The TSUBAME Now and Future--- Running a 100TeraFlops-Scale Supercomputer for Everyone as a NAREGI Resource and Its Future

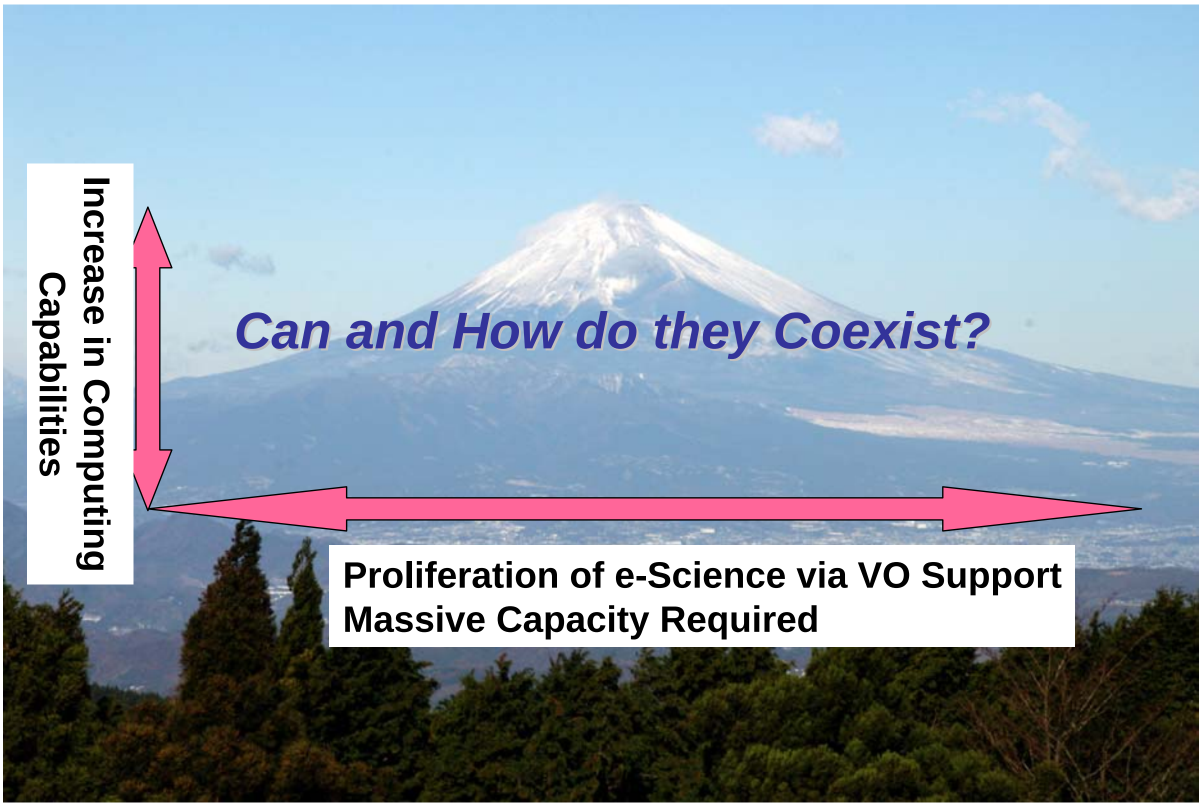
Satoshi Matsuoka, Professor/Dr.Sci.

Global Scientific Information and Computing Center

Tokyo Inst. Technology

& NAREGI Project National Inst. Informatics

Capacity vs. Capability



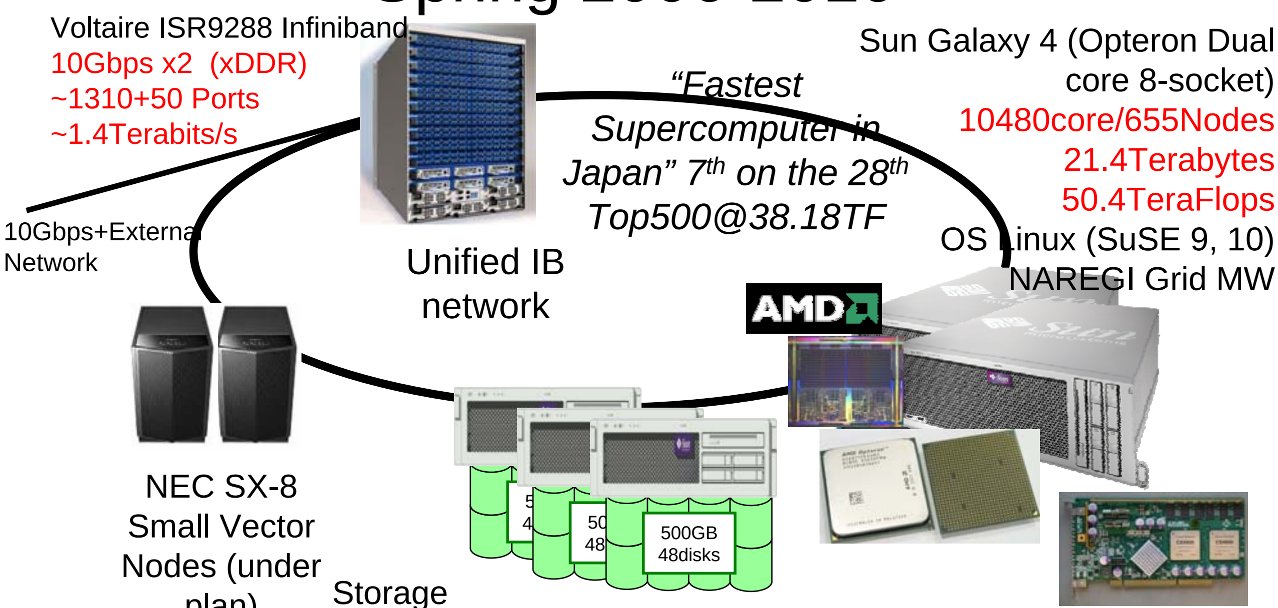
TSUBAME "Grid" Cluster Supercomputer

- **T**okyo-tech
- **S**upercomputer and
- **UB**iquitously
- **A**ccessible
- **M**ass-storage
- **E**nvironment

燕

TSUBAME means "a swallow" in Japanese, Tokyo-tech (Titech)'s symbol bird, and its logo (but we are home to massive # of parakeets)

The TSUBAME Production "Supercomputing Grid Cluster" Spring 2006-2010

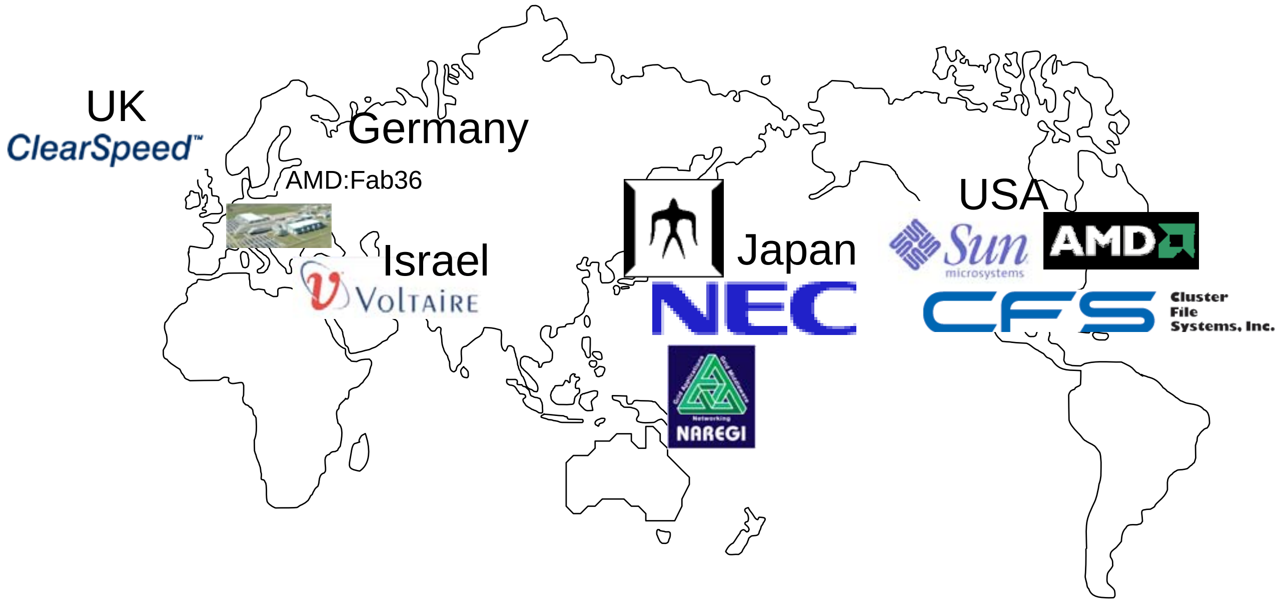


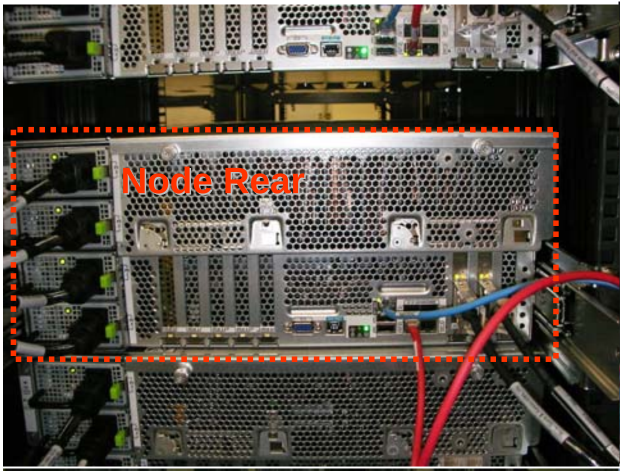
Storage
1 Petabyte (Sun "Thumper")
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV
50GB/s aggregate I/O BW

ClearSpeed CSX600
SIMD accelerator
360 boards,
35TeraFlops(Current))

TSUBAME Global Partnership

NEC: Main Integrator, Storage, Operations
SUN: Galaxy Compute Nodes, Storage
AMD: Opteron CPU
ClearSpeed: CSX600 Accel.
NAREGI: Grid MW
Voltaire: Infiniband Network
CFS: Parallel FSCFS
Titech GSIC: us





TSUBAME Architecture =

Commodity PC Cluster

+

Traditional FAT node Supercomputer

+

The Internet & Grid

+

(Modern) Acceleration

Design Principles of TSUBAME(1)

- Capability and Capacity : have the cake and eat it, too!
  - **High-performance, low power x86 multi-core CPU**
    - High INT-FP, high cost performance, Highly reliable
    - Latest process technology – high performance and low power
    - Best applications & software availability: OS (Linux/Solaris/Windows), languages/compilers/tools, libraries, Grid tools, all ISV Applications
  - **FAT Node Architecture (later)**
    - Multicore SMP – most flexible parallel programming
    - High memory capacity per node (32/64GB)
    - Large total memory – 21.4 Terabytes
    - Low node count – improved fault tolerance, easen network design
  - **High Bandwidth Infiniband Network, IP-based (over RDMA)**
    - (Restricted) two-staged fat tree
    - High bandwidth (10-20Gbps/link), multi-lane, low latency (< 10microsec), reliable/redundant (dual-lane)
    - Very large switch (288 ports) => low switch count, low latency
    - Resilient to all types of communications; nearest neighbor, scatter/gather collectives, embedding multi-dimensional networks
    - IP-based for flexibility, robustness, synergy with Grid & Internet

Design Principles of TSUBAME(2)

- PetaByte large-scale, high-perfomance, reliable storage
  - **All Disk Storage Architecture (no tapes), 1.1Petabyte**
    - Ultra reliable SAN/NFS storage for /home (NEC iStore), 100GB
    - Fast NAS/Lustre PFS for /work (Sun Thumper), 1PB
  - Low cost / high performance SATA2 (500GB/unit)
  - High Density packaging (Sun Thumper), 24TeraBytes/4U
  - Reliability thru RAID6, disk rotation, SAN redundancy (iStore)
    - Overall HW data loss: once / 1000 years
  - High bandwidth NAS I/O: ~50GBytes/s Livermore Benchmark
  - **Unified Storage and Cluster interconnect**: low cost, high bandwidth, unified storage view from all nodes w/o special I/O nodes or SW
- Hybrid Architecture: General-Purpose Scalar + SIMD Vector Acceleration w/ ClearSpeed CSX600
  - 35 Teraflops peak @ 90 KW (~ 1 rack of TSUBAME)
  - General purpose programmable SIMD Vector architecture

TSUBAME Timeline

- 2005, Oct. 31: TSUBAME contract
- Nov. 14th Announce @ SC2005
- 2006, Feb. 28: stopped services of old SC
  - SX-5, Origin2000, HP GS320
- Mar 1~Mar 7: moved the old machines out
- **Mar 8~Mar 31: TSUBAME Installation**
- Apr 3~May 31: Experimental Production phase 1
  - 32 nodes (512CPUs), 97 Terabytes storage, free usage
  - Linpack 38.18 Teraflops May 8th, #7 on the 28th Top500
  - **May 1~8: Whole system Linpack, achieve 38.18 TF**
- June 1~Sep. 31: Experimental Production phase 2
  - 299 nodes, (4748 CPUs), still free usage
- **Sep. 25-29 Linpack w/ClearSpeed, 47.38 TF**
- Oct. 1: Full production phase
  - ~10,000CPUs, several hundred Terabytes for SC
  - Innovative accounting: Internet-like Best Effort & SLA

TSUBAME as No.1 in Japan

東京工業大学 Tokyo Institute of Technology

GSIC Global Scientific Information and Computing Center

>>

>85 TeraFlops
1.1Petabyte
4 year procurement cycle

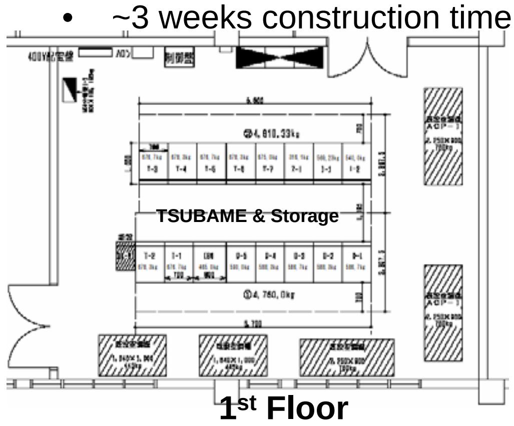
Has beaten the Earth Simulator

Has beaten all the other Univ. centers combined

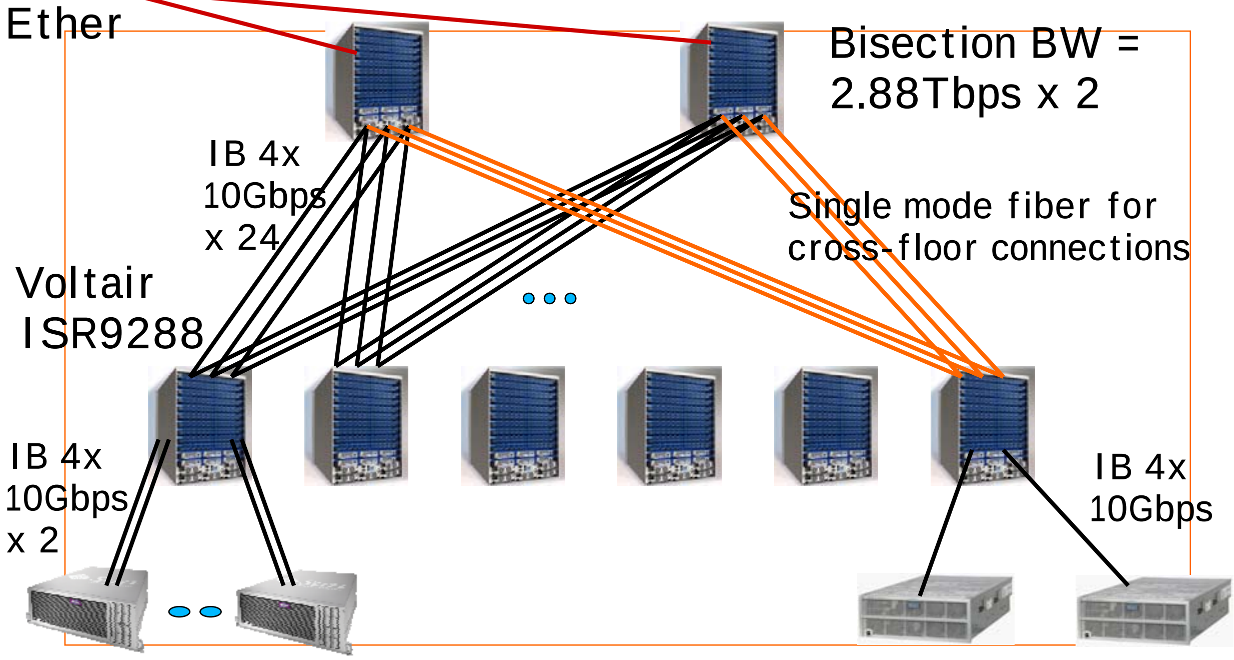
Total 45 TeraFlops, 350 Terabytes

TSUBAME Physical Installation

- 3 rooms (600m^2), 350m^2 service area
- 76 racks incl. network & storage, 46.3 tons
  - 10 storage racks
- 32 AC units, 12.2 tons
- Total 58.5 tons (excl. rooftop AC heat exchangers)
- Max 1.2 MWatts
- ~3 weeks construction time



TSUBAME Network: (Restricted) Fat Tree, IB-RDMA & TCP-IP



X4600 x 120nodes (240 ports) per switch
=> 600 + 55 nodes, 1310 ports, 13.5Tbps

X4500 x 42nodes (42 ports)
=> 42ports 420Gbps

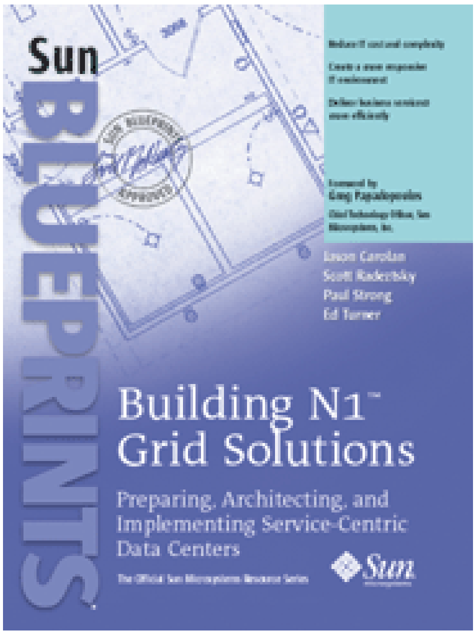
The Benefits of Being “Fat Node”

- Many HPC Apps favor large SMPs
- Flexble programming models---MPI, OpenMP, Java, ...
- Lower node count – higher reliability/manageability
- Full Interconnect possible --- Less cabling & smaller switches, multi-link parallelism, no “mesh” topologies

	CPUs/Node	Peak/Node	Memory/Node
IBM eServer (SDSC DataStar)	8, 32	48GF~217.6GF	16~128GB
Hitachi SR11000 (U-Tokyo, Hokkaido-U)	8, 16	60.8GF~135GF	32~64GB
Fujitsu PrimePower (Kyoto-U, Nagoya-U)	64~128	532.48GF~799GF	512GB
The Earth Simulator	16	128GF	16GB
TSUBAME (Tokyo Tech)	**16**	**76.8GF+ 96GF**	**32~64GB**
IBM BG/L	2	5.6 GF	0.5~1GB
Typical PC Cluster	2~4	10~40GF	1~8GB

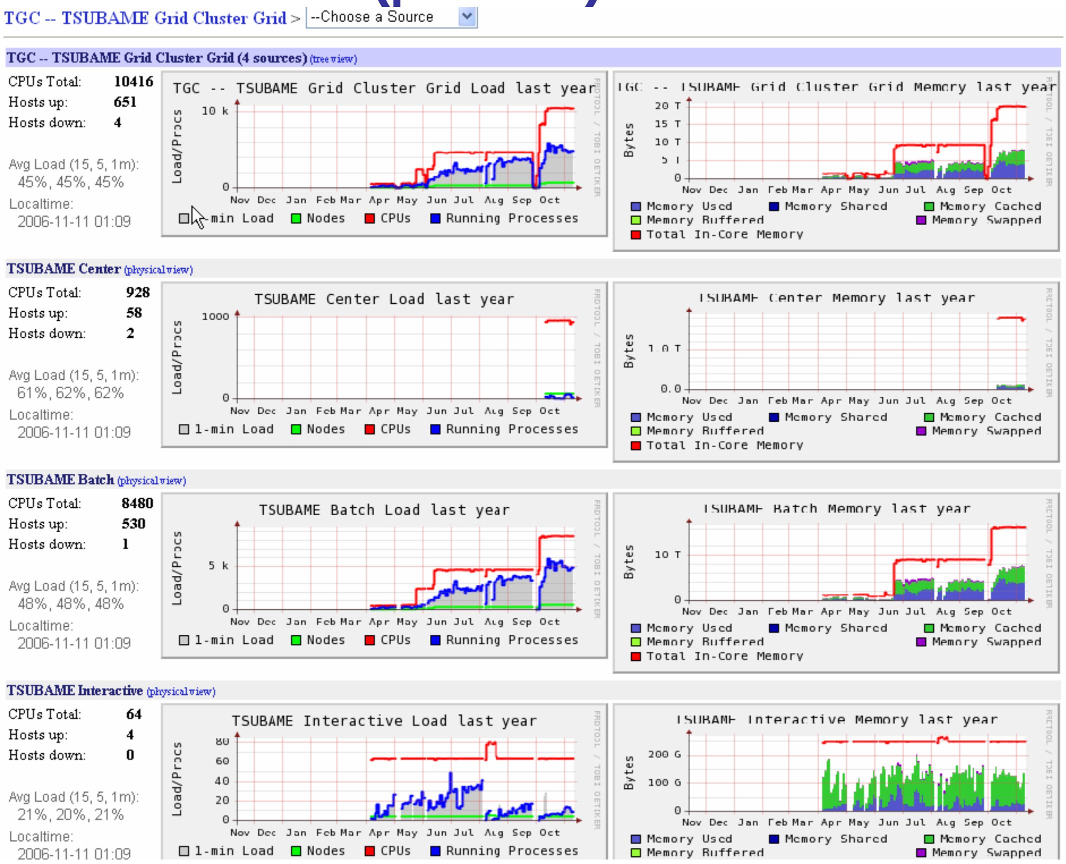
Sun Tsubame Technical Experiences to be Published as Sun Blueprints

- Coming RSN
- About 100 pages
- Principally authored by Sun’s On-site Engineers

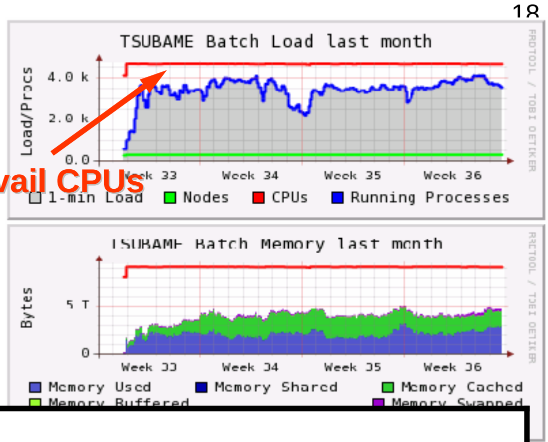


TSUBAME in Production

Oct.1 2006 (phase 3) ~10400 CPUs



TSUBAME Reliability

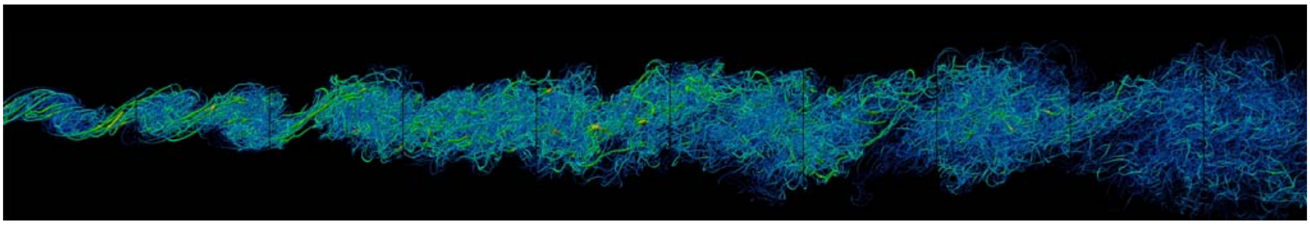


- Very High Availability (over 99%)
- Faults frequent but localized effect only
  - Jobs automatically restarted by SGE
- Most faults *NOT* HW, mostly SW
  - Fixed with reboots & patches

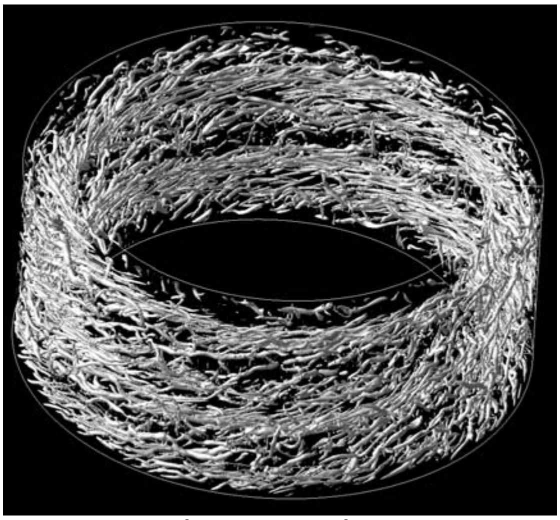
TSUBAME Fault Overview 8/15/2006 - 9/8/2006

Date	Faults	Compute Nodes (655 nodes)			Thumper HDD Faults (2016 HDDs)	Total HW Faults
		Overall Compute Node faults	Possible HW Faults (incl. unknowns)	HW Breakage Faults (excl. unknowns)		
Total						
24 Days	39	34	12	3	4	7
Per Day	1.63	1.42	0.50	0.13	0.17	0.29
Over Year	593.1	517.1	182.5	45.6	60.8	106.5
Unit MTBF (Y)	1.1043	1.26672	3.589041	14.356164	33.13973	
Unit MTBF (H)	9,674	11,096	31,440	125,760	290,304	

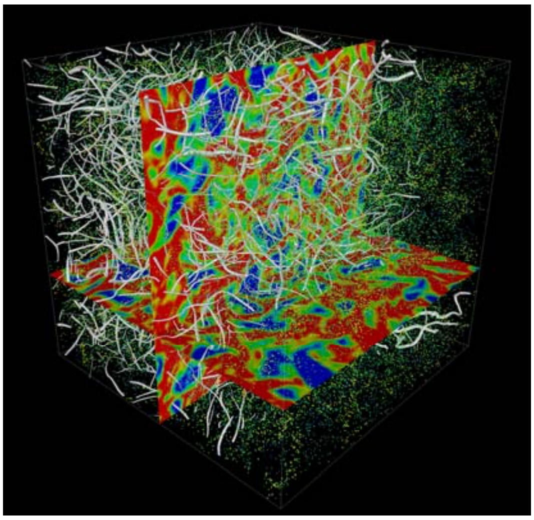
TSUBAME Applications---Massively Complex Turbulant Flow and its Visualization (by Tanahashi Lab and Aoki Lab, Tokyo Tech.)

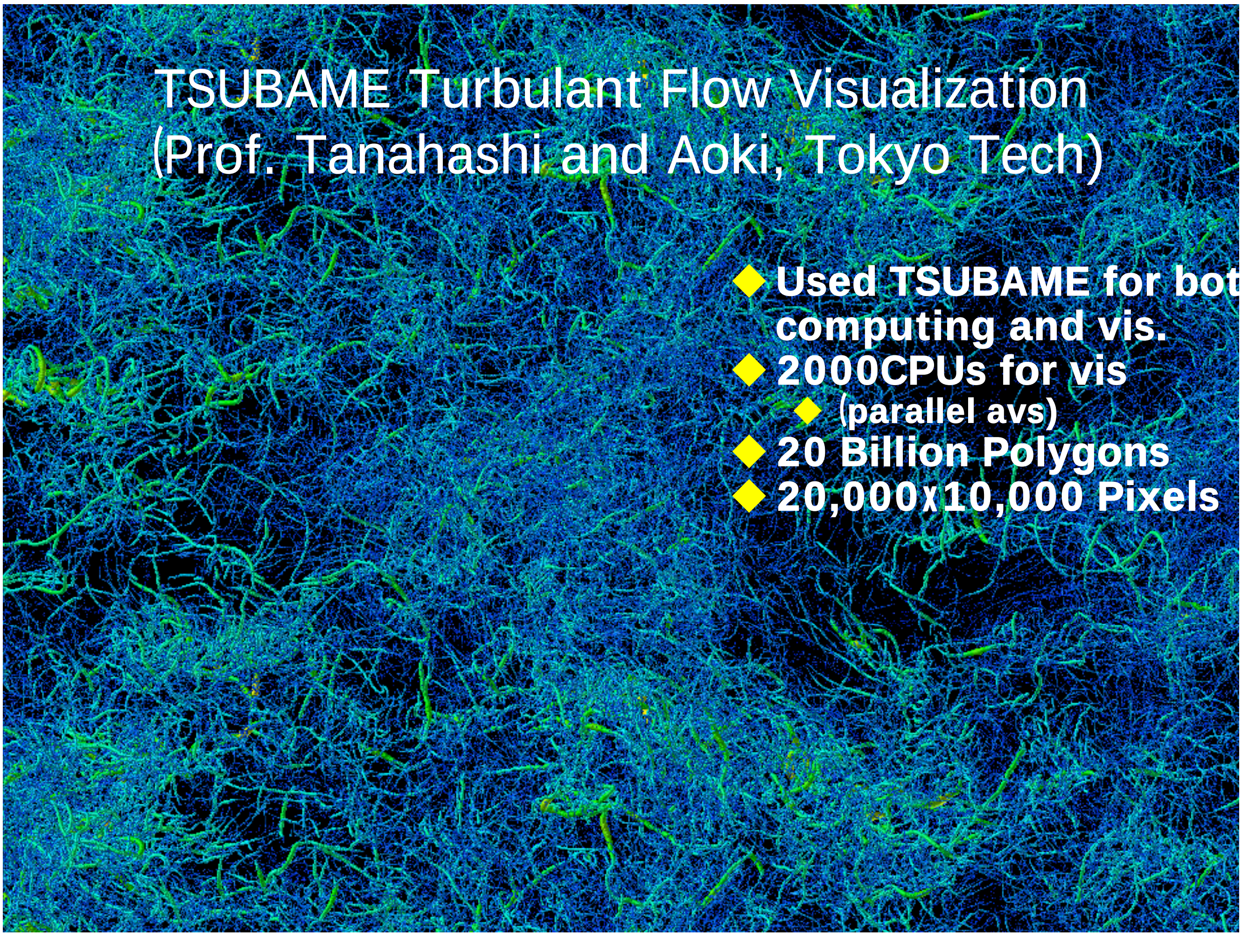


Turbulant Flow from Airplane

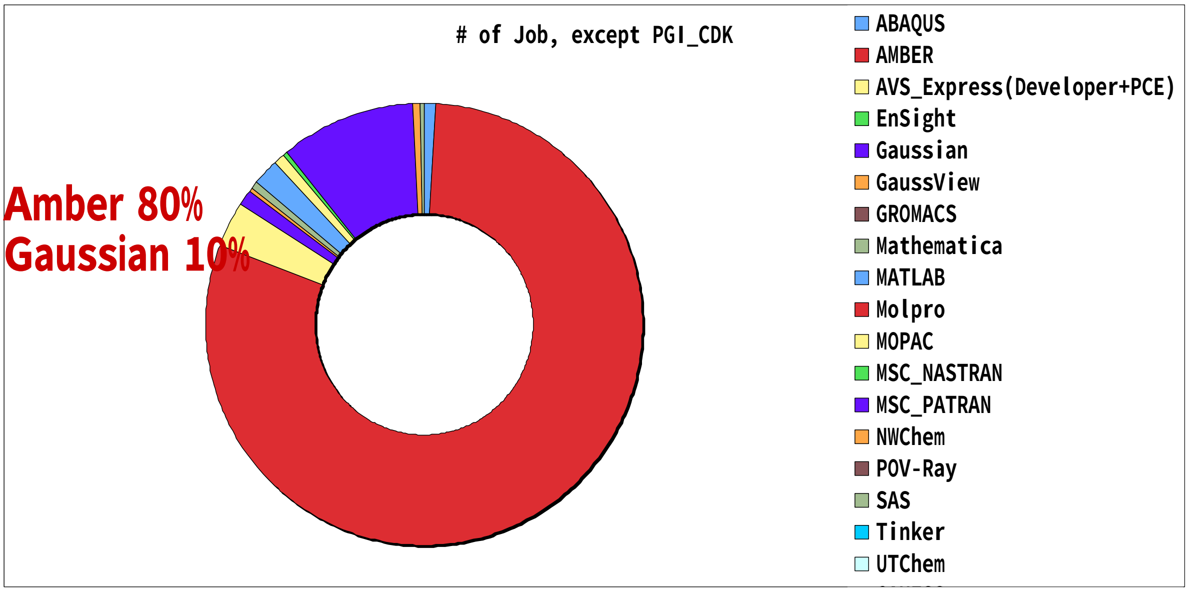


Taylor-Couette Flow



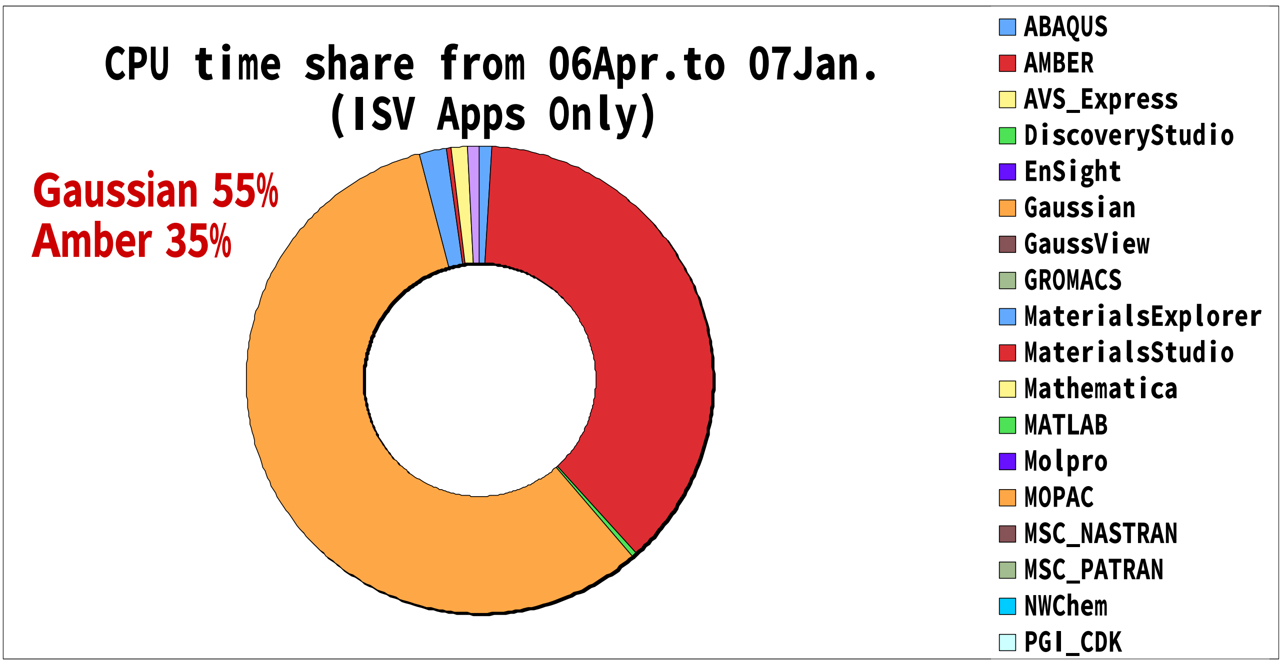


TSUBAME Job Statistics for ISV Apps (# Processes)



1,363,374 Processes (ISV only, excl. PGI_CDK)
Approx. 5000/day (via Sun GridEngine)

TSUBAME Job Statistics for ISV Apps (# CPU Timeshare)



Status as of Mar 13th, 2007

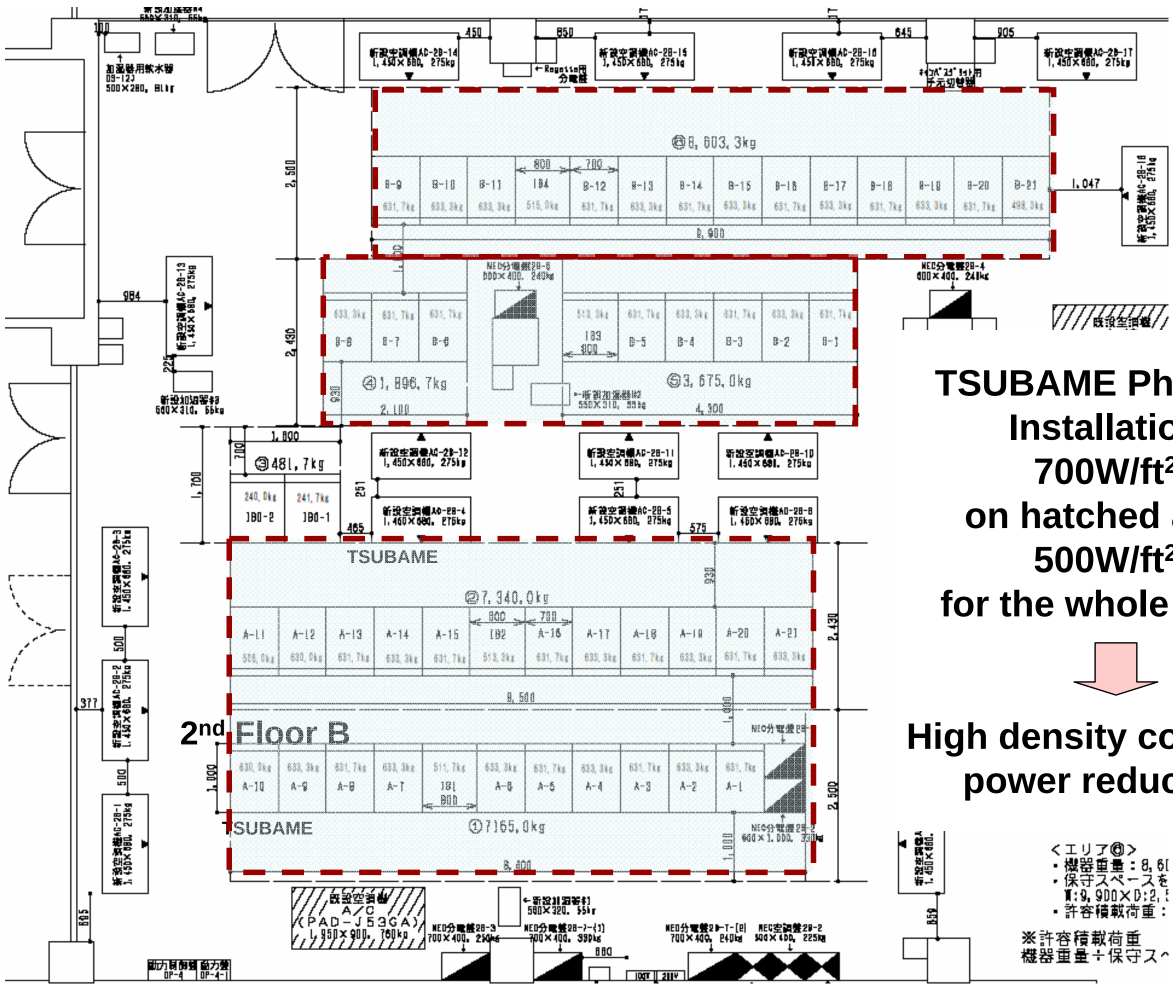
```
QUEUE              FREE NODE      FREE CPU       FREE MEMORY
-------------------------------------------------------------
TOTAL                    260      1676 CPU           5240 GB
- bes1                  107       783 CPU           2786 GB
- bes2                  110       606 CPU           1526 GB
- default                26       112 CPU            412 GB
- gridMathem              8       128 CPU            256 GB
- high                    9        47 CPU            260 GB
- sla1                    0         0 CPU              0 GB
- sla2                    0         0 CPU              0 GB
```


Performance/Watt of TSUBAME Comparisons with other leading Supercomputers

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS /Watt	Watts/ CPU
TSUBAME(Opteron)	10480	800,000	50,400	63	76.336
TSUBAME(w/ClearSpee	11,200	810,000	85,000	104.94	72.321
Earth Simulator	5120	6,000,000	40,000	6.7	1171.9
ASCI Purple (LLNL)	12240	6,000,000	77,824	12.971	490.2
AIST Supercluster	3188	522,240	14400	27.574	163.81
LLNL BG/L (rack)	2048	25,000	5734.4	229.38	12.207
Next Gen BG/P (rack)	4096	30,000	16384	546.13	7.3242
TSUBAME Next Gen (2010)	40000	800,000	1000000	1250	20

TSUBAME Cooling Density Challenge

- Room 2F-B
  - 480 nodes, 1330W/node max, 42 racks
  - Rack area = 2.5m x 33.2m = 83m^2 = 922ft^2
    - Rack spaces only---Excludes CRC units
  - Max Power = x4600 nodes 1330W x 480 nodes + IB switch 3000W x 4 = <u>650KW</u>
  - Power density ~= 700W/ft^2 (!)
    - Well beyond state-of-art datacenters (500W/ft^2)
  - Entire floor area ~= 14m x 14m ~= 200m^2 = 2200 ft^2
  - But if we assume 70% cooling power as in the Earth Simulator then total is 1.1MW – still ~500W/ft^2

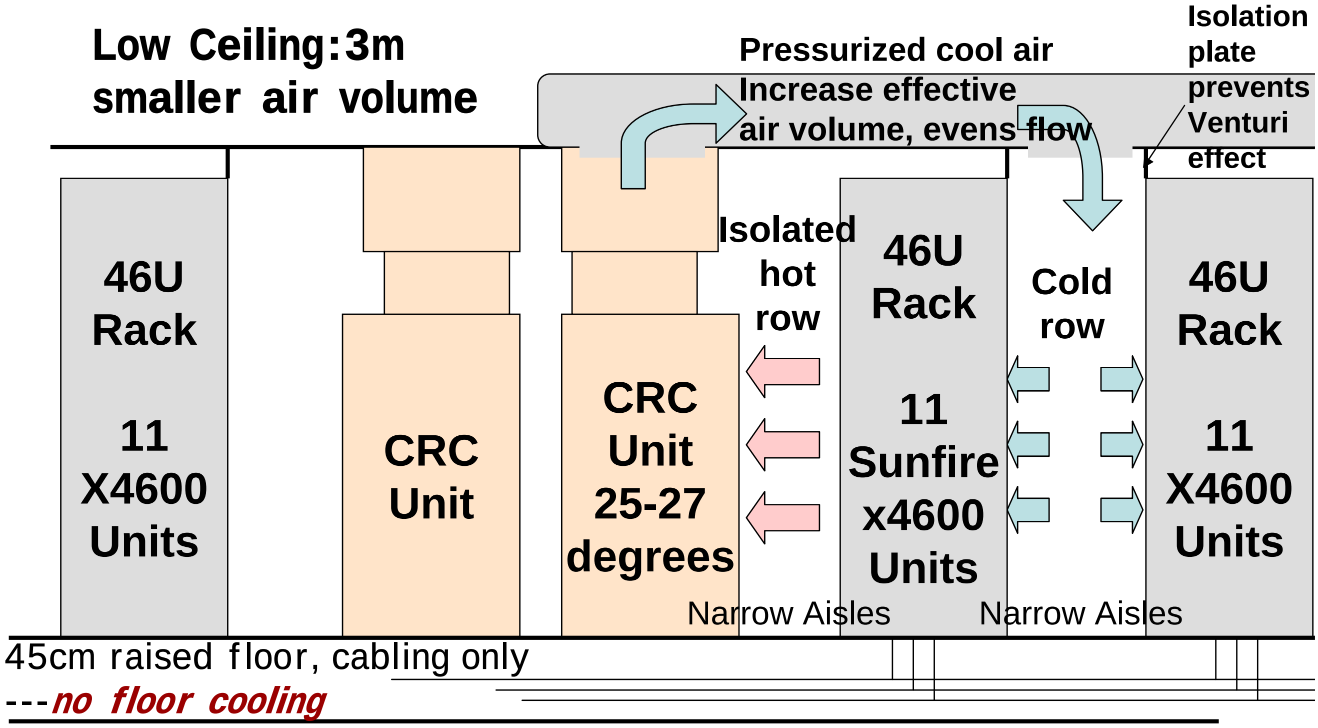


TSUBAME Physical Installation 700W/ft^2 on hatched area 500W/ft^2 for the whole room

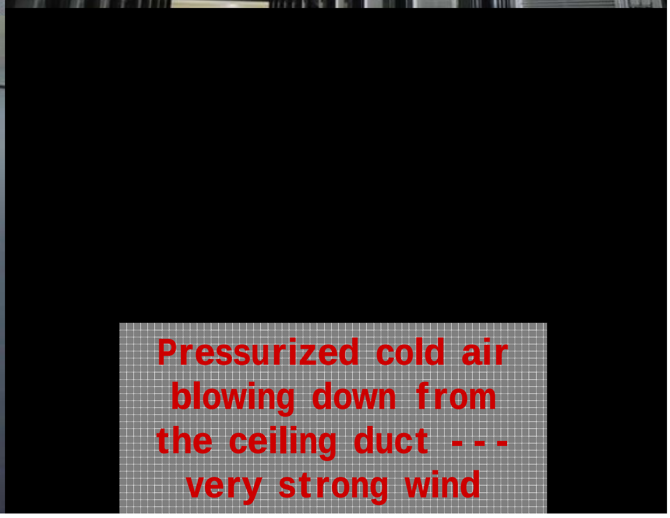
High density cooling & power reduction

Cooling and Cabling 700W/ft^2

--- hot/cold row separation and rapid airflow---



no turbulant airflow causing hotspots



Everybody's Supercomputer TSUBAME as a Grid Resource

Breaking the Traditional Supercomputer and Grid Economics

みんなのスパコン
"Everybody's Supercomputer"

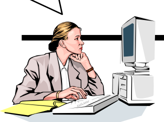
TITECH GRID

Isolated High-End

Massive Usage Env. Gap

- •Different usage env. from
- •No HP sharing with client's PC
- •Special HW/SW, lack of ISV support
- •Lack of common development env. (e.g. Visual Studio)
- •Simple batch based, no interactive usage, good UI

Might as well use my Laptop



IT Consolidation: Seamless integration of supercomputers with *end-user and enterprise* environment

Hmm, it's like my personal machine

"Everybody's Supercomputer"

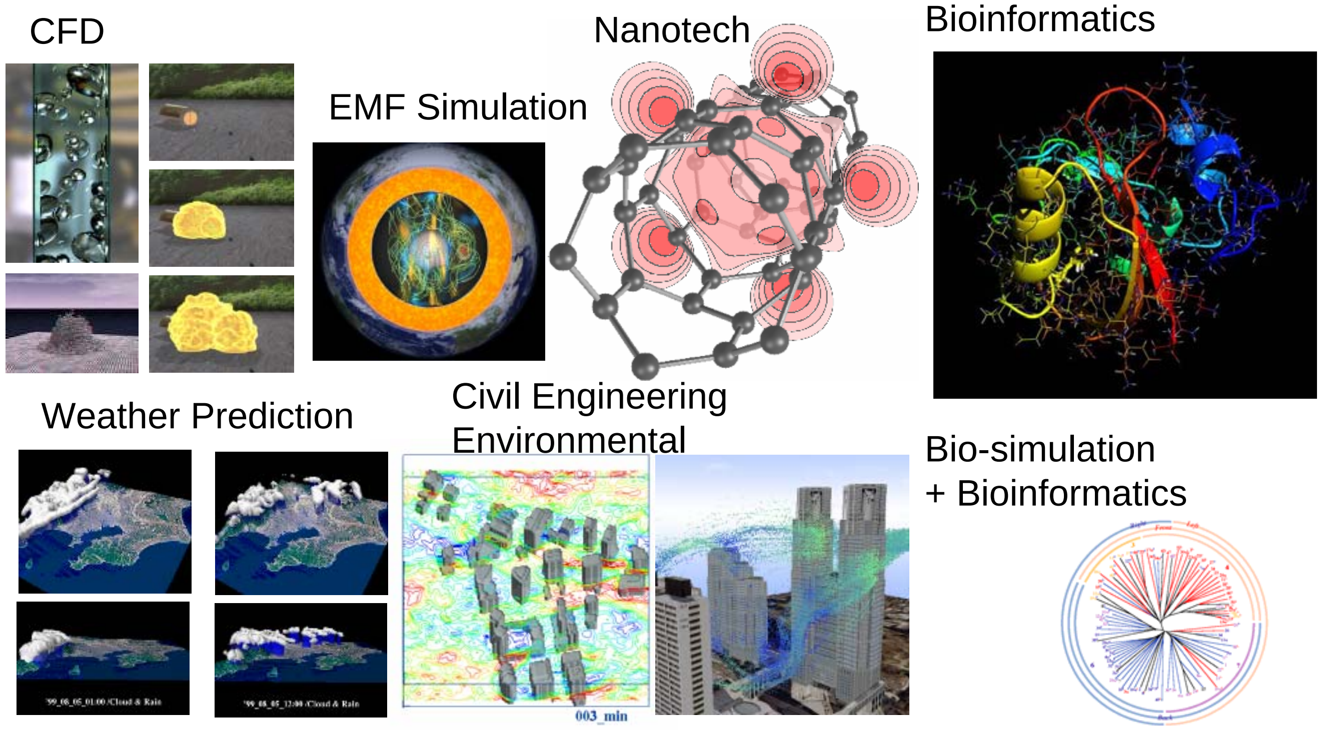
Seamless,Ubiquitous access and usage

=>Breakthrough Science through Commoditization of Supercomputing and Grid Technologies

みんなのスパコン

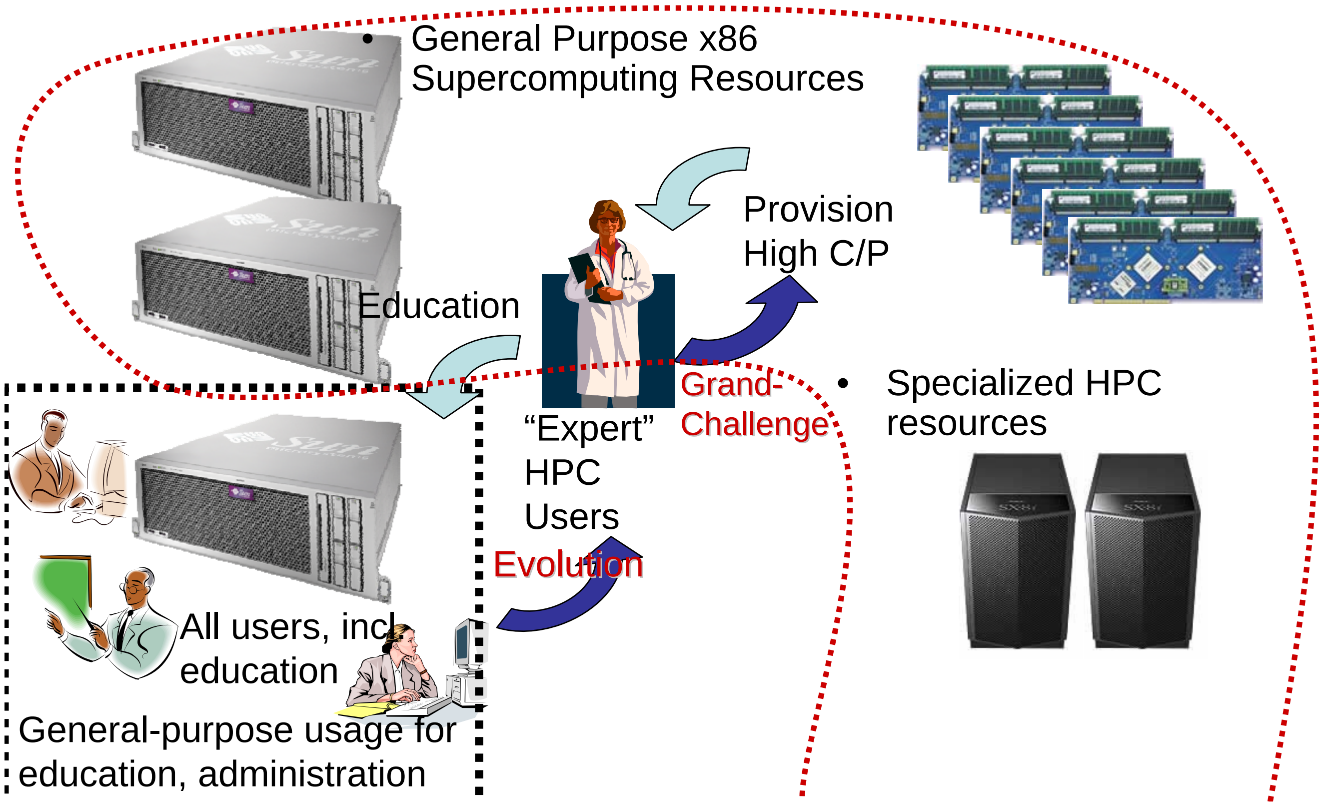
Grand Challenge Supercomputing @ Titech

100 Teraflops-scale computing with Petascale Storage



みんなのスパコン

Incubating the Next-Generation HPC Users



みんなのスパコン

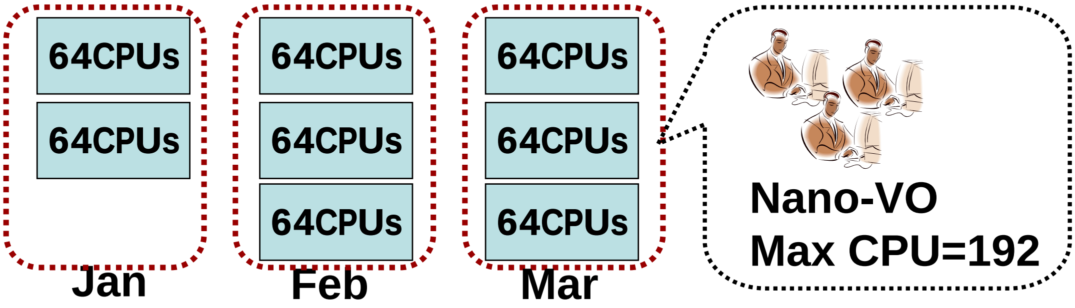
VO-Based Scheduling/Accounting

Dynamic machine-level resource allocation

SLA > BES > Small

- Q. How do you make capability and capacity coexist?
- Three account types
  - Small Usage: no prior allocations, small, ubiquitous resource usage (up to 16 CPUs, etc.), *free*
  - Service Level Agreement: Exclusive use of each SMP node, allocation charge on node-time basis, expensive
  - Best Effort (new): Internet-Inspired, Inexpensive
    - Flat allocation fee per each UNIT
    - Each UNIT is max 64 CPU usage at any given time
    - Group/VO-based accounting, multiple UNITs purchasable

Over 1300 users



みんなのスパコン

TITECH GRID

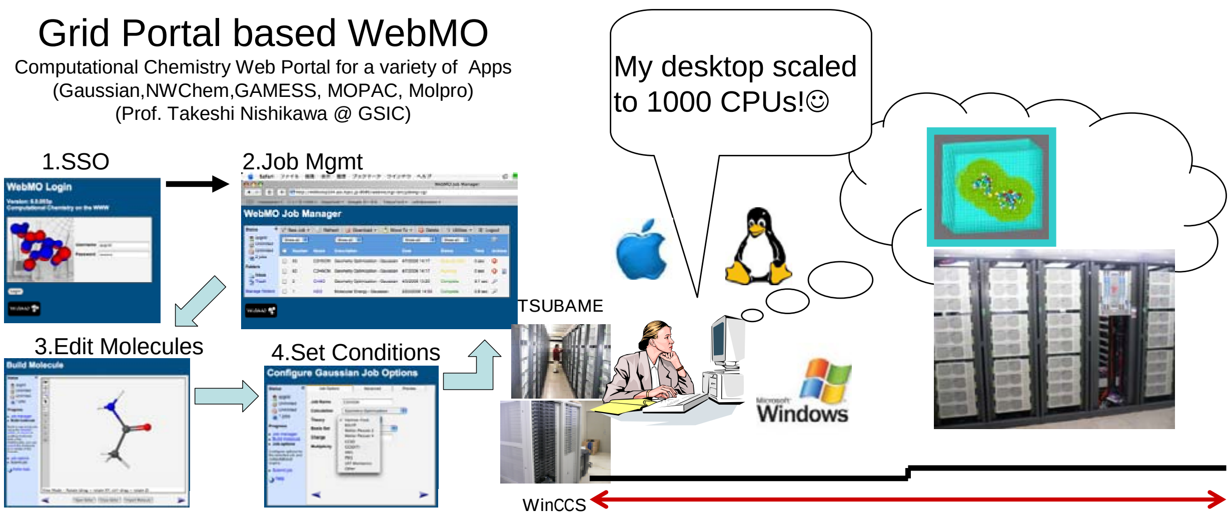
Supercomputing in All Educational Activities

Over 10,000 users

- High-End education using supercomputers in undergrad labs
  - High end simulations to supplement "physical" lab courses
- Seamless integration of lab resources to SCs w/grid technologies
- Portal-based application usage

Grid Portal based WebMO

Computational Chemistry Web Portal for a variety of Apps (Gaussian,NWChem,GAMESS, MOPAC, Molpro)
(Prof. Takeshi Nishikawa @ GSIC)



みんなのスパコン

TSUBAME General Purpose DataCenter Hosting

As a core of IT Consolidation

All University Members == Users

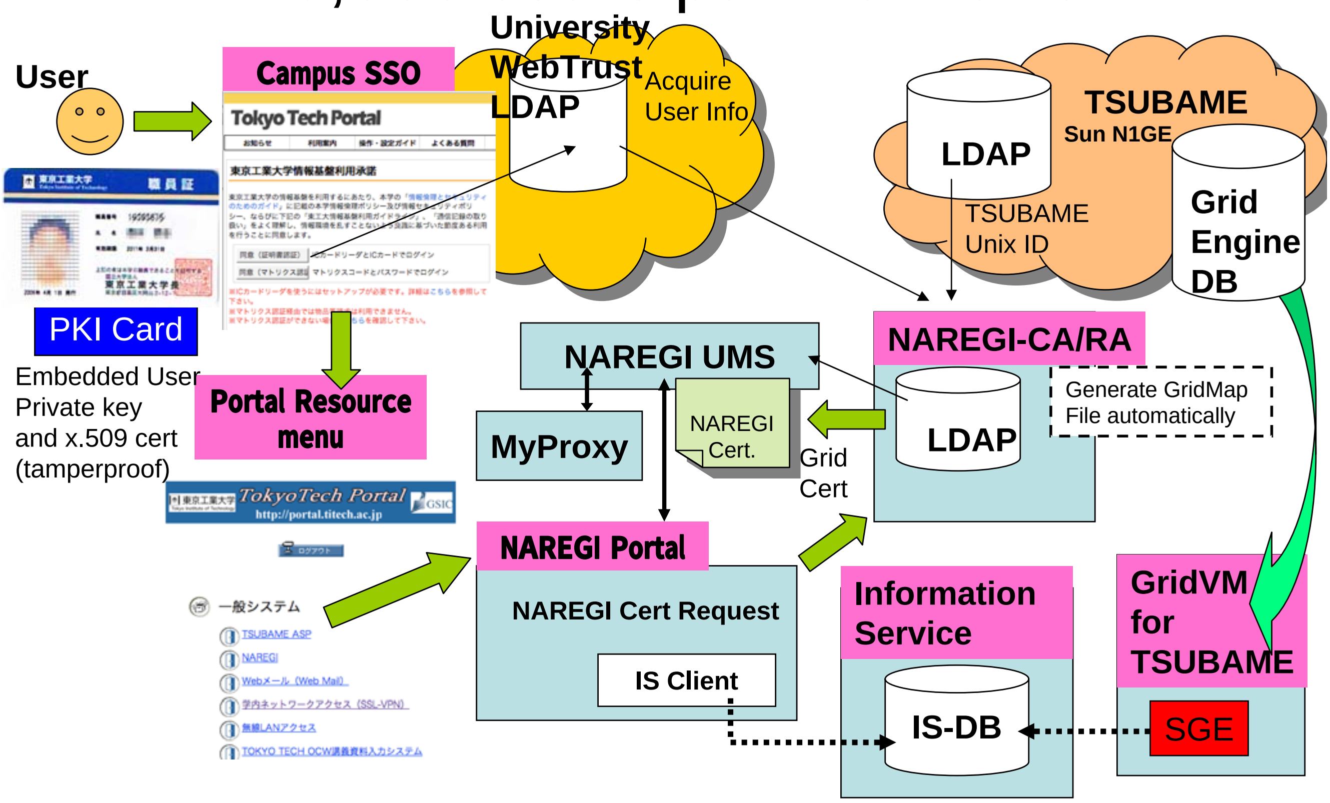
- Campus-wide AAA Sytem (April 2006)
  - 50TB (for email), 9 Galaxy1 nodes
- Campus-wide Storage Service (NEST)
  - 10s GBs per everyone on campus
    PC mountable, but accessible directly from TS
  - Research Repository
- CAI, On-line Courses
  (OCW = Open CourseWare)
- Administrative Hosting (VEST)

36

Titech PKI-based SSO/AAA & IT Services

SSO WebMO Portal Access

NAREGI Beta 2 Deployment@ Titech
> 10,000 users per institution



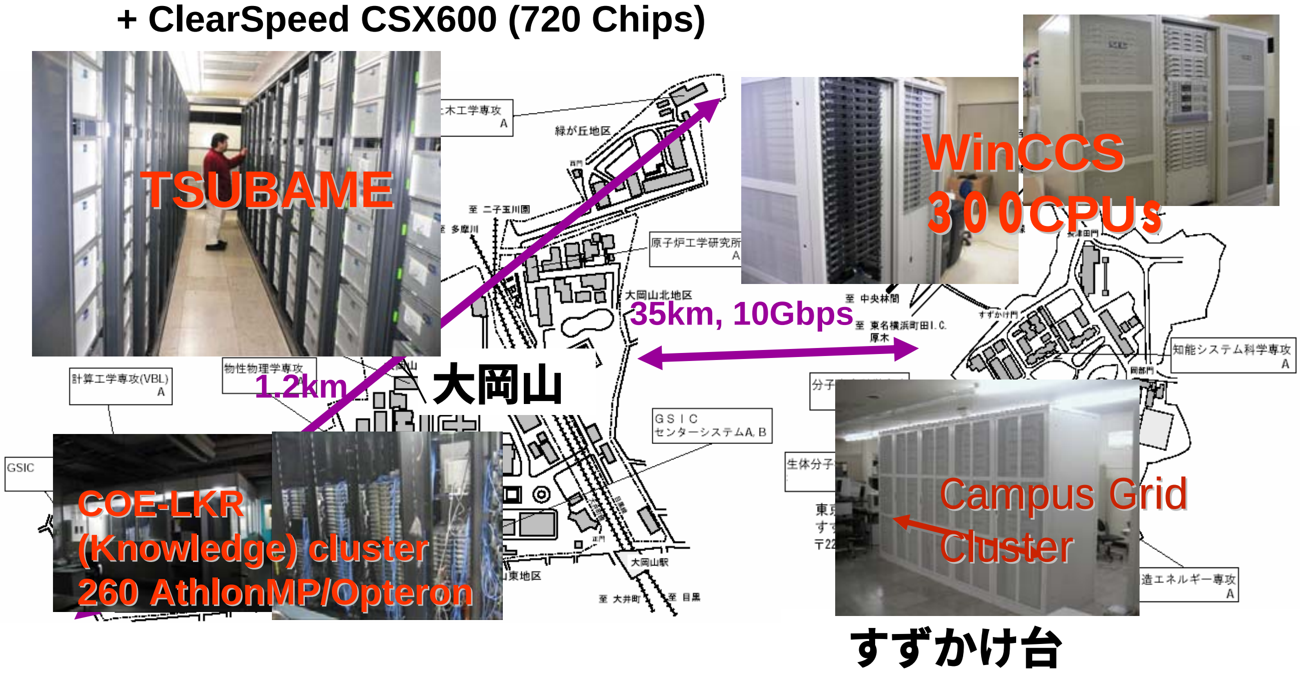
Titech Supercomputer Contest
"The 12th SuperCon"

- High-school students (~10 out of 50 team apps)
- Since 1995: Cray => Origin => TSUBAME
- 700 CPUs allocated for 1 week

Multple Testimonies
"TSUBAME was so easy to use, just like my PC, but much faster!"

Titech Campus Grid 2006
- An x86 "DataCenter" Grid -

- **~13,000 CPUs, 90 TFlops, ~26 TBytes Mem, ~1.1 PBytes HDD**
- **CPU Cores: x86: TSUBAME (~10600), Campus Grid Cluster (~1000), COE-LKR cluster (~260), WinCCS (~300)**
  + ClearSpeed CSX600 (720 Chips)

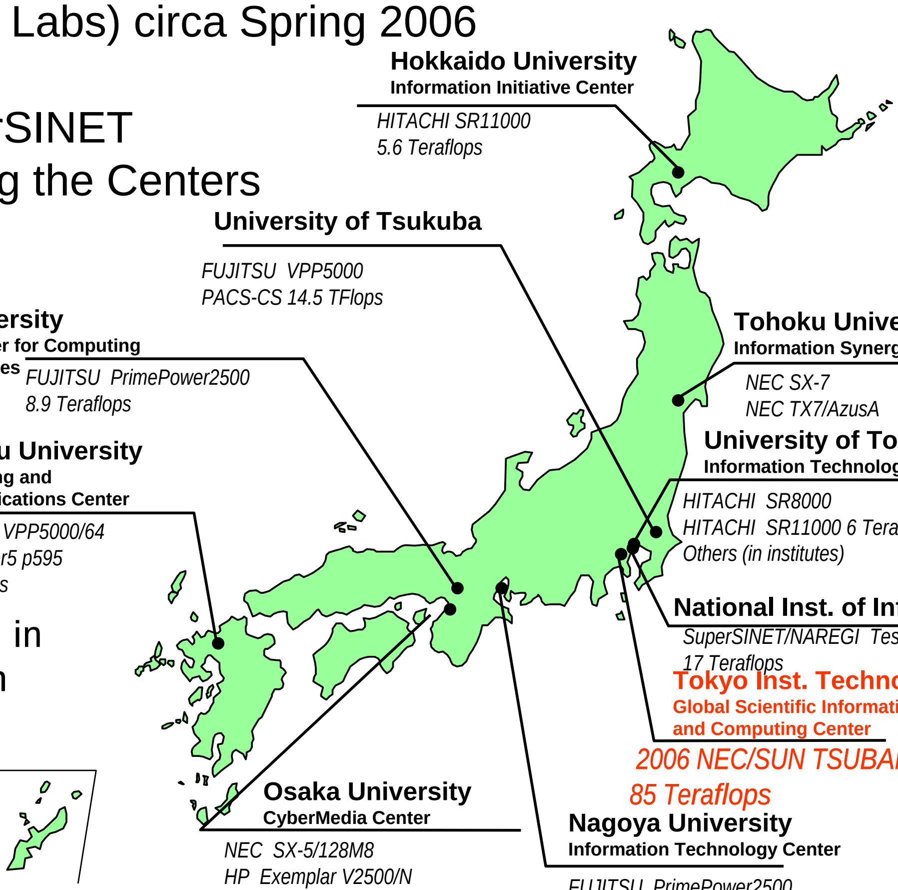


TSUBAME Siblings ---The Domino Effect on Major Japanese SCs

- Sep. 6th, U-Tokyo, Kyoto-U, and U-Tsukuba announced "common procurement procedure" for the next gen SCs in 1H2008
  - 100-150 TFlops
  - HW: x86 cluster-like SC architecture
  - NW: Myrinet10G or IB + Ethernet
  - SW: Linux+SCore, common Grid MW
- Previously, ALL centers ONLY had dedicated SCs
- Other centers will likely follow...
  - No other choices to balance widespread usage, performance, and prices
  - Makes EVERY sense for University Mgmt.
- (VERY) standardized SW stack and HW configuration
  - Adverse architecture diversity has been *impediment* for Japanese Grid Infrastructure

Japan's 9 Major University Computer Centers (excl. National Labs) circa Spring 2006

10Gbps SuperSINET Interconnecting the Centers

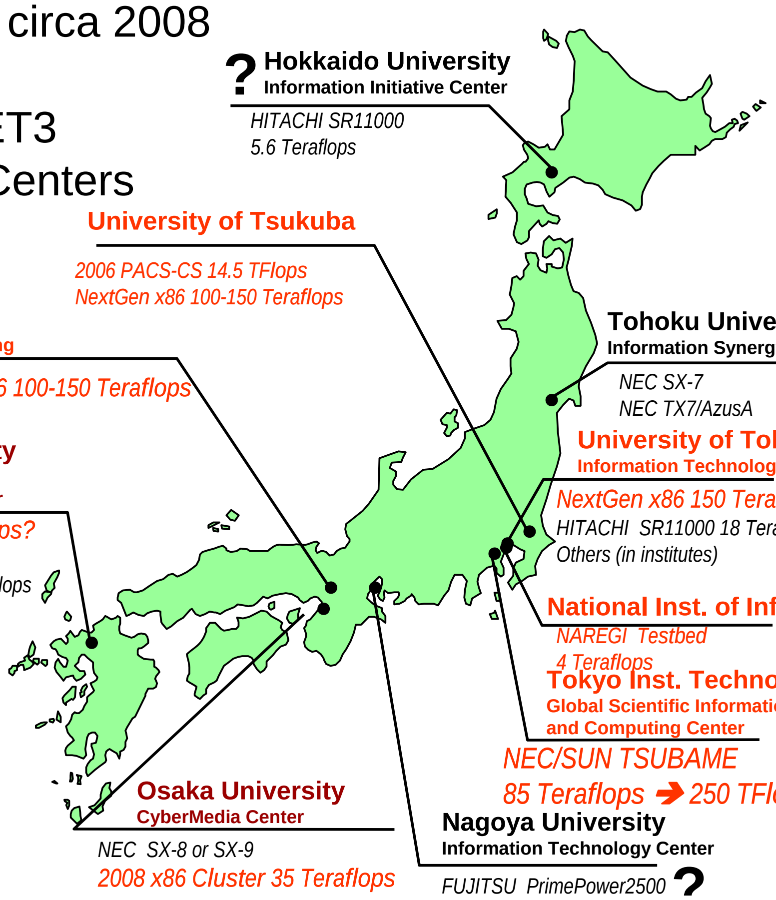


~60 SC Centers in Japan incl. Earth Simulator

- 10 Petaflop center by 2012

Japan's 9 Major University Computer Centers (excl. National Labs) circa 2008

>40Gbps SuperSINET3 Interconnecting the Centers



Hokkaido University ?
Information Initiative Center
HITACHI SR11000
5.6 Teraflops

University of Tsukuba
2006 PACS-CS 14.5 TFlops
NextGen x86 100-150 Teraflops

Kyoto University
Academic Center for Computing and Media Studies
NextGen x86 100-150 Teraflops

Tohoku University ?
Information Synergy Center
NEC SX-7
NEC TX7/AzusA

University of Tokyo
Information Technology Center
NextGen x86 150 Teraflops
HITACHI SR11000 18 Teraflops
Others (in institutes)

Kyushu University
Computing and Communications Center
2007 x86 50 TeraFlops?
Fujitsu Primequest?
IBM Power5 p595 5 Teraflops

National Inst. of Informatics
NAREGI Testbed
4 Teraflops

Tokyo Inst. Technology
Global Scientific Information and Computing Center
NEC/SUN TSUBAME
85 Teraflops ➔ 250 TFlops?

x86 TSUBAME sibling domination

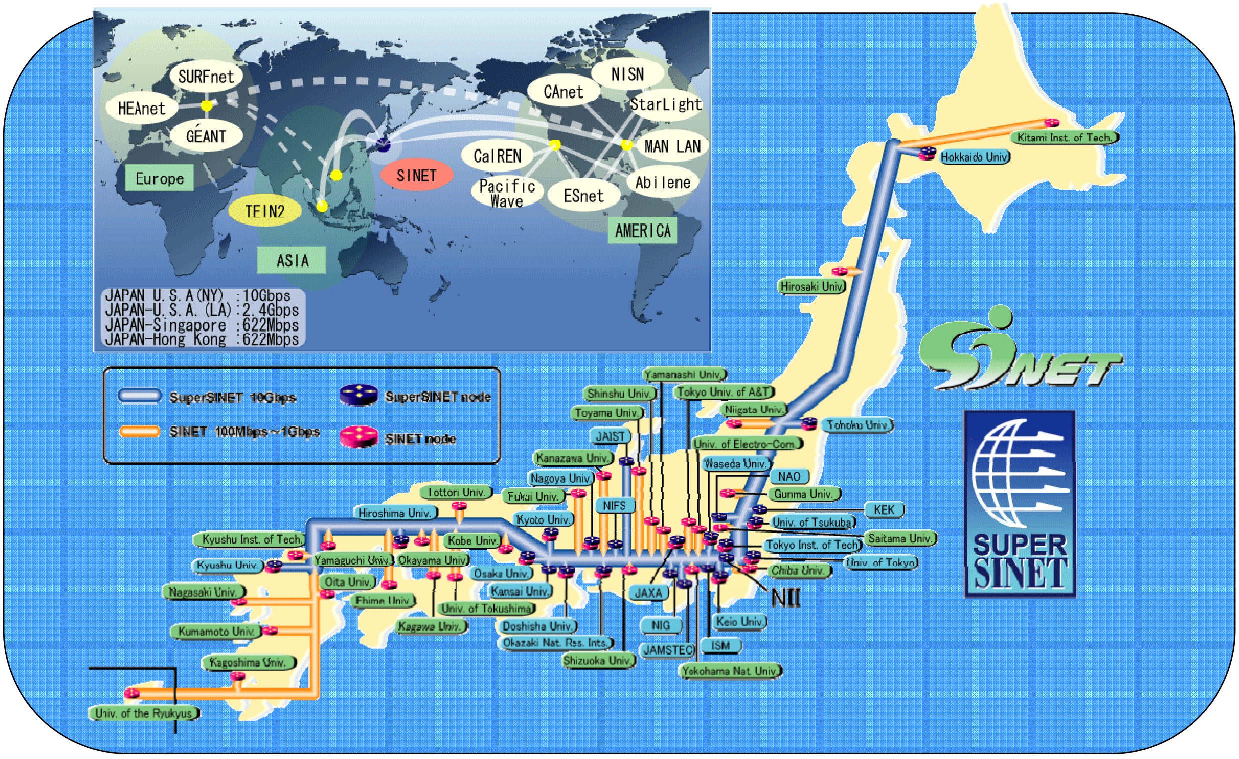
Osaka University
CyberMedia Center
NEC SX-8 or SX-9
2008 x86 Cluster 35 Teraflops

Nagoya University
Information Technology Center
FUJITSU PrimePower2500 ?
11 Teraflops

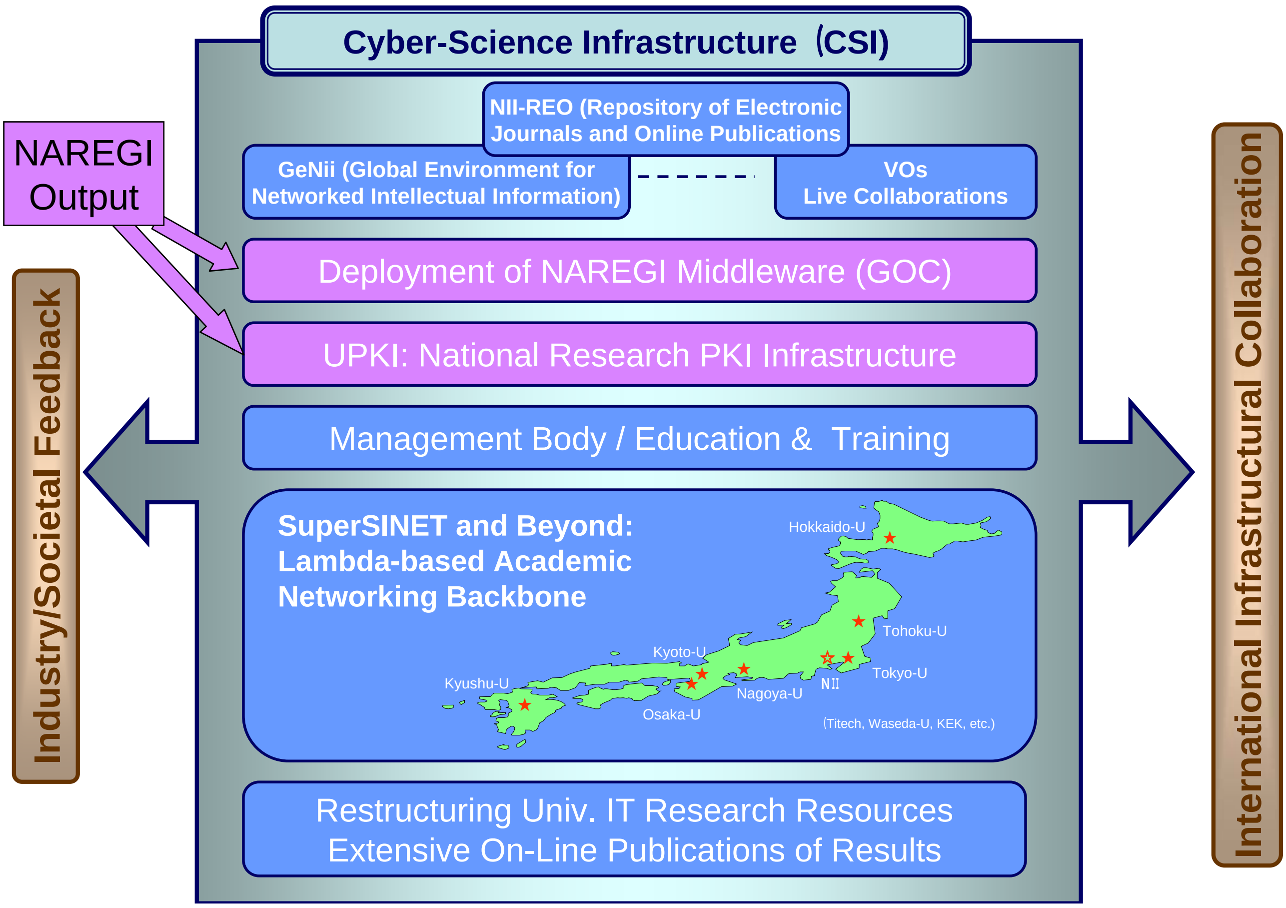
Still - 10 Petaflop center by 2012

Super SINET3 (new!)

Dynamic L1/L2/L3 provisioning 40 Gbps Backbone



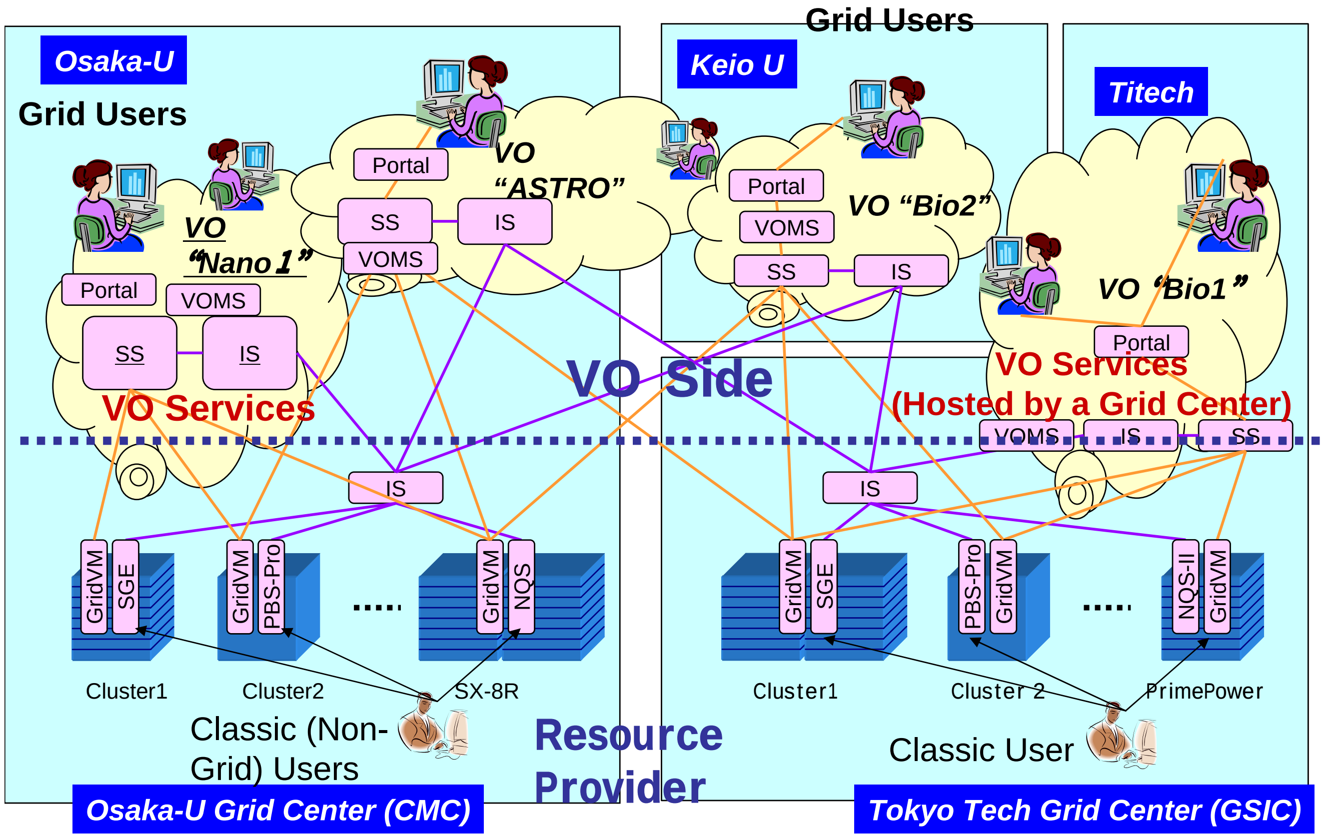
Japanese CyberScience Infrastructure Project



NAREGI Beta 2 - v.1.0 Highlights

- Production Release Candidate (2Q 2007)
- Lots of bug, performance & stability fixes
- Stable WS(RF) components and APIs (+ Globus 4.0.3)
- RPM and Dynamic, VM-based deployment
- VO and “Resource Provider” decoupling for multiple VO management by VOs and Centers
- Integration of NAREGI WF and Ninf-G GridRPC
- More BQ and systems support
  - NEC SX-NQS, SGE, Fujitsu NQS… (Condor?)
- Flexible Job submission and WF management
  - Non-grid jobs, non-reserved jobs, various WF tools
- EGEE-GIN Interoperation (new)
- Various Administration and Logging Tools
- Support from dedicated NAREGI support team

NAREGI β2 Operational Model



GIN (Grid Interoperation Now)

- An activity of OGF for interoperation among production grids
- Major grid projects are participating
  - EGEE, NAREGI, UK National Grid Service, NorduGrid, OSG, PRAGMA, TeraGrid, ...
- Trying to identify islands of interoperation between production grids and grow those islands
- Areas
  - GIN-auth: Authorization and Identity Management
  - GIN-data: Data Management and Movement
  - GIN-jobs: Job Description and Submission
  - GIN-info: Information Services and Schema
  - GIN-ops: Operations Experience of Pilot Test Applications

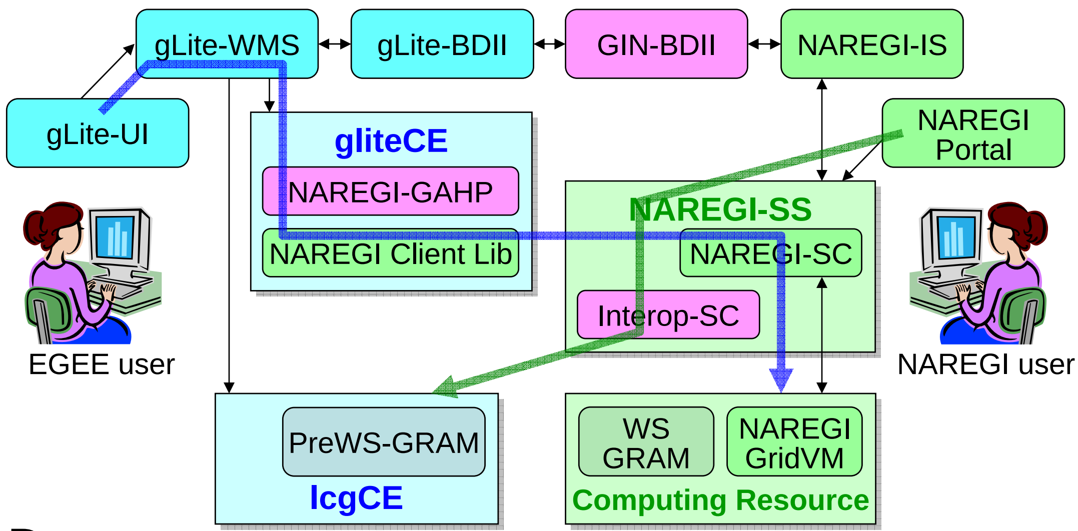
GIN

NAREGI GIN Activities

- Developing an interoperation island with EGEE
- Developing an Interoperation island with WS-GRAM based grids
- JSDL interoperability (for Phase-2)

GIN-jobs: NAREGI-EGEE Architecture

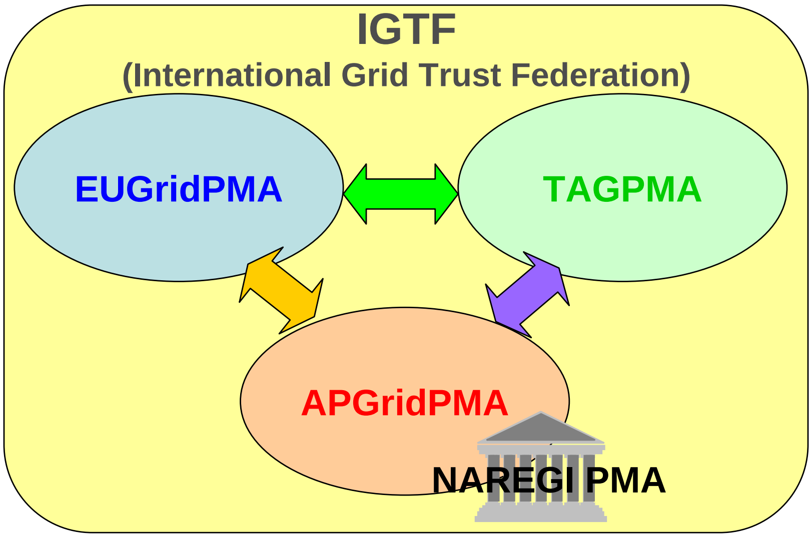
- Architecture



- Demo
  - NAREGI → EGEE: using NAREGI Workflow
  - EGEE → NAREGI: using glite WMS commands

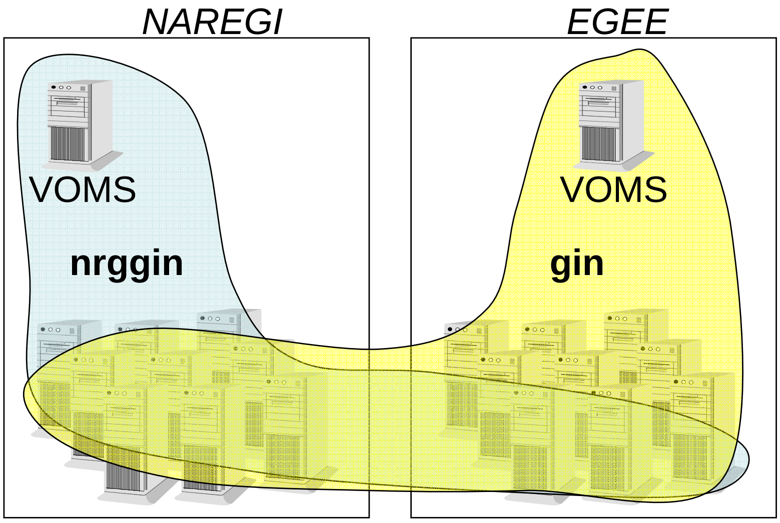
Authentication

- IGTF is framework of International Grid Trust Federation.
- IGTF consists of APGridPMA, EUGridPMA and TAGPMA.
- NAREGI CA joined the APGrid PMA.
- NAREGI CA has been approved as a production-level CA by APGridPMA.
- GSI compliant with x.509 proxy certificates for authentication.
- It has become available to use grid computing easily on the worldwide Internet by IGTF.



VO Management

- The GIN VO is a VOMS service.
- NAREGI uses VOMS as VO management system.
- Transport of supported authorization attributes via VOMS extensions.
- VO names are expected to abide by the VO naming conventions described in GIN VO Naming in order to avoid name conflicts between grids.
- All members of GIN VO should observe AUP(Acceptable Use Policy).

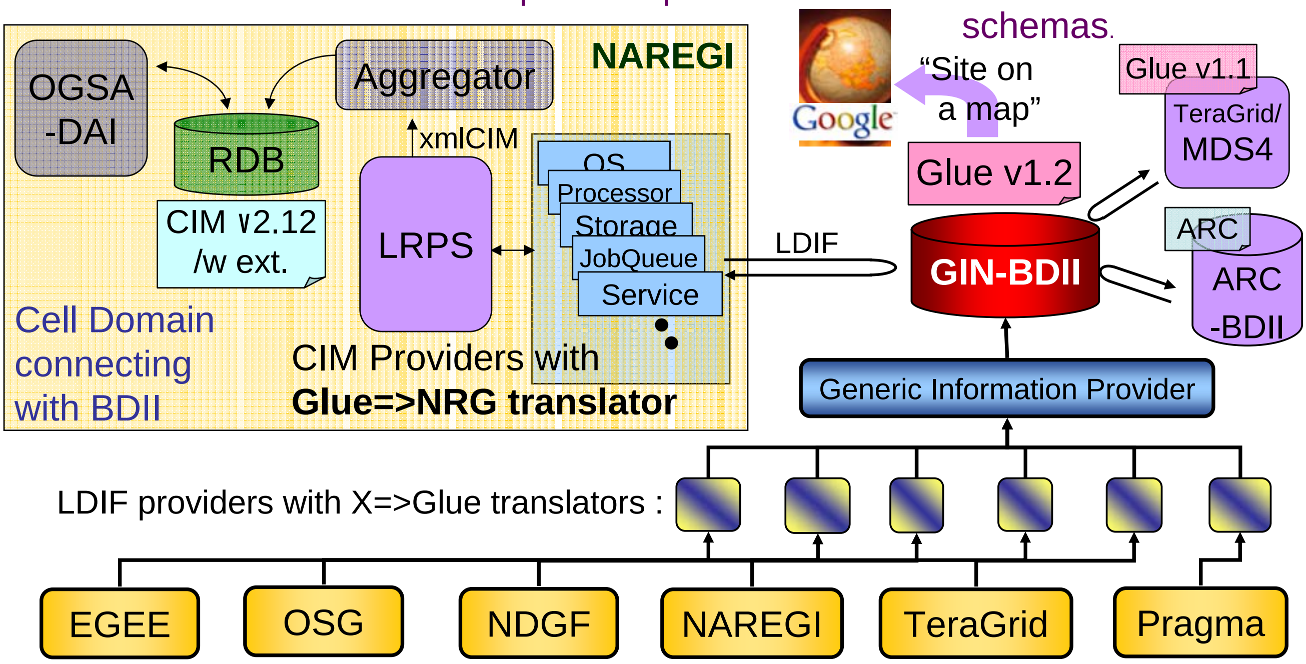


reference
http://forge.gridforum.org/sf/wiki/do/viewPage/projects.gin/wiki/GINAuth

GIN-info: Architecture

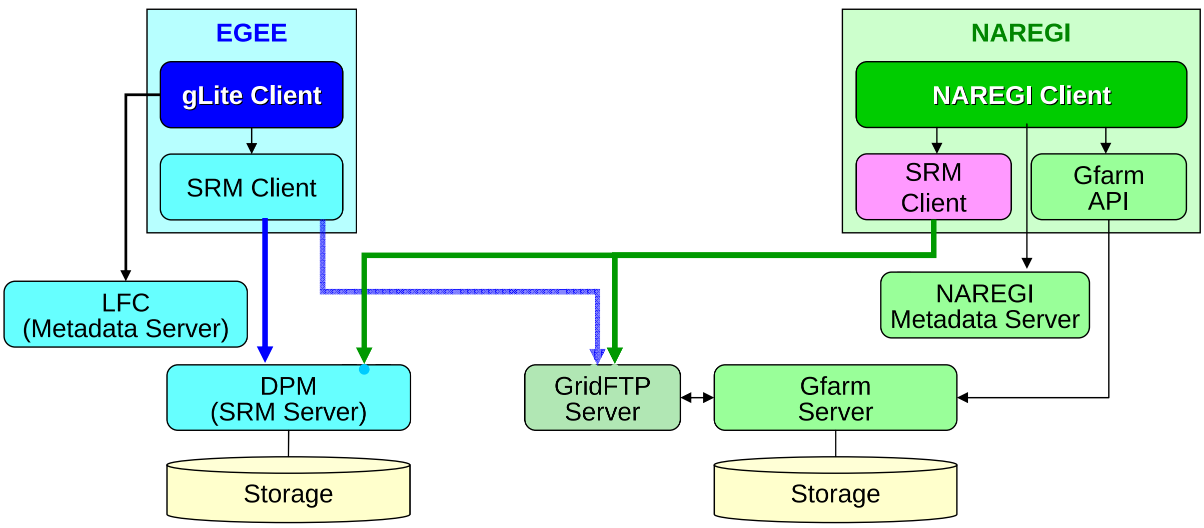
All of grid information can be retrieved by each of grid in its fashion WRT resource description schema, data format, query language, client API, …

Each information service in grid acts as an information provider for the other and translator embedded in the provider performs conversion between different schemas.



GIN-data: Architecture

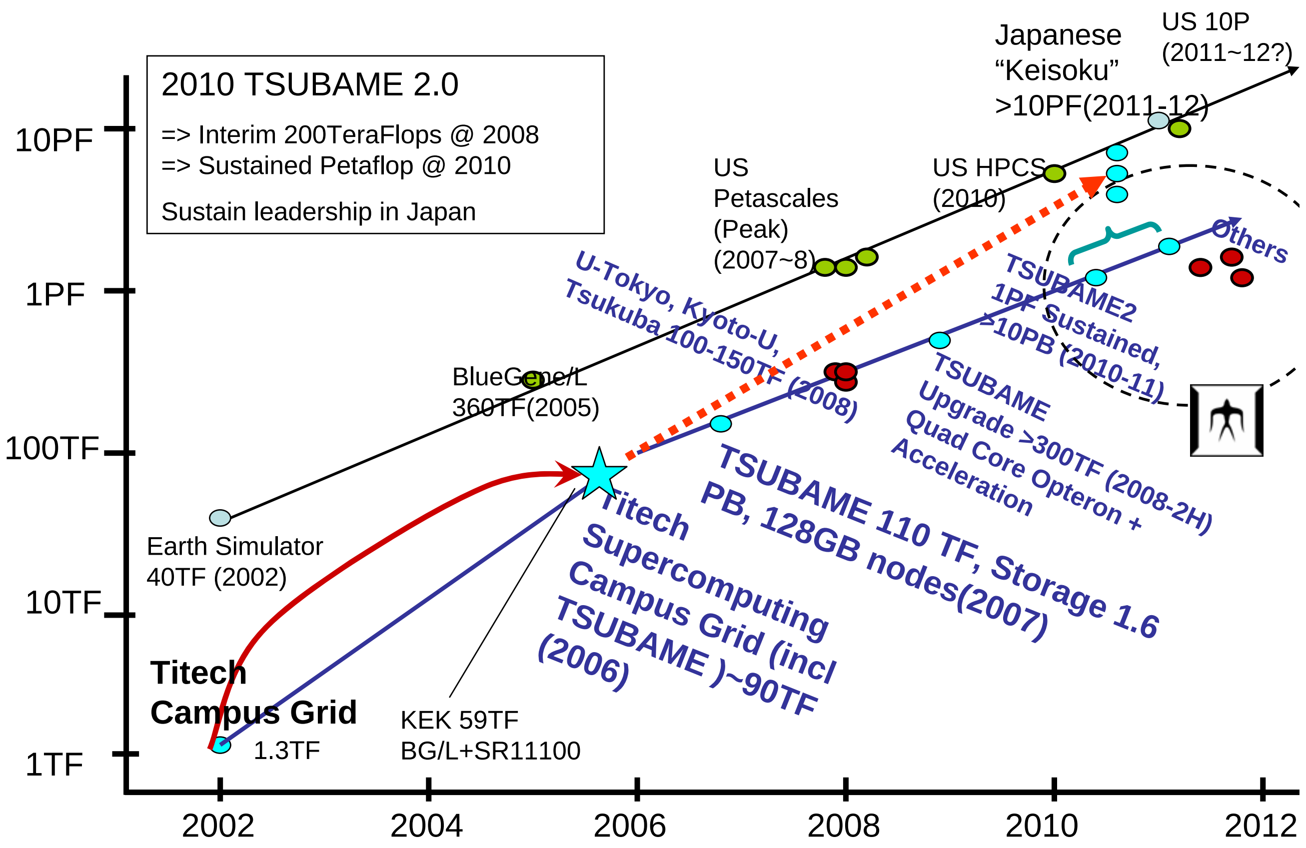
- NAREGI and EGEE gLite clients can access to both data resources (e.g., bi-directional file copy) using SRM interface.
- GridFTP is used as its underlying file transfer protocol.
- File catalog (metadata) exchange is planned.



NAREGI GIN Summary

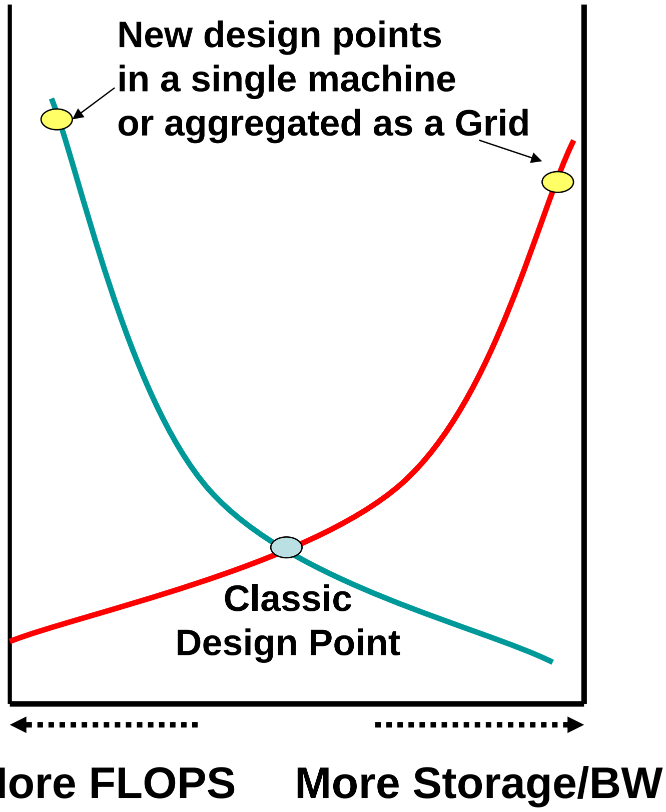
- NAREGI developed EGEE-NAREGI island as an activity of GIN
  - Bilateral information exchange
  - Bilateral job submission
  - Bilateral file exchange
  - Interoperable security properties
- Next steps
  - Improve interoperation interfaces and functions
    - WS-GRAM, BES, JSDL, …
  - Grow the island with other EGEE partners
  - KEK will use NAREGI-EGEE interoperation environment for their high energy physics calculations

Scaling Towards Petaflops



Future Petascale Designs

- Assuming Upper bound on Machine Cost
- A single machine entails compromises in all applications
- Heterogeneous Grids of Large Resources would allow multitple design points to coexist
- And this also applies to a single machine as well



Upscaling the Resources to a Petascale Grid