Towards a Petascale Research Grid Infrastructure

Satoshi Matsuoka

Professor, Global Scientific Information and Computing Center

Tokyo Institute of Technology

TOKYO INSTITUTE OF TECHNOLOGY

GLOBAL SCIENTIFIC INFORMATION AND COMPUTING CENTER

GSIC

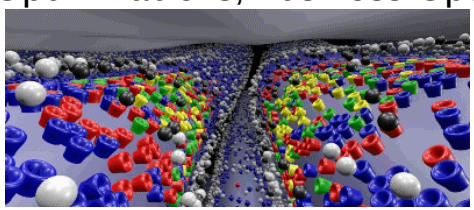
Outline of Talk

1. Introduction to High Performance Grid Computing
2. High Performance Grid Computing with Commodity Clusters
3. High-Performance Parallel Programming on the Grid Using Commodity Clusters
4. Large-Scale International Grid Projects and Collaborative Efforts Based on Clusters
5. Grid Services into a Component Model-the OGSA Story
6. National Research Grid Initiative (NAREGI)
7. The Japanese Large-scale Grid-Cluster Infrastructure of March 2004
8. DataGrid and the Grid DataFarm Project -Scaling to PetaBytes-
9. Scaling to the Millions - The MegaScale Federation of Low Power, Very Dense, High Performance Clusters

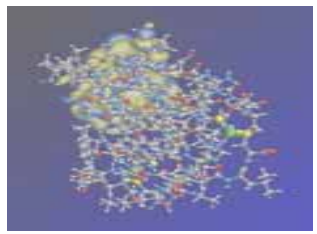
1. Introduction to High-Performance Grid Computing

New Computing Challenges Tera~PetaFlops, Tera~PetaBytes, Giga~TeraBps E-Science Apps

-High Energy Physics (e.g. LHC), QCD
-Astronomy, Planetary Science, Earth Science (Petascale Data Analysis)
-Genome Informatics, Protein-level Analysis, Personalized Medicine
-Bioorganism Simulation, Brain Simulation
-Nanotechnology, e.g., Molecular Computing
-Earthquake Simulation, Disaster Simulation
-Weather Simulation, Multi-physics Fluid Dyanmcs
-Operations Research, Engineering Design Optimizations, Business Optimizations



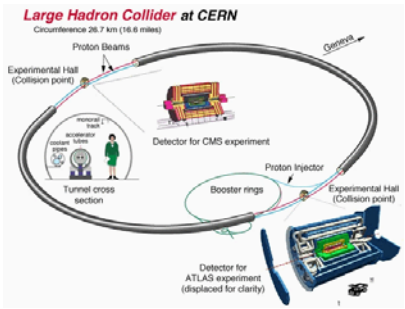
MCell Brain Cell Simulation



Genome and Protein Informatics

VLBI: Virtual Observatory

3D Eartquake Simulator in MIKI



CERN LHC and Patascale Data Analysis

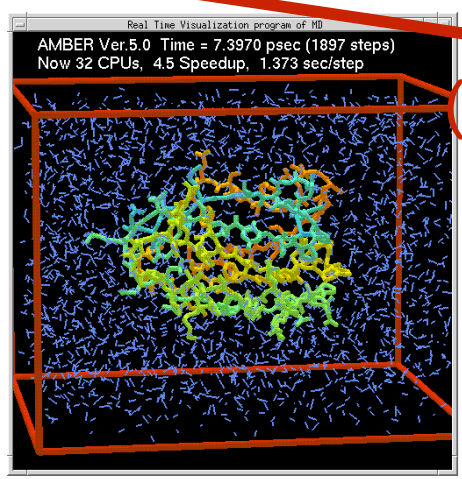
Trad. Supercomputing DOESN'T SCALE to such needs

-Problems "unfit" for Vector parallelism
-Unrealistic Cost Performance
-Collaborations amongst multiple sites and people

The limitations of physic and economy in computing digresses competitiveness of Japan (or, US, EU, or whatever hat you may wear…)

Example: Ab-initio Molecular Dynamics Protein Folding Simulation(AIST CBRC MolTreC)

# of CPUs	Length of Protein (Basis)	
	Small ~100	Standard(~400
1	10 years	160 years
250 (now)	2 weeks	8 months
4000 (large now)	21 hours	2 weeks
10,000 (future)	1 hour	½ day

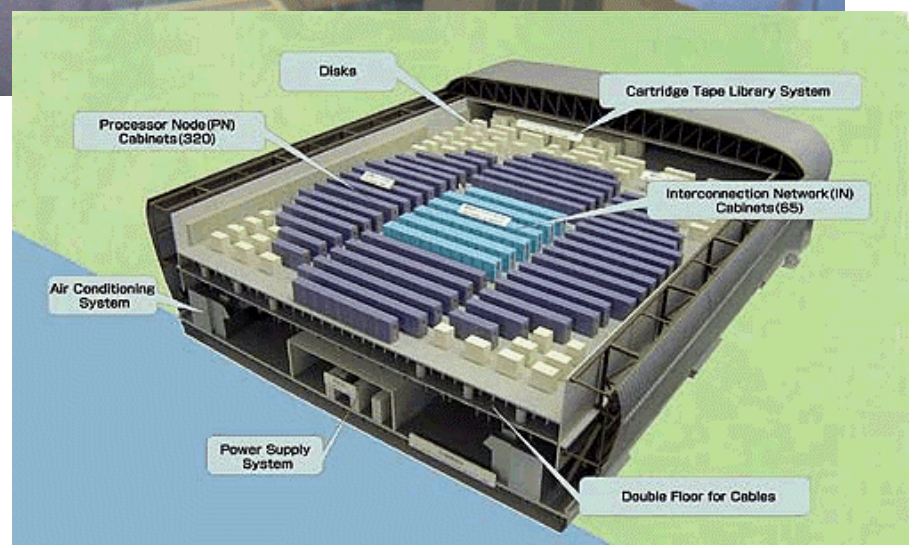


Petaflops
(1000 trillion FP calculations/s)

Note: acceleration includes algorithmic improvements

Extreme Supercomputing (The Big Iron)

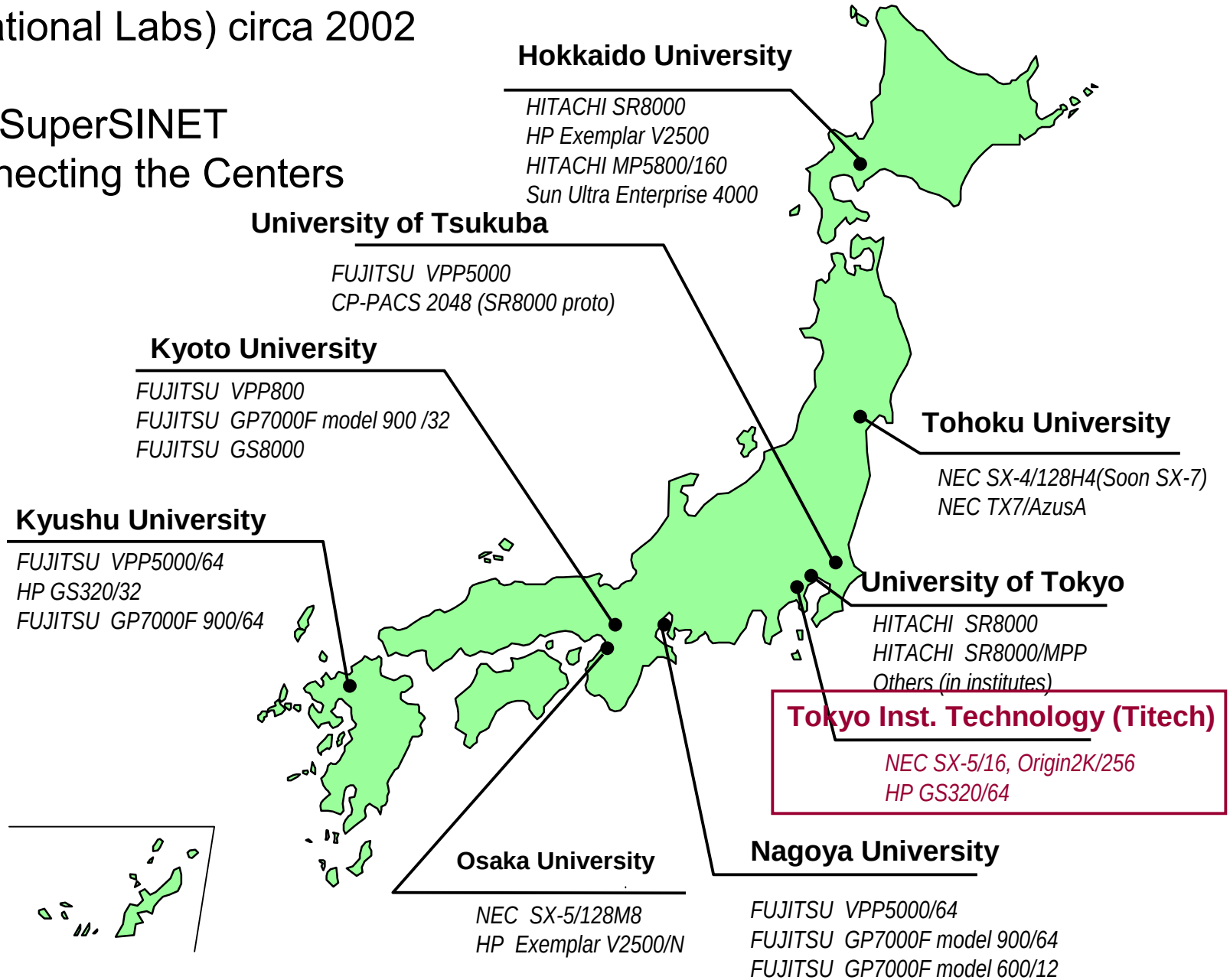
- Japanese Earth Simulator (2002)
- Target Application: CFD-Weather, Climate, etc.
- 640 NEC SX/6 (mod)
  - 5120 Vector CPUs
- 40TeraFlops (peak), 35TeraFlops (Linpack)
- #1 on Top500 (by far)
- $400 million
  - $20-$30mil/y maintenance
- Size of a large concert hall
- Expect to be on top of Top500 until 60-100TeraFlop ASCI machines arrive in 2004-5



(Earth Simulator Picture from JAERI web page)

Inter-university Computer Centers (excl. National Labs) circa 2002

10Gbps SuperSINET Interconnecting the Centers



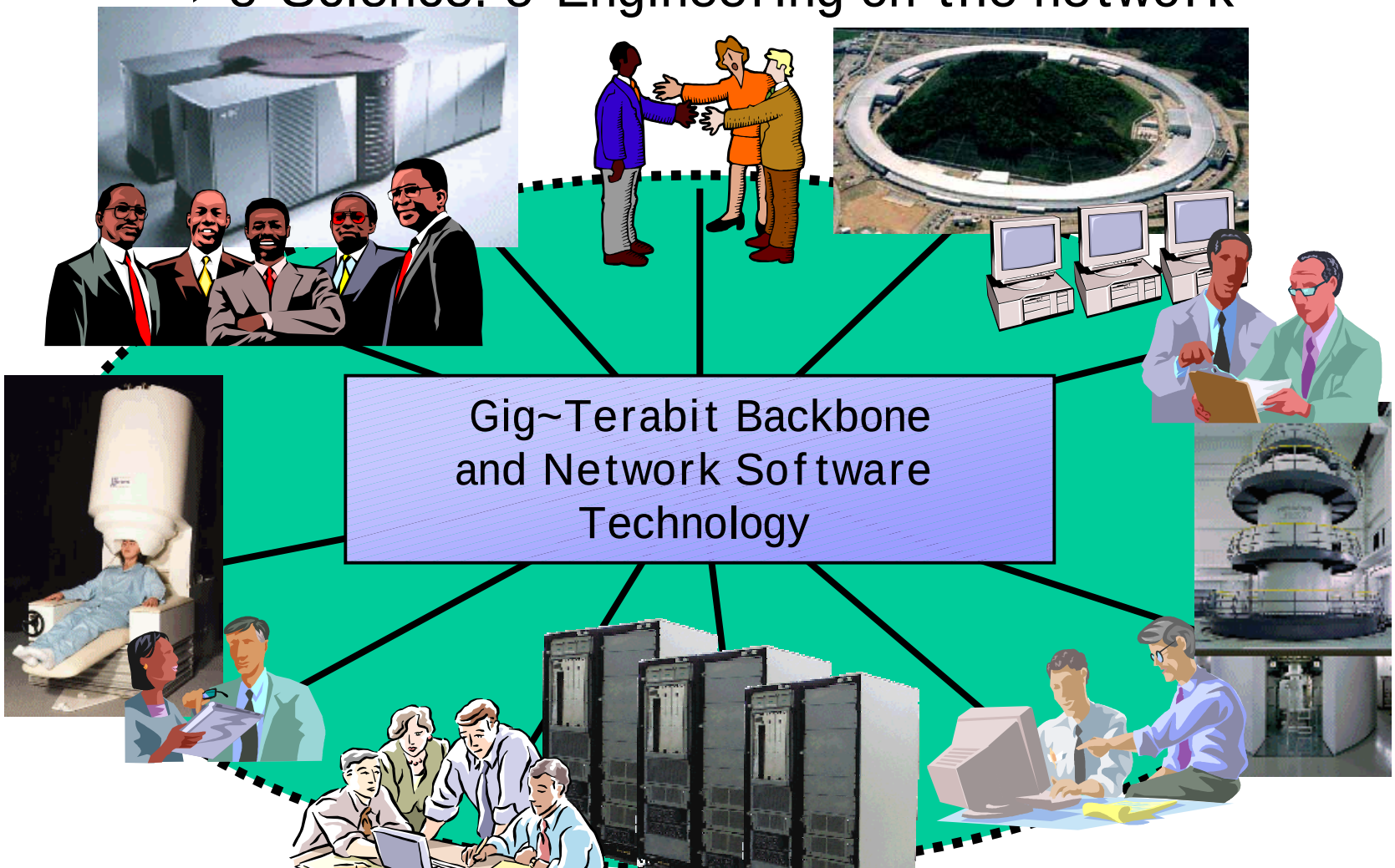
Moreover, Not Everone Runs Big Jobs

- Example: Titech GSIC operates 3 supercomputers
  - ~1000 users
  - 16 proc SX-5, 256 proc Origin2K, 64 proc AlphaServer GS320, 400GFlops total (1/100 of ES)
  - All heavily utilized (99% for SX-5) – all filled up
  - Annual *rental* budget: $6 million (~$35 million / 6 years total)
  - 3 years left on 6-year rental contract
  - All machines disappeared from Top500 (Nov, 2002)

- Not everyone runs big jobs---but gets filled up!
  - 4-8 proc jobs (e.g. Gaussian) consume 50% cycles
  - Smaller % of users run "big" jobs (but sum is large)
  - Machines are needlessly partitioned and used for Gaussian, etc.

- Need to federate thousands of small-to-medium resources
  - Not just big iron

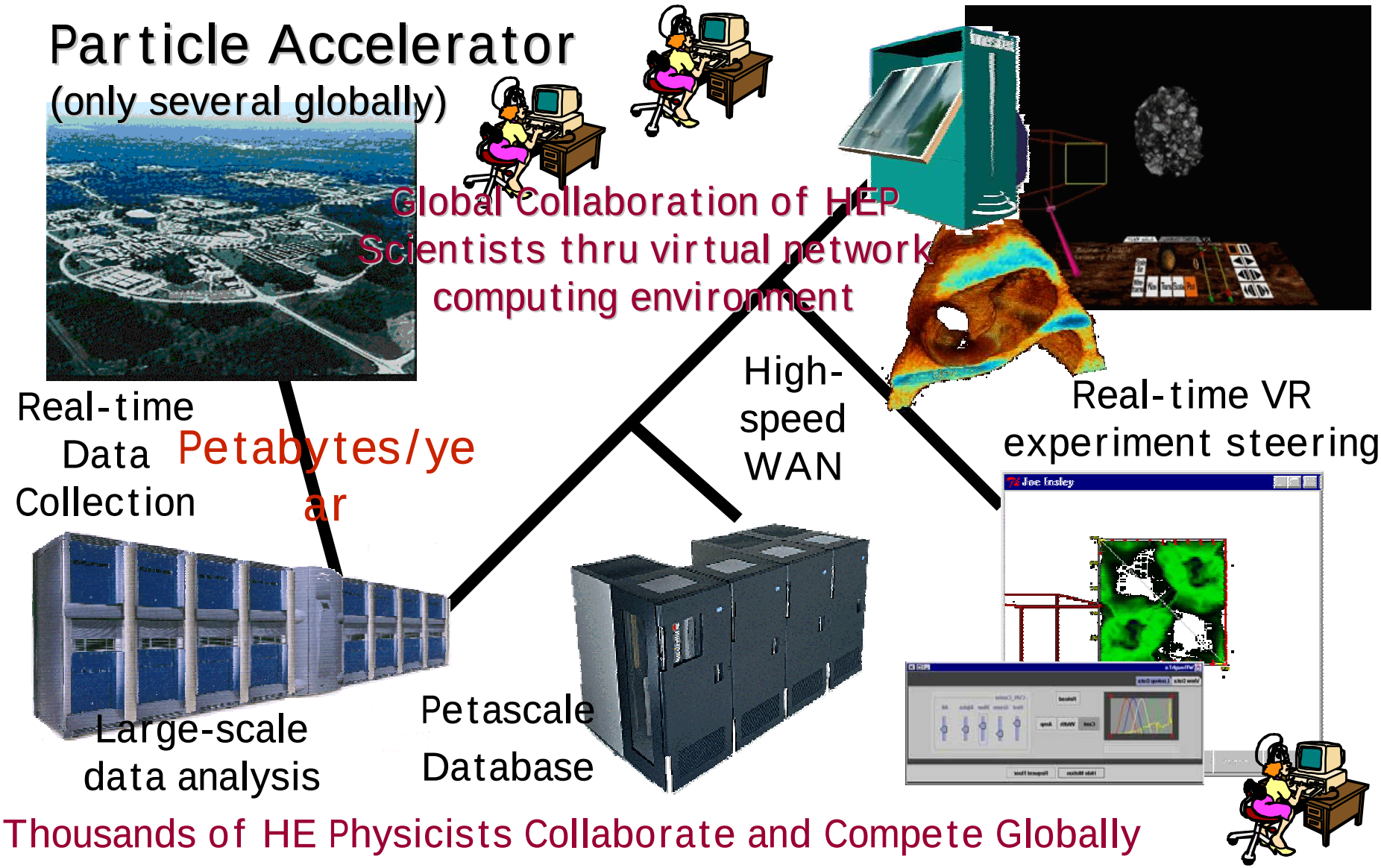
Global Collaboration/Coordination Requirement for Modern-day Science and Technology => e-Science. e-Engineering on the network



Collaboration of Distributed Computing, Storage Sensor as well as Human Resources

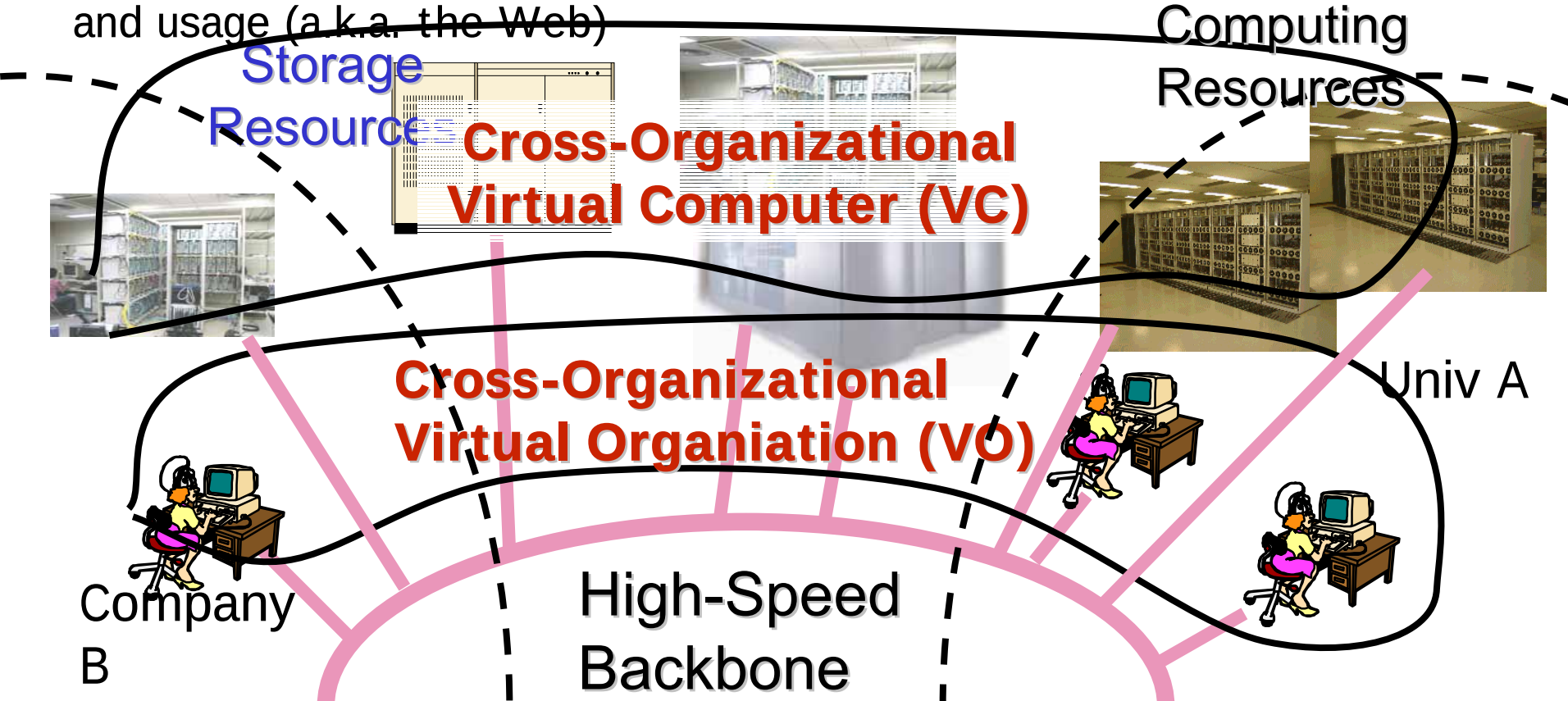
Simple SC Speedup is insufficient

E-Science Global Collaboration Example: High-Energy Particle Physics



Grid: a network infrastructure that allows dynamic creation of virtual computers that crosses administration boundaries

- "Virtual Computer": Hides where the resource is => QoS determined by Grid
- "Cross admin boundaries, dynamic": Scales to Internet, various technical challenges
- "Network Infrastructure": Sharing of resources, standardized access and usage (a.k.a. the Web)



Common Misconceptions on Grids

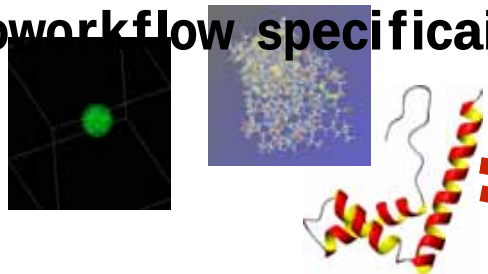
- *"Grid is a Collection of Networked Supercomputers"*
  - => Aggregates a variety of resources including PC, Supercomputers, Storage, Sensors, on a network
- *"A Grid is to employ abundunt idle cycles on a PC like Seti@Home"*
  - => Seti is technically not a Grid since a user merely volunteers his PC without previledge to use the infrastructure as a VC. Does not have to be "idle"
- *"Grid = Web Services"*
  - => Actually Unification via WSRF/OGSA efforts. Web Services are currently stateless whereas a Grid is fundamentally stateful.

Benefits of the Grid

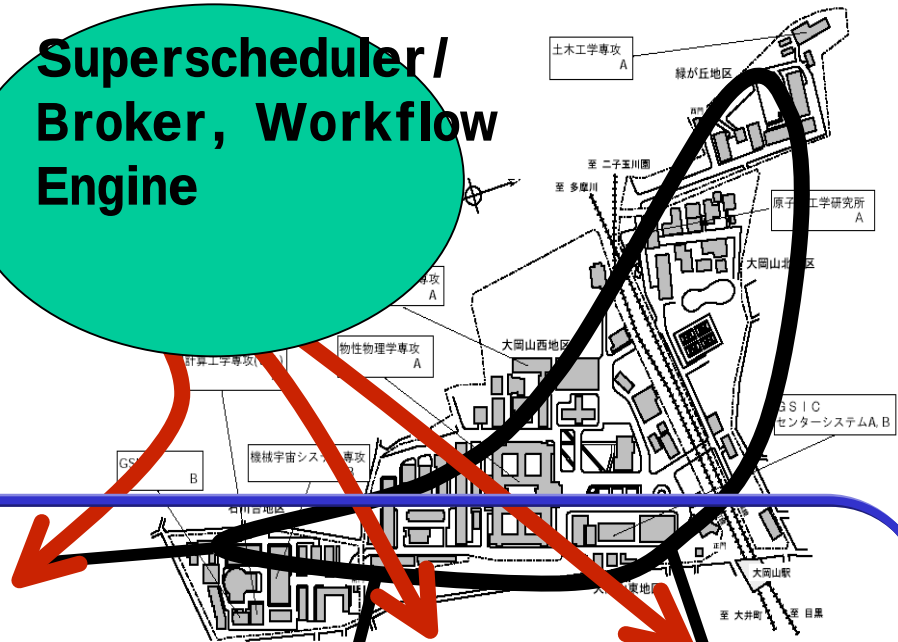
- 1. Higher Utilization of Distributed Resources
  - E.g., Supercomputing Center Grid, Nationwide Virtual Supercomputers
  - No increase in the overall “pie”
- 2. Higher Reliability and Upgradability of compute resources
  - Same objectve as the Internet (or, ARPANet)
- 3. “Collaboratory Science”: tight collaboration of virtual organizations over the network
  - E.g., EU DataGrid w/3000 worldwide high-energy physicists
- 4. Tight Integration of Data, Sensors, Human Resources
  - VLBI (Astronomy) Project
- 5. Ultra-Scaling of Resources
  - Distributed placement of (otherwise oversized) resources
- 6. Exploitation of Idle Cycles/Storage on non-dedicated, commodity resources
  - Peer-to-Peer(P2P), Voluntary Computing

Sever Grid and Metacomputing in HPC Grid

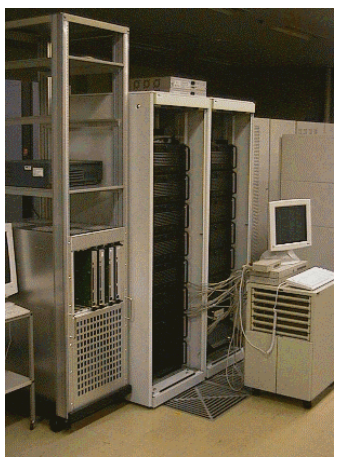
HPC Apps etc., as is or slightly modified, soworkflow specificaiton



Superscheduler / Broker, Workflow Engine



Resource allocation on Server Grid, automatic scheduling of exisiting jobs



Distributed HPC Servers

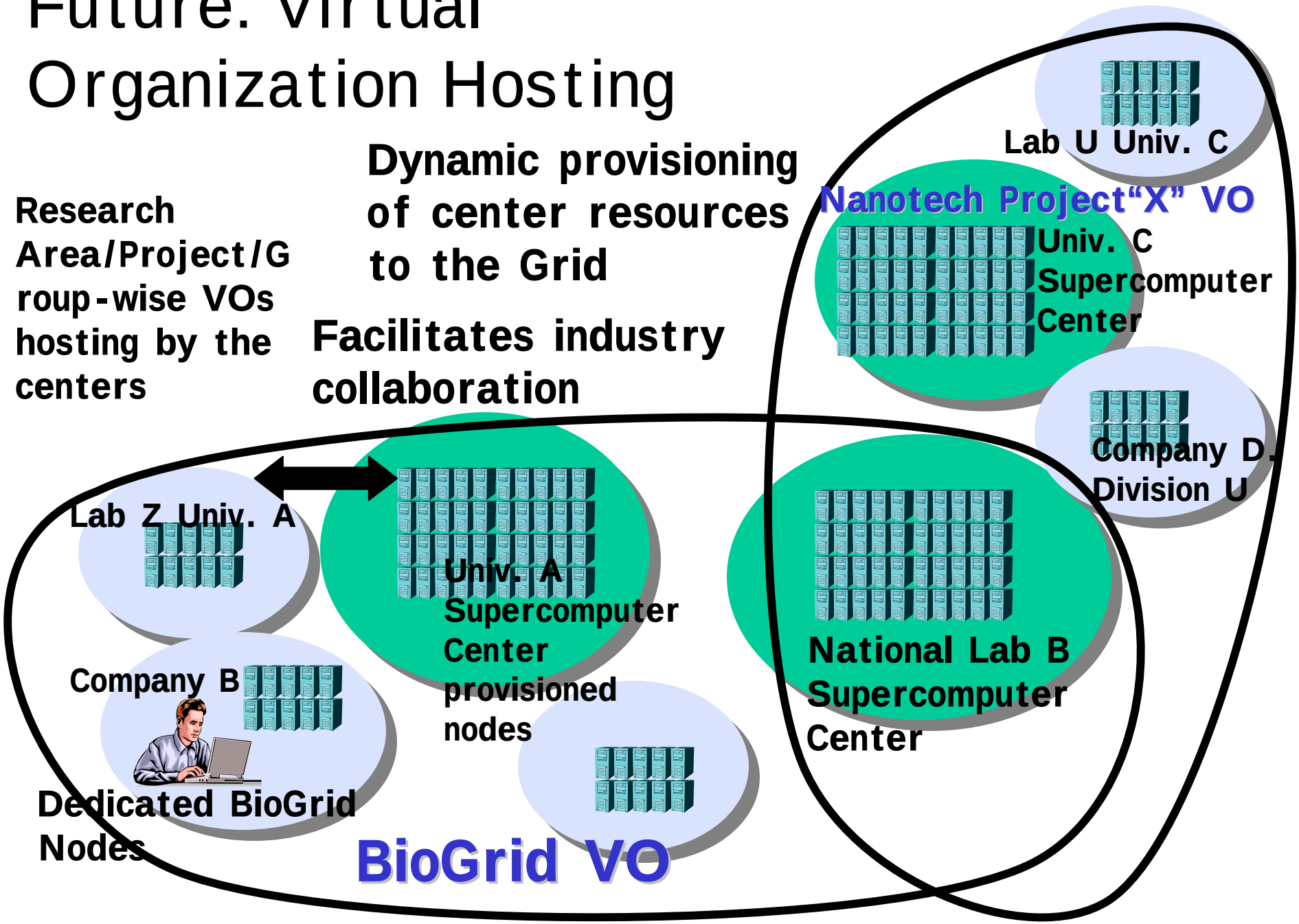
Metacomputing

Future: Virtual Organization Hosting

Research Area/Project/Group-wise VOs hosting by the centers

Dynamic provisioning of center resources to the Grid

Facilitates industry collaboration



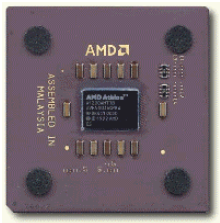
2. High Performance Grid Computing with Commodity Clusters

Next Generation TeraFlops Supercomputing with "Commodity Grid"

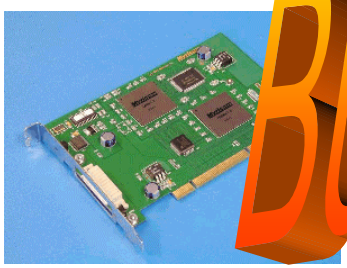
Rapid Rise of Commodity Computing Power
-High-perf. commodity PC technology
-Ultra High-Speed (LAN/WAN) technology

New Generation PC Clusters
- x10 ~ x100 cost performance
-Upgrade follows technology curve

Ultra High-Speed Backbone + Grid Software Technology => Widespread National Infrastructure



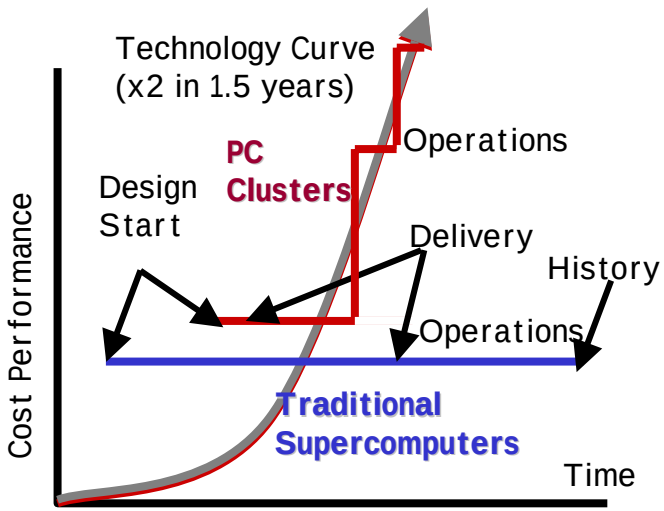
AMD Athlon
1.6Ghz
3.2 GigaFlops peak



"Commodity Grid"

But is this vision True?

Myrinet 2000 Multi-Gigabit Network



"Commodity Supercomputing Everywhere"
- x 100 increase in compute & storage power to scientists
- Direct access to and processing of sensory and archived petascale data
- Multi-Tier comllaboration (Virtual Lab)

Various projects in Japan are springing up rapidly FY 2002-3

The “Field of Dreams”

- “Systems research needs to be apps driven”
- But how?
- The “field of dreams” – one shall build an infrastructure, and they shall come and play
- The infrastructure must be seriously large
  - 10 node cluster – “OK, that’s interesting”
  - 500 node cluster – “How do we work with you?”
  - 10,000 node Grid of clusters – “You can have my house!”
- Does this scenario really work? – users, HW/SW, management?

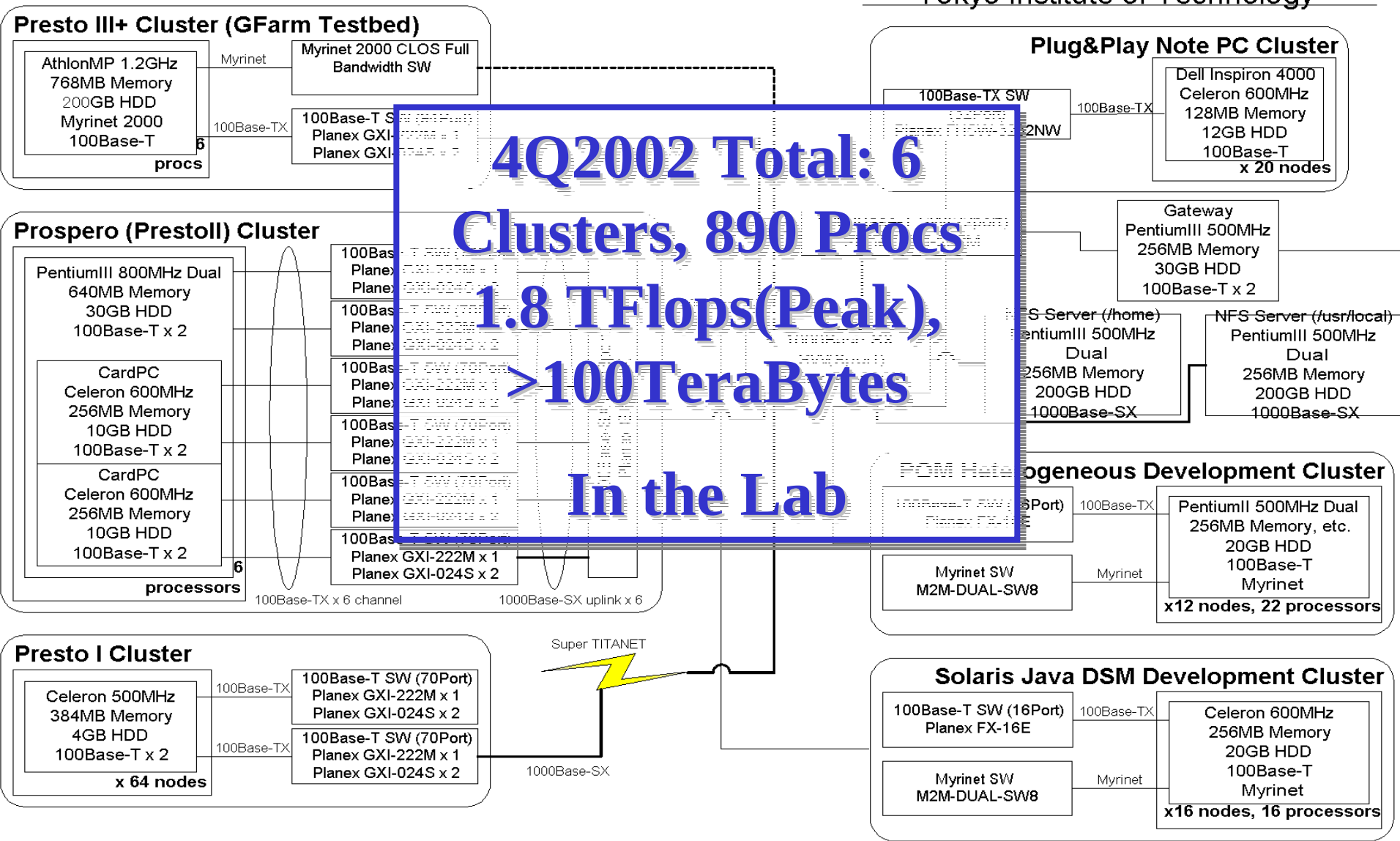
Lab Experimental Grid Cluster Projects

- Presto III Cluster Experiment
  - *Q: Can a Single Univ. Lab maintain a Terascale computing facility?*
  - Collaboration with AMD, Bestsystems Co., Tyan, Appro, Myricom
  - Dual 256 node/512 proc AthlonMP 1900+ Rpeak 1.6 TeraFlops
  - Full Myrinet 2K network
  - June 2002 47th Top 500, 716GFlops
    - 2nd Fastest PC cluster at the time
  - 100Terabyte Storage for storage intensive/ DataGrid apps
  - $1 million
- Applications
  - "Field of Dreams"
  - DataGrid Gfarm/LHC/ATLAS
  - Bioinfomatic Apps
  - Advanced nonlinear optimizations
  - Of course, Grid&Cluster software research
- Nov. 2002: 68th (760.1 GigaFlops)
  - The Livermore Cluster achieves 5th, 5.8TeraFlops (Linux Networx)
- June 2003 Clockup the CPU
  - 1.6Ghz -> 2Ghz peak (planned)
  - Replace and send older 128node/256procs to KEK (planned)
    - 384 nodes, 768 CPUs
    - 2.8 TeraFlops combined Titech/KEK

Titech GSIC Matsuoka Lab Grid Cluster Infrastructure (4Q2003)

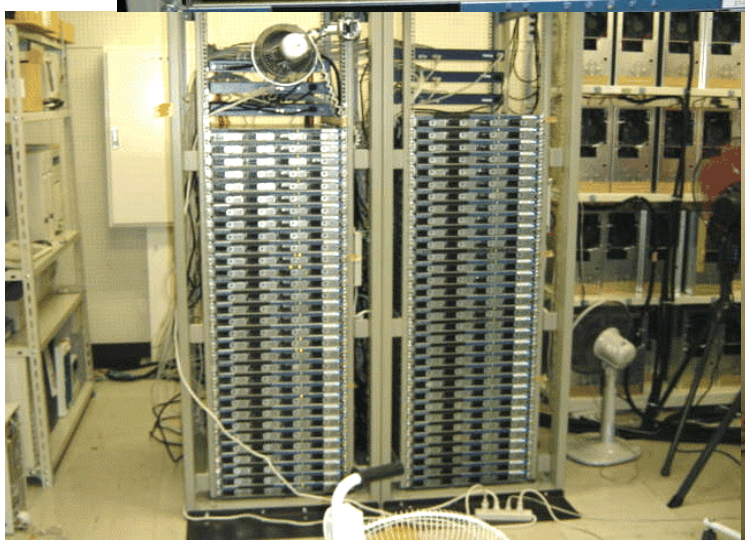
Total: 6 Clusters, 634 Processors
~858 GFLOPS (Peak), 29.5 TeraBytes (Storage)

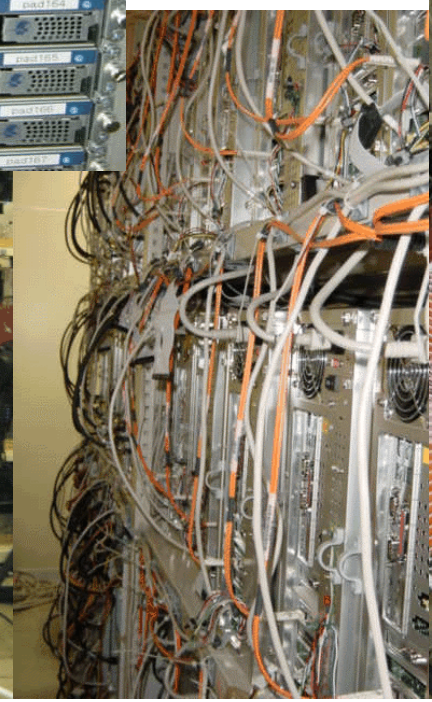
Grid Cluster Infrastructure
Matsuoka Laboratory (3Q2001)
Tokyo Institute of Technology

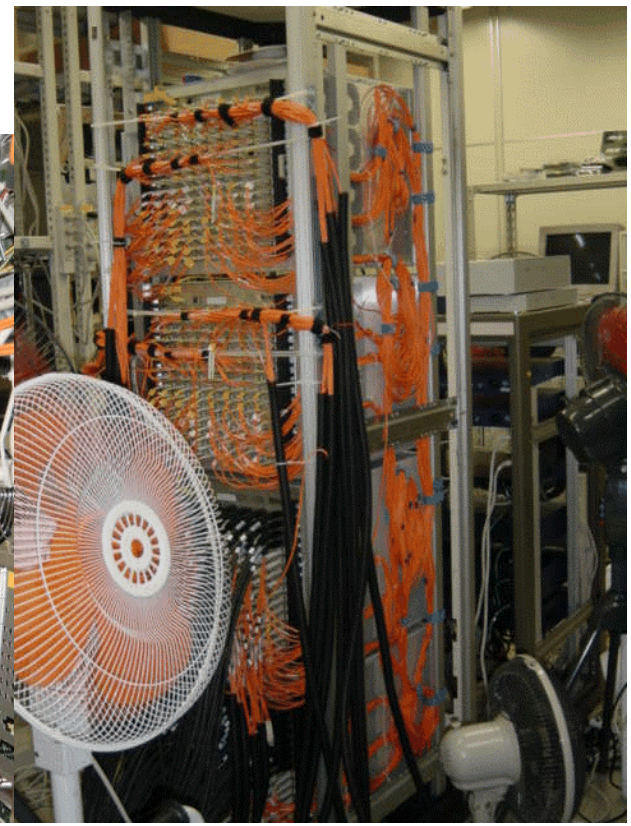


4Q2002 Total: 6 Clusters, 890 Procs
1.8 TFlops(Peak), >100TeraBytes
In the Lab

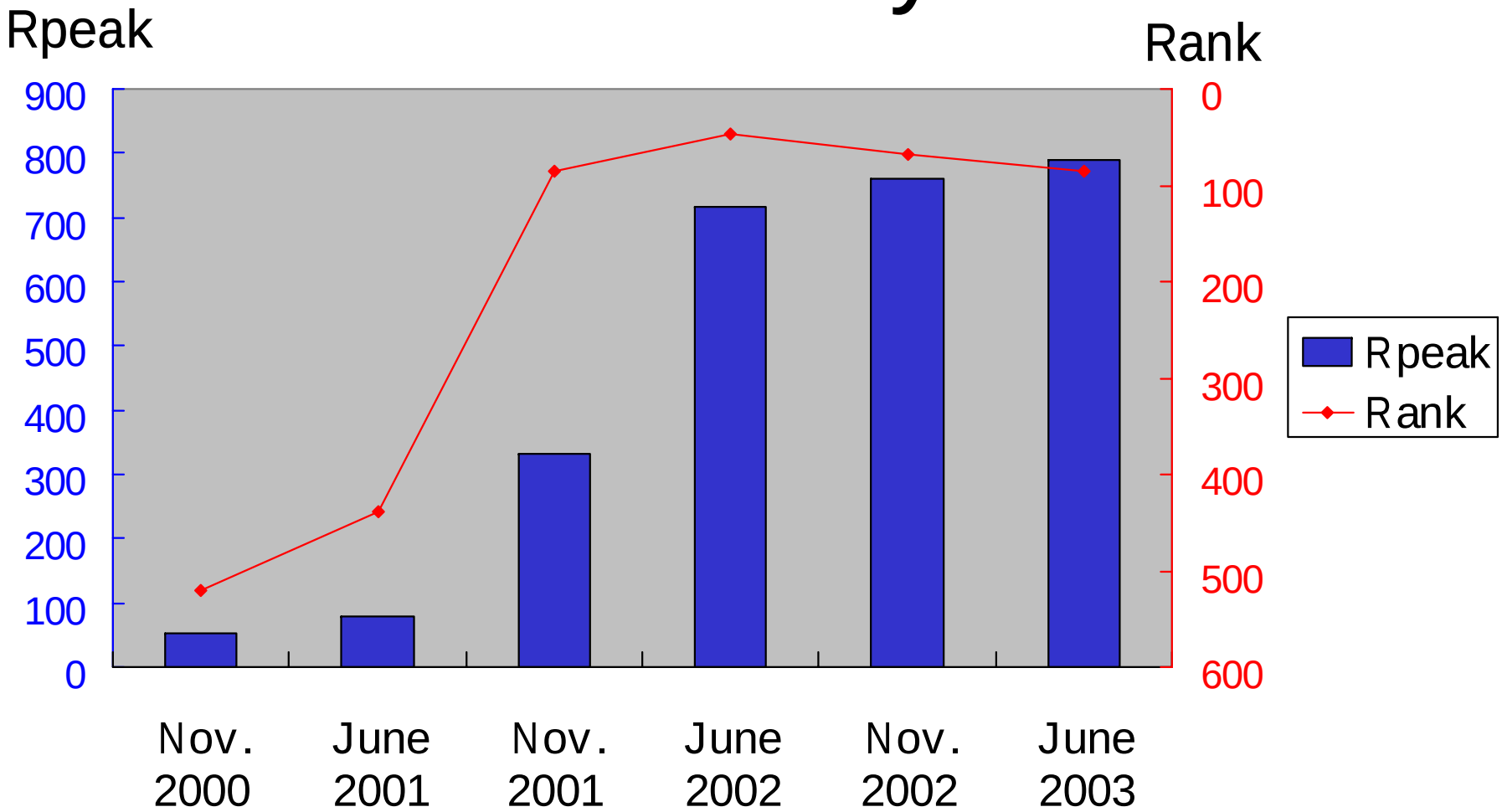
Presto III Cluster in the Lab Computing Room







Presto Cluster Top500 History

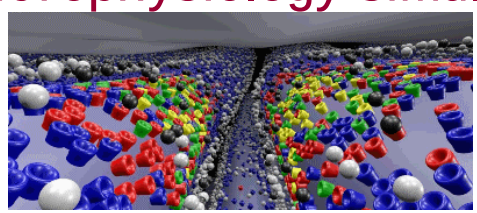


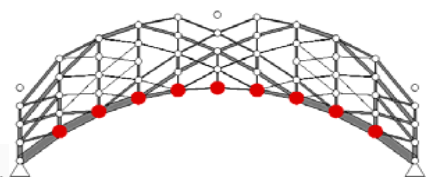
Lab Grid Clustering Software Projects

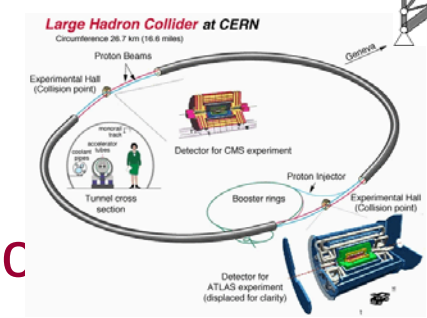
- Dependable, Fault-tolerant clustering for the Grid
  - Parakeet Fault Tolerant MPI
  - Fault Tolerant GridRPC
- Plug&Play Clustering
  - Extended Parakeet
  - Lucie Dynamic Cluster Management
- Heterogeneous Clustering
  - Hetergeneous Omni OpenMP
  - Heterogeneous High Performance Linpack (HPL) and NAS PB
  - Muliple Cluster Coupling on the Grid
- Grid Projects on Clusters
  - Ninf-G GridRPC
    - SOAP/XML GridRPC prototype
  - GFarm – middleware for Petascale data processing
    - Automated data transfer
    - Fast&Secure Job Launch
    - Integrated Replication and FT
  - Grid Performance Benchmarking and Monitoring
  - Bricks Grid Simulator
    - DataGrid Simulation
- Java for Cluster and the Grid
  - JiPANG Jini-based Grid Portall
  - JavaDSM – Secure and Portable Java DSM System for Clusters
  - JPoP – Parallal Optimization Framwork

Field of Dreams in Action
Grid Cluster Applications Collaboration

- Advanced Operations Research (Prof. Kojima et. al., TITECH)
  - Cyclic Polynomial Solution Enumeration, World Record Problem Size
  - Successive Relaxation Method for Non-Convex Quadratic Programming, World Record
- Bricks Grid Simulator (our own app)
- UCSD/Salk MCell (cellular microphysiology simulation, F. Berman, H. Casanova et. al.)
  - Over TransPAC/APAN
- High-Energy Physics
  - LHC/ATLAS Gfarm collaboration
- Various Bioinformatics
  - Open BioGrid (RIKEN) and others
- National Astronomical Observatc
- ...



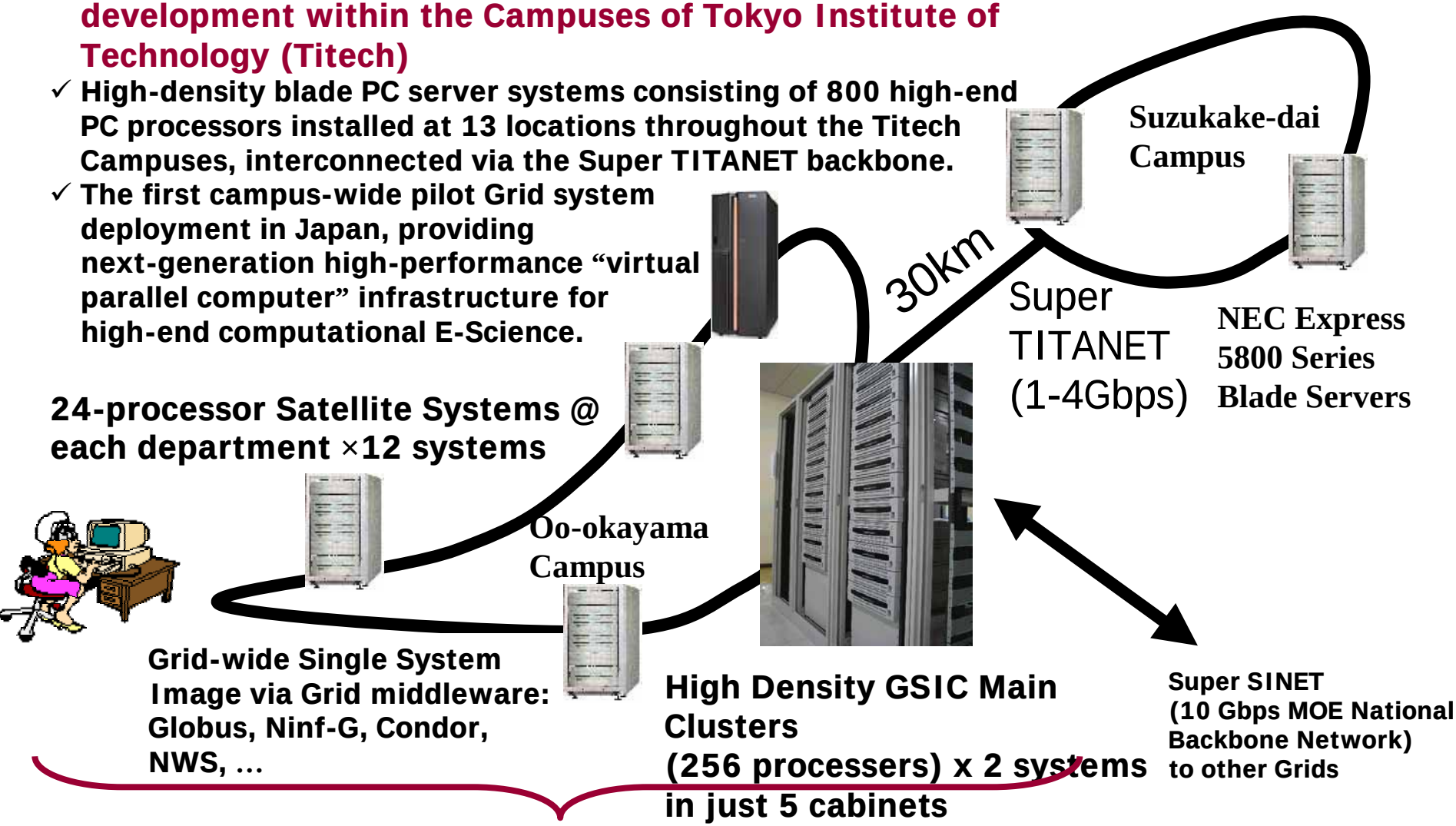




Titech Campus Grid - System Image

Titech Grid is a large-scale, campus-wide, pilot commodity Grid deployment for next generation E-Science application development within the Campuses of Tokyo Institute of Technology (Titech)

- ✓ **High-density blade PC server systems consisting of 800 high-end PC processors installed at 13 locations throughout the Titech Campuses, interconnected via the Super TITANET backbone.**
- ✓ **The first campus-wide pilot Grid system deployment in Japan, providing next-generation high-performance "virtual parallel computer" infrastructure for high-end computational E-Science.**



800-processor high-perf blade servers, > 1.2 TeraFlops, over 25 Terabytes stoarge

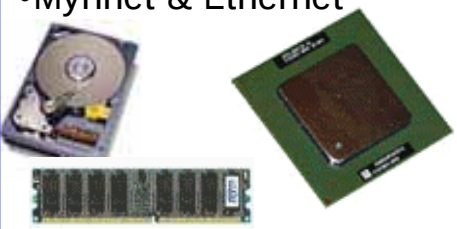
How we implement the Titech Grid

- Fact: Departments lack space, power, air-conditioning, maintenance expertise etc.
- Technological solutions
  - High density, high performance blade design by NEC
    - x2 density c.f., 1U rack server design
- And of course, all the Cluster & Grid middleware
  - Phase 1 (May – Aug 2002): RWCP Score Cluster-OS single user & multiuser deployment
  - Phase 2 (Sep – Oct): Condor, utilize other dept. clusters in background
  - Phase 3 (Oct – Dec): Globus deployment, test CA management, Ninf-G
  - Phase 4 (Dec - 2003): Global Grid Queue, Sun GE + Score multiuser
  - Deployed Software: SCore, Condor, Globus, Sun GridEngine, NWS, Ninf-G, MPICH-G, Avaki, Cactus Gfarm, Lucie…etc.
- It works!
  - People are starting to flood the infrastructure

Titech Grid - Details

Commodity High-End PC Componets

- Intel Pentium III-S 1.4Ghz, 512KB L2 Cache × 2
- 80GigaByte HDD
- 512MB~1GB DIMM
- Myrinet & Ethernet



NEC Express 5800 BladeServer

Department Satellite Cluster

-12 nodes, 24CPU

-512MB mem 80GB HDD/node

-Trunked 100Base-T network

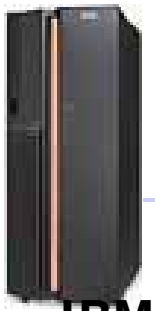
-Small 13U Cabinet

=> Small Desk, 20A/100V Power

-**Participating Departments/Labs**

×12 sites

On 2 campuses (Oo-okayama and Suzukakedai)



IBM p670 High-end resource

GSIC Cluster

- **128Nodes/256CPUs x 2 machines**

-Myrinet2000 Gigabit Networking

--1GB Memory/node

- Very dense, 4CPUs / 1U, 4 racks

-**Installed at GSIC**

GSIC

Grid Operation +R&D

-Operations

-Middleware Deployment

-Application Porting

-Portal Management

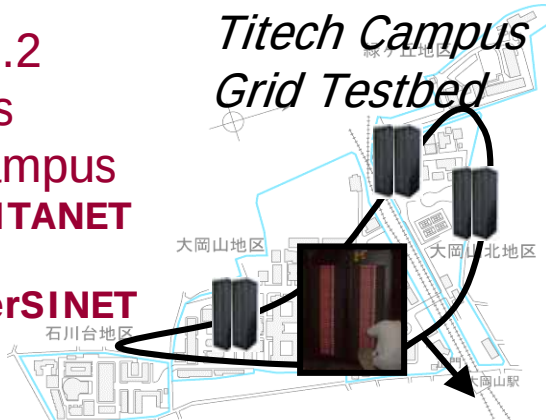
Virtual Terascale Supercomputer via Grid Middleware

Total over 800CPUs, 1.2 TeraFlops, 25TeraBytes installed throughout campus

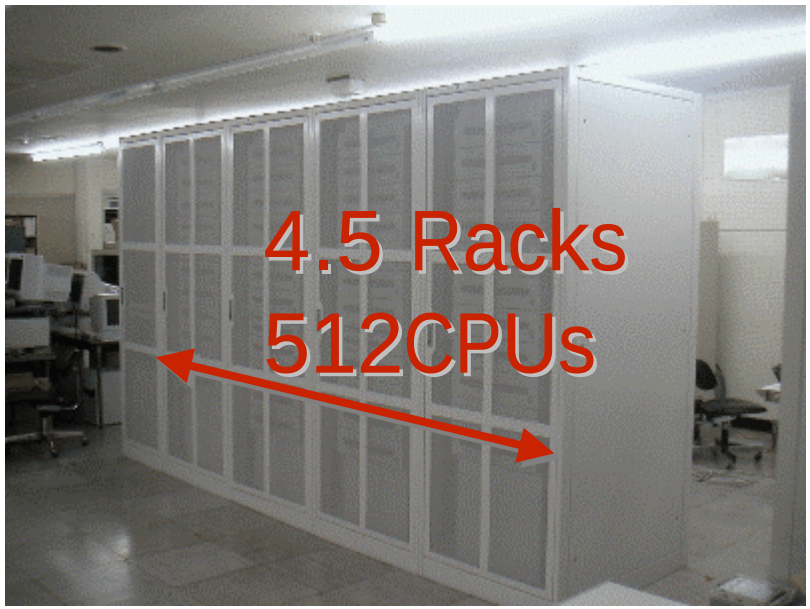
-Intraconnection via **SuperTITANET** campus network (1-4Gbps)

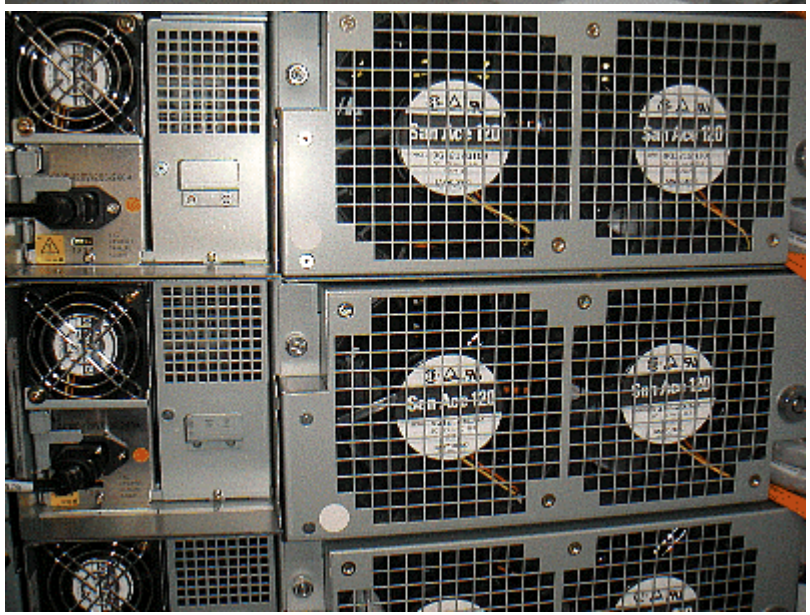
-Network to outside via **SuperSINET** (10Gbps National backbone)

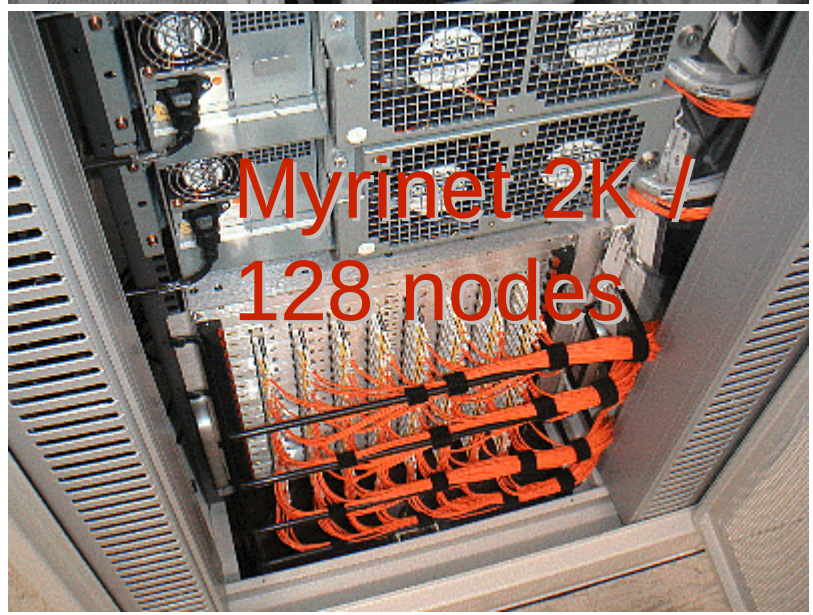
Titech Campus Grid Testbed



Titech Grid GSIC Center Clusters A, B (256CPU x 2)

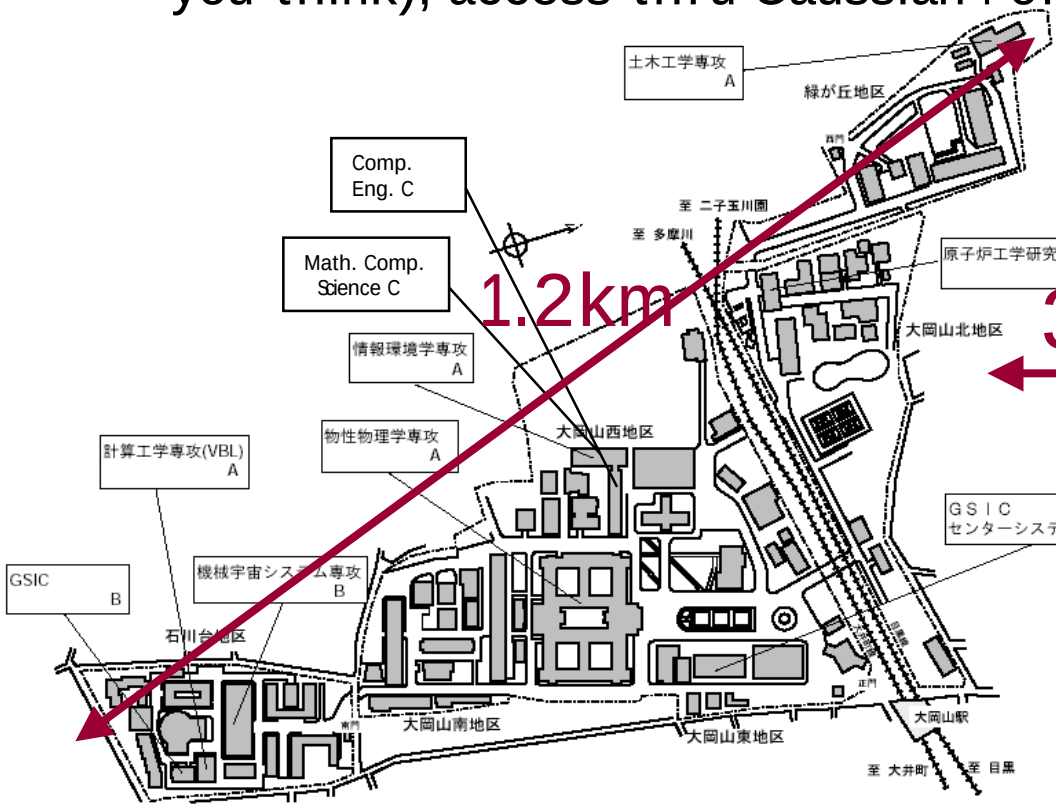






Titech Grid Campus Sites

- 14 installations, 18 participating departments
  - Univ.-wide solicitation and applications thereof
  - Department listed apps – Bioinfo, CFD, Nanotech, Env. etc.etc.
  - Ubiquitous Condor-Gaussian, GAMESS being planned (harder than you think), access thru Gaussian Portal (by AIST-GTRC)

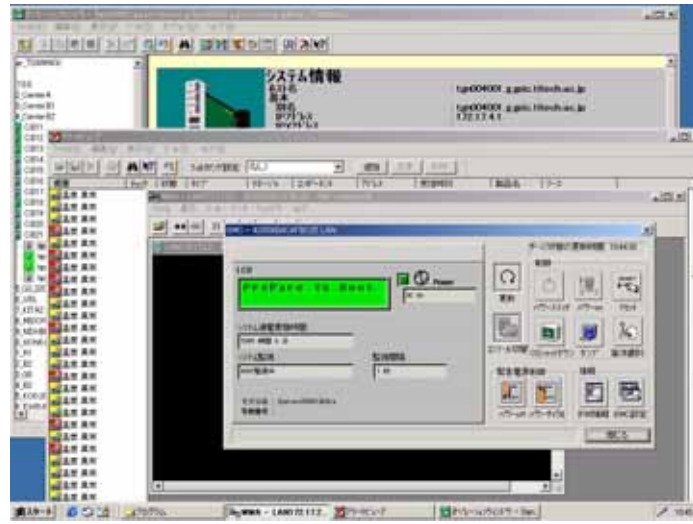


Oo-okayama campus (10 sites)

Suzukakedai campus (5 sites)

Titech Grid Usage Scenario

- Extensive adoption by users, production-level load
- But, observed usage scenario:
  - Single cluster (MPI) jobs
  - Metacomputing: Param sweep and B&B search via Condor and Ninf
  - No cross-machine MPI, etc.
- Dealing w/ distributed HW/SW harder than expected
- Compute/Datacenter better?
  - A set of large clusters and storage servers @ GSIC only w/ Hi-BW networks
  - 50TF, 3 PB in a building y2005



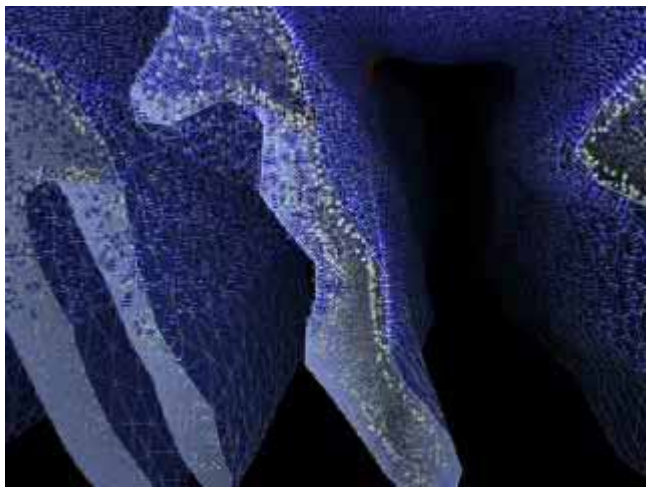
Onto a Production Grid – Reality

- Bootstrapping Problem
  - User Side
    - People not used to sharing compute resources
    - People not used to using various Grid middleware
    - People want to share federated data but don't know how or have time to learn to use tools
    - People do not have idea or experience of coupling applications on the Grid
  - Center (Operations) Side
    - Do not (yet) have skills to manage clusters
    - Do not (yet) have skills to manage large machines in distribution
    - Do not (yet) have skills to manage Grid middleware
    - Do not (yet) have skills to facilitate campus-wide security
  - Research Side
    - Do not have skills to manage a center with over 1000 users
    - Do not know if middleware or tools will scale

So currently the Titech Infrastructure is:

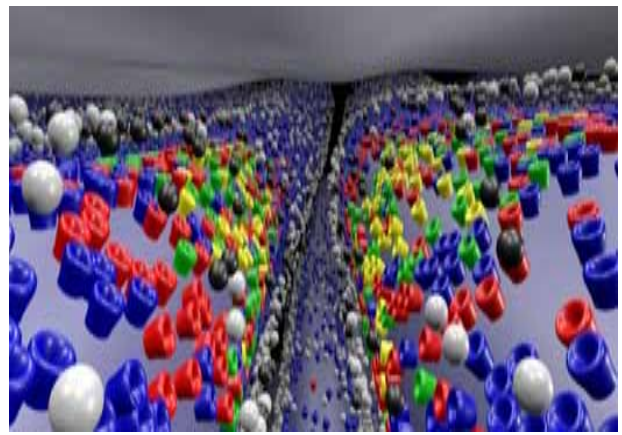
- Traditional Supercomputer
  - 336 procs, 450 GigaFlops
- Semi-Production Titech Grid Clusters & Servers
  - 816 procs, 1.26 TeraFlops (would be 4 TeraFlops today)
- Lab Experimental Clusters (incl. PrestoIII)
  - ~900 procs, 1.8 TeraFlops (> 2 TeraFlops 2003)
- Total 2000 procs, 3.5-4 TeraFlops
  - Still modest-where do we go from here?
  - Replacement Jan 2006 -> 50-150 TeraFlops Campus Grid

SDSC/Titech Collaboration: MCell: Cellular Microphysiology over APAN/TransPAC

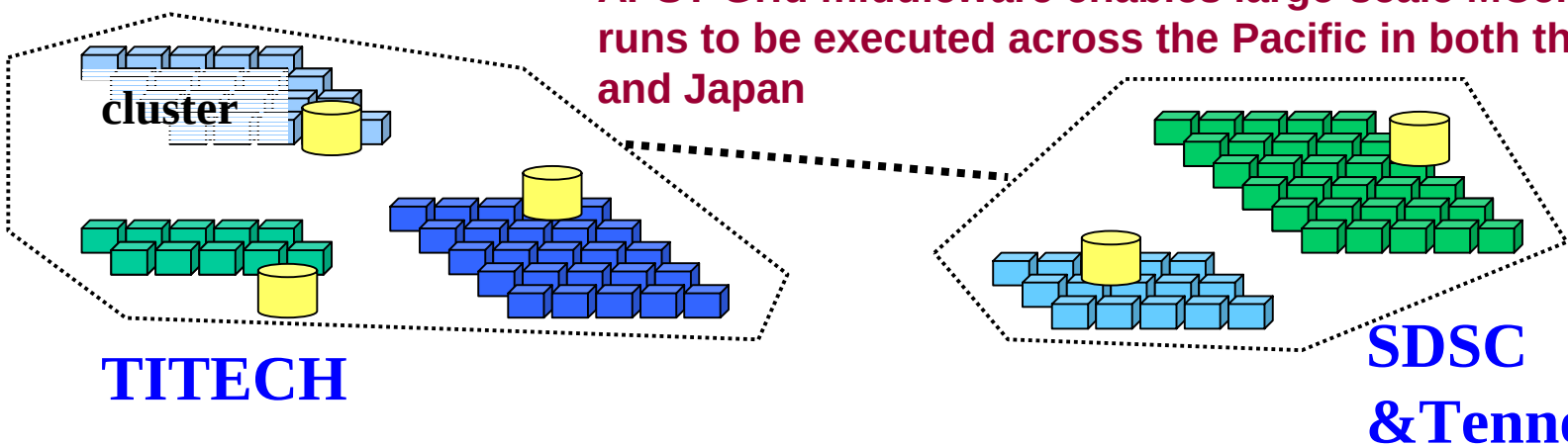


Mcell animation

MCell simulation of the process of synaptic transmission at the neuromuscular junction of a rat diaphragm muscle.



APST Grid middleware enables large-scale MCell runs to be executed across the Pacific in both the US and Japan



Experimental Results

Experimental Setting:

Mcell simulation with 1,200 tasks:

- composed of 6 Monte-Carlo simulations
- input files: 1, 1, 20, 20, 100, and 100 MB

Result: With proper scheduling, (a)-(c) performed competitively to (d)

4 scenarios:

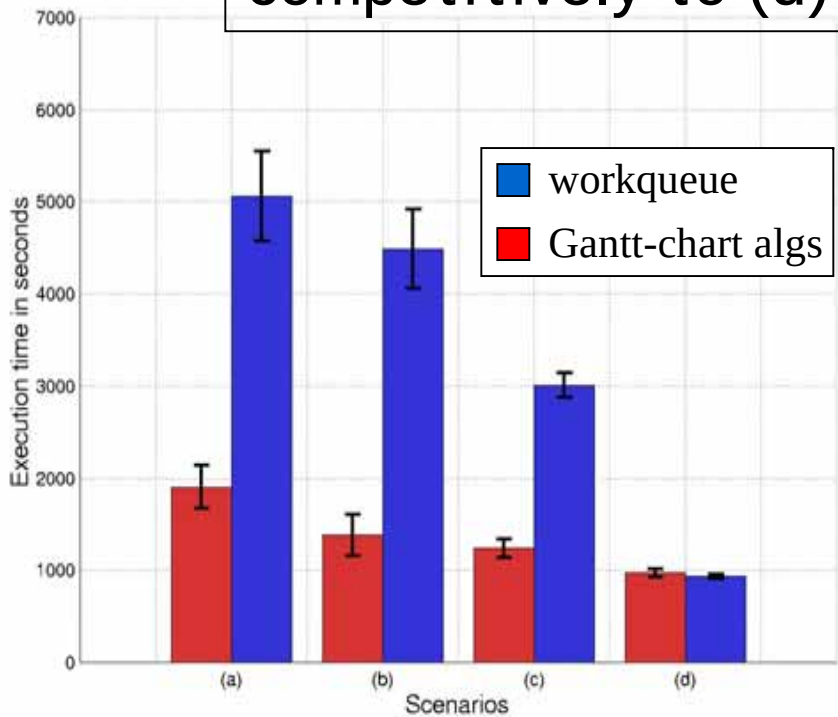
Initially

(a) all input files are only in Japan

(b) 100MB files replicated in California

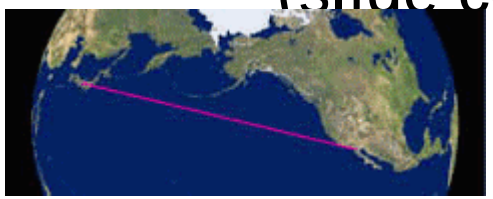
(c) in addition, one 100MB file replicated in Tennessee

(d) all input files replicated everywhere



APAN Trans-Pacific Telemicroscopy Collaboration, Osaka-U, UCSD, ISI

(slide courtesy of Mark Ellisman@UCSD)



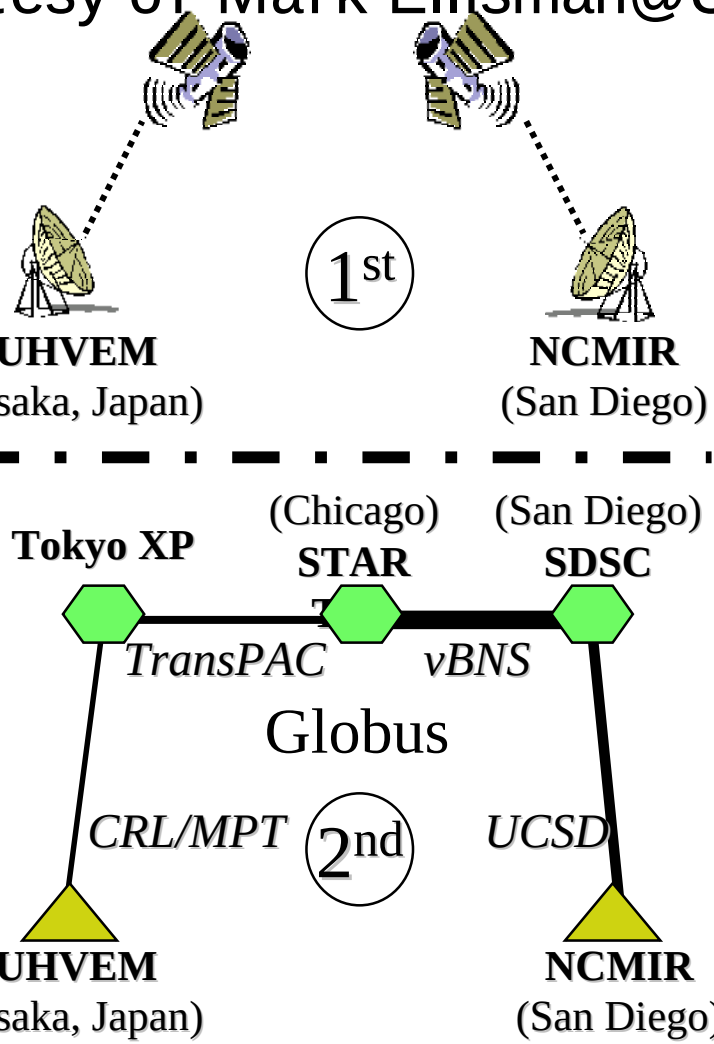
NET NEWS

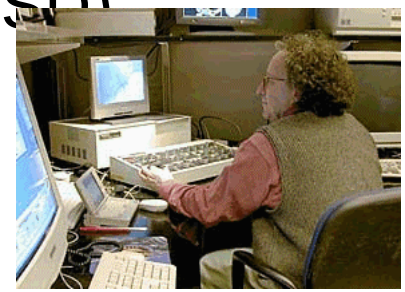
Microscopy Across an Ocean

A big push by biologists to use computer networks to operate rare instruments from afar passed a major milestone on 25 June: Scientists took a spin on the world's most powerful electron microscope in Japan—while sitting 6000 kilometers away in California.

Six years ago, University of California, San Diego, neuroscientist Mark Ellisman thrilled audiences at a conference in Chicago by using the Internet to control an electron microscope in San Diego. Several U.S. agencies jumped in to fund projects for operating microscopes by remote control, and by now at least a dozen groups are doing so in the United States. Ellisman's team has since moved on to the Mount Everest of microscopes: Osaka University's Ultra High Voltage Electron Microscope, a 3,000,000-volt behemoth that can create three-dimensional images from much thicker samples (such as biological cells) than ordinary microscopes can. Ellisman and his U.S. and Japanese colleagues wondered if they could operate this instrument's roomful of controls from across the Pacific Ocean.

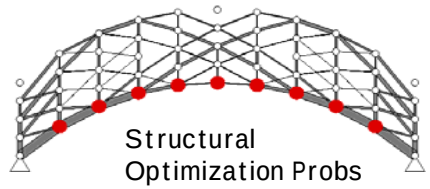
They showed they could. Over 5 hours, San Diego scientists imaged nerve cells from a rat and a frog without setting foot in Japan, controlling things like focus and specimen position across a private data line while the images came in across a satellite video link. Ellisman says this lays the groundwork for researchers all over the United States and Japan to borrow each others' specialized microscopes, probably via a high-speed Internet link, "within a year or two."





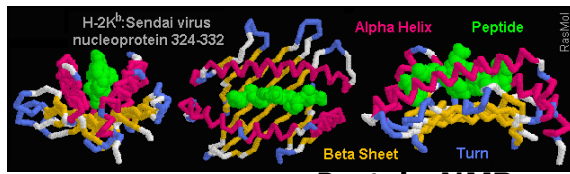
ACT-JST Commodity Optimization Grid Project

3) Grid-Enabled, Terascale Mathematical Optimization Libraries and Apps
- **Non-Convex Quadratic Optimizaion using SCRM**
- **Higher-order polynomial solving w/Homotopy meth.**
- **BMI optimization for control theory apps**
- **Parallel GA for Genome Informatics apps**

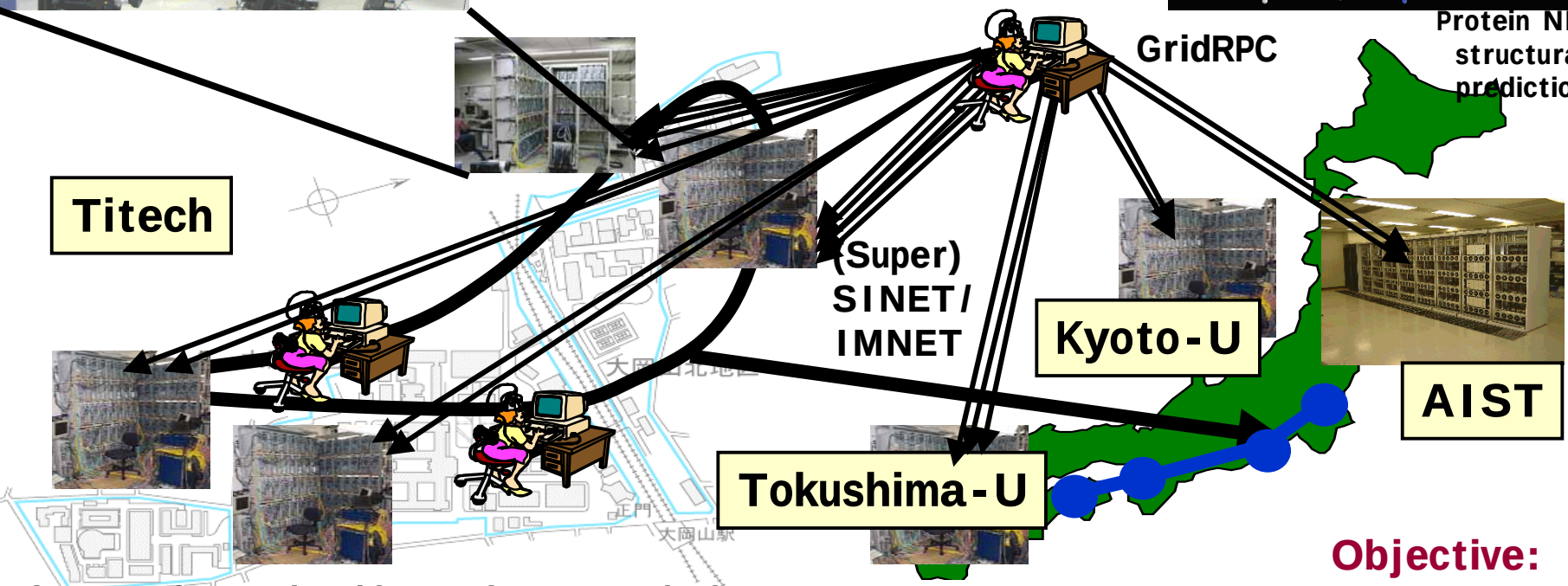


$x_1+x_2+x_3=c_1$
$x_1x_2+x_1x_3+x_2x_3=c_2$
$x_1x_2x_3=c_3$
Cyclic Polynomial
All-Solutions

2) Scalable extensions for GridRPC
- **a) >million task parallelism**
- **b) FT under various fault models**
- **c) High-level and generalized GridRPC-API**



Protein NMR structural prediction



1) Commodty Nationwide PC Cluster Testbed
> 1000 processors, multi-teralops

Objective:
sustain 1 Teraflop for a week at 1/100 cost

4) App1: Successive Convex Relaxation Method (SCRM)

$$\begin{aligned} &\text{Maximize} \quad : \quad c^T x \\ &\text{Constrants} : \quad f_l + 2q_l^T x + x^T Q_l x \le 0, \quad l = 1, \ldots, m \\ &\qquad\qquad c \in R^n,\ Q_l \in R^{n \times n},\ q_l \in R^n,\ x \in R^n \end{aligned}$$

Characteristics

- Serve as base solver for Quadratic Integer Programming, Multi-level Linear Programming, etc.
- Need to solve tons of small SDP (Semi-Definite Programming) problems in parallel
- SCRM is well suited for parallelization on the Grid (x90 speedup w/128 processors achieved)

4) All-Solution Finding of Higher-Order Polynomials using the Homotopy Method

Engineering Apps

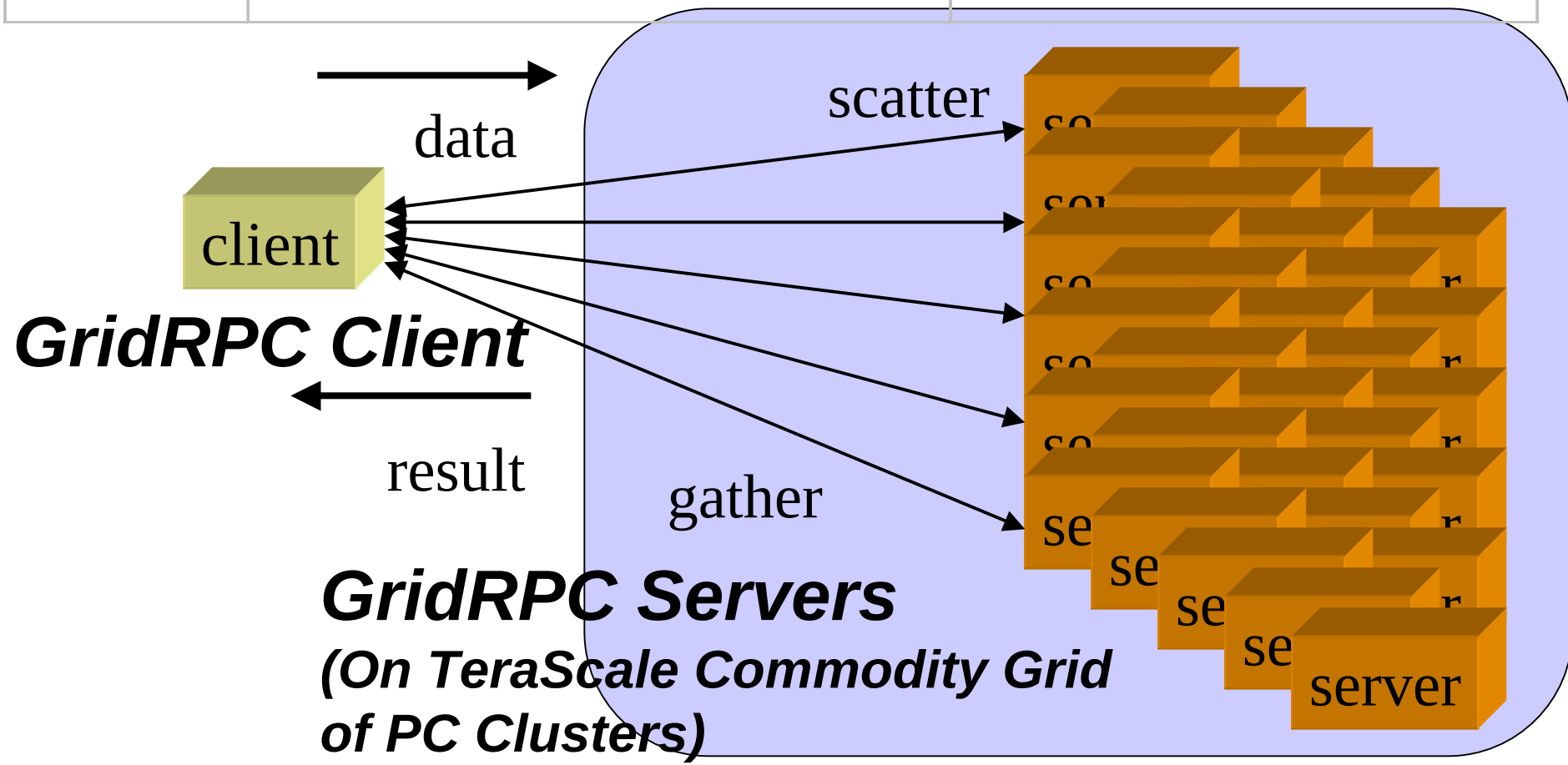
- Construct Interpolation Functions in Numerical Integration
- Compute Geometric Join-Points of Higher-order curves
- Numerous applications in the real world

$$\begin{aligned} x_1 + x_2 + x_3 &= c_1 \\ x_1 x_2 + x_2 x_3 + x_3 x_1 &= c_2 \\ x_1 x_2 x_3 &= c_3 \end{aligned}$$

World Record for Cyclic Polynomial Benchmark – 13th order polynomial
Expect 3 week compute time with 1000 processors on the Commodity Grid

4) Parallelization of Advanced OR Applications

	SCRM	Homotopy
GridRPC Client	Control of Search & Iteration, Generation of Sibling SDP Problems	Control of Search & Iteration, Distribution of Subsolver Data
GridRPC Server	Execute SDPA Subsolver	Execute Homotopy "Path following method"

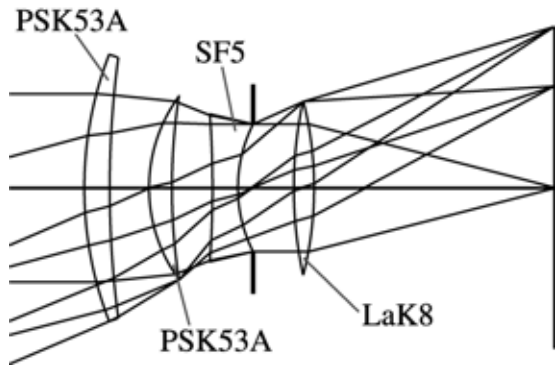


Commodity Optimization Grid:
TeraScale GA Optimization Challenges on the Grid

Apply to difficult problem domain where experts spend considerable time in try&error in optimal solution finding

Example1: Lens Design

Search for curvature, spacing, material, etc. to achieve optimal image ➔ takes an expert weeks, GA-based design in few minutes



NMR Protein Structure Analysis

Optimize protein structure according to the observed NMR signal (takes months ~ year by an expert)

Evaluation of Solution Extremely costly ➔ require massive parallelization (millions of GridRPC calls)
Plan to determine structures of up to 30,000Da class proteins, limit of current NMR scan

3. High-Performance Parallel Programming on the Grid Using Commodity Clusters

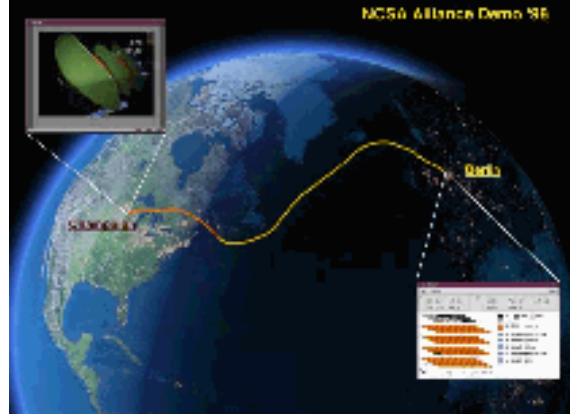
Grid and Existing Distributed Computing Technologies

- Existing Distributed Computing Technologies
  - Common IP tools - telnet, ftp, mail, etc.
  - Distributed OS ('80s)
  - Message Passing - PVM, MPI, Concurrent Objects, etc.
  - Traditional WWW Technologies - HTTP/HTML/CGI
  - ORB (Object-Request Broker)- CORBA, DCOM, Java RMI
  - Agent Technologies (Plangent, Aglets, Voyager, etc.)
  - XML + Web Services (SOAP/WSDL/UDDI/.NET, etc.)
  - Etc. etc.
- Partially useable but insufficient
  - Ex. ORBs cannot constitute single identity and security over multiple organizational domains
  - OGSA (Open Grid Services Architecture) Component Model

Some Necessary Grid Technology Features

- Cross-administrative-domain user/resource authentication, naming, authorization, tracking/auditing, accounting, etc.
- Scalable resource matching and scheduling
- Secure and robust peer-to-peer connectivity (w/firewall)
- "Autonomic Computing" – Self organizing, self-repairing, self-monitoring, fault-tolerant, etc.
- Maitain high-performance and high-throughput --- computation, data, networking
- Classic DC issues such as remote consistency, distributed deadlock handling, etc.

Existing Grid Tools

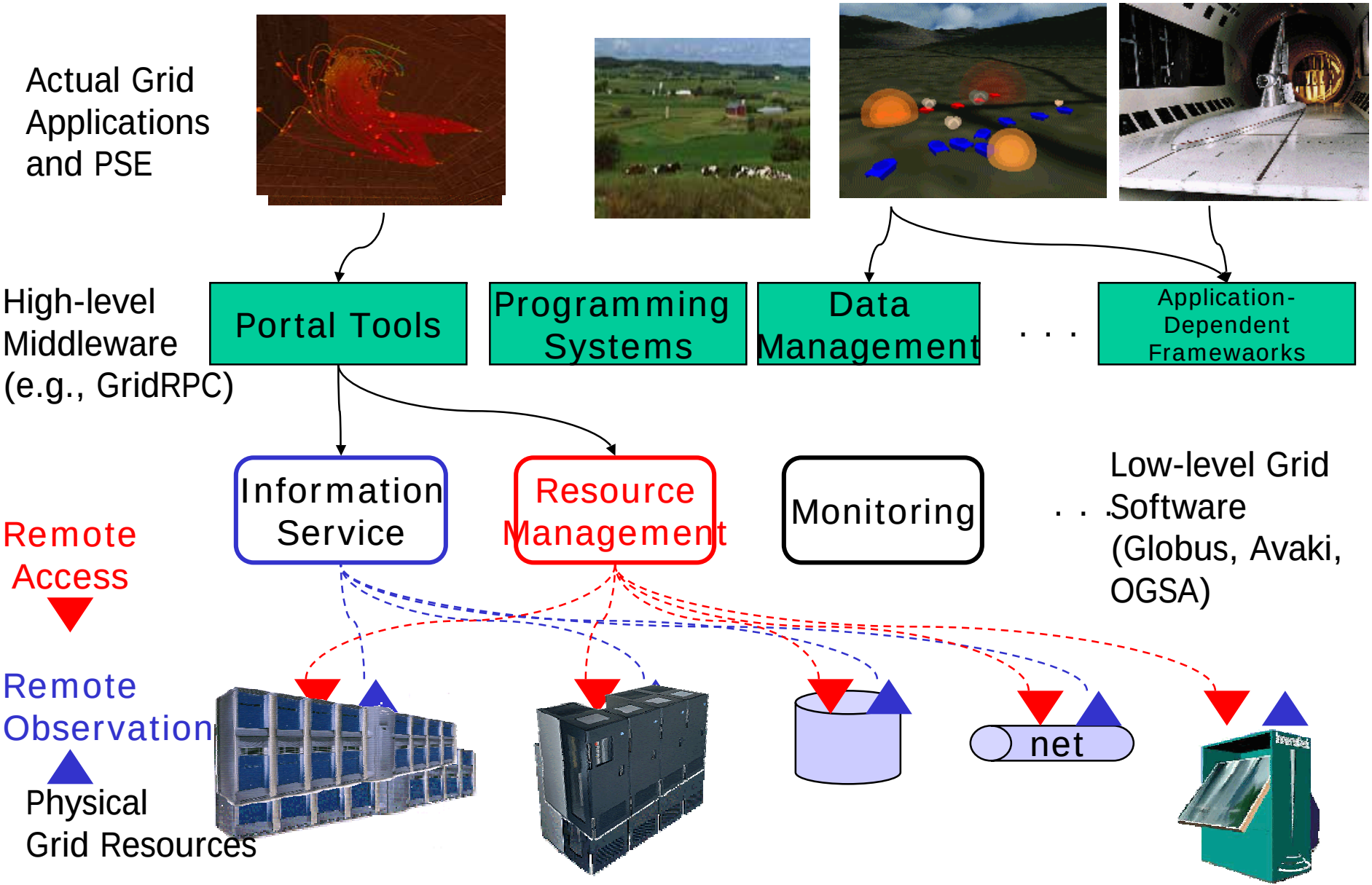


- "GridOS"
  - Globus, Legion (Avaki)
  - OGSA(various companies)
- High-throughput computing
  - Condor-G, Nimrod-G
- Message Passing/ GridRPC/ Programming Tools
  - MPICH-G2
  - GridRPC
    - Ninf-G, Netsolve, Nimrod
  - AppLeS, APST, GAT
- Data Intensive (DataGrid)
  - Globus DataGrid Layer
  - KEK/AIST/Utokyo/Titech Gfarm

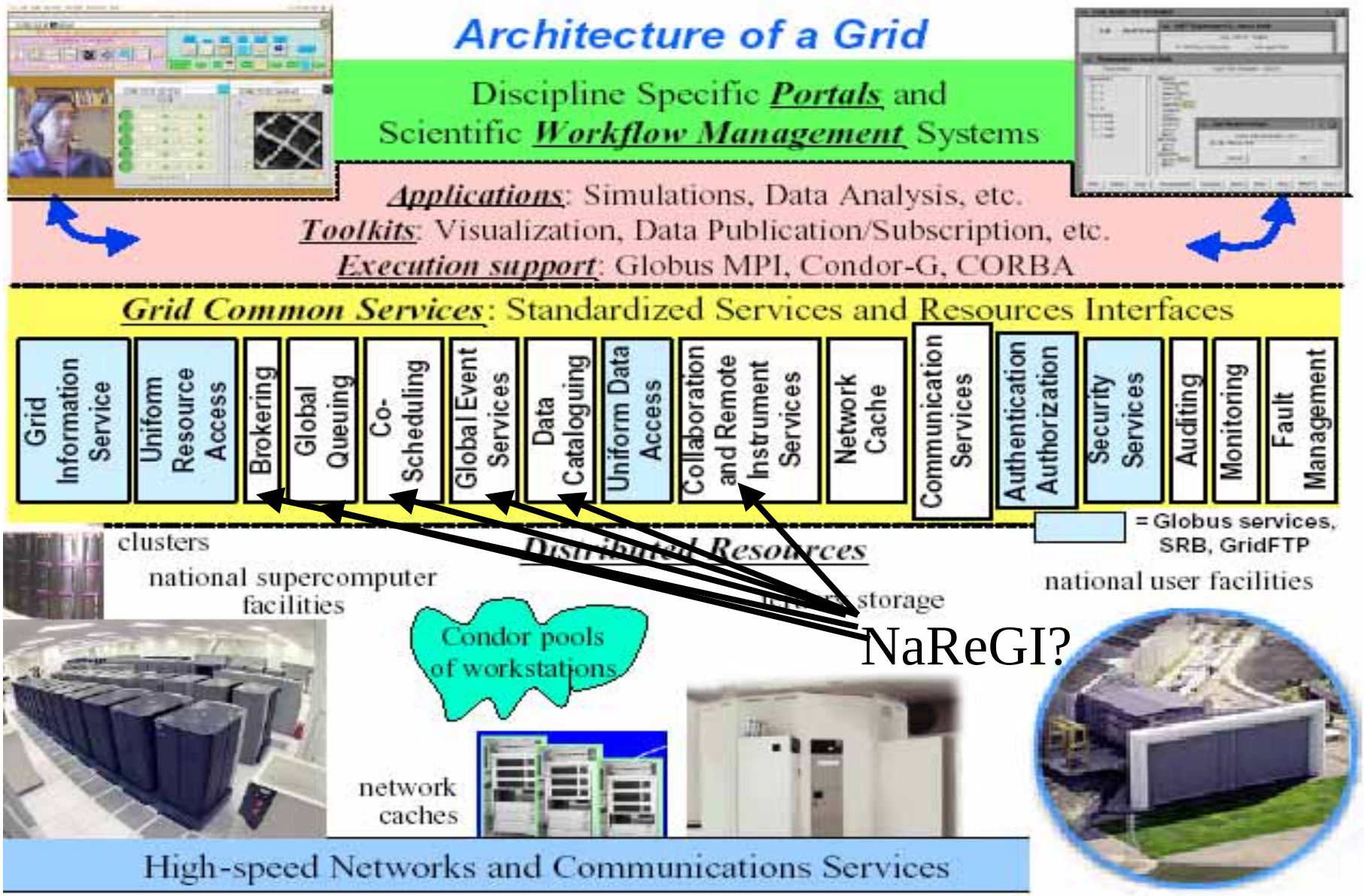
Titech/AIST projects in Red

- Perf Monitoring/Simulation
  - NWS, GMA, Bricks (Titech)
- Desktop access to Grid Resources, Portals
  - Commodity Grid Toolkits (CoG Kits), NPACI GridPort, GrdiSpeed
- Distributed collaboration
  - CAVERNsoft, Access Grid
- Commercial P2P Grid Tools
  - United Devices
  - Entropia
  - Platform Computing
  - Avaki
  - Sun GridEngine
  - Etc.

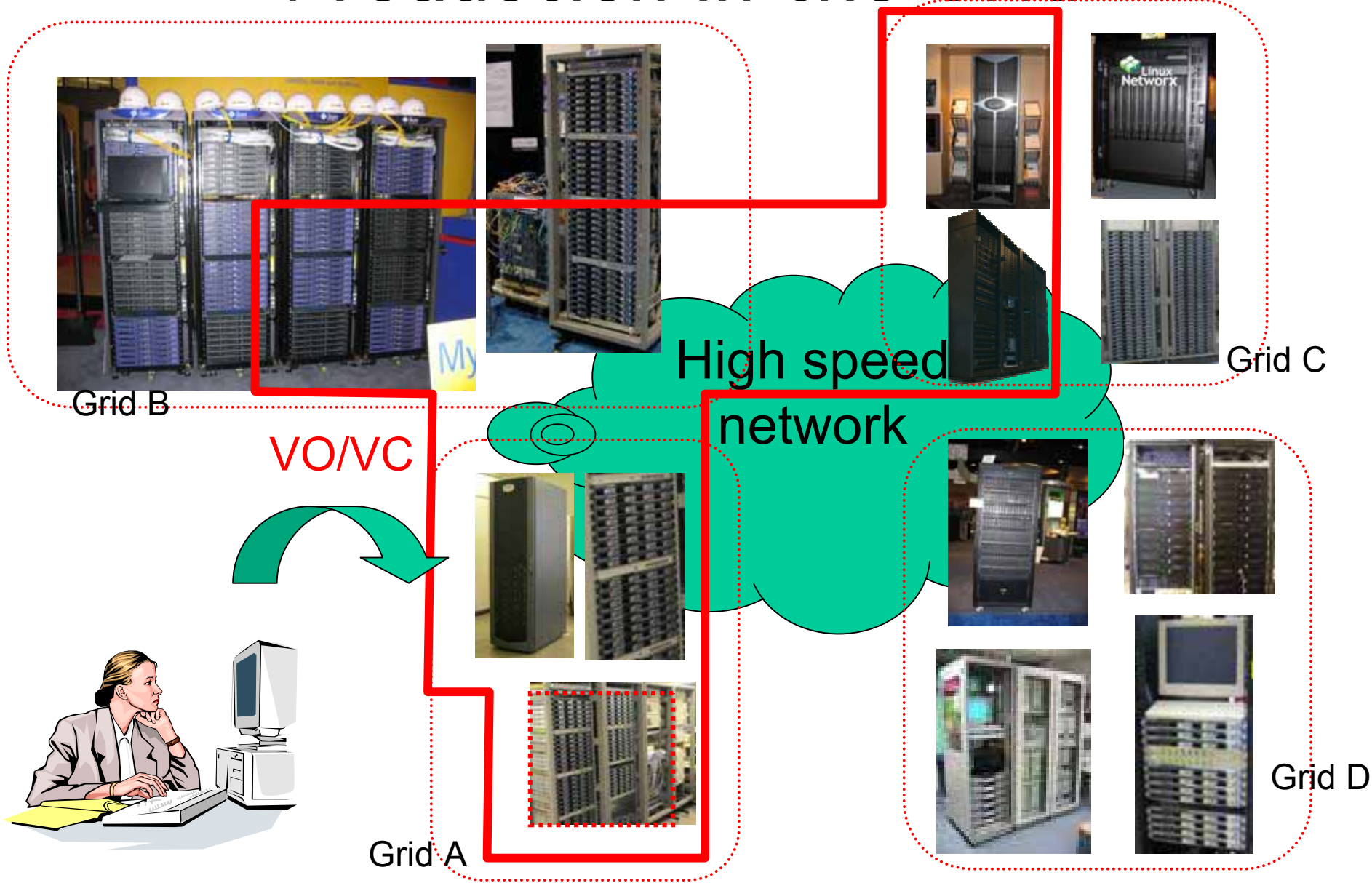
Grid Software Hierarchy



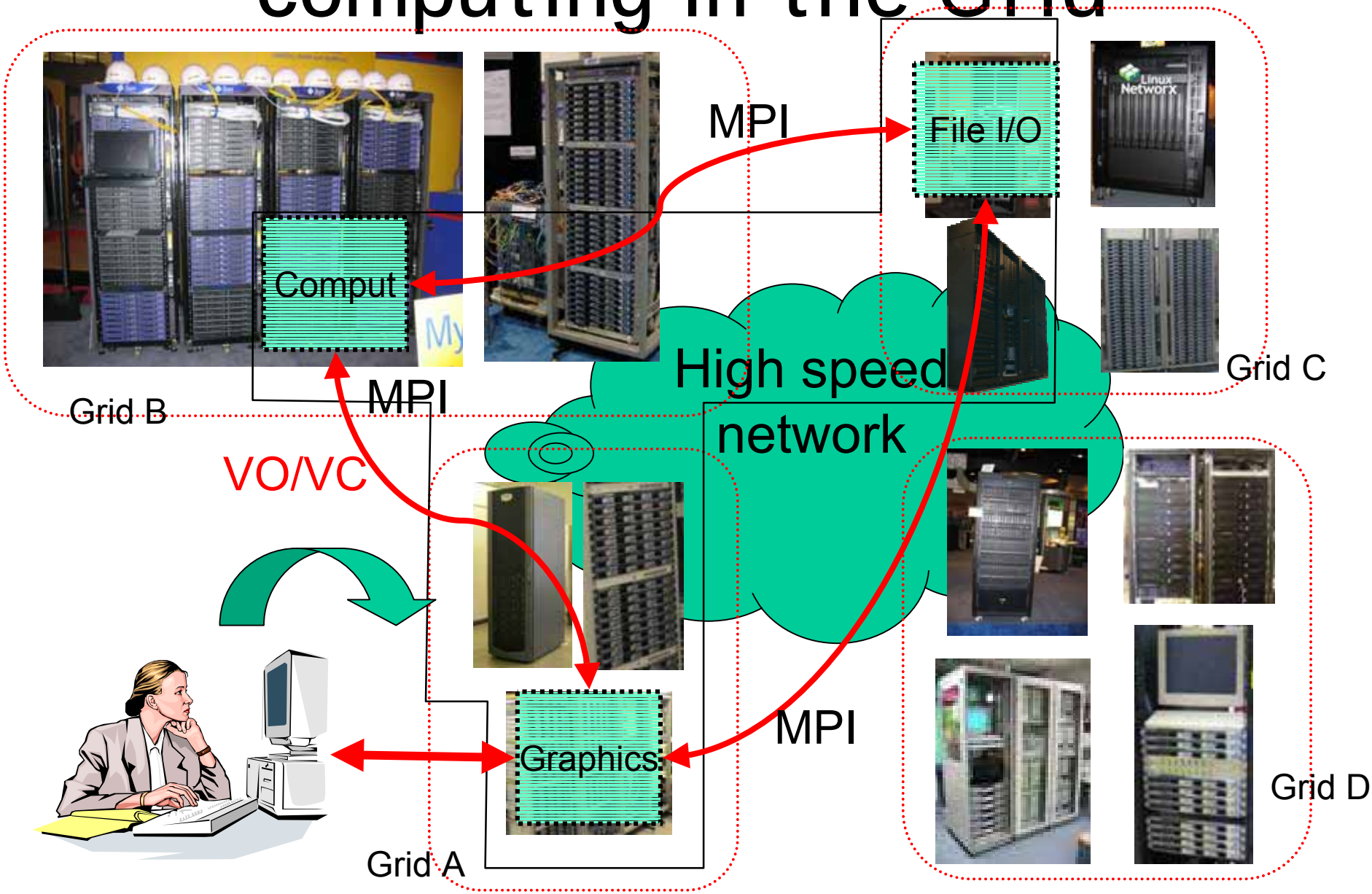
Architecture of the Grid (Slide Courtesy of Bill Johnston, NASA&LBL)



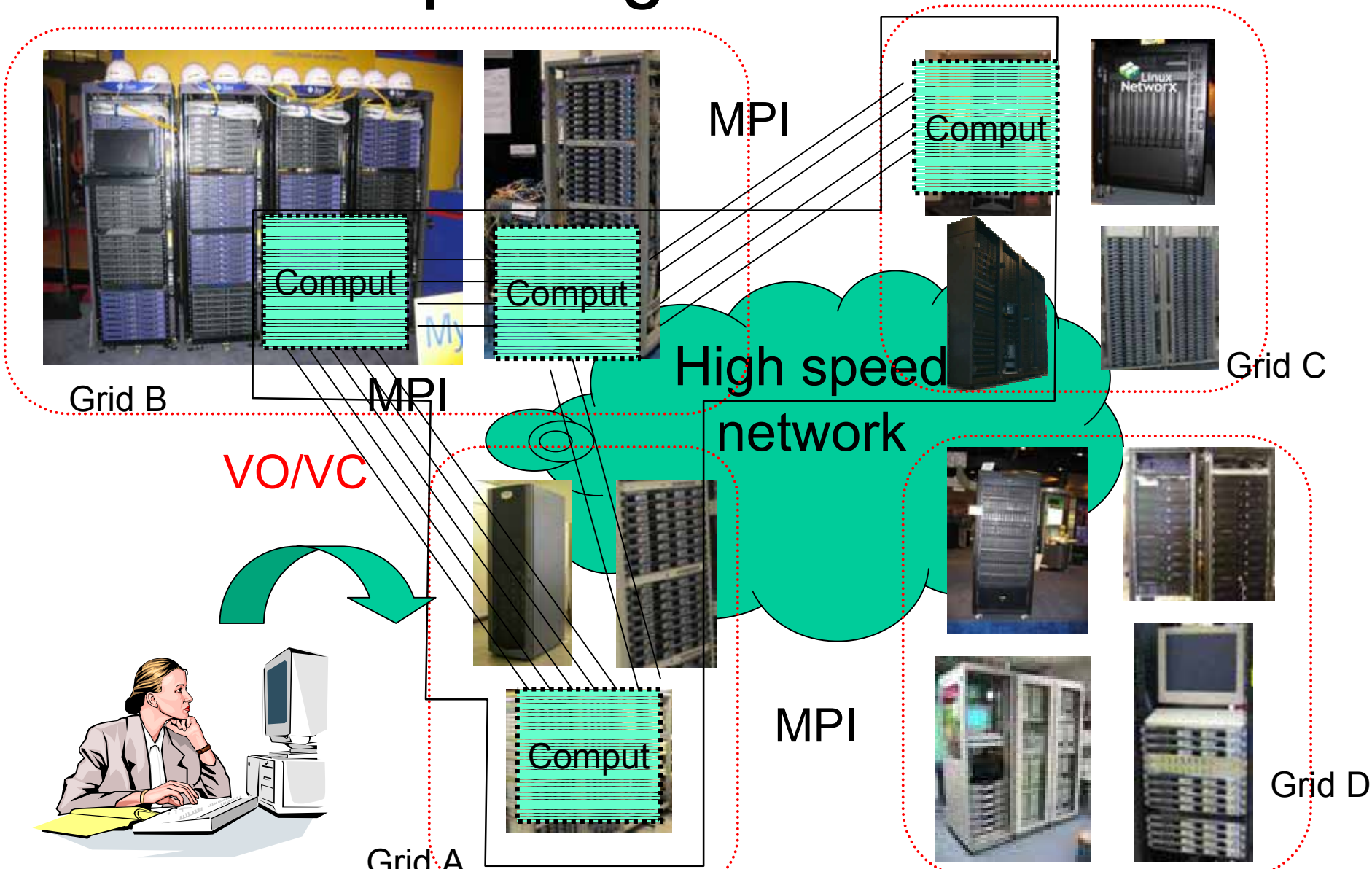
Scenario A: Develop at local, Production in the Grid



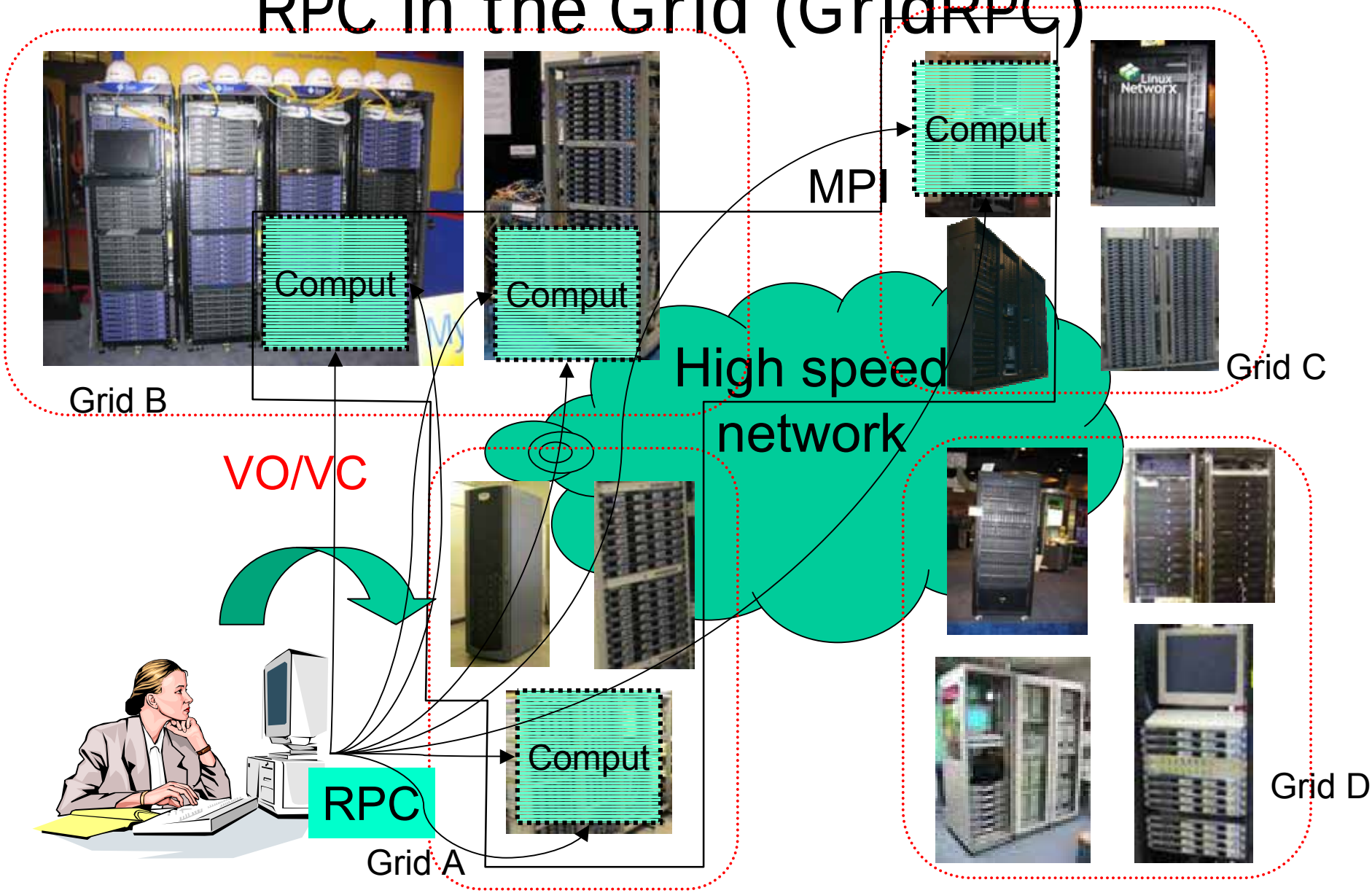
Scenario B: Distributed computing in the Grid



Scenario C: Full scale computing in the Grid



Scenario D: More flexibility with RPC in the Grid (GridRPC)



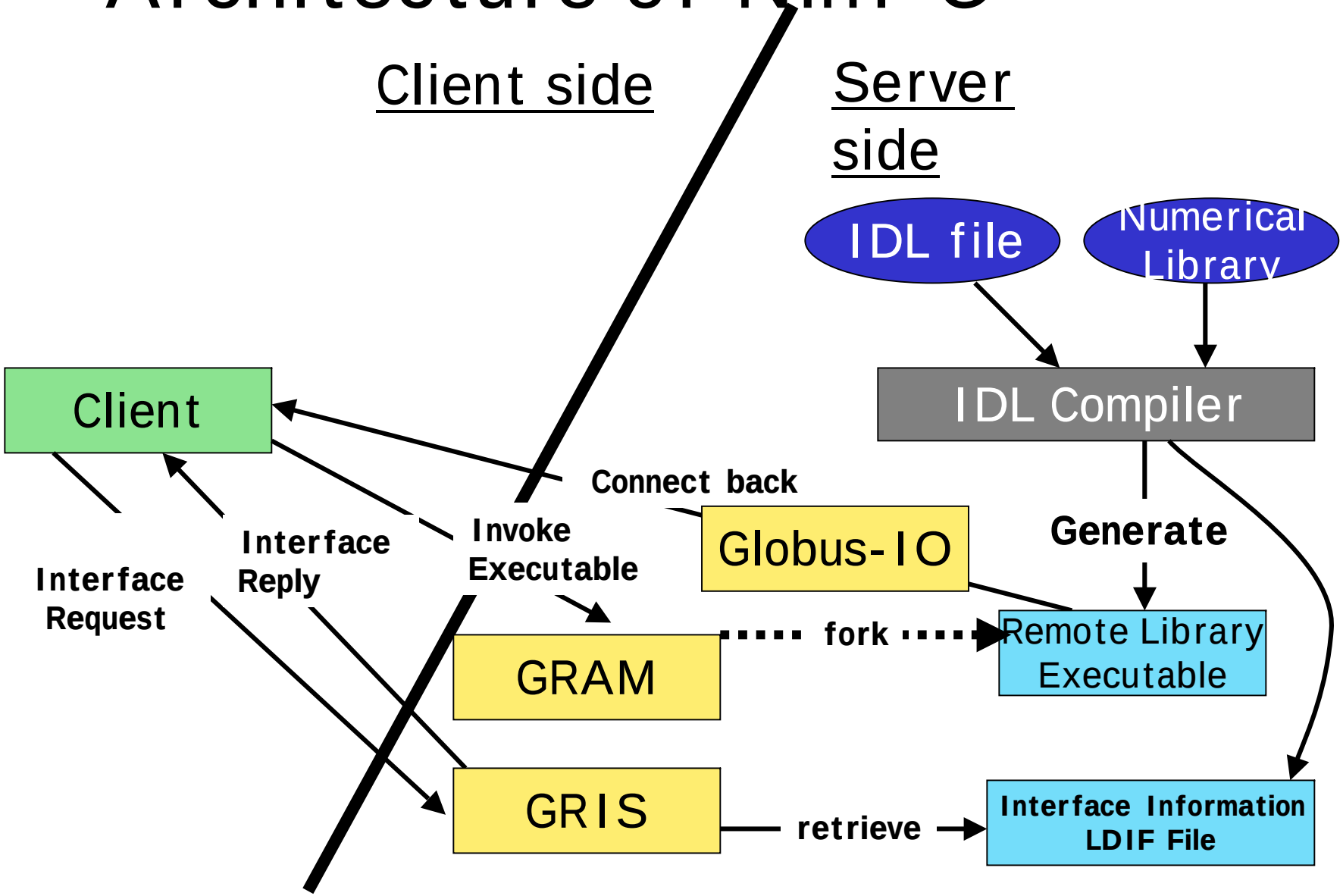
What is Ninf-G GridRPC?

- A software package which supports programming and execution of Grid applications using GridRPC.
- Ninf-G includes
  - C/C++, Java APIs, libraries for software development
  - IDL compiler for stub generation
  - Shell scripts to
    - compile client program
    - build and publish remote libraries
  - sample programs
  - manual documents

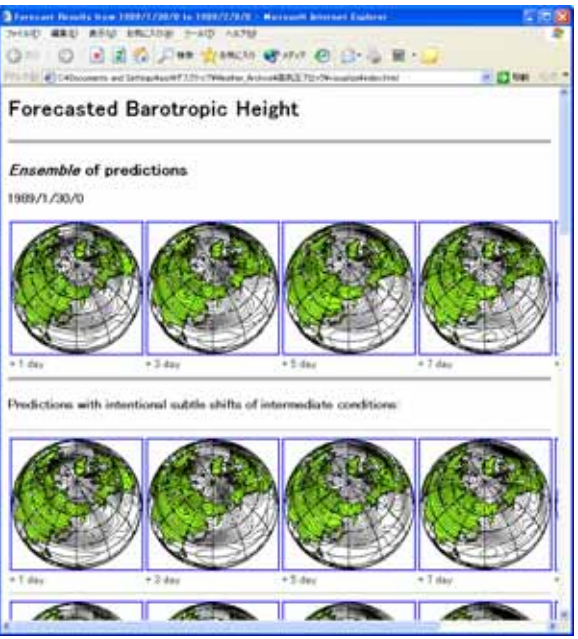
Ninf-G: Features At-a-Glance

- Ease-of-use, client-server, Numerical-oriented RPC system
- No stub information at the client side
- User's view: ordinary software library
  - Asymmetric client vs. server
- Built on top of the Globus Toolkit
  - Uses GSI, GRAM, MDS, GASS, and Globus-IO
- Supports various platforms
  - Ninf-G is available on Globus-enabled platforms
- Client APIs: C/C++, Java

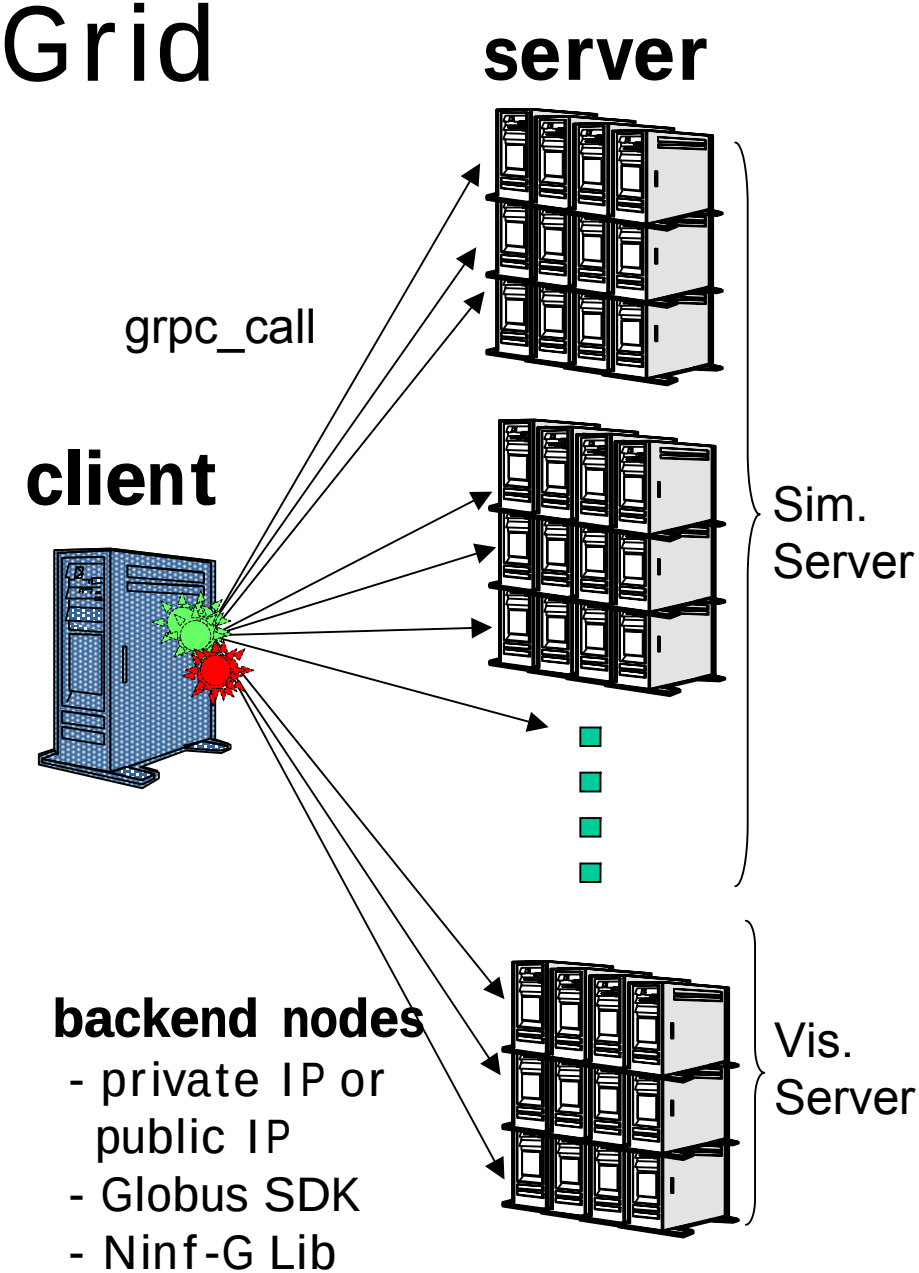
Architecture of Ninf-G



Climate Simulation using GridRPC over APGrid



Frontend node
- public IP
- Globus
  - gatekeeper
  - jobmanager
    - pbs, grd, sqms
- NAT

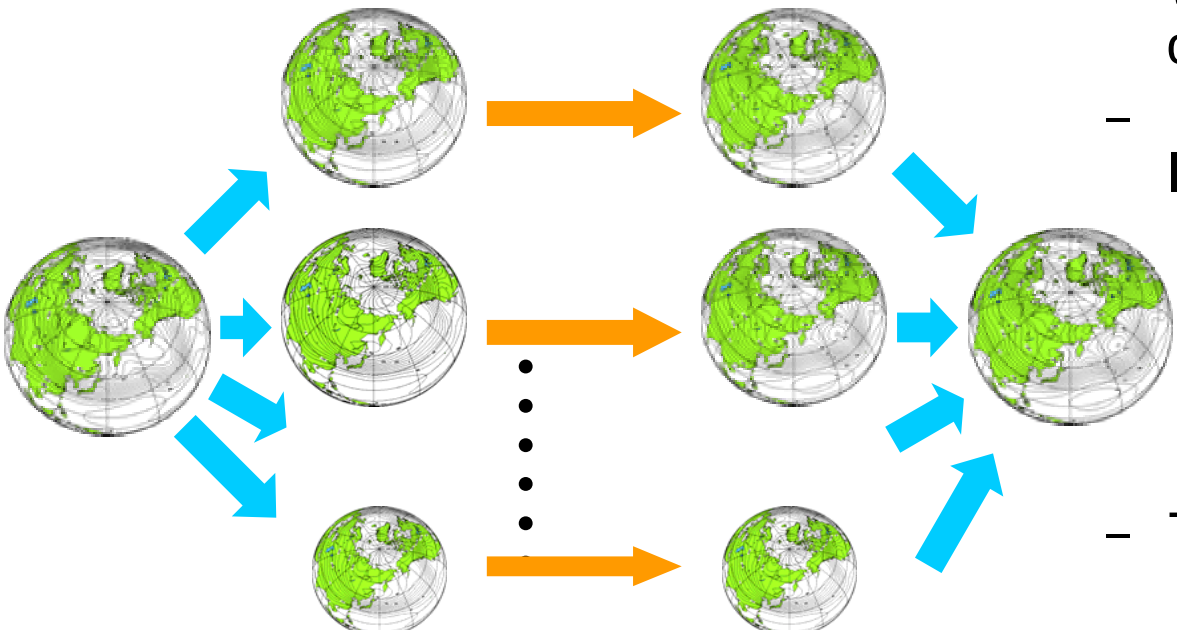


backend nodes
- private IP or public IP
- Globus SDK
- Ninf-G Lib

Sequential Run: 8000 sec
Execution on Grid: 300 sec (100cpu)

Application: Climate Simulation

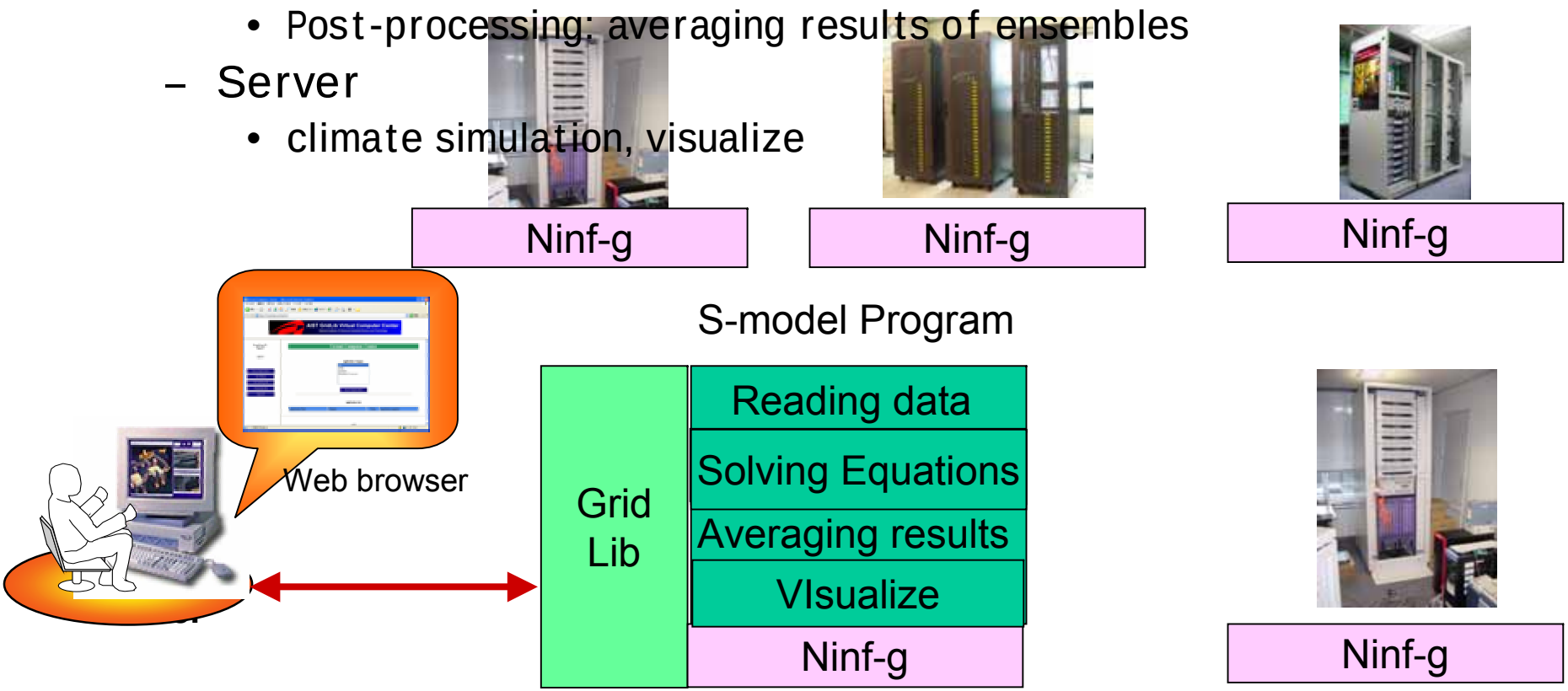
- Goal
  - Short- to Middle- term, global climate simulation
    - Winding of Jet-Stream
    - Blocking phenomenon of high atmospheric pressure



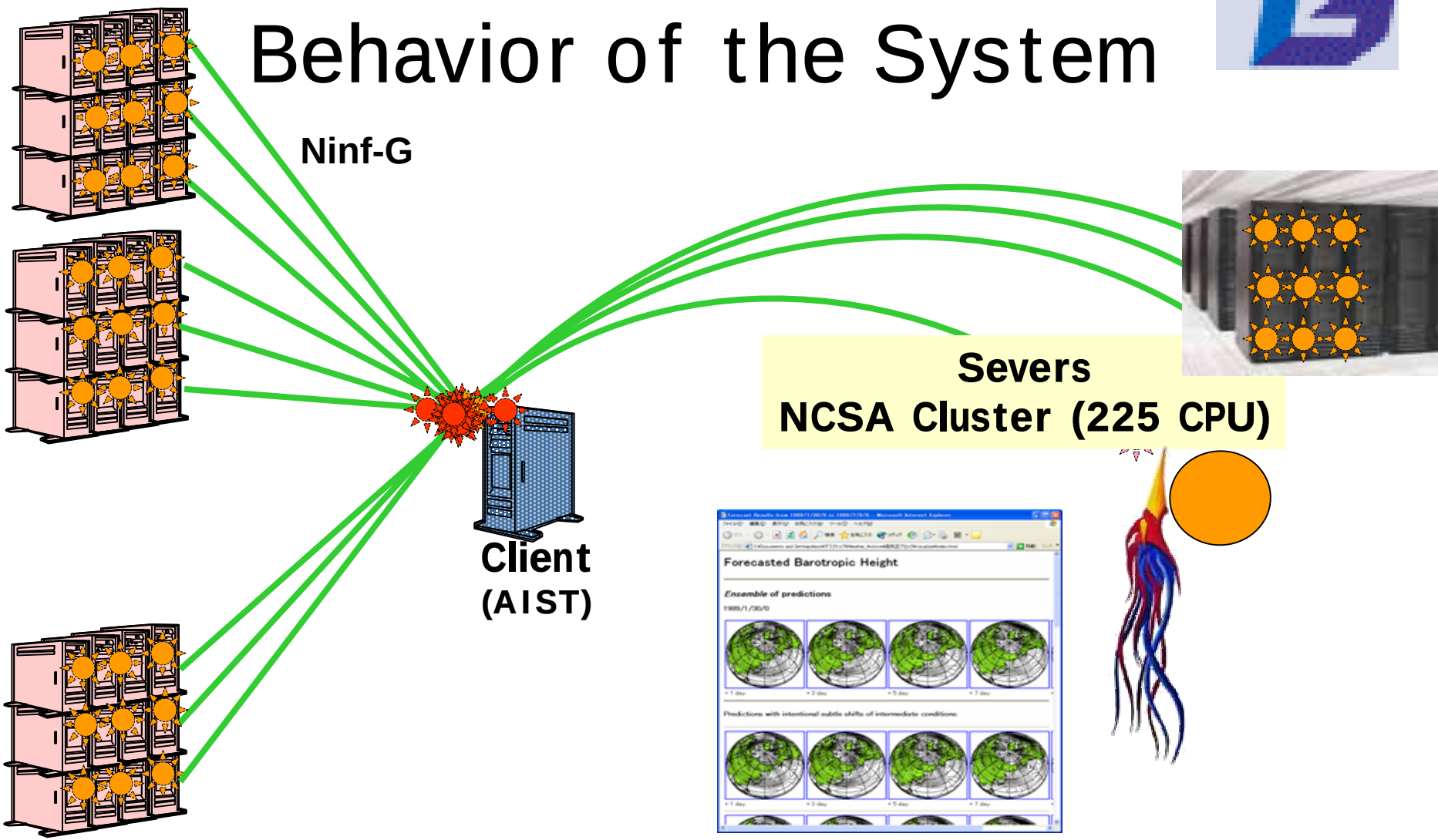
- Barotropic S-Model
  - Climate simulation model proposed by Prof. Tanaka (U. of Tsukuba)
  - Simple and precise
  - Modeling complicated 3D turbulence as a horizontal one
  - Keep high precision over long periods
    - Taking a statistical ensemble mean
      - ~ several 100 simulations
    - Introducing perturbation at every time step
  - Typical parameter survey

“Ninfy” the original (seq.) climate simulation

- Dividing a program into two parts as a client-server system
  - Client:
    - Pre-processing: reading input data
    - Post-processing: averaging results of ensembles
  - Server
    - climate simulation, visualize



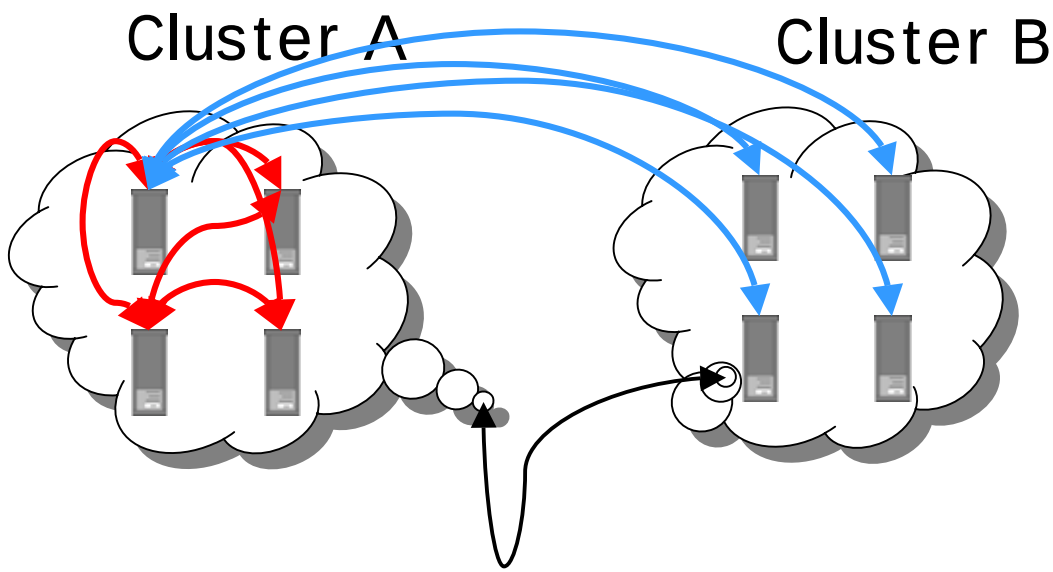
Behavior of the System



Preliminary Evaluation

- Testbed: 500 CPU
  - TeraGrid: 225 CPU (NCSA)
  - ApGrid: 275 CPU (AIST, TITECH, KISTI)
- Ran 1000 Simulations
  - 1 simulation = 20 seconds
  - 1000 simulation = 20000 seconds = 5.5 hour (if runs on a single PC)
- Results
  - 150 seconds = 2.5 min
- Insights
  - Ninf-G2 efficiently works on large-scale cluster of cluster
  - Ninf-G2 provides good performance for fine grain task-parallel applications on large-scale Grid.

Grid MPI

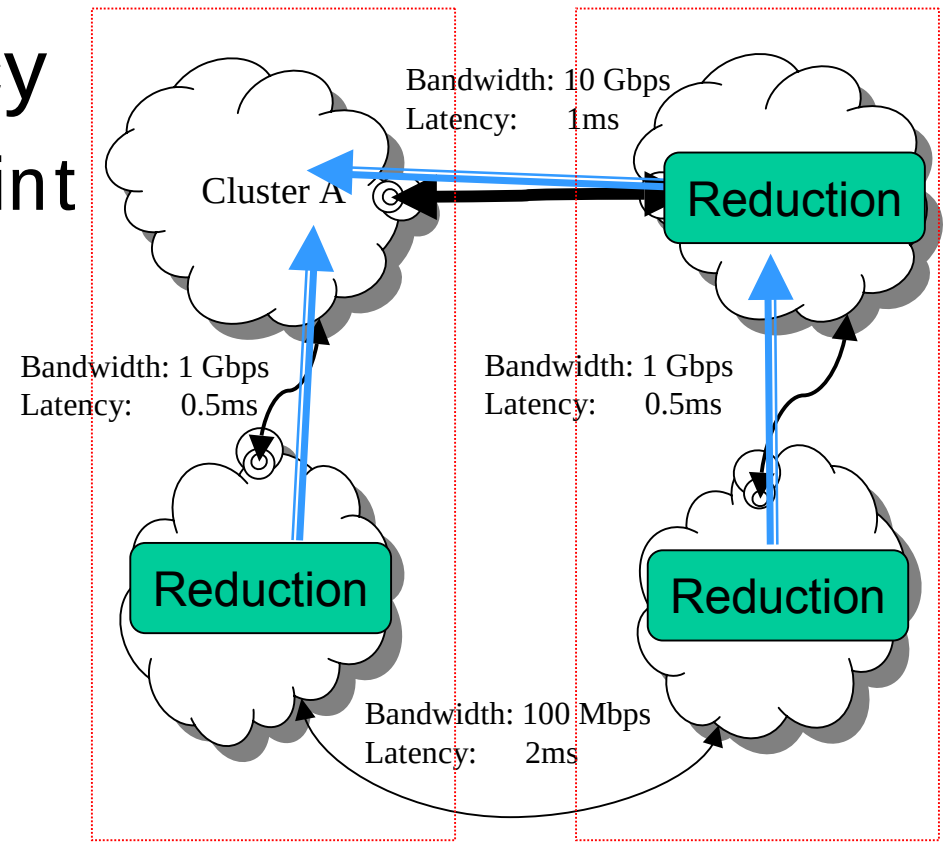


- Intra-cluster
  - Vendor MPI
  - SCore
  - IMPI compatible
- Inter-cluster
  - GRAM (Globus) & IMPI compatible

<table>
<tr><td colspan="9">MPI Core</td></tr>
<tr><td colspan="4">RPIM</td><td rowspan="3">IMPI</td><td colspan="4">Grid ADI</td></tr>
<tr><td rowspan="3">SSH</td><td rowspan="3">RSH</td><td rowspan="3">GRAM</td><td rowspan="3">Vendor MPI</td><td colspan="3">Latency-aware Communication Topology</td><td rowspan="3">Other Comm. Library</td></tr>
<tr><td colspan="2">P-to-P Communication</td><td rowspan="2">Vendor MPI</td></tr>
<tr><td colspan="2">TCP/IP</td><td>PMv2</td><td>Others</td></tr>
</table>

LACT (Latency-Aware Communication Topology)

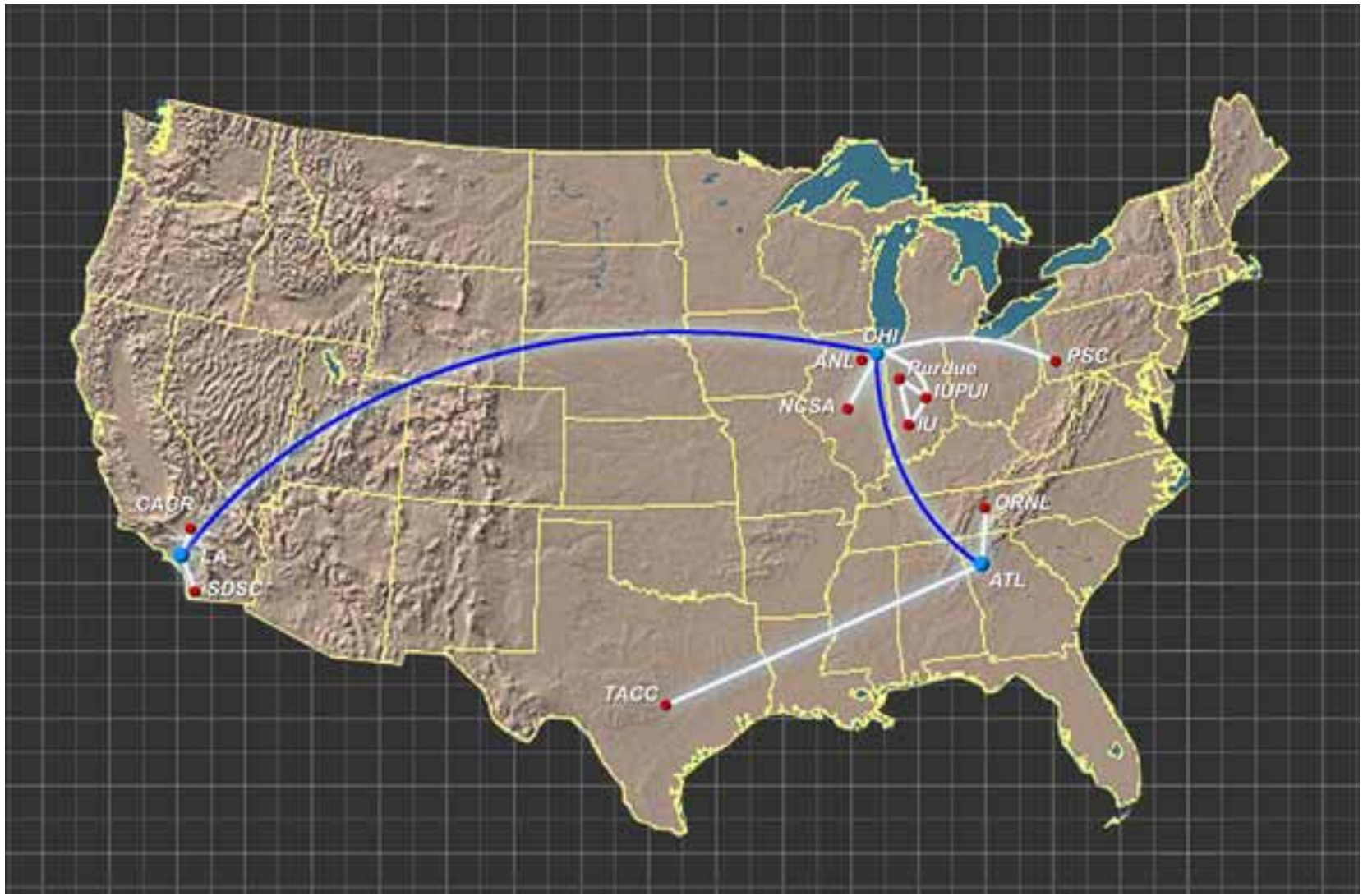
- Bandwidth and Latency
  - Routing of Point-to-Point message
    - Based on routing table
    - Message forwarding at intermediate node
  - Routing of collective communications
    - Communication pattern adapted to the network topology



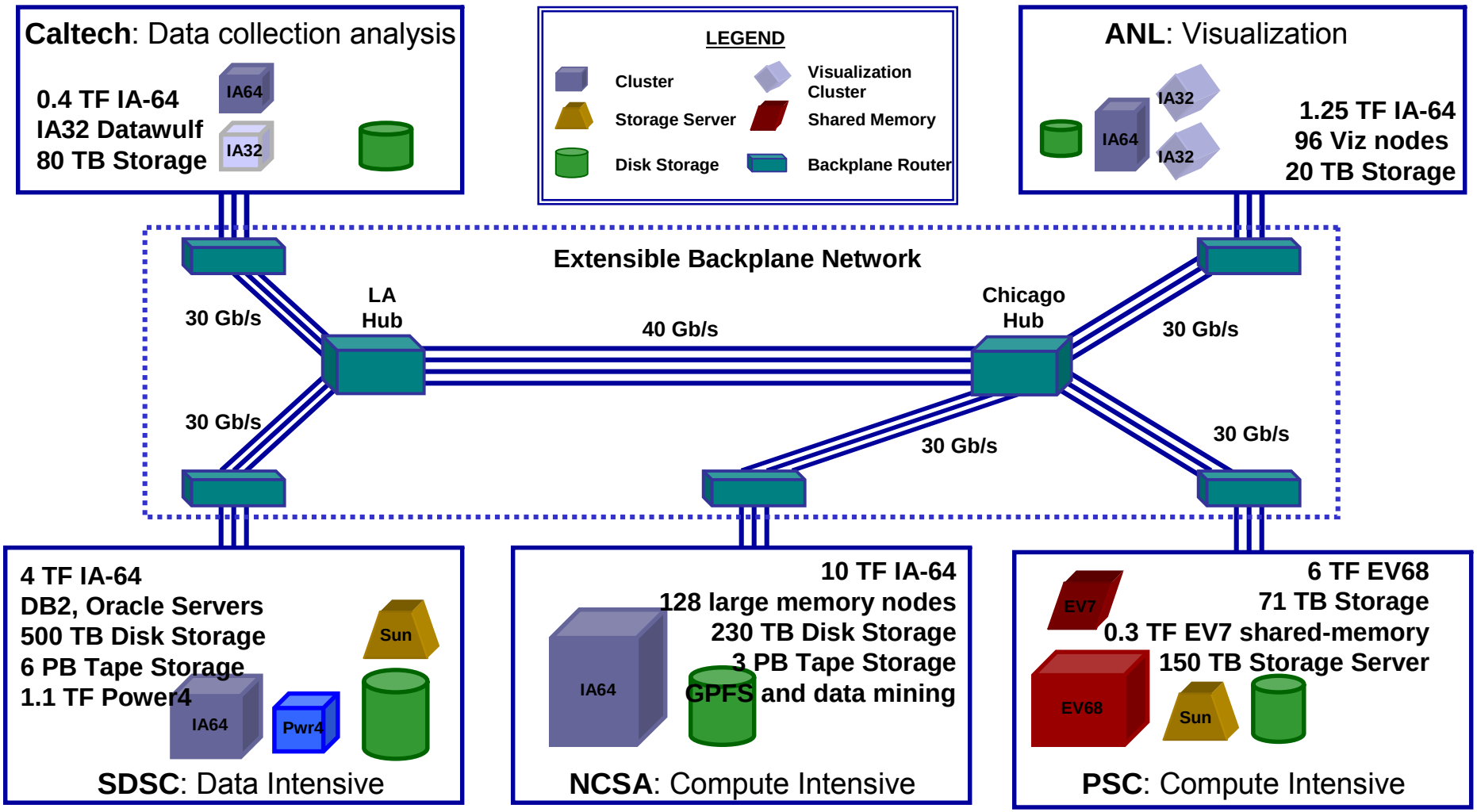
Example Reduction

Large-Scale International Grid Projects and Collaborative Efforts Based on Clusters

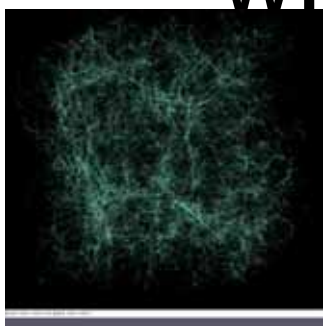
NSF TeraGrid: Extensible Terascale Facility 2004



TeraGrid: Resource Providers (+ 4 New Sites)

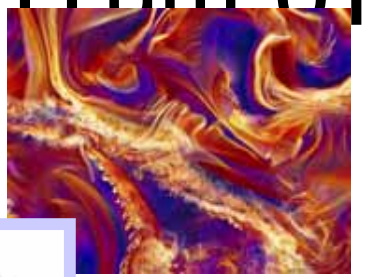


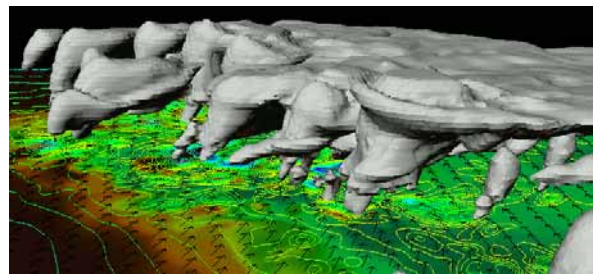
TeraGrid Applications span a wide spectrum of Science



ENZO
(Astrophysics)

PPM
(Astrophysics)

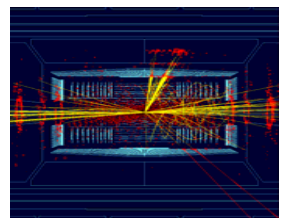




MEAD (Atmospheric Sciences)

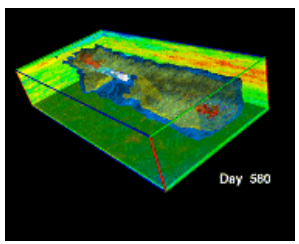
GridSAT
(Computer Science)

AtlasMaker
(Astronomy)

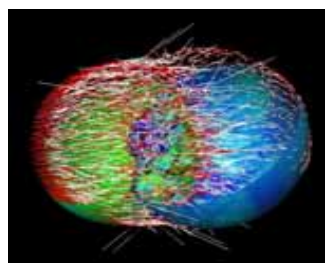


CMS/ GriPhyN
(Physics)

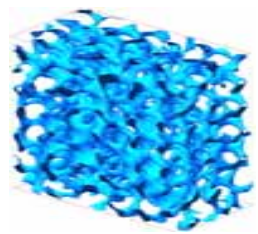
GAFEM
(Ground-water modeling)



Encyclopedia of Life
(Biosciences)



VTF
(Shock Physics)



TeraGyroid
(Condensed Matter Physics)

BioCoRE
Biological Collaborative Environment
(Biomedicine)

EGEE:
Enabling Grids for E-science in Europe

Goal

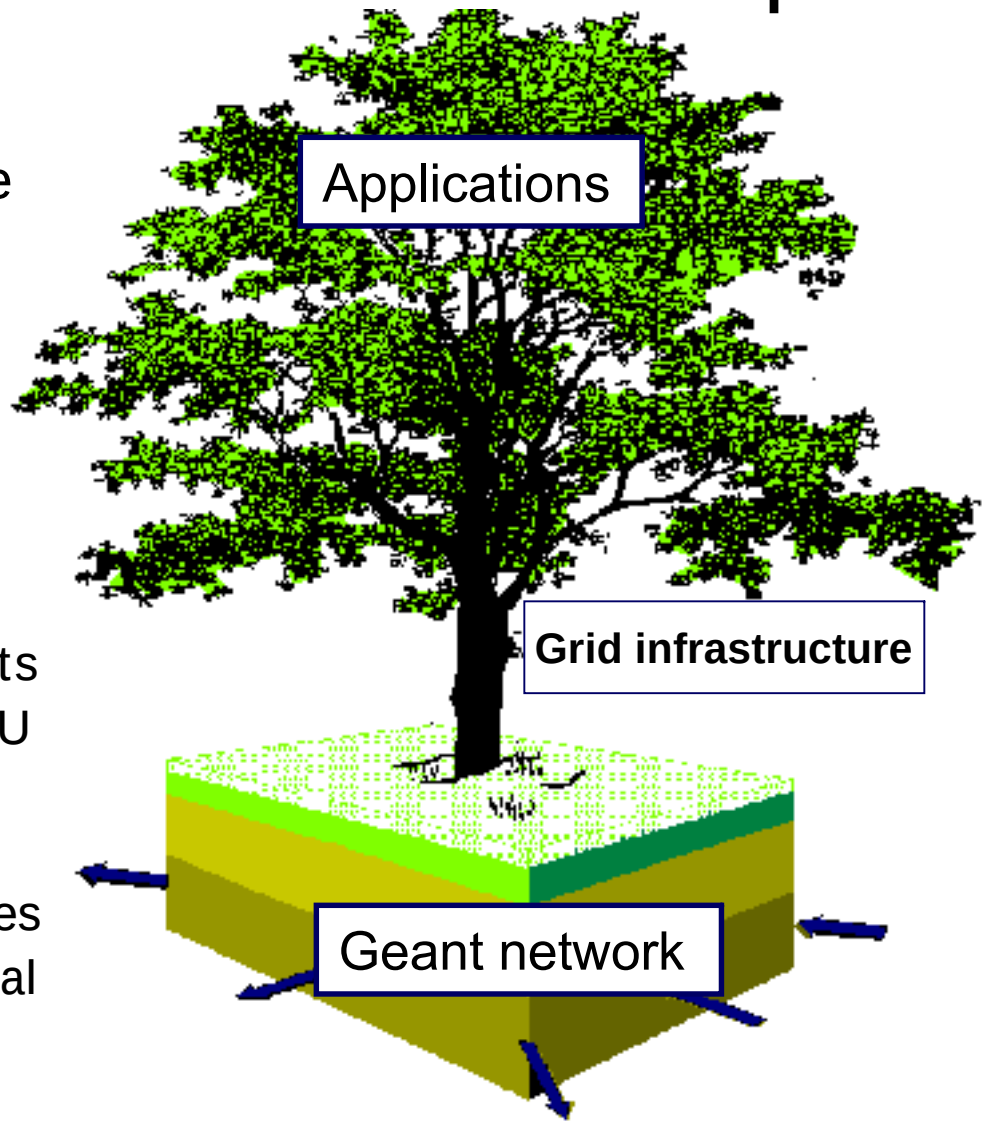
- Create a wide European Grid production quality infrastructure on top of present and future EU RN

Build on

- EU and EU member states major investments in Grid Technology
- International connections (US and AP)
- Several pioneering prototype results
- Large Grid development teams in EU

Approach

- Leverage current and planned national and regional Grid programmes
- Work closely with relevant industrial Grid developers, NRENs and US-AP projects



Slide **courtesy: EGEE Consortium**

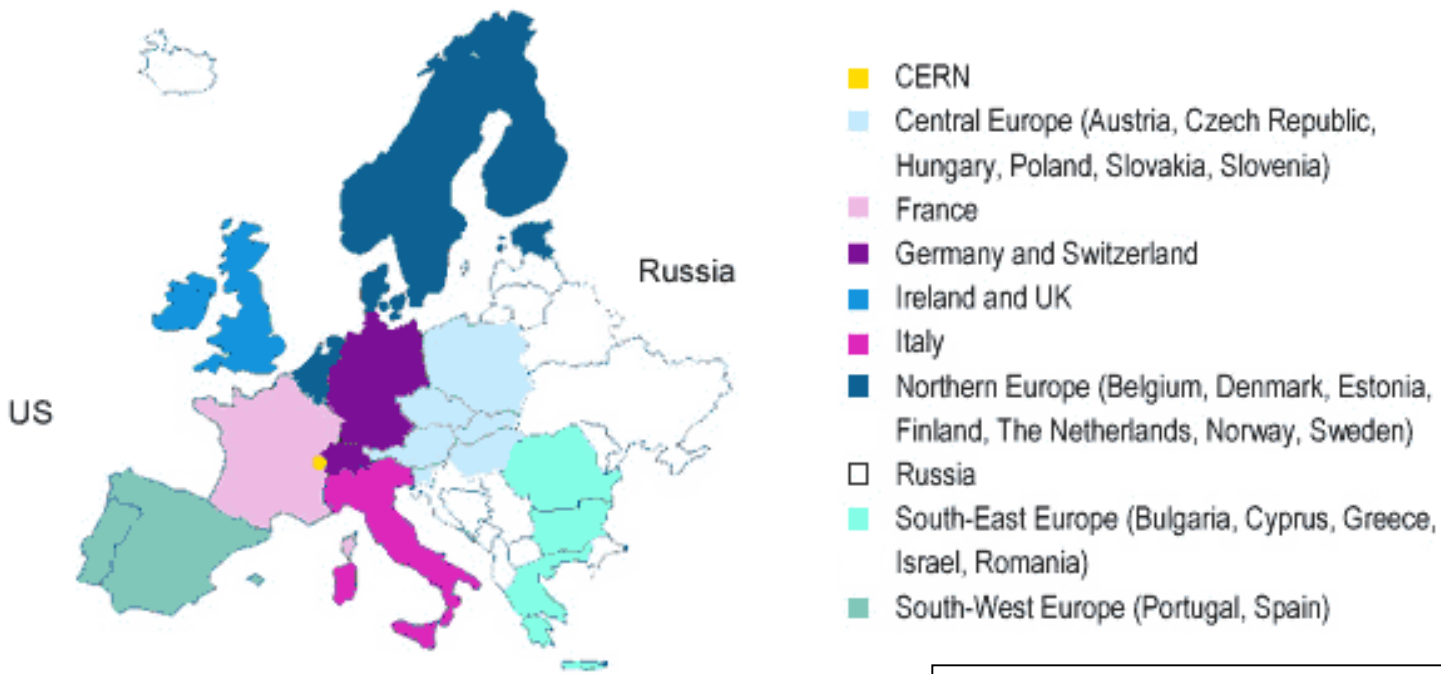
EGEE: Why?

- Access to a production quality grid will change the way science and business is done in Europe
- Benefits for several application fields:
  - Bioinformatics *(study of the human genome and proteome to understand genetic diseases)*
  - Engineering *(design optimization, simulation, failure analysis and remote Instrument access and control)*
  - Medical/Healthcare (*imaging, diagnosis and treatment)*
  - Natural Resources and the Environment *(weather forecasting, earth observation, modeling and prediction of complex systems)*

Slide **courtesy: EGEE Consortium**

EGEE: Partners

- Leverage national resources in a more effective way for broader European benefit
- 70 leading institutions in 27 countries, federated in regional Grids



Picture **courtesy: EGEE Consortium**

OSAKA UNIVERSITY

TransPAC

ISDL 同志社

NCSA

東京大學

USM UNIVERSITI SAINS MALAYSIA

AIST

PRAGMA

Bioinformatics Institute A*STAR

ApGrid / PRAGMA Testbed

- 10 countries
- 21 organizations
- 22 clusters
- 853 CPUs

KISTI www.kisti.re.kr

DSTC

NCHC

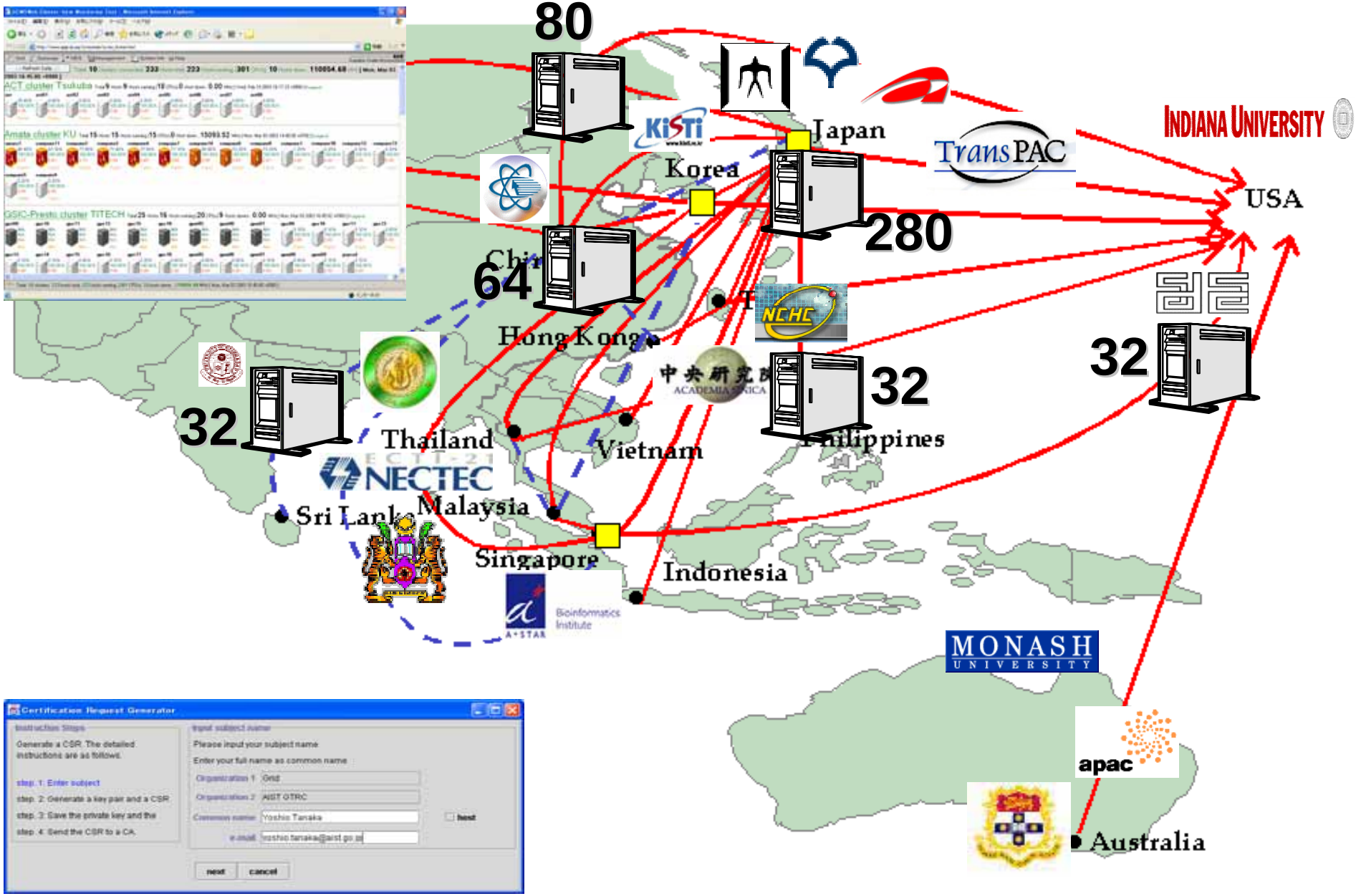
apgrid Asia-Pacific Grid

The University of Hong Kong

APAN Asia-Pacific Advanced Network

The ApGrid Testbed

http://www.apgrid.org/



APGrid Resources used in the SC02 experiment

- KOUME Cluster (AIST)
  - Client
- UME Cluster (AIST)
  - jobmanager-grd, (40cpu + 20cpu)
  - AIST GTRC CA
- AMATA Cluster (KU)
  - jobmanager-sqms, 6cpu
  - AIST GTRC CA
- Galley Cluster (Doshisha U.)
  - jobmanager-pbs, 10cpu
  - Globus CA
- Gideon Cluster (HKU)
  - jobmanager-pbs, 15cpu
  - HKU CA
- PRESTO Cluster (TITECH)
  - jobmanager-fork, 4cpu
  - TITECH CA
- VENUS Cluster (KISTI)
  - jobmanager-pbs, 60cpu
  - KISTI CA
- ASE Cluster (NCHC)
  - jobmanager-fork, 8cpu
  - NCHC CA
- Handai Cluster (Osaka U)
  - jobmanager-pbs, 20cpu
  - Osaka CA

- Total: 183

Grid Standardization Activities
The Global Grid Forum & The Japanese Grid Consortium

http://www.ggf.org http://www.jpgrid.org

- GGF: International Grid Standardization and R&D Forum
  - Established 1999 as Grid Forum (Modeled after IETF)
  - 2002H, merged with European eGrid Forum ➔ The Global GF
    - Chair: Dr. Charlie Catlett (Argonne NL)
    - Over 60 Working Groups/Research Groups, Workshops
    - 3 times/year GGF meeting
    - The speaker (Matsuoka) is an "Area Director" & member of the Steering Group
    - 2002, Absorbed NPI (New Productivity Initiative) and the Intel P2P Consortium
  - 2003/3/4-7 Tokyo, > 800 participants
  - GGF10 March 2004 Berlin, GGF11 June 2004 Hawaii, …
- Grid Consortium (Japan): Largely industrial consortium for commecial applications of Grids
  - Established 2002, > 30 Coorporate members
  - GGF reports, lectures, tutorials, GGF franchise

Grid Consortium Japan
グリッド協議会

Grid Services into a Component Model-the OGSA Story

Following Slides Courtesy of:

Dennis Gannon (Indiana University)

Steve Tuecke (Argonne National Lab)

Web Services

- A distributed computing infrastructure based on XML and some basic Web Ideas
  - A web service is a server process that exposes typed ports to the network
  - Described by the Web Services Definition Language. An XML document that contains
    - The type of messages the service understands and the types of responses and exceptions it returns
    - The associated “methods” are bound together as “port types”
    - Port types are bound to protocols and endpoints as “ports”
  - A WSDL document completely defines a service and how to access it.

What is OGSI in 1 slide

- “Open Grid Services Interface”
- It is a “component model” for web services.
- It defines a set of behavior patterns that each OGSI service must exhibit.
- Every “Grid Service” port type extends a common base type.
  - Defines an introspection model for the service
  - You can query it (in a standard way) to discover
    - What methods/messages a port understands
    - What other port types does the service provide?
    - If the service is “stateful” what is the current state?

On-Demand e-Business Usage Scenario Observation

- No business will depend on an IT outsourcing infrastructure that is proprietary, and may go South.
- E.g., the Internet, Web hosting Datacenter, etc., work because there are standardized services offered by multiple service providers
  - Web Services may succeed for the same reason
- Will OGSA achieve this level of commonality and necessity?

OGSI/A = The 3rd (4th?) Generation Distributed OO

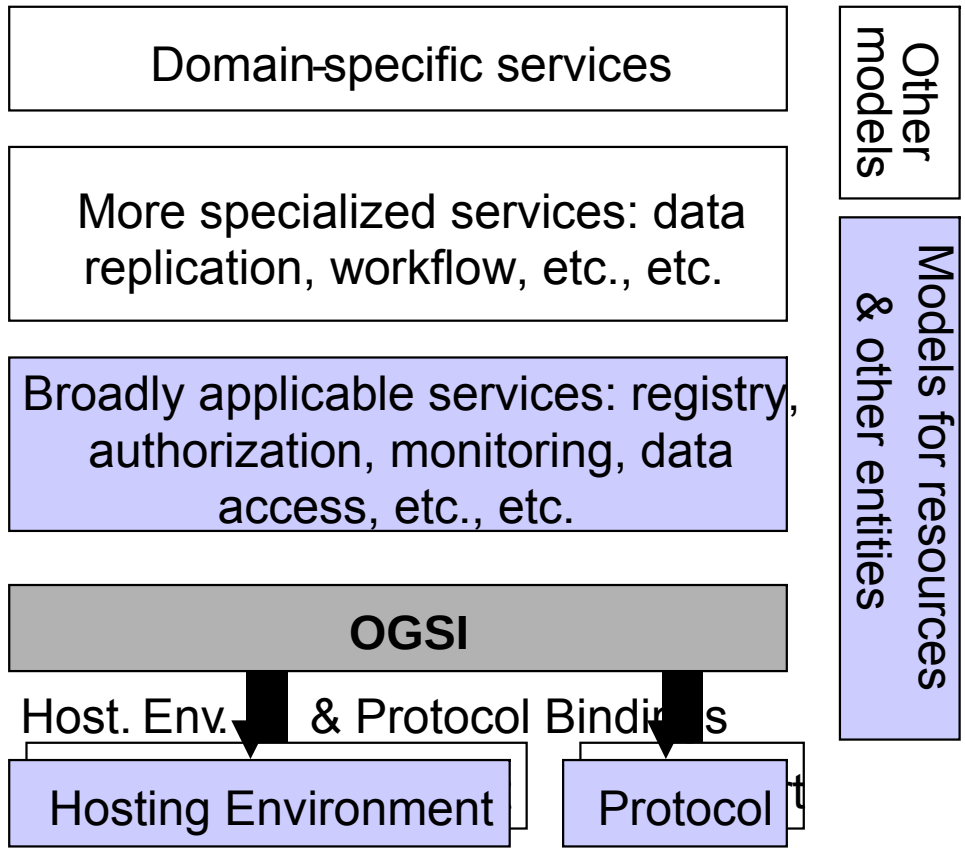
1980	Distributed OS, Parallel/Distributed Prog. Languages	(Distributed) OO-OS, Distributed OO Prog. Languages, etc.
1990	RPC-based Systems, e.g., SunDCE	CORBA ORB
	Java, COM	Java RMI, DCOM, Servelets
2000	Web Services	OGSI/OGSA

Open Grid Service Architecture (OGSA)

- OGSA-WG in Global Grid Forum chaired by
  - Ian Foster, ANL and Univ. of Chicago
  - Jeff Nick, IBM
  - Dennis Gannon, IU
- Active Members from
  - IBM, Fujitsu, NEC, SUN, Hitachi, Avaki
  - Univ. of Mich, Chicago, Indiana

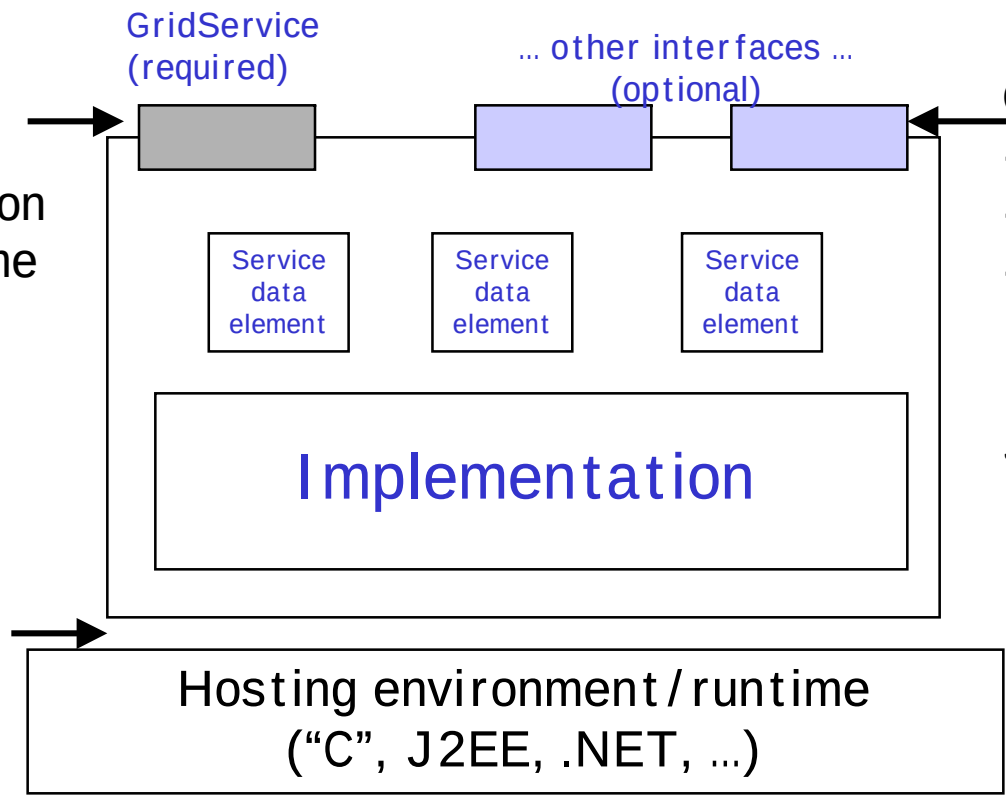
OGSA Top Level View

- OGSA is the set of "core" Grid services
  - Stuff you can't live without
  - If you built a Grid you would need to invent these things



The Grid Service = Interfaces/Behaviors + Service Data

Required:
- Introspection (service data)
- Explicit destruction
- Soft-state lifetime

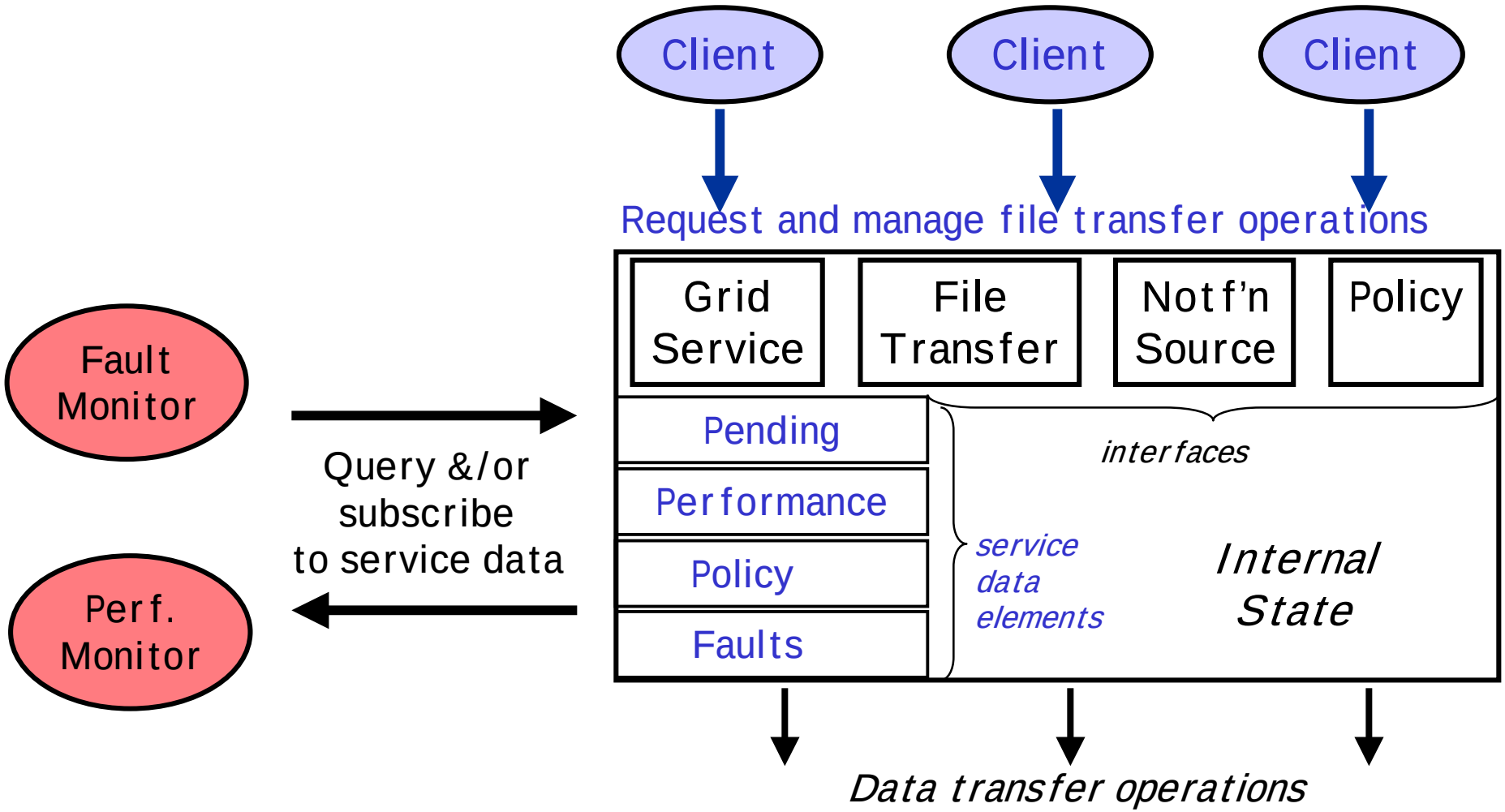


Optional:
- Service creation
- Notification
- Service group

+ application-specific interfaces

Binding properties:
- Authentication
- Reliable invocation
- Transactions
- QoS

Example: File Transfer Service



Core Services

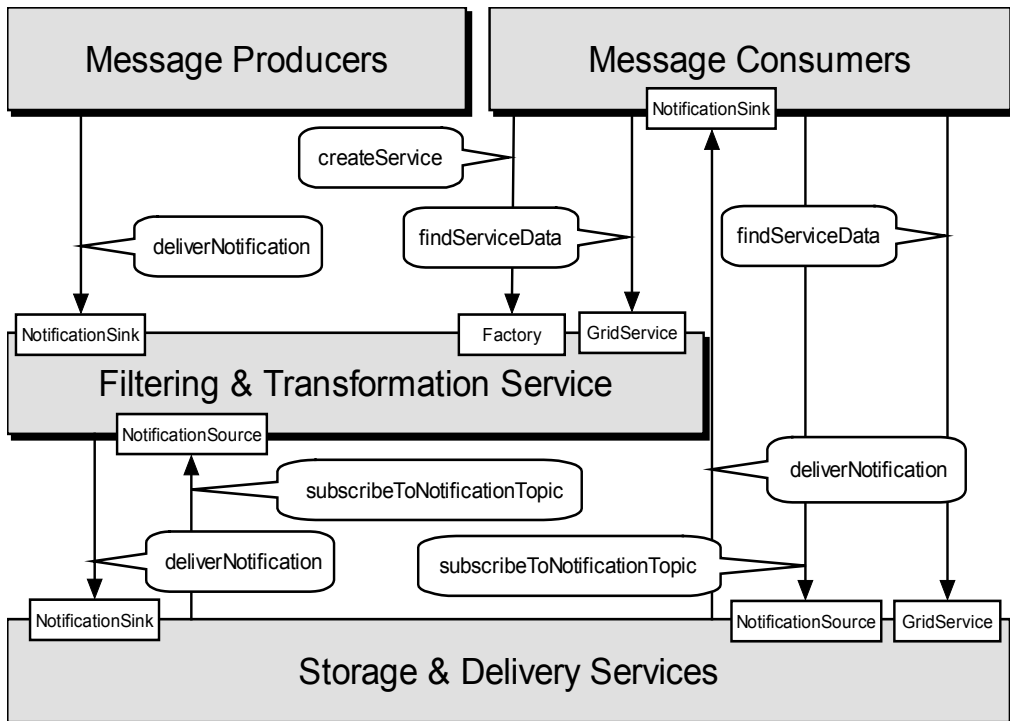
- Registries, and namespace bindings
  - Registry is a collection of services indexed by service metadata.
    - “find me a service with property X.”
  - Directory is a map from a name space to GSHs.
- Queues
  - For building schedulers and resource brokers
  - (not clear)

Security

- Authentication
  - 2-way. Who are you and who am I?
- Authorization
  - What am I authorized to use/see/modify
- Accounting/Billing
  - (not really security – see monitoring)
- Privacy
- Group Access
  - Easily create a group to share access to a virtual Grid.
- Very complex issues related to services and message delivery.

Messaging, Events, Logging

- Messaging
  - Delivery Model
    - Queuing and Pub/Sub message delivery
- Events
  - Standard XML schemas
- Standard Logging

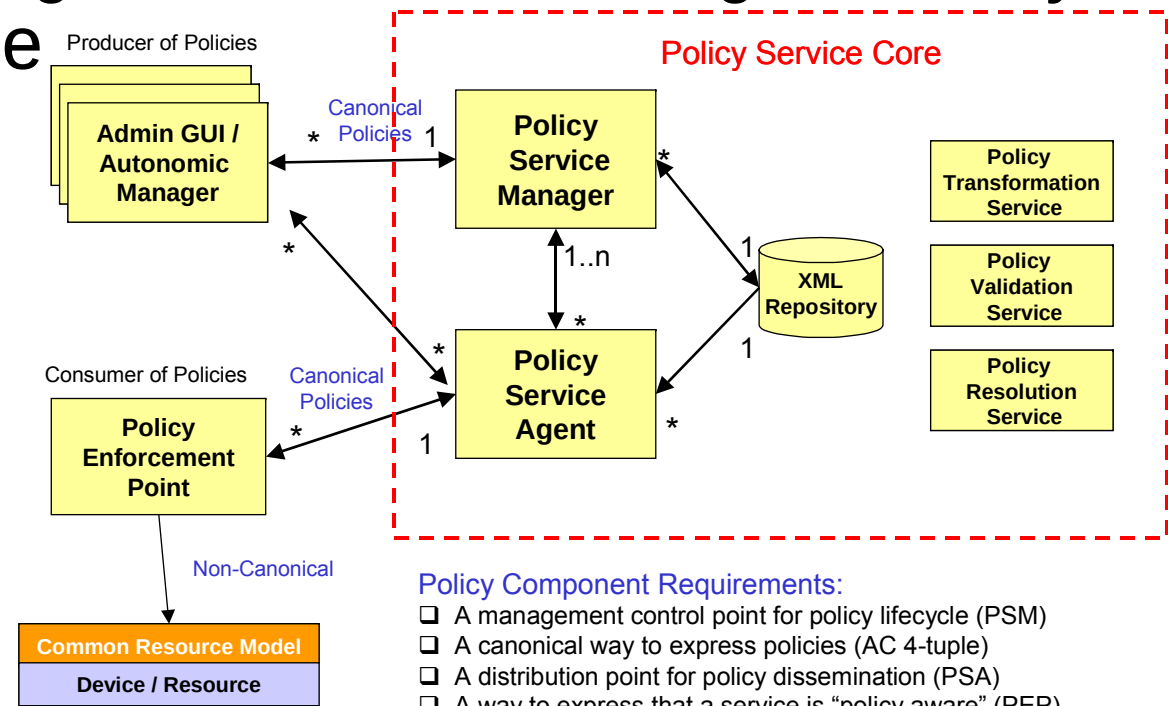


Common Resource Model

- Every resource on the grid that is manageable is represented by a service instance
  - CRM is the Schema hierarchy that defines each resource
  - Service for a resource presents its management interface to authorized parties.

Policy Management

- Policy management services
  - Mechanism to publish policy and the services it applies to.
  - Policy life-cycle mgmt.
- Policy languages exist for routing, security, resource use



Policy Component Requirements:

- ❑ A management control point for policy lifecycle (PSM)
- ❑ A canonical way to express policies (AC 4-tuple)
- ❑ A distribution point for policy dissemination (PSA)
- ❑ A way to express that a service is “policy aware” (PEP)
- ❑ A way to effect change on a resource (CRM)

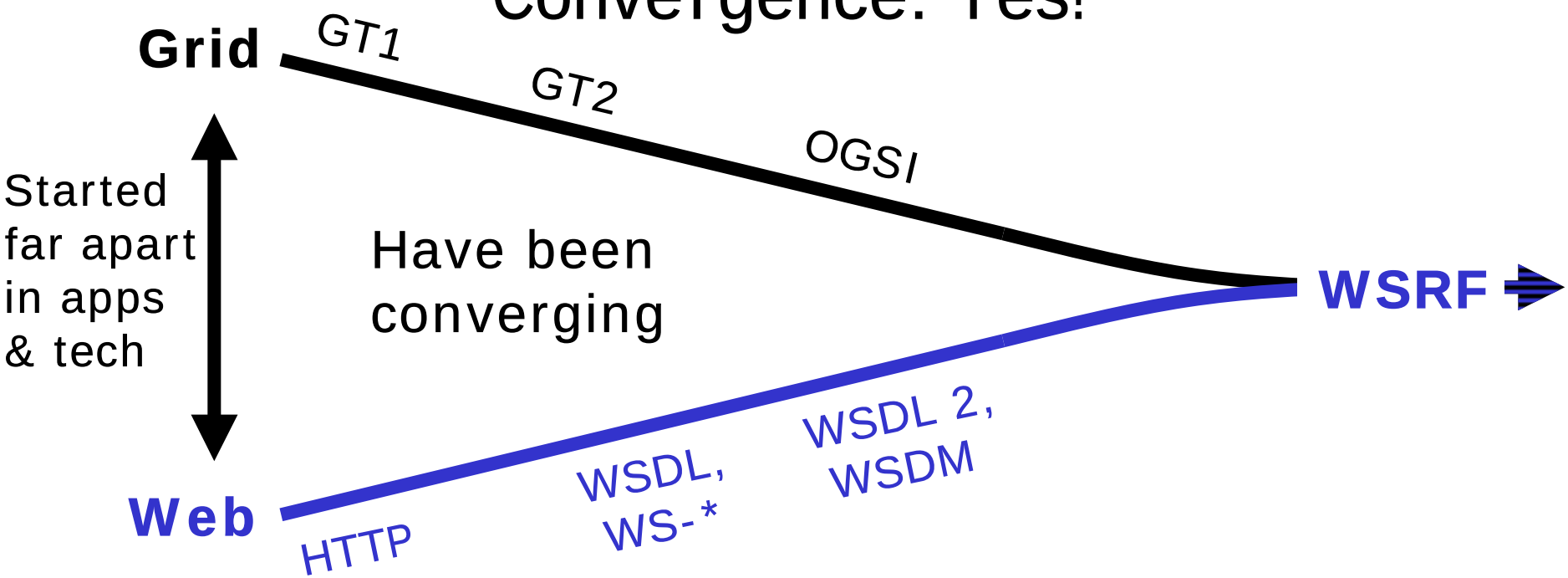
Grid Service Orchestration

- Creating new services by composing other services
- Two types of Orchestration
  - Composition in space
    - One services is directly invoking another
  - Composition in time
    - Managing the workflow
      - First do this.
      - Then do this and that
      - When that is done do this
        - » If something goes wrong do this
      - And so on…

Data Services

- Distributed Data Access
- Data Caching
- Data Replication Services
- Metadata Catalog Services
- Storage Services

From OGSI to Web Services Resource Framework (WSRF)
Grid and Web Services: Convergence: Yes!



The definition of WSRF means that Grid and Web communities can move forward on a common base

OGSI Concerns Addressed with WSRF

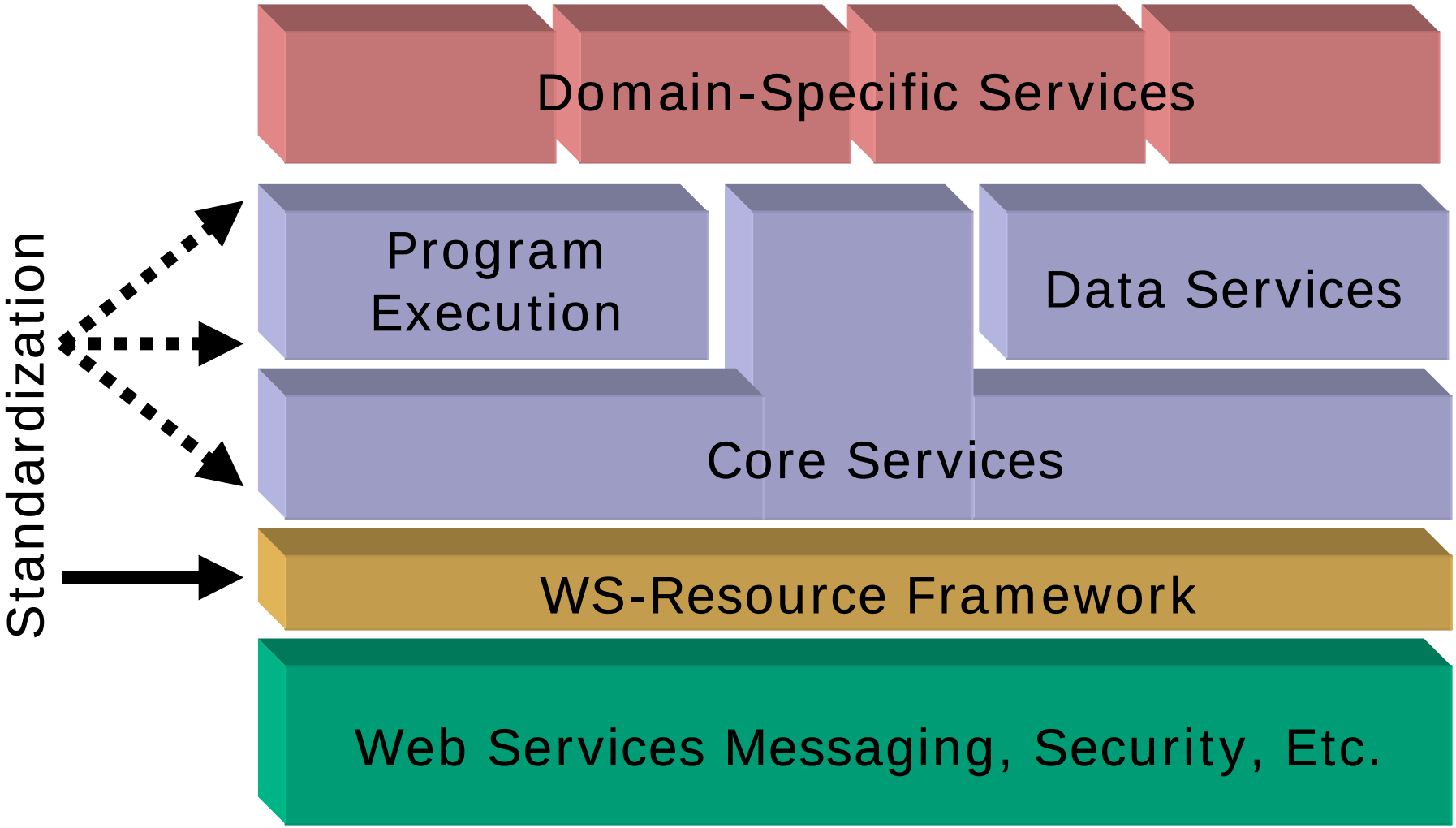
- Too much stuff in one specification
  - ➔WSRF partitions OGSI v1.0 functionality into a family of composable specifications
- Does not work well with existing Web services tooling
  - ➔WSRF tones down the usage of XML Schema
- Too object oriented
  - ➔WSRF makes an explicit distinction between the “service” and the stateful “resources” acted upon by that service

From OGSI to WSRF: Refactoring and Evolution**

OGSI	WSRF
Grid Service Reference	*WS-Addressing* Endpoint Reference
Grid Service Handle	*WS-Addressing* Endpoint Reference
HandleResolver portType	**WS-RenewableReferences**
Service data defn & access	**WS-ResourceProperties**
GridService lifetime mgmt	**WS-ResourceLifeCycle**
Notification portTypes	**WS-Notification**
Factory portType	Treated as a pattern
ServiceGroup portTypes	**WS-ServiceGroup**
Base fault type	**WS-BaseFaults**

Draft document at www.globus.org/wsrf this week

Open Grid Services Architecture with WSRF



National Research Grid Initiative (NAREGI)

Satoshi Matsuoka

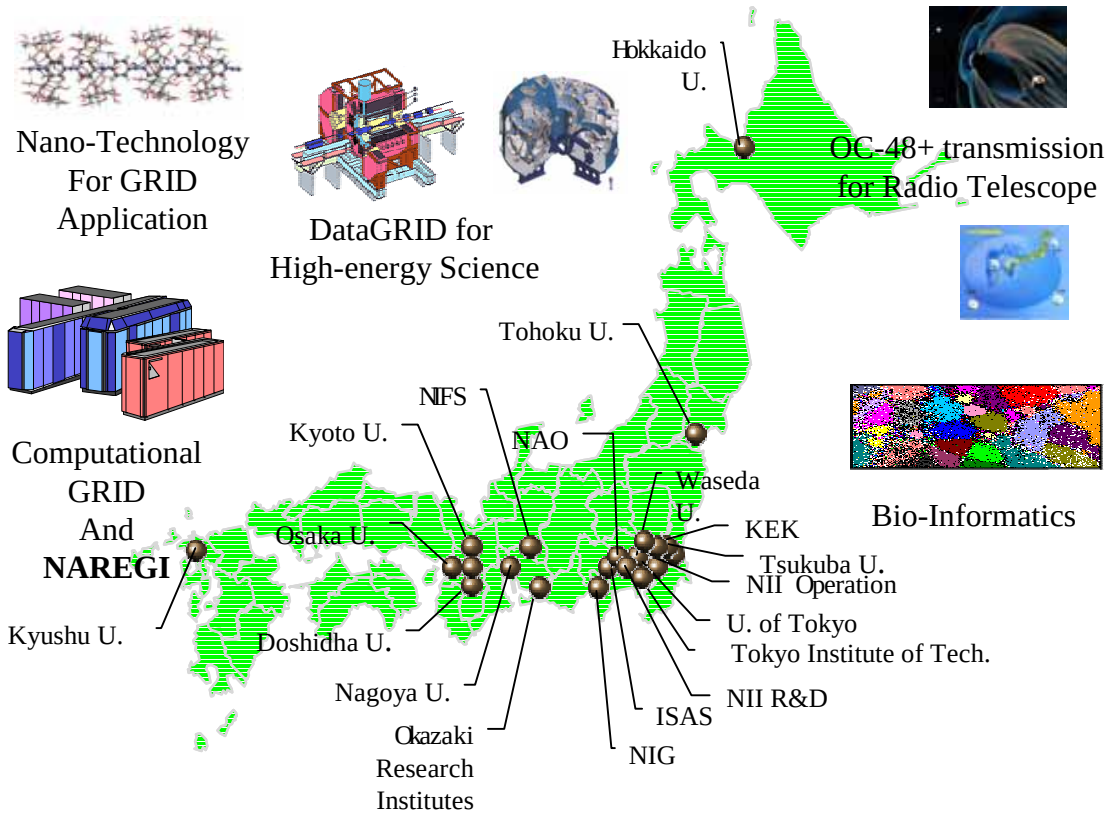
Project Deputy Leader, NAREGI Project
Visiting Professor, National Institute of Informatics
Professor, Tokyo Inst. Technology

Grid Related Projects in Japan

- Information Technology Based Lab (ITBL
- Super-SINET(NII)
- VizGrid (Prof. Matsuzawa, JAIST)
- BioGrid (Prof. Shimojo, Osaka-U)
- Campus Grid(Prof. Matsuoka, Titech)
- National Research Grid Initiative (NAREGI)
- Grid Technology Research Center(Mr. Sekiguchi)
- Japan Virtual Observatory (JVO)

SuperSINET: All Optical Production Research Network (separate funding)

- 10Gbps Photonic Backbone
- GbEther Bridges for peer-connection
- Very low latency – Titech-Tsukuba 3-4ms roundtrip
- Operation of Photonic Cross Connect (OXC) for fiber/wavelength switching
- 6,000+km dark fiber, 100+ e-e lambda and 300+Gb/s
- Operational since January, 2002



National Research Grid Initiative (NAREGI) Project:Overview

- A new Japanese MEXT National Grid R&D project
  ~$(US)17M FY'03 (similar until FY'07)
($125million total planned)

- One of two major Japanese Govt. Grid Projects

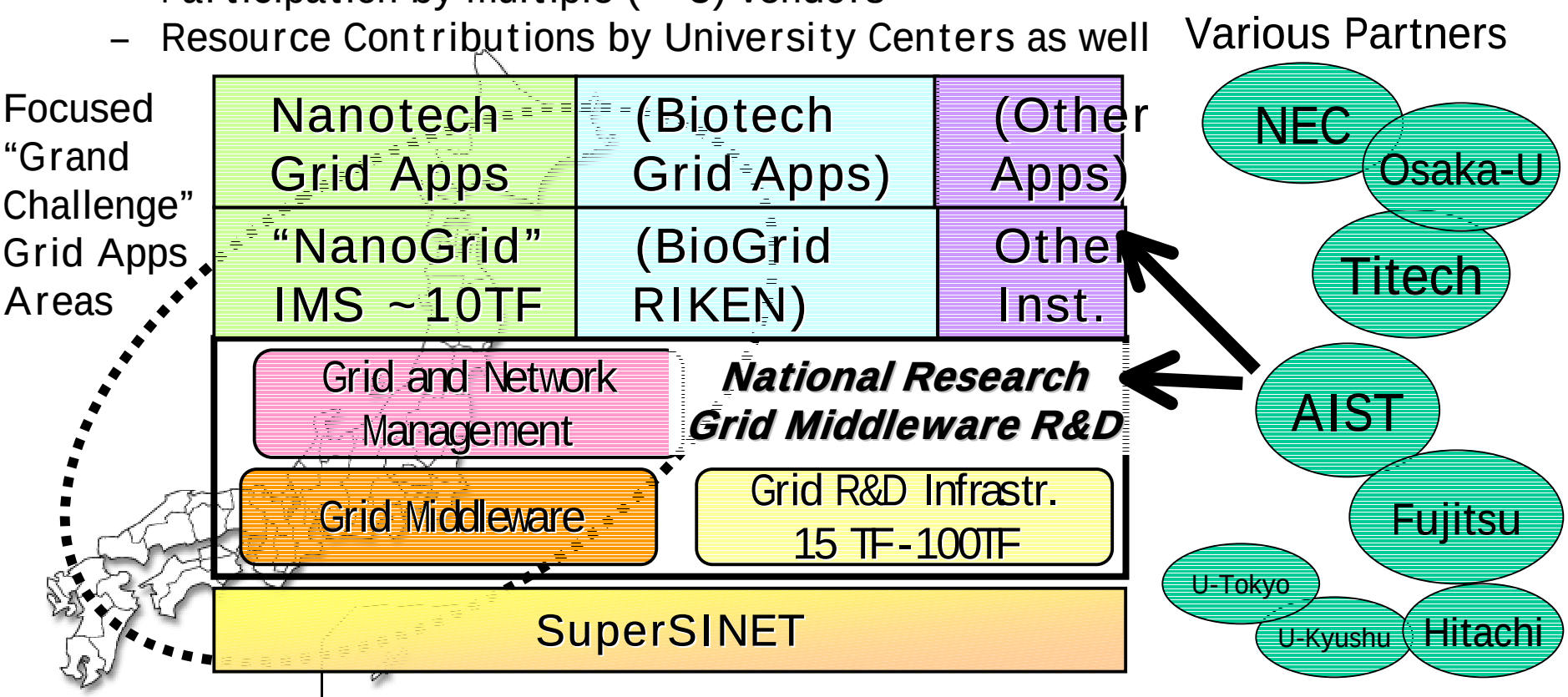
- Collaboration of National Labs. Universities and Major Computing and Nanotechnology Industries

- Acquisition of Computer Resources underway (FY2003) – 17 Teraflop Dedicated Testbed

MEXT:Ministry of Education, Culture, Sports,Science and Technology

National Research Grid Infrastructure (NAREGI) 2003-2007

- Petascale Grid Infrastructure R&D for Future Deployment
  - $45 mil (US) + $16 mil x 5 (2003-2007) = $125 mil total
  - Hosted by National Institute of Informatics (NII) and Institute of Molecular Science (IMS)
  - PL: Ken Miura (Fujitsu➔NII)
    - SLs Sekiguchi(AIST), Matsuoka(Titech), Shimojo(Osaka-U), Hirata(IMS)…
  - Participation by multiple (>= 3) vendors
  - Resource Contributions by University Centers as well



NAREGI R&D Assumptions & Goals

- Future Research Grid Metrics
  - 10s of Institutions/Centers, various Project VOs
  - > 100,000 users, > 100,000 CPUs/machines
    - Machines very heterogeneous, SCs, clusters, desktops
  - 24/7 usage, production deployment
  - Server Grid, Data Grid, Metacomputing…
- Do not reeinvent the wheel
  - Build on, collaborate with, and contribute to the "Globus, Unicore, Condor" Trilogy
  - Scalability and dependability are the key
- Win support of users
  - Application and experimental deployment essential
  - However not let the apps get a "free ride"
  - R&D for production quality (free) software

National Research Grid Initiative (NAREGI) Project:Goals

(1) To develop a Grid Software System (R&D in Grid Middleware and Upper Layer) as the prototype for future Grid Infrastructure in scientific research in Japan

(2) To provide a Testbed to prove that the High-end Grid Computing Environment (100+Tflop/s expected by 2007) can be practically utilized in the Nano-science Simulations over the Super SINET.

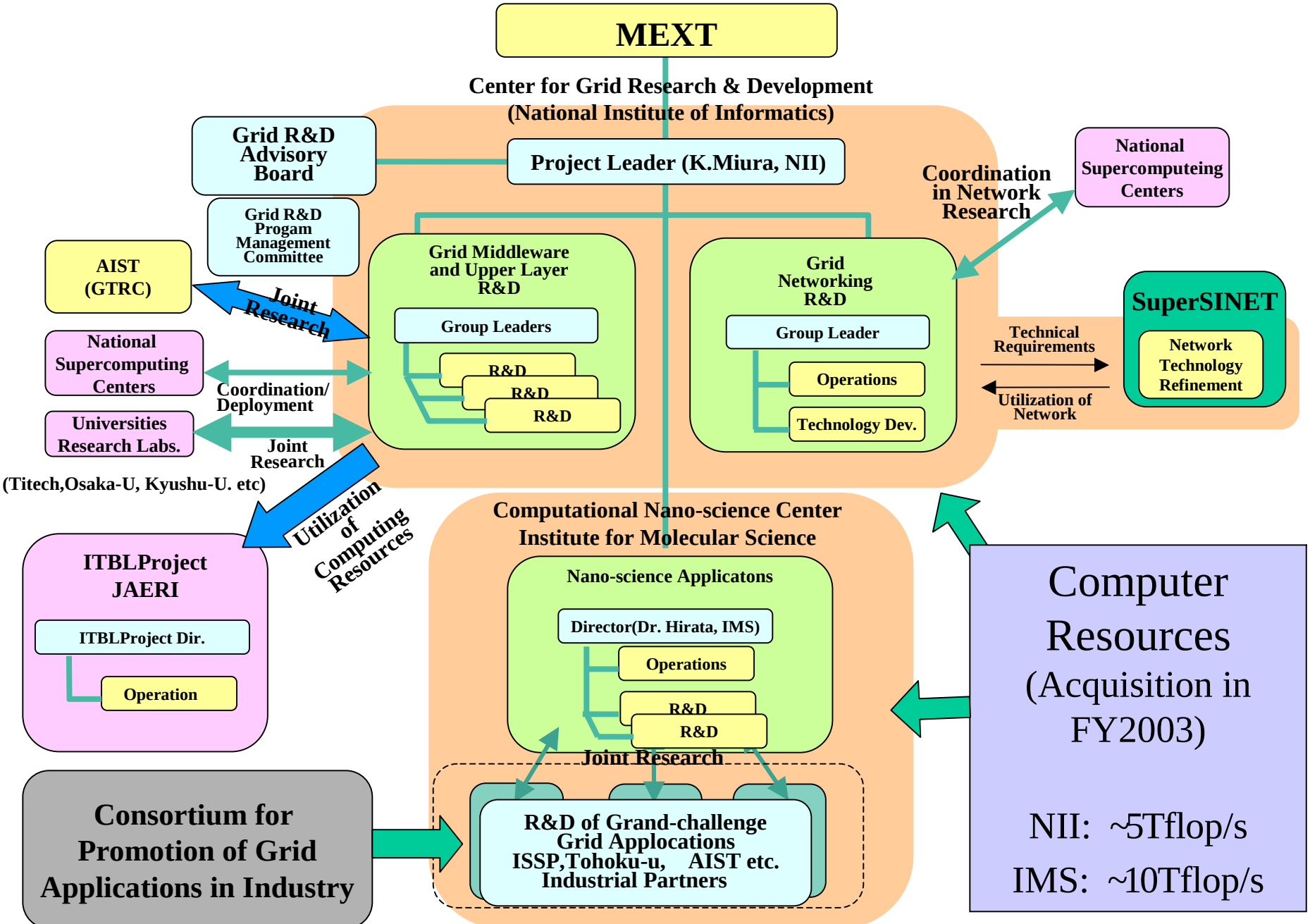
(3) To Participate in International Collaboration (U.S., Europe, Asian Pacific)

(4) To Contribute to Standardization Activities, e.g., GGF

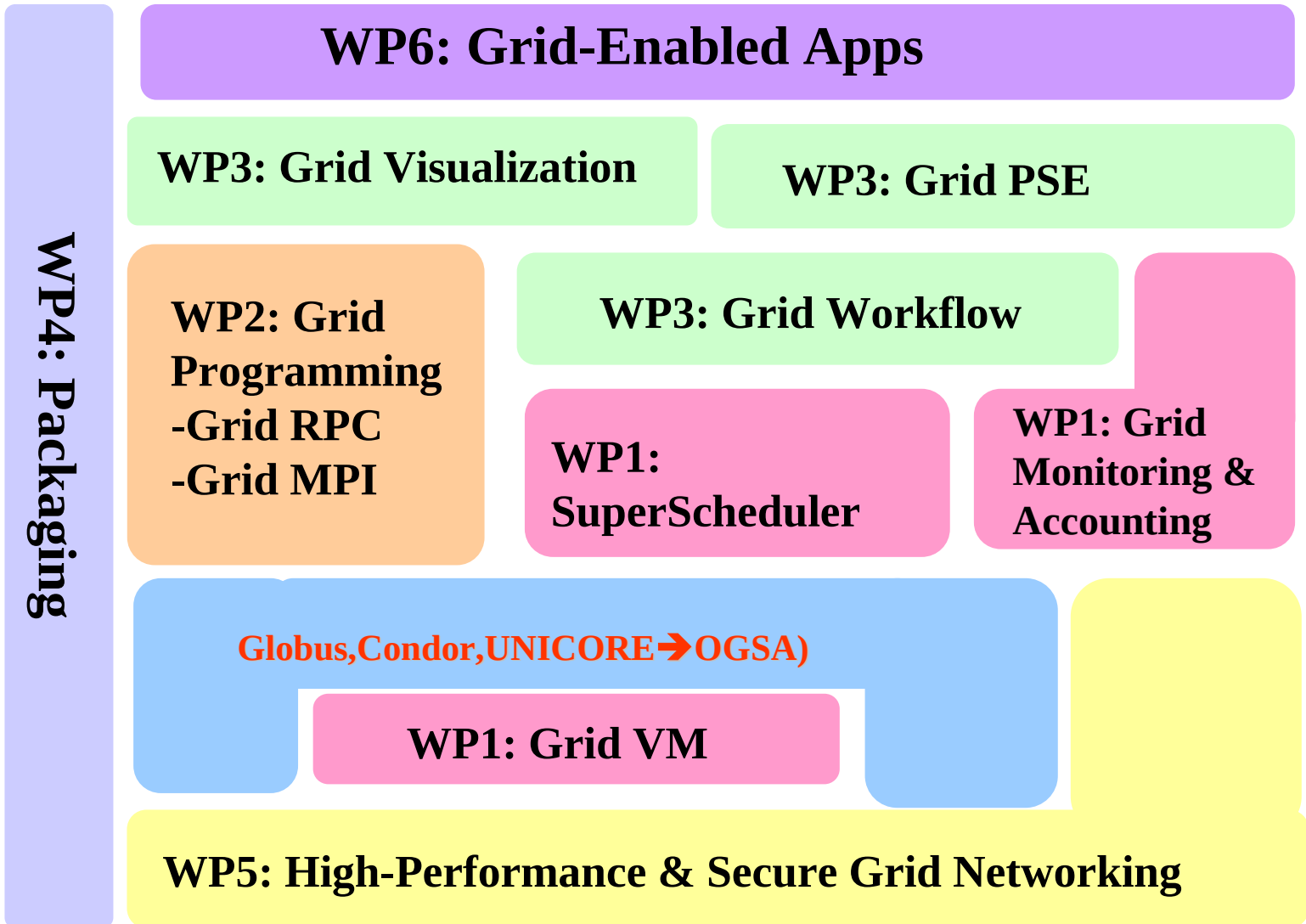
Participating Organizations

- National Institute of Informatics (NII)
  (Center for Grid Research & Development)
- Institute for Molecular Science (IMS)
  (Computational Nano science Center)
- Universities and National Laboratories(Joint R&D)
  (AIST, Titech, Osaka-u, Kyushu-u, Kyushu Inst. Tech., Utsunomiya-u, etc.)
- Research Collaboration
  (ITBL Project, National Supecomputing Centers etc.)
- Participating Vendors (IT and Chemicals/Materials)
- Consortium for Promotion of Grid Applications in Industry

NAREGI Research Organization and Collaboration



NAREGI Software Stack



R&D in Grid Software and Networking Area (Work Packages)

- WP-1: Lower and Middle-Tier Middleware for Resource Management:
  Matsuoka (Titech), Kohno(ECU), Aida (Titech)
- WP-2: Grid Programming Middleware:
  Sekiguchi(AIST), Ishikawa(AIST)
- WP-3: User-Level Grid Tools & PSE:
  Miura (NII), Sato (Tsukuba-u),
  Kawata(Utsunomiya-u)
- WP-4: Packaging and Configuration Management:
  Miura (NII)
- WP-5: Networking, Security & User Management
  Shimojo (Osaka-u), Oie (Kyushu Tech.)
- WP-6: Grid-enabling tools for Nanoscience Applications :
  Aoyagi (Kyushu-u)

WP-1: Lower and Middle-Tier Middleware for Resource Management

- Unicore ⇔ Condor ⇔ Globus Interoperability
  - Adoption of ClassAds Framework
- Meta-scheduler
  - Scheduling Schema
  - Workflow Engine
  - Broker Function
- Grid Information Service
  - Attaches to multiple monitoring framework
  - User and job <u>auditing</u>
  - CIM-based node information schema
  - <u>Accounting</u> based on user/job audit

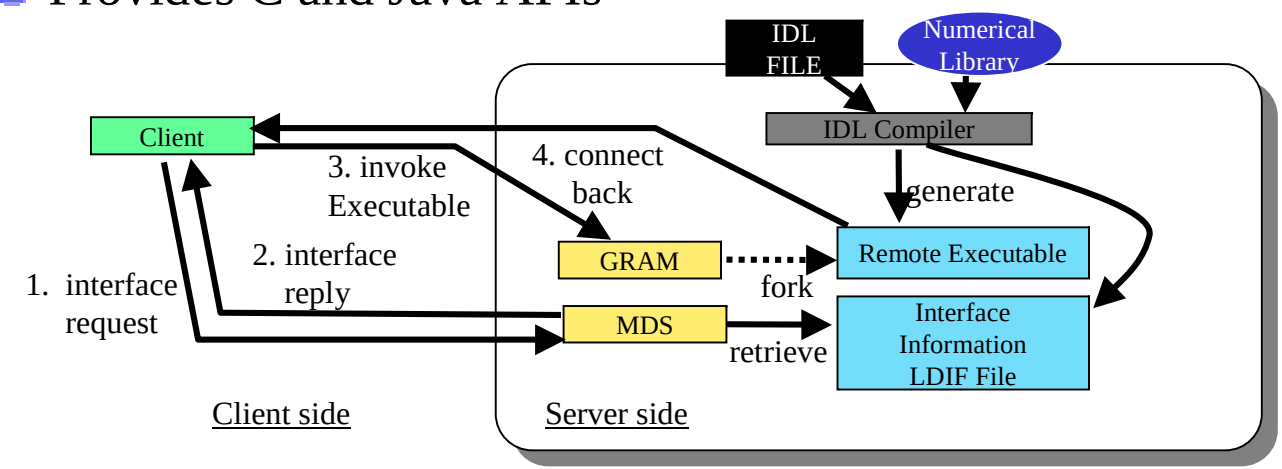
WP-1: Lower and Middle-Tier Middleware for Resource Management(Continued)

- Self-Configurable Monitoring
- Grid Self-Configuration Management (including packaging)
- GridVM(Lightweight Grid Virtual Machine)
  - Support for Co-scheduling
  - Fine-Grained Resource Control
  - Node (IP) virtualization
  - Interfacing with OGSA

WP-2:Grid Programming – GridRPC/Ninf-G2 (AIST/GTRC)

GridRPC

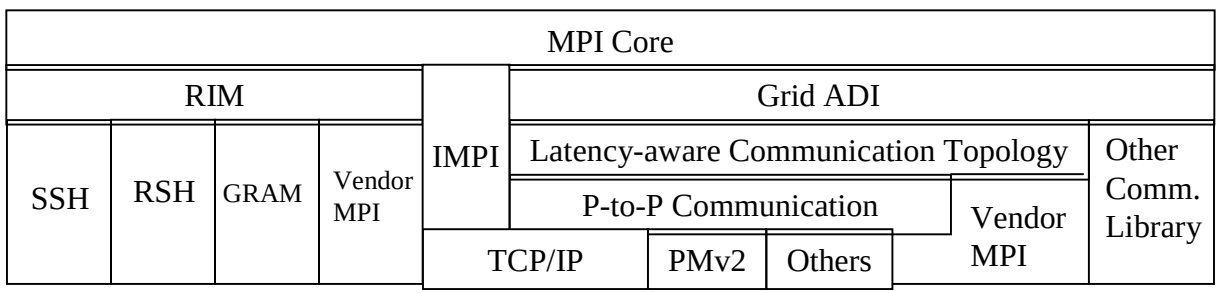
- Programming Model using RPC on the Grid
- High-level, taylored for Scientific Computing (c.f. SOAP-RPC)
- GridRPC API standardization by GGF GridRPC WG
- Ninf-G Version 2
  - A reference implementation of GridRPC API
  - Implemented on top of Globus Toolkit 2.0 (3.0 experimental)
  - Provides C and Java APIs



WP-2:Grid Programming -GridMPI (AIST and U-Tokyo)

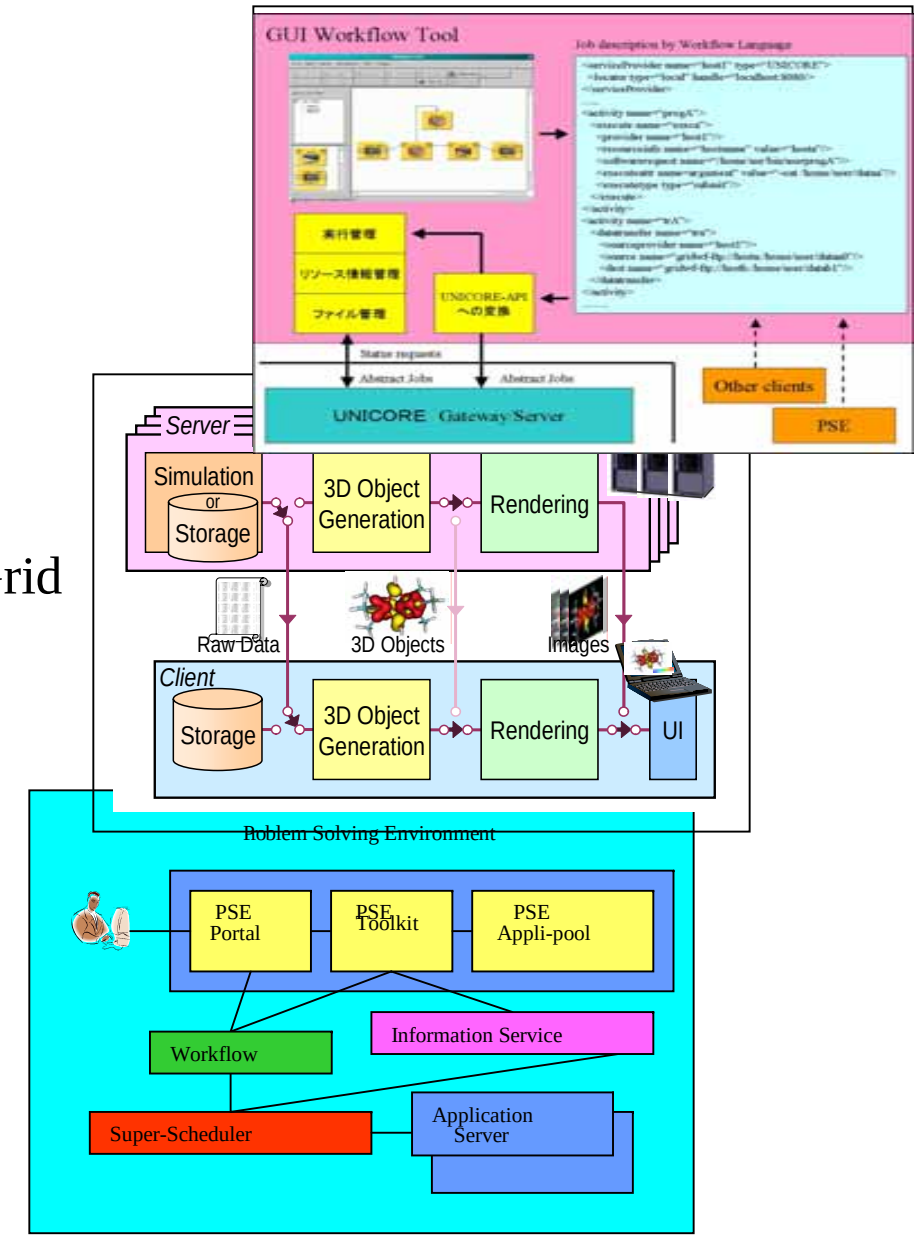
GridMPI

- Provides users an environment to run MPI applications efficiently in the Grid.
- Flexible and hterogeneous process invocation on each compute node
- GridADI and Latency-aware communication topology, optimizing communication over non-uniform latency and hides the difference of various lower-level communication libraries.
- Extremely efficient implementation based on MPI on Score (Not MPICHI-PM)



WP-3: User-Level Grid Tools & PSE

- Grid Workflow
  - Workflow Language Definition
  - GUI(Task Flow Representation)
- Visualization Tools
  - Real-time volume visualization on the Grid
- PSE /Portals
  - Multiphysics/Coupled Simulation
  - Application Pool
  - Collaboration with Nano-science Applicatons Group



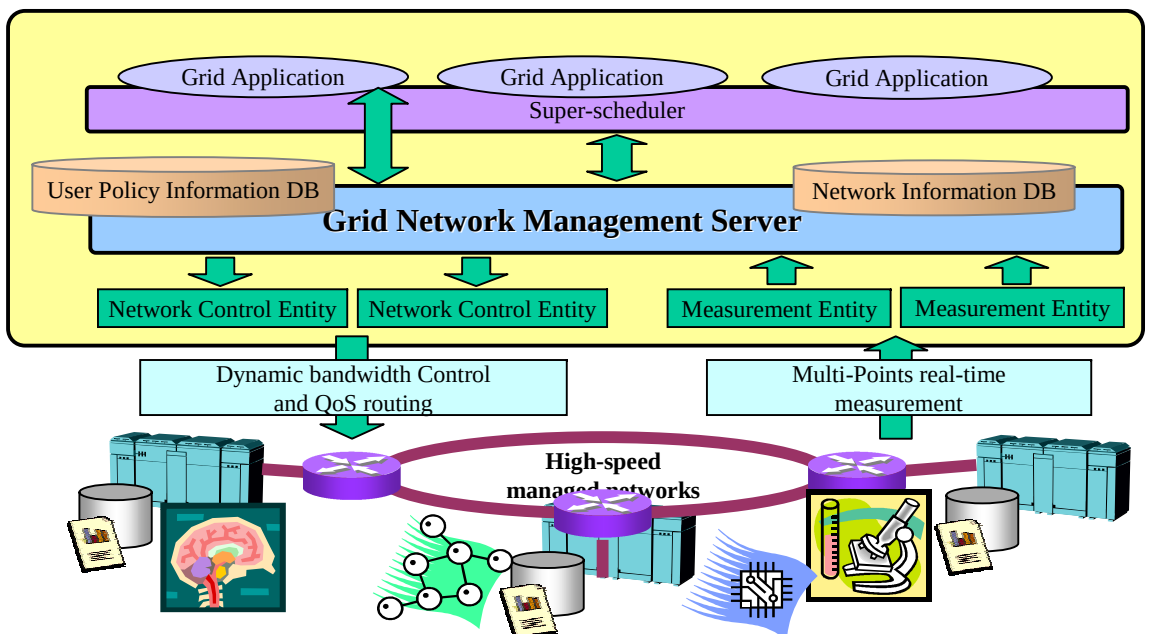
WP-4: Packaging and Configuration Management

- Collaboration with WP1 management
- Activities
  - Selection of packagers to use
  - Interface with autonomous configuration management (WP1)
  - Test Procedure and Harness
  - Testing Infrastructure

    c.f. NSF NMI packaging and testing

WP-5: Network Measurement, Management & Control for Grid Environment

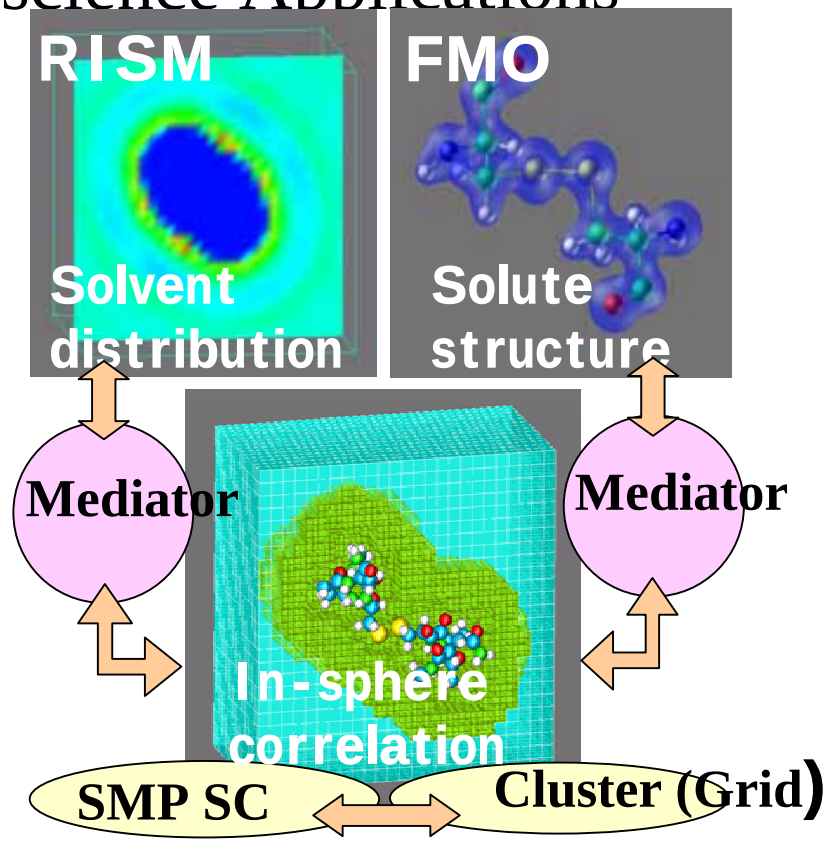
- **Traffic measurement on SuperSINET**
- **Optimal QoS Routing based on user policies and network measurements**
- **Robust TCP/IP Control for Grids**
- **Grid CA/User Grid Account Management and Deployment**



WP-6:Adaptation of Nano-science Applications to Grid Environment

- Analysis of Typical Nanoscience Applications
  - Parallel Structure
  - Granularity
  - Resource Requirement
  - Latency Tolerance
- Coupled Simulation

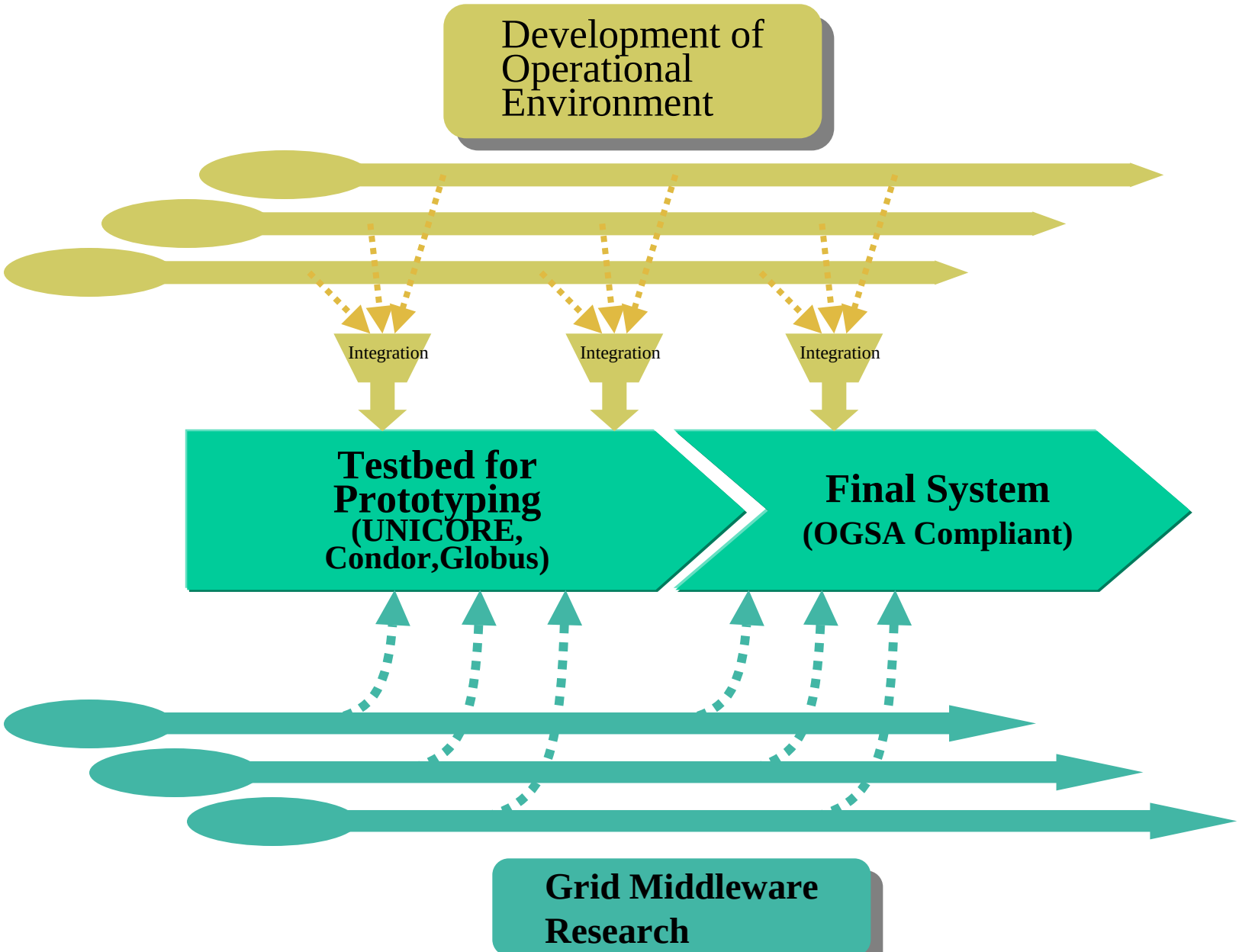
  (e.g.,FMO & RISM)
- Collaboration with IMS



RISM: Reference Interaction Site Model

FMO: Fragment Molecular Orbital Method

Development Concept



Nano-science and Technology Applications Targeted

Participating Organizations:

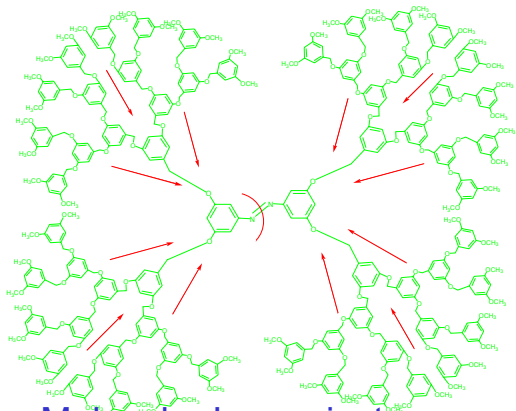
-Institute for Molecular Science
-Institute for Solid State Physics
-AIST
-Tohoku University
-Kyoto University
-Industry (Materials, Nano-scale Devices)
-Consortium for Promotion of Grid Applications in Industry

Research Topics and Groups:

-Functional nano-molecules(CNT,Fullerene etc.)
-Nano-molecule Assembly(Bio-molecules etc.)
-Magnetic Properties
-Electronic Structure
-Molecular System Design
-Nano-simulation Software Integration System
-Etc.

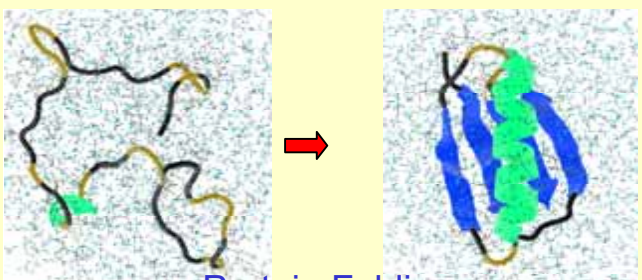
Examples of Nano-Applications Research (1)

Functional Nano-Molecules



Molecular Isomerizaton
by Low Energy Integration

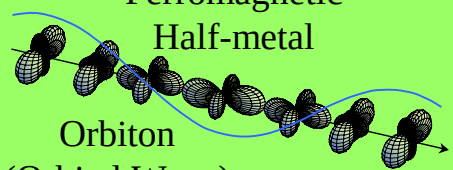
Nano-Molecular Assembly



Protein Folding

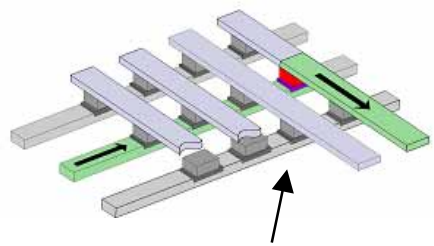
Nano-Electronic System

Manganese-Oxide
Ferromagnetic
Half-metal



Orbiton
(Orbital Wave)

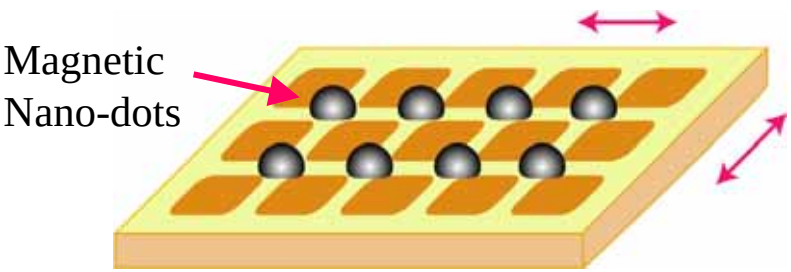
Apps:Half-metal Magnetic
Tunneling Device



Memory Device

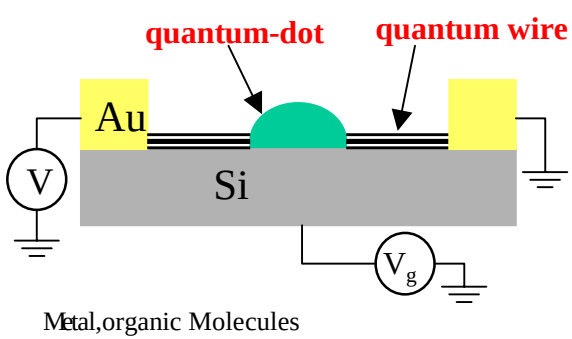
Examples of Nano-Applications Research (2)

Nano-Magnetism



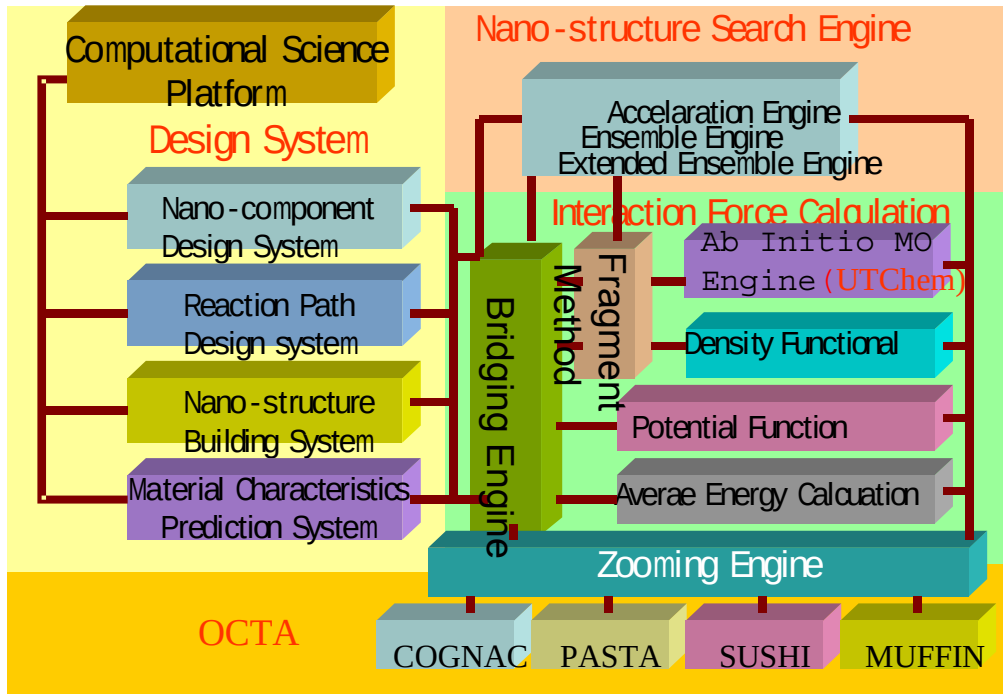
Controling Arrangement of
Nano-dots by Self Organization

Nano-system Design



Nano-device, Quantum Transport

Nano-simulation Software Integration System



NII Center for Grid R&D (Jimbo-cho, Tokyo)

Mitsui Office
Bldg. 14th Floor

700m^2 office space

(100m^2 machine room)

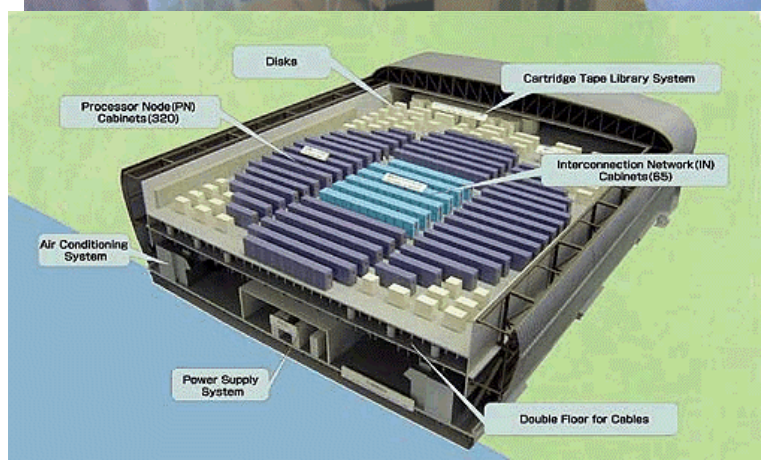
Summary

- We regard Grid as one of the fundamental technologies of the IT infrastructure in 21st century
- In the NAREGI project, seamless federation of heterogeneous resources is the primary objective
- Computations in Nano-science/technology applications over Grid is to be promoted, including industrial participation.
- International Co-operation is essential.

The Japanese Large-scale Grid-Cluster Infrastructure of March 2004

We can make pretty big clusters now for Grids

- Earth Simulator – 5120 procs, 40TF, PB storage
- ASCI Red Storm - ~10,000 procs (onto 20,000 and beyond)
- > 20 machines over 2000 procs in Top 500
- Virginia Tech “Big Mac” - 2200 proc, 17.6TF, $6 million, Ranked 3rd on Top500 (Nov. 2003)



New Japansesae Grid Infrastructures March 2004

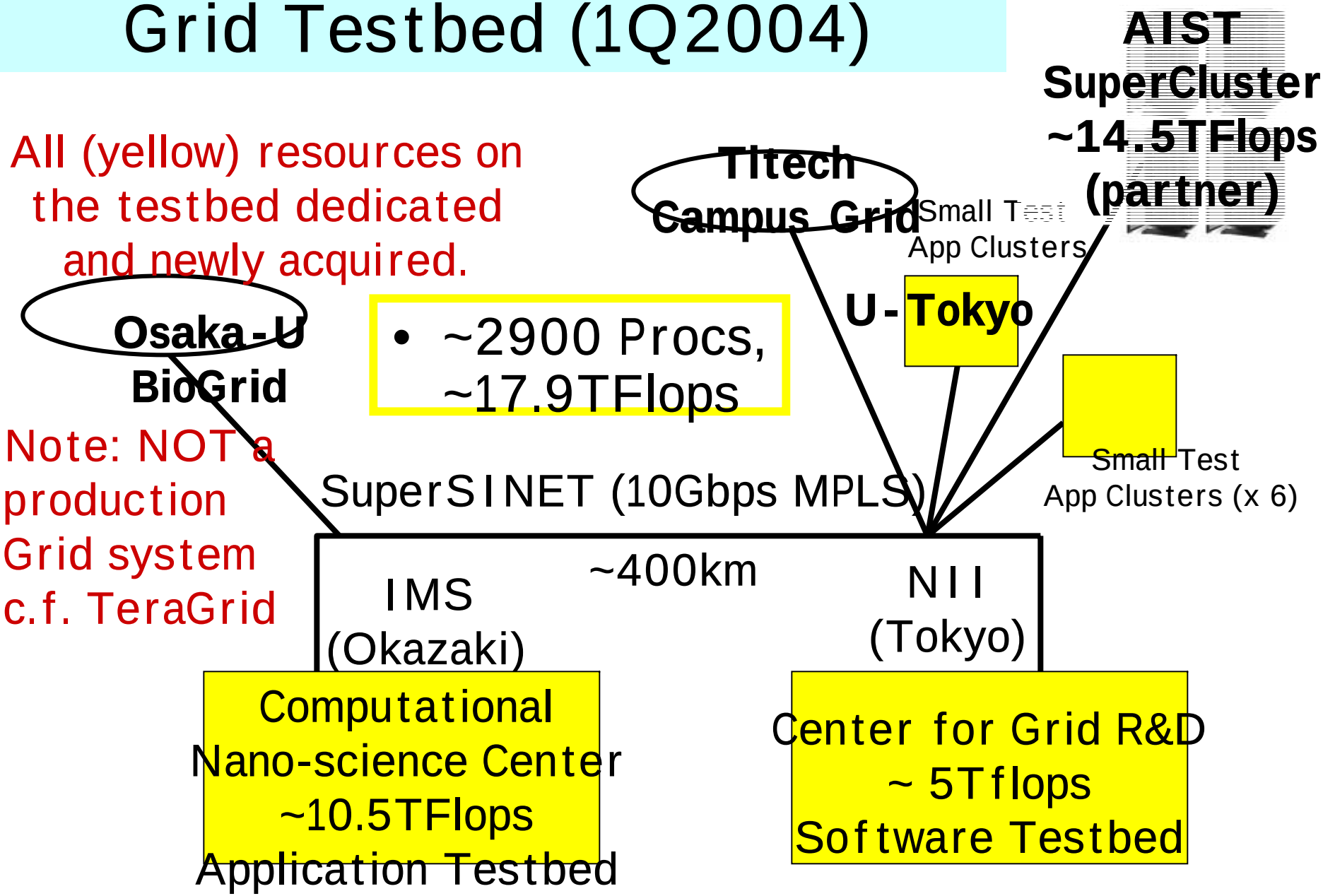
- All Infrastructures Grid-enabled, to be operational March, 2004, interconnected via SuperSINET
- *NAREGI (National Research Grid Initiative) Grid Testbed, 17.9TF, experimental*
  - IMS (Institute for Molecular Science) 2002 CPU 12.8TF
  - NII (National Institute for Informatics), 896 CPU 5.1TF
- *AIST GTRC (Grid Technology Research Ctr) Grid SuperCluster, 14.5 TF, experimental & production*
  - Opteron(2148)+Itanium(528)+Xeon(512)
- *Riken (Institute for Scientific Research) Super Combined Cluster, 12.5TF, production*
  - Xeon(2048) + MD-Grape + SX-7
- Exisiting Infrastructures
  - *SuperSINET National AON, production*
  - *Titech Campus Grid, 800CPU, 1.3TF (since Mar. 2003), semi-production*

Total 8934CPU
46.2TFlops

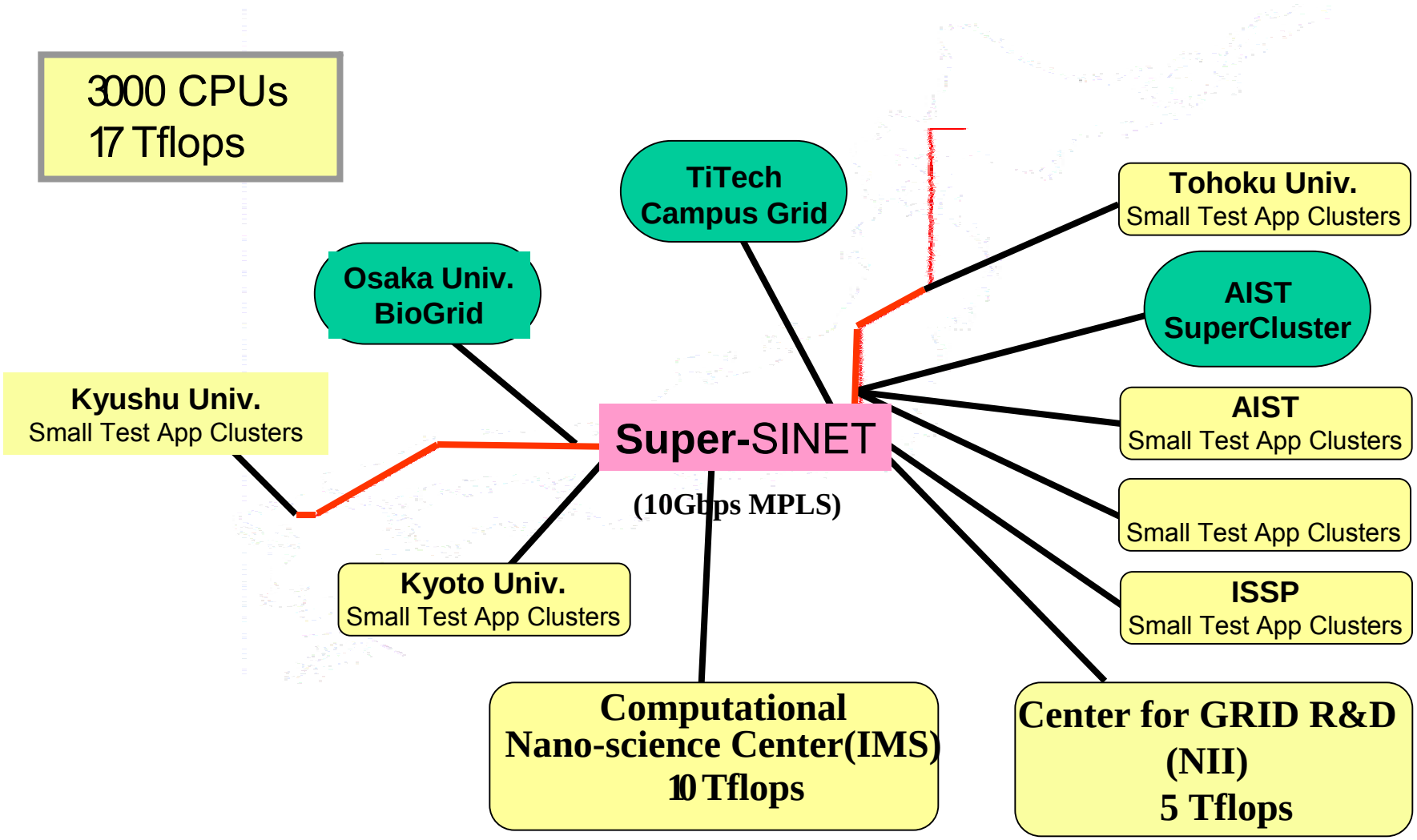
NAREGI Phase 1 Dedicated Grid Testbed (1Q2004)

All (yellow) resources on the testbed dedicated and newly acquired.

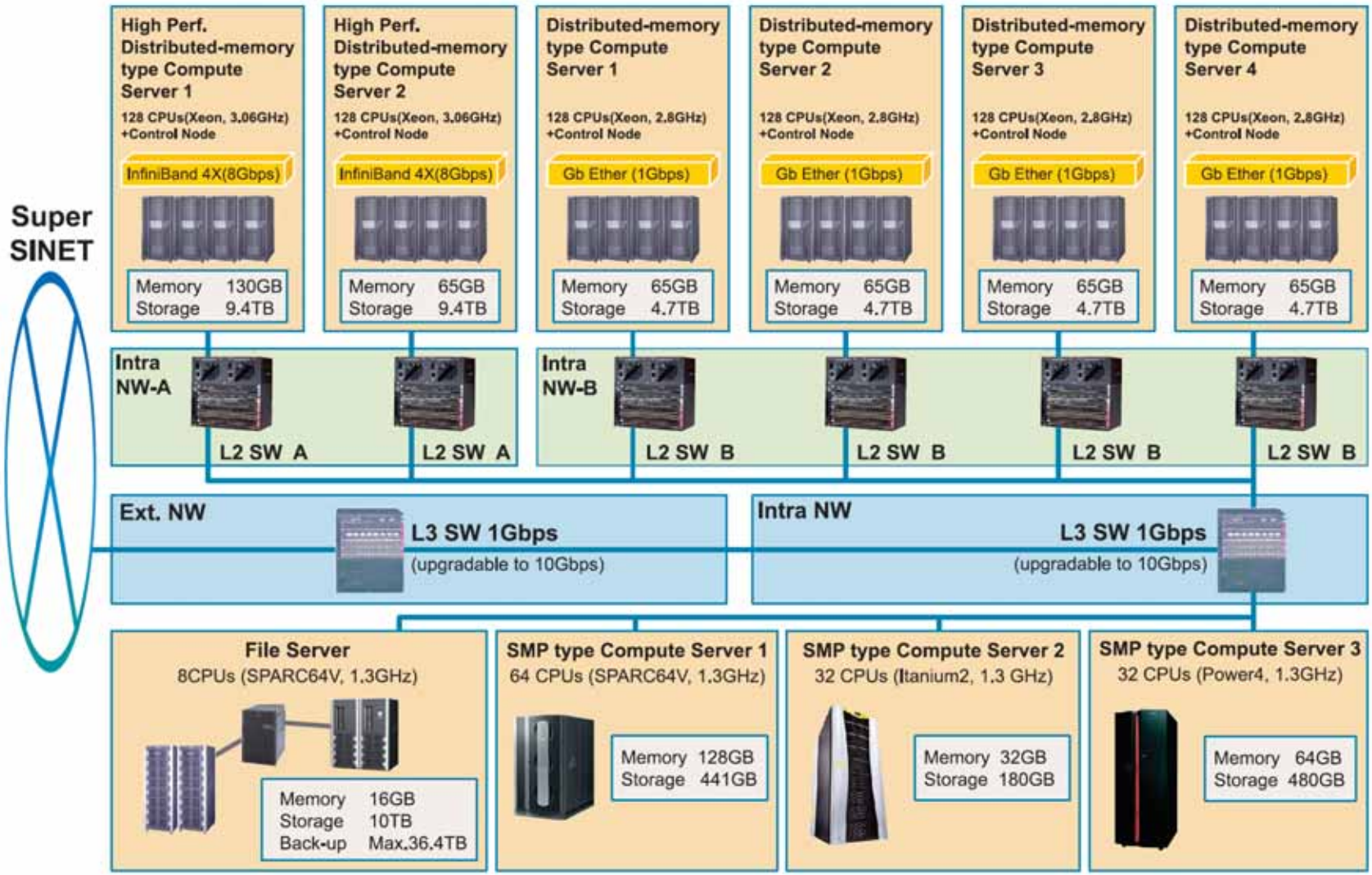
Note: NOT a production Grid system c.f. TeraGrid



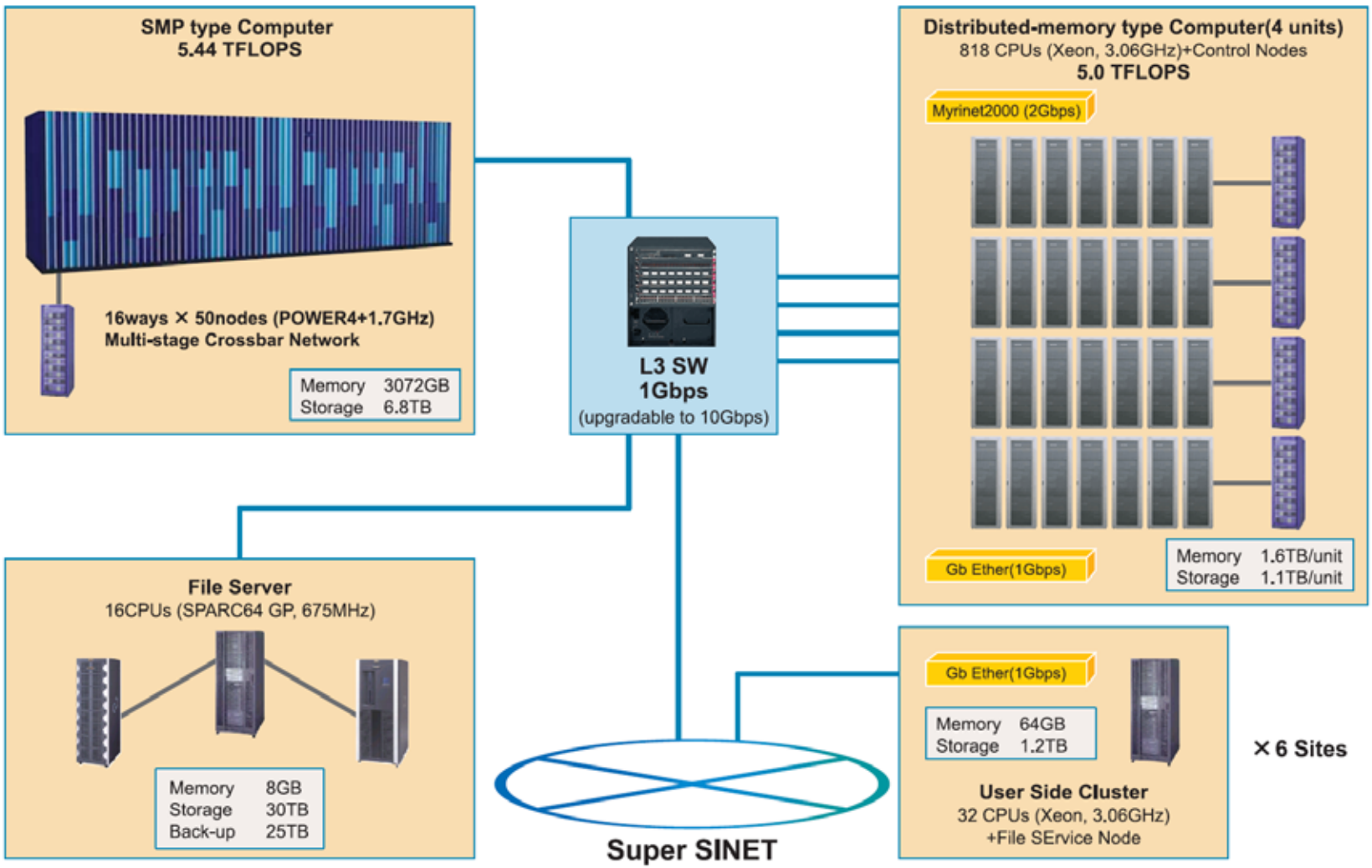
NAREGI Phase 1 Testbed



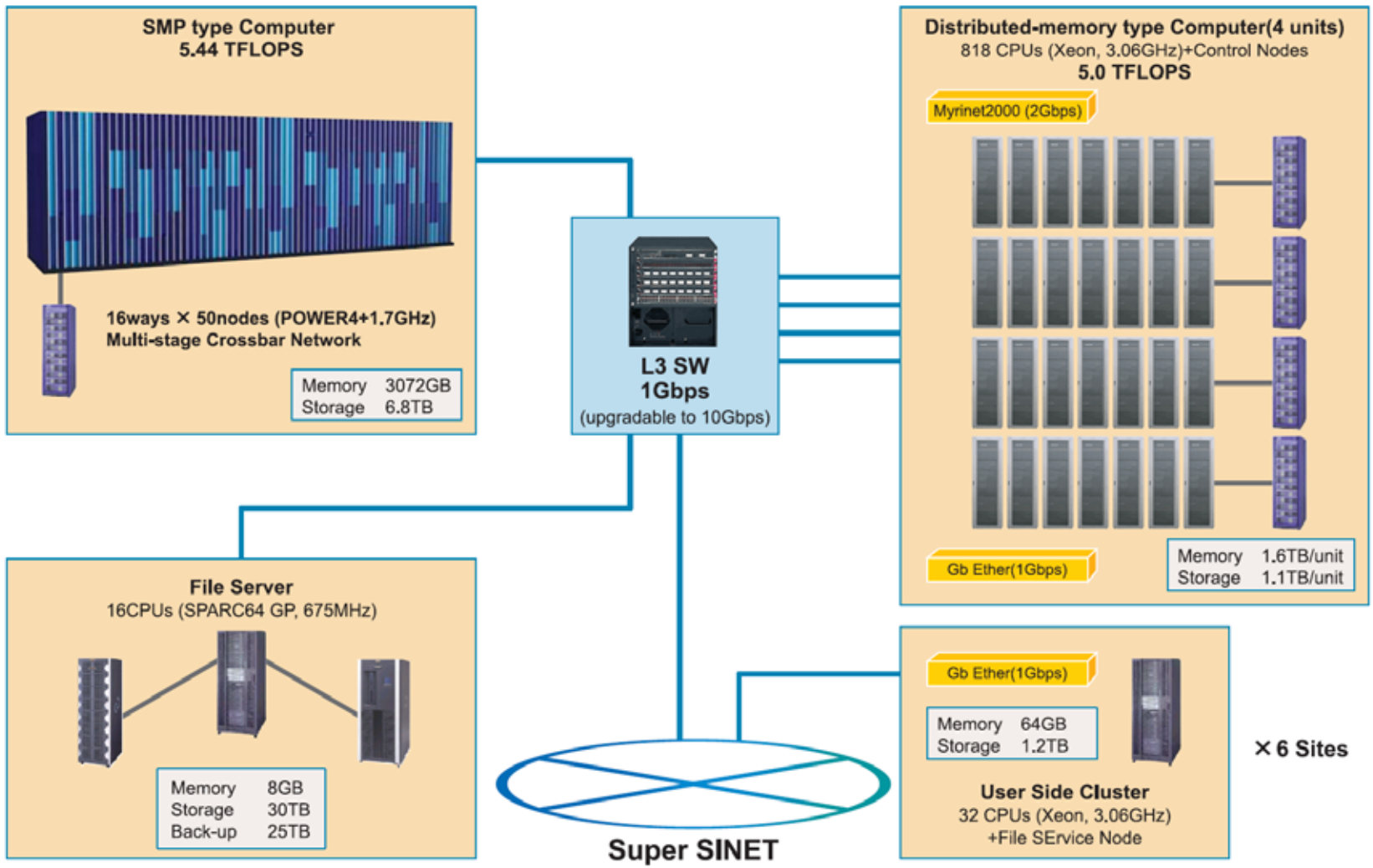
Computer System for Grid Software Infrastructure R&D (at NII) (5 Tflops, 700GB)



Computer System for Grid Software Infrastructure R&D (at IMS) (10 Tflops, 5TB)

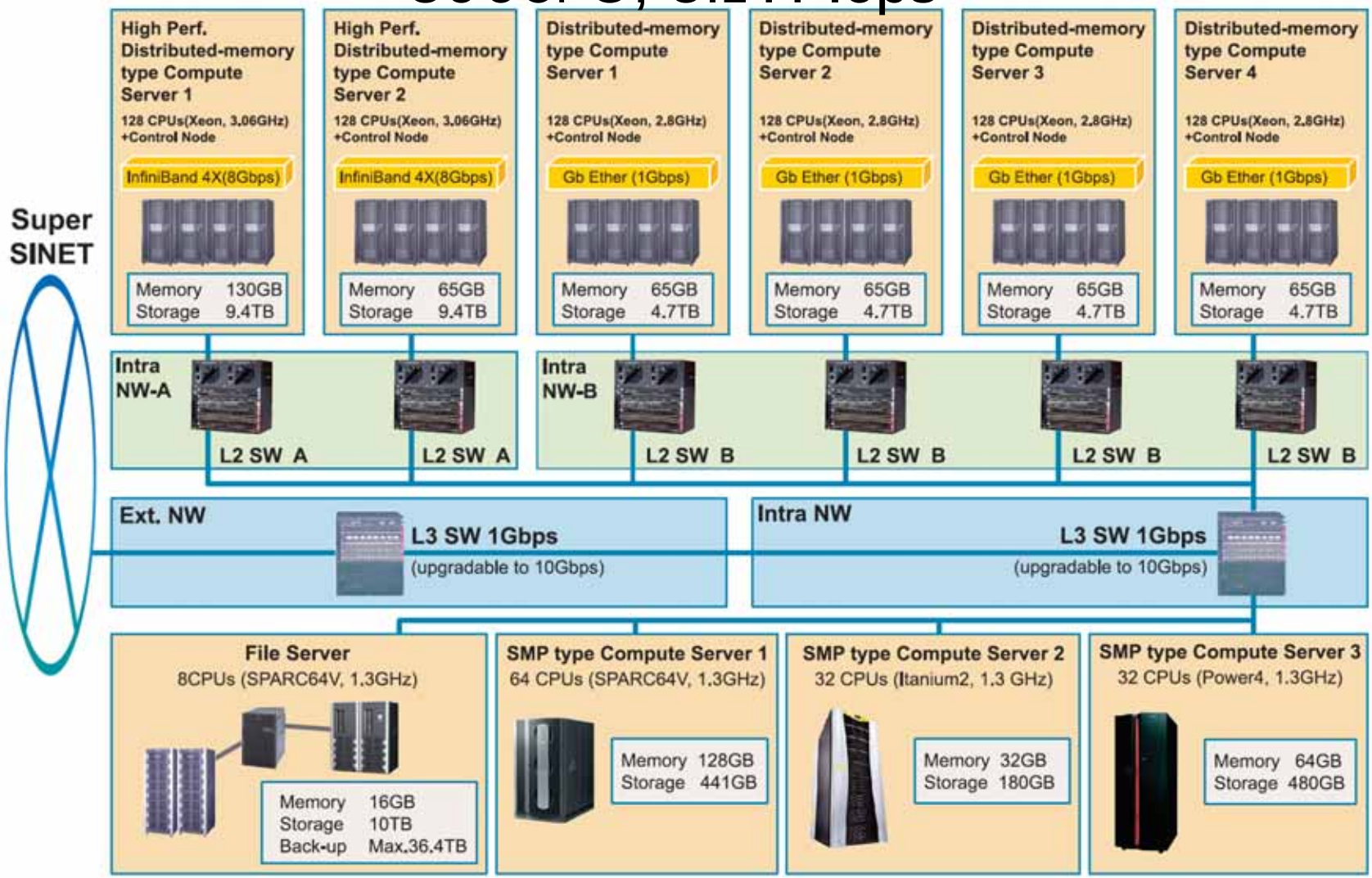


NAREGI Application Grid Testbed @ IMS (Institute for Molecular Science) 2002CPU, 12.8TFlops

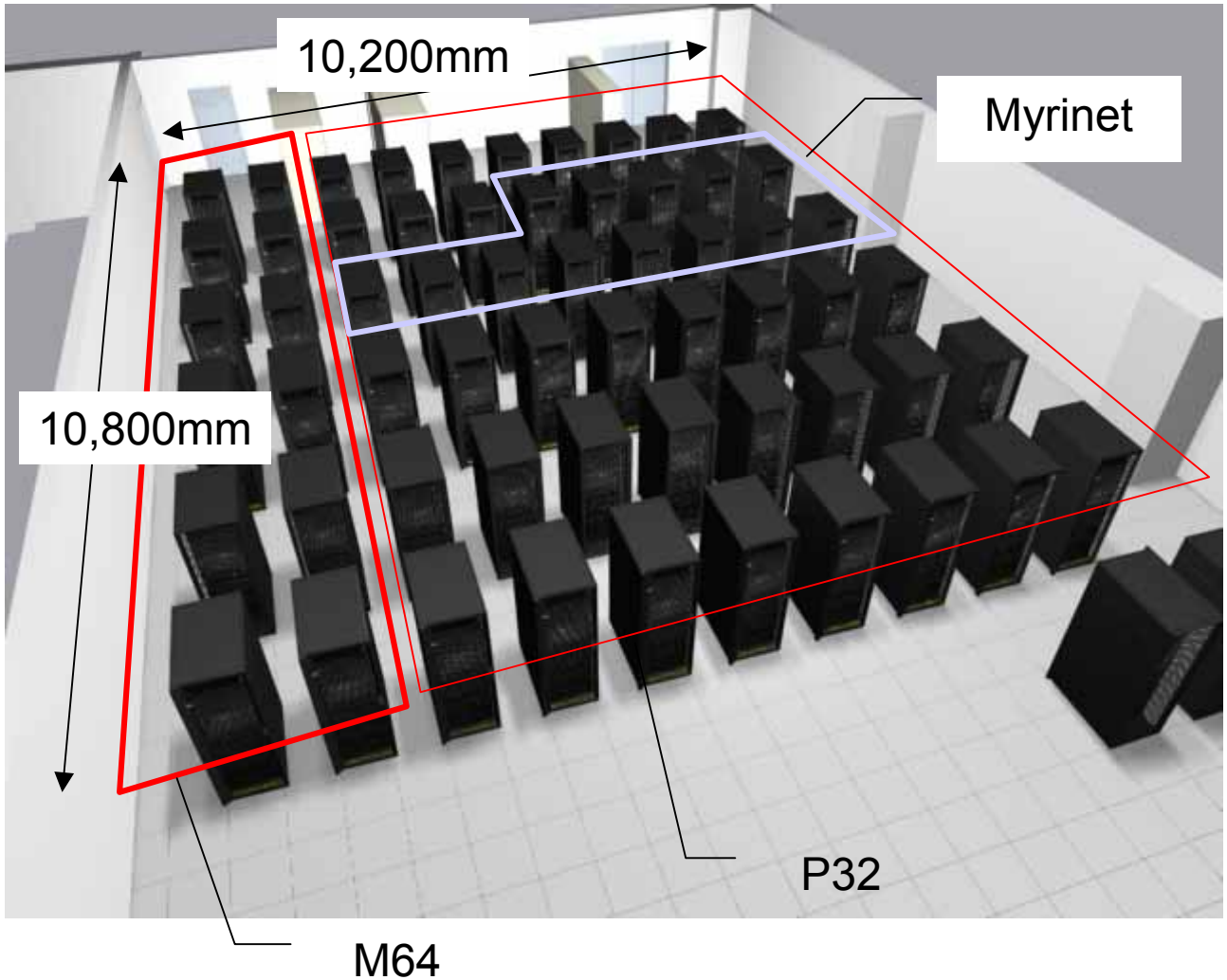


NAREGI Middleware R&D Grid Testbed @ NII (National Institute of Informatics)

896CPU, 5.1TFlops



AIST GTRC Grid) Super Cluster



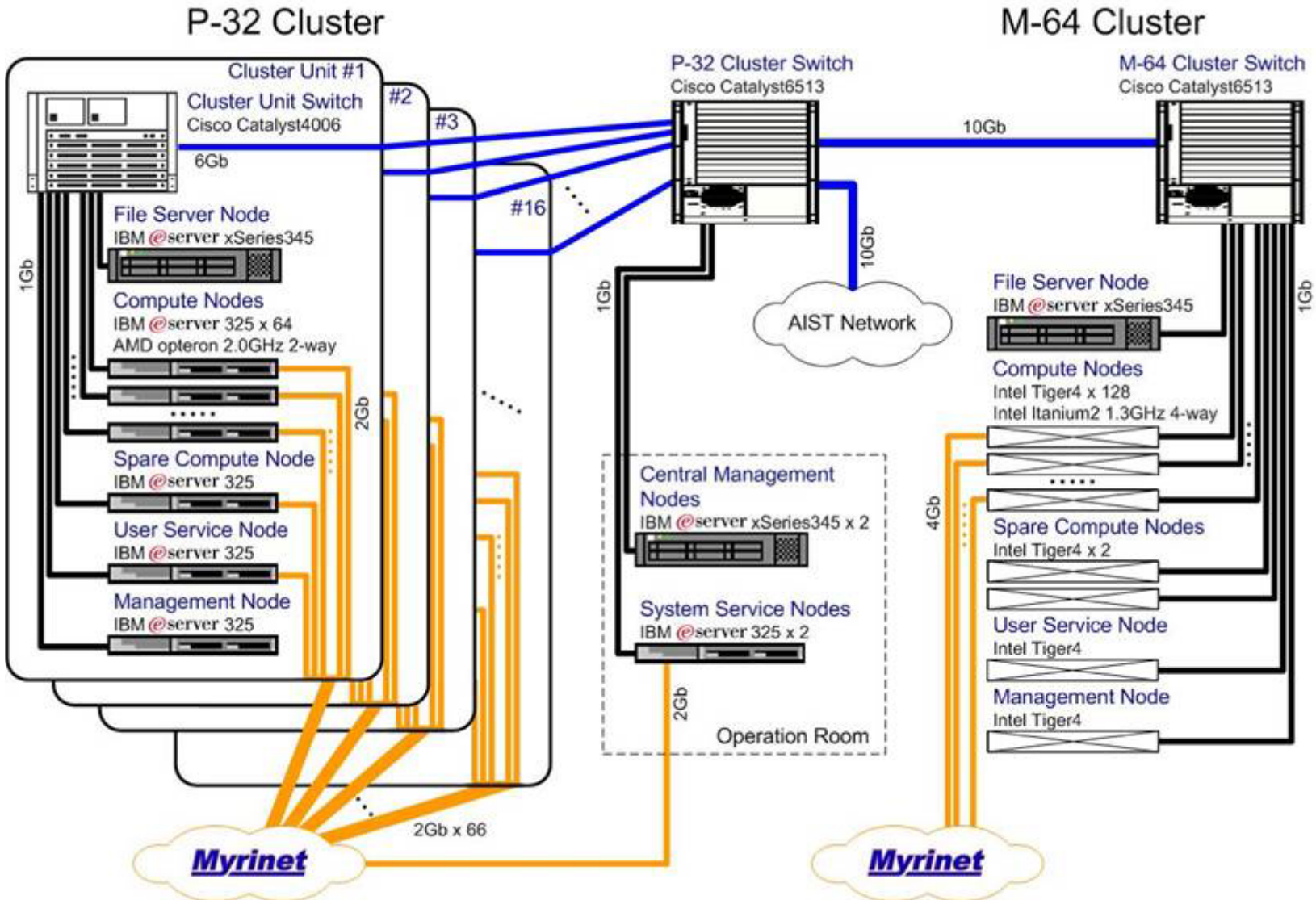
P32: IBM eServer325
Opteron 2.0GHz, 6GB
2way x 1074 node
Myrinet 2000
8.59TFlops/peak

M64: Intel Tiger 4
Madison 1.3GHz, 16GB
4way x 131 node
Myrinet 2000
2.72TFlops/peak

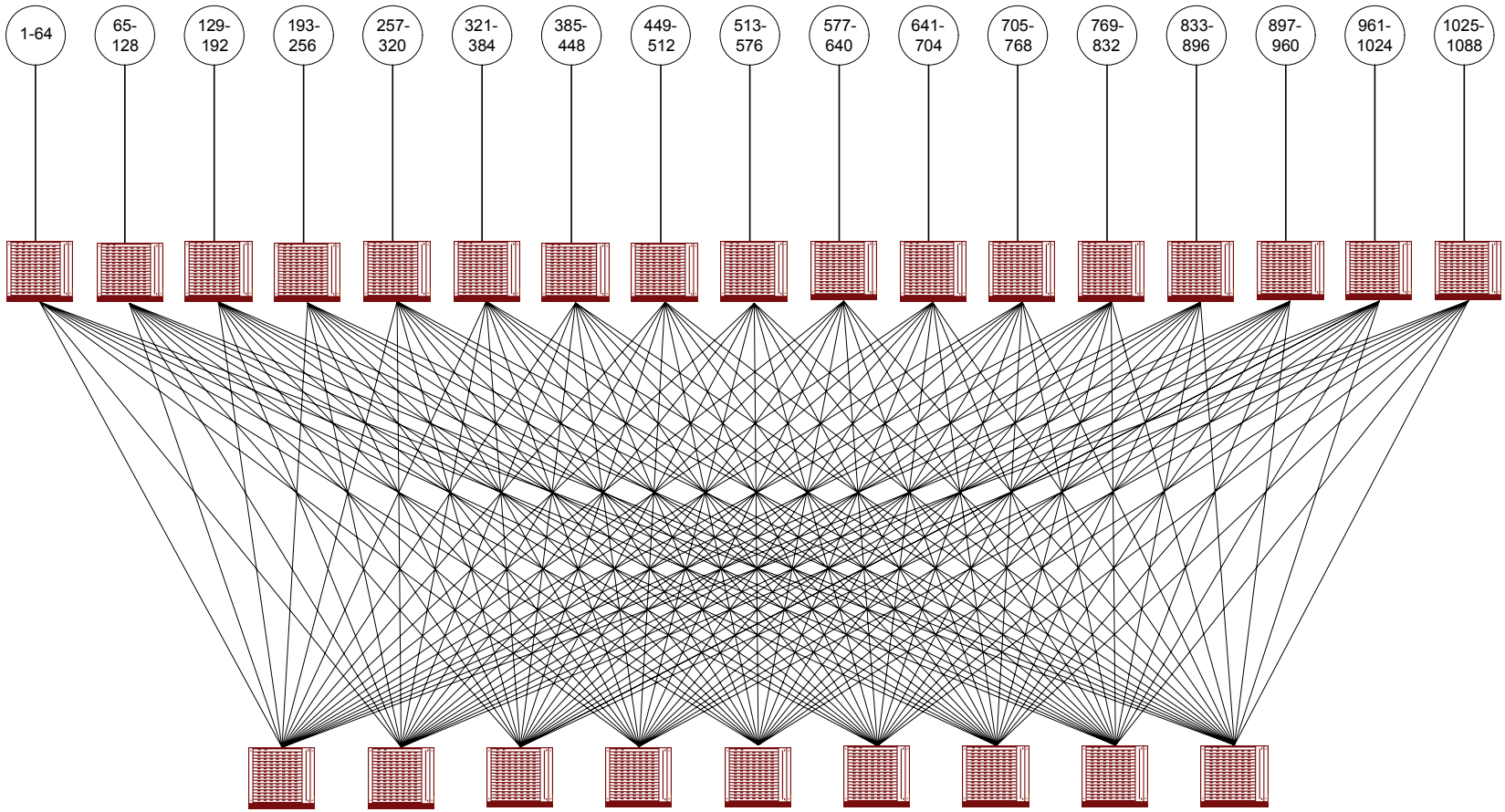
F32: Linux Networx
Xeon 3.06GHz, 2GB
2way x 256+ node
GbE
3.13TFlops/peak

total 14.5TFlops/peak, 3188 CPUs

AIST GTRC (Grid) SuperCluster 3188CPU, 14.5TF



Computational Network for P-32 Cluster



Riken (Grid-enabled) Super Combined Cluster 2048procs, 12.5TFLops (excl. MD Grape and SX)

- **PC Cluster System**
  - **1024nodes, 2048CPU**
  - **CPU Intel Pentium Xeon 3.06GHz**
  - **Mmeory 4GB/node, 2GB/node**
  - **HDD 100GB/node**

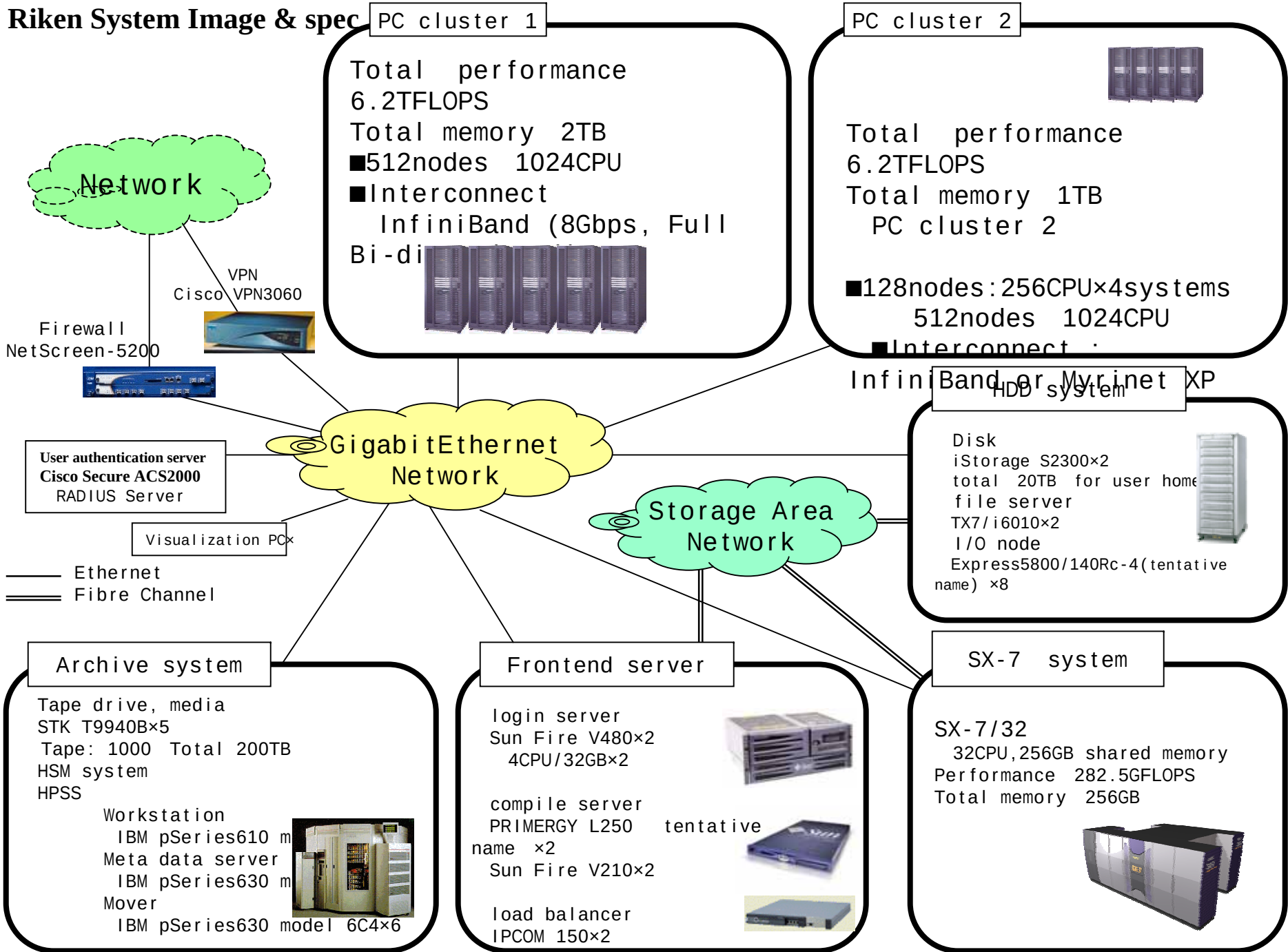
- **PC Cluster for MD**
  - **20 MD-Grape board computers**
  - **64GFLOPS/board**
  - **1.2TFLOPS**

- **SX-7 system**
  - **32CPU**
  - **282.5GFLOPS**
  - **memory 256GB**

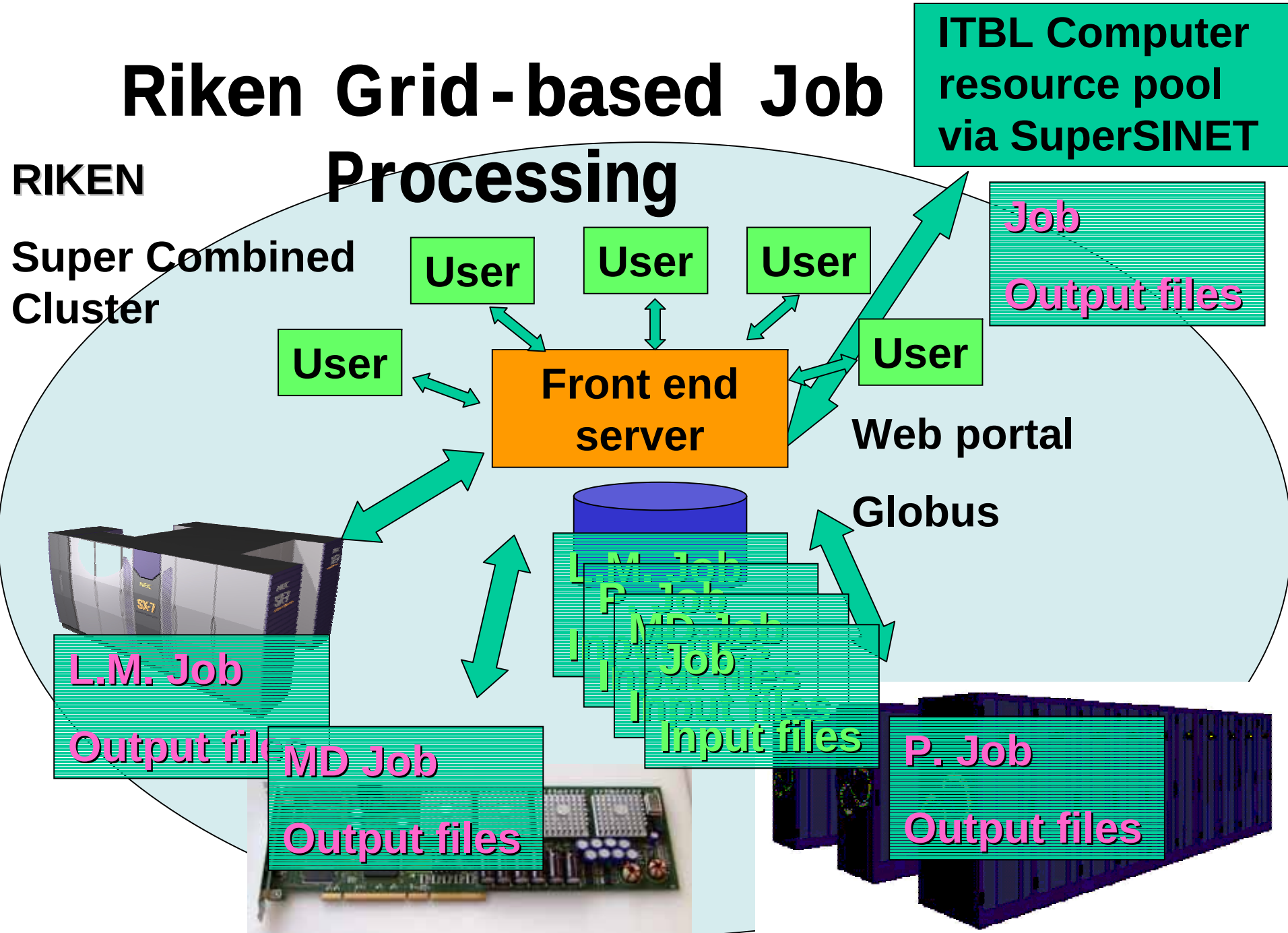
- HDD
  - **20TB**

- Tape storage
  - HSM system HPSS
  - STK T9940B
  - Tape: 1000 Total 200TB

Riken System Image & spec

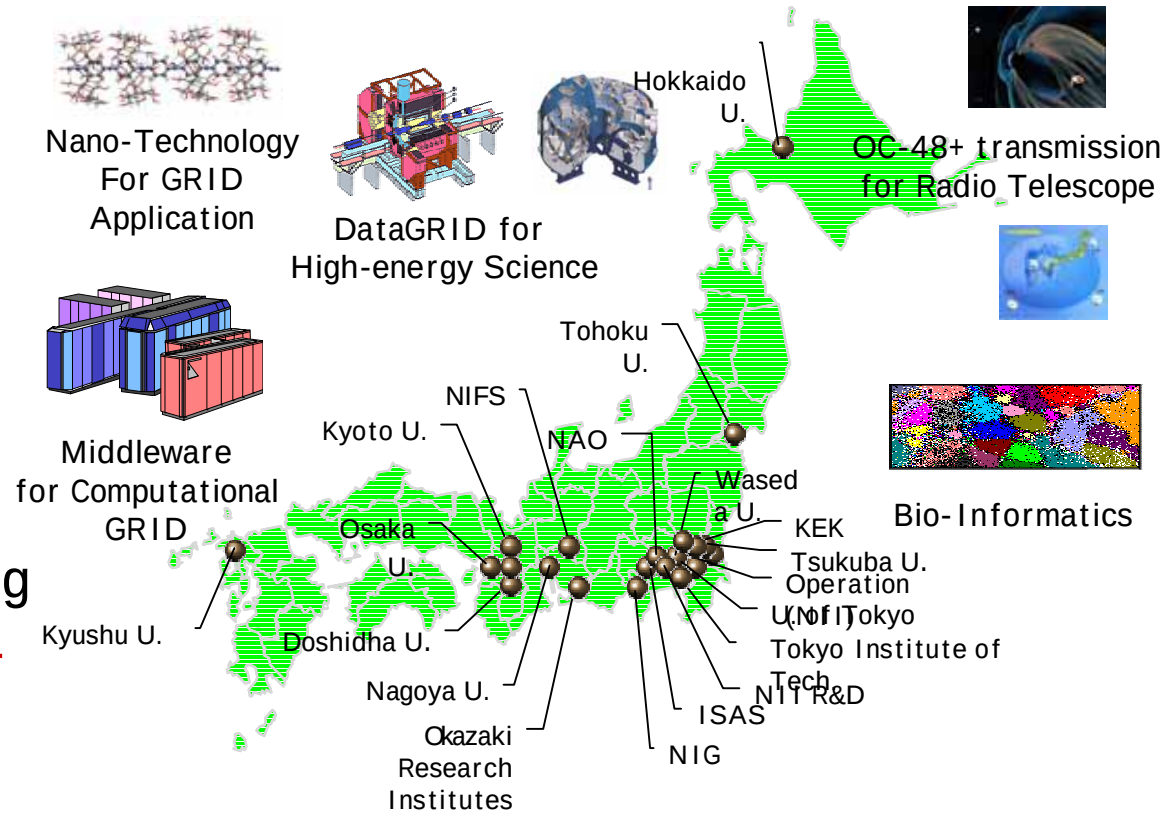


Riken Grid-based Job Processing



SuperSINET: AON Production Research Network (separate funding) interconnect for NAREGI, ITBL, etc. and future National Grid Infrastructure

- 10Gbps General Backbone
- GbE Bridges for peer-connection
- Very low latency – Titech-Tsukuba 3-4ms roundtrip
- Operation of Photonic Cross Connect (PXC) for fiber/wavelength switching
- 6,000+km dark fiber, 100+ e-e lambda and 300+Gb/s
- Operational from January, 2002 until March, 2005



Sample Campus Grid Infrastructure (2006)

CPU Farm
Cluster 3000CPU/
1500Core/750node
18~30TFlops IB 4x
or 1GbE low bisection
25 racks

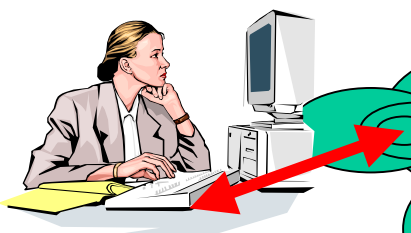
Infiniband 300Gbps

Infiniband 300Gbps

Tightly-Coupled Cluster
1600CPU/800Core/
400node 13~16TFlops
IB 4x high bisection
15 racks

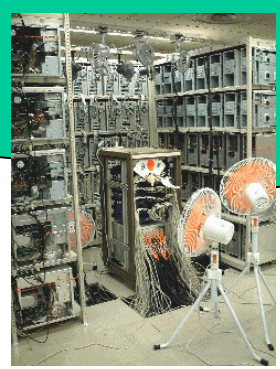
Storage Farm
2400 1TB HDD RAID 5
Direct HDD Backup (No tapes)
2 Petabyte, iSCSI over IB
10 racks

Infiniband/10GbE/1GbE General Purpose Grid Interconnect



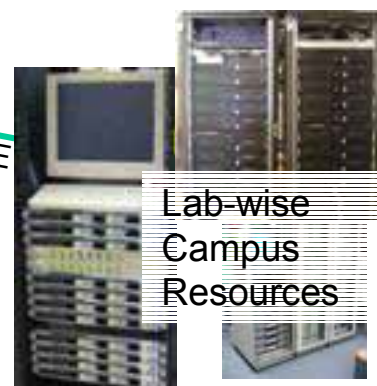
Grid Users
~6000CPU+GPU
49~64TFlops
2PBytes 70 racks
~1MegaWatt

Tightly-Coupled
1000CPU/500Core
8TFlops, 3GB/s proprietary
interconnect
10 racks



Graphics Acceleration
250GPU 10TFlops(single precision)
Infiniband 4x high bisection
10 racks

Intra-Campus Fiber
10Gbps via IB/10GbE



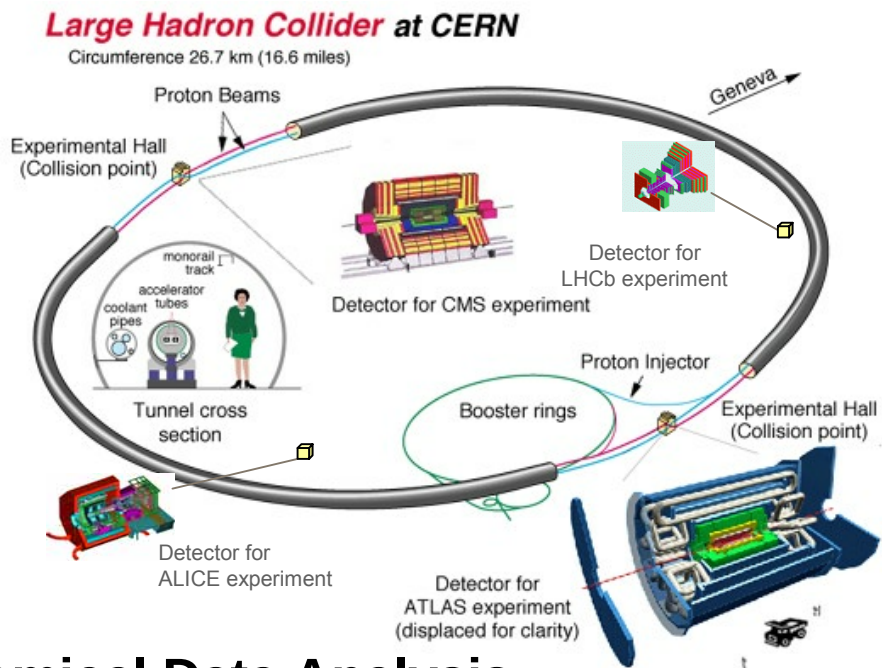
Lab-wise
Campus
Resources

DataGrid and the Grid
DataFarm Project
-Scaling to PetaBytes-

[Background] Petascale Data Intensive Computing

High Energy Physics

- CERN LHC, KEK Belle
  - ~MB/collision,
    100 collisions/sec
  - ~PB/year
  - 2000 physicists, 35 countries



Astronomical Data Analysis

- data analysis of whole the data
- TB~PB/year/telescope
- SUBARU telescope
  - 10 GB/night, 3 TB/year

[Background 2] Large-scale File Sharing

- P2P – exclusive and special-purpose approach
  - Napster, Gnutella, Freenet, …
- Grid technology – file transfer, metadata management
  - GridFTP, Replica Location Service
  - Storage Resource Broker (SRB)
- Large-scale file system – general approach
  - Legion, Avaki [Grid, no replica management]
  - **Grid Datafarm** [Grid]
  - Farsite, OceanStore [P2P]
  - AFS, DFS, . . .

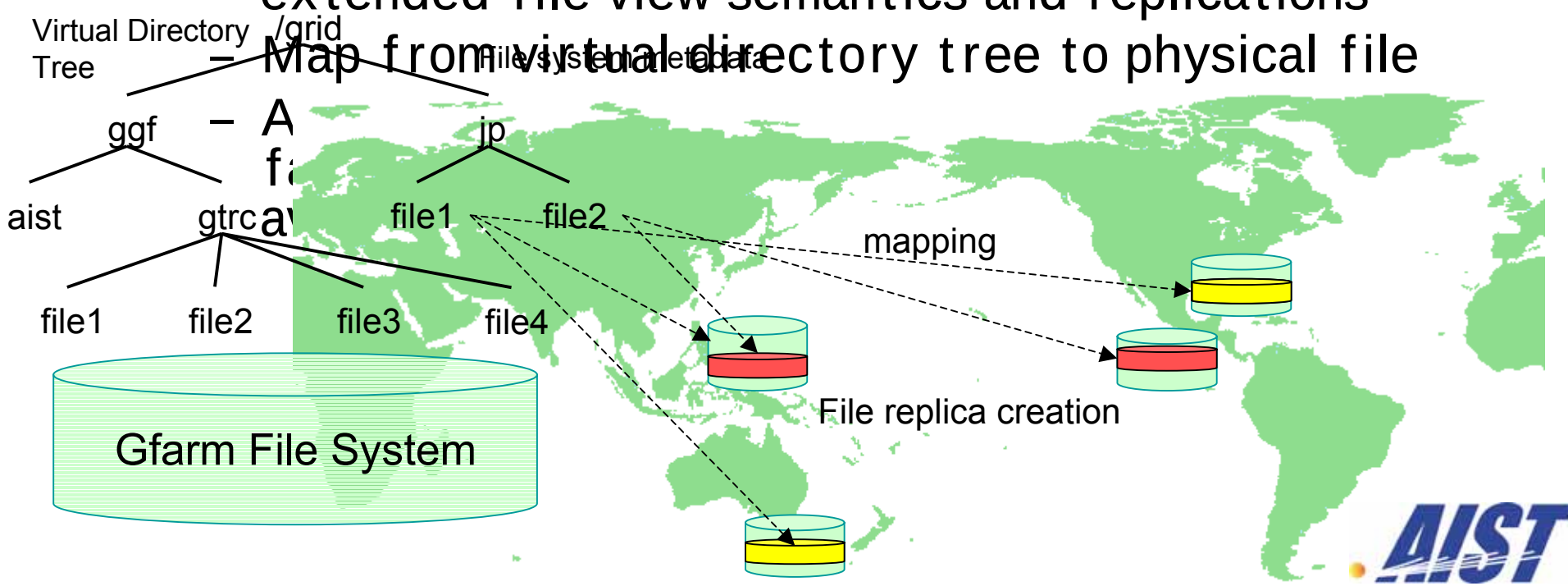
Goal and feature of AIST/Titech/KEK Grid Datafarm

- Goal
  - Dependable data sharing among multiple organizations
  - High-speed data access, High-speed data processing

- Grid Datafarm
  - Grid File System – Global dependable virtual file system
    - Integrates CPU + storage
  - Parallel & distributed data processing
    - similar to MPI-IO
- Features
  - Secured based on Grid Security Infrastructure
  - Scalable depending on data size and usage scenarios
  - Data location transparent data processing (file affinity scheduling)
  - Automatic and transparent replica access for fault tolerance
  - High-performance data access and processing by accessing multiple dispersed storages in parallel

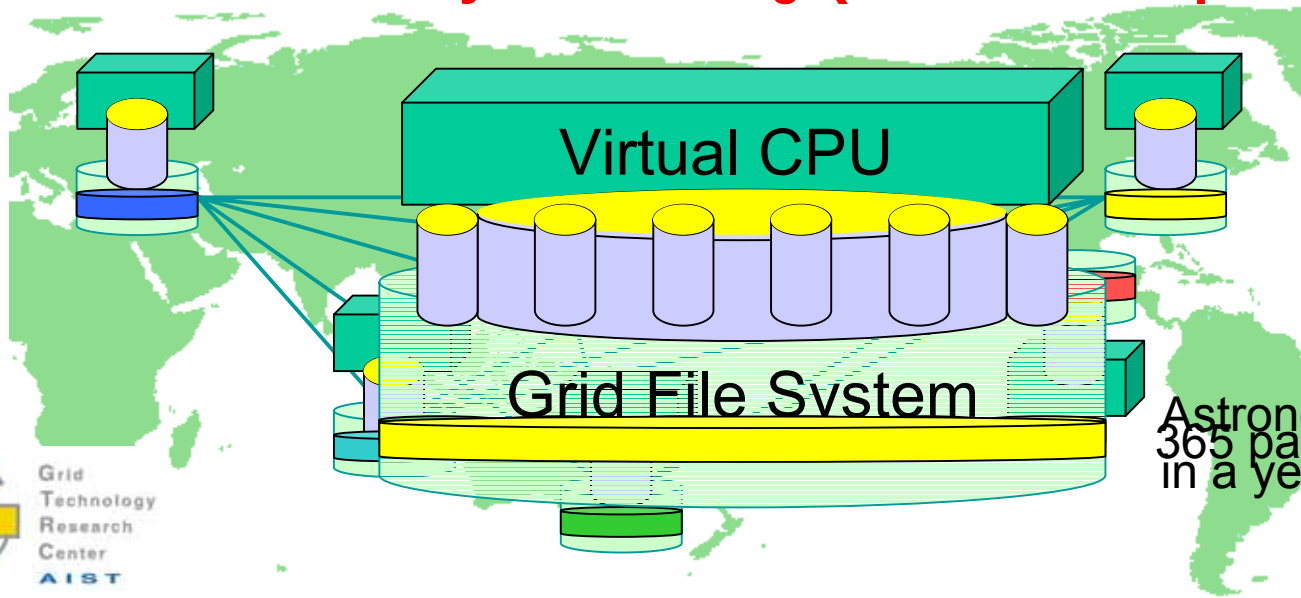
Grid Datafarm (1): Gfarm file system - World-wide virtual file system [CCGrid 2002]

- Transparent access to dispersed file data in a Grid
  - POSIX I/O APIs, and native Gfarm APIs for extended file view semantics and replications
  - Map from virtual directory tree to physical file
  - A
    f
    a



Grid Datafarm (2): High-performance data access and processing support [CCGrid 2002]

- World-wide parallel and distributed processing
  - Aggregate of files = superfile
  - Data processing of superfiles = parallel and distributed data processing of member files
    - **Local file view (SPMD parallel file access)**
    - **File-affinity scheduling ("Owner-computes")**



World-wide Parallel & distributed processing

Astronomic archival data 365 parallel analysis in a year (superfile)

Grid Technology Research Center AIST

AIST

Example: Gfarm replication command

- gfwhere – inquiry of replica catalog
- gfrep – parallel third-party file replication
  - Target domain name
  - Target host list

```
% gfwhere gfarm:host3
0: hpc01.hpcc.jp
1: hpc02.hpcc.jp
2: hpc03.hpcc.jp
```

```
% gfrep -D apgrid.org gfarm:host3
% gfwhere gfarm:host3
0: hpc01.hpcc.jp gfm01.apgrid.org
1: hpc02.hpcc.jp gfm02.apgrid.org
2: hpc03.hpcc.jp gfm03.apgrid.org
```

Transfer technology in long fat networks

- Bandwidth and latency between US and Japan
  - 1~10 Gbps, 200~300ms RTT
- TCP acceleration
  - Adjustment of congestion window
  - Multiple TCP connections
  - HighSpeed TCP Scalable TCP FAST TCP
  - XCP (not TCP)
- UDP based acceleration
  - Tsunami UDT RBUDP atou . . .

Bandwidth prediction without packet loss

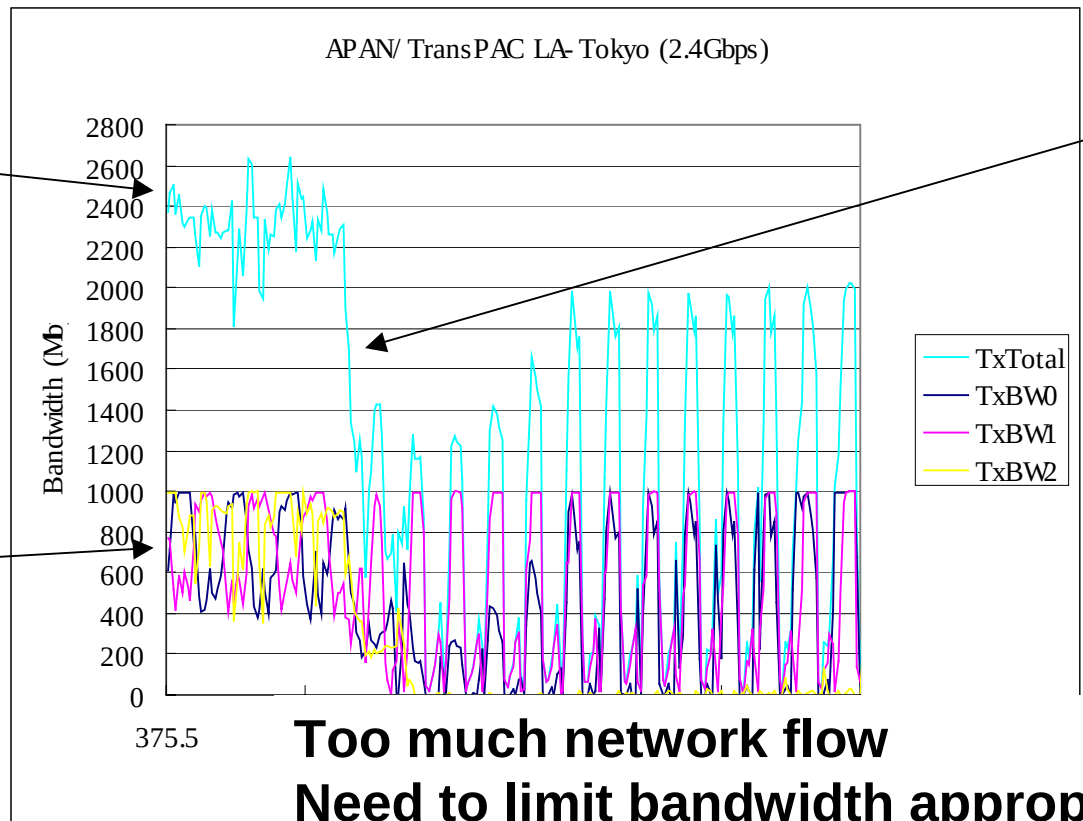
Multiple TCP streams sometimes considered harmful…

- Multiple TCP streams achieve good bandwidth, but excessively congest the network. In fact would "shoot oneself in the foot".

High oscillation
Not stable!

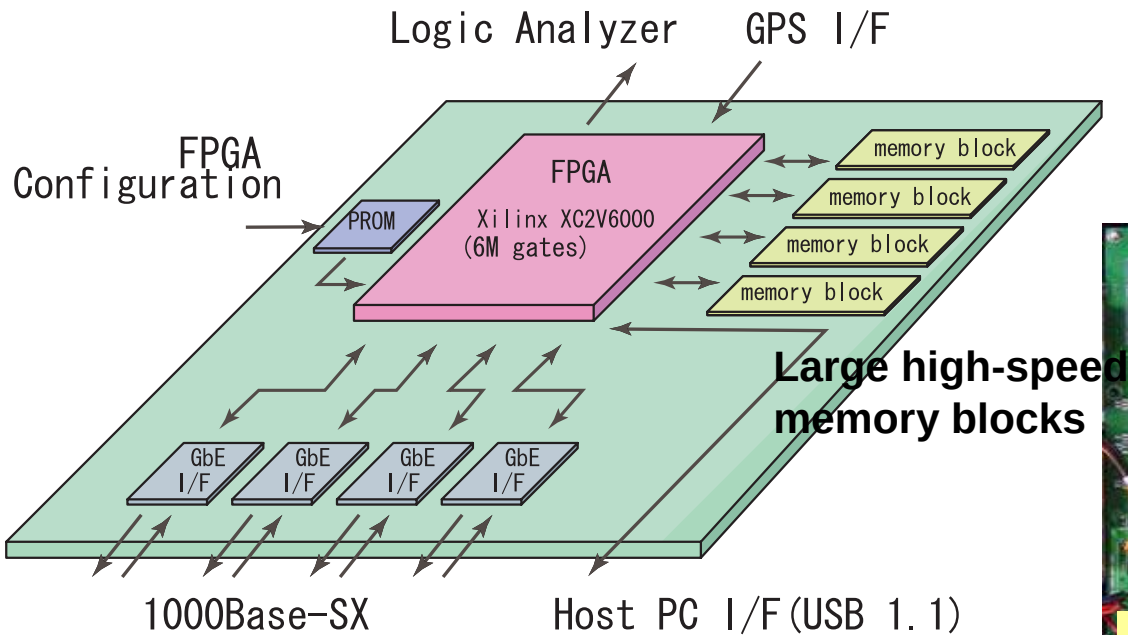
Compensate
each other

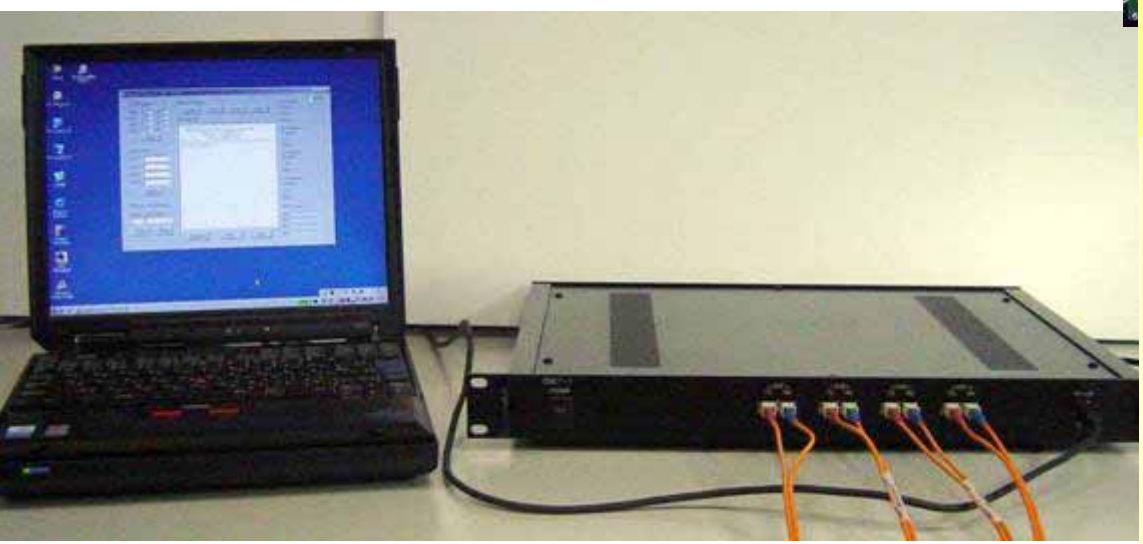
Too much
congestion



Too much network flow
Need to limit bandwidth appropriately

A programmable network testbed device GNET-1 @ AIST-GTRC





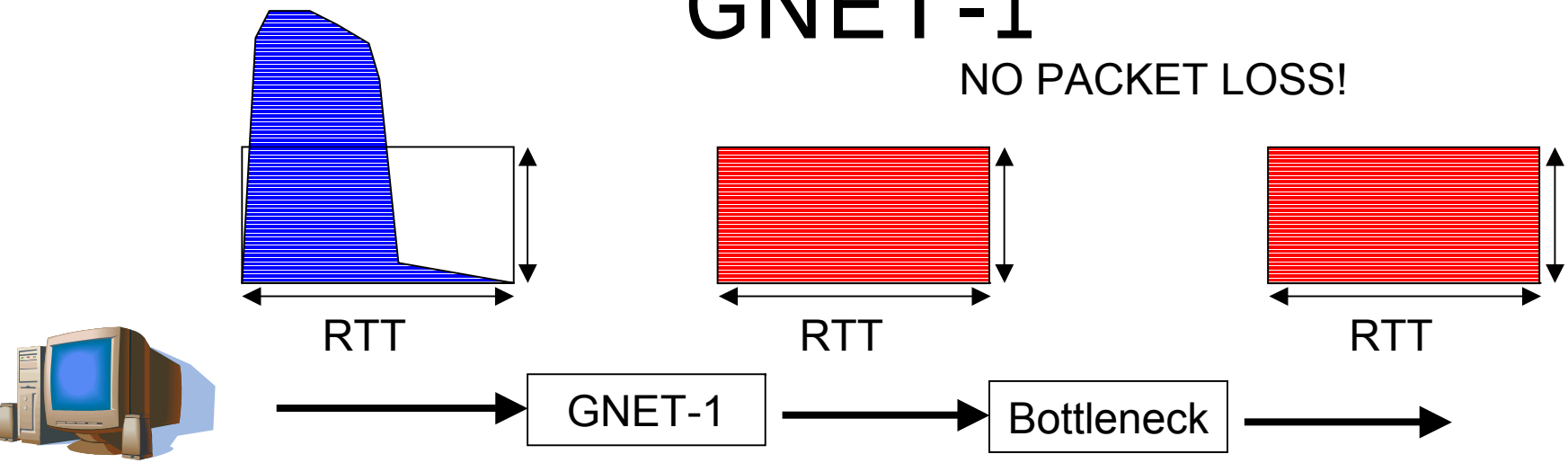
Programmable hardware network testbed

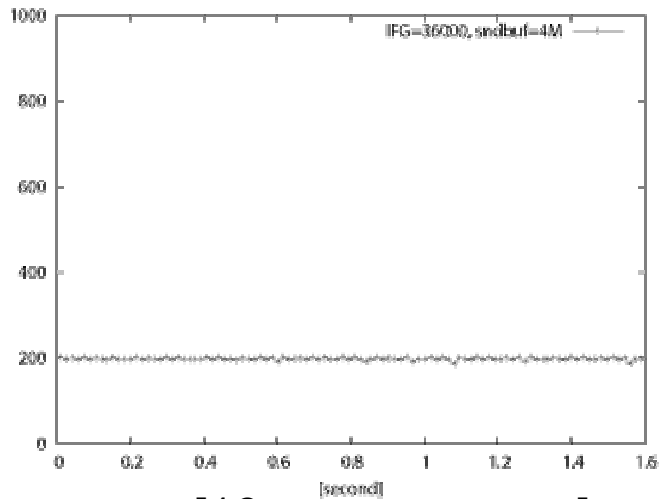
WAN emulation
- latency, bandwidth, packet loss, jitter, . . .

Precise measurement
- bandwidth in 100 usec
- latency, jitter between 2 GNET-1

General purpose, very flexible!

IFG-based pace control by GNET-1

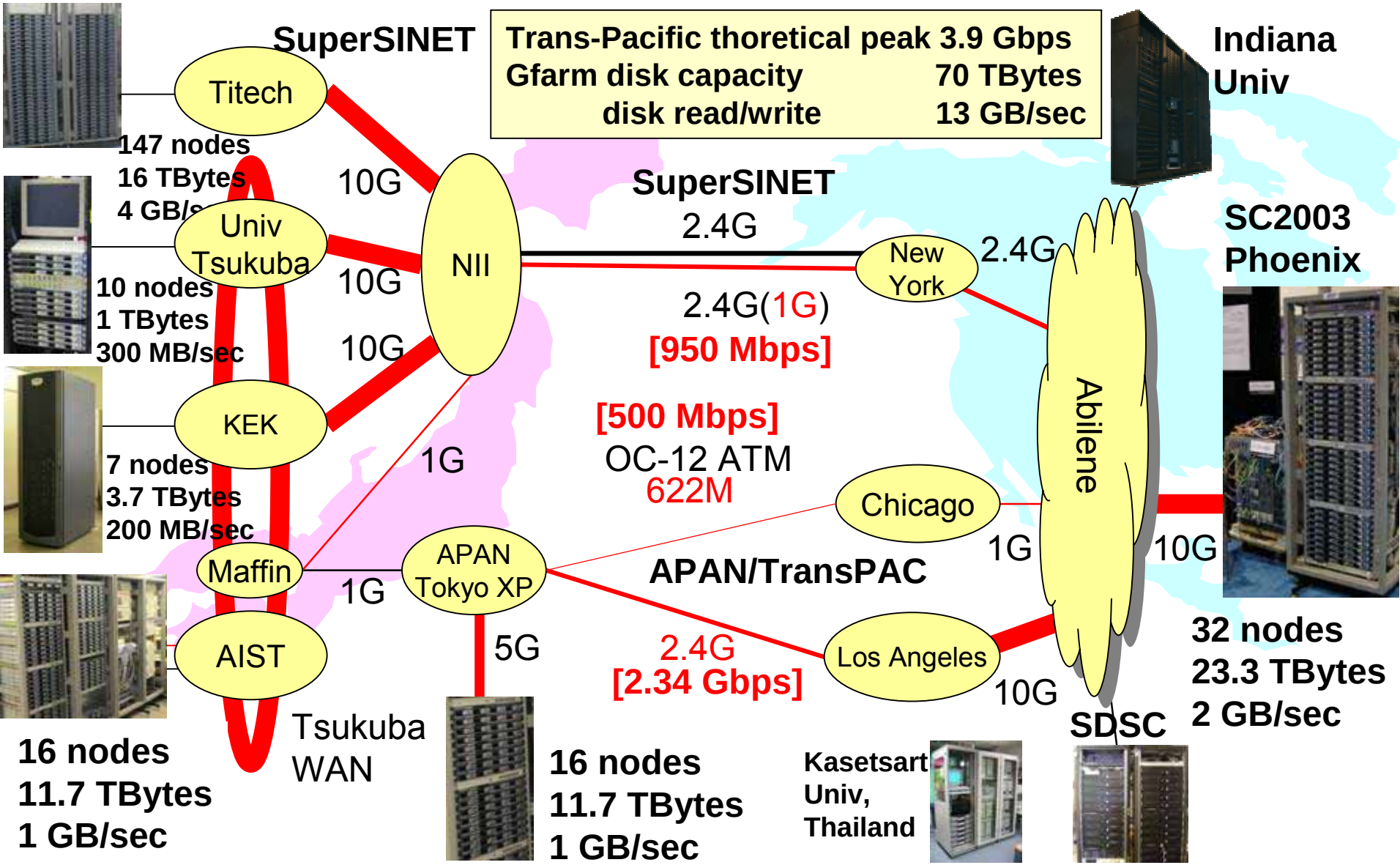




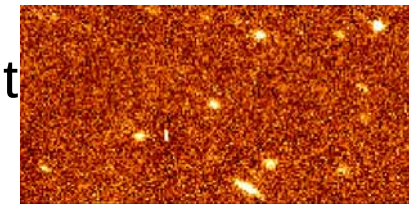
[16-msec average]

- GNET-1 provides
  - Precise traffic pacing at any data rate by changing IFG (Inter-Frame Gap)
  - Packet loss free network using large input buffer (16MB)

SC2003 Trans-Pacific Grid Datafarm testbed: Network and cluster configuration



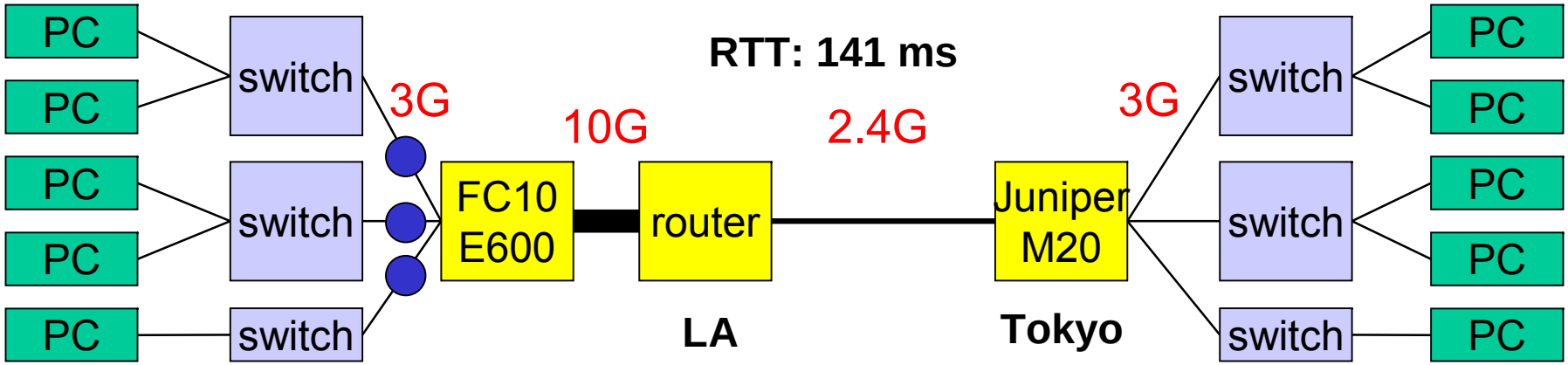
Scientific Data for Bandwidth Challenge

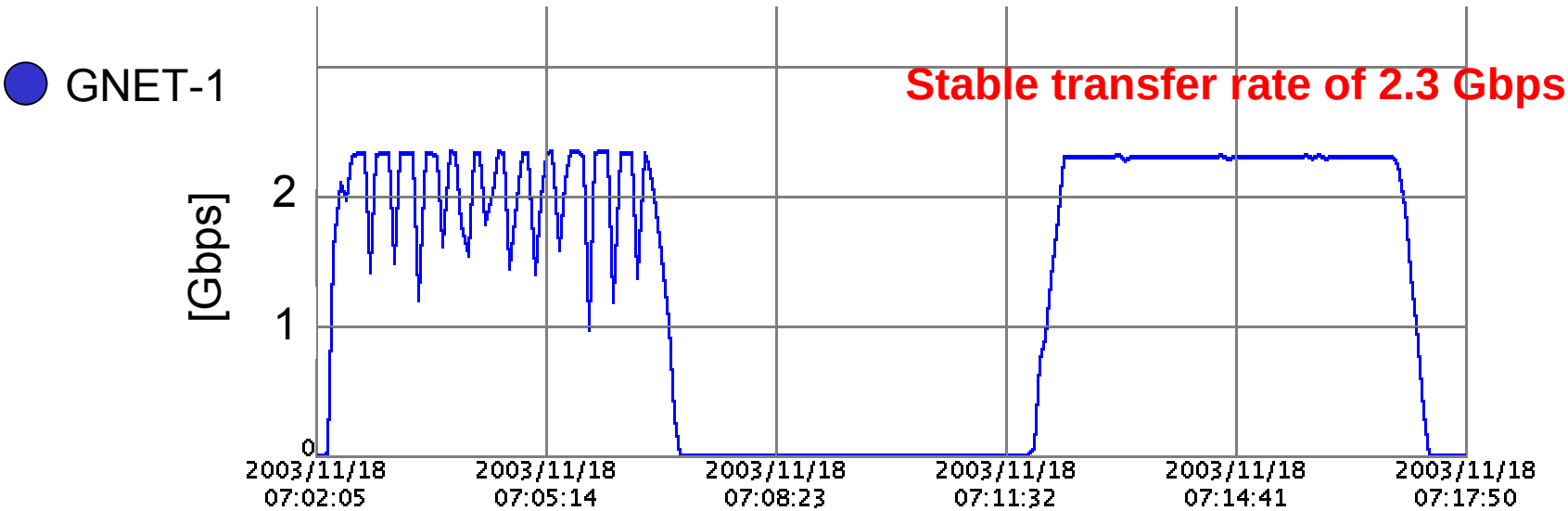


- Trans-Pacific File Replication of scientific data
  - For transparent, high-performance, and fault-tolerant access
- Astronomical Object Survey on Grid Datafarm [HPC Challenge participant]
  - World-wide data analysis on whole the archive
  - **652 GBytes** data observed by SUBARU telescope
  - N. Yamamoto (AIST)

- Large configuration data from Lattice QCD
  - Three sets of hundreds of gluon field configurations on a 24^3*48 4-D space-time lattice (3 sets x 364.5 MB x 800 = **854.3 GB**)
  - Generated by the CP-PACS parallel computer at Center for Computational Physics, Univ. of Tsukuba (300Gflops x years of CPU time) [Univ Tsukuba Booth]

Network bandwidth in APAN/TransPAC LA route **[file replication]**

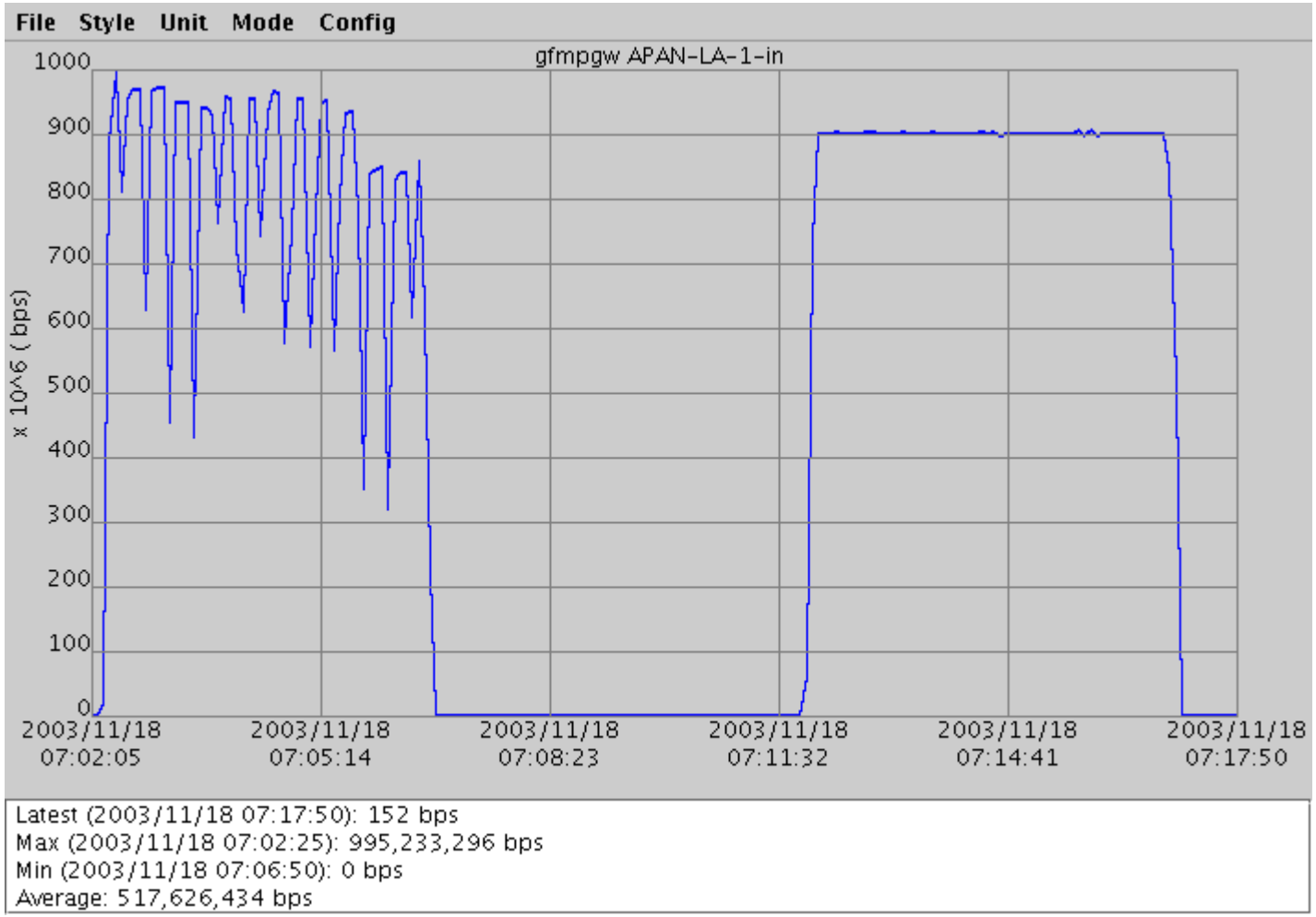




No pacing

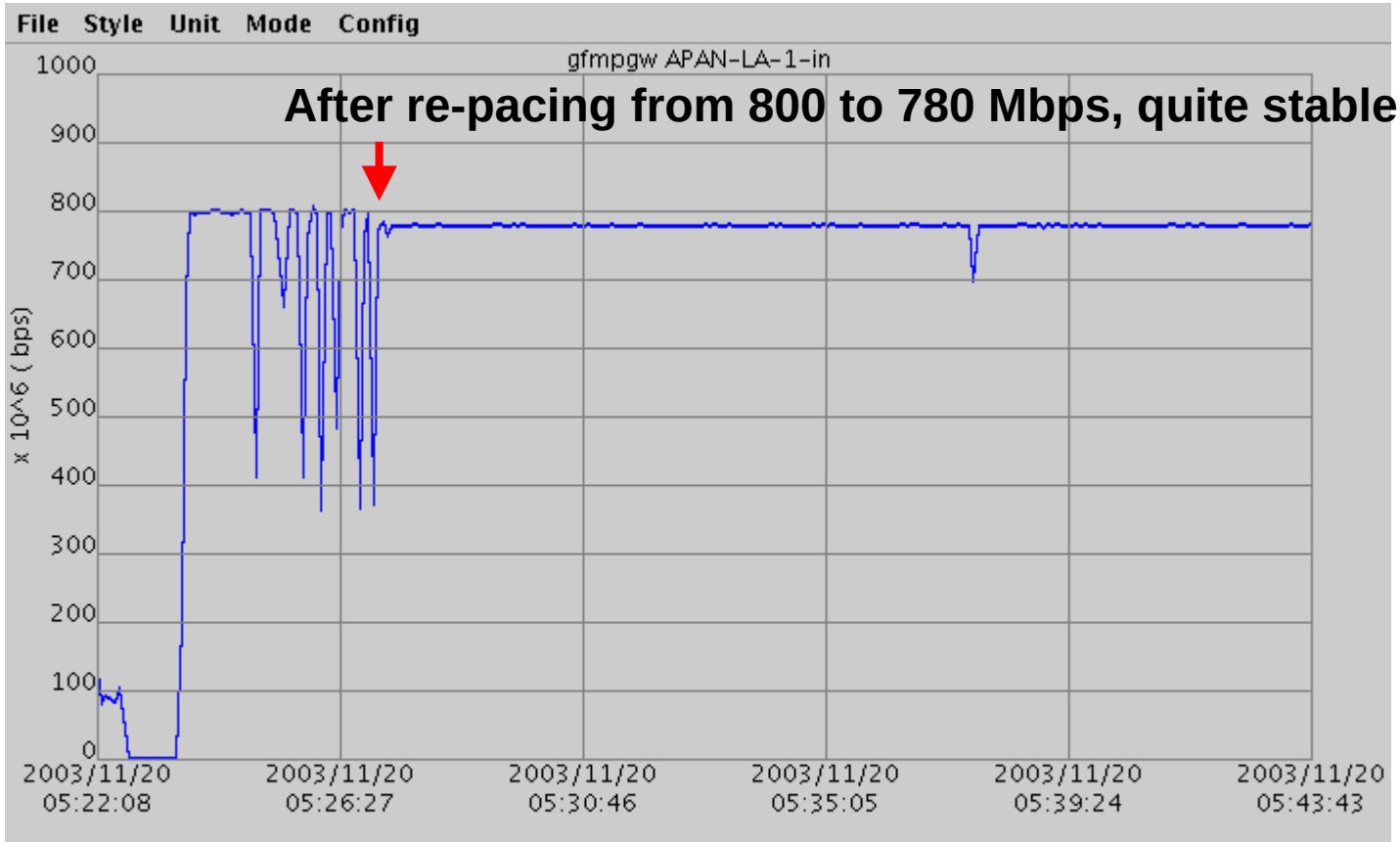
Pacing in 2.3 Gbps (900 + 900 + 500)

APAN/TransPAC LA route (1)



Latest (2003/11/18 07:17:50): 152 bps
Max (2003/11/18 07:02:25): 995,233,296 bps
Min (2003/11/18 07:06:50): 0 bps
Average: 517,626,434 bps

APAN/TransPAC LA (2)



Latest (2003/11/20 05:43:43): 781,499,352 bps
Max (2003/11/20 05:25:58): 806,293,096 bps
Min (2003/11/20 05:22:43): 0 bps
Average: 717,879,096 bps

Scaling to the Millions
The MegaScale Federation of Low Power, Very Dense, High Performance Clusters

The JST-CREST "Megascale" Project

- Q: How do we achieve "PetaFlops" "Megascale CPUs?"
- PI: Hiroshi Nakashima (Toyohashi IT) – Megascale Cluster Federation (Grid) programming
- Co-Pis
  - Hiroshi Nakamura (U-Tokyo) – Low power processor architecture
  - Mitsuhisa Sato (Univ. Tsukuba) – Low power compiler and runtime
  - Taisuke Boku (Univ. Tsukuba) – Dependable multi-way interconnect
  - Satoshi Matsuoka (Titech) – Dependable and Autonomous Cluster Middleware

Ideas:

1. Low power, dense cluster design thru architecture and software efforts

2. Cluster scalability and dependability and autonomy up to 1000 "dense" nodes

3. Federate 1000 clusters via Grid-level programming

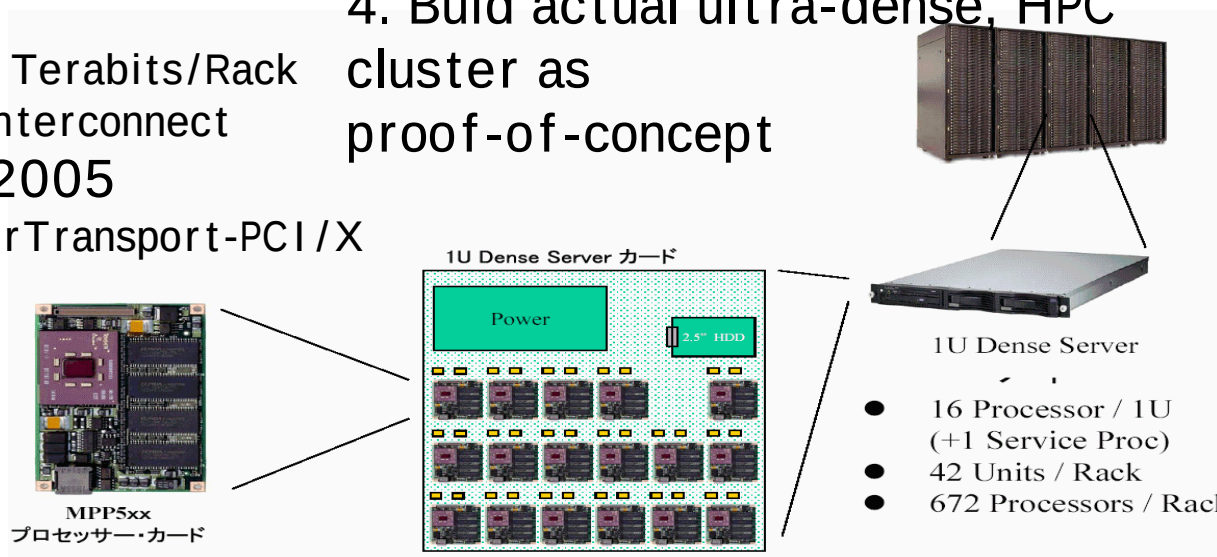
4. Buid actual ultra-dense, HPC cluster as proof-of-concept

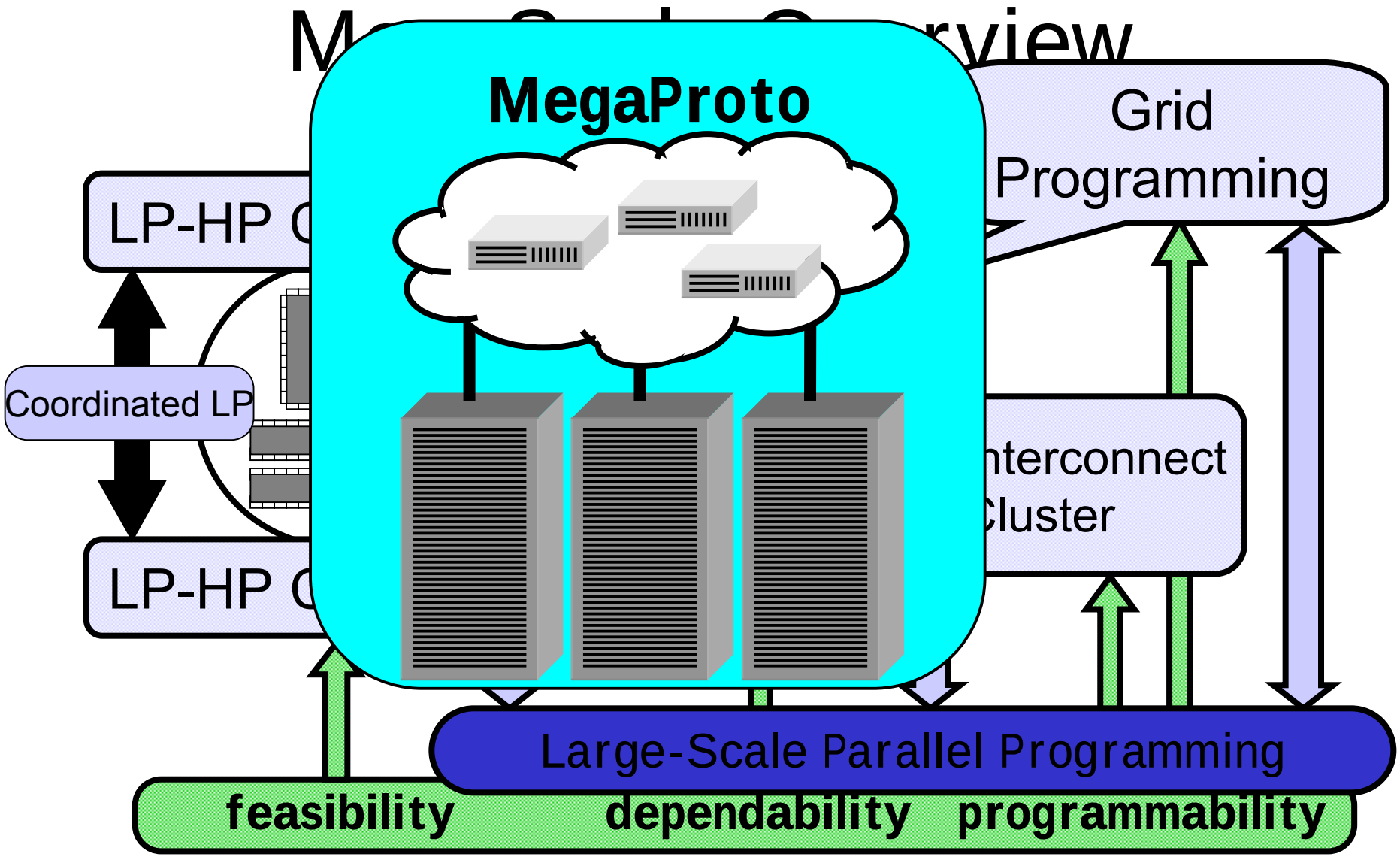
Megascale Cluster Prototype 1Q 2004

- 670GigaFlops, 150GB/Rack
- 2Gbps Interconnect/node, 0.6 Terabits/Rack
- Combined SAN/Cluster/Grid interconnect

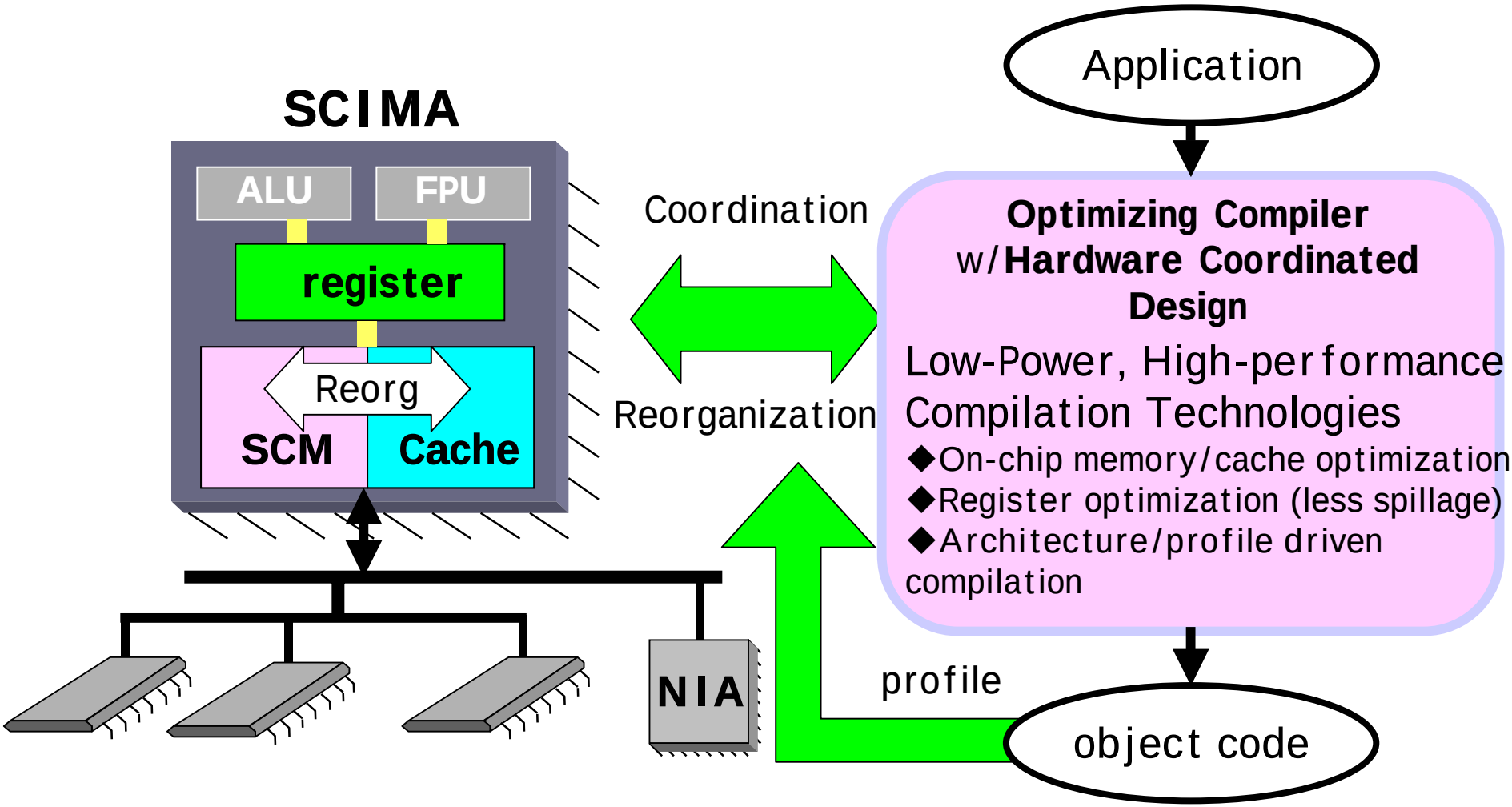
2nd Megascale Prototype 2005

- Processor Card Upgrade, HyperTransport-PCI/X
- 1.5 TeraFlops, 300GB/Rack
- 1.7 TeraByte/s Mem BW/Rack

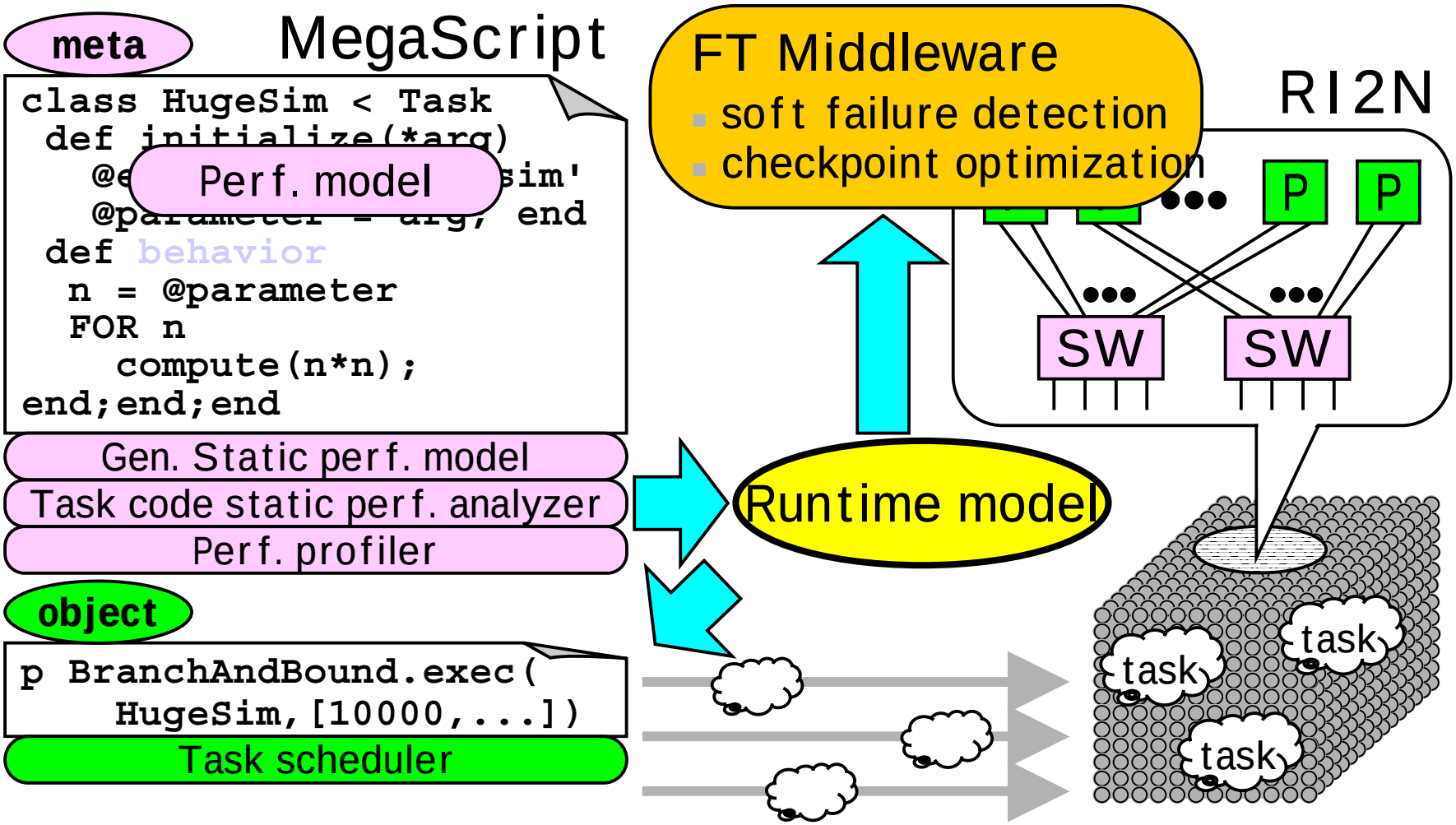



MegaProto
Grid
Programming
LP-HP C
Coordinated LP
nterconnect
Cluster
LP-HP C
Large-Scale Parallel Programming
feasibility
dependability
programmability

MegaScale Overview (2)



MegaScale Overview (3)



MegaProto Design Objectives (1)

- **MegaProto** Objectives
  - Project Claim Low Power, High Density
    ➔ MegaProto as Proof of Concept
  - As testbed for middleware R&D
    - Low-Power, High-Performance Compilation, RI2N Dependable Networking, Advanced Cluster Fault Modeling/Detection/Tolerance, MegaScript Grid programming
  - How do we design a prototype ASAP?

MegaProto Design Objectives (2)

- Target Power & Perf @ **(19" x 42U)**
  - Total Power **10KW**
  - Peak Performance **1TFlops**
  - ➔ Perf/Power **100MFlops/W**

ES x 10

- **How do we scale to 1PFlops**
  - As is **10MW** ➔ Within Reason
  - **X10 improvement 1MW ➔ Common Computing Center**

MegaProto Design Objectives (4)

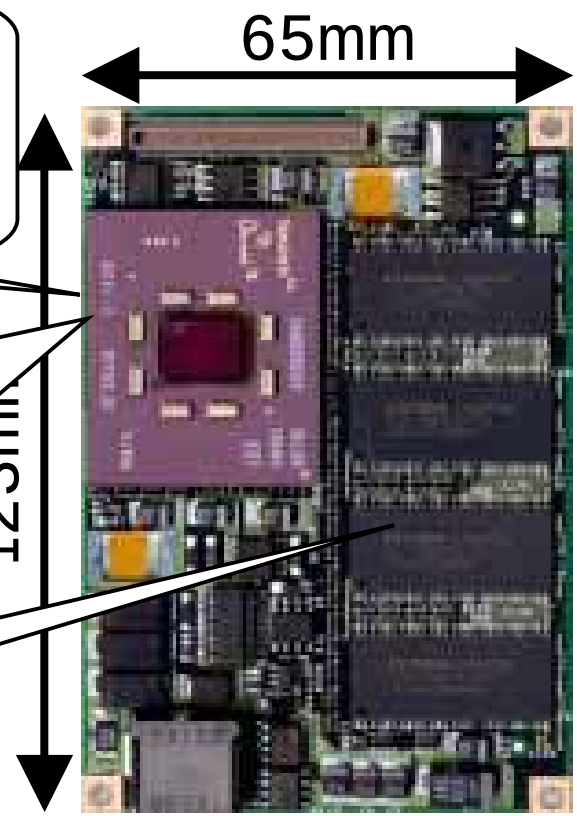
- CPU **1GFlops/7W**
  140MFlops/W
  - System **15W 66MFlops/W**
  - Rack **10KW 10KW/15W 660 proc**
  - **1U** board **16~24 proc**
    - **16proc 16 x 42U 672 MFlops @ 10KW**
    - **24proc 24** x **42U 1008 MFlops @ 15KW**
- Plan for x2 improvement in perf/power

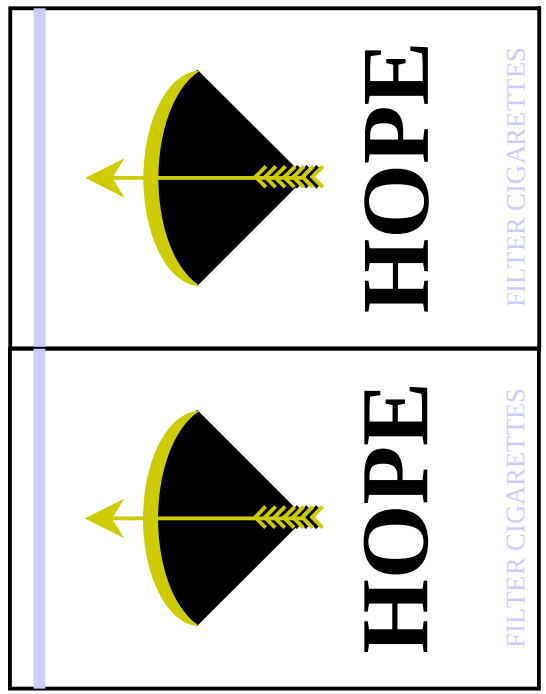
MegaProto Computing Element

- V1: TM5900
- 0.9GFlops@8W
  112.5MFlops/W
- L2C=512KB

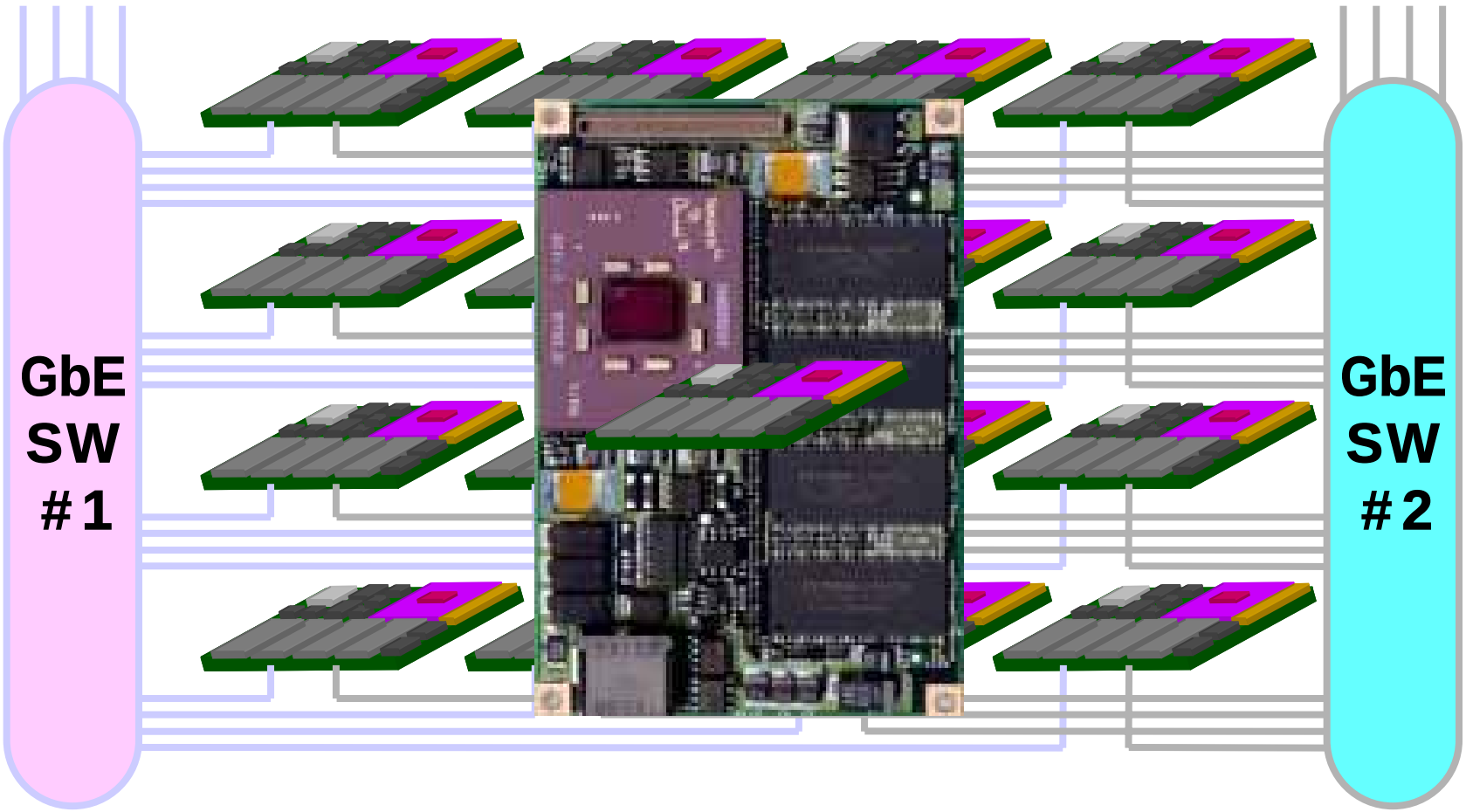
- V2: Effecion TM8000
- 2.4GFlops@8W
  300MFlops/W
- L2C=512KB~1MB

- 256MB SDRAM
- 512KB flash

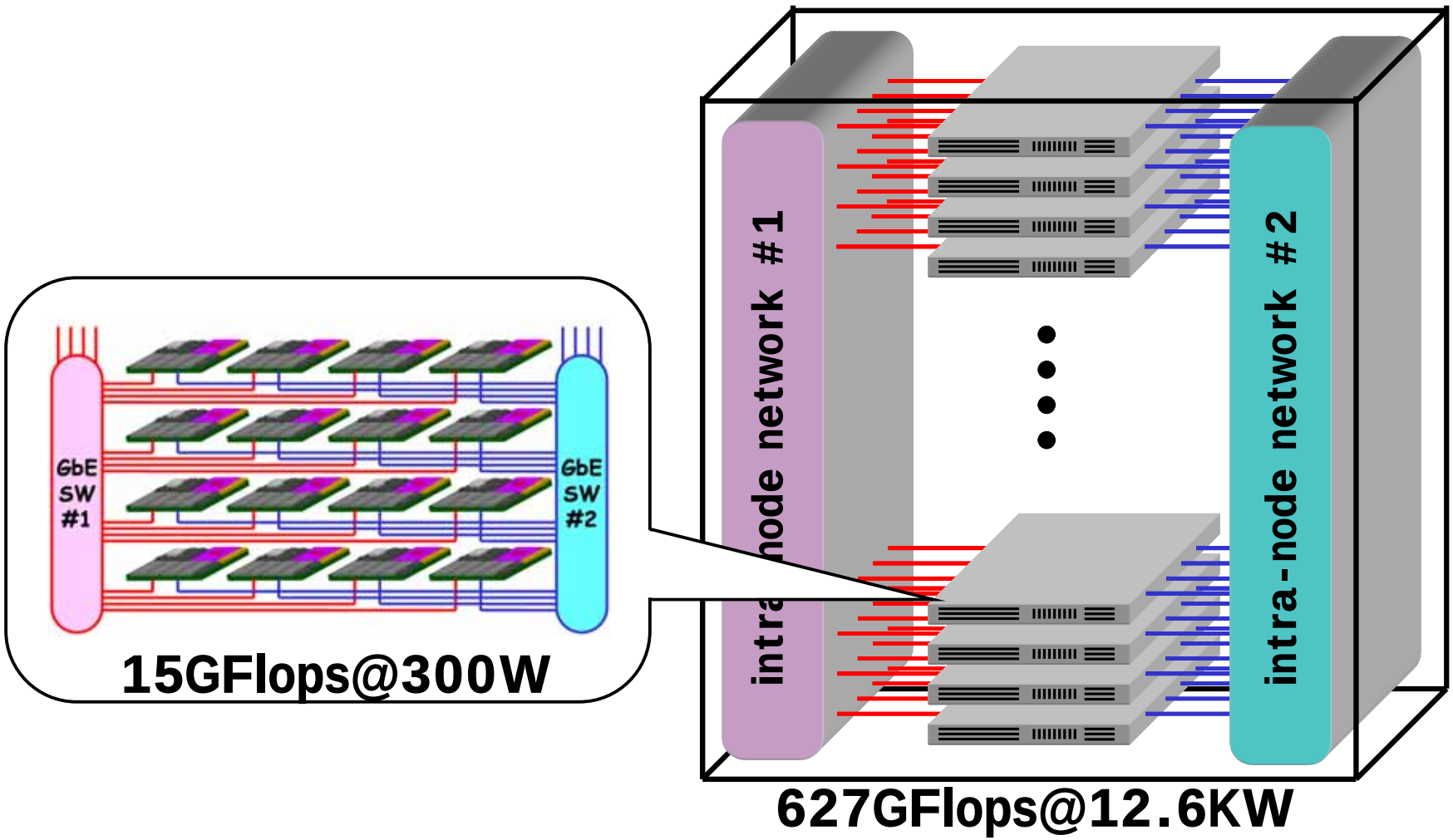




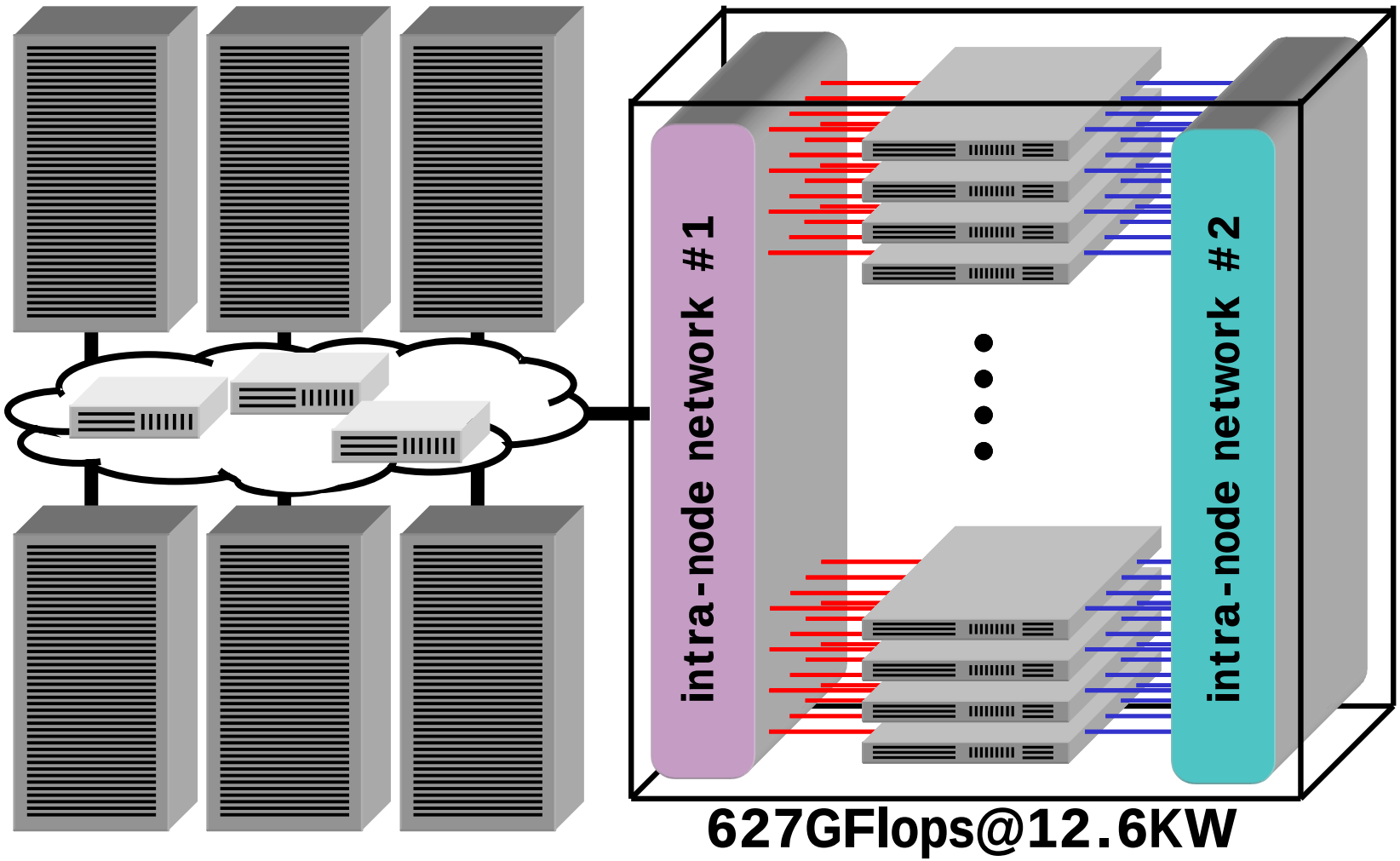
MegaProto System Layout (1U)



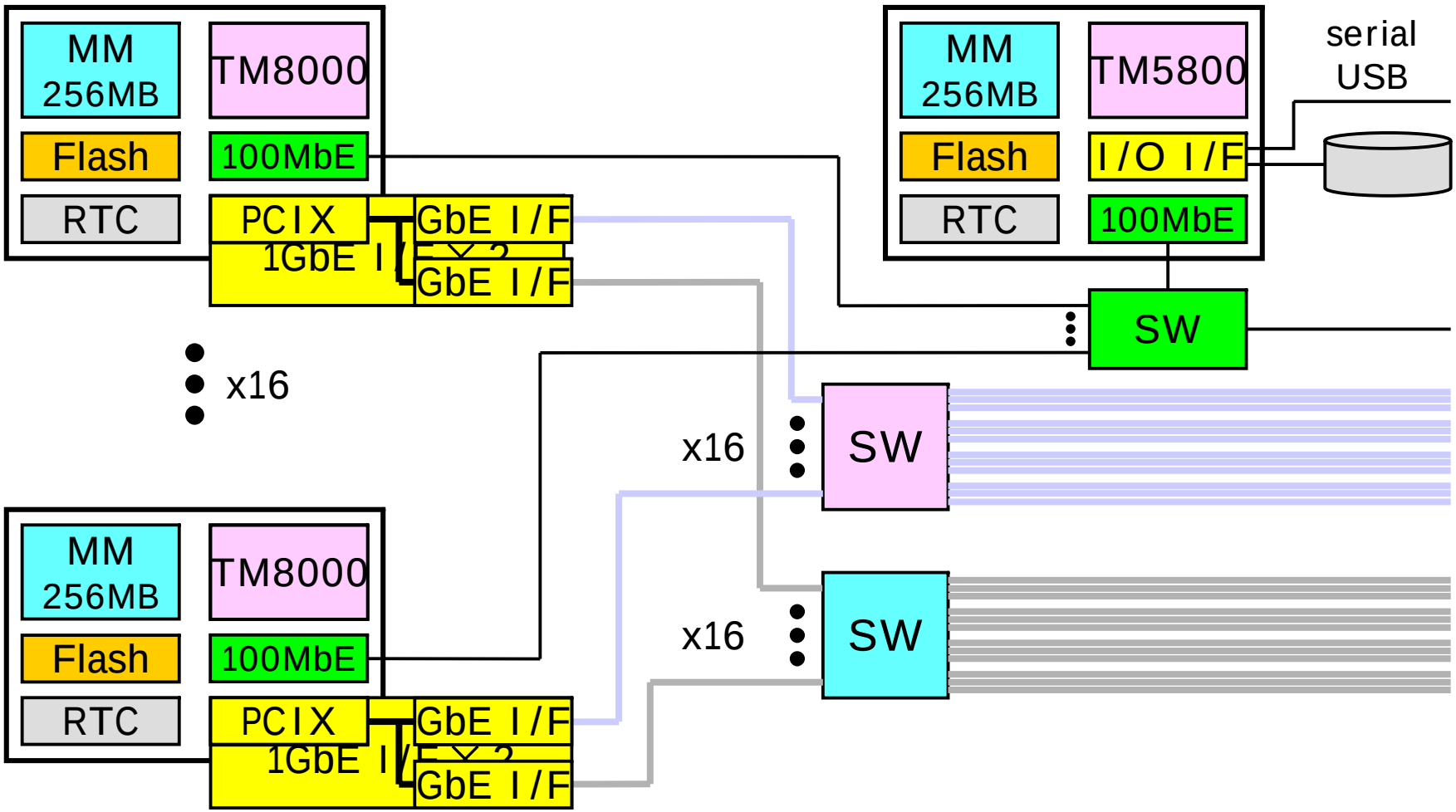
MegaProto System Layout (Rack)

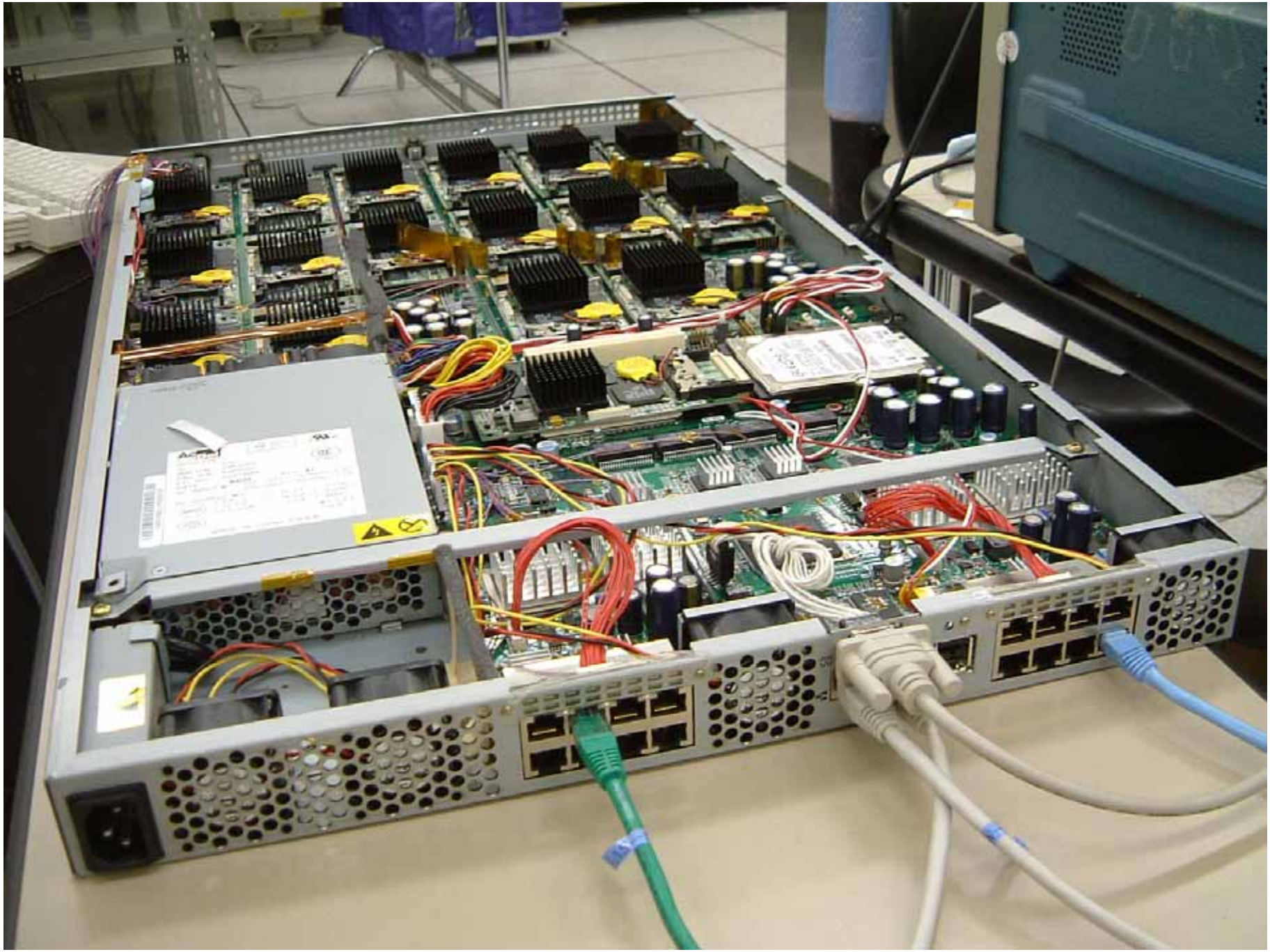


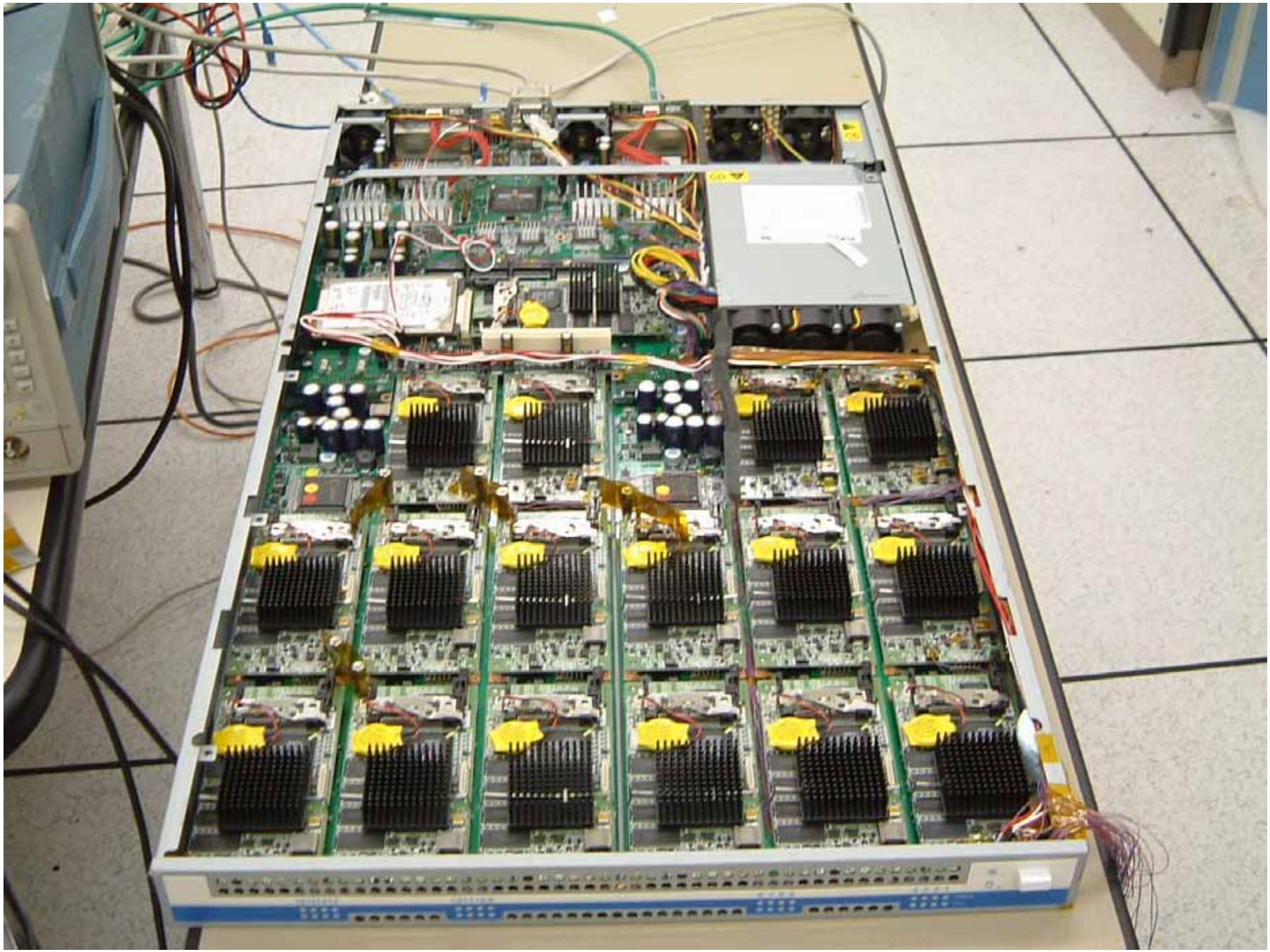
MegaProto System Layout (Multi-Rack)

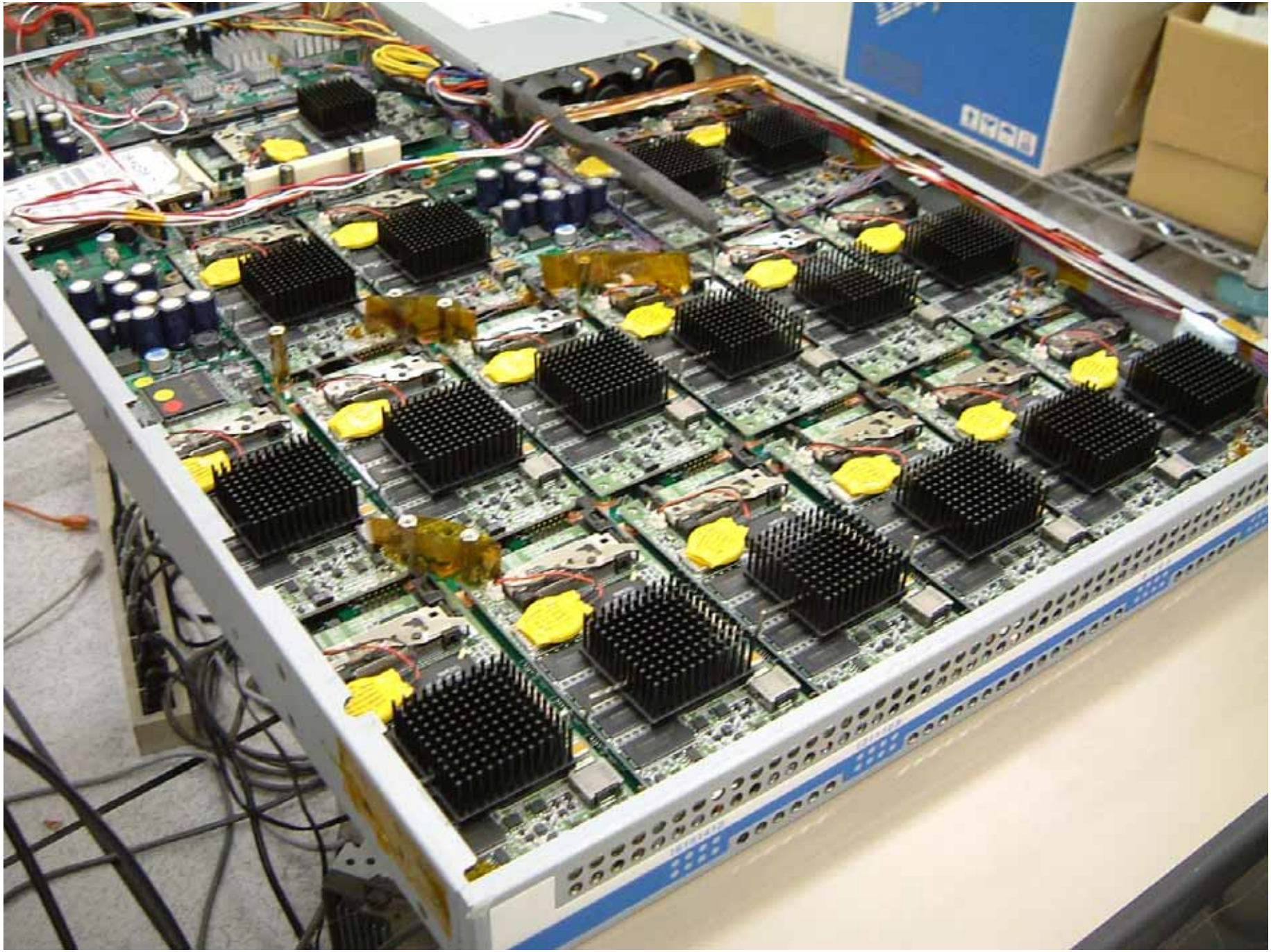


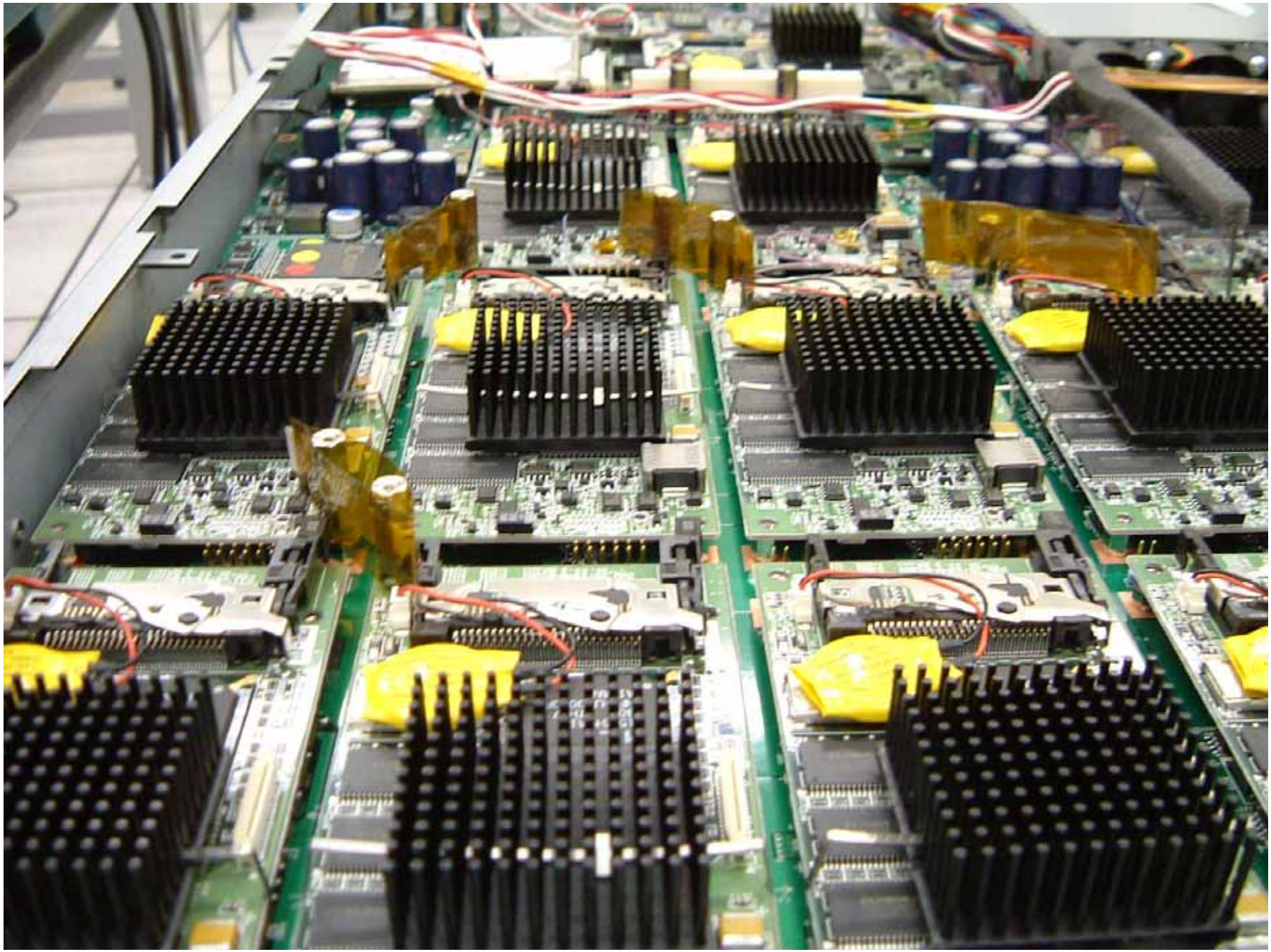
MegaProto 1U Block Diagram











Current Status (1)

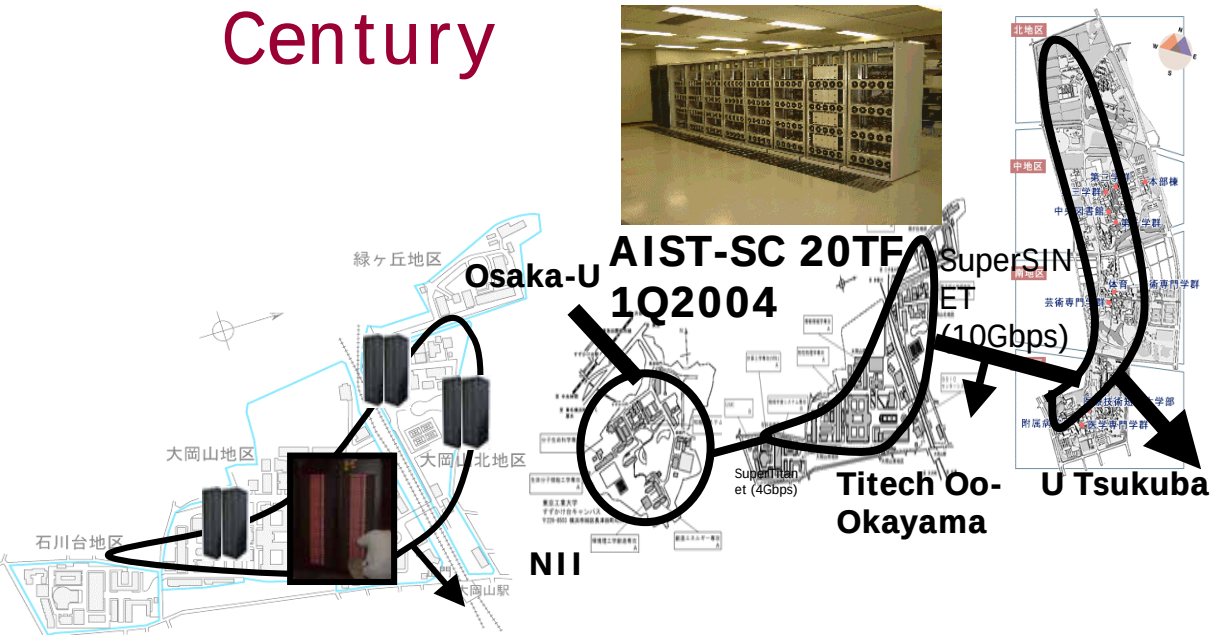
- **MegaProto#0**
  - **(TM5800@933MHz 2 x 1GbE) x 16**
    15GFlops@300W (50MFlops/W)
  - **uplink (8 x 1GbE) x 2 32Gb/s per 1U**
  - **Fabrication by IBM Japan**
  - **March 2004** Initial **2 Units (32 PE) just delivered! → used initially for compiler development**

Current Status (2)

- **MegaProto #1~2 @ PY2004~5**
  - **(TM8000 2 x 1GbE) x 16**
    30GFlops@300W (100MFlops/W)
    1.3TFlops@13KW/42U-rack
  - **#1@PY2004 16U 256PE**
  - **#2@PY2005 32U 512PE**

Summary: Towards National Grid Infrastructure

- National HEC Infrastructure for the 21st Century



2002-3
Titech Campus Grid+lab

~2000CPUs, 3-4TeraFlops
130 TeraBytes
16 sites on Campus
SuperTITANET (1-4Gbps)

2003-7
National Research Grid Infrastructure Project

10K-40K CPUs, 100-200 TeraFlops total
4-6 campuses, SuperSINET (10-40Gbps)

2007-
National Grid Infrastructure (speculated)

- **100 TeraFlops per Center**
- **> 100,000CPUs, >PetaFlop**
  10 Univ. Centers Nationwide
  (Next Generation SuperSINET)
- **Focus on and Grid R&D as plateaus towards ES**

High Performance Computing and Grid in Asia

- Submission deadline
  - January 31, 2004
  - No automatic extension
- Conference
  - July 20-22, 2004
- Venue
  - Omiya Sonic City, Tokyo Area, Japan
  - by 20-30 min train from major terminals in Tokyo
- Sponsor
  - High Performance Computing SIG, Information Processing Society of Japan
  - Co-sponsored by IEEE CS Japan Chapter

HPC Asia 2004