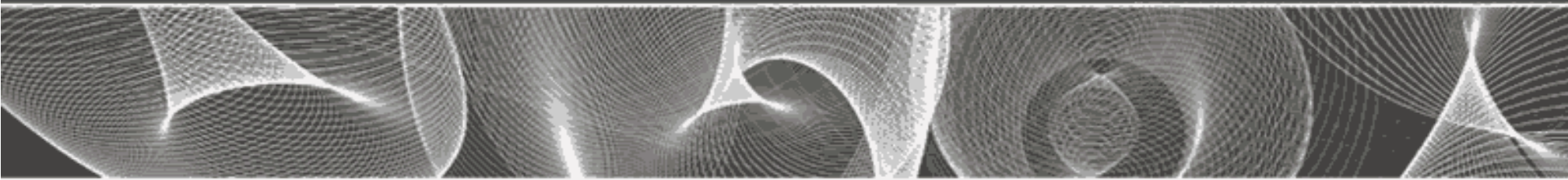


# QUADRICS IN LINUX CLUSTERS

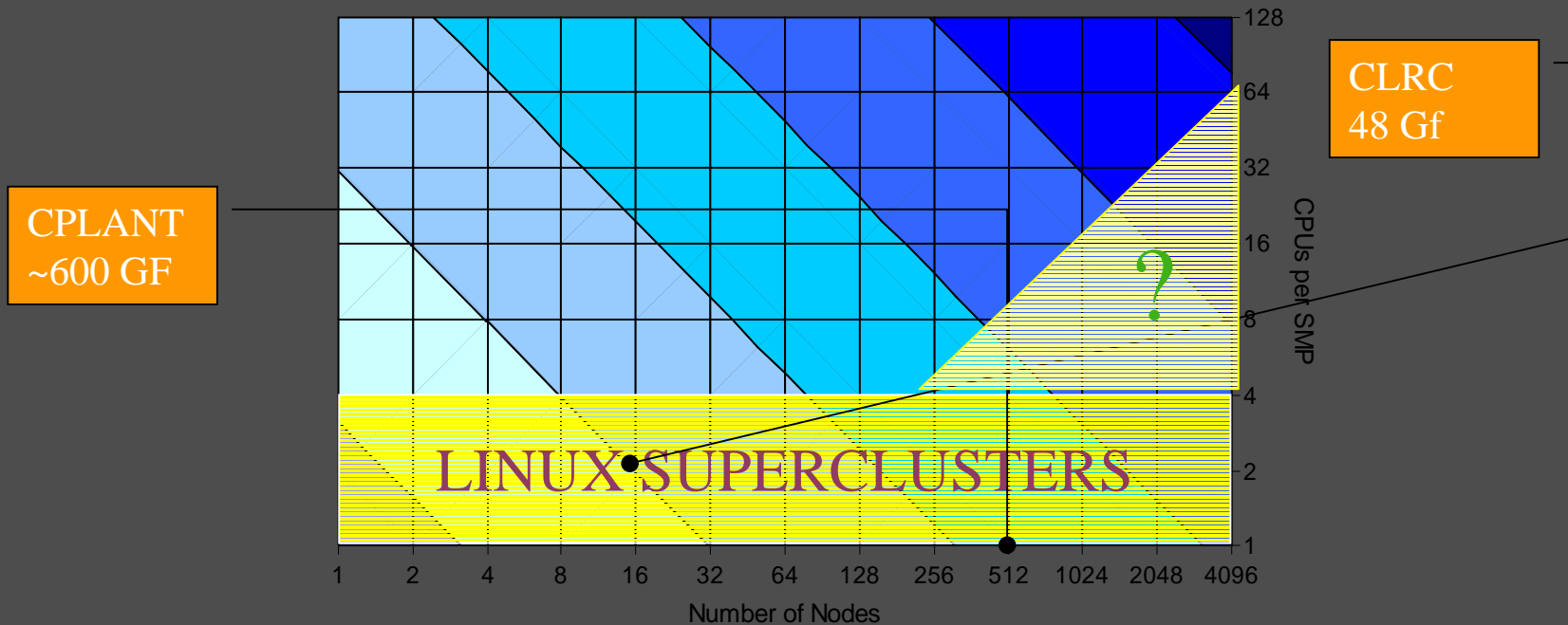
John Taylor



## QLC 21/11/00

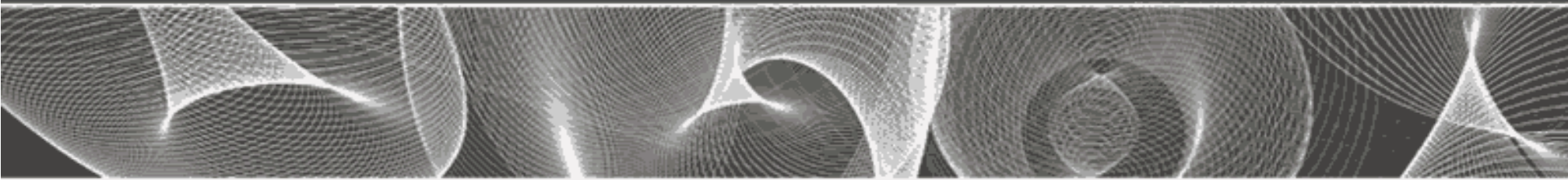
- Motivation
- Quadrics Cluster Products
- Performance
- Case Studies
- Development Activities

## Super-Cluster Performance Landscape



## Where is the HPC Market

- HPC has migrated from MPP to:
  - Clustered Shared Memory Systems
    - AlphaServer SC, IBM SP and Vector Machines
  - LINUX Clusters (Alpha or x86)
    - “Commoditized Network”
  - Quadrics solving the differential
    - degree of coherence/SSI - programming model, manageability, administration



## LINUX Pros and Cons

- Open Source
- Wide Availability
- Early Availability
- Cheaper
- Leverage MPP S/W

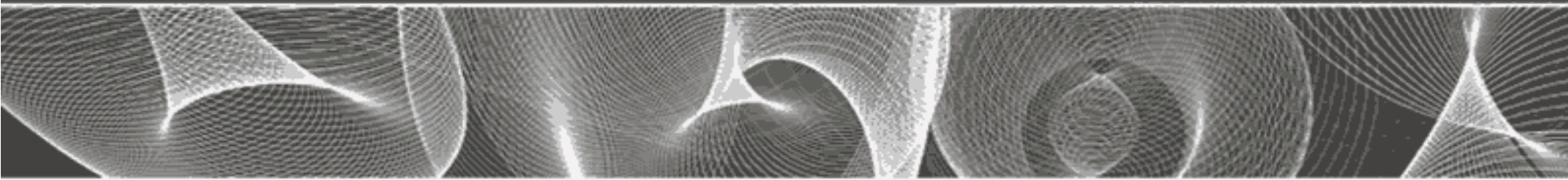
- Support
- Security
- Scalability
- TPSW Availability
- Performance



## Business Strategy

Technology Leadership in High Performance  
Interconnect and Cluster Management Software (QsNet  
and RMS)

Technology and Business Partnerships  
Creation of Channels  
HPC Services e.g Integration of v. large LINUX clusters



## QSW extensible HPC Cluster Components

- Generic Technology for tightly coupled clusters of SMP's
- Where tightly coupled means:
  - A hardware interconnect capable of scaling in both the number of SMP's and the number of CPU's per node
  - A "SSI" providing a coherent view of the system as a single entity.
  - The provision of application development environments consistent with the DSM model



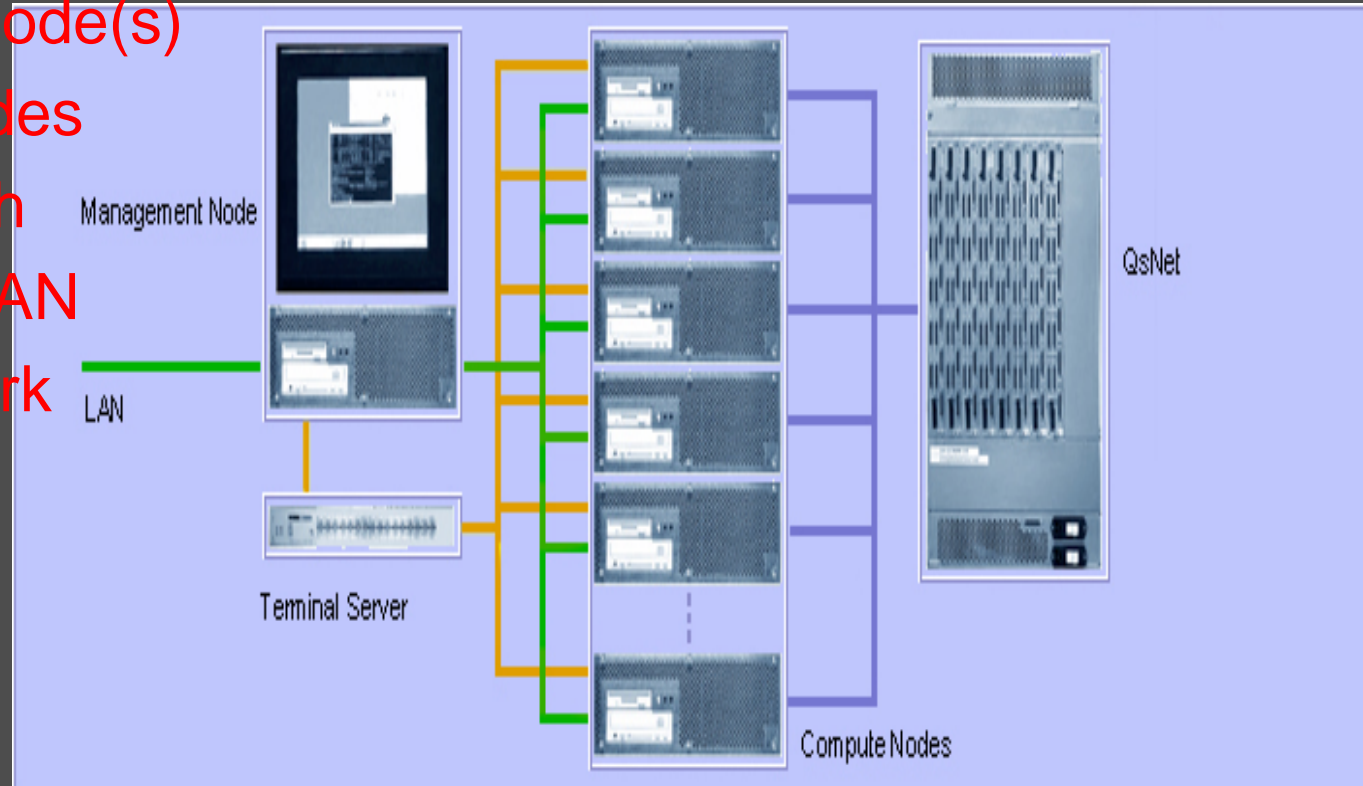
## High Performance Cluster Products from Quadrics

High performance interconnect  
Resource management system  
Parallel application development  
tools  
Integrated TPSW Support



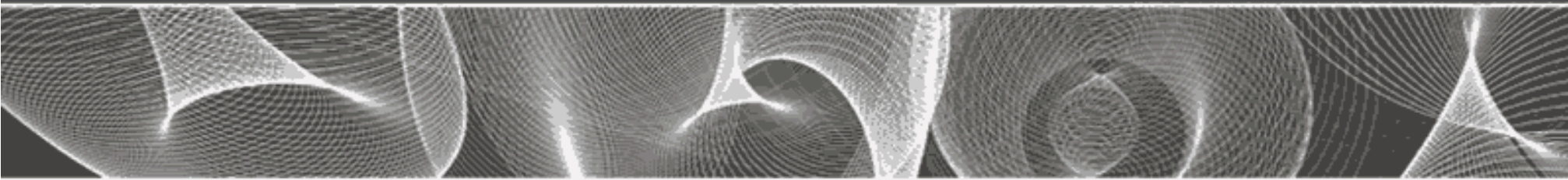
## Architecture Overview

- Management Node(s)
- Processing Nodes
- Quadrics Switch
- Management LAN
- Console Network
- Disk Array



## Quadrics Interconnect (QsNet)

- Two Custom Design ASICs make up the network
  - Quadrics Network Adapter (elan)
    - 2nd Generation - 64bit/66 MHz. PCI -Based
    - Very Low Latency , High Bandwidth
  - QSW Multi-Stage Network (elite)
    - Modular Design , Fat Tree Topology
- Combined to provide high scalability, flexibility and tolerance



## Network Adapter **Network Components**



16 way Switch Card



128 way Switch Chassis

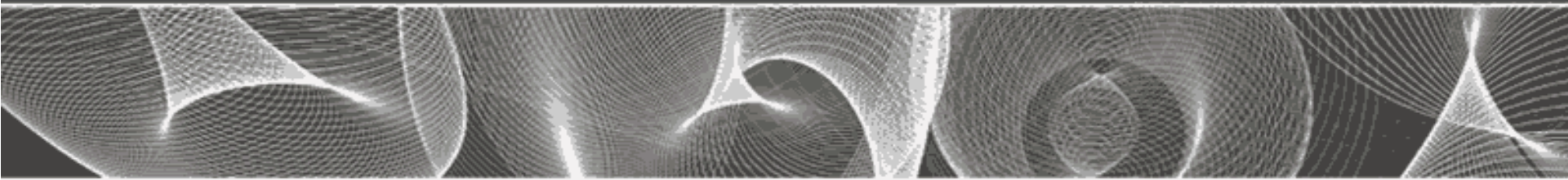


## QsNet Adaptor

- Intelligent PCI adapter
  - DMA engine
  - microprocessor w/64MB SDRAM
- One Sided Communications
  - Get/Put
  - Send/Receive (TPORTS or Queued DMAs)
- OS Bypass with Virtual Addressing
  - no page locking or copying
  - full protection

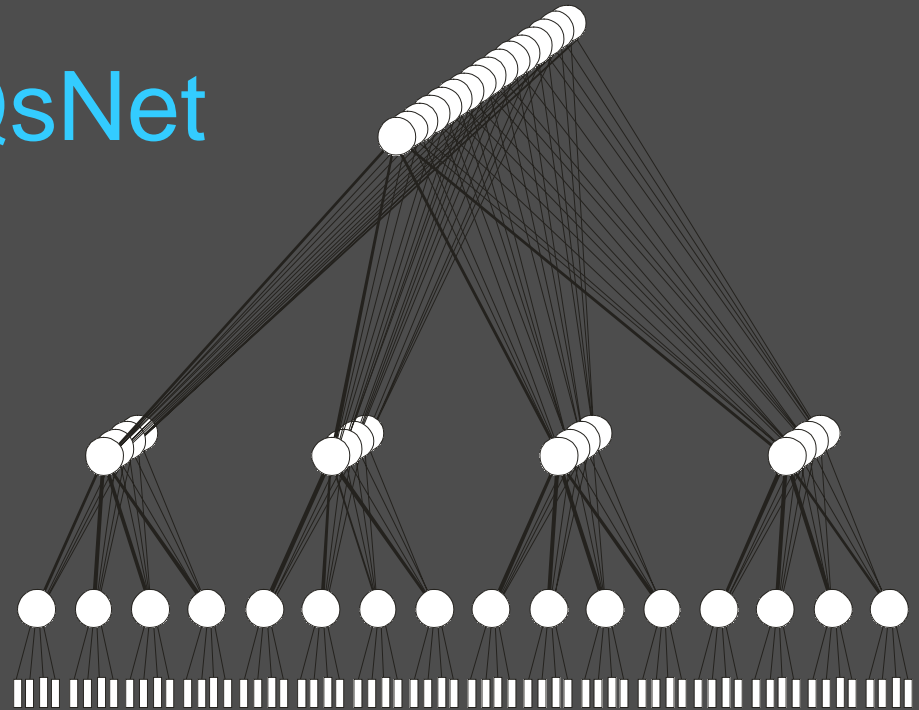






## QsNet

- Full bi-sectional bandwidth
- Logarithmic cost
- Multiple routes
- Hardware broadcast
- General purpose topology

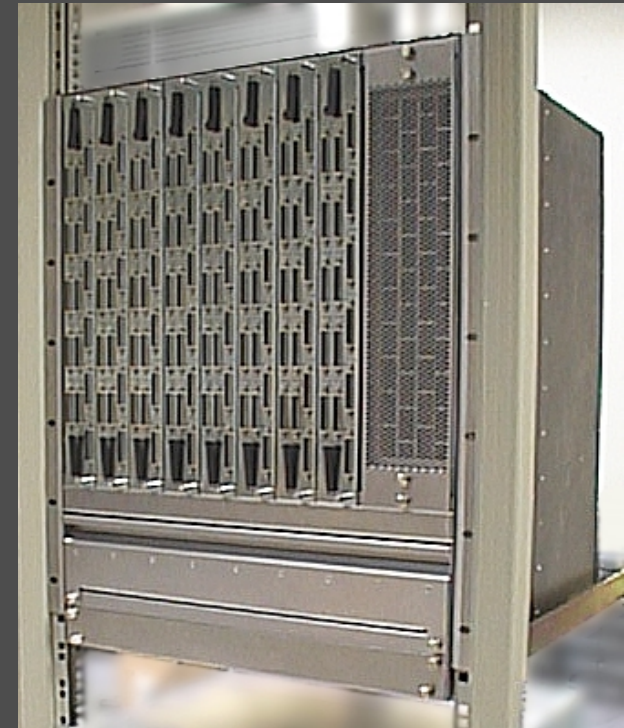




## Quadrics 128-way Switch

### Features

- 1-8 × 16 way switch cards
- 16 × 8 top switch cards
- Passive mid-plane
- 2 from 3 fault tolerant 48V PSU
- Live insertion
- Temperature, fan and PSU status
- Full JTAG boundary scan
- **Performance**
  - 42.5 Gbytes/sec bi-sectional bandwidth
  - 175 ns latency



## Quadrics Cluster Software

- Standard Software Hierarchy + Enhancements to Couch Parallelism
  - Inter Processor Communication
  - Single Point Cluster Management (Switch, Console)
  - Scheduling of Parallel Programs
  - Scalable File System
  - Accounting and Monitoring
  - High Availability Strategies
  - TPSW Support

## Quadrics Software

### RMS Products

Baseline (“RMS-lite”) free with hardware

Value Added Product

### Operating Systems

Tru64 UNIX V5.0 and V5.1

Alpha Linux 2.2.14 and 2.4

Intel Linux 2.4

Solaris 2.6

## Quadrics Software Components (1)

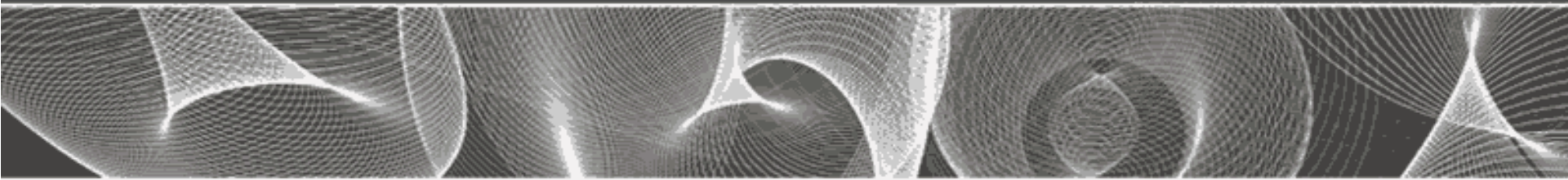
- Baseline Product
- QsNet Linux drivers
- IP over QsNet
- MPI/SHMEM optimized for QsNet
- QsNet diagnostics
- Documentation (electronic)



## Quadrics Software Components (2)

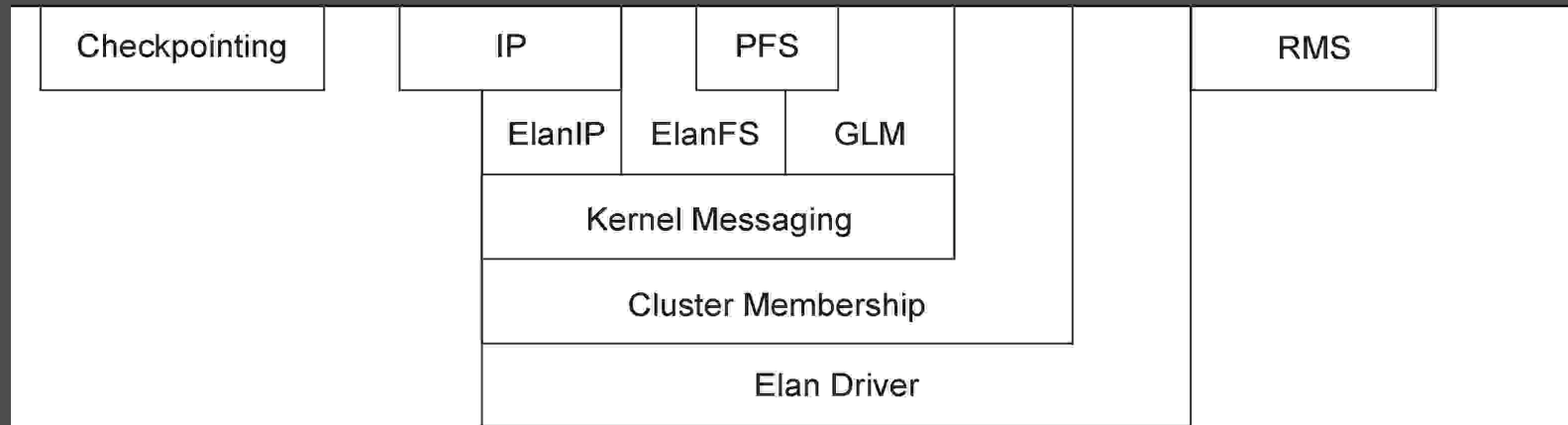
- Full Product
  - Single point installation and system management
  - RMS parallel job scheduler
  - Pandora Graphical User Interface
  - Filesystem over QsNet
  - TotalView support
  - Documentation (electronic plus one paper copy)
- The full product will be supplied as RPMs for the current product release of RedHat Linux fully qualified on a range of platforms and licensed using flexlm. Sources will be available to customers under a "no commercial reuse" license.





## Software Overview

- Kernel Services



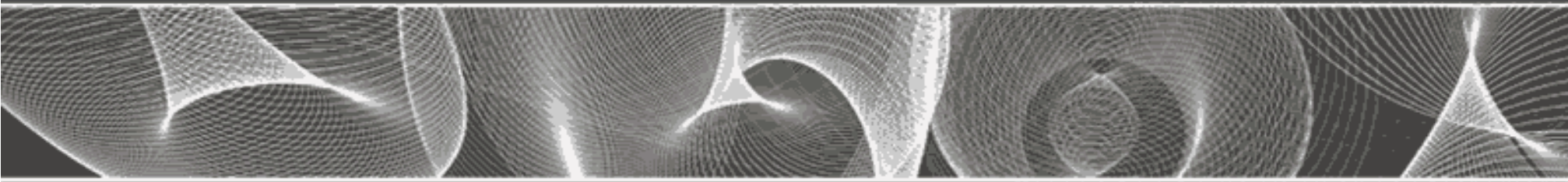
## Cluster Services

- Node status monitoring
  - bitmask of the functioning set of nodes
- Console logging
- Automated installation (Linux)
- Graphical User Interface – Pandora



## Application Development

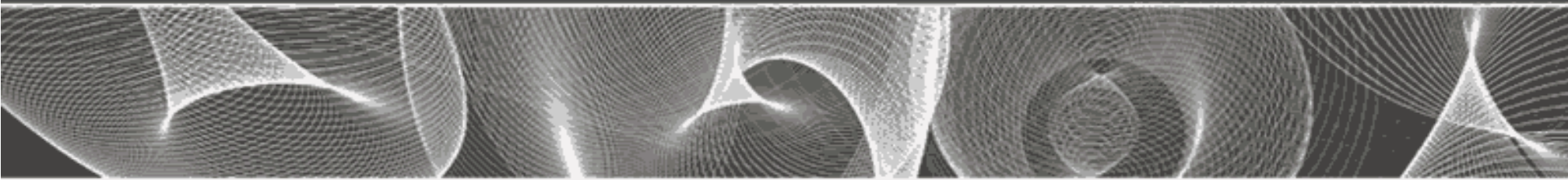
- Standard compilers
- ATLAS Blas libraries
- MPI and Shmem
- Totalview
- VAMPIR (soon)
- PBS and LSF (future)



## Performance

### Standard Benchmarks

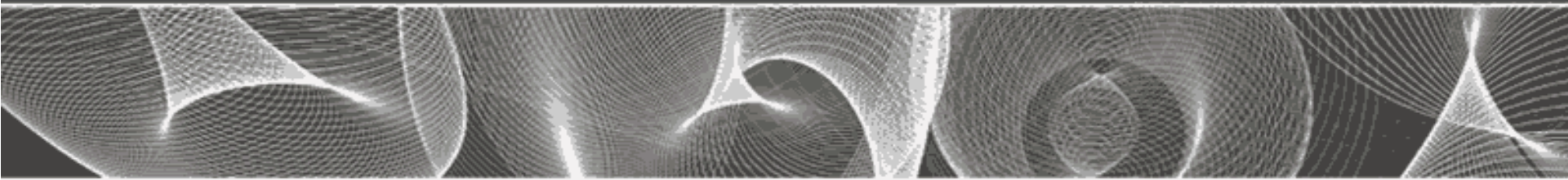
### Application Specific



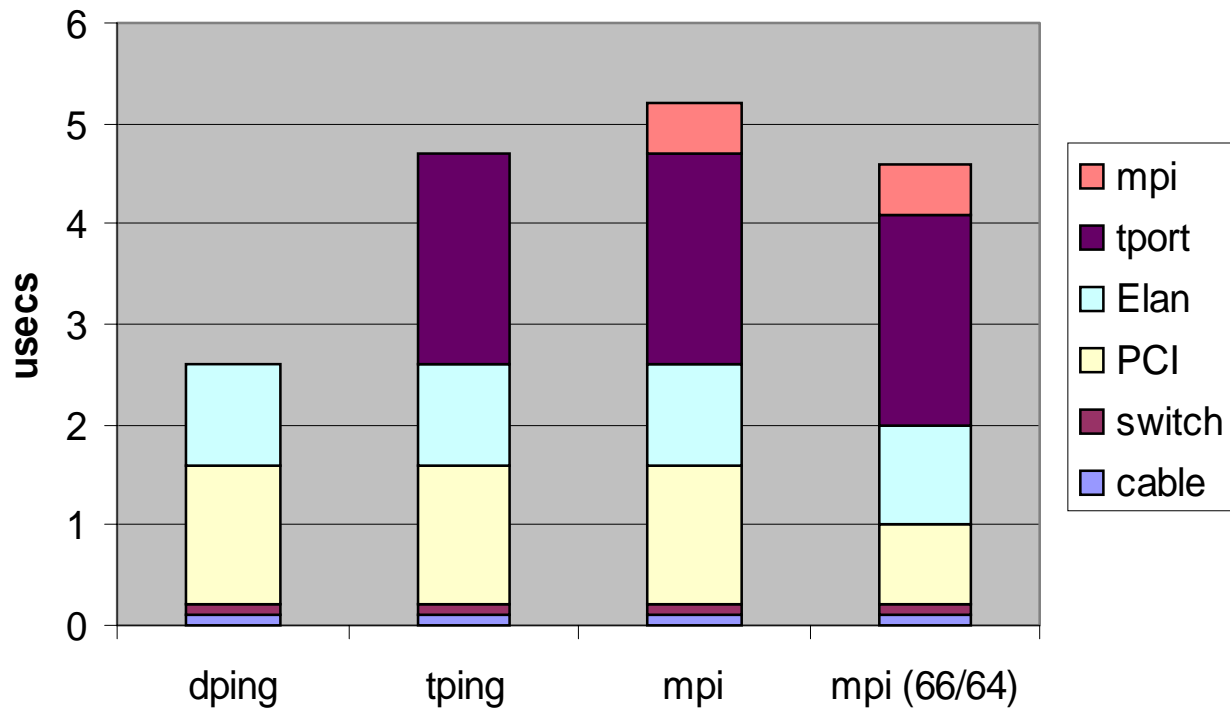
## Performance Overview

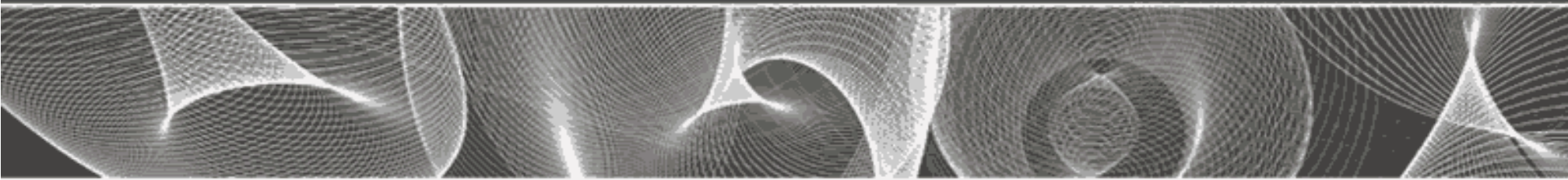
- Line rate 400 Mbytes/s
- Peak data rate (adapter memory) 340 Mbytes/s
- MPI Send (33Mhz/64bit) 200 Mbytes/s
- DMA 2.5  $\mu$ s
- MPI send 5  $\mu$ s
- MPI Send (66MHz/64bit) 307 Mbytes/s
- DMA write 1.7 usec
- MPI send 4.5 usec





## Rough Latency Budget





Computational Science and Engineering Department

Daresbury Laboratory

## Molecular Modelling on High-End and Commodity-Type Computers: Status and Perspectives

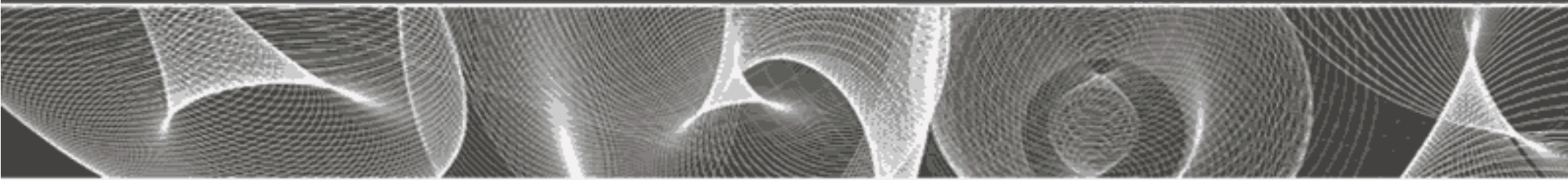
Martyn F. Guest and Paul Sherwood  
CCLRC Daresbury Laboratory

m.f.guest@daresbury.ac.uk

<http://www.cse.clrc.ac.uk/Activity/QUASI>

## Development Activities

- Porting to IA-32 and IA-64 Intel Systems
- Extending Current Generation network
  - Tracking the increase in Node “fatness”
  - Increasing Node Count (Distributed Switch)
- Next generation network
  - Tracking the increase in CPU performance
  - Fibre Interconnect - EMC is better. Copper is cheaper
- Software
  - High Availability Strategies
  - High Performance File Systems



## References

- [Http://www.quadrics.com](http://www.quadrics.com)
- <http://www.compaq.com/hpc>
- [http://www.c3.lanl.gov/cic3/teams/par\\_arch/Publications.html](http://www.c3.lanl.gov/cic3/teams/par_arch/Publications.html)
- <http://www.cse.clrc.ac.uk/Activity/QUASI>
- <http://www.psc.edu>