TSUBAME---A Year Later

Satoshi Matsuoka, Professor/Dr.Sci.

Global Scientific Information and
Computing Center

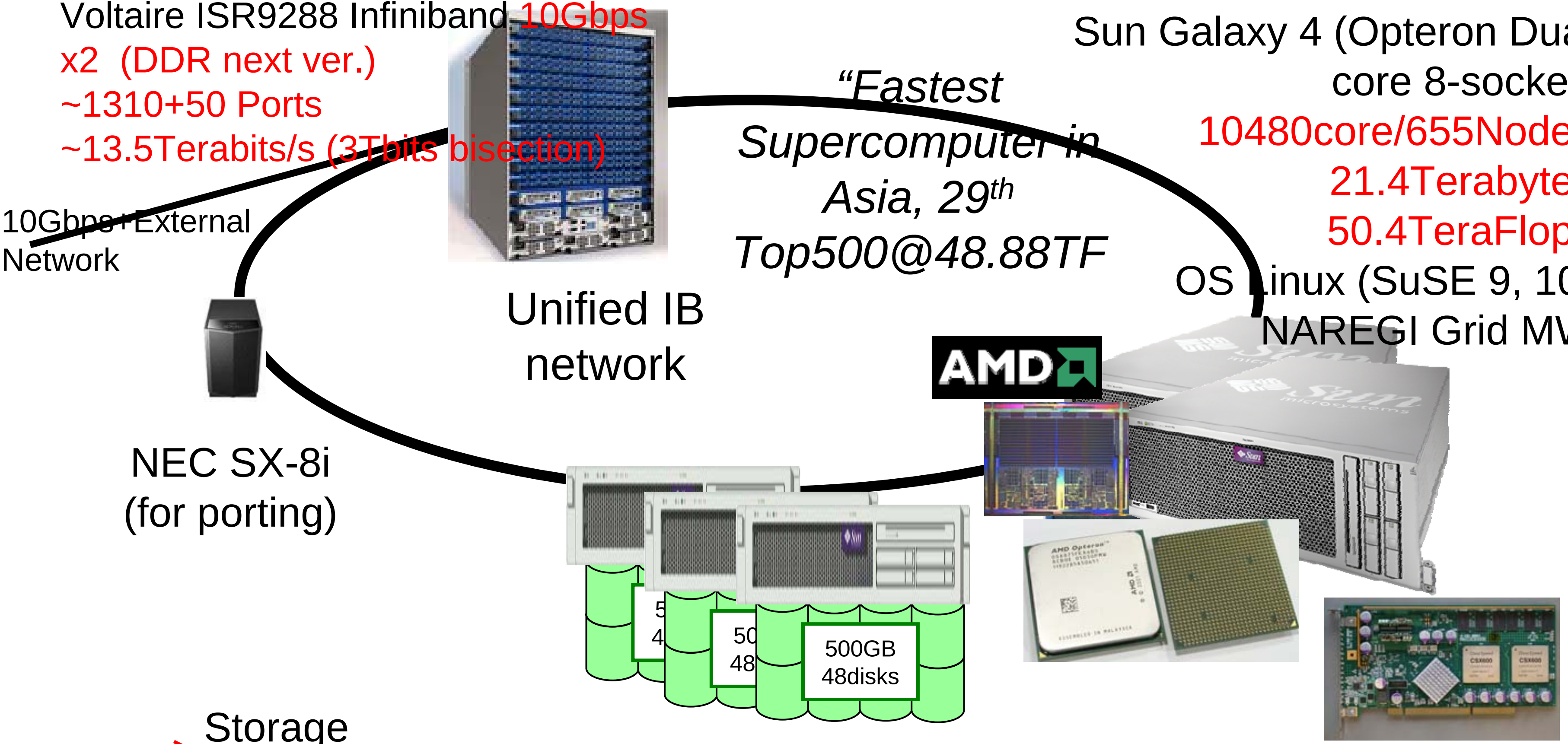
Tokyo Inst. Technology
& NAREGI Project National Inst. Informatics

EuroPVM/MPI, Paris, France, Oct. 2, 2007

Topics for Today

- Intro
- Upgrades and other New stuff
- New Programs
- The Top 500 and Acceleration
- Towards TSUBAME 2.0

The TSUBAME Production “Supercomputing Grid Cluster” Spring 2006-2010



Voltaire ISR9288 Infiniband 10Gbps x2 (DDR next ver.)
~1310+50 Ports
~13.5Terabits/s (3Tbits bisection)

10Gbps+External Network

“Fastest Supercomputer in Asia, 29th Top500@48.88TF

Sun Galaxy 4 (Opteron Dual core 8-socket)
10480core/655Nodes
21.4Terabytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

Unified IB network

NEC SX-8i
(for porting)

Storage
1.5PB ~~1.0~~ Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
70GB/s ~~50~~GB/s aggregate I/O BW

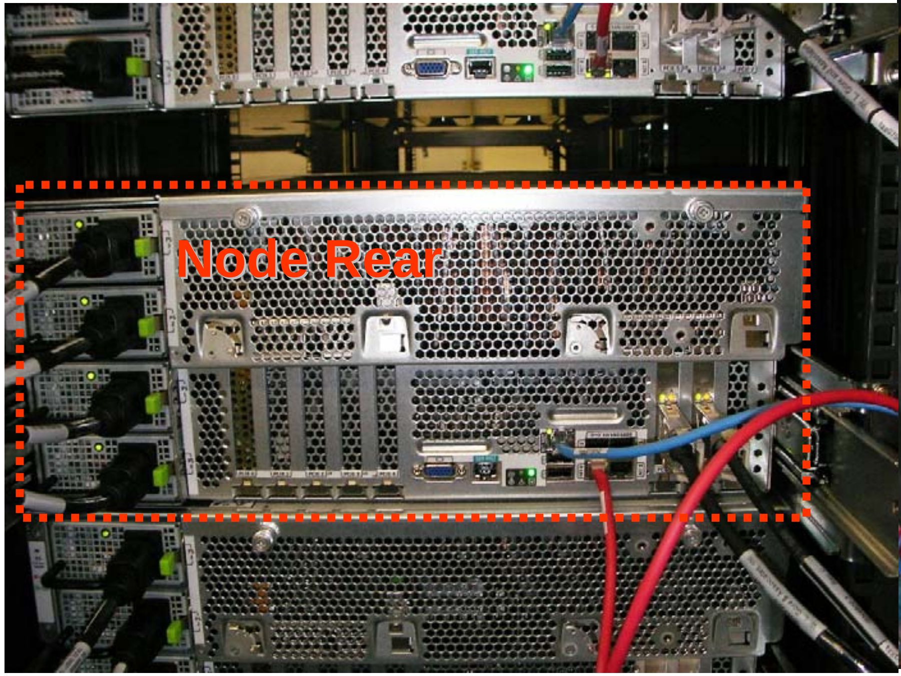
ClearSpeed CSX600
SIMD accelerator
360 boards,
35TeraFlops(Current))

Titech TSUBAME

~76 racks

350m2 floor area

1.2 MW (peak)



Node Rear

Local Infiniband Switch (288 ports)

Currently 2GB/s / node Easily scalable to 8GB/s / node

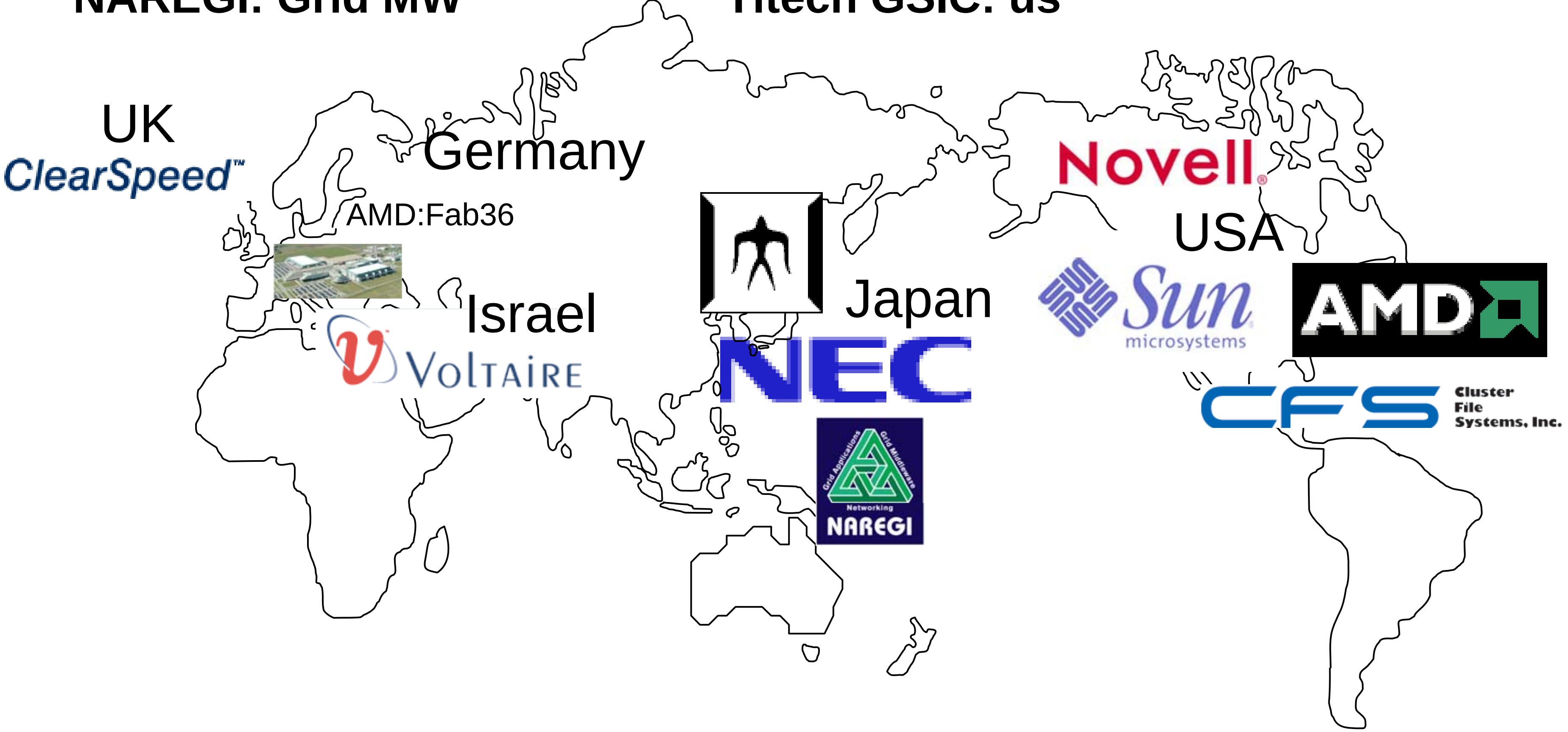
~500 TB out of 1.1PB

Cooling Towers (~32 units)

TSUBAME assembled like iPod…

NEC: Main Integrator, Storage, Operations
SUN: Galaxy Compute Nodes, Storage, Solaris
AMD: Opteron CPU
ClearSpeed: CSX600 Accel.
Novell: Suse 9/10
NAREGI: Grid MW

Voltaire: Infiniband Network
CFS: Parallel FSCFS
Titech GSIC: us



The racks were ready
Nodes arrives in mass

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 (32/64/128(new)GB)
    - 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 Architecture =

Commodity PC Cluster

+

Traditional FAT node Supercomputer

+

The Internet & Grid

+

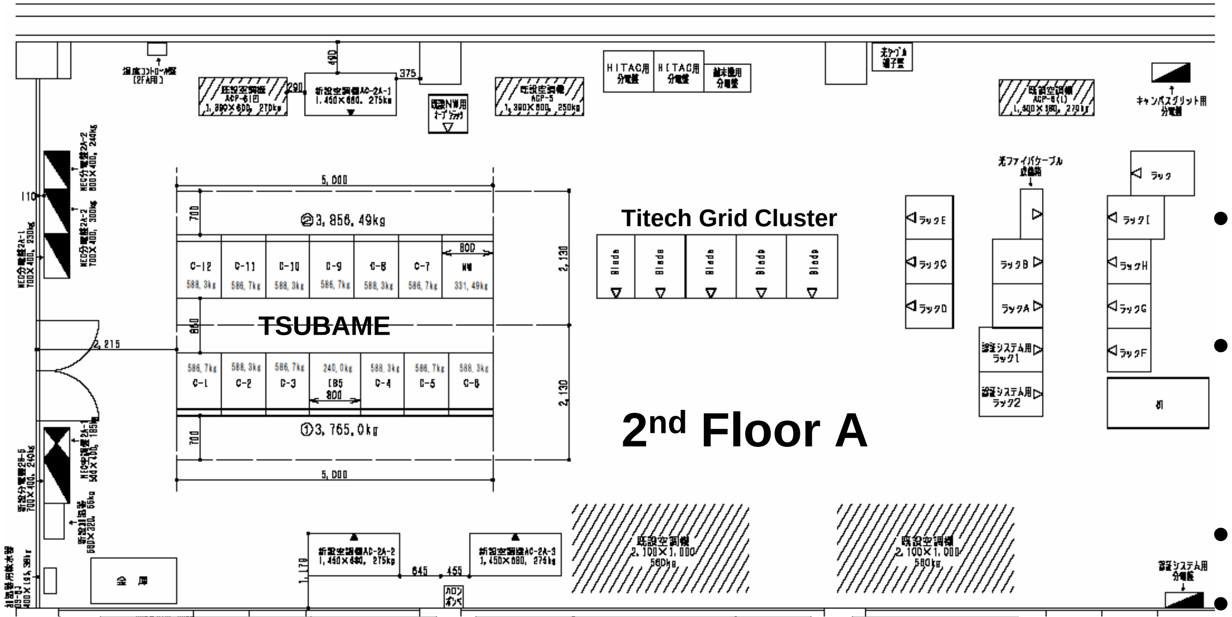
(Modern) Commodity SIMD-Vector Acceleration

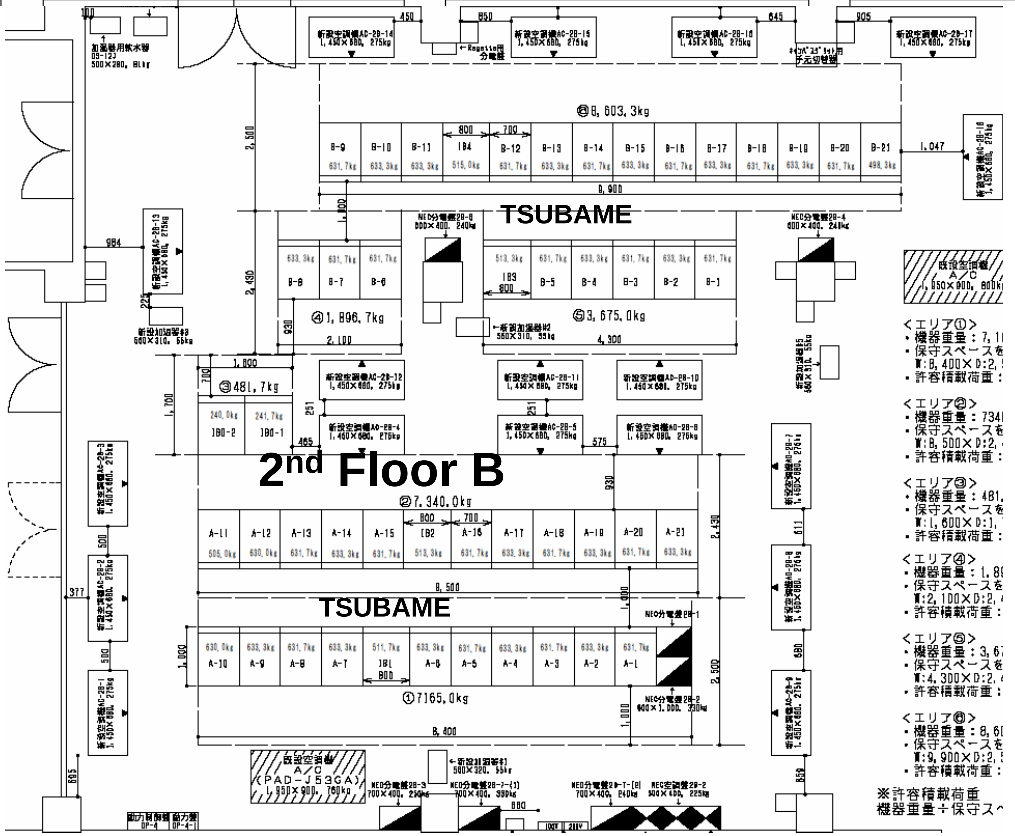
+

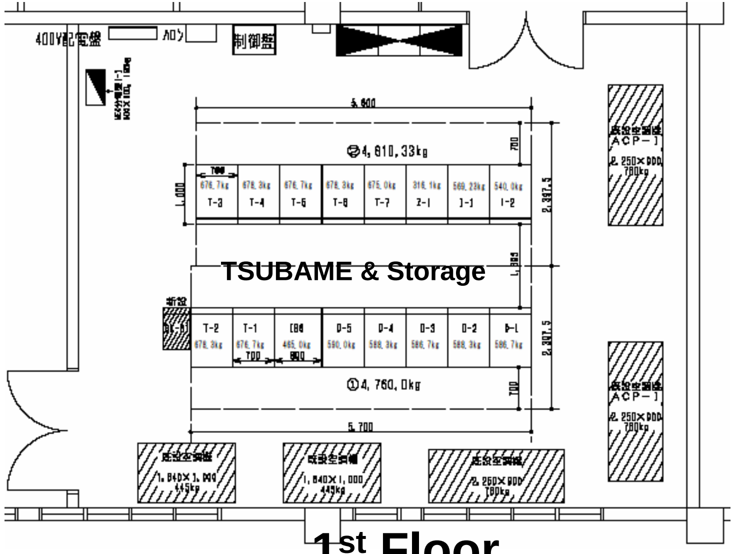
iPod (HW integration & enabling services)

TSUBAME Physical Installation

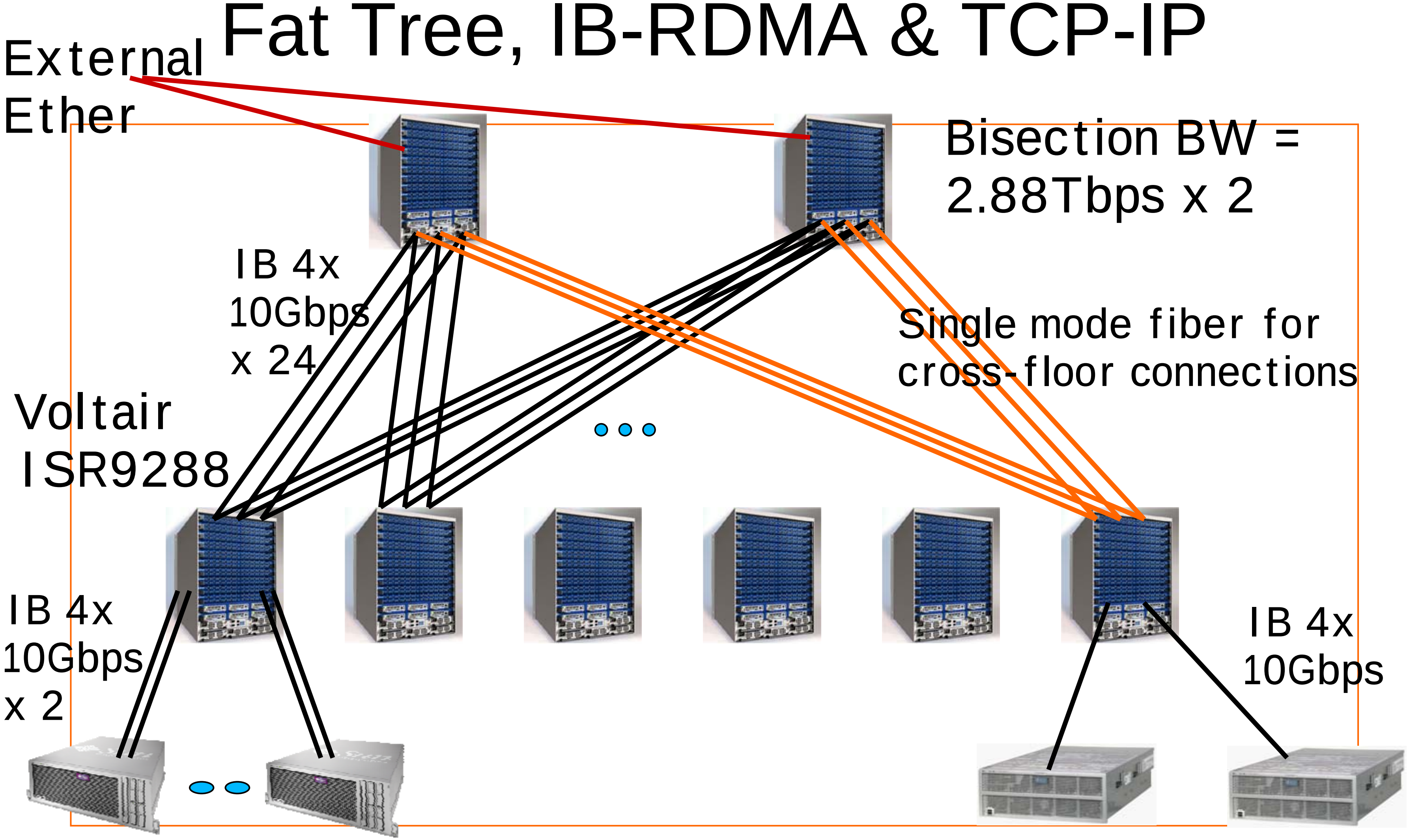
- 3 rooms (600m^2), 350m^2 service area
- 76 racks incl. network & storage, 46.3 tons
  - – 10 storage racks
- 32 AC units, 12.2 tons
- Total 58.5 tons (excl. rooftop AC heat exchangers)
- Max 1.2 MWatts
- ~3 weeks construction time







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

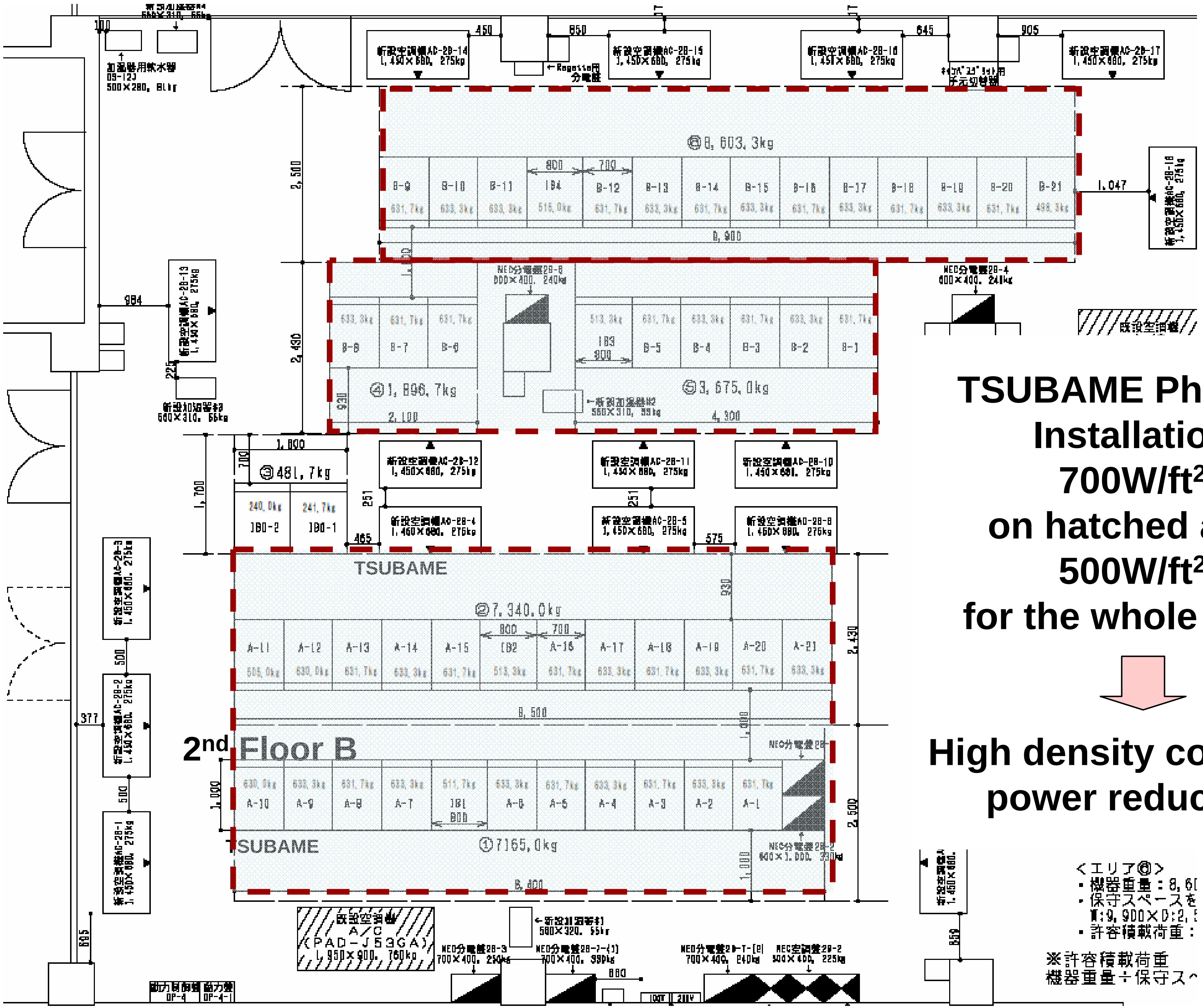
The Benefits of Being “Fat Node”

- 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**	**32~128(new)GB**
IBM BG/L	2	5.6 GF	0.5~1GB
Typical PC Cluster	2~4	10~40GF	1~8GB

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 = 650KW
  - 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

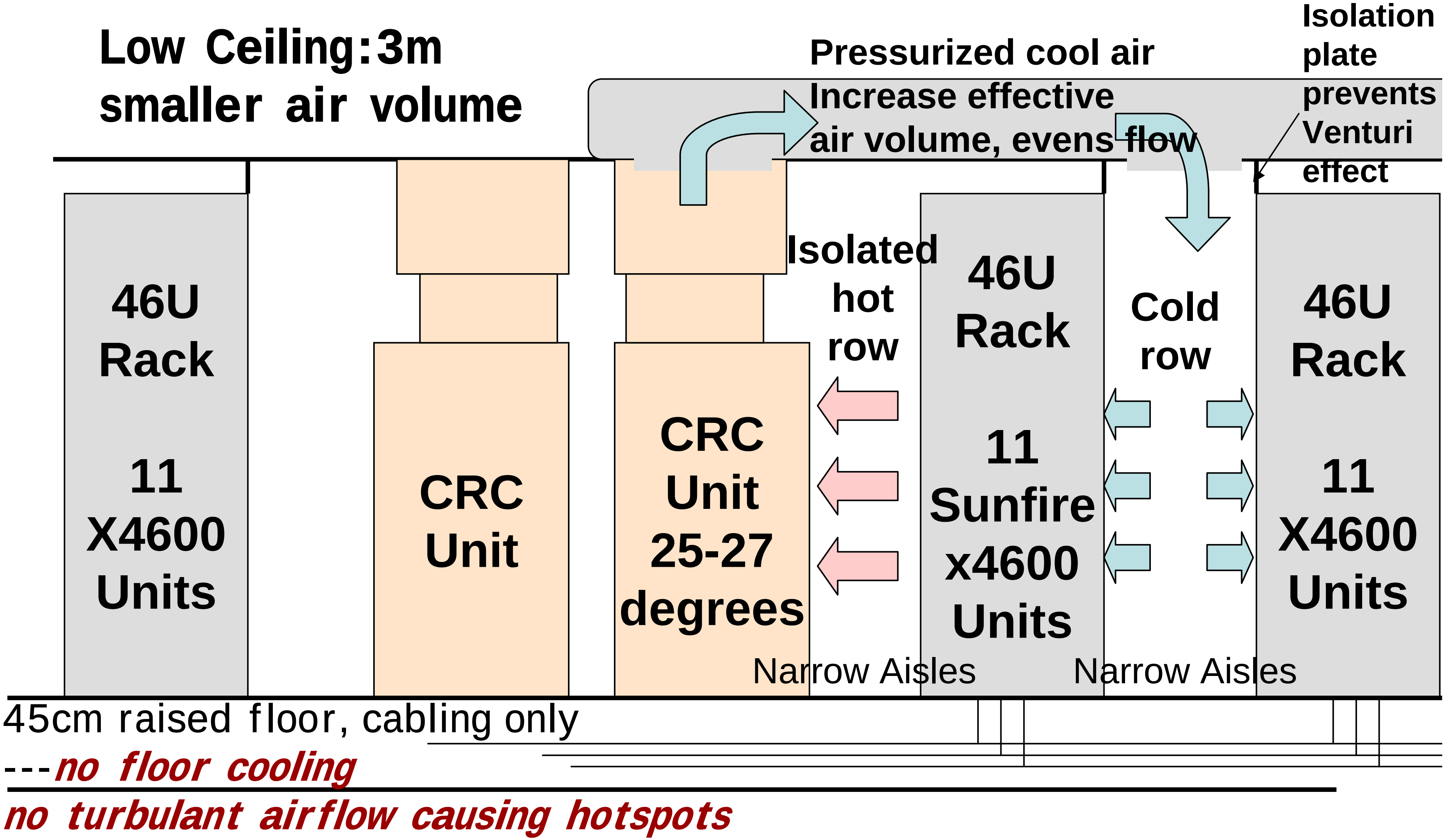
for the whole room

⇩

High density cooling & power reduction

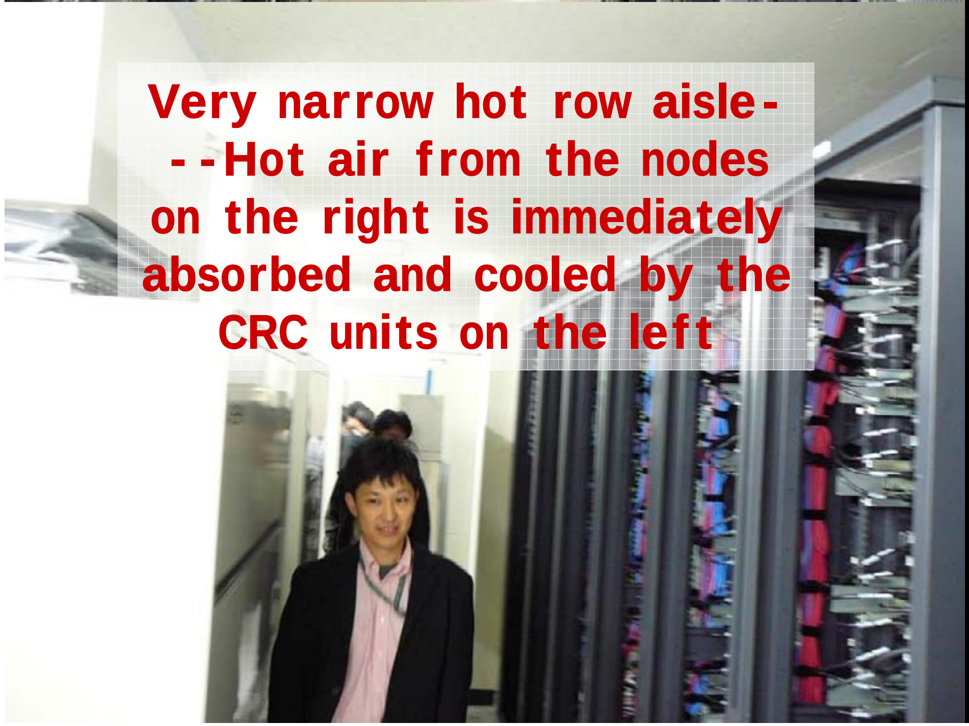
Cooling and Cabling 700W/ft^2

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



Narrow Cold Row Aisle---
no floor cooling, just cables
underneath

Duct openings on the ceiling,
and the transparent
isolation plates to prevent
hot-cold mixture



Very narrow hot row aisle-
--Hot air from the nodes
on the right is immediately
absorbed and cooled by the
CRC units on the left

Pressurized cold air
blowing down from
the ceiling duct ---
very strong wind

TSUBAME as No.1 in Japan circa 2006

TITECH GRID

東京工業大学
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

>>

All University National Centers

>85 TeraFlops
1.1Petabyte
4 year procurement cycle

Has beaten the Earth Simulator in both peak and Top500

Has beaten all the other Univ. centers combined

Total 45 TeraFlops, 350 Terabytes (circa 2006)

みんなのスパコン

"Everybody's Supercomputer"

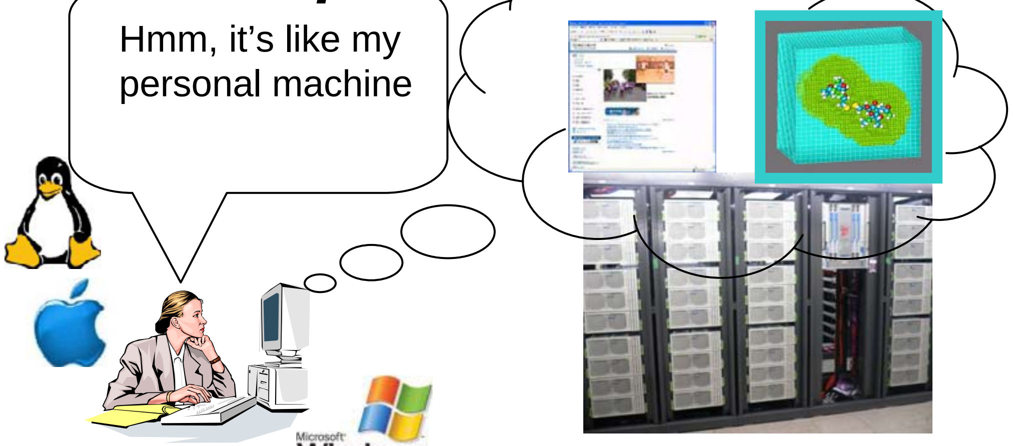
Isolated
High-End

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

Service Oriented Idealism of Grid: Seamless integration of supercomputer resource with *end-user and enterprise* environment



"Everybody's Supercomputer"

Seamless,Ubiquitous access and usage

=>Breakthrough Science through Commoditization of Supercomputing and Grid Technologies

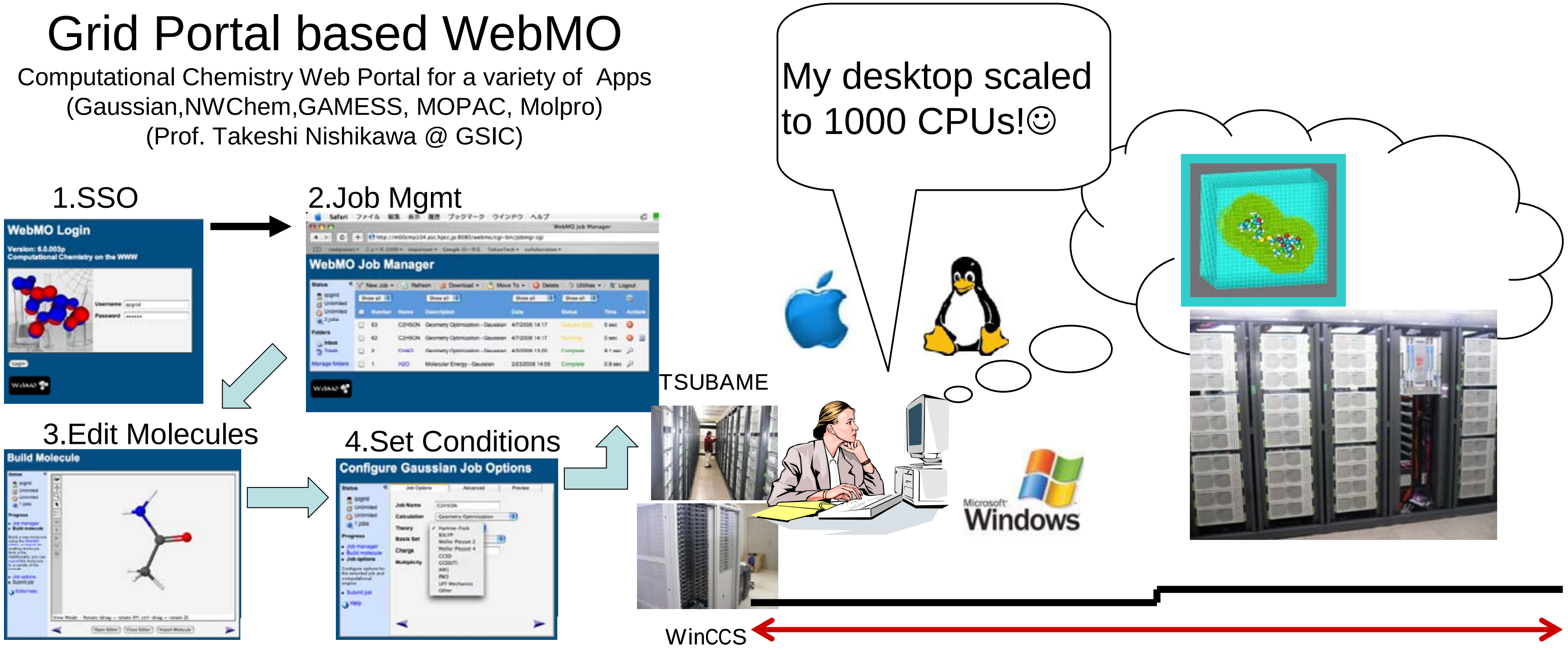
みんなのスパコン

HPC Services in Educational Activities to 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 TSUBAME
  - Research Repository
- CAI, On-line Courses
  (OCW = Open CourseWare)
- Administrative Hosting (VEST)

東京工業大学
Tokyo Institute of Technology
学生証

学籍番号 012345678
所属 工学部
2006年 4月 1日 入学
氏名 大岡 太郎
1985年 5月 1日 生
有効期限 2007年 3月31日

上記の者は本学の学生であることを証明する
東京工業大学長
東京都目黒区大岡山2-12-1
2006年 4月 1日 発行

Tsubame Status

How it’s flying about…
(And doing some research too)

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 Scheduling and Accounting --- Synonimity w/ Existing Social Infrastructures

- Three account/queue types (VO-based) (REALY MONEY!)
  - Small FREE Usage: ***"Promotion Trial (Catch-and-bait)"***
  - Service Level Agreement: ***"Cell Phones"***
    - Exclusivity and other high QoS guarantees
  - Best Effort (new): ***"Internet ISP"***
    - Flat allocation fee per each "UNIT"
- Investment Model for allocation (e.g. ***"Stocks&Bonds"***)
  - Open & extensive information, fair policy guarantee
  - Users make their own investment decisions---collective societal optimization (Adam Smith)

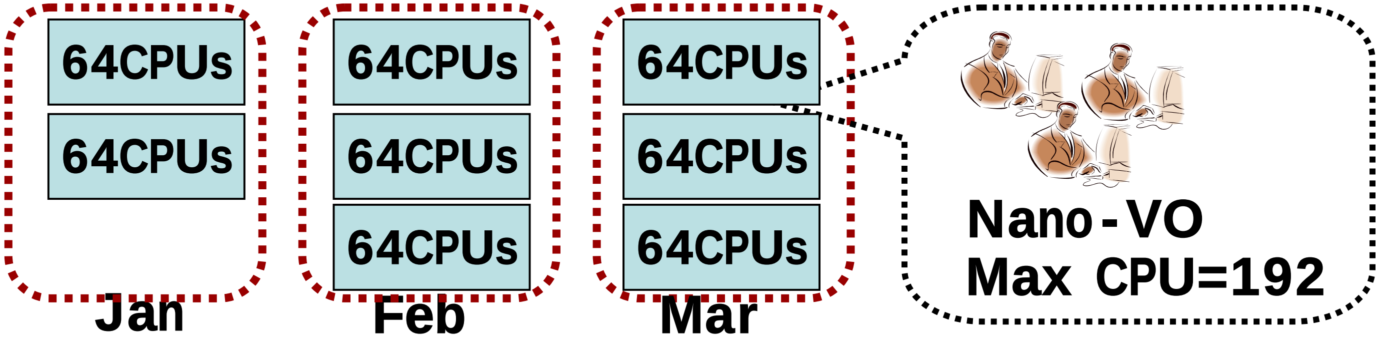
C.f. Top-Down planned allocation (planned economy)

10,000 accounts

Over 1300 SC users

Dynamic machine-level resource allocation

SLA > BES > Small



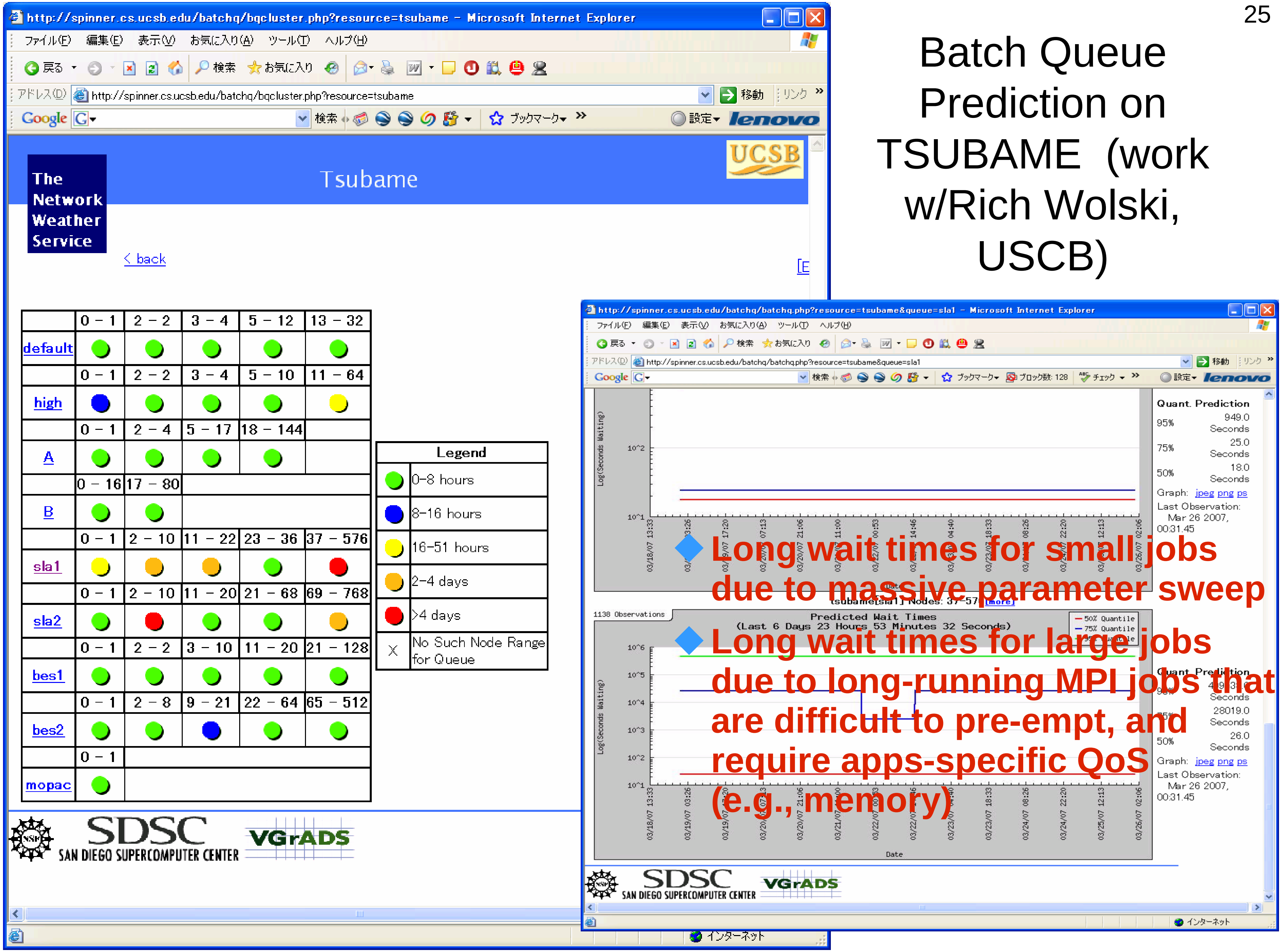
64CPUs
64CPUs
Jan

64CPUs
64CPUs
64CPUs
Feb

64CPUs
64CPUs
64CPUs
Mar

Nano-VO
Max CPU=192

Batch Queue Prediction on TSUBAME (work w/Rich Wolski, USCB)



	0 – 1	2 – 2	3 – 4	5 – 12	13 – 32
default	🟢	🟢	🟢	🟢	🟢
	0 – 1	2 – 2	3 – 4	5 – 10	11 – 64
high	🔵	🟢	🟢	🟢	🟡
	0 – 1	2 – 4	5 – 17	18 – 144	
A	🟢	🟢	🟢	🟢	
	0 – 16	17 – 80			
B	🟢	🟢			
	0 – 1	2 – 10	11 – 22	23 – 36	37 – 576
sla1	🟡	🟠	🟠	🟢	🔴
	0 – 1	2 – 10	11 – 20	21 – 68	69 – 768
sla2	🟢	🔴	🟢	🟢	🟠
	0 – 1	2 – 2	3 – 10	11 – 20	21 – 128
bes1	🟢	🟢	🟢	🟢	🟢
	0 – 1	2 – 8	9 – 21	22 – 64	65 – 512
bes2	🟢	🟢	🔵	🟢	🟢
	0 – 1				
mopac	🟢				

Legend	
🟢	0-8 hours
🔵	8-16 hours
🟡	16-51 hours
🟠	2-4 days
🔴	>4 days
X	No Such Node Range for Queue

Quant. Prediction

95%	949.0 Seconds
75%	25.0 Seconds
50%	18.0 Seconds

Graph: jpeg png ps

Last Observation: Mar 26 2007, 00:31.45

tsubame[sla1] Nodes: 37-57 [more]

1138 Observations

Predicted Wait Times
(Last 6 Days 23 Hours 53 Minutes 32 Seconds)

Quant. Prediction

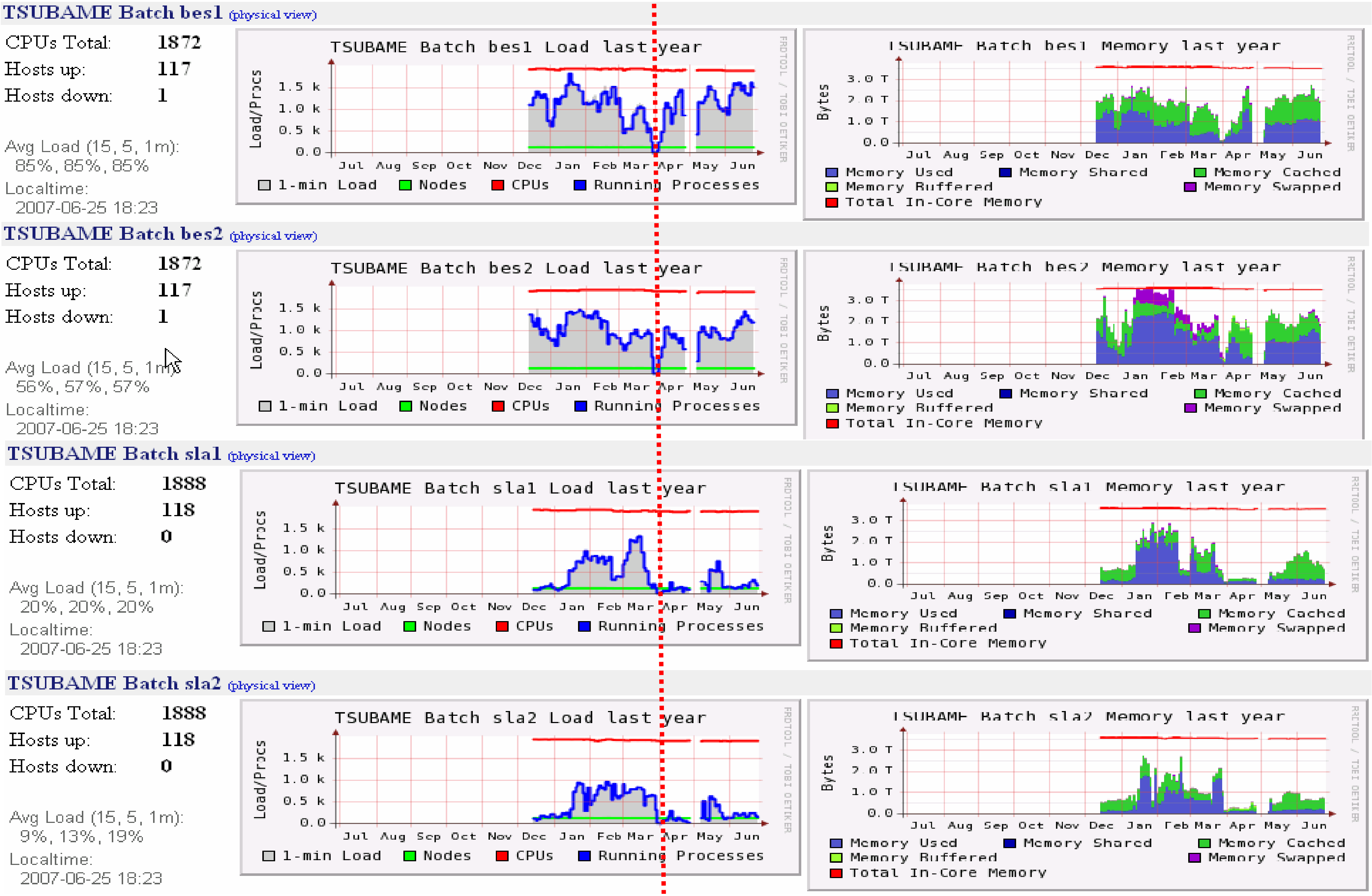
95%	[illegible] Seconds
75%	28019.0 Seconds
50%	26.0 Seconds

Graph: jpeg png ps

Last Observation: Mar 26 2007, 00:31.45

- Long wait times for small jobs due to massive parameter sweep
- Long wait times for large jobs due to long-running MPI jobs that are difficult to pre-empt, and require apps-specific QoS (e.g., memory)

New School Year



TSUBAME Batch bes1 (physical view)

CPUs Total: 1872
Hosts up: 117
Hosts down: 1

Avg Load (15, 5, 1m):
85%, 85%, 85%
Localtime:
2007-06-25 18:23

TSUBAME Batch bes2 (physical view)

CPUs Total: 1872
Hosts up: 117
Hosts down: 1

Avg Load (15, 5, 1m):
56%, 57%, 57%
Localtime:
2007-06-25 18:23

TSUBAME Batch sla1 (physical view)

CPUs Total: 1888
Hosts up: 118
Hosts down: 0

Avg Load (15, 5, 1m):
20%, 20%, 20%
Localtime:
2007-06-25 18:23

TSUBAME Batch sla2 (physical view)

CPUs Total: 1888
Hosts up: 118
Hosts down: 0

Avg Load (15, 5, 1m):
9%, 13%, 19%
Localtime:
2007-06-25 18:23

TSUBAME Batch bes1 Load last year
TSUBAME Batch bes1 Memory last year
TSUBAME Batch bes2 Load last year
TSUBAME Batch bes2 Memory last year
TSUBAME Batch sla1 Load last year
TSUBAME Batch sla1 Memory last year
TSUBAME Batch sla2 Load last year
TSUBAME Batch sla2 Memory last year
Load/Procs
1.5 k
1.0 k
0.5 k
0.0
Bytes
3.0 T
2.0 T
1.0 T
0.0
Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun
□ 1-min Load
🟩 Nodes
🟥 CPUs
🟦 Running Processes
🟦 Memory Used
🟦 Memory Shared
🟩 Memory Cached
🟩 Memory Buffered
🟪 Memory Swapped
🟥 Total In-Core Memory
RRDTOOL / TOBI OETIKER

Tsubame in Magazines
(e.g., Unix Magazine, a 20 page special)

UNIX magazine 2007 Summer 7

Webサービスを支える
アプリケーションサーバのすべて

みんなのスパコンを掲げる
日本最速の
スーパーコンピュータ
「TSUBAME」の技術

●Debian GNU/Linux 4.0
●Ubuntu Linux
●iptables再入門

「みんなのスパコン」を掲げる
日本最速のスーパーコンピュータ

「TSUBAME」の技術

「計算機」という訳語が示すように、コンピュータの本分は計算であり、その最先端に位置する存在がスーパーコンピュータである。スーパーコンピュータは世界各国で開発競争が続けらており、その性能は年々向上し続けている。現在、日本国内で最速のスーパーコンピュータは、東京工業大学の「TSUBAME（つばめ）」である。本稿ではTSUBAMEに投入されているさまざまな技術や工夫を探り、併せてTSUBAMEの果たすべき役割を東京工業大学の松岡教授に伺った。

p96 TSUBAMEへ至る
スーパーコンピュータの歴史

p98 ～インタビュー～ 松岡教授が語る
TSUBAMEが担う役割

p102 TSUBAMEのシステム構成
佐藤通敏

UNIX magazine 2007 Summer 95

For Details…

- A ~70 Page Document that describes the policy, the implementation, and every other little detail…
(by M. Hamakawa @Sun Services, Japan)

Sun microsystems

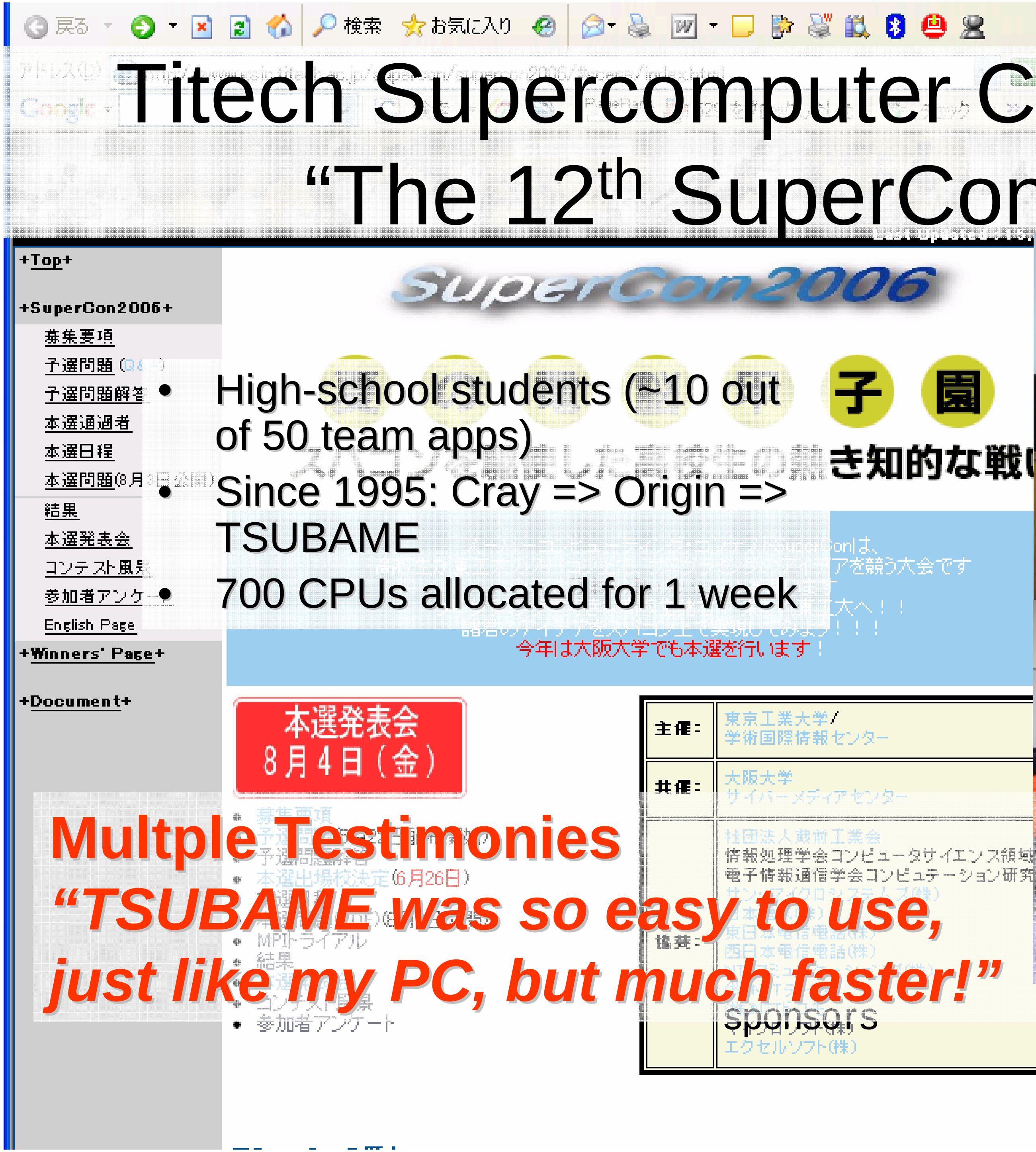
SUN N1™ GRID ENGINE SOFTWARE AND THE
TOKYO INSTITUTE OF TECHNOLOGY SUPERCOMPUTER GRID

Minoru Hamakawa, Sun Services, Japan

Sun BluePrints™ On-Line — June 2007

Part No 820-1695-10
Revision 1.0, 5/23/07
Edition: June 2007

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

sponsors

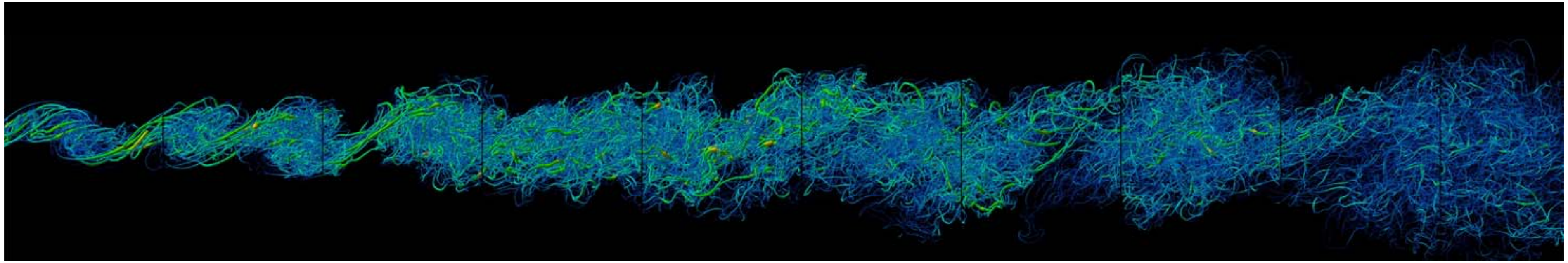
Multple Testimonies

“TSUBAME was so easy to use, just like my PC, but much faster!”

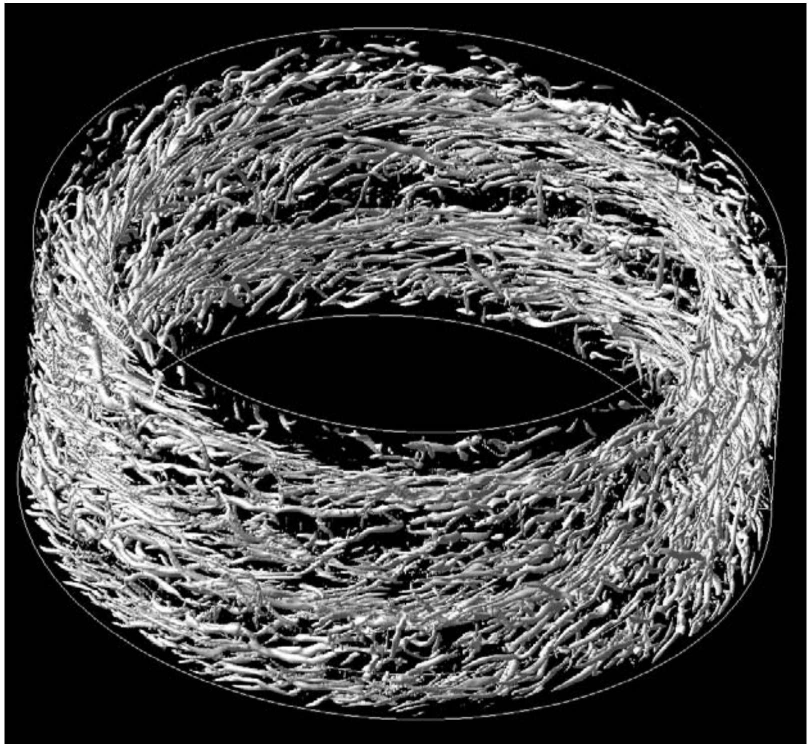
TSUBAME Application Profile

- Large scale codes, e.g. port from the Earth Simulator
  - Simple porting is easy
  - Tuned Vector code into cache-friendly “normal code” takes more time.
- Large-Scale (>1,000~10,000 instances) Parameter Survey, Ensemble, Optimization, …
- Lots of ISV Code---Gaussian, Amber, …
- Storage-Intensive Codes --- Visualization
- => Often Limited by Memory, not CPUs
- Must Give users both EASE and COMPELLING REASON to use TSUBAME

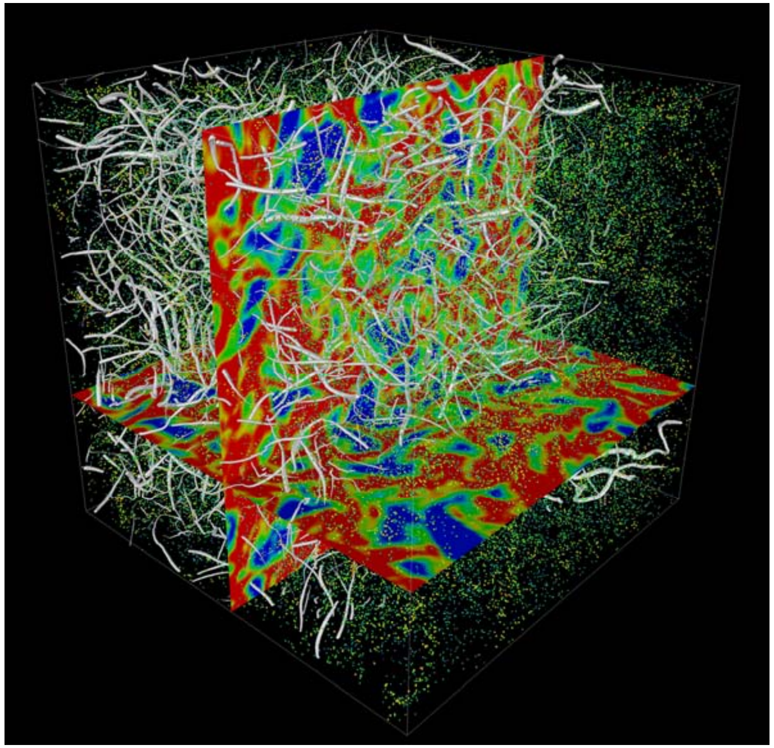
TSUBAME Applications---Massively Complex Turbulant Flow and its Visualization (by Tanahashi Lab and Aoki Lab, Tokyo Tech.)



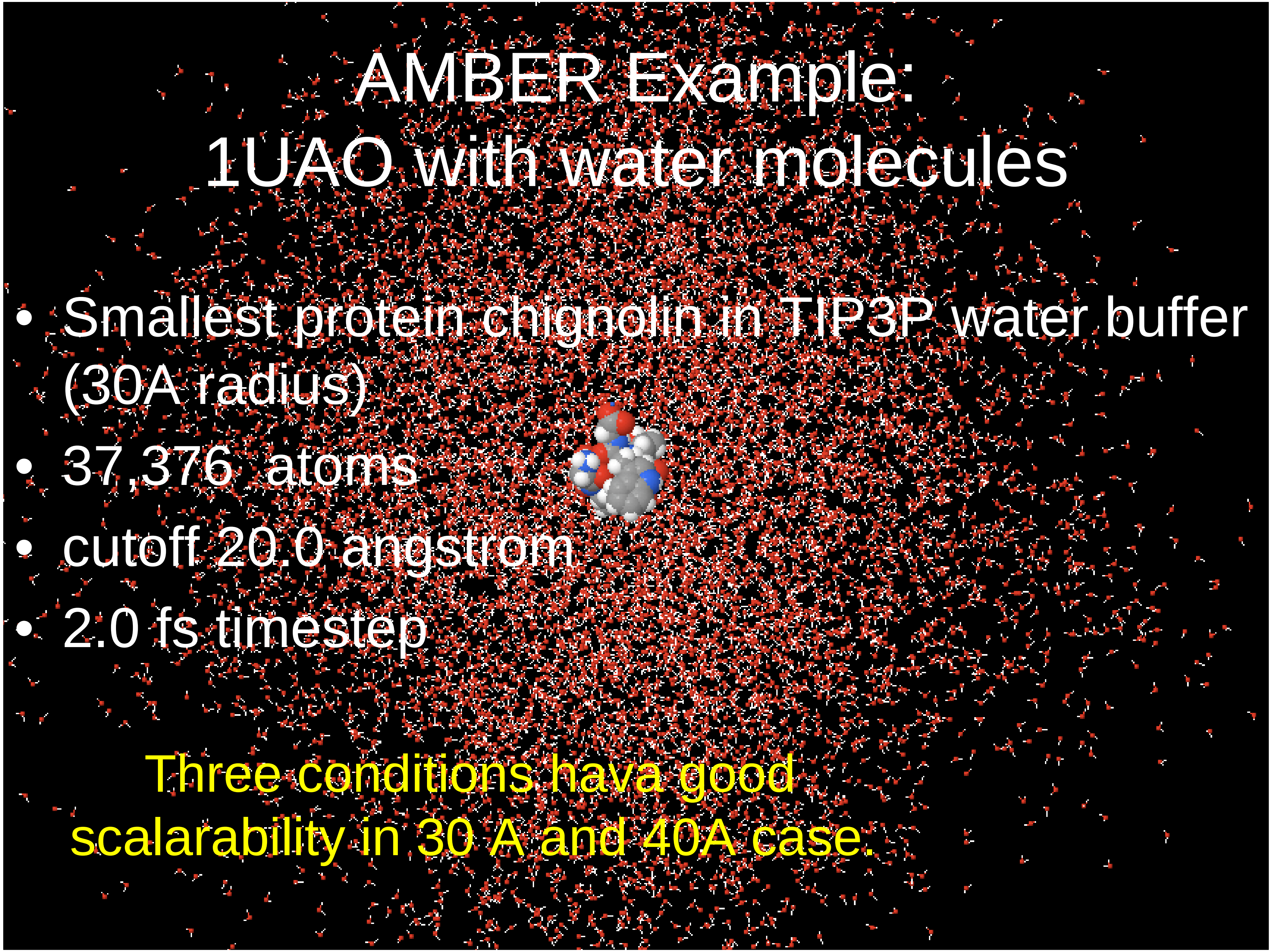
Turbulant Flow from Airplane



Taylor-Couette Flow



AMBER Example: 1UAO with water molecules



- Smallest protein chignolin in TIP3P water buffer (30A radius)
- 37,376 atoms
- cutoff 20.0 angstrom
- 2.0 fs timestep

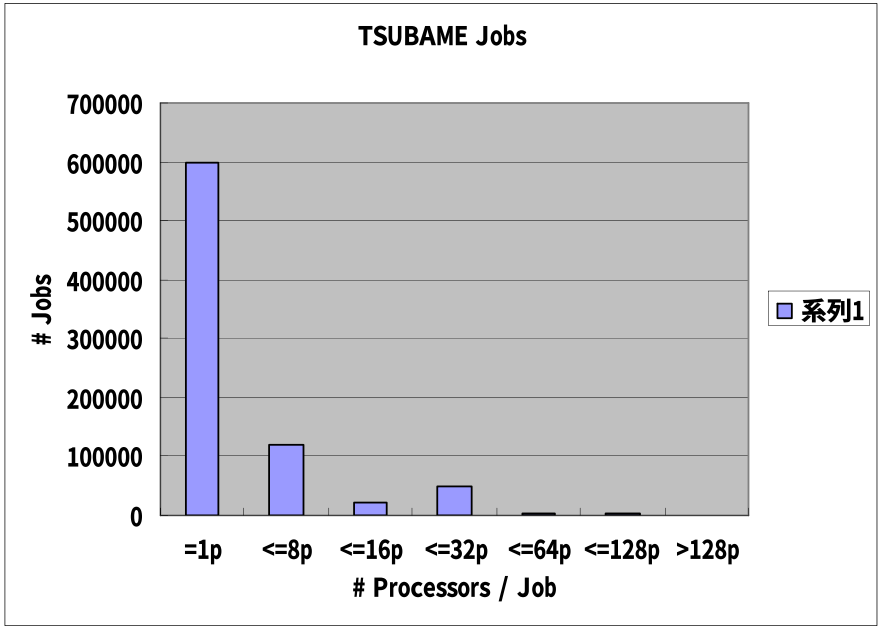
Three conditions hava good scalarability in 30 A and 40A case.

TSUBAME Job Statistics Dec. 2006-Aug.2007 (#Jobs)

- 797,886 Jobs (~3270 daily)
- 597,438 serial jobs (74.8%)
- 121,108 <=8p jobs (15.2%)

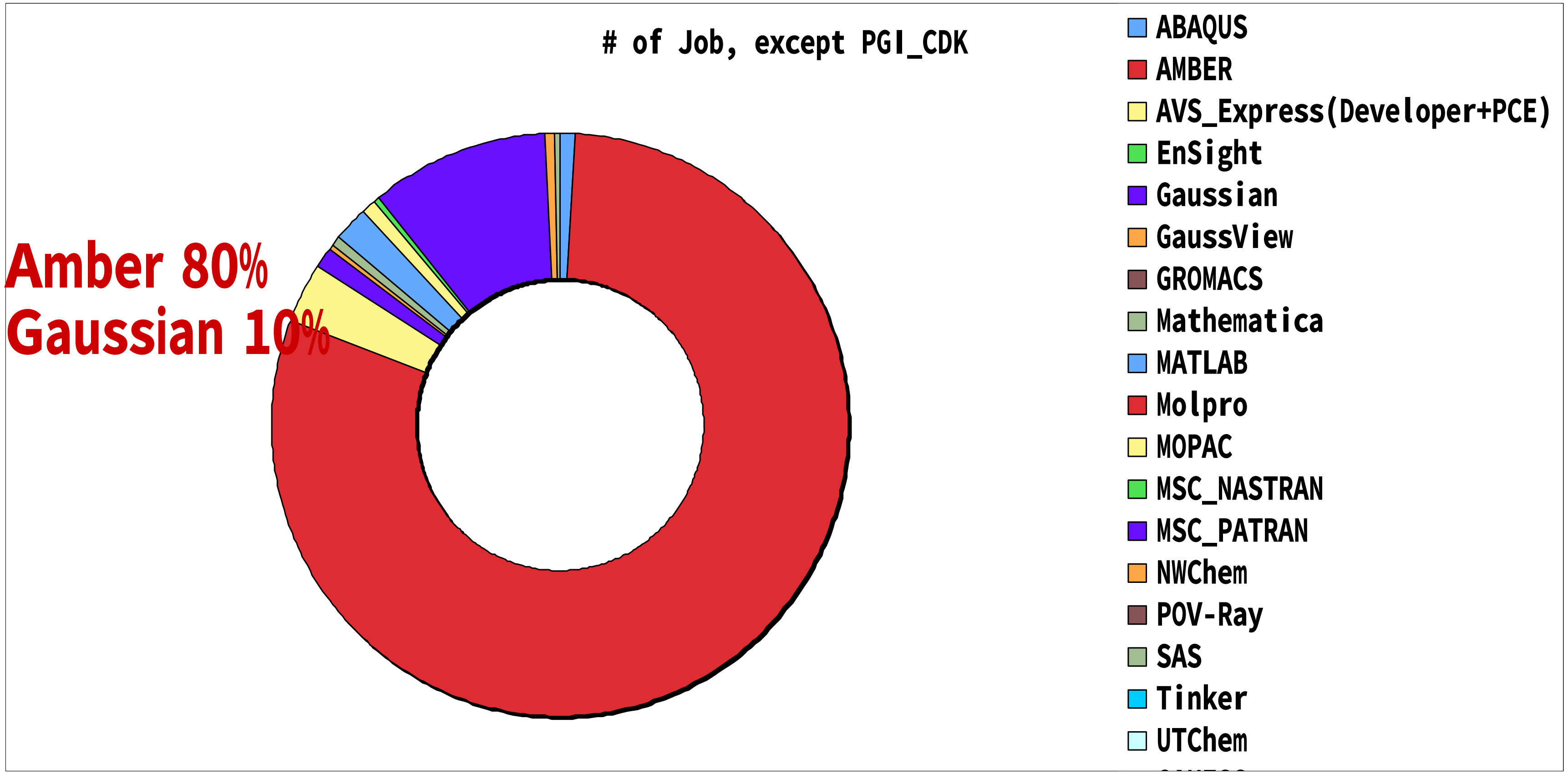
90%

- 129,398 ISV Application Jobs (16.2%)
- *However, >32p jobs account for 2/3 of cumulative CPU usage*



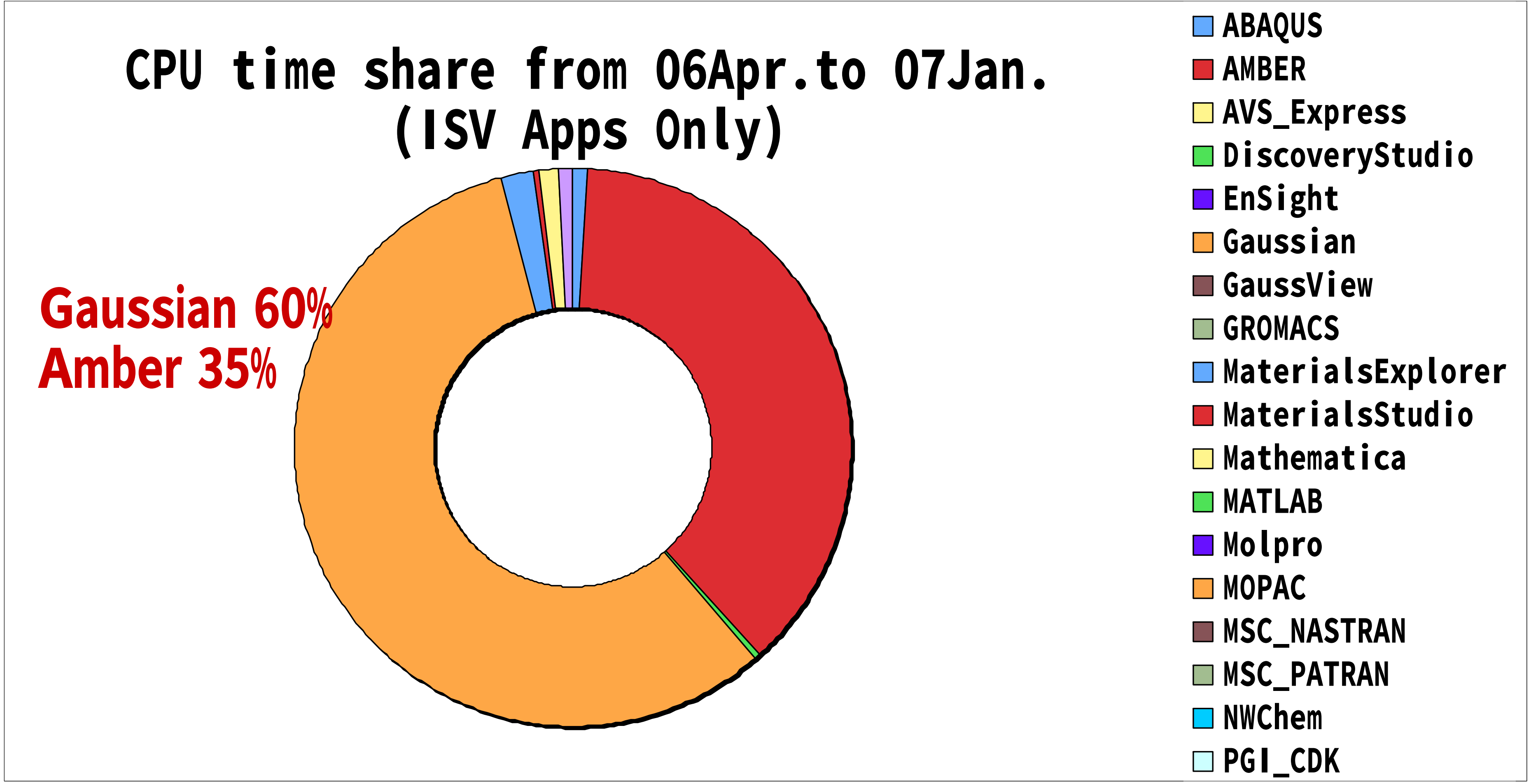
Coexistence of ease-of-use in both
- *short duration* parameter survey
- large scale MPI
(Both are hard for *physically* large-scale distributed grid)

TSUBAME Job Statistics for ISV Apps (# Processes)



Amber 80%
Gaussian 10%

Reprisal: TSUBAME Job Statistics for ISV Apps (# CPU Timeshare)



Gaussian 60%
Amber 35%

Multi-User and Ensemble! (60,000-way Gaussian ensemble job recorded on TSUBAME) => Throughput(!)

TSUBAME Draws Research Grants

- “Computationism” Global Center-of-Excellence (Global COE) Program
  - Incubating Math/Computer Science/HPC Experts
  - $2~2.5 mil x 5 years
- “Center of (Industrial) Innovation Program”
  - Industrial Collaboration w/ High-End Facilities
  - ~$1 mil x 5 years
- More Coming…

Compuationism Approach to Science

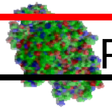
Non-traditional computational modeling

⇨ Apply non-traditional mathematical approaches

⇨ Making the Impossible (Infeasible) Possible

Example) Proteomic Interactions

1000x1000 mutual interactions of proteins



	P1	P2	P3	P4	P5	....	P1000
P1		☆					
P2	☆						
P3					☆		
P4							
P5							
…							
P1000							

Complex & Large Scale

Drug Design

Narrowing the Candidate

Complexity：1000 ⇒ 1000 x 1000

Infeasible with traditional ab-initio approaches

100s of years on a Petascale supercomputer

Structural Matching [Y. Akiyama]

⇒ Non-traditional modeling and approach

Possible in a few months

Educating “Computatism Experts”

Incubating Computing Generalists

Target Profile

Theory of Computing & Applied Math

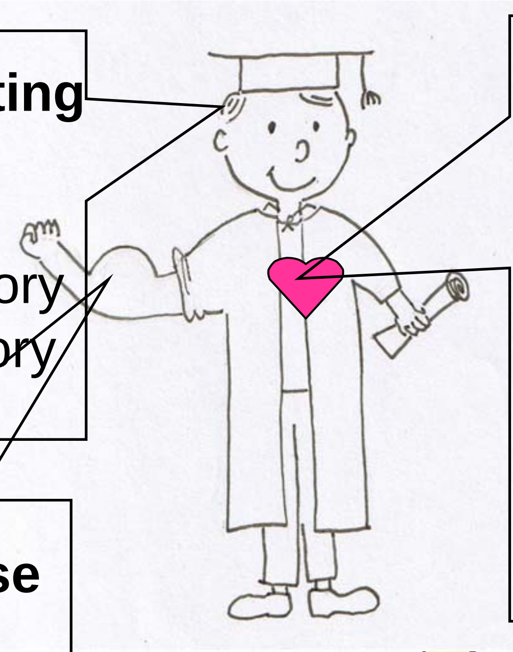
- Algorithms
- Optimization Theory
- Probabilistic Theory

…

HPC & CS Expertise

- Modeling
- Programming
- Systems

…



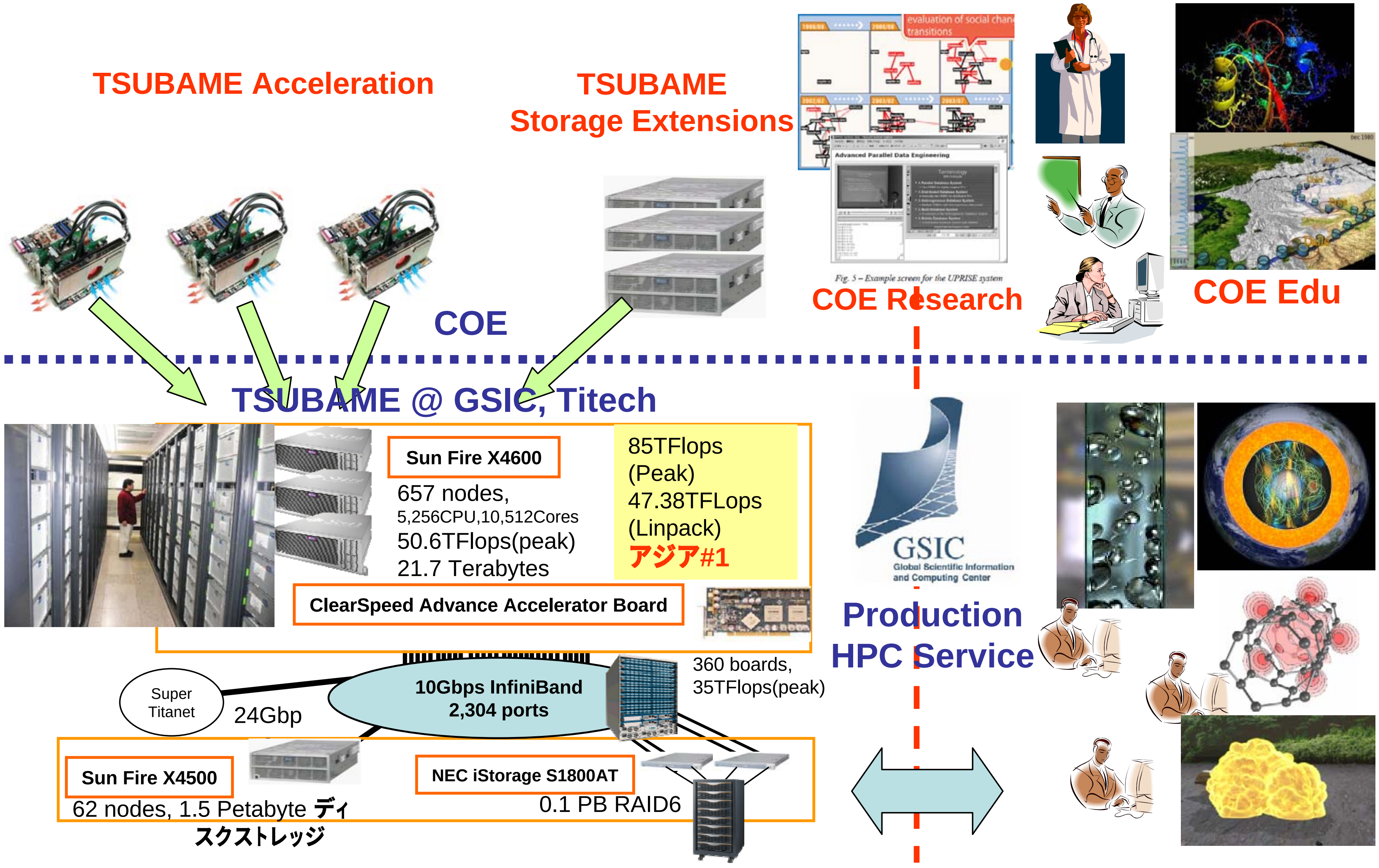
Computationism Ideology

- **Work with domain scientists**
- **Willing to** Study and understand the Science and the discipline

Collaborate

Domain Scientist Counterpart

Building the COE on TSUBAME



TSUBAME Acceleration

TSUBAME Storage Extensions

evaluation of social change transitions
Advanced Parallel Data Engineering

Fig. 5 – Example screen for the UPRISE system

COE Research

COE Edu

COE

TSUBAME @ GSIC, Titech

Sun Fire X4600

657 nodes,
5,256CPU,10,512Cores
50.6TFlops(peak)
21.7 Terabytes

85TFlops (Peak)
47.38TFLops (Linpack)
アジア#1

ClearSpeed Advance Accelerator Board

360 boards,
35TFlops(peak)

GSIC
Global Scientific Information and Computing Center

Production HPC Service

Super Titanet

24Gbp

10Gbps InfiniBand
2,304 ports

Sun Fire X4500

62 nodes, 1.5 Petabyte ディスクストレッジ

NEC iStorage S1800AT

0.1 PB RAID6

Ministry of Edu. “Center of Innovation Program”

Industrial Collaboration w/ High-End Facilities
Provide industrial access to TSUBAME (via Grid)

- (x86) PC&WS Apps in industry *directly* execute at x10~x100 scale

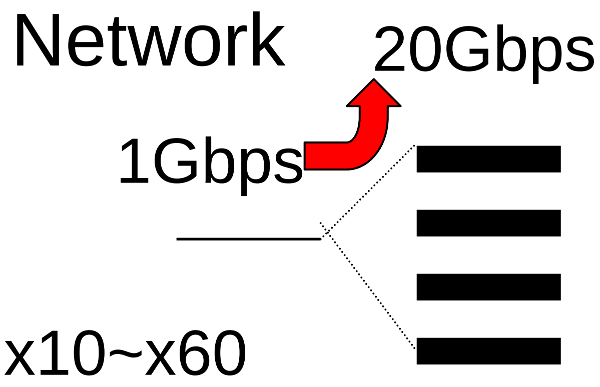
  Not just CPu power but memory/storage/network, etc.

- HPC-Enabling non-traditional industries ---ICT, Financials, Security, Retail, Services, …)
- **E.g. Ultra Large-scale portfolio risk analysis by a Megabank (ongoing)**

Why Industries are interested in TSUBAME?

- Standard Corporate x86 Cluster Env. vs. TSUBAME -

	CPU Core	Network	RAM	Disk(Cap, BW)
Std.	2~4 (node) 32~128 (job)	1Gbps 32Gbps	2~8GB 128GB	500GB, 50MB/s 10TB(NAS), 100MB/s
TSUBAME	16 (node) 1920 (job)	20Gbps 2.5Tbps	32~128GB 3840GB	120TB, 1GB/s 120TB, 3GB/s

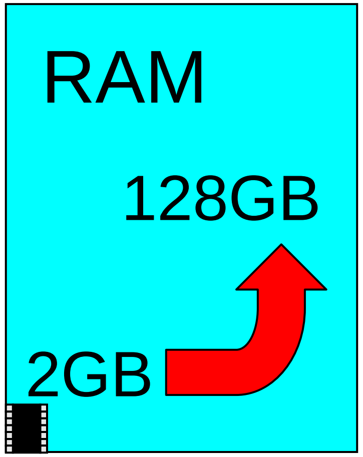


Network

20Gbps

1Gbps

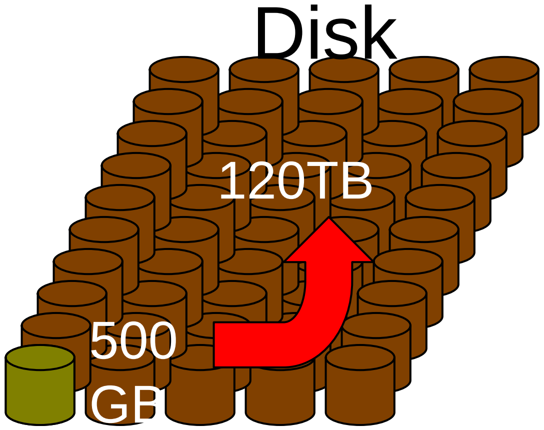
x10~x60



RAM

128GB

2GB



Disk

120TB

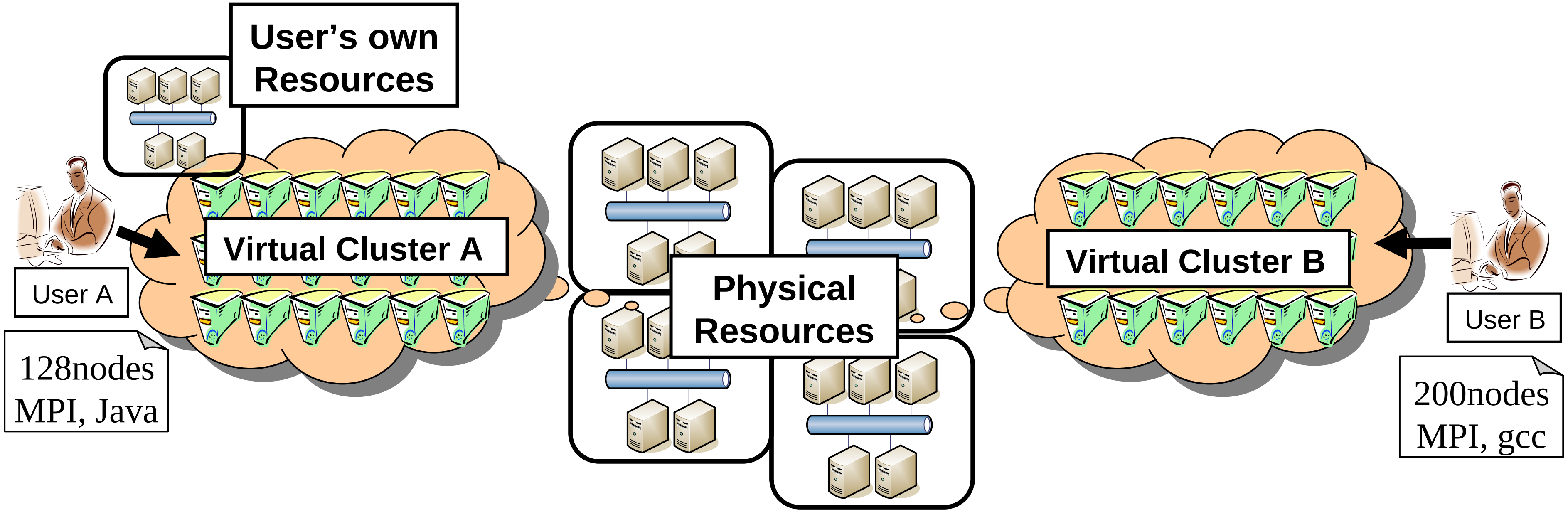
500 GB

The Industry Usage is Real(!!!) and will be Stellar (!!!)

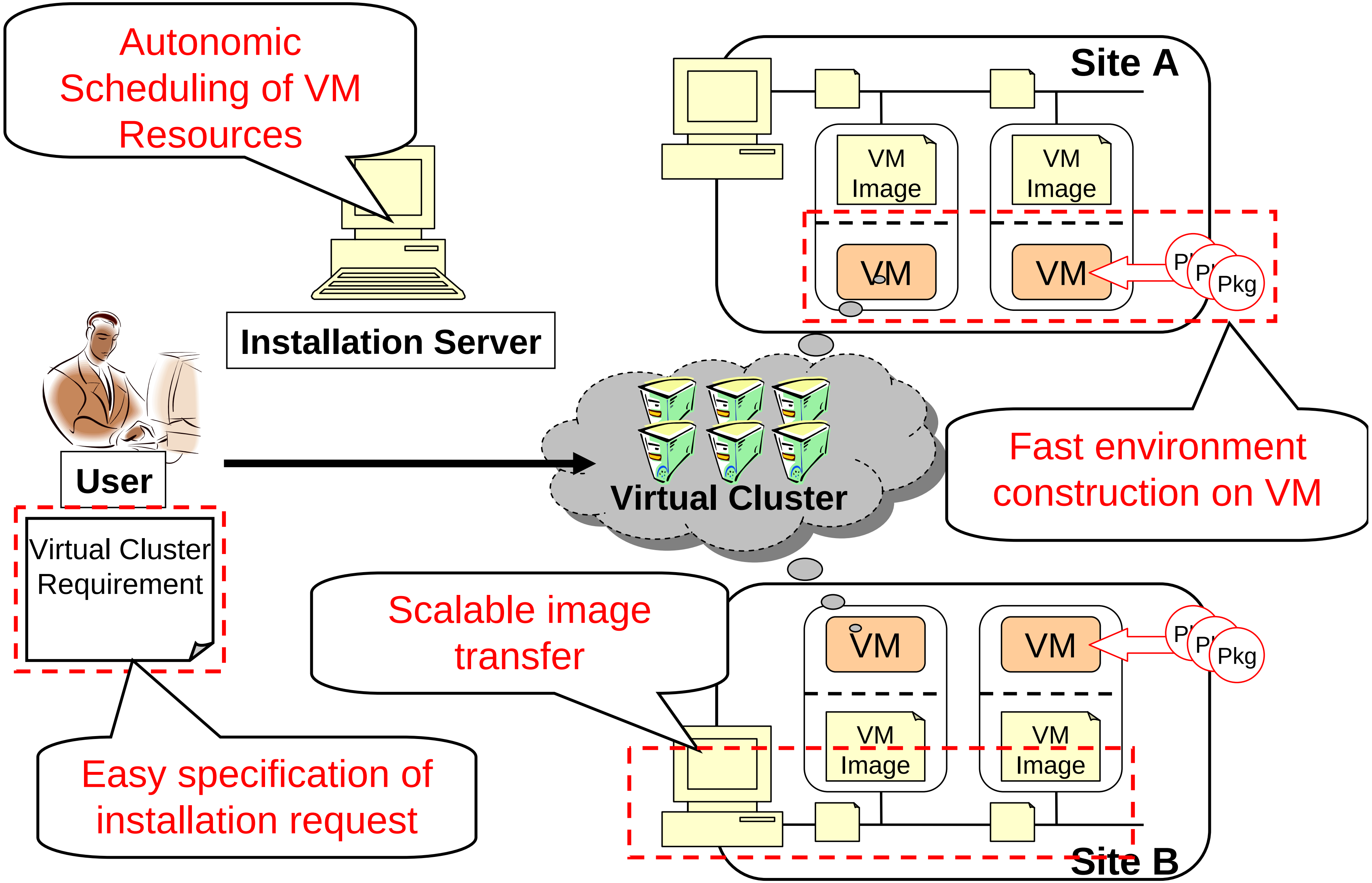
- Two calls since July: 8 real industry apps for TSUBAME (and 18 others for Nat’l Univ. Centers coalition)
- Example: a Japanese Megabank has run a real financial analysis app. on 1/3 of TSUBAME, and is EXTREMELY happy with the stellar results.
  - Only runnable with >20GB mem, IB-based I/O
  - Stay tuned for follow-on announcements…
- Big booster for non-dedicated commercial usage
  - The overall grid must be as such

Research: Grid Resource Sharing with Virtual Clusters ([CCGrid2007] etc.)

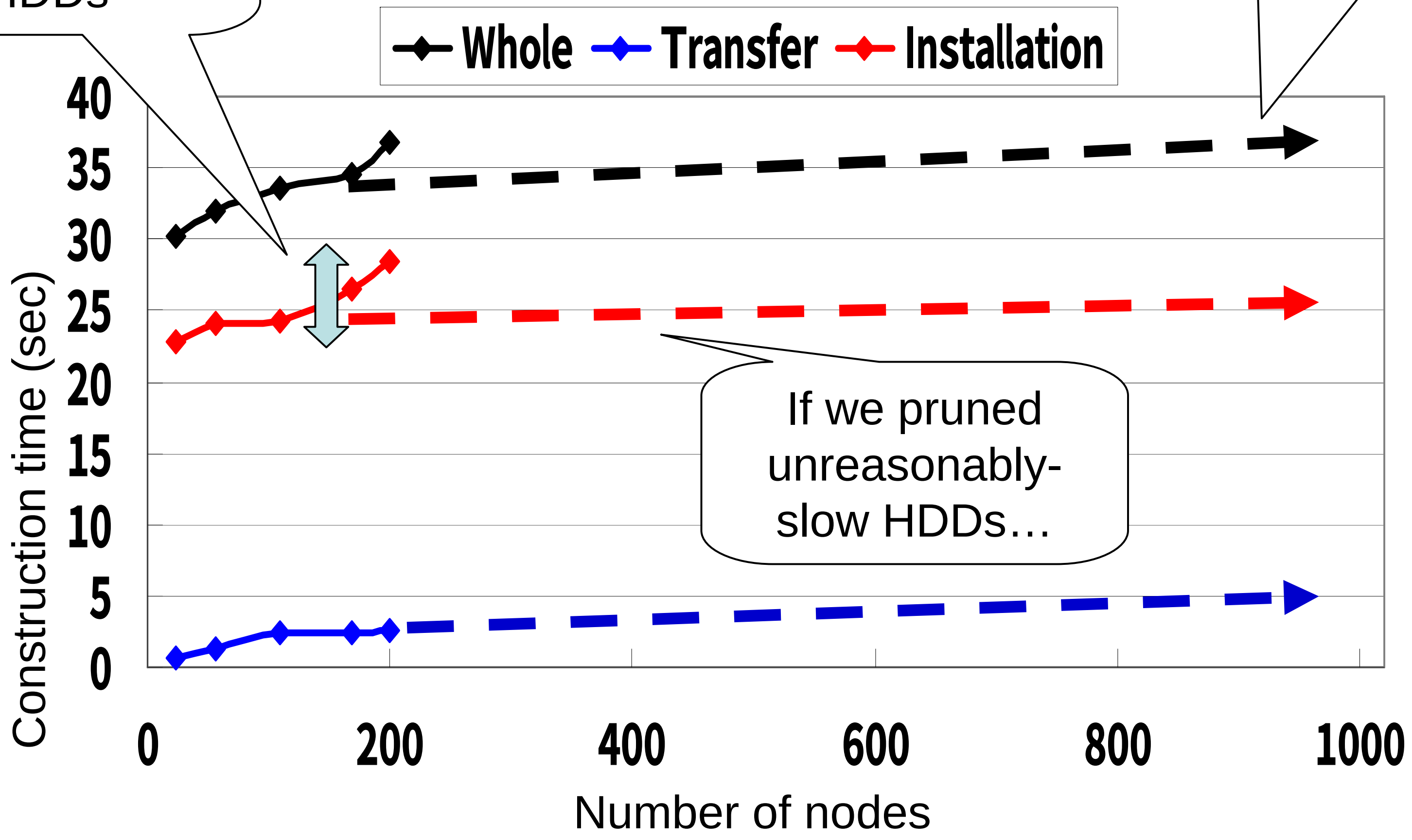
- Virtual Cluster
  - Virtual Machines (VM) as computing nodes
    - Per-user customization of exec environment
    - Hides software heterogeneity
    - Seamless integration with user's own resources
  - Interconnected via overlay networks
    - Hides network asymmetry
    - Overcomes private networks and firewalls



Our VPC Installer Architecture


Autonomic Scheduling of VM Resources
Installation Server
User
Virtual Cluster Requirement
Easy specification of installation request
Virtual Cluster
Scalable image transfer
Fast environment construction on VM
Site A
VM Image
VM Image
VM
VM
Pkg
Site B
VM
VM
Pkg
VM Image
VM Image

Scalability w/# of VPC nodes: Optimistic Extrapolation to 1000 VMs



Likely to be due to some unstable HDDs

1000-VM virtual cluster in less than 1 minute!

If we pruned unreasonably-slow HDDs…

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

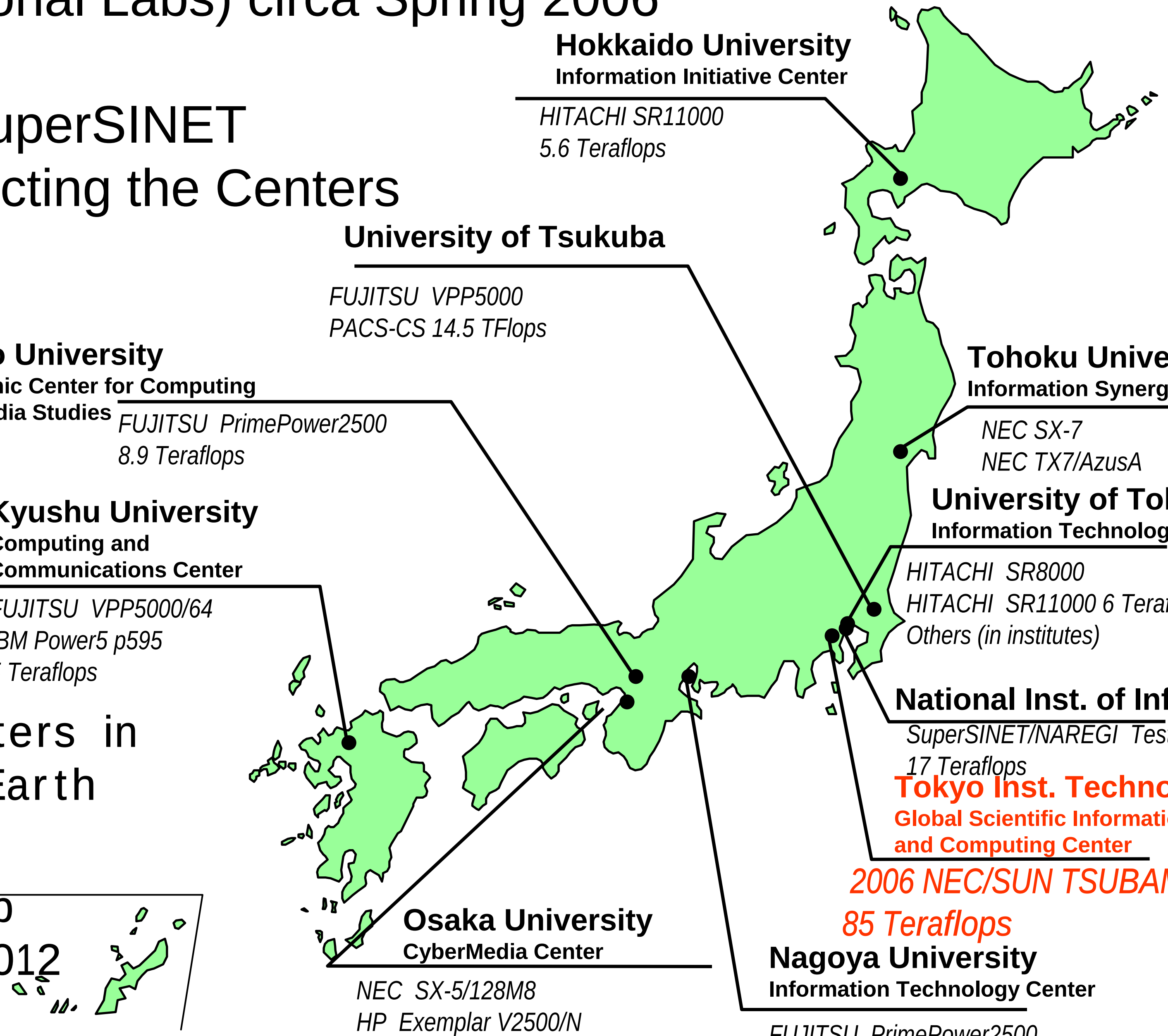
- Sep. 6th, 2006---U-Tokyo, Kyoto-U, and U-Tsukuba announced “common procurement procedure” for the next gen SCs in 1H2008
  - 100-150 TFlops
  - HW: x86 cluster-like SC architecture
  - NW: Myrinet10G or IB + Ethernet
  - SW: Linux+SCore, common Grid MW
- Previously, ALL centers ONLY had dedicated SCs
- 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



Hokkaido University
Information Initiative Center
HITACHI SR11000
5.6 Teraflops

University of Tsukuba
FUJITSU VPP5000
PACS-CS 14.5 TFlops

Kyoto University
Academic Center for Computing and Media Studies
FUJITSU PrimePower2500
8.9 Teraflops

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

Kyushu University
Computing and Communications Center
FUJITSU VPP5000/64
IBM Power5 p595
5 Teraflops

University of Tokyo
Information Technology Center
HITACHI SR8000
HITACHI SR11000 6 Teraflops
Others (in institutes)

National Inst. of Informatics
SuperSINET/NAREGI Testbed
17 Teraflops

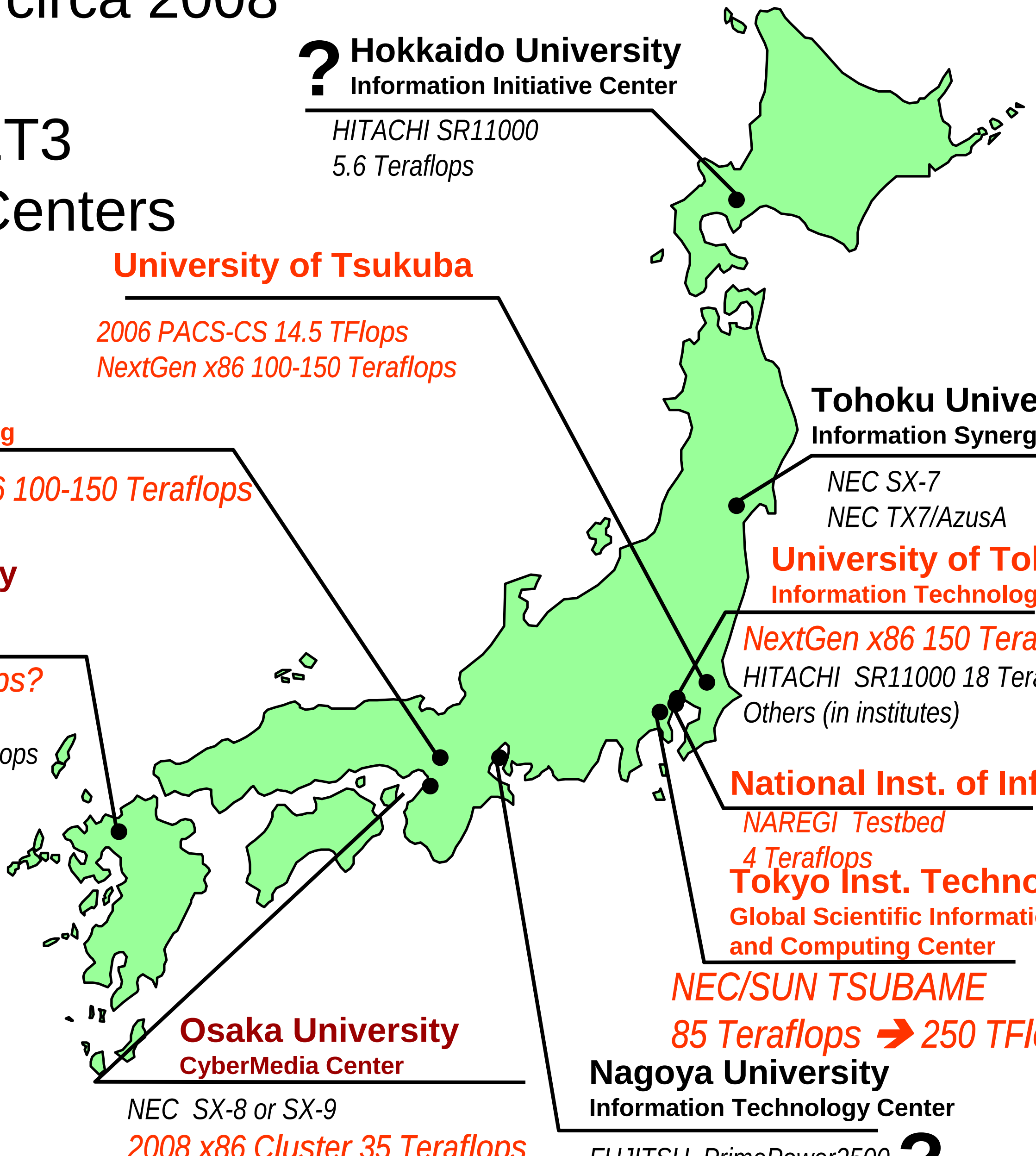
Tokyo Inst. Technology
Global Scientific Information and Computing Center
2006 NEC/SUN TSUBAME
85 Teraflops

Osaka University
CyberMedia Center
NEC SX-5/128M8
HP Exemplar V2500/N
1.2 Teraflops

Nagoya University
Information Technology Center
FUJITSU PrimePower2500
11 Teraflops

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

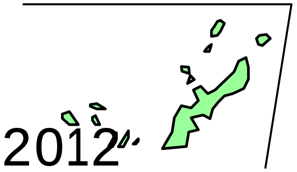
Tokyo Inst. Technology
Global Scientific Information and Computing Center
NEC/SUN TSUBAME
85 Teraflops ➔ 250 TFlops?

x86 TSUBAME sibling domination

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

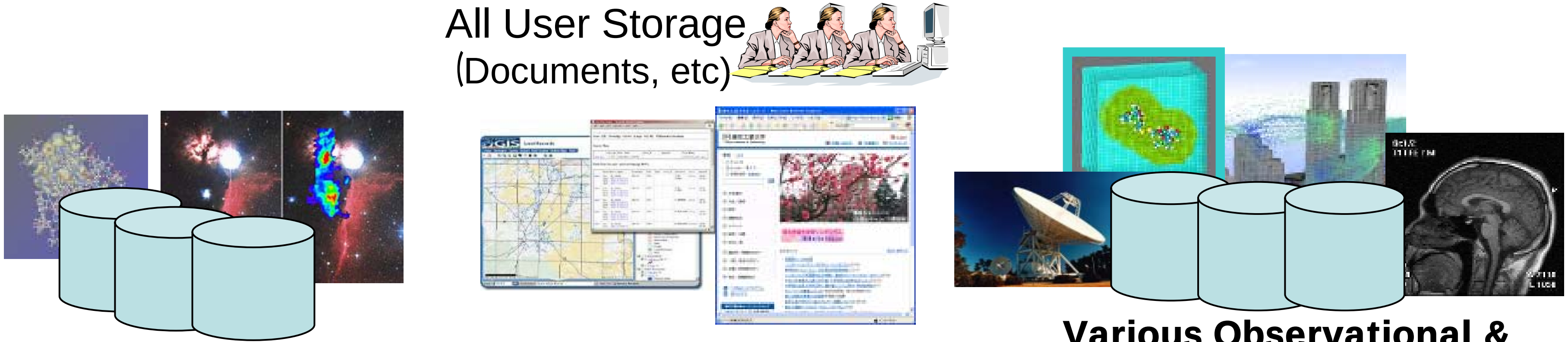
Nagoya University
Information Technology Center
FUJITSU PrimePower2500 ?
11 Teraflops

Still - 10 Petaflop center by 2012



TSUBAME Upgrades

Towards Multi-Petabyte Data Grid Infrastructure based on TSUBAME



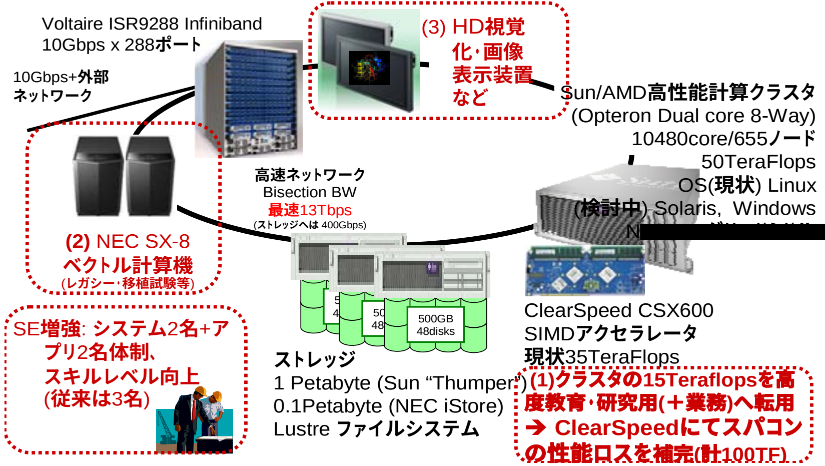
All User Storage
(Documents, etc)

Various public research DBs and Mirrors---Astro, Bio, Chemical

All Historical Archive of Research Publications, Documents, Home Pages,

Various Observational & Simulation Data

Archival & Data Grid Middleware



Voltaire ISR9288 Infiniband
10Gbps x 288ポート

10Gbps+外部
ネットワーク

(3) HD視覚化·画像表示装置など

Sun/AMD高性能計算クラスタ
(Opteron Dual core 8-Way)
10480core/655ノード
50TeraFlops
OS(現状) Linux
(検討中) Solaris, Windows
N

高速ネットワーク
Bisection BW
最速13Tbps
(ストレッジへは 400Gbps)

(2) NEC SX-8
ベクトル計算機
(レガシー·移植試験等)

SE増強: システム2名+アプリ2名体制、
スキルレベル向上
(従来は3名)

ストレッジ
1 Petabyte (Sun "Thumper")
0.1Petabyte (NEC iStore)
Lustre ファイルシステム

ClearSpeed CSX600
SIMDアクセラレータ
現状35TeraFlops

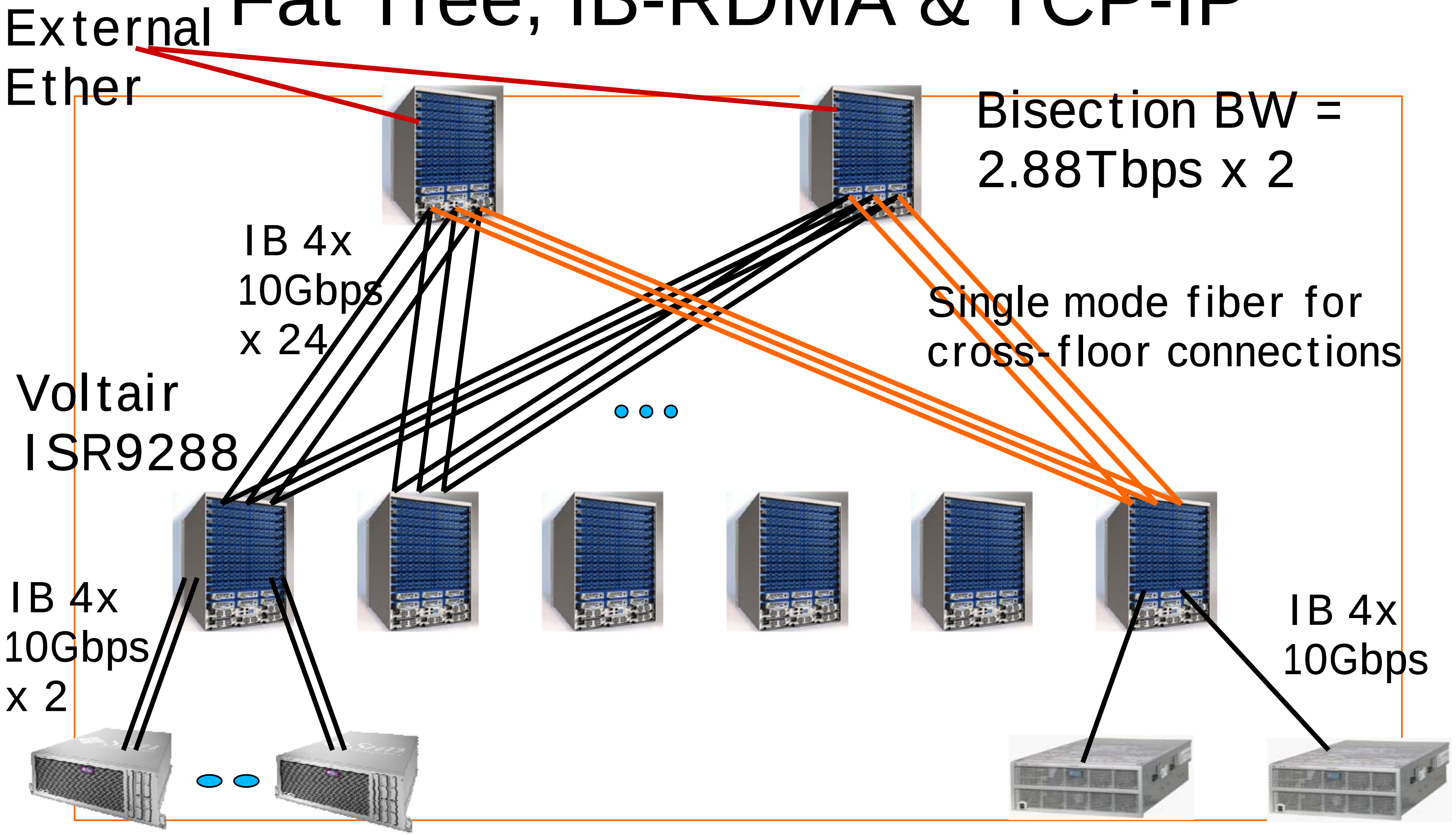
(1)クラスタの15Teraflopsを高度教育·研究用(＋業務)へ転用
➔ ClearSpeedにてスパコンの性能ロスを補完(計100TF)

NESTRE System

Petabytes, Stable Storage
Data Provenance
"Archiving Domain Knowledge"

TSUBAME
～100 TeraFlops, Petabytes Storage

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



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

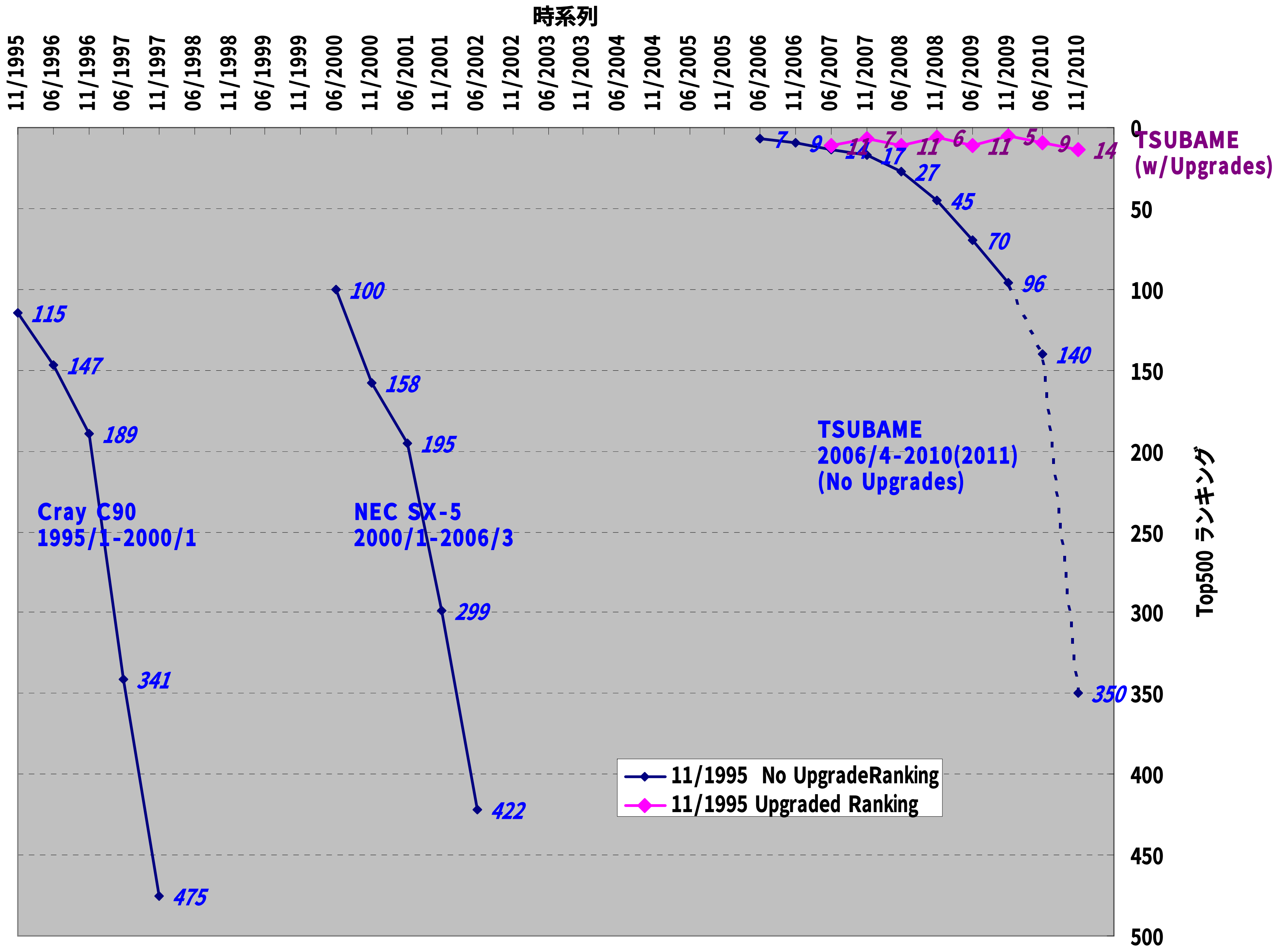
X4500 x 42nodes (42 ports)
=> 42ports 420Gbps

NESTRE (and the old cluster nodes it replaced)

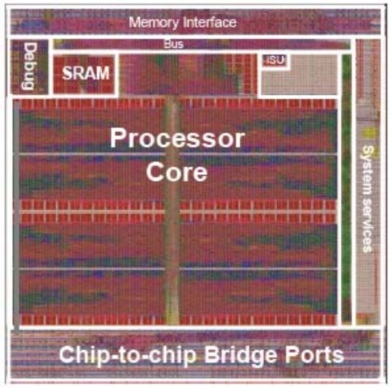
TSUBAME
Linpack and Acceleration

Heterogeneity both Intra- and
Inter- node

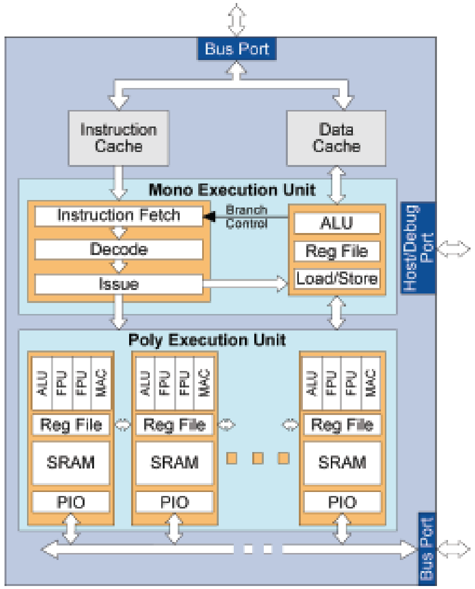
GSIC 過去のスパコンおよびTSUBAME Top500 性能の歴史および予測



ClearSpeed Advance Accelerator Board



ClearSpeed



Hardware

- 25W Max Power
- CSX600 processor x2 (96GFLOPS Peak)
- IEEE 754 64bit Double-Precision Floating Point
- 133MHz PCI-X Host Interface
- On board memory： 1GB (Max 4 GB)
- Internal memory bandwidth ：200 Gbytes/s
- On-board memory bandwidth：6.4Gbytes/s

Software

- Standard Numerical Libraries
- ClearSpeed Software Development Kit (SDK)

Applications and Libraries

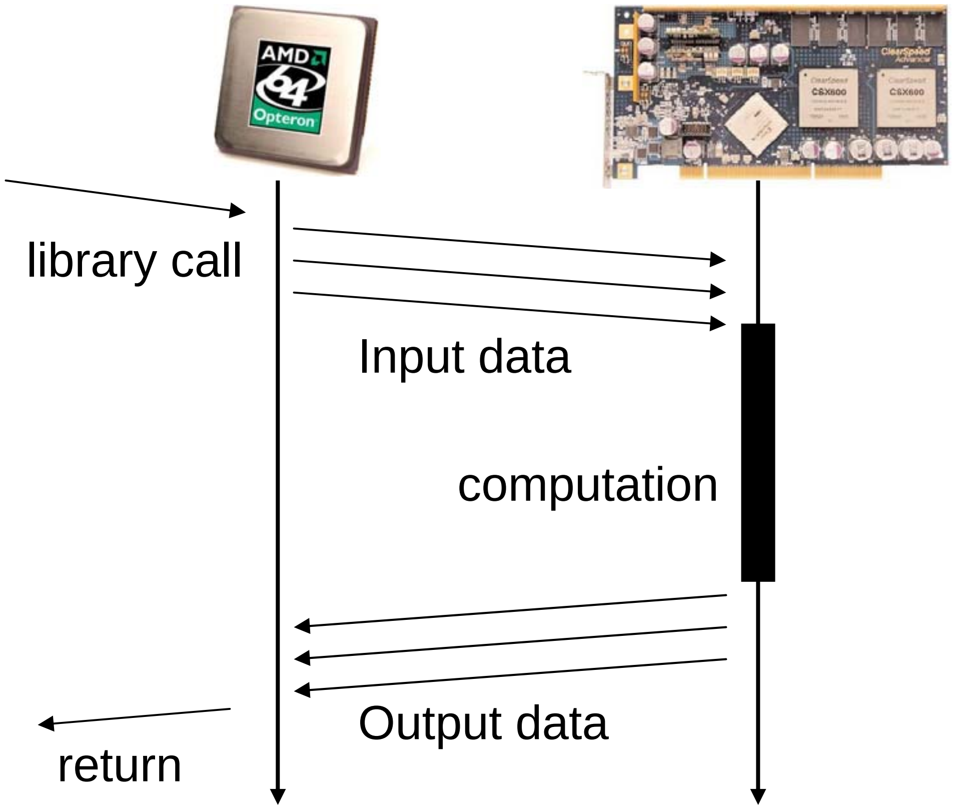
- Linear Algebra- BLAS, LAPACK
- Bio Simulations- AMBER, GROMACS
- Signal Processing - FFT (1D, 2D, 3D), FIR, Wavelet
- Various Simulations - CFD, FEA, N-body
- Image Processing - filtering, image recognition, DCTs
- Oil&Gas - Kirchhoff Time/Wave Migration

ClearSpeed Mode-of Use

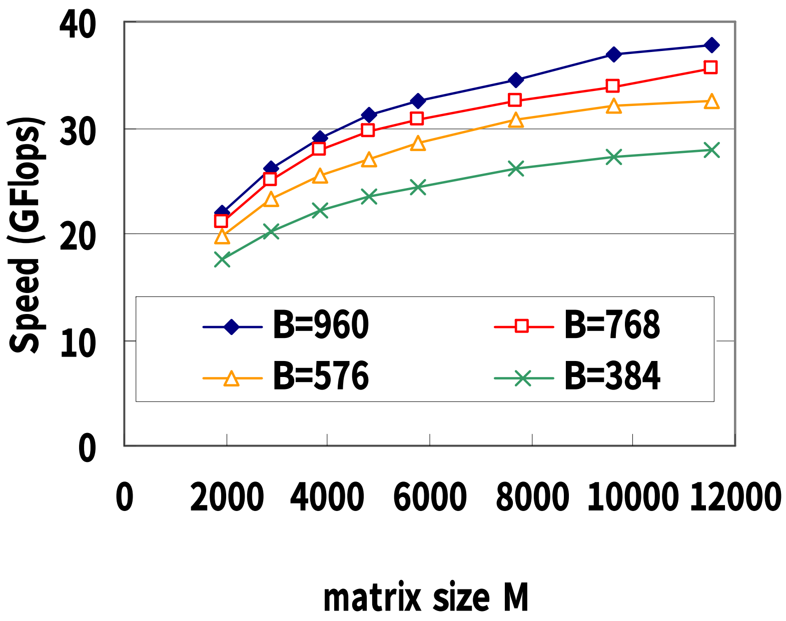
- 1. User Application Acceleration
  - Matlab, Mathematica, **Amber, Gaussian…**
  - Transparent, offload from Opterons
- 2. Acceleration of Standard Libraries
  - BLAS/DGEMM, LAPACK, FFTW…
  - Transparent to users (Fortran/C bindings)
- 3. User Applications
  - Arbitrary User Applications
  - Need MPI-like programming with C-dialect

Note: Acceleration is "Narrow Band"=> Hard to Scale

ClearSpeed Matrix Library



(MxB) x (BxM) multiplication speed



- About 40 GFlops DGEMM w/old library
  - 70GFlops with new beta(!)
- Performance heavily depends on matrix size

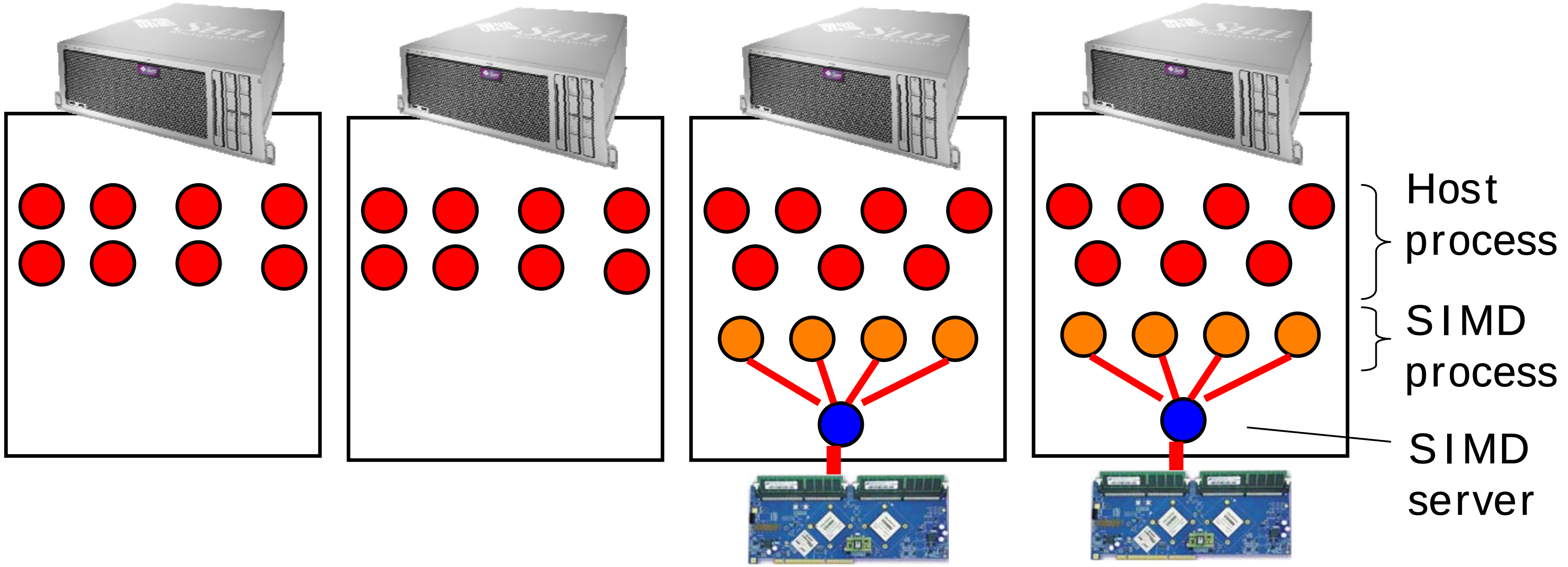
Issues in a (VERY) Heterogeneous HPL w/Acceleration

- How can we run HPL efficiently under following conditions?
  - Need to use efficiently both Opteron and ClearSpeed
    - About 70 GFlops by 16 Opteron cores
    - 30-40 GFlops by ClearSpeed (current)
  - Only (360/655) TSUBAME nodes have ClearSpeed
  - Modification to HPL code for heterogeneity
- Our policy:
  - Introduce HPL processes (1) that compute with Opterons and (2) that compute with ClearSpeed
  - Make workload of each HPL process (roughyl) equal by oversubscription

Our Heterogeneous HPL Algorithm

Two types of HPL processes are introduced

- Host processes use GOTO BLAS's DGEMM
- SIMD processes throw DGEMM requests to accelerator



Additional SIMD server directly calls CSXL DGEMM

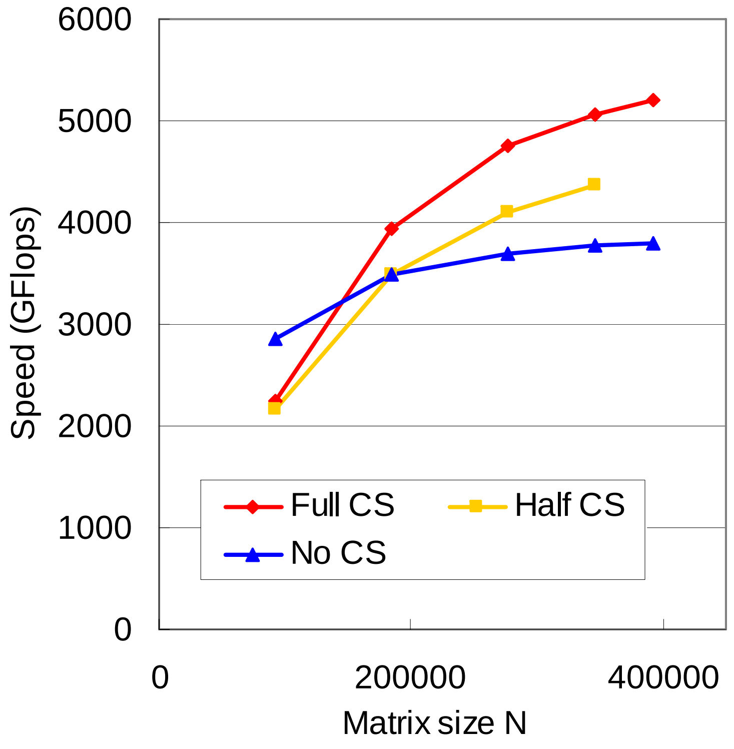
- mmap() is used for sharing matrix data

Linpack Details

- SunFire X4600 nodes in TSUBAME
  - Each has 16 Opteron cores, 32 GB memory
- Three measurements:
  - Full CS: ClearSpeed boards on all nodes are used
  - Half CS: # of ClearSpeed boards is the half of nodes
    - Heterogeneous in both intra and inter node
  - No CS: Only Opteron CPUs are used
- Numbers of processes per node are
  - With CS: 3 host processes (x4thread) + 3 SIMD processes
  - W/o CS: 4 host processes (x4thread)

Results(2)

Speed vs matrix size on 60 nodes



Block size NB is

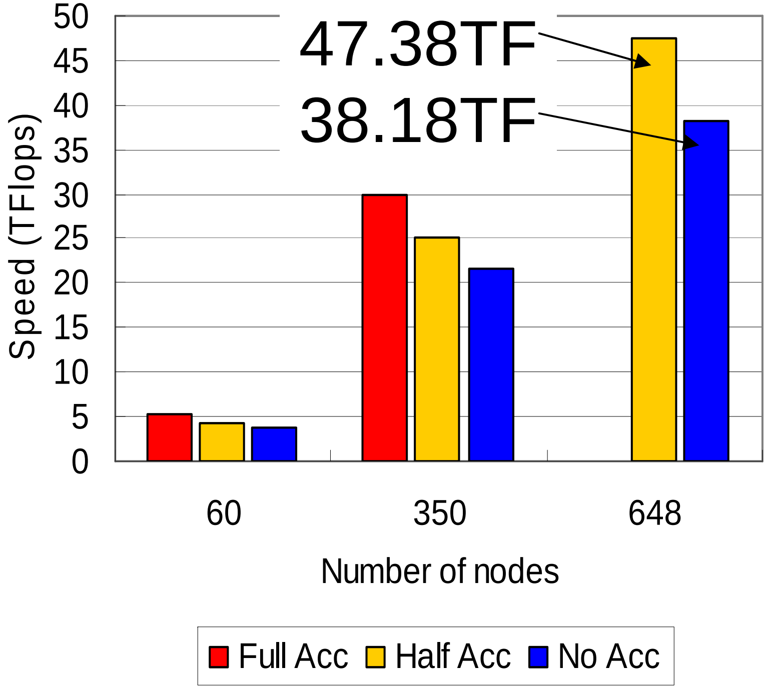
- 960 in Full CS/Half CS
- 240 in No CS

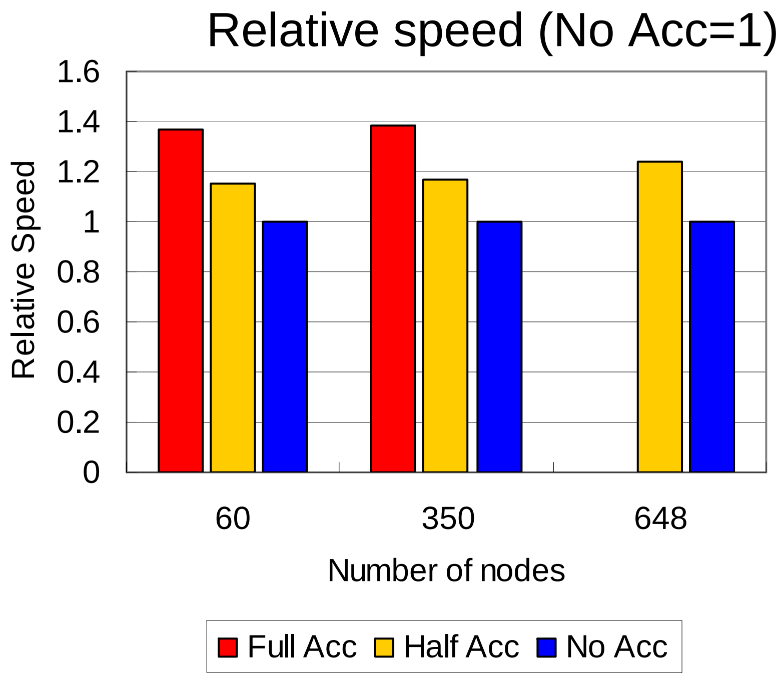
Peak speeds are

- Full CS: 5.203TFlops (N=391680)
- Half CS: 4.366TFlops (N=345600)
- No CS: 3.802TFlops (N=391680)

Note: Half CS doesn't work (very slow) with N=391680, because of the memory limitation

Experimental Results



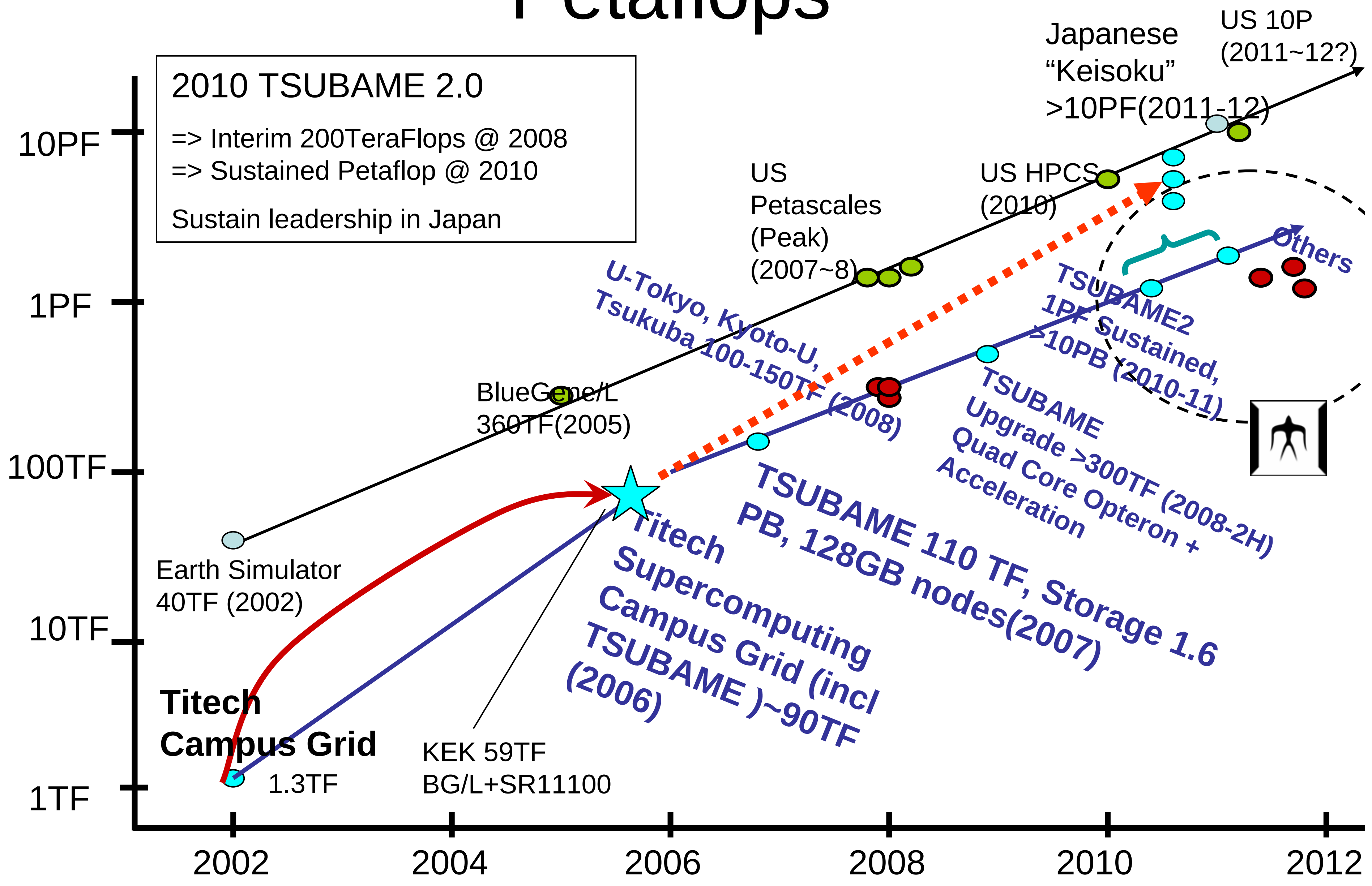


- 47.38TF with 648 nodes and 360 Accelerators Sep.
  - +24 % improvement over No Acc (38.18TF)
  - +25.5GFlops per accelerator
  - Matrix size N=1148160 (It was 1334160 in No Acc)
  - 5.9hours
- NEW(!) With new DGEMM, 48.88 TFlops / 62% Efficiency

Onto TSUBAME 2.0

Petascale and Beyond-but how?

TSUBAME Upgrades Towards Petaflops



2010 TSUBAME 2.0

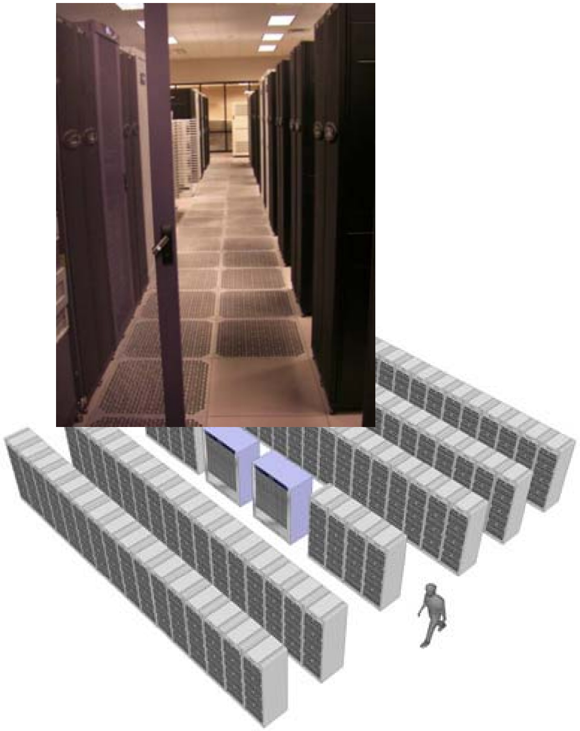
=> Interim 200TeraFlops @ 2008
=> Sustained Petaflop @ 2010

Sustain leadership in Japan

In the Supercomputing Landscape, Petaflops class is already here… in early 2008

Other Petaflops 2008/2009
- LANL/IBM “Roadrunner”
- JICS/Cray(?) (NSF Track 2)
- ORNL/Cray
- ANL/IBM BG/P
- EU Machines (Julich…)

…



2008 LLNL/IBM “BlueGene/P”
~300,000 PPC Cores, ~1PFlops
~72 racks, ~400m2 floorspace
~3MW Power, *copper* cabling

2008Q1 TACC/Sun “Ranger”
~52,600 “Barcelona” Opteron
CPU Cores, ~500TFlops
~100 racks, ~300m2 floorspace
2.4MW Power, 1.4km IB cx4
copper cabling
2 Petabytes HDD

> 10 Petaflops
> million cores
> 10s Petabytes
planned for 2011-2012
in the US, Japan, (EU),
(other APAC)

Scaling to a PetaFlop in 2010 is Easy, Given Existing TSUBAME

Year	2003	2006	2008	2010	2012	2014	2015
Microns	0.09	0.065	0.045	0.032	0.022	0.016	0.011
Scalar Cores	1	2	4	8	16	32	64
GFLOPS/Socket	6	24	48	96	192	384	768
Total KW for 1 PF (200W/Socket)	3.3E+05	83333	41667	20833	10417	5208	2604
SIMD/Vector	-	96	192	384	768	1536	3072
GFLOPS/Board	-	96	192	384	768	1536	3072
Total KW for 1 PF (25W/Board)	-	260.4	130.2	65.1	32.6	16.3	8.14

2009 Conservatively Assuming 0.065-0.045 microns, 4 cores, 48 GFlops/Socket=>200Teraflops, 800 Teraflop Accelerator board

“Commodity” Petaflop *easily* achievable in 2009-2010

In fact we can build one now (!)

- @Tokyo---One of the Largest IDC in the World (in Tokyo...)
- Can fit a 10PF here easy (> 20 Rangers)
- On top of a 55KV/6GW Substation
- 150m diameter (small baseball stadium)
- 140,000 m2 IDC floorspace
- 70+70 MW power
- Size of entire Google(?) (~million LP nodes)

Commodity Scaling to 2~10 PFs Circa 2011 (Cont'd)

- Loosely coupled apps scale well
- Impractical to assume memory intensive, large message apps (such as spectral methods) to scale to Petaflops
  - Strong technological scaling limits in memory size, bandwidth, etc.
  - Physical limits e.g., power/cooling, $$$
  - Impracticality in resolution (because of chaotic nature of physics, etc.)
- Why ensemble methods and coupled methods (which are scalable) are good
  - => Apps that worked "well on grids" (small scale)

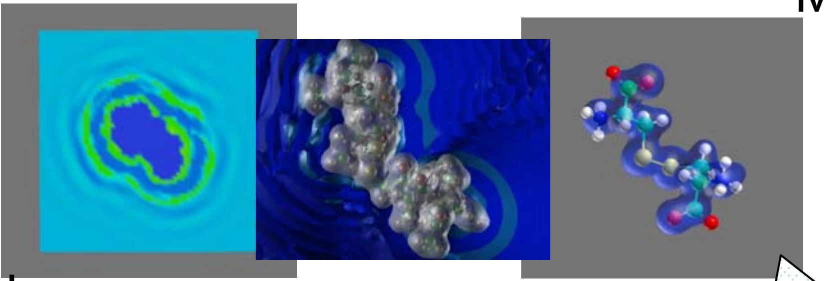
Nano-Science : coupled simluations on the Grid as the sole future for true scalability

… between Continuum & Quanta.

Material physics
(Infinite system)

- Fluid dynamics
- Statistical physics
- Condensed matter theory

…



Molecular Science

- Quantum chemistry
- Molecular Orbital method
- Molecular Dynamics

…

10^{-6} 10^{-9} m

E.g., Advanced MD, req. mid-sized tightly-coupled SMP (#CPU not the limit, but memory and BW)

Multi-Physics

E.g. Fragmented MO, Could use 100,000 loosely-coupled CPUs in pseudo paramter

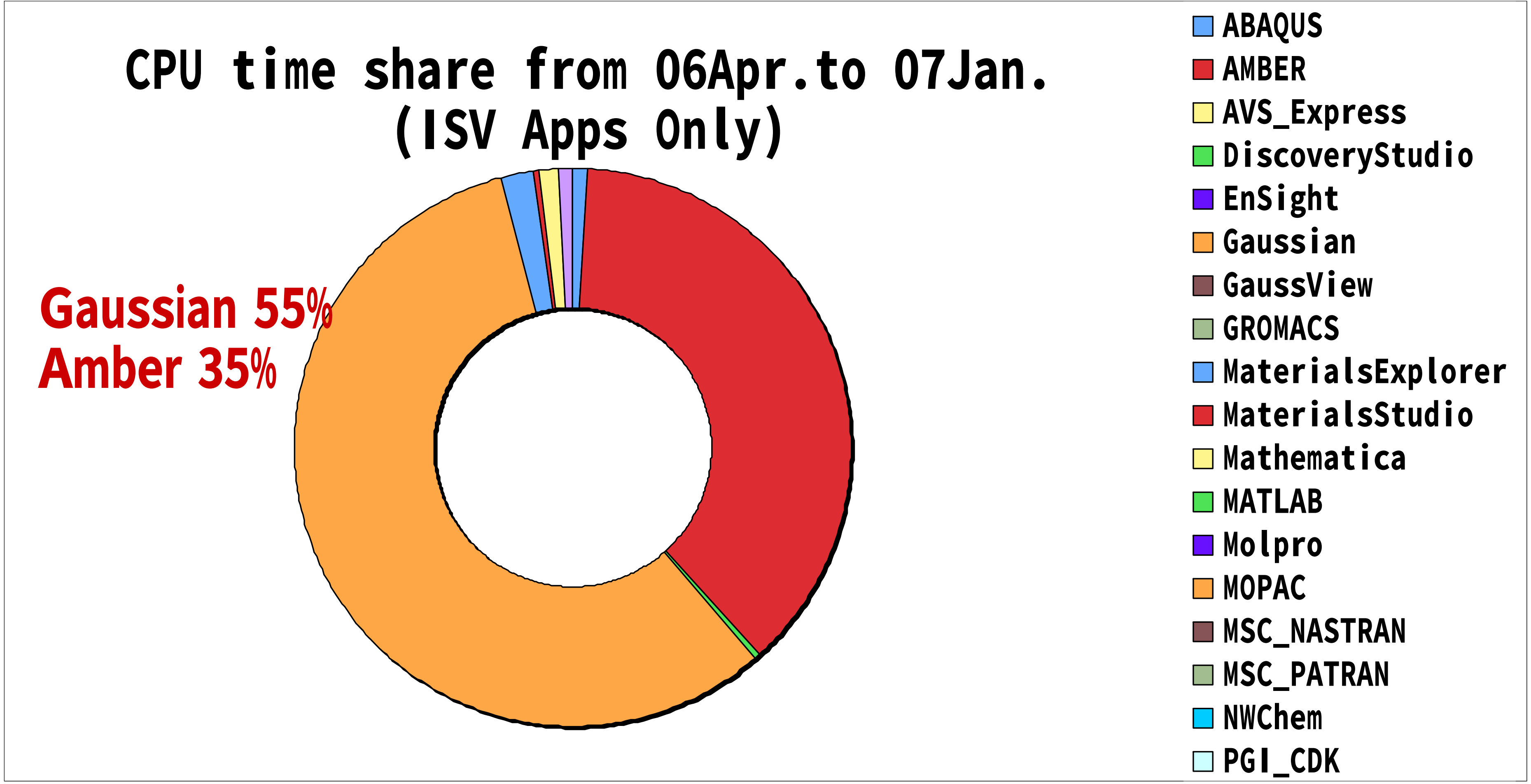
Old HPC environment:

- decoupled resources,
- hard to use,
- special software, ...
- **Too general-purpose(!)**

The only way to achieve true scalability!

Slide stolen from my NAREGI Grid Slide Stack => Tightly-coupled "Grid" as future Petascale machine

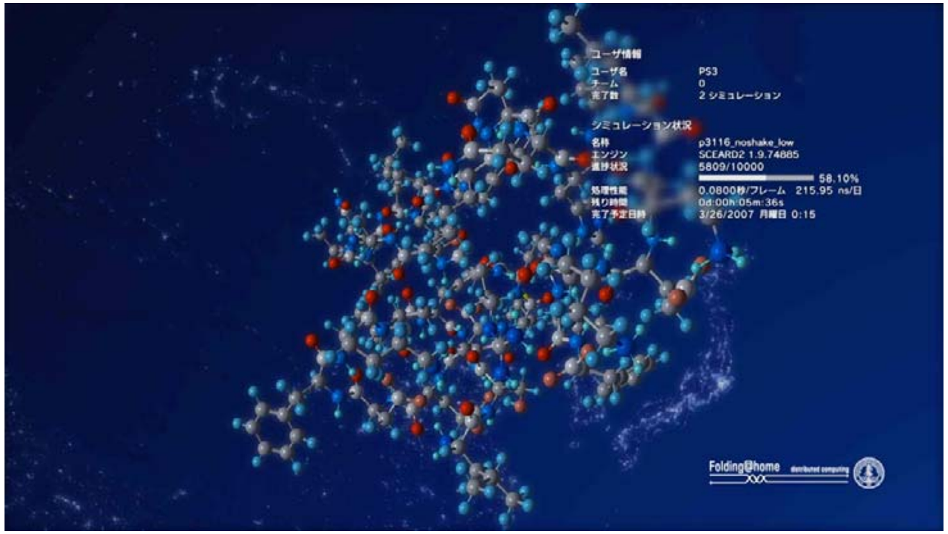
Reprisal: TSUBAME Job Statistics for ISV Apps (# CPU Timeshare)



Multi-User and Ensemble! (20,000-way Gaussian ensemble job recorded on TSUBAME) => Throughput(!)

Standford Folding@Home

- (Ensemble) GROMACS, Amber etc. on Volunteer Grid
- PS3: 1/2 (effective) Petaflops and growing (in standard OS(!))
- Accelerator (GPGPU) most Flops/CPU/unit
- Combined, 71% effective FLOPS @ 14% CPUs
- 7 Petaflops Peak (SFP), 10% efficiency
  - Feasible *NOW* to build a *useful* 10PF machine



ユーザ情報

ユーザ名 PS3
チーム 0
完了数 2 シミュレーション

シミュレーション状況

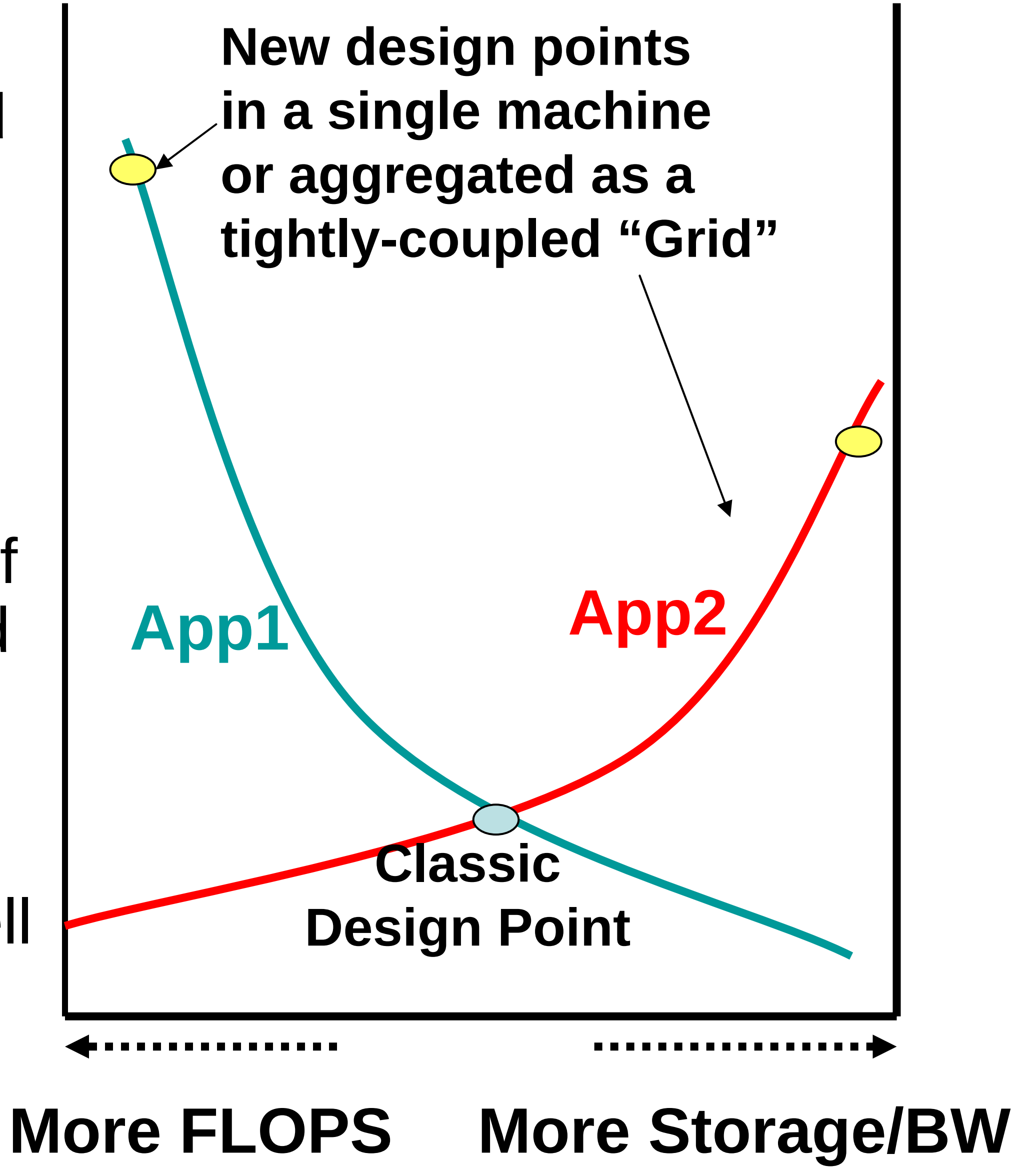
名称 p3116_noshake_low
エンジン SCEARD2 1.9.74885
進捗状況 5809/10000
58.10%
処理性能 0.0800秒/フレーム 215.95 ns/日
残り時間 0d:00h:05m:36s
完了予定日時 3/26/2007 月曜日 0:15

Folding@Home 2007-03-25 18:18:07

OS Type	TFLOPS	Active CPUs	GFLOPS/CPU
Windows	154	161,586	0.95
Mac/PPC	7	8,880	0.79
Mac/Intel	9	3,028	2.97
Linux	43	25,389	1.69
GPGPU	**44**	**749**	**58.74**
PS3	**482**	**30,294**	**15.91**
Total	739	229926	3.21

Future Multi-Petascale Designs

- Assuming Upper bound on Machine Cost
- A homogeneous 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



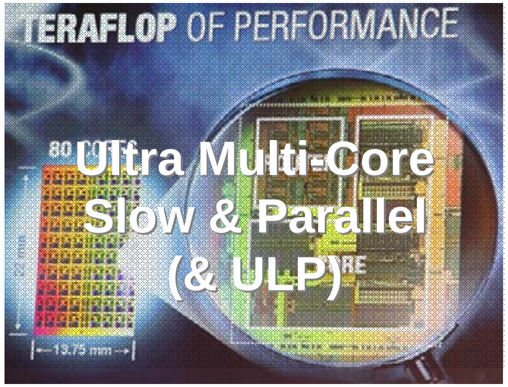
Biggest Problem is Power…

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS/Watt	Watts/CPU Core	Ratio c.f. TSUBAME
TSUBAME(Opteron)	10480	800,000	50,400	63.00	76.34	
TSUBAME(w/ClearSpeed)	11,200	810,000	85,000	104.94	72.32	1.00
Earth Simulator	5120	6,000,000	40,000	6.67	1171.88	0.06
ASCI Purple (LLNL)	12240	6,000,000	77,824	12.97	490.20	0.12
AIST Supercluster	3188	522,240	14400	27.57	163.81	0.26
LLNL BG/L (rack)	2048	25,000	5734.4	229.38	12.21	2.19
Next Gen BG/P (rack)	4096	30,000	16384	546.13	7.32	5.20
TSUBAME 2.0 (2010Q3/4)	160,000	810,000	2,048,000	2528.40	5.06	24.09

TSUBAME 2.0 x24 improvement in 4.5 years…? ➔ ~ x1000 over 10 years

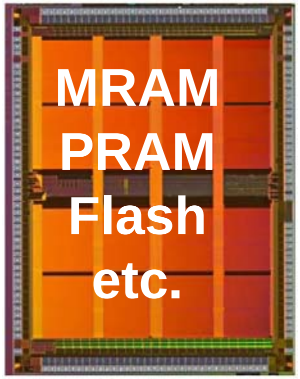
The new JST-CREST "Ultra Low Power HPC" Project 2007-2012

- x1000 Flops/W improvement @ 10 years -



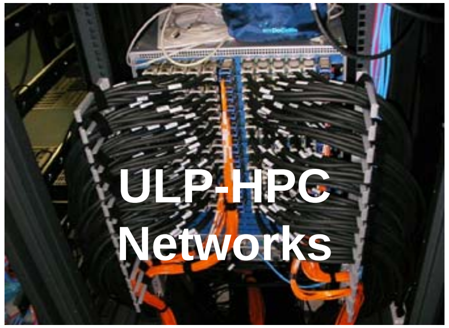
Ultra Multi-Core
Slow & Parallel
(& ULP)

ULP-HPC
SIMD-Vector
(GPGPU, etc.)



MRAM
PRAM
Flash
etc.

Zero Emission
Power Sources

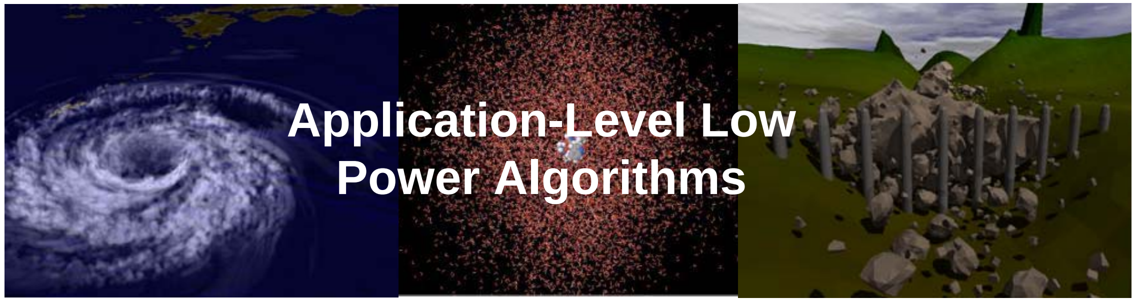


ULP-HPC
Networks

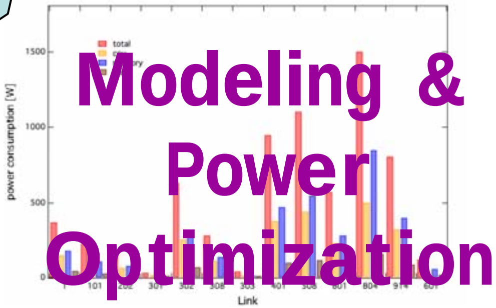
New Massive & Dense
Cooling Technologies

Xen vmware
VPC
Tokyo Tech

VM Job Migration
Power Optimization



Application-Level Low
Power Algorithms



Modeling & Power Optimization

TSUBAME in Retrospect and Future

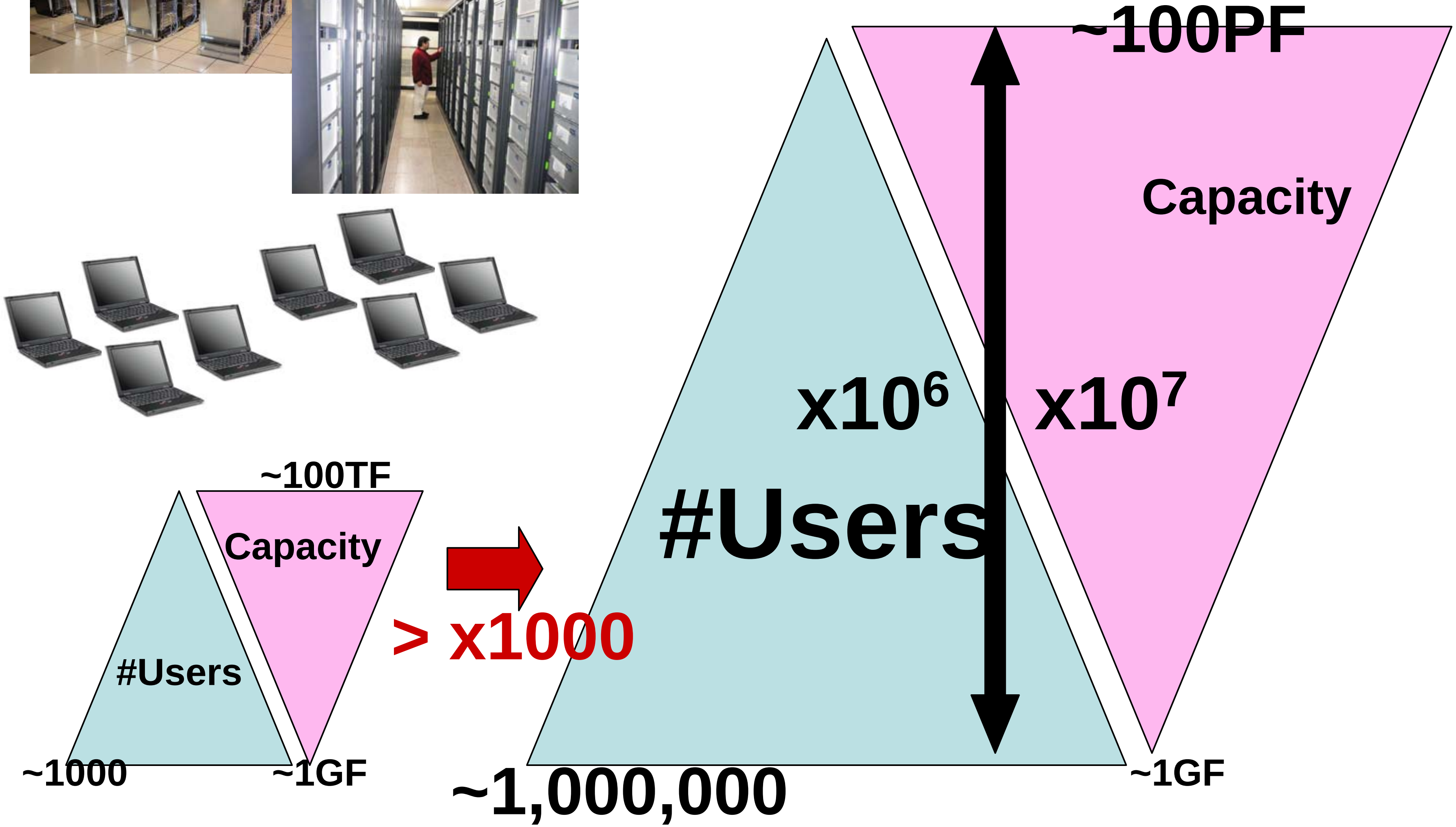
- Increasing Commoditization of HPC Space
  - CPUs (since Beowulf, ASCI Red, …)
  - High BW memory, Large-memory SMP
  - Very Fast I/O (PCI-E, HT3, …)
  - High BW Interconnect (10GbE, IB => 100Gb)
  - Now SIMD-Vector (ClearSpeed, GPGPU, Cell…)
  - Next: Extreme Many-Core, Optical Chip-Chip interconnect, 3-D Chip Packaging, …

Timeline

TSUBAME

- Technology => Software Stack & the right apps & meta-application schema
  - The same software stack on your laptop + Grid
  - DON'T focus on a single app or user efficiency – meta-application schema, multi-user, infrastructue design
  - Learn from the Grid (!)
- proprietary architectures makes no sense
  - Ecosystems and Economics THE KEY of future HPC(!)

Beyond Petascale "Grid"
Scalability is the key



~100TF

Capacity

#Users

~1000

~1GF

> x1000

~100PF

Capacity

$x10^6$

$x10^7$

#Users

~1,000,000

~1GF

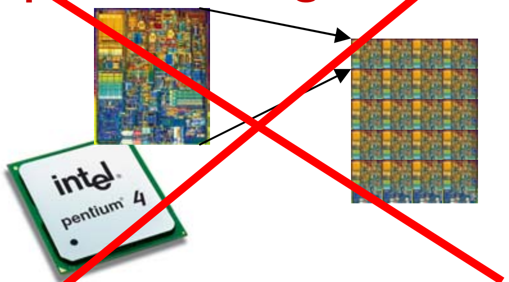
2016A.D. Deskside Petascale

2006A.D. Titech Supercomputing Grid
#1 in Asia: 100TeraFlops,
> 10,000 CPU, 1.5 MegaWatt, 300m^2

1000 times scaling down of a SC: but how?

2016 Deskside Workstation
>100TeraFlops, 1.5KiloWatt, 300cm^2

Simple scaling will not work



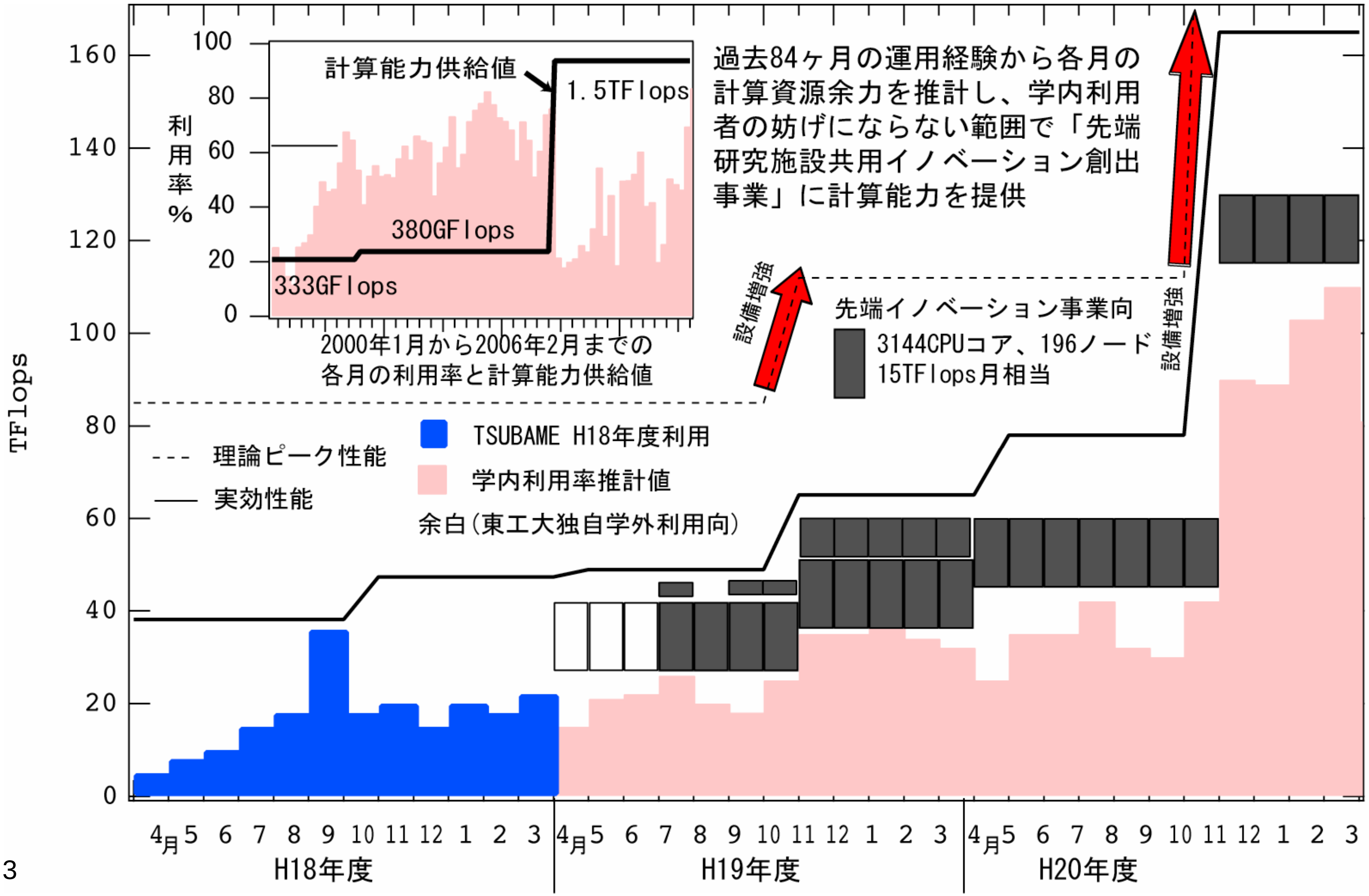
No more aggressive clock increase
Multi-core works but less than x100

Need R&D as “Petascale Informatics” in CS and Applications to achieve x1000 breakthrough

+

What can a scientist or an engineer achive with daily, personal use of petascale simulation?

Seasonal Corporate Usage



過去84ヶ月の運用経験から各月の計算資源余力を推計し、学内利用者の妨げにならない範囲で「先端研究施設共用イノベーション創出事業」に計算能力を提供

申請書p.3