Dawning4000A

A 10TFlops Linux Cluster

Ninghui Sun
snh@ncic.ac.cn
Institute of Computing Technology
Chinese Academy of Sciences

Dec 3th, 2003. Cluster2003

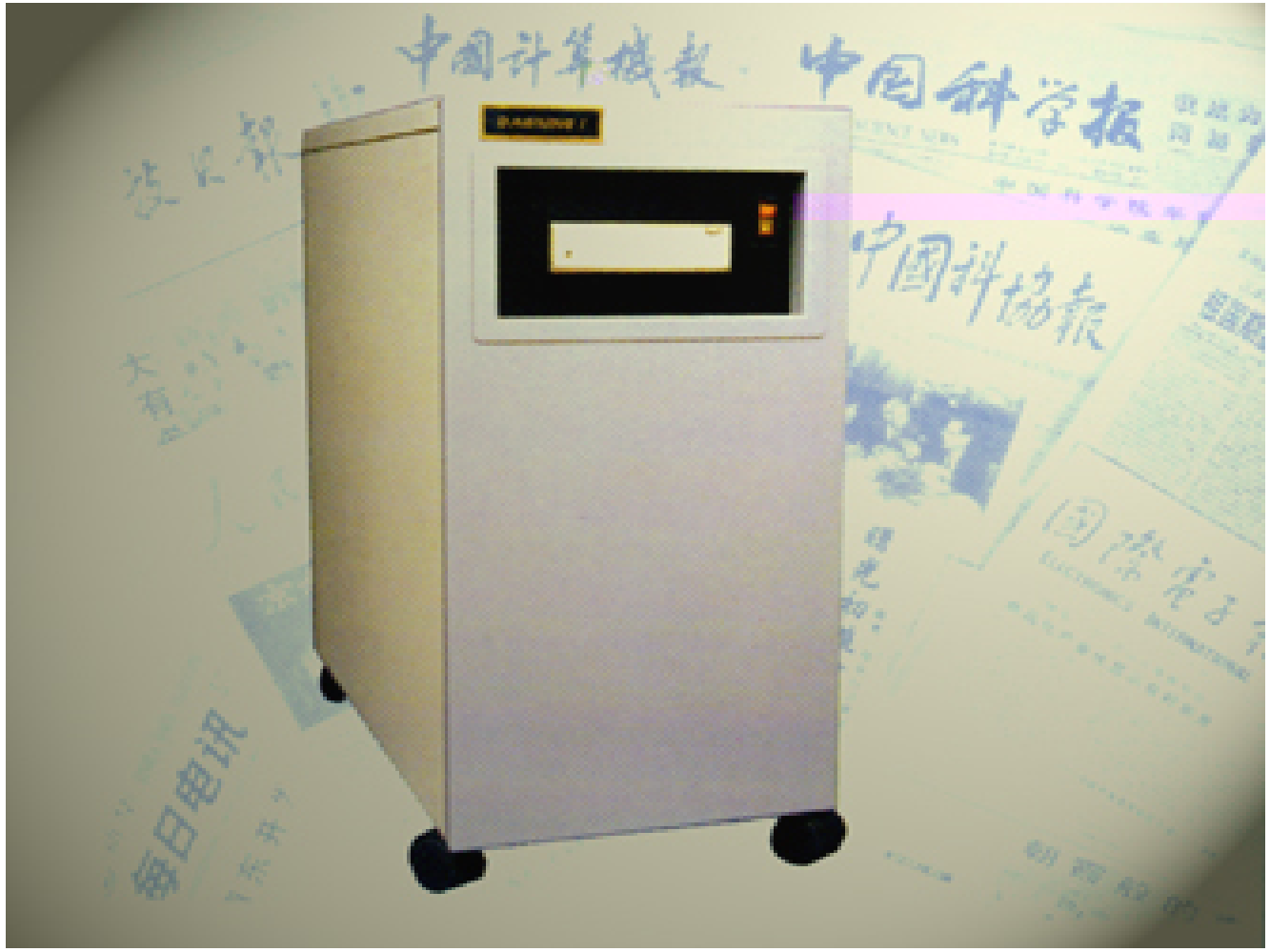
Background

中国科学院计算技术研究所

Institute of Computing Technology,Chinese Academy of Sciences

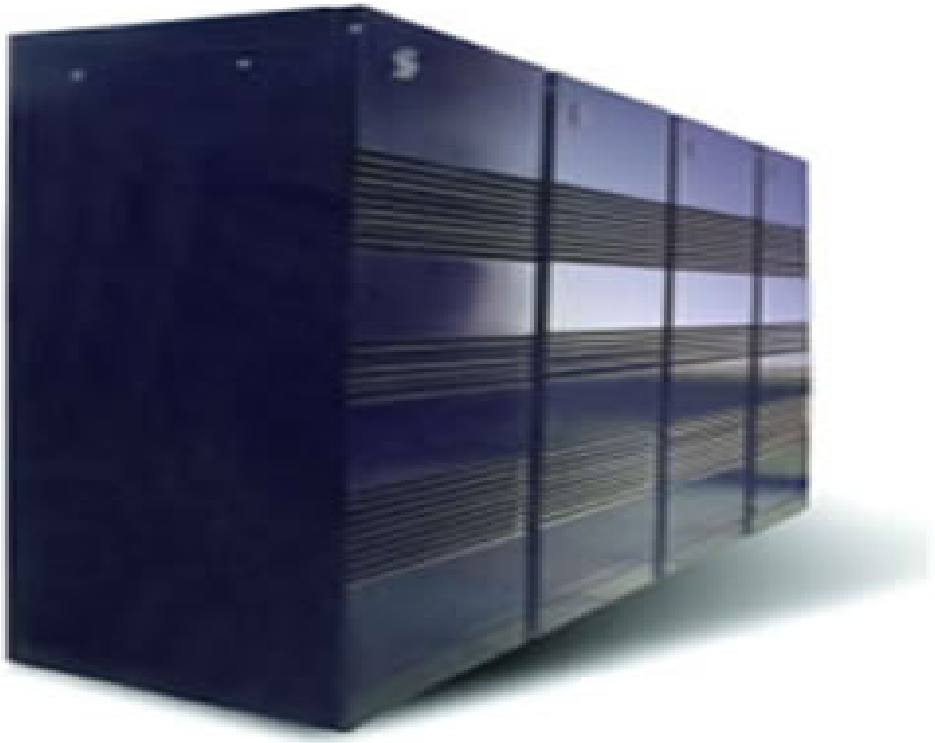
Established in 1956

First computing research institute in China



Dawning-1
First SMP Computer in China
1993

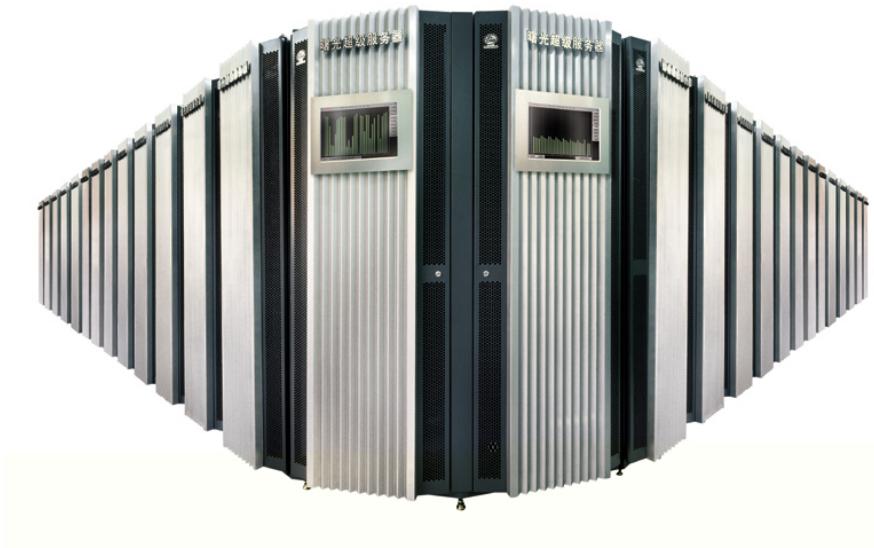
Dawning1000
First MPP Computer in China
1995



Dawning2000-I
First Cluster Computer in China
1998

Dawning2000-II
First SMP Cluster in China
1999

Dawning3000
SUMA Cluster in China
2001



Dawning4000-L
3Tflops Linux Cluster in China
2003

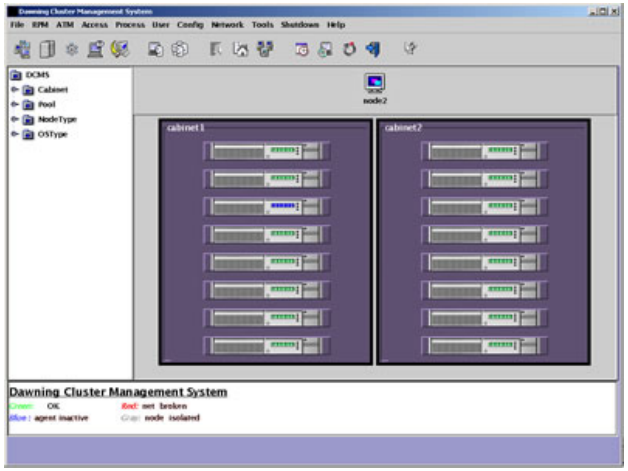
DAWNING TC4000L SUPER COMPUTER

Dawning4000-A
Grid-enabling Cluster in China
2004

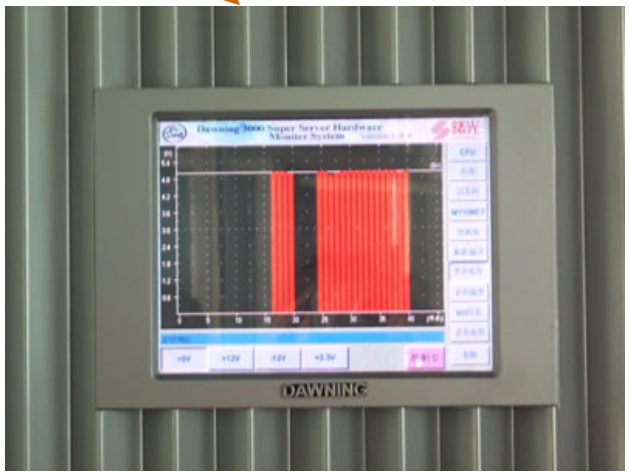
Industry Impact

Inside Console

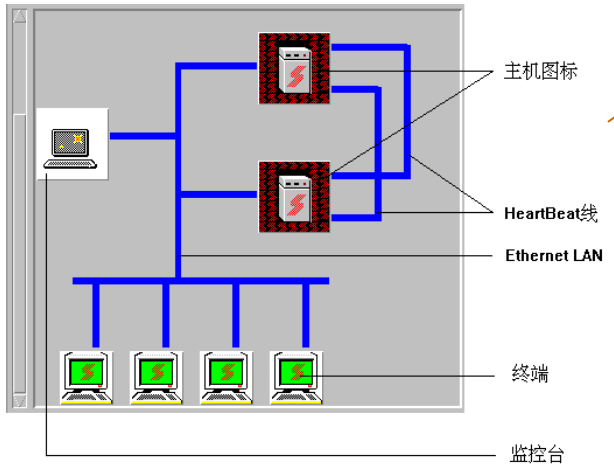
Cluster Management



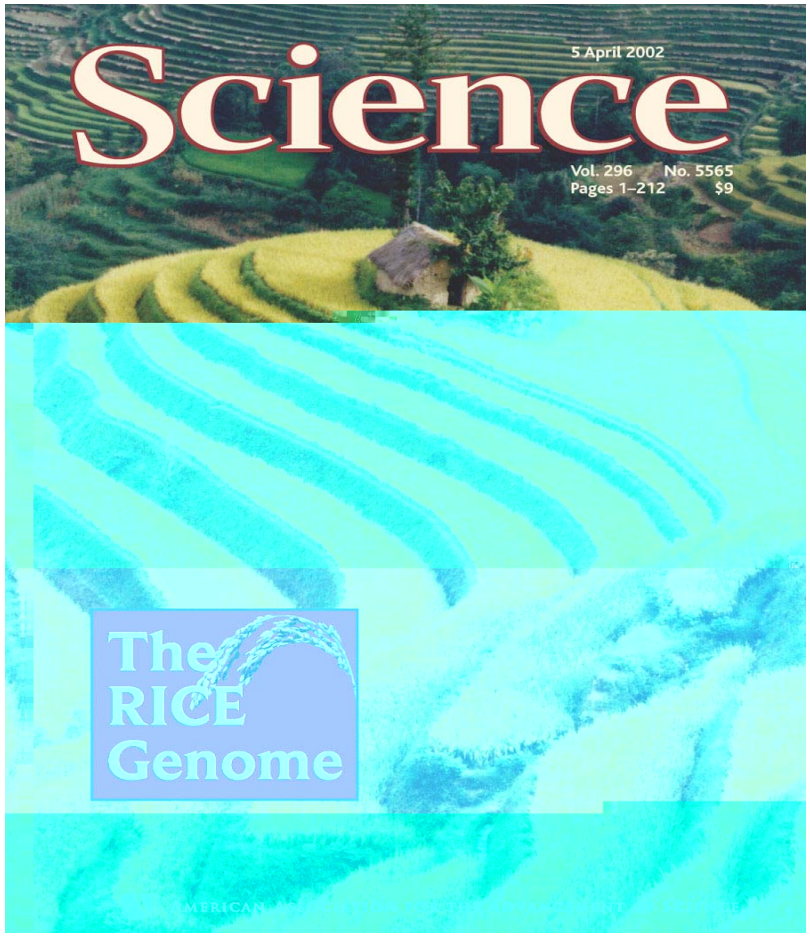
Hardware Monitor



HA Software



Draft Sequence of Rice Genome



National High Performance Computing Environment

Petroleum Industry

Changing the business rules

曙光服务器
DAWNING SERVER

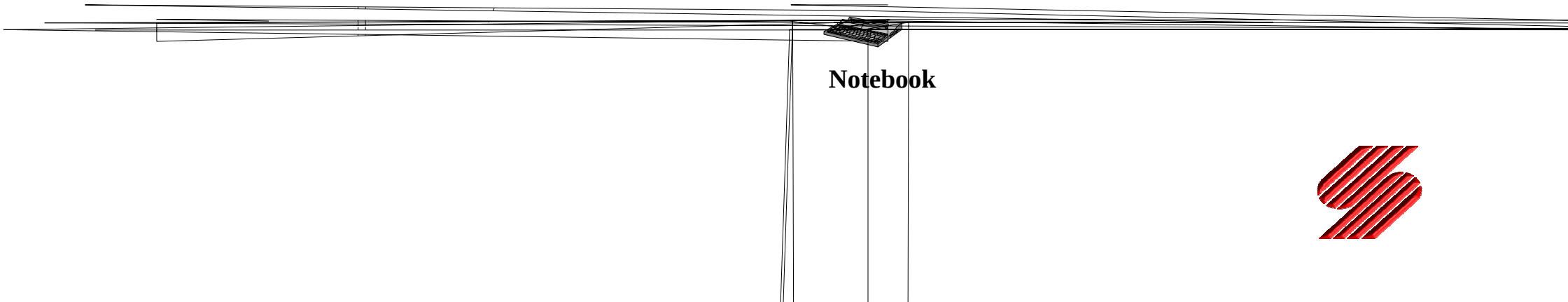
有科技 就有奇迹

Dawning Server Family

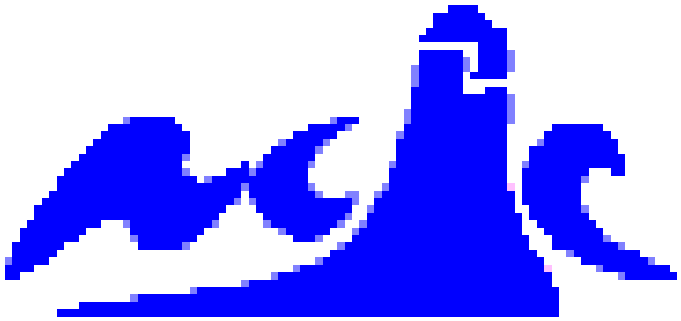
Dawning4000A

CNGrid

Resource-sharing

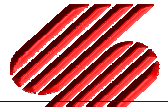


Joint Project

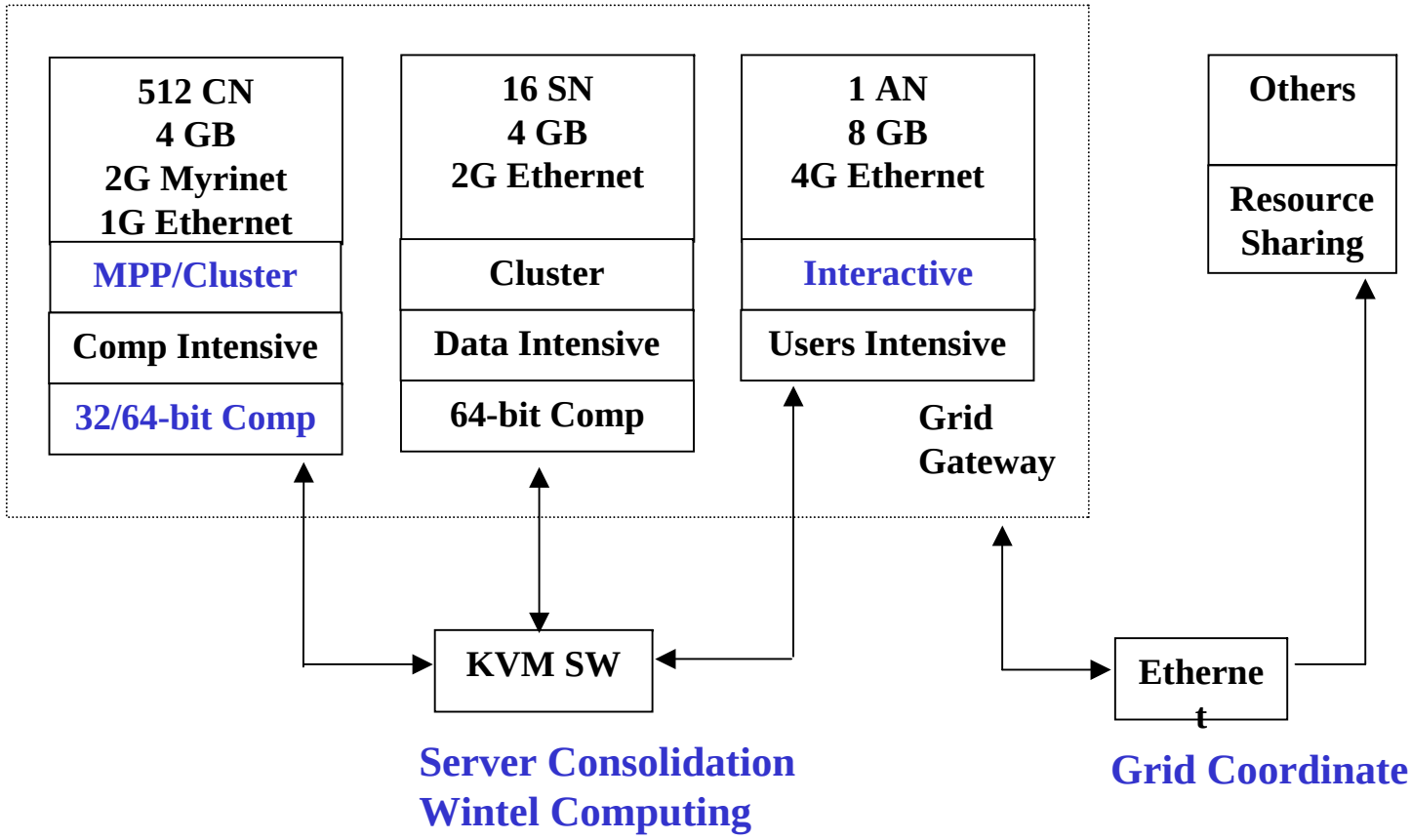


Hardware

- 512 Computing Nodes (4 CPU 4GB 36GB
- 16 Storage Nodes (overlap)
- 1 Access Node (4 CPU 8GB 36GB
- 16 RAIDs 1.3TB
- 4 Ethernet Switches
- 12 Myrinet Switches, 512 Adapters
- 2 Management Switch 512 Management Cards
- 4 Management Consoles
- 40 Cabinets
- 1 Grid Router
- 4 Grid Keys
- 1 System Monitor



Application Models



Industry Standard Cluster

Industry Standard Architecture

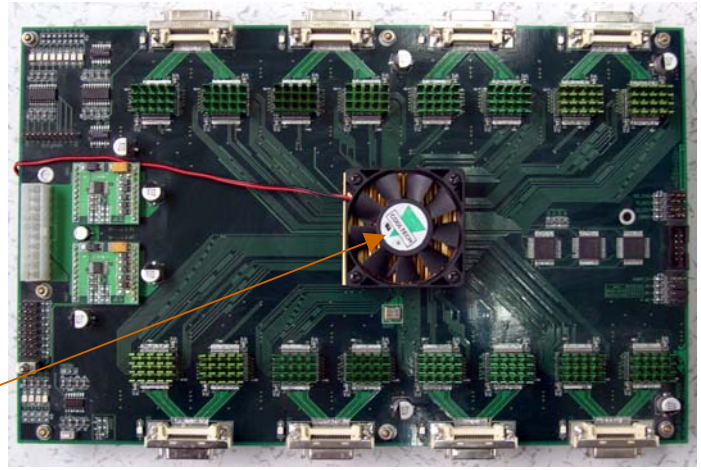
- Xeon/Opteron/Itanium
- SRAM/DDR
- HyperTransport/PCI Express
- SCSI/FC/SATA/iSCSI
- Myrinet/Quatric/Infiniband/Ethernet
- Chassis/Cabinet/Cable
- Linux/GNU Compiler/Java
- GM/VIA/Verb/MPI/PVM/uDAPL
- MKL/ESSL/ACML/Scalapack/Gauss
- CMS/PBS/LSF/Luster
- Paradiam/LS-DYNA/Cerius/MM5/Blast/Oracle RAC
- LinuxNetworx/Scali

Cluster Technology

- Cluster Switch
- Parallel Environment
  - BCL4
  - DMPI/DPVM
  - JobManager
  - SVM
  - Debugger
- DCOS
  - CSMS
  - DCIS
  - DCMM
  - MultiTerm
- DCFS
- Clustone

DCNet

- 5Gbps Switch
- 8x8 Crossbar
- 250 ns switching
- Up to 1024 Nodes
- Xscale Communication Processor



UX8 Chip

Basic Communication Layer

BCL-1

- Kernel level communication protocol for Dawning2000-I

BCL-2

- 0-copy user-level communication protocol for Dawning2000-II

BCL-3

- Semi-user level communication for Dawning3000

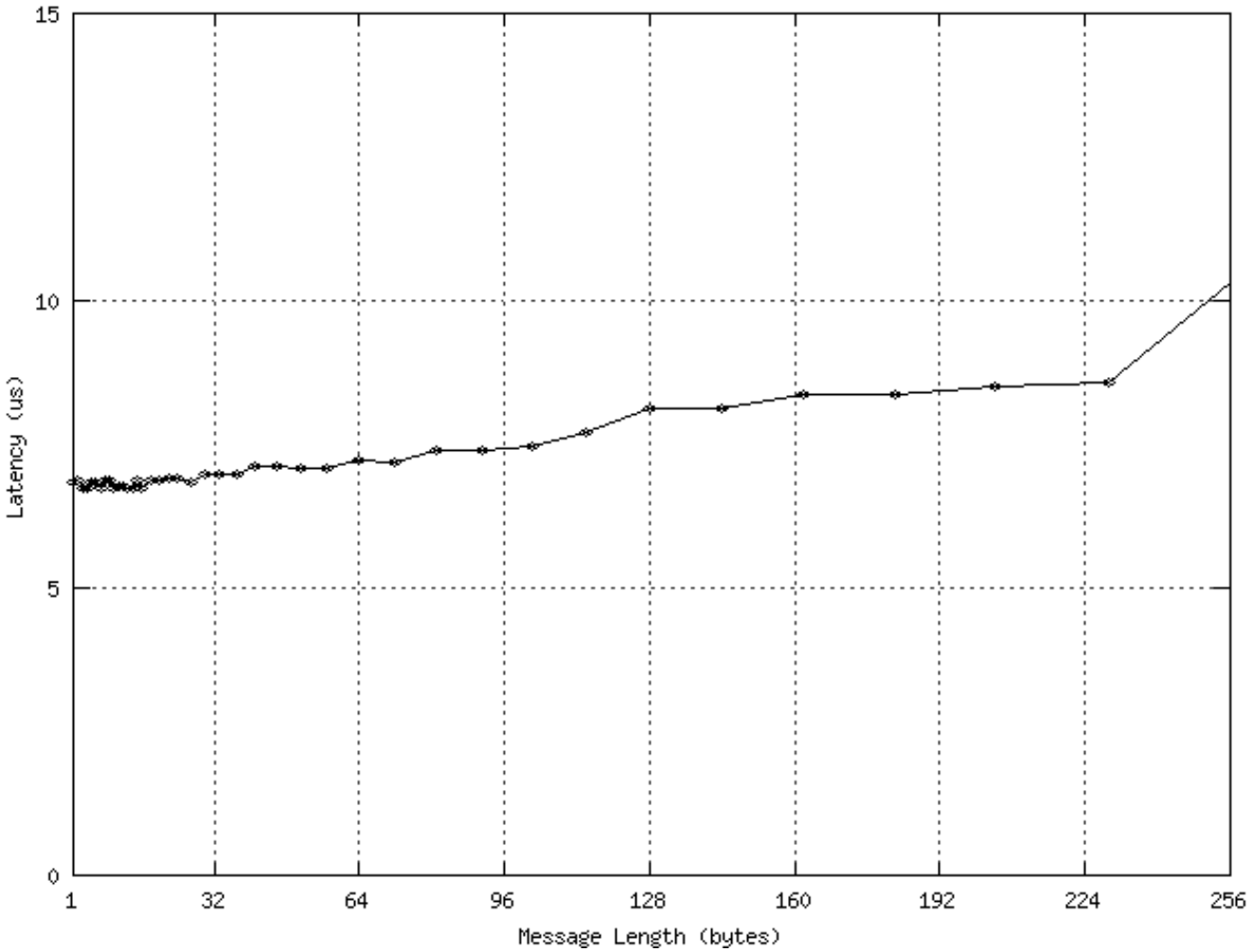
BCL-4

- Parallel and Grid communication protocol for Dawning4000

BCL4

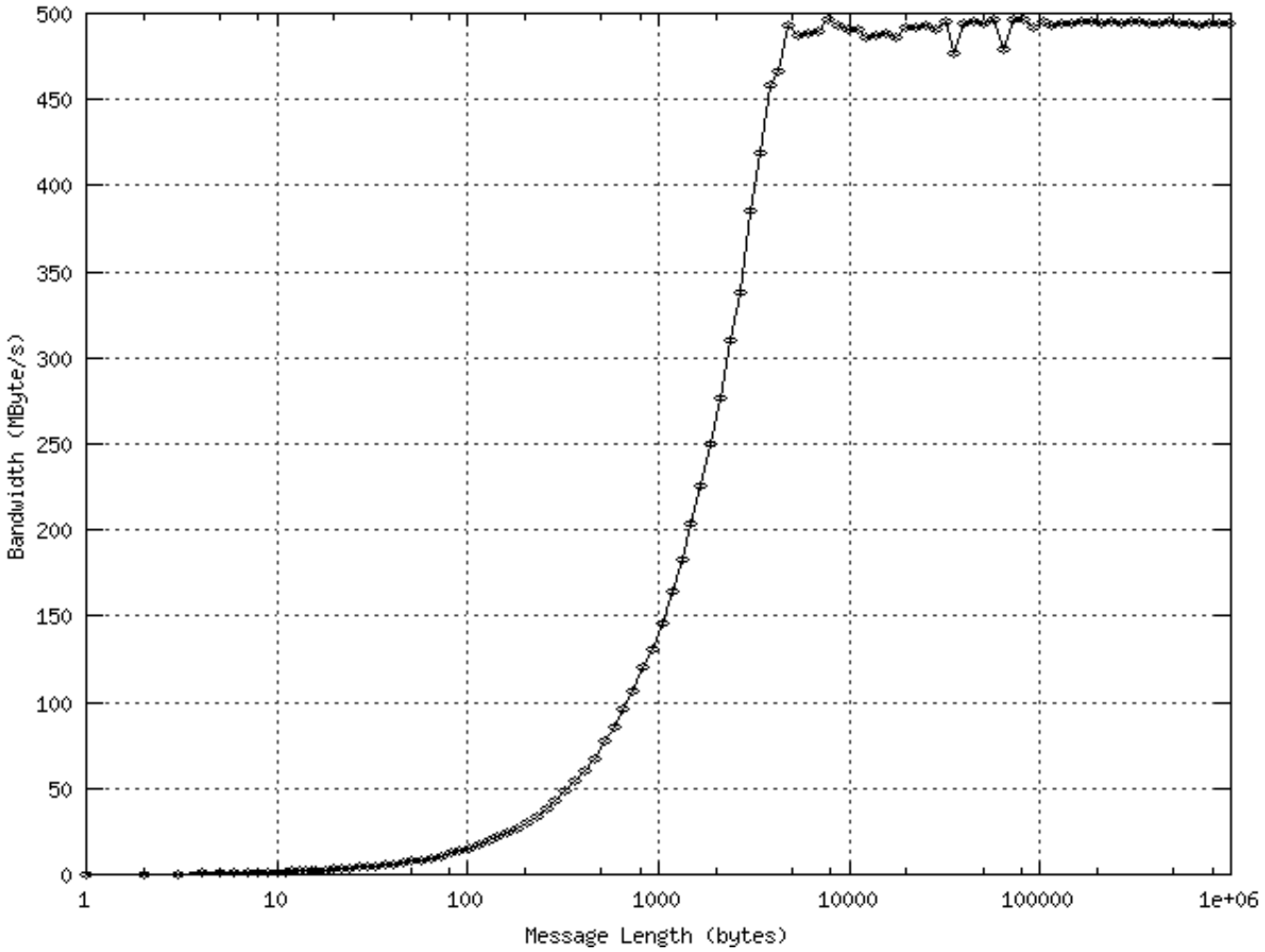
- Based on gm 2.0.4
- Parallel communication to enable high throughput
- Support heterogeneous interconnection
- Support VIA and MPI-2 Interfaces
- Special support for inter-cluster communication to enable Grid computing
- Special support for high performance socket

BCL-4 Latency



1.8GHz AMD Opteron
Myricom M3-E32/M3F-PCIXD-2

BCL-4 Bandwidth



1.8GHz AMD Opteron
Myricom M3-E32/M3F-PCIXD-2

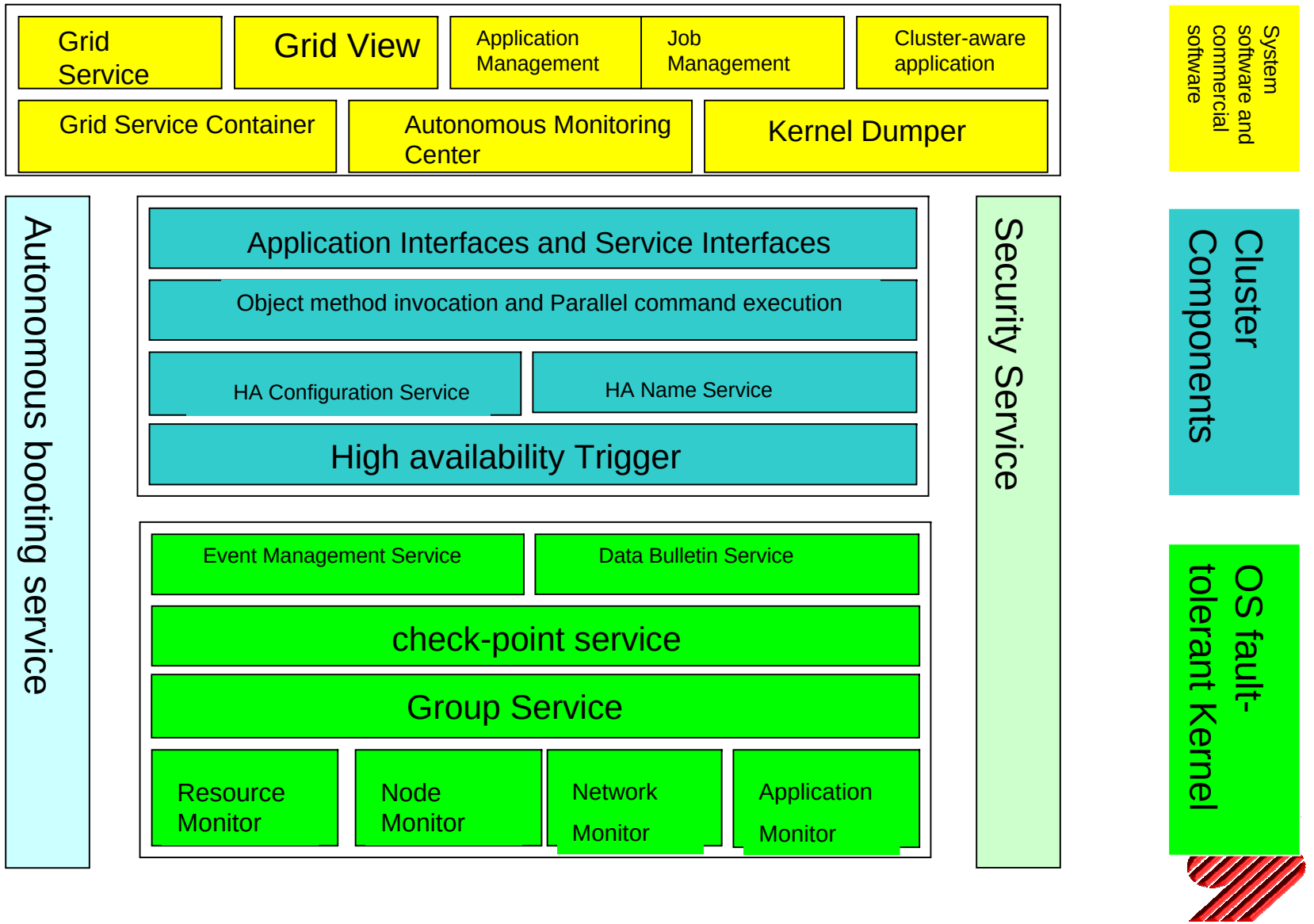
DCFS

- Shared global file system
- Standard system call interface
- Multiple storage servers
- Multiple metadata servers
- High scalability
- High aggregate I/O bandwidth
- High metadata performance
- Easy management

Clustone

- Computation Stack
  - VIA/GM/BCL
  - MPI/PVM/OpenMP
  - C/Fortran
  - Debugger
- Commercial Computing
  - Load Dispatcher
  - Cluster Web
  - HACMP/Wolfpack
  - Oracle RAC
- Kernel of ClusterOS
  - Component Framework
  - Detector
  - DataBase Service
  - Checkpoint Service
  - HA Service
  - Membership
  - ClusterORB

Cluster OS Hierarchy



Massively Cluster Computer

Features

Cluster

- ❑ Following Moore's Law
- ❑ Volume COTS
- ❑ Easily Porting
- ❑ Desktop Image
- ❑ Switch Interconnection

MPP

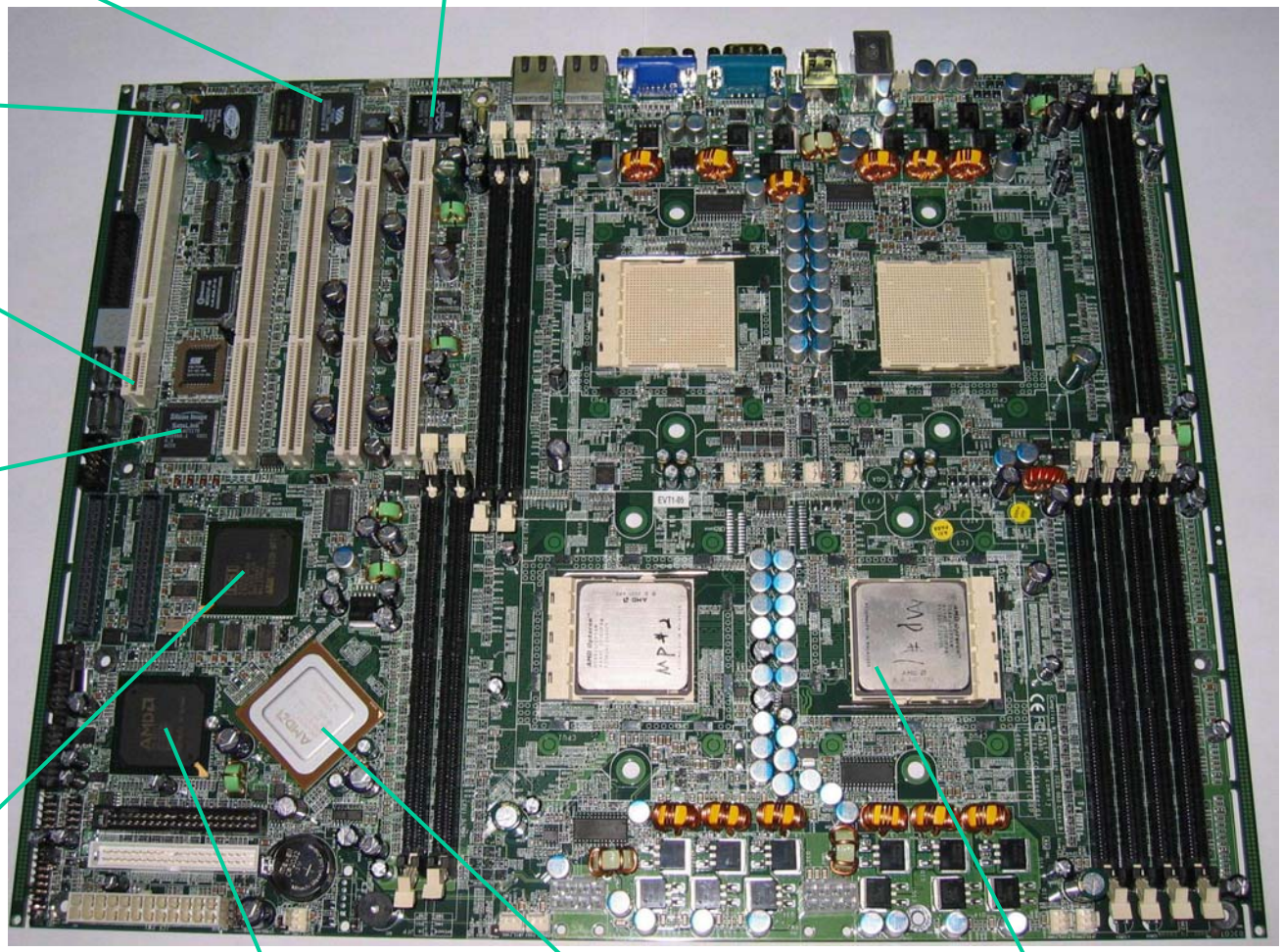
- ❑ Petaflops Scalability
- ❑ Sustained Performance
- ❑ Reliability
- ❑ Easily Management
- ❑ Single system image

Achieve to PetaFlops

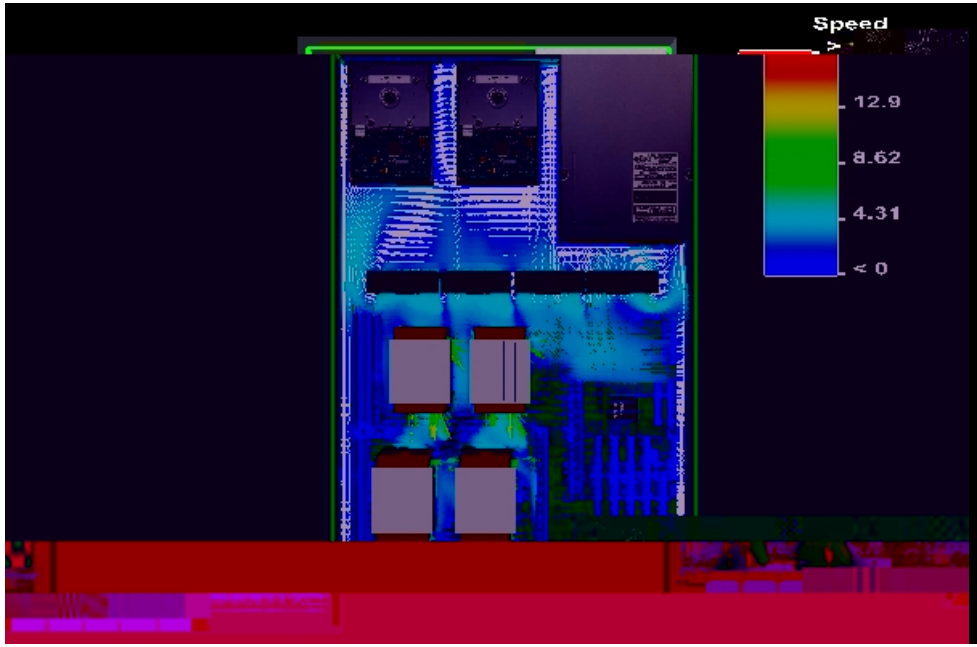
- 4-Years Gap
  - 2004: 10Tflops (Dawning4000A)
  - 2000: 12.8Tflops (IBM ASCI White)
- 5-Times Gap
  - Dawning4000: up to 20Tflops (Industry Standard Cluster)
  - IBM ASCI Purple: 100Tflops (Specific Technology)
- Initiative Effort

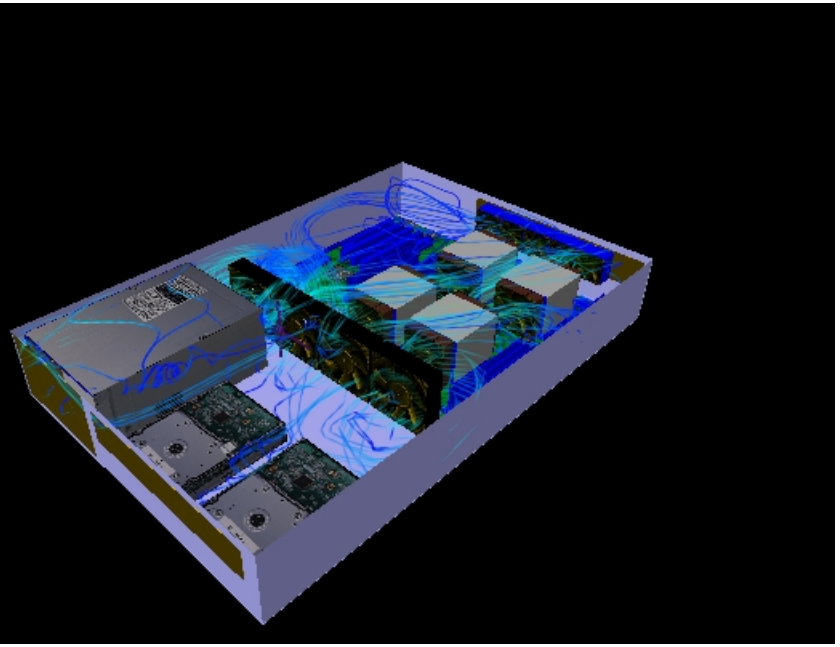
Motherboard

- Designed by Dawning
- Manufactured by Tyan
- Heat Dissipation Simulation by AVC
- SWTX footprint (13”x16”)
- Dedicated Management Slot
- Integrated KVM Switch
- 4P2U High-density Node

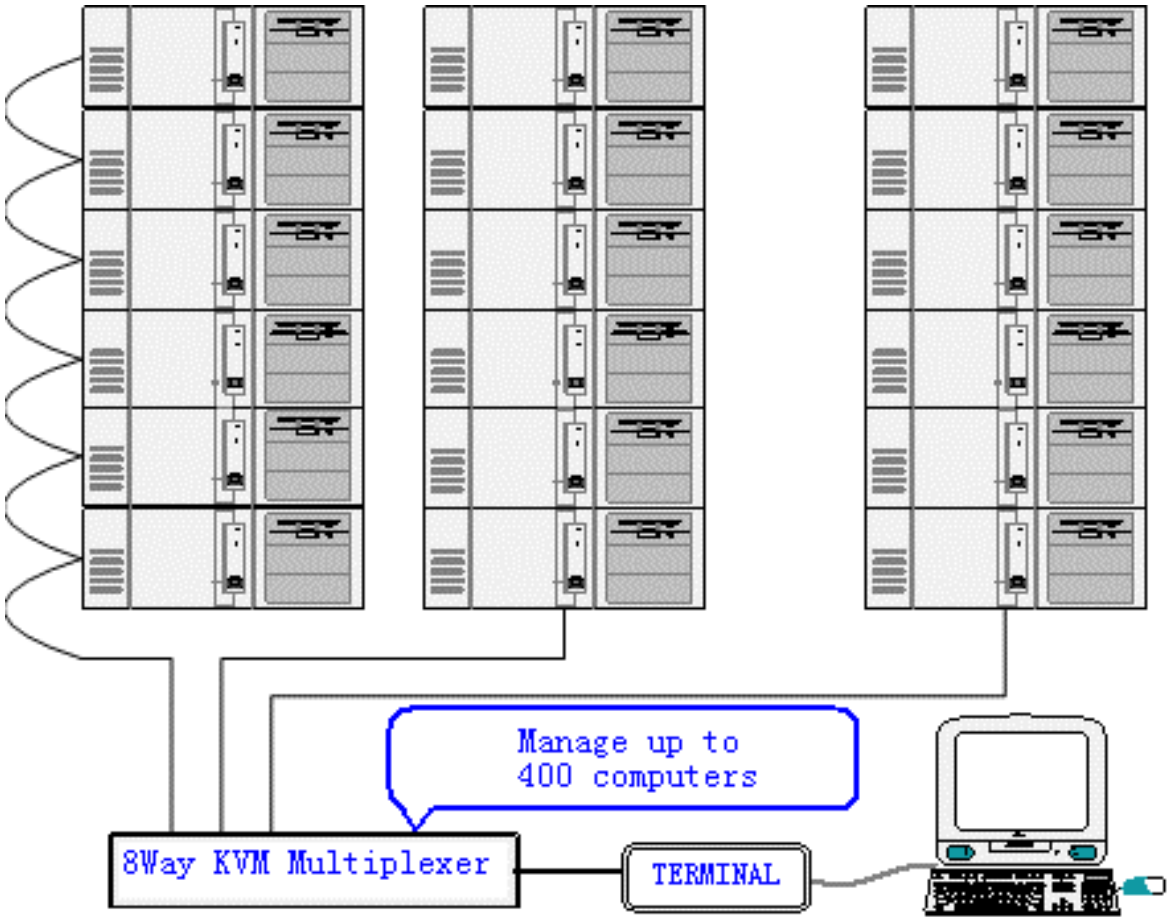

VIA USB2.0
Brocom Dual GE
ATI VGA
Dawning Slot
Silicon SATA
LSI Dual U3 SCSI
AMD 8111
AMD 8131
AMD Opteron

2U4P

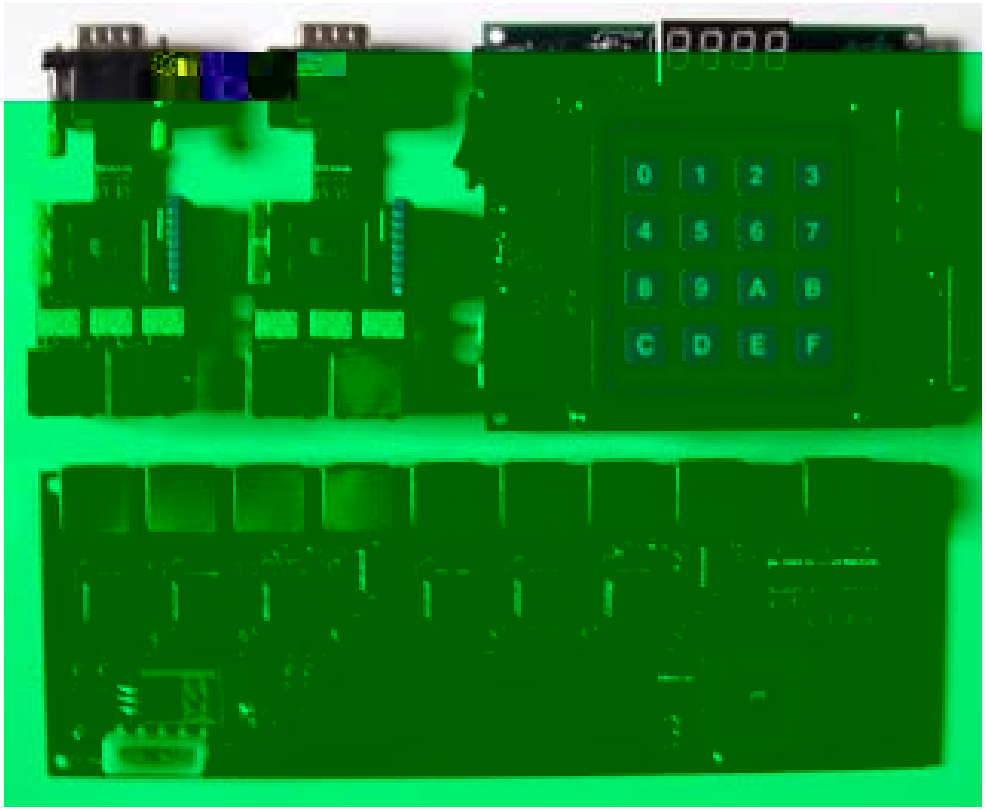




Standalone Solution



KVM Switch



MCC LWK

- Target: improve the saturated performance of computing-intensive application
- Light Weight Kernel Linux-32
  - Hardware CPU Memory KVM Ethernet SCSI320
  - Function System call Multi-process Virtual Memory Ext2 TCP/IP NFS
- Optimization
  - Multi-User/Single-Job
  - Static Schedule of Process gang NUMA
  - Simplified MMU NUMA Memory Mapping
  - No Swap, Dynamic Library, Device Management
  - Cut out: OS Services, Daemons
  - Customized High Performance Communication
  - MPI I/O of Cluster File System

MCC Loader

- Target: Dynamically loading multiple OSs from remote image servers for large scale cluster
- Key Techniques
  - Net Booting
  - Environment Building
  - BIOS Setup
  - Multicasting
  - Image Management

Spreader

- Target Management of large scale files
- Features
  - Database of metadata
  - Logical View
  - Parallel file operations
  - Dynamic update

AutoAdministrator

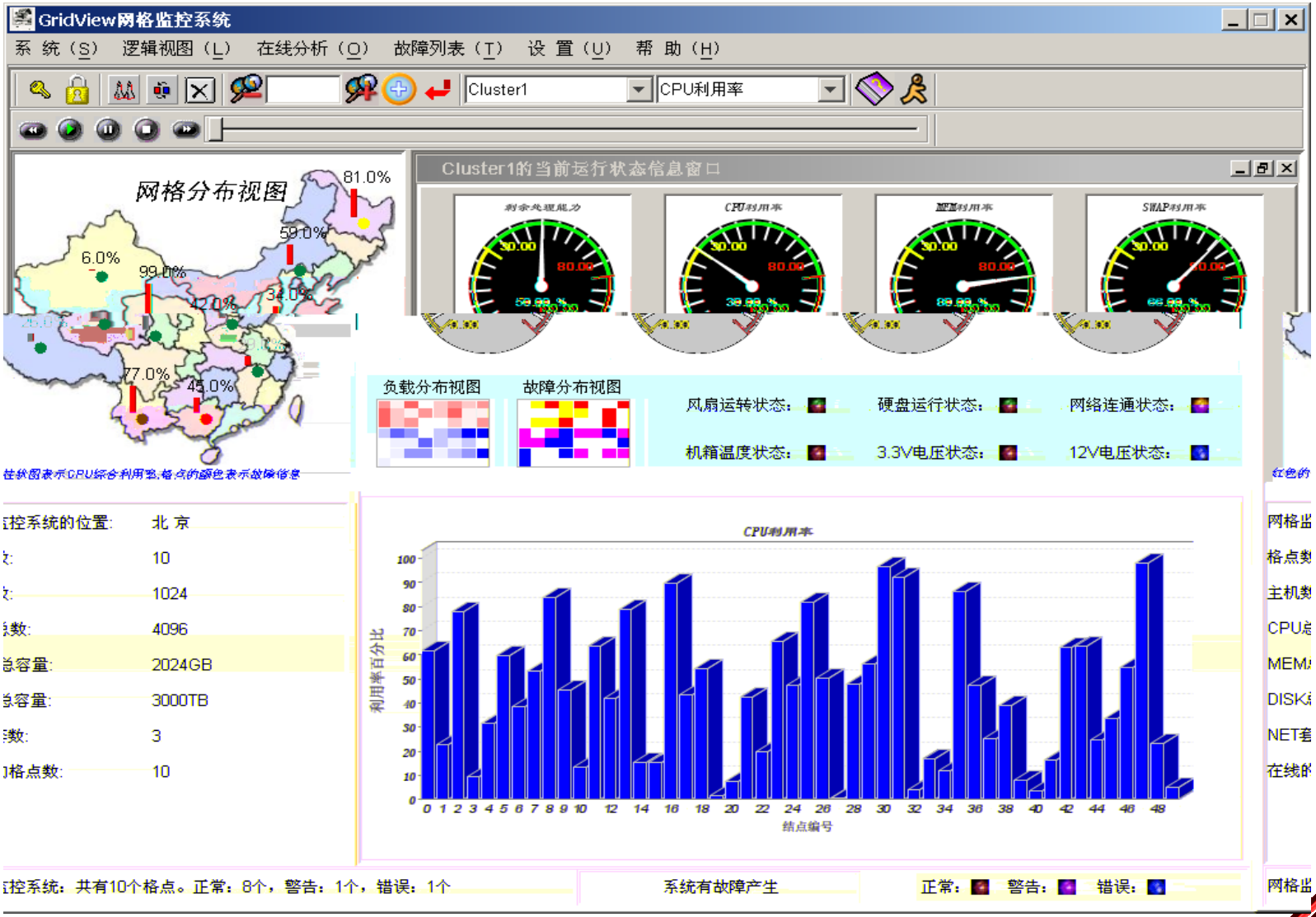
- Target: Improve the availability of system
- Features
  - Memory Leak
  - System Overload Protection
  - Cluster Software Aging
  - Node Crash
- Technique
  - Failure Check
  - Dynamic Prediction
  - Recovery
  - Intelligent Administration Policy


grid

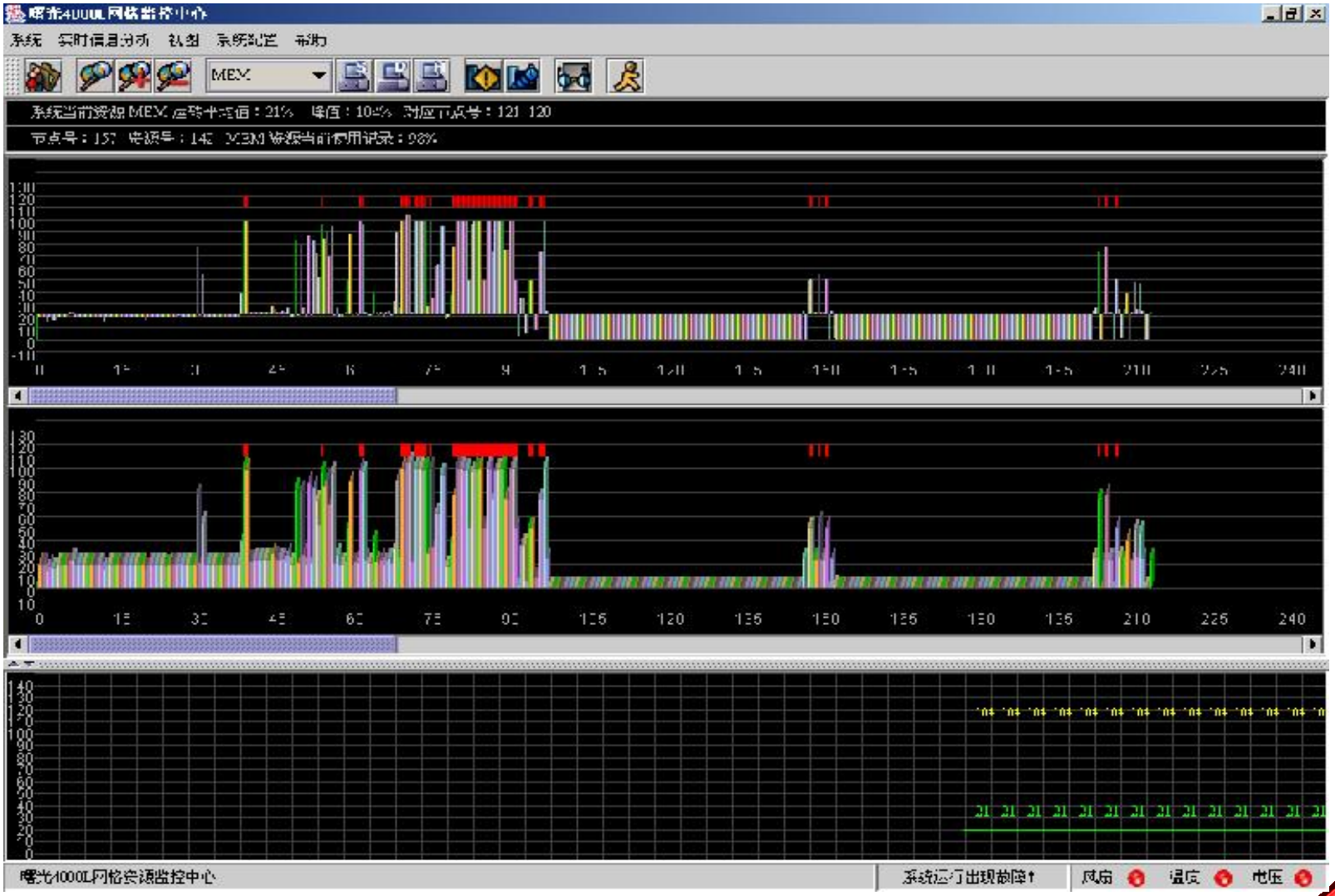
Grid-enabling Components

- GridView
- Grid Router
- Grid Key
- Grid Gateway
- Grid I/O
- GridOS

Grid View


GridView网格监控系统
系 统（S） 逻辑视图（L） 在线分析（O） 故障列表（T） 设 置（U） 帮 助（H）
Cluster1
CPU利用率
网格分布视图
81.0%
59.0%
6.0%
99.0%
42.0%
34.0%
77.0%
45.0%
Cluster1的当前运行状态信息窗口
剩余处理能力
CPU利用率
内存利用率
SWAP利用率
负载分布视图
故障分布视图
风扇运转状态：
硬盘运行状态：
网络连通状态：
机箱温度状态：
3.3V电压状态：
12V电压状态：
北 京
10
1024
4096
2024GB
3000TB
3
10
CPU利用率
利用率百分比
结点编号
系统有故障产生
正常：
警告：
错误：

Snapshot



Grid Router

- Router between System and Grid Environment
- 2Gbps Smart Adapter
- Remote Management
- Access Internet Resource
  - Parallel NAT
- System Interconnection in Grid
  - Parallel VPN
- Access Internal Service
  - Load Balancing
  - Full-life Authentication
  - Parallel Firewall
  - Parallel Ftp, Telnet, X-win Proxy
- High Availability

Grid Key

- A USB Security Device on Client
- Features
  - Access resources in Grid
  - Authentication
  - Service Setup
  - Digital Signature
  - Account

Grid I/O

- Grid ftp
- Grid MPI-IO
- VegaFS

Vega Grid

GOS Version 1.0 (Alpha)

Linux (~ 44M) Win32 (~ 43M)

User Guide API Manual

http://vega.ict.ac.cn

THANKS