

Logistical Computing and Internetworking in Data-Intensive Environments

Micah Beck, *Assoc. Prof. & Director*

*Logistical Computing &
Internetworking (LoCI) Lab*

Computer Science Department

University of Tennessee

mbeck@cs.utk.edu

Fall Creek Falls Workshop

Oct 28, 2003



LoCI

LOGISTICAL COMPUTING AND
INTERNETWORKING LAB



UNIVERSITY OF TENNESSEE

Logistical Computing and Internetworking Laboratory

- » Part of the University of Tennessee's Computer Science Department
 - Co-Directed by two CS faculty members, 20 lab members total
- » Micah Beck, Assoc. Professor
 - Networking, scalability, computation
- » James S. Plank, Assoc. Professor
 - Fault tolerance, performance



Communication: Foundation of Collaboration

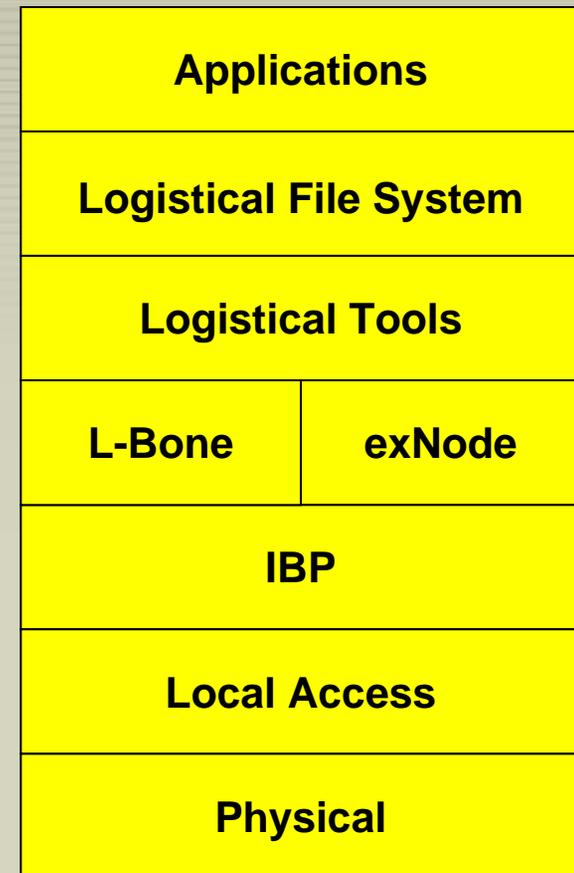
- Synchronous communication
 - » Conferencing, distributed computing
- Asynchronous communication (storage)
 - » Caching/staging/replication
 - » Messaging, single source multicast
 - » Disconnected operations
- Support for Distributed Applications
 - » State management
 - » Extensible network functionality

What is Logistical Networking

- » A scalable mechanism for deploying shared storage resources throughout the network
- » An general store-and-forward overlay networking infrastructure
- » A way to break long transfers into segments and employ heterogeneous network technologies
- » P2P storage and content delivery that doesn't using endpoint storage or bandwidth

The Network Storage Stack

- Our adaption of the network stack architecture for storage
- Like the IP Stack
- Each level encapsulates details from the lower levels, while still exposing details to higher levels



IBP: The Internet Backplane Protocol

- » Storage provisioned on community “depots”
- » Very primitive service (similar to block service, but more sharable)
 - Goal is to be a common platform (exposed)
 - Also part of end-to-end design
- » Best effort service – no heroic measures
 - Availability, reliability, security, performance
- » Allocations are time-limited!
 - Leases are respected, can be renewed
 - Permanent storage is too strong to share!

The Network Storage Stack

LoRS: The Logistical Runtime System:
Aggregation tools and methodologies

The L-bone:
Resource Discovery
& Proximity queries

The exNode:
A data structure
for aggregation

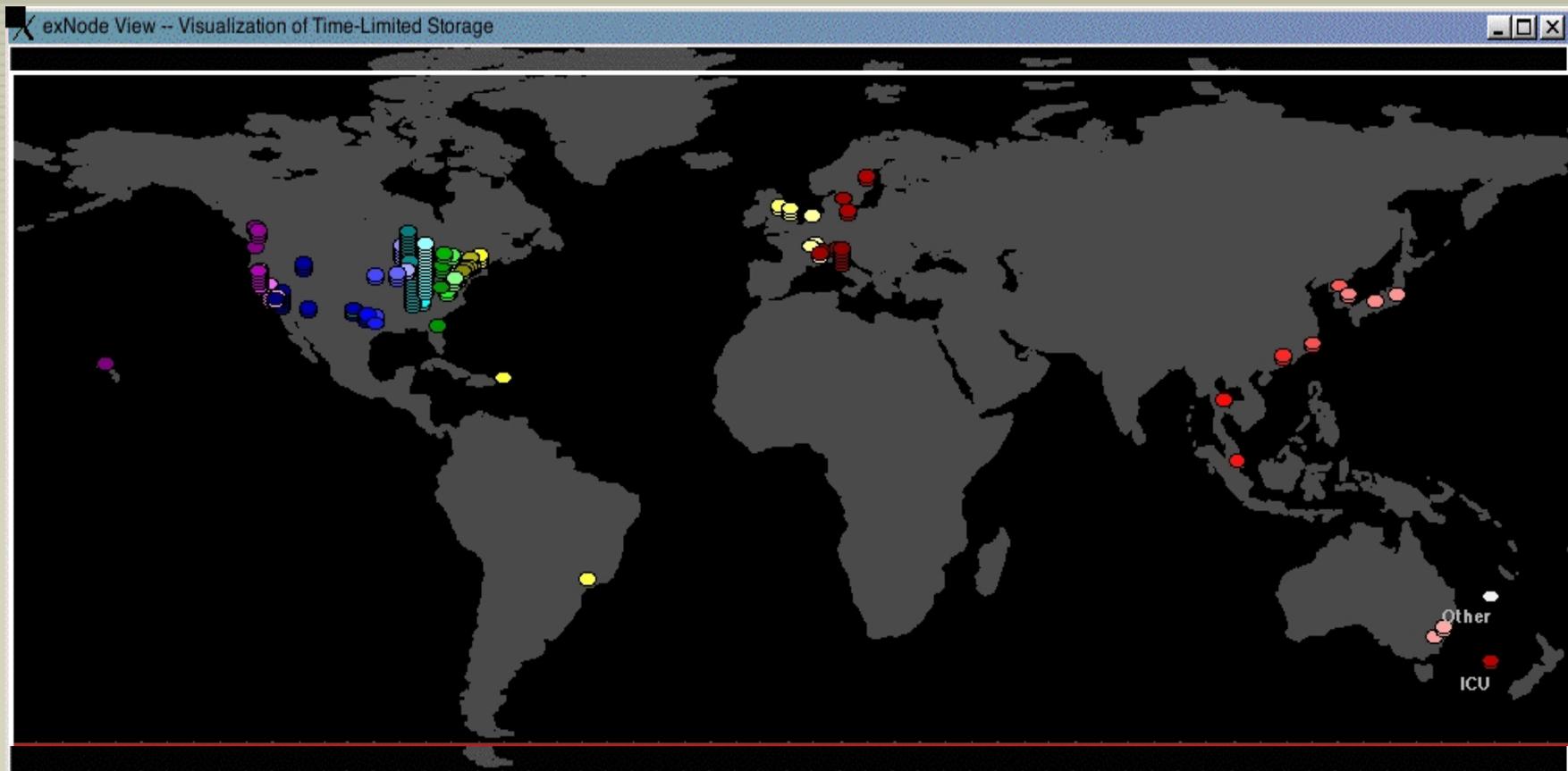
IBP: Allocating and managing network
storage (like a network malloc)

The Backbone Storage Resources (L-Bone)

- » LDAP-based storage resource discovery.
 - Query by capacity, network proximity, geographical proximity, stability, etc.
 - Periodic monitoring of depots.

- » Multiple shared storage pools
 - Nat. Logistical Networking Testbed (NSF)
 - » 22TB today; 50TB by 2005, 100TB goal
 - Energy Sci. Logistical Net. Testbed (DOE)
 - » 8TB in 2003 to support SciDAC projects

L-Bone: August 2003 (20TB)



IBP Deployment

- » Depots/collaborations supporting DOE projects
 - ORNL, NC State, SUNY Stony Brook, UCSD
 - NERSC (security issues)
- » Initial Configuration:
 - Dell Server Running Linux Red Hat
 - SAN-attached IDE RAID arrays (1.6 TB each)
 - 2 GigE NICs used where available
- » Direct connectivity to 10Gb/s router planned in NC
- » 4 TB available at Starlight (NSF NLNT)
 - Another 7 TB purchased

The Network Storage Stack

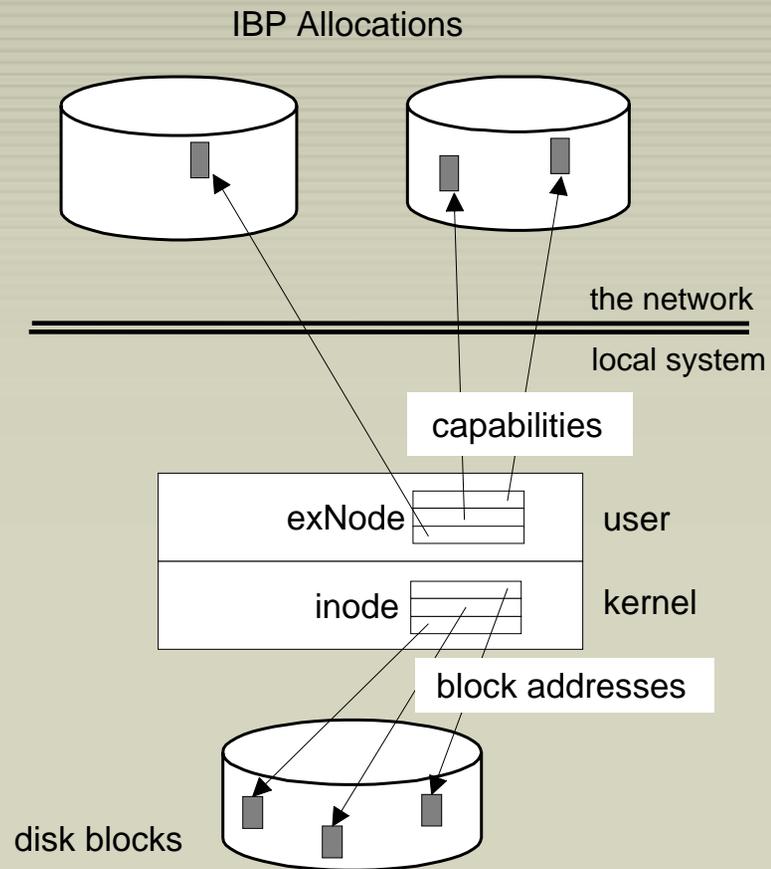
LoRS: The Logistical Runtime System:
Aggregation tools and methodologies

The L-bone:
Resource Discovery
& Proximity queries

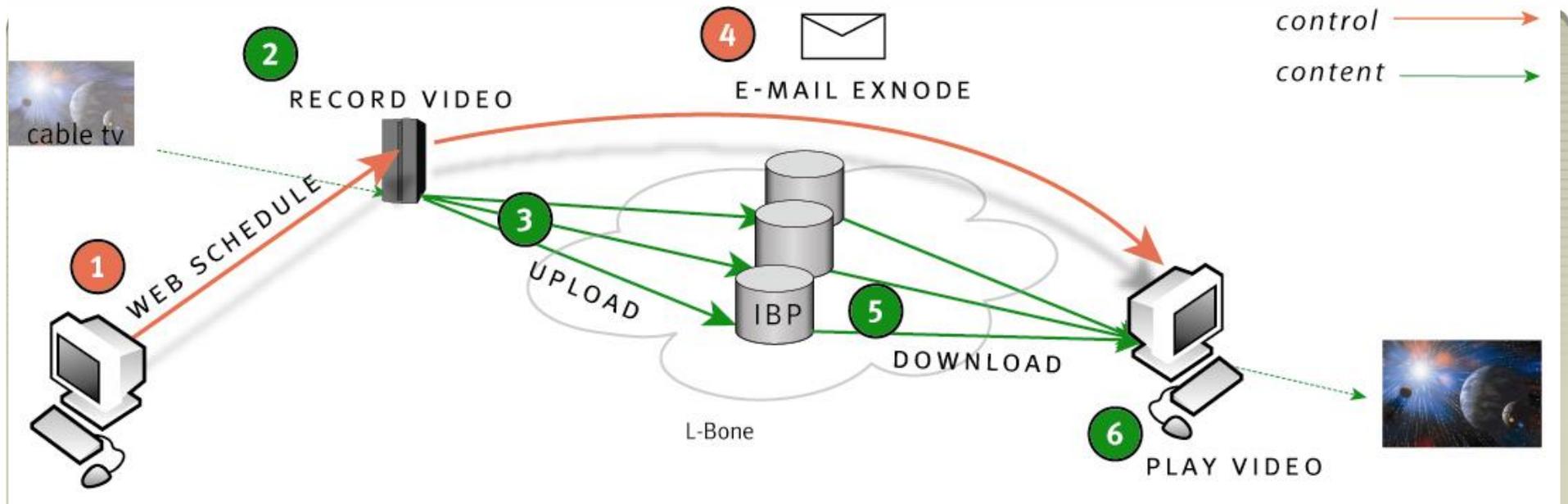
The exNode:
A data structure
for aggregation

IBP: Allocating and managing network
storage (like a network malloc)

ExNode vs inode



IBPvo: PVR with the ExNode



- Video recorded at source (1 GB = 1 hour)
- Upload to IBP depots, exNode created
- exNode mailed to recipient
- Download or streamed to recipient

The Network Storage Stack

LoRS: The Logistical Runtime System:
Aggregation tools and methodologies

The L-bone:
Resource Discovery
& Proximity queries

The exNode:
A data structure
for aggregation

IBP: Allocating and managing network
storage (like a network malloc)

Logistical Runtime System

- » Basic Primitives:
 - Upload, Download, Augment, Refresh
- » End-to-end Services
 - Checksums, Encryption, Compression
- » Other Things We Can Do
 - Routing through an intermediate depot to reduce IP RTT, speeding up TCP transfers
 - Overlay multicast using either multiple TCP streams or IP multicast at tree nodes

TSI Site Deployment: ORNL, NCSU, SUNY Stony Brook, NERSC, UCSD (8TB)

The screenshot displays the LoRS Command interface, which is used for managing data distribution across various sites. The interface is divided into several sections:

- LoRS Command Window:** Contains a menu bar with options: Upload, Download, Add_Copy, Refresh, Delete, Other.
- Necessary Parameters:** Fields for selecting an exNode file to augment (ORNL.xnd) and saving the new exNode to (NEW.xnd). Location is set to state= NC and city= Ralleigh.
- Optional/Advanced Parameters:** Includes exNode Structure (Copies: 1, Blocksize: 2M), Data Condition (Duration: 1d, Allocation Type: soft), and Augment Performance (Threads: 40, Max Depots: 8). A checkbox for "Use TCP Datavers (MCOPI)" is present.
- Buttons:** Quit, Stop, List, Add_Copy Now, and LoRS Command 0.81.

The right side of the interface features a map of the United States with numerous data points represented by small cylinders. A red bar at the bottom of the map area displays the following information:

- Filename 200MB.txt
- Size 209715200
- SetCopy
- Copying..
- 138.1559 Mbps

A vertical progress bar on the right side of the map area shows the status of the copy operation, with a red bar indicating progress. At the bottom of the interface, there are buttons for Quit, Clear, Draw Arrows, and 1 Active Connection(s).



Multithreaded Download (Streaming)

The screenshot displays the XDarwin environment with several windows. The main window, titled "exNode View - Visualization of Time-Limited Storage", shows a map of the United States and Europe with various exNodes marked. Blue arrows indicate active connections from nodes like UCSD, TAMU, IUPUI, and UNC to a central download point. Below the map, a "Downloading" section shows a progress chart with 26 colored bars representing individual threads, each labeled with a number and a date (e.g., "00 Aug 21", "01 Aug 21", etc.).

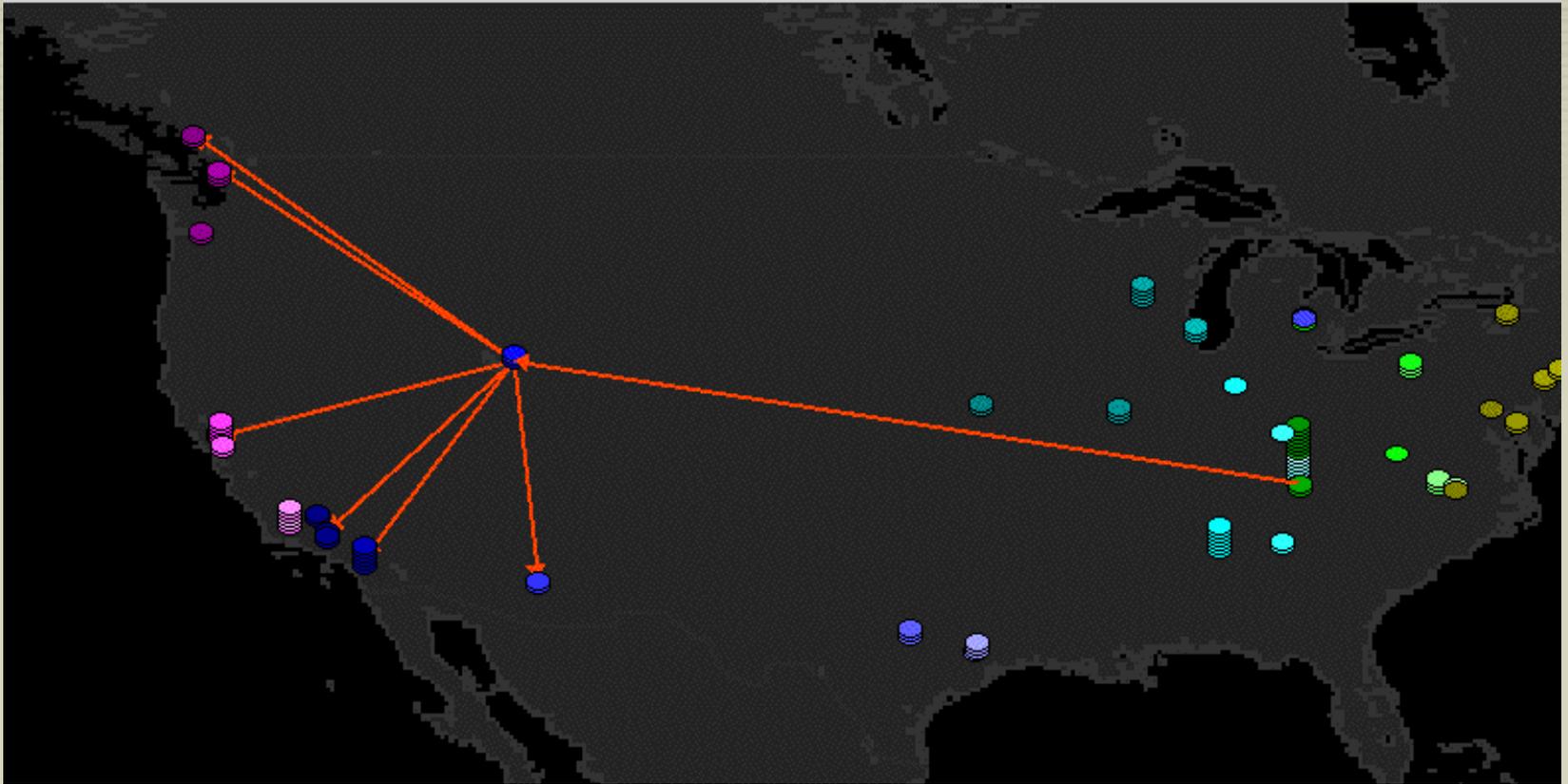
Other windows include a terminal window showing download progress for multiple threads, an email list, and an "exNode Command" window with the following settings:

- Mode: download
- Select an exNode to download: /Users/atchley/Movies/o- (Browse..)
- Download the file to: /dev/null (Browse..)
- Threads: 6
- Prebuffer: 1
- Blocksize: 5120 K
- Buttons: Kill, List, Run

At the bottom of the interface, there are control buttons: Clear, Slow Redraw, Quick Redraw, Draw Arrows, Active Connection, and Quit. The UT logo is visible in the bottom right corner.



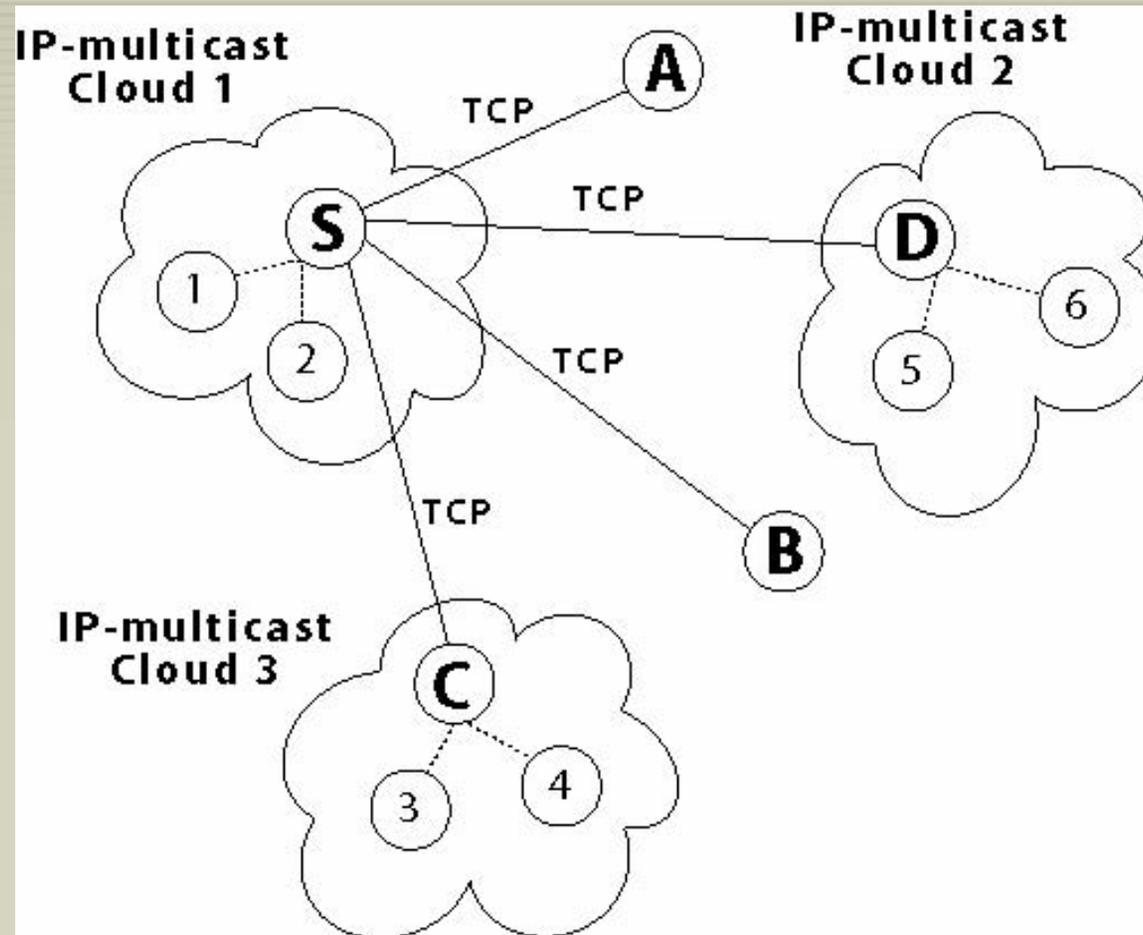
Point-to-Multipoint



LDCI



Heterogeneous Asynch. Multicast



IBP Enables Data Intensive Collaboration

- » Large files can be uploaded to nearby depots, then managed by movement between depots
 - End systems are not involved in long distance transfers
- » Data can be moved near to distant collaborator without being downloaded into their end system
 - Direct access to collaborators private storage is not required
- » Depot-to-depot transfers can take advantage of multithreading, UDP transfer, Net/Web 100, other high-performance optimizations



The Rest of the Talk

- » Goals and Objectives
 - Design and development of the IBP depot
 - Design and development of Logistical Networking middleware
 - Application Impact
- » Technical Approach & Accomplishments
 - IBP depot
 - Middleware Services
 - Application and User Support
- » Terascale Supernova Initiative
- » Global research participation
- » Plans and Futures



IBP depot

- » IBP as RPC over TCP
- » Data Movers Plug-in Modules
- » Multi-resource depots
- » Encapsulated Data Movers
- » Persistent sockets for optimization, security and pipelining
- » Porting depot and client code
- » Computation in place (Network Functional Unit)

Middleware Services

- » Logistical Runtime System (LoRS)
- » Logistical Backbone (L-Bone)
with NWS integration
- » Latency hiding through aggressive prestaging
- » File services using Logistical Networking
infrastructure
- » Exposed multicast and routing
- » Integration of Network Weather Service



Latency hiding through aggressive prestaging

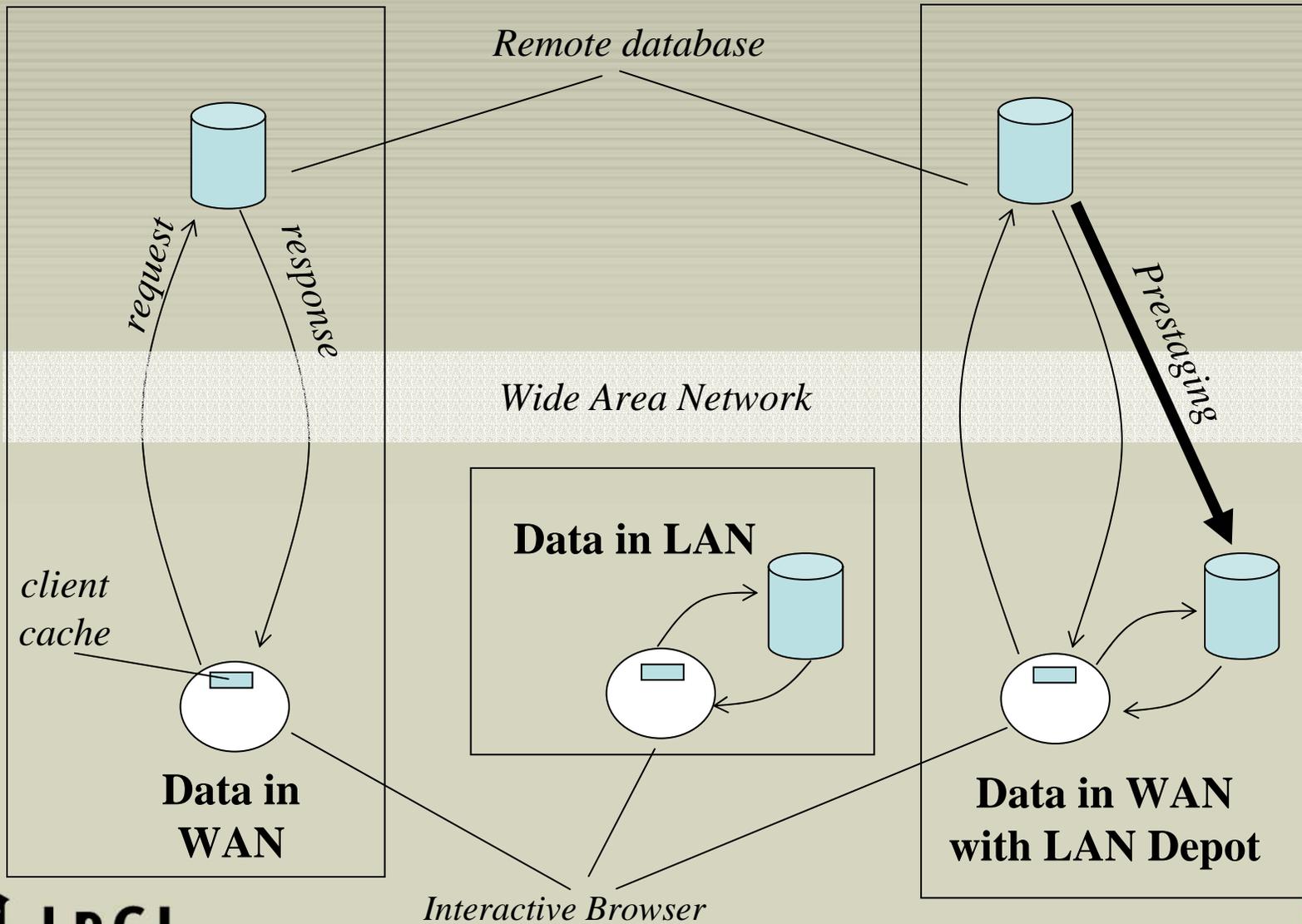
Remote Visualization by Browsing Image Based Databases with Logistical Networking

*Jin Ding, Jian Huang, Micah Beck, Shaotao Liu, Terry
Moore, and Stephen Soltesz*

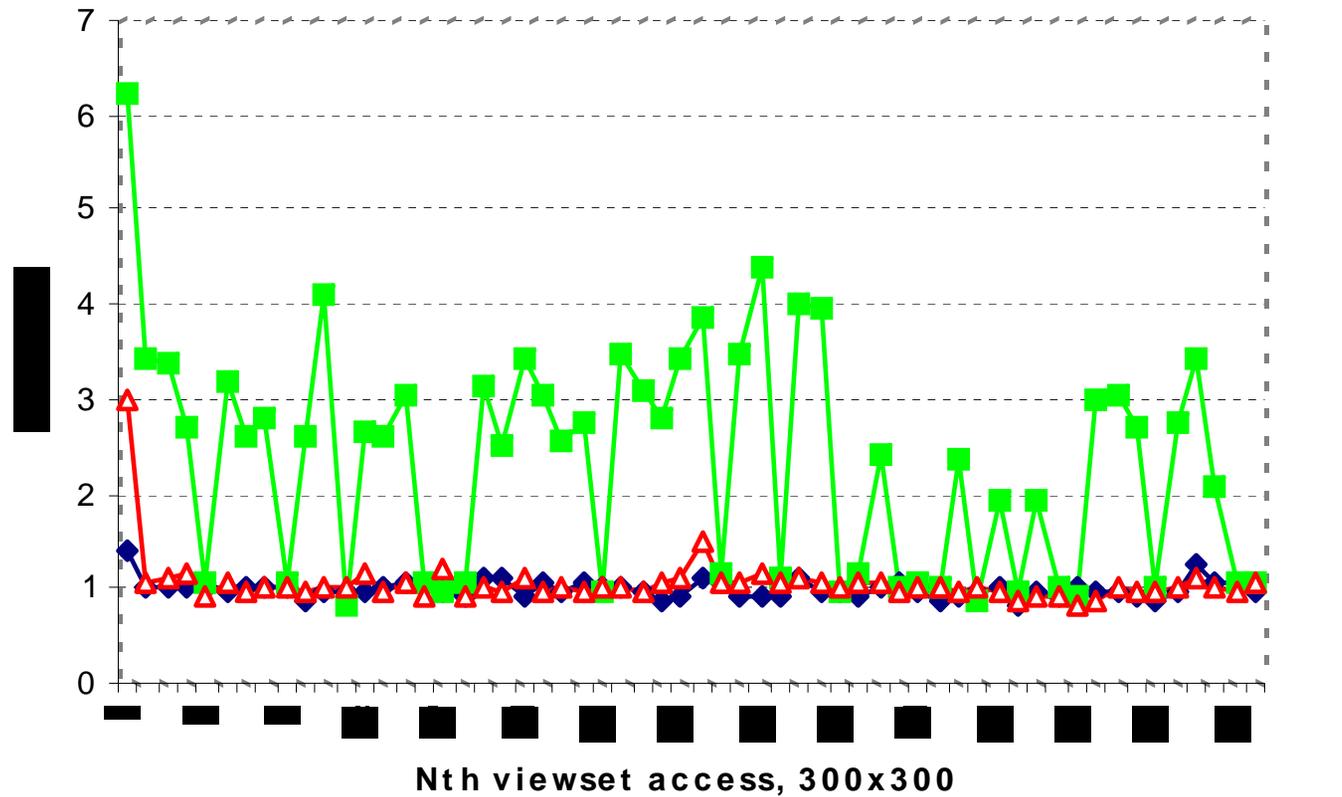
To appear in SC 2003, Phoenix, AZ, November, 2003



Latency hiding through aggressive prestaging

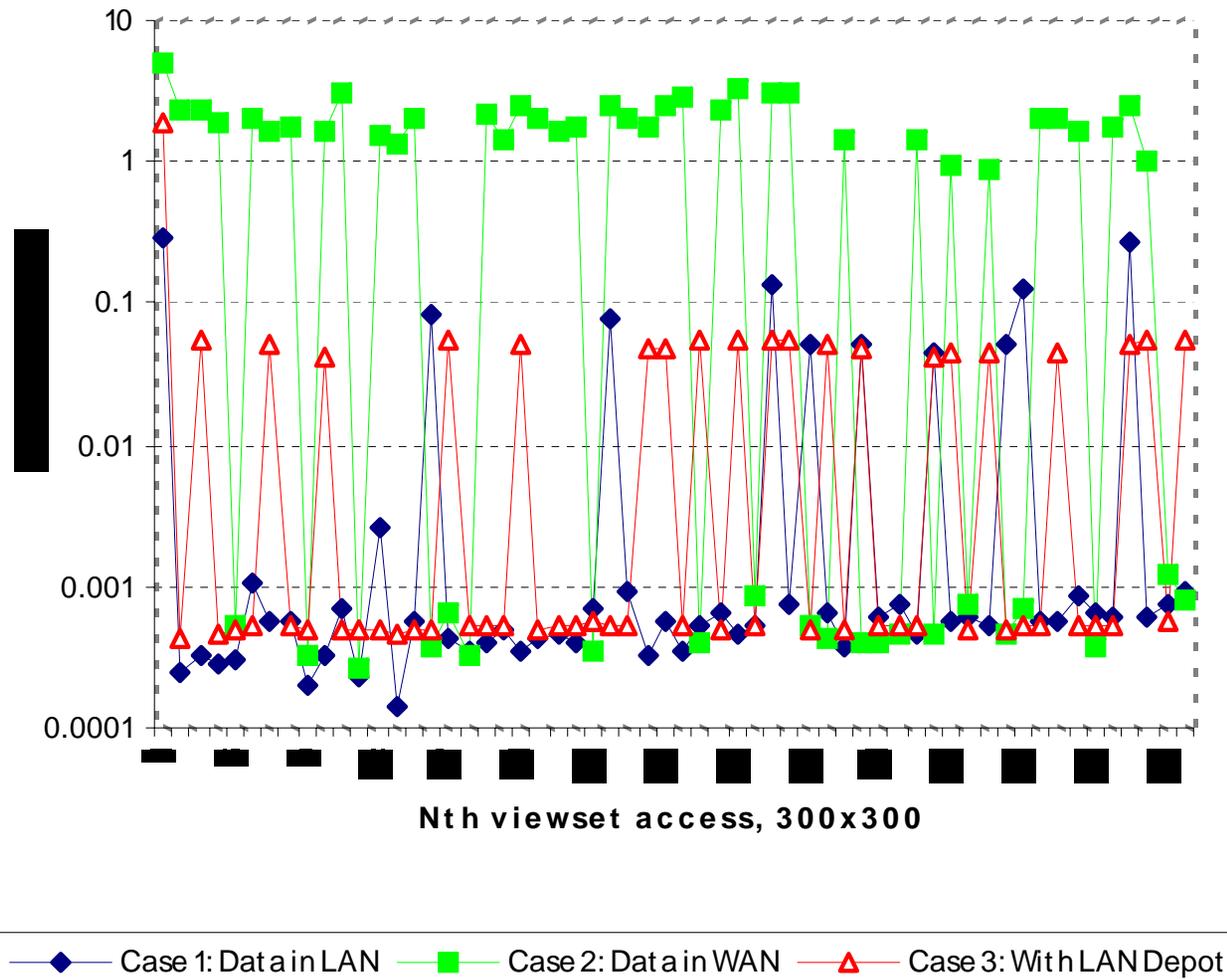


User Latency, 300x300 resolution

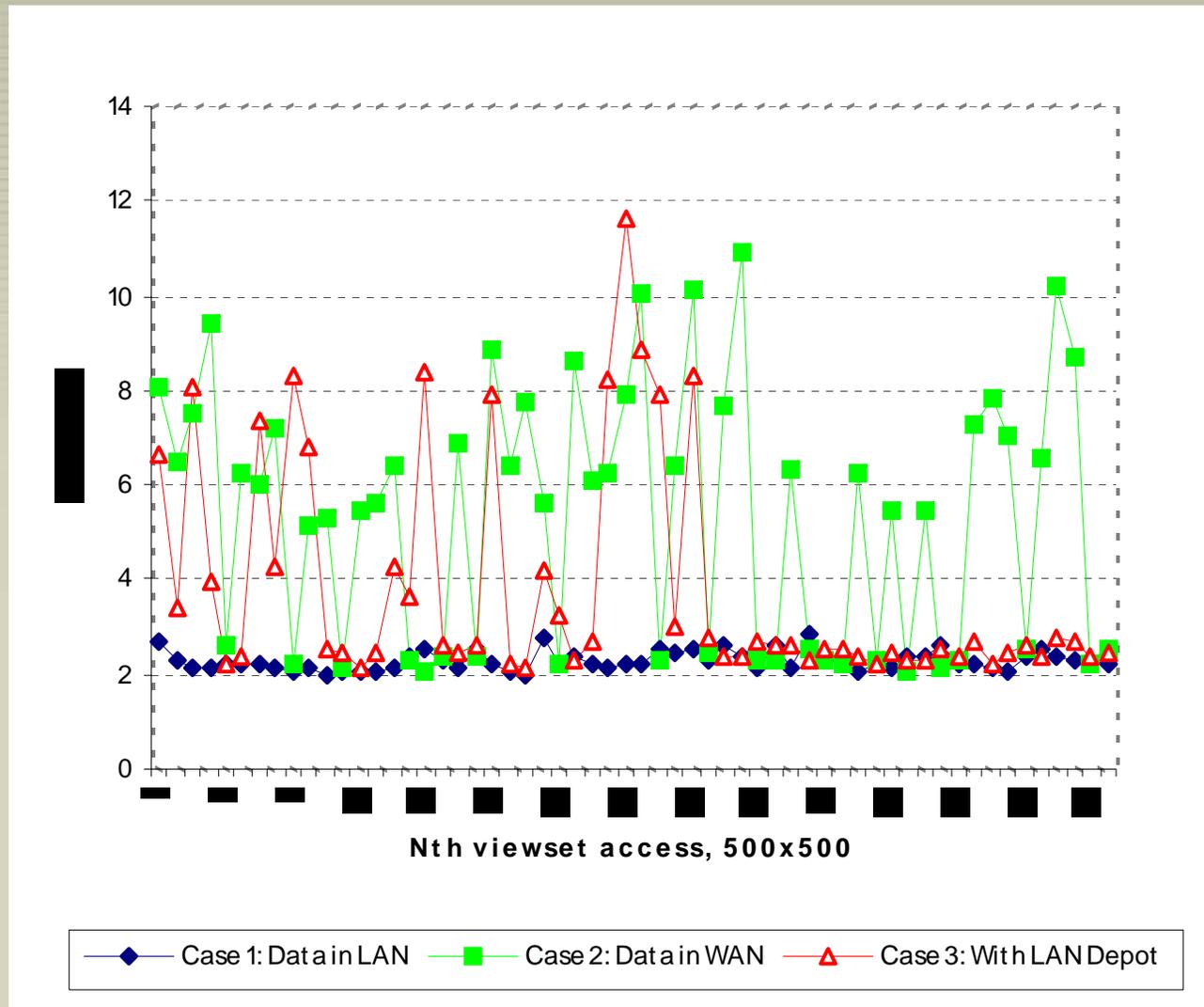


—◆— Case 1: Data in LAN —■— Case 2: Data in WAN —△— Case 3: With LAN Depot

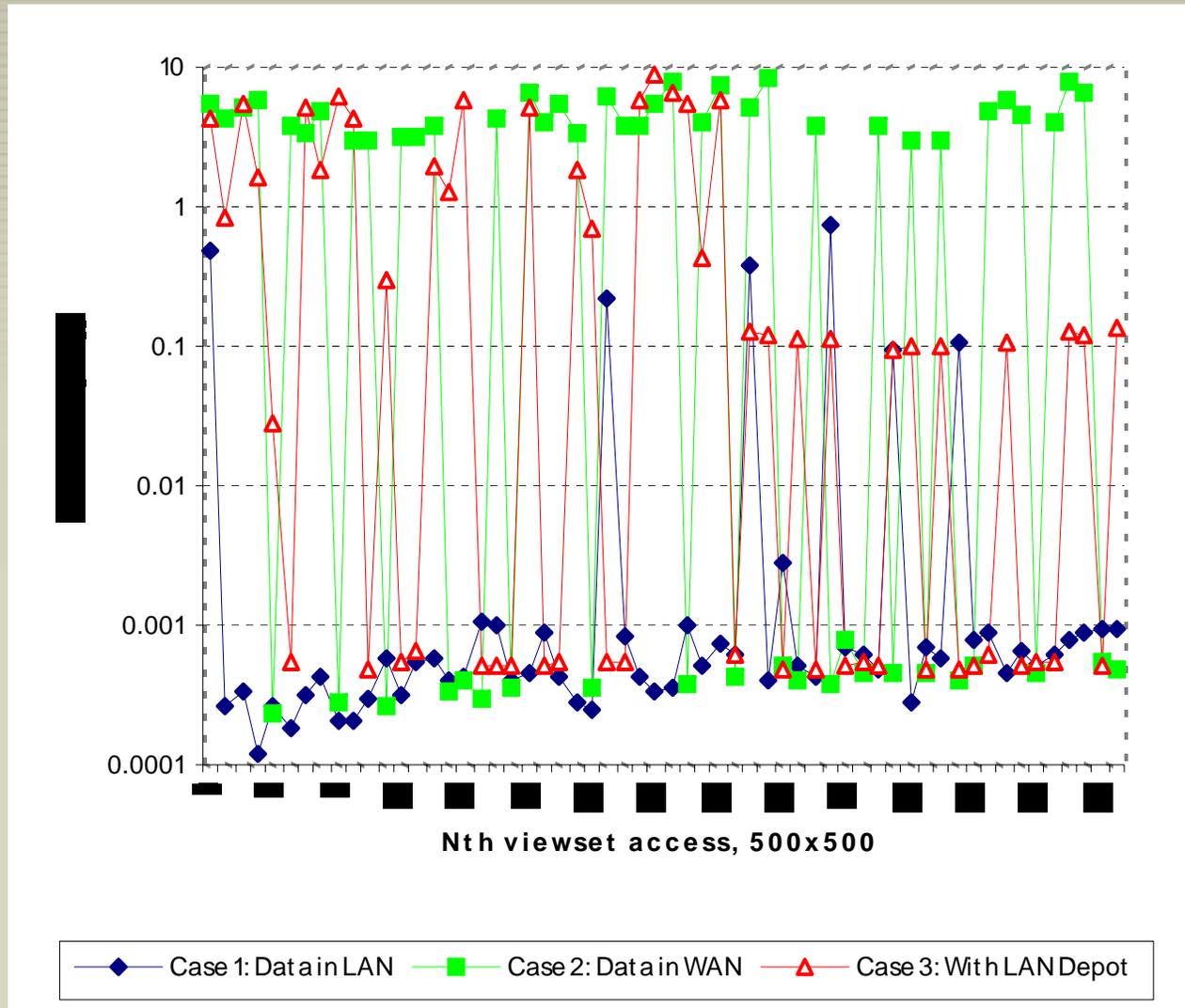
Network Latency, 300x300



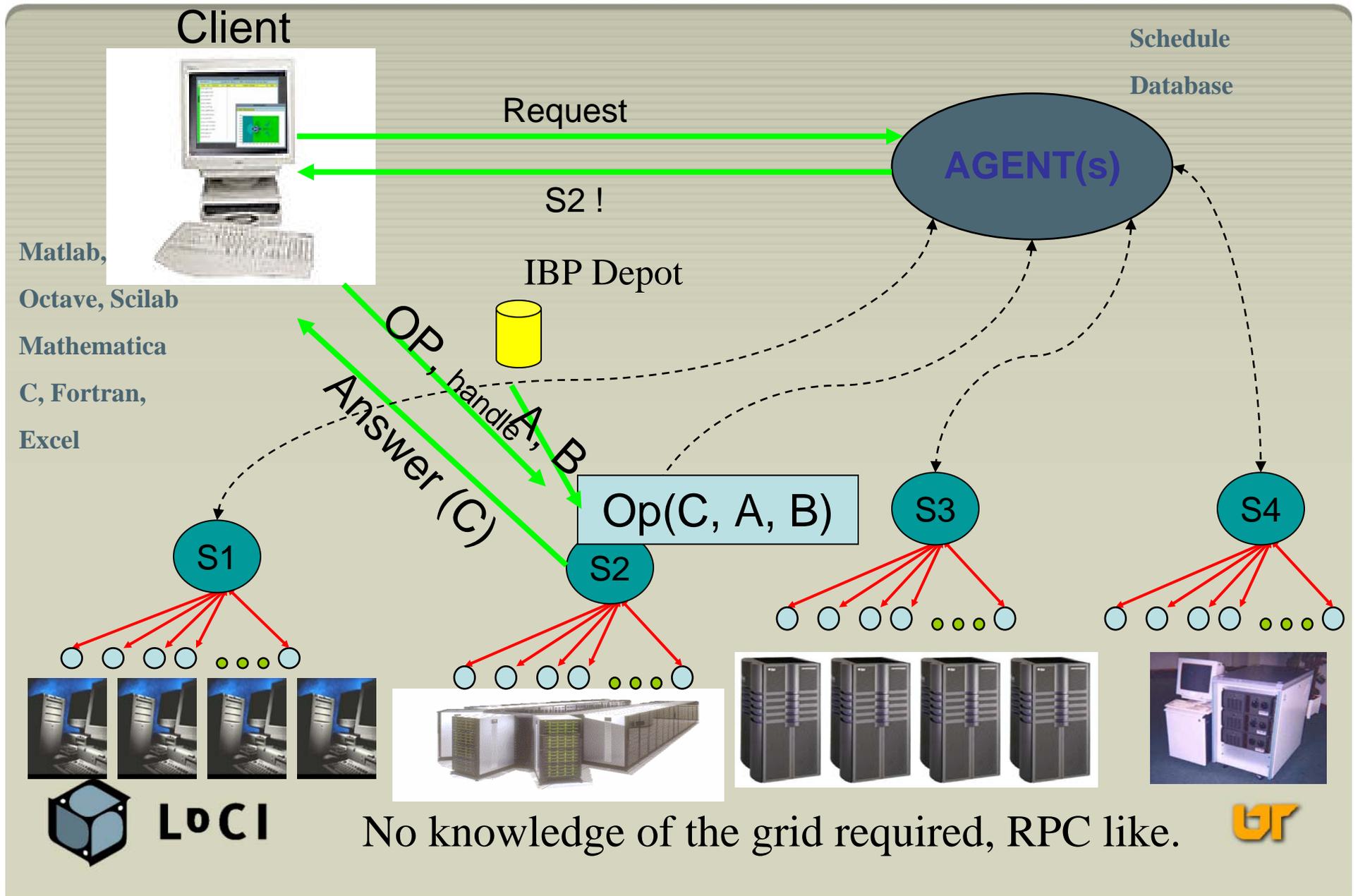
Client Latency, 500x500 Resolution



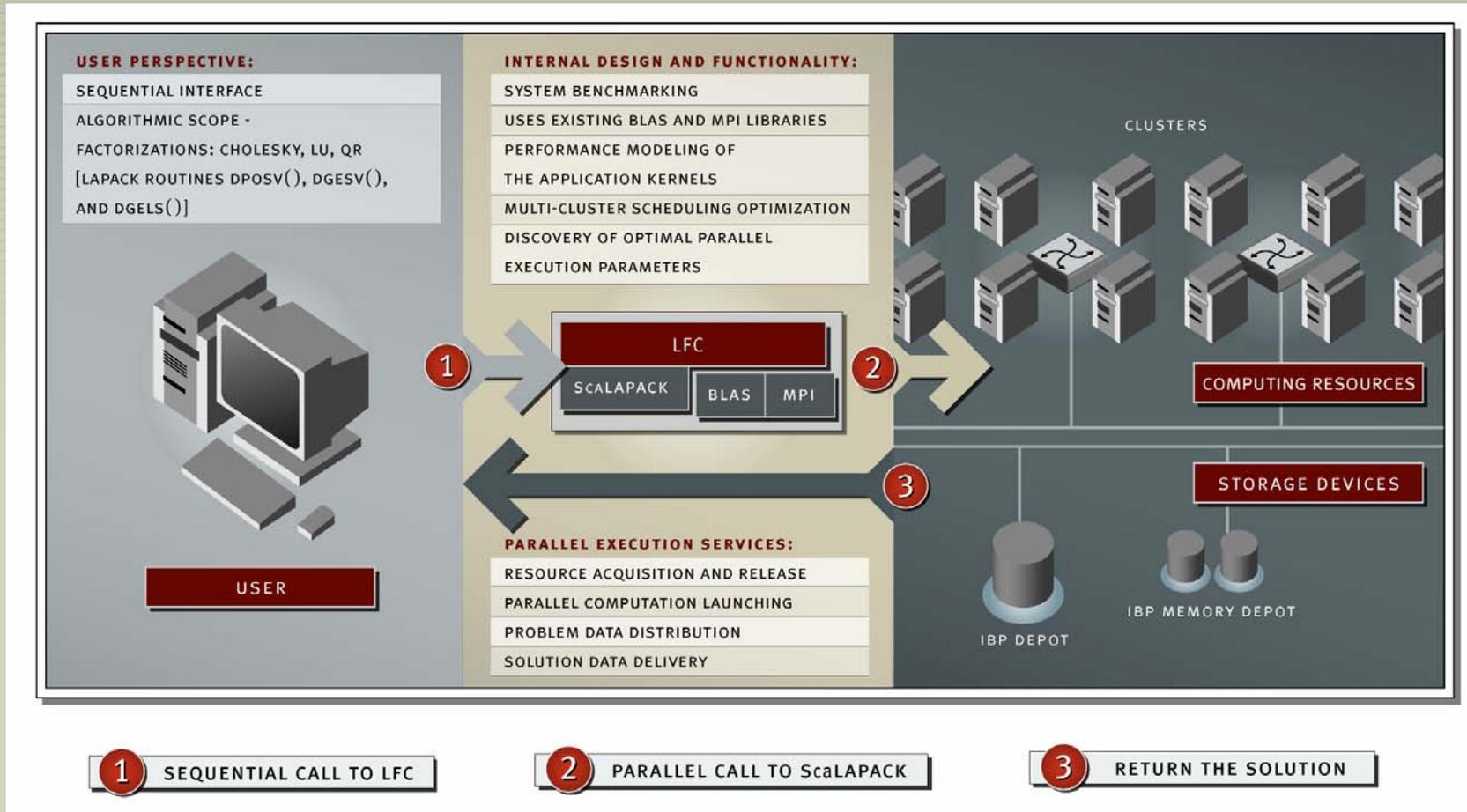
Network Latency, 500x500



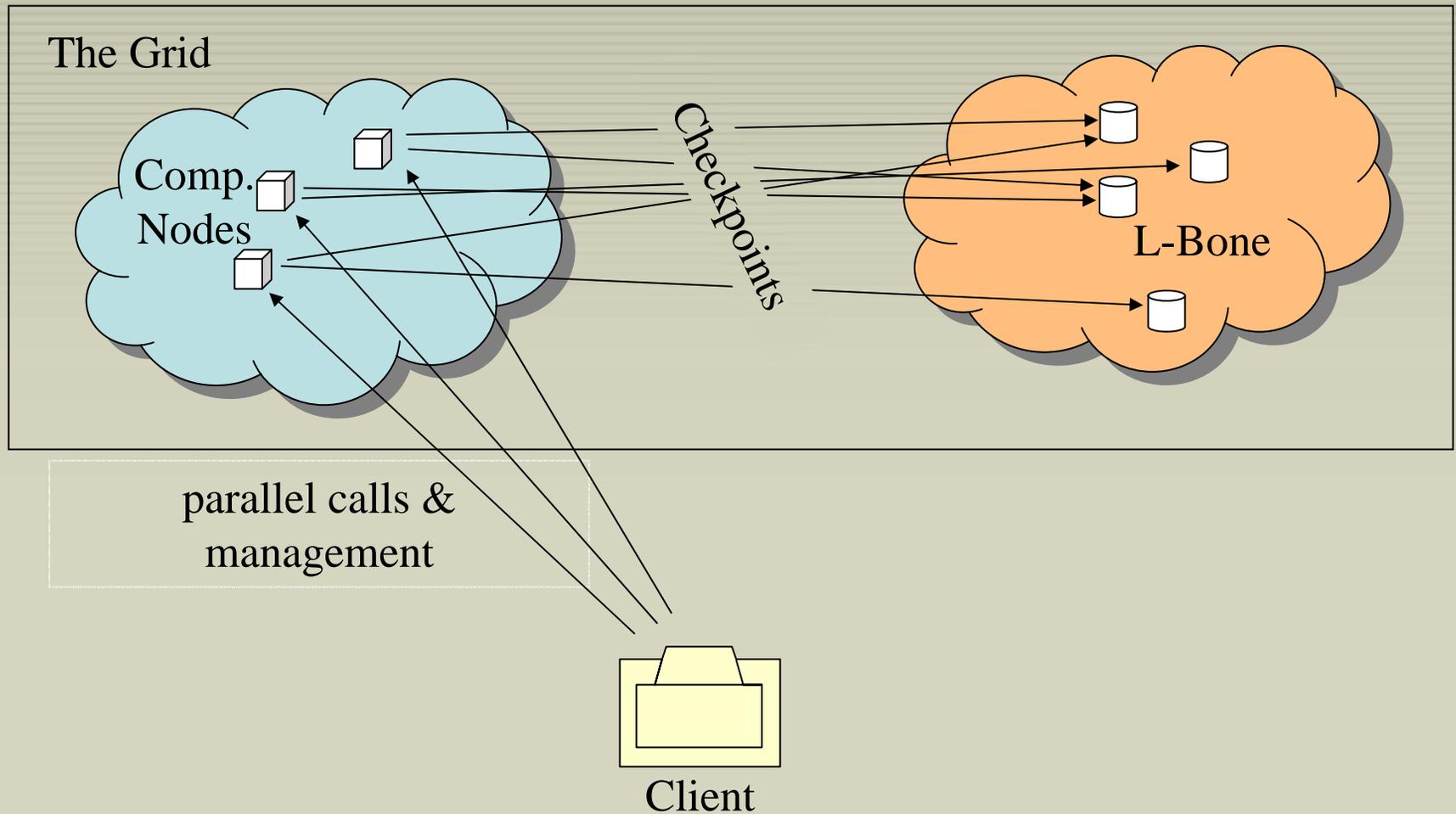
Integration with NetSolve/GridSolve



LAPACK for Clusters



Checkpointing in GridSAT



Terascale Supernova Initiative (TSI)

- » Exploring alternative architectural approaches
- » Project with many participants, spanning DOE Labs and universities
- » Initial impact: Move data between ORNL and NCSU at 2-300Mbps
- » Wider impact: Deployment at all major TSI collaboration sites (8TB)
- » Future impact: Integration with and optimization of TSI workflow



A Logistical Visualization Scenario

- » Terascale datasets at ORNL (TSI)
- » Visualization researchers at Tennessee
 - Postprocessing to compute correlations
 - Cluster computing used at Tennessee
 - No terascale storage at Tennessee
- » Data movement overshadows computation
- » Computation can process datasets sequentially
- » Solution:
 - Upload to IBP, process in batches
 - Use available storage for intermediate buffering

Engagement with Climate Modelling Community

- » Post-Doc Researcher to engage with Climate Modeling Projects
 - UT Science Alliance funding in support of UTK/ORNL Joint Institute for Computational Science (JICS)
- » Initial focus: satellite data from Southeast U.S.
 - Capture and distribution of data in support of researchers at Oak Ridge Associated Univ.
 - Indexing and management issues
- » Leverage DOE & public depot infrastructure for international collaborations

Plans and Futures

- Further integration with TSI workflow reaching other SciDAC applications
- Extensible depot services
- Layer 2 and optical depot connectivity
- Logistical toolkit for remote and distributed visualization
- Address "Data Grids"
- Exploit clusters (Feng; LANL Green Destiny)
- Adoption of networking best practices
- Standardization and commercial adoption
- Peering between distinct logistical networks
- 100TB deployed throughout DOE labs and partner institutions
- Client integration with all common DOE platforms



Conclusions

- » Logistical Networking is a new architectural approach at the fabric layer
- » Emphasis on Internet-like scalability of shared storage resources
- » A major challenge is convincing collaborations that architectural innovation is required
- » Stable open source software has been delivered and integration with applications is increasing
- » Resource deployment is proceeding successfully
- » Policy integration has begun