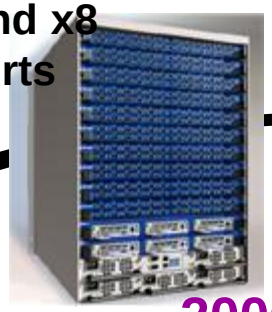
TSUBAME2.0 がリードする GPUベクトルスーパーコンピューティング

東京工業大学 学術国際情報センター
教授
松岡　聡

GPUコンピューティング講習会＠東工大
2010年9月13日

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

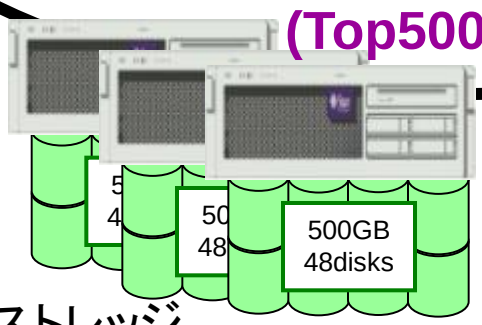
Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)



10Gbps+外部
ネットワーク

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

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



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

TSUBAME1の4年の運用成果

1. 東工大のシンボル：世界トップレベルの情報インフラ

Top500
4期連続日本一
7期連続性能向上
(世界初)

3. 産学連携等の推進、大型プロジェクトへの呼び水、アライアンスを組む他大学計算ニーズホスティング

MITSUBISHI CHEMICAL
NUMERICAL TECHNOLOGIES
SUMITOMO CHEMICAL
AMD
NEC
Sun microsystems
Microsoft

文部科学省先端研究施設共用イノベーション創出事業【産業戦略利用】

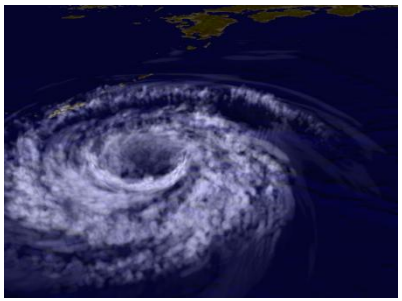
CompView
Global COE
「計算世界観」

企業との包括 collaboration

Research Organization of Information and Systems
国立情報学研究所
National Institute of Informatics
NAREGI

NAREGI/NII-CSI
全国サイバーサイエンスインフラ
＊NAREGI 開発への貢献
＊阪大-東工大 NAREGI β 2連携

2. 研究推進：莫大な計算パワー・ストレッジ(1ペタバイト以上)・みんなのスパコン



4. 学内の分散した情報基盤の集約化・ホスティング

- TSUBAME「みんなのスパコン」
  - 新概念の課金利用法によるユーザ数増加 => 2000人へ3倍増
  - SE運用業務の追加(アプリ・性能評価・グリッド試験運用など)
- 各種ITサービスのホスティング

TSUBAME 1.2への進化 (GPU等拡張) 2008年10月

GSIC
Global Scientific Information and Computing Center

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

NEC SX-8i

Storage
1.5 Petabyte (Sun x4500 x 60)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s aggregate I/O BW

10Gbps+External NW

Unified Infiniband network

Sun x4600 (16 Opteron Cores)
32~128 GBytes/Node
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

10,000 CPU Cores
300,000 SIMD Cores
~900TFlops-SFP,
~170TFlops-DFP
80TB/s Mem BW (x2 ES)

GCOE TSUBASA
Harpertown-Xeon
90Node 720CPU
8.2TeraFlops

NEW: co-TSUBAME
72Node 586CPU (Low Power)
~5TeraFlops

PCI-e

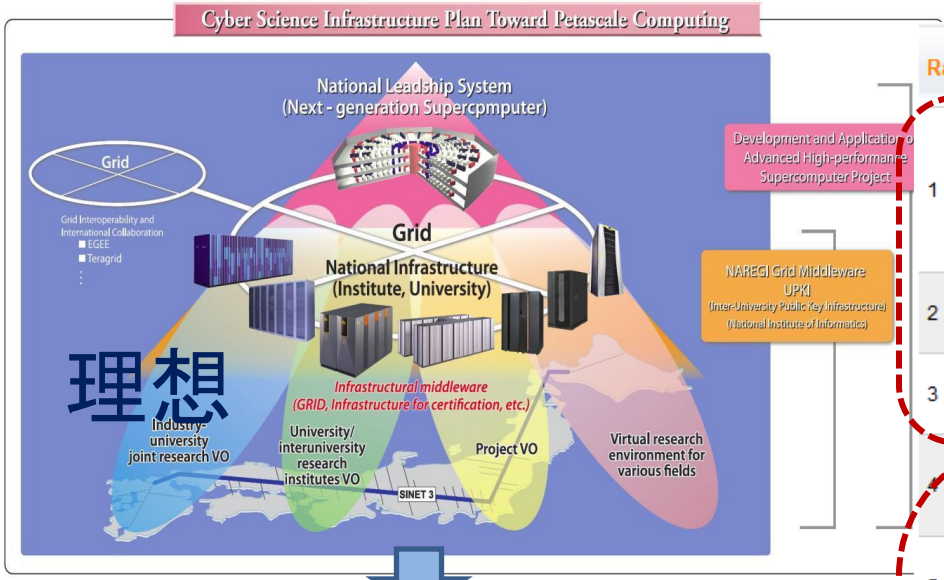
ClearSpeed CSX600
SIMD accelerator
360 648 boards,
35
52.2TeraFlops

NVIDIA

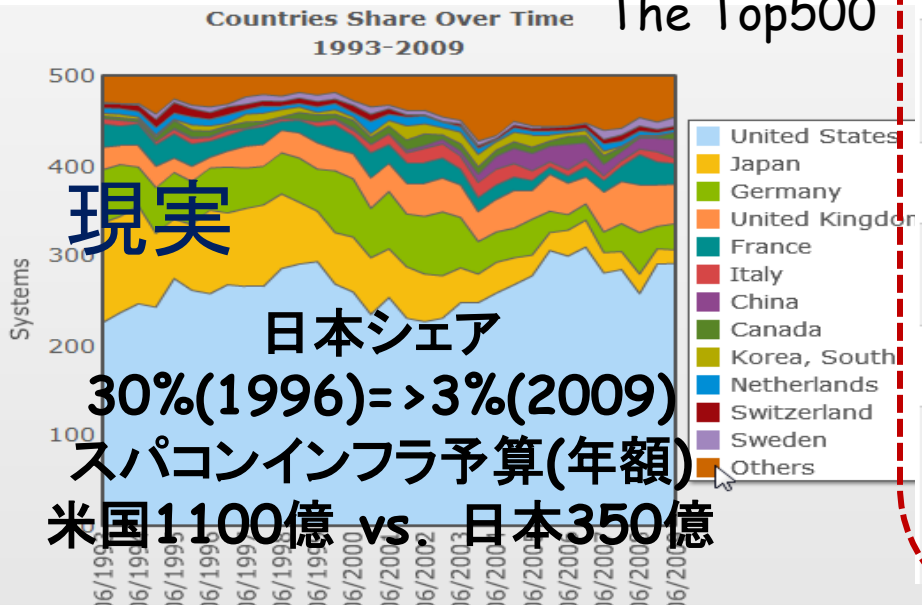
Nvidia Tesla S1070: 170台, 総計 680カード
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)


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

理研NLPと(8+1)基盤センターを結ぶ世界トップレベルのe-Scienceインフラ、のはずだが。。。



The Top500



現実

日本シェア
30%(1996)=>3%(2009)
スパコンインフラ予算(年額)
米国1100億 vs. 日本350億

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2008 IBM	129600	1105.00	1456.70	2483.47
2	Oak Ridge National Laboratory United States	Jaguar - Cray XT5 QC 2.3 GHz / 2008 Cray Inc.	150152	1059.00	1381.40	6950.60
3	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
4	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0/2.66 GHz / 2008 SGI	51200	487.01	608.83	2090.00

1ペタ級

1/2ペタ級

DoE ORNL/Cray XT5 "Jaguar"
~180,000 AMD Opteron CPU Cores,
1.64 Petaflops Peak

DoE LANL/IBM "Roadrunner"
> 100,000 Cell SPE Cores
> 1.3 Petaflops Peak

米国は異なるアーキテクチャによる達成数万~10万CPUコアへのスケーラビリティで世界を圧倒的にリード

DoE LLNL/IBM BG/L
~212,000 IBM PowerPC Cores
~0.59 Petaflops Peak

NSF TACC "Ranger" Cluster
~63,000 AMD Opteron Cores
0.57 Petaflops Peak

今後のペタ級マシン(主要サンプル)

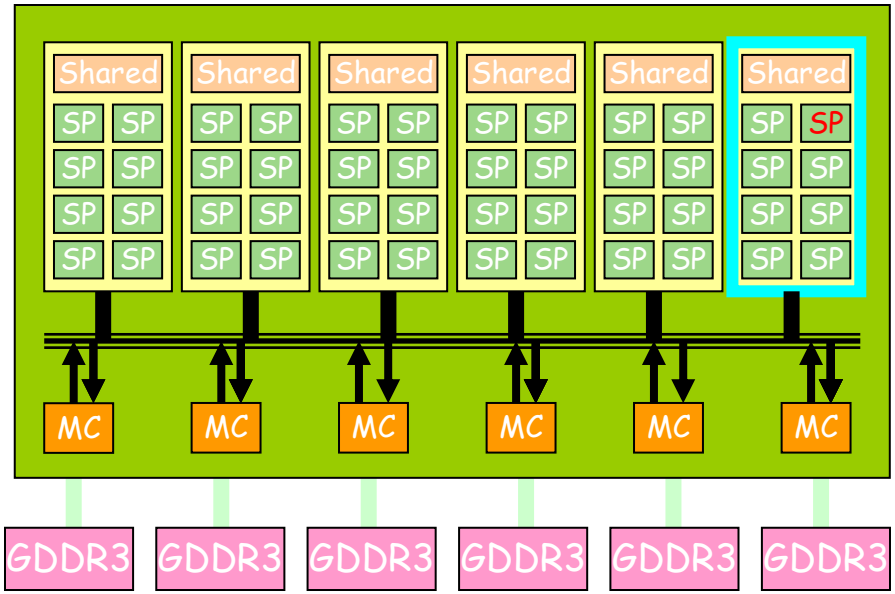
Inst/Agency/Country(	Name	Machine	Perf
ORNL/DoE/US	Jaguar Upgrade	Cray XT5/Istanbul	~2PF
Tennessee大学/NSF/US	Cracken	Cray XT5/Istanbul	1PF
Julich/欧州(ドイツ)	Jugene	IBM BG/P	1PF
中国・防衛大学	天河	GPU Cluster	1.2PF
LLNL/DoE/US	Sequoia Proto	IBM BG/P	~1PF
Tokyo tech./MEXT/JP	TSUBAME2.0	GPU Cluster/ NEC-HP	2.4PF
中国	星雲(Nebulae)	GPU Cluster/Dawning	3PF
LBNL/DoE/US	Franklin 6	Cray XT6	1.2PF
France CEA	Tera 100	Bull Nehalem EX	1.2PF
LANL/DoE/US	???	???	???
欧州ペタ/PRACE計画	???	IBM/Cray/Sun/Bull…	1-2PF?
ORNL/DoE/US	Jaguar Upgrade 2	Cray XT6 +GPU	20PF
NCSA/NSF/US	Blue Waters	IBM Power7 server	10-20PF
LLNL/DoE/US	Sequoia	IBM BG/Q / PERCS	22PF
ArgonneNL/DoE/US	???	IBM BG/Q / PERCS	10PF
神戸ペタ-Riken/MEXT/JP	KEI	富士通 Venus 専用設計	10PF
欧州ペタコン郡/PRACE計画	???	IBM, Cray等	~10PF x 4~5
中国	???	???Dawning?	30PF以上?

2009

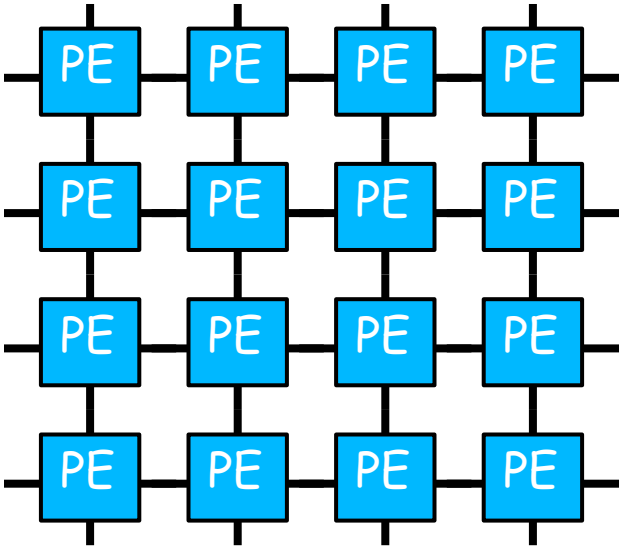
2010

2011-12

GPU（Multithreaded Vector）vs. Standard Many Cores?



vs.



- 今後スパコンが巨大化するにつれ、強スケーリング（*strong scaling*）が選択肢において重要な決定要素に

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?

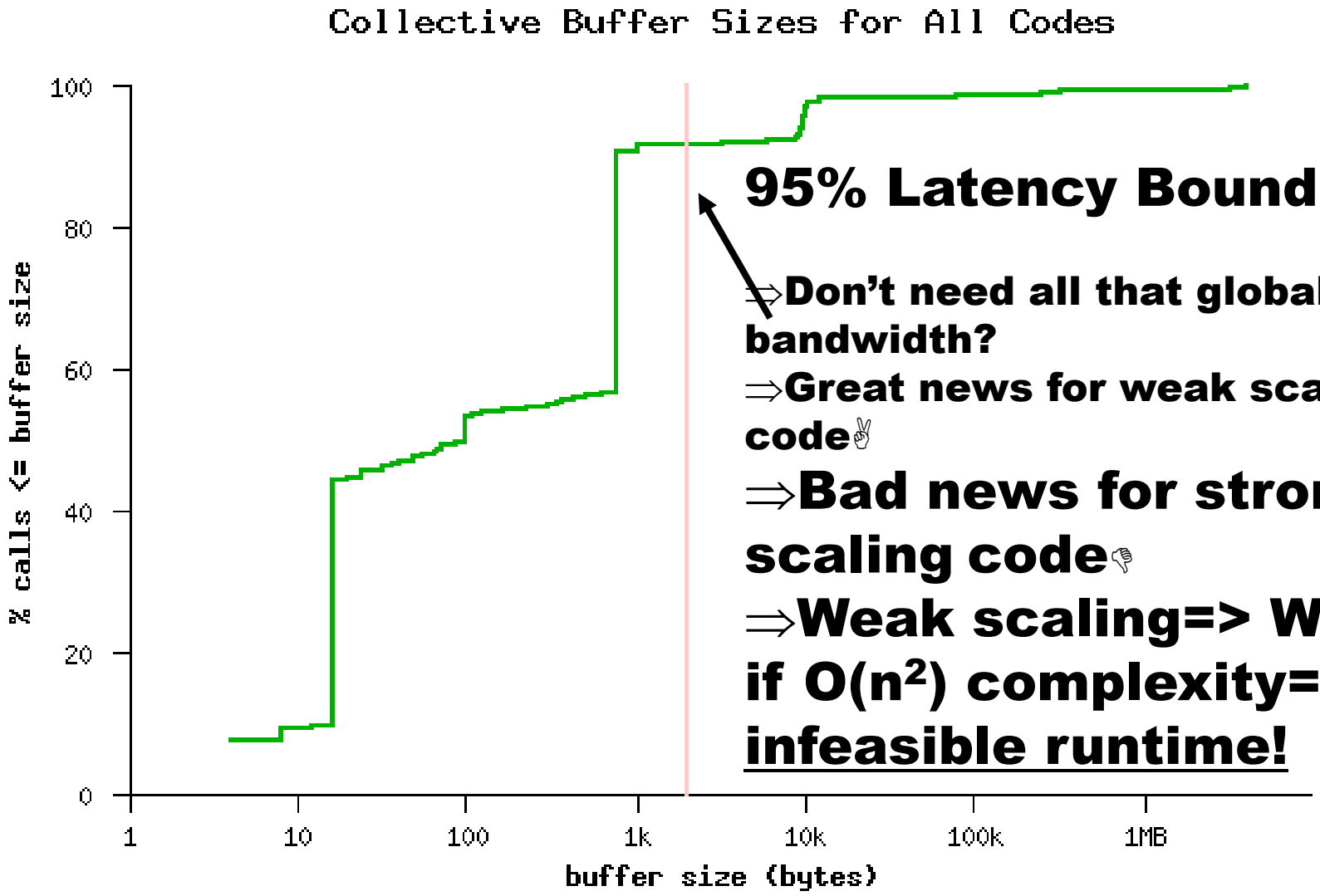
- How large does a message have to be in order to saturate a dedicated circuit on the interconnect?
  - $N^{1/2}$ from the early days of vector computing
  - Bandwidth Delay Product in TCP

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

Petaを達成した米国は2018-2020 Exaflopへ驀進を開始

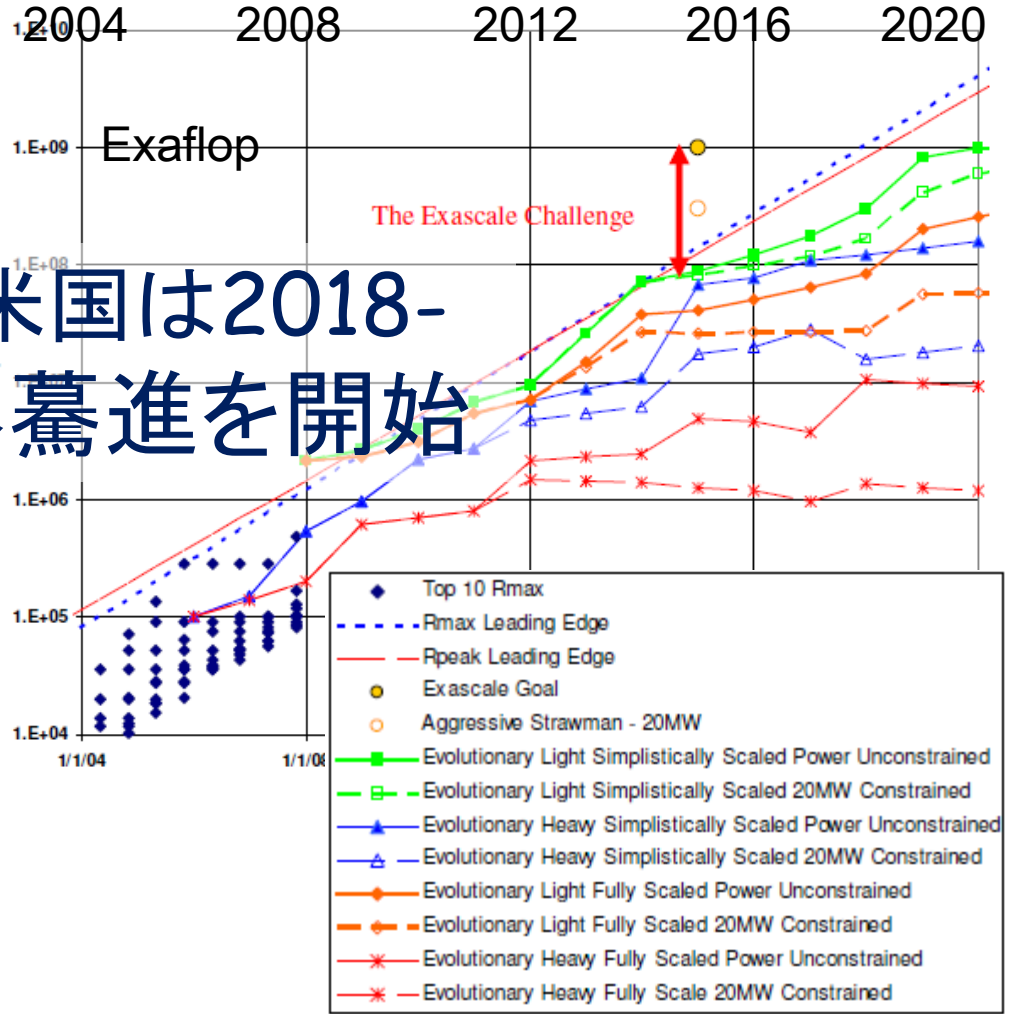
Peter Koggeらによる300ページのDoD Exascaleシステムのレポート

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

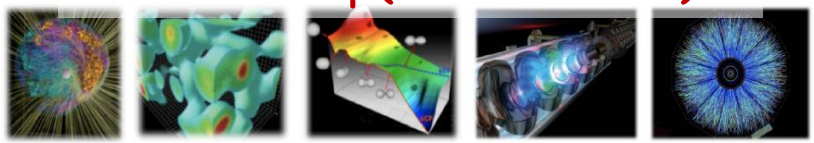


Exa-scale Computational Resources

(slide courtesy Martin Savage)

- Meeting structured around present INCITE physics areas of effort

6アプリ分野のExascale Workshop(2008-2009)

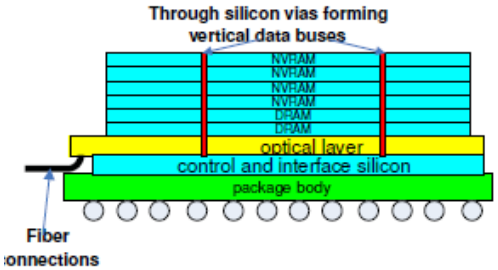


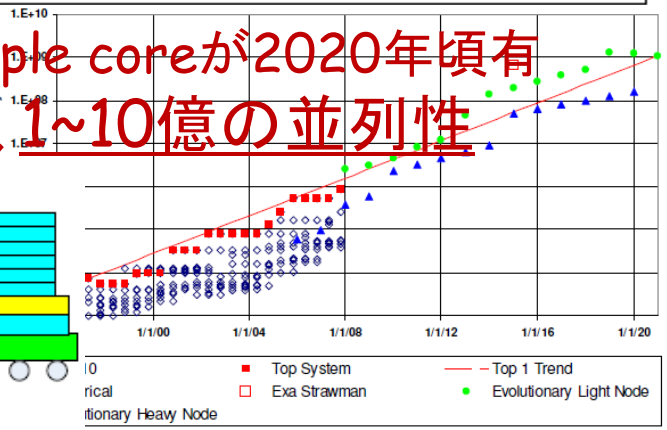
Nuclear Astrophysics | Cold QCD and Nuclear Forces | Nuclear Structure and Reactions | Accelerator Physics | Hot and Dense QCD

- 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

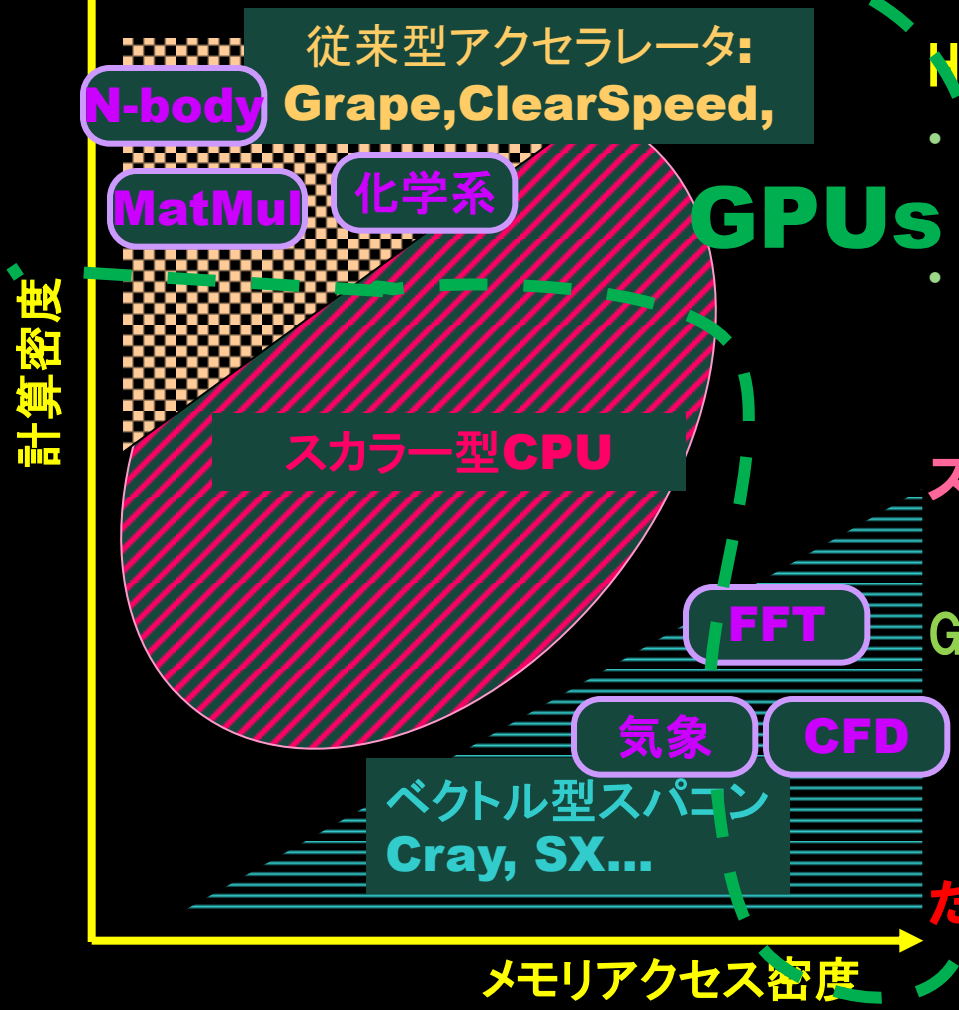




ペタからエクサへのスケーリング
強スケーリング達成のためには

- ***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)
- ***又はレーテンシ隠し(高バンド幅＋大量のスレッド)***
  - 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

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

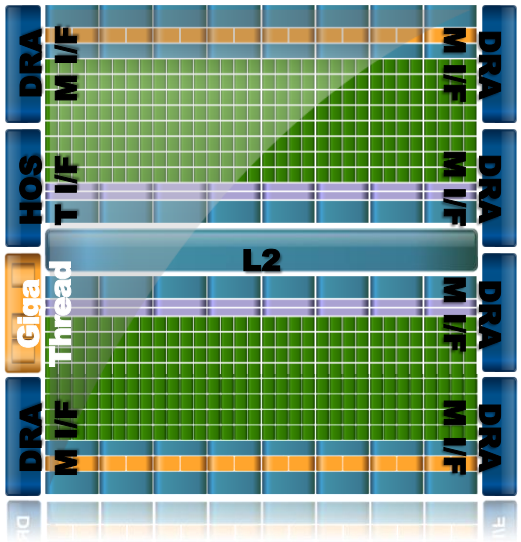
スカラー型CPUはどちらもそこそこ
→高性能を得るために巨大並列化

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

TSUBAME2.0のノードアーキテクチャ
GPU+CPUによるスカラー・ベクトル混合型アーキテクチャ



NVIDIA Tesla M2050 GPU
515GFlops/ 448 CUDA Cores
3GBメモリ容量、**150GB/s**メモリバンド幅
新世代のベクトルプロセッサ

GPU Vector

GDDR5 150GB/s

Core Dataset Vector-Parallel

CPU Scalar
- **OS**
- **Services**
- **Legacy**
- 非均質疎行列系

PCIe 2 x16 8GB/s x n

IOH

IB QDR HCA

40Gbps IB x 2

Flash 200GB >450MB/s

ネットワーク・I/O強化のためのカスタムMB設計

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の特徴

1. 世界一クラスのペタコン: 倍精度2.4ペタフロップス

- 最新型GPU・CPUによるベクトル・スカラー混合アーキテクチャ:高い計算性能とバンド幅
  - 2.4 Petaflops, メモリバンド幅0.72ペタバイト/s (地球シミュレータの4.3倍)
- 世界最高速クラスの200テラビット級のバイセクションバンド幅を実現した光ネットワーク
- 最新のSSDなどを活用した多階層ストレージによる15PBの大規模化と高速化 (0.66TB/s)

2. 世界一環境「グリーンスパコン」

- TSUBAME1同等のエネルギー消費・30倍の電力性能比・PUE=1.28・「Green500」世界一？
  - GPU+マルチコアCPUの大幅活用による高効率化
  - 最先端の冷却法: 密閉型の水冷ラック, 負荷集中での高熱勾配, 夏季の負荷キャップ
  - JST-CREST Ultra Low Power-HPC等の研究成果の応用
- => PUE 1.28以下(他の国内のスパコンセンターは1.7~2.0程度)

3. 「クラウド型スパコン」:総合的学内ITホスティング

- Windows HPC/Linuxなど複数OS, 複数環境のサポート
- 仮想化による種々のデータセンターホスティング機能のサポート
- 教育用/Kioskシステムのバックエンド化、全学アカウント・総合的学内ITの集中化・費用削減

4. 東工大GSICでの種々の基礎研究・メーカー共同開発の成果

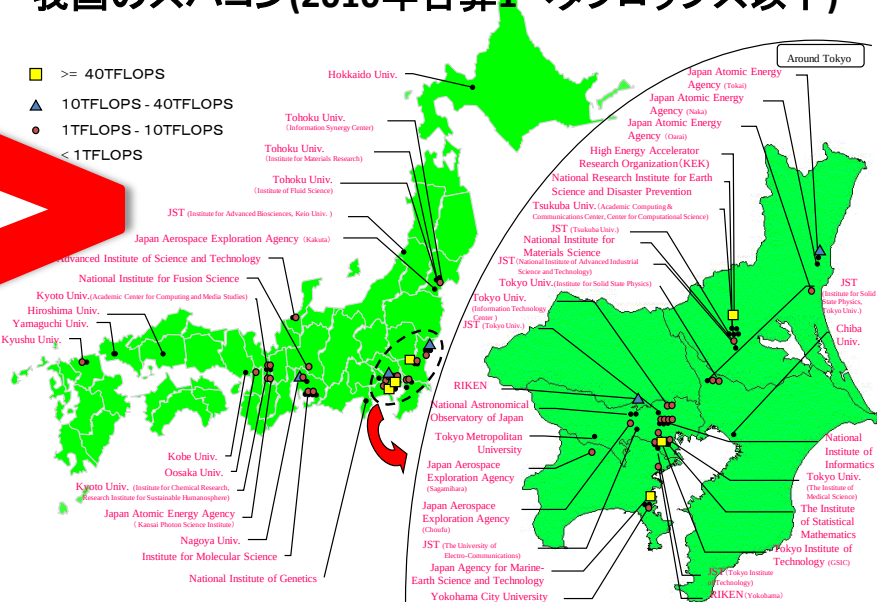
- **JST-CREST "Ultra Low Power HPC",** 科研特定領域「情報爆発」, 文科省-国立情報研 **NAREGI / e-Science**等、多くの基礎研究
- 海外企業とも: **NVIDIA CUDA CoE** (日本初), **Microsoft TCI** 包括共同研究契約
- **NEC, HP, NVIDIA, Microsoft, Voltaire, DDN** 等との共同開発体制

東工大TSUBAME2.0
2010年11月稼働予定

東京工業大学
Tokyo Institute of Technology

日本全土のスパコン全て

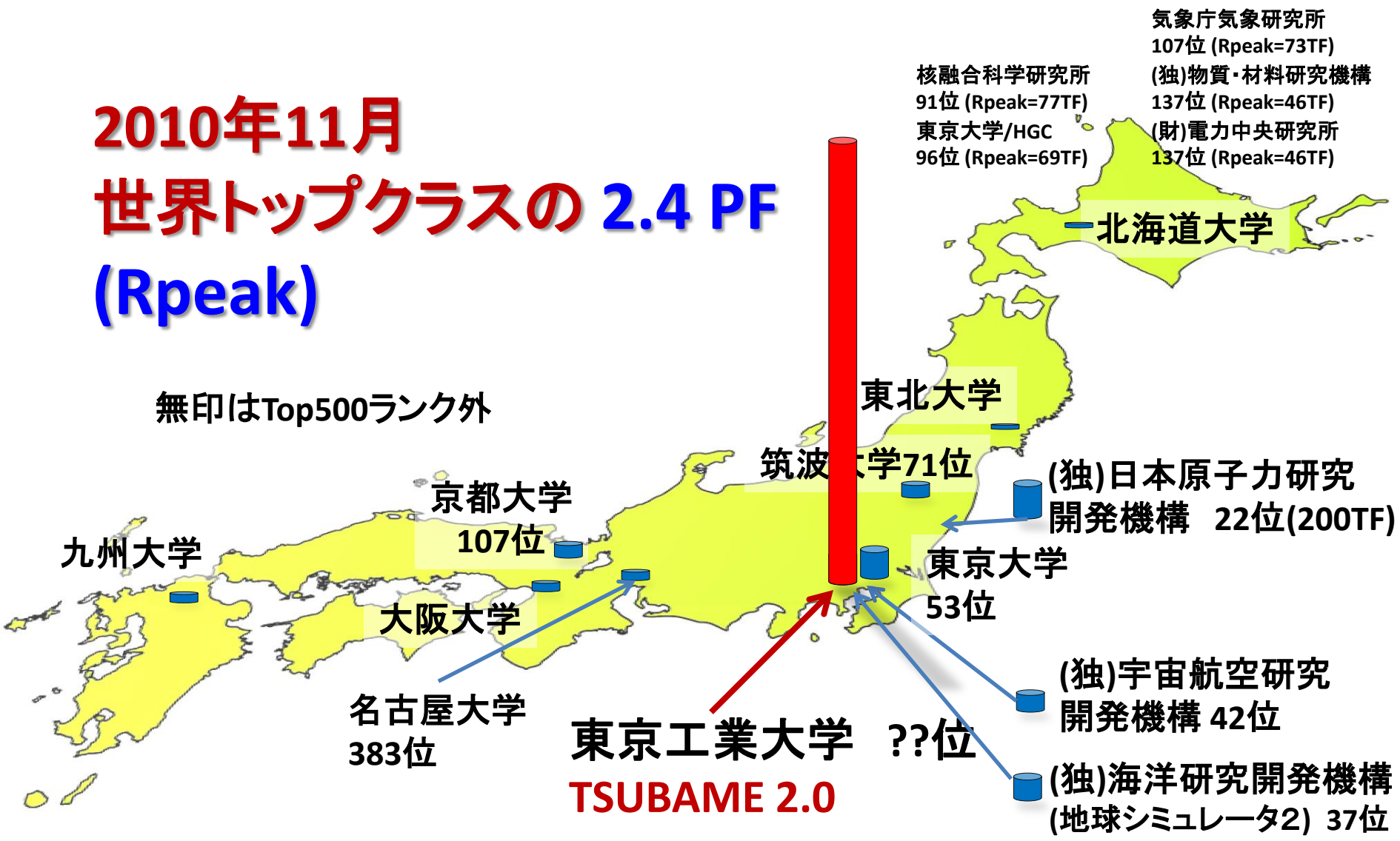
我国のスパコン(2010年合算1ペタフロップス以下)



2.4 ペタフロップス(単精度4.8ペタ)
0.72 ペタバイト/秒メモリバンド幅
世界トップ超高速光結合網-200テラビット/秒
総合15ペタバイト階層ストレージ
60ラック (200m2, TSUBAME1以下)
グリーン (1MW程度、PUE=1.28高効率水冷)
クラウド(WindowsHPC+Linux+VM)

2010年合算 1ペタフロップス以下
全国60か所
年額300 億円以上

日本主要スパコンとTSUBAME2.0



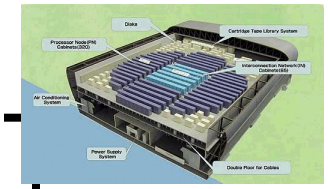
順位はTOP500 (2010 June)

TSUBAME2.0技術パートナーベンダー群

NEC: 主幹・全体設計及びインテグレーション・クラウド管理
HP: ノード開発・全体設計・グリーン; Microsoft: WindowsHPC・クラウド仮想化
NVIDIA: Fermi GPU (ベクトル) CUDA;　Voltaire: QDR Infiniband Network
DDN: 大規模ストレージ; Intel: Westmere & Nehalem-EX CPU (スカラー)
PGI: GPU用ベクトル化コンパイラ
東工大GSIC: 基礎設計・各種GPU技術・スパコン省電力技術・クラスタ運用

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?)

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

Others

TSUBAME 1.2
Upgrade
(3Q2008) 170TF

BlueGene/L
360TF(2005)

世界一ライン

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

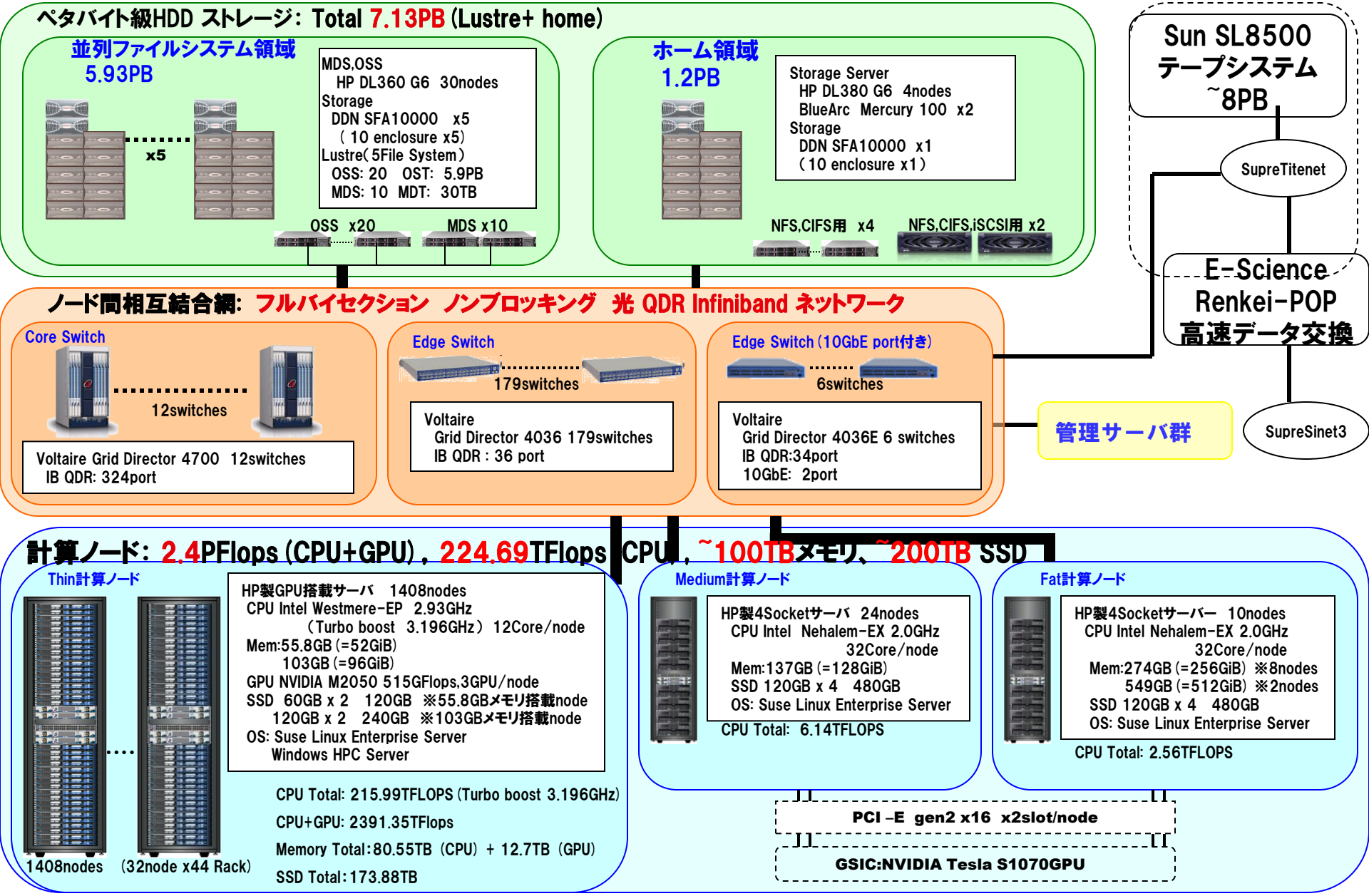
KEK 59TF
BG/L+SR11100

10PF
1PF
100TF
10TF
1TF

2002 2004 2006 2008 2010 2012


Push Here
Do Not
Push

TSUBAME2.0 システム概念図



	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 計算ノード群

HPと「共同開発」した新型のThin計算ノード（NVIDIA M2050を搭載）、および大容量メモリを搭載したMedium計算ノード，Fat計算ノードにより構成された計算環境

Thin計算ノード

IB QDR x2

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

HP製 GPU搭載用新設計サーバ
CPU: Intel Westmere-EP 2.93GHz x2 (12core/node)
※Turbo boost：3.196GHz
Memory: 55.8GB(=52GiB) DDR3 1333MHz
103GB(=96GiB) DDR3 1333MHz
SSD：60GB x2 (120GB/node) ※Memory 55.8GB搭載ノード
120GB x2 (240GB/node) ※Memory 103GB搭載ノード

1408nodes：215.99TFlops ※Turbo boost
4224GPUs：2175.36TFlops
Total： 2391.35TFLOPS
Memory：80.6TB（CPU） + 12.7TB（GPU）
SSD：173.9TB

Medium計算ノード

IB QDR

PCI-e Gen2x16 x2
※NVIDIA Tesla
S1070 GPU接続

HP製 4ソケットサーバ
CPU: Intel Nehalem-EX 2.0GHz x4 32core/node
Memory: 137GB(=128GiB) DDR3 1066MHz
SSD：120GB x4 (480GB/node)

24nodes: 6.14TFlops
Memory: 3.0TB+GPU
SSD 11.5TB

Fat計算ノード

IB QDR

PCI-e Gen2x16 x2
※NVIDIA Tesla
S1070 GPU接続

HP製 4ソケットサーバ
CPU: Intel Nehalem-EX 2.0GHz x4 (32core/node)
Memory: 274GB(=256GiB) DDR3 1066MHz
549GB(=512GiB) DDR3 1066MHz
SSD：120GB x4 (480GB/node)

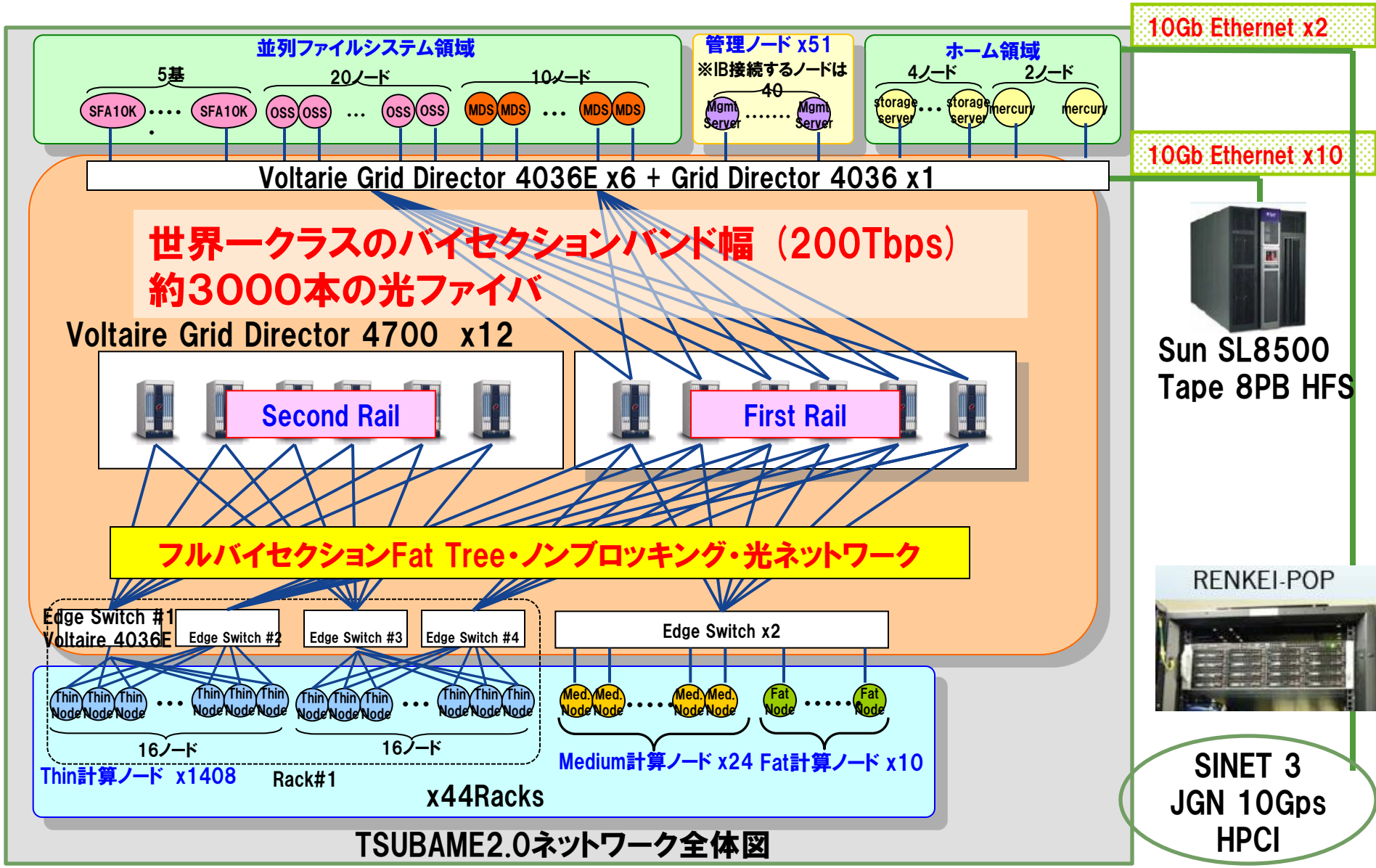
10nodes: 2.56TFlops
Memory: 3.0TB+GPU
SSD: 4.8TB+

CPU：224.69TFlops
GPU：2175.36TFlops

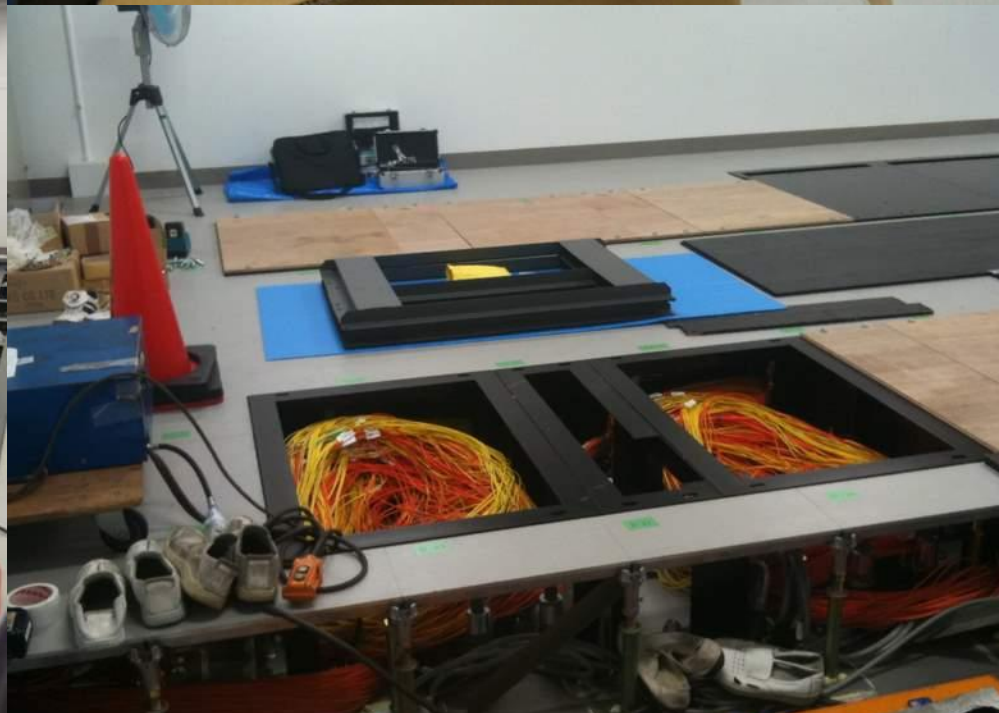
計算環境として
2.4PFlops
メモリ： 約100TB
SSD: 約200TB

2010年10月6日公開予定

TSUBAME2.0ノード間相互結合網



TSUBAME2.0ネットワーク全体図



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

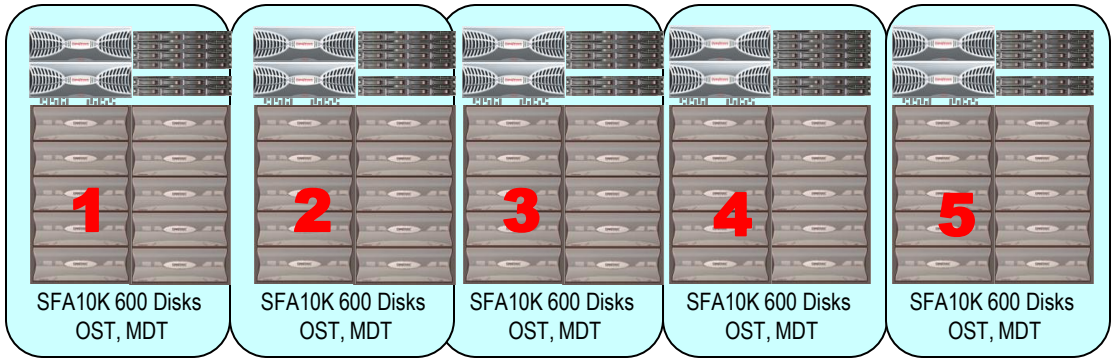
Lustre 並列ファイルシステム領域

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

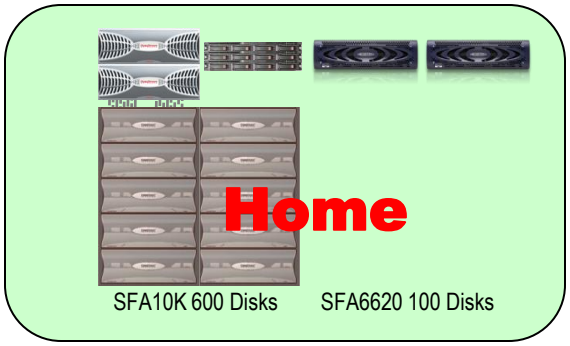
ホーム・クラウドサービス用領域

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


DataDirect

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

【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

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)

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

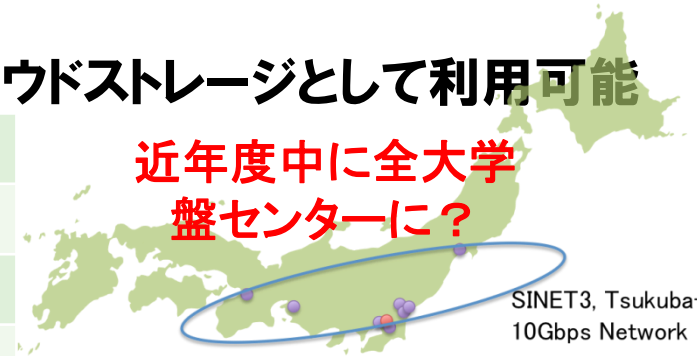
東工大 e-Science RENKEI-POP による分散ストレージ・HPCIへの貢献

- 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

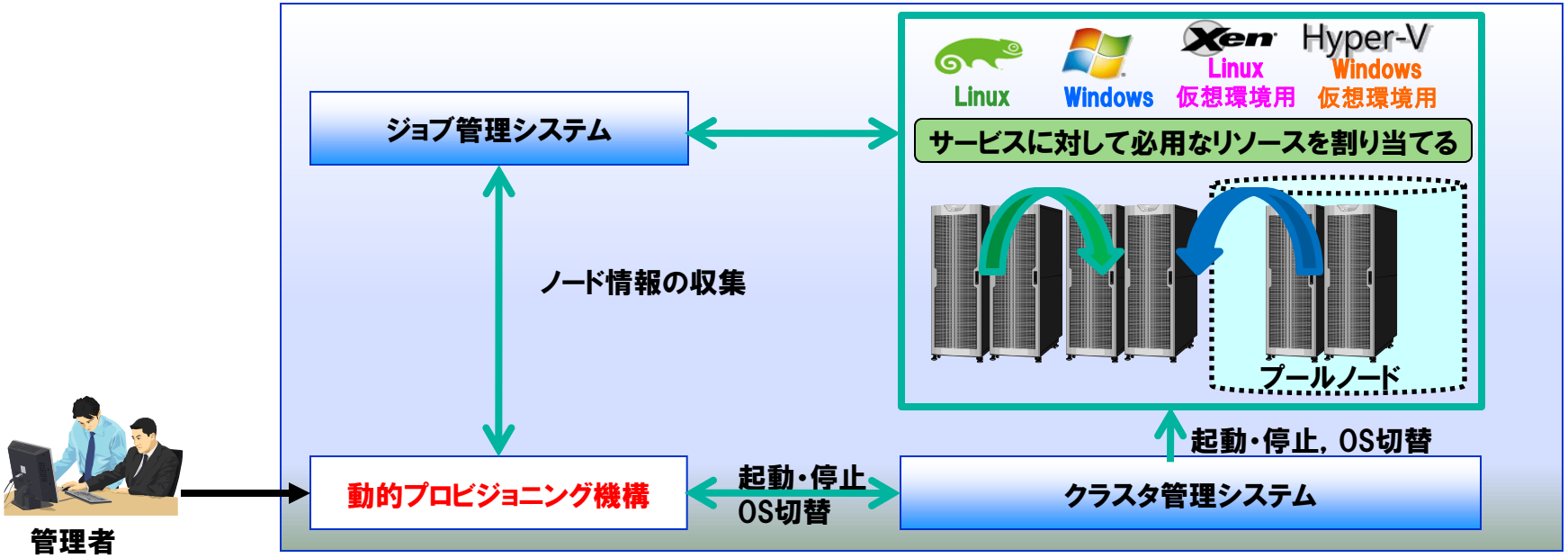
東京工業大学	大阪大学
国立情報学研究所	高エネルギー加速器研究機構
名古屋大学	筑波大学
産業技術総合研究所	東北大学



TSUBAME2.0クラウド運用形態

OSを動的に変更する「動的プロビジョニング機構」

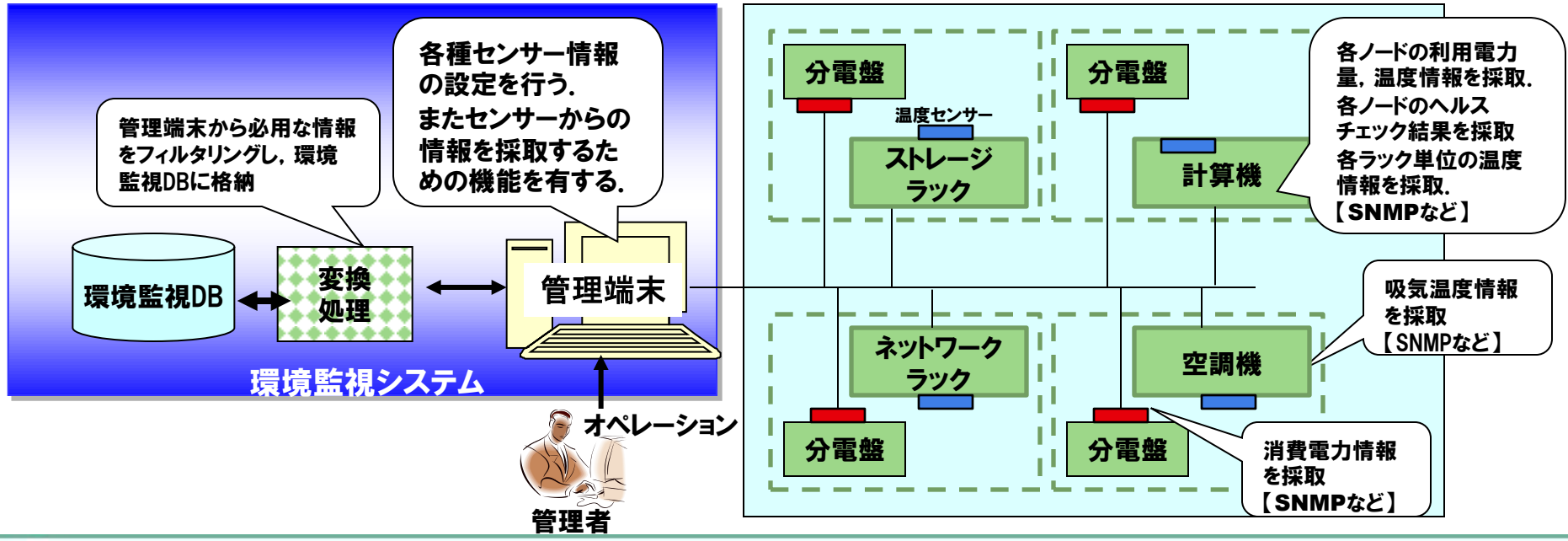
- 「ジョブ管理システム」「クラスタ管理システム」と連携
- プールノード(サービスに割り当てられていない計算ノード)を利用し，余剰の計算リソースをサービスに割り当て
- Linux,Windows双方のバッチスケジューラにより計算ノードを協調管理
- Linuxノード，Windowsノードの動的な増減が可能
- バーチャルマシン上の仮想計算ノードもスケジュール可能な資源として動的にバッチスケジューラの管理対象



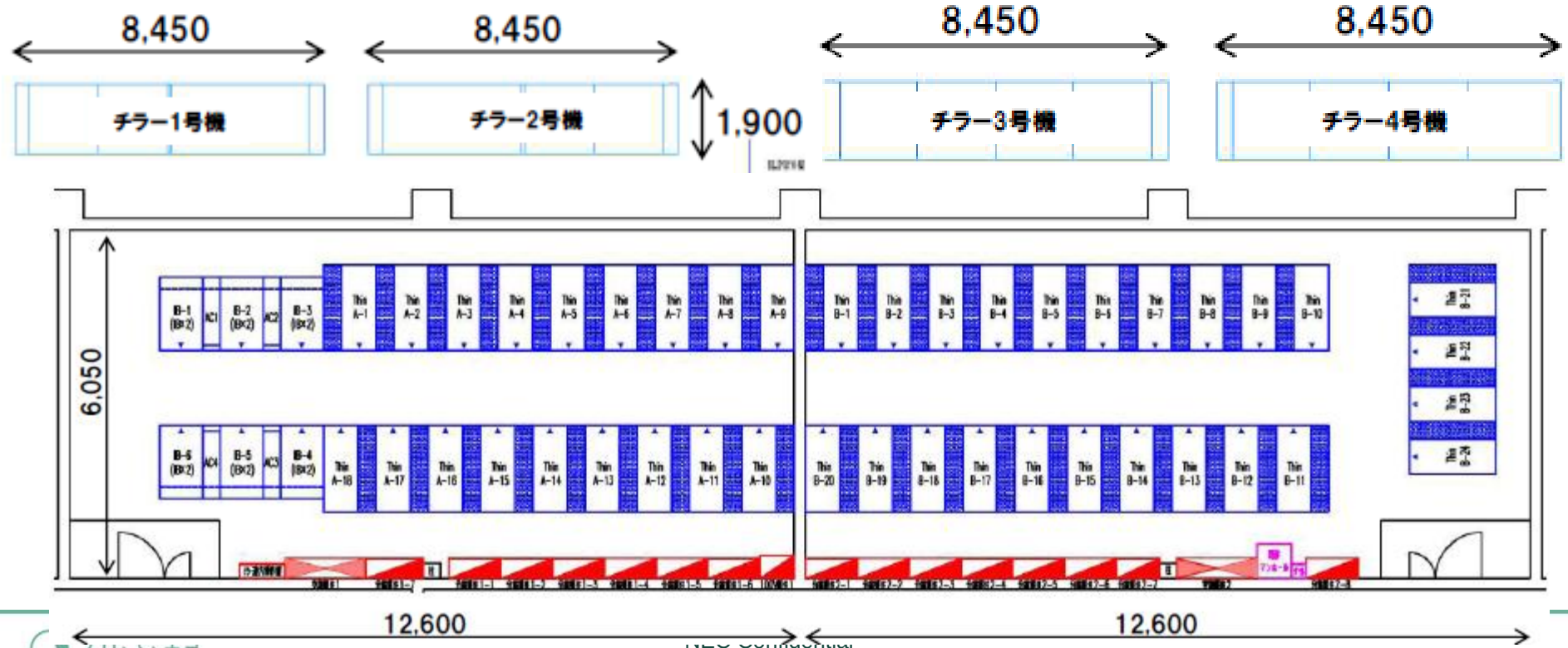
グリーンスパコン: 環境監視システム

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）

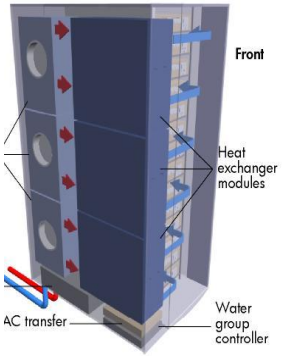


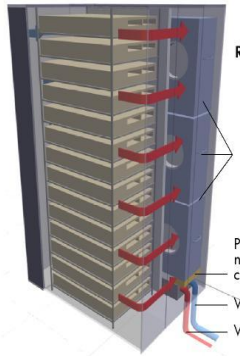
TSUBAME2.0 レイアウト
(全体で200m2程度, TSUBAME1の2/3以下)



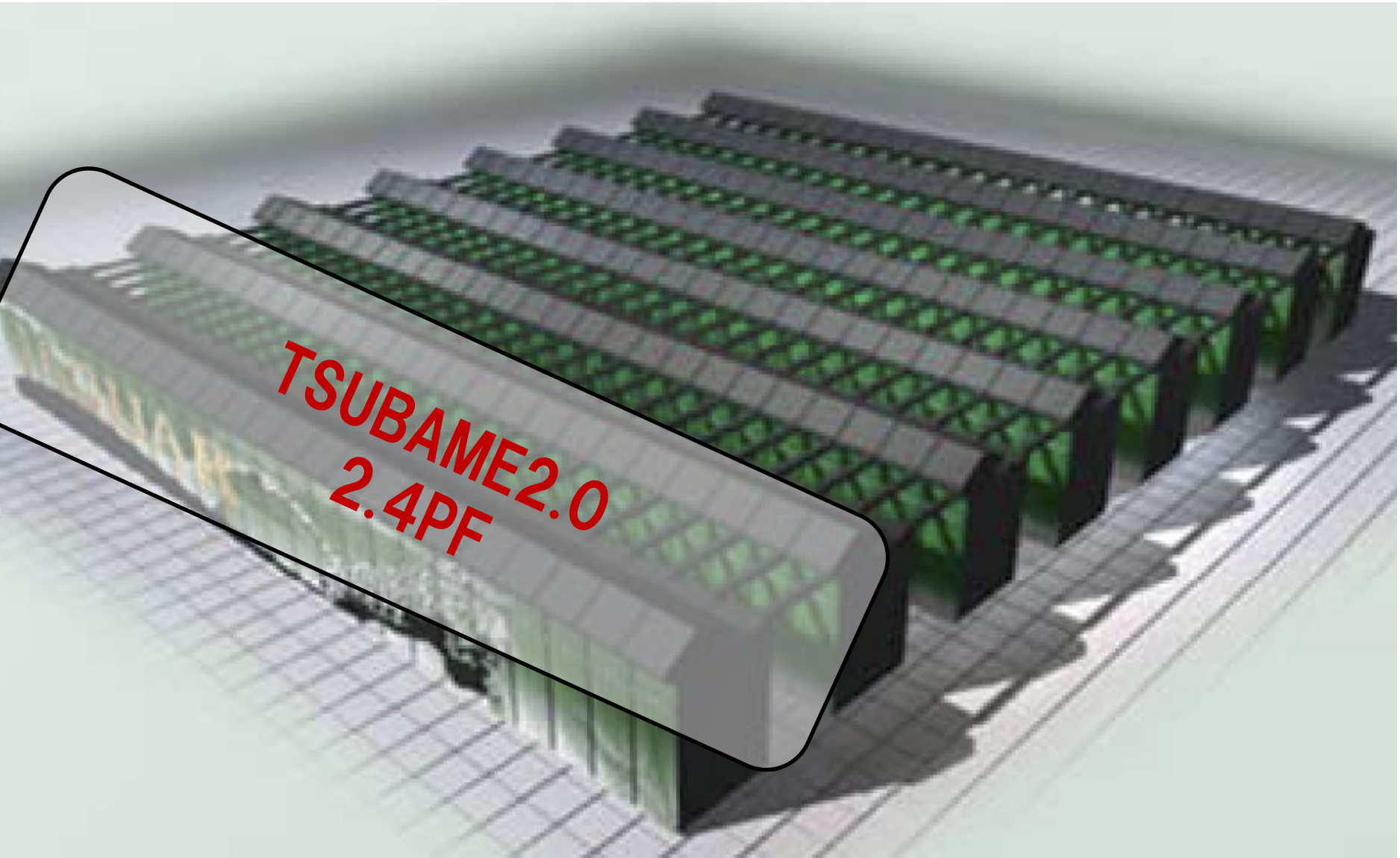
グリーンスパコン：HP Modular Cooling System G2 による高密度実装・水冷キャビネット冷却

- ラック内に熱交換システムを内蔵した密閉型水冷システム
- 高密度な冷却が可能・ラックあたり最大35kW（世界最高）
  通常のデータセンターの10倍!!
- サーバの吸入口に均質な冷却風を提供
- ドア平開は自動化・加湿不要
- 完全自動温度制御による最適な消費電力点の制御
- 95% から 97% の熱を水冷で除去
- ポリカーボネート製のドアにより大幅なノイズ削減



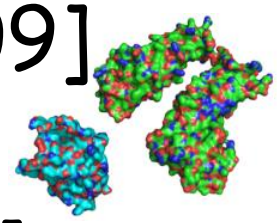


ORNL Jaguar and Tsubame 2.0
Similar Peak Performance, 1/4 the Size and Power

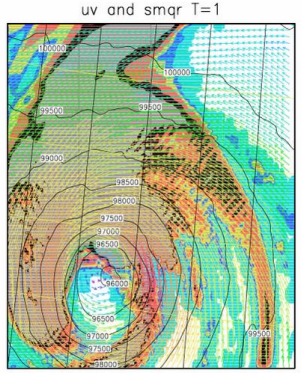


TSUBAME2.0 アプリケーション性能予測

- ~1.4 PFlops Linpack [IEEE IPDPS 2010]
- ~0.5 PFlops **タンパク質ドッキング** [SC08,09]
  - **本学秋山泰教授との共同研究**(Microsoft-TCI)



- 100-150 TFlops *ASUCA* **気象予測**[SC10]
- HPC-Challenge Performances
  - **テネシー大** Jack Dongarra**教授ら**GPU**用を開発へ**
- **その他多くのアプリでベクトル性能発揮**
  - QCD, Lattice-Boltzmann CFD,
    分子シミュレーション、ゲノム解析、
    大規模サーチなどでペタフロップス級



青木G（東工大）

TSUBAMEにおける次世代気象予測

メソスケール大気モデル:

雲の解像： 3次元非静力学平衡モデル

Compressible equation taking consideration of sound waves.

メソスケール

2000 km

台風

a few km

竜巻, ダウンバースト
集中豪雨

従来の研究: WRF の GPU Computing

- **WRF (Weather Research and Forecast)**

  Community Code developed by 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 scales

  ⇒ x 41.1 increase

•Michalakes, J. and M. Vachharajani: GPU Acceleration of Numerical Weather Prediction. *Parallel Processing Letters Vol. 18 No. 4. World Scientific. Dec. 2008. pp. 531—548*

**John C. Linford, John Michalakes, Manish Vachharijani, and Adrian Sandu. Multi-core acceleration of chemical kinetics for simulation and prediction, *proceedings of the 2009 ACM/IEEE conference on supercomputing (SC'09), ACM, 2009.*

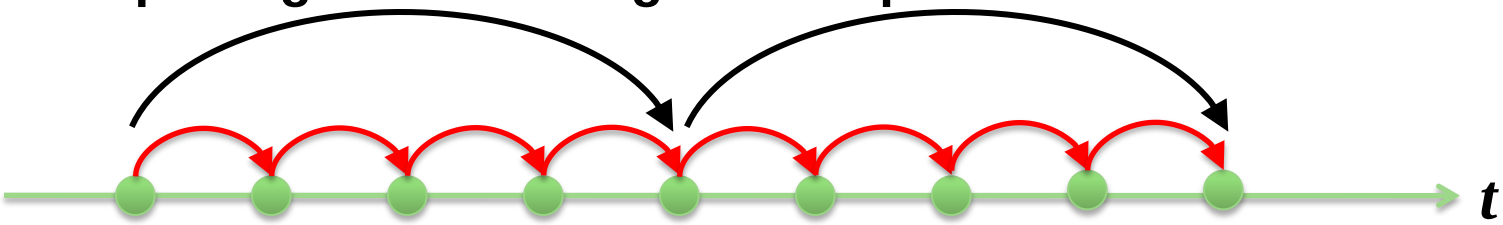
ASUCA versus WRF

- **ASUCA：気象庁が開発している次世代気象モデル**

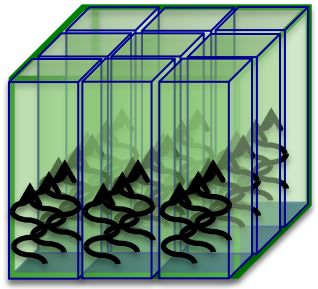
次期気象予報の現業モデルとして期待

メソスケール・非静力学平衡モデル

Time-splitting method: long time step for flow



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



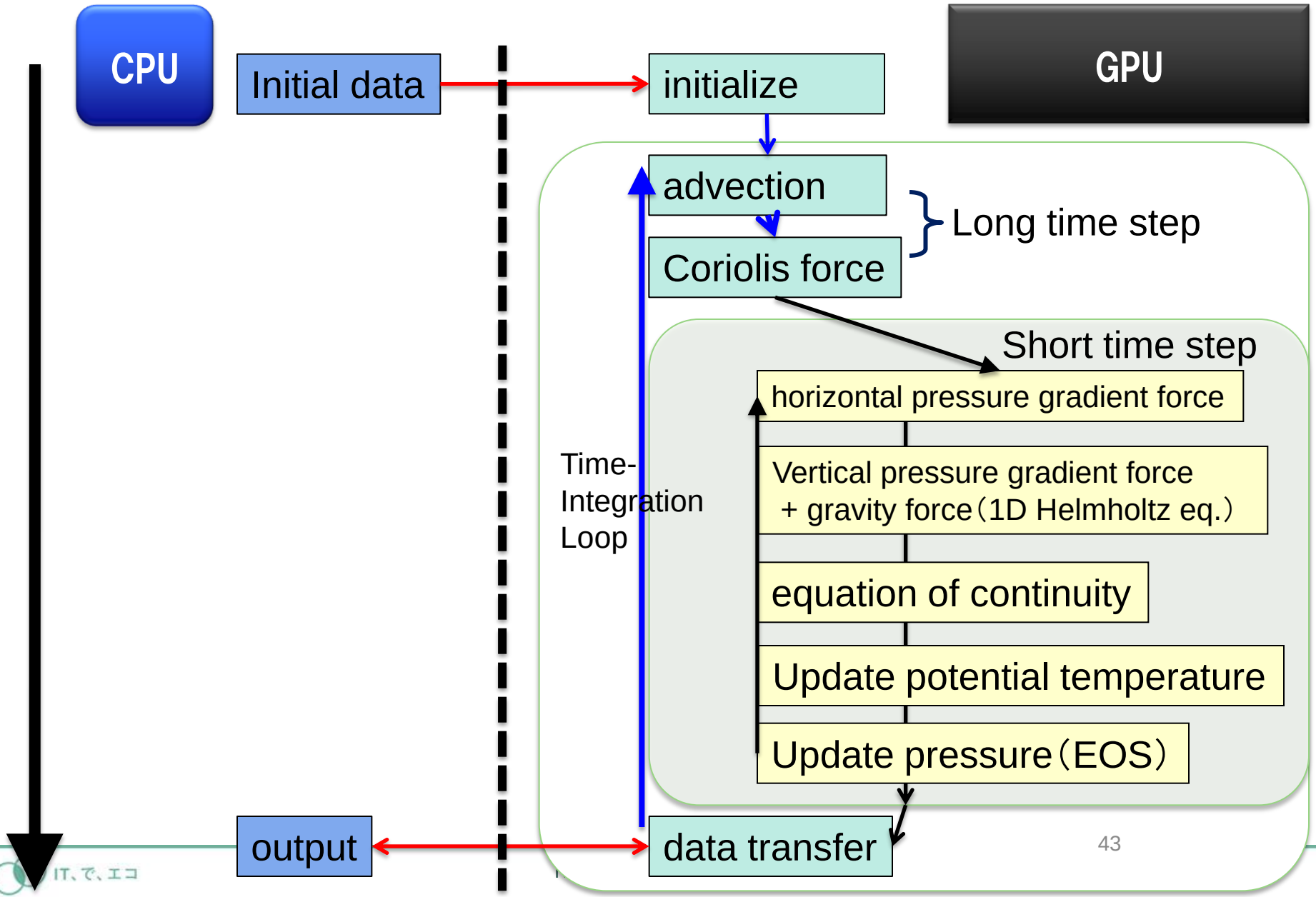
HEVI (Horizontally explicit Vertical implicit) scheme

Horizontal resolution ~ 1 km

Vertical resolution ~ 100 m

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

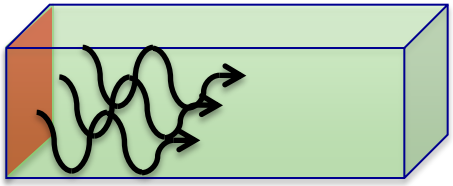
ASUCAにおける計算の流れ



Implementation

Thread

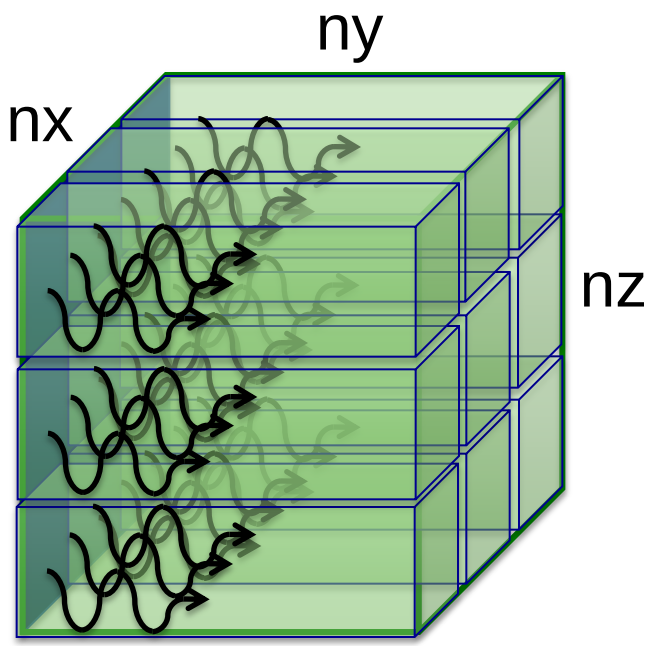
Block



64 x 4 threads (2D) in a block

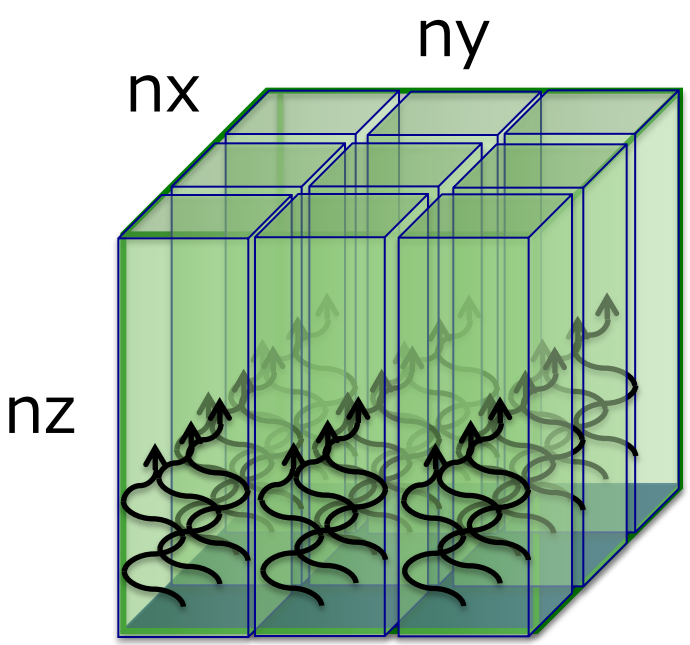
- **3D Advection equation**

  Each thread specifies a (x, z) point, marching in y

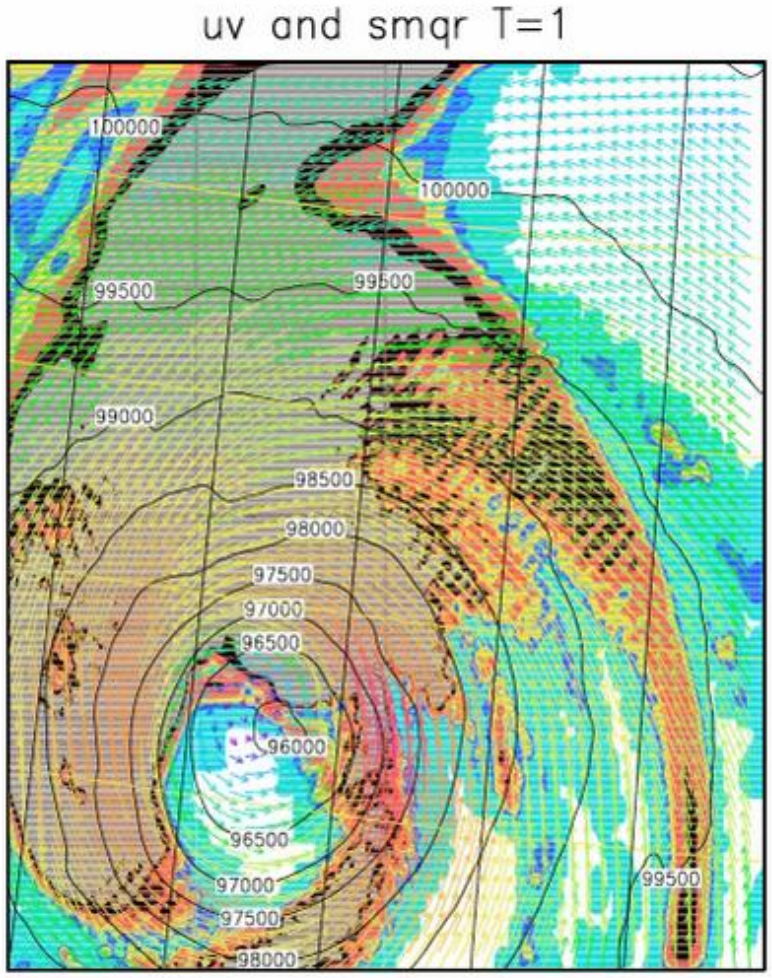


- 1D Helmholtz equation

  Element in k depends on elements in k+/- 1, marching in z direction



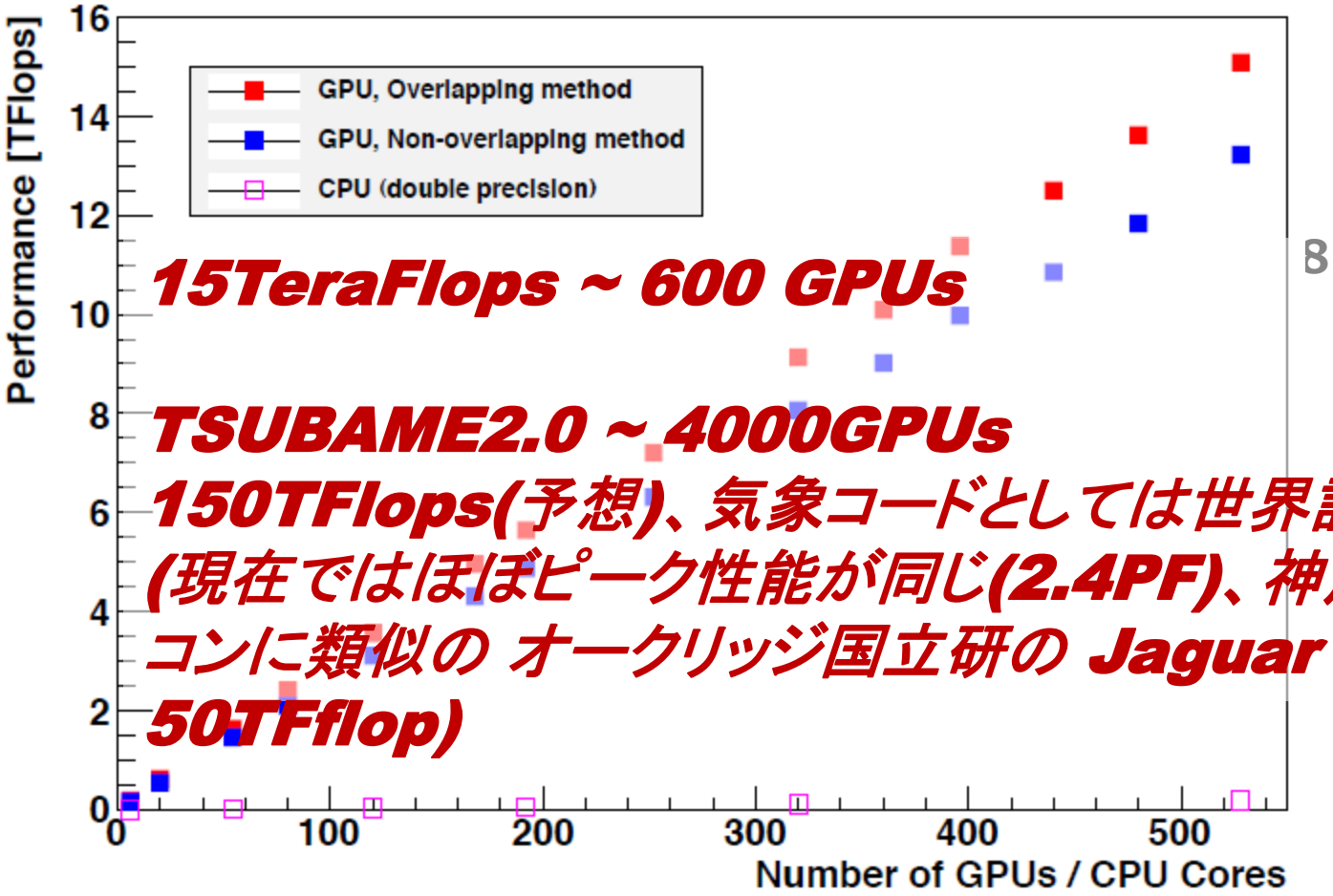
ASUCA 台風シミュ 2km mesh 3164×3028×48



70分の実行時間で、6時間のシミュレーション時間(実時間の5倍)

TSUBAME2.0だと0.5kmメッシュが実時間で可能に

ASUCA Multi GPU Performance (TSUBAME1.2) Supercomputing 2010で発表



15TeraFlops ~ 600 GPUs

TSUBAME2.0 ~ 4000GPUs
150TFlops(予想)、気象コードとしては世界記録(現在ではほぼピーク性能が同じ(2.4PF)、神戸ペタコンに類似の オークリッジ国立研の Jaguar の 50TFflop)

Mountain Wave Test in single precision
NVIDIA Tesla S1070 on TSUBAME

対**Jaguar**電力性能比 **8-12**倍
コスト性能比**10**倍以上

TSUBAME の共同利用

平成19年度～ **民間企業へ TSUBAMEを提供**

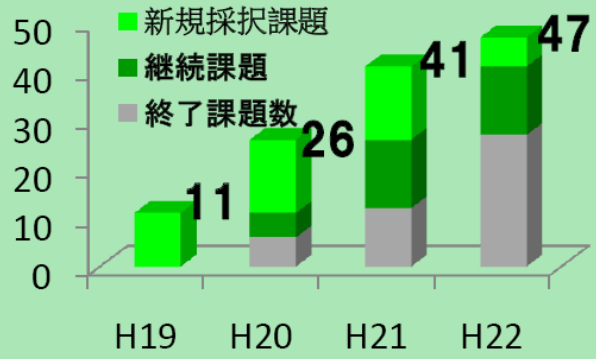
- 文部科学省 先端研究施設共用促進補助事業

平成21年度～ **TSUBAME 共同利用**

- 東工大独自事業として、外部利用制度の確立

「産業利用」にて大きな成果

4年間で民間企業の47課題を採択・実施



成果報告会 6/29(火) 13:00～ 蔵前会館

東工大・TSUBAME共用促進シンポ

文部科学省より最高の評価

(22機関中、3機関のみ)

平成22年度～ **学際大規模情報基盤共同利用・共同研究拠点**

- 「ネットワーク型」の共同利用・共同研究拠点

日本最先端のスパコン環境を提供し、学術・産業・社会へ貢献

最新のGPU環境を提供し、 新世代ベクトル計算の技術普及