GPUは専用アクセラレータではない−ポストペタスケールへのスケーリングへの本質的な変容

東京工業大学 学術国際情報センター
教授
松岡　聡

理研シンポジウム
2011年02月16日

「アクセラレータ」の定義

- 高速計算用専用ハードウェア
  - ▶少量生産・特殊マーケット・高価格
  - ▶ソフトウェアの世代間の継承の困難さ
- アプリケーション分野の限定
  - ▶特殊な計算処理のみ高速
  - ▶「嵌らない」計算はできないか、CPUよりかなり遅い
- 一般的なプログラムの不動作
  - ▶特殊なプログラミング・言語等
  - ▶ポインタ、リカージョン、構造体などの制限
  - ▶OS等、システムソフトウェアは動かない・連動しない

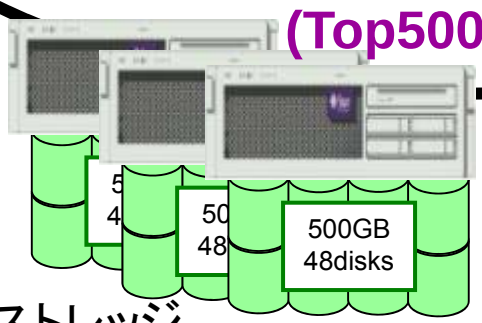
2006年4月東工大 "TSUBAME1.0"
日本一の「みんなのスパコン」

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

10Gbps+外部
ネットワーク

Sun/AMD高性能計算クラスタ
(Opteron Dual core 8-Way)
10480core/655ノード
50.4TeraFlops
OS(現状) Linux
(検討中) Solaris, Windows
NAREGIグリッドミドル

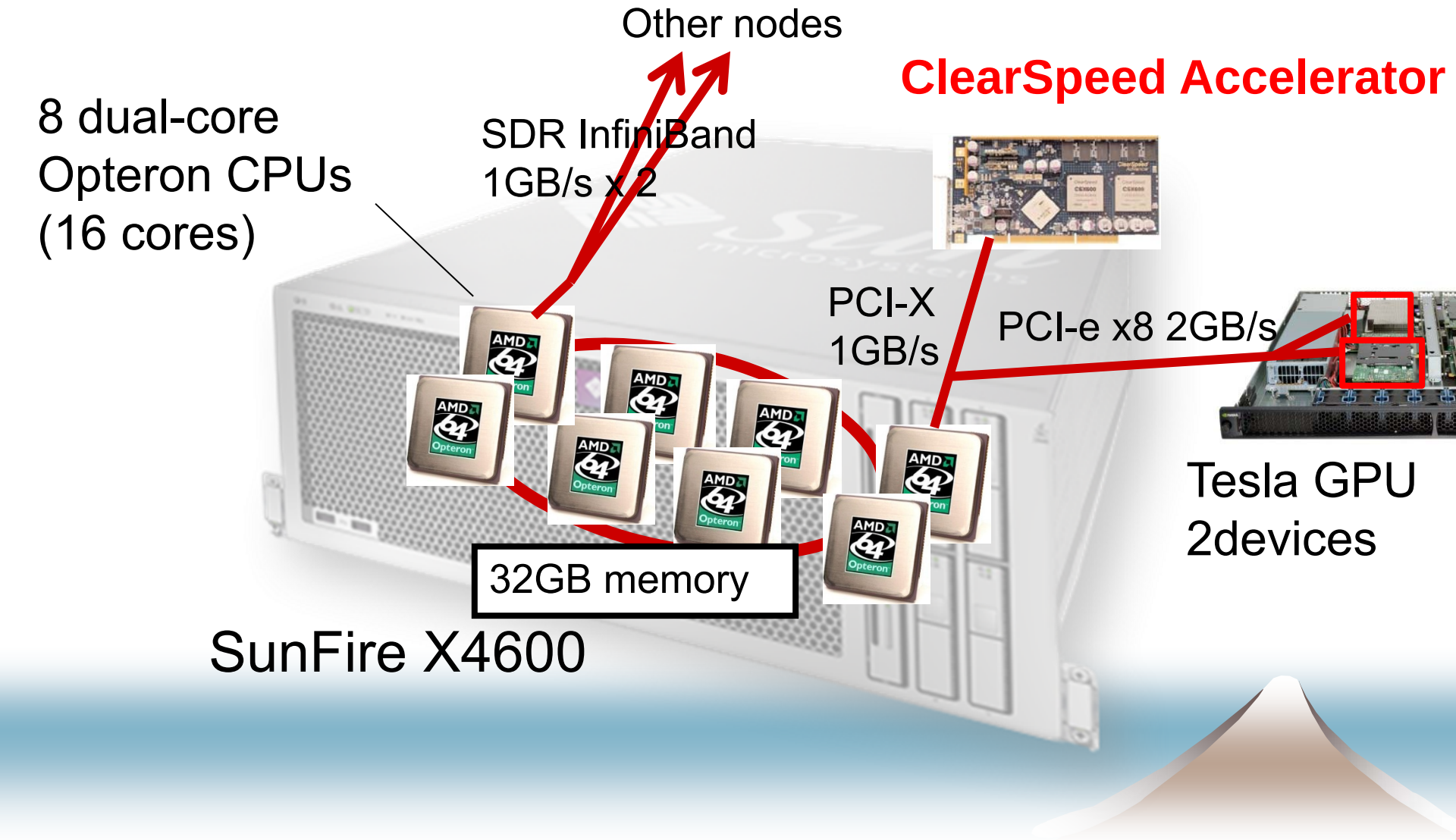
2006年6月
アジア No.1, 世界No.7
38.18Teraflops
(Top500)



ストレッジ
1 Petabyte (Sun "Thumper")
0.1Petabyte (NEC iStore)
Lustre ファイルシステム
>400Gbps

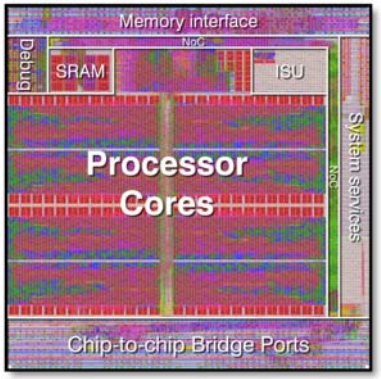
ClearSpeed CSX600
SIMD accelerator
360 boards,
30TeraFlops

異種プロセッサを持つ TSUBAMEノードの構成



Advance™ Dual CSX600 PCI-X accelerator board

360 boards in TSUBAME1.0
-> increase to ~648 in 2007
(>50TeraFlops)

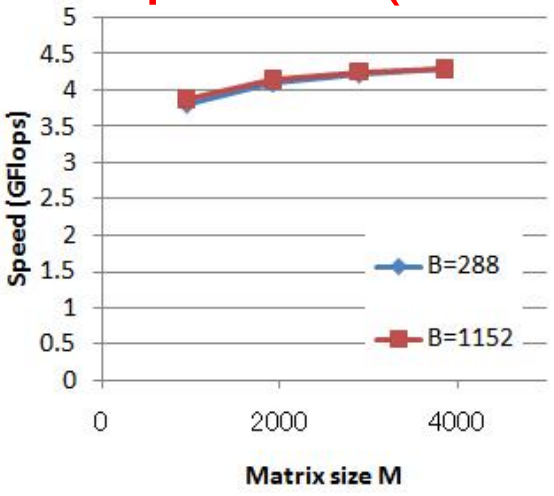


- 50 DGEMM GFLOPS sustained
- 0.4 M 1K complex single precision FFTs/s (20 GFLOPS)
- ~200 Gbytes/s aggregate B/W to on-chip memories
- **6.4 Gbytes/s aggregate B/W to local ECC DDR2-DRAM**
- 1 Gbyte of local DRAM (512 Mbytes per CSX600)
- ~1 Gbyte/s to/from board via PCI-X @133 MHz
- < 25 watts for entire card (8” single-slot PCI-X)

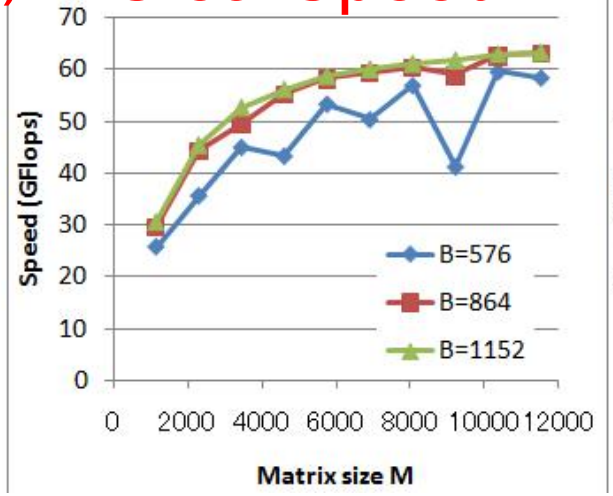
Measured Kernel Performance

Performance of Multiply of (MxB) x (BxM)

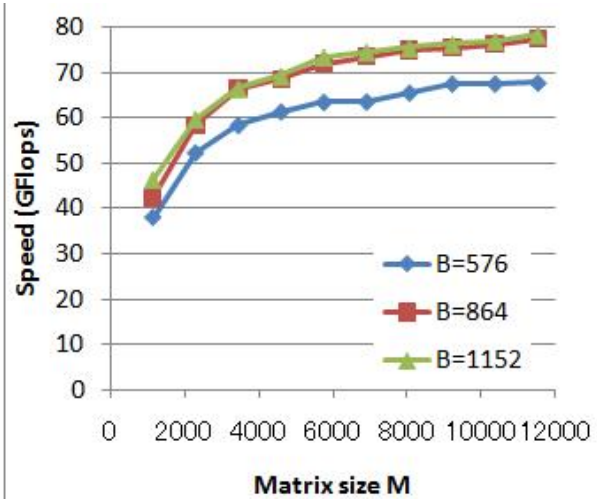
Opteron (1core)



ClearSpeed



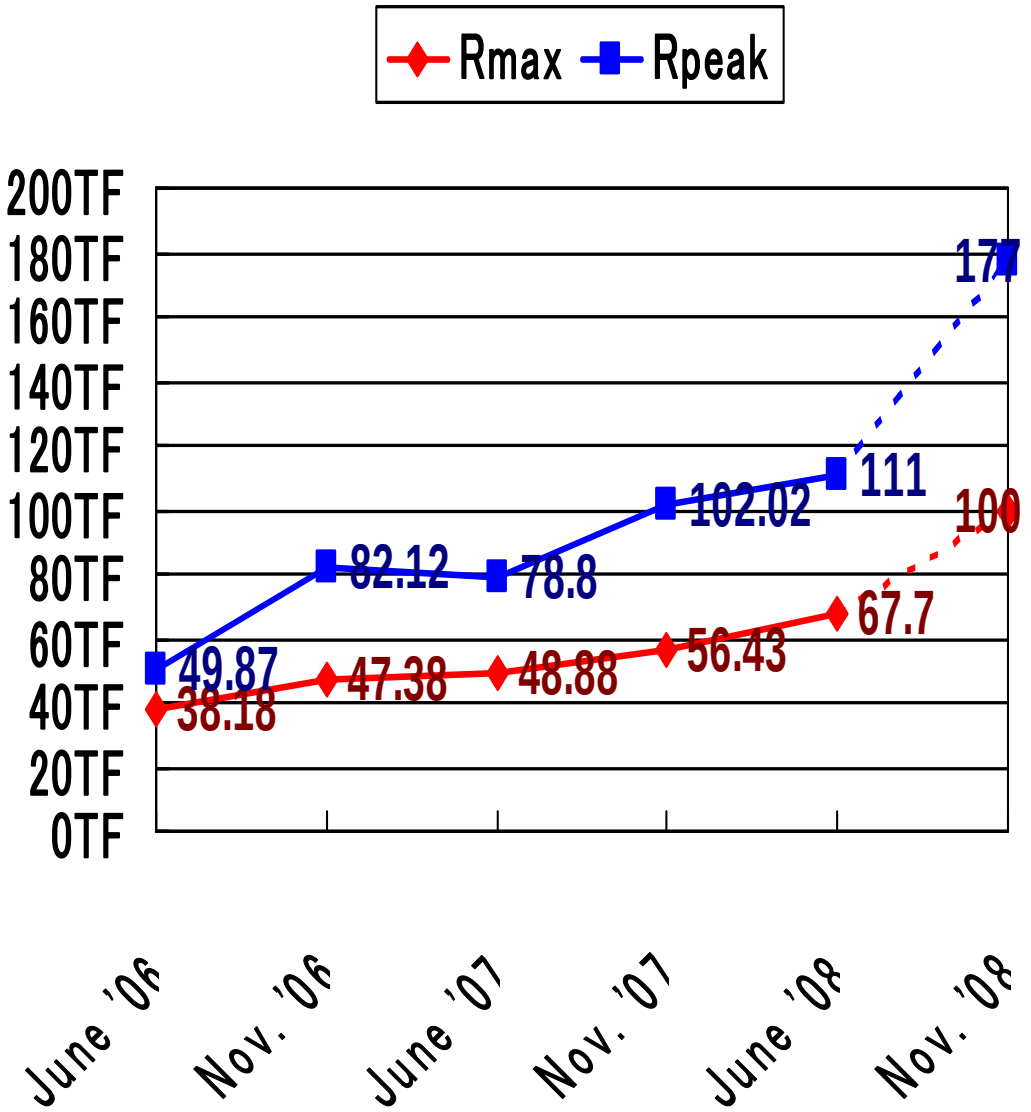
G200 Tesla



- Opteron: GotoBLAS by Kazushige Goto
- ClearSpeed: CSXL by ClearSpeed
  - Matrix size should be multiple of 288
- Tesla: NUBLAS from our group

Results---5 consecutive Top500 Performance Increase

- First ever "Heterogeneous" Architecture on Top500 w/47.38TF (#9 Nov. 2006 28th Top 500)
  - 648 nodes 360 Accelerators
- "Heterogeneous HPL" algorithm published @ IEEE IPDPS 2008
- Continued improvements via
  - Adding more Clearspeed boards (648)
  - Algorithmic Improvements
  - Various Tuning
- Now at 67.7 TF for the June 2008 (31st) Top500



ClearSpeed Experience: Mode-of Use

- 1. Numerical Library Acceleration
  - Transparent to users (Fortran/C bindings)
  - BLAS/LAPACK, IMSL (Dense LA) – often used
    - Joint CPU-CS acceleration – x2 for TSUBAME
  - BW-intensive kernels very slow – not used
- 2. User Application Acceleration
  - Matlab, Mathematica, Amber10, …
  - Slow: Perf(Opteron Node) > Perf(CS) – not used
- 3. User Applications
  - Need MPI-like programming with C-dialect Cn
  - *Hardly any users here…*

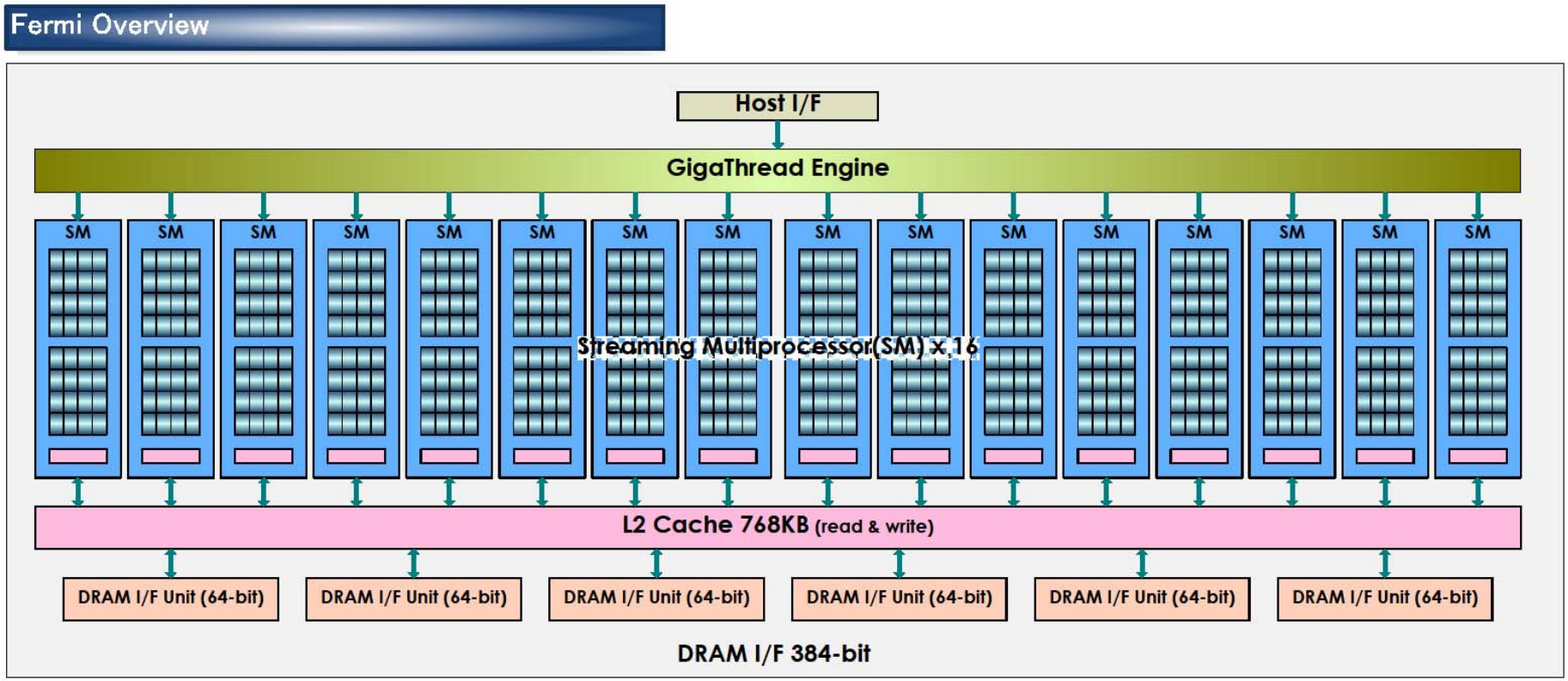
CS Acceleration is extremely "Narrow Band"=> Hard to Scale

GPUs as Commodity Massively Parallel Vector Processors

- E.g., NVIDIA Tesla, AMD Firestream
  - High Peak Performance > 1TFlops
    - Good for tightly coupled code e.g. Nbody
  - **High Memory bandwidth (>100GB/s)**
    - **Good for sparse codes e.g. CFD**
  - **Low latency over shared memory**
    - **Thousands threads hide latency w/zero overhead**
  - **Slow and Parallel and Efficient vector engines for HPC**
  - Restrictions: Limited non-stream memory access, PCI-express overhead, programming model etc.

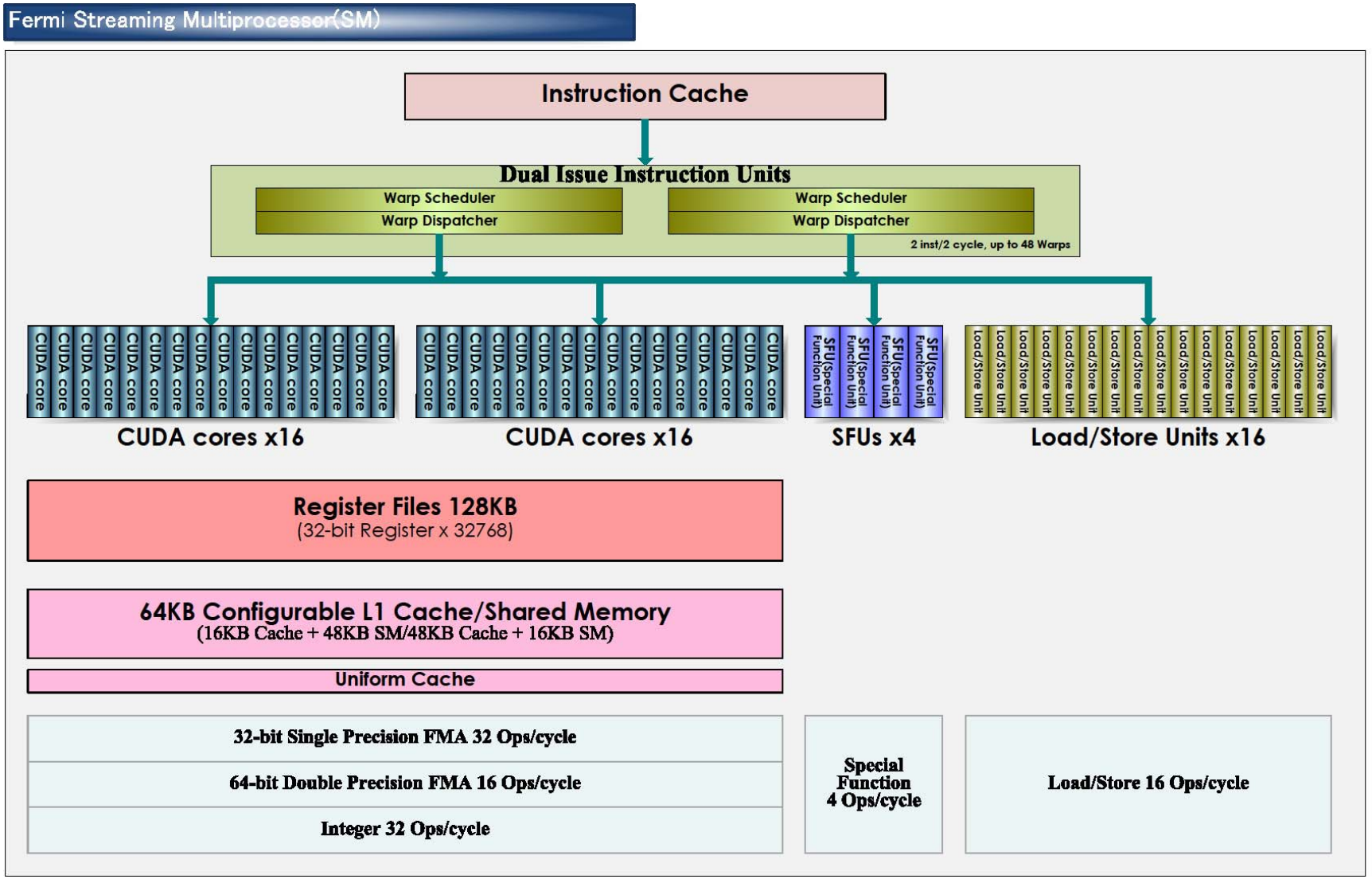
How do we exploit them given vector computing experiences?

NVIDIA Fermi
Many Core, Multhreaded, SIMD-Vector, MIMD Parallel Architecture



(Figure by Kazushige Goto)

NVIDIA Fermi SM "Core"

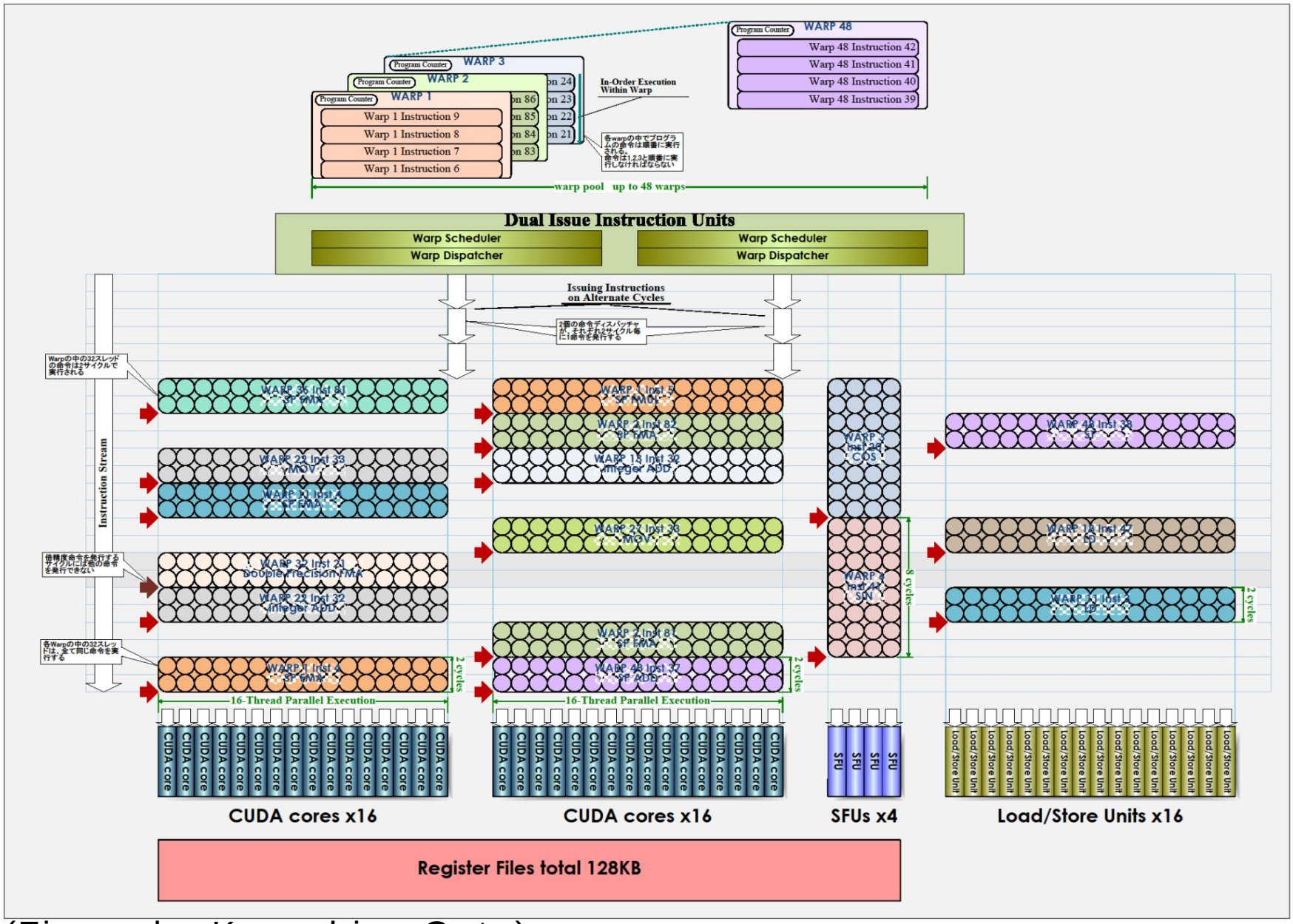


(Figure by Kazushige Goto)

Parallelism in CUDA GPUs

- SIMD-Vector Parallelism (WARP)
  - 32 (16) way vector parallelism
- SPMD Thread Parallelism
  - Thousands of threads possible
  - Cyclic pipeline, hides memory latency, like vector processor but allows out-of-order execution
- MIMD Parallelism (Kernel)
  - Up to #SM (16 in Fermi) Kernels with independent instruction streams
- GPU-CPU Heterogeneous Parallelism
  - Streaming Data Transfer Enginer via PCI-e
  - Massive Parallelism: GPU, Short Latency: CPU
  - In single chip sharing memory in Denver generation

Fermi Warp Scheduling



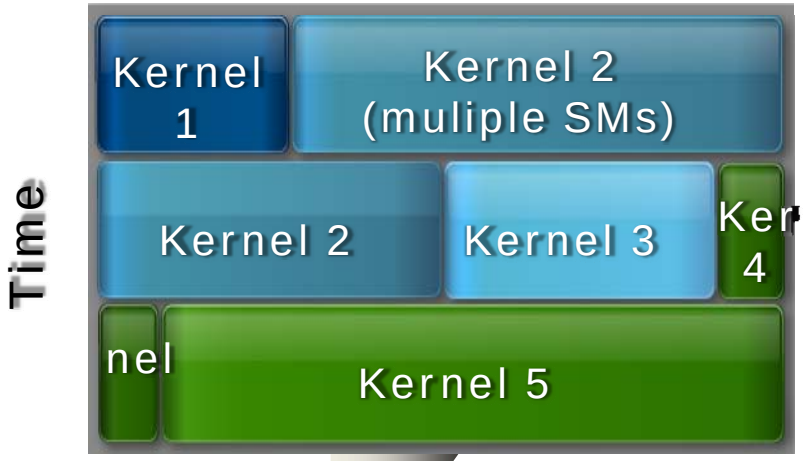
(Figure by Kazushige Goto)

Fermi: MIMD Execution of CPU threads and Multiple GPU Kernels

複数Kernel同時実行 + 高速コンテキストスイッチ

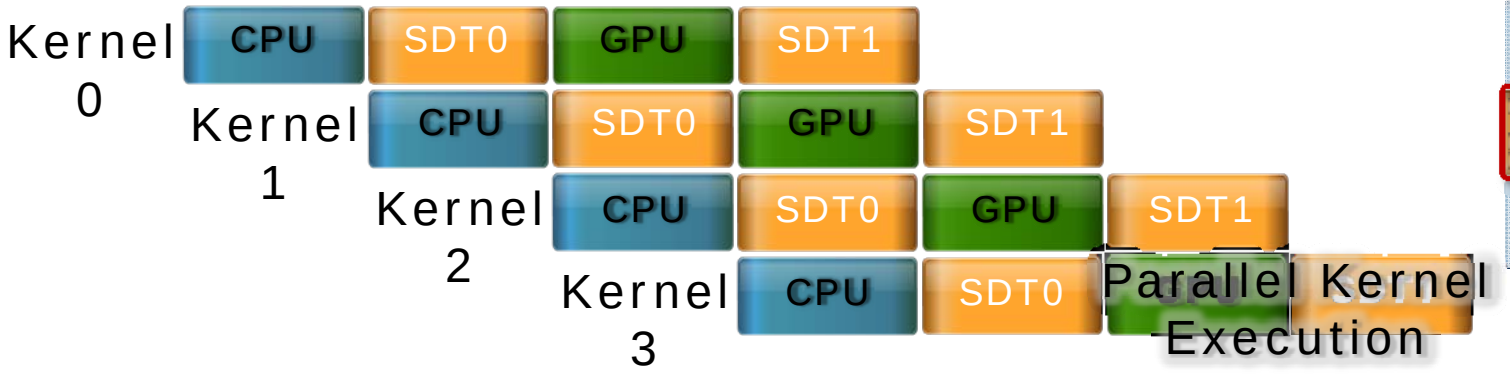
SM毎に異なる命令ストリームに割り当て可能

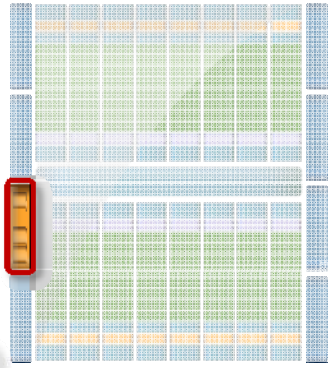
⇒16並列のタスクパラレル+内部の細粒度SIMD並列



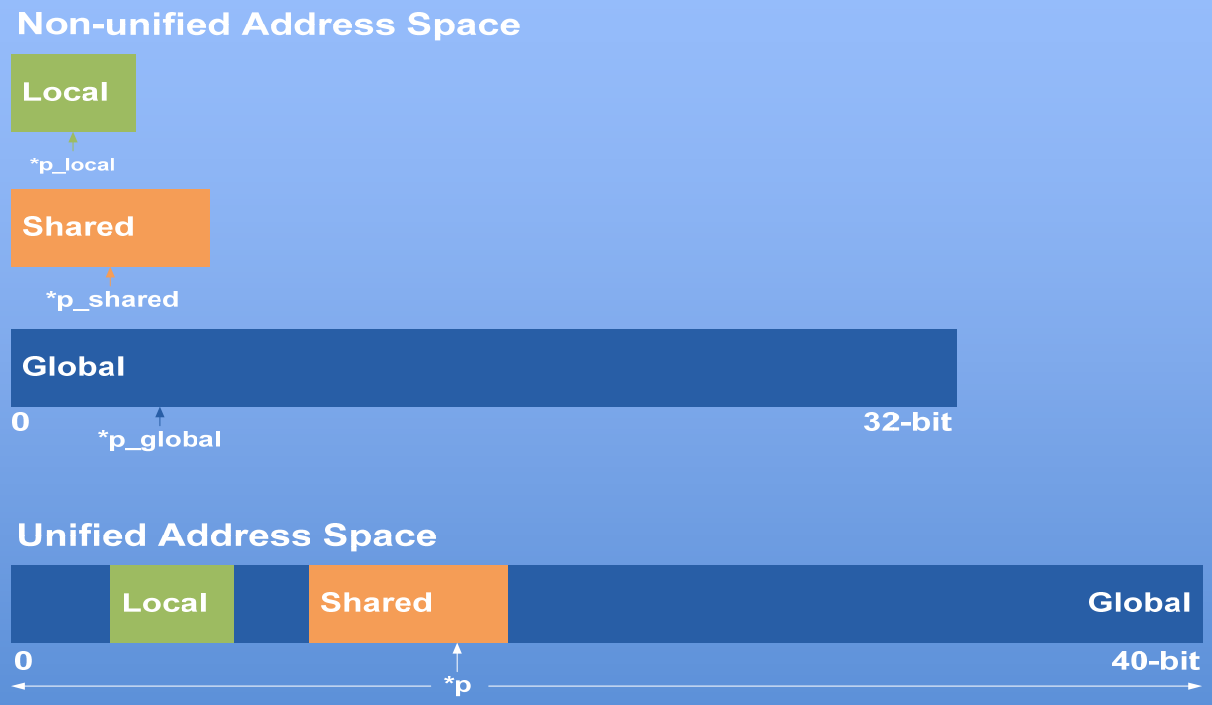
GigaThread Streaming Data Transfer Engine

- CPU→GPU 及び GPU→CPU データ転送の同時実行
  - CPU と GPUの完全に並列実行をオーバラップ可能



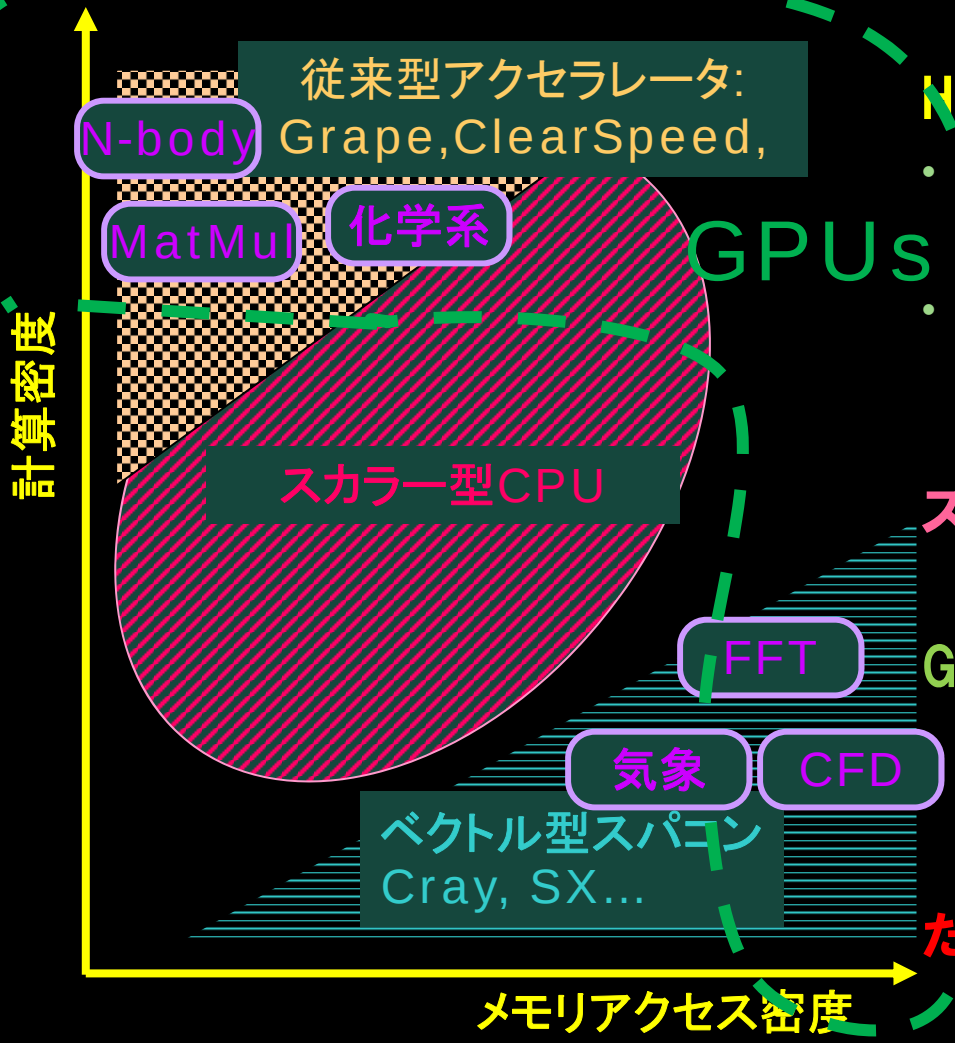


- Fermi における統合アドレススペース
  C／C++ ポインターのフルサポート



将来は、GPUs/CPUs の統合アドレッシング
- Barcelona SC: GMA (非対称共有メモリ)
- CUDA XXX (ごめんなさい言えません)
- NVIDIA Denver, HPC用 AMD Fusion

新世代のベクトル計算機としてのGPU



NPCのワークロードとして、二つのタイプ

- 計算密度が高い「密問題」
  →従来型のアクセラレータが得意
- メモリアクセス密度が高い「粗問題」
  →ベクトル型スパコンが得意

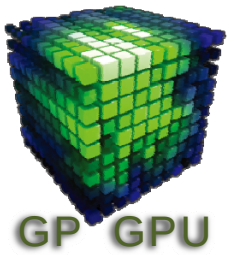
スカラー型CPUはどちらもそこそこ
→高性能を得るために巨大並列化

GPUは、新世代のベクトルプロセッサと、計算密度が高いアクセラレータとしての両面を持つ
→効率の良いスパコンの主要素

ただし、少ないメモリ量・CPUや他GPUとの通信・GPU向アルゴリズムやアプリ・ベクトル並列のプログラミング手法・システムソフトウェア等の技術課題

東工大GSICでの研究開発

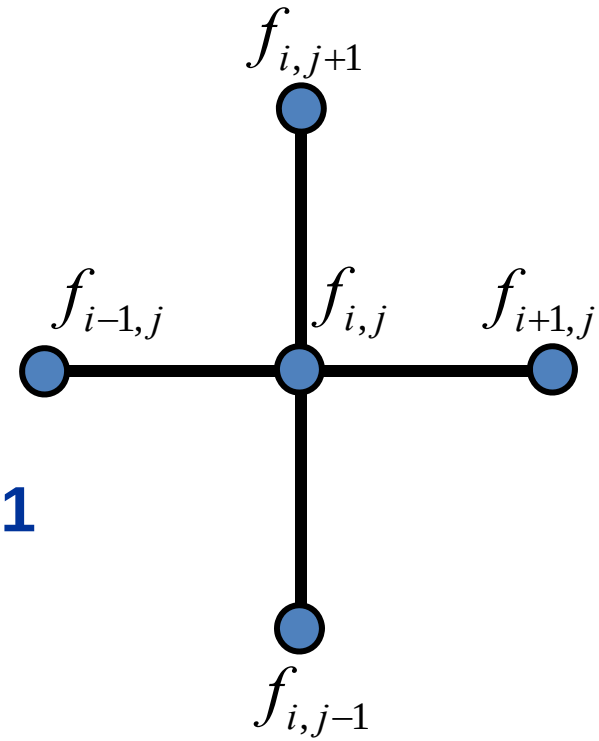
Compute Intensive or Memory Bound ?



2-dimensional diffusion Equation

$$\frac{f_{i,j}^{n+1}-f_{i,j}^{n}}{\Delta t}=\kappa\left(\frac{f_{i+1,j}^{n}-2f_{i,j}^{n}+f_{i-1,j}^{n}}{\Delta x^2}+\frac{f_{i,j+1}^{n}-2f_{i,j}^{n}+f_{i,j-1}^{n}}{\Delta y^2}\right)$$

$$f_{i,j}^{n+1}=c_0 f_{i,j}^{n}+c_1 f_{i+1,j}^{n}+c_2 f_{i-1,j}^{n}+c_3 f_{i,j+1}^{n}+c_4 f_{i,j-1}^{n}$$

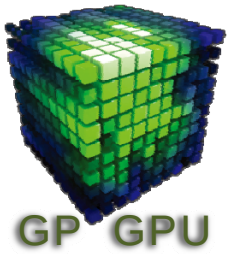


FLOP = 9

Byte = 4*6 = 24 byte : read 5, write 1

FLOP/Byte = 9/24 = 0.375

Arithmetic INTENSITY: FLOP/Byte



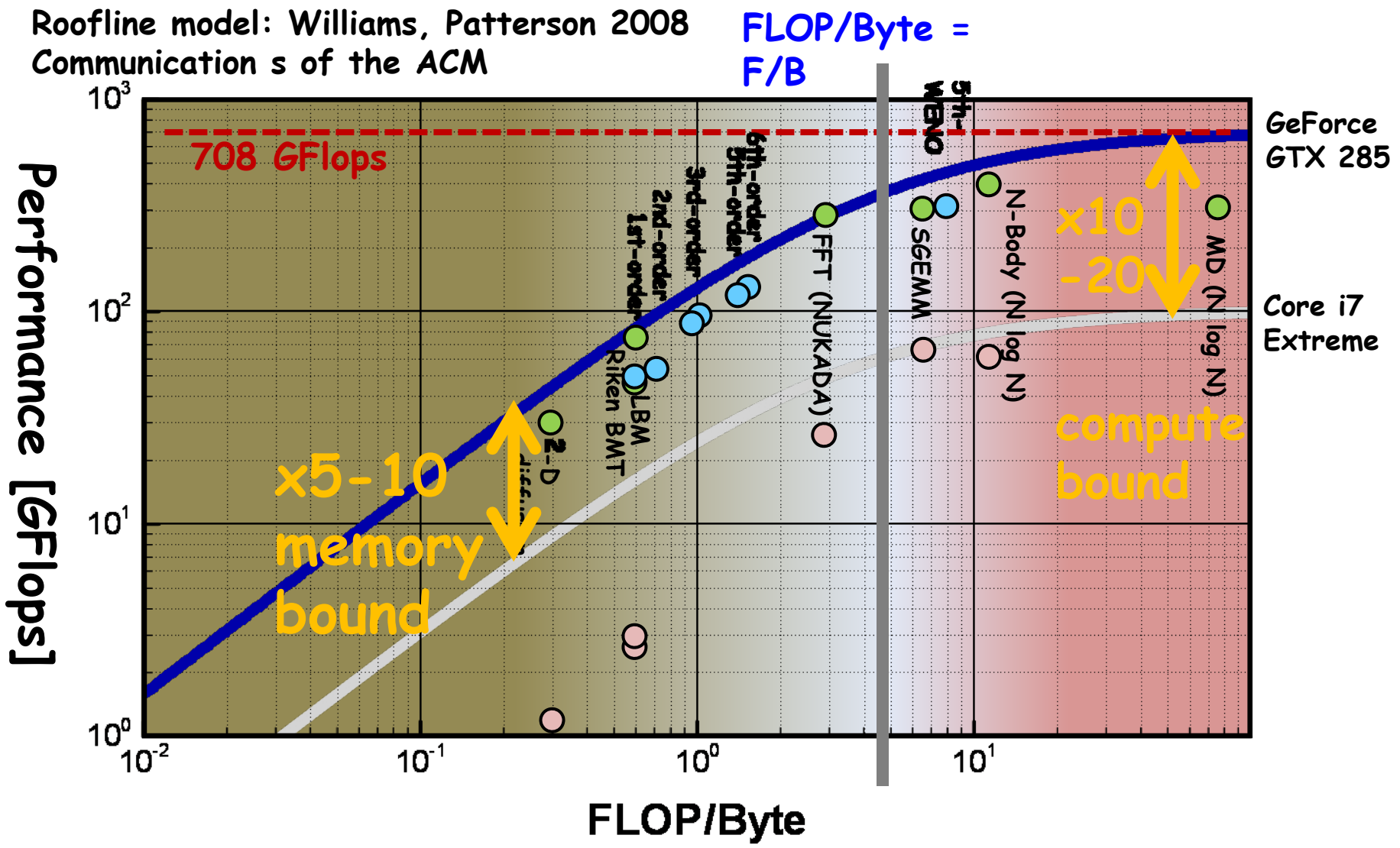
FLOP = number of FP operation for applications
Byte = Byte number of memory access for applications

F = Peak Performance of floating point operation
B = Peak Memory Bandwidth

$$\text{Performance} = \frac{\text{FLOP}}{\text{FLOP/F} + \text{Byte/B} + \alpha}$$

$$= \frac{\text{FLOP/Byte}}{\text{FLOP/Byte} + \text{F/B} + \alpha} \text{F}$$

GPU vs. CPU Performance



DOE のキーアプリケーション群

(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

アプリケーションにはバンド幅？レーテンシ？ Latency Bound vs. Bandwidth Bound?

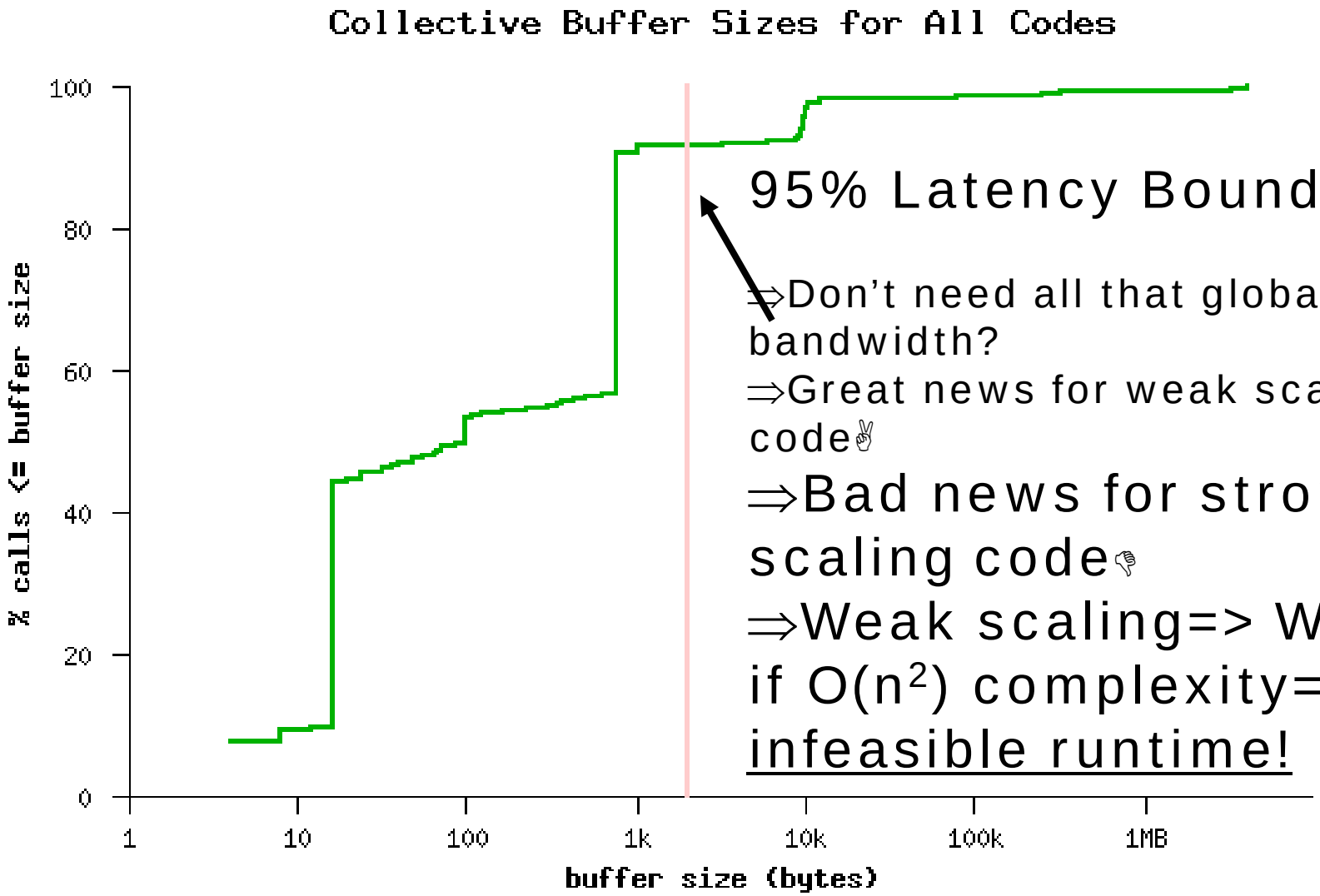
- How large does a message have to be in order to saturate a dedicated circuit on the interconnect?
  - $N^{1/2}$ from the early days of vector computing
  - Bandwidth Delay Product in TCP

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)

多くの実問題は実はレーテンシバウンド
–小規模メッセージパッシングプロセッサの問題–



95% Latency Bound!!!

⇒Don't need all that global NW bandwidth?

⇒Great news for weak scaling code✌

⇒Bad news for strong scaling code👎

⇒Weak scaling=> What if $O(n^2)$ complexity=> infeasible runtime!

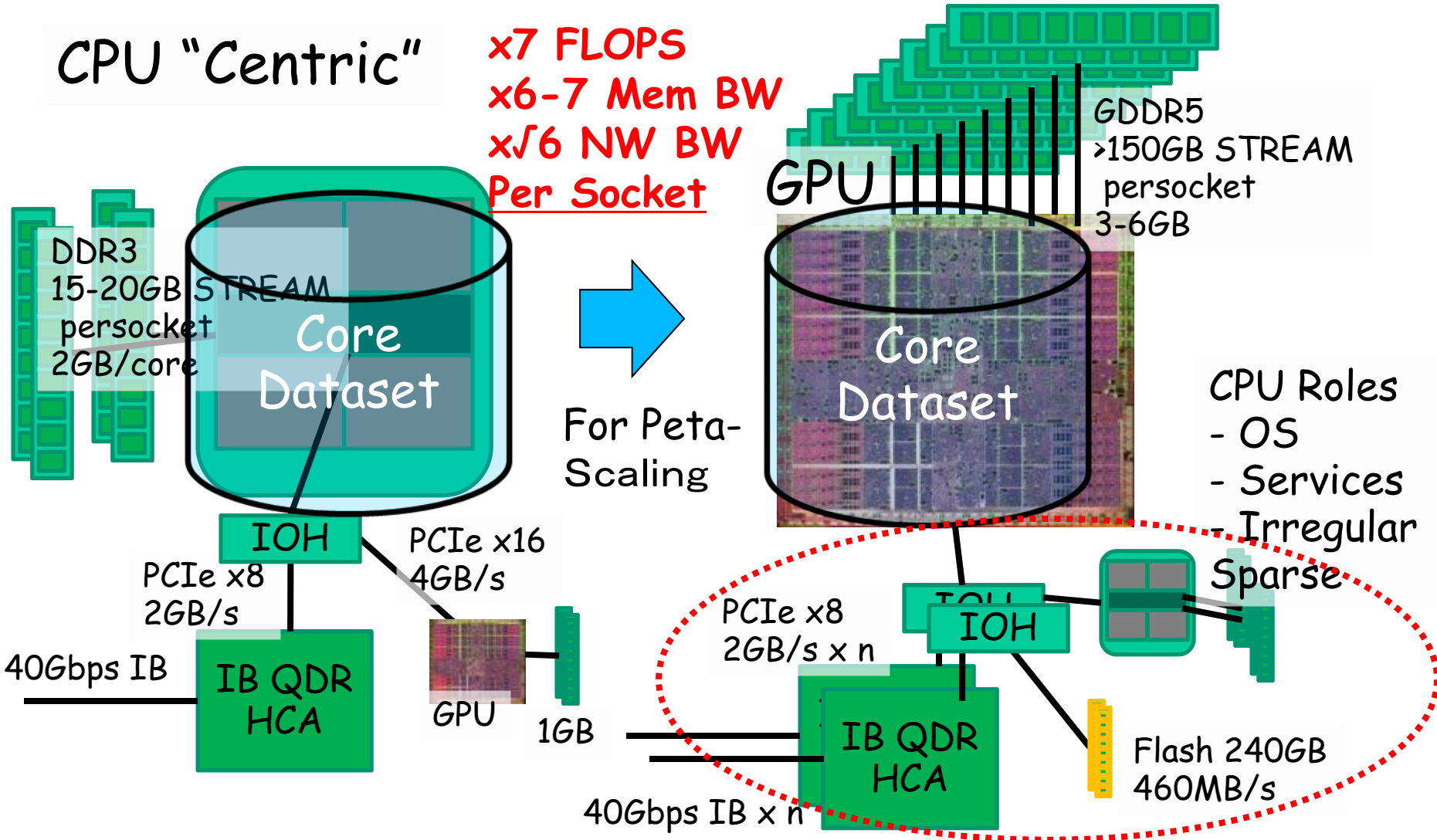
(Original slide courtesy John Shalf @ LBL)

ペタからエクサへのスケーリング

強スケーリング達成のためには

- レーテンシをなるべく短く
  - Extreme multi-core incl. vectors
  - "Fat" nodes, exploit short-distance interconnection
  - Direct cross-node DMA (e.g., put/get for PGAS)
- でなければ、レーテンシを隠す
  - Dynamic multithreading (Old: dataflow, New: GPUs)
  - Trade Bandwidth for Latency (so we do need BW…)
  - Departure from simple mesh system scaling
- レーテンシに敏感なアルゴリズムを変更
  - From implicit Methods to direct/hybrid methods
  - Structural locality, extraploation, stochastics (MC)
  - Still may require global bandwidth for implicit solvers

TSUBAME2.0のノードアーキテクチャ GPU+CPUによるベクトル・スカラー混合アーキテクチャ



TSUBAMEの歴史

2006
TSUBAME 1.0
85TFlops/1.1PB

アジアNo.1
「みんなのスパコン」
x86 CPU+アクセラレータ

2007
TSUBAME 1.1
100TFlops/1.6PB

ストレージ・アクセラレータ
増強

2008
TSUBAME 1.2
160TFlops/1.6PB

GPUアクセラレータ
680枚増強

2010/11/01
TSUBAME 2.0
2.4PFlops/7.1PB
日本初のペタコン、TSUBAME1.0の30倍の性能

- TSUBAME初の完全リプレース

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(!)**

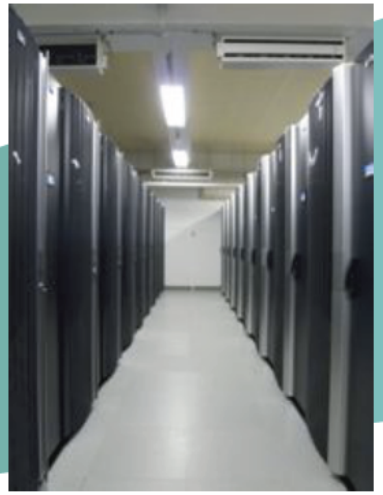
- 0.72 Petabyte/s aggregate mem BW,
  - Effective 0.3-0.5 Bytes/Flop, restrained memory capacity (100TB)
- Optical Dual-Rail IB-QDR BW, full bisection BW(Fat Tree)
  - 200Tbits/s, Likely fastest in the world, still scalable
- Flash/node, ~200TB (1PB in future), 660GB/s I/O BW
  - >7 PB IB attached HDDs, 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.

TSUBAME2.0 2010年11月1日稼働開始

TSUBAME2.0: A GPU-centric Green 2.4 Petaflops Supercomputer

Tsubame 2.0: "Tiny" footprint, very power efficient
- Floorspace less than 200m² (2,100 ft²)
- Top-class power efficient machine on the Green 500

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes



2.4 PFLOPS
80 TB

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB

Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB

Chip
(CPU ,GPU)

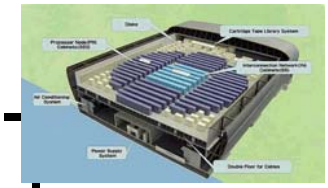
CPU(Westmere EP)
76.8 GFLOPS

GPUs(Tesla M2050)
515 GFLOPS
3 GB

Integrated by NEC Corporation

TSUBAME2.0の性能向上

地球シミュレータ ⇒ TSUBAME 4年間
30-40倍のダウンサイズ



TSUBAME ⇒
TSUBAME 2.0
4年間で30-40倍

TSUBAME2.0
2.4PF (2010)

Japanese NLP
>10PF(2012)

US >10P
(2011~12?)

Others

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

TSUBAME 1.2
Upgrade
(3Q2008) 170TF

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

TSUBAME 1.0 80TF (1Q2006)

BlueGene/L
360TF(2005)

世界一ライン

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


~50 compute racks + 6 switch racks
Two Rooms, Total 160m²
1.4MW (Max, Linpack), 0.48MW (Idle)

TSUBAME2.0システム概念図

Thin計算ノード

HP SL390z G7 x1408

*CPU:215.99TFlops
*CPU+GPU
2391.35TFlops
*Memory 80.55TB
*SSD 173.88TB

Medium計算ノード

HP DL580 G7 x24

*CPU:6.14TFlops
*Memory 3.07TB
*SSD 11.52TB

Fat計算ノード

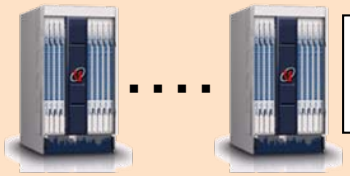
HP DL580 G7 x10

*CPU:2.56TFlops
*Memory 3.07TB
*SSD 4.8TB

計算ノード

2.4PFlops (CPU+GPU)

種別	ノード数	CPU Clock	Memory	SSD
Thin	1367 41	2.93GHz 2.93GHz	54GiB 96GiB	120GB 240GB
Medium	24	2.0GHz	128GiB	480GB
Fat	8 2	2.0GHz 2.0GHz	256GiB 512GiB	480GB 480GB



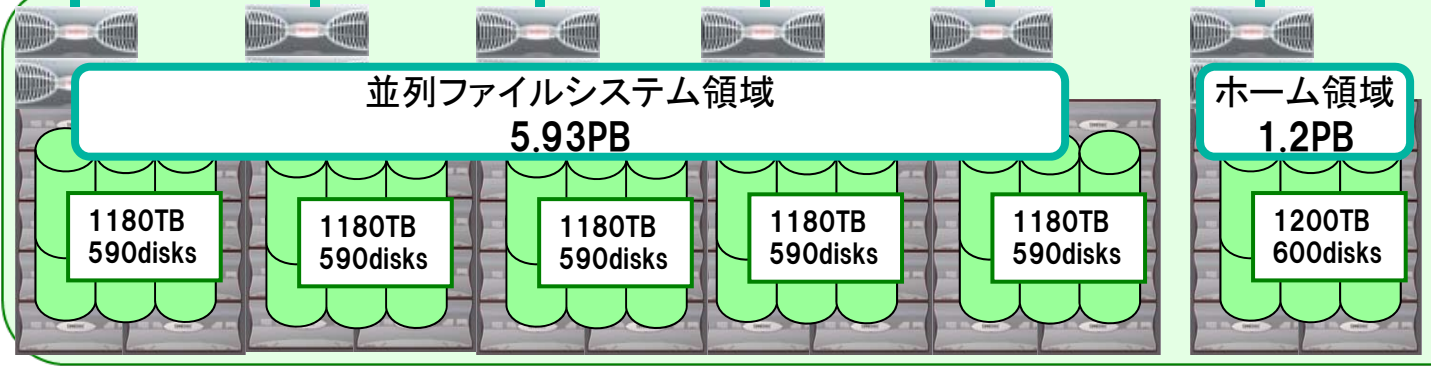
InfiniBand Network
Voltaire GridDirector 4700
12 switches

ノード間相互結合網

FatTree接続　ノンブロッキング・フルバイセクション

ペタバイト級ストレージ

7.13PB



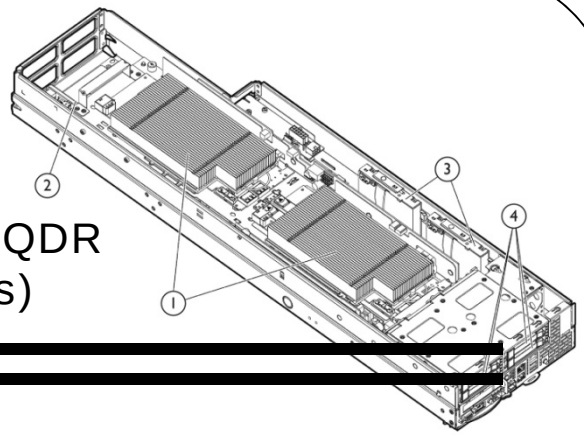
	TSUBAME 1 (2006年, 22億円)	T2K東大 (2008年, 90億円)	TACC Ranger (2008年. 60億円?)	TSUBAME2.0 (2010年, 32億円)
Cores/Node	16	16	16	12(CPU)+1344(GPU)
Node Mem BW(GBytes/s)	20	20	20	64(CPU)+450(GPU)
Node Network BW (Gbps)	20	40	10	80
#Nodes	655	952	3,936	1408(Thin) + 34(Med/Fat)
#Cores (Total)	10,480(CPU)	15,232	62,976	17,664(CPU)+189万(GPU)
# GPUs/Accelerators	360 (ClearSpeed)	0	0	4224 (Tesla M2050)
理論 Peak TFLOPS (倍精度)	80	141	579	2400
合算メモリバンド幅(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 (倍精度-TFLOPS)	48	102	433	>1000
合算 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/年	~1MW/年	2.4MW Year	~1MW/年
Compute Racks	65	70?	~100	~44

40倍以上

高バンド幅 ベクトル・スカラー混合

TSUBAME2.0 Compute Nodes

Thin Node



Infiniband QDR x2 (80Gbps)

HP SL390G7 (Developed for TSUBAME 2.0)
GPU: NVIDIA Fermi M2050 x 3
515GFlops, 3GByte memory /GPU
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node)
Memory: 54, 96 GB DDR3-1333
SSD:60GBx2, 120GBx2

1408nodes:

4224GPUs:59,136 SIMD Vector Cores, 2175.36TFlops (Double FP)

2816CPUs, 16,896 Scalar Cores: 215.99TFlops

Total: 2391.35TFLOPS

Memory: 80.6TB (CPU) + 12.7TB (GPU)

SSD: 173.9TB

IB QDR

PCI-e Gen2x16 x2
NVIDIA Tesla S1070 GPU

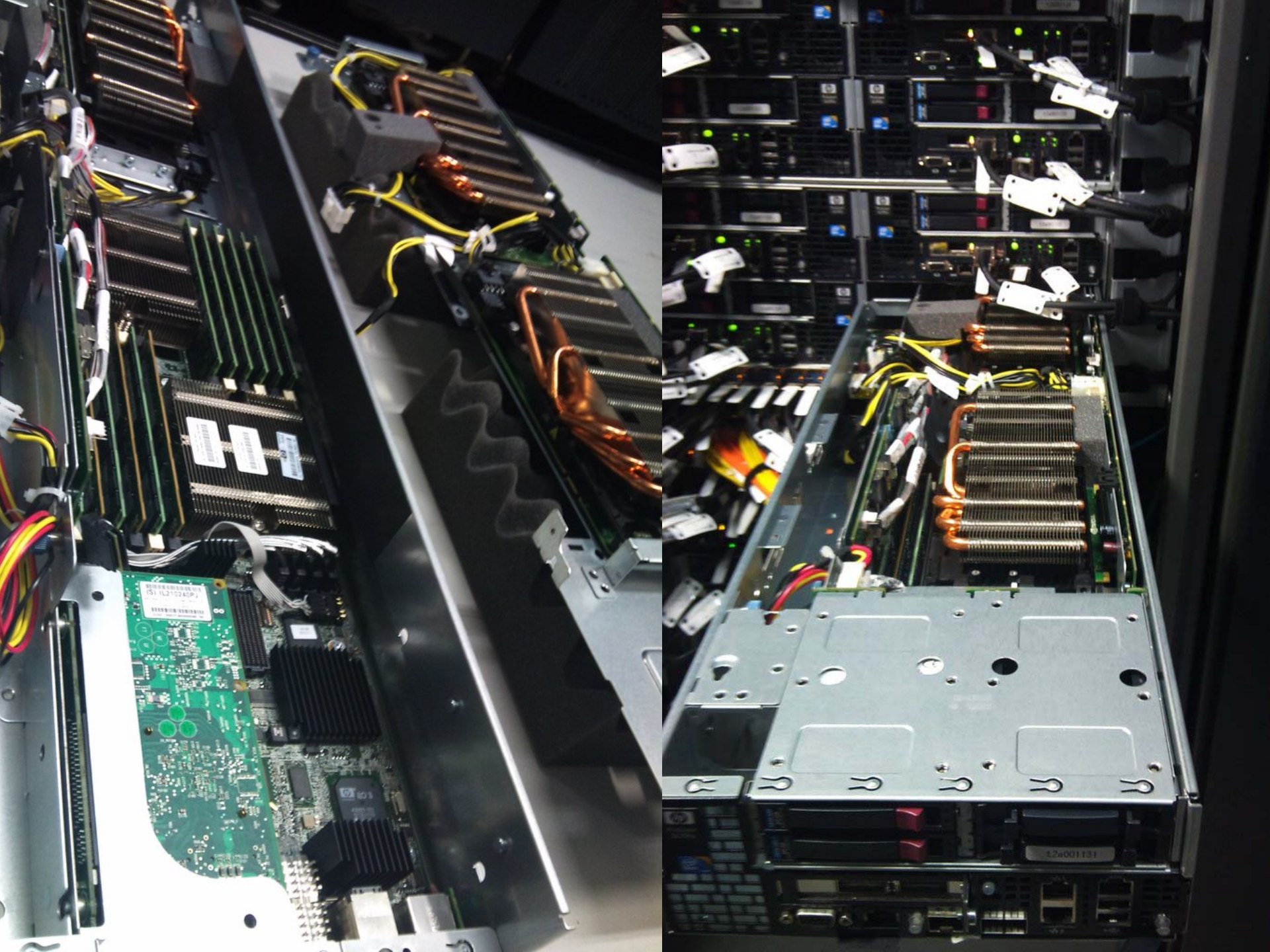
HP 4 Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 (32cores/node)
Memory: 128, 256, 512GB DDR3-1066
SSD:120GB x4 (480GB/node)

34 nodes:
8.7TFlops

Memory:
6.0TB+GPU

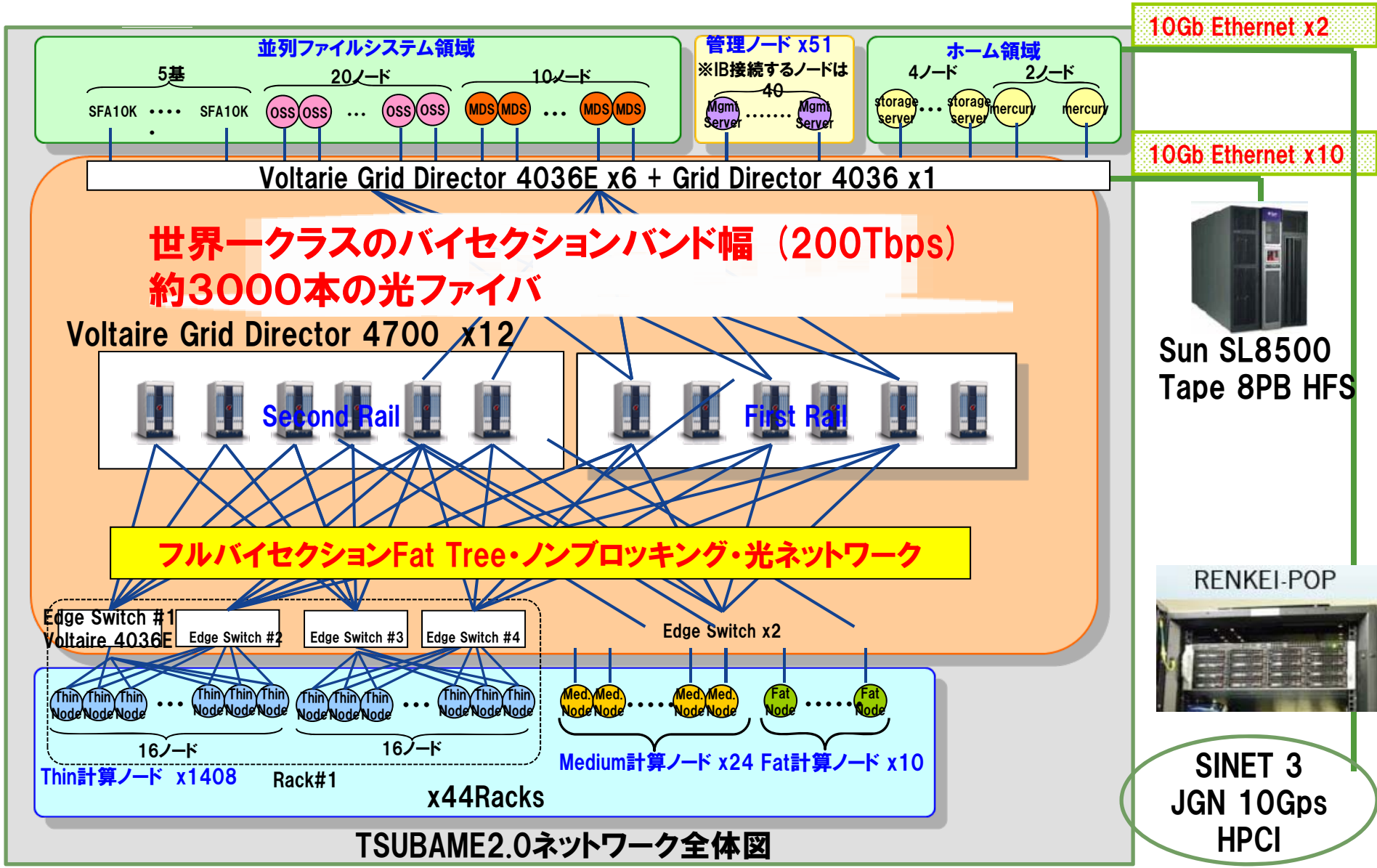
SSD: 16TB+

Total Perf
2.4PFlops
Mem: ~100TB
SSD: ~200TB




MADE IN TOKYO

TSUBAME2.0ノード間相互結合網



TSUBAME2.0ネットワーク全体図


3500 Fiber Cables > 100Km

TSUBAME 2.0ペタバイト級ストレージ

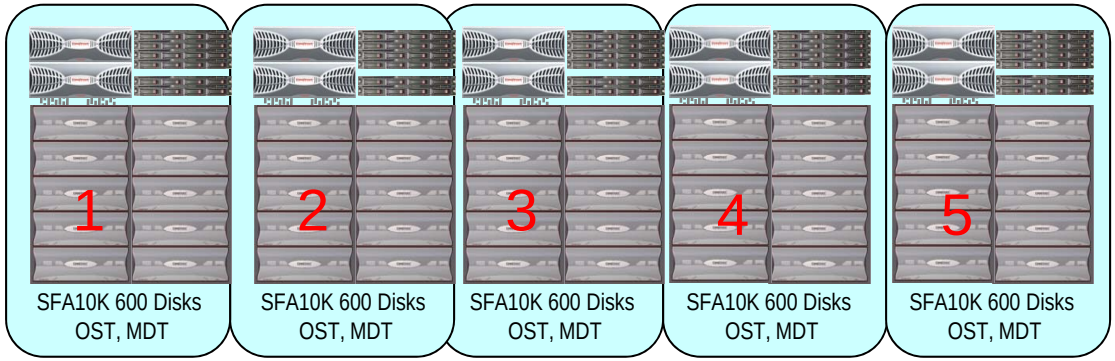
1) 各ノードの短期記憶用SSD、2) Lustre/GPFSを利用した「並列ファイルシステム領域」, NFS,CIFS,iSCSIを備えた「ホーム・クラウドサービス用領域」のHDD群, および 3) 長期保存用テープシステムで構成

Lustre 並列ファイルシステム領域

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

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

ストレージ:DDN SFA10000 x5
- -Total容量:5.93PB
  2TB SATA x 2950 Disks + 600GB SAS x 50 Disks



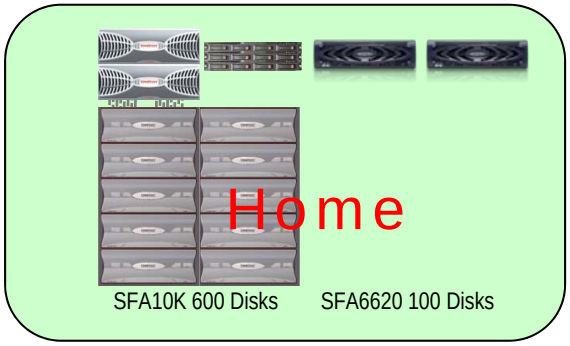
並列ファイルシステム領域 5.93PB

ホーム・クラウドサービス用領域

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

NFS/CIFS/iSCSI アクセラレーション:BlueArc Mercury100 x2
- -10GbE x2

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



ホーム領域 1.2PB

約200TB SSD+7.13PB HDD + 約8PBテープ(予定)の大容量階層ストレージ
合算15ペタバイト: 全国大学基盤センター群合算の数倍の容量

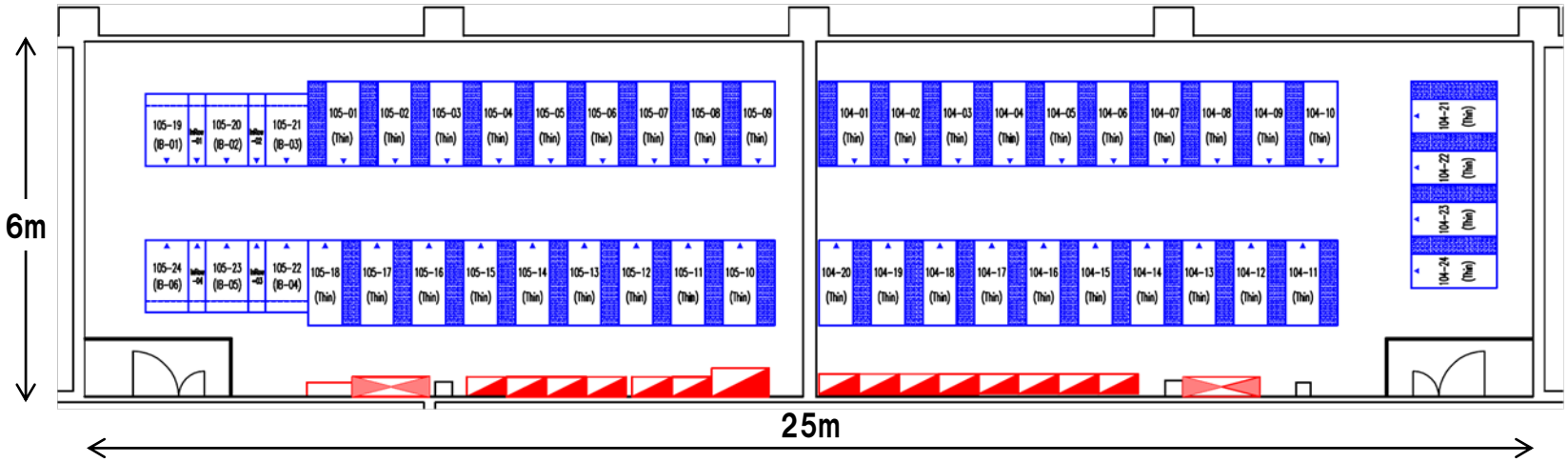

11ペタ(10^{15})バイトの
ストレージ

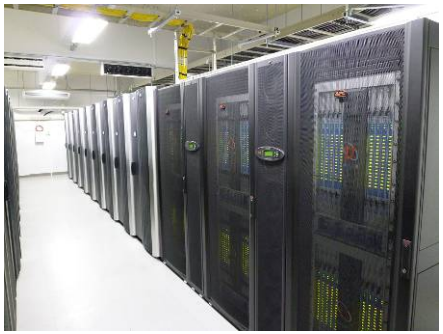
TSUBAME2.0(Green:施設関連)

超省スペース化

- TSUBAME1.2よりも狭いスペースに，2倍以上のNode数を設置
  約2PFlops越えの計算能力を約１５０㎡のスペースに収容

【１階マシン室レイアウト】

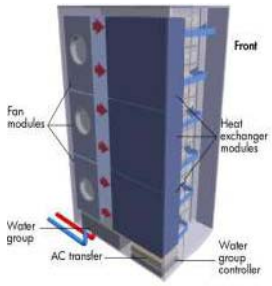


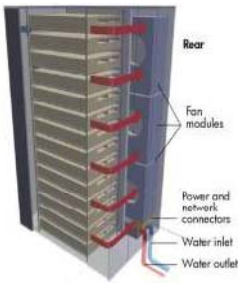


TSUBAME2.0(Green:施設関連)

超高密度実装Rackの冷却

- 水冷ラックによる，35kW/Rackの冷却を実現
  -HP社製MCSラック導入により，サーバ吸気口に均質な冷却風を提供



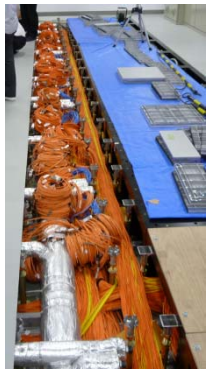


年間PUE1.28の実現

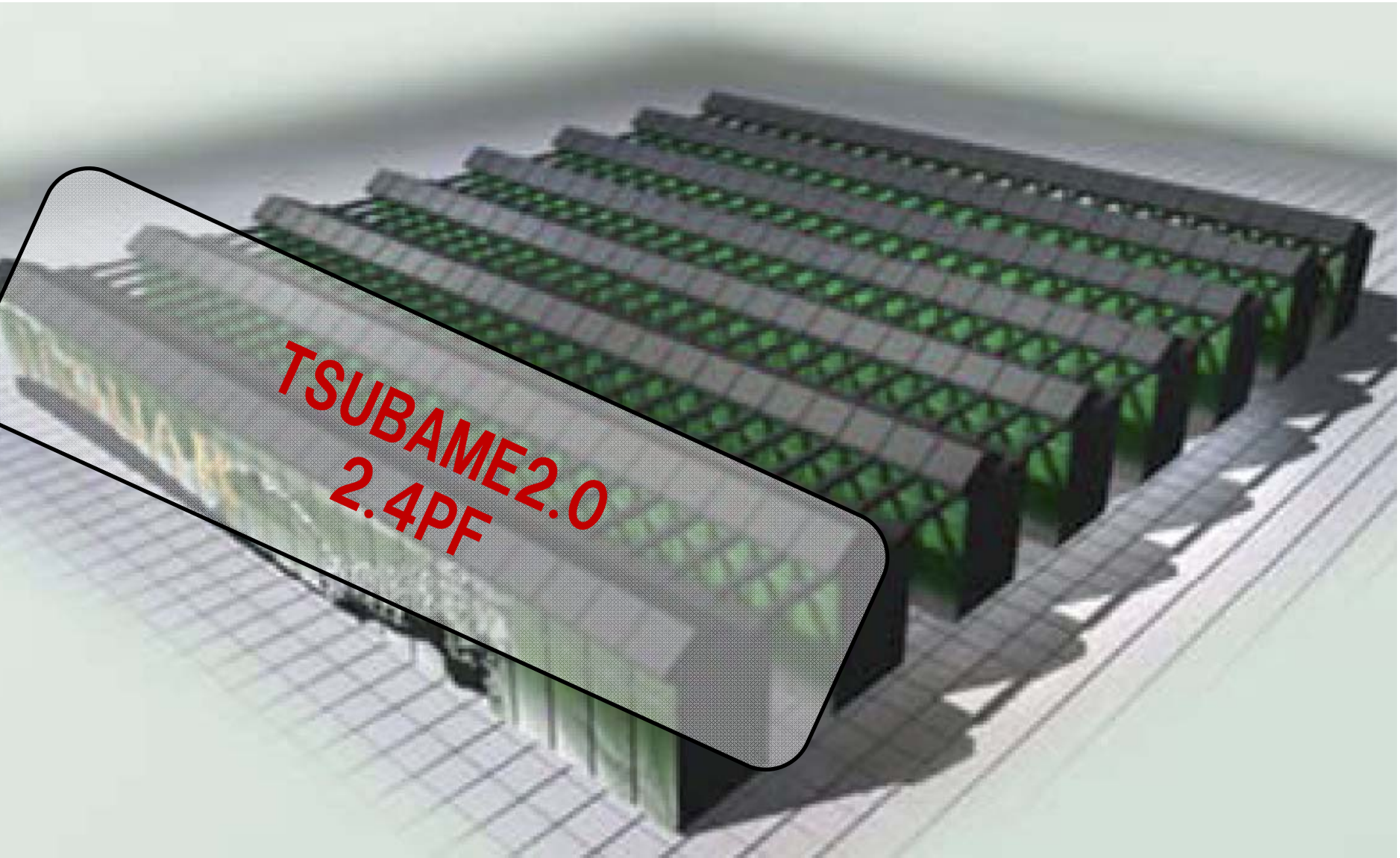
- 大規模水冷システム構築
  -高効率チラー採用により，付帯設備電力を削減
  -効率の良い冷却設備を使用することにより平均電力を削減
  -施設関連のデータを業務管理システムにリアルタイムで供給し、施設管理も含めたITシステム統合管理

大量ケーブルの配線

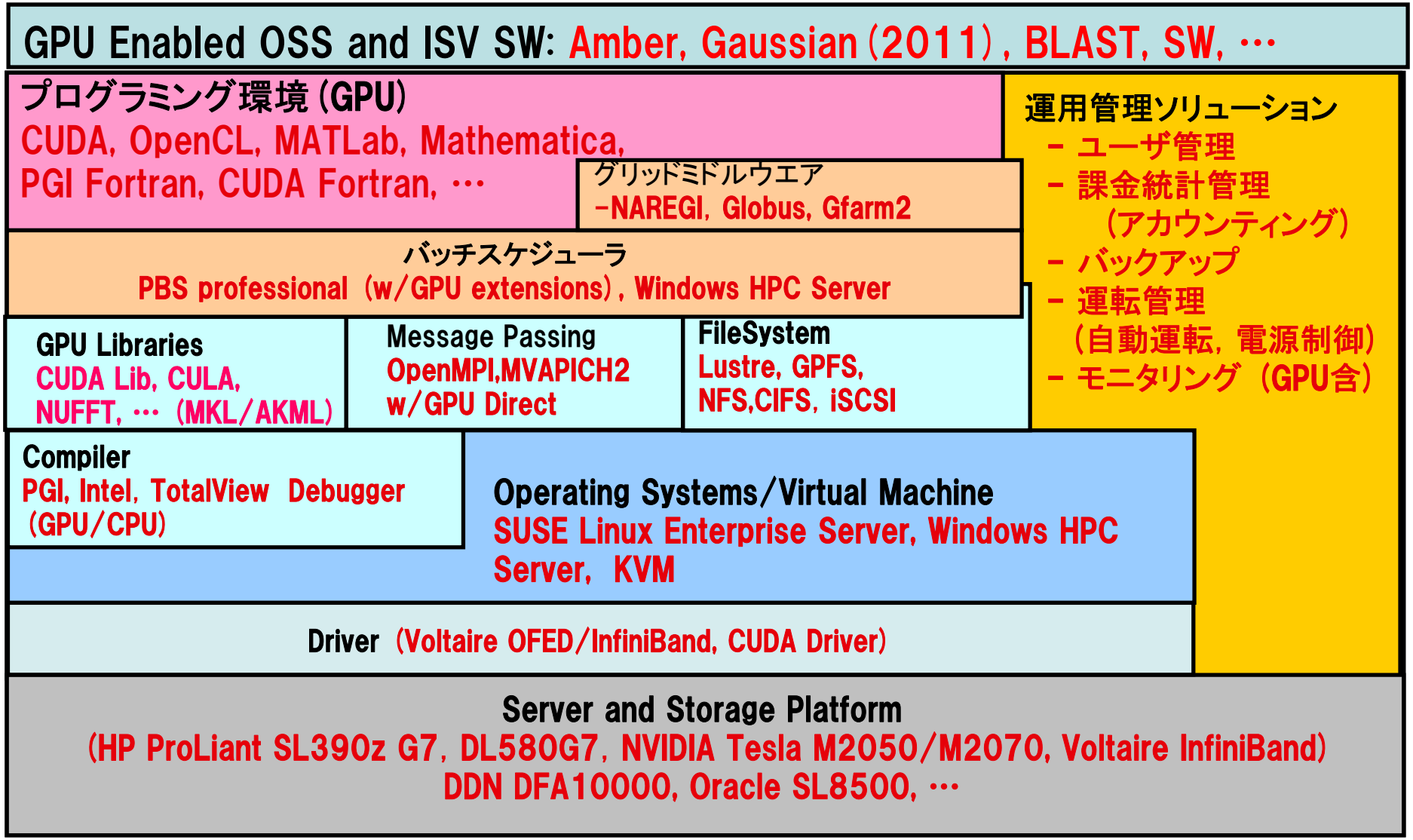
- ラック間IBケーブル:約3,500本(約100,000m)を配線
  ※TSUBAME1.2:約1,700本(20,000m)



ORNL Jaguar and Tsubame 2.0
Similar Peak Performance, 1/4 the Size and Power

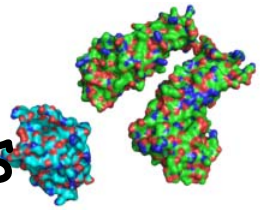


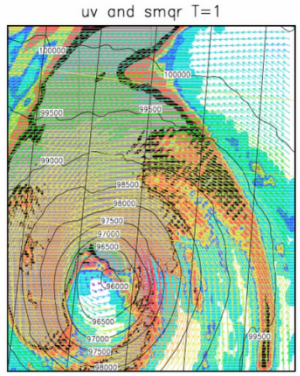
TSUBAME2.0（システムスタック）



TSUBAME2.0 アプリケーション性能予測

- 1.192 TFlops Linpack [IEEE IPDPS 2010]
  - Top ranks Green 500?
- ~0.5 PF 3D Protein Rigid Docking (Node 3-D FFT) [SC08, SC09]
- 145Tlops ASUCA Weather Forecast [SC10 Best Student Paper Finalist]
- Multiscale Simulation of Cardiovascular flows [SC10 Gordon Bell Finalist]
- Various FEM, Genomics, MD/MO (w/IM Apps: search, optimization, …





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

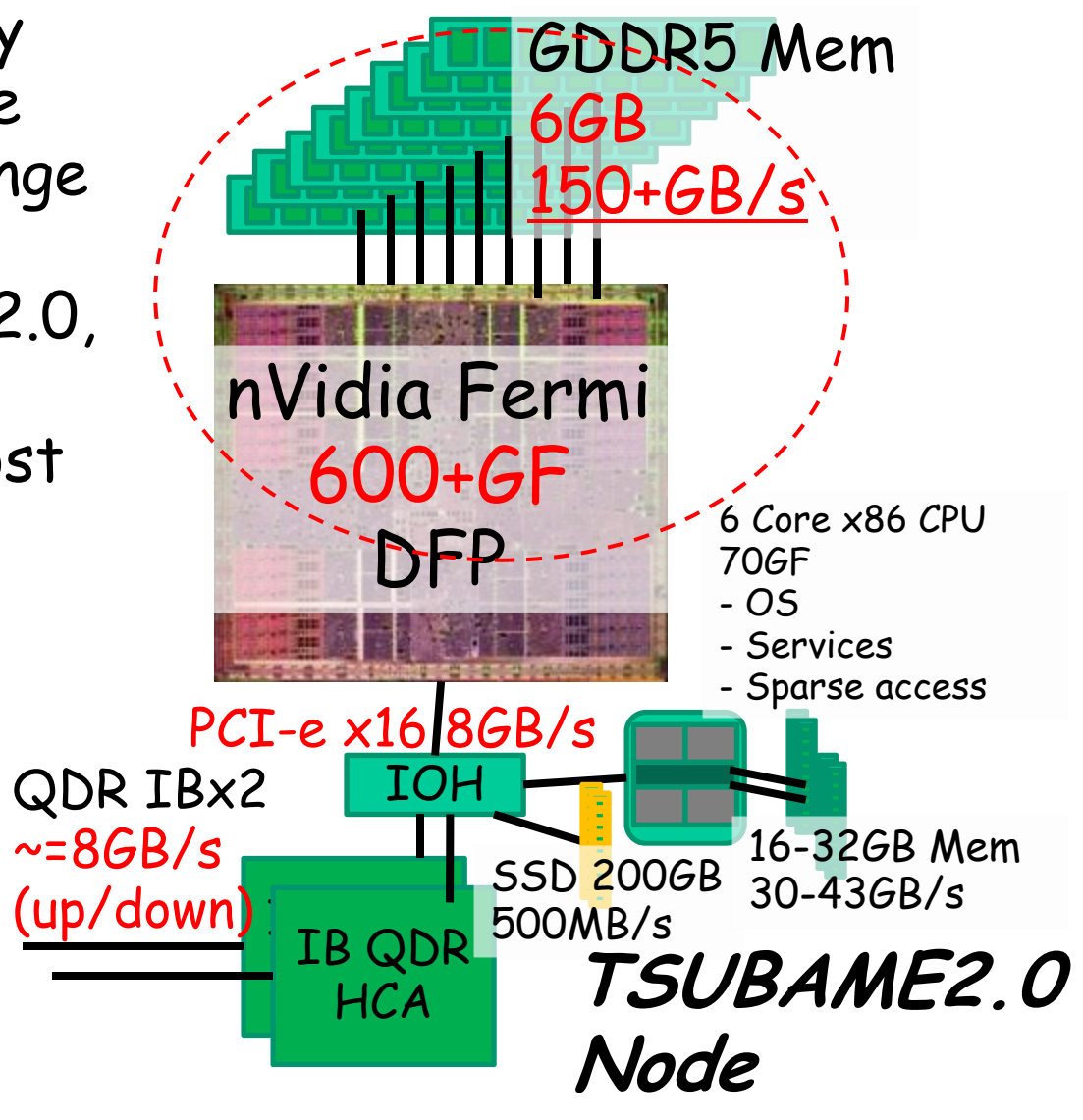
200m2

The "IDEAL TSUBAME2.0"

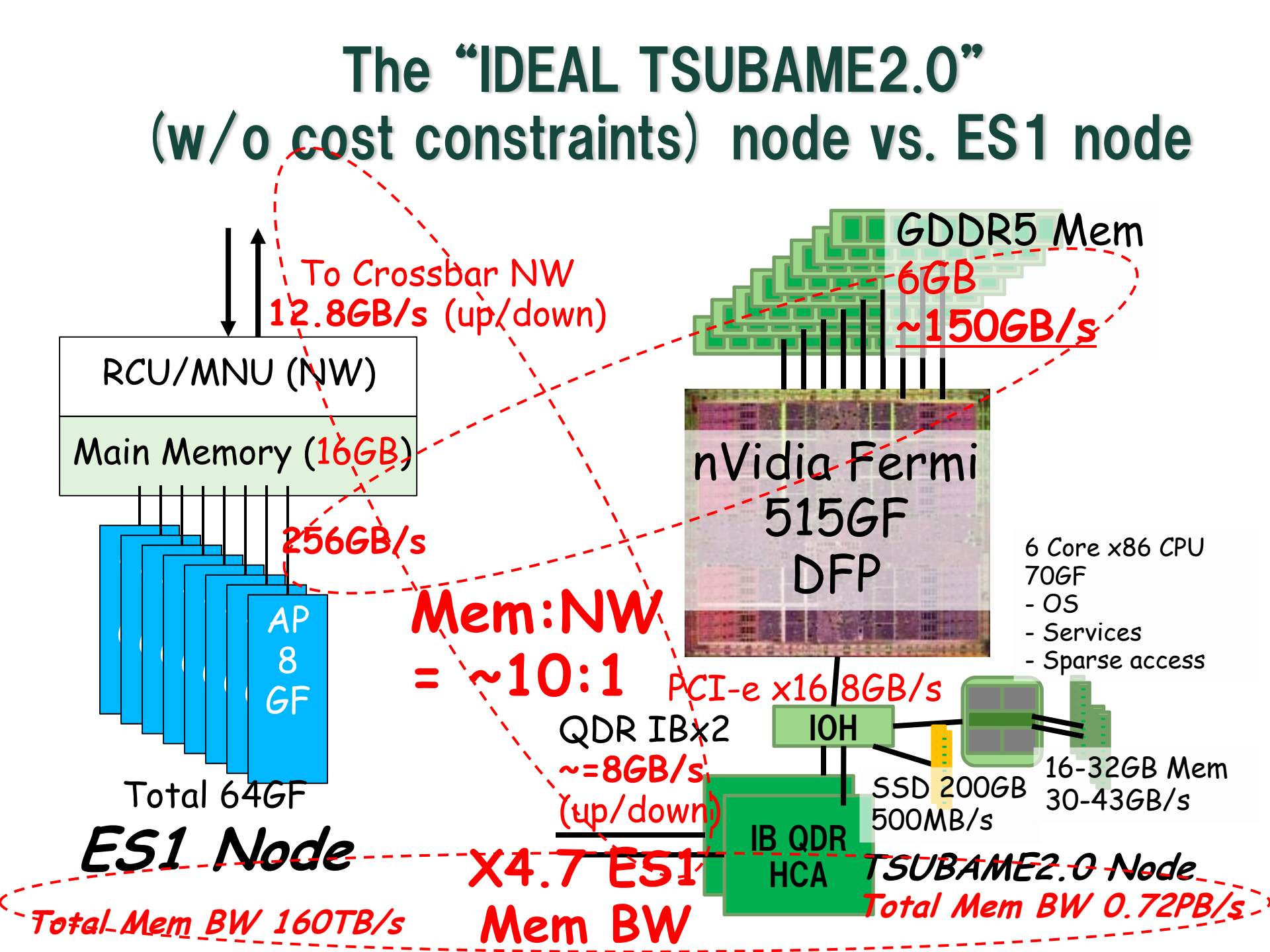
- What are architecturally possible without excessive design, power, or SW change

- In the REAL TSUBAME2.0, will have to compromise various parameters for cost and other reasons

- Almost Equal to "High Bandwidth" TSUBAME2/SL390 Config



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



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

DDR3-1066 Mem
16GB 64GB/s

Fujitsu Venus
128GF
DFP

Bytes/Flop
=0.3~0.5

GDDR5 Mem
6GB
~150+GB/s

nVidia Fermi
500+GF
DFP

6 Core x86 CPU
70GF
- OS
- Services
- Sparse access

PCI-e x16 8GB/s

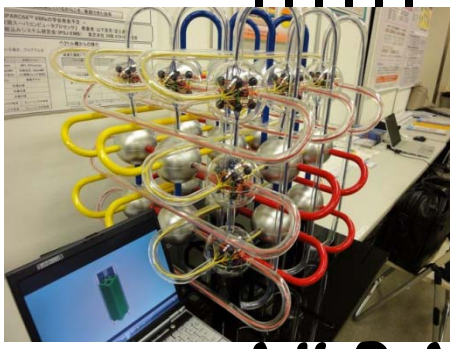
IOH

16-32GB Mem
30-43GB/s

SSD 200GB
500MB/s

QDR IBx2
~=8GB/s
(up/down)

IB QDR
HCA

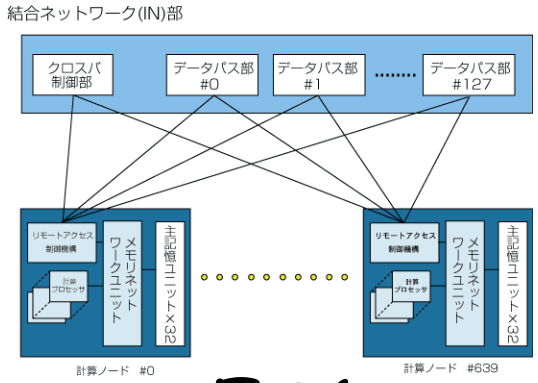


6-D Torus
5GB/s / link
up to 4
simultaneous
transfers

NLP Node

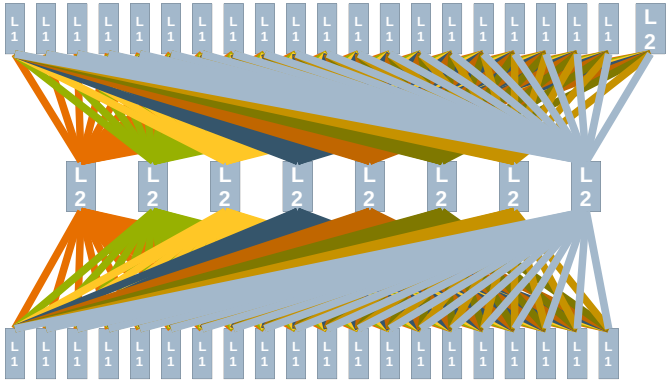
TSUBAME2.0
Node

Comparing the Networks



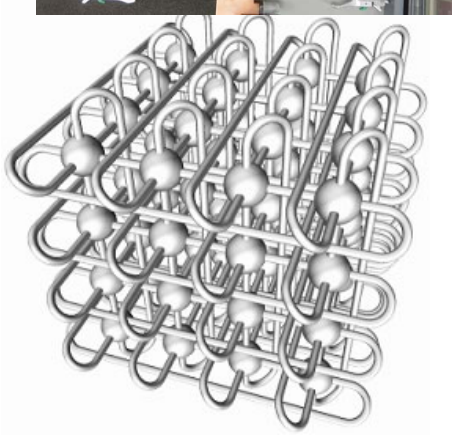
ES1

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 ΣBW : TSUBAME2 ΣBW = 1 : 6
  - ⇒ BW-bound apps (e.g. CFD) should scale equally on both w.r.t. ΣBW (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)

TSUBAME2.0世界ランキング
スパコン二大リスト (2010年11月)

The Top 500 (ベンチマーク絶対性能、ペタフロップス)

1位: 2.566 中国防衛大 Tianhe 1-A (11)
2位: 1.758 :米国オークリッジ国立研究所 Cray Jaguar (81)
3位: 1.271 :中国深圳国立スパコンセンター Dawning Nebulae (13)
4位: 1.192 :日本 東工大/HP/NEC TSUBAME2.0 (2)
5位: 1.054 :米国ローレンスバークレー国立研究所 Cray Hopper (30)
6位: 1.050 :仏CEA国立研究所 Bull Bullx (97)
7位: 1.042 :米国オークリッジ国立研究所 IBM Roadrunner (16)
33位(日本2位): 0.1914:日本原子力研究開発機構/富士通 (95)

(Green500 rank)

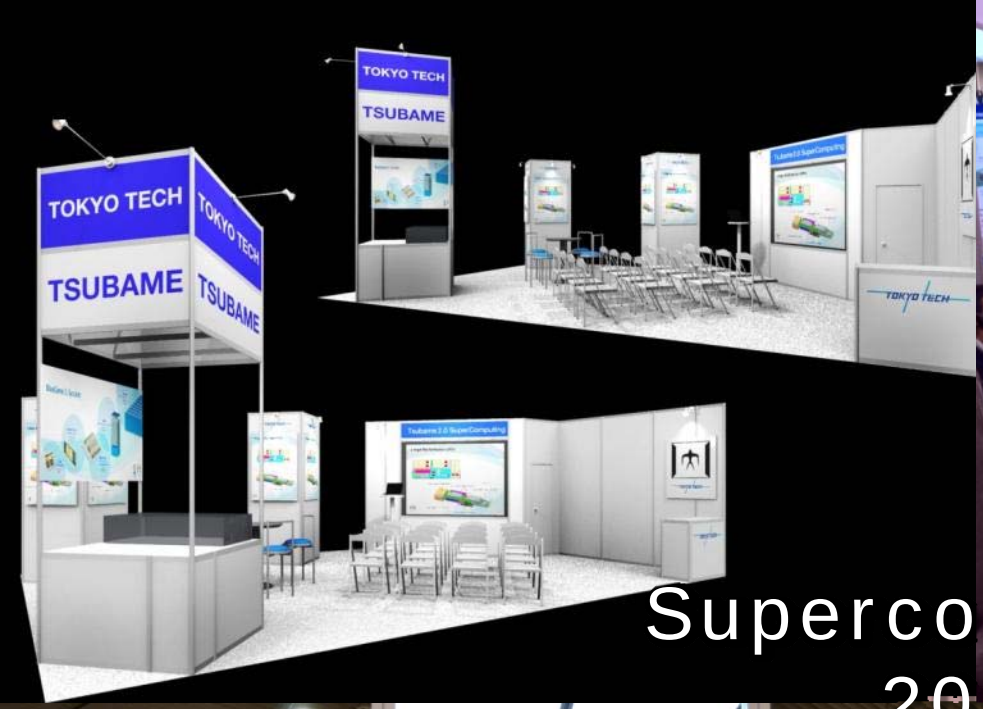
The Green 500 (ベンチマーク電力性能、メガフロップス/W)

1位: 1684.20 : 米国 IBM研究所 BlueGene/Q プロトタイプ (116)
2位: 958.35 :日本 東工大/HP/NEC TSUBAME2.0 (4)
3位: 933.06 :米国 NCSA Hybrid Cluster実験機 (403)
4位: 828.67 : 日本 理研 京 (170)
5-7位: 773.38 : ドイツ ユーリッヒ大等 IBM QPACE SFB TR (207-209)
10位(日本3位): 636.36 :日本 環境研 (102)

(Top500 rank)

"Little Green 500" では TSUBAME2.0の実験構成が
1.037 Gigaflops/W 達成 (米Microsoftとの共同研究)


THE GREEN 500™
sponsored by
SUPERMICRO
This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.
GSIC Center, Tokyo Institute of Technology
Is recognized as the
Greenest Production Supercomputer in the World
on the world's Green500 List of computer systems as of
November 2010
Wu-chun Feng, Co-Chair
Kirk Cameron, Co-Chair



Supercomputing 2010 @ New Orleans 東工大ブース

GPUで最速スパコン

で高速計算

中国のスーパーコンピューターが計算速度で初めて世界一になったスパコンランキング「TOP500」では、中央演算処理装置（CPU）に加え、「GPU」＝🔑＝と呼ばれる画像処理装置を多く積んだスパコンが、上位5台のうち3台を占めた。高速の割に小型で消費電力も少ないのが特徴だ。スパコンは大規模化が限界に来ており、GPUや専用演算素子を駆使した新世代へと、急速に多様化が進んでいる。（東山正宣）

世界一の「天河1号A」（中国国防科学技術大）は、2位の「ジャガー」（米オークリッジ研）の約1・5倍の計算速度をたたき出した。

天河の部品のほとんどは米国製だ。TOP500などによると、米インテル製のCPU1万4千個と、米エヌビディア製のGPU7千個からなる。これまでのほとんどのスパコンがCPUだけを使っていたのに対して、天河はGPUをCPU代わりに使っているのが特徴だ。

本職は画像処理

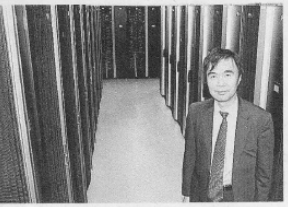
3位の星雲（中国深圳スパコンセンター）や4位のツバメ2・0（東工大）も構成や部品はほとんど同じ。3台の計算速度の違いはCPUとGPUの数の違いを反映している。東工大の松岡聡教授は「数を増やしても、使いこなして性能を出すのは簡単ではない。中国もなかなかやるようになってきた」と話す。

GPUの本来の仕事である画像処理では、簡単な計算を数多くこなす必要がある。このため、GPUには単純な計算を担う小さな演算素子がたくさん詰め込まれている。一方、CPUは条件ごとに場合分けをするなど複雑な計算をさせるため、それに対応した大型の素子を積んでいる。二つを使い分け、計算をうまく割り振れれば効果は絶大だ。

エヌビディアによると、天河の計算速度（1秒間に2566兆回）の8割をGPUが担っているという。同社日本法人の林憲一マーケティングマネジャーは「CPUだけで同じ計算速度を出そうとすると5万個必要になる。床面積は倍、消費電力は3倍になっていただろう」と試

🔑 **GPU**

Graphics Processing Unit(画像処理装置)の略。コンピューターの画面に写真や動画を表示する装置。ゲームの3D表示やハイビジョンが当たり前になった2000年代後半から急速に高性能化が進んだ。最新の機種では切手大の中に512個



開発費も安く

市販品を使っているので開発費も安い。天河は約80億円、ツバメは30億円だった。2002年に世界一をとった「地球シミュレータ」が600億円、理化学研究所が建設している「京」が、建屋も含めて1100億円かかるのに比べ、けた違いに安い。

市販化も早い。ツバメを開発したNECはすでに部品ごとのリース販売を始めた。ツバメは大型冷蔵庫ほどのラック50台からなる。その1台ずつが地球シミュレータと同等の計算速度を持つ。価格は3年で6千万円ほど。02年から8年間で大きさも価格も千分の1になった。

クラウドで普及、省エネに貢献

計算速度を競う「TOP500」と並んで注目されるコンテストが、18日に発表される「GREEN500」だ。TOP500に選ばれたスパコン500台が消費電力あたりの計算速度を競う。2007年に始まった。

松岡教授によると、世界の電力の1割程度がIT（情報技術）関連で使われ、そのうちスパコンの消費電力は4〜5%程度。この割合は上がると予想される。クラウドコンピューティングを通じてスパコンが普及するからだ。

クラウドコンピューティングでは、計算を手元のパソコンで

ーにやってもらう。ネットの動画閲覧サイト「ユーチューブ」が代表的だ。膨大な数のパソコンから任された計算を、スパコンが担う流れが進んでいる。

省エネ性能の進歩は、スパコンがパソコンを上回る。ツバメ2・0の先代機種は、同じ計算をするのに必要な電力が、パソコンよりも多かったが、今は効率が20倍上がり、パソコンの3分の1になった。クラウド化が進めばITの電力消費が減らせると、松岡教授はみる。

GPUを使ったツバメ2・0は、GREEN500で上位入りが確実視されている。

朝日20101119

省エネスパコン 日本2位

東工大「ツバメ」、理研4位

スーパーコンピューターの省エネ性能を競う世界ランキングが18日に発表され、東京工業大の「ツバメ2・0」が2位、理化学研究所が神戸に建設している「京」が4位になった。1位は米IBMが開発中の「ブルージーンQ」、3位は米国立スーパーコンピューター応用研究所の試験機。すでに運用されているスパコンとしてはツバメが世界一だった。

ランキングは「グリーン500」で消費電力当たりの計算速度を競う。1ワット当たりの計算速度は、ブルージーンが毎秒16億8400万回、ツバメが9億5800万回。ツバメは、計算速度を競う世界ランキング「TOP500」では4位だった。京は現在、全体の0・5%しかできていないが、高性能が示された。

スパコンは消費電力の問題で大型化が限界に近づいている。ブルージーンやツバメは従来の演算装置だけに頼らない新世代のスパコンだ。（東山正宣、ニューオーリンズ＝小宮山亮磨）

朝日20101119(夕刊)

スパコン省エネ性能

東工大、世界2位

日本製 10位内に3つ

米バージニア工科大学が集計したスーパーコンピューターの省エネ性能ランキング「グリーン500」で、東京工業大学のスパコン「TSUBAME2・0」が世界2位となった。日本のスパコンは10位内に3つがランクイン。演算性能の高さでは中国などに抜かれたが、省エネ技術で日本勢の強さを示した。

グリーン500は消費電力当たりの演算能力の高さを競う。1位は米IBM製のスパコンだが、同機種は試作機の扱いで、実際に運用しているスパコンの中では東工大のTSUBAMEが1位。消費電力を抑えやすい画像処理チップ（GPU）を採用したほか、冷却機構を工夫するなどして省エネ性能を高めた。

日本勢では、次世代スパコン「京(けい)」の試作機が4位、国立環境研究所のスパコンが10位に入った。

日経20101121

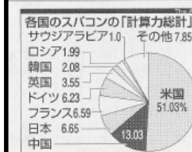
2010-11-18 朝刊 3面

中国スパコン 軍の影

計算速度 世界一

米、技術力を注

中国のスーパーコンピューターが計算速度のランキングで初めて世界一になった。上位を独占してきた米国との激しい開発競争が幕を開けた。日の丸スパコンの世界一奪還の夢は、遠のくばかりだ。



■スパコン世界ランキング

順位	国	名称	設置機関	計算速度
1	中	天河1号	国防科学技術大	2566
2	米	ジャガー	オークリッジ研究所	1759
3	中	星雲	深圳スパコンセンター	1271
4	日	ツバメ2.0	東京工業大	1192
5	米	ホッパー	ローレンス・バークリー研究所	1054
33	日	BX900	日本原子力研究開発機構	191
54	日	地球シミュレータ	海洋研究開発機構	122
169	日	京（開発中）	理化学研究所	48

米「TOP500」コンテスト（2010年11月）から。計算速度は兆回／秒

世界のスパコンの性能を調べる研究者組織「TOP500」が16日、米ニューオーリンズで開いた表彰式。頂点に立った「天河1号」を開発した中国国防科学技術大の劉光明教授が登壇し、「天河1号は中国の科学、技術、経済の発展を加速する」と語った。計算速度を合計した計算力総計でも、中国は世界全体の13%を占めて2位に入った。

中国中央テレビは17日夜、全国ニュースのトップで表彰式を取り上げたほか、人民日報も16日、「中国の速度が米国を驚愕させる」とする記事を掲載し成果をたたえた。

「世界一」は中国の科学技術力の向上を意味するだけでなく、応用を通じて産業などへ与える影響も大きい。天河1号は1秒間に約2500兆回計算する能力がある。より現実に近い条件でシミュレーションができる。劉教授は用途に石油探査、医学、新素材開発、気候予測、金融リスク分析など10項目をあげた。

巨大な計算力は軍事と切り離せない。国防科技大は人民解放軍直属の研究機関で、学生や研究者の大部分を軍人が占める。劉教授は朝日新聞の取材に軍事利用を否定したが、開発の背景に軍事がある

とみる関係者は少なくな

世界一を奪われたと

え、なお圧倒的な優位に

米国の受け止め方はどう

「TOP500」を運営

テネシー大のジャック

朝日20101118

ペタフロップス？ギガフロップス/W?

6.6万倍高速
3倍省エネ
<<
4.4万倍データ

Laptop: SONY Vaio type Z (VPCZ1)
CPU: Intel Core i7 620M (2.66GHz)
MEMORY: DDR3-1066 4GBx2
OS: Microsoft Windows 7 Ultimate 64bit
HPL: Intel(R) Optimized LINPACK Benchmark for Windows (10.2.6.015)
256GB HDD

18.1 ギガ(10^9)フロップス
369 メガ(10^6)フロップス / Watt

Supercomputer: TSUBAME 2.0
CPU: 2714 Intel Westmere 2.93 Ghz
GPU: 4071 nVidia Fermi M2050
MEMORY: DDR3-1333 80TB + GDDR5 12TB
OS: SuSE Linux 11 + Windows HPC Server R2
HPL: Tokyo Tech Heterogeneous HPL
11PB Hierarchical Storage

1.192 ペタ(10^{15})フロップス
1037 メガ(10^6)フロップス / Watt

気象計算

GP GPU

気象庁数値予報課との共同研究:

メソスケール大気シミュレーション:

雲解像非静力平衡モデル

Compressible equation taking consideration of sound waves.

Meso-scale

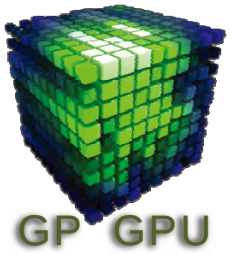
2000 km

台風

1 km 以下

竜巻、集中豪雨、ダウンバースト

WRF の GPU Computing



WRF (Weather Research and Forecast)

オープンソース・コミュニティコード（NCAR, NCEP, OU, NOAA/FSL, AFWA）

WSM5 (WRF Single Moment 5-tracer) Microphysics*

Represents condensation, precipitation and thermodynamic effects of latent heat release

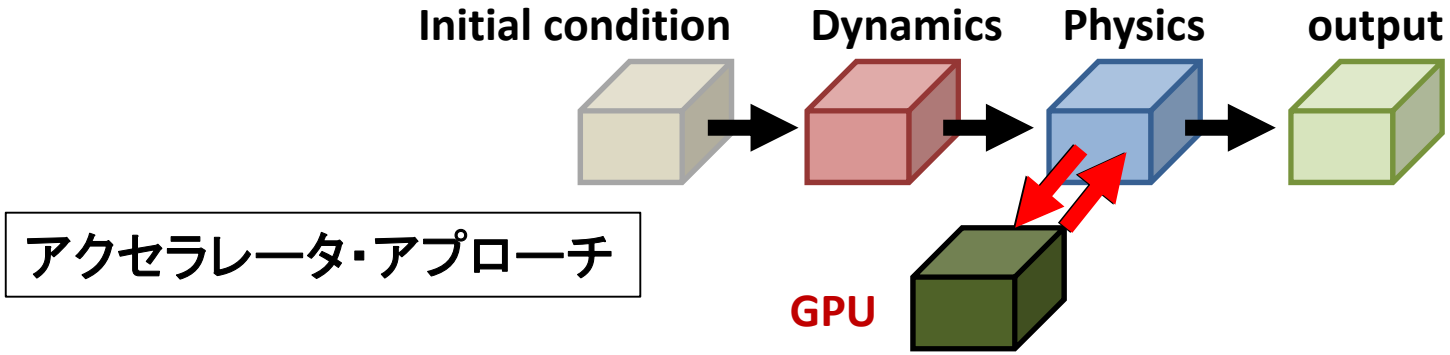
1 % of lines of code, 25 % of elapsed time

⇒ 20 x boost in microphysics （1.2 - 1.3 x overall improvement）

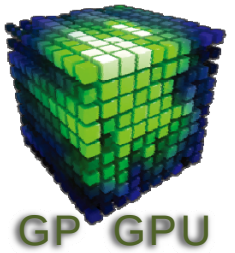
WRF-Chem**

provides the capability to simulate chemistry and aerosols from cloud scales to regional

⇒ x 8.5 increase



気象庁の次期気象コードASUCA



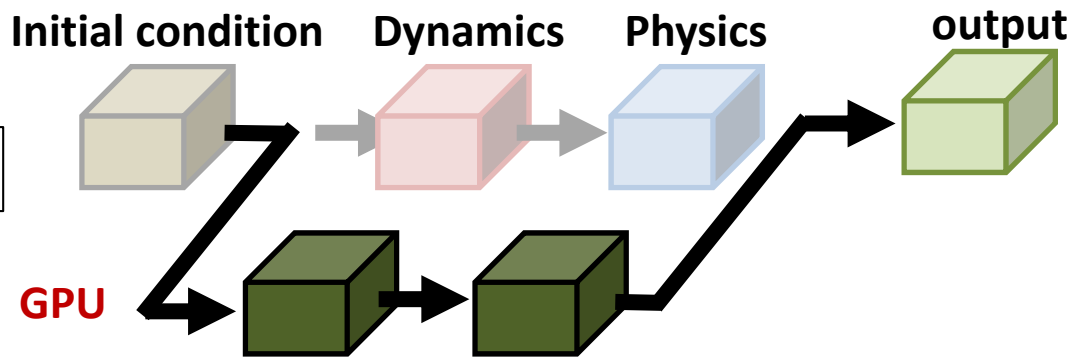
■*ASUCA Production Code*

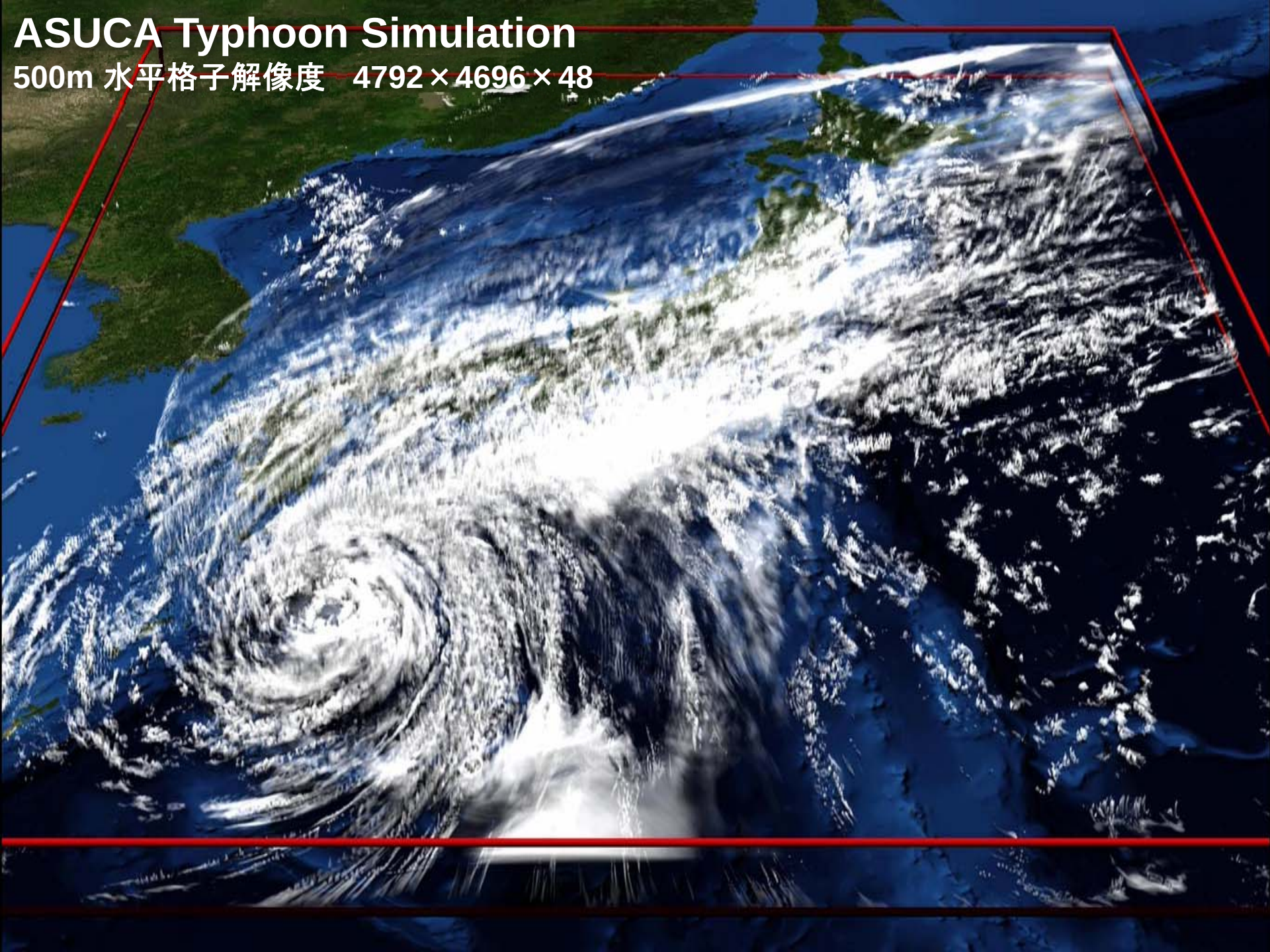
- ✓ *A next-generation high resolution weather simulation code that is being developed by Japan Meteorological Agency (JMA)*
- ✓ *ASUCA succeeds the JMA-NHM as an operational non-hydrostatic regional model at JMA*

■*Similar Structure as WRF*

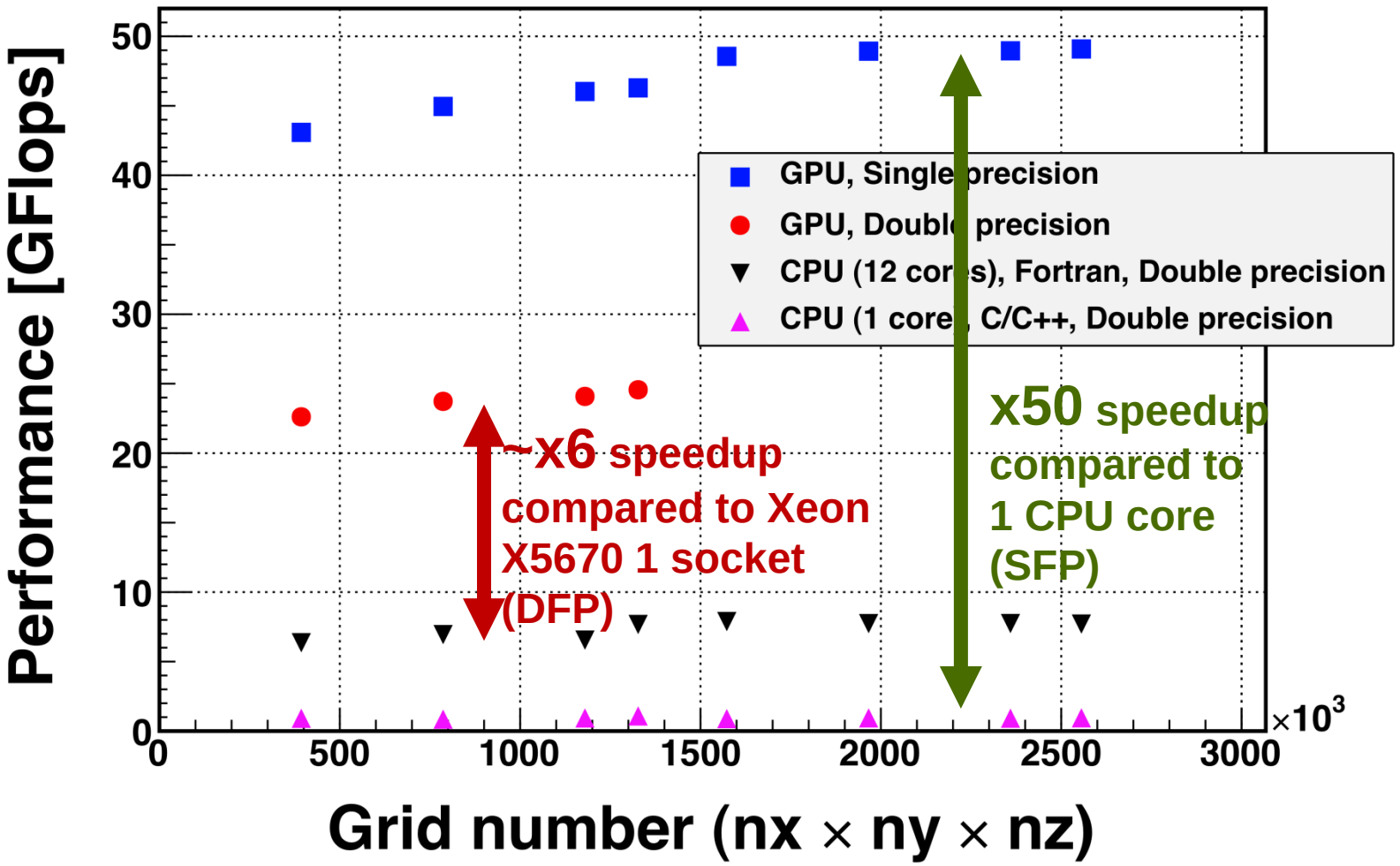
- ✓ *HEVI (Horizontally explicit Vertical implicit) scheme*
- ✓ *Dynamical Core uses a numerical scheme with 3rd-order accuracy in time and space*
  - *Flux-form non-hydrostatic compressible equation*
  - *Generalized coordinate*

フルGPU アプローチ

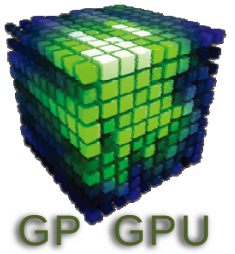



ASUCA Typhoon Simulation
500m 水平格子解像度　4792×4696×48

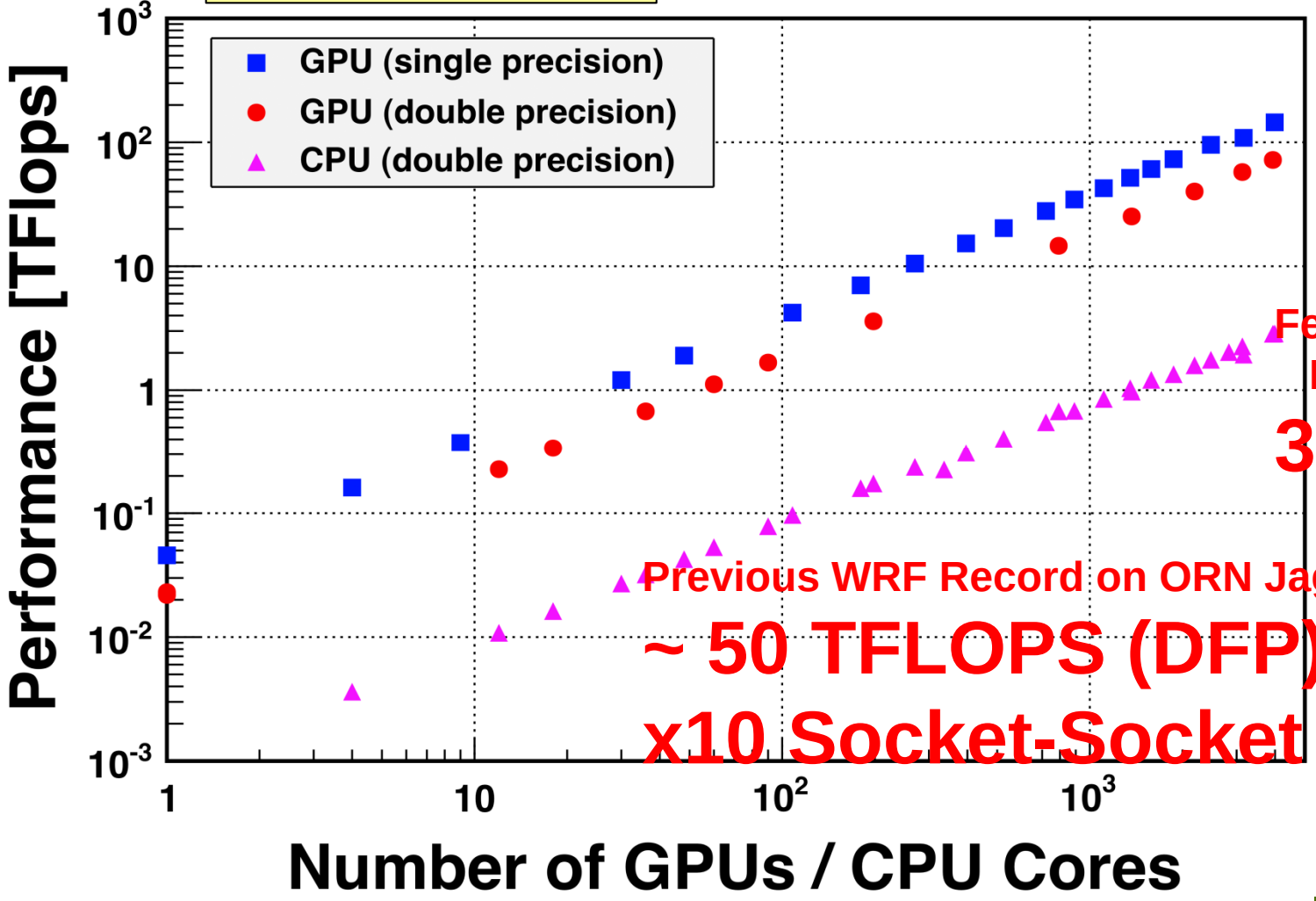
TSUBAME 2.0 (1 GPU)



TSUBAME 2.0での実行性能



弱スケーリング



145.0 Tflops
Single precision

76.1 Tflops
Doublele precision

Fermi core Tesla M2050

3990 GPUs

Previous WRF Record on ORN Jaguar

~ 50 TFLOPS (DFP)

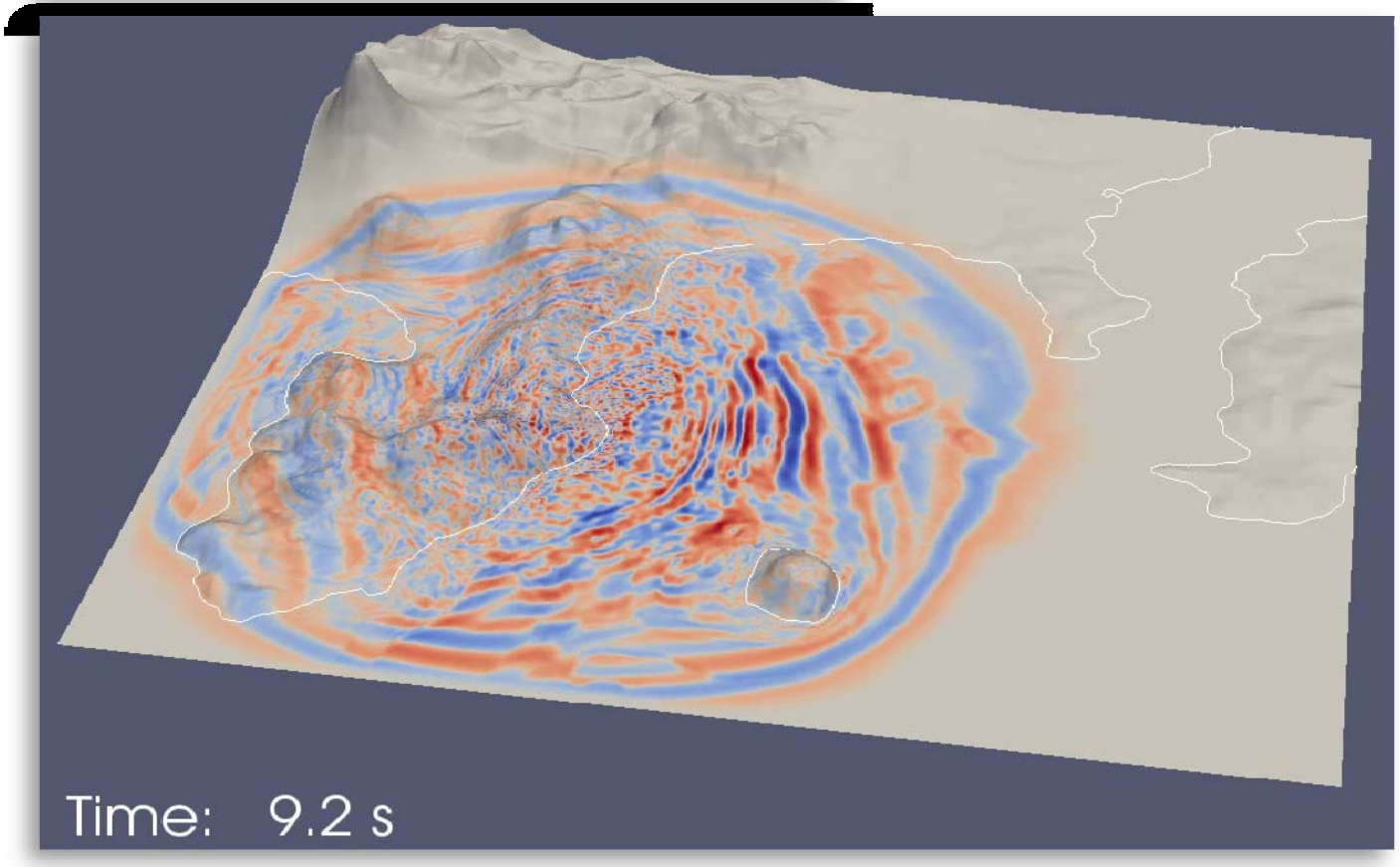
x10 Socket-Socket

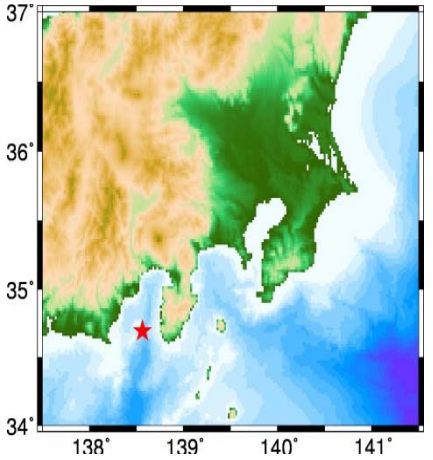
地震波伝播シミュレーション

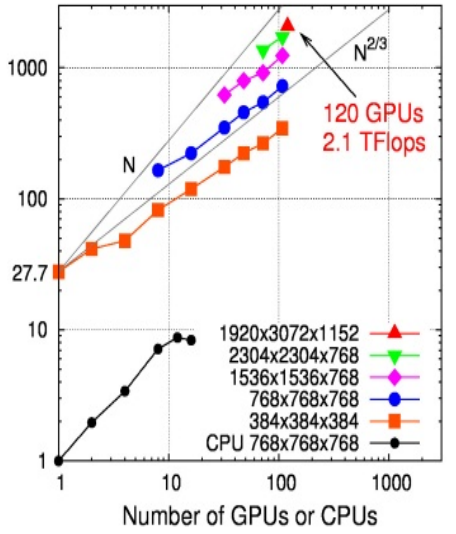
地球惑星科学専攻　岡元太郎氏　提供

GSIC
Global Scientific Information and Computing Center

TSUBAME 1.2 120 GPUs
for 1920x3072x1152　2.1 TFlops

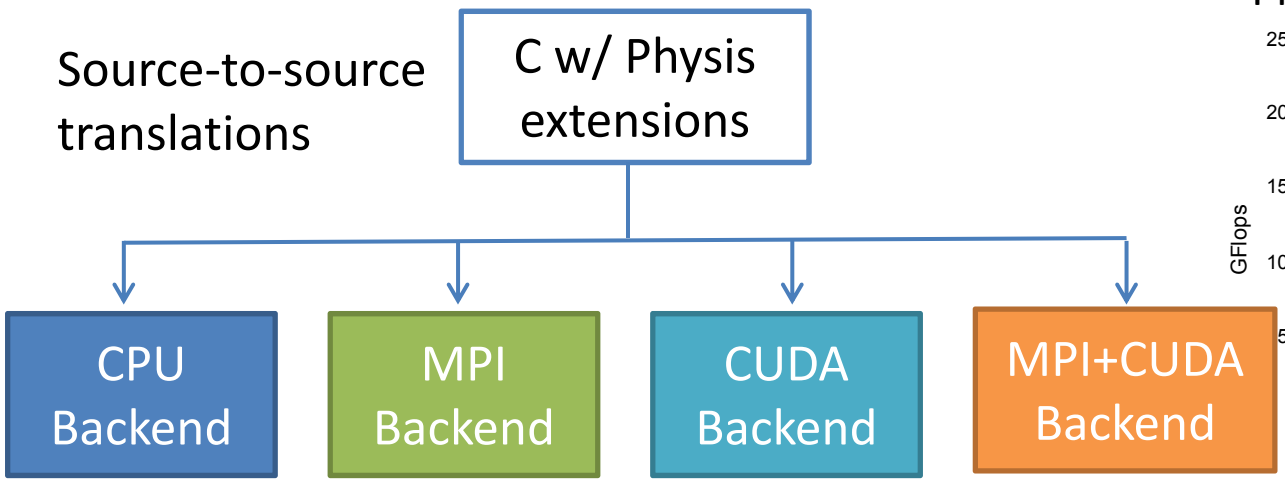




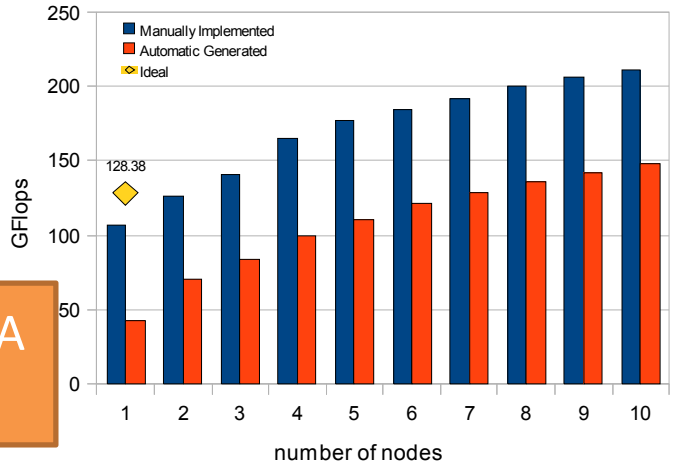


Physis: A Domain Specific Framework for Large-Scale GPU Clusters

- Portable programming environment for stencil computing on structured grids
  - Implicitly parallel
  - Distributed shared memory
- Provides concise, declarative abstractions for stencil computing

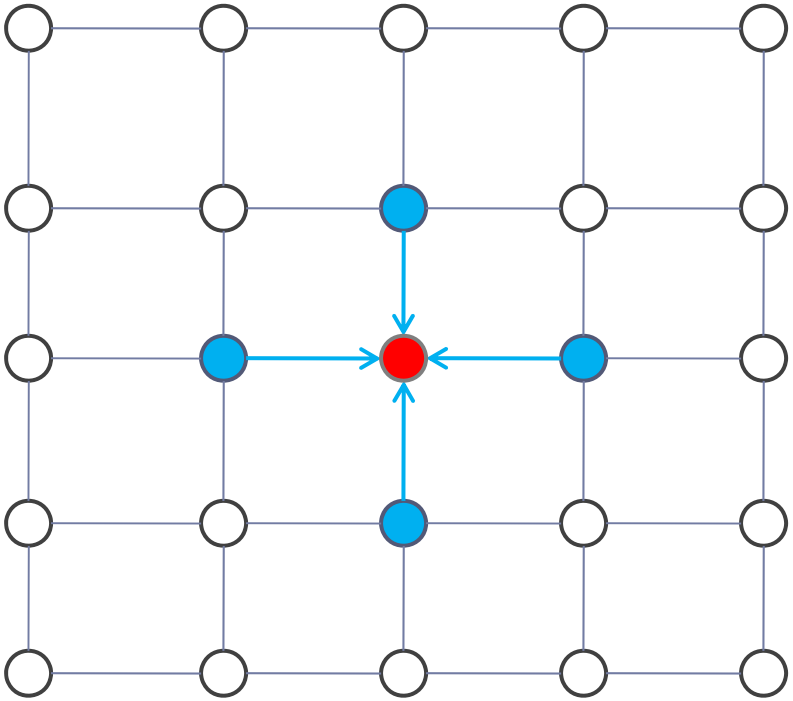


Preliminary Performance Results



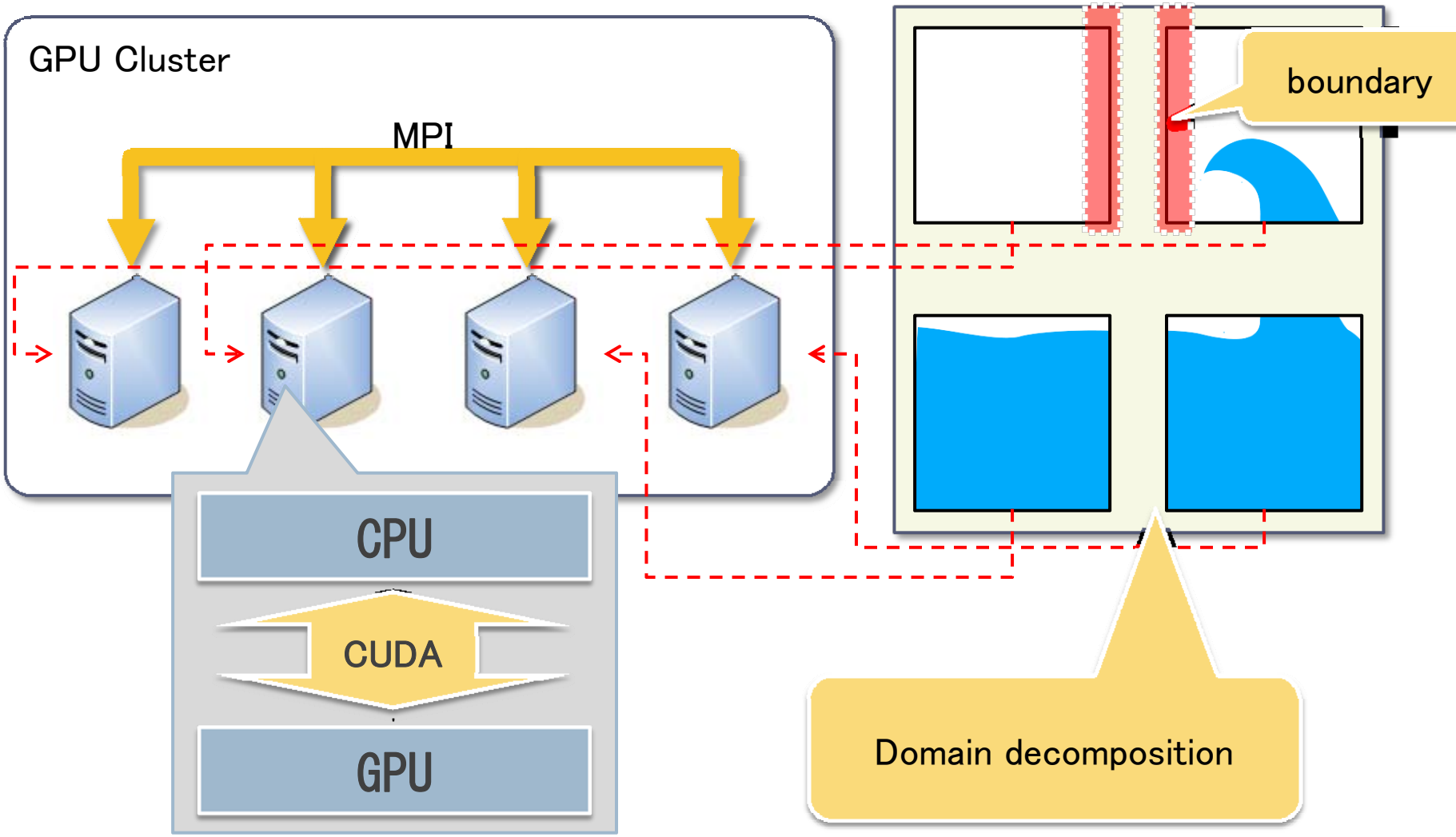
A five-point stencil

- T : time
- X,Y : space



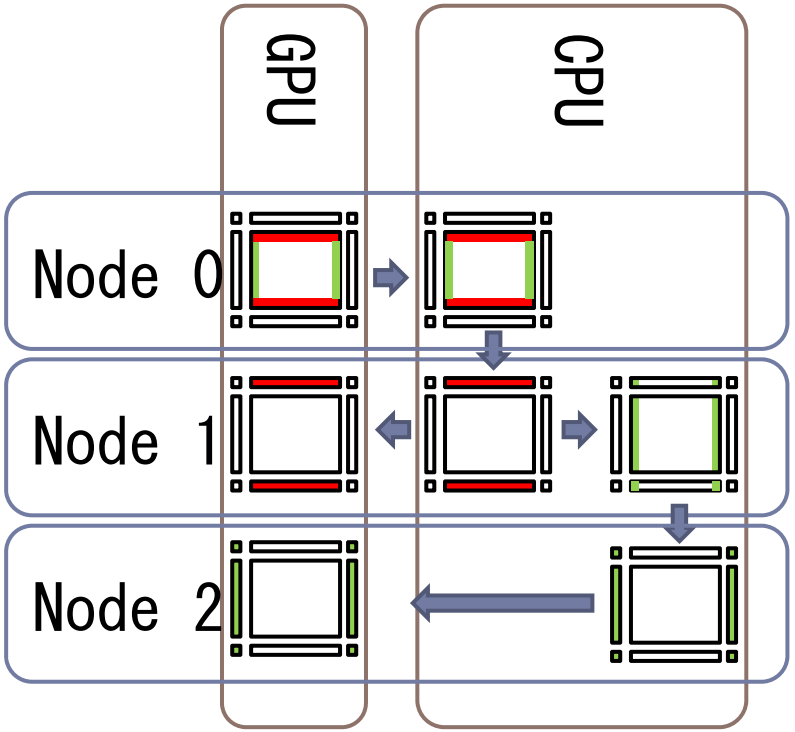
```
for (t = 0.0; t < T; t += dt) {
  for (y = 0; y < Y; y++) {
    for (x = 0; x < X; x++) {
      new_f(x,y) = e1*old(y,x)
       + e2*old(y,x-1) + e3*old(y,x+1)
       + e4*old(y-1,x) + e5*old(y+1,x);
    }
  }
}
```

Implementation on GPU clusters



Complexity of implementation

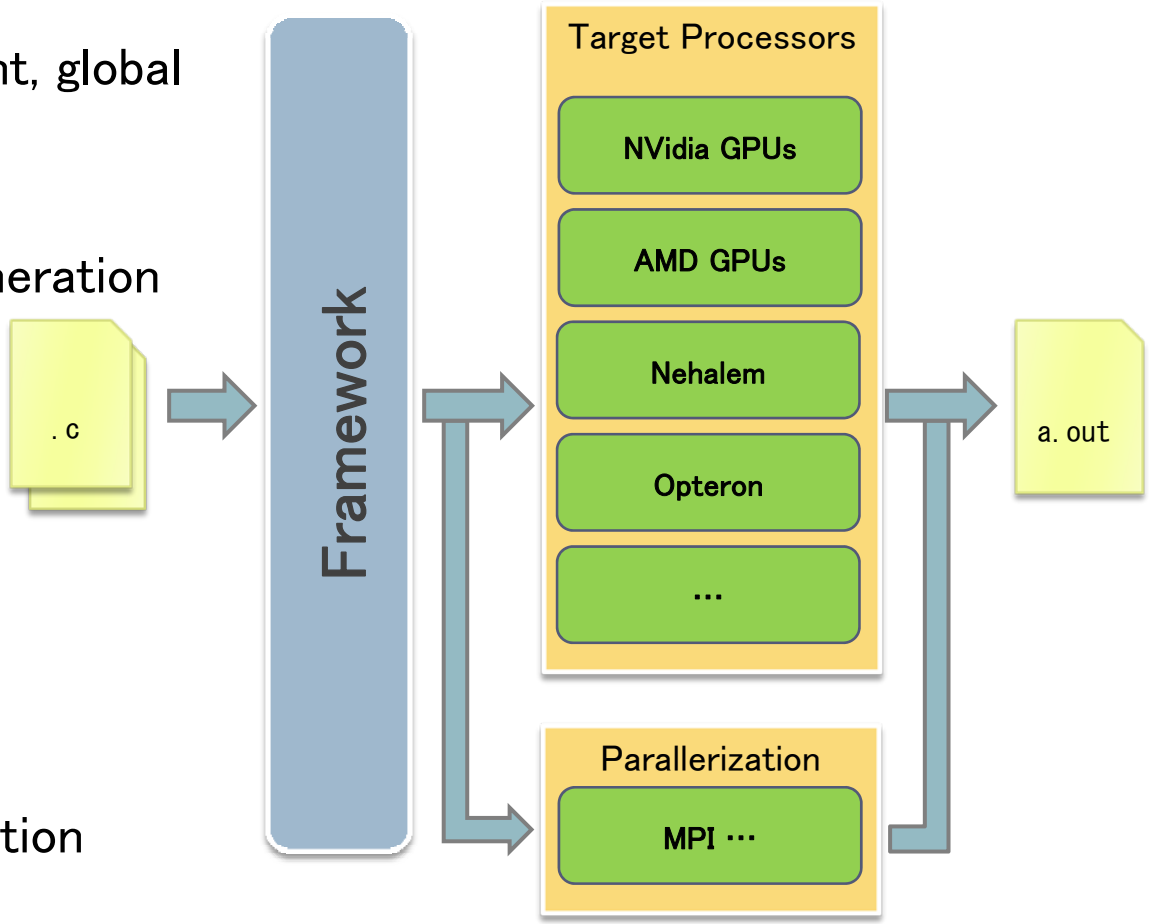
- CPU,GPU,MPI
- Code for omputation is concise, code for parallelization isn't.
  - Problem decomposition
  - Boundary exchange
  - GPUs cannot communicate directly
  - GPU->CPU -> CPU->GPU
- Code for optimization brings more complicacy
- Most difficult parts are non-essential for stencil computation



Procedure of boundary exchange

Overview of the framework

- Archtecture independent, global view description
- CPU,GPU,MPI code generation
- Code 2 Code
- Optimization
- Auto-tuning
- Checkpointing
- Optimal resource allocation



An example of a 7-point stencil

```
__stencil__  void average(int x, int y, int z, grid3d_real g) {
    float ret = psGridGet(g,0,0,0)
       + psGridGet(g,-1,0,0) + psGridGet(g,1,0,0)
       + psGridGet(g,0,-1,0) + psGridGet(g,0,1,0)
       + psGridGet(g,0,0,-1) + psGridGet(g,0,0,1);
    psGridEmit(g, ret / 7.0);
}
```

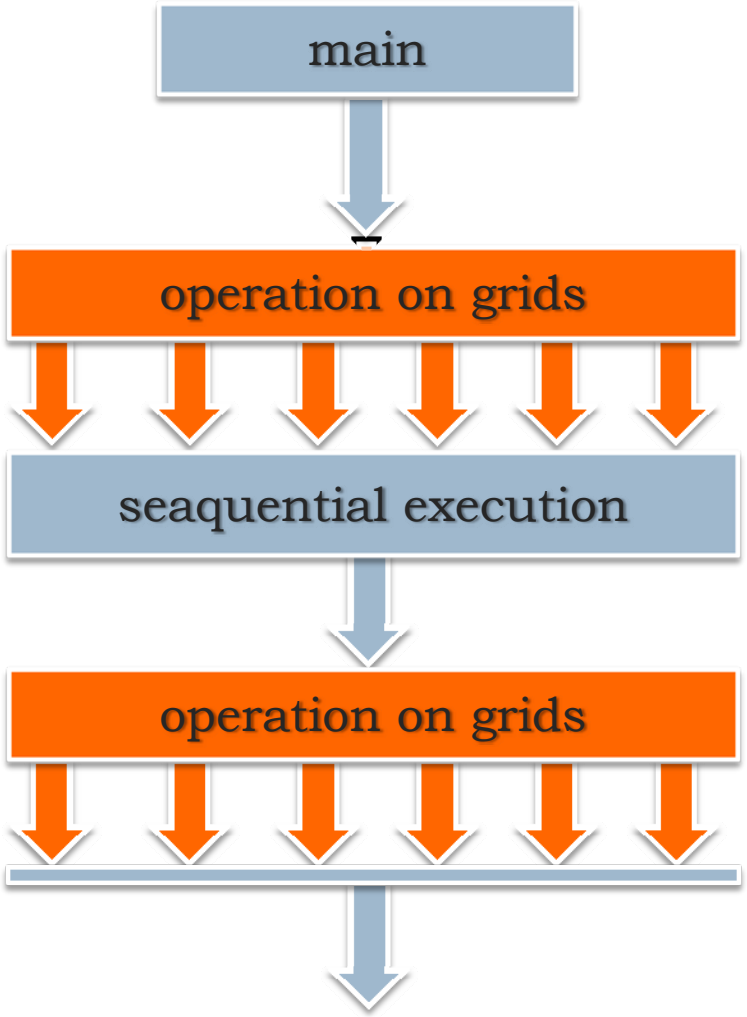
Parallel execution

```
void computation(float *inbuff, float *outbuff) {
  PS_Grid g = psGridNew(float, N,N,N);
    psGridCopyIn(g, inbuff);
    for (int t = 0; t < T; t += dt) psStencilMap(average, g);
    psGridCopyOut(g, outbuff);
}
```

Apply stencil kernel to a grid

Execution Model

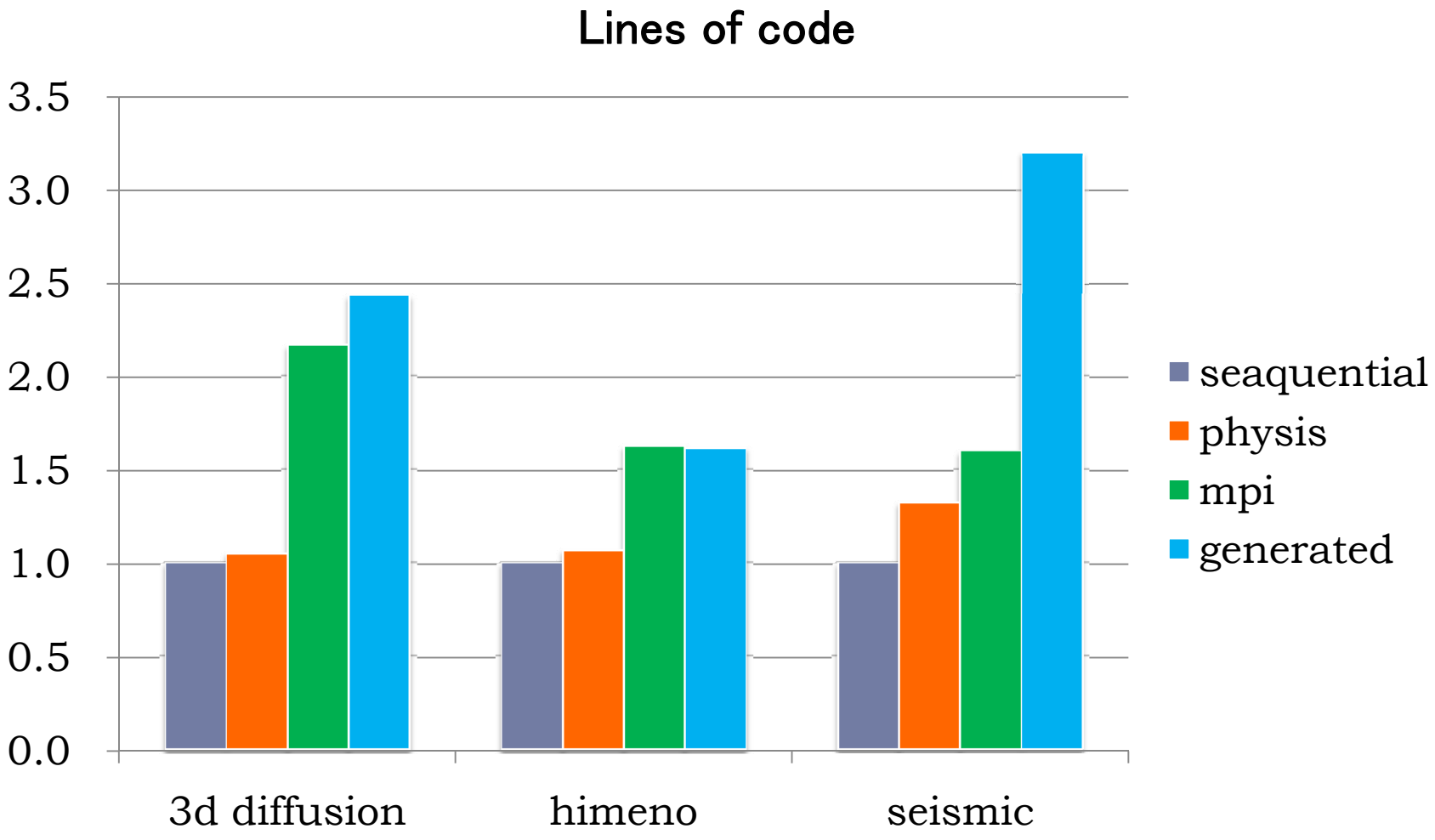
```
__stencil__
void kernel(int x, int y, int z, grid g) {
  float val = (psGridGet(-1,0,0)
            + psGridGet(1,0,0)) / 2.0;
  psGridEmit(g, val));
}
int main(int argc, char *argv[]) {
  psInit(&argc, &argv);
  int i;
  PS_Grid g =psGridNew(float, 256, 256, 256);
  printf( “start¥n” );
  int *buff = (int *)malloc(256*256*256*4);
  for (i = 0; i < 256; i++) {
    buff[i] = rand();
  }
  psGridCopyIn(g, buff);
  PS_Dom dom = psDom(0,255,0,255,0,255);
  for (i = 0; i < 100000; i++) {
    psStencilMap(kernel, dom, g);
  }
  psGridCopyOut(g, buff);
  printf( “end¥n” );
  psFinalize();
  reutrn 0;
}
```



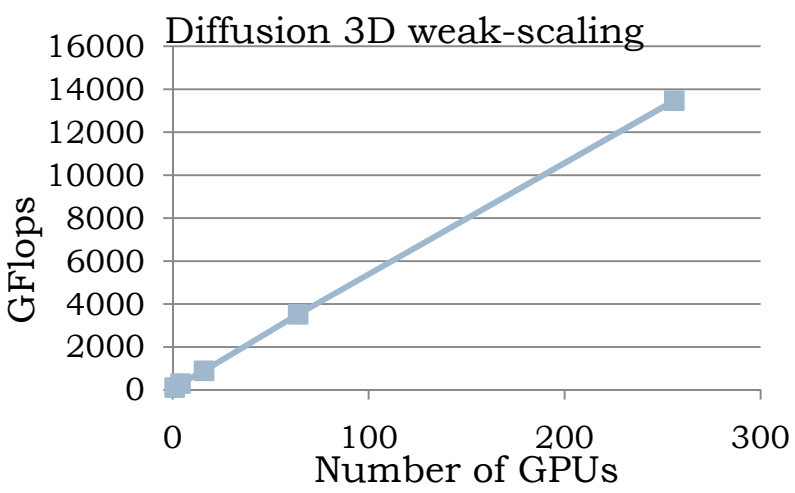
Evaluation

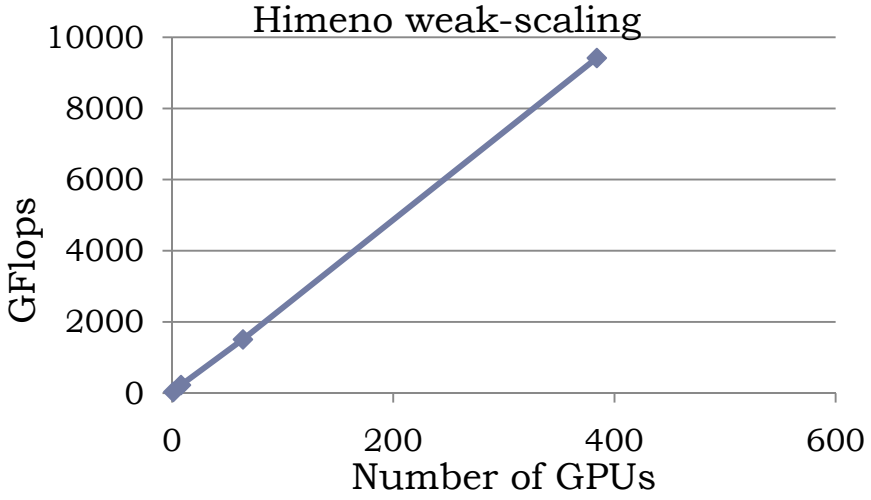
- Evaluated the performance and scalability of Physis on TSUBAME 2.0
  - 1D/2D decomposition
  - Overlapping of boundary exchange and computation
- Target applications
  - 3D diffusion equation
  - Himeno benchmark
  - Seismic wave simulation code
    - Free surface boundary condition is not supported

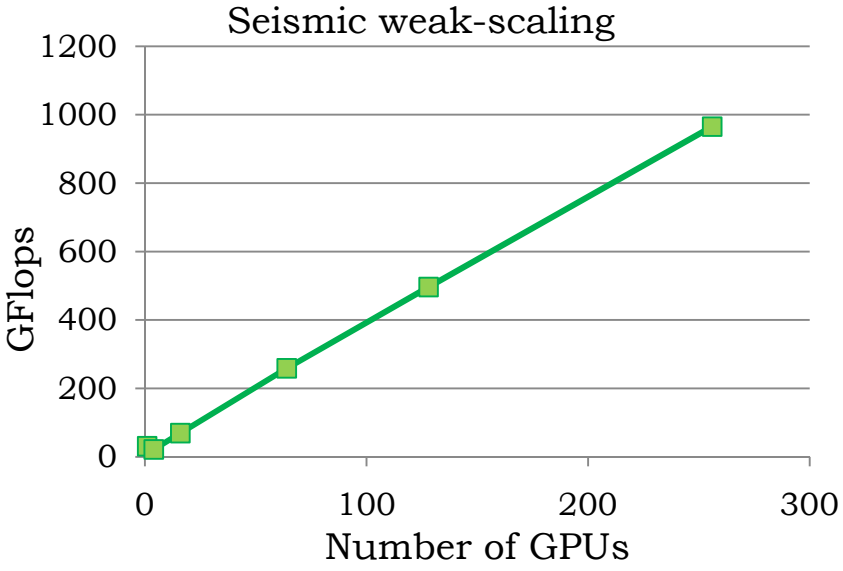
生産性的な


Lines of code
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3d diffusion
himeno
seismic
seaquential
physis
mpi
generated

Weak Scalability



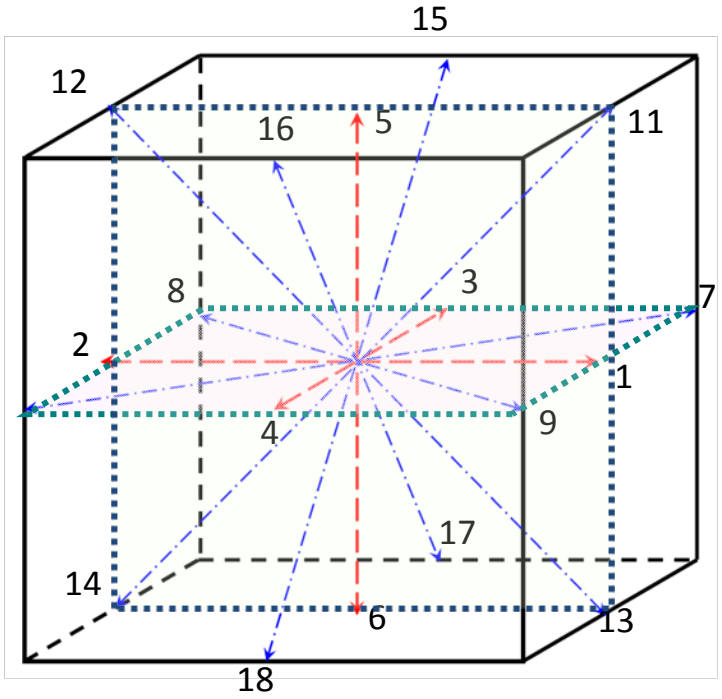




格子ボルツマン法による流体計算

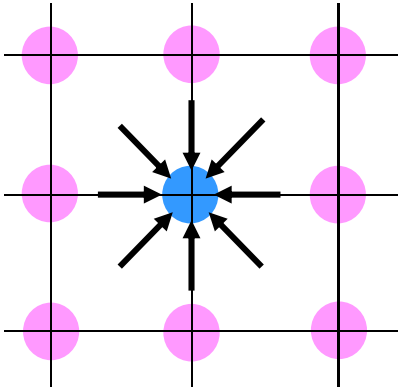
$$\frac{\partial f_i}{\partial t} + \mathbf{e}_i \cdot \nabla f_i = -\frac{1}{\lambda}\left(f_i - f_i^{eq}\right)$$

$$f_i^{eq} = \rho w_i \left[1 + \frac{3}{c^2}(\mathbf{e}_i \cdot \mathbf{u}) + \frac{9}{2c^4}(\mathbf{e}_i \cdot \mathbf{u})^2 - \frac{3}{2c^2}(\mathbf{u} \cdot \mathbf{u})\right]$$

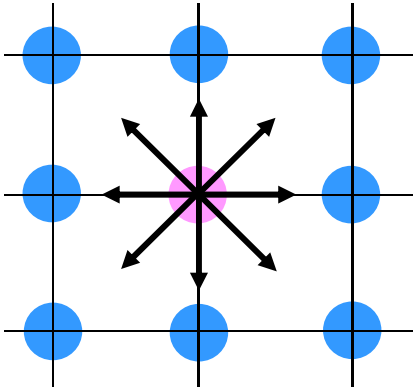


メモリアクセスが支配的な計算:

Collision step:



Streaming step:



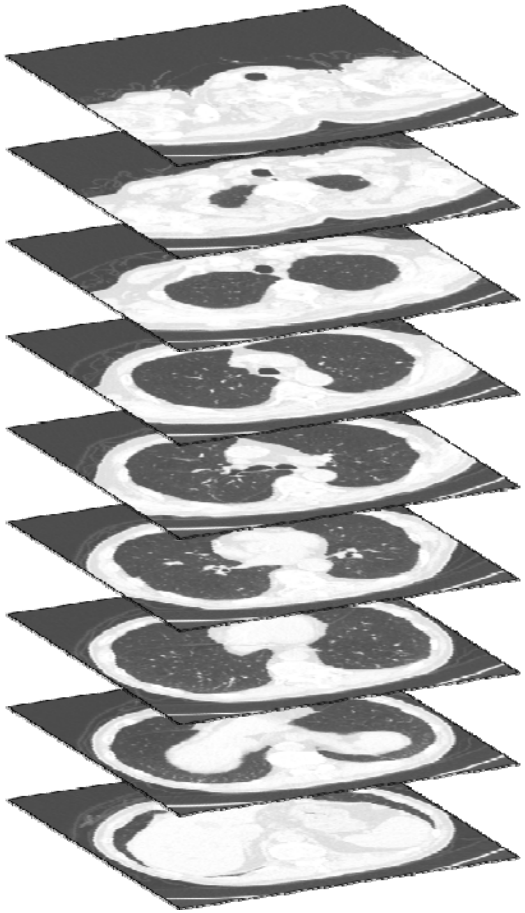
i is the value in the direction of *ith* discrete velocity
e_i is the discrete velocity set;
w_i is the weighting factor
c is the particle velocity
u is the macroscopic velocity

肺気管の呼気流解析

東北大学･医工系　山口研究室との共同研究

GP GPU

X線 CTスキャン
512×512×512



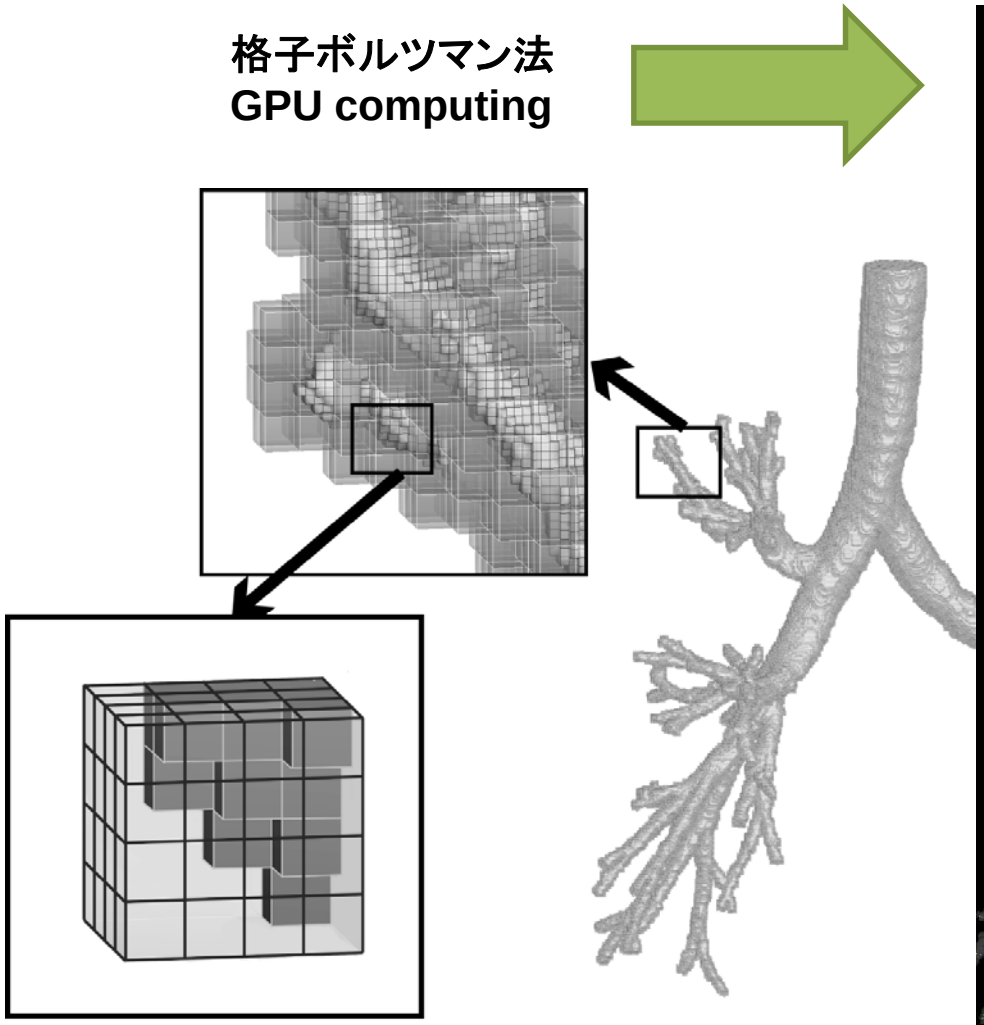
気管構造の抽出

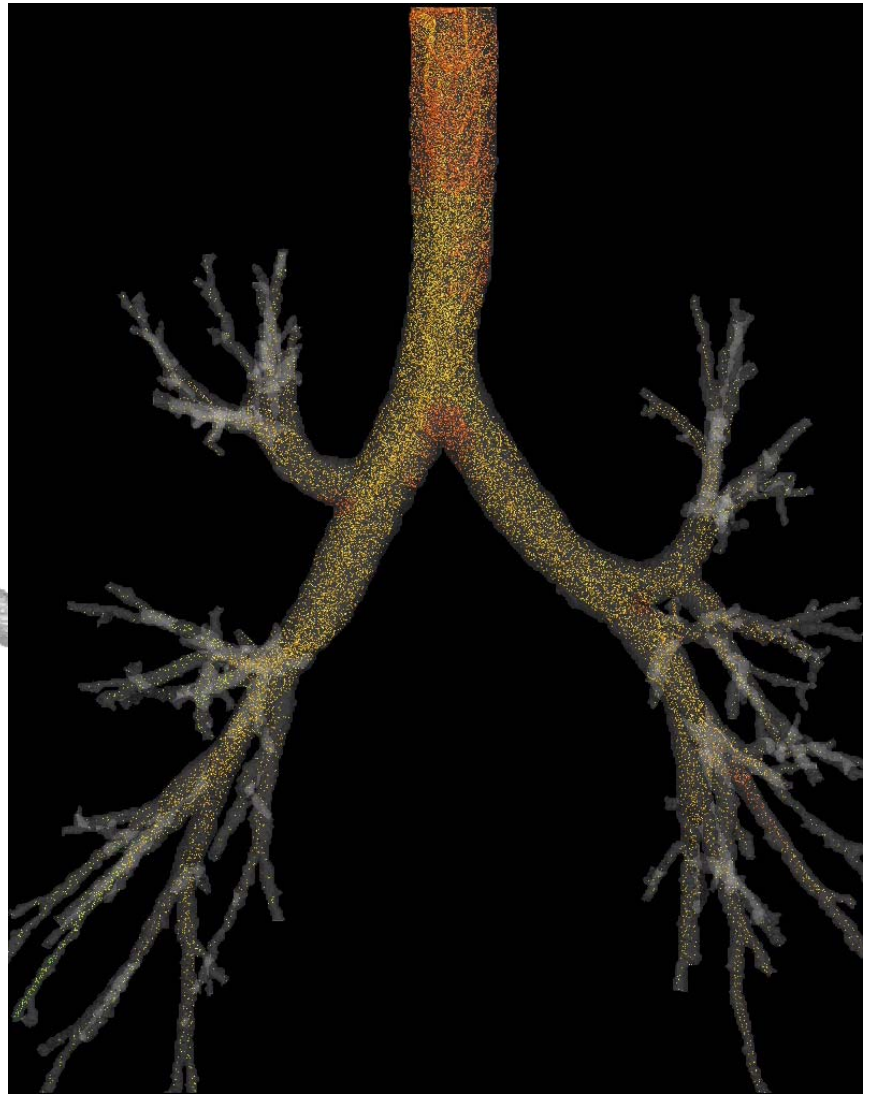
肺気管の呼気流解析

東北大学・医工系　山口研究室との共同研究

GP GPU

格子ボルツマン法
GPU computing





気液二相流

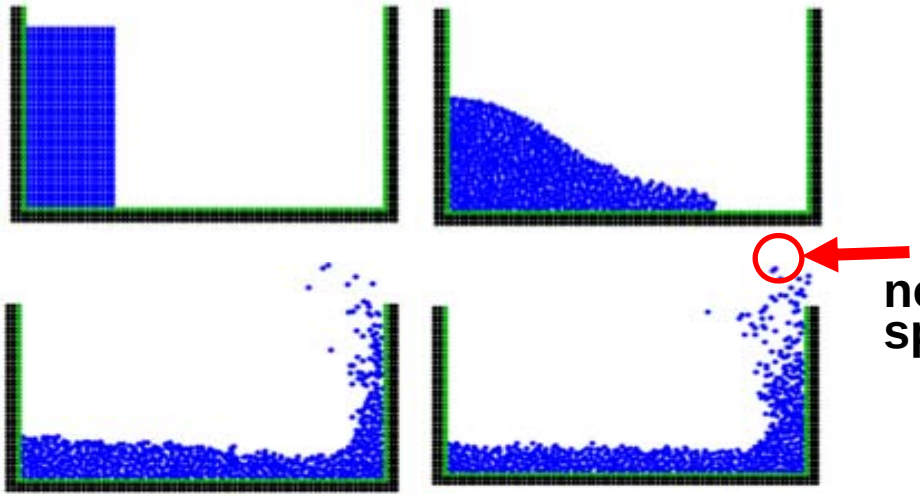
GP GPU

粒子法
例：SPH

Low accuracy
$< 10^6$ particles

Mesh Method

- Navier-Stokes solver：Fractional Step
- Time integration：3rd TVD Runge-Kutta
- Advection term：5th WENO
- Diffusion term：4th FD
- Poisson：AMG-BiCGstab
- Surface tension：CSF model
- Surface capture：CLSVOF(THINC + Level-Set)



not splash

Numerical noise and unphysical oscillation

High accuracy　$> 10^8$ mesh points

MUPHY : Multi-Physics Simulator

[Massimo Bernaschi]

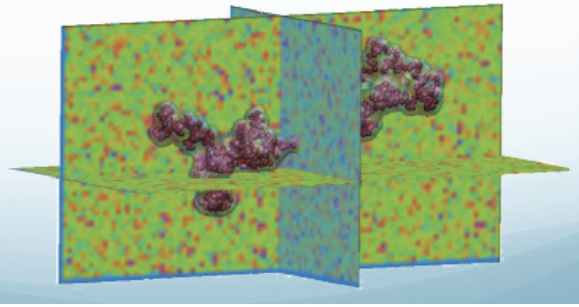
road spectrum of ***fluid-particle*** coupling mechanisms

brary of particle and molecular representations

brary of fluid types

o-polymers, colloidal suspensions, gels, biofluids, …

ulti-Platform

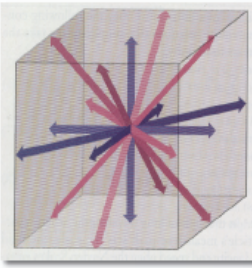


Blood Plasma via Lattice Boltzmann

From (minimal) Boltzmann equation

$$(\partial_t + \vec{v}\partial_x) f(\vec{x},\vec{v},t) = -\omega(f - f^{eq})(\vec{x},\vec{v},t)$$

Collision + Streaming of a set of discrete velocities



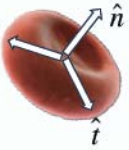
Superset of the Navier-Stokes dynamics

Exact streaming (no self-convected): uniform mesh

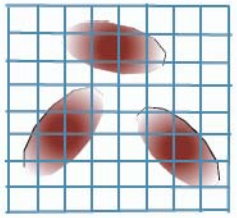
Complexity O(N)

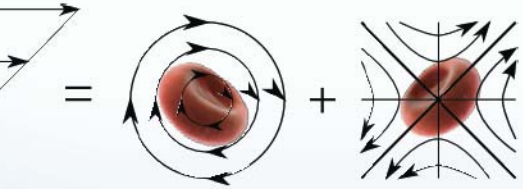
Enable complex geometries

Ellipsoidal Suspensions of RBC



$$\tilde{\delta}_{\xi_\parallel \xi_\perp}(x) = \prod_{\alpha=x,y,z} \tilde{\delta}_{\xi_\alpha}((Qx)_\alpha)$$





✓Tensorial drag

✓Tank Treading

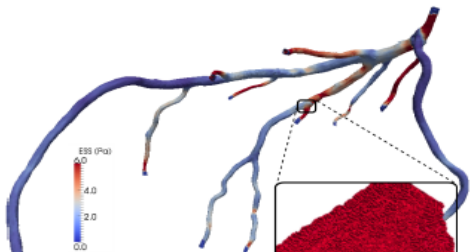
✓Complex Rheology

Strip off DOFs: O(10) per RBC

MUPHY Performance on the TiTech GPU cluster

The performance of 1536 GPUs for the Lattice Boltzmann kernel is the same as the whole Jülich Bluegene/P system (294712 cores!)

# of GPUs	time	efficiency
256	76.16	N.A.
512	38.52	98.86%
1024	19.95	95.37%
1536	13.43	94.43%



x3 40 rack BG/P

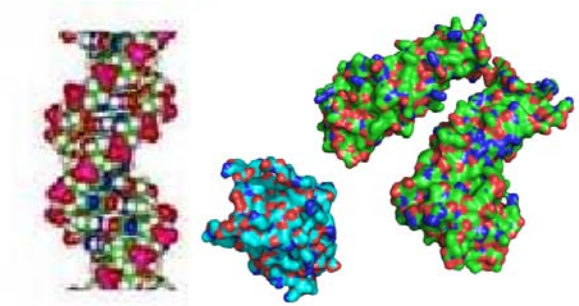
# of GPUs	time
256	648.23
512	327.97

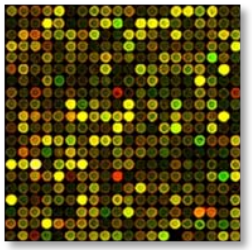
LB and Molecular Dynamics

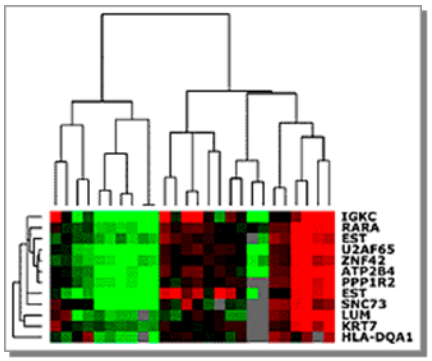
バイオインフォマティクス

コンピュータで生物学に関する
大量データを解析・理解・予測する学問。

「生命情報科学」







大量データ ＋ 傾向発見 → 予測 （データ マイニング）
系のモデル ＋ 数値代入 → 予測 （シミュレーション）

ゲノム科学の爆発的発展

Genome project件数
EMBL DB (bp)

100Gb/1run
200Gb/1run
500Mb/1run

細菌
酵母
線虫
ハエ
ヒト
マウス
メタゲノム
ニワトリ
イヌ
サル
ウマ

比較ゲノム解析
新型シークエンサーの登場
大規模ゲノム計画の幕開け

1,000人ヒトゲノム計画
10,000哺乳類ゲノム計画
10,000微生物ゲノム計画

300
250
200
150
100
50
0

1.6兆塩基
1.4兆塩基
1.2兆塩基

1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009

西暦年

提供： 黒川 顯 教授（生命理工）

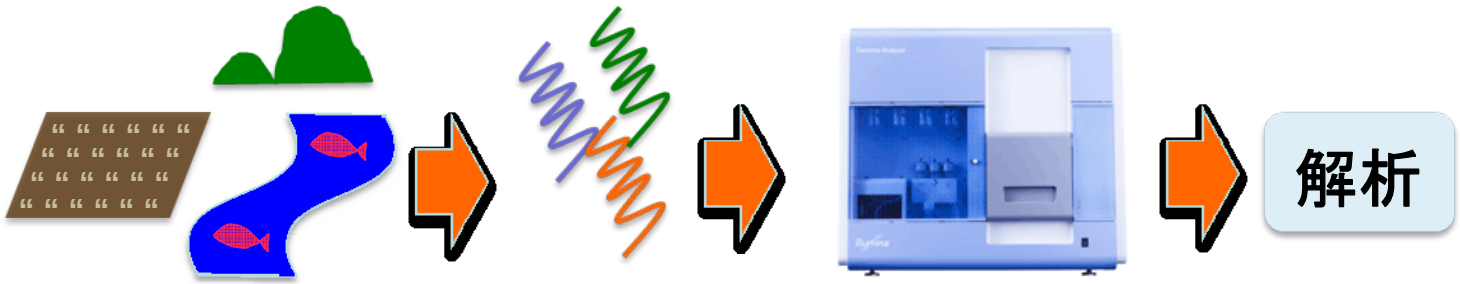
メタゲノム解析

細菌は地球上のあらゆる環境に存在し、環境は細菌の遺伝子によって満たされている。

➡ 環境は巨大な遺伝子プールだとみなすことができる。

新型シークエンサーの登場およびコンピューター技術の発達により、ゲノム情報をより多く、より早く手に入れることが可能になった。

➡ 遺伝子プールを解明する唯一の手段、メタゲノム解析が可能になりつつある。

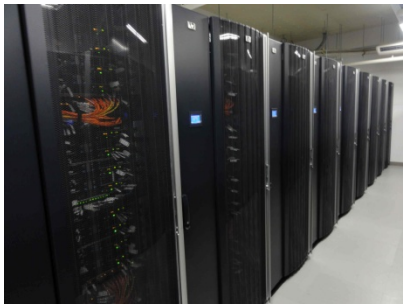


メタゲノム解析

提供： 黒川 顕 教授（東工大生命理工）

土壌メタゲノムプロジェクト – KEGG DBへのBLASTX –

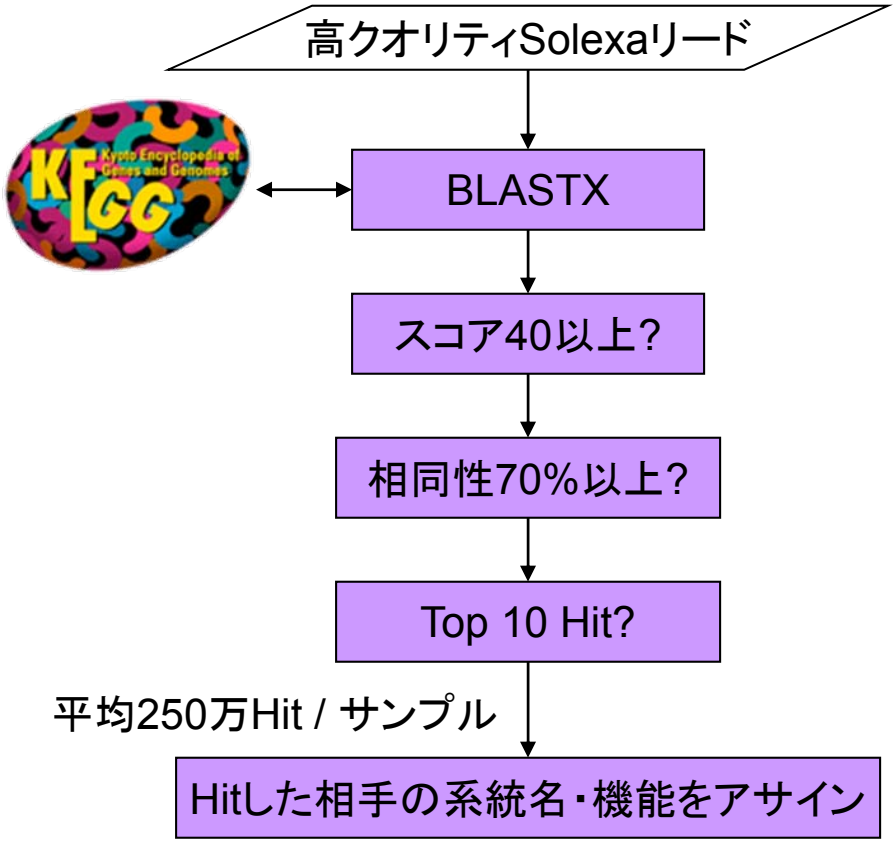
黒川　顕　教授
（東工大生命情報専攻）



解析に使用したSolexa配列の本数

Samples	Total Reads
0week	12,717,083
1week_C	13,595,042
1week_M	11,728,687
3week_C	8,695,063
3week_M	10,405,877
6week_C	12,073,087
6week_M	11,493,280
12week_C	11,188,195
12week_M	7,360,481
24week_C	12,979,727
24week_M	14,516,215
Total	126,752,737

計算には東工大のTSUBAMEを使用



KEGG DBとアミノ酸配列レベルで比較し、微生物系統と遺伝子機能を推定する。

マルチGPU環境による大規模バイオインフォマティクス

秋山　泰　教授
（東工大計算工学専攻）

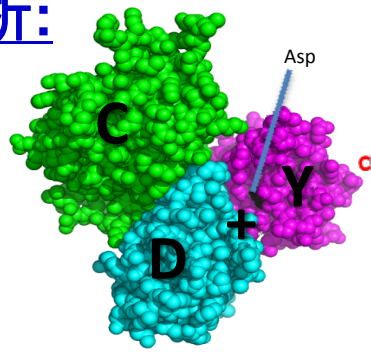
1) 大量DNA断片配列向けの相同性検索:

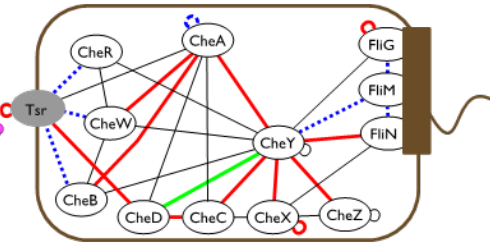
- 従来のRMAP等の低感度マッピング法とは一線を画するBLAST並みの高感度検索を行う"***GHOST***" システムを開発。
- メタゲノム解析で必要となるアミノ酸ベースの比較が可能。
- 4GPUのTesla S1070を1基用いた場合、単一CPUコア上でBLASTPを実行するのに比べて**約200倍の高速化を達成。**
  - 次世代シーケンサ Illumina社 Genome Analyzerによって決定された土壌中の微生物メタゲノム断片(60塩基) 10万本 を KEGG DBと比較。

"GPU challenge 2010"
自由課題部門 第1位受賞

2) タンパク質立体構造の形状相補性解析:

- 1000 × 1000 ＝百万ペア級の、総当たりの**網羅的タンパク質間ドッキング計算**に挑戦する独自並列ソフトウェア "***MEGADOCK***" を開発。
（MEXT　次世代生命体統合シミュレーションPJ）
- バクテリアの化学走性の制御機構の解析
- 肺がんに関係するEGFRシグナル伝達系の解析





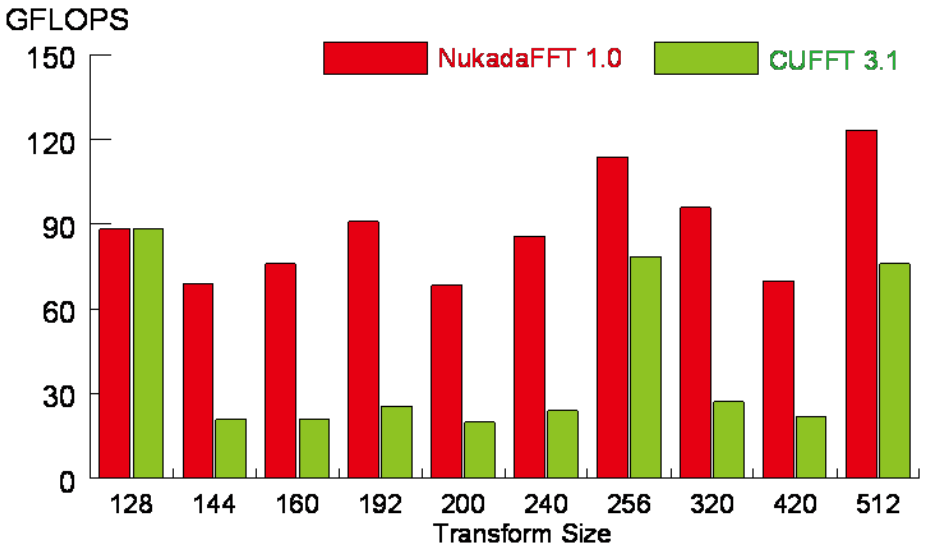
大腸菌走化性系の網羅的ドッキング予測（JBCB 2009）

- CheY, CheD, CheCの3つのタンパク質間の未知の相互作用の可能性を計算で示唆した。

NukadaFFT 1.0 release !? [SC08,SC09]

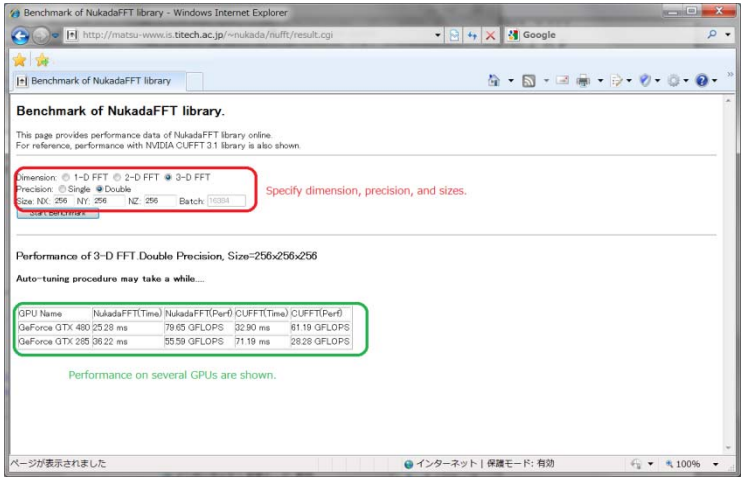
NukadaFFT library is a very fast auto-tuning FFT library for CUDA GPUs.

Tuning Parameters:
(1) Factorization of transform size.
(2) # of threads launched per SM.
(3) Padding insertion patters for shared memory
The library generates many kinds of FFT kernels and selects the fastest one. (exhaustive)



Performance of 1-D FFT.
(Double Precision, batch=32,768, GPU=GeForce GTX 480.)

For more details, you can see GTC 2010 Research Poster, or catch the author.



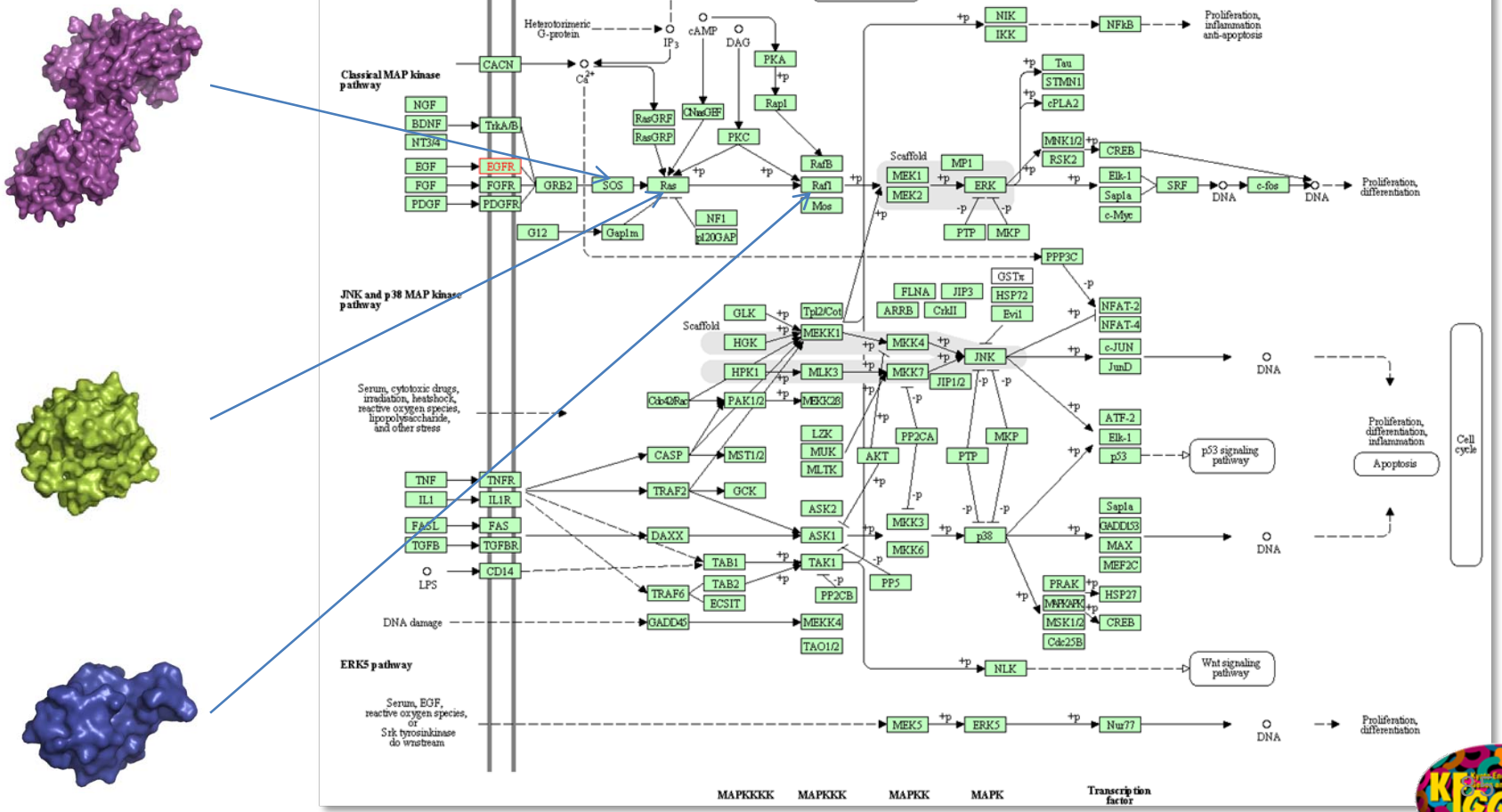
You can try online benchmarking as for the size you are interested in.

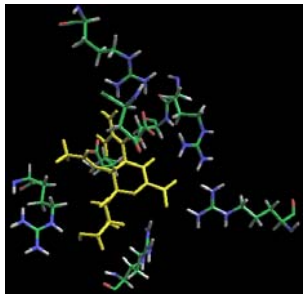
http://matsu-www.is.titech.ac.jp/~nukada/nufft/

肺がん関連（EGFR系）シグナル伝達系

497×497 = 247,009 pairsを解析済 (TSUBAME1)

1000 x 1000（1メガ）級の相互作用予測に挑戦

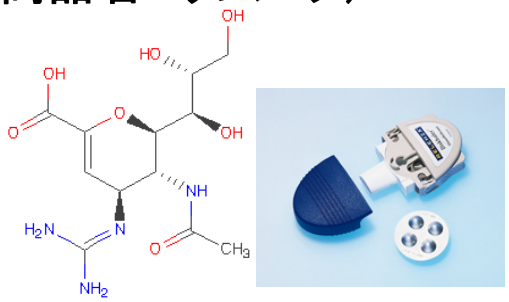




ab initio フラグメント分子軌道法計算

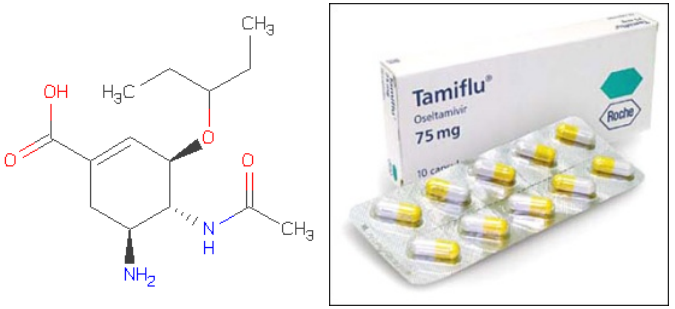
ノイラミニダーゼ酵素　(NA)　のポケット部分の立体構造を
計算機内で変位させて、薬剤結合性の変化を予測したい。

ザナミビル
（商品名　リレンザ）



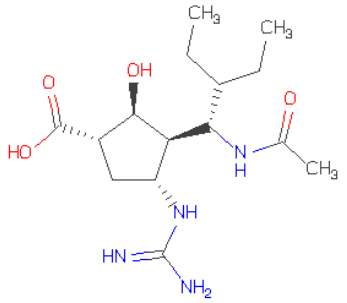
ビオタ社　(1989)

オセルタミビル
（商品名　タミフル）

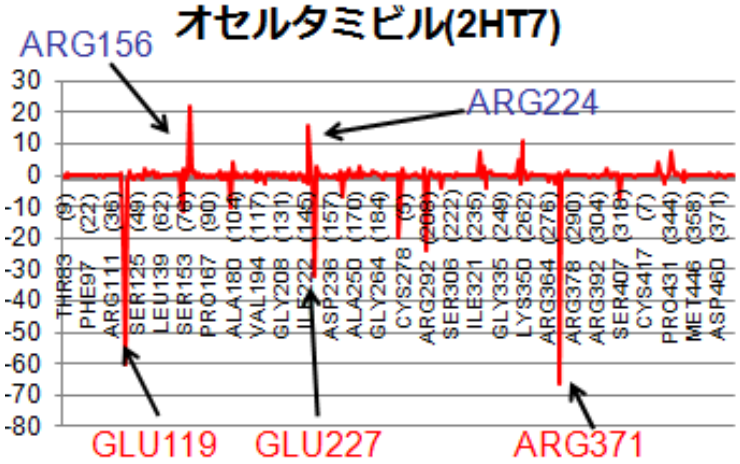


ギリアド・サイエンス社　(1996)

ペラミビル（注射薬）



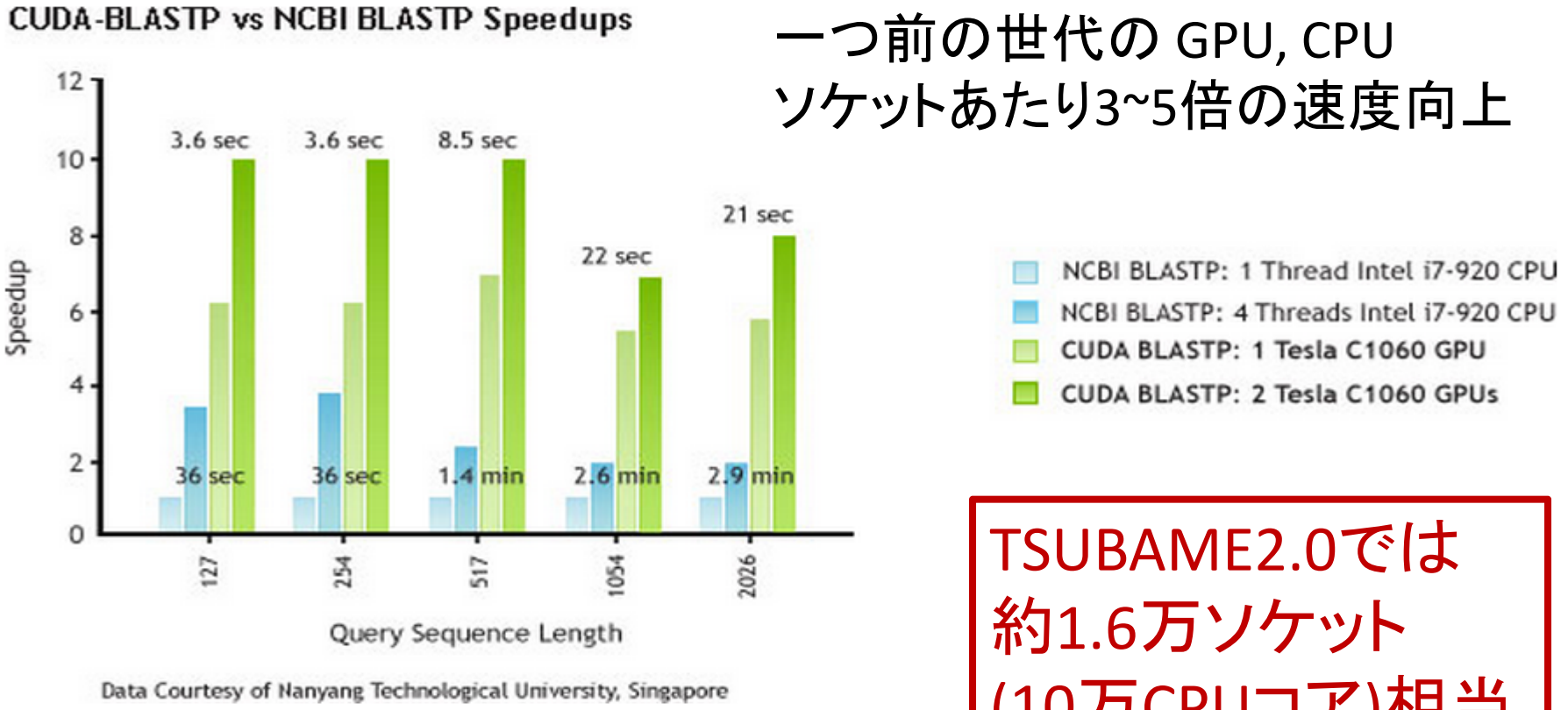
バイオクリスト社



各阻害薬とノイラミニダーゼ酵素の結合プロファイル解析　（尾渡・秋山，2009）

バイオインフォ：遺伝子相同性検索BLAST on GPUs

- CUDA-BLASTP (NTU)
  http://www.nvidia.com/object/blastp_on_tesla.html



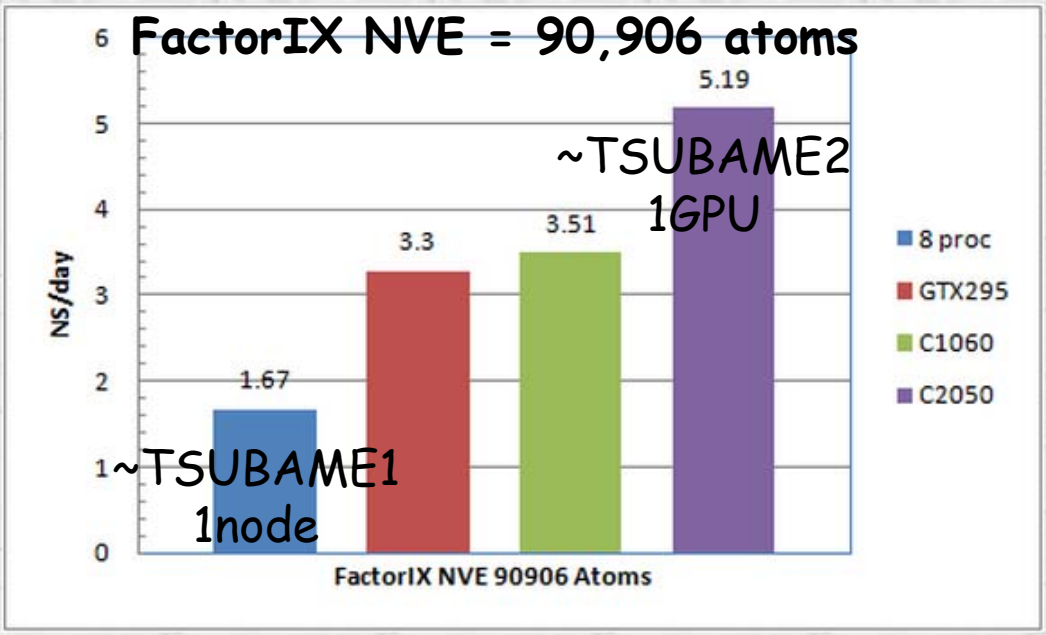
一つ前の世代の GPU, CPU
ソケットあたり3~5倍の速度向上

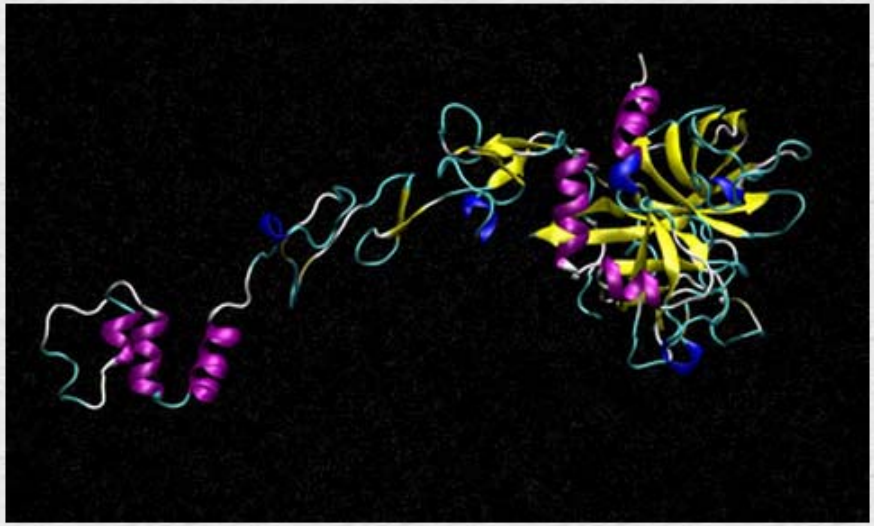
TSUBAME2.0では
約1.6万ソケット
(10万CPUコア)相当

- GPU-BLAST (CMU)
  http://eudoxus.cheme.cmu.edu/gpublast/gpublast.html
  “4 times speedup on Fermi GPUs”

分子動力学：GPU Amber 11

- ベンチマークでGPU1枚あたり8コアのCPUノードの4-20倍高速
  - マルチGPU版は「9月末に出る」はずだったがまだ出ていない。
- 精度は「問題ない」そうだが、今後検討の余地あり
- 平均10倍とすると、TSUBAME2.0では**30万コア**相当





GPUは「アクセラレータ」か？

- 高速計算用専用ハードウェア
  - 少量生産・特殊マーケット・高価格　×
  - ソフトウェアの世代間の継承の困難さ　×に近い
- アプリケーション分野の限定
  - 特殊な計算処理のみ高速　△⇒×
  - 「嵌らない」計算はできないか、CPUよりかなり遅い　×
- 一般的なプログラムの不動作
  - 特殊なプログラミング・言語等　△⇒×
  - ポインタ、リカージョン、構造体などの制限　△⇒×
  - OS等、システムソフトウェアは動かない・連動しない
    △⇒×　(CUDA xxx, Denver/Fusion, …)

以下参考資料

今後のペタ級マシン

Inst/Agency/Country(	Name	Machine	Peak Perf
ORNL/DoE/US	Jaguar Upgrade	Cray XT5	2.3PF
Tennessee大学/NSF/US 2009	Cracken	Cray XT5	1PF
Julich/欧州(ドイツ)	Jugene	IBM BG/P	1PF
中国・防衛大学	天河 (Tihanhe 1)	GPU Cluster/Dawning	1.2PF
中国・深圳国立スパコン	星雲 (Nebulae)	GPU Cluster/???	3PF
日本・東工大	TSUBAME2.0	GPU Cluster/HP-NEC	2.4 PF
LBNL/DoE/US 2010	Hopper	Cray XE6	1.3PF
中国・防衛大学	天河 (Tihanhe 1-A)	GPU Cluster/Dawning	5 PF
欧州PRACE計画・仏CEA	Tera 100	Nehalem-EX Cluster/Bull	1.25PF
ORNL/DoE/US	Jaguar Upgrade 2	Cray XE6 +GPU	20PF
NCSA/NSF/US	Blue Waters	IBM Power7 server	10PF
LLNL/DoE/US	Sequoia	IBM BG/Q	20PF
ArgonneNL/DoE/US 2011-12	???	IBM BG/Q	10PF
日本・理研	「京」	富士通 Venus 専用設計	10PF
日本・筑波大	HA-PACS	GPU Cluster/HP-NEC	1PF
欧州ペタコン群/PRACE計画	???	IBM, Cray等	~PF x 4~5
中国	4~6個所	???Dawning?	合算数十PF以上

ペタ～エクサへのスケーリングのロードブロック

- 「10億並列へ」は勇ましいが。。。
  - 電力・エネルギー
  - (強)スケーリングの欠落
  - N^2 vs. N 問題により深まるメモリ階層 (I/O含む)
  - 極端に低まる信頼性と実行不能性
  - プログラミングや実行モデル

Petaを達成したが中国に抜かれた米国は2018-2020 Exa(10^{18})flopへ驀進を開始

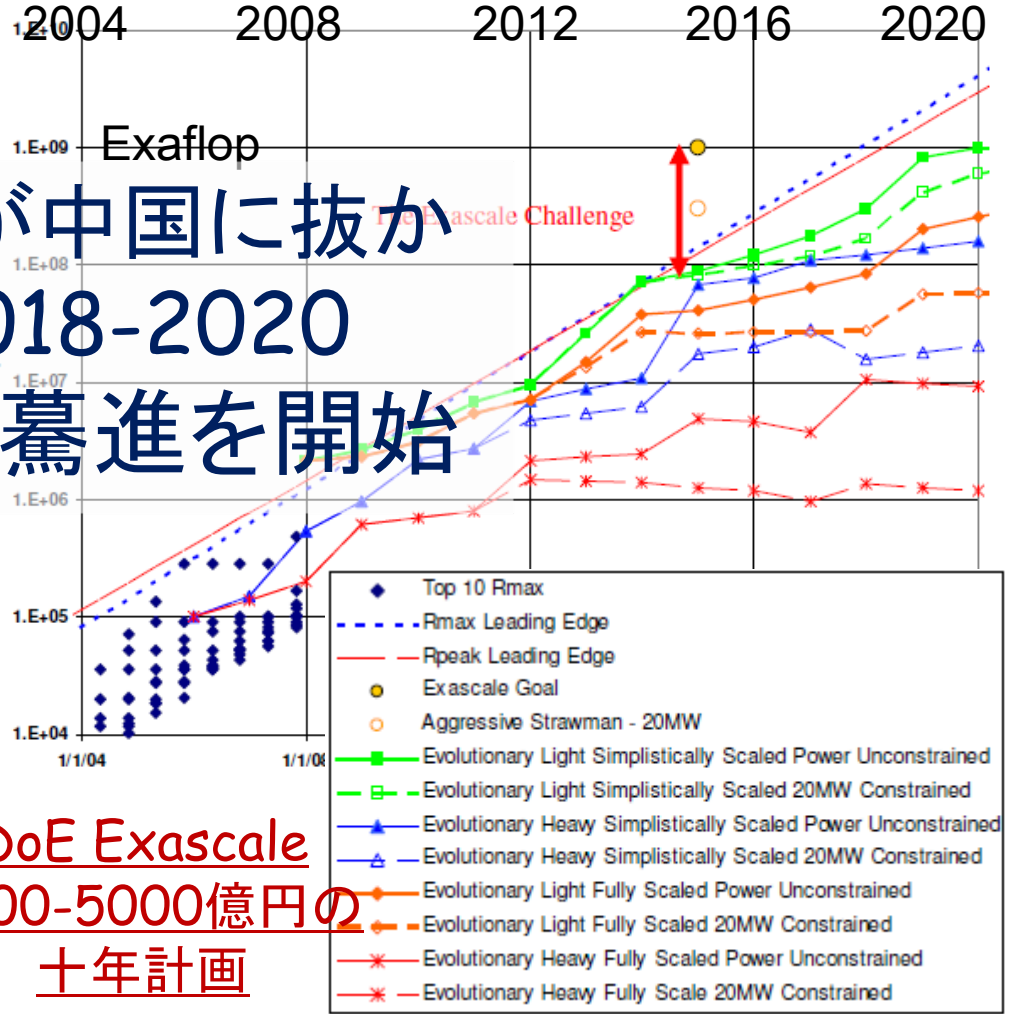
ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

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

Peter Koggeらによる300ページのDoD Exascaleシステムのレポート



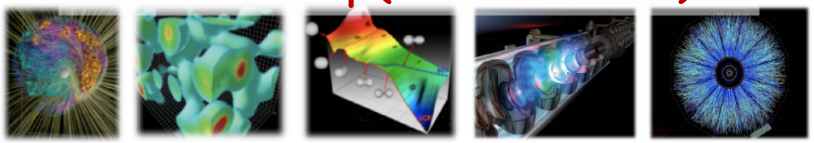
DoE Exascale 2000-5000億円の十年計画

Exa-scale Computational Resources

(slide courtesy Martin Savage)

- Meeting structured around present DOE Office of Science areas of effort

6アプリ分野のExascale Workshop(2008-2009)

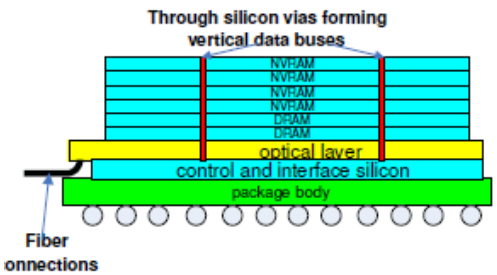


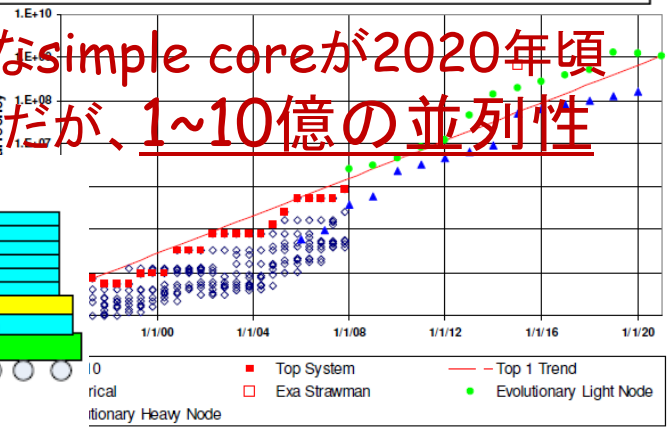
Nuclear Astrophysics | Cold QCD and Nuclear Forces | Nuclear Structure and Reactions | Accelerator Physics | Hot and Dense QCD

- Exa-scale computing is REQUIRED to accomplish the Nuclear Physics mission in each area
- Staging to Exa-flops is crucial :
  - 1 Pflop-yr to 10 Pflop-yrs to 100 Pflop-yrs to 1 Exa-flop-yr (sustained)

Paul Messina June 28, 2009

軽量なsimple coreが2020年頃有望だが、1~10億の並列性





DoE Exascale 性能指標

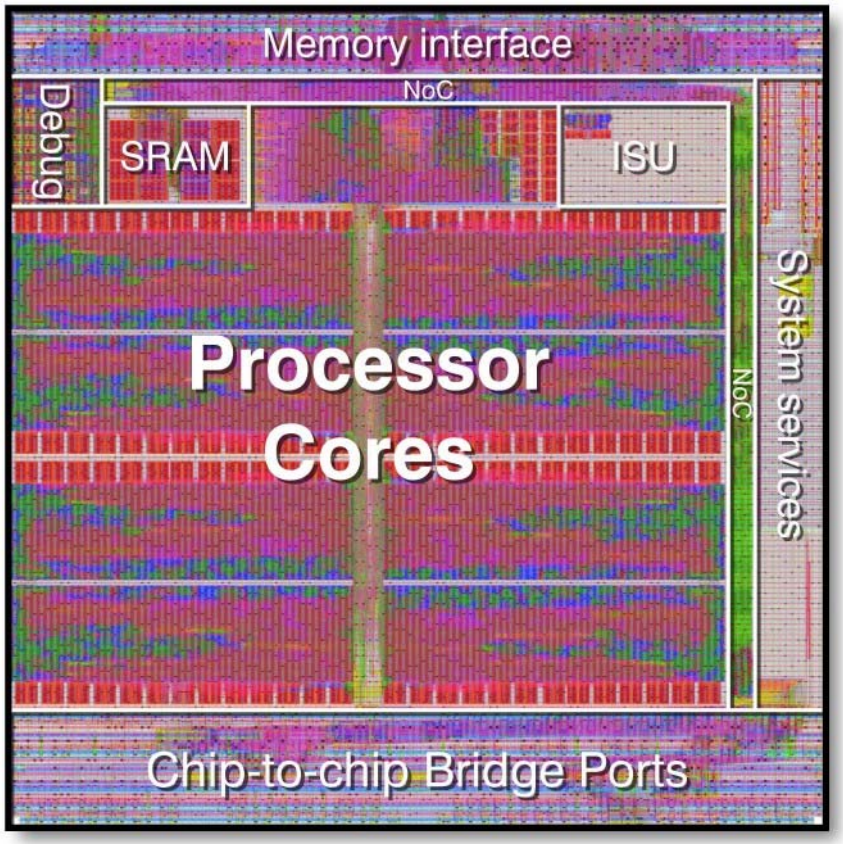
System attributes	"2010"		"2015"		"2018-20"	
System peak	2 PetaFlops		100-200 PetaFlops		1 ExaFlop	
Power	Jaguar 6 MW	TSUBAME 1.3 MW	15 MW		20 MW	
System Memory	0.3PB	0.1PB	5 PB		32-64PB	
Node Perf	125GF	1.6TF	0.5TF	7TF	1TF	10TF
Node Mem BW	25GB/s	0.5TB/s	0.1TB/s	1TB/s	0.4TB/s	4TB/s
Node Concurrency	12	O(1000)	O(100)	O(1000)	O(1000)	O(10000)
#Nodes	18,700	1442	50,000	5,000	1 million	100,000
Total Node Interconnect BW	1.5GB/s	8GB/s	20GB/s		200GB/s	
MTTI	O(days)		O(1 day)		O(1 day)	

Exaflopsへ向けた我が国のeScience インフラ・基盤センターへの提言

- 世界レベルのLeadership Computing and Data Facilityとしての性能ロードマップ策定と、基盤センター群による競争・協調・多様性を伴った年次の着実な遂行：HPCIのあるべき型
  - 国家レベルでの年次スケーリングの目標設定
  - 多様性と競争原理の適用➔センター毎の定額予算からの脱却
- HPCIの資源センターが自らの計算科学・計算機科学・応用数学を連動するHPC研究開発・人材育成・産学連携体制
  - 理研NLPや研究所を中心とした連携・人材交流
  - 基礎研究=>運用実験=>(リーダーシップ)実運用
  - 「カタログからマシンを買うだけのセンター」はいらない
- 「ガラパゴス」から「国際連携」へ
  - LHCや国際宇宙ステーションに学べ

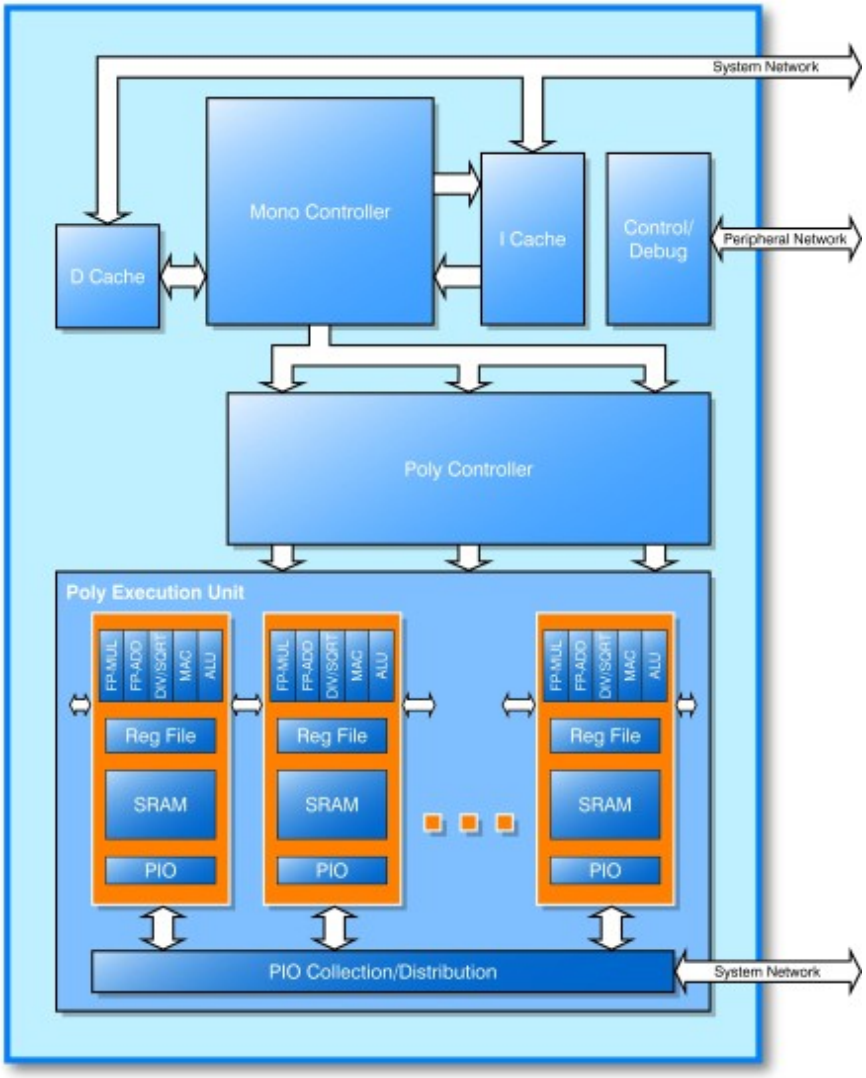
国際Exascale SW Project
http://www.exascale.org/
筑波 10/19-21/2009

CSX600 coprocessor layout



- Array of 96 Processor Elements
- 250 MHz, 48 GFlops (Double FP c.f. Grape SFP)
- IBM 0.13µm FSG process, 8-layer metal (copper)
- 47% logic, 53% memory
  - More logic than most processors!
  - About 50% of the logic is FPUs
  - Hence around one quarter of the chip is floating point hardware
- 15 mm x 15 mm die size
- 128 million transistors
- Approx. 10 Watts

CSX600 processor core



Multi-Threaded Array Processing

- Programmed in high-level languages
- Hardware multi-threading for latency tolerance
- Asynchronous, overlapped I/O
- Run-time extensible instruction set
- Bi-endian (compatible with host CPU)

Array of 96 Processor Elements (PEs)

- **Each is a Very Long Instruction Word (VLIW) core, not just an ALU**
- **Flexible data parallel processing**
- Built-in PE fault tolerance, resiliency

High performance, low power dissipation

TSUBAME2.0(ファイルサービス基盤:全体概要)

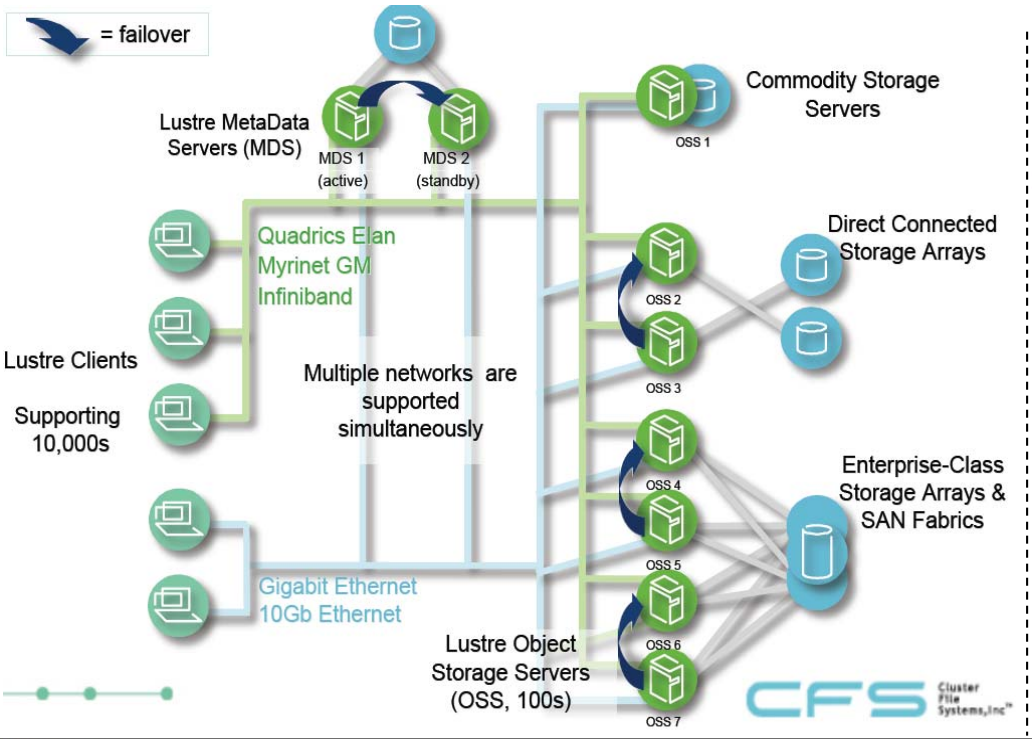
ファイルサービス基盤は3つのサービス基盤から構成

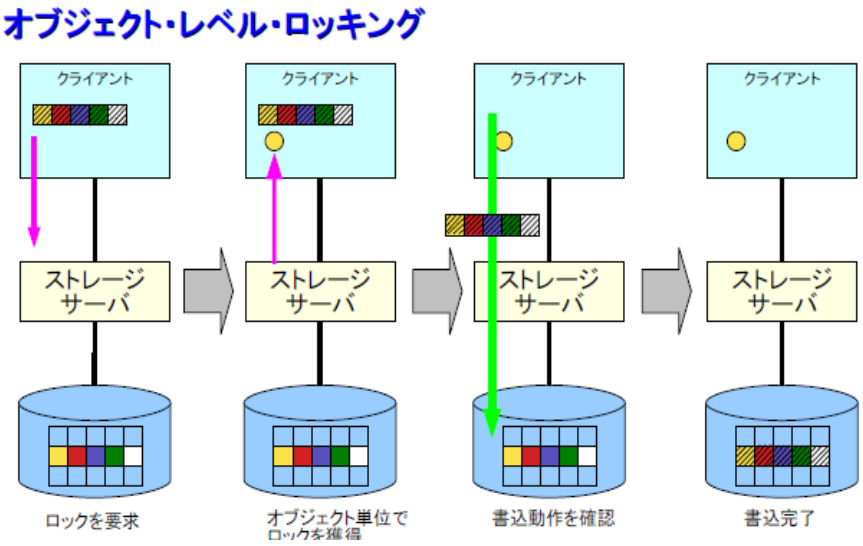
- 並列ファイルサービス基盤
  - Lustreを構築し，並列ファイルシステム領域として利用
  - LustreのRedundancy機能により信頼性を確保
- ホームファイルサービス基盤
  - NFS，CIFSを構築し，ホーム領域のマウントポイントとして利用
  - iSCSIによりネットワーク経由で仮想ブロック，オブジェクトストレージの作成が可能
- Gridファイルサービス基盤(今後サービス提供予定)
  - 学外のファイルサービスに対して，ファイル転送
  - グリッドのファイル共有システムを構築するため，GridFTP，Gfarm2を構築
  - Gfarm2により，学外から並列ファイルシステム領域の利用が可能

TSUBAME2.0（並列ファイルシステムサービス基盤）

大規模グリッド環境で、容量，アクセス性能の両面において優れたスケーラビリティを発揮し、豊富な稼働実績を有するLustre並列ファイルシステムで構築

- **オブジェクトストレージサーバ・ベースの並列（クラスター）ファイルシステム**
  - **オブジェクトストレージサーバの導入によりクライアントとストレージ装置をSAN等で直接接続することなく，大規模な並列ファイルシステムを構築することが可能**
  - **メタデータサーバは，オブジェクトの生成等にのみ関与し実I/Oには関与しないため高いスケーラビリティを実現**

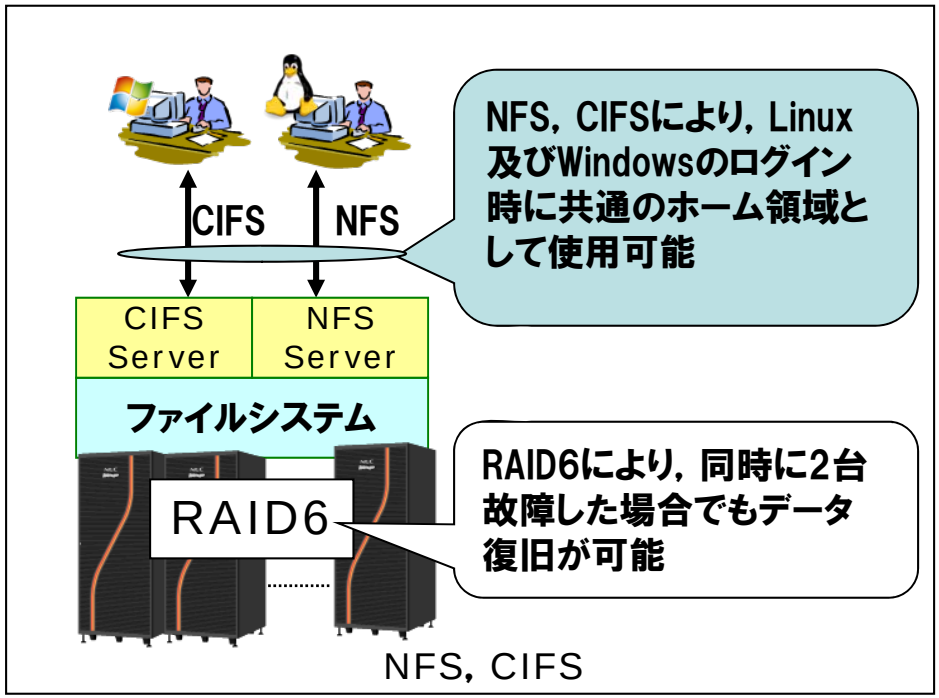


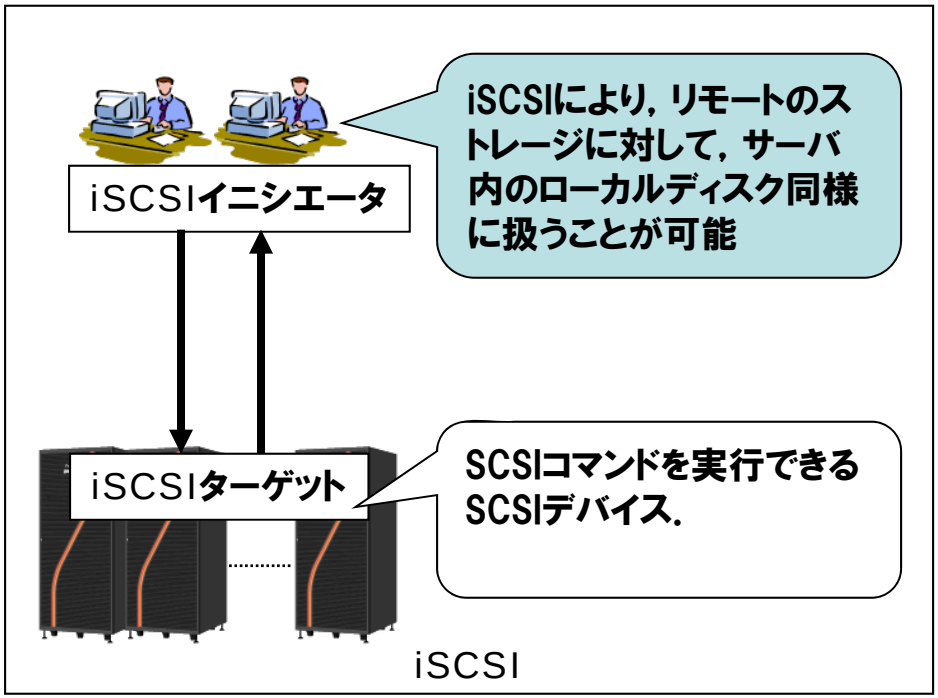


TSUBAME2.0(ホームファイルシステムサービス基盤)
−NFS, CIFS, iSCSI−

Linux及びWindowsからマウント可能なホーム領域を堅牢性を高めて構成

- データ保存サービスの特長
  - LinuxノードからNFSにより利用可能
  - WindowsノードからCIFSにより利用可能
  - RAID6により2台故障した場合でも復旧可能 ※RAID5,RAID1+0でも構成可能
  - iSCSIにより，リモートのストレージに対して，サーバ内のローカルディスク同様に扱うことが可能





TSUBAME2.0 テープシステム（別調達）
合計15PB以上、階層ファイルシステムの構築

スイッチ（Brocade BI-RX-4-AC）
10G-SR x6ポート
10G-LR x2ポート
1G-Ethernet x24ポート

10GBase-SR(MMF) x6

バックアップ用サーバ x3

SunFire X4275 3nodes
ディスク 300GB(RAID1時)
メモリ 24GB

階層型ストレージ用
サーバ追加予定

FC 4GBps x12 48Gbps

FCスイッチ（Brocade 5300）
FC x80ポート

FC 4GBps x20 80Gbps

100/1000Base-T x1

ACSLSサーバ
（ロボット制御用）

SunFire X4140 1nodes

テープライブラリ SL8500 2台をPTP構成

本年度~8PB拡張予定

Sun SL8500 2台
-SL8500 + ストレージエクスパンジョンモジュール3台(5.3PB構成, 6,632スロット)
-SL8500 + ストレージエクスパンジョンモジュール2台(3.9PB構成, 4,904スロット)

【全体】
容量:4PB(非圧縮時) LTO4 5000巻
(9.2PBまで拡張可能)
ドライブ数:20 drive (128driveまで拡張可能)
ドライブ転送速度:120MB/s(非圧縮時)
スロット数:11,536 slot

GSIC国際棟

20Gbps

SuperTitanet3

既設スイッチ
(Foundry BigIron RX8 Switch)

TSUBAME

Lustre
ファイルシステム領域

ホーム領域
（NFS）

GSIC情報棟

東工大 e-Science RENKEI-POP による分散ストレージ・HPCIへの貢献

- 目的: 高速SINET網を活用・スパコンセンター間データ共有基盤の構築
  - RENKEIプロジェクト(文科省e-Science委託事業)と連携
- ストレージサーバRENKEI-PoP (Point of Presence)の開発・全国に配備
  - 大容量、高速IO性能を備えたデータ転送用サーバアプライアンス
  - SINET3上に広域ファイルシステムGfarm等によりRENKEI-クラウド構築
  - TSUBAME2.0や他の機関のスパコン間の大規模データ交換

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)

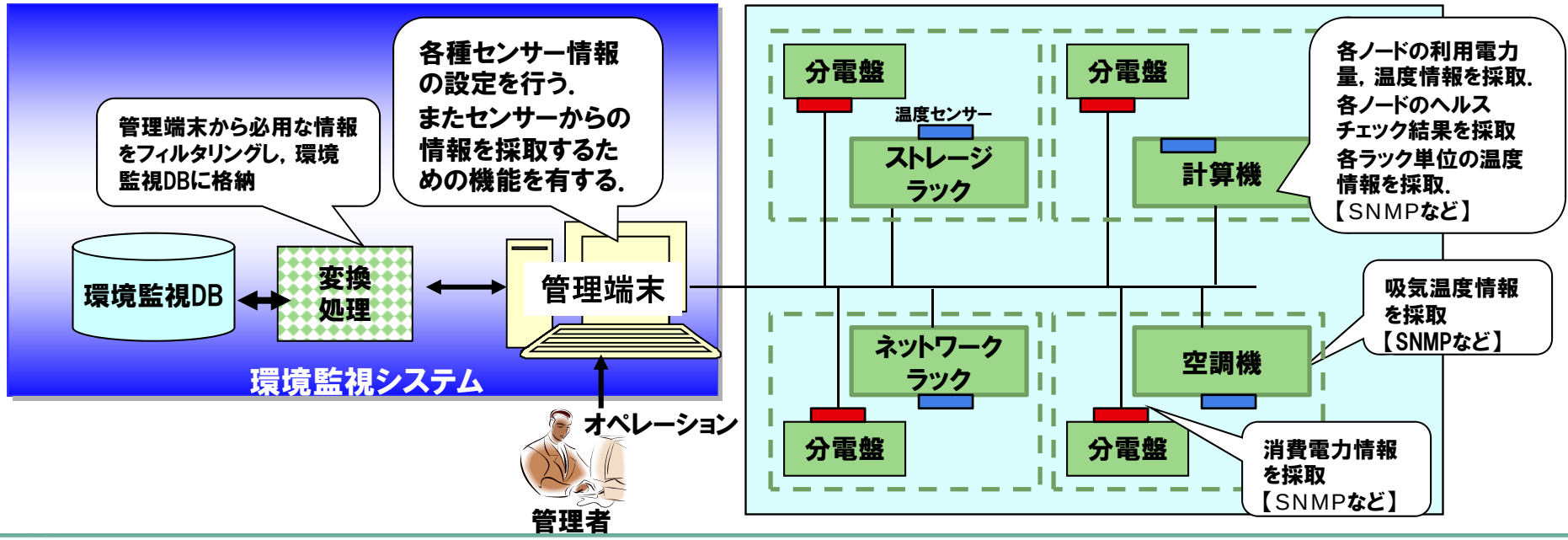
- 現在9拠点に配備、110TBの高速分散クラウドストレージとして利用可能

東京工業大学	大阪大学
国立情報学研究所	高エネルギー加速器研究機構
名古屋大学	筑波大学
産業技術総合研究所	東北大学

グリーンスパコン: 環境監視システム

各計算ノード，ラック，及び計算機室の温度情報・消費電力等を監視する「環境監視システム」.

- センサー情報及び各計算ノードの情報をオンラインでモニタリング
  - 温度情報（温度センサーから取得）
  - ヘルスチェック結果，サービス提供状況，故障の有無
  - 消費電力（各ノード・及び各分電盤から取得）



TSUBAME2.0(Green:ピーク電力抑制制御)

「ジョブ管理システム」と「運転管理システム」により，電力抑制警報発令時に手動で縮退運転へ移行。これにより，Peak電力を削減することが可能

- 電力抑制警報発令時の特徴
  - 現行ユーザのチェックポイントの取得が可能なジョブはチェックポイントを取り，ノードを停止することにより縮退運転に移行．
  - チェックポイントを取得したジョブはリスタート機能，取得できないジョブは再投入により，ユーザジョブを直接的に妨げること防ぎます．