

Cluster and Grid Superservers: The Dawning Experiences in China

Zhiwei Xu, Ninghui Sun, Dan Meng, Wei Li

Institute of Computing Technology

Chinese Academy of Sciences

P.O. Box 2704

Beijing 100080, China

{zxu, snh, md}@ncic.ac.cn, liwei@ict.ac.cn

China I T Market

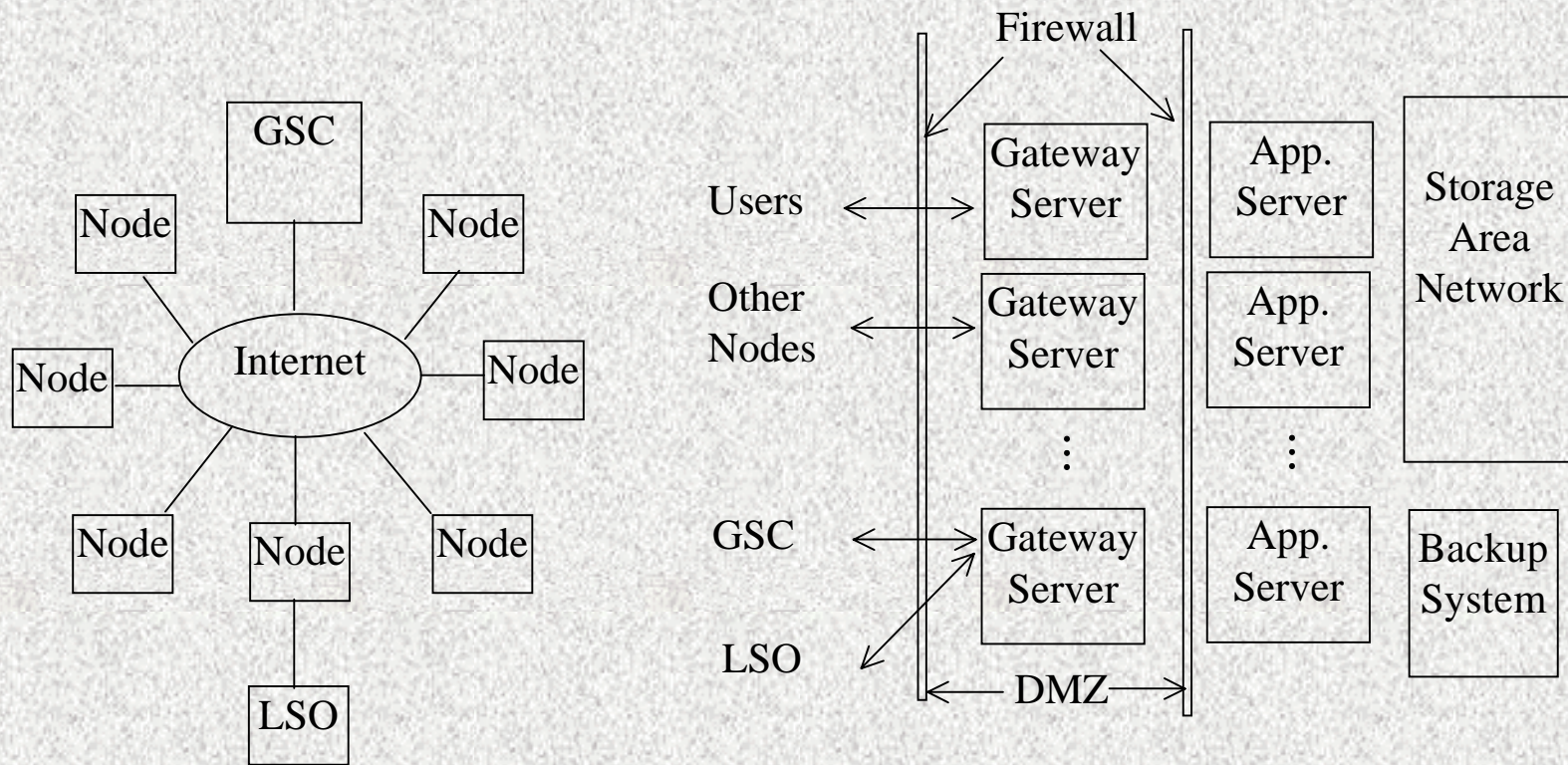


China is an expanding market for IT and high-performance computers:

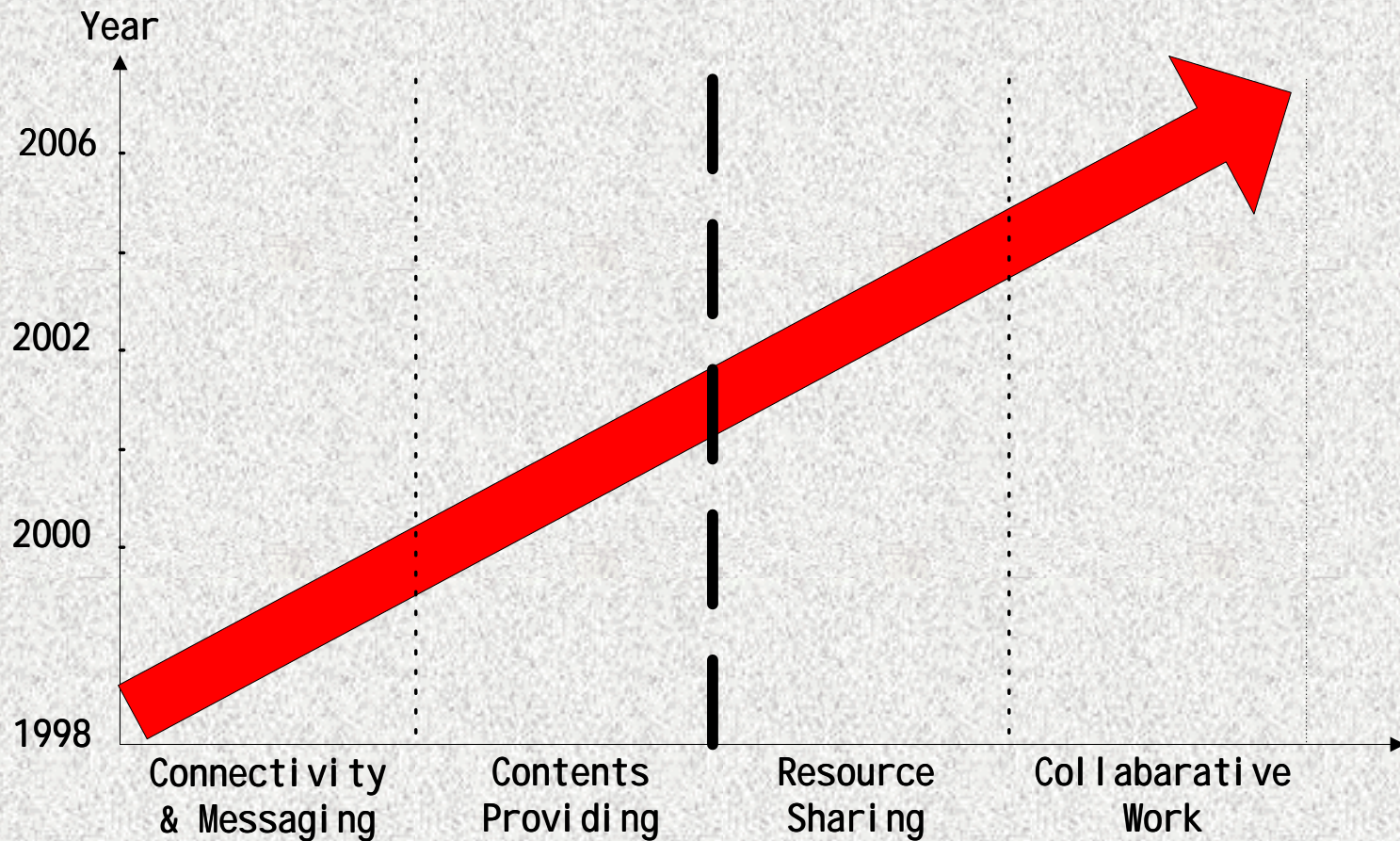
- Double-digit growth rates observed for several years
- Still a small percentage of the world market

Year	1997	1998	1999	2000
PCs Sold	3.50 million	4.08 million	4.94 million	7.17 million
Computers Connected to the Internet	0.299 million	0.747 million	3.50 million	8.92 million
Internet Users	0.062 million	2.10 million	8.90 million	22.5 million
“www” Web sites	1,500	5,300	15,153	265,405
Mobil Phones Sold			43.29 million	85.26 million
Telephones Sold			163 million	260 million

A Country-Wide ASP Platform in China



E-Government Project of Beijing City



Introduction to ICT



Institute of Computing Technology (ICT)

- Established in 1956 as the first computing institute in China
- 50 professors, 150 research associates, 360 graduate students
- Research areas
 - CPU micro-architecture (Godson microprocessor)
 - High-performance computers (Dawning Superservers and Vega Grid)
 - Internet devices, computer network
 - Operating systems, compilers (Intel IA-64/Linux, open source)
 - Middleware, application software
 - Information security
 - Intelligent information processing

Superserver Trends



☞ Application Trend:

High-performance computers are increasingly used for applications other than technical computing

☞ Value Trend:

The performance concept will be augmented to total performance of ownership (TPO), while the cost concept will be augmented to total cost of ownership (TCO)

☞ Architecture Trend:

The I/O subsystem will become a focus point for the high-performance computer research

☞ Networking Trend:

A trend towards a pervasive/grid architecture, where the client side consists of many ubiquitous Internet devices, while the server side consists of all servers on the Internet organized as a single logical grid

Dawning High-Performance Computers

Name Year	Architecture	Memory Disk	Gflop/s Peak, MM, Linpack
Dawning 1 (1993)	SMP, Unix, 8 CPUs, Motorola 88100 25 MHz	1 GB, 4 GB	0.06, 0.035, N/A
Dawning 1000 (1995)	MPP, Unix, 36 CPUs, Intel i860 40 MHz	1 GB, 5 GB	2.5, 1.6, 1.2
Dawning 2000-I (1998)	Cluster, AIX, 34 CPUs, PowerPC 300 MHz	8 GB, 152 GB	20, N/A, 5.59
Dawning- 2000-II (1999)	SMP Cluster, AIX, 160 CPUs, PowerPC 333 MHz, Power 3 200 MHz	50 GB, 662 GB	110.4, 46.3, 39
Dawning 3000 (2000)	SMP Cluster, AIX, 280 CPUs, PPC RS-64 400 MHz, Power 3 375 MHz	168 GB, 3630 GB	403.2, 279.6, 233.6

Dawning 3000 Design Principles



☞ The Superserver Principle

☞ The Commodity Principle

☞ The SUMA Principle

- Scalability

- Usability

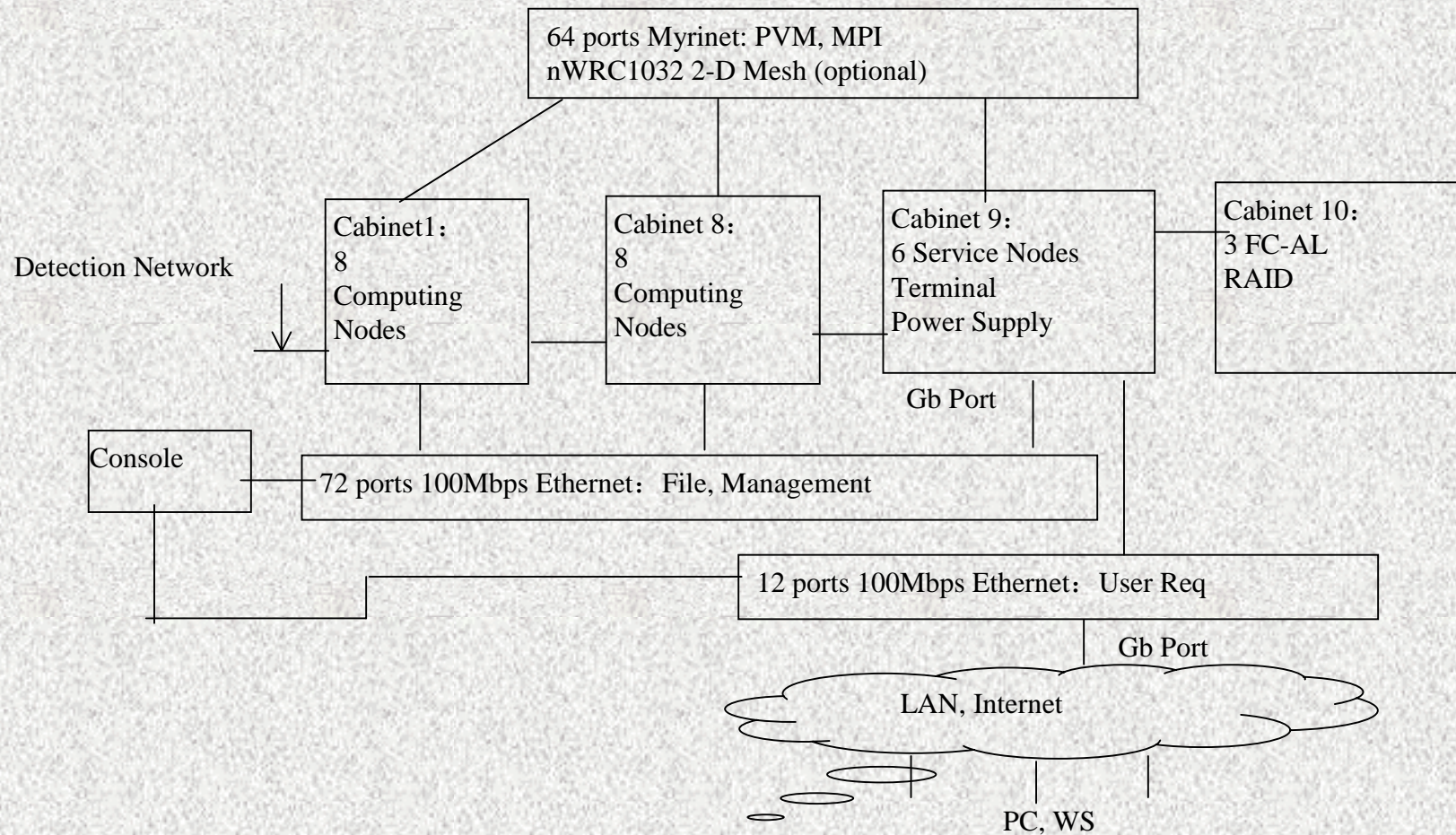
- Manageability

- Availability

Dawning 3000 Architecture

- ☞ Utilizing an SMP cluster architecture
 - 70 nodes with 280 microprocessors, 168 GB memory, 630 GB internal disks, 3 TB external disks
 - 64 computing nodes, 6 I/O service nodes, each node is a 4-CPU SMP system running IBM AIX 4.3.3
 - Interconnected by five networks
- ☞ System area network (SAN)
 - Three Options: Myrinet by Mricom, a NCI C-designed 2-D mesh, or a Gigabit Ethernet
- ☞ Parallel programming software
 - Supports C, Fortran, Java, and open source software such as GNU C, perl, tcl/tk, etc
 - BCL3, ADI -2 (Abstract Device Interface), PVM, MPI, JI AJI A, OpenMP, Autopar, DCDB(Dawning Cluster Debugger)
- ☞ Cluster file system - COSMOS
 - A single image distributed file system
- ☞ Management software
 - RMS, JOSS, CSMS, MONITOR, DSC, SEPS, PowerRouter

Dawning 3000 Architecture



Dawning 3000 Performance



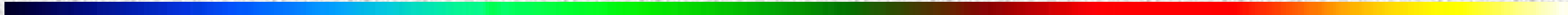
☞ Standard Benchmarks

- Communications benchmarks
- Matrix multiplication
- Linpack
- SPECWeb99
- Lotus NotesBench
- TPC-C
- Andrew file system benchmark.

☞ Real-case Applications

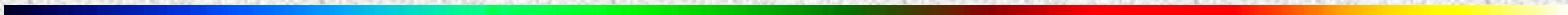
- MM5 short-term forecast
- Oil reservoir simulation application called PRI S
- CPMD benchmark for drug design

Latency and Bandwidth of Point-to-Point Communication on Dawning 3000




	BCL3	MPI	PVM
Intra-node Latency(μ s)	3.6	7.7	7.2
Intra-node Bandwidth(MB/s)	384.6	413.4	362.9
Inter-node Latency(μ s)	24.3	32.3	31.2
Inter-node Bandwidth(MB/s)	130.5	124.9	112.9

MPI Collective Communications Performance on Dawning 3000



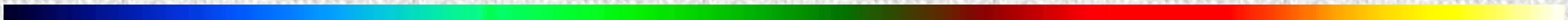
CPUs	4	8	16	32	64	128	256
Barrier(μ s)	32.0	78.2	135.2	184.2	256.6	341.3	764.0
Broadcast(MB/s)	581.2	532.9	623.7	975.5	1549.0	2489.0	4387.0
Reduce(MB/s)	241.9	429.6	586.8	920.6	1425.0	2427.0	4159.0
Scatter(MB/s)	307.9	159.7	137.9	121.1	117.1	110.5	103.1
Gather(MB/s)	153.2	87.1	90.9	83.8	108.8	113.9	114.1
Shift(MB/s)	487.1	489.0	912.1	1816	4017	6814.0	14928
All-to-All(MB/s)	463.6	238.1	264.7	442.5	748.2	748.7	1585.0

Performance of Matrix Multiplication Benchmark on Dawning 3000



CPU's	4	8	16	32	64	128	256
ProblemSize	10080	16000	23040	32256	46080	64768	109568
Performance(Gflops)	5.4	10.5	20.3	39.7	76.0	146.6	279.6
Utilization	90%	87%	85%	83%	79%	76%	73%

Performance of Linpack Benchmark on Dawning 3000

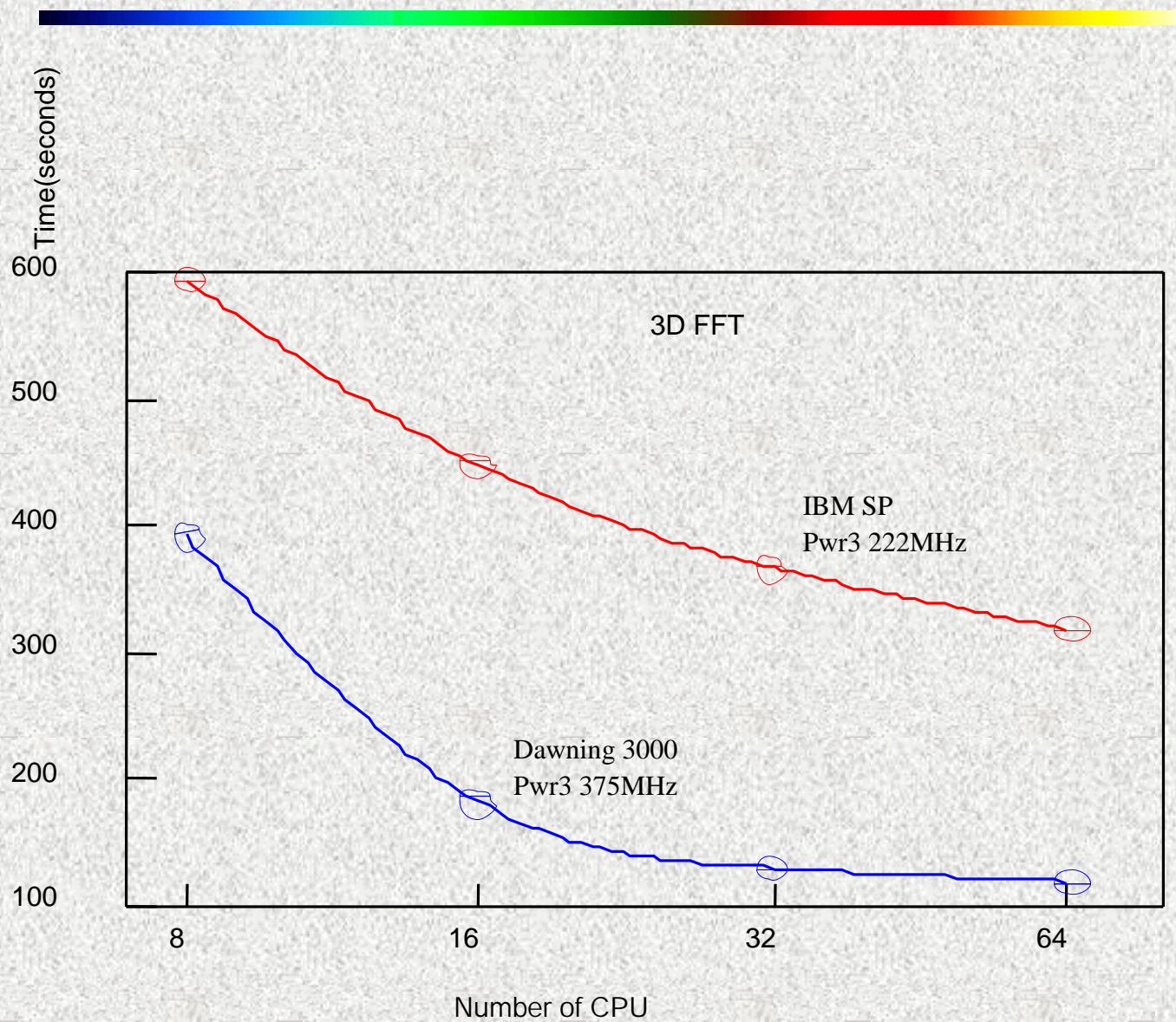


CPU's	4	8	16	32	64	128	256
ProblemSize	20480	25600	36864	51200	71680	112640	163840
Performance(Gflops)	5.2	9.8	18.9	36.5	68.1	128.9	233.6
Utilization	86%	82%	79%	76%	71%	67%	61%

Performance of Weather Forecast Application

Program	Problem Size	CPU	Time(sec)	Speedup	SP Time(sec)
MM5	12 hours	16	4485	1	5278
		32	2605	1.72	
		40	2393	1.87	2803
		64	1616	2.77	
		128	1221	3.67	
	48 hours	128	5902	—	
T213	6 hours	8	920	1	
		16	471	1.95	720
		32	285	3.22	
		64	175	5.24	360
		128	133	6.89	
	10 days	64	6600	—	

Performance Comparison of CPMD



Grid-Related Research



☞ Design goals:

- Design and implement Grid-level system software (Grid middleware)
- Build a national-scale infrastructure supporting high performance computing
- Develop applications running on the testbed

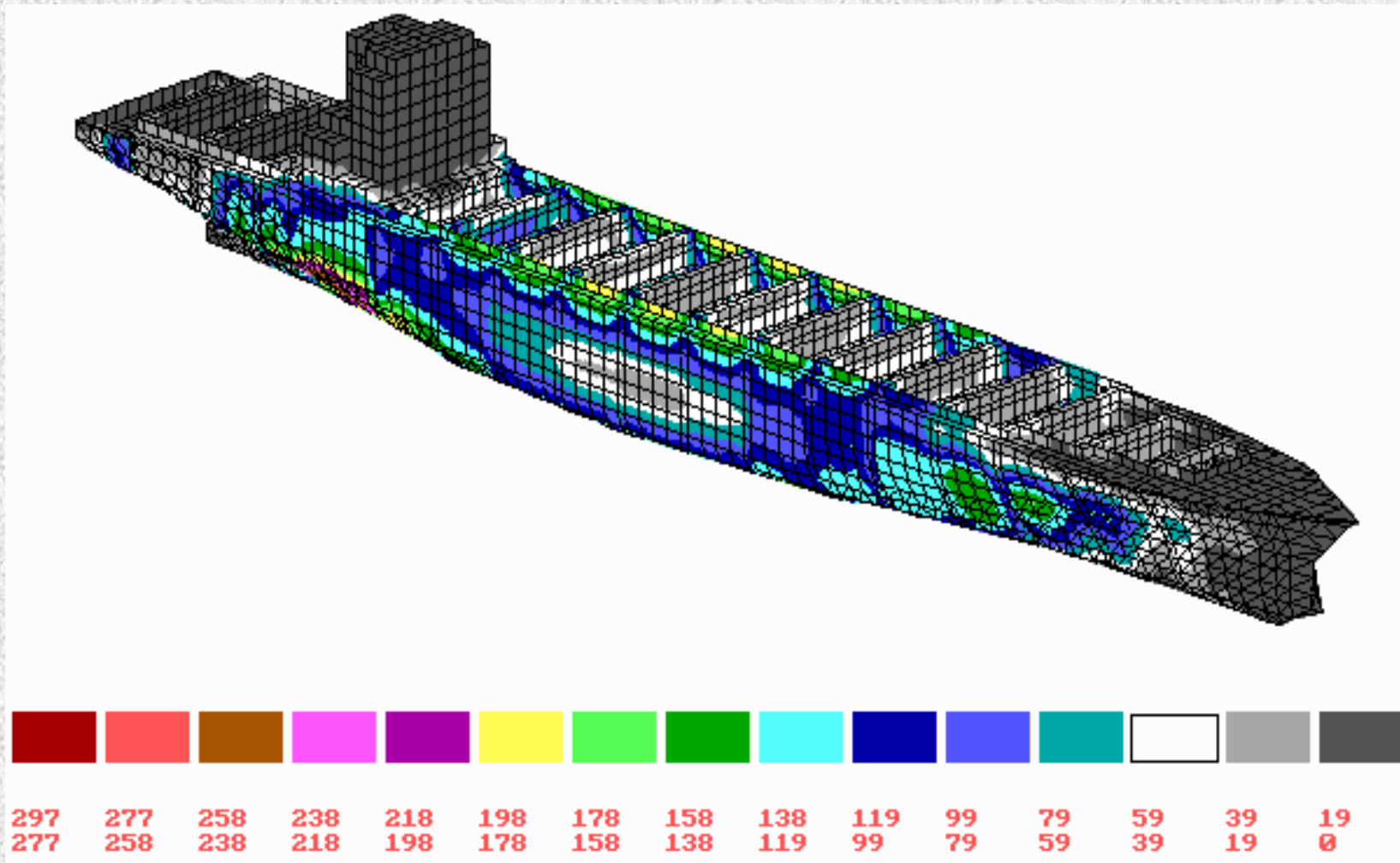
☞ Grid-Oriented Superservers

The next generations of Dawning systems, Dawning 4000 and Dawning 5000, to be completed in 2002 and 2005, respectively, will not only provide higher performance than Dawning 3000, but also offer much more capable support for grid

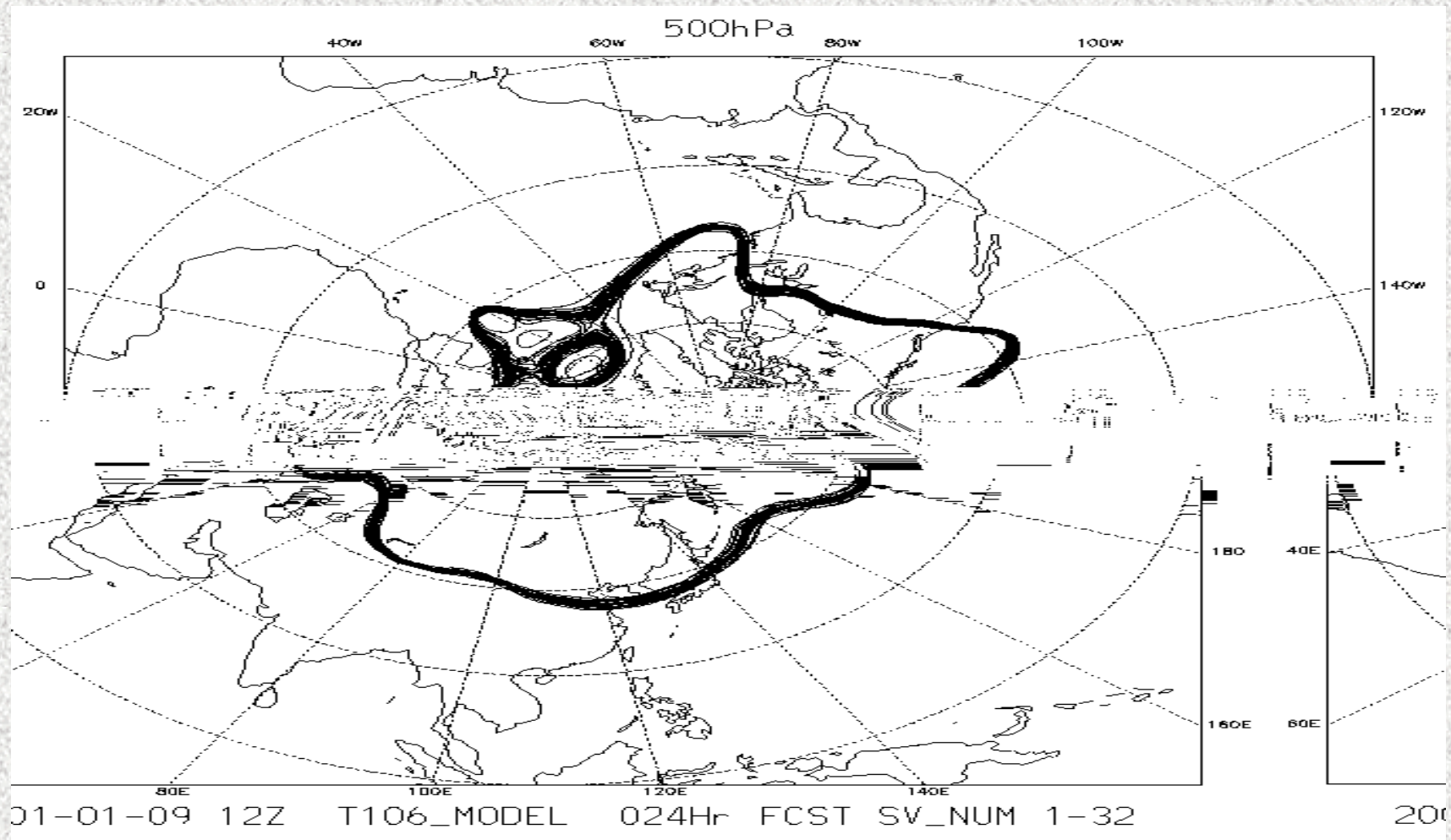
☞ Service Grid

We call our Grid framework a service grid, to emphasize that a grid is itself a superserver, providing various services to the users. The service types should include messaging, computation, contents, transactions, and even knowledge services

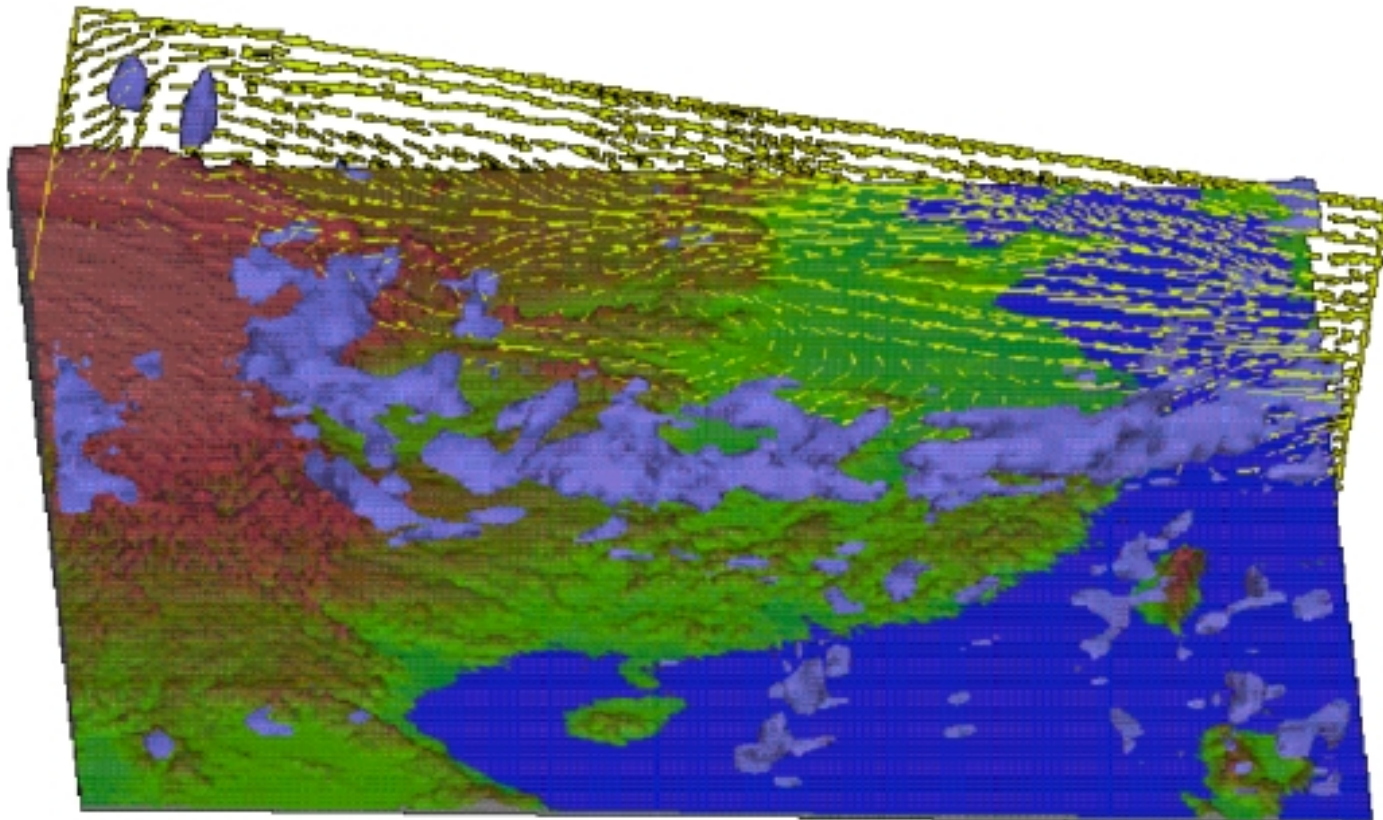
Stress Distribution of 3800TEU



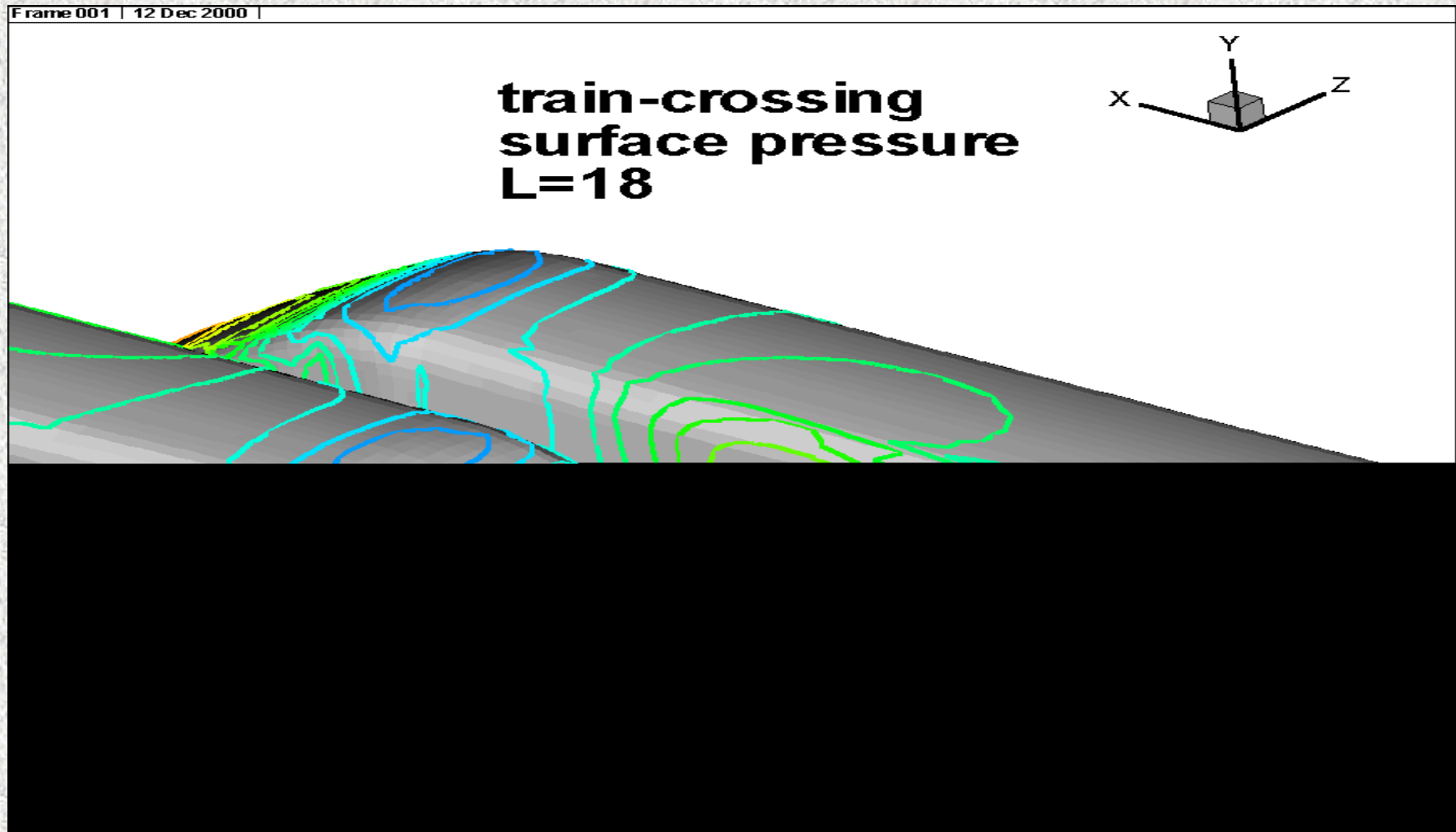
500hPa Contours in Weather Forecasting



High-resolution Mesoscale Numerical Weather Prediction



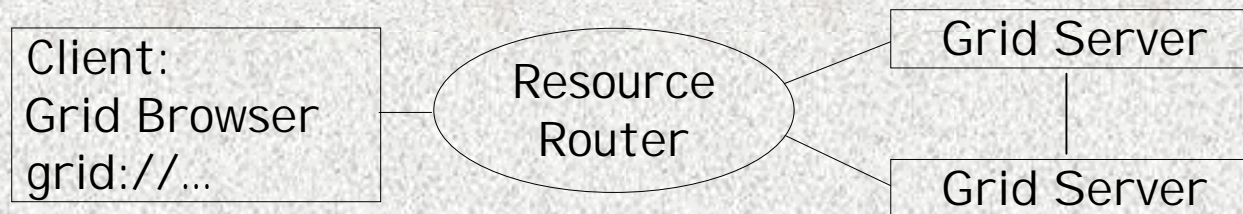
Numerical Wind Tunnel: Train-Crossing Surface Pressure



The Vega Grid Project

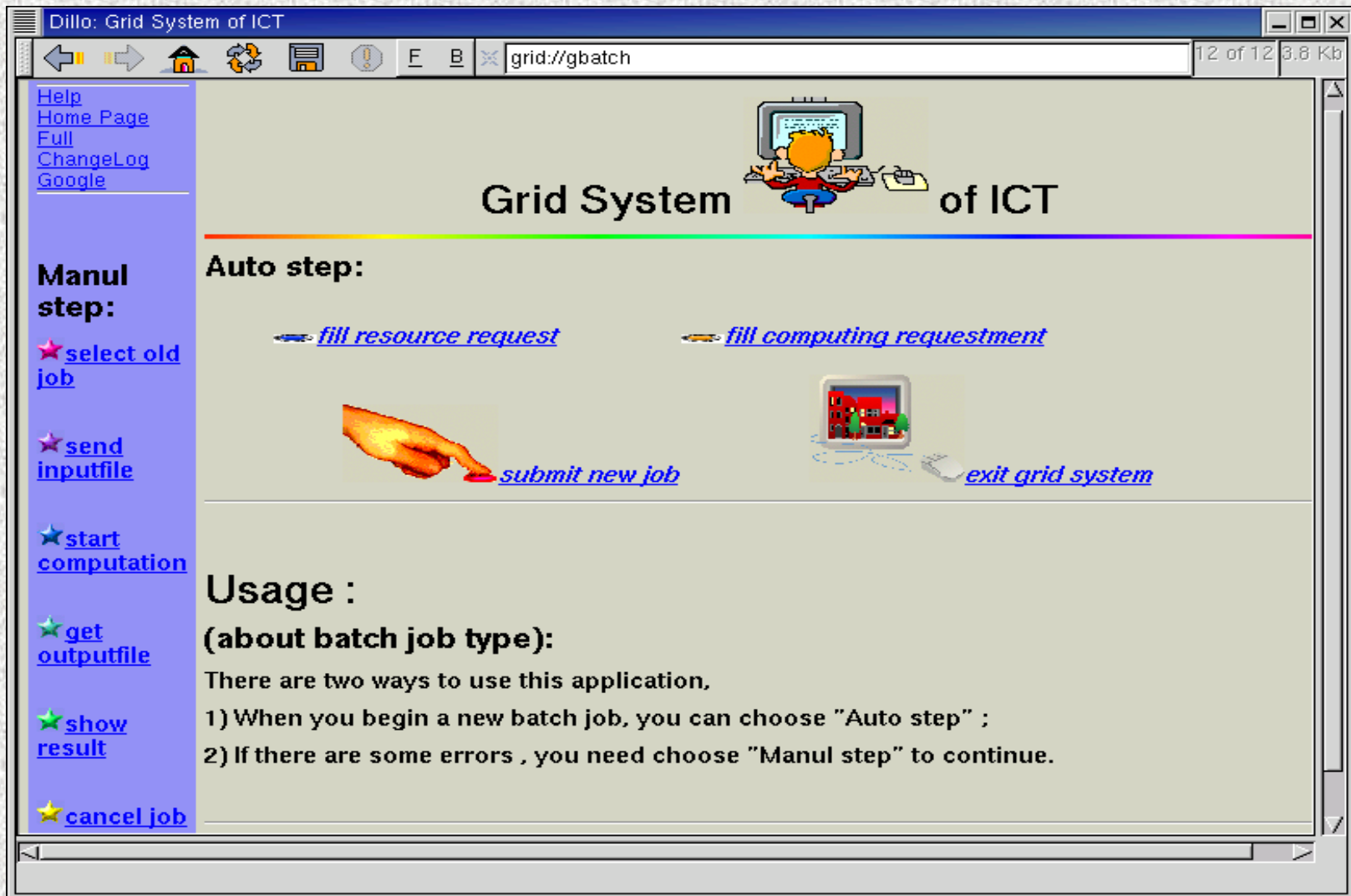
Knowledge Grid	TBD
Information Grid	GPI P
Computation Grid (Vega)	GCP GSI P

Grid Protocol Architecture based on Internet Protocol



Vega Grid Architecture

Global Batch Process System in Vega Grid



Traval - An online transaction example in Vega Grid (1)



Traval - An online transaction example in Vega Grid (2)

