Hetero-Acceleration: the Yellow Brick Road onto Exascale?

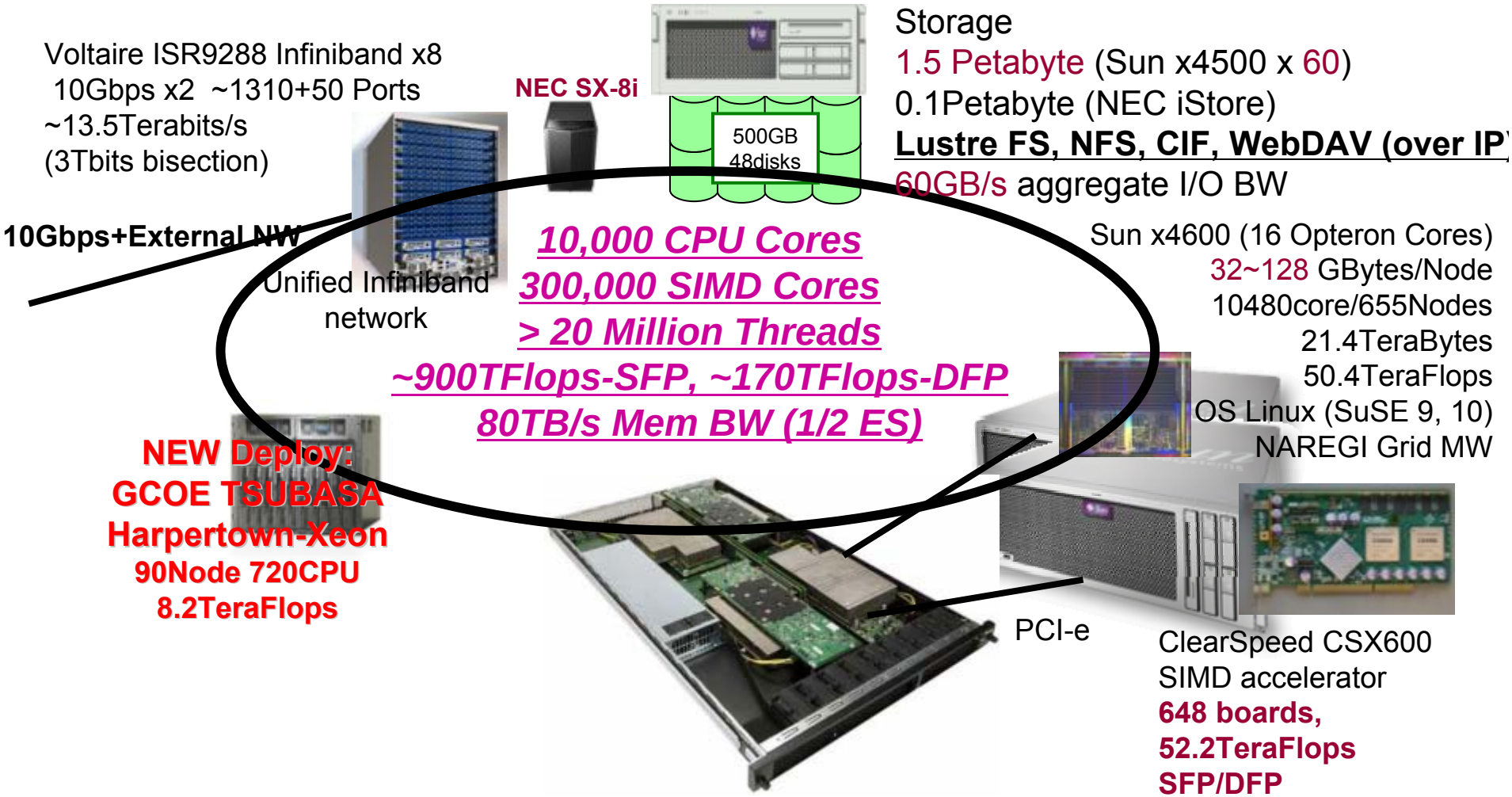
(Based on Overview of TSUBAME 2.0 and Beyond)

Satoshi Matsuoka
Global Scientific Information and Computing Center (GSIC)
Tokyo Institute of Technology (Tokyo Tech.)

HPC 2010, Cetraro, Italy
20/Jun/2010

TSUBAME 1.2 Experimental Evolution (Oct. 2008)

The first "Petascale" SC in Japan



170 Nvidia Tesla 1070, ~680 Tesla cards
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0


680 Unit Tesla Installation…
While TSUBAME in Production Service (!)

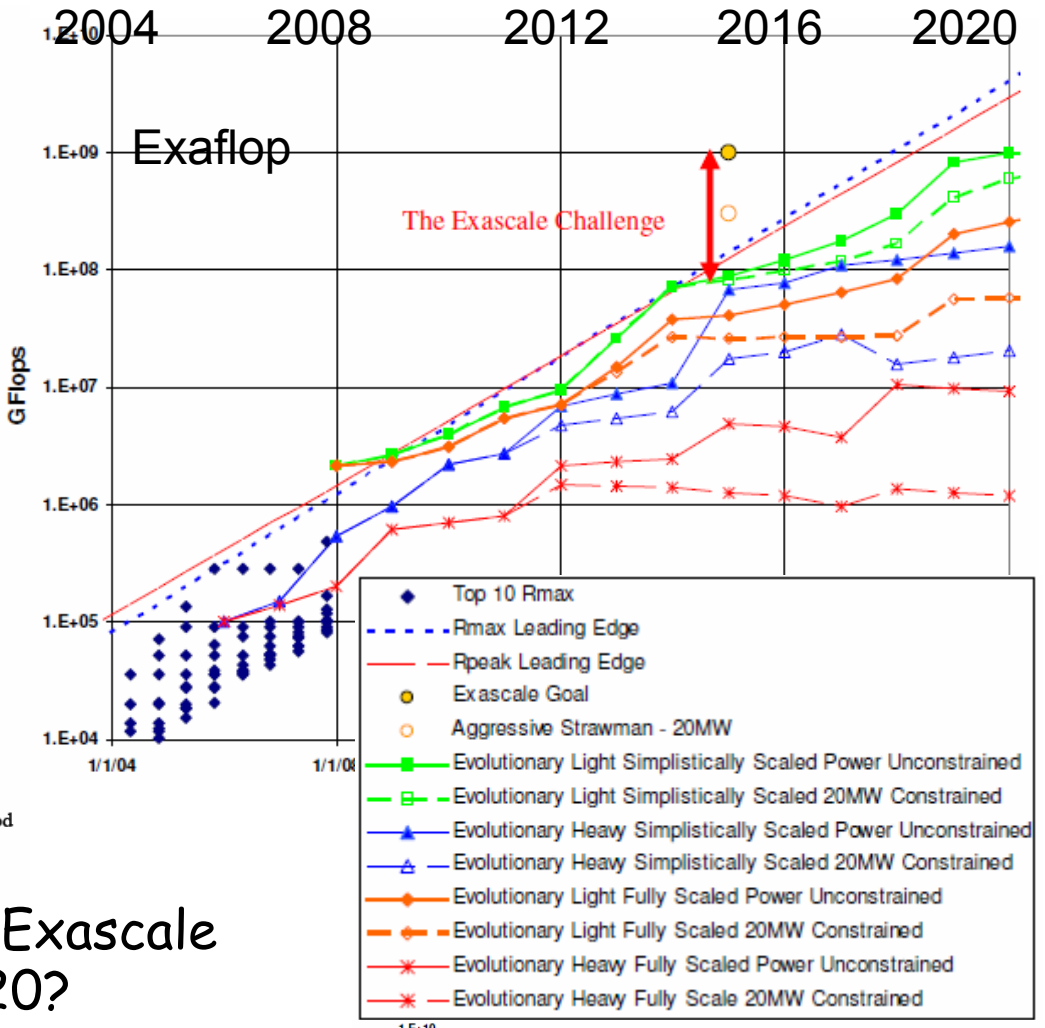
ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

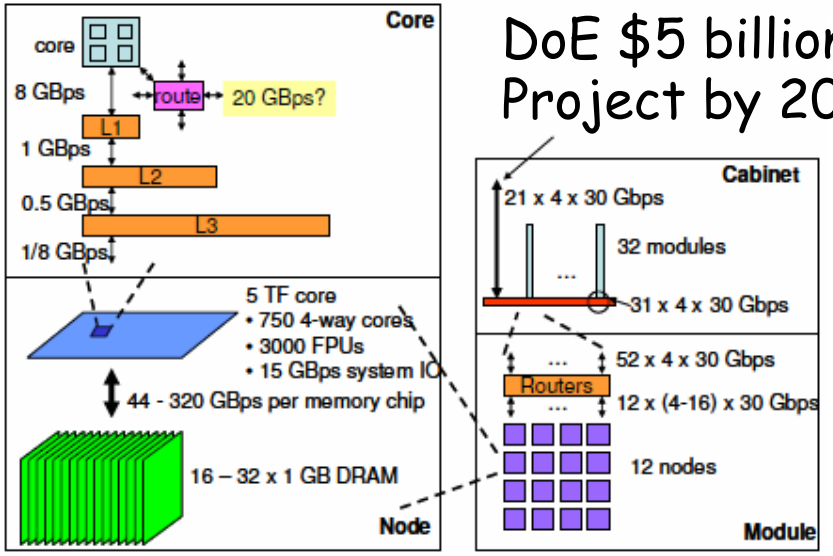
September 28, 2008

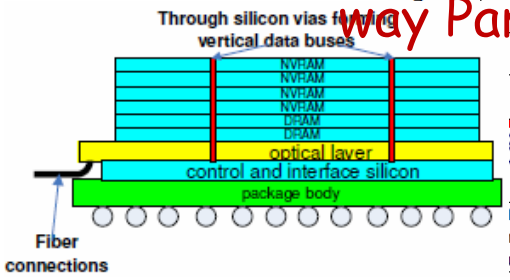
This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the

Peter Kogge 300page DoD Exascale Report

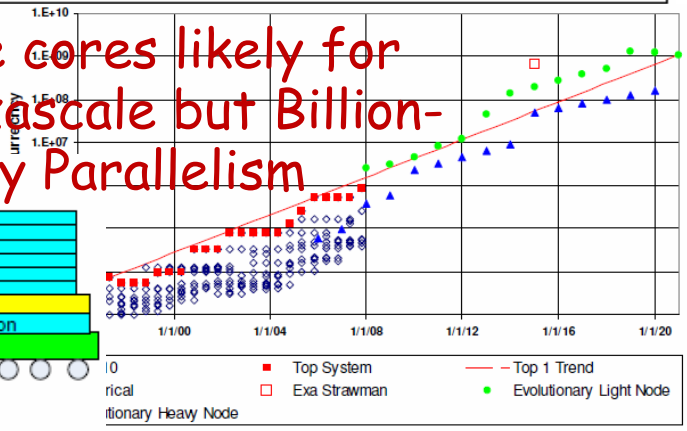


DoE $5 billion Exascale Project by 2020?

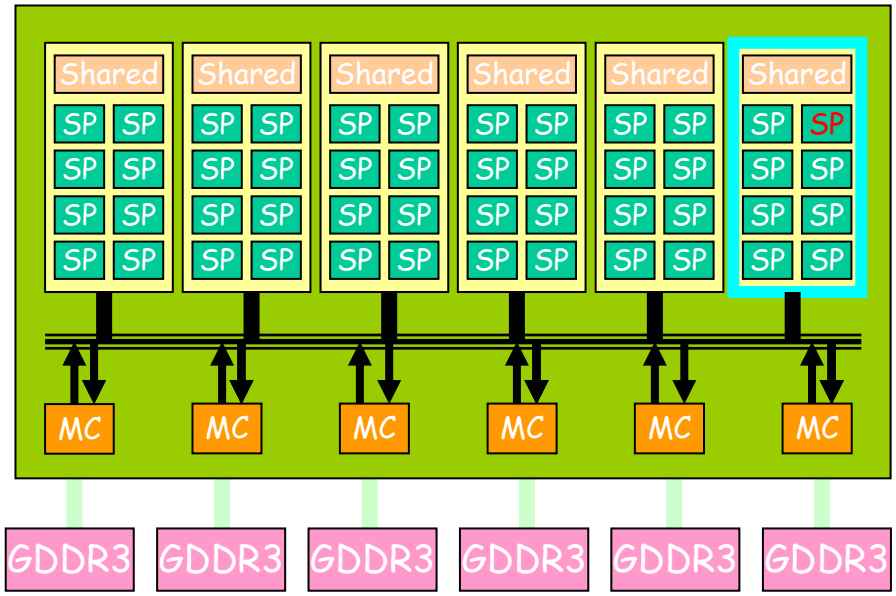




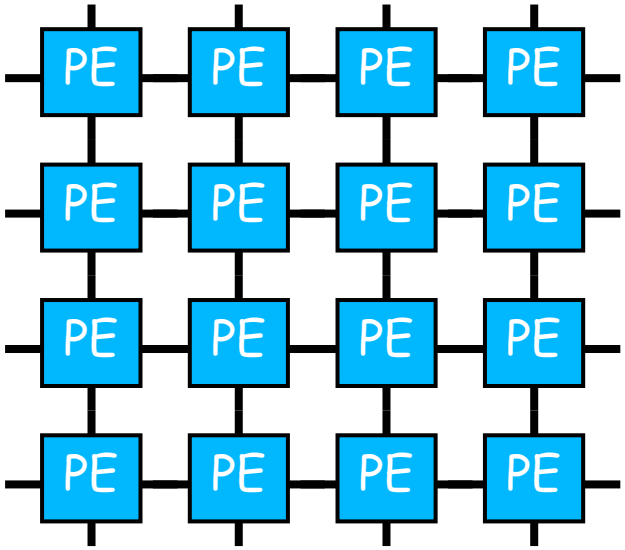
Simple cores likely for 2020 exascale but Billion-way Parallelism



GPU (Multithreaded Vector) vs. Standard Many Cores?

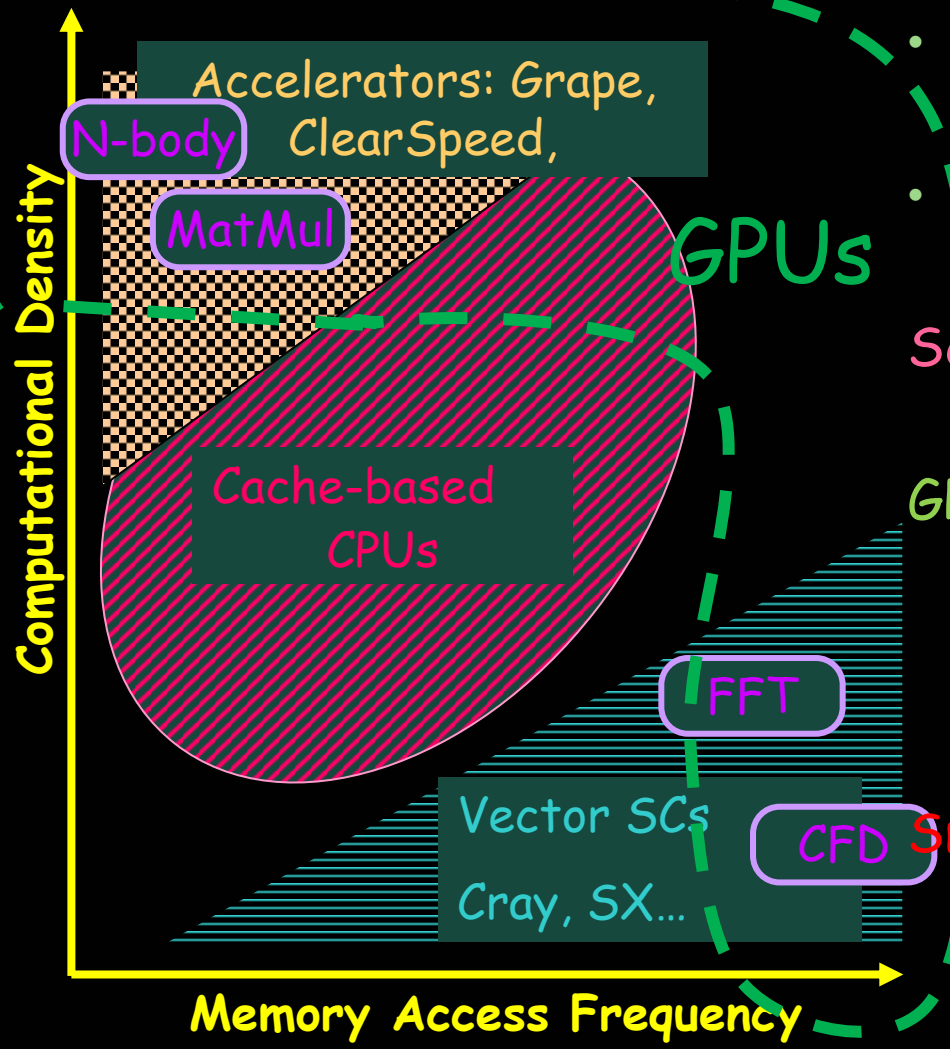


vs.



- Consider the choices in terms of ***strong scaling*** of the system, software, and applications (co-design)

GPUs as Modern-Day Vector Engines



Two types of HPC Workloads

- High Computational Density
  => Traditional Accelerators
- High Memory Access Frequency
  => Traditional Vector SCs

Scalar CPUs are so-so at both
=> Massive system for speedup

GPUs are both ***modern-day vector engines and high compute density accelerators***
=> Efficient Element of Next-Gen Supercomputers

Small memory, limited CPU-GPU BW, high-overhead communication with CPUs and other GPUs, lack of system-level SW support incl. fault tolerance, programming, …

Our Research@ Tokyo Tech

DOE SC Applications Overview

(following slides courtesy John Shalf @ LBL NERSC)

NAME	Discipline	Problem/Method	Structure
MADCAP	Cosmology	CMB Analysis	Dense Matrix
FVCAM	Climate Modeling	AGCM	3D Grid
CACTUS	Astrophysics	General Relativity	3D Grid
LBMHD	Plasma Physics	MHD	2D/3D Lattice
GTC	Magnetic Fusion	Vlasov-Poisson	Particle in Cell
PARATEC	Material Science	DFT	Fourier/Grid
SuperLU	Multi-Discipline	LU Factorization	Sparse Matrix
PMEMD	Life Sciences	Molecular Dynamics	Particle

Latency Bound vs. Bandwidth Bound?

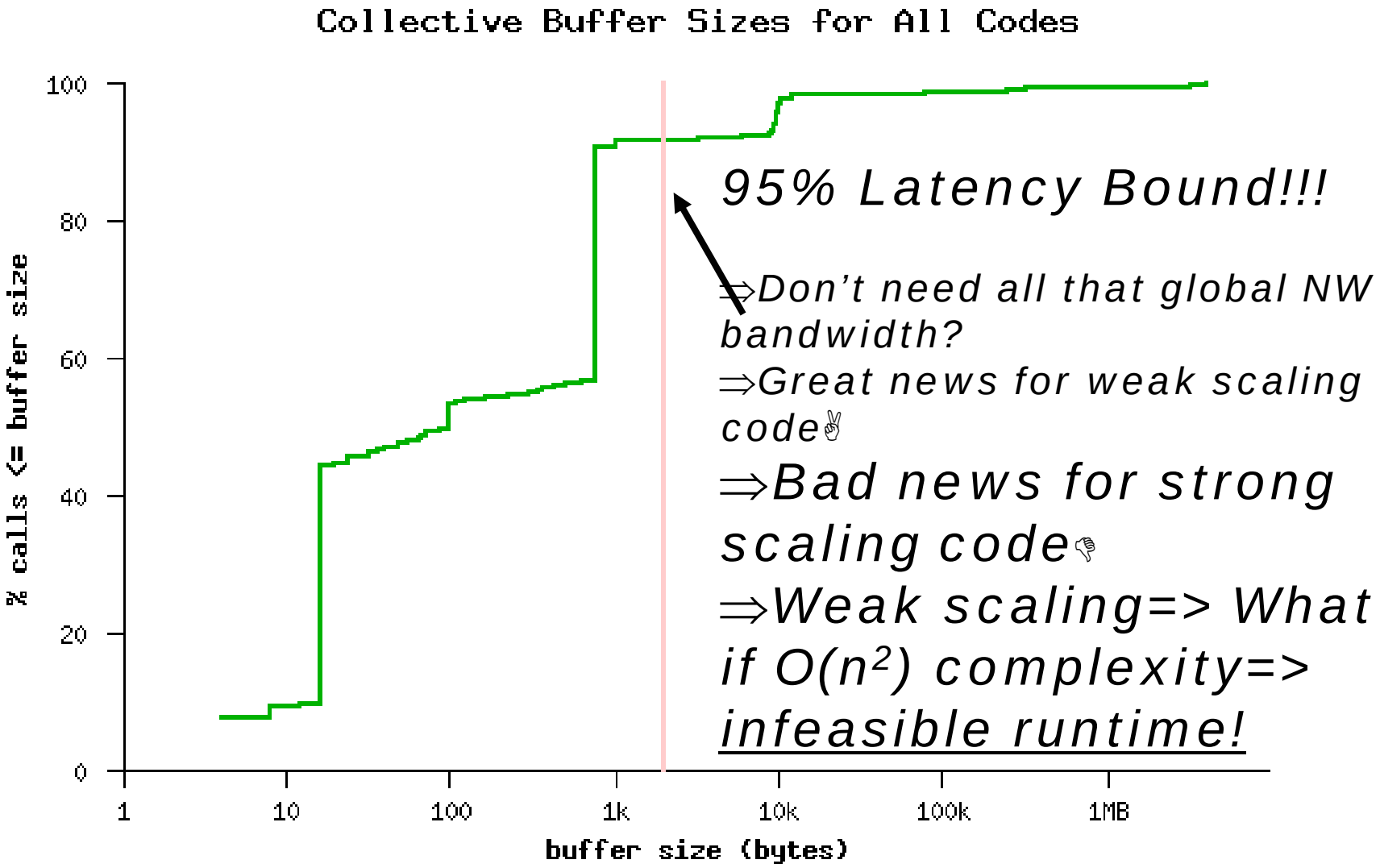
- How large does a message have to be in order to saturate a dedicated circuit on the interconnect?
  - $N^{1/2}$ from the early days of vector computing
  - Bandwidth Delay Product in TCP

System	Technology	MPI Latency	Peak Bandwidth	Bandwidth Delay Product
SGI Altix	Numalink-4	1.1us	1.9GB/s	2KB
Cray X1	Cray Custom	7.3us	6.3GB/s	46KB
NEC ES	NEC Custom	5.6us	1.5GB/s	8.4KB
Myrinet Cluster	Myrinet 2000	5.7us	500MB/s	2.8KB
Cray XD1	RapidArray/IB4x	1.7us	2GB/s	3.4KB

- Bandwidth Bound if msg size > Bandwidth*Delay
- Latency Bound if msg size < Bandwidth*Delay
  - Except if pipelined *(unlikely with MPI due to overhead)*
- W/HW DMA a few 100ns but not much more

(Original slide courtesy John Shalf @ LBL)

Collective Buffer Sizes are Small(!)
- demise of message passing in strong scaling -

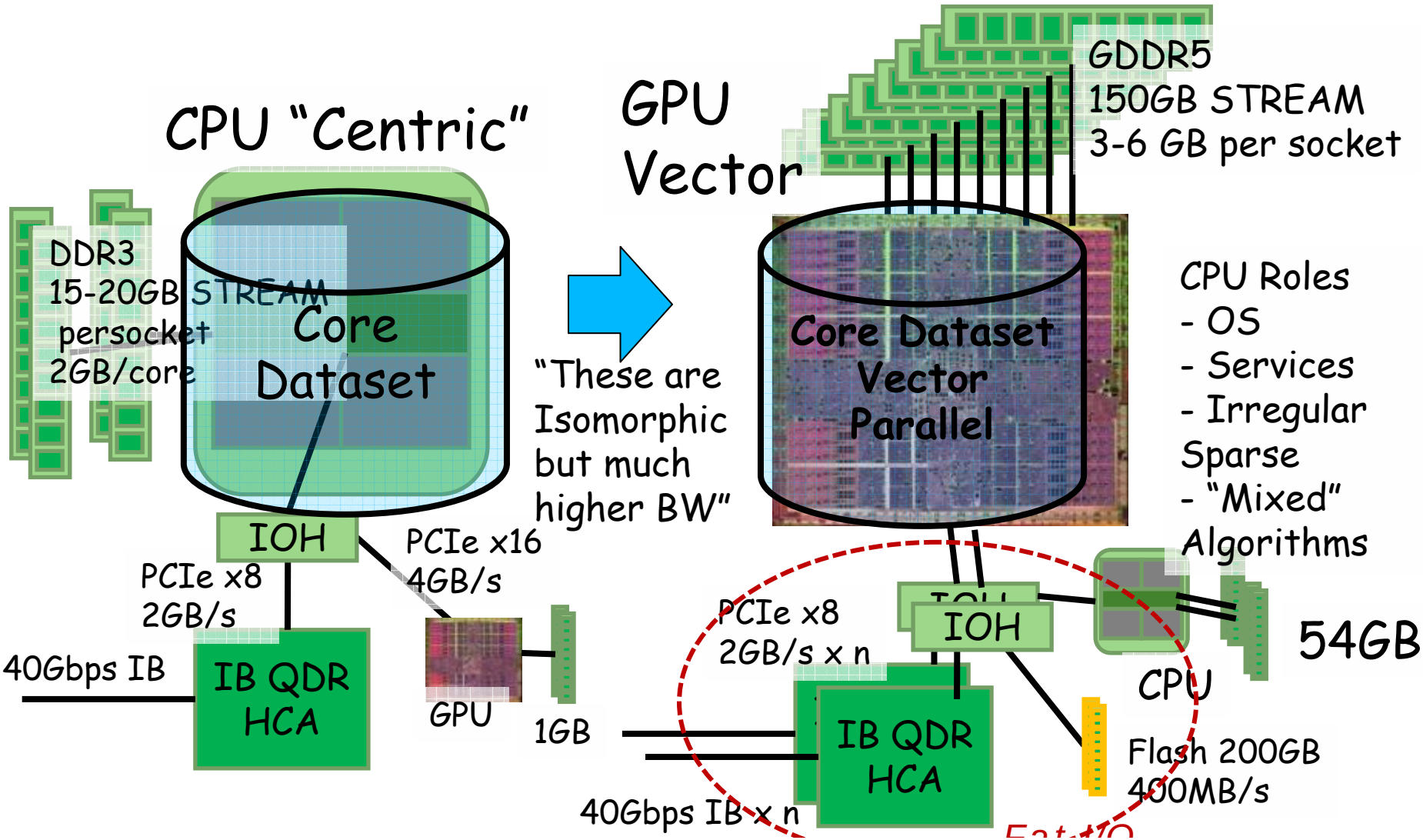


(Original slide courtesy John Shalf @ LBI

Scaling Peta to Exa Design?

- ***Shorten*** latency as much as possible
  - Extreme multi-core incl. vectors
  - "Fat" nodes, exploit short-distance interconnection
  - Direct cross-node DMA (e.g., put/get for PGAS)
- ***Hide*** latency if cannot be shortened
  - Dynamic multithreading (Old: dataflow, New: GPUs)
  - Trade Bandwidth for Latency (so we do need BW…)
  - Departure from simple mesh system scaling
- ***Change*** Latency-Starved Algorithms
  - From implicit Methods to direct/hybrid methods
  - Structural locality, extraploation, stochastics (MC)
  - Need good programming model/lang/system for this…

From TSUBAME 1.2 to 2.0: From CPU Centric to GPU Centric Nodes for Scaling



Highlights of TSUBAME 2.0 Design (Oct. 2010) w/NEC-HP

- 2.4 PF Next gen multi-core x86 + next gen GPGPU
  - 1432 nodes, Intel Westmere/Nehalem EX
  - 4224 NVIDIA Tesla (Fermi) M2050 GPUs
  - ~100,000 total CPU and GPU "cores", High Bandwidth
  - **1.9 million "CUDA cores", 32K x 4K = 130 million CUDA threads(!)**

- 0.72 Petabyte/s aggregate mem BW,
  - Effective 0.3-0.5 Bytes/Flop, restrained memory capacity (100TB)
- Optical Dual-Rail IB-QDR BW, full bisection BW(Fat Tree)
  - 200Tbits/s, Likely fastest in the world, still scalable
- Flash/node, ~200TB (1PB in future), 660GB/s I/O BW
  - >7 PB IB attached HDDs, 15PB Total HFS incl. LTO tape
- Low power & efficient cooling, comparable to TSUBAME 1.0 (~1MW); PUE = 1.28 (60% better c.f. TSUBAME1)
- Virtualization and Dynamic Provisioning of Windows HPC + Linux, job migration, etc.

TSUBAME 2.0 Performance

Earth Simulator ⇒ TSUBAME 4years
x 40 Downsizing

TSUBAME ⇒
TSUBAME 2.0
x40 downsizing
again?

TSUBAME2.0
2.4 PF (2010-2H)

Japanese NLP
>10PF(2012)
US >10P
(2011~12)

US/EU
Petascales
(Peak)
(2008)

US NSF/DoE
(2009 2H)

U-Tokyo, Kyoto-U,
Tsukuba T2K 61-140TF (2008)

Others

BlueGene/L
360TF(2005)

TSUBAME 1.2
Upgrade
(3Q2008) 170TF

TSUBAME 1.1 Upgrade 111TF, 1.6
PB, 128GB nodes (1Q2008)

TSUBAME 1.0 80TF (1Q2006)

Earth Simulator
40TF (2002)

Titech
Campus Grid
1.3TF

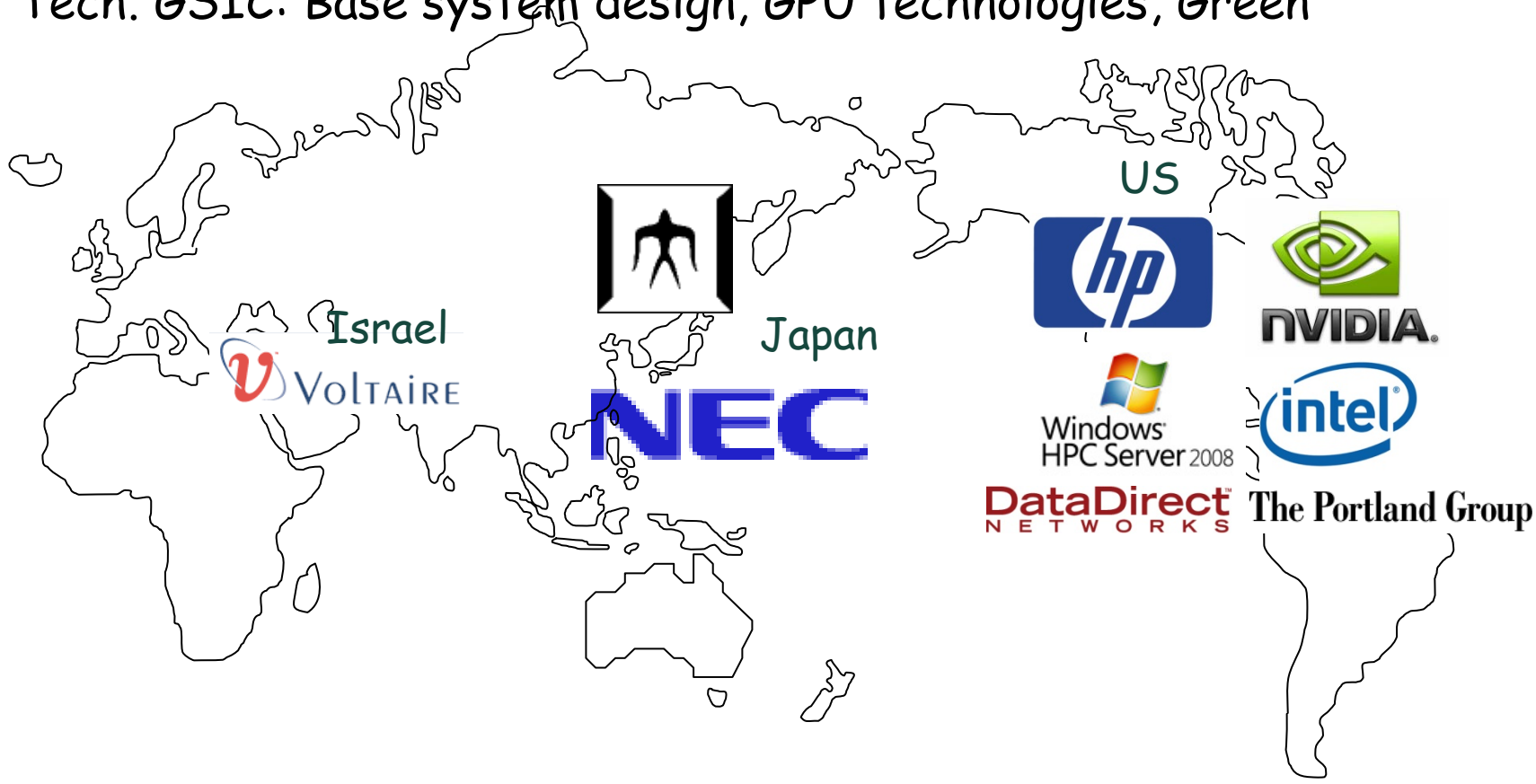
KEK 59TF
BG/L+SR11100

10PF
1PF
100TF
10TF
1TF

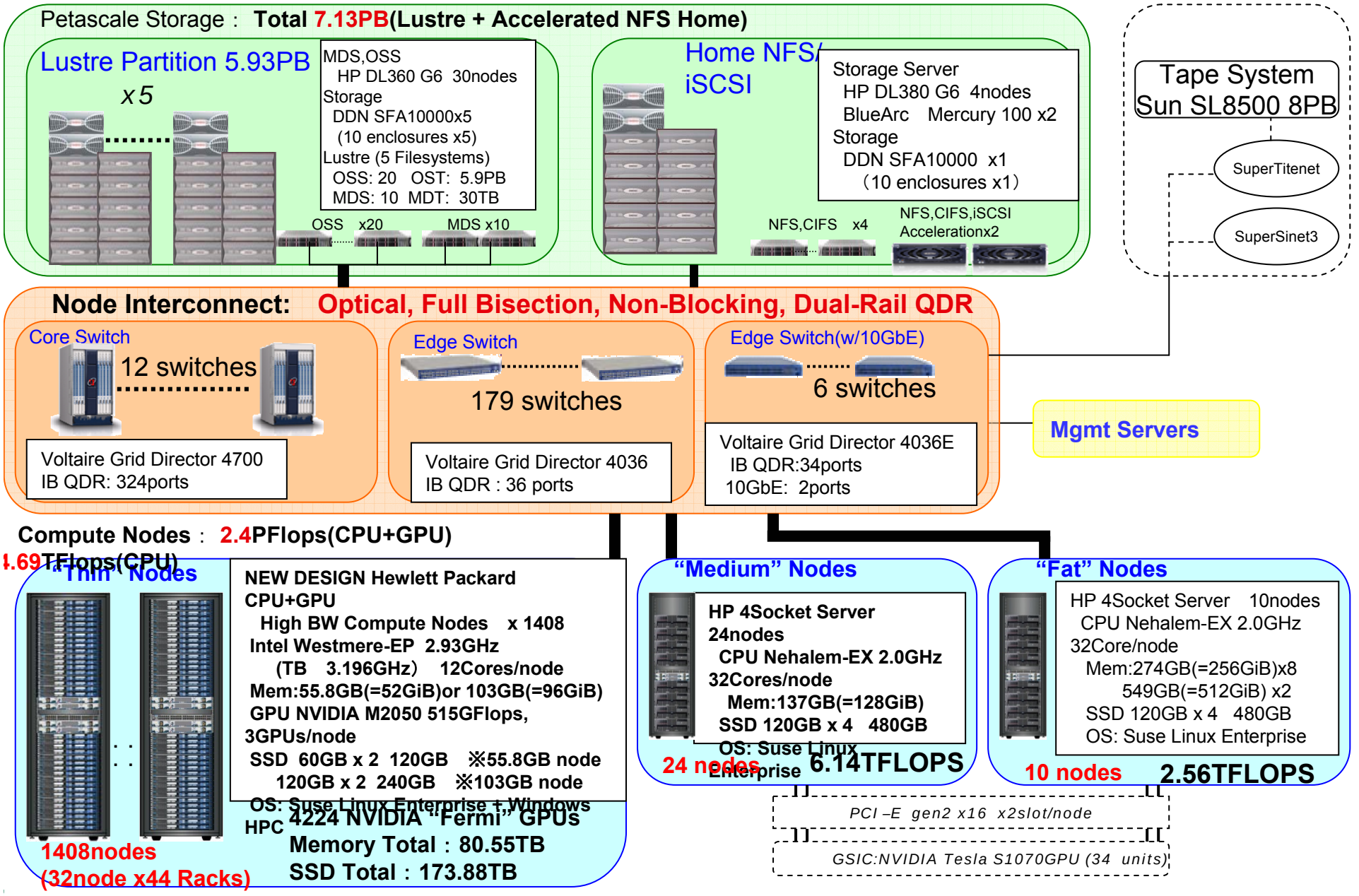
2002
2004
2006
2008
2010
2012

TSUBAME2.0 Global Partnership

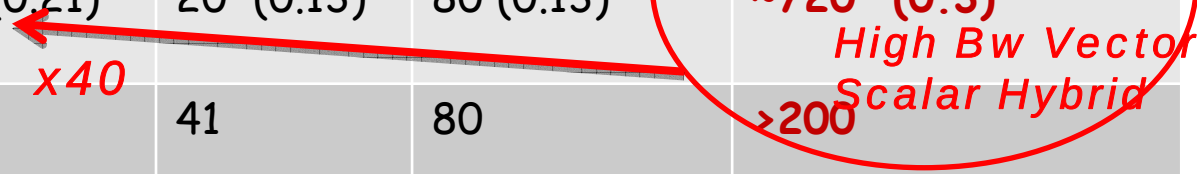
NEC: Main Contractor, overall system integration, Cloud-SC mgmt.
HP: Node R&D, Green; Microsoft: WindowsHPC, Cloud & VM
NVIDIA: Fermi GPUs, CUDA; Voltaire: QDR Infiniband Network
DDN: Large-Scale Storage; Intel: Westmere & Nehalem-EX CPUs
PGI: GPU Vectorizing Compiler
Tokyo Tech. GSIC: Base system design, GPU technologies, Green



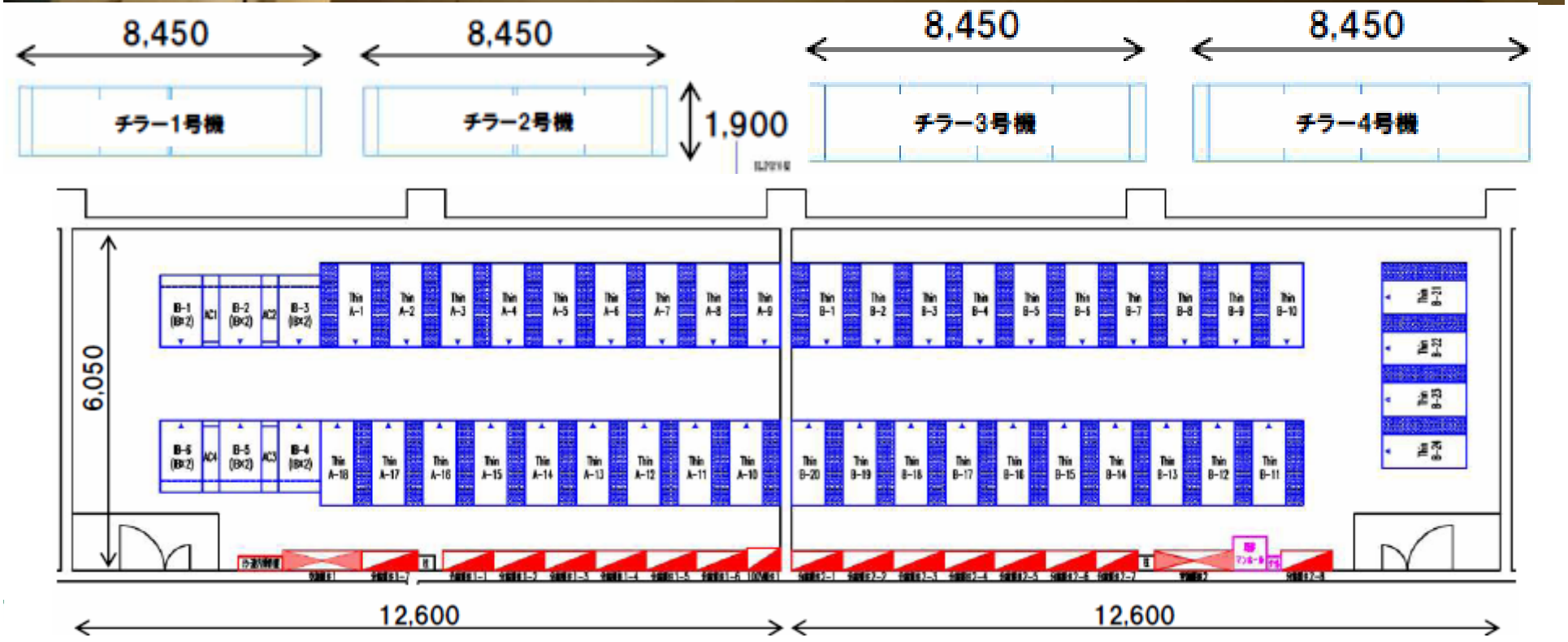
TSUBAME2.0 System Overview (2.4 Pflops/15PB)



	TSUBAME 1 (2006)	T2K Tokyo U (2008)	TACC Ranger (2008)	TSUBAME2.0 (2010)
Cores/Node	16	16	16	12(CPU)+1344(GPU)
Node Mem BW(GBytes/s)	20	20	20	64(CPU)+450(GPU)
Node Network BW (Gbps)	20	40	10	80
#Nodes	655	952	3,936	1408(Thin) + 34(Med/Fat)
#Cores (Total)	10,480(CPU)	15,232	62,976	17,664(CPU)+189万(GPU)
# GPUs/Accelerators	360 (ClearSpeed)	0	0	**4224 (Tesla M2050)**
Peak TFLOPS (DFP)	80	141	579	**2400**
Aggregate Mem BW (TB/s) (Flops/Byte)	17 (0.21)	20 (0.13)	80 (0.13)	**~720 (0.3)**
ネットワークバイセクション (Tbps)	6	41	80	**>200**
Memory (Tbytes)	21	30	126	100
Linpack (DFP-TFLOPS)	48	102	433	**>1000**
Aggregate 3D-FFT 256^3 (TFLOPS)	~13	~20	~80	**~700 (GPU only)**
HDD Storage (Raw TBytes)	1100	1500	1700	**7130**
Local SSD Storage/BW (Raw TBytes) (Bandwidth TByte/s)	0/0	0/0	0/0	**~200 (0.66PByte/s)**
Energy(Incl. Cooling)	850KW/Year	~1MW/Year	2.4MW Year	**~1MW/Year**



TSUBAME2.0 Layout (200m2 for main compute nodes)



TSUBAME 2.0 Compute Nodes Details

"Thin" nodes implement the "vector-scalar" hybrid high-bandwidth design combining NVIDIA "Fermi" Tesla GPUs + Intel Westmere CPU in a new, customly architected & designed high bandwidth nodes for Multi-GPU computing

Highly dense, efficient power & thermals, extremely reliable, extensive monitoring & mgmt

Thin Compute Nodes

IB QDR x2

NVIDIA M2050 (Fermi)
515GFLOPS/GPU
3GPU/node

Custom Designed with Hewlett Packard
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node) TB : 3.196GHz
Memory: 55.8GB(=52GiB) DDR3 1333MHz
103GB(=96GiB) DDR3 1333MHz
SSD : 60GB x2 (120GB/node) ※Memory 55.8GB nodes
120GB x2 (240GB/node) ※Memory 103GB nodes

1408nodes : 215.99TFlops ※Turbo boost

4224GPUs : 2175.36TFlops

Total : 2391.35TFLOPS

Memory : 80.55TB

SSD : 173.88TB

Medium Compute Nodes

IB QDR

PCI-e Gen2x16 x2
※for NVIDIA Tesla S1070 GPUs

HP 4-Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 32core/node
Memory: 137GB(=128GiB) DDR3 1066MHz
SSD : 120GB x4 (480GB/node)

24nodes:6.14TFlops

Total : 6.14TFlops

Fat Compute Nodes

IB QDR

PCI-e Gen2x16 x2
※for NVIDIA Tesla S1070 GPUs

HP 4-Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 (32core/node)
Memory: 274GB(=256GiB) DDR3 1066MHz
549GB(=512GiB) DDR3 1066MHz
SSD : 120GB x4 (480GB/node)

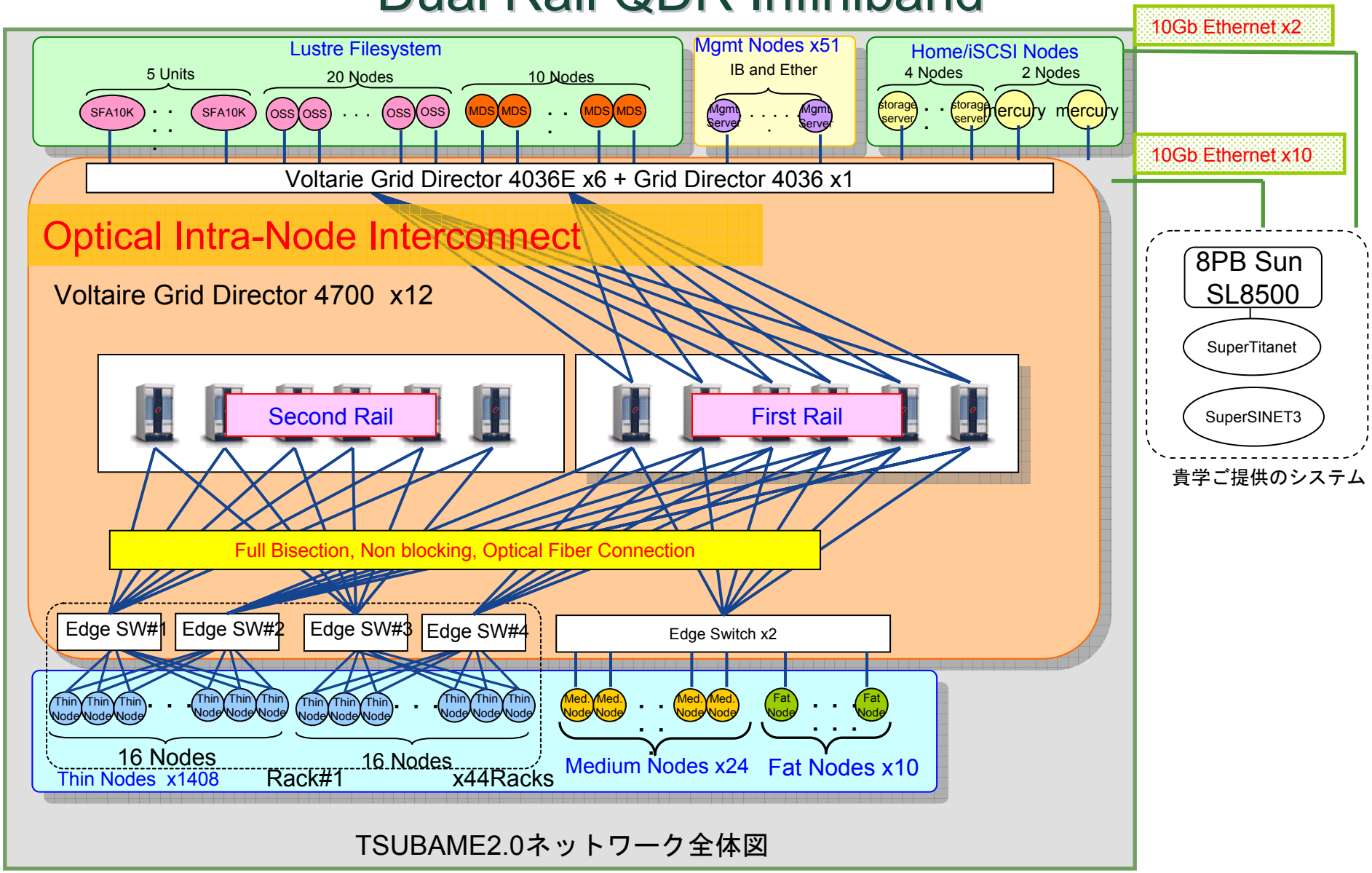
10nodes:2.56TFlops

Total: 2.56TFlops

CPU : 224.69TFlops
GPU : 2175.36TFlops

Total
2.4PFlops
(200TB SSD)

TSUBAME 2.0 Full Bisection Fat Tree, Optical, Dual Rail QDR Infiniband



TSUBAME2.0ネットワーク全体図

TSUBAME2.0 Storage

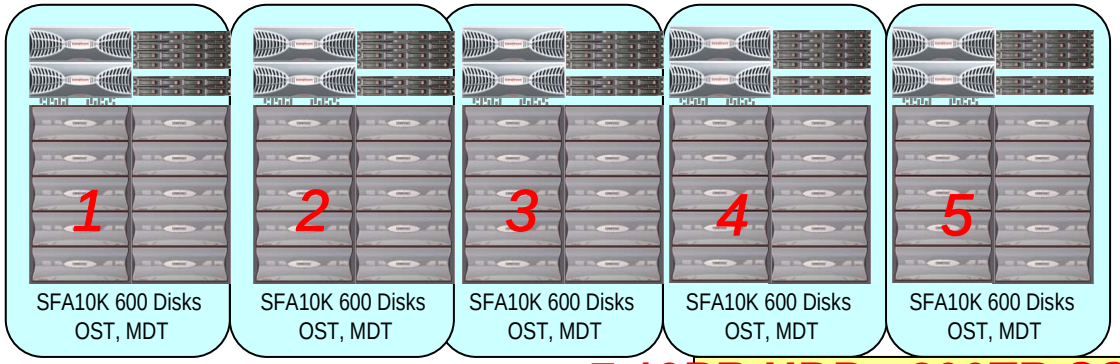
Multi-Petabyte storage consisting of Luster Parallel Filesystem Partition
and NFS/CIFS/iSCSI Home Partition + Node SSD Acceleration

Lustre Parallel Filesystem Partition, 5.93PB

MDS：HP DL360 G6 x10
- -CPU:Intel Westmere-EP x2 socket(12 Cores)
- -Memory：51GB(=48GiB)
- -IB HCA:IB 4X QDR PCI-e G2 x1port

OSS：HP DL360 G6 x20
- -CPU:Intel Westmere-EP x2 socket(12 Cores)
- -Memory：25GB(=24GiB)
- -IB HCA：IB 4X QDR PCI-e G2 x2port

Storage：DDN SFA10000 x5
- -Total Capacity：5.93PB
  2TB SATA x 2950 Disks + 600GB SAS x 50 Disks

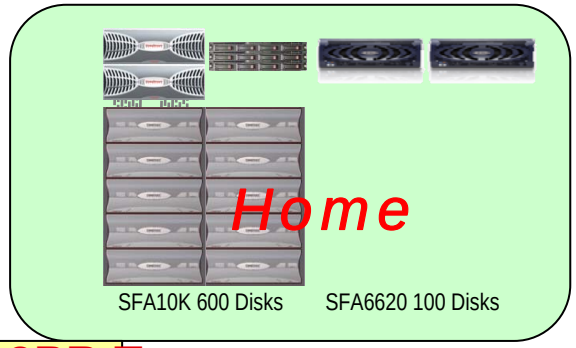


Home Partition 1.2PB

NFS/CIFS：HP DL380 G6　x4
- -CPU:Intel Westmere-EP x2 socket(12 Cores)
- -Memory：51GB(=48GiB)
- -IB HCA:IB 4X QDR PCI-e G2 x2port

NFS/CIFS/iSCSI：BlueArc Mercury100 x2
- -10GbE x2

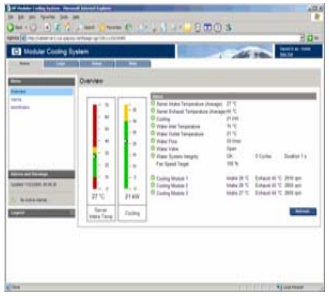
ストレージ：DDN SFA10000 x1
- -Total Capacity：1.2PB
  2TB SATA x 600 Disks



7.13PB HDD +200TB SSD + 8PB Tape

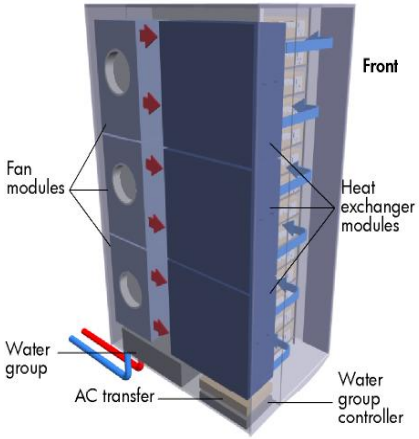
HP Modular Cooling System G2

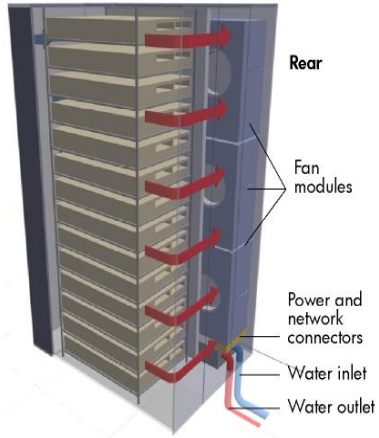
- HP's water-cooled rack solution.
- Completely closed racks with their own heat exchanger.
- 1.5 x width of normal rack+rear extension
- Cooling for high density deployments
- 35kW of cooling capacity in a single rack
- 17.5kW of cooing when using two racks.
- up to 2000 lbs of IT equipment
- Uniform air flow across the front of the servers
- Automatic door opening mechanism controlling both racks
- Adjustable temperature set point
- Removes 95% to 97% of heat inside racks
- Polycarbonate front door reduces ambient noise considerably



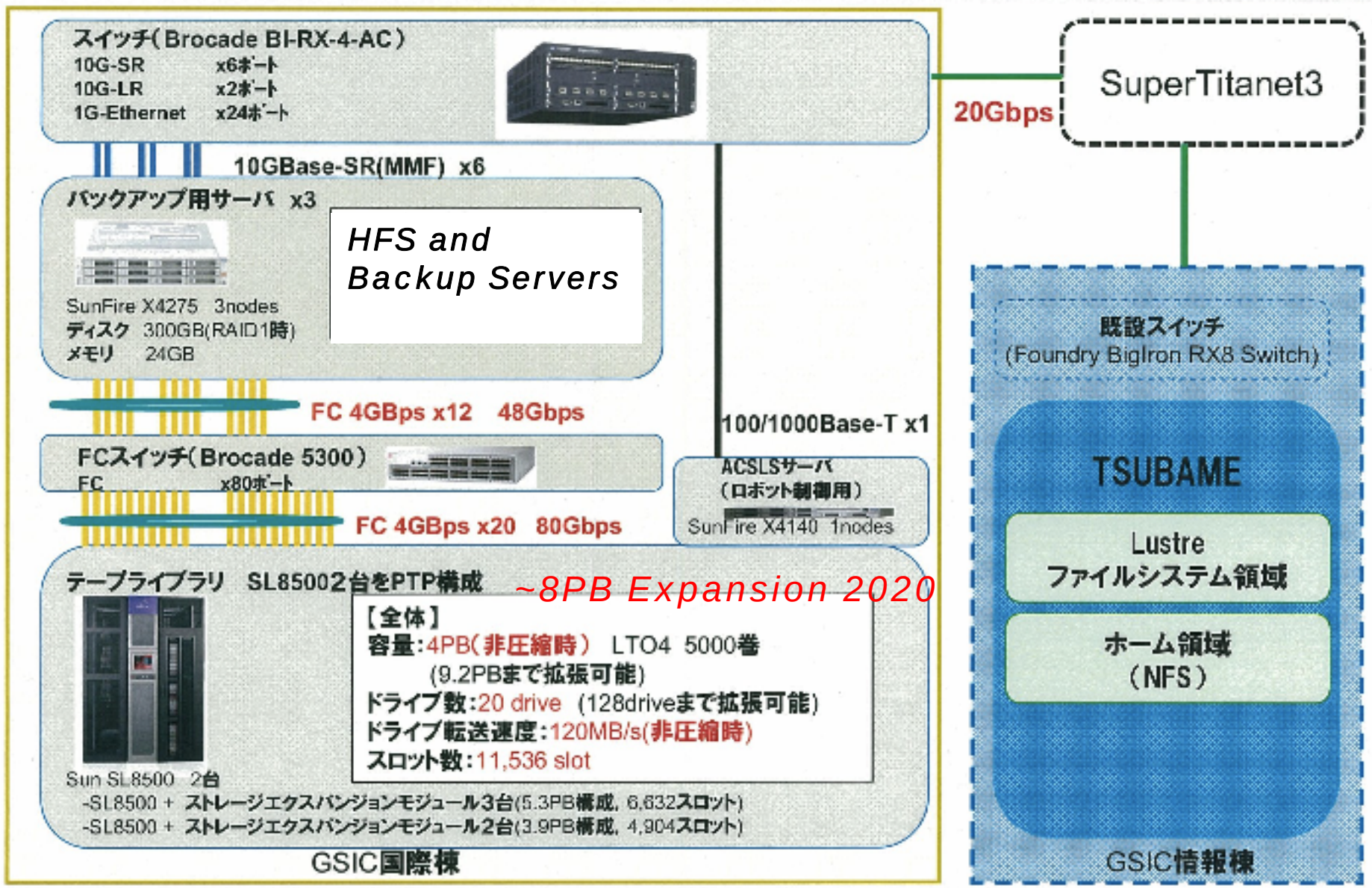
Web GUI

A summary of all monitored values on one page





TSUBAME2.0 Tape System
HFS of over15PB and Scalable



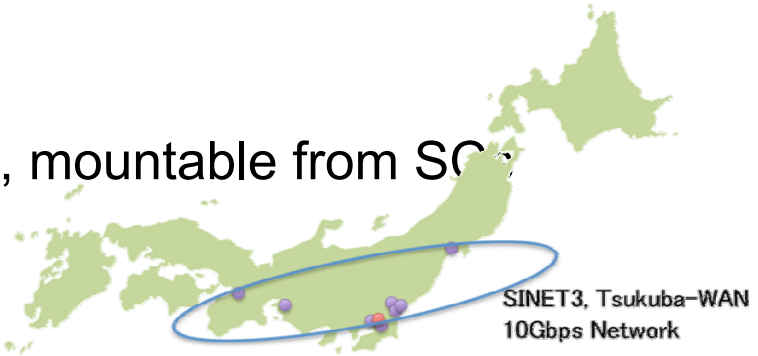
RENKEI-POP Distributed Storage

- Objective: SC Centers-Wide Distributed Storage on SINET National NW
  - RENKEI (MEXT e-Science)Project and R&D and nationwide deployment of RENKEI-PoPs(Point of Presence)
  - Data transfer server engine with large capacity and fast I/O
  - “RENKEI-Cloud” on SINET3 with various Grid/Cloud Services including Gfarm distributed filesystem

CPU	Core i7 975 Extreme (3.33 GHz)
Memory	12GB (DDR3 PC3-10600 , 2GB*6)
NIC	10GbE (without TCP/IP Offload Engine)
System Disk	500GB HDD
SSD RAID	30TB (RAID 5, 2TB HDD x 16)

Tokyo Inst. Tech.	Osaka Univ.
National Inst. Inform.	KEK
Nagoya Univ.	Univ. Tsukuba
AIST	Tohoku Univ.

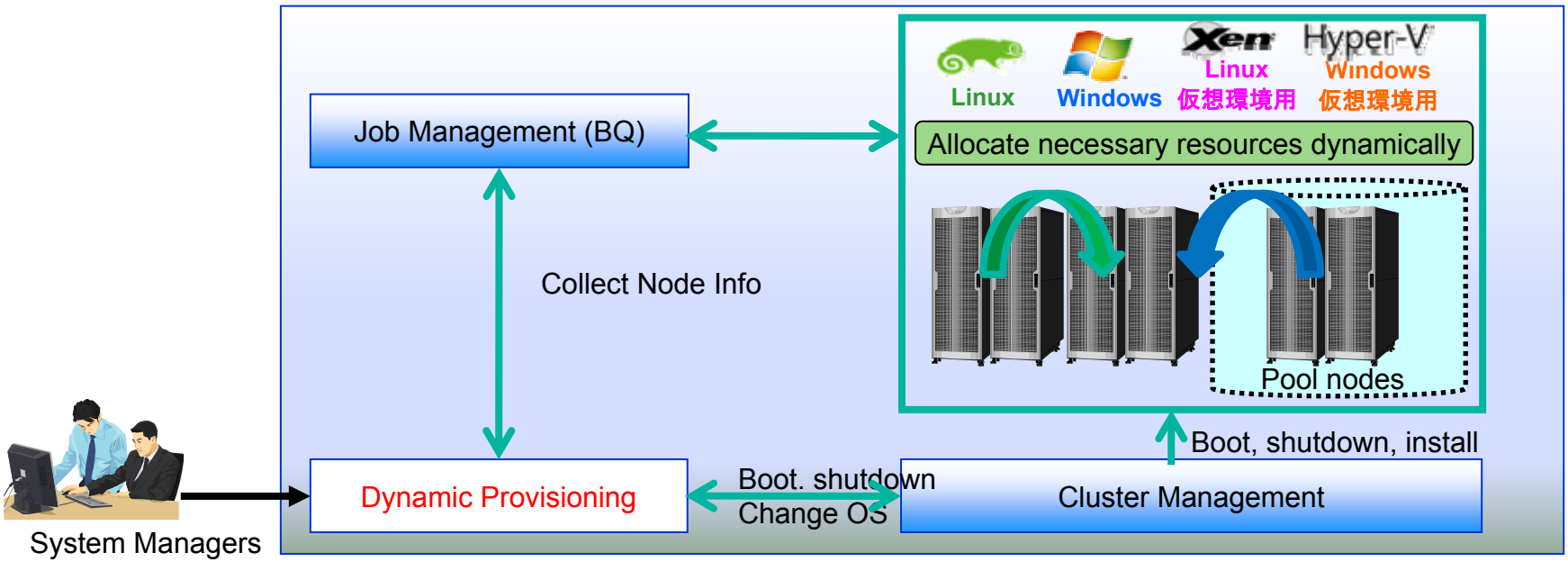
e, mountable from SC



Cloud Provisioning Supercomputing Nodes

Tsubame 2.0 will have a dynamic provisioning system that works with batch queue systems and cluster management systems to dynamically provision compute resources

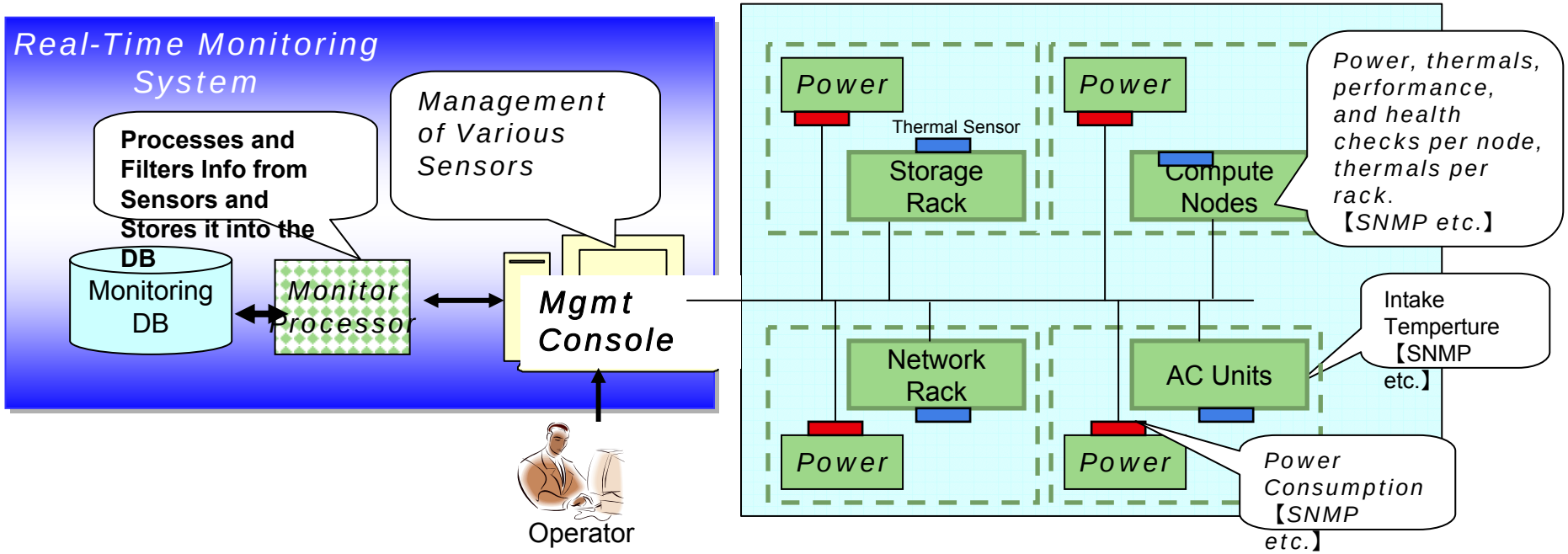
- Pool nodes can be dynamicaly allocated to various supercomputing services
- Linux and Windows batch queue systems coordinate to manage the nodes
- Dynamic increase/decrease of nodes allocated to particular services
  ※Increse / Decrease of Windows HPC and Linux nodes of different configurations
- Virtual Nodes on VMs on respective OSes can be subject to dynamically allocated and scheduled.



Green SC: Environmental Monitoring

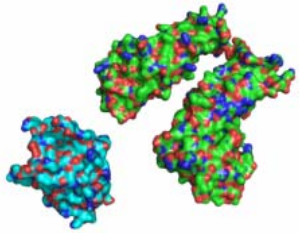
Real-time monitoring and processing of every node, rack, and the room

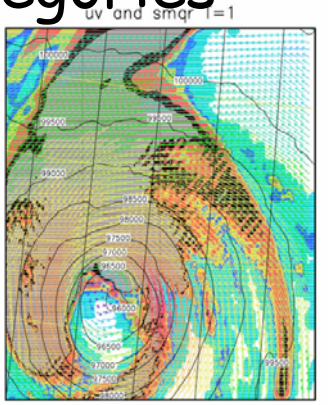
- Thermals
- Power Consumption
- HW and SW Health checks



TSUBAME2.0 Estimated Performances

- ~1.4 PFlops Linpack [IEEE IPDPS 2010]
- ~0.5 PF 3D Protein Rigid Docking (Node 3-D FFT) [SC08, SC09]
- ~150 TFlops ASUCA Forecast [SC10]
- Top-level HPC-Challenge Performances
  - Work with Jack's group on GPU-Specific Categories
- Lattice-Boltzmann, Various FEM, Genomics, MD/MO (w/IMS), …
- CS-Apps: search, optimization, …





(Faster Than) Real-time Tsunami Simulation

(Prof. Takayuki Aoki, Tokyo Tech.)

ADPC : Asian Disaster Preparedness Center

Early Warning System:

Data Based Extrapolation → (Faster than) Real-time CFD

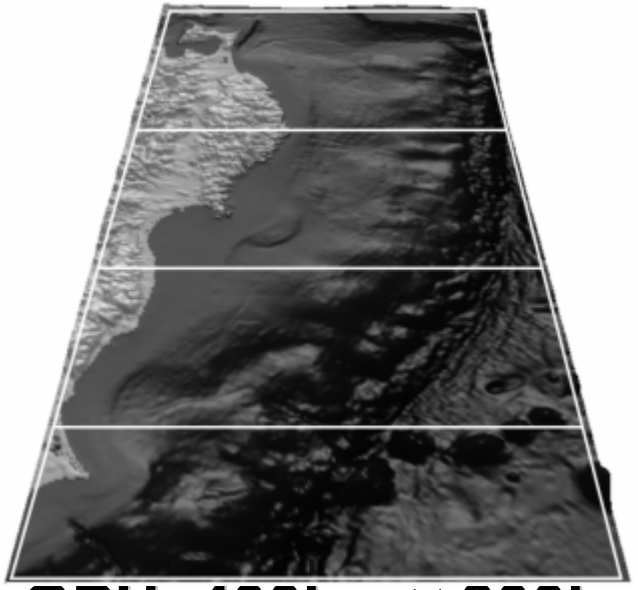
high accuracy

Shallow-Water Equation

Conservative Form:

Assuming hydrostatic balance in the vertical direction,

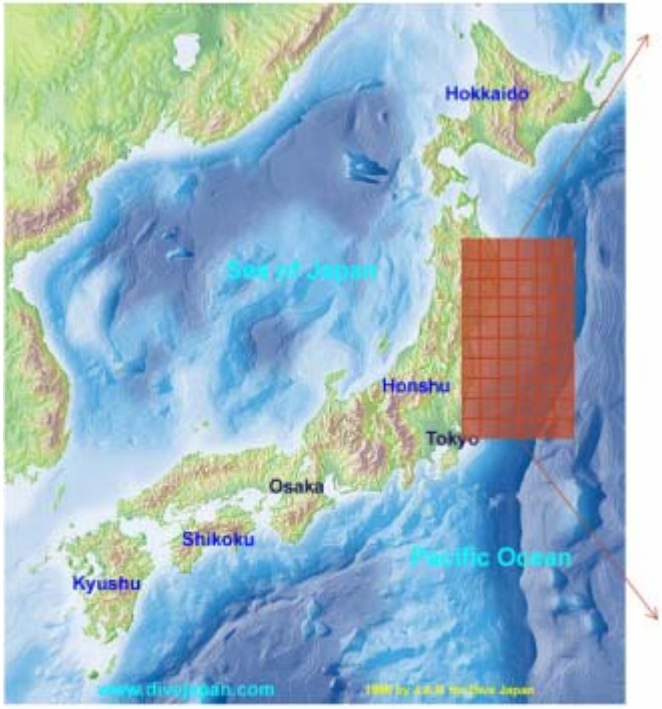
3D → 2D equation



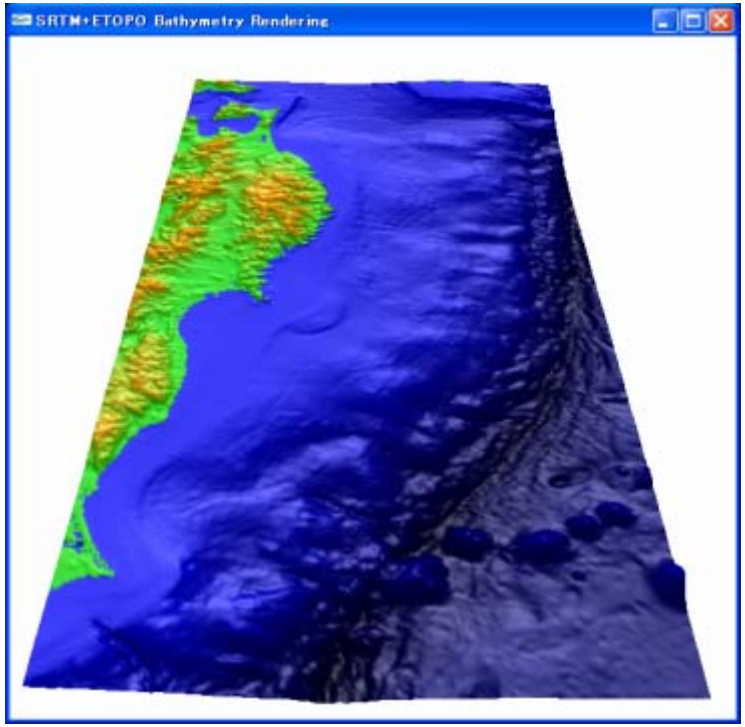
8 GPU 400km × 800km
(100m mesh)

Tsunami Prediction of Northern Japan Pacific Coast

- Bathymetry



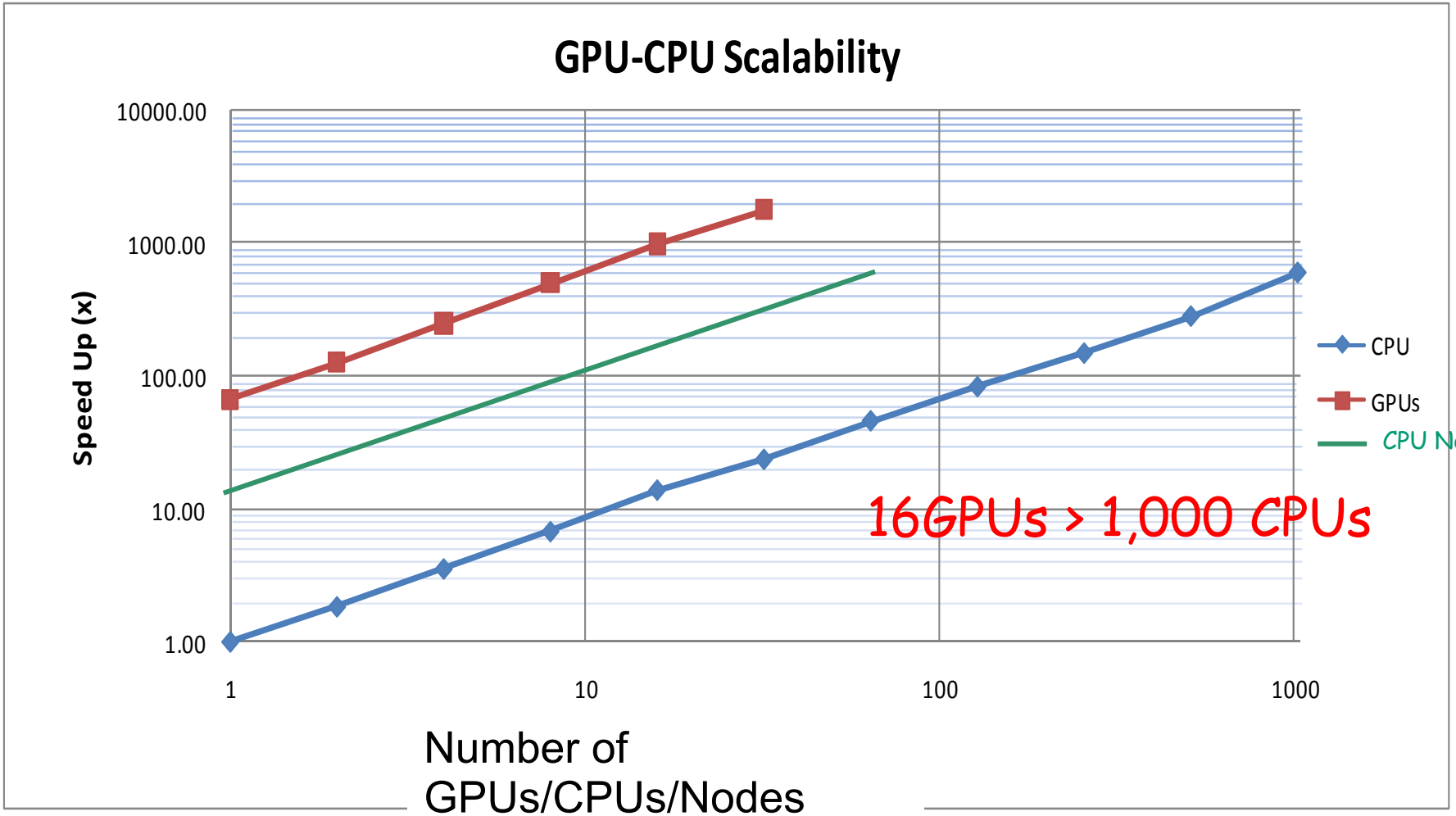
- Grid Size
  4096x8192
- Latitude
  $N35° - N42°$
- Longitude
  $E140°30' - E144°18'$
- Length
  370km x 740km

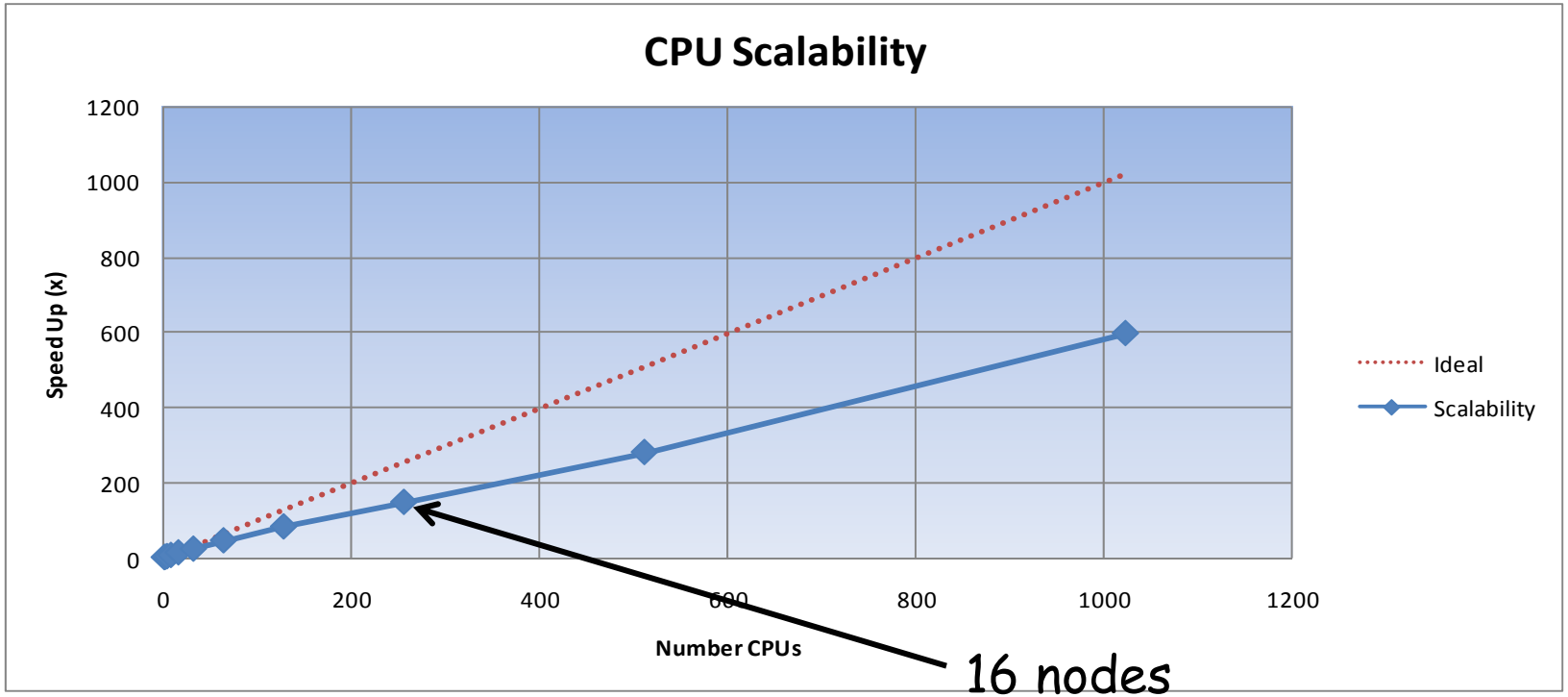


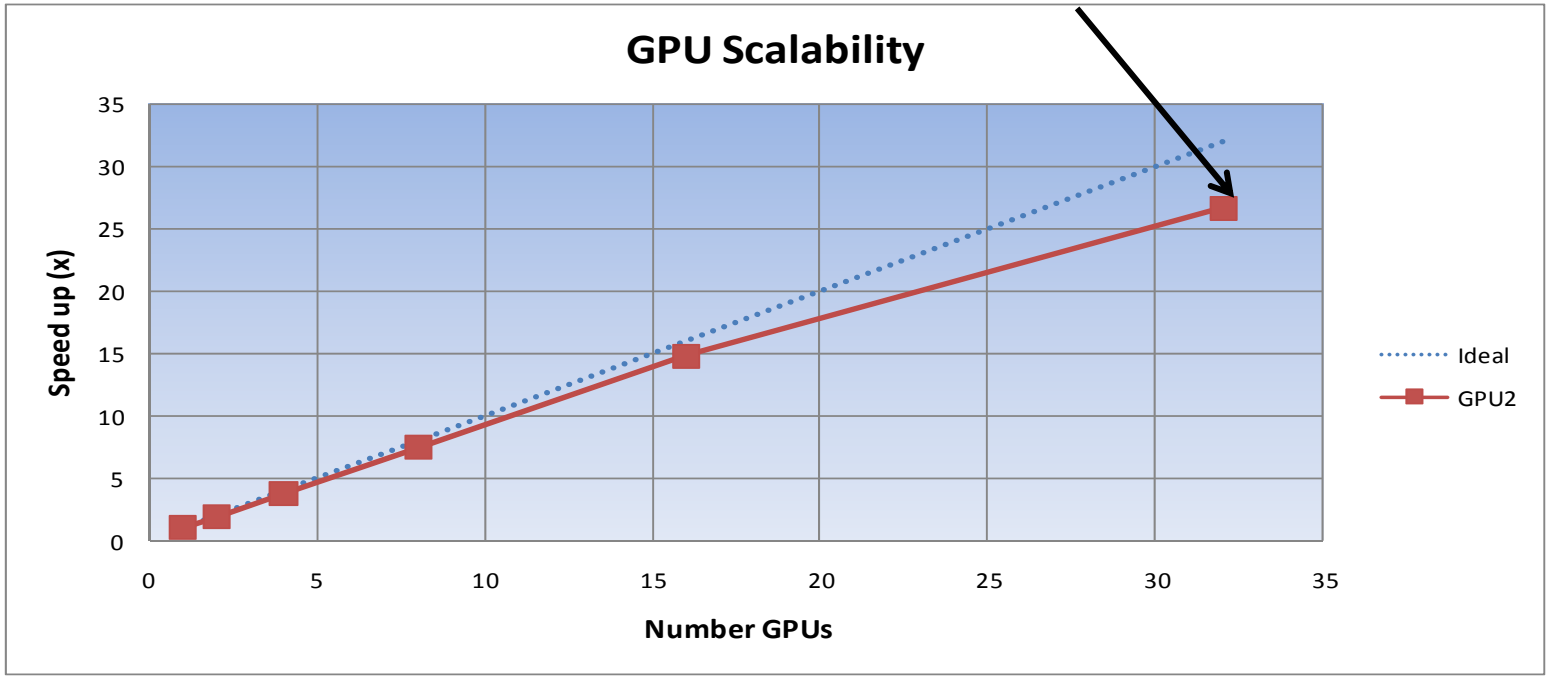
Multi-node CPU/GPU Comparison

*** Strong Scaling ***

- Results on TSUBAME1.2




CPU Scalability
Speed Up (x)
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
Number CPUs
Ideal
Scalability
16 nodes

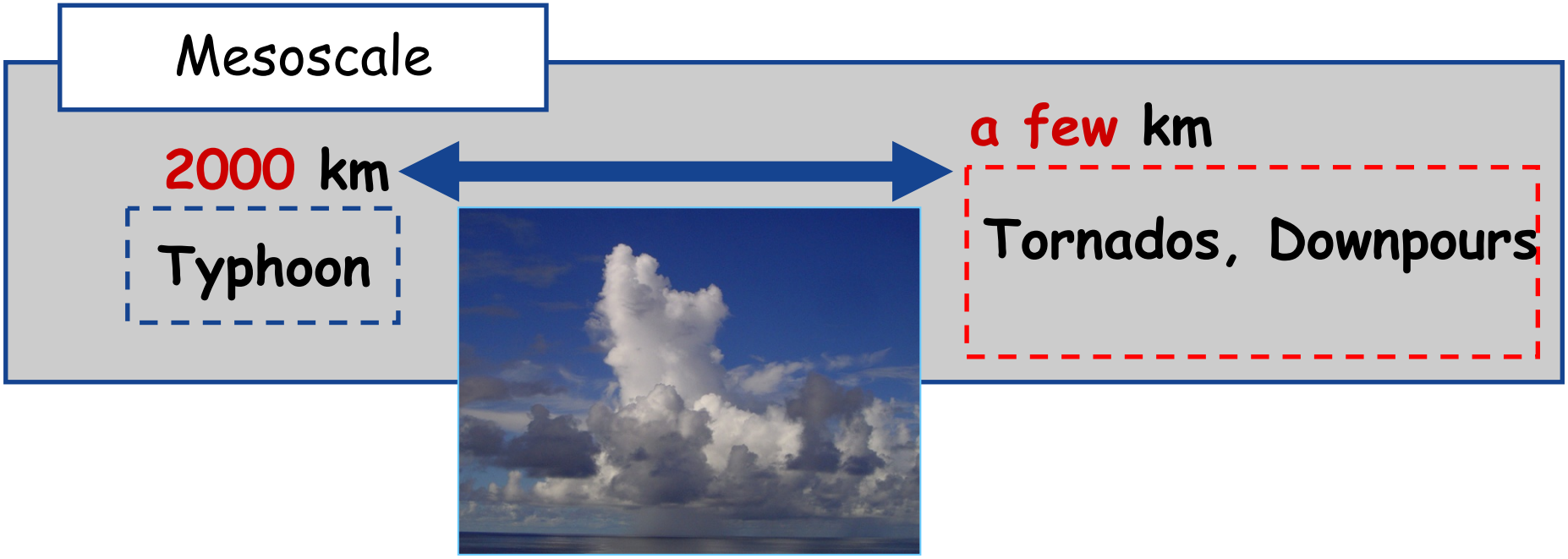

GPU Scalability
Speed up (x)
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35
Number GPUs
Ideal
GPU2

Next Gen Weather Forecast

Mesoscale Atmospheric Model:

Cloud Resolution: 3-D non-static

Compressible equation taking consideration of sound waves.



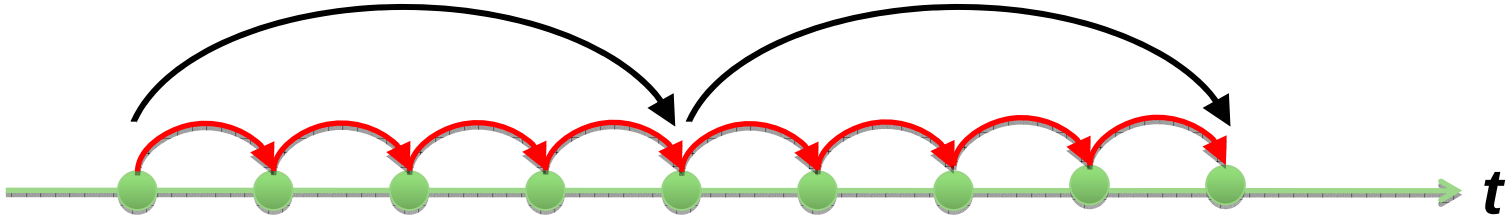
GPU enabling ASUCA [SC10]

- **ASUCA : Next Generation Production Weather Forcast Code (by Japan's National Meteorlogical Agency)**

  Mesoscale production code for real weather forcast

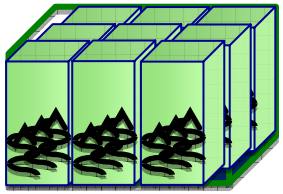
  Very similar to NCAR's WRF

  Time-splitting method: long time step for flow



u, v (~ 100 m/s), w (~ 10 m/s) << sound velocity (~300/ms)

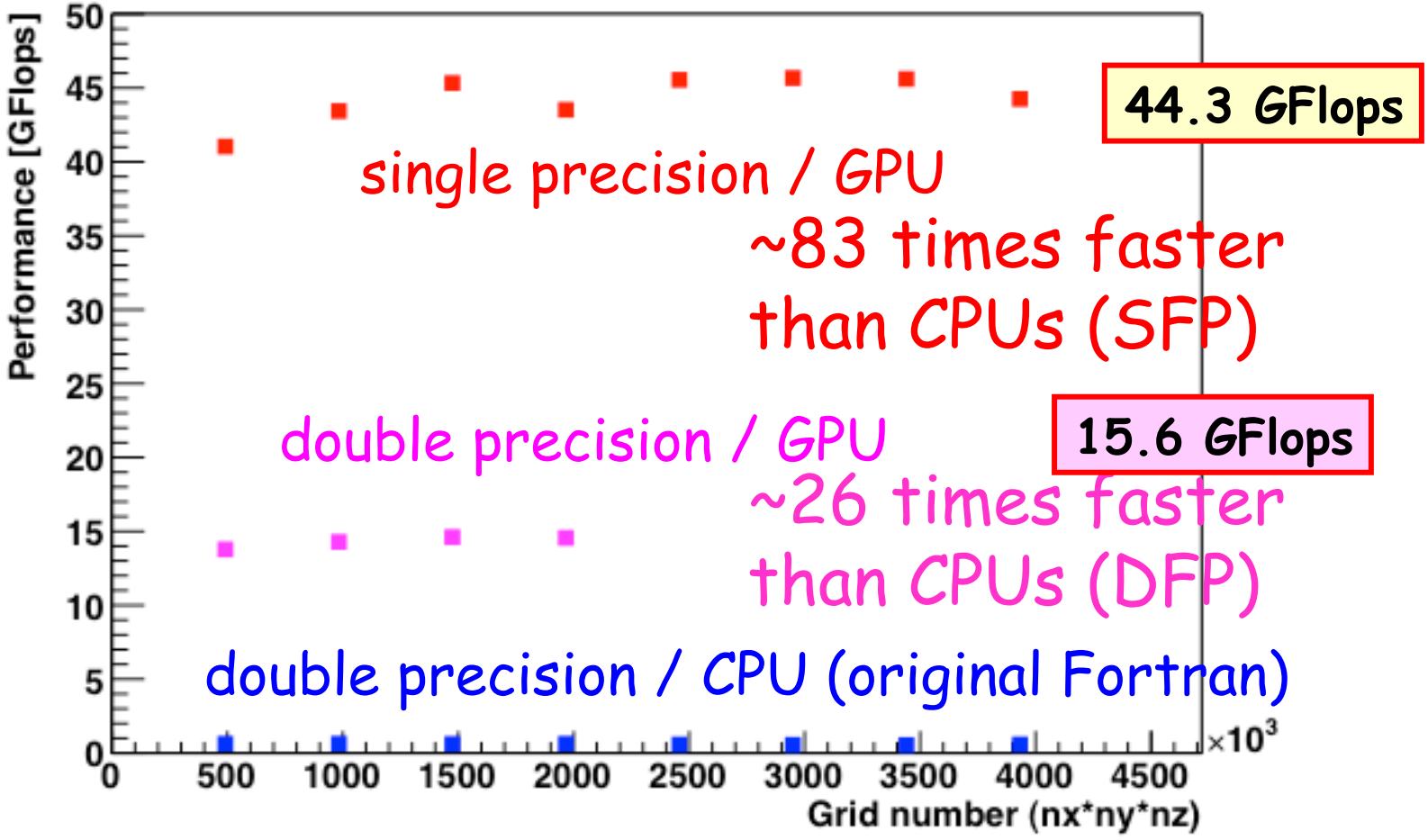
HEVI (Horizontally explicit Vertical implicit) scheme
Horizontal resolution ~ 1 km
Vertical resolution ~ 100 m



1-D Helmholtz equation (like Poisson eq.) **sequential process**

Entire "Core" of ASUCA now ported to GPU (~30,000 lines)
By Prof. Aoki Takayuki's team at Tokyo Tech.

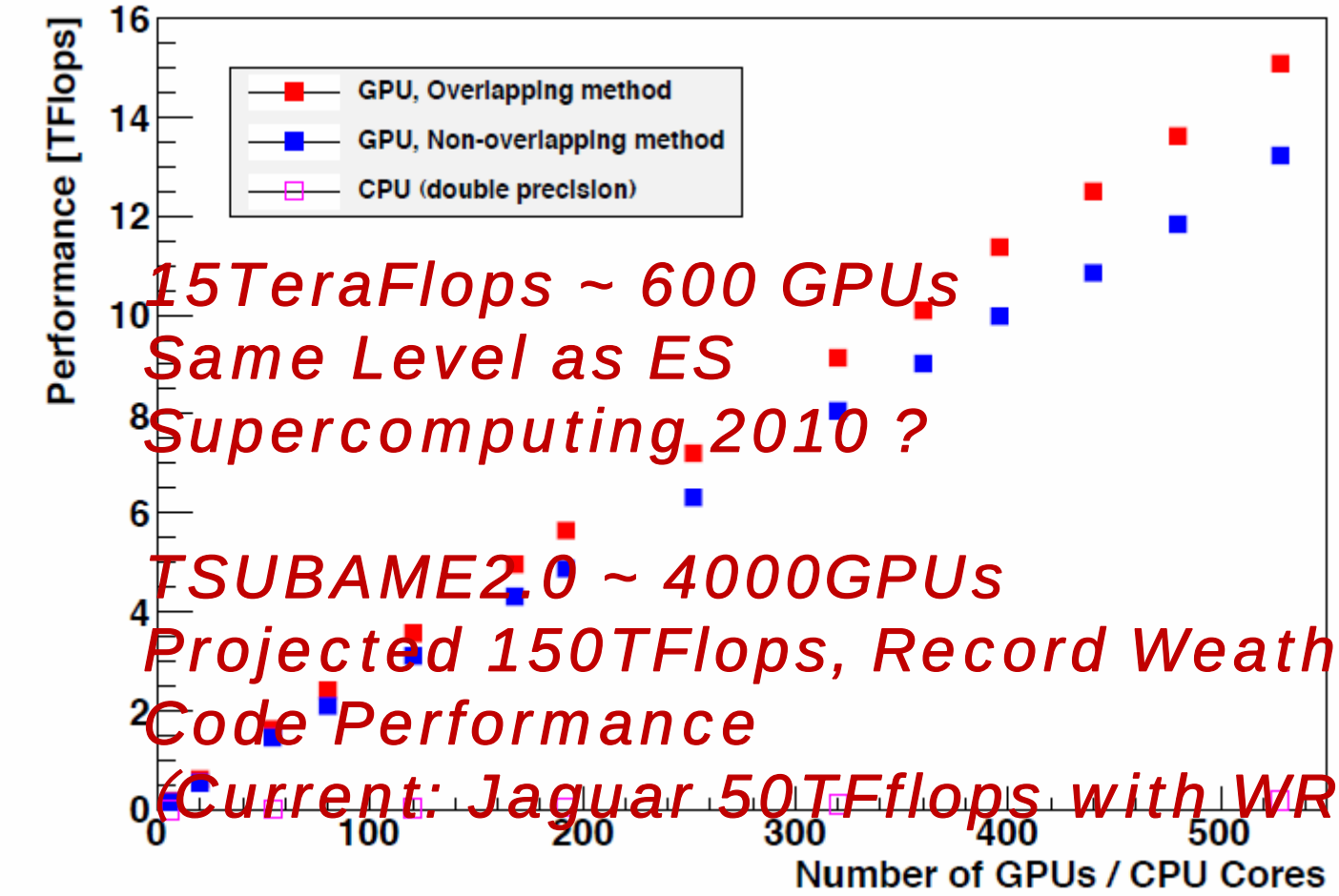
ASUCA Single GPU Perf. (TSUBAME1.2. Tesla s1070)



Mountain Wave Test
NVIDIA Tesla S1070 card

320 x 256 x 64

ASUCA Multi GPU Performance (TSUBAME1.2)

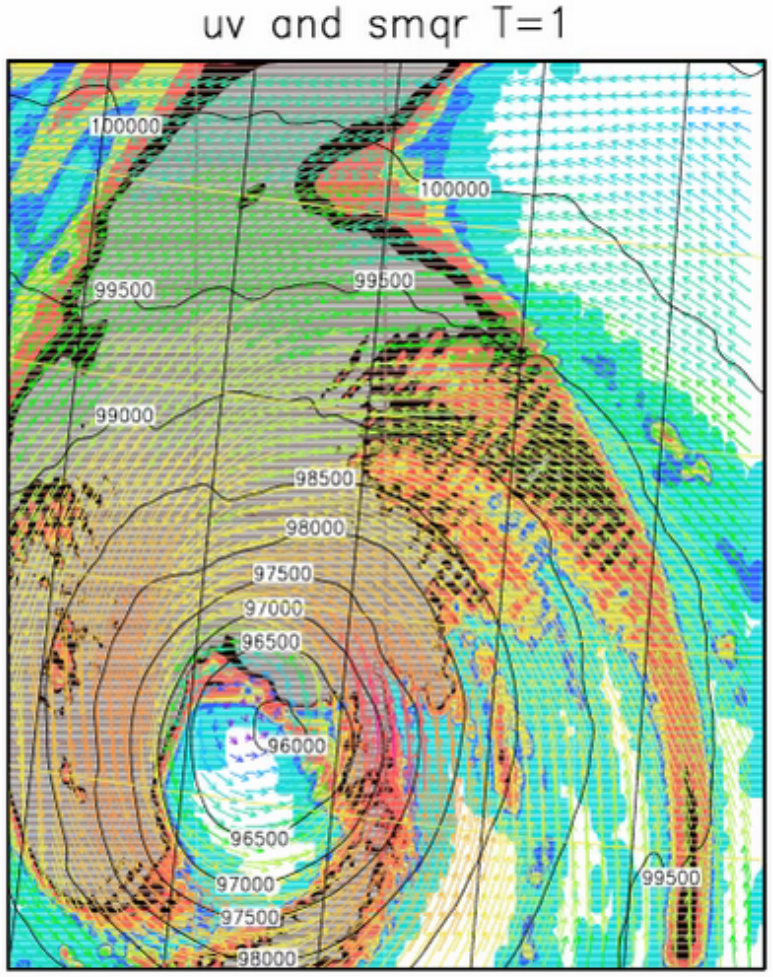


15TeraFlops ~ 600 GPUs
Same Level as ES
Supercomputing 2010 ?

TSUBAME2.0 ~ 4000GPUs
Projected 150TFlops, Record Weather Code Performance
(Current: Jaguar 50TFflops with WRF)

Mountain Wave Test in single precision
NVIDIA Tesla S1070 on TSUBAME

ASUCA Typhoon Simulation 2km mesh 3164×3028×48

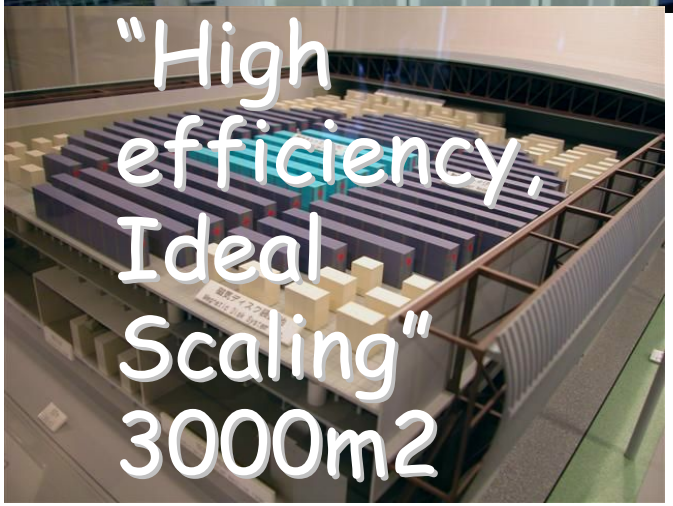


70 minutes wallclock time for 6 hour simulation time (x5 faster)

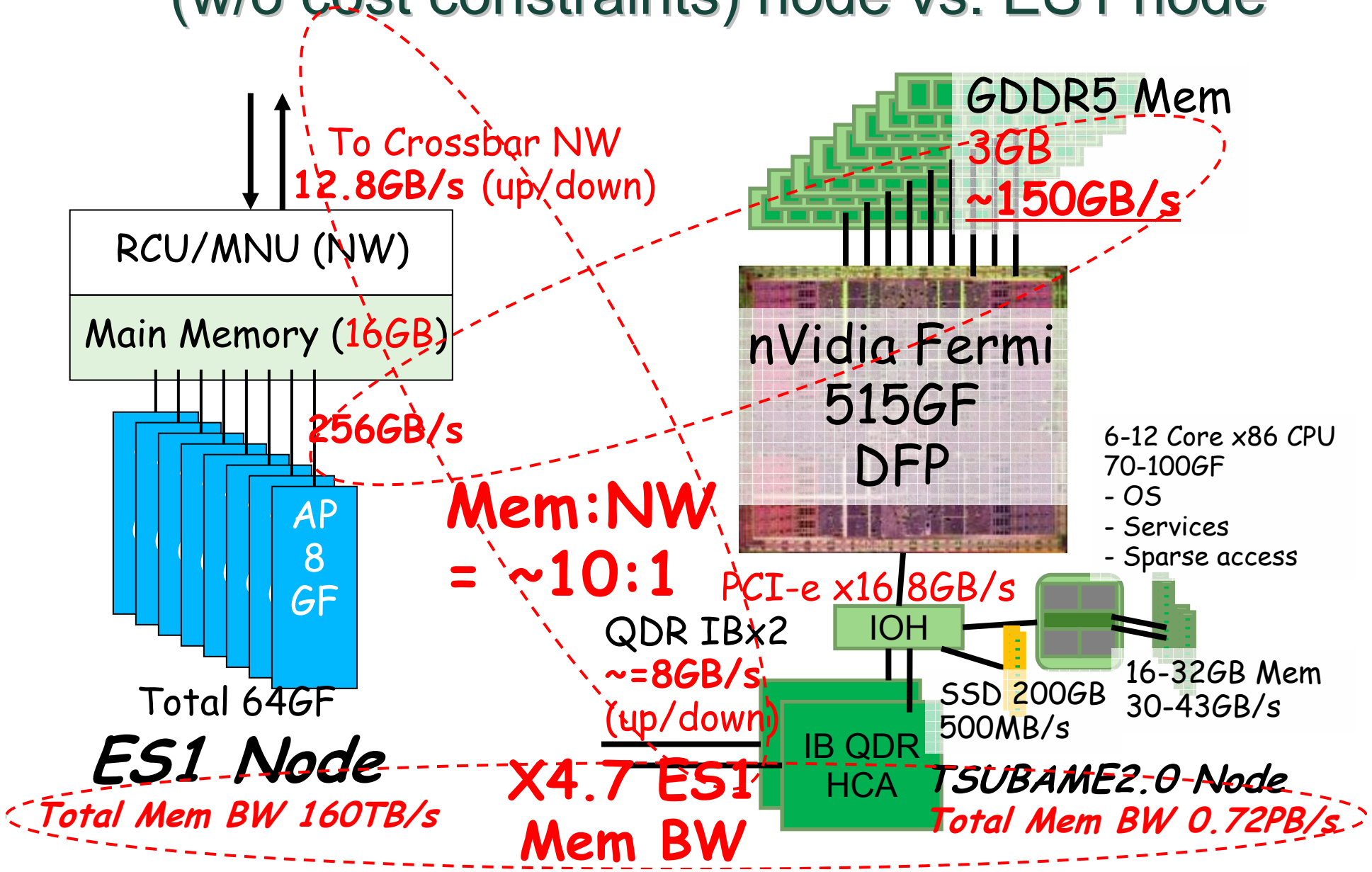
Expect real-time with 0.5km mesh (for Clouds) on TSUBAME2.0

TSUBAME2.0 (2010) vs. Earth Simulator1 (ES) 2002) vs. Japanese 10PF NexGen @Kobe (2012)

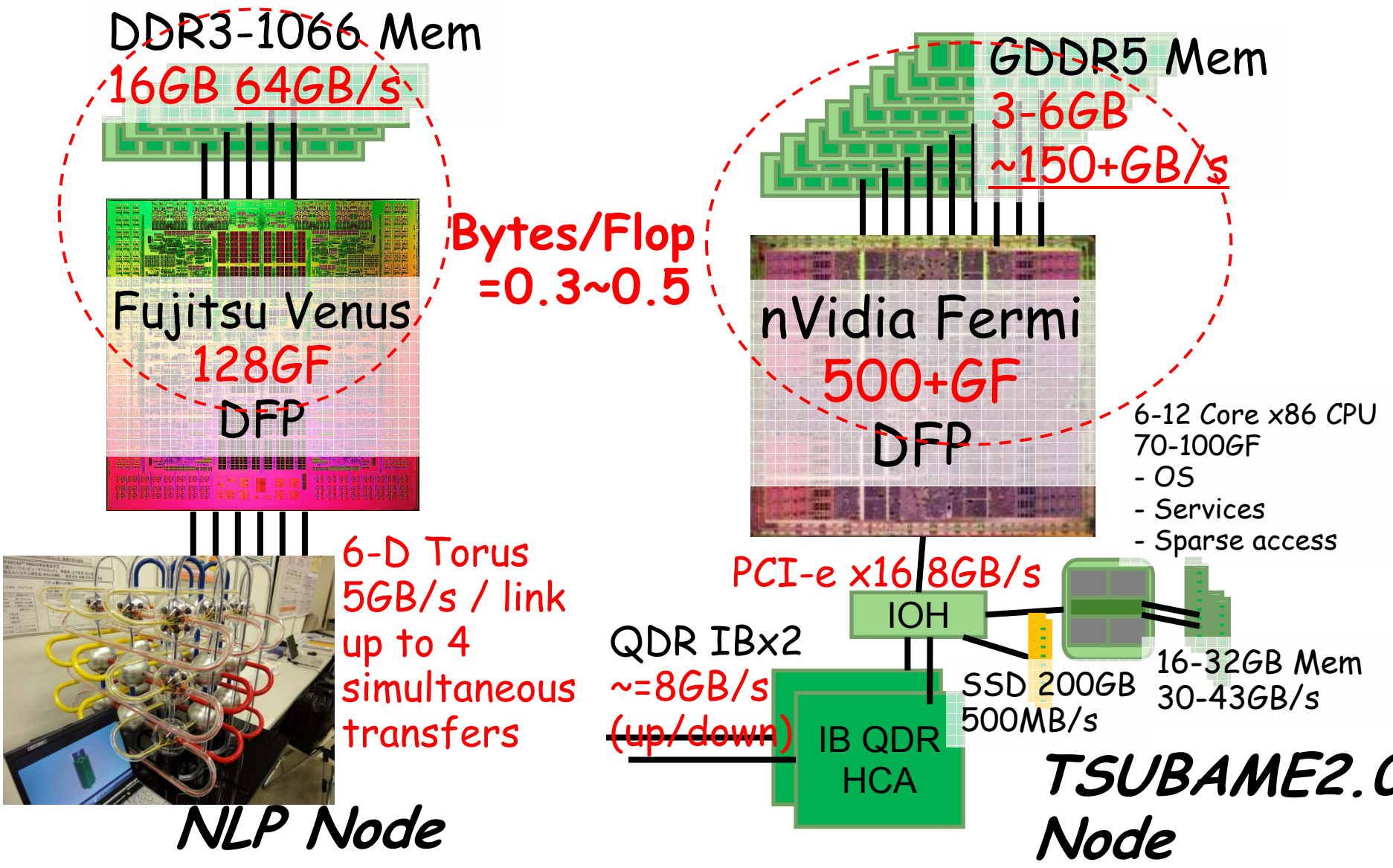
200m2



The "IDEAL TSUBAME2.0" (w/o cost constraints) node vs. ES1 node

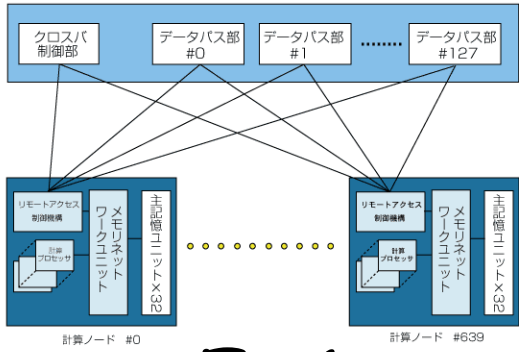


The "IDEAL TSUBAME2.0" node vs. 10PF NLP Node (2012)



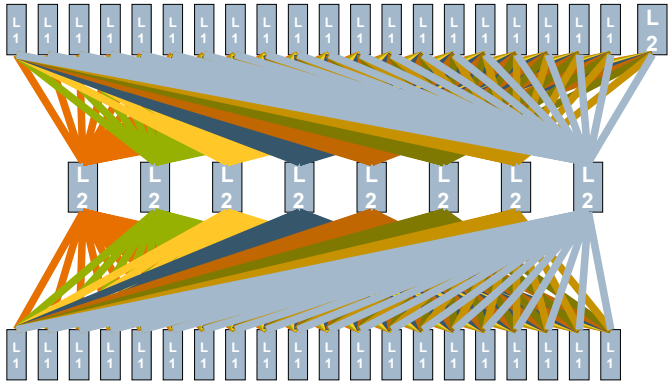
Comparing the Networks

結合ネットワーク(IN)部



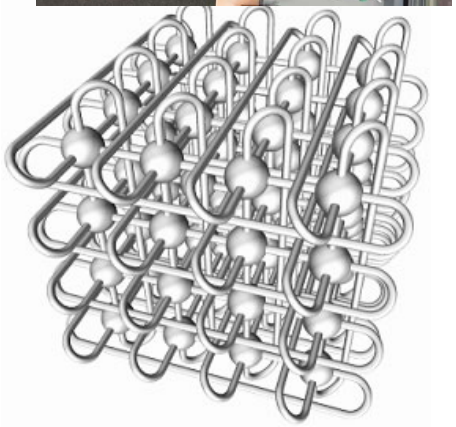
ES1

12.8GB/s Link
5us latency
Full Crossbar
~8TB/s
Bisection BW



Ideal
TSUBAME2.0

(4+4)GB/s Link
2us latency
Full Bisection Fat Tree
~60TB/s Bisection BW



10PF NLP

5GB/s Link
?us latency
6-D Torus
~30TB/s?
Bisection BW

Summary of Comparisons

- (1) ES1 vs. Ideal TSUBAME2.0
  - Similar (Mem BW : Network BW), full bisection NW
  - ES1 ∑BW : TSUBAME2 ∑BW = 1 : 6

  ⇒ **BW-bound apps (e.g. CFD) should scale equally on both w.r.t. ∑BW (TSUBAME2.0 6 times faster), Other apps *drastically* faster on TSUBAME2.0**
- (2) 10PF NLP vs. Ideal TSUBAME2.0
  - Similar Memory Bytes/Flop (0.3~0.5)
  - NLP x2 superior on Mem BW : Network BW
  - TSUBAME2.0 x2 better on Bisection BW?

  ⇒ **Most apps similar efficiency and (strong) scalability NLP ~4 times faster on full machine (weak scaling)**

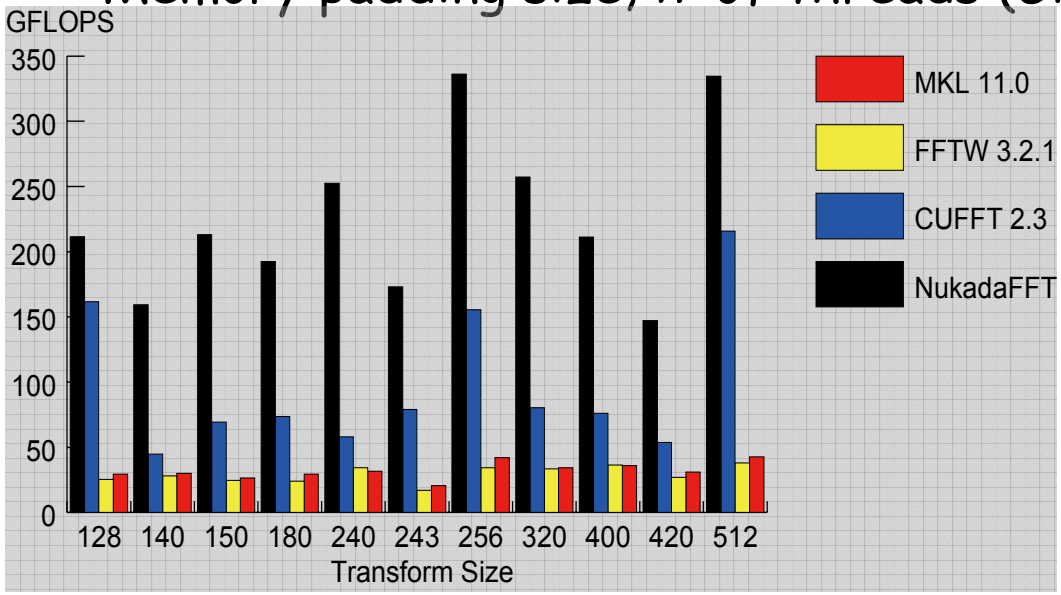
But it is not just HW Design…SOFTWARE(!)

- The ***entire software stack*** must preserve bandwidth, hide latency, lower power, heighten reliability for Petascaling
- Example: TSUBAME1.2, Inter-node GPU ⇔ GPU achieves only 100-200MB/s in real apps
  - c.f. 1-2GB/s HW dual rail IB capability
  - Overhead due to unpinnable buffer memory?
  - New Mellanox driver will partially resolve this?
  - Still need programming model for GPU ⇔ GPU
- Our SW research as CUDA CoE (and other projects such as Ultra Low Power HPC)

Auto-Tuning FFT for CUDA GPUs [ACM/IEEE SC09]

HPC libraries should support GPU variations:

- # of SPs/SMs, Core/memory clock speed,
- Device/shared memory size, Property of memory bank, etc…
- Auto-tuning of various parameters!
- Selection of kernel radices (FFT-specific)
- Memory padding size, # of threads (GPU-specific)

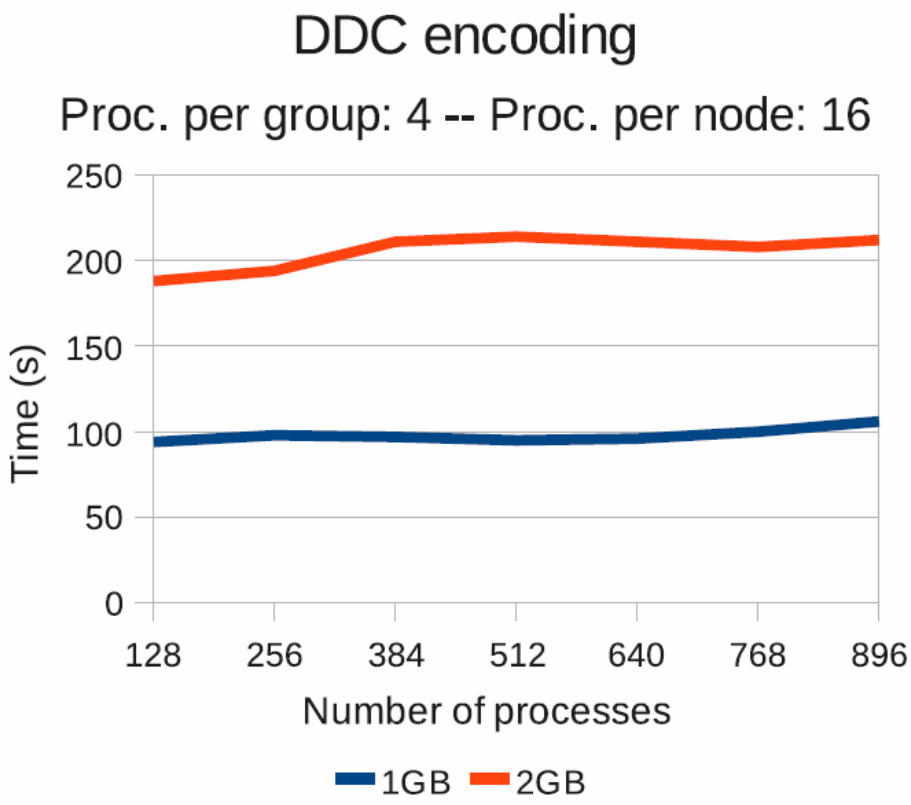


Our auto-tuned library is

- x4 power efficient than libraries on CPUs
- x1.2 -- 4 faster than vendor's library
- RELEASE RSN!

Distributed Diskless Checkpoint for Large Scale Systems (to 100,00s of nodes + GPU) [IEEE CCGrid 2010]

- *Global I/O for FT will not scale and major Energy/BW waste > TBs for peta~exascale syste*
- Exploit fast *local* I/O w/NVMs (aggr. TB/s) for checkpoints
- Group-based redunduncy + RS encoding technique for *O(1) scalability*
- Combine with our ***ongoing work on checkpointing on GPUs***



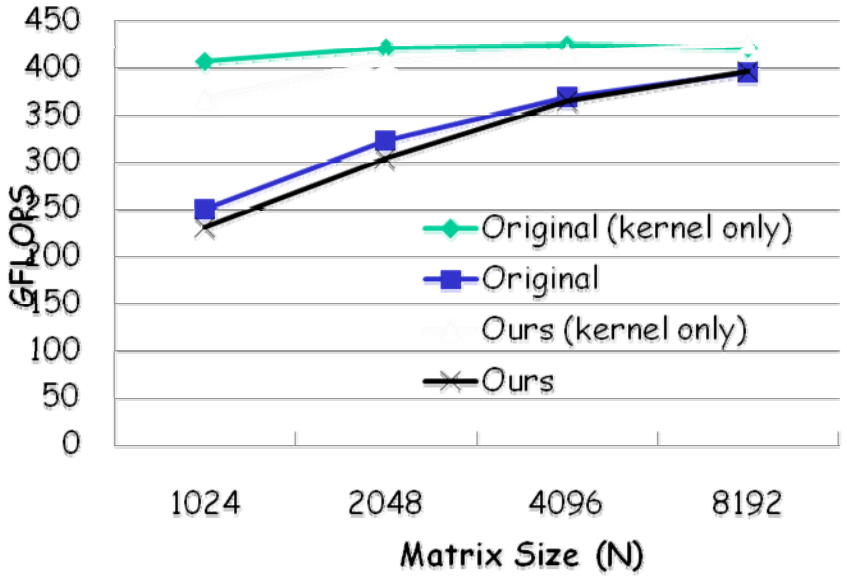
Software Framework for GPGPU Memory FT [IEEE IPDPS 2010]

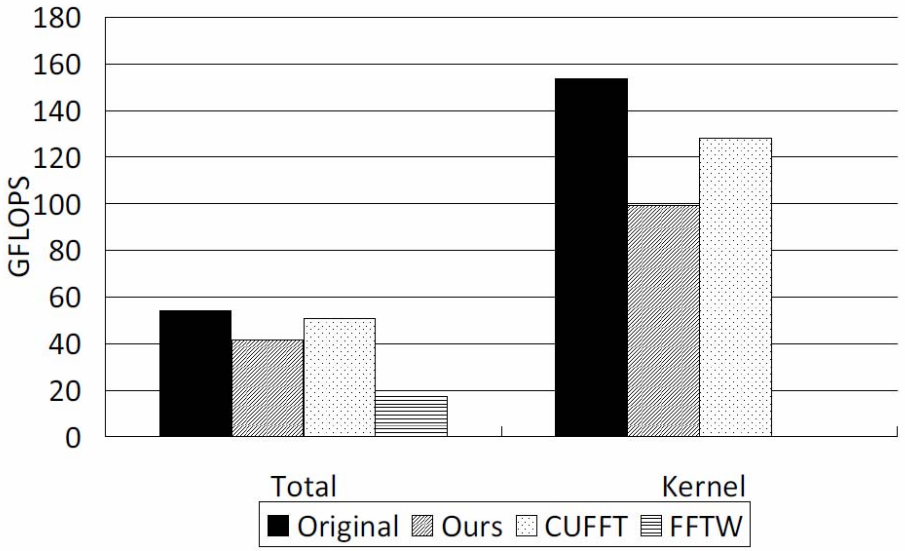
- Error detection in CUDA global memory + Checkpoint/Restart
- Works with existing NVIDIA CUDA GPUs

Lightweight Error Detection

- *Cross Parity* for 128B blocks of data
  - Detects a single-bit error in a 4B word
  - Detects a two-bit error in a 128B block
- No on-the-fly correction → Rollback upon error

Exploit the latency hiding capability of GPUs for data correctness





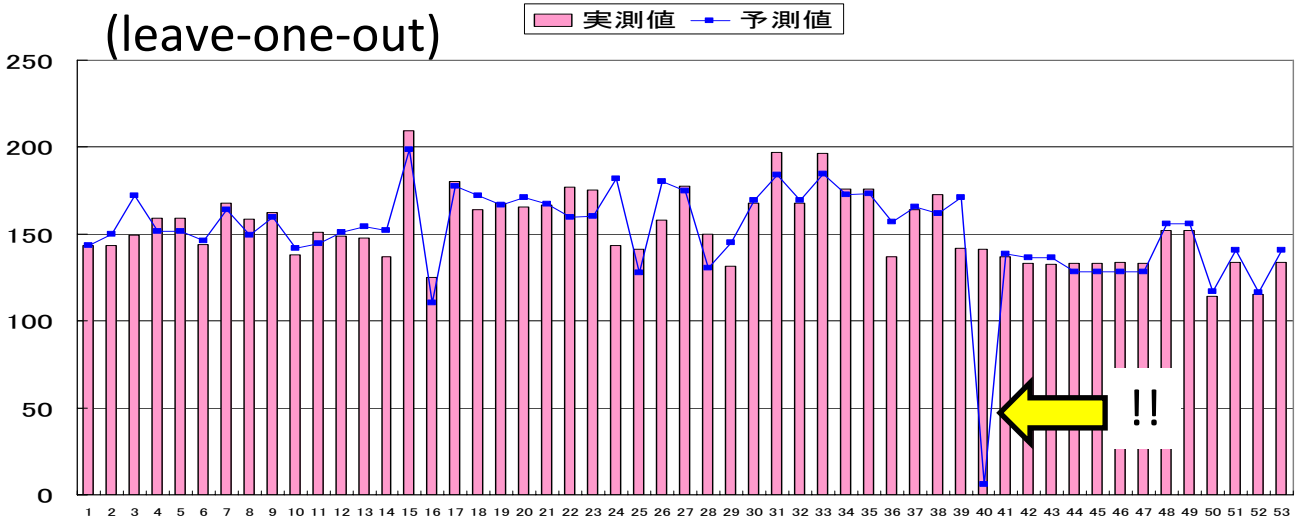
GPU Power Consumption Modeling

Employ learning theory to predict GPU power from performance counters

$$\text{GPU power} = \sum c_i \times p_i$$

p_i: GPU perf. Counter (12 types)、c_i: learning constant

54 GPU kernels: prediction vs. measurement (leave-one-out)



Average error 7%

- Do need to compensate for extraneously erroneous kernel (!!)

TODO: Fewer couters for real-time prediction
higher-precition, non-linear modeling

Low Power Scheduling in GPU Clusters

[IEEE IPDPS-HPPAC 091

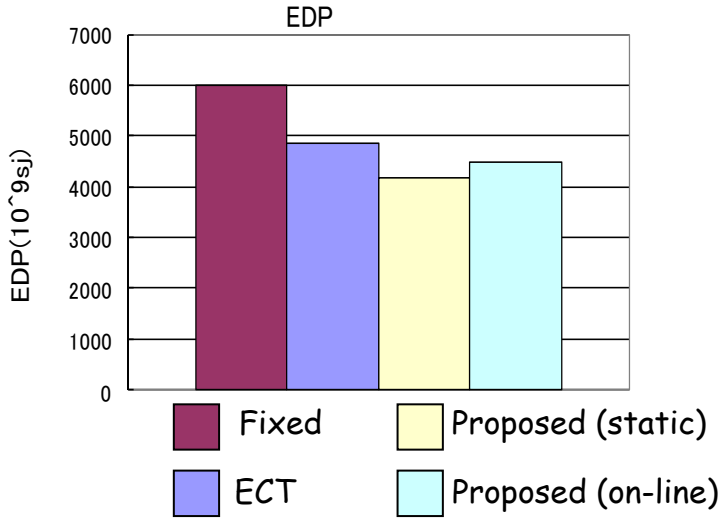
Objective: Optimize CPU/GPU Heterogeneous

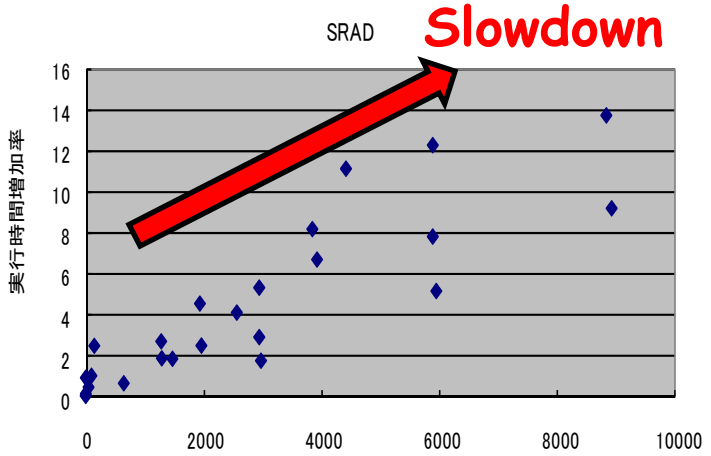
- Optimally schedule mixed sets of jobs executable on either CPU or GPU but w/different performance & power
- Assume GPU accel. factor (%) known

  30% Improvement Energy-Delay Product

TODO: More realistic environment

- Different app. power profile
- PCI bus vs. memory conflict
  - GPU applications slow down by 10 % or more when co-scheduled with memory-intensive CPU app.



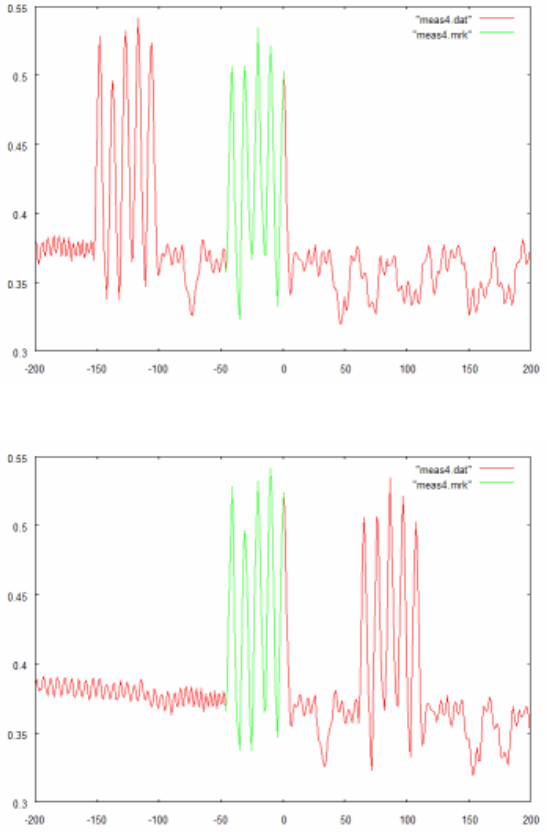


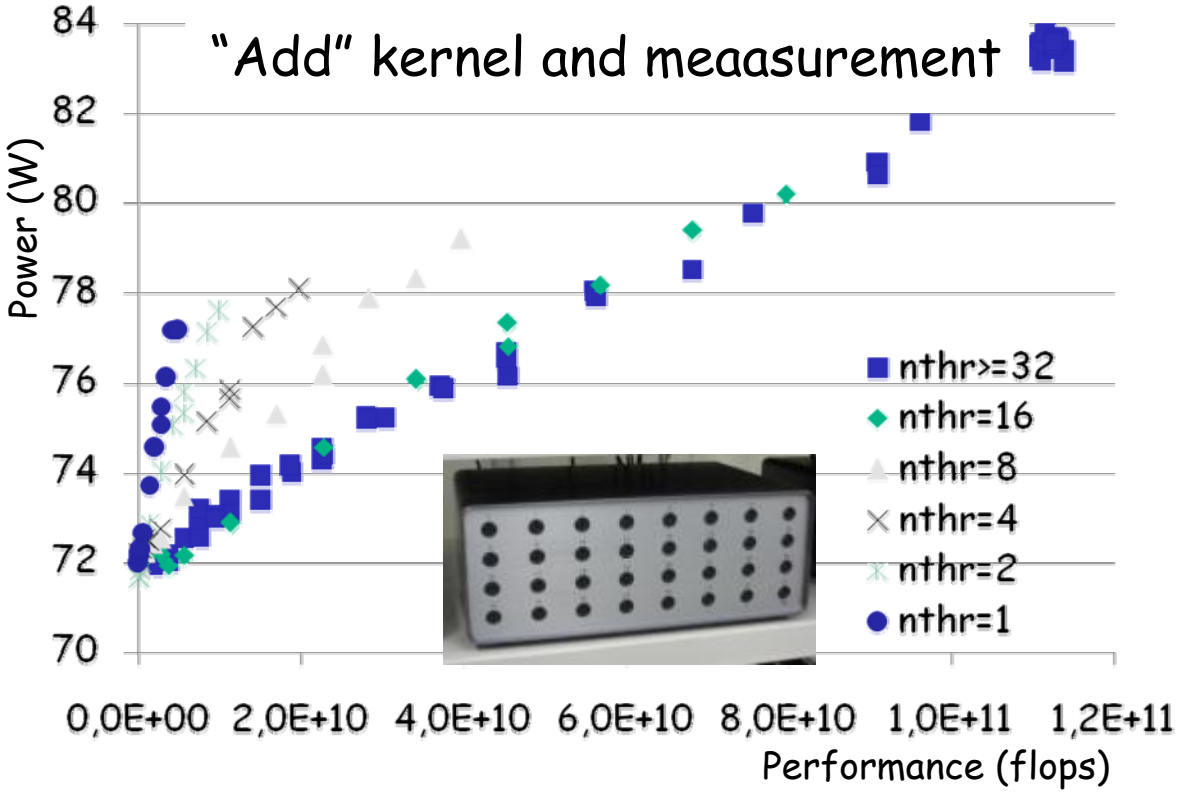
Background CPU process BW (MB/s)

High Precision GPU Power Measurement (Reiji Suda, U-Tokyo, ULPHPC)

- "Marker" = Compute + Data transfer
- Automatically detect Markers
- Sample 1000s execution in 10s seconds

マーカーの自動検出例

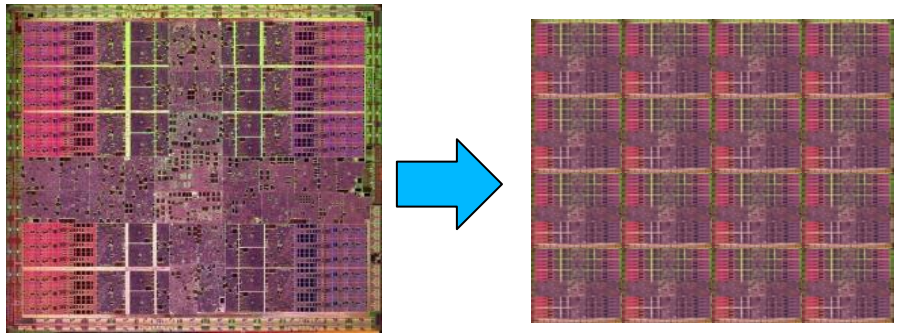




Scaling TSUBAME2.0 to Exaflop

- TSUBAME2.0: 40-45nm, 2~3PF, ~60 racks, 1.5MW = x320 scaling?
- x20 physical scaling now (1000 racks, 30MW)
  - >3000-4000m^2, 1000 tons
- x16 semiconductor feature scaling 2016~2017

2008	2009	2010	2011	2012	2014	2016-17
45nm	40nm	32nm	28nm	22nm	15nm	11nm



45nm

11nm, x16 transistors & FLOPs

Other innovations such as 3-D memory / flash packaging, optical chip-chip connect, multi-rail optical interconnect etc.

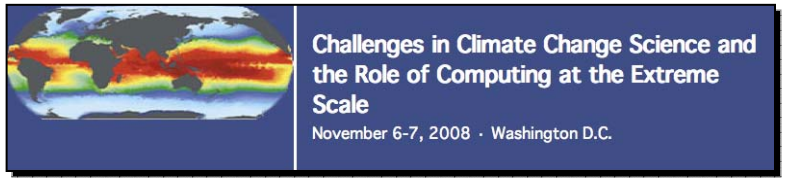
But what about the network? 3-40,000 nodes?

(From US DoE Exascale PPT by Rick Stevens@ANL) Uncertainty quantification is critical and requires exascale resources.

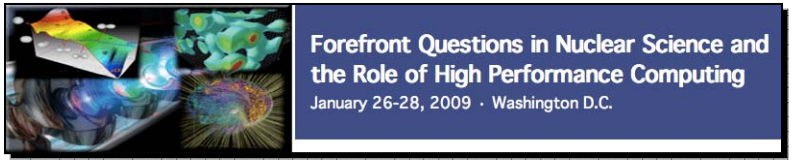
EXASCALE RESOURCES

Large ensembles — ***Embedded UQ***

Large ensembles	Embedded UQ
Response surface	Adjoint enabled forward models
Posterior exploration	Data extraction from model
Finding least favorable priors	Local approximations, filtering
Bounds on functionals	Stochastic error estimation



"We need to be able to make quantitative statements about the predictability of regional climatic variables that are of use to society."



"computational techniques and needs complement the scientific areas that will be pursued with extreme scale computing. Examples include … verification and validation issues for extreme scale computations "

"scientists must create new suites of application codes, Integrated Performance and Safety Codes (IPSCs) that incorporate …integrated uncertainty quantification.."

Japanese 10 PF Facility @ Kobe, Japan

Construction: started in March, 2008 and will complete in May, 2010, Machine operation late 2011 ~ early 2012

Computer Wing

Total Floor Area:17,500m^2
2 Computer rooms: 6,300m^2 each
4 Floors (1 underground floor)
=> x4 ES, 2000~4000 racks

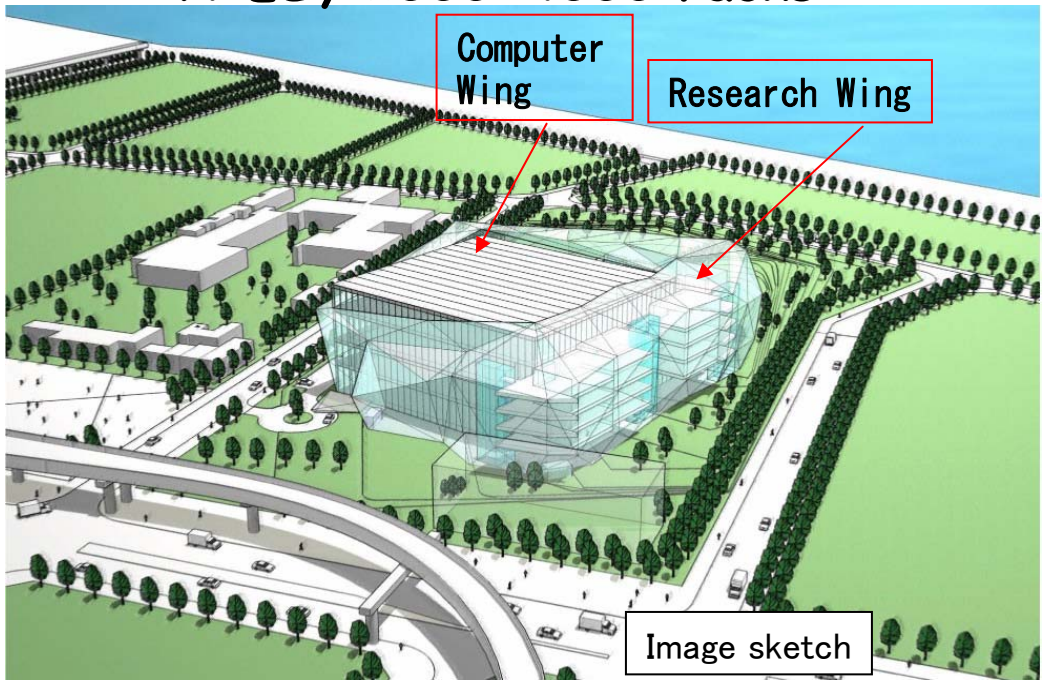


Image sketch

Research Wing

Total floor area: 8,500m^2
Area of 1floor: 1,900m^2
7 floors (1 underground floor)
(cafeteria is planned on the 6th floor)

Other Facilities

Co-generation System
Water chiller system
Electric Subsystem

Initially 30MW capability@2011