

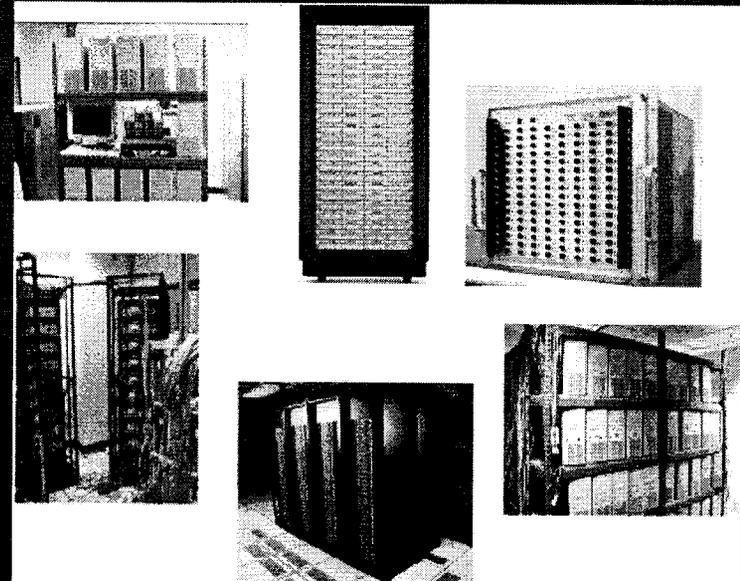
ESTO Computational Technologies Project



# Advanced Topics in Cluster Computing

**Charles D. Norton, Viktor K. Decyk,  
Zach Isaacs, Gerhard Klimeck,  
Nooshin Meshkaty, and Paul Springer**

Applied Cluster Computing Technologies  
Earth Science Data Systems Section  
Jet Propulsion Laboratory  
California Institute of Technology



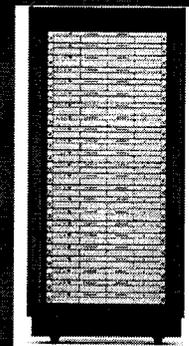
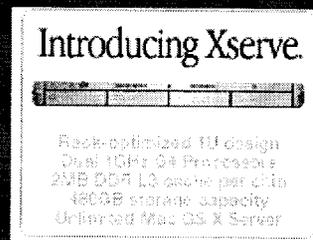
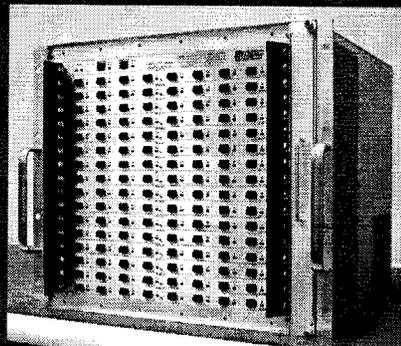
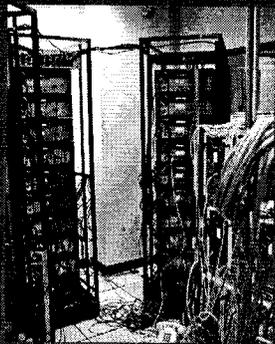
2002 JPL Information Technology Symposium

# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## Evaluating Technology for ESTO Applications



### **Task Objective**

Evaluate applicability of algorithms for cluster programming, software, administration, networking, I/O, performance, and new technology

### **Technical Approach**

- Investigate stability and performance of Gigabit networking systems for ESTO/CT-style computations
- Characterize and study additional areas important to large calculations such as parallel I/O and system management software
- Support and maintain our installed cluster base
- Provide consulting for emerging projects using clusters

### **NASA and ESTO/CT Relevance**

- Researching & resolving problems, with recommendations to vendors, impacts bringing the best technology forward for ESTO/CT application teams
- Characterize bandwidth performance problems in Myrinet-2000 for large numbers of processors and evaluate other systems, such as Dolphin NICs
- Examine emerging non-traditional systems such as Apple Xserve (G4 processors) and Racksaver (Blade systems)
- Pursue issues in parallel I/O and management of large systems

Clusters support task based data processing, optimization, physics-based modeling and more...

# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## Multiple Generations of Cluster Systems

### Single CPU Pentium III System (Nimrod)

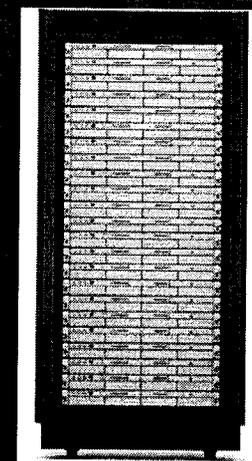
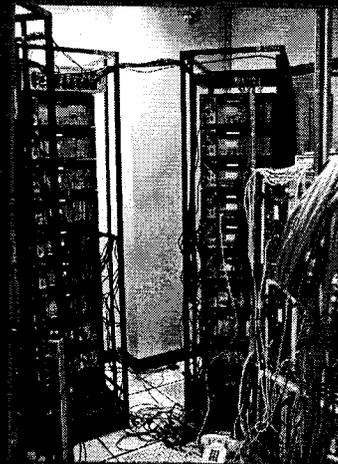
- 32 processors @ 450 MHz with 8 GB disk
- 512 MB RAM per node
- 100 Mbit/s Ethernet, RedHat Linux, MPI
- Fortran 90/95 and C/C++ Compilers
- 14.4 Gflops peak

### Dual-PE Pentium III System (Pluto)

- 32 dual-PE nodes @ 800 MHz with 10 GB disk
- 2 GB RAM per node
- 2 Gbit/s Myricom Switch and 100 Mbit/s Ethernet
- RedHat Linux, MPI (MPICH and LAM)
- Multiple Fortran 77/90/95 and C/C++ Compilers
- 51.2 Gflops peak
- RAID System (to be installed)

### Dual G4 System (Apple Xserve) Arrives in FY'03

- 33 dual-PE nodes @ 1 GHz clk with 60 GB disk
- 1 GB RAM per node
- 100/1000 Mbit/s Ethernet, OS X Server, MPI
- Fortran 90/95 and C/C++ Compilers
- Multiple Funding Sources
- ~2 TB storage
- ~500 Gflops peak (with Velocity Engine)

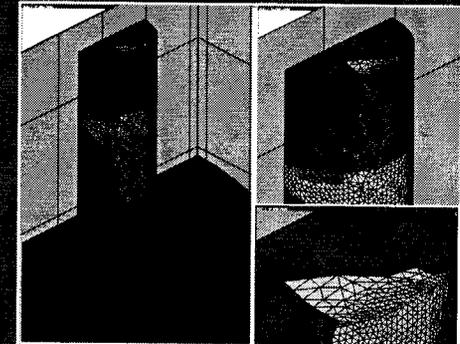
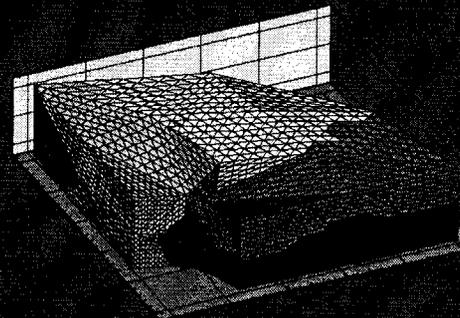
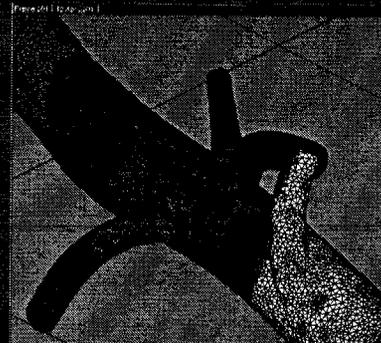


# ESTO Computational Technologies Project

PYRAMID: Parallel Unstructured Adaptive Mesh Refinement Library



Modern... Simple... Efficient... Scalable...



## ***Task Objective***

Develop an advanced software library supporting parallel unstructured adaptive mesh refinement for large-scale scientific & engineering simulations

## ***Features***

- Efficient object-oriented design in Fortran 90/95 and MPI
- Automatic mesh quality control, dynamic load balancing, mesh migration, partitioning, integrated mathematics and data accessibility routines, easy solver integration
- Scalable to hundreds of processors and millions of elements using triangles (2D) and tetrahedra (3D)
- Power, completeness, and ease of use

**Charles D. Norton and John Z. Lou**

Applied Cluster Computing Technologies, Earth Science Data Systems

## ***NASA and ESTO/CT Relevance***

- Large scale modeling and simulation applications with complex geometry including support of ESTO/CT Round III teams such as earthquake fault modeling and more

## **Context For This Discussion**

## ***Very Demanding on Computer Architecture***

- Irregular application in communication and computation
- Memory intensive requiring large volumes of data

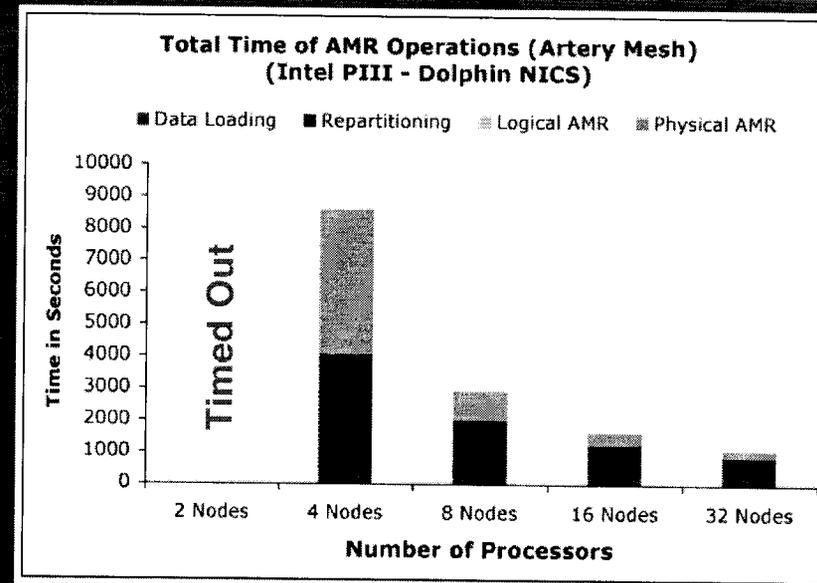
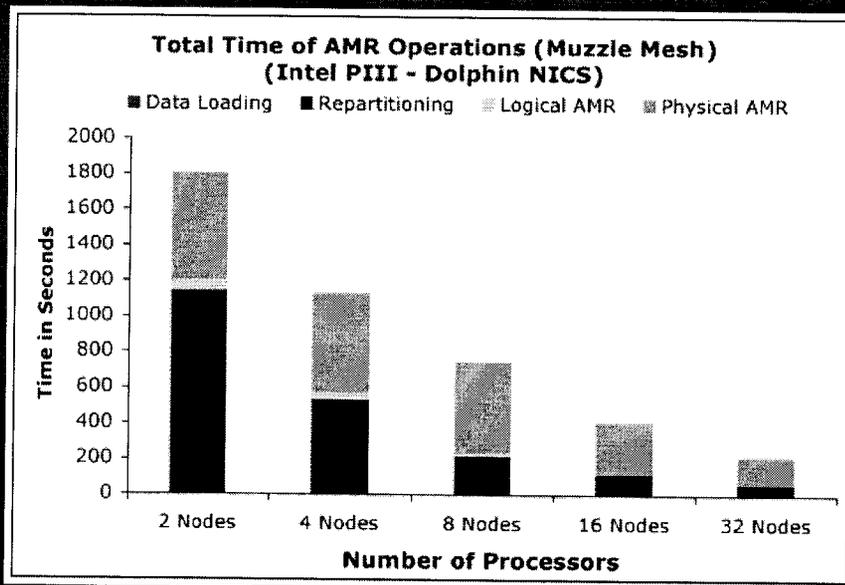
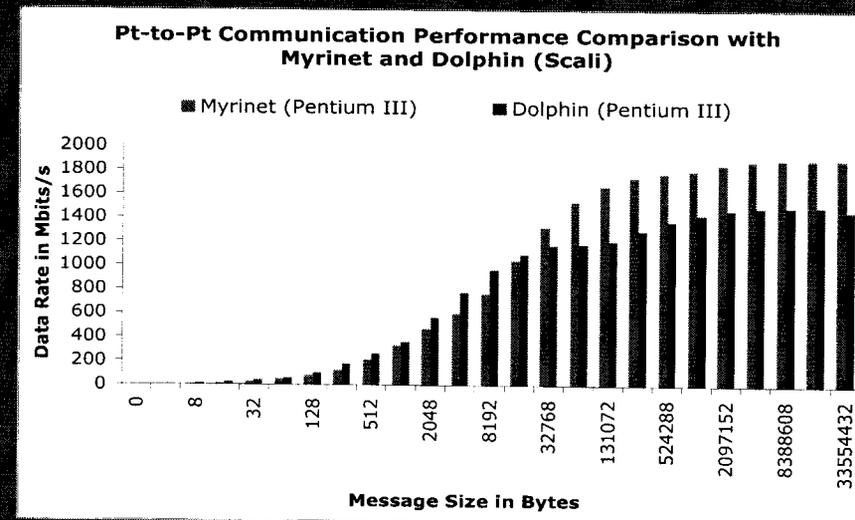
# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## Performance Analysis on Dolphin Network

- Non-switch based network with technology integrated into the network interface card
- Communication time decreases with an increase in the number of processors
- Only 32 processors available, but percentage time in major communication is relatively low



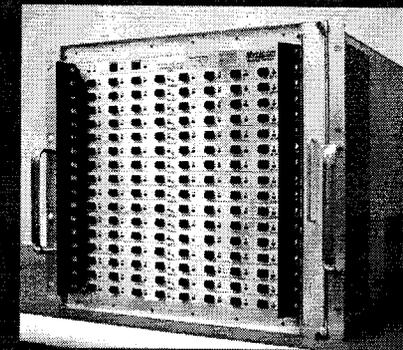
# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development

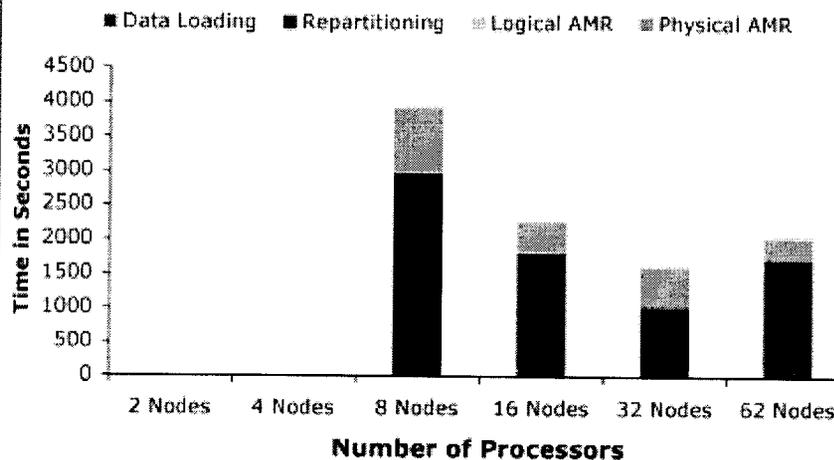


## Performance Analysis on Pentium III with Myrinet-2000

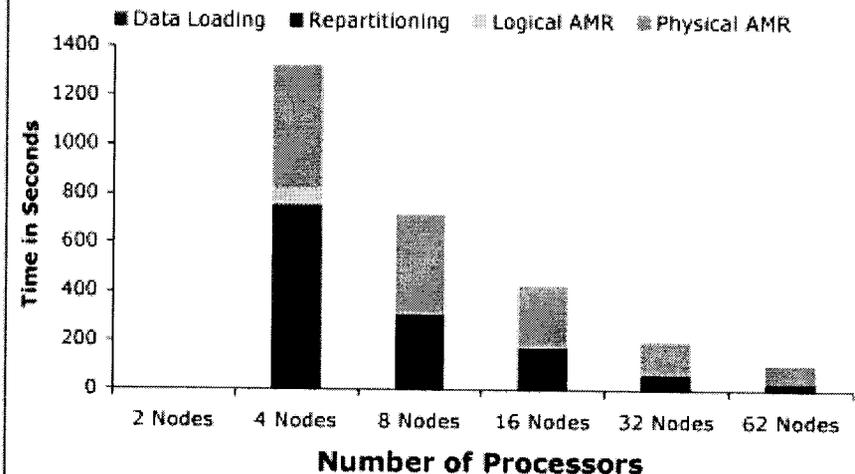
- Always seem to exhibit some problems with Myrinet, but generally useful
- Recent Linux Upgrade may indicate that performance tuning is needed since application performance is less than expected.
- Currently the “only game in town” for large scale cluster networking (thousands of processors)



**Total Time of AMR Operations (Artery Mesh)  
(Intel Pentium III - Myrinet)**



**Total Time of AMR Operations (Muzzle Mesh)  
(Intel Pentium III - Myrinet)**



# ESTO Computational Technologies Project

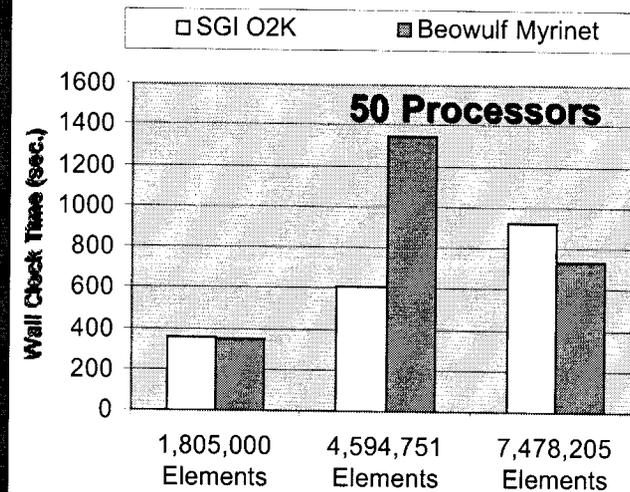
Cluster Computing Technology Research and Development



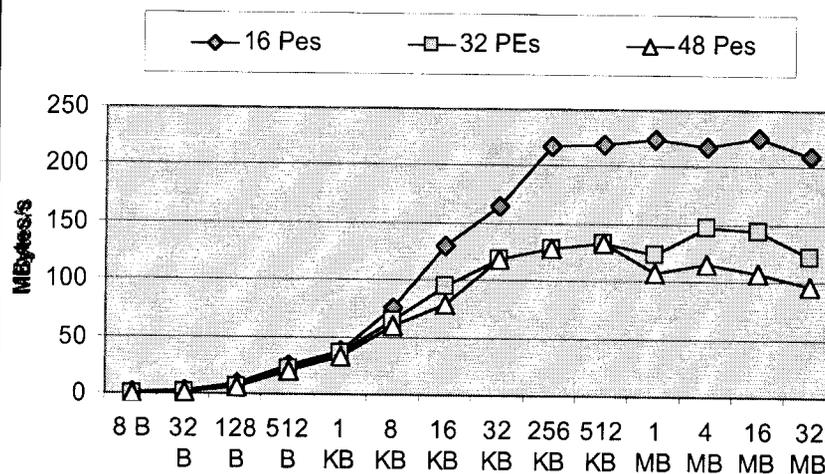
## Three Views of Performance with Myrinet

- 4.5 million element case is much slower in refinement and repartitioning
- 7.5 million element case is faster in refinement
- Normalized aggregate bisection bandwidth
- Cray T3E network remains the benchmark

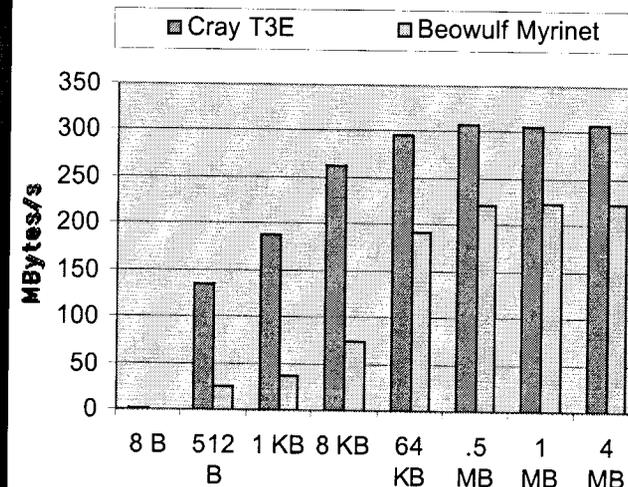
Performance for Artery Mesh Refinement



Normalized Bisection Bandwidth Across Nodes



Cray T3E Network and Myrinet 2000



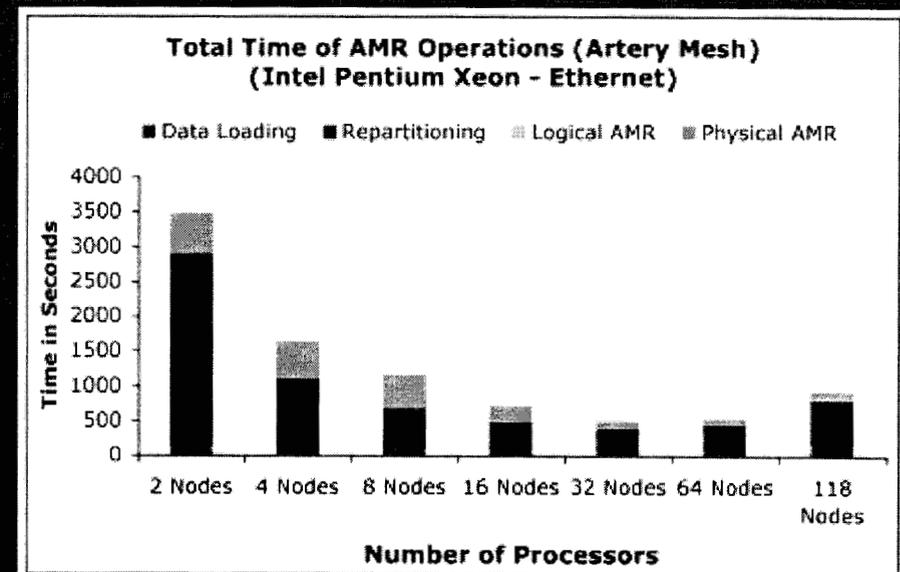
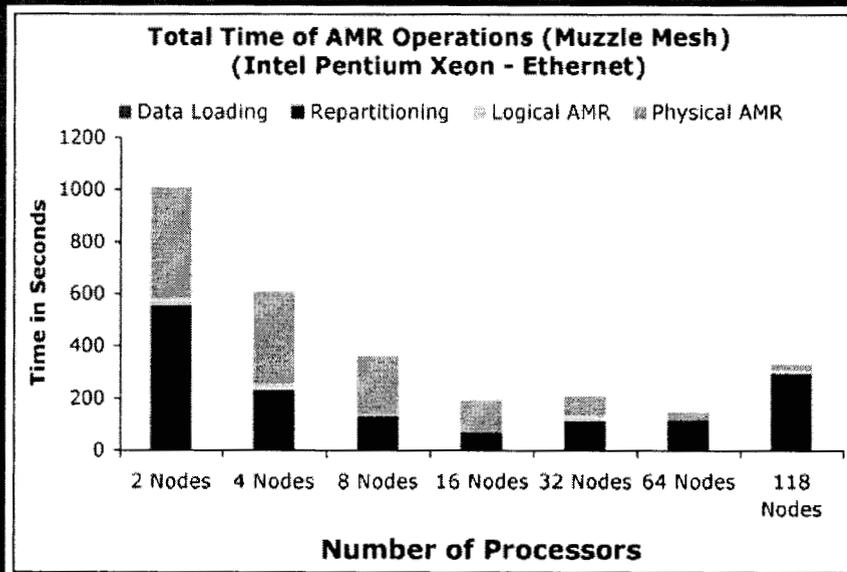
# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## Performance Analysis on Racksaver Pentium Xeon

- Ethernet system that integrates cooling and numerous processors in a single rack mounted system
- Ethernet communication is a bottleneck for dual-2.4 GHz processors where some form a gigabit networking is needed for larger numbers of processors for communication intensive applications



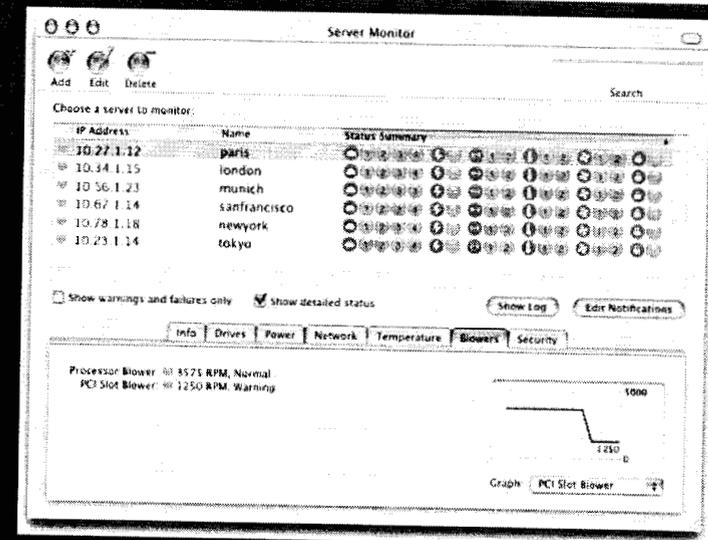
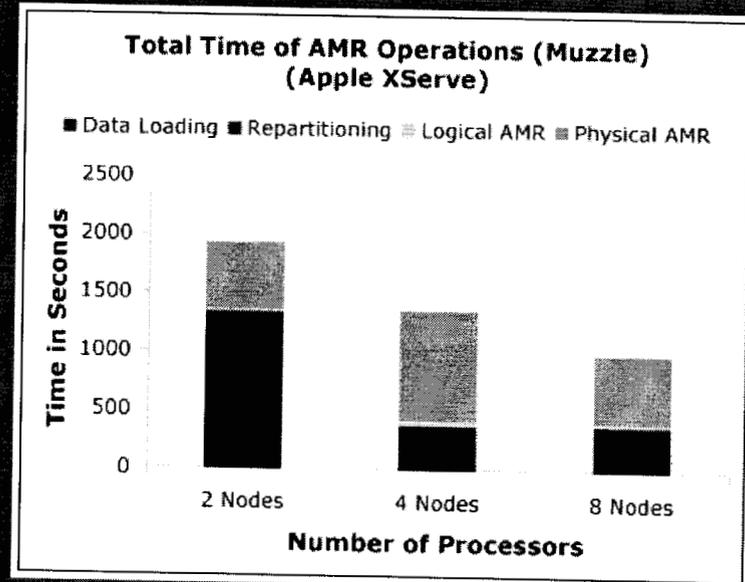
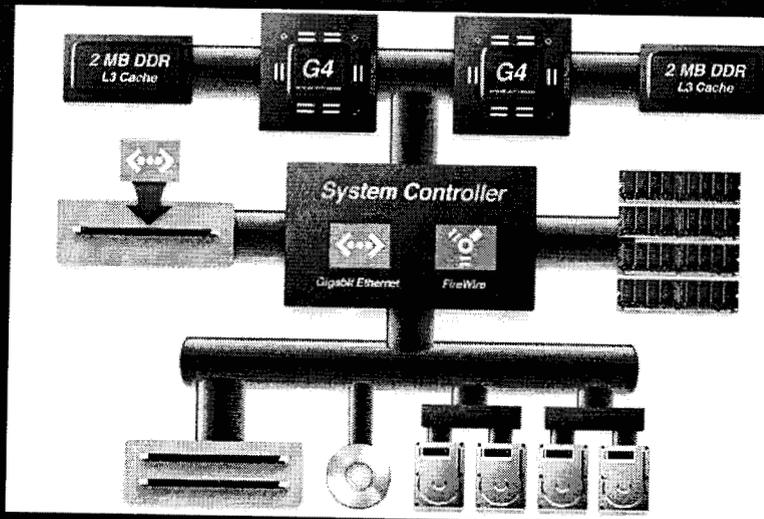
# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## Largest Xserve Cluster Installed

- 20 TB and 630 Gflops (peak) possible in a full 42 U rack
- No tools required for parts replacement
- Integrated (remote) monitoring tools
- OS X with BSD Unix (Mach Kernel) underneath
- Processes migrate among SMP nodes uncontrollably
- Many apple file services steal CPU cycles during paging for large memory intensive jobs, slowing execution



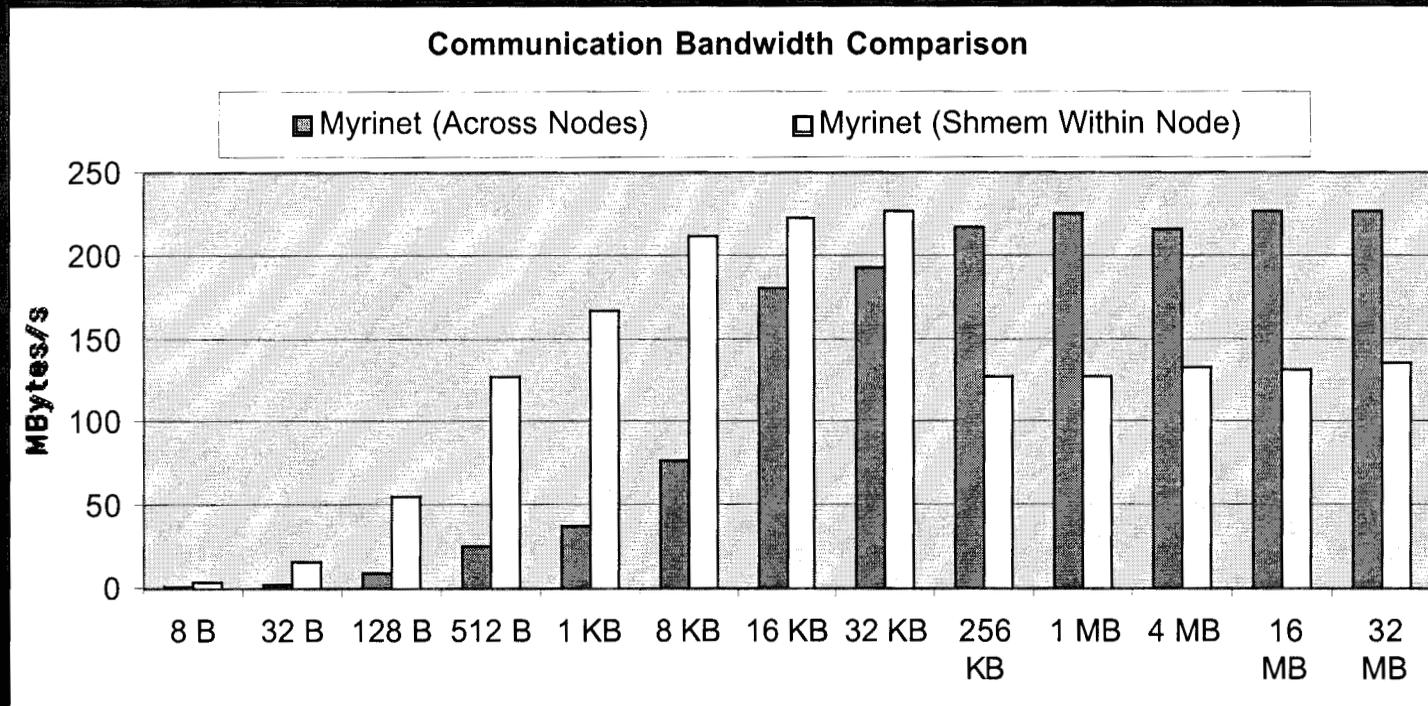
# ESTO Computational Technologies Project

Cluster Computing Technology Research and Development



## *The Effects of SMP Processors*

- Shared-Memory message passing allows fast data access from cache, but performance is cache limited.
- Performance may be improved with SMP messaging, but due to bus saturation this is generally unlikely for most applications.



# ESTO Computational Technologies Project

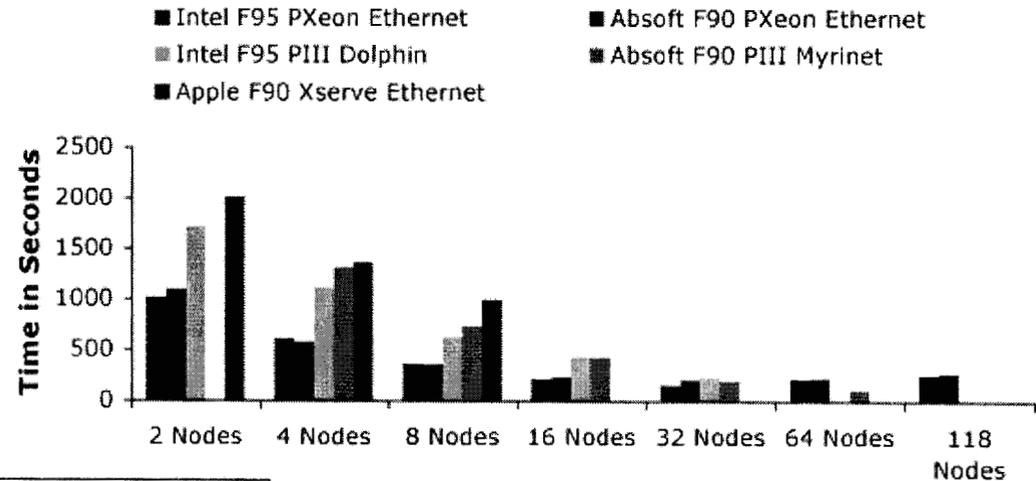
Cluster Computing Technology Research and Development



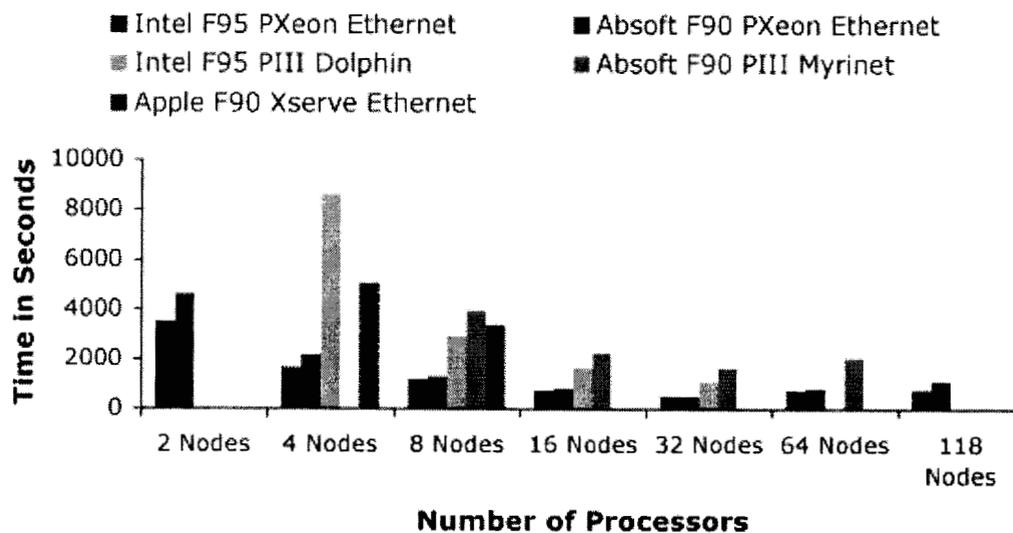
## System Performance Comparisons

- 2.2 Ghz processors on slow networks can outperform slower processors on fast networks
- Need to examine overhead of Apple system processes and tools
- Compiler issues remain a problem, explaining data gaps

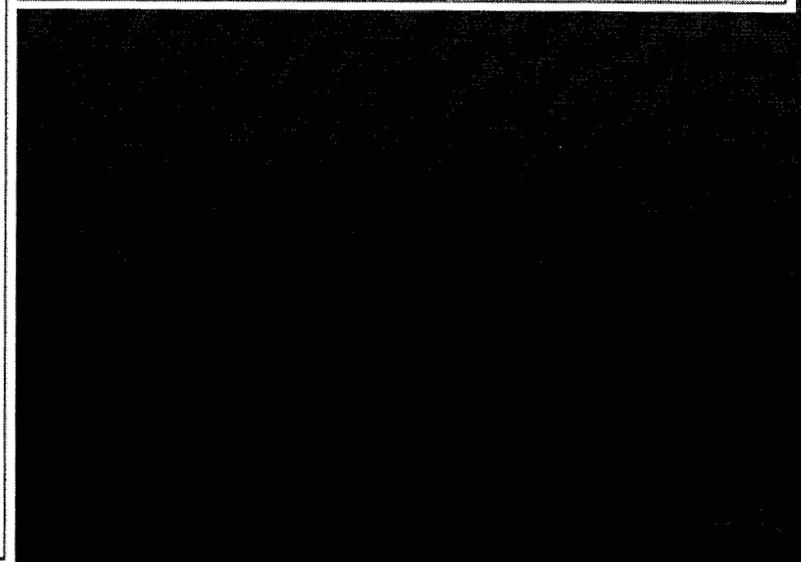
### Muzzle Mesh 3 Level Adaptive Refinement



### Artery Mesh 1 Level Refinement



### Number of Processors



ESTO Computational Technologies Project



# Identifying JPL's Cluster Community

**Email:**

**[Charles.D.Norton@jpl.nasa.gov](mailto:Charles.D.Norton@jpl.nasa.gov)**

Applied Cluster Computing Technologies  
Earth Science Data Systems Section  
Jet Propulsion Laboratory  
California Institute of Technology

2002 JPL Information Technology Symposium

