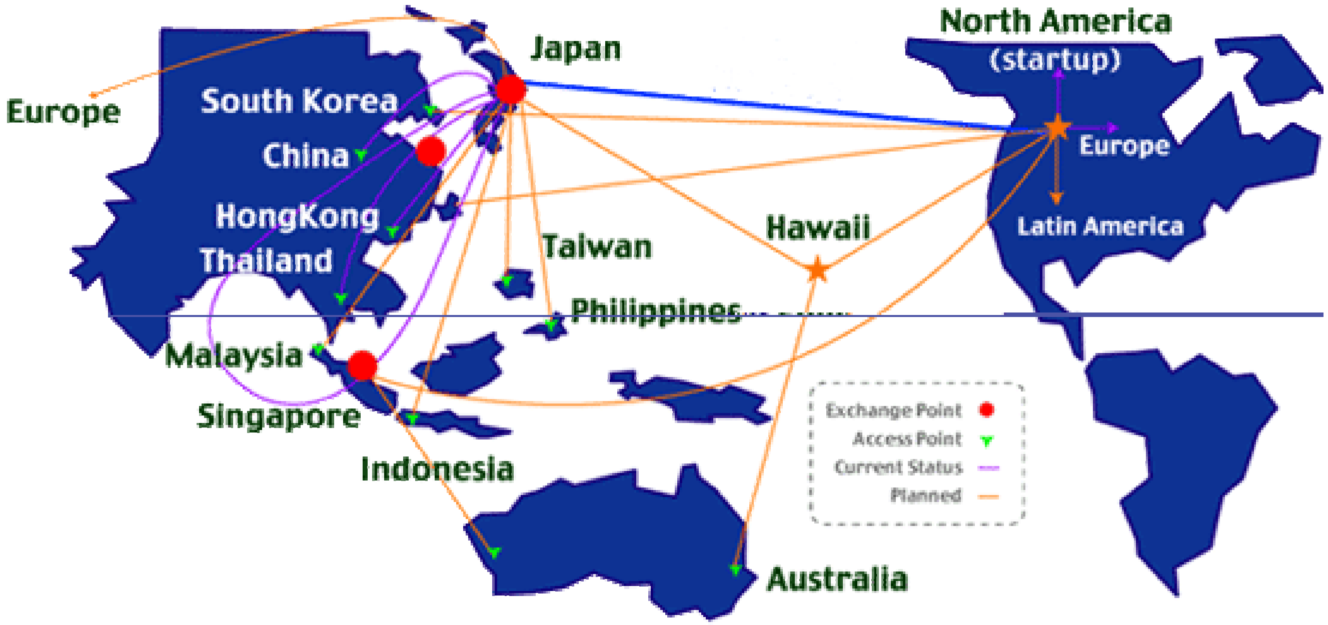
“Towards Grid and Cluster Federations”



Satoshi Sekiguchi

Director, Grid Technology Research Center,

Advanced Industrial Science and Technology, Japan

Talk Contents

- **Back ground – who am I**
- **Grids and clusters**
  - Typical usage scenario
  - Perfect test bed: AIST Super cluster
- **Challenges:**
  - Grid RPC – Ninf-G2
  - Grid MPI – GNET-1
  - Grid File System – gfarm
- **Conclusion**

Grid Technology Research Center, AIST

- **Establishment**
  - Since Jan. 1, 2002
  - 7 years term
  - 24th Research Center of AIST
- **Location**
  - Tsukuba Central
    Umezono 1-1, Tsukuba
  - Tokyo Office
    - Ueno area
    - 30 people for software development
- Engaged in developing grid middleware, applications and system technologies
- Research $$ approx. 1000M JPY

		2002/ 1H	2002/ 2H	2003 /1H
Researchers				
	Full time	**14**	**16**	**19**
	Fellowship	**1**	**8**	**9**
Collaborators		**7**	**21**	**32**
	Sub total	22	44	60
Staff				
	Administration	**2**	**1**	**1**
	Support	**5**	**7**	**9**

One of the world's foremost GRID Research Center, the largest in Japan

AIST

Historical Background

SIGMA-1 (1987-)
Dataflow machine 128PE

ELECTROTECHNICAL Laboratory

1994 Poorman's supercomputer
SunMicrosystems SS2 25MHz x 32

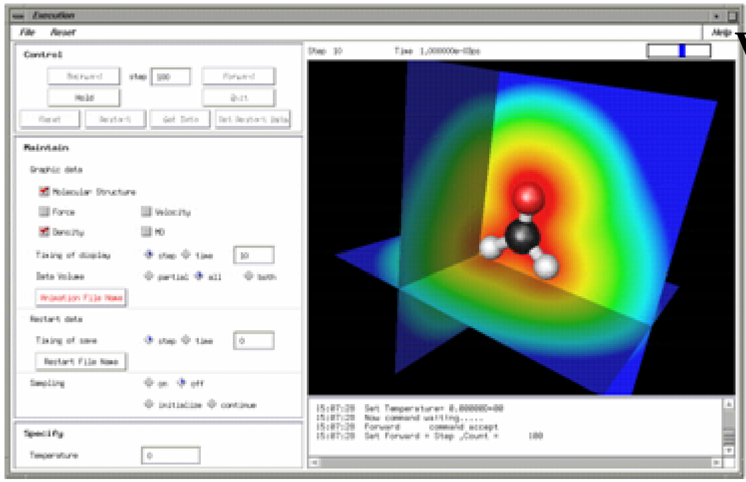
1996- Wiz cluster
DEC Alpha Station 333MHz x 33

Ninf

1994-

1997

Virtual Microscope



2000-

[redacted]

Asia-Pacific Grid

GRID DATA FARM

2001-

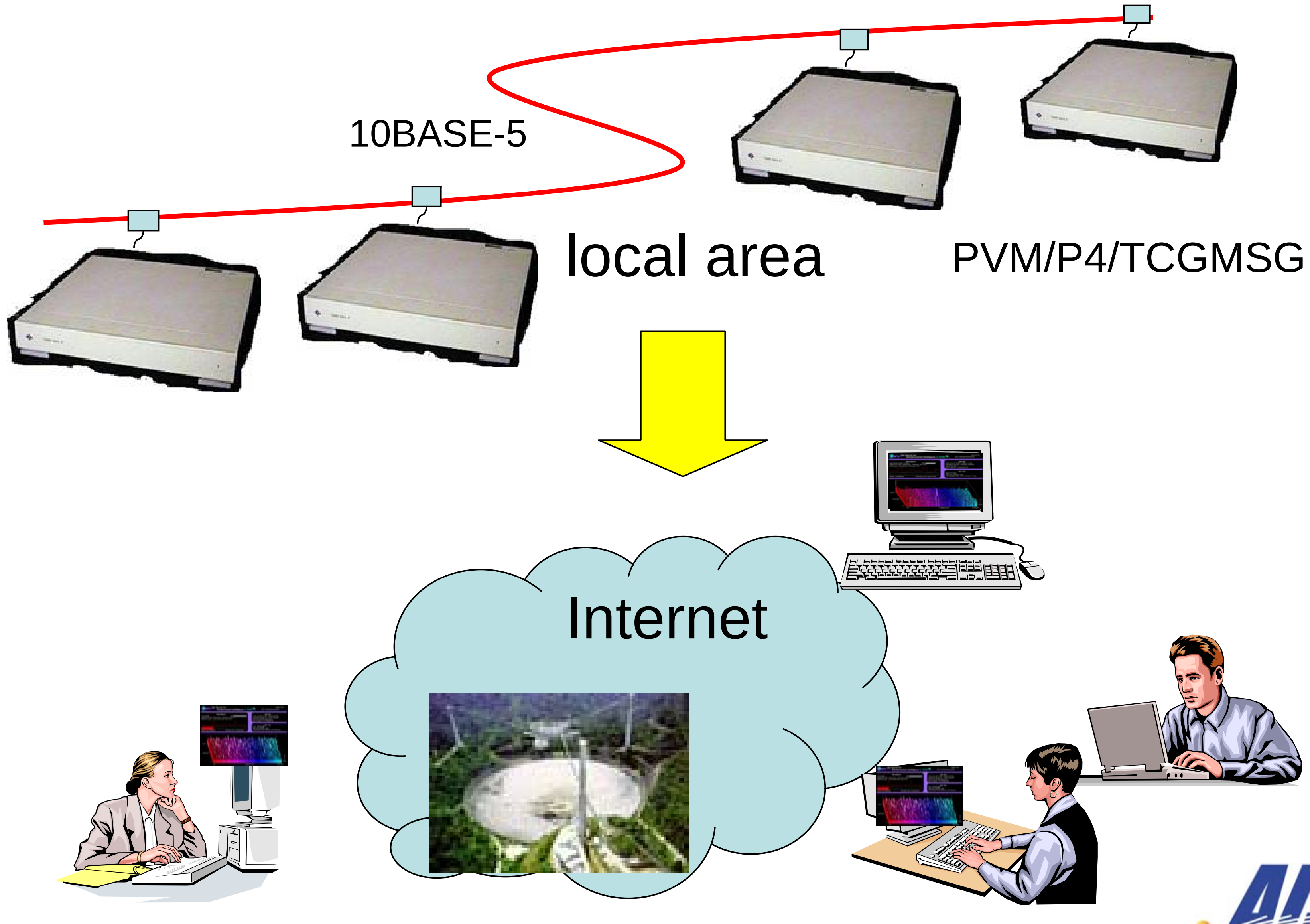
2002-
gfarm cluster

AIST

Japan Grid Cluster Federation (SC2002)

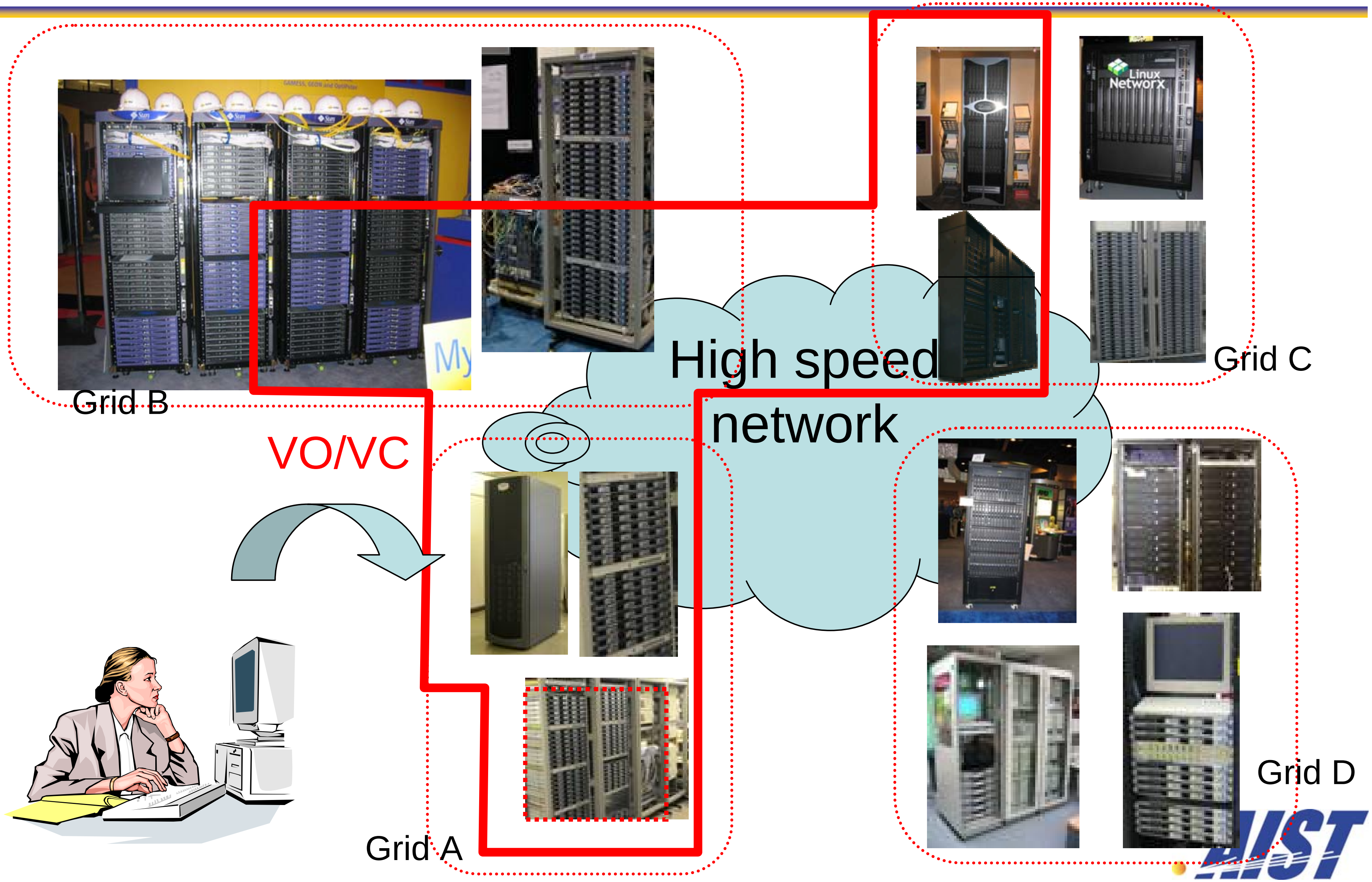
aims to build grids powered by clusters with performance aware middleware.

Cluster to Grid as a poor man's supercomputer

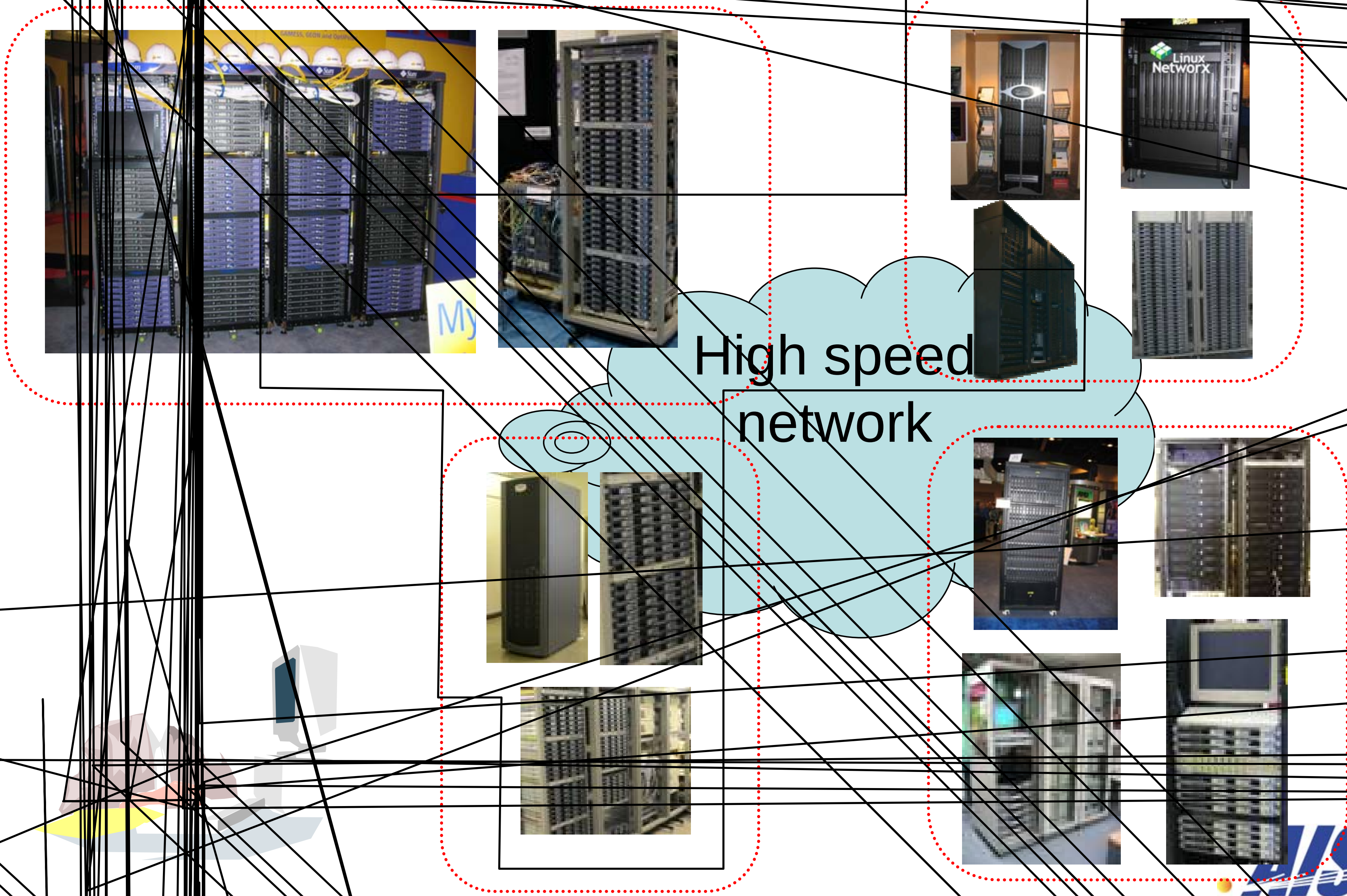


Cluster to Grid across a campus

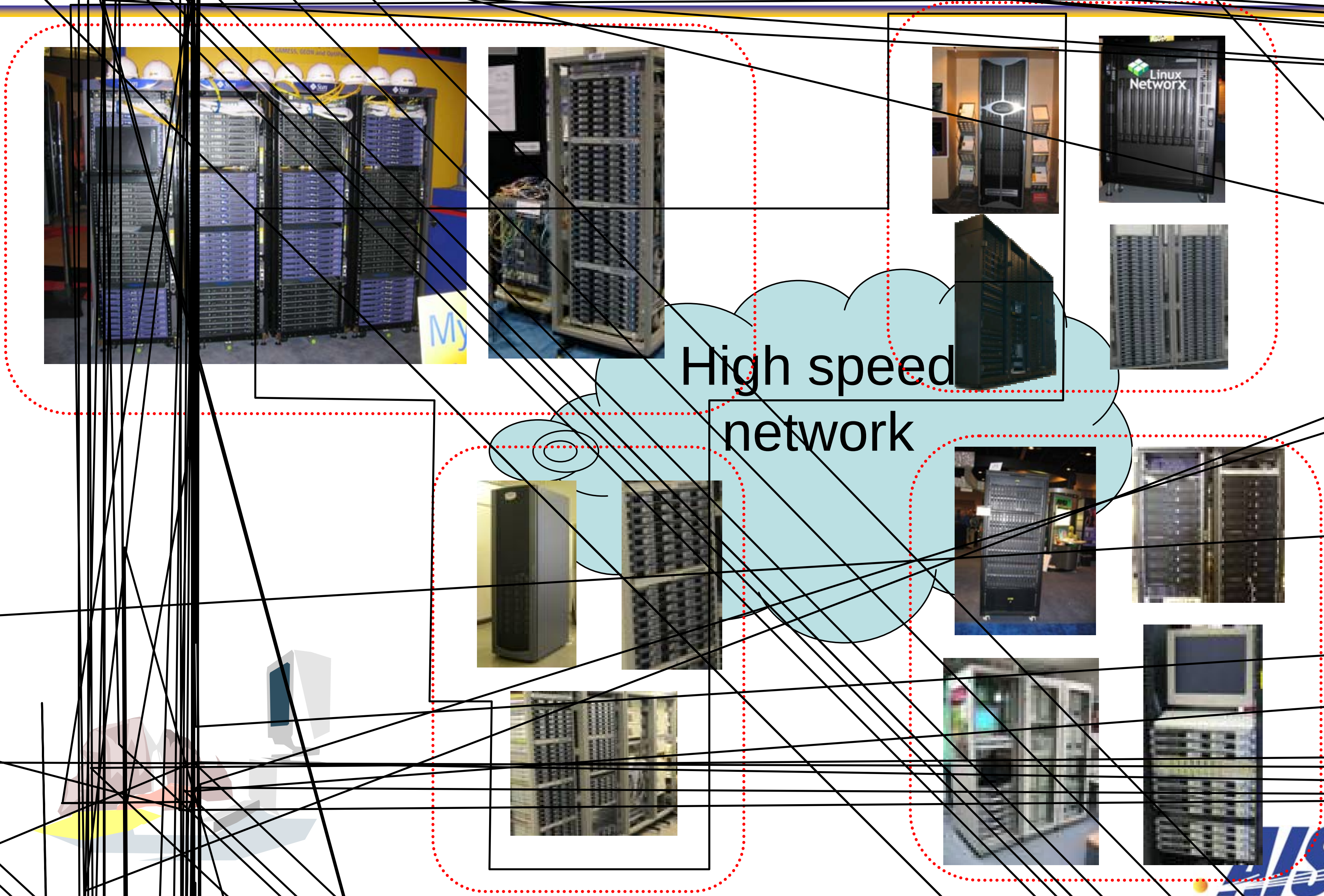
Scenario A: Develop at local, Production in the grid



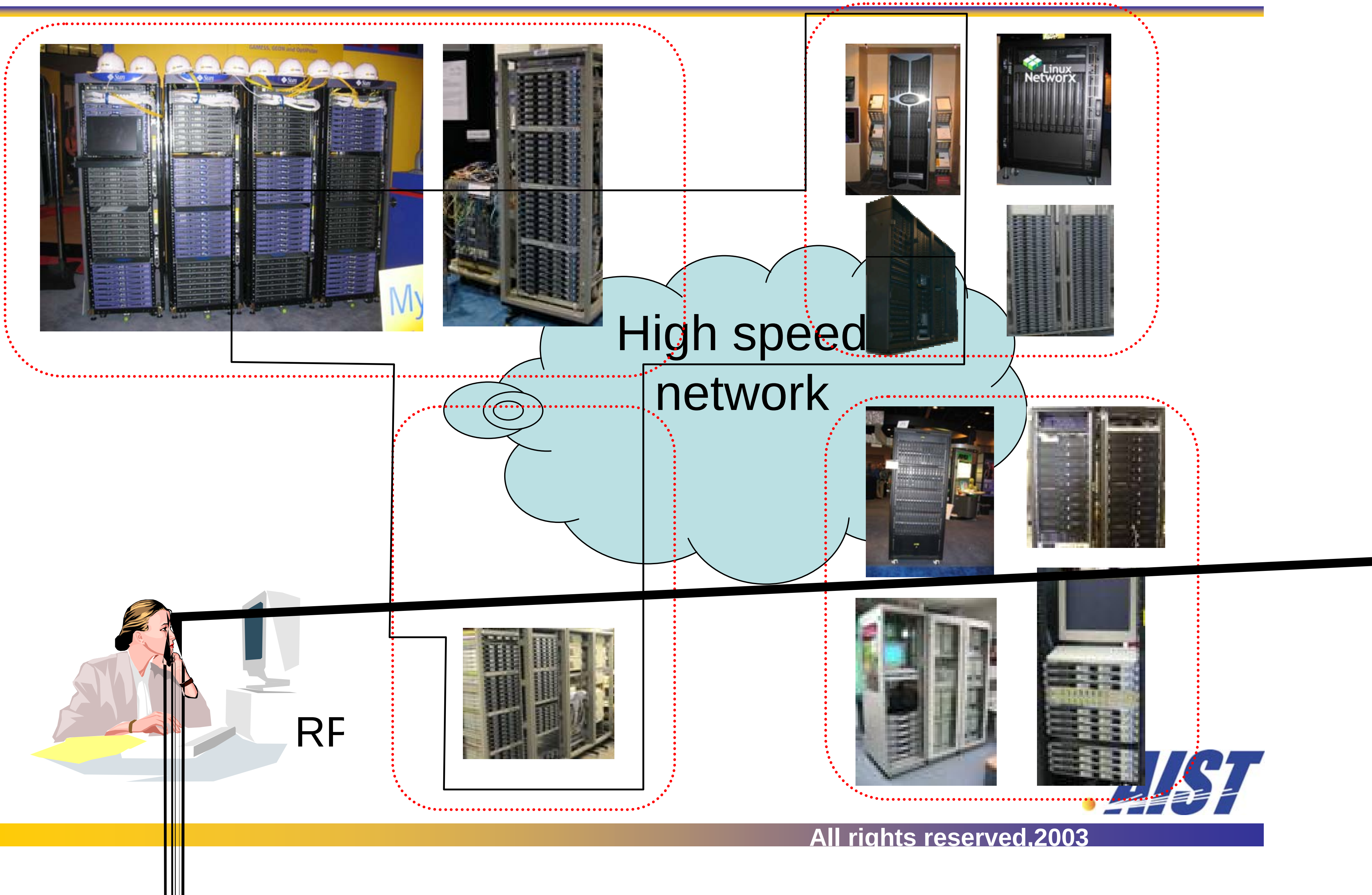
Scenario B: Distributed computing in the grid



Scenario C: Full scale computing in the grid



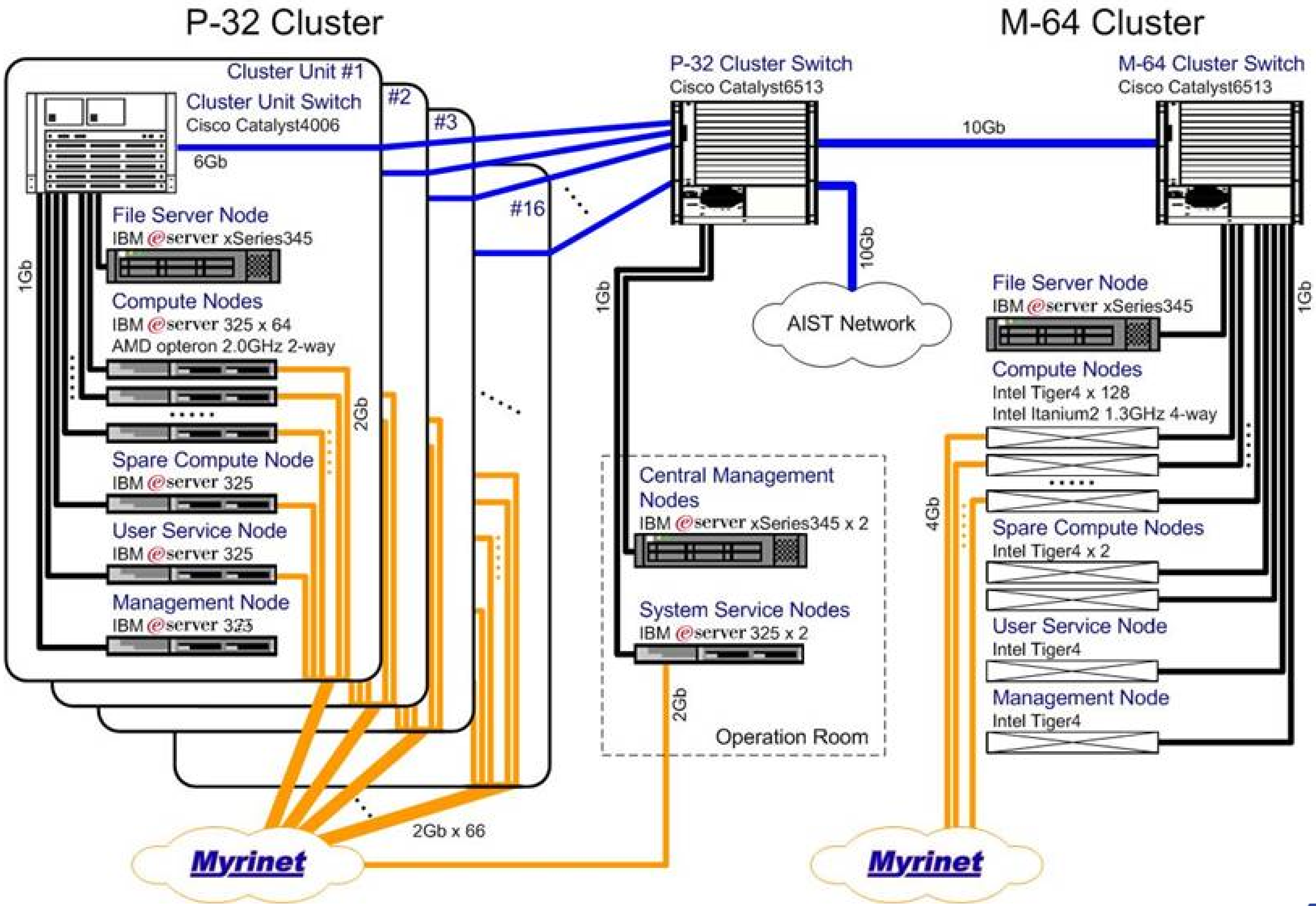
Scenario D: More flexibility with RPC in the grid



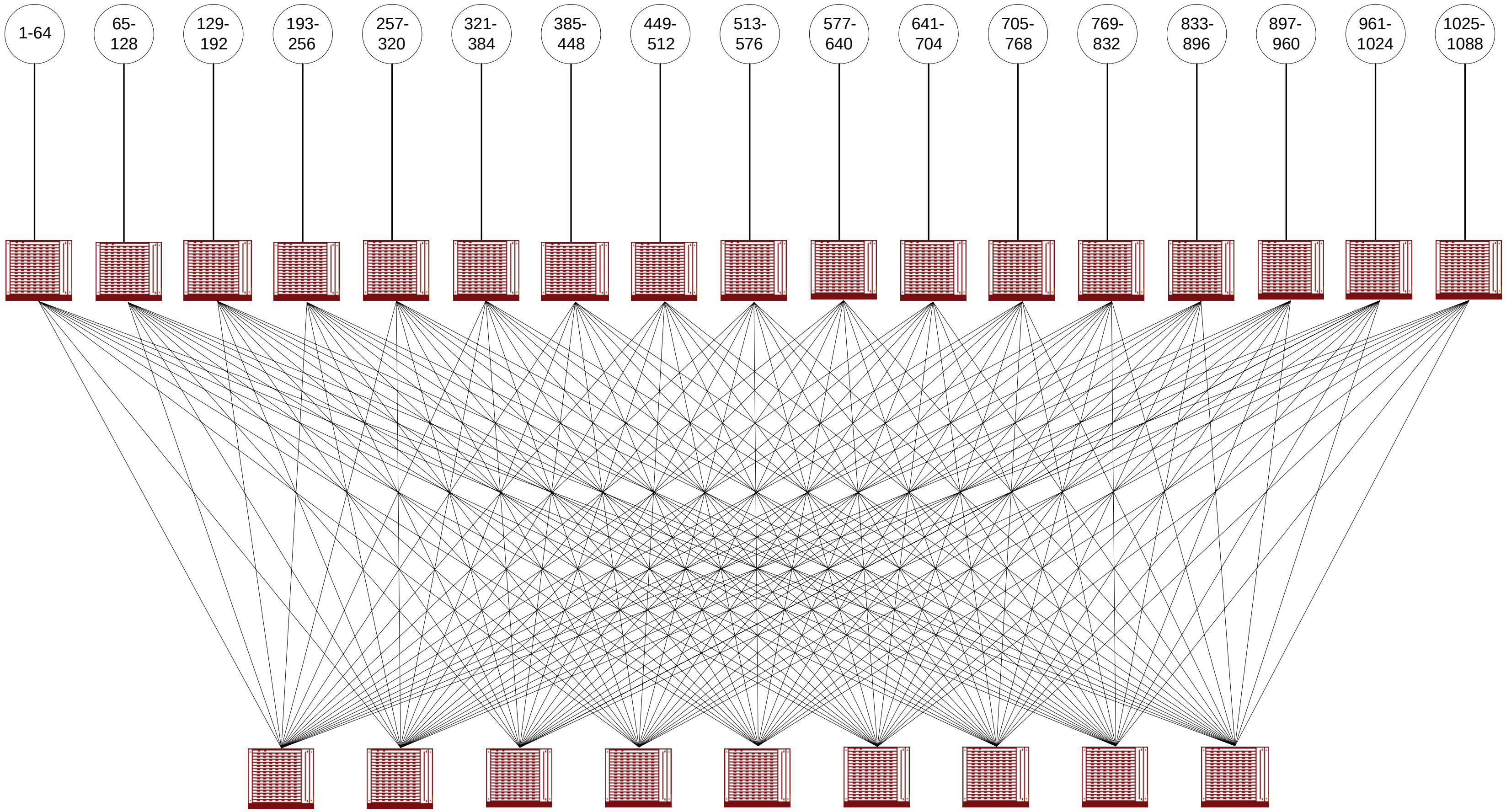
Make the all scenarios possible

- **Our solutions towards grid and cluster federations are:**
  - AIST super cluster
    - Perfect test bed
  - Grid MPI
    - Extremely keen on communication performance
    - GNET-1 provides “pure grid”
  - Grid RPC
    - Easy application deployment
    - Ninf-G2 – it works everywhere
- **Also, ...**
  - Grid Data farm
    - Cluster enables high I/O bandwidth

AIST Super cluster P32 & M64 network config.

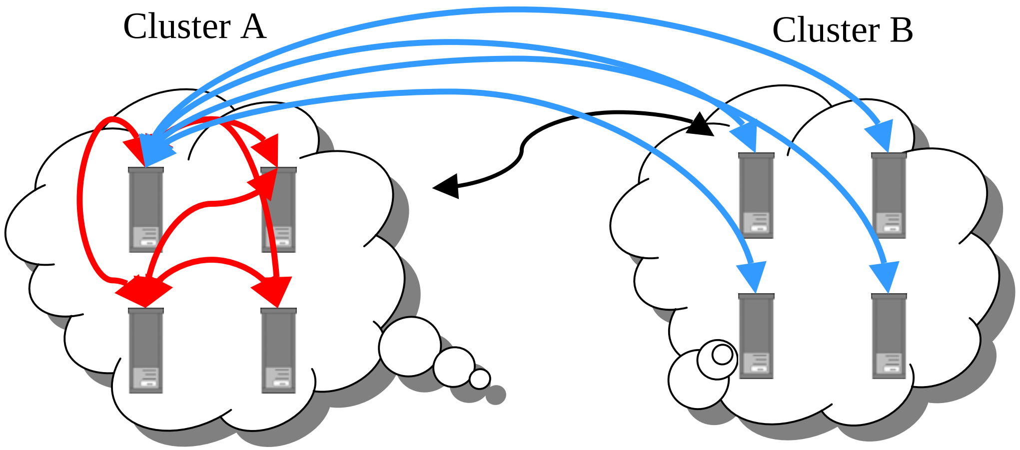


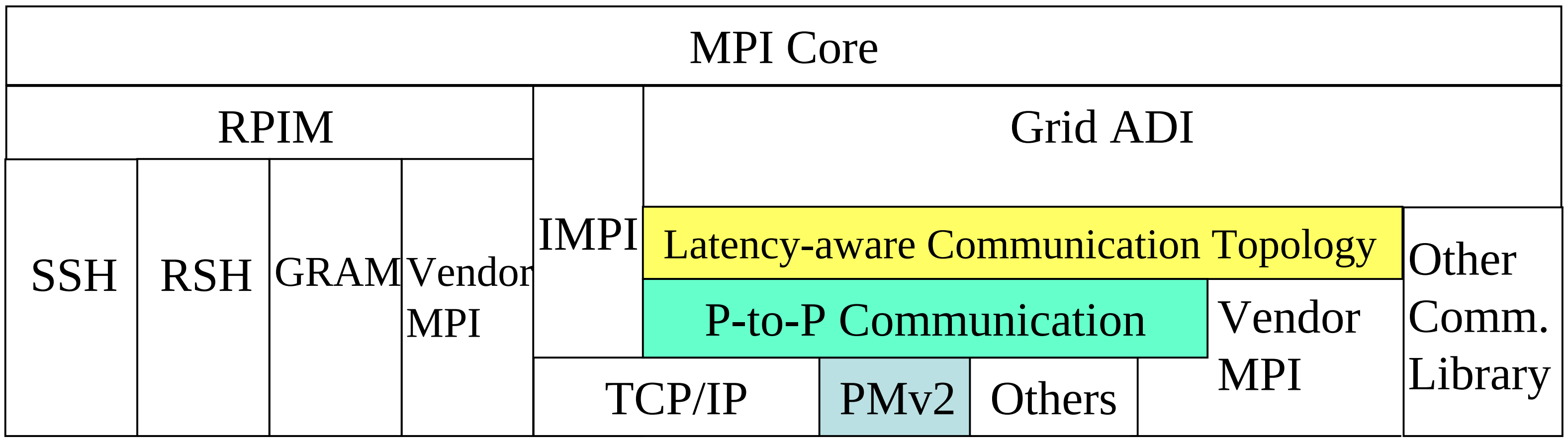
Computational Network for P-32 Cluster



Grid MPI

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

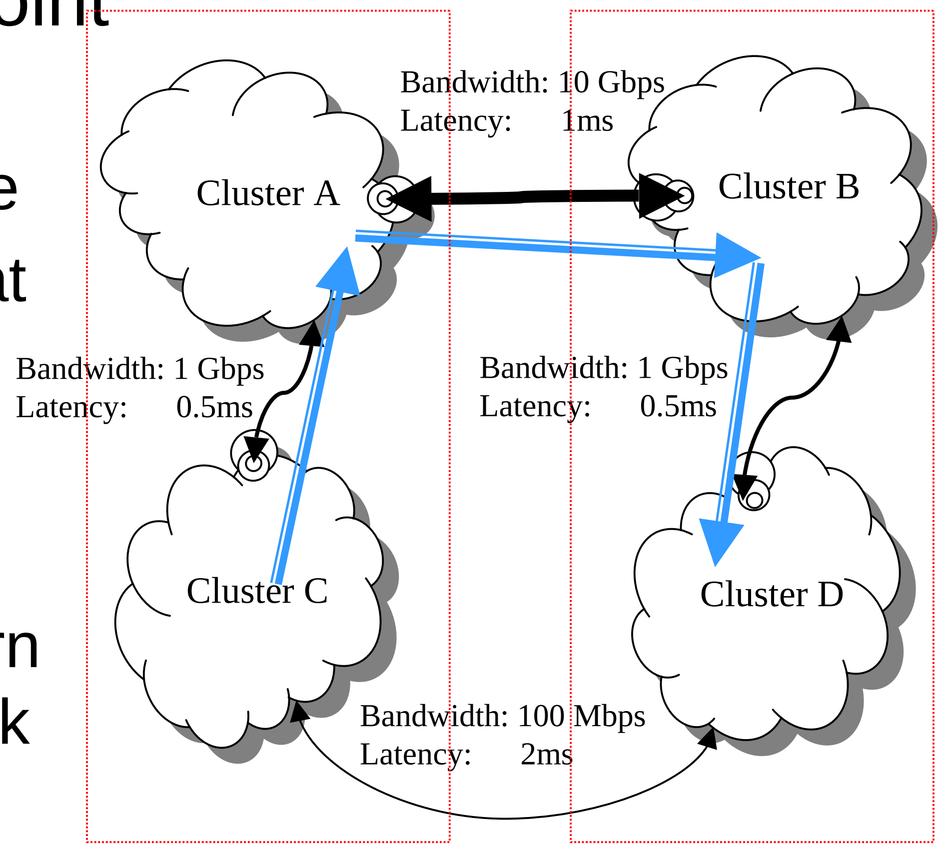




LACT (Latency-Aware Communication Topology)

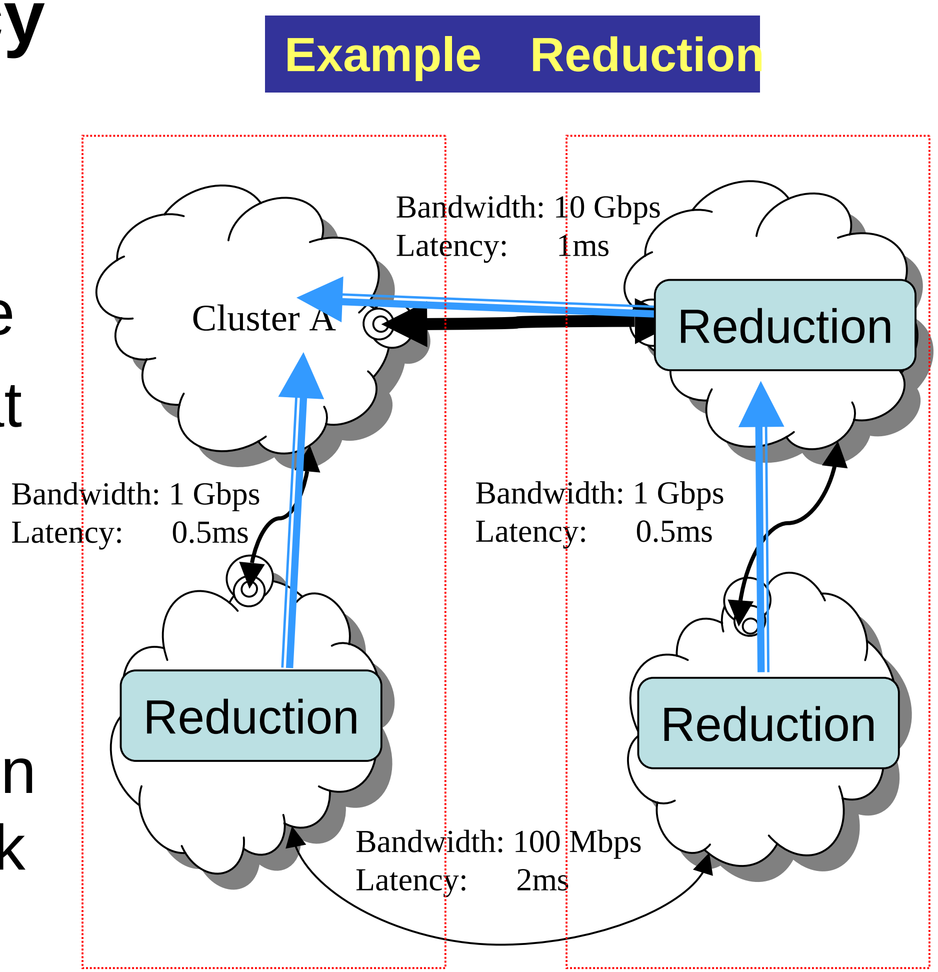
Bandwidth and Latency

- 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



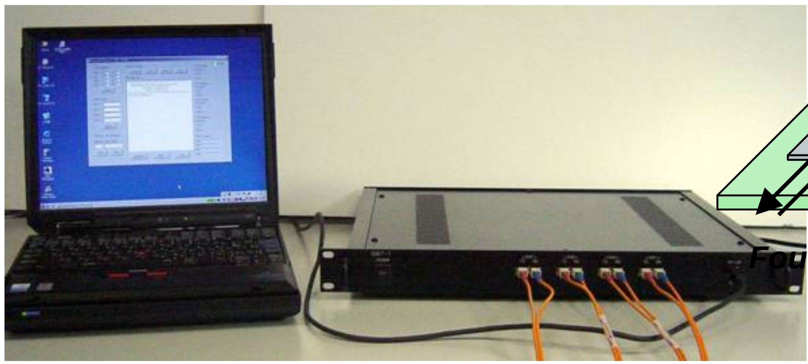
LACT (Latency-Aware Communication Topology)

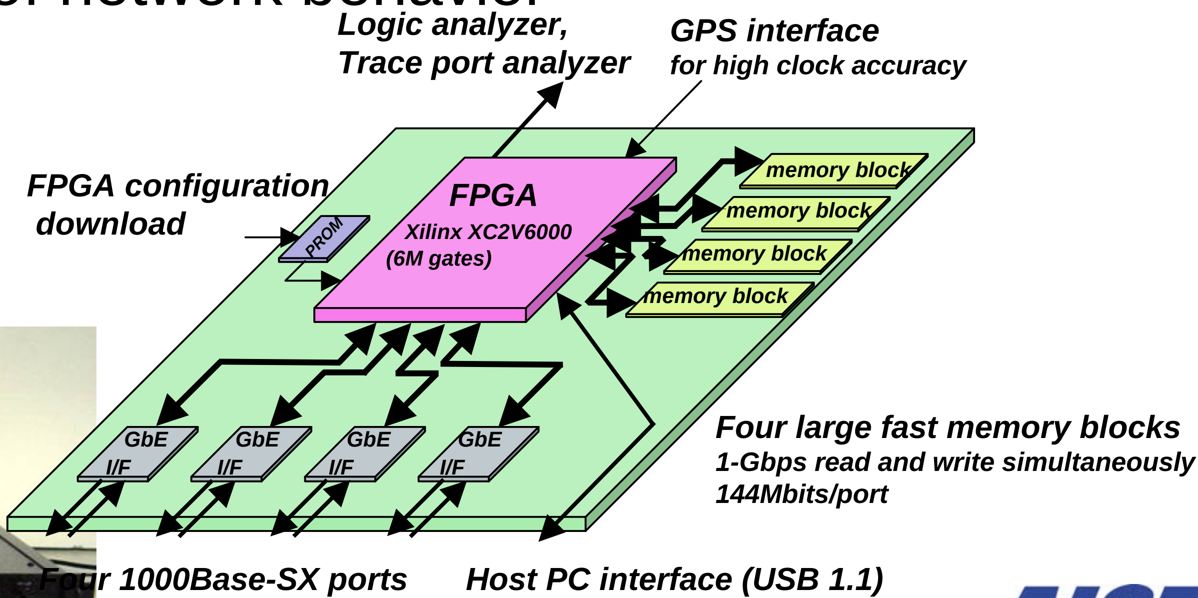
- **Bandwidth and Latency**
  - Routing of Point-to-Point message
    - Based on routing table
    - Message forwarding at intermediate node
  - Routing of collective communications
    - Communication pattern adapted to the network topology



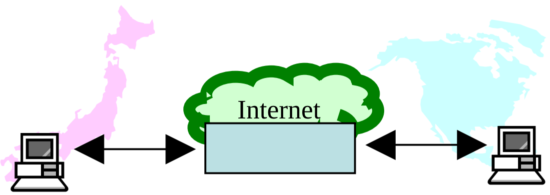
GNET-1: a fully programmable network testbed

- **GNET-1 provides functions by programming the core FPGA**
  - wide area network emulation,
  - network instrumentation,
  - traffic shaping, and
  - traffic generation at gigabit Ethernet wire speeds
  - precise control of network behavior

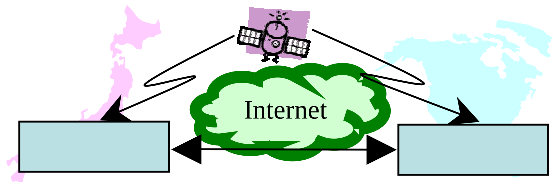




GNET-1 Current functions

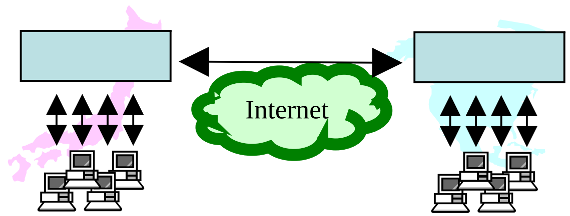


a) Can emulate network with one-way latency up to 134 ms and with traffic shaping, errors, and jitter.



c) Can measure latency and jitter between GNET-1s with μs precision using GPS.

b) Can measure throughput at an arbitrary sampling rate from 100 μs to 1 s.

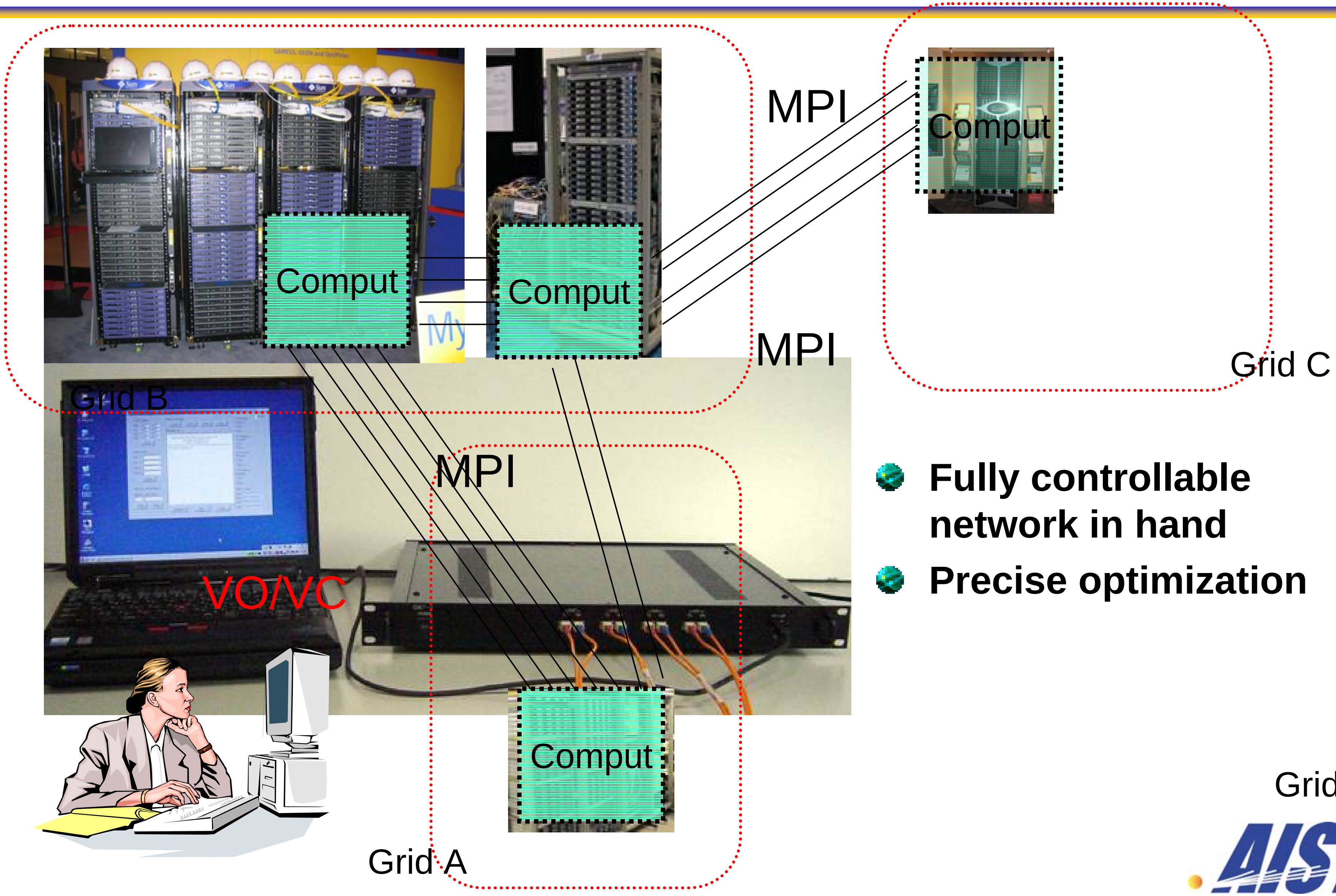


d) Can control transfer rate by adjusting IFG.

GNET-1 provides any functions you require!

AIST

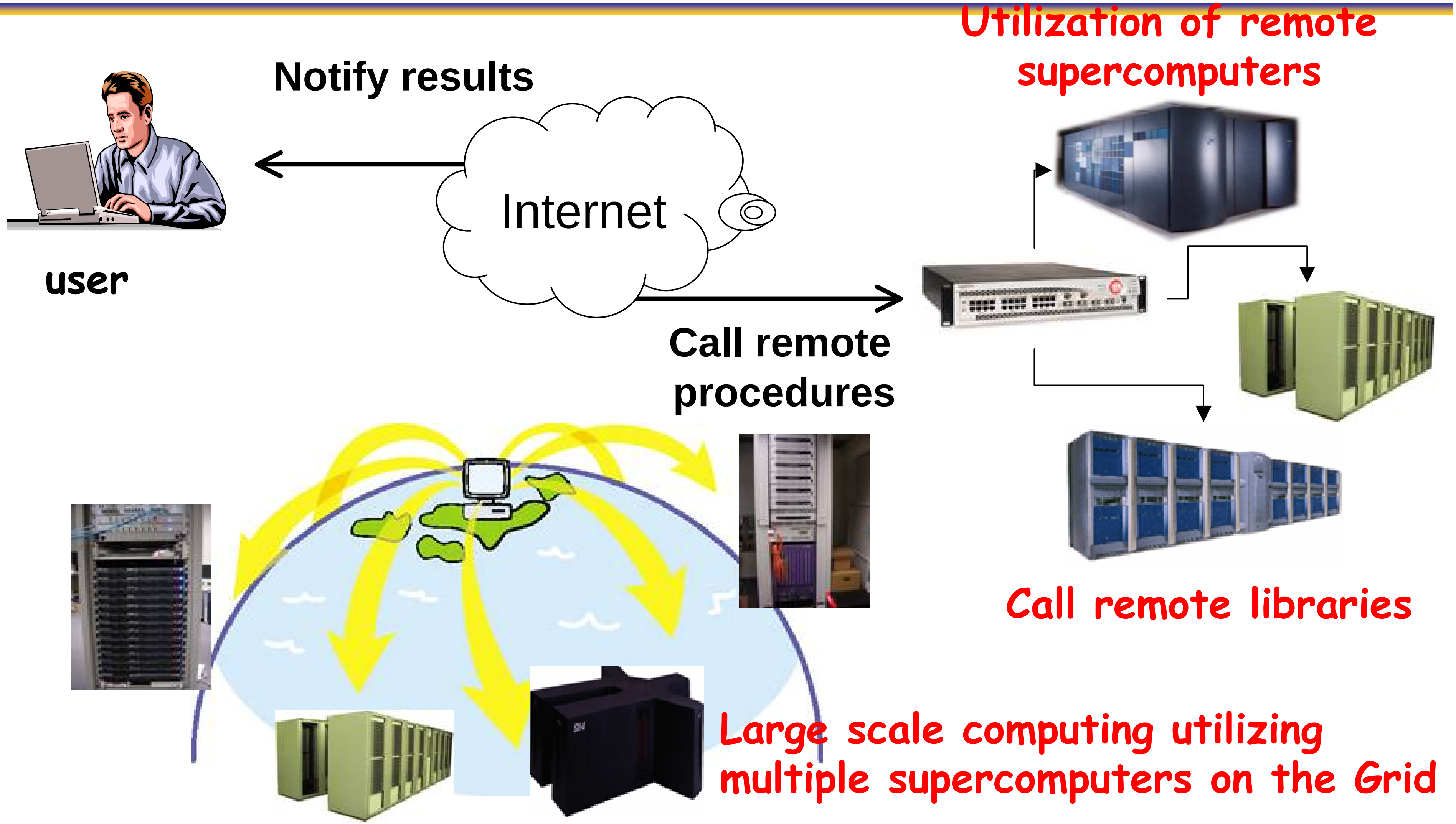
"Pure Grid" test bed – no uncertain noise



- **Fully controllable network in hand**
- **Precise optimization**

Grid D

Grid RPC and Ninf-G

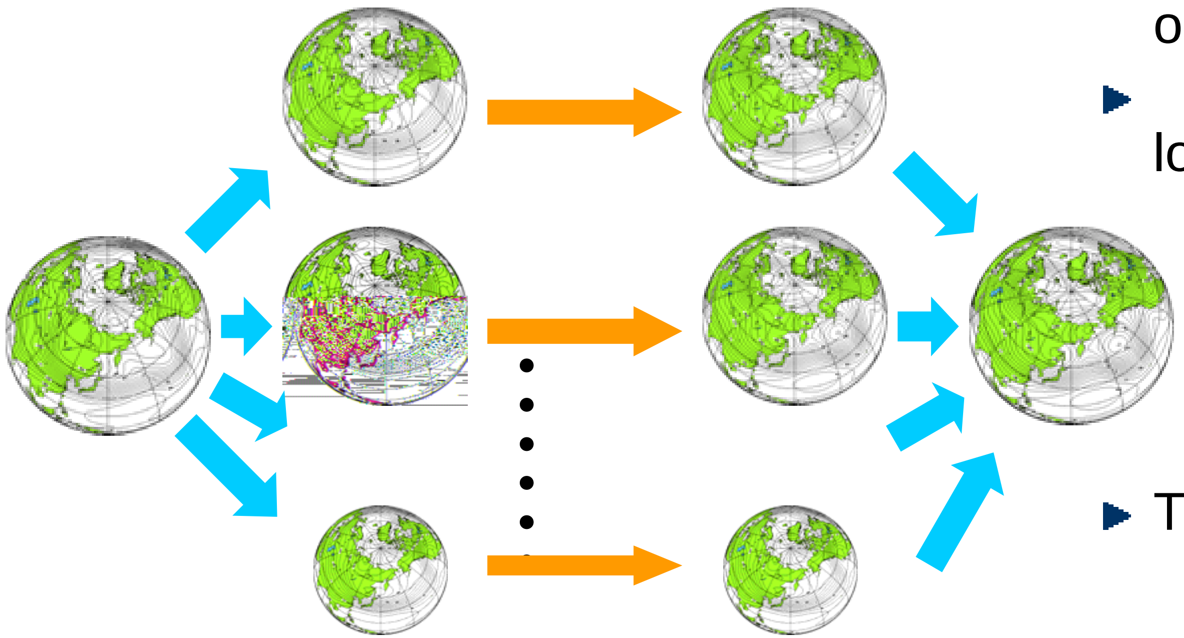


GridRPC (cont'd)

- **Compare to MPI**
  - Client-server programming is suitable for task-parallel applications.
  - Does not need co-allocation
  - Able to use nodes with private IP address if NAT is available (at least when using Ninf-G)
  - Better fault tolerancy - retry
- **Standard GridRPC API is proposed at the GGF GridRPC WG**
  - Define standard GridRPC API
    - later deal with protocol
  - Standardize minimal set of features
    - higher-level features can be built on top
  - Provide several reference implementations
    - Ninf-G, NetSolve
    - Ninf-G2 is available at http://ninf.apgrid.org/
    - As a part of NaReGI project

Application: Climate Simulation

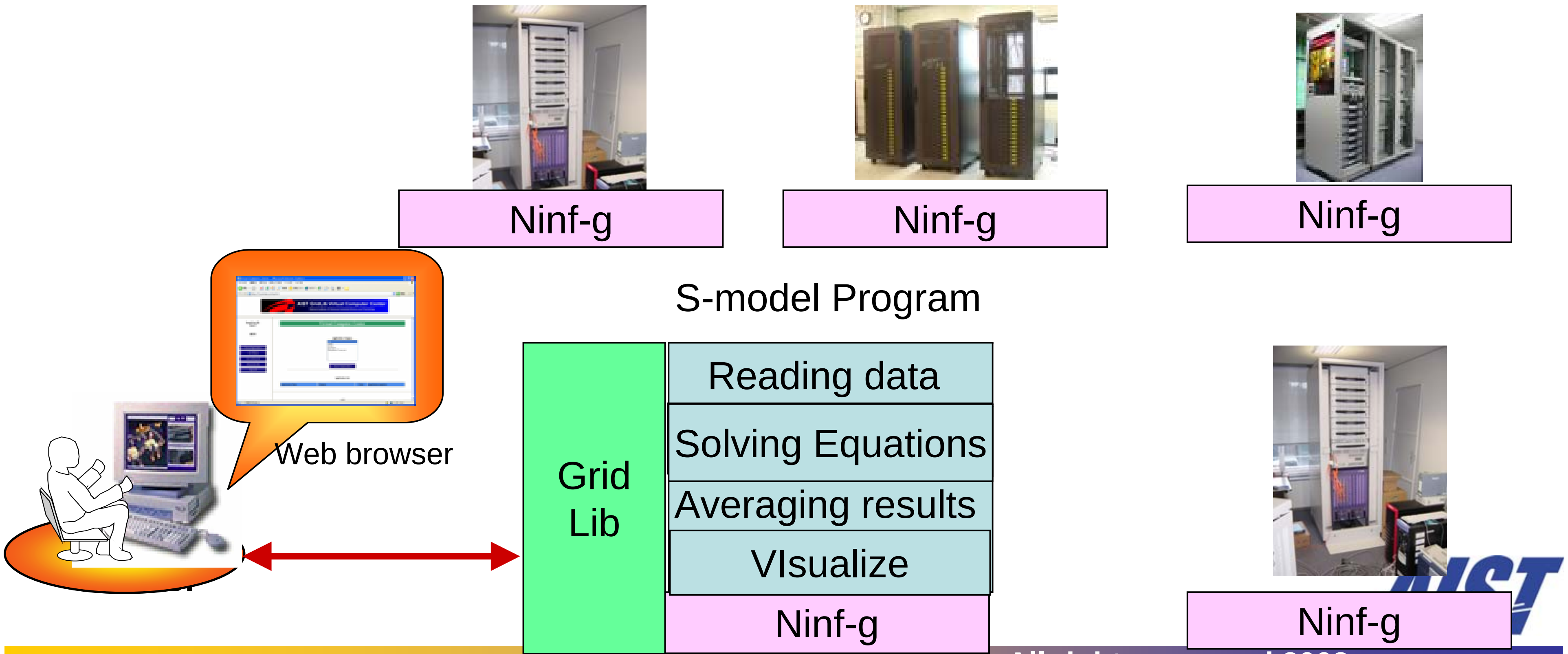
- **Goal**
  - Short- to Middle- term, global climate simulation
    - Winding of Jet-Stream
    - Blocking phenomenon of high atmospheric pressure



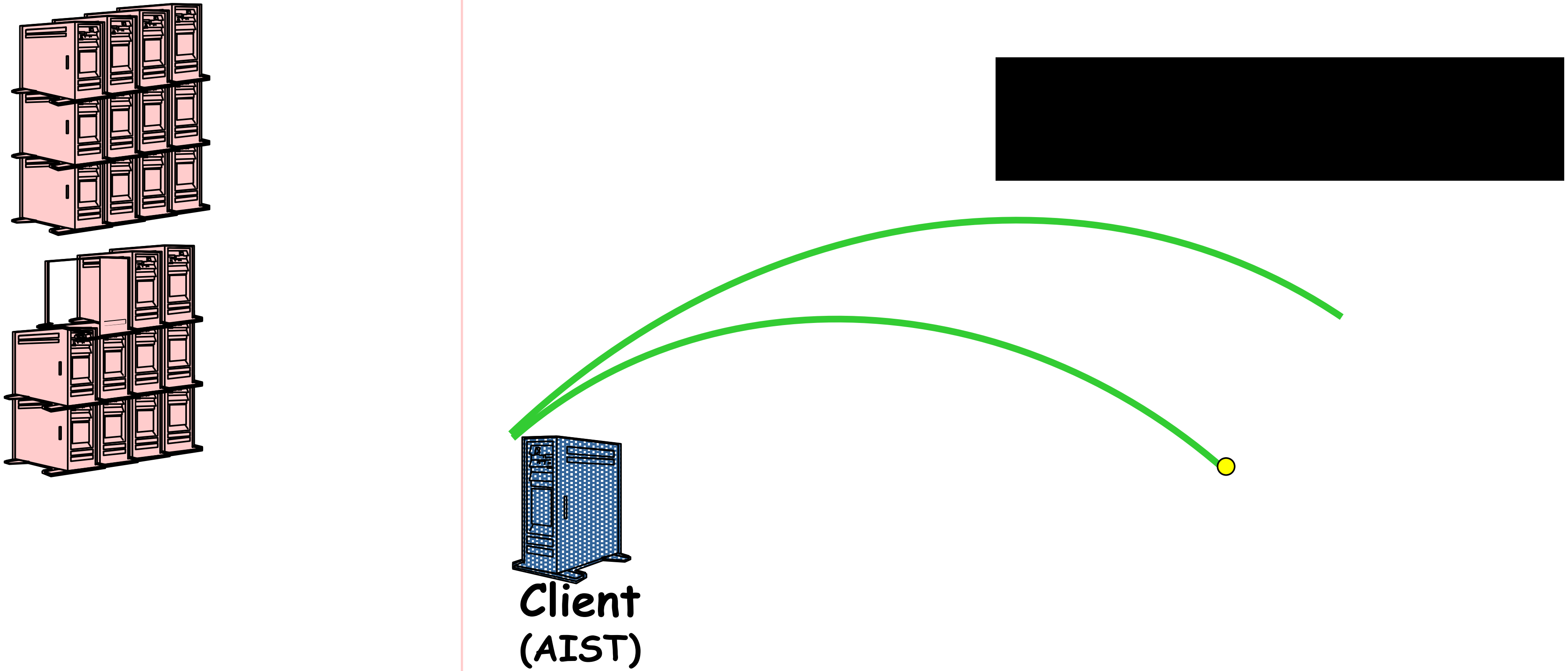
- **Barotropic S-Model**
  - Climate simulation model proposed by Prof. Tanaka (U. of Tsukuba)
  - Simple and precise
  - Modeling complicated 3D turbulence as a horizontal one
  - Keep high precision over long periods
    - Taking a statistical ensemble mean
      - ~ several 100 simulations
    - Introducing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



ApGrid / PRAGMA Testbed

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

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.

Special acknowledgement on this study

- **TeraGrid EC (esp. Pete Beckman @ ANL)**
- **TeraGrid Help Team**
- **Resource Contributors**
  - NCSA, TITECH, KISTI, AIST, SDSC
- **Ninf-G developer Team**

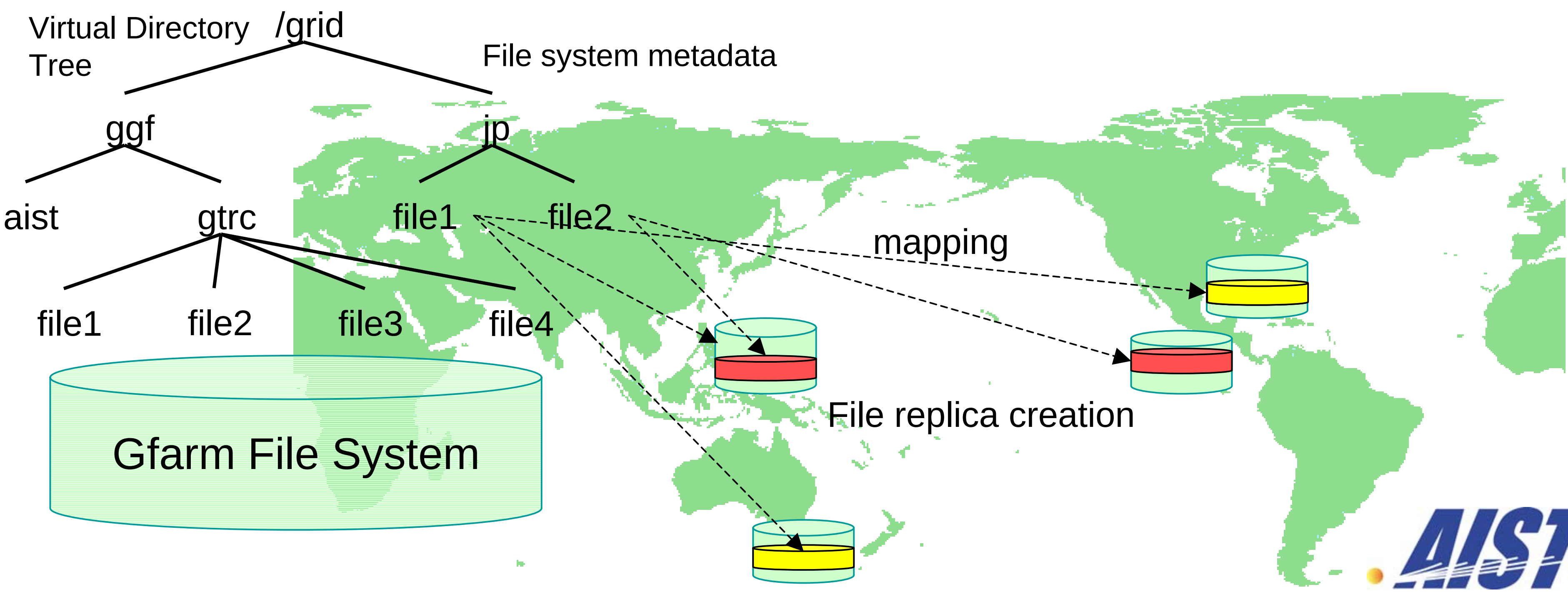
Univ. of Hong Kong OPEN Campus, Oct 18, 2003

Goal and feature of 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
  - Global parallel and distributed processing
- **Features**
  - Secured based on Grid Security Infrastructure
  - From small scale to world wide scale depending the data size and usage scenarios
  - Data location transparent data access
  - Automatic and transparent replica access for fault tolerance
  - High-performance data access and processing by accessing multiple dispersed storages in parallel

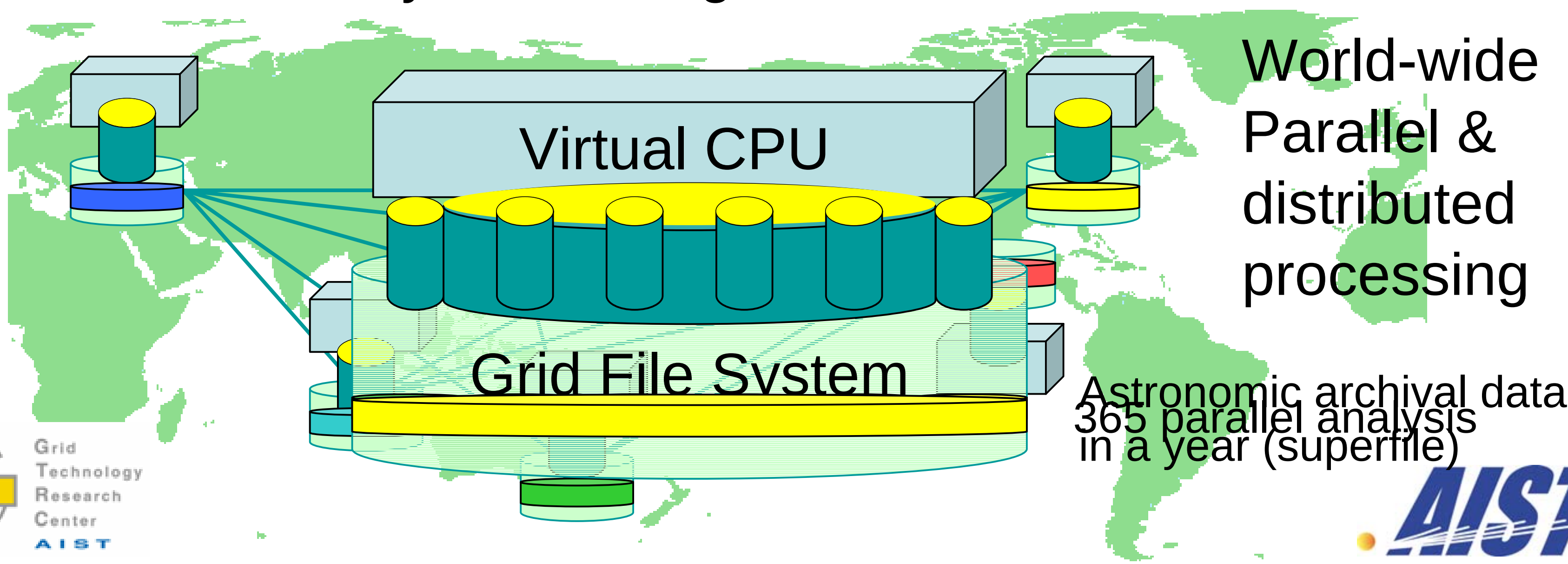
Grid Datafarm (1): Gfarm file system - World-wide virtual file system [CCGrid 2002]

- **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
  - Automatic and transparent replica access for fault tolerance and access-concentration avoidance

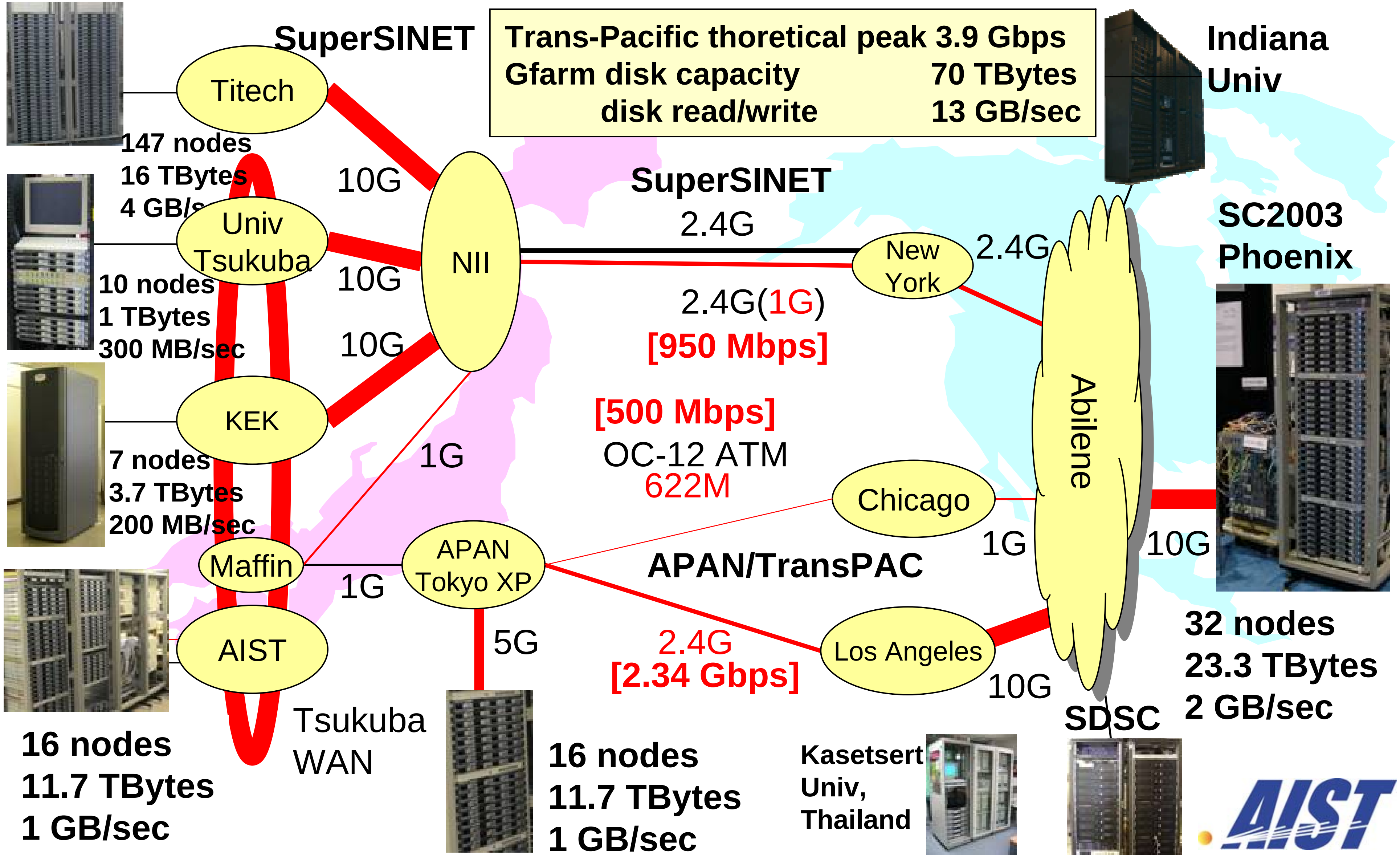


Grid Datafarm (2): High-performance data access and processing support [CCGrid 2002]

- **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
    - File-affinity scheduling

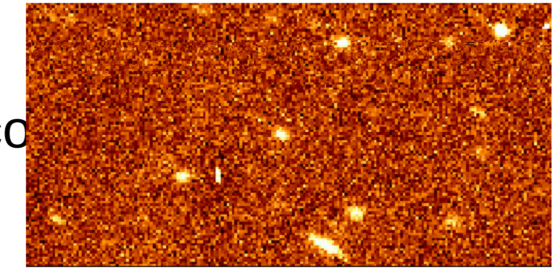


Trans-Pacific Gfarm Datafarm testbed: Network and cluster configuration

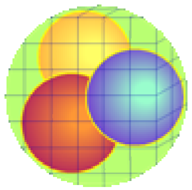


Scientific Data for Bandwidth Challenge

- **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 telesco
  - N. Yamamoto (AIST)

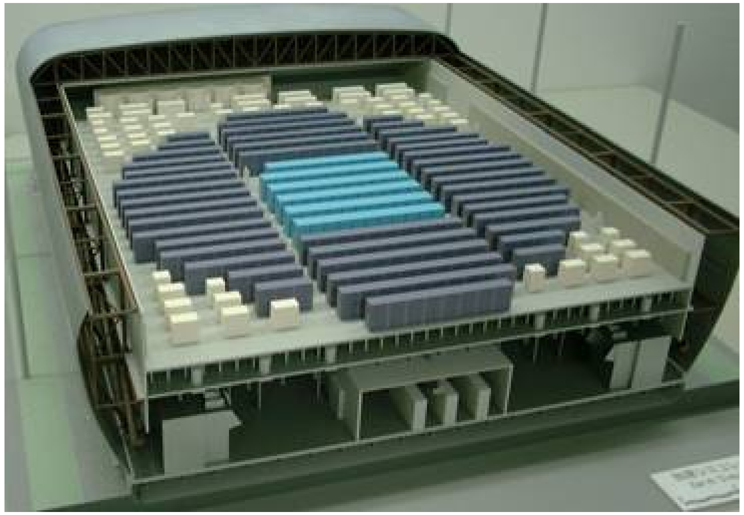


- **Large configuration data from Lattice QCD**
  - Three sets of hundreds of gluon field configurations on a 24^3*48 4-D space-time lattice (3 sets x 364.5 MB x 800 = **854.3 GB**)
  - Generated by the CP-PACS parallel computer at Center for Computational Physics, Univ. of Tsukuba (300Gflops x years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



Earth simulator impact

- **Top 500 list (Nov.,2003)**
  - 33 systems listed, only 6 systems in top100
  - ES (=35.8TFlops) > Σall others = (27.4TFlops)
    - Σ53 others = (21.4TFlops) as of June 2002
    - 18 systems in top 100
- **Clusters**
  - 6 clusters in June, 2002
    - 2.64TFlops out of 5.47TFlops
  - 8 clusters in Nov., 2003
    - 5.76TFlops out of 9.12TFlops
    - AIST owns 3 clusters WoW
  - +25TFlops (peak) in June, 2004
    - Riken + AIST super cluster + + +



Titech Campus Grid - System Image (Pseudo grid)
(since April 2002)

Slide: courtesy of S. Matsuoka (Titech),

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.**

24-processor Satellite Systems @ each department ×12 systems

Suzukake-dai Campus

30km

Super TITANET (1-4Gbps)

NEC Express 5800 Series Blade Servers

Oo-okayama Campus

Grid-wide Single System Image via Grid middleware: Globus, Ninf-G, Condor, NWS, …

High Density GSIC Main Cluster (256 processers) x 2 systems in just 5 cabinets

Super SINET (10 Gbps MOE National Backbone Network) to other Grids

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

Onto a Production Grid – Reality

- **Bootstrapping Problem**

Slide: courtesy of S. Matsuoka (Titech),

  - 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

Summary

- **Grid Cluster federation is the way to go for achieving high performance with low cost.**
  - And is real computing environment near future
  - We don't wait for a long time to touch a monster
- **Scenario A & D are ready to go by grid RPC and existing grid tools, however**
  - CA operation, policy issues, etc.
- **Scenario B & C are not simple**
  - A lot of more work to be done
  - Scheduling, resource mgnt, FT, etc.
  - "pure grid" is useful in this development

Top 10 Cluster Ranking in Asia (preliminary)

rank		CPU	GHz	#node	#ways	#proc	TFlops
1	Riken, JP	Xeon	3.06	1024	2	2048	12.40
2	CAS, CN	Opteron	2.4	512	4	2048	9.83
3	AIST, JP	Opteron	2.0	1074	2	2148	8.59
4	CAS, CN	Itanium2	1.3	265	4	1024	5.32
5	KISTI, KR	Xeon	2.4	512	2	1024	4.92
6	AIST, JP	Xeon	3.06	256	2	512	3.13
7	AIST, JP	Itanium 2	1.3	132	4	528	2.75
8	AMSS, CN	Xeon	2.0	256	2	512	2.05
9	UHK, HK	Xeon	1.8	256	2	512	1.84
9	Doshisha, JP	Opteron	1.8	256	2	512	1.84

Disclaimer: This table may contain wrong numbers due to uncertain source,
DO NOT MAKE ANY COPY OF THIS

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