Peeking into 21st Century Distributed Data Intensive Science

Sudharshan Vazhkudai

Network and Cluster Computing, CSMD

Oak Ridge National Laboratory

http://www.csm.ornl.gov/~vazhkuda

vazhkudaiss@ornl.gov

Acknowledgments: ORNL, U of M, ANL, Globus Toolkit Project, GriPhyN

UTK Computer Science
February 13, 2004

Talk Summary

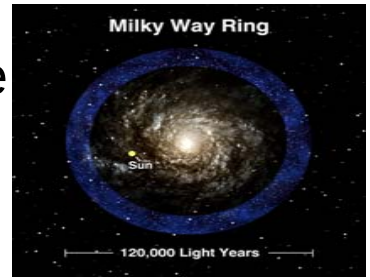
- Orchestrating bulk, Grid data transfers (part of my doctoral work)
  - Challenges:
    - Scalable scheduling architecture (design and experimentation problem)
    - Synthesizing information necessary for scheduling (statistical modeling problem)
- Constructing Distributed Storage Facility through Scavenging (work-in-progress at ORNL)
  - Challenges:
    - Commoditizing the Grid storage fabric
    - Scalable, transparent system with good performance

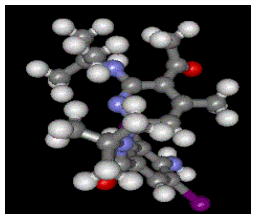
Outline: Orchestrating Grid Data Transfers

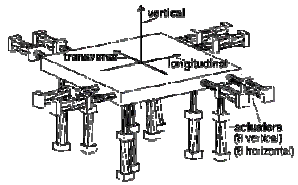
- The Good & “Not-so-good” News!
- Data Grids – The Solution…?
- Contributions

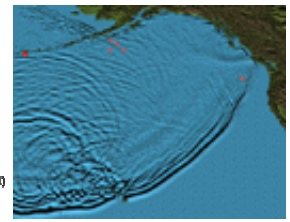
The Good News is…

- All the kinds of problems we can solve using distributed data-intensive science
  - High Energy Physics (GriPhyN, EUDataGrid, PPDG)
  - Digital Sky Survey (SDSS)
  - Earthquake Simulations (NEESgrid)
  - SETI@Home, FightAIDS@Home, Folding@Home Evolution@Home
  - Bio Terrorism (Small Pox Research Grid)
  - BioInformatics, Geo Spatial Analysis
  - Medical data, digitizing patient records (Health Grid)



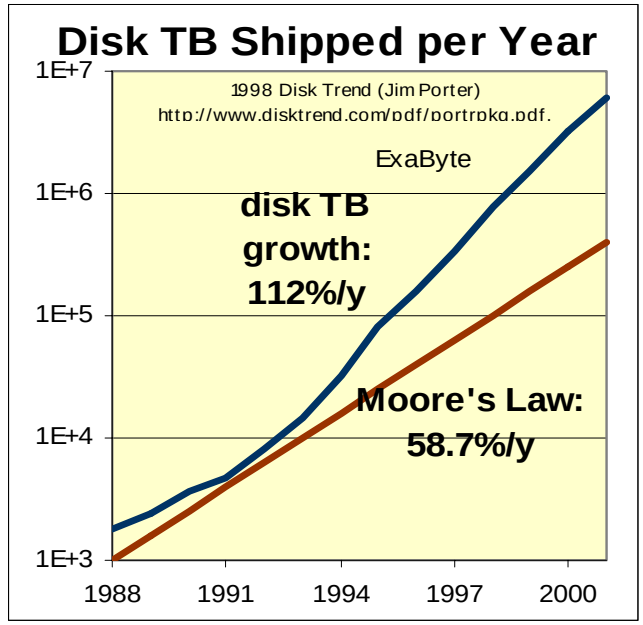






Technology Trends

- Moore's Law: Price/Performance doubles every 18 months
- Storage capacity beating Moore's Law – doubles every 12 months
- What does it mean…?
  - New Storage = sum of all old storage
- Gilder's Law: Networks are improving every 8 months
  - WANs are becoming faster than LANs!
- So, Distributed Data-Intensive Science is the Next Step in Evolution…
- E. coli doubles every 20 minutes!



Meta-Message: Should we spend millions building petaflop supercomputers or spend it in building distributed software infrastructure?

So, What is the “Not-so-good” News?

- The Data Deluge…
  - 2000 ~0.5 Petabyte
  - 2005 ~10 Petabytes
  - 2010 ~100 Petabytes
  - 2015 ~1000 Petabytes?
- Problem:
  - Massive data collections
  - Computationally intensive analyses
  - Rapid access to large subsets
  - Data, computers, and users distributed across networks of varying capability
  - How to collect, manage, access and interpret such huge quantities of data?

Digitize everything!!

Factoids

10^{24} Yotta

Zetta

Exa

10^{15} Peta

Tera

Giga

Mega

Kilo

Meta-Message: “Remember this: With great power comes great responsibility!!” – Spiderman

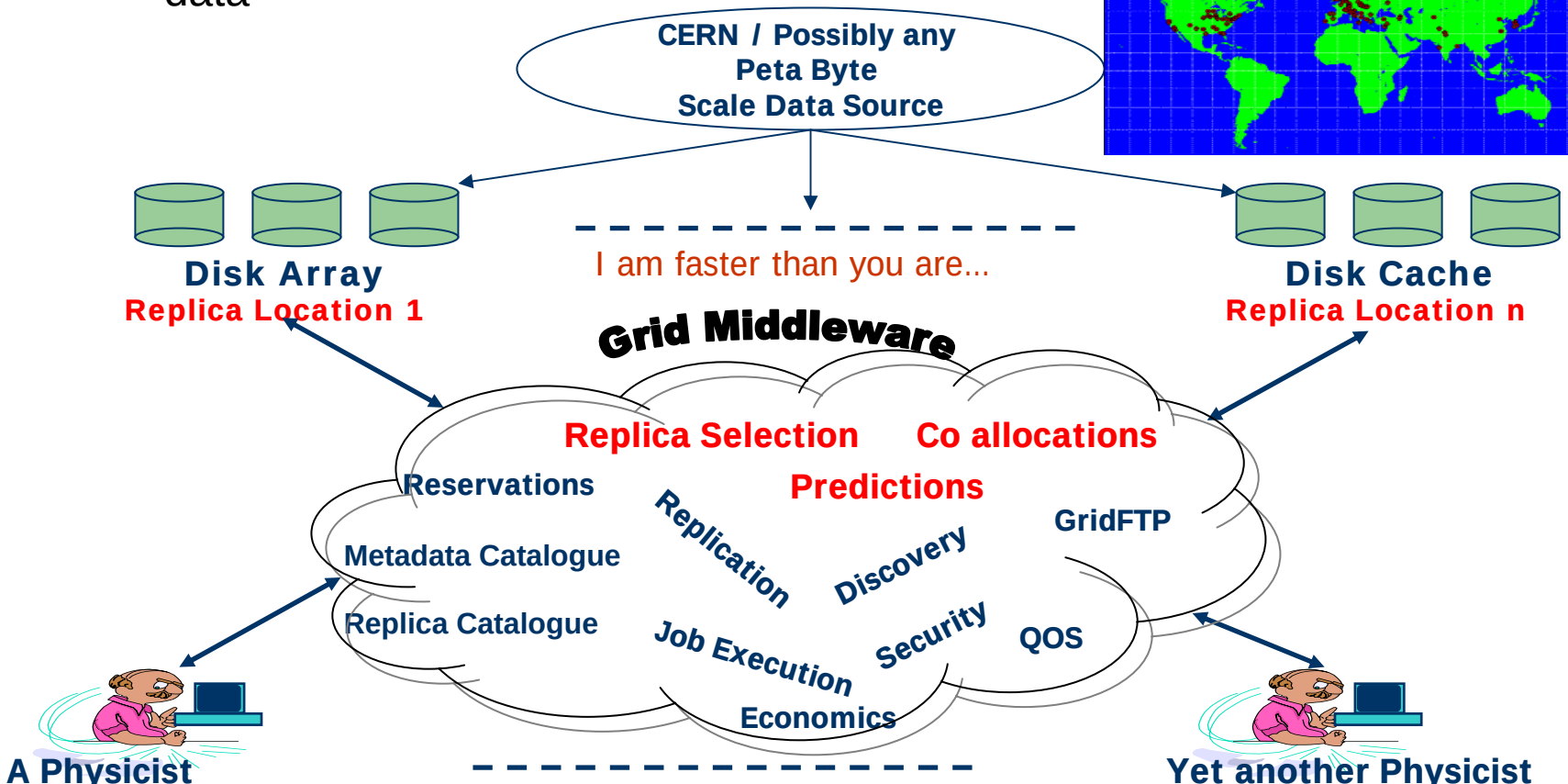
Outline

- The Good & “Not-so-good” News!
- Data Grids – The Solution…?
- Contributions

Problem Scope

- Problem:
  - Efficient access to massively replicated, bulk data

2000 Physicists, 150 Institutes, 32 Countries



Outline

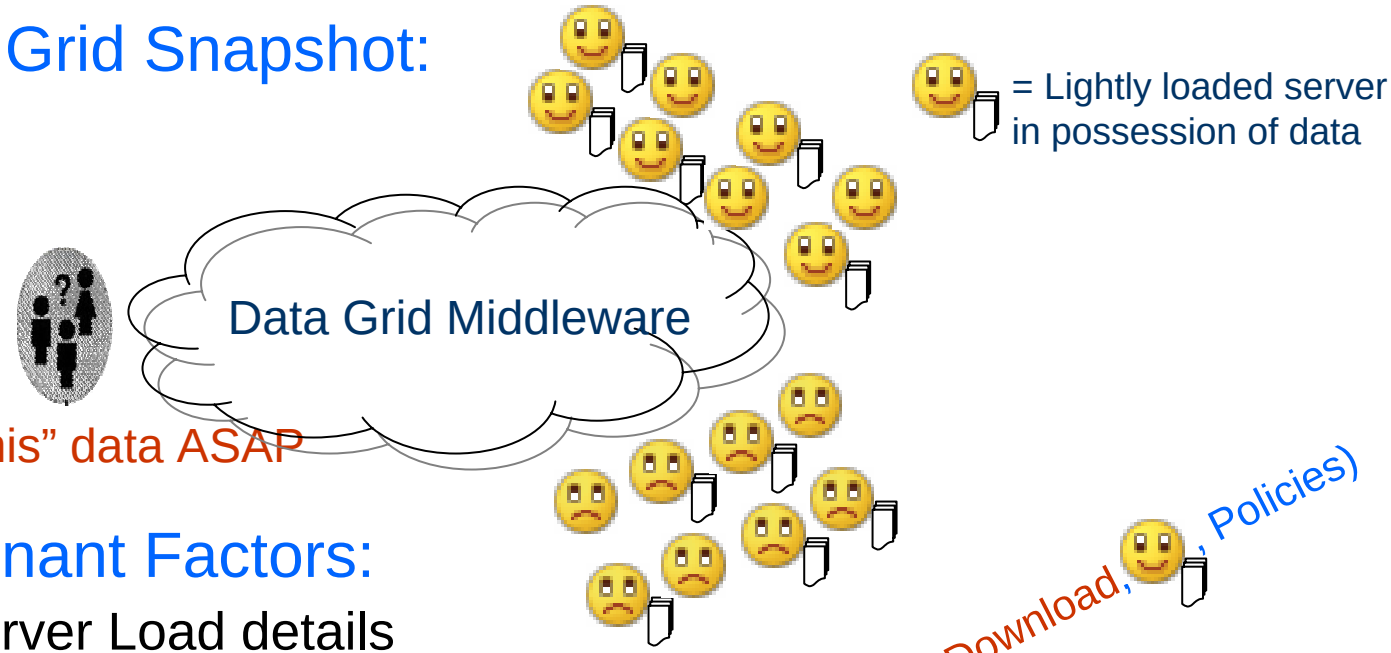
- The Good & “Not-so-good” News!
- Data Grids – The Solution…?
- Contributions

Replica Selection Background

- Problem:
  - To select datasets from among numerous alternatives
- Use case:
  - Scientific data sharing communities
  - Content distribution networks
  - Widely-distributed file systems
- Related Work:
  - Aqua
  - Coda
  - Akamai, OceanStore

Heuristics for Replica Selection [CCGRID'01]

- Data Grid Snapshot:



Get me "this" data ASAP

- Dominant Factors:
  - Server Load details
  - Usage Policies
  - Total Time-to-Download

A good heuristic = f(Time-to-Download, , Policies)

Scheduling Architecture [SC'00]

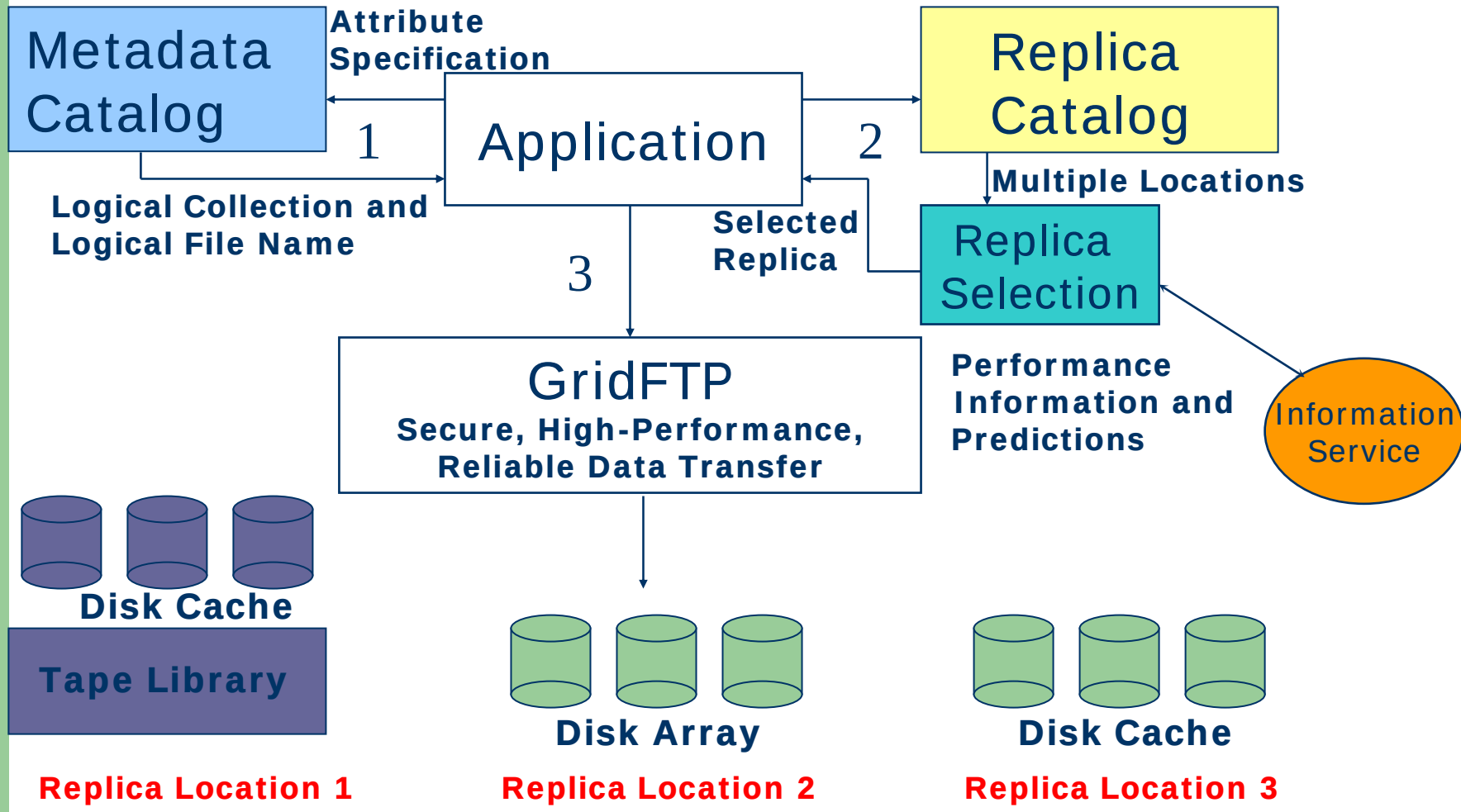
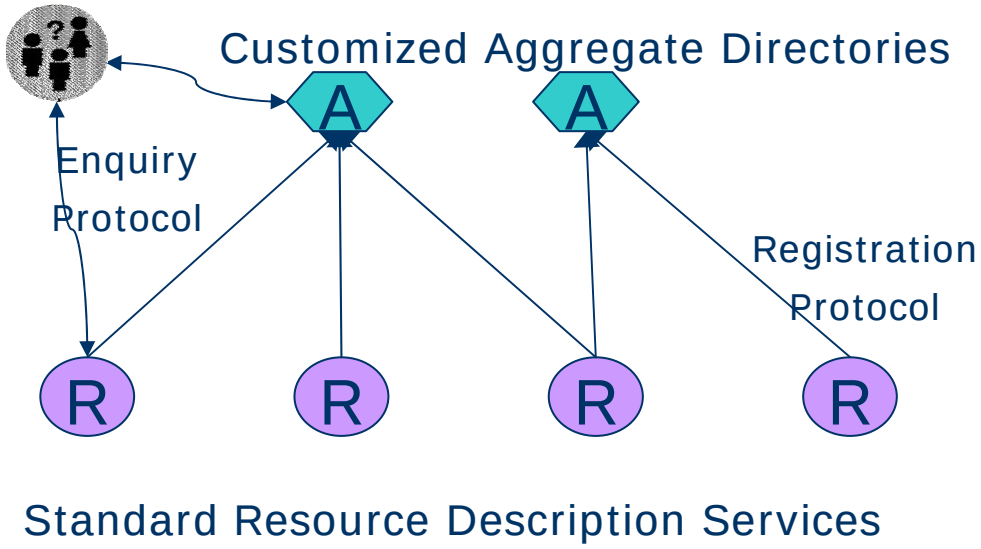


Image Courtesy: Globus Project

Attributed-based Storage Resource Management [SC'00, CCGRID'01]

- Mechanisms to expose:
  - (Time-to-Download, , Policies)
  - Identify, collect and publish server load details
  - Discovery in the context of an information service

- Publish/Query Infrastructure



Storage Information Provider

dn:"140.221.65.69,
hostname=dpsslx04.lbl.gov,dc=lbl,dc=gov,o
=grid"
cn:"140.221.65.69"
gridftpurl:"gsiftp://dpsslx04.lbl.gov:61000"
minrdbandwidth:1462K
maxrdbandwidth:12800K
avgrdbandwidth:6062K
……
Policy: only these sites & at this time

Results and Summary

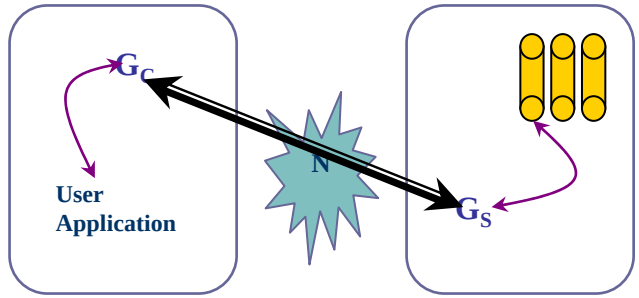
- A decentralized scheduler architecture
  - Demonstrated a working prototype for Data Grids (C and LDAP)
- Use of cost-functions
  - Matching storage capabilities and application requirements using Classified Advertisements
- Attribute-based Storage
  - Managing storage in Grids
  - Exposing and discovering capabilities (LDAP and scripting)

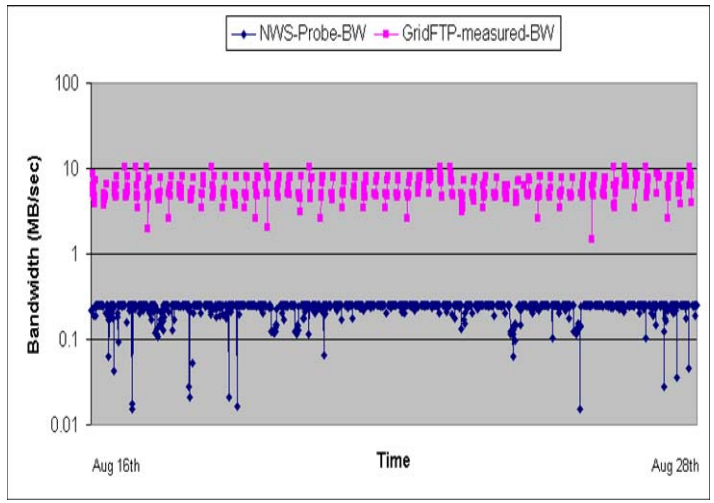
Outline

- The Good & “Not-so-good” News!
- Data Grids – The Solution…?
- Contributions – Predictions

Forecasting Background

- Problem:
  - An estimate for GridFTP end-to-end future data transfer performance
- Use Case:
  - Data Grids, Networking community
  - Internet downloads
- Related Work:
  - ~~Option1: Model components~~
  - Option2: Whole system evolution





Predictions on Past History – Univariate Forecasting [IPDPS'02]

- Predictors feed on GridFTP performance logs
- Mean based:
  - **Pros:** *simple*
  - **Cons:** *smoothes out fluctuations*
- Median based:
  - **Pros:** *simple, excludes extreme values*
  - **Cons:** *may not be representative of entire dataset*
- Autoregressive models $Y_t = a + bY_{t-1}$
  - **Pros:** *weighted averages*
  - **Cons:** *computational cost, larger datasets*
- Context Sensitive & Insensitive Variants
  - File size categories
  - Training sets
  - Sliding windows
    - Count based
    - Temporal windows

Sample Log

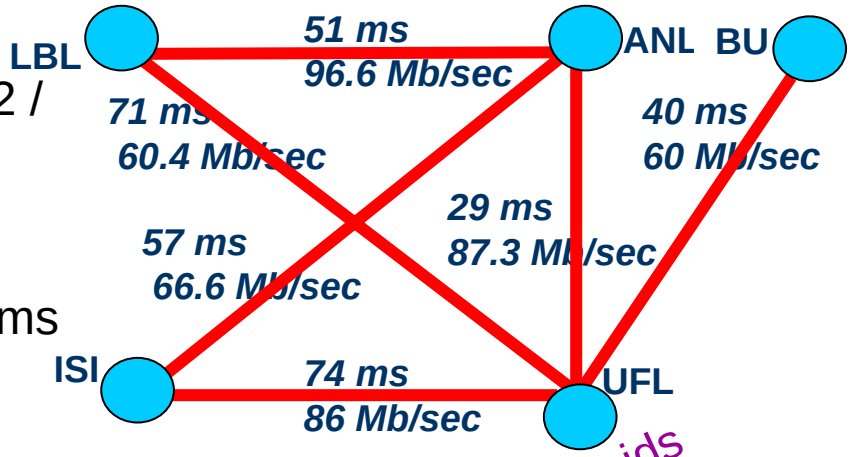
DATE=1007992862.934798 HOST=dpsslx04.lbl.gov PROG=wuftpd NL.EVNT=FTP_INFO START=1007992861.890575 FILE=/home/ftp/vazhkuda/READ-DIR/10MB BUFFER=1000000 BLOCK=8192 NBYTES=10240000 VOLUME=/home/ftp STREAMS=8 STRIPES=1 DEST=1[128.227.64.235] TYPE=RETR CODE=226

Predictors

	Mean	Median	ARIMA
All data	AVG	MED	AR
Last 1 Value	LV		
Last 5 Values	AVG5	MED5	
Last 15 Values	AVG15	MED15	
Last 25 Values	AVG25	MED25	
Last 5 Hours	AVG5hr		
Last 15 Hours	AVG15hr		
Last 25 Hours	AVG25hr		
Last 5 Days			AR5d
Last 10 Days			AR10d

Experiment Setup

- Testbed
  - Several sites connected by OC-12 / OC-48 links, Internet2
  - RAID Disks at servers and clients
  - Symmetric multi processors
  - Linux and Solaris operating systems
  - Prediction middleware in “C”



- GridFTP Experiments
  - Bulk transfers ranging from 10MB to 1GB
  - Transfers at random intervals
  - Tuned buffer sizes and parallel flows
  - ~ 100 to 200 transfers in various sizes during three distinct 2 week periods

Factoids

3 DVDs?

OC192	10Gbps
OC48	2.4Gbps
OC12	622Mbps
OC3	155.5Mbps
OC1	51.8Mbps
T3	44.7Mbps
T1	1.54Mbps

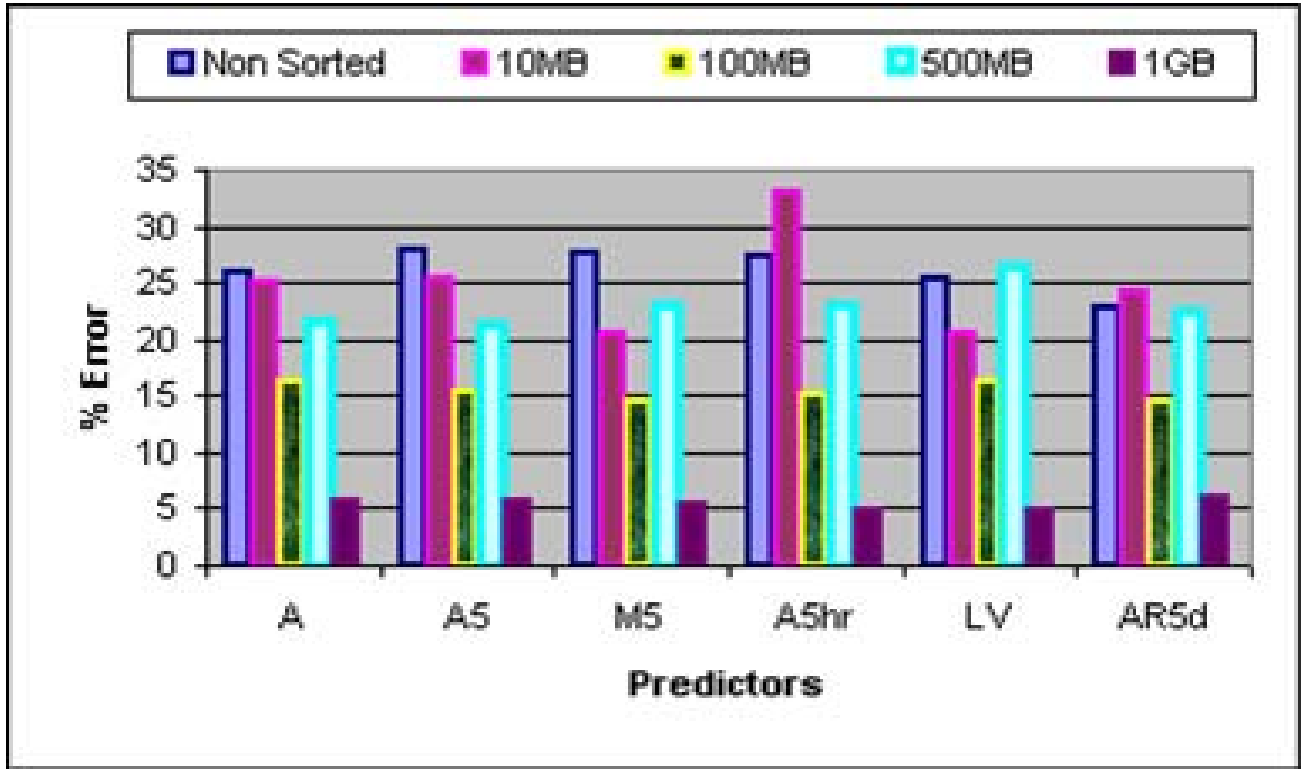
Meta-Message: Optical links are coupled by electronic devices. Computing is still the bottleneck!!

Univariate Predictor Performance

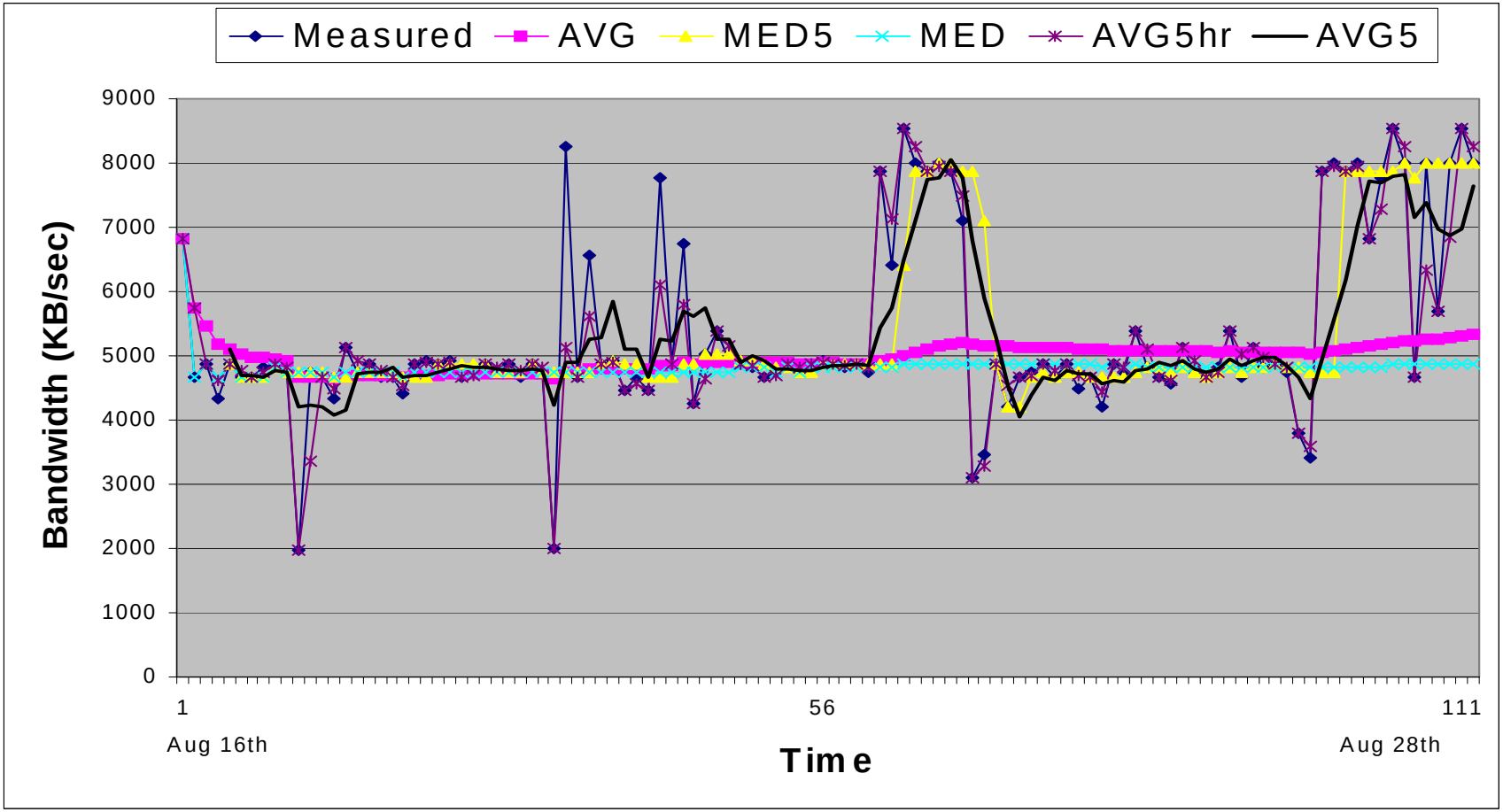
- Normalized Error:

$$\% \text{ Error} = \frac{\sum |\text{Measured}_{BW} - \text{Predicted}_{BW}|}{(\text{size} * \text{Mean}_{BW})} * 100$$

- Results [IPDPS'02]:
  - 15 – 25% error
  - 10% reduction due to classification
  - Moving variants perform better



Univariate Predictor Performance



Meta-Message: Simple averages and medians do not capture trends.

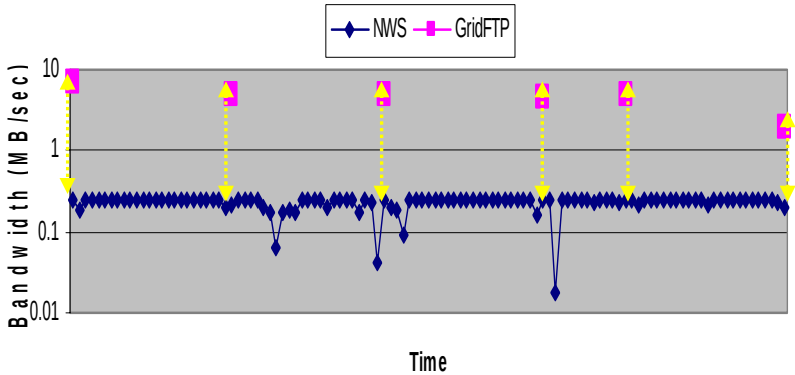
Multivariate Forecasting [HPDC'02, GRID'02]

- Drawbacks of the Univariate Model:
  - Grid Data Transfers can be quite sporadic
  - Log data alone not reflective of current system behavior
  - Component load variations can affect eventual throughput achieved

- Why multivariate Forecasting?
  - Mitigate the adverse effects of sporadic transfers
  - Improve prediction accuracy

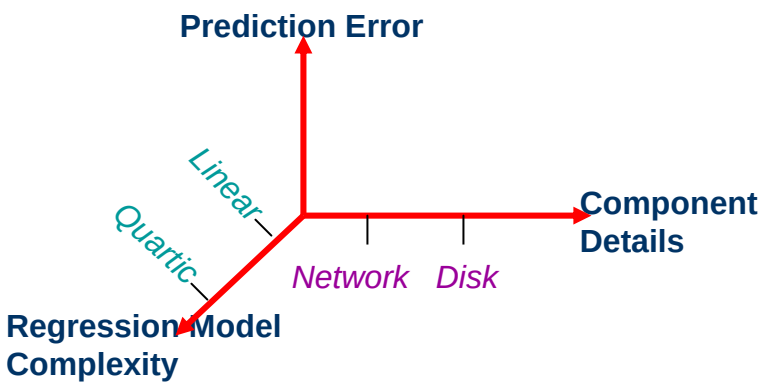
- Desired Functional Properties of additional variables:
  - As lightweight as possible
  - Correlated to the end-to-end GridFTP throughput

- Networks & Disks:
  - 70% and 30% of the transfer
  - "regular" NWS (64K) and iostat probes
  - Captures ambient load during transfers



Regression Analysis

- Why Regression?
  - Allows us to study the effects of several independent variables on a dependent variable
  - GridFTP throughput is dependent on several component load variations
- Prediction Model:



- Linear Regression:
  - $G^l = a + bN$
  - Coefficients obtained using previous network & GridFTP values (least squares)

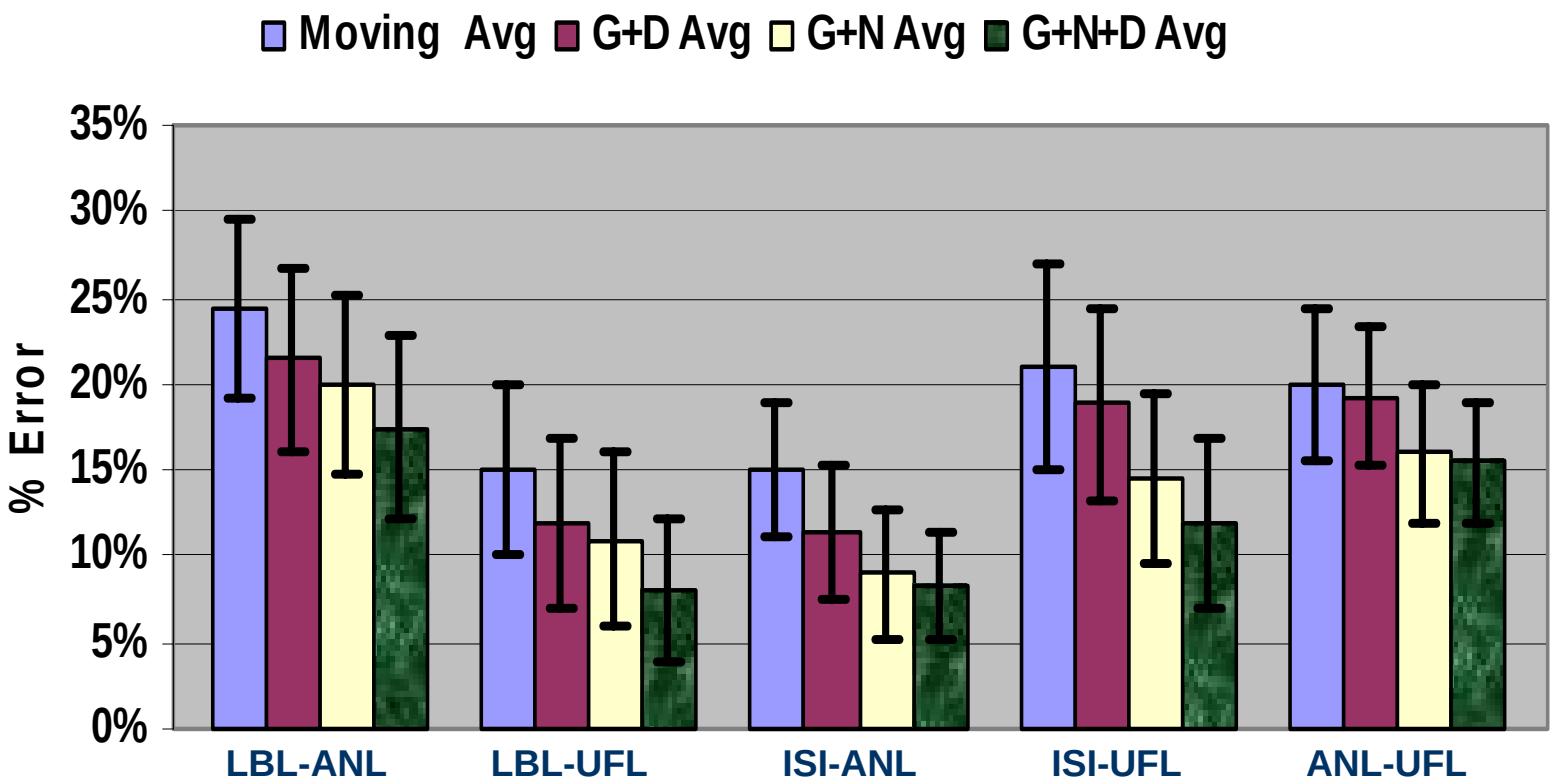
$$a = \mathrm{Mean}(G) - b * \mathrm{Mean}(N)$$

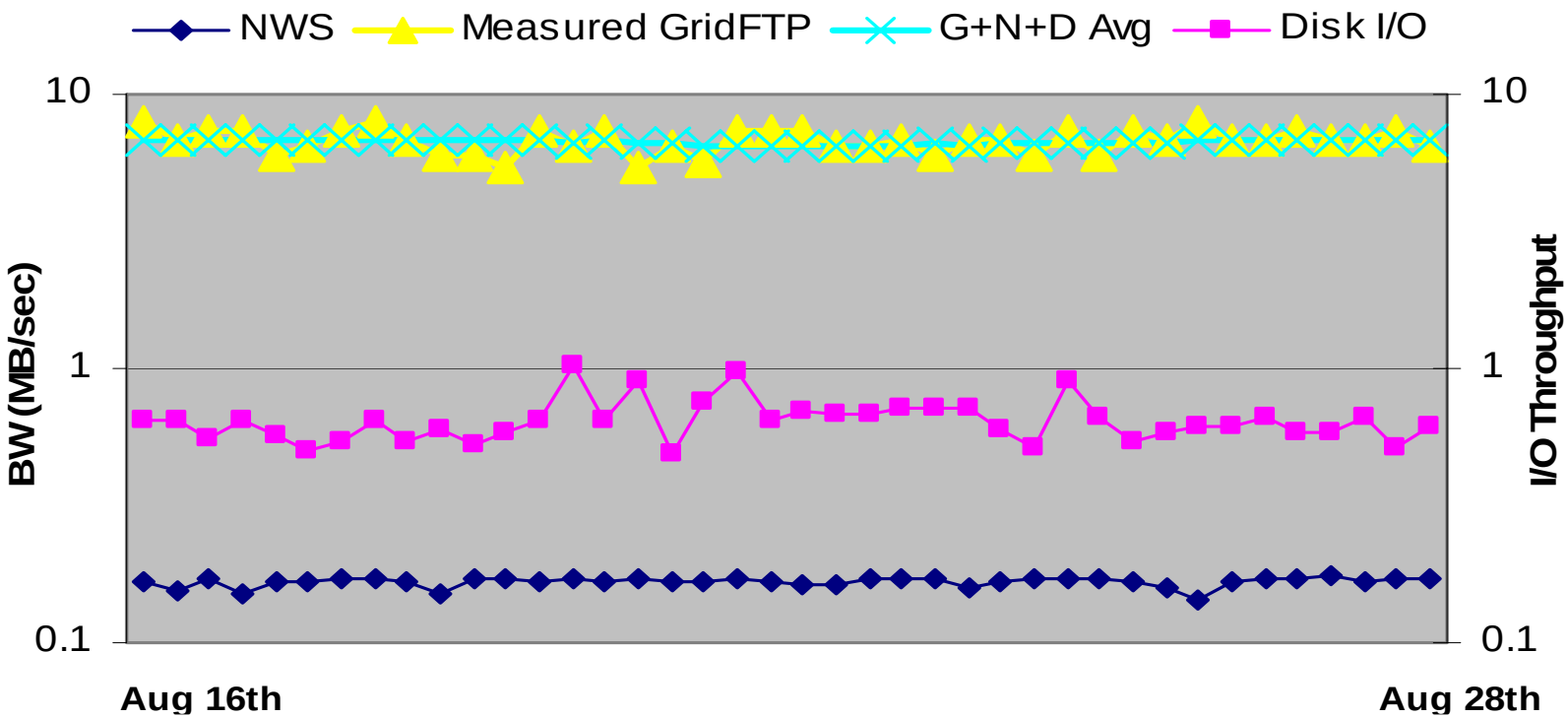
$$b = \frac{\sum NG - (\sum N \sum G/size)}{\sum G^2 - (\sum G)^2/size}$$

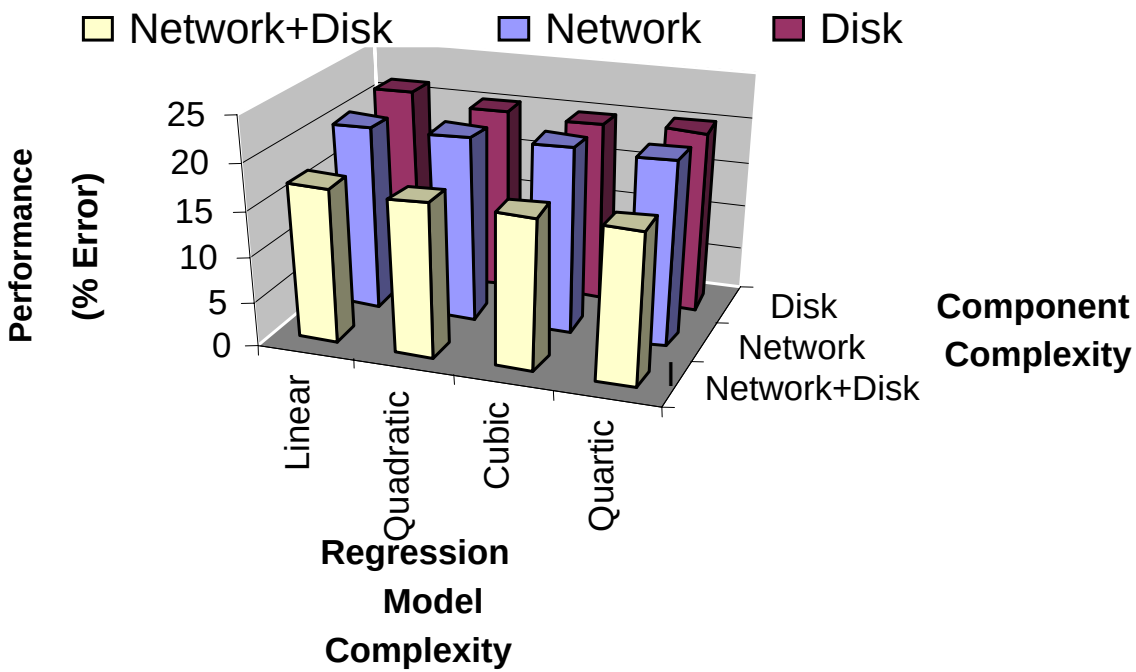
- Polynomial Models:
  - Help Improve accuracy
  - $G^l = a + b_1N + b_2N^2$
- Multiple Regression:
  - Account for several variables
  - $G^l = a + b_1N + b_2D$

Multivariate Performance

- Results [JHPCA'03]:
  - Within 15% error
  - Moving Avg < G+D Avg < G+N Avg < G+N+D Avg




NWS
Measured GridFTP
G+N+D Avg
Disk I/O
10
1
0.1
BW (MB/sec)
I/O Throughput
Aug 16th
Aug 28th


Network+Disk
Network
Disk
Performance
(% Error)
25
20
15
10
5
0
Linear
Quadratic
Cubic
Quartic
Regression
Model
Complexity
Disk
Network
Network+Disk
Component
Complexity

Summary

- Prediction Architecture for bulk data transfer times
- Univariate Tools
  - Prediction within 25% error
  - 10% reduction due to file groupings
- Multivariate Tools
  - Capturing network and disk factors in the prediction model
  - Several statistical imputation techniques to mitigate sporadic transfers
  - Prediction within 15% error
- Synthesizing meaningful information:
  - In the face of transient conditions, failure & unavailability
  - Scheduling is only as good as the information in hand
  - **Be Agnostic!!**
- Very pragmatic for a production Grid system
- Replica sites, once selected and ranked, can then be used to deliver the data collectively using co-allocation techniques [GRID'03]

Outline: Revisiting Storage Fabric for the Grid

- Grid Storage Fabric Background
- The Evolving Computing Landscape—An Analogy
- Storage Scavenging of User Desktop Workstations
- Use Cases and Design Choices
- Morsels, Morsels Everywhere…
- User Autonomy, Space Reclaim and Availability
- Other Services

Work in progress…

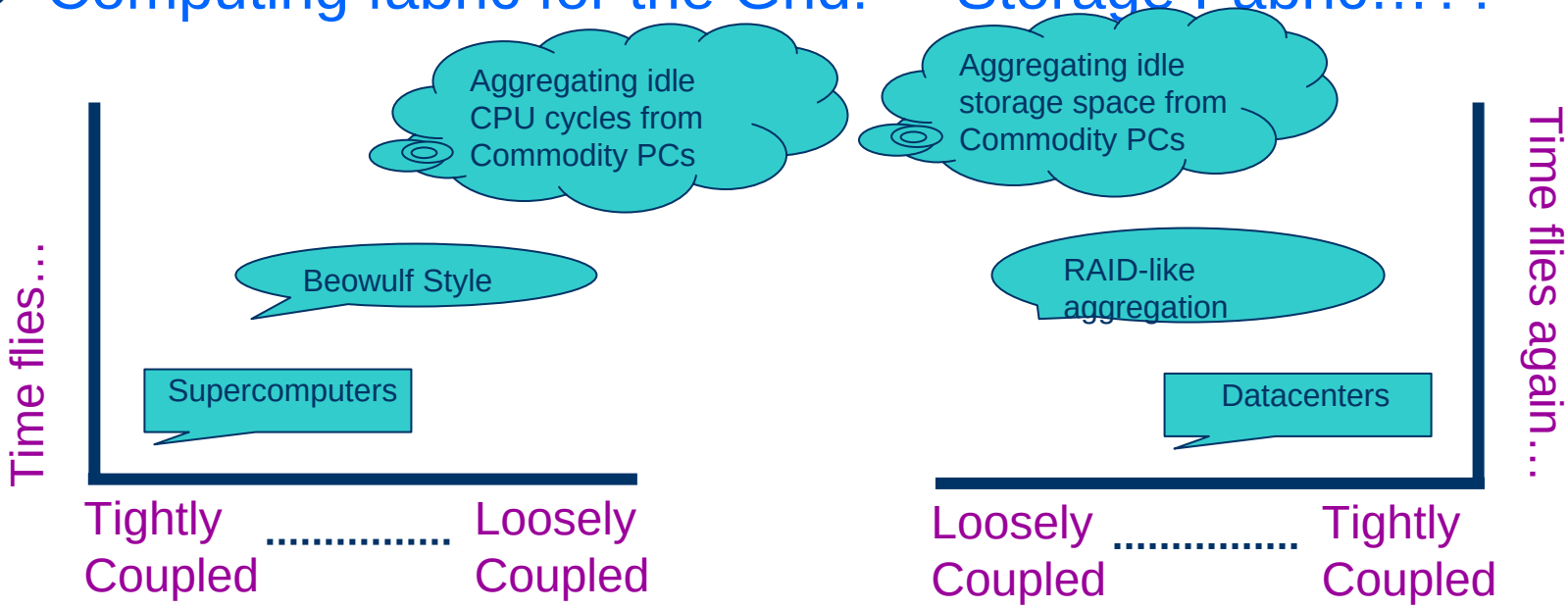
Grid Storage Fabric Background

- Scientific discoveries driven by analyses of massively distributed, bulk data.
  - Proliferation of high-end mass storage systems, SANs and datacenters
  - Providers such as IBM, HP, Panasas, etc.
- Merits:
  - Excellent price/performance ratio
  - Good storage speeds and access control
  - Support intelligent parallel file systems
  - Optimized for wide-area, bulk transfers
  - Reliability!
  - Successful demonstration in production Grids: DOE Science Grid, Earth System Grid, TeraGrid, etc.
- Drawbacks:
  - Increasing deployment/maintenance/administrative costs
  - Specialized software and central points of failure
  - Costs and specialized features prohibit wider acceptability and limits to select few research labs & organizations
  - Aforementioned production Grids are hardly half-a-dozen sites…!!

Meta-Message: If grids are to become prevalent and grow beyond the confines of a few organizations, exploiting commodity fabric features is absolutely essential!

The Evolving Computing Landscape—An Analogy

- Grids often quoted to have grown out of the traditional supercomputer or “computer-center” model
- Computing fabric for the Grid: Storage Fabric…??



Meta-Message: Proprietary systems are being replaced with commodity clusters, delivering new levels of performance and availability at dramatically affordable price point.

Storage Scavenging of User Desktop Workstations

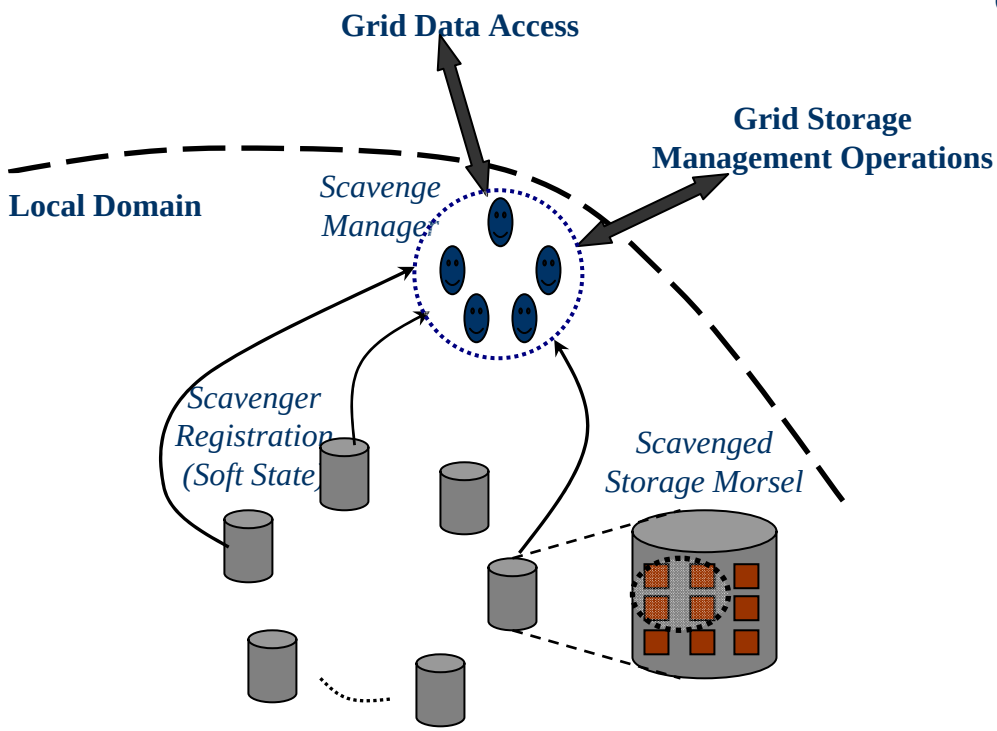
- Harnessing collective storage potential of individual workstations ~ Harnessing idle CPU cycles
- Why Storage Scavenging can be viable?
  - Economics of buying gigabytes of storage is increasingly affordable
  - Space usage to Available storage ratio is significantly low
  - Increasing numbers of workstations are online most of the time
  - Even a meager contribution (*Contribution << Available*) can amass collective, staggering aggregate storage!
- Concerns:
  - Vagaries of volatility…
  - Question of Trust: datasets on arbitrary user workstations
  - Performance of such aggregate storage

Meta-Message: Despite the high maintenance and administrative costs, a factor that attracts the Grid community to high-end storage and data centers is their ability to deliver sustained high-throughput for data operations.

Use Cases and Design Choices

- Use Cases:
  - Grid Data Replication & Mirroring
  - Data Staging
  - In-transit Data
- Related Work:
  - Network/Distributed File Systems (NFS, LOCUS)
  - Parallel File Systems (PVFS, XFS)
  - Serverless File Systems (FARSITE, xFS, GFS)
  - Peer-to-Peer Storage (OceanStore, PAST, CFS)
  - Grid Storage Services (LegionFS, SRB, IBP, SRM, GASS)
- Design Choices & Assumptions:
  - Scalability: O(100) or O(1000)
  - Commodity Components: Quality & Quantity
  - User Autonomy
  - Well connected & Secure
  - Heterogeneity
  - Large, “write once read many” datasets
  - Transparent
  - Grid Aware

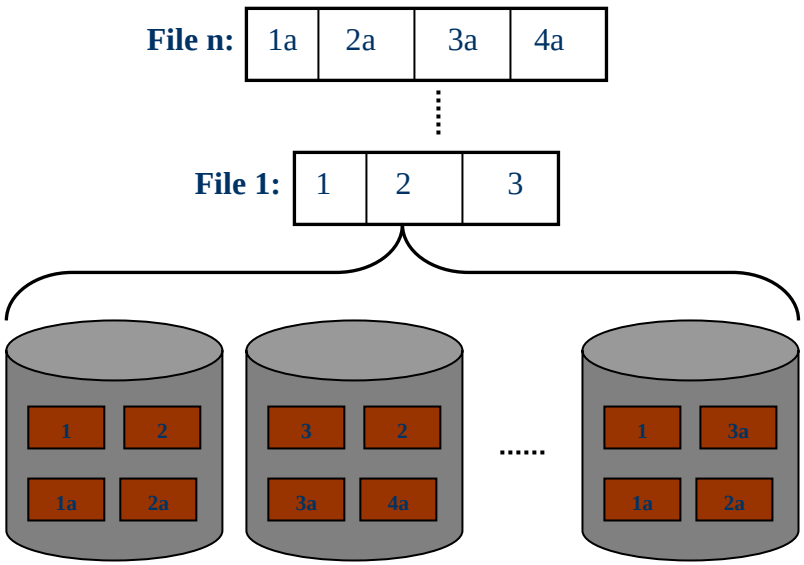
Morsels, Morsels Everywhere…



An architecture for aggregating idle workstation storage space using scavenging.

- Basic Architecture
  - Soft-State Registration
  - Manager Configuration
  - Scavengers
  - File Management
  - Morsel Management
  - Metadata Management
  - Distributed Garbage Collection

User Autonomy, Space Reclaim and Availability



File/Morsel management using the piece-by-piece strategy and availability through morsel replication.

- User Autonomy:
  - Passive Scavenging
  - Proactive Scavenging
- Space Reclaim:
  - User withdrawal
  - Which morsels to relocate/evict?
  - Which scavenger workstations to relocate to?
- Availability:
  - Redundant Array of Replicated Morsels
  - Minimum replication factor for morsels
  - Where to replicate?
  - Which morsel replica to choose from in response to user file fetches?

Other Services

- Grid Awareness:
  - Information Providers
  - Protocol Agnostic
  - Distributed Space Reservations/Pinning
  - Single sign-on
- Performance Optimizations:
  - Aggregate I/O bandwidth
  - Grouping scavengers based on access rates and service history
- Hierarchical Hot & Cold Storage:
  - Hierarchy of high-end storage data centers, aggregated commodity storage and archival storage
  - Move data back and forth in the hierarchy based on metrics
    - Data Aging
    - Data Access Rates

Philosophical Musings...

- Broad Grid Trends:
  - Potentially massive amounts of computing, storage and network BW
  - Resource unavailability is the Law; Not an exception!
  - "Democratizing" data access is definitely the wave of the future

Questions…?

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.