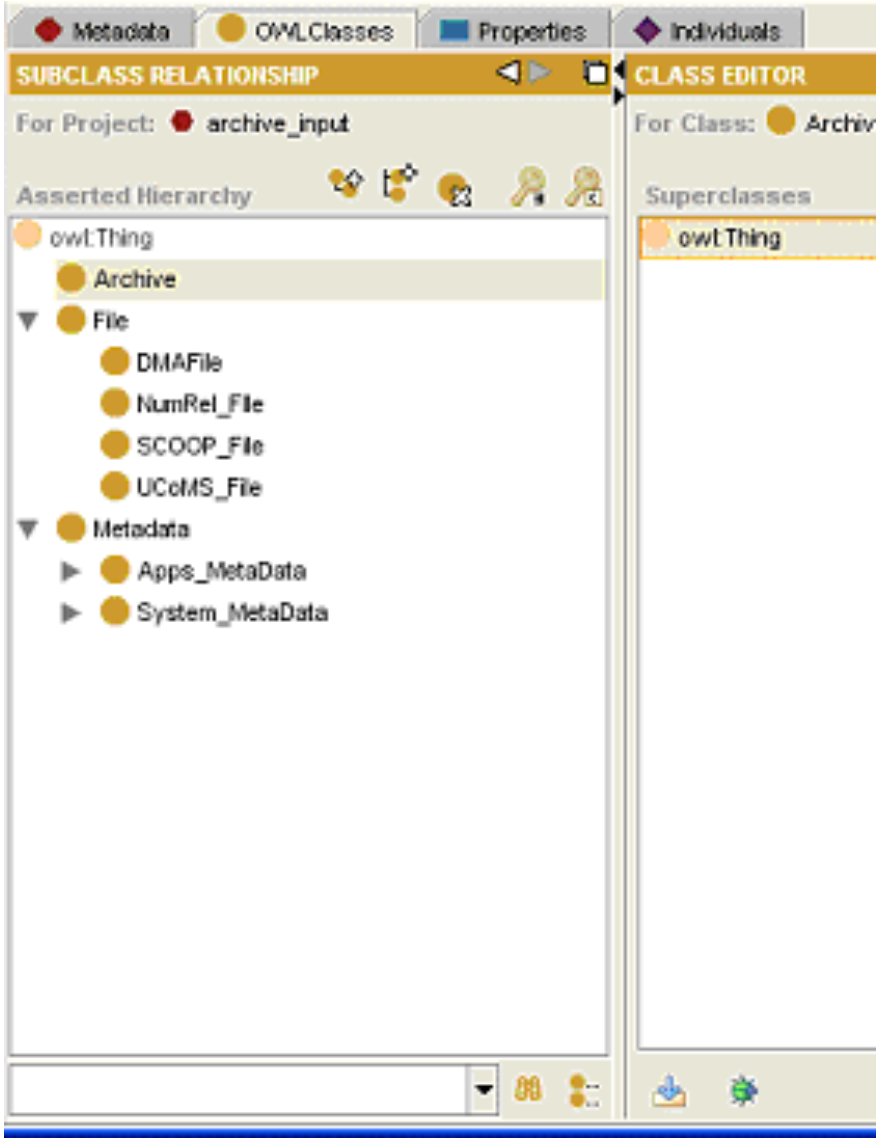
- Ontology Definition
  - Done in Protégé, stored as flat file
  - Define metadata structures
- Web interface development
  - Enable user to interact with ontology metadata store and access SRB files
- SRB registration
  - Register the physical files at local storages to SRB master
  - It is possible to bind the metadata with registration progress, metadata will be stored in MCAT
- Ontology metadata ingestion
  - Ingest the metadata into the ontology store

Ontology definition



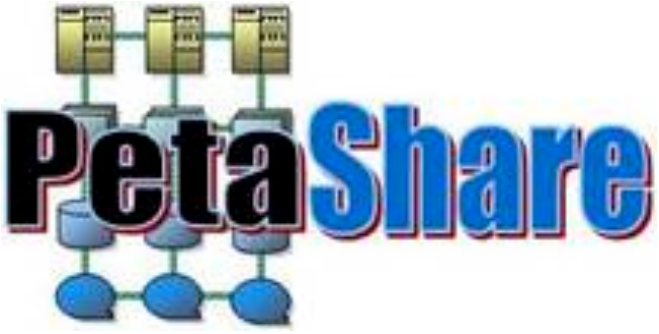
Actual ontology code/view

- <?xml version="1.0"?>
  <rdf:RDF
  xmlns:p1="http://www.owl-ontologies.com/Ontology1177880024.owl#ANA:"
  xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:p3="http://www.owl-ontologies.com/Ontology1177880024.owl#hasAppsMeta:ANA:"
  xmlns:p2="http://www.owl-ontologies.com/Ontology1177880024.owl#hasAppsMeta:"
  xmlns="http://www.owl-ontologies.com/Ontology1177880024.owl#"
  xmlns:p4="http://www.owl-ontologies.com/assert.owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xml:base="http://www.owl-ontologies.com/Ontology1177880024.owl">
  <owl:Ontology rdf:about="">
  <owl:imports rdf:resource="http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl"/>
  <owl:imports rdf:resource=""/>
  </owl:Ontology>
  <owl:Class rdf:ID="UTChem_Output">
  <rdfs:subClassOf>
  <owl:Class rdf:ID="Reservoir_UTChem_Metadata"/>
  </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Trans">
  <rdfs:subClassOf>
  <owl:Class rdf:ID="SCOOP_MetaData"/>
  </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Metadata"/>
  <owl:Class rdf:ID="Wind_FirstTime">
  <rdfs:subClassOf>
  <owl:Class rdf:ID="WIND"/>
  </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#WIND">
  <rdfs:subClassOf>
- … …



Summary

- PetaShare aims to enable data intensive collaborative science across state, by providing
  - Additional storage
  - Cyberinfrastructure to access, retrieve and share data
    - Data-aware storage
    - Data-aware schedulers
    - Cross-domain metadata



A system driven by the local needs (in LA), but has potential to be a generic solution for the broader community!

Acknowledgment: This work was supported by NSF grant CNS-0619843.