HPC Software Futures

Rama K Govindaraju
HPC Software Architect
CINECA, SP Scicomp 2004
ramag@us.ibm.com

Special Notices

This presentation was produced in the United States. IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the products, programs, services, and features available in your area. Any reference to an IBM product, program, service or feature is not intended to state or imply that only IBM's product, program, service or feature may be used. Any functionally equivalent product, program, service or feature that does not infringe on any of IBM's intellectual property rights may be used instead of the IBM product, program, service or feature.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. Sources for non-IBM list prices and performance numbers are taken from publicly available information including D.H. Brown, vendor announcements, vendor www Home Pages, SPEC Home Page, GPC (Graphics Processing Council) Home Page and TPC (Transaction Processing Performance Council) Home Page. IBM has not tested these products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

BM may have patents or pending patent applications covering subject matter in this presentation. The furnishing of this presentation does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of a specific Statement of General Direction.

The information contained in this presentation has not been submitted to any formal IBM test and is distributed "AS IS". While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM is not responsible for printing errors in this presentation that result in pricing or information inaccuracies.

The information contained in this presentation represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

All prices shown are IBM's suggested list prices; dealer prices may vary.

IBM products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Information about non-IBM products was obtained from suppliers of those products. IBM makes no representations or warranties regarding these products. Non-IBM products are offered and warranted by third-parties, not IBM.

Special Notices

Outline

- Cluster and Parallel File systems (GPFS)
  - Data Grid Enablement
    - GPFS multi-cluster
    - SANergy
    - NFS V4
- Protocols
  - RDMA (Remote DMA) enablement for pHPS
  - Striping Options
- Scheduler and Resource Manager (LL)
    - Advanced Features
    - Grid Enablement
    - Integration with the Globus toolkit
- AIX, Libraries (ESSL, PESSL), CSM

File System Enhancements

- Grid Engagements
- GPFS Multi-cluster
- SANergy
- NFS V4
- File systems and databases
- Security issues

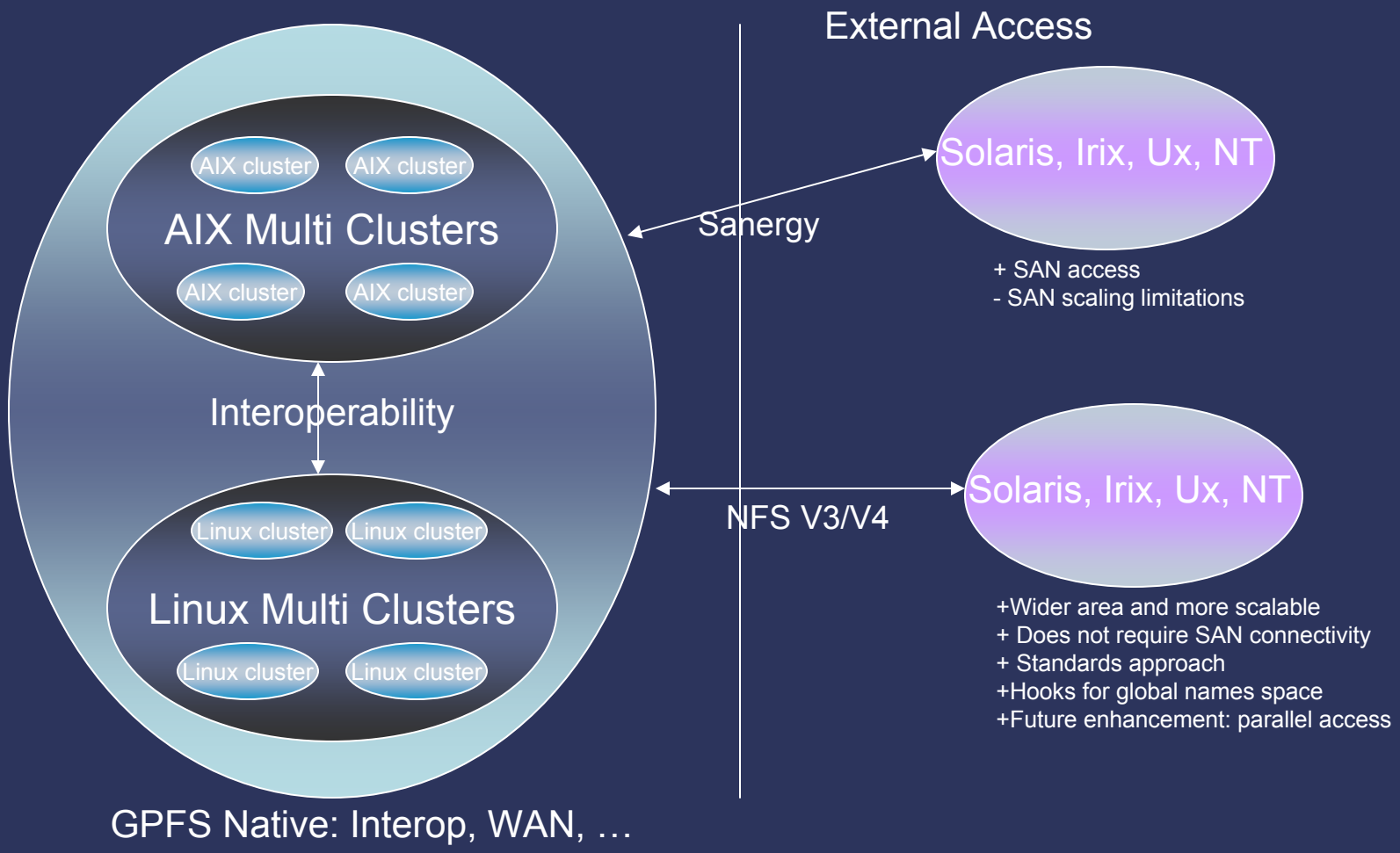
Grid Collaborative Environments

- DTF (Distributed Tera-scale Facility)
- DEISA (Distributed European Infrastructure for Scientific Apps)
- DOE Science Grid (NERSC, …)
- UK Science Grid
- Charles Schwab
- University of Pennsylvania
  - NDMA
- GSA (Global Storage Architecture)
- The Cancer Grid
- MIT Bio Grid
- Ford
- Many other actual engagements
  - Helping us gain insight into problems
  - Helping us reshape our system architecture and designs to meet these challenges

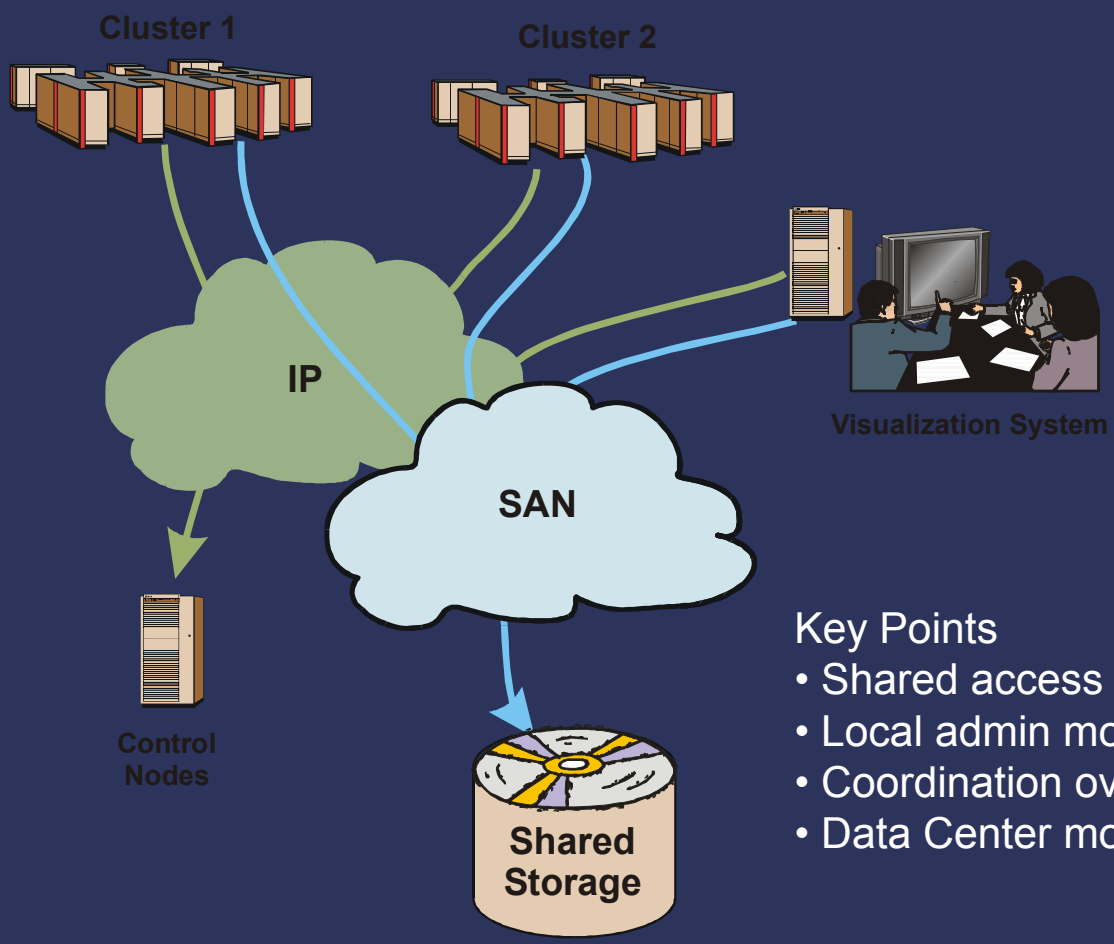
Data Grid - Requirements

- Need data accessible from anywhere in the Grid
  - Enabling Technology: GPFS: Cluster and Parallel File System
    - AIX/Linux Interoperability
    - GPFS multi-cluster (single and multiple admin domains)
    - GPFS over a WAN – cross cluster mounts
- Global name space
  - Enabling Technology: NFS V4 referral service
- High Speed Access
  - Enabling Technologies: GPFS parallel access, NFS V4 replication services, NFS V4 Compound RPC mechanisms
- Load Balancing
  - Enabling Technology: NFS V4 redirection services
- Security and administration
  - Enabling Technologies
    - **Globus GSI, uid field enhancements, EIM, LDAP, GSS-APIs**

GPFS Data Grid: Landscape



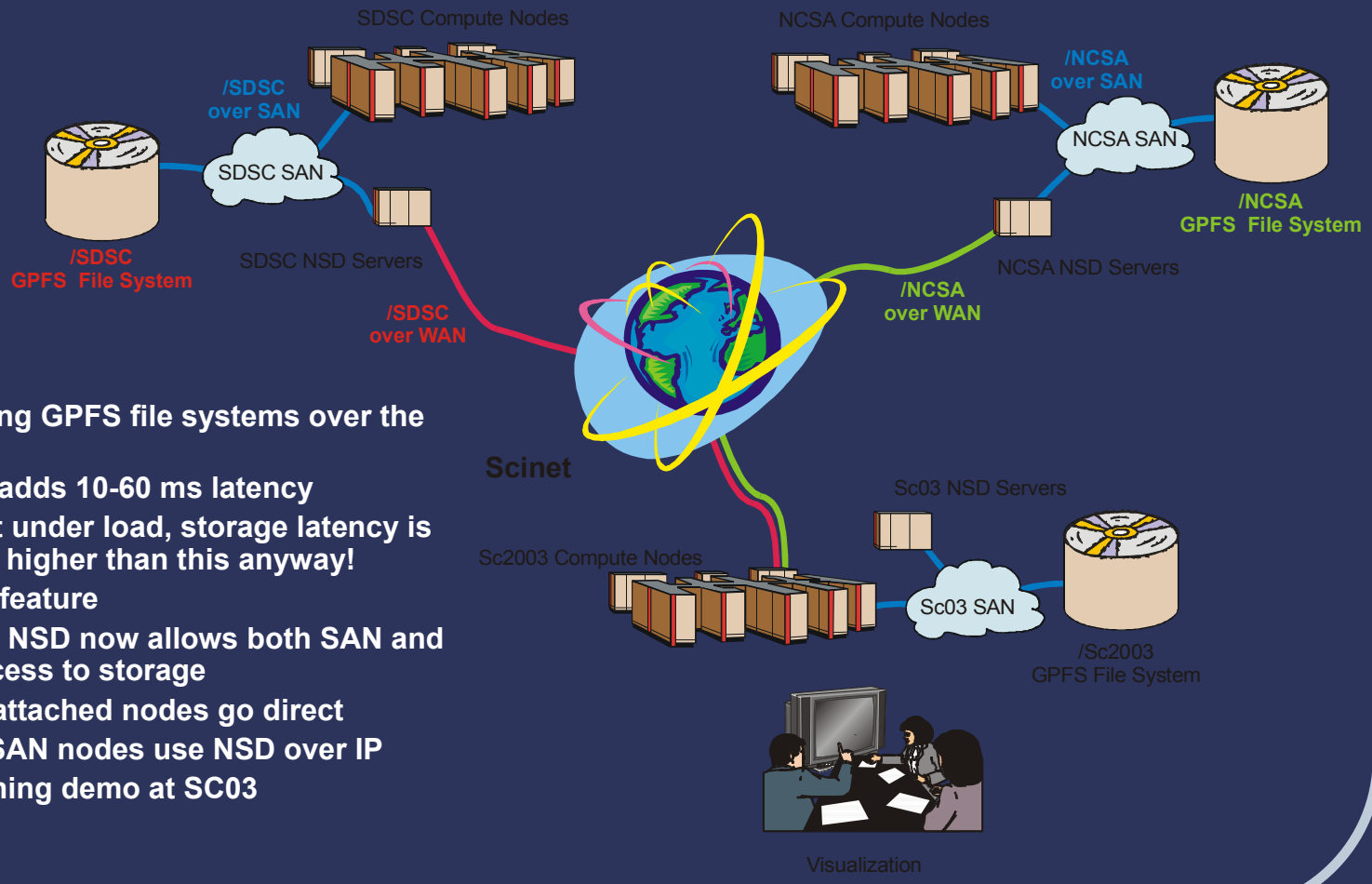
GPFS multi-cluster model



Key Points

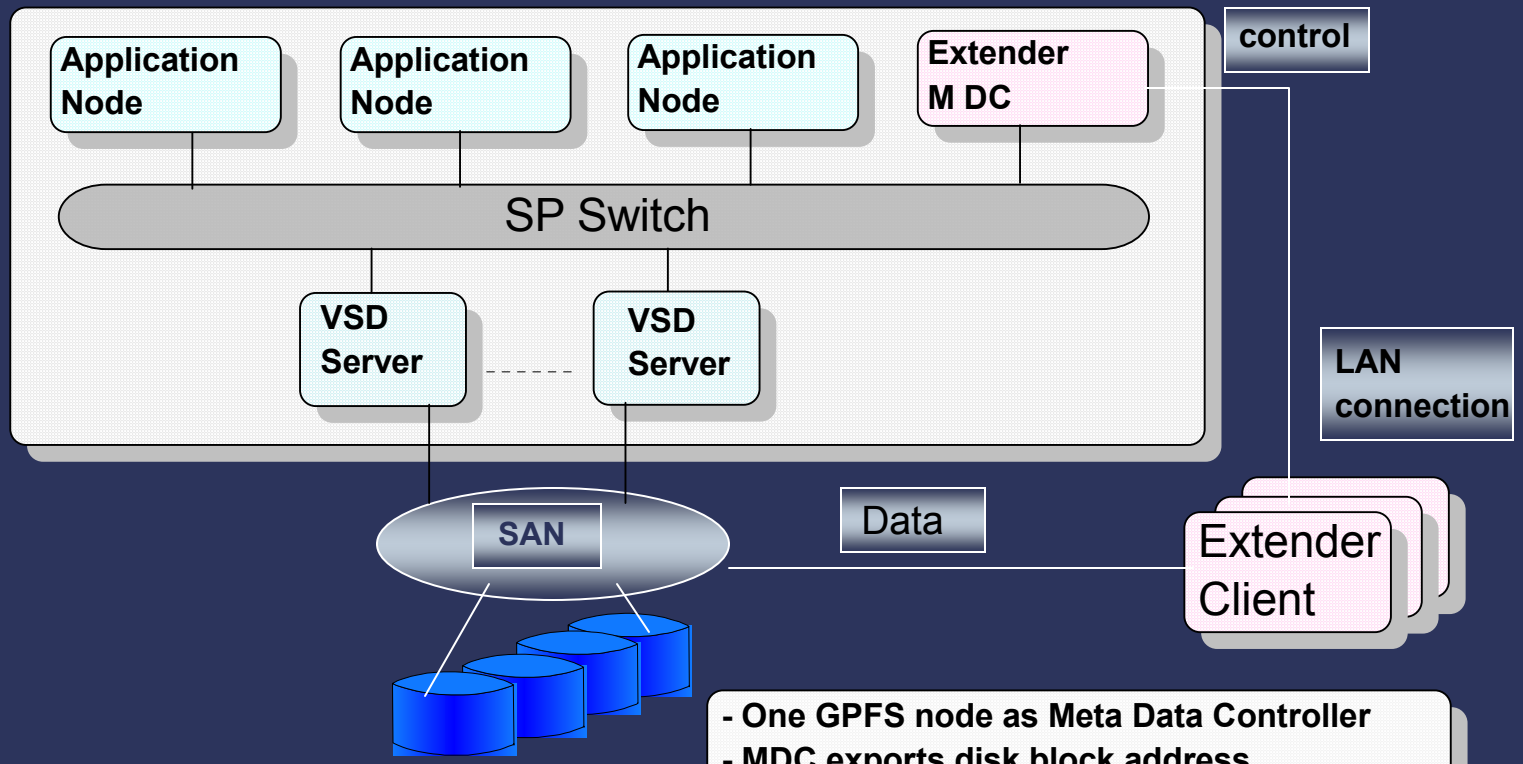
- Shared access to data over SAN (SW)
- Local admin model
- Coordination over IP network
- Data Center model extending to WAN

Access to GPFS over the WAN



- Goal: sharing GPFS file systems over the WAN
  - WAN adds 10-60 ms latency
  - … but under load, storage latency is much higher than this anyway!
- New GPFS feature
  - GPFS NSD now allows both SAN and IP access to storage
  - SAN-attached nodes go direct
  - Non-SAN nodes use NSD over IP
- Award winning demo at SC03

GPFS Enhancements: Sanergy



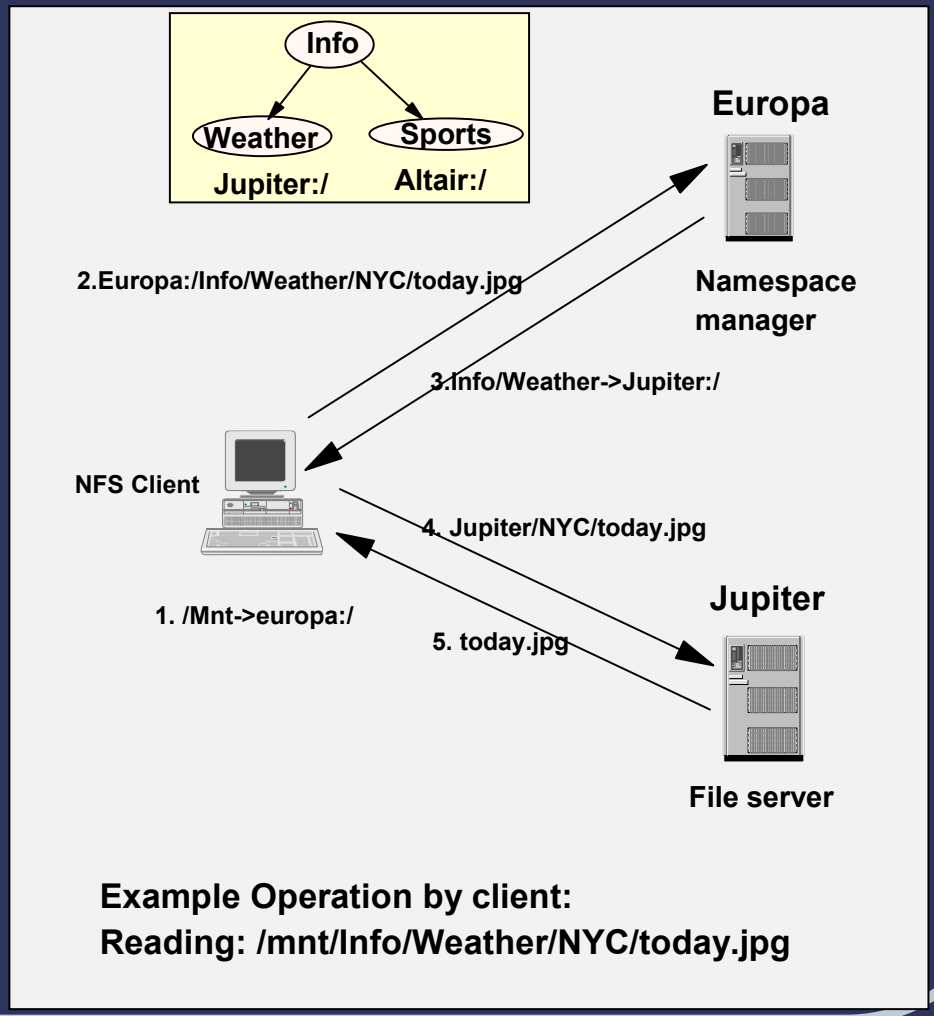
SANERGY support In GPFS 2.2

Data Grid: NFS V4 key attributes

- Attributes
  - Named attributes enabling a capability handshake
  - Referral Service (also called redirection)
    - Load balancing
    - Hook to federate a global name space across multiple cluster file systems
  - Compound RPC mechanisms
  - Replication and parallel access
  - Security enhancements
- AIX enablement first: GPFS exploitation subsequent to that
  - Advanced features come in later releases

Data Grid: Global Name Space

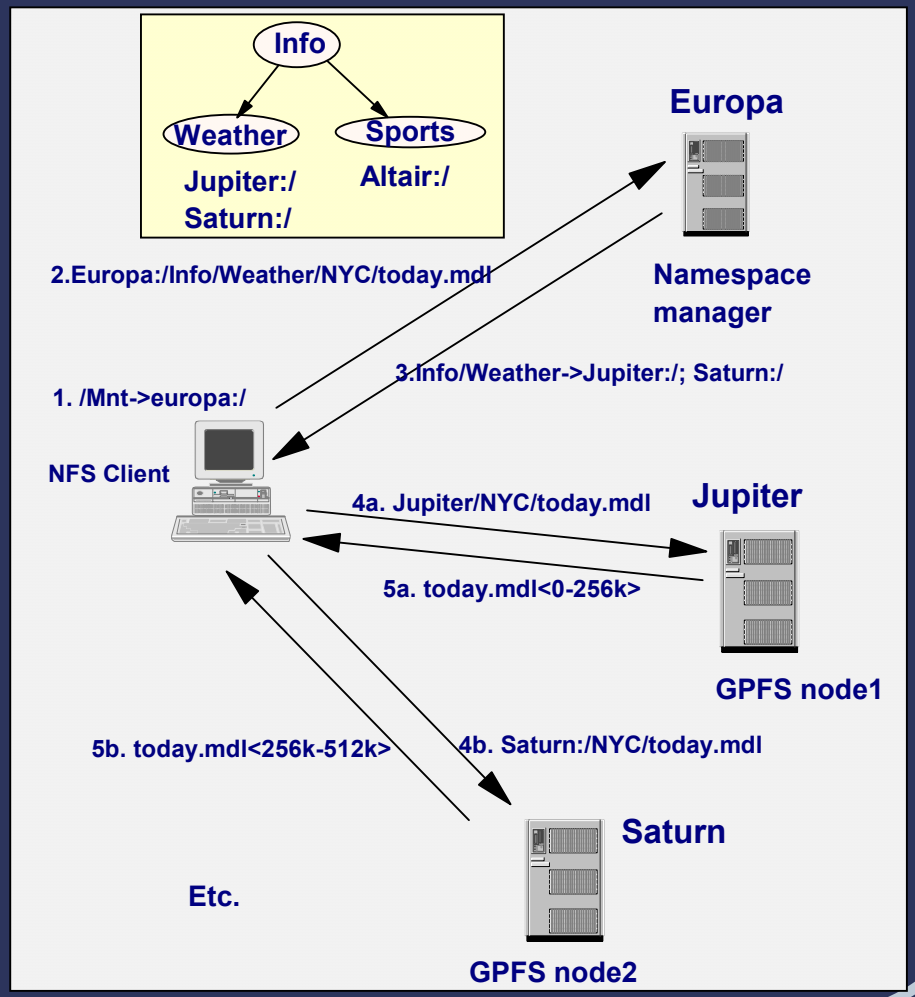
- Idea: Build global name space using the V4 migration/referral feature
  - replaces automounter
  - client nodes mount namespace manager instead of file servers
  - namespace manager looks like a file server with all its file systems migrated elsewhere
  - namespace manager redirects file requests to the correct server



NFS V4: Parallel Access

- Idea: use NFS V4 replication feature to access large GPFS files in parallel through multiple V4 servers
  - multiple NFS servers in cluster export same GPFS file system
  - name lookup returns multiple servers
  - extended attribute defines striping permutation for reads & writes
  - parallel aware client honors striping permutations for R/Ws
  - parallelism for multiple clients or fast individual clients
  - eliminates NFS server bottleneck

Example Client Operation
Reading: /mnt/Info/Weather/NYC/today.mdl



Data Grid: GPFS and DB2-CM

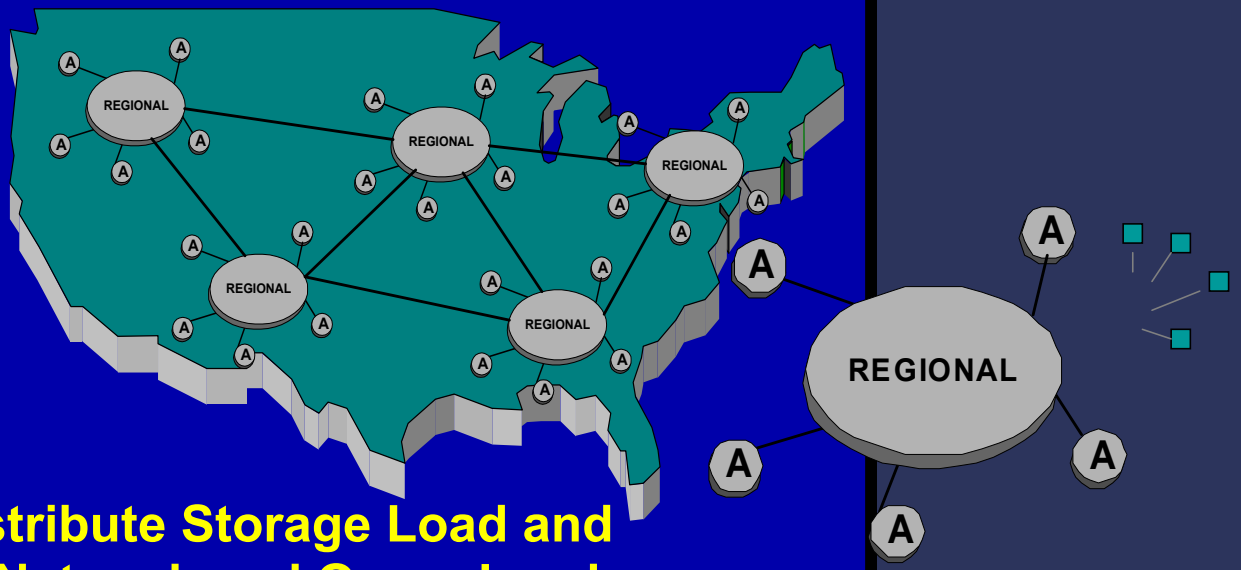
- Need tie in between the multi-cluster file system and databases
- Concept: Database has a http link to a file in the FS
  - VFS traps to ensure consistency
- Example: University of Pennsylvania NDMA
  - NDMA (National Database Mammography Archive)
    - Efficient archiving: Digital Archives & libraries (File systems)
    - Easy access from any participating location: Advanced Networking and secure file sharing
    - Diagnosis tools – consultation experts: Computer Aided Diagnosis, Image Analysis
    - Research: Education and Training
    - Innovative teaching tools: Education and Training
    - Privacy and confidentiality: Information Security

Data Grid Example: NDMA

Potential for 28 Petabytes/year over 2000 Hospitals, in full production

- 7 Regional Archives @ 4,000 TB/yr
- 20 Area @ 100 TB/yr
- 15 Hospitals @ 7 TB/yr

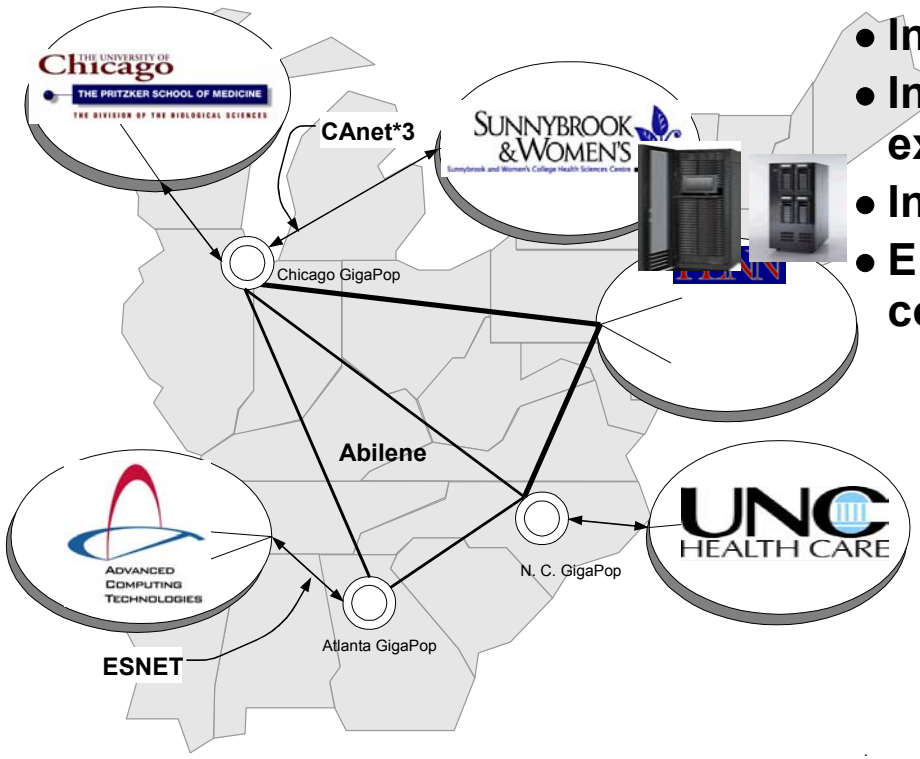
Proposed Hierarchical Layout



Goal: Distribute Storage Load and Balance Network and Query Loads

Data Grid Example: NDMA Configuration

Current NDMA Configuration



Testbed to demonstrate feasibility

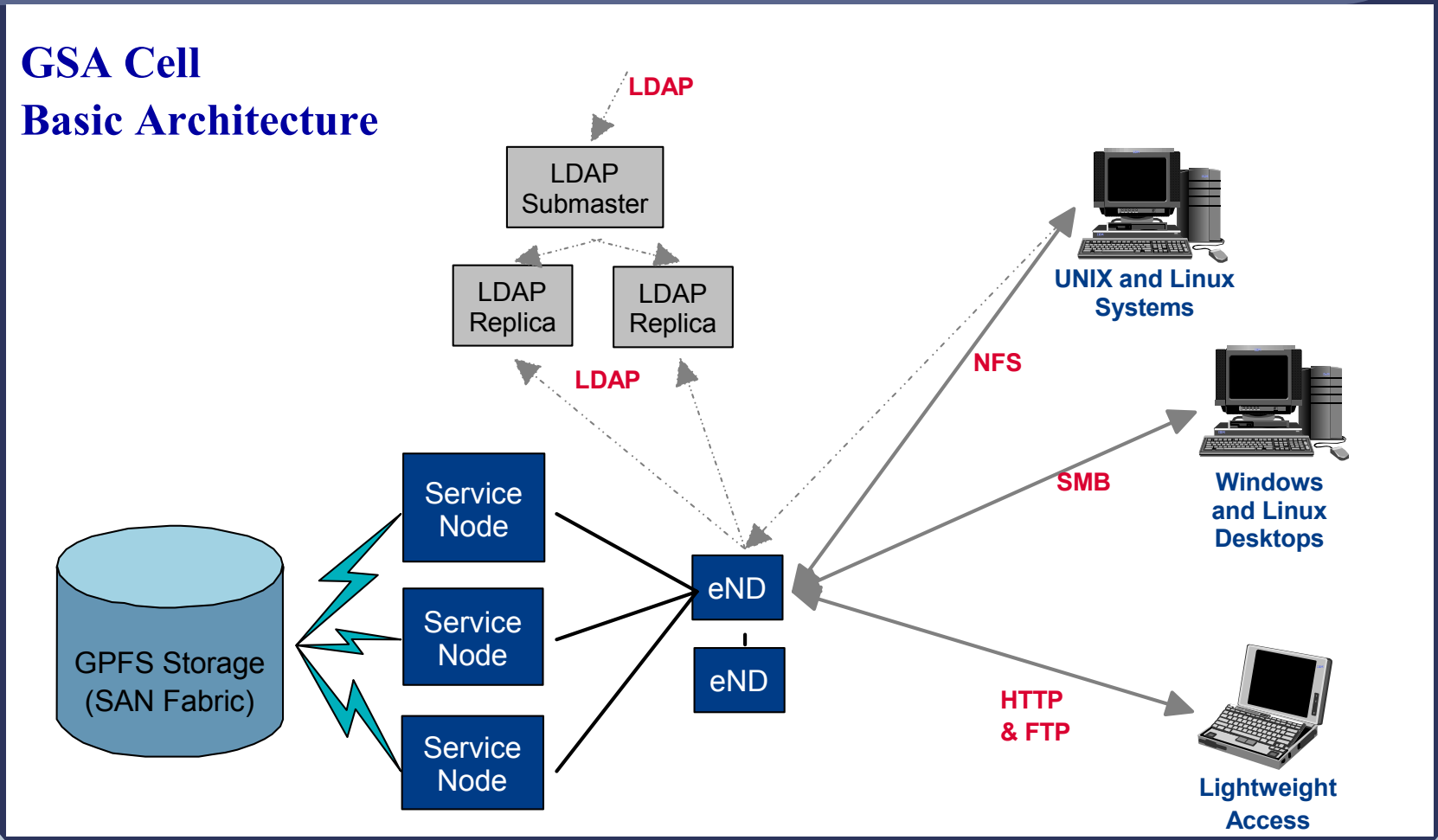
- Storage and retrieval
- Infrastructure for access
- Instant consultation with experts
- Innovative teaching tools
- Ensure privacy and confidentiality

http://nscp.upenn.edu/NDMA

Data Grid Example: GSA

- **Use internally what we sell**
- **IBM's internal solution for sharing file based data**
  - **Replacing legacy solutions**
    - **AFS & DFS, OS/2 Lan Server, Windows NT File Server**
  - **Built from multiple products and technologies**
    - **off the shelf components**
    - **developed administration tools**
  - **NFS is a core component**
    - **NFS V2/V3 now**
    - **NFS V4 to be added in the future**
- **IBM's goals**: Use readily available technology and products
  - **Strong security**
  - **High availability & performance**
  - **Global name space**
  - **Lower TCO than current solutions**
    - **reduced H/W costs, near-zero administration**
  - **Multi-protocol access**
    - **NFS, CIFS, HTTP, FTP**
  - **Support a wide range of client platforms**
- Components: GPFS, SAN, Edge Server, LDAP, NFS, CIFS, HTTP, Executable NFS auto-mounter maps, Custom Administration Tools and GUIs

Data Grid Example: GSA



GSA Status

- History:
  - Entered pilot late in 2001
  - Limited production in 2002
  - Became a rated service in 2003
- GSA space and growth
  - 13 TB of used space in September 2003
  - Just over 1% of total IBM file system space
  - Monthly growth rate around 20%
  - 23,000 active user accounts
- Currently 12 GSA cells in production
  - 6 more underway

Data Grid: Security

- Problem: Admin domains may not be identical
  - Examples:
    - char@washington == lewis@pok
    - bill@watson == pulleyblank@pok
    - dave@nersc == skinner@doe
- Solutions
  - GSS-API
  - LDAP, EIM (enterprise identity management)
  - GSI from globus
  - GPFS multi-cluster credentials management

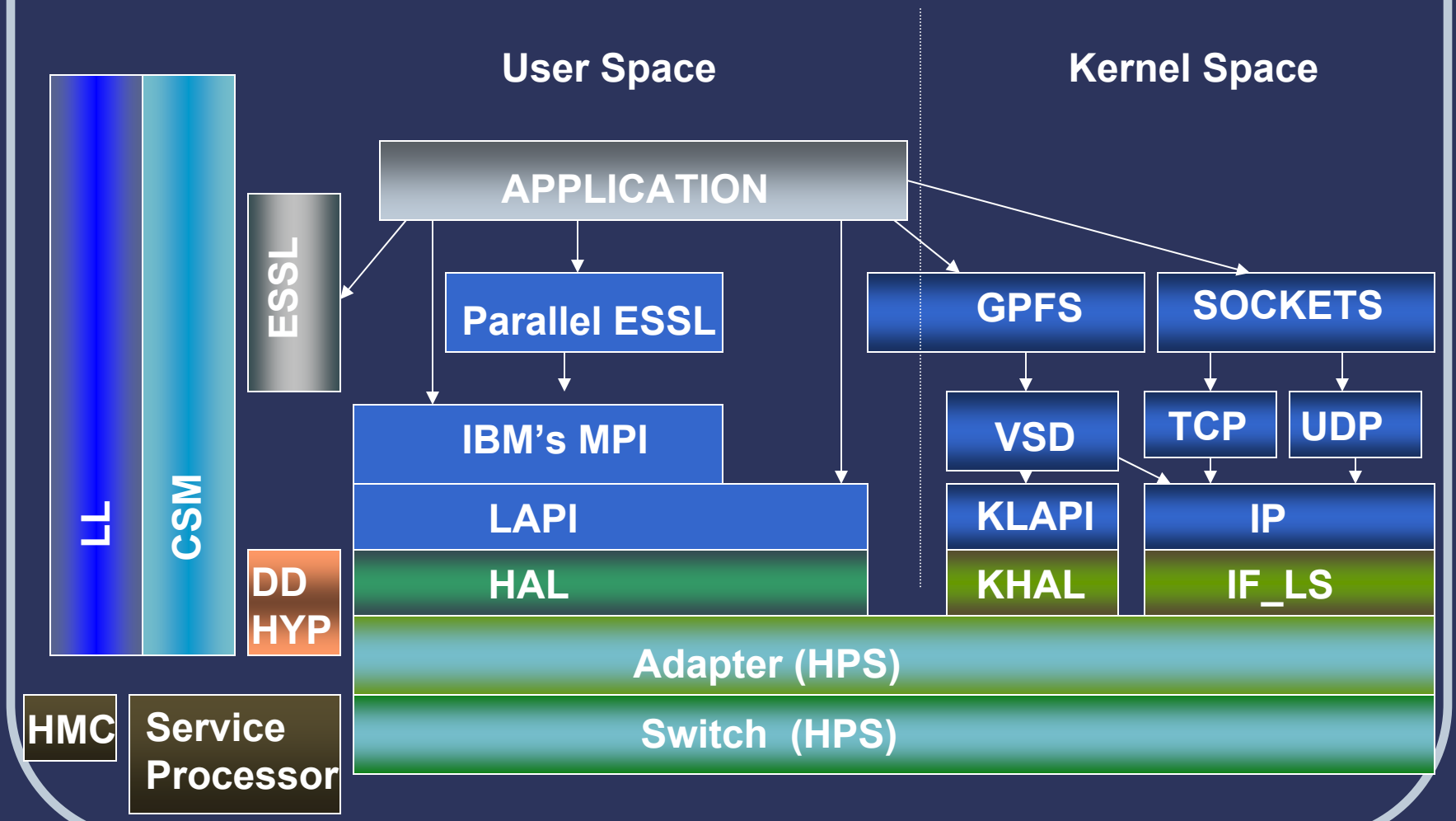
GPFS Roadmap

- GPFS v2.2 on pBlades
- GPFS Multi-cluster and WAN support
- Power 5 and pHPS support
- Performance/Scaling Improvements
- 100 GB/s to a single file
- 2PB file system scaling
- Improvements for incremental recovery in the event of a RAID failure
- Utility to ID File with Physical Device
- Performance monitoring GUI

Protocols

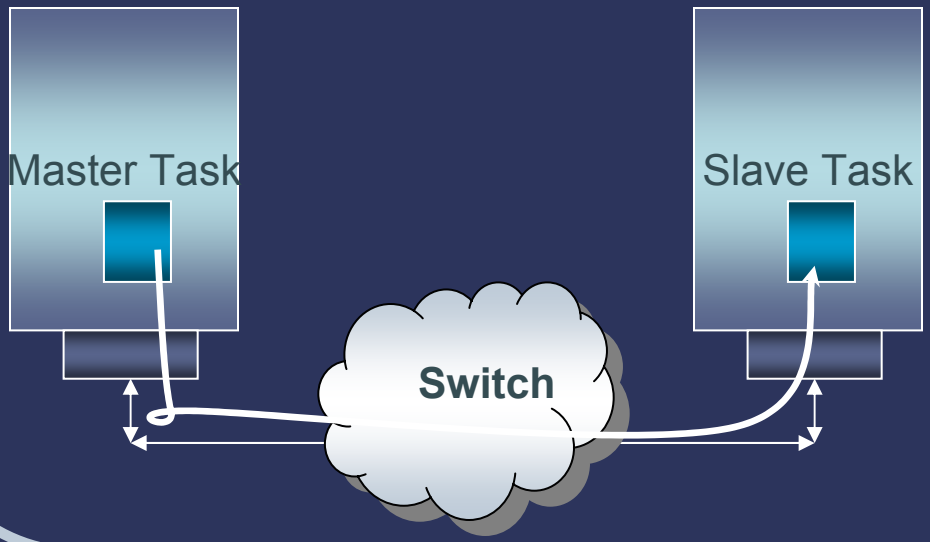
- What is RDMA?
- RDMA value proposition
- Caching Effects
- Striping across multiple interfaces

Communication Software Architecture



What is RDMA?

- Remote Direct Memory Access
- Memory Semantics for Remote Access
- No slave side protocol processing
  - Assumes pinned/mapped transfers

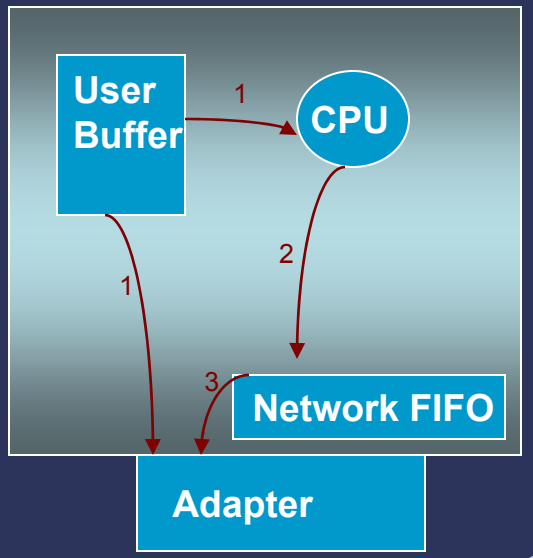


Possible Values

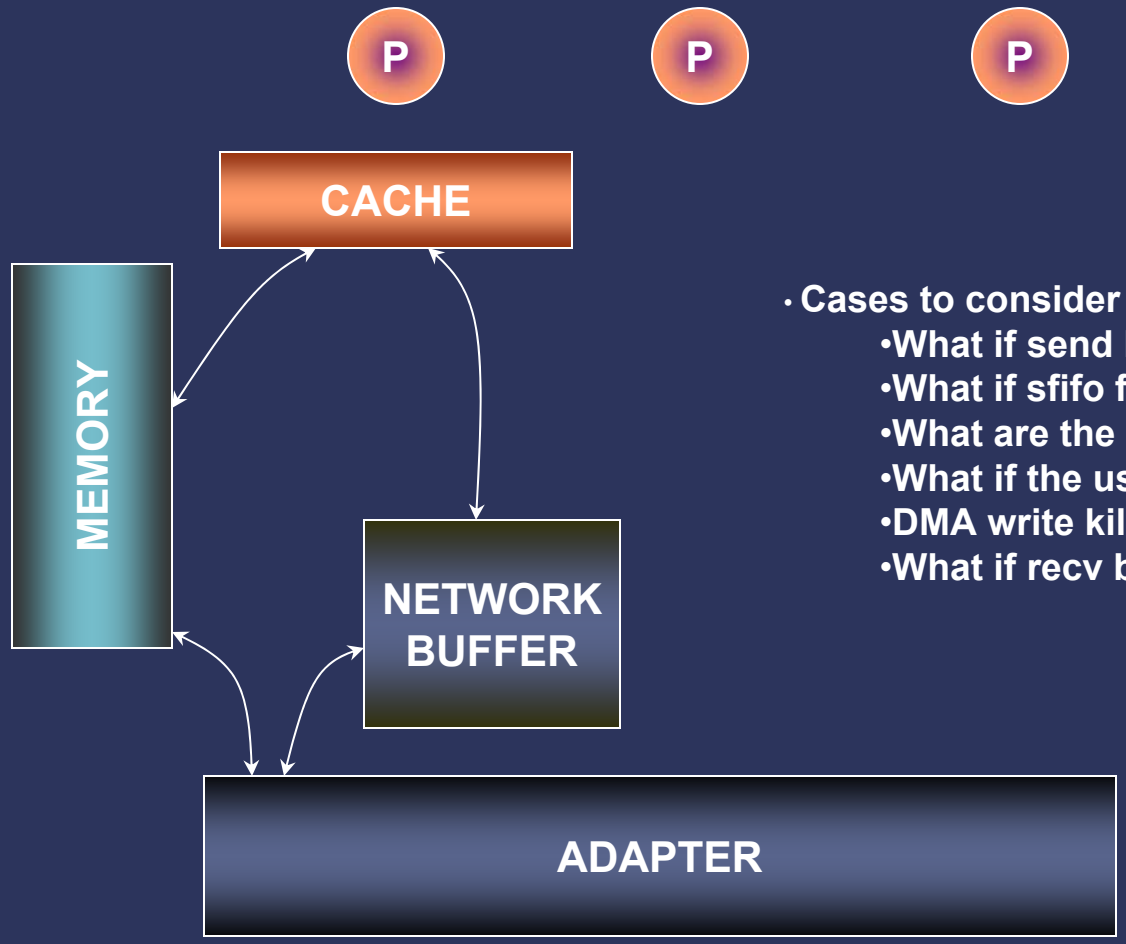
1. **Overlap of Computation and Communication/IO**
2. **Packetization and Unpacking offloaded & interrupts reduction**
3. **Reduce memory subsystem load**
4. **One sided programming model**
5. **Striping of a single message across multiple network interfaces**
6. **Improved raw bandwidth**

Communication Modes

- Packet Mode (now called FIFO mode)
  - Message chopped into 2K packet chunks on the host and copied by CPU (user buffer <-> network buffer)
  - Memory bus crossing depends on caching. At least 1 IO bus crossing
- Zero Copy
  - minimize bus crossings
  - 1 IO bus crossing
  - CPU offload (no copy)
- RDMA enablement
  - No slave side protocol
  - CPU offload
  - Enhanced Programming model
  - 1 IO bus crossing



Bus Crossings Discussion

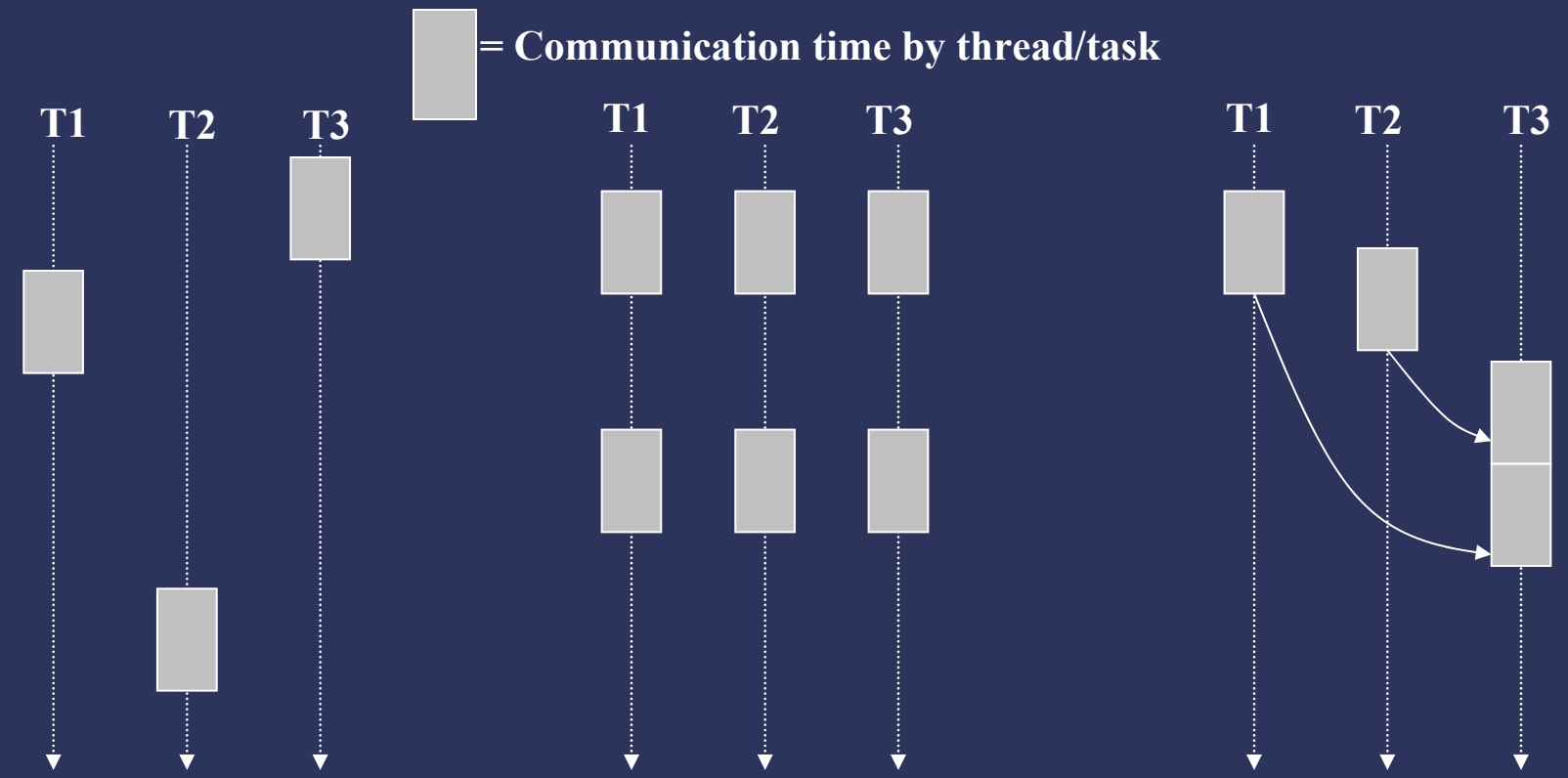


- Cases to consider
  - What if send buffer is already in cache?
  - What if sfifo fits in cache?
  - What are the other processors doing?
  - What if the user makes a blocking call?
  - DMA write kills the receive caching
  - What if recv buffer already in cache?

MPI User Space RDMA

- Implementation Options
  - Pin and map on the fly
  - Lazy deregistration model
  - User exposed RDMA
- Overlap: depends on usage
- Memory subsystem load: difficult to gauge
  - Caching effects
- One sided programming model:
  - MPI one sided (unclear)
  - LAPI or MPI hints based approaches
    - Needs application changes
- Striping: Depends on the communication pattern
- Raw Performance: questionable

Striping Options: Usage Models

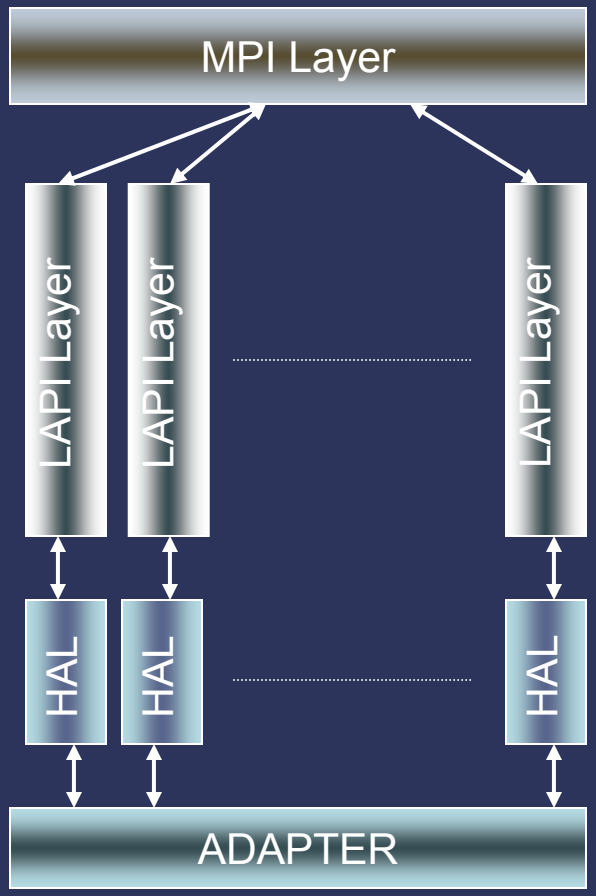


a) Asynchronous Model

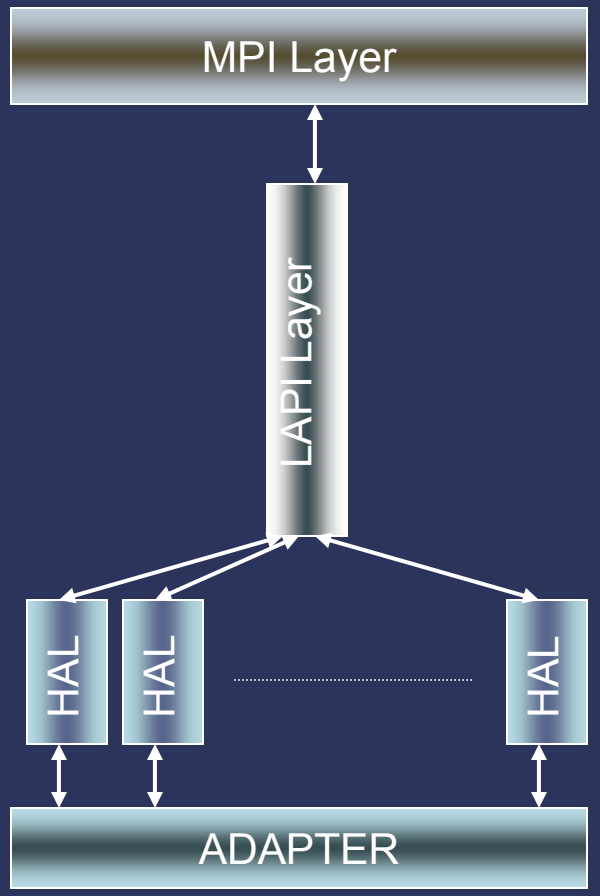
b) Synchronous Model

c) Aggregate Comm Thread Model

Striping Implementation Models



Multiple threads doing copies model



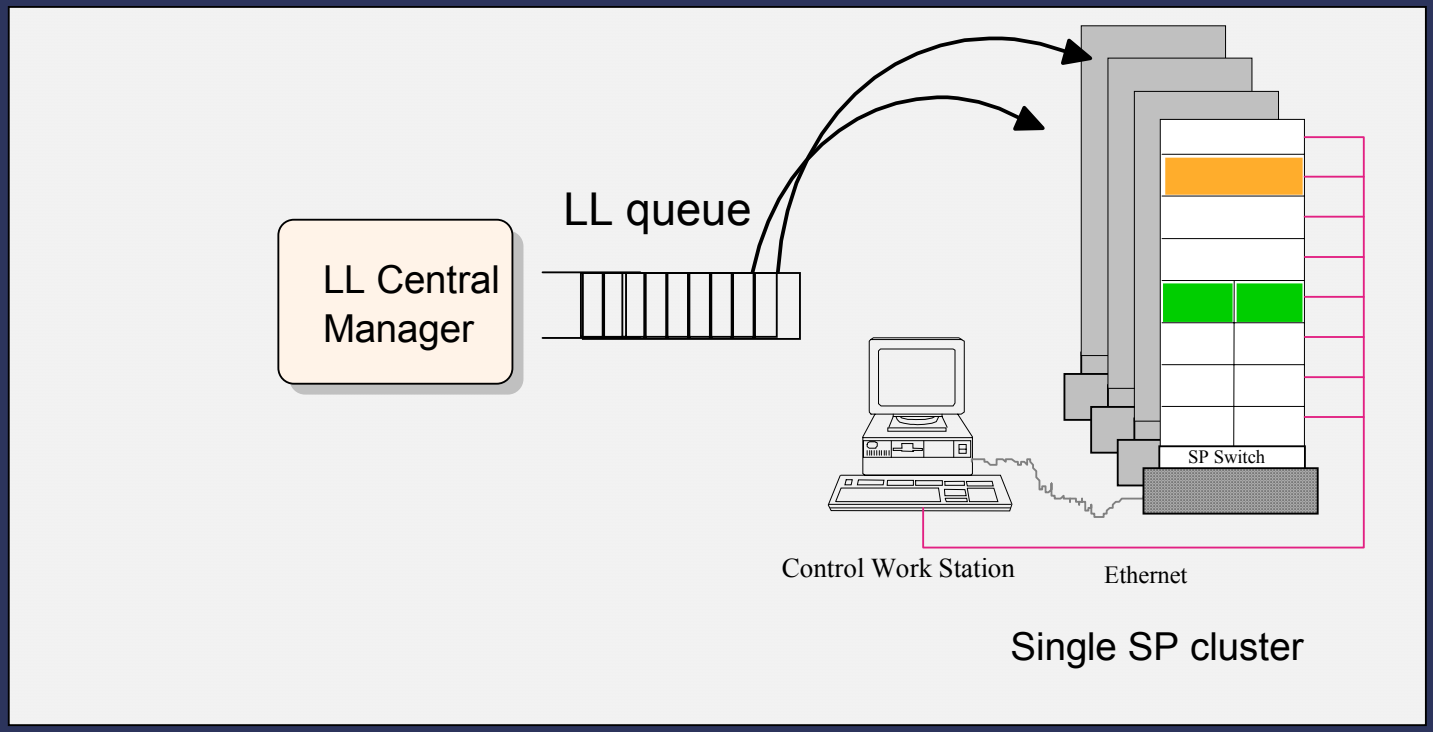
Single Thread with Pipelined RDMA model

Job Scheduling

- Multi-cluster scheduling models
- Advance Reservation
- Global Resource Broker
- Backfill Scheduling
- Persistent Services

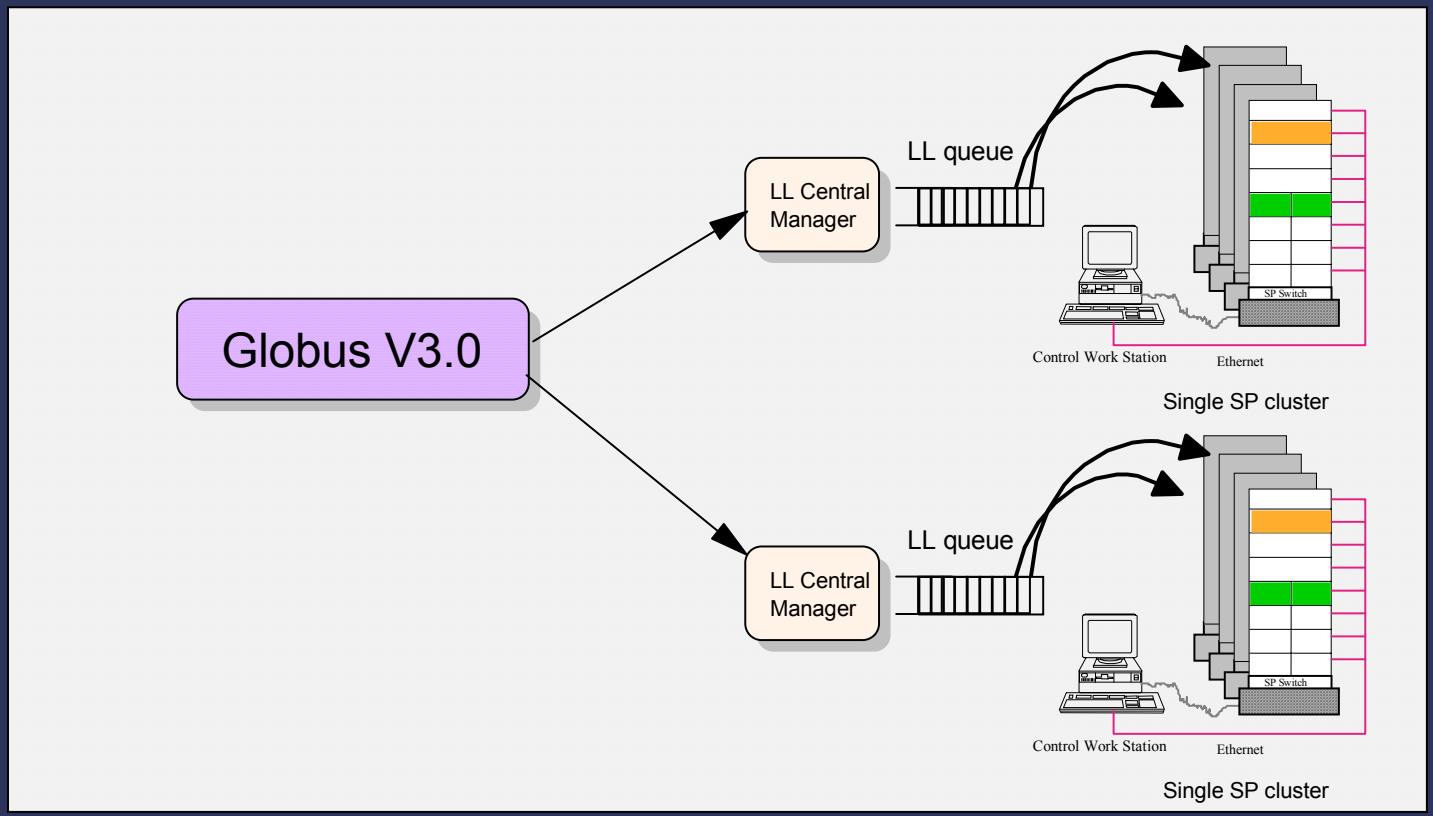
Compute Grid: LL Scheduling

- Use LL to schedule jobs on a single SP cluster
  - POE to launch application on the single SP cluster (run time environment)



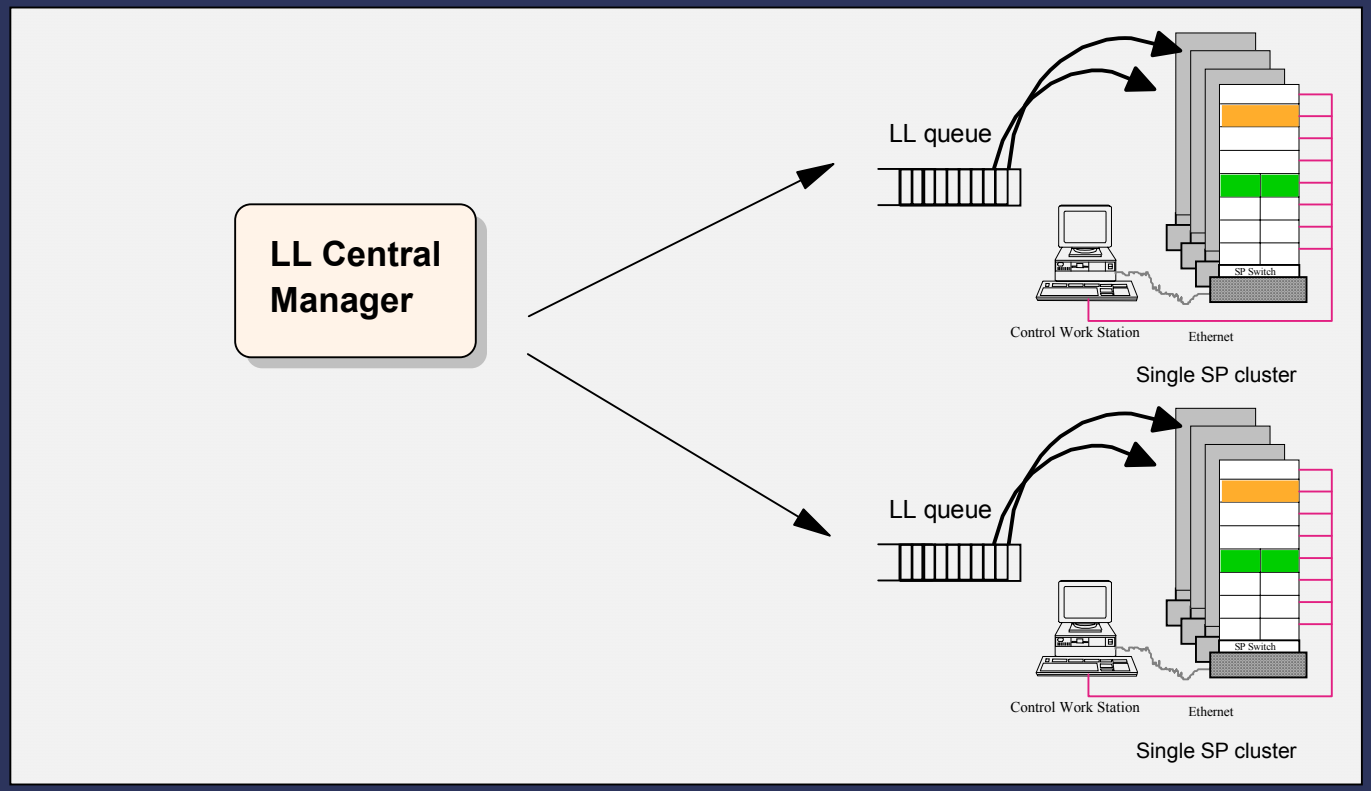
Compute Grid: Job Scheduling

- Globus 3.0 & LL integrated to enable a Grid SchedulingSolution



Compute Grid: Job Scheduling

- LL is capable of scheduling over multiple SP clusters
  - the SPs must be part of the same admin domain (users, etc.)



Compute Grid: Advance Reservation

- Problem: Certain applications have to be scheduled at very specific times
  - The time and the resource needs are known in advance
  - Examples:
    - Weather model to run daily forecast
    - Credit Card billing to run once every month
    - Deadline scheduling requirements
  - How do we factor such advanced reservation requirements while maximizing throughput
- Solution: LL Advanced reservation function
  - Currently definition under way in the GGF standards body
  - IBM is an active participant
  - Pluggable design
  - Work in conjunction with the preemption capability

Compute Grid: Global Resource Broker

- Problem:
  - Multiple clusters that make up the grid domain
  - Each cluster with different capabilities and admin domains
  - How do we route each job to the most appropriate cluster?
- Solution:
  - LL's central manager as the global resource broker
  - Needs tie in with the Globus toolkit
    - Currently a major deficiency of Globus-GRAM
  - Will need LL ported to other platforms
    - Work under way

Compute Grid: Backfill Scheduling

- Problem:
  - Head of the queue needs more resources than currently available
  - Maximize throughput while maintaining the scheduling priorities
- Solution:
  - LL backfill scheduling with preemption (or C/R)
    - Scan the back of the queue to find the first job that can make use of available resources
    - When resources become available to be able to schedule the front of the queue job, preempt the backfilled job
      - Alternately checkpoint the backfilled job if the capability exists

Grid Services: Persistent Services and Affinity

- Problem:
  - Certain services are transactional in nature
  - Need to avoid the job launch and shutdown overheads
  - Examples: financial modeling, simple stochastic modeling
- Solution:
  - LL/PE persistent parallel service (work under way)
  - Supports MPI applications
    - Single admin domain
    - Multiple admin domains
  - Can render solutions via web services

Compute Grid: Programming Model

- Extend MPI to work across cluster boundaries
  - Data center multi clusters are easier
  - WAN models need faster IP backbones with channel aggregation
    - Latency sensitivity issues need to be addressed
    - Topology aware MPI implementation needed
    - Tie in with the scheduling affinity issues
  - Needs some more evolution of networks to address HPC application issues

Operating Systems: AIX

- Virtualization
- WLM enhancements
- VLAN capability
- Capacity on Demand
- Medium sized pages
- Affinity scheduling

ESSL & PESSL Enhancements

- (ESSL v4.1)
  - Power 4 Tuning (GQ), new LAPACK Subroutines & FFT SMP Algorithms
  - One Dimensional FFTs for large N
  - Compiler Support: XLF 8.1 && C for AIX v6
  - VAC++ for AIX v6
  - Parallel ESSL v3.1 support of the HPS Switch and New ScaLAPACK subroutines

  - ESSL 4.1.1 and Parallel ESSL 3.1.1 JS20 Blades & Myrinet SLES8
  - ESSL 4.1.1 and Parallel ESSL 3.1.1 pSeries & Myrinet RedHat 3
  - ESSL 4.1 and Parallel ESSL 3.1 Squadrons As is Standalone AIX 5.2
- (ESSL v4.2) - Squadrons 64W SMP Tuning, Additional LAPACK subroutines
  - ESSL 4.1.1 and Parallel ESSL 3.1.1 JS20 Blades & Myrinet RedHat 3
  - Parallel ESSL 3.1 HPS Software Update
  - ESSL 4.2 Squadrons Tuned SLES9 and RedHat 3
  - ESSL 4.2 and Parallel ESSL 3.1.1 JS20 Blades & Myrinet AIX 5.2
  - ESSL 4.2 Squadrons H 64-way AIX 5.3
  - Parallel ESSL 3.1.1 Squadrons HPS AIX 5.2
- Parallel ESSL 3.2 AIX 5.3, SLES9, RedHat

CSM Roadmap

	CSM 1.3.1 (2Q03)	CSM 1.3.2 (4Q03)	CSM 1.3.3 (2Q04)	CSM 1.4 (4Q04)
Function				
	• Additional Linux Distribution Support • BladeCenter support • Multiple NFS servers for install scaling • Additional xSeries hardware supported • Additional pSeries hardware supported • CSM / XCAT Transition support • CSM/Director Integration Phase 1	• Additional Linux Distribution Support • Secondary Adapter Support • Additional xSeries hardware supported • **CSM/Director Integration Phase 2** • Interoperability with xSeries Linux and pSeries Linux • Additional diagnostic probes • Backup and restore scripts	• Non-node device support • Additional xSeries hardware supported • Additional pSeries hardware supported • Highly available management server • **PSSP to CSM Transition Utilities** • ASCI Purple Support	• **Kerberos V5 for remote commands** • Additional pSeries hardware supported • Additional xSeries hardware supported • Install enhancements • Additional diagnostic support • WebSM GUI Enhancements • Least privilege command access mechanism • Additional monitoring support
Platforms				
	xSeries - Redhat 7.2 , Redhat 7.3, RedHat AS 2.1, RedHat 8.0, SuSE 8.0, SuSE 8.1, SLES 7, SLES 8 pSeries - AIX 5.2, AIX 5.1, SLES 8	xSeries - Redhat 7.3, SuSE 8.1, SLES 8, Redhat 8.0, RedHat 8.1, Redhat AS 2.1, RedHat AS 3.0 pSeries - AIX 5.2, AIX 5.1, SLES 8, RedHat AS 3.0	xSeries - SuSE 8.x, SLES 8,SLES 9, RedHat 8.x, Redhat AS 3.0 pSeries - AIX 5.2, AIX 5.1, SLES 8, SLES 9, Redhat AS 3.0	xSeries - SuSE 8.x, SLES 9, RedHat 8.x, Redhat AS 3.0 pSeries - AIX 5.3, AIX 5.2 SLES 9, Redhat AES 3.0
Scaling				
	xSeries - 512 nodes pSeries - 128 nodes	xSeries - 512 nodes pSeries - 128 nodes	xSeries - 1024 nodes pSeries - 256 nodes	xSeries - 1024 nodes pSeries - 512 nodes

Challenges Ahead

- Making Grid environments more easily manageable and seamless as an “On Demand Infrastructure”
- Programming Models and tools for developing Grid applications
- LL usability enhancements
  - Missing features
  - C/R and preemption capabilities need to be more robust
  - Lot more research investigation necessary
- Standards definitions critical
  - Data access and sharing (caching protocols)
  - Programming models
  - Scheduling
- Stronger linkages between file systems and databases

Applause☺

Data Grid: GPFS and Globus

- Conventional Globus Toolkit model for data access:
  - Replica Location Service, Grid FTP
  - User explicitly moves data from storage site to compute site
  - GPFS is at one or both endpoints
- GPFS fits well with Globus Toolkit model
  - RLS physical file location can refer to GPFS files
  - The Grid FTP parallel/striped transfers map well to GPFS
- Providing Globus Toolkit support for GPFS
  - A truly parallel version of GridFTP

Data Grid: Grid Data Access

- RLS/Grid FTP model has limitations
  - Replica management is cumbersome
  - Moving files is a separate, user-initiated step
  - No single name space
- High-speed networks enable direct file access across the Grid
  - File server model: NFS V4 with parallel extensions
  - Cluster file system model: treat the Grid as a “virtual SAN”
- Research activity to prototype both Grid file server and Grid “virtual SAN” for GPFS
- Interested parties: LLNL (NFS4), Teragrid (“virtual SAN”)