

A Case Study of Using MPI on Clusters in Grid Environments

EEL6763 Parallel Computer Architecture
Class Project
Spring 2005

Ming Zhao
ming@acis.ufl.edu



- Background
- Approach
- Experiments
- Conclusion

- **Cluster Computing**
 - Cost-effective alternative to a traditional supercomputer, built from COTS employing fast inter-connects
 - Growing acceptance in high performance computing
- **MPI Programming**
 - A de facto standard for parallel computing on distributed memory systems
 - Increasingly popular in cluster computing environments
- **Grid Technologies**
 - *“Coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations”*
 - Federate existing resources to solve large applications
 - Enable wide-area, multi-institutional cooperation

Reference: I. Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations . In International J. Supercomputer Applications , 2001.

- **Heterogeneous Environments**
 - Hardware, operating system, MPI library etc.
- **Cross-Domain Issues**
 - Trust relationship, user identity, security enforcement
- **Grid Data Provision**
 - Performance, security, deployment, integration
- **Objectives of this Project**

■ For Heterogeneous MPI Environments

— PArallel Computer eXtension (PACX)

- Supports NEC SX4/5, Cray T3E, Hitachi SR2201/SR8000, IBM SP2/SP3, SGI Origin 2000/3000, Alpha cluster, Linux cluster, Sun cluster
- Supports up to 1024 computers to form a meta-computer
- Supports the full MPI-1.2 and parts of MPI-2.0 standard

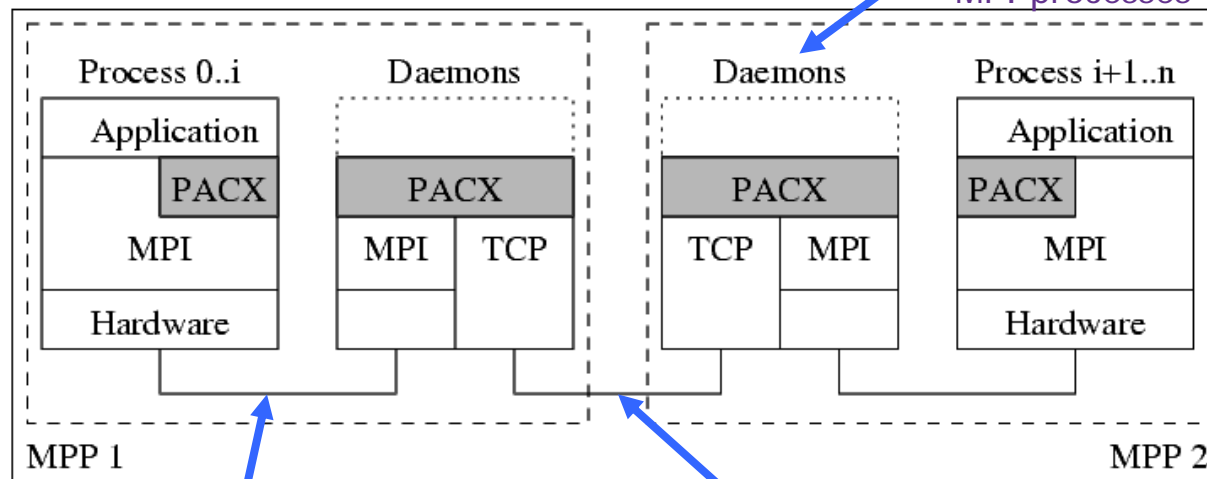
■ For Cross-Domain High-Performance Data Access

— Grid Virtual File System (GVFS)

- Based on a de facto, widely deployed standard, supports generic file system interface
- Supports on-demand cross-domain data access for unmodified application binaries
- Performance and security enhancements for wide-area environments

- Optimized two-level communication
 - Internal communication: Vendor MPI library
 - External communication: TCP/IP via daemons

Two daemons on each system;
Intercept inter-cluster
communication;
Implemented as local, additional
MPI processes

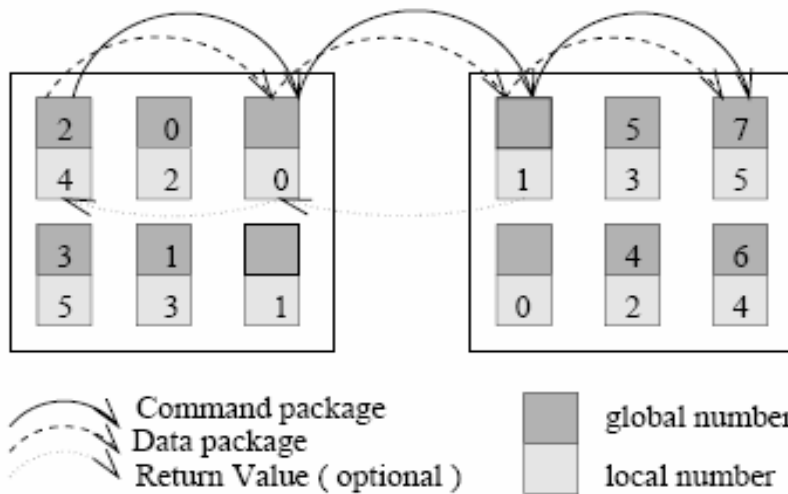


Optimized vendor-MPI library for
intra-cluster communication

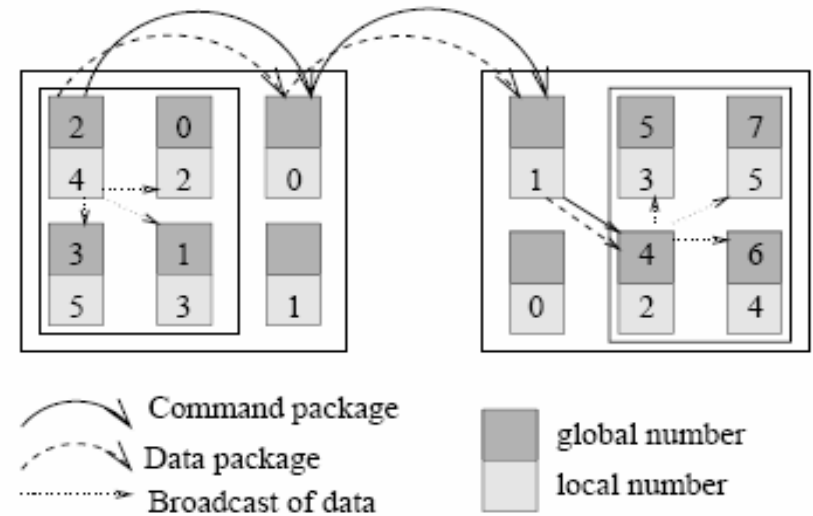
TCP/IP for inter-cluster
communication via daemons

Reference: E. Gabriel, M. Resch, T. Beisel, and R. Keller. Distributed computing in a heterogenous computing environment. In EuroPVMMPI'98 Liverpool/UK, 1998.

PACX-MPI

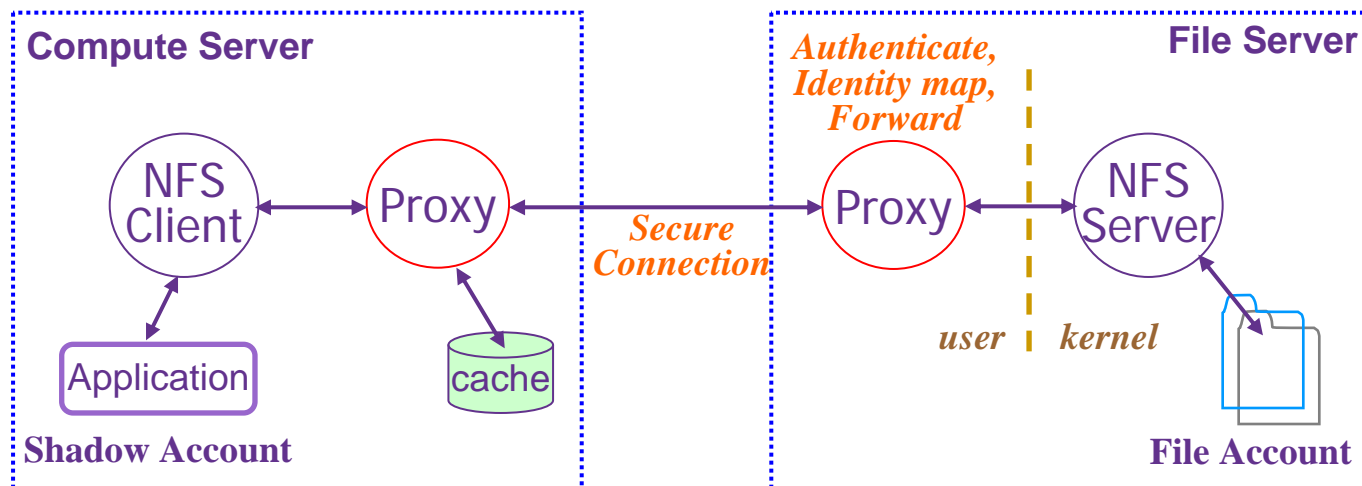


Point-to-point communication in PACX-MPI



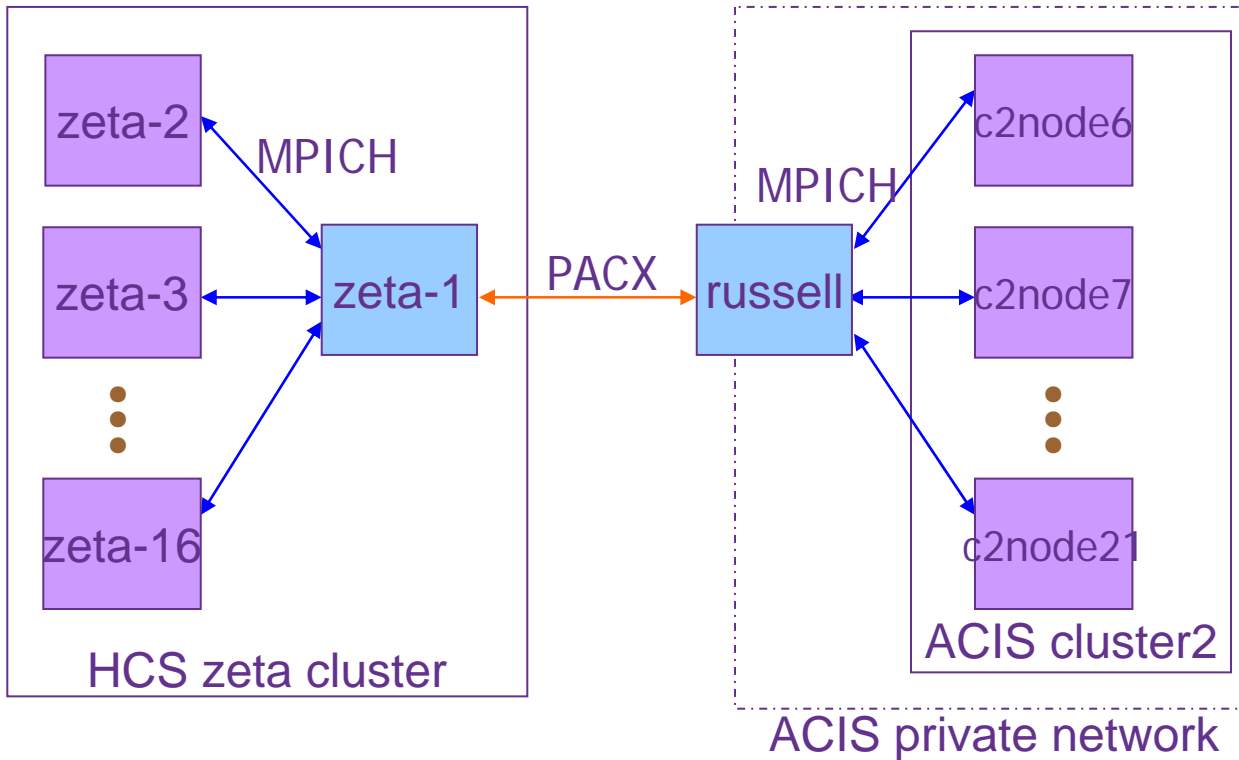
Global communication in PACX-MPI

- Distributed file system virtualization for unmodified applications, based on Network File System (NFS)
- User-level proxy maps user identity, provides on-demand data access
- Caching, prefetching, secure tunneling and authentication



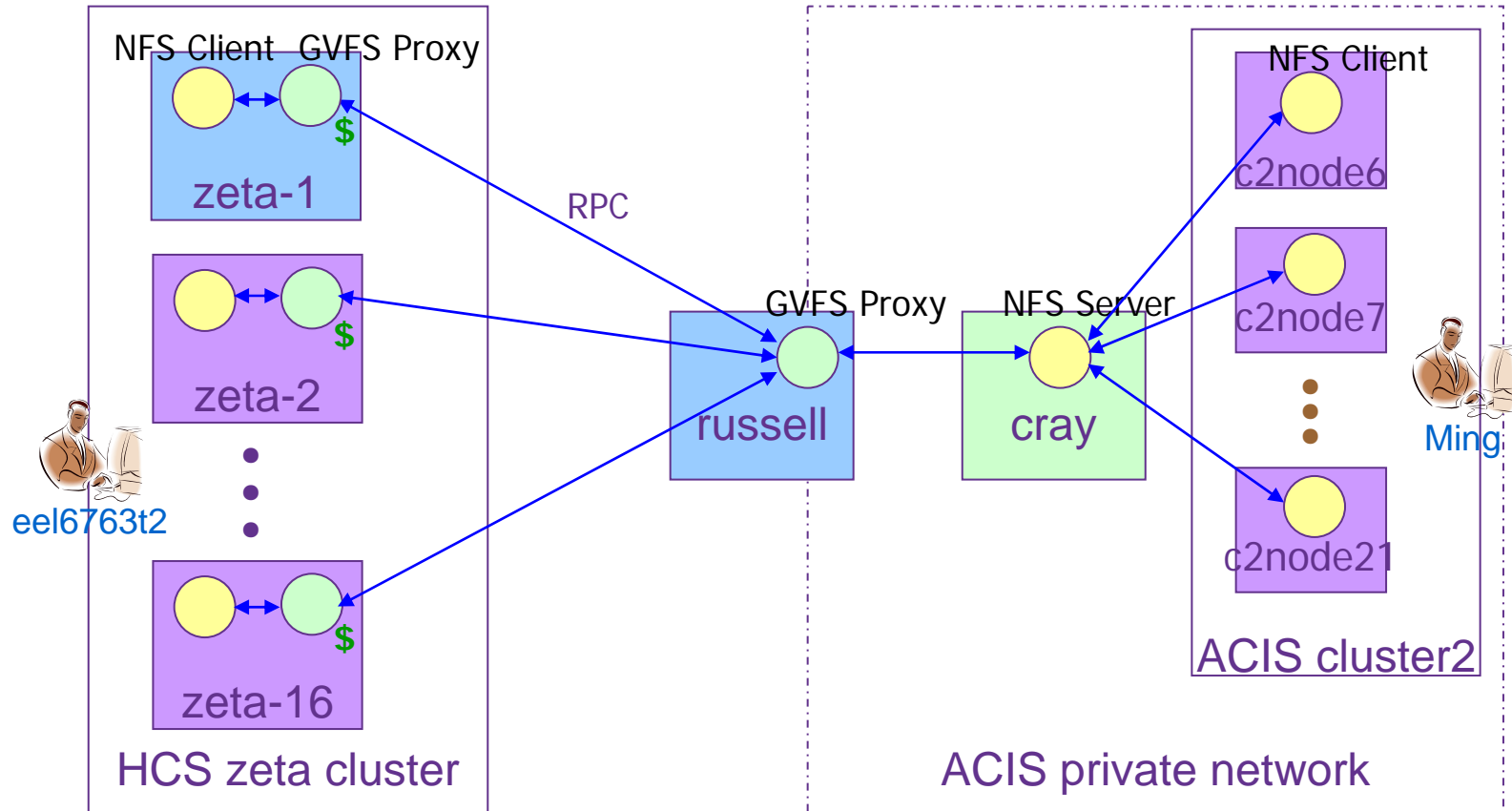
Reference: Figueiredo, Renato J., Kapadia, Nirav, Fortes, Jose A. B. Seamless Access to Decentralized Storage Services in Computational Grids via a Virtual File System . In Cluster Computing, 2004.

Meta-Cluster Setup

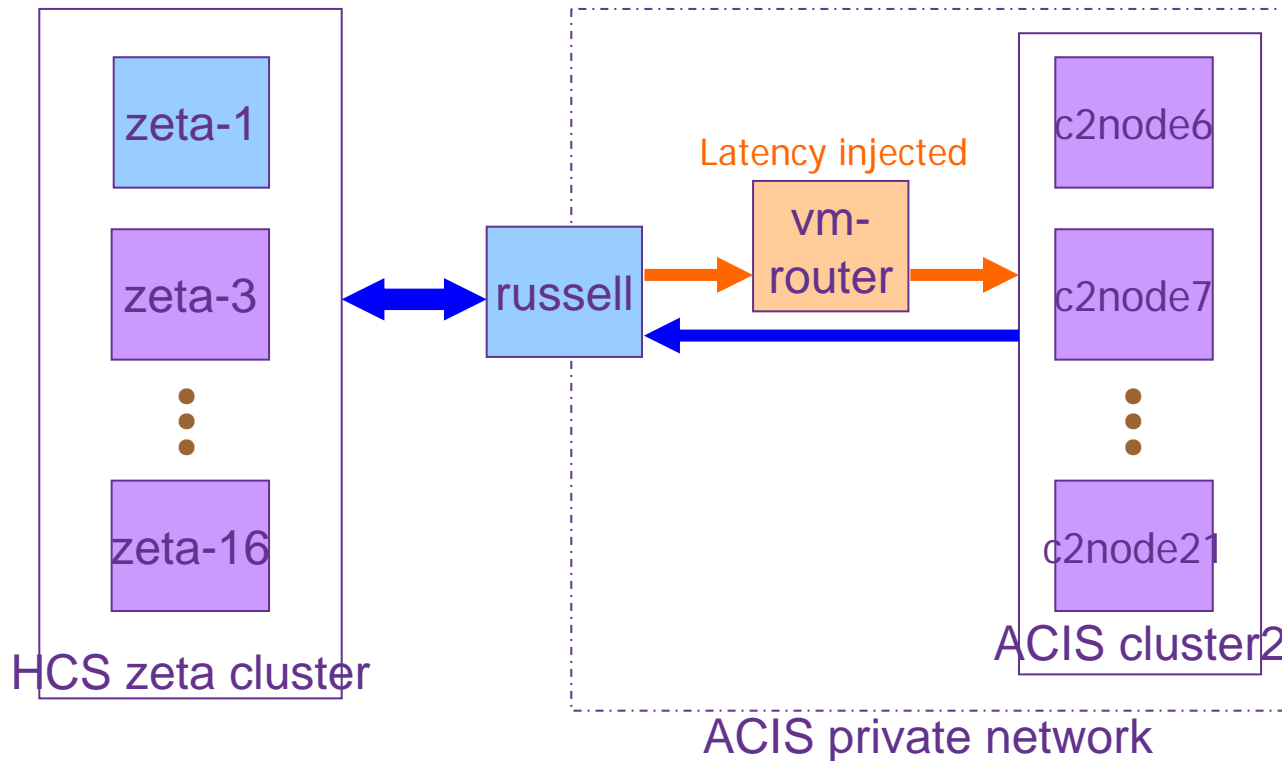


- **HCS zeta cluster**
 - 32 nodes (dual 733MHz Pentium III, 256MB memory)
 - 100Mbps Ethernet
- **ACIS cluster 2**
 - 32 nodes (dual 2.4GHz hyper-threaded Xeon, 1.5GB memory)
 - Gbps Ethernet
- **Gateway russell**
 - VMware ESX virtual machine (2GHz Xeon, 256MB memory)
 - Connected to both clusters via 100Mbps Ethernet

Data Provision



Wide-Area Latency Simulation

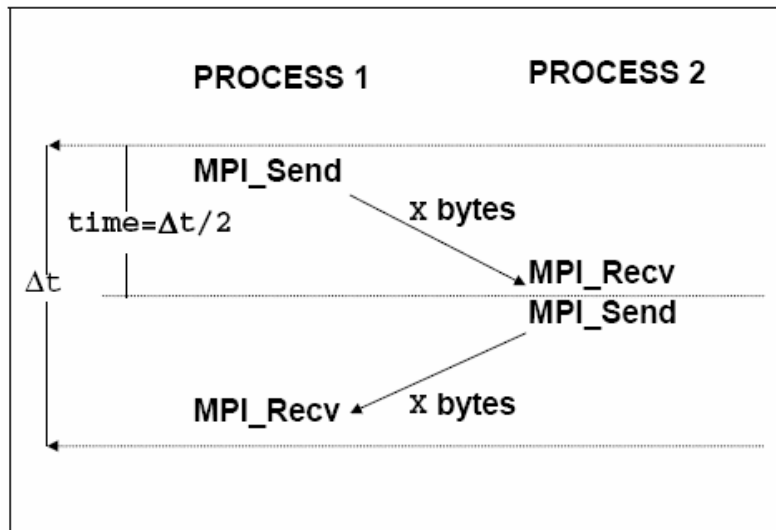


- **VM-Router**
 - VMware ESX virtual machine (2GHz Xeon, 256MB memory)
 - Simulates wide-area latency via NIST Net
- **NIST Net Network Emulator**
 - A general-purpose tool for emulating performance dynamics in IP networks
 - Works on a Linux machine set up as a router

Reference: Mark Carson, Darrin Santay. NIST Net - A Linux-based Network Emulation Tool. In ACM SIGCOMM Computer Communication Review, 2003.

Micro Benchmarks

- Intel MPI Benchmarks (the former Pallas MPI Benchmarks)
 - A concise set of elementary MPI benchmark kernels
 - Single transfer (e.g. PingPong), parallel transfer (e.g. Sendrecv), collective (e.g. Alltoall)



Ping-Pong pattern

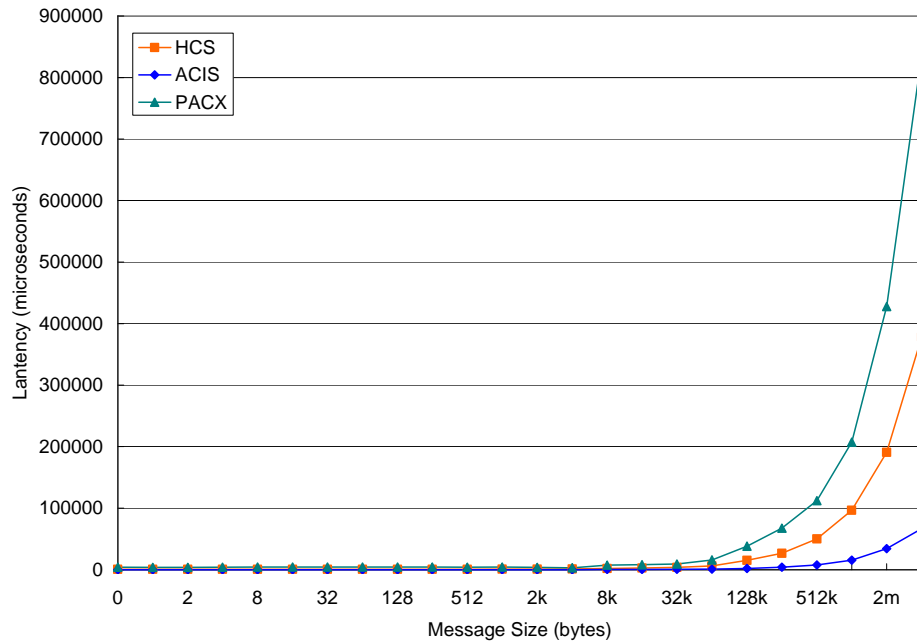
Reported latency = $\Delta t / 2$

Reported throughput = $x / 1.048576 / \text{time}$

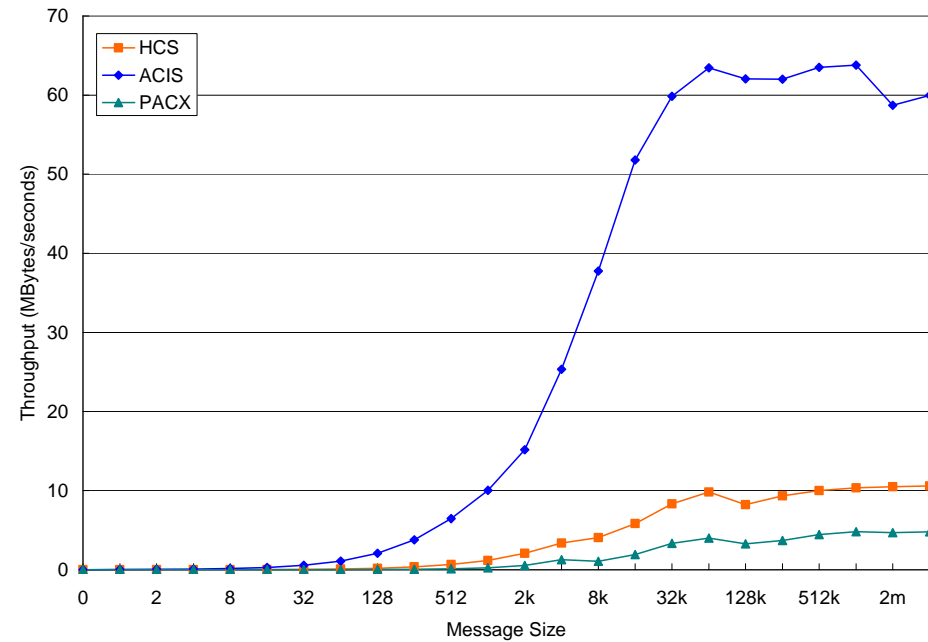
Results

PACX: meta-cluster, ACIS: ACIS cluster, HCS: HCS cluster

Ping-Pong Latency



Ping-Pong Throughput

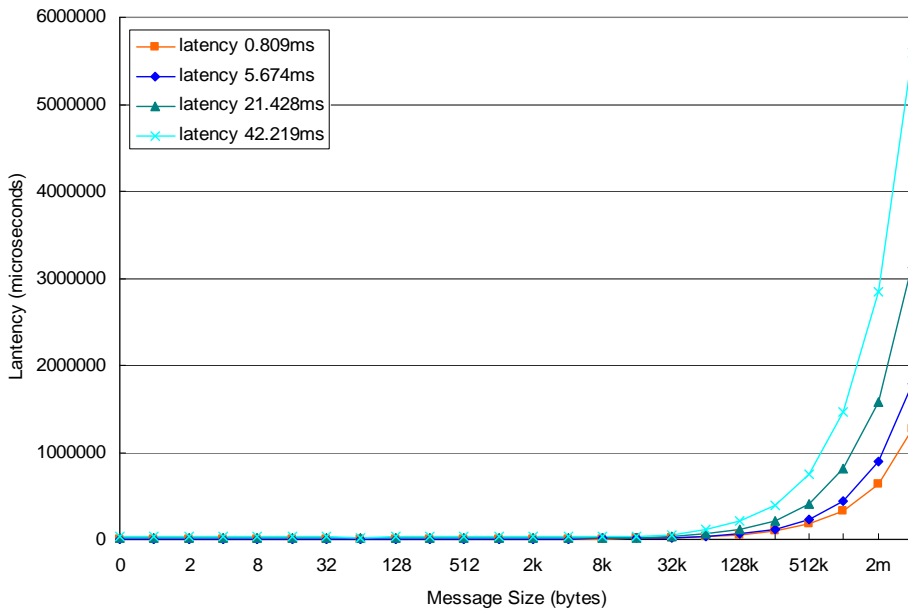


- Inter-cluster communication incurs more latency (longer distance, narrower bandwidth, more routing delay etc.)

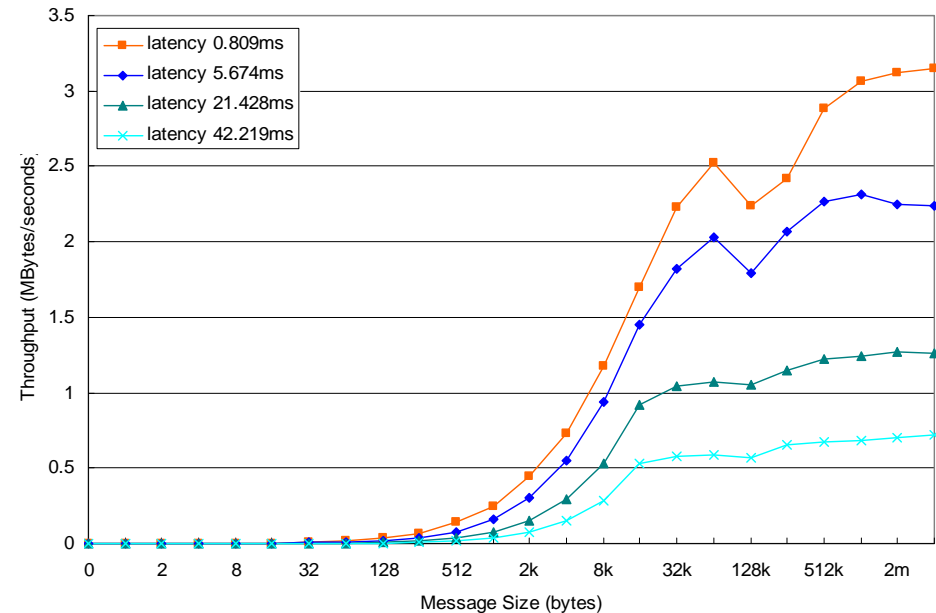
Results

PACX: meta-cluster, ACIS: ACIS cluster, HCS: HCS cluster

Ping-Pong Latency



Ping-Pong Throughput

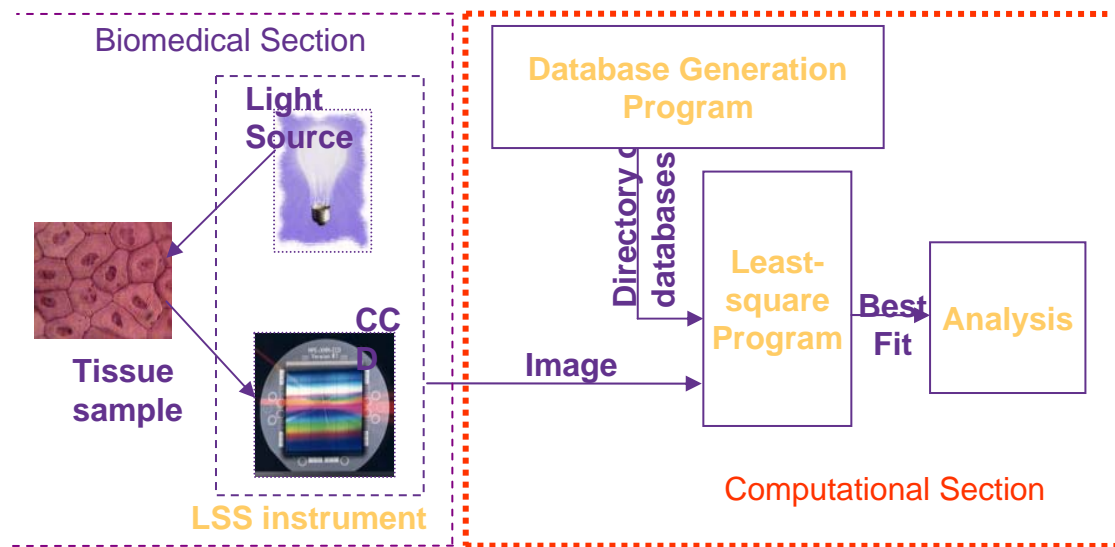


- As the emulator injects more latency on the router, inter-cluster communication overhead further increases.

Light Scattering Spectroscopy (LSS)

■ LSS Analysis

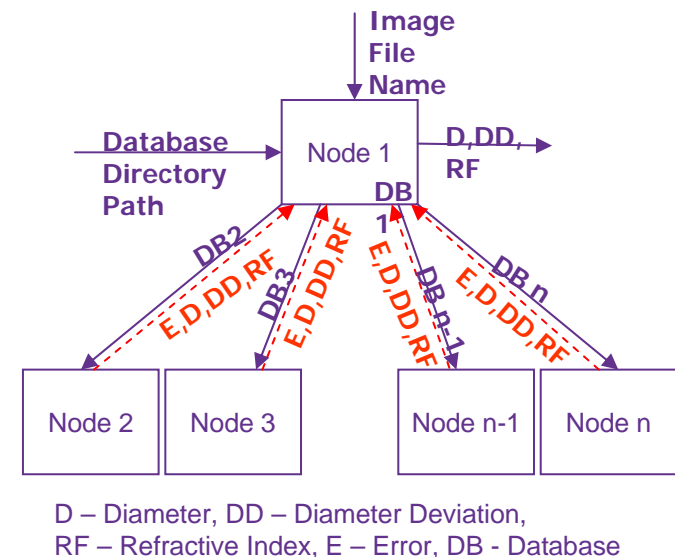
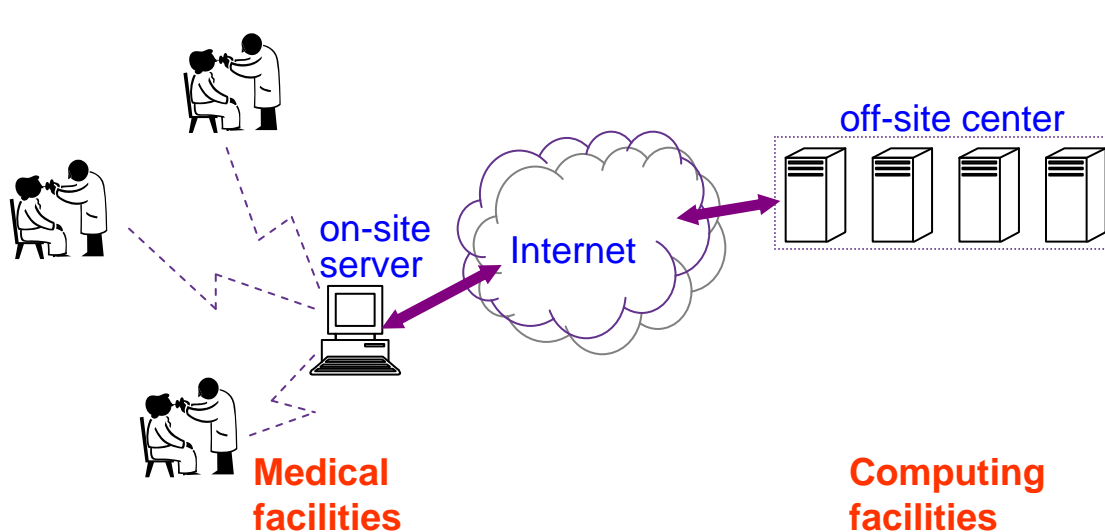
- Probes the structure of living cells without tissue removal, helps in non-invasive detection of precancerous changes in human epithelium
- Obtains parameters (size and refractive index) from spectrum, approximated using lookup on Mie-theory spectra database



Reference: Backman V, et al. Detection of preinvasive cancer cells in situ, Nature, 2000.

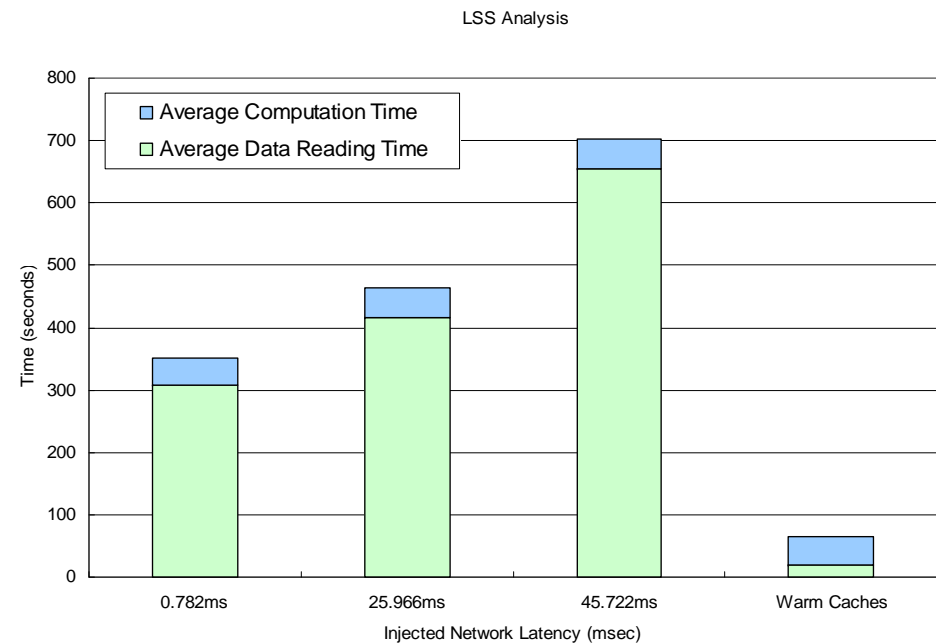
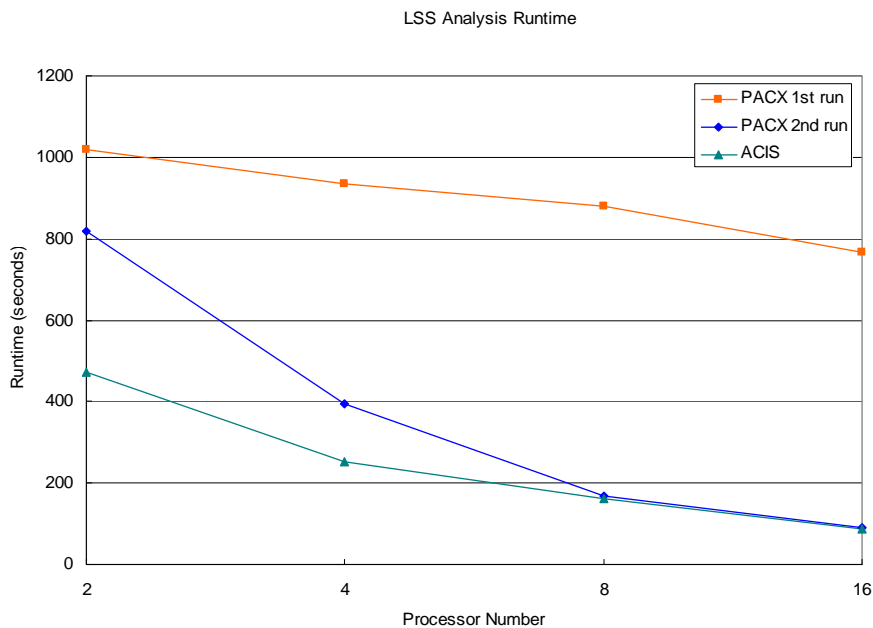
LSS Analysis

- Needs for Grid-Enabled Parallel Computation
 - Very large databases (Gbytes), highly data intensive computation
 - Database generation and lookup can be parallelized
 - Quasi-real time analysis allows instant clinical feedback



Reference: J. Paladugula, M. Zhao and R. J. Figueiredo. Support for Data-Intensive, Variable-Granularity Grid Applications via Distributed File System Virtualization - A Case Study of Light Scattering Spectroscopy, In Proceedings of CLADE/2004.

ACIS: ACIS cluster; PACX 1st run: meta-cluster, cold GVFS caches, 2nd run: warm GVFS caches

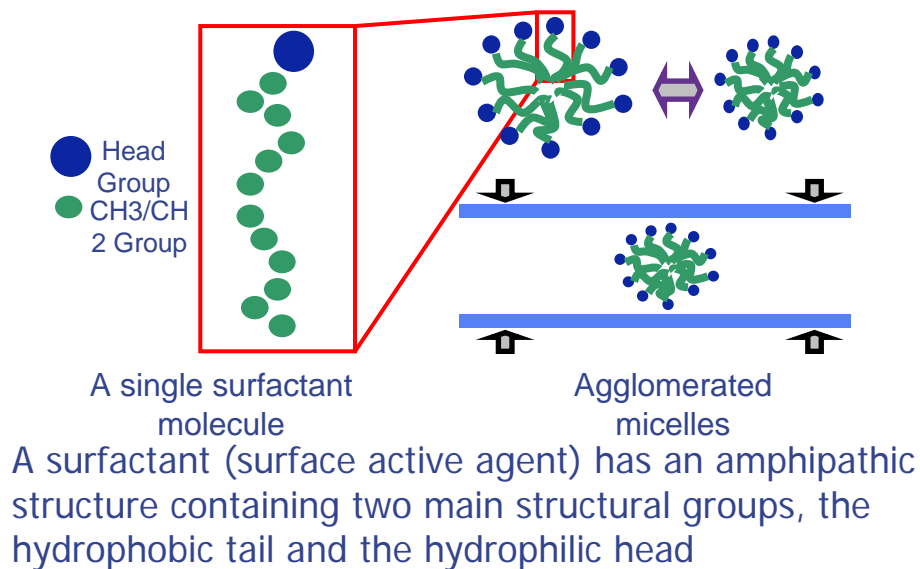


GVFS caching effectively leverages data locality and helps the application perform in WAN even better than in LAN

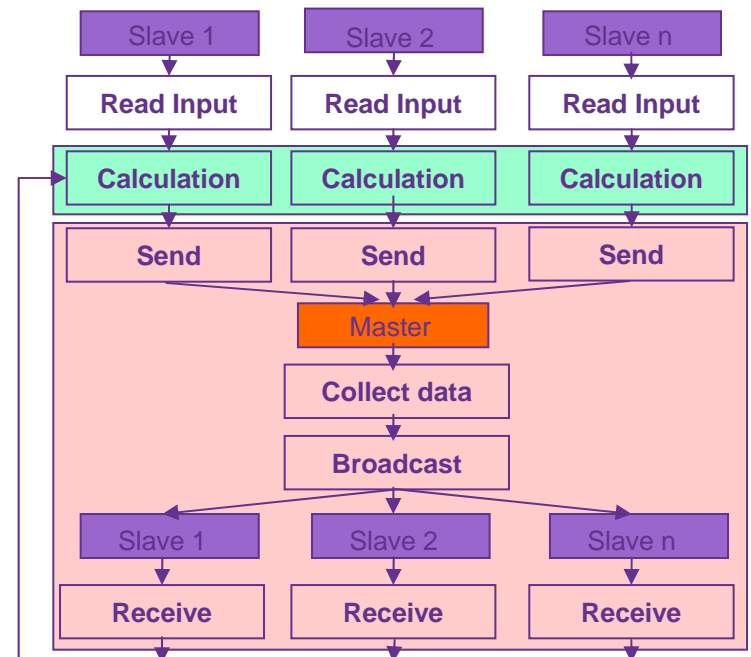
Molecular Dynamics Simulations

Parallel Computer Simulation

- Study the motion of solids, liquids and gases at atomistic level.
- Divide atoms and assign each processors a subset of the simulation space
- Computation intensive but not much communication

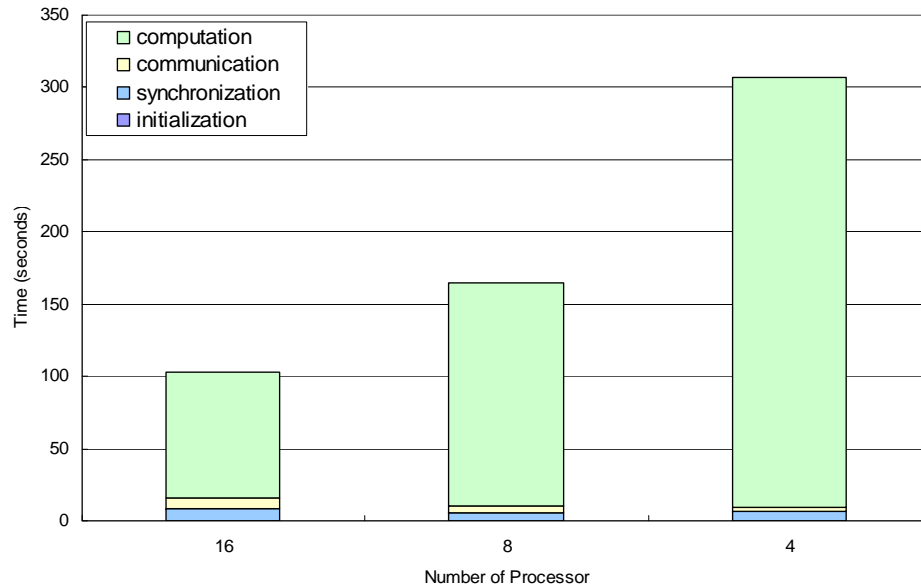


Reference: I. Jang and S.B. Sinnott. Support for Data-Intensive, Molecular Dynamics Simulations of the Chemical Modification of Polystyrene through CxFy+ Beam Deposition, In Journal of Physical Chemistry, 2004.



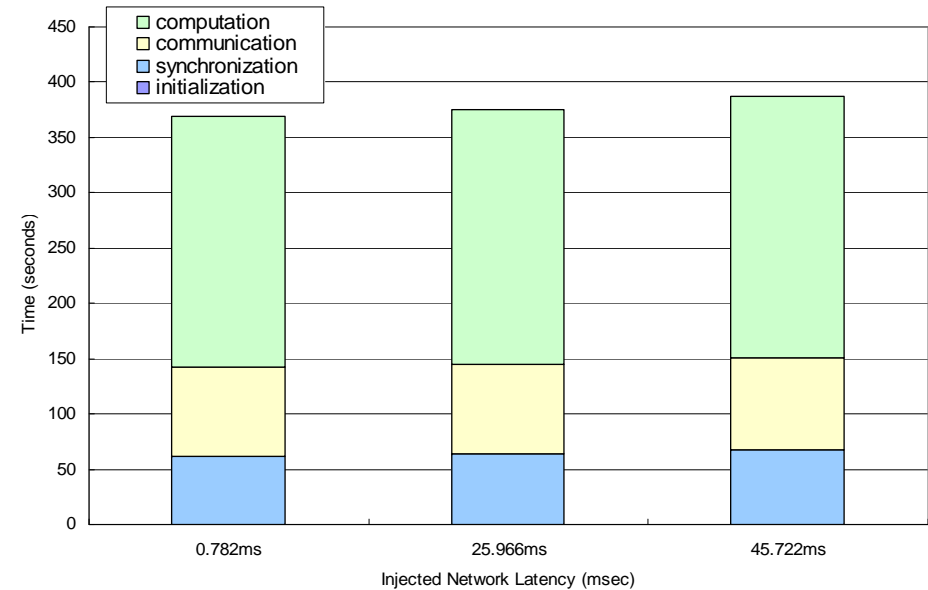
Results

Parallel Molecular Dynamics Simulations



More processes incur more communication and synchronization overhead, but achieve much greater computation speedup

Parallel Molecular Dynamics Simulations



The latency is injected on a unidirectional link, while the communication mostly happens on the other unaffected links, so the performance does not degrade much as the injected latency increases.

Conclusions

- PACX and GVFS enable coupling of clusters in Grids to perform message passing based parallel computing
- PACX supports meta-cluster communication, including clusters in private networks
- GVFS provides a shared file system in Grid environments and leverages caching to hide high network latency
- Mapping algorithm to network topology can be useful to avoid excess use of the slow inter-cluster connection