JST ULP-HPC

次世代テクノロジのモデル化・最適化による超低消費電力ハイパフォーマンス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
研究基盤部門　教授

代表者　松岡 聡

2012/11/30 ULP最終最終シンポジウム発表資料

過去のスパコンのイメージ…
10年で1000倍の向上→電力を犠牲に

今や電力がスパコンの速度向上の主律束に

2016年*ULP-HPC*技術により
デスクサイド・ペタスケールコンピューティングへ
(Mid-size to Deskside Scaling)

2006年東工大TSUBAME
アジア一位：85TeraFlops,
> 10,000 CPU, 1.2 MegaWatt, 350m^2

ペタコン技術の
1000倍のダウン
スケール：
どうやって？

2016年デスクサイドワークステーション
~100TeraFlops, 1.5KiloWatt,
350cm^2

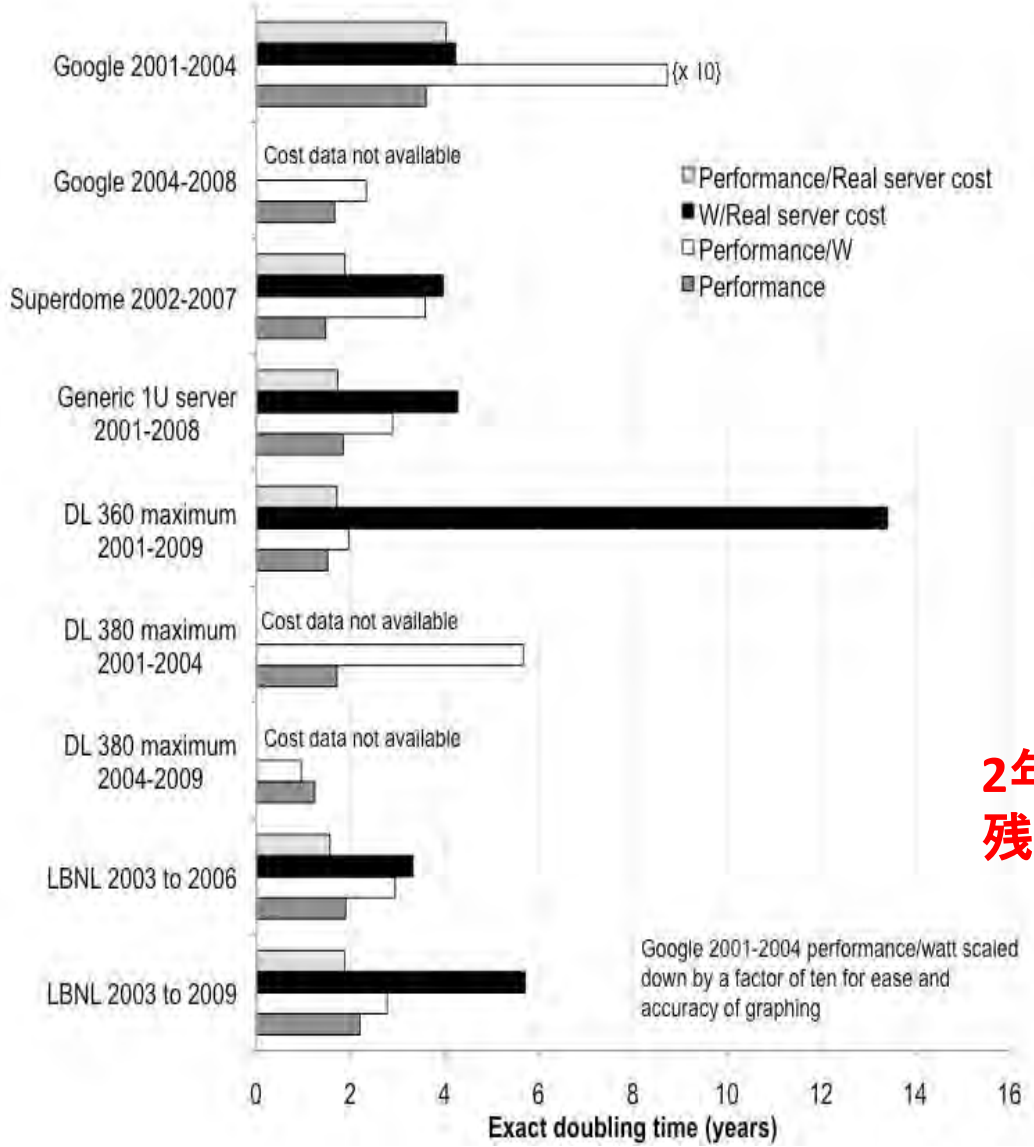
ULP-HPC
1000倍(あと×10)の技術ブレークスルーを行う情報科学研究

「ペタスケールシミュレーションがエンジニアにとって日常的に」

単純なスケーリングでは達成不可

クロック上昇不可
「マルチコア」でもせいぜい数十~100倍

現代版Mooreの法則では「2年で2倍」

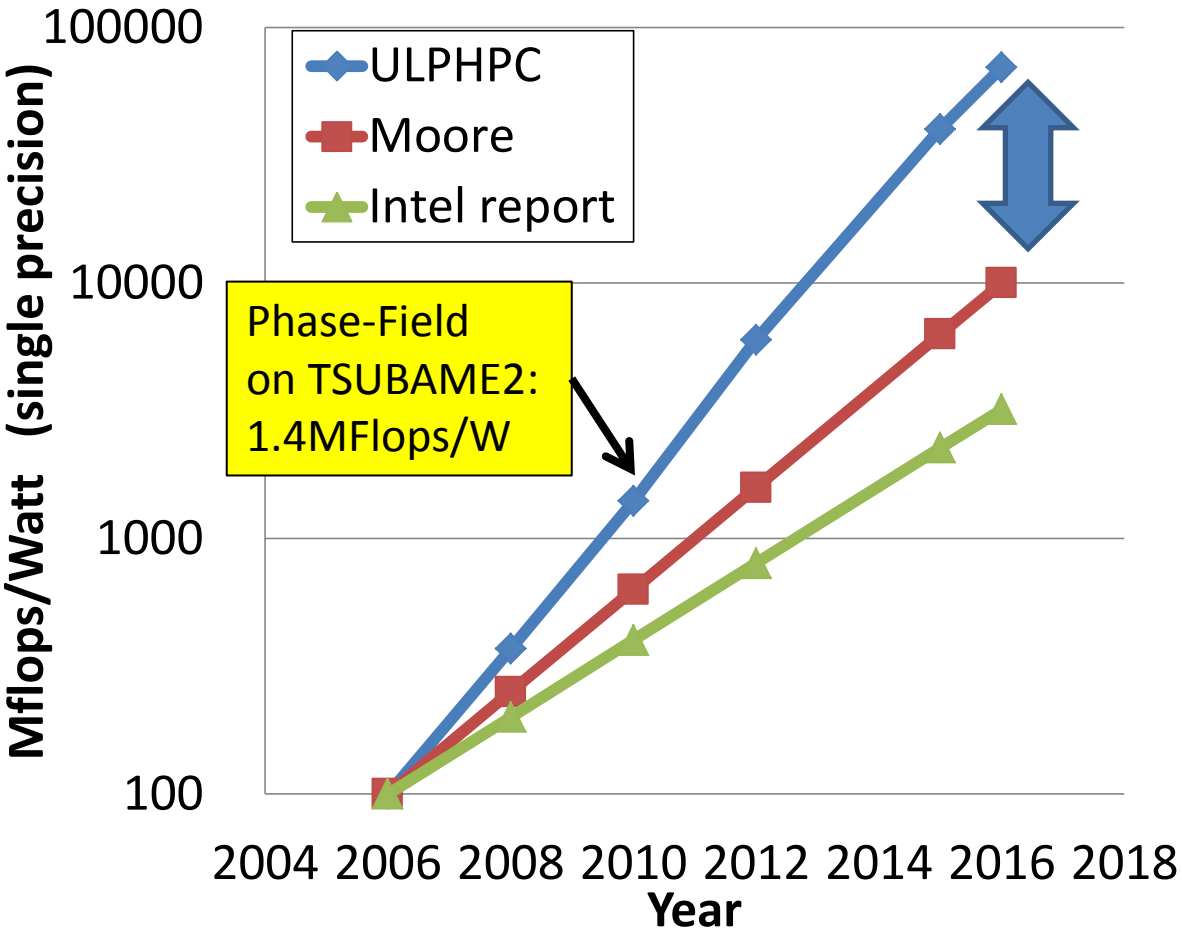


Source: **Assessing trends over time** in **performance**, **costs**, and **energy use** for **servers**, **Intel, 2009.**

"Performance per server and performance per thousand dollars of purchase cost double every two years or so, which tracks the typical doubling time for transistors on a chip predicted by the most recent incarnation of Moore's law"

2年間2倍→10年で高々32倍
残りの30倍のギャップをどう埋める？

10年で電力性能比1000倍の目標



プロセスルール改良・Moore's Lawだけでは10年で100倍

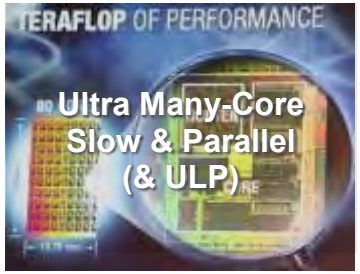
1000倍とのギャップを埋めるのがULP-HPC技術

(グラフは冷却電力向上分を除く)

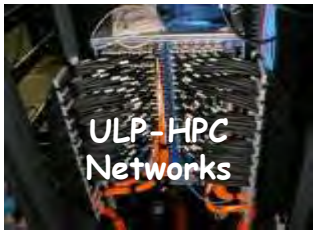
省電力化の手法と有効性

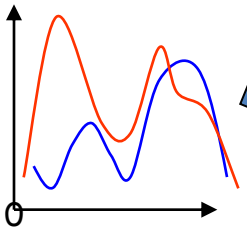
情報基盤/手法	エンタープライズ・ビジネス・クラウド	HPC
仮想化による統合化 (Server Consolidation)	○	×
DVFS (Dynamic Voltage/Frequency Scaling)	○	△
新デバイス （と、それを生かしたアーキテクチャ)	△ (コストや継続性)	○
新アーキテクチャ （と、それを生かすソフトウェア）	△ (コストや継続性)	○
冷却技術 (排熱利用やデバイス省電力化含)	△	○ (ただし高熱密度)

我々のCREST *ULP-HPC*全体スキーム



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





省電力高性能
システムソフト機構・モデル化

(提案1) ULP-HPC用次世代SW/HW要素利用
技術およびモデル化
(松岡@東工大・本多@電通大、鯉渕@NII)

(提案3) 自動チューニング共通基盤
(須田@東大)

モデルと実測の Bayes 的融合

- Bayes モデルと事前分布

$$y_i \sim N(\mu_i, \sigma_i^2)$$

$$\mu_i \mid \beta, \sigma_i^2 \sim N(x_i^T \beta, \sigma_i^2 / \kappa_0)$$

$$\sigma_i^2 \sim \mathrm{Inv}\text{-}\chi^2(\nu_0, \sigma_0^2)$$

モデルによる
所要時間の推定

- n 回実測後の事後予測分布

$$y_i \mid (y_{i1}, y_{i2}, \cdots, y_{in}) \sim t_{\nu_n}(\mu_{in}, \sigma_{in}^2 \kappa_{n+1} / \kappa_n)$$

$$\nu_n = \nu_0 + n, \quad \kappa_n = \kappa_0 + n, \quad \mu_n = (\kappa_0 x_i^T \beta + n\bar{y}_i) / \kappa_n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + \sum (y_m - \bar{y}_i)^2 + \kappa_0 n (\bar{y}_i - x_i^T \beta)^2 / \kappa_n$$

所要時間の実測データ

ABCLibScript: アルゴリズム選択

```
!ABCLib$ static select region start
!ABCLib$ parameter (in CacheS, in NB, in NPrc)
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$      (2.0d0*CacheS*NB)/(3.0d0*NPrc)
!ABCLib$    対象1（アルゴリズム1）
!ABCLib$    select sub region end
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$      (4.0d0*ChcheS*dlog(NB))/(2.0d0*NPrc)
!ABCLib$    対象2（アルゴリズム2）
!ABCLib$    select sub region end
!ABCLib$ static select region end
```

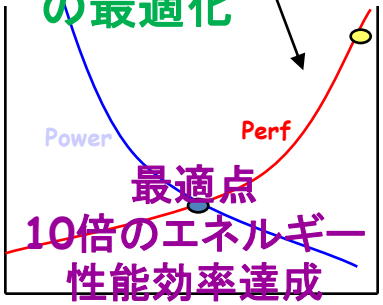
実行起動前自動チューニング指定、アルゴリズム選択処理の指定

コスト定義関数で使われる入力変数

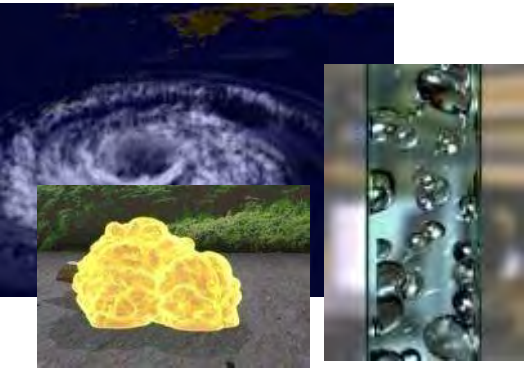
コスト定義関数

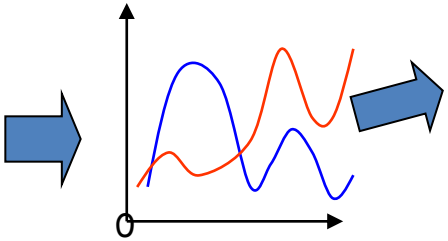
対象領域1、2

電力性能比
の最適化



最適点
10倍のエネルギー
性能効率達成





省電力高性能
アルゴリズム・モデル化

(提案2) 超省電力型のHPCアプリケーション
及びアルゴリズム(青木@東工大)

2016年 TSUBAME
1/1000に

Total System FLOPS/W, MemBW/W considered significant

- "Power" (and cost) is THE constraining factor in large scale supercomputers
  - Determines the size of the system
- Large scale apps performance limited by <u>TOTAL system FLOPS</u>, or <u>TOTAL Memory Bandwidth</u>
  - (Traditional Byte/Flop argument not very useful)
- FLOPS/W = Total Sys FLOPS / Total Sys W
  - ~"Green 500"
- MemBW/W = Total Sys MemBW / Total Sys W
  - ~"Green Graph 500"
- Exascale 50GFlops/W, 5GByte/s / W (B/F=0.1)
  - x20 improvement in 8 years?

ULP-HPC 遂行計画・マイルストーン

2007Q3

基礎開発

各種要素技術の研究開発

- マルチコア・アクセラレータ関係の種々の電力最適化アルゴリズム・プログラミング機構・コンパイラ等
- 新世代メモリ利用を含むプログラム
- VMベースのシステム全体の電力最適化
- 低消費電力・高性能ネットワークの設計活用
- 流体構造および分子科学における省電力高性能アプリケーションアルゴリズム
- 上記の電力性能モデル化、自動チューニング適用
- 電力性能自動チューニング基盤(理論・システム)

2009Q1

第一次統合化

各種要素技術の連携・統合開始

- 各要素のモデルベースを統合化した自動チューニングシステムを開発
- 結果をTSUBAME2.0の設計に反映

-各種テストベッドでの実際のソフトウェア機構開発・モデルの制定・自動チューニング
-モデルベースのシミュレーション遂行・大規模性能電力最適化シミュレーション
-一部HW小プロトタイプの開発

TSUBAME2.0の設計

他の関連マシン

2010Q3

TSUBAME 2.0 完成・設置: 本研究の成果を用い20倍以上の性能向上を達成

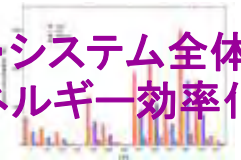
新基盤上での更なる開発・最終統合化

2012Q4

HWの要素活用による
ポテンシャル x3~x30

アプリ最適化(x5)・
システム最適化 x1.5(平均)
冷却最適化 (x1.4)

自動チューニングで
両ポテンシャル最大化・システム全体
平均で10倍程度のエネルギー効率化



TSUBAME3.0の設計
(2015年度完成)

1000倍の想定内訳

プロセスルール縮小　100倍

×

メニーコア/GPU等活用
ソフトウェア技術　5倍

×

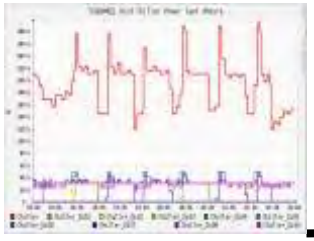
動的電力制御技術　1.5倍

×

冷却電力の削減　1.4倍

↓

1000倍

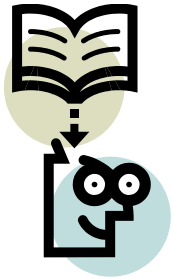


ULP-HPCの
主研究対象

ウルトラグリーン
スパコン(後述)

ULP-HPC 査読論文・講演等

- **査読論文: 122件**
  - 国内（和文）誌33件、国際（欧文）誌 89件
- **基調講演・招待講演：合計152件**
  - 国内会議 67件、国際会議 85件
  - Satoshi Matsuoka, the First Petaflop/s System in the World and Its Impact on Supercomputing, Opening keynote talk, ISC 08 (International Supercomputing Conference) @ Dresden, Germany, June 2008.
  - Satoshi Matsuoka. Petascaling Commodity onto Exascale: GPUs as Multithreaded Massively-Parallel Vector Processors - the Only Road to Exascale. Keynote Talk, IEEE Cluster Computing Conference 2009, New Orleans, USA, Sep.3, 2009.
  - Takayuki Aoki: Large-scale Stencil Applications on GPU-rich Supercomputer TSUBAME2.0, The annual IEEE International Conference on High Performance Computing (HiPC 2011), Bangalore, India, December 21, 2011
- **口頭発表・ポスター：262件**
  - 国内会議　185件、国際会議　77件
- **派生プロジェクト**：丸山ポストペタCREST,遠藤ポストペタCREST、科研基盤(S),日仏FP3C, G8気候シミュレーション、グリーンスパコン概算要求、TSUBAME2.5…



ULP-HPC受賞等 (国際賞10件)

- **松岡　聡,2010年11月発表のGreen500グリーンスパコンランキングにおいて世界2位，およびthe Greenest Production Supercomputer in the World賞を獲得**
- 松岡　聡, 2010年11月発表のTop500スパコンランキングにおいて世界4位
- **額田　彰, IEEE Japan Chapter Young Author Award 2010, 2010年12月**
- **下川辺・青木・額田・遠藤・丸山・松岡ら：2011年11月SC11において，ACM Gordon Bell Prizes: Special Achievement Award in Scalability and Time-to-Solution受賞**
- **イタリアCNR Massimoらと共同で，遠藤・松岡ら：SC11において，ACM Gordon Bell Prizes: Honorable Mention**
- 松岡ら：HPCwire Annual Award:
  - Reader’s Choice Award - Best application of "green computing" in HPC
  - Editor’s Choice Award - Best application of "green computing" in HPC
  - Reader's Choice Award - Best HPC collaboration between government and industry
- 鯉渕: Best Paper Award, the Second International Conference on Networking and Computing(ICNC), 2011年12月
- 遠藤：情報処理学会山下記念研究賞，2012年3月
- 松岡ら：HPC Wire “Number Crunching, Data Crunching and Energy Efficiency: the HPC Hat Trick” (TSUBAME2.0のGraph500, Top500, Green500上位に対して)
- **松岡・遠藤・青木：文部科学大臣表彰科学技術賞（開発部門）「運用世界一グリーンペタスパコンの開発」**

報道等

- **TSUBAME2.0を用いた成果・受賞・グリーンスパコン等に関する多数報道**
  - 読売新聞(11/20), 朝日新聞(11/10), 日経産業新聞(2/24)
  - 日経コンピュータ, WIRED
  - クラウドWatch (Gordon Bell賞受賞に関して)
  - その他asahi.com, HPCwire, IT media等

日経産業新聞

10

テクノトレンド

先端技術

「京」より100倍速い次世代スパコン

米欧中は開発推進

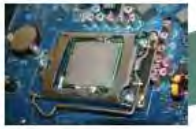
TSUBAME2.0 2010年11月1日稼働開始
世界最小のペタフロップス・省電力スパコン

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

ULP-HPC
メーカー新規共同開発

Chip
(CPU ,GPU)



CPU(Westmere EP)
76.8 GFLOPS

GPUs(Tesla M2050)
515 GFLOPS
3 GB

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB

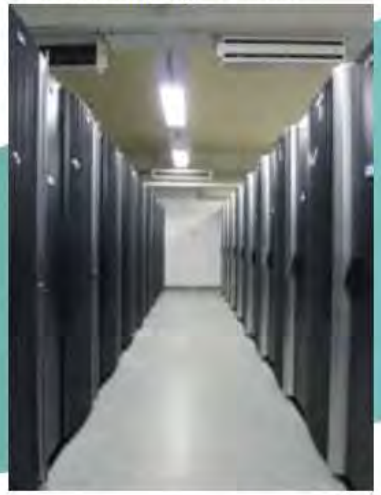
Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB

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

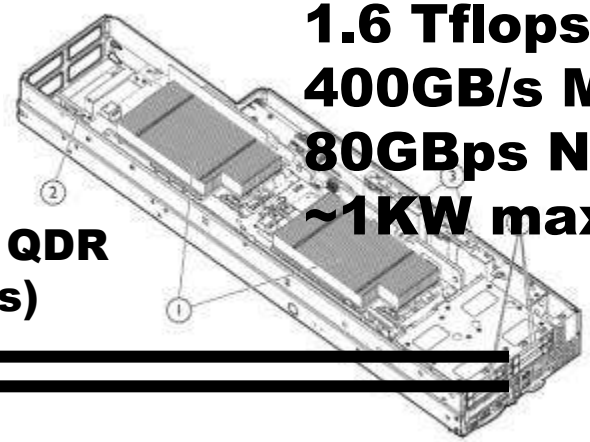


2.4 PFLOPS
80 TB

Integrated by NEC Corporation

TSUBAME2.0 Compute Node

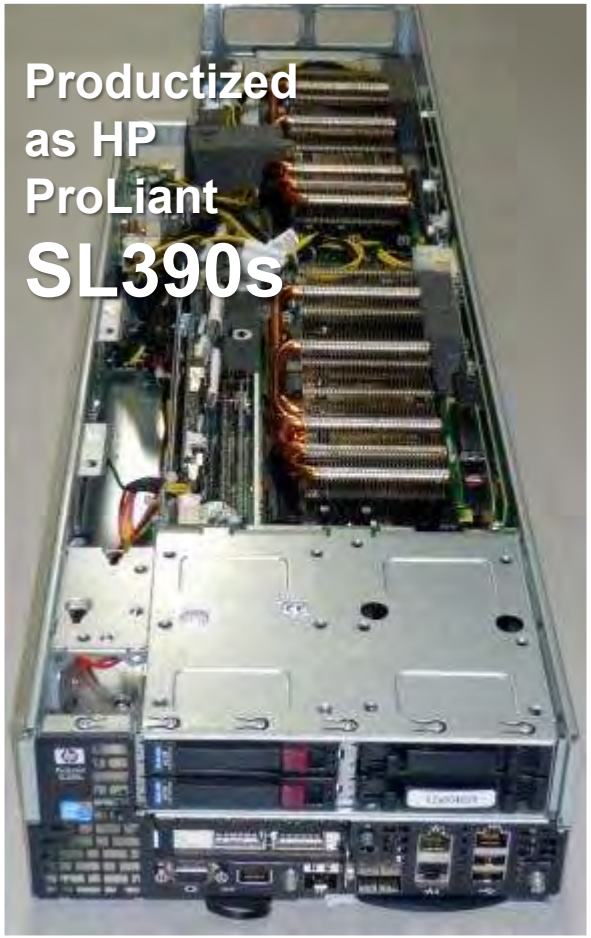
Thin Node



1.6 Tflops Peak
400GB/s Mem BW
80GBps NW BW
~1KW max

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)
Multi I/O chips, 72 PCI-e (16 x 4 + 4 x 2) lanes --- 3GPUs + 2 IB QDR
Memory: 54, 96 GB DDR3-1333
SSD：60GBx2, 120GBx2



TSUBAME2.5
4PF Peak, 800GB/s Mem BW

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

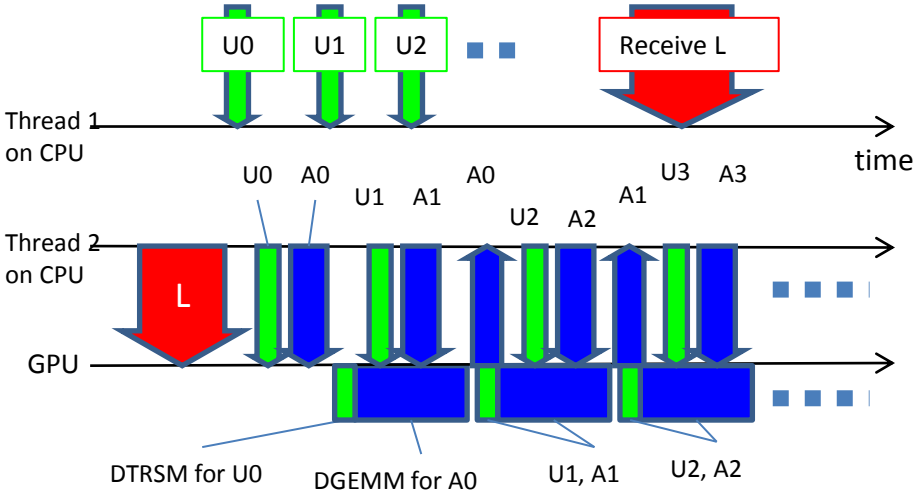
TSUBAME2.0へのULP-HPCの成果の反映

- GPU中心の2.4 PF我が国最速・世界4位のスパコン
  - 1432 nodes, Intel Westmere/Nehalem EX
  - 4224 NVIDIA Tesla (Fermi) M2050 GPUs
  - ~76,600 total CPU and GPU "cores", High Bandwidth
  - 活用するソフト・アプリ・最適化の研究⇒電力効率世界一？(後述)
- 省電力運用のための仕組みが随所
  - I/O・ネットワークの新デバイス採用(マルチDFB発光素子一体型ケーブル・SSDデバイス等)による省電力化
  - ノード・ラック・配電版など随所の電力センサーネットワーク
  - 大量の温度センサー(ノード18個=>全体で2万個以、ファンセンサー…)
  - ノード単位の電力キャップ・高効率冷却ノード設計
  - 密閉型水冷ラック/チラー (35KW => PUE=1.28以下)
- 今後さらなるULP-CRESTの成果の適用
  - 開発された高性能・省電力GPUライブラリ等の利用
  - 省電力自動チューニング
  - 省電力スケジューリング(温度感知=>マイグレーション)

TSUBAME2.0の4000超GPUを用いたLinpack実行 [IPDPS10, SACSIS11]

数百～数千アクセラレータを効率利用するLinpackソフトウェアを継続開発

- TSUBAME1.0→1.1→1.2→2.0で連続的に性能向上を実現
- アクセラレータやノードのヘテロ性・PCI通信の影響・アクセラレータメモリの小ささに対応する最適化



TSUBAME2.0全体を用いて

1.192PFlops (実行時間2.4h)

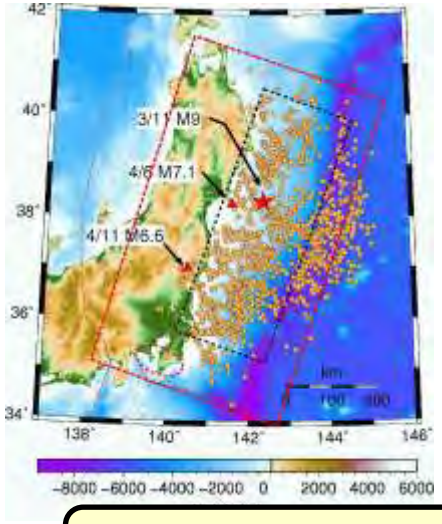
⇒　Top500世界4位

958MFlops/Watt

⇒　Green500世界2位

運用スパコン1位

TSUBAME2.0スパコンの電力危機対応



2011・3・11東日本大震災

- 目黒区に設置されたTSUBAME2.0は、地震自体の影響はほとんど無し
- だが、原発事故に伴う**電力危機**のため、運用を大幅に変更する必要が発生
- 1.0~1.5MWattを消費するTSUBAME2.0に対し、大学より節電要請あり

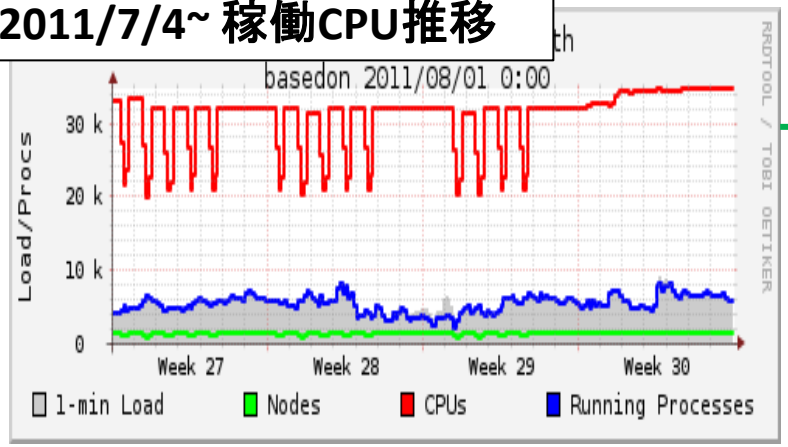
計画停電への対応

Webで電力見える化

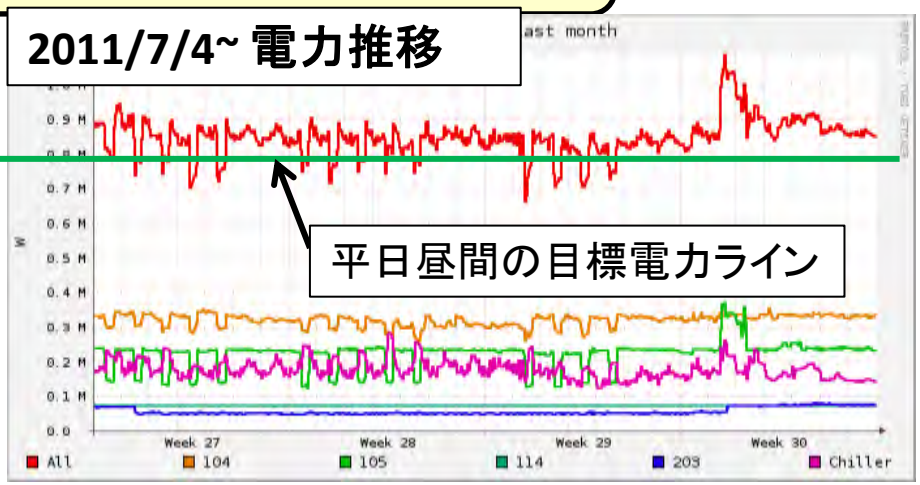
キュー構成変更

ピークシフト運用　(2011夏・2012夏)
平日昼間70%運用、夜間週末100%運用に自動切り替え

2011/7/4~ 稼働CPU推移



2011/7/4~ 電力推移



平日昼間の目標電力ライン

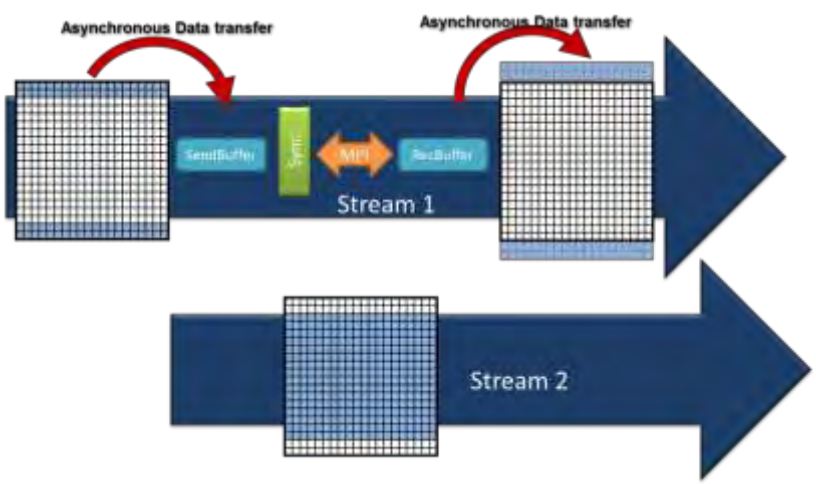
超低消費電力・数値計算アルゴリズム開発とHPCアプリケーション

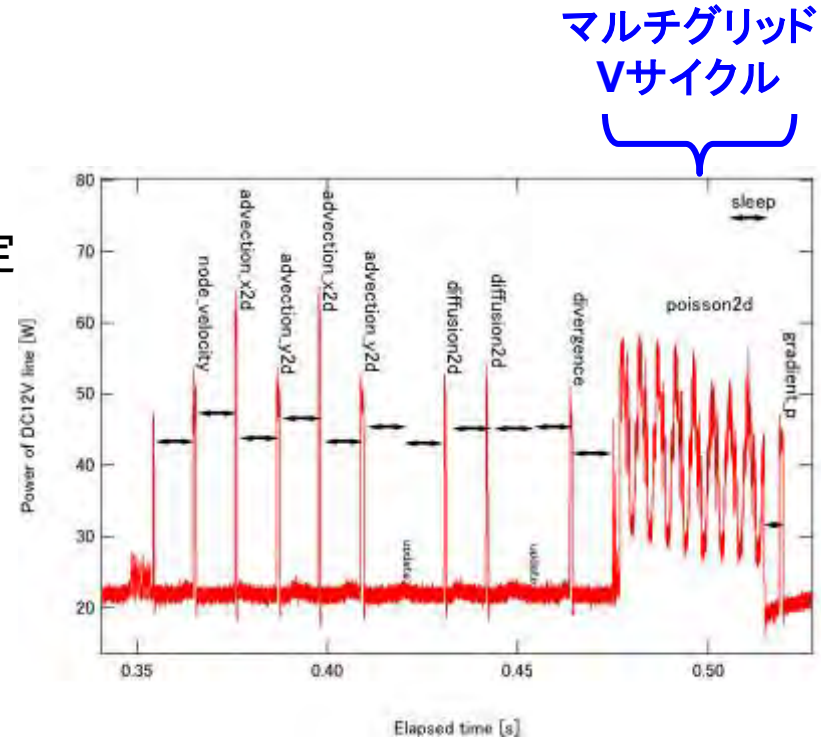
東工大　青木グループ

HPCアプリの関数毎の詳細電力測定

- GPUアプリのメモリ・アクセスの消費電力測定
- GPUアプリの浮動小数点演算の消費電力測定

低速メモリへのアクセス低減が有効





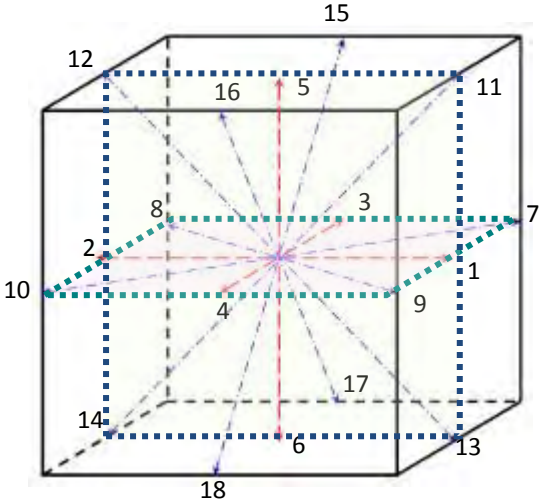
- GPUの Shared Memory を Software Managed Cache として利用
- 通信と計算のオーバーラップ

1m格子解像度による10km×10kmエリアの都市気流のLESシミュレーション

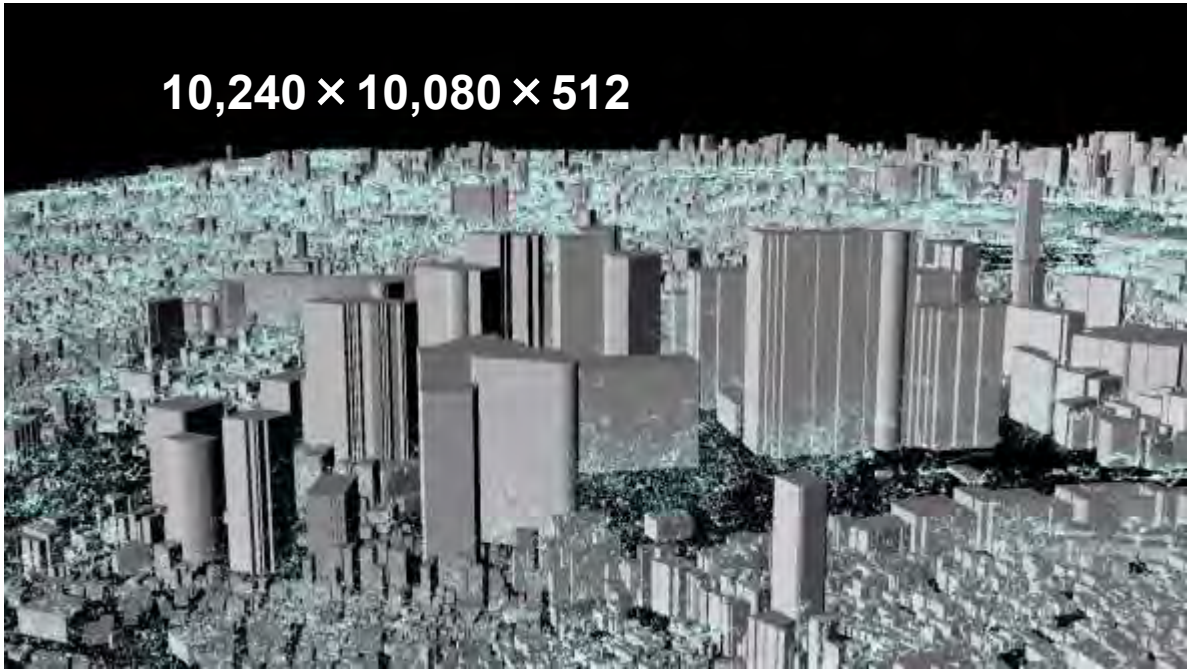
格子ボルツマン法： D3Q19 モデル

■ LES (Large-Eddy Simulation) モデル：
Coherent-Structure Smagorinsky model

■ メモリアクセス律速な計算にも関わらず 4000 GPU で 600 TFlops，高電力比性能 545 Mflops/W



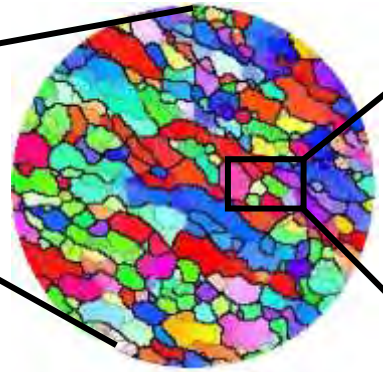
TBSテレビ, 朝日新聞，読売新聞, 日経新聞，PC Watch等で
2012年10月, 11月に報道

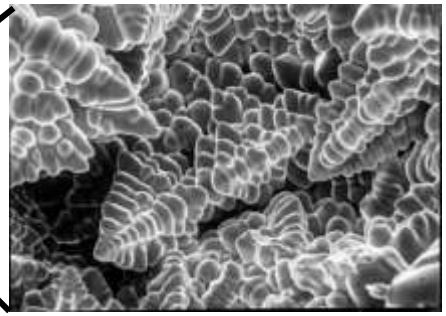


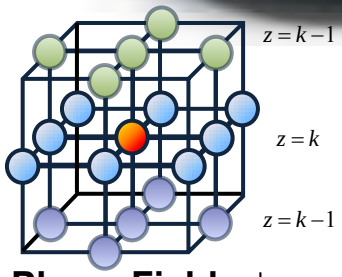
TSUBAME2.0による高電力比効率 合金材料の樹枝状凝固シミュレーション

新材料開発

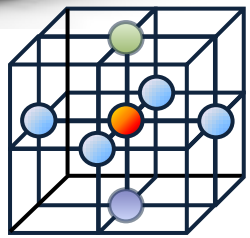
Material Microstructure





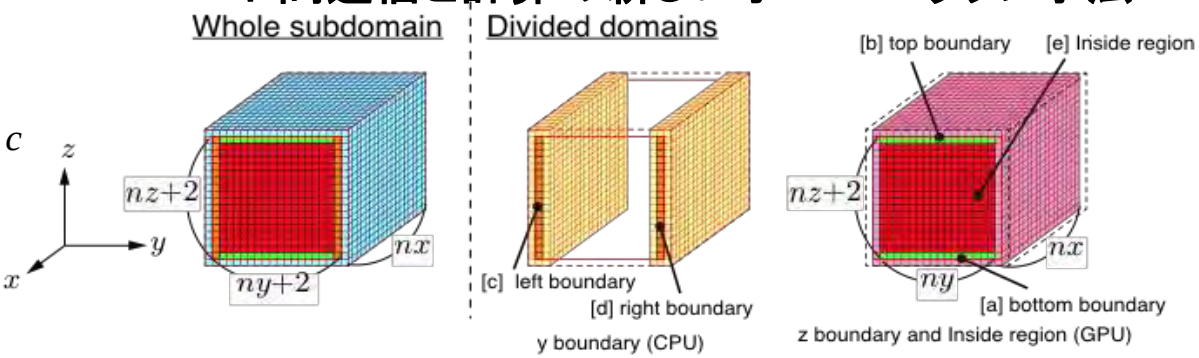


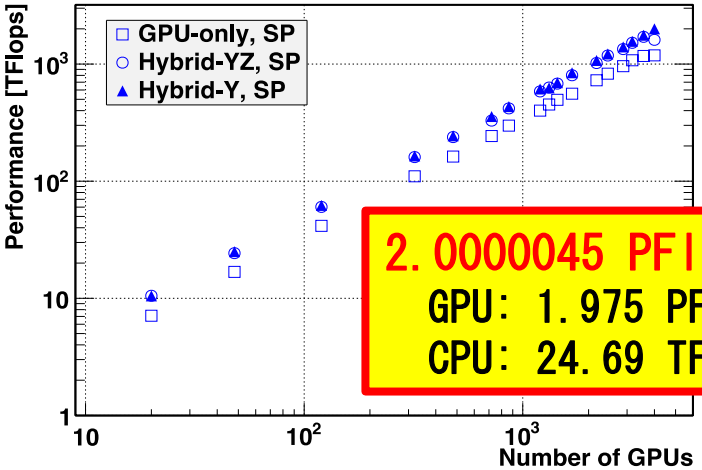
Phase Field : ϕ



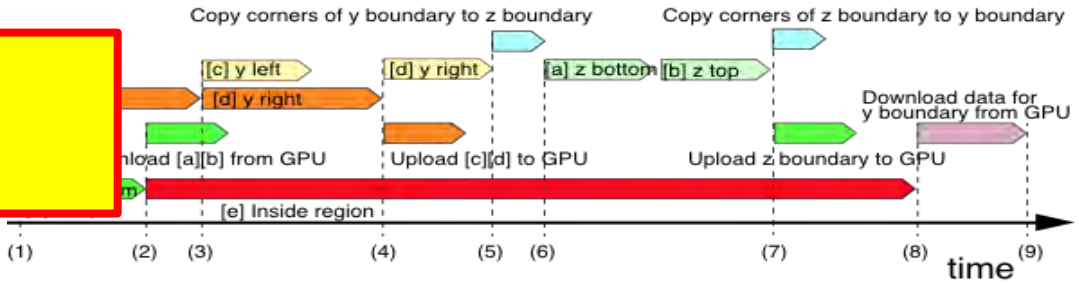
Condensation: c

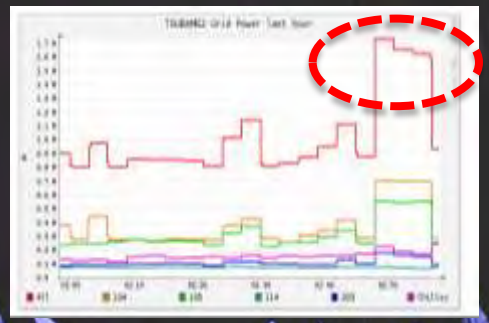
ノード間通信と計算の新しいオーバーラップ手法





2.0000045 PFlops
GPU: 1.975 PFlops
CPU: 24.69 TFlops





~1.36 MW

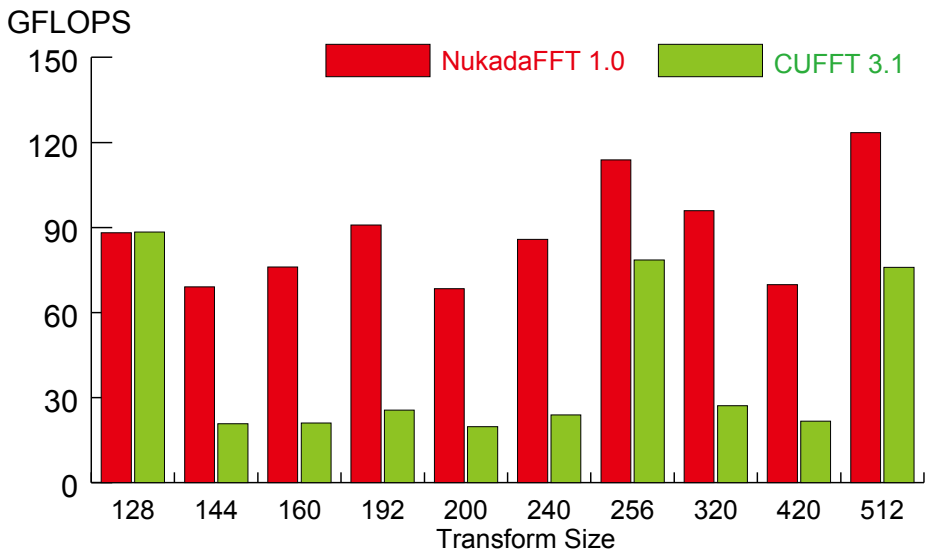
高電力比性能: 1468 MFlops/W

(c.f. Green500: 957 MFlops/W)

High Performance FFT on GPU

自動チューニング機能搭載 NukadaFFT library [SC08,SC09]

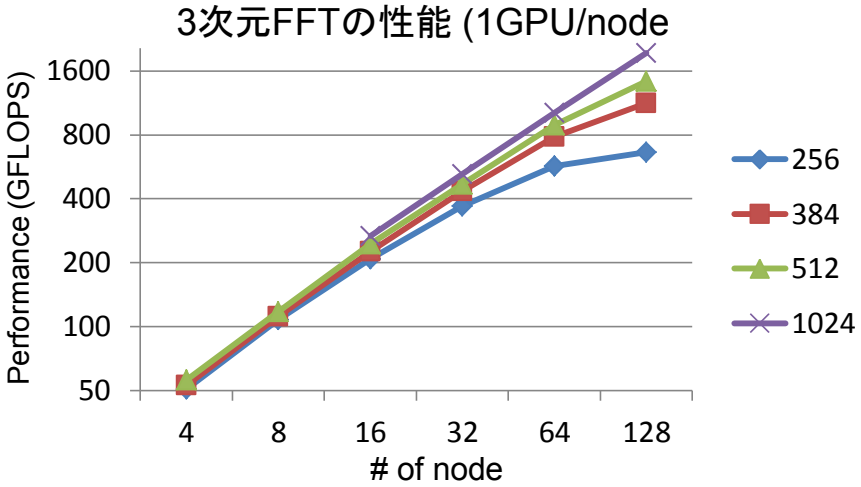
チューニング項目：(1) 基底の分解、(2) スレッド数、(3) バンクコンフリクト回避のためパディング自動挿入



Performance of 1-D FFT.
(Double Precision, batch=32,768, GPU=GeForce GTX 480.)
http://matsu-www.is.titech.ac.jp/~nukada/nufft/

NVIDIA純正CUFFTライブラリより高性能
AMD RADEON GPUでもOpenCL版で同程度の性能

TSUBAME2.0の複数GPUによるFFT計算 [SC12]



GPU間のAll-to-all通信⇒スケールしにくい

(1) Ibverbs APIを用いてオーバヘッド削減
(2) 複数の相手ノードと同時通信で競合発生時のペナルティ軽減
(3) 複数のリンクに適切に割り当てることによって競合発生率を低減

最大256ノード(768GPU)で4.8TFLOPS!

TSUBAME2.0 電力消費 (Jaguar比　4-5倍)

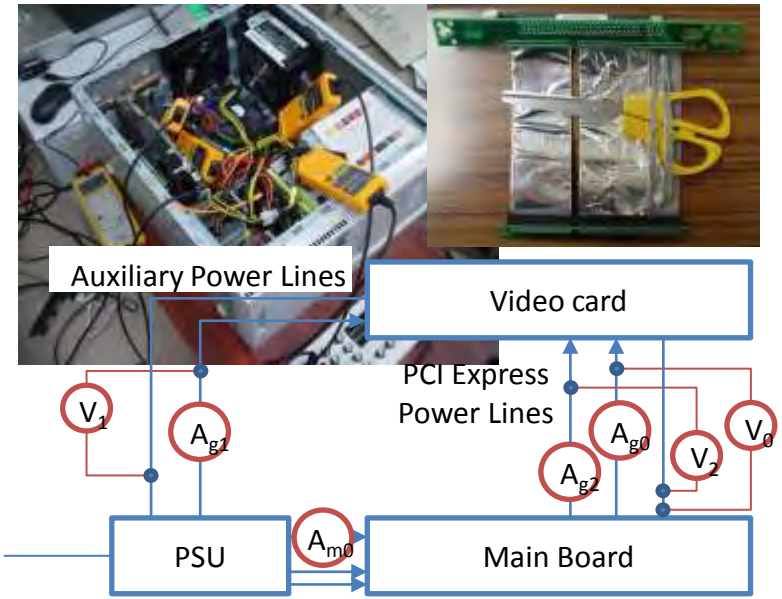
	Compute nodes & Interconnect (kW)	Storage (kW)	Cooling (kW)	Total (kW)
アイドル時	530	70	200	800
平均的運用時	680	0	230	980
Graph500 (CPU)	902	5	346	1323
Earthquake sim. (700nodes)	550/903	2	345	1320
ASUCA Weather	9			1308
NAMD (700nodes)	706,			1527
Turblence Sim.	1190	2	240	1502
Phase-field	1362	3	294	1729
Linpack	1417		-	-
GPU DGEMM	1538	72	410	2020

利用方法によって消費電力に大きな開き

制御の必要性

GPU の消費電力測定

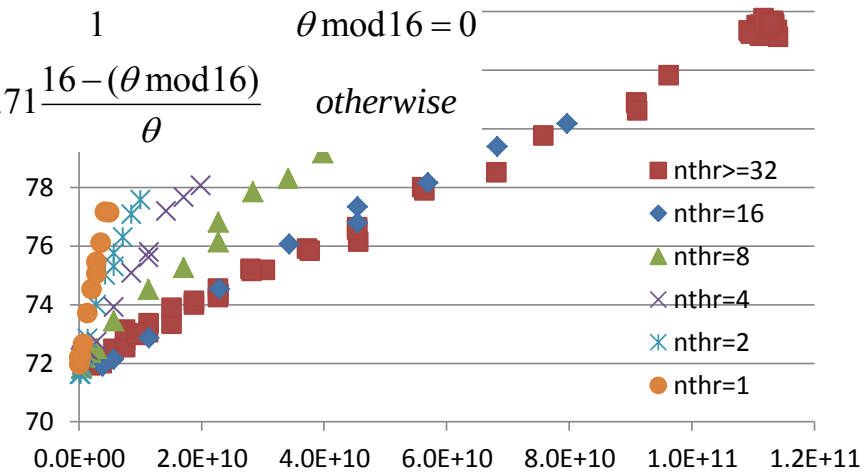
GPU 電力の測定装置



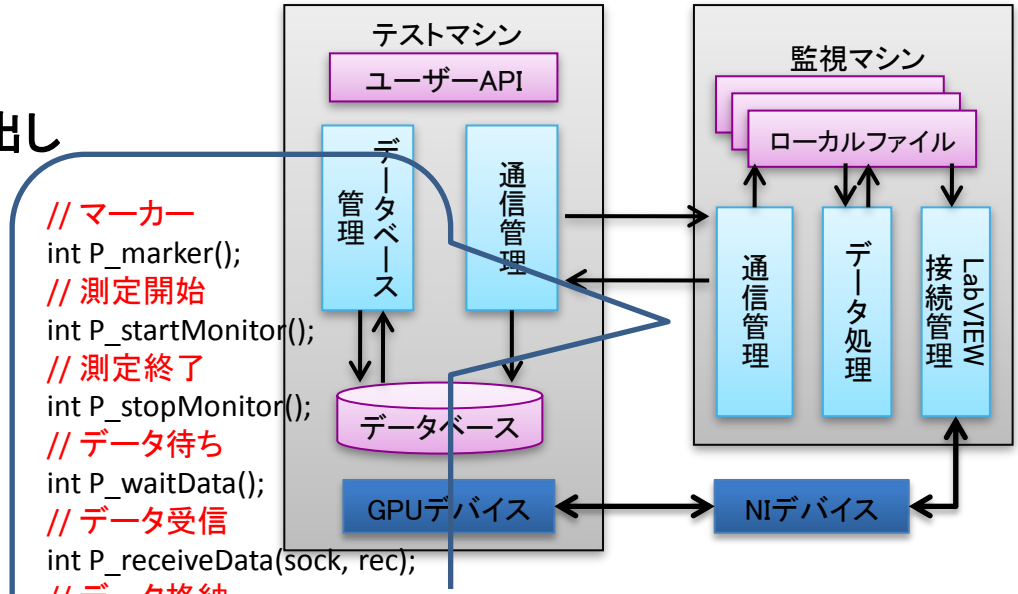
CUDA カーネルの精密な電力効率測定

$$w \approx 72.0 + 1.02 \times 10^{-10} \rho f$$

$$\rho = \begin{cases} 1 & \theta \bmod 16 = 0 \\ 1 + 0.71 \dfrac{16 - (\theta \bmod 16)}{\theta} & otherwise \end{cases}$$



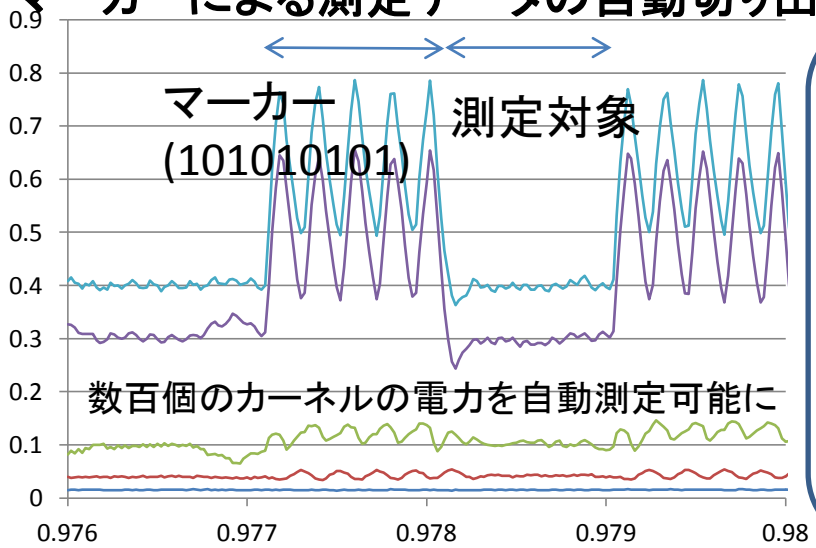
電力測定 API 策定・ライブラリ実装




```
// マーカー
int P_marker();
// 測定開始
int P_startMonitor();
// 測定終了
int P_stopMonitor();
// データ待ち
int P_waitData();
// データ受信
int P_receiveData(sock, rec);
// データ格納
int P_storeData(rec);
```

左上図の PC，本日のデモシステムで動作
ppOpen-AT 電力拡張に組み込み

マーカーによる測定データの自動切り出し



統計的GPU電力モデリング [IEEE IGCC10]

GPUパフォーマンスカウンタ

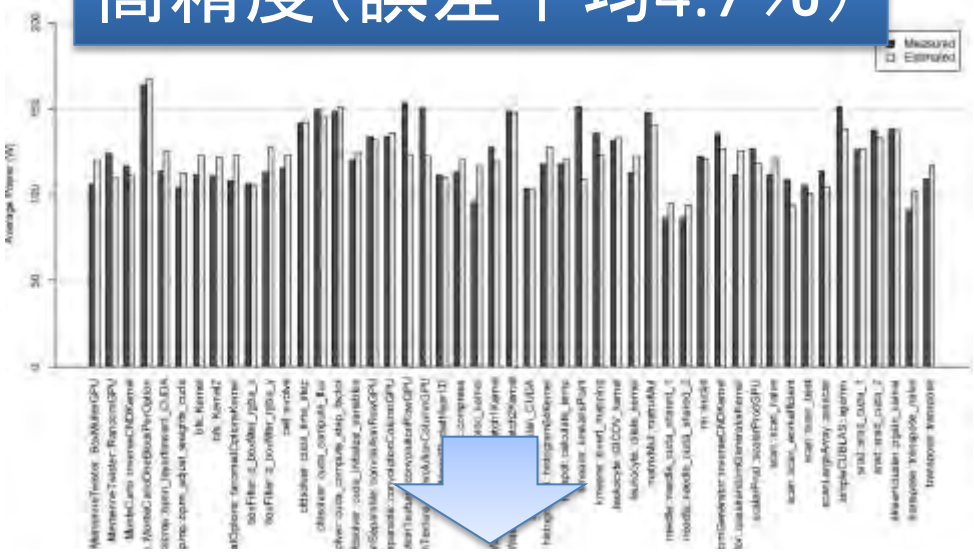
$$p = \sum_{i=1}^{n} \alpha_i c_i + \beta$$

平均消費電力

- リッジリグレッションによる過学習の防止
- クロスフィッティングによる最適パラメータの決定

- GPUの消費電力を統計的に推定
- 性能プロファイル（パフォーマンスカウンタ）を説明変数とした線形回帰モデル

高精度（誤差平均4.7%）



DVFSを導入しても高精度を確認

今後：電力モデルによる電力最適化

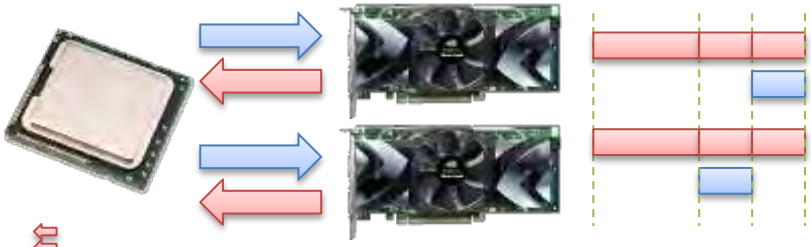
線型モデルでも十分な精度

↓

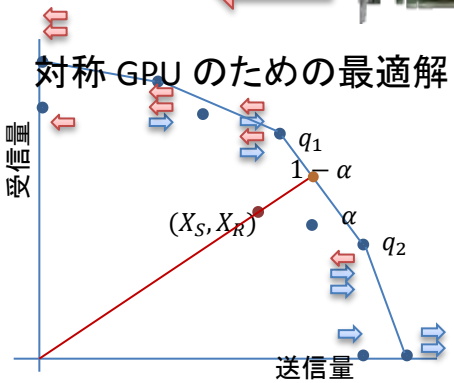
億単位のプロセッサからなる**エクサスケール**における最適化の実現可能性

GPU 高性能化手法

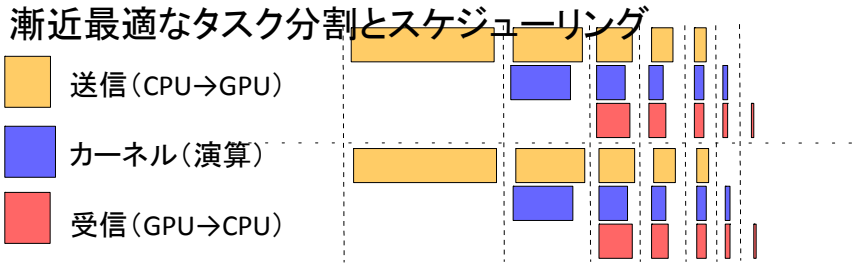
CPU-GPU 複数データ転送の最適解



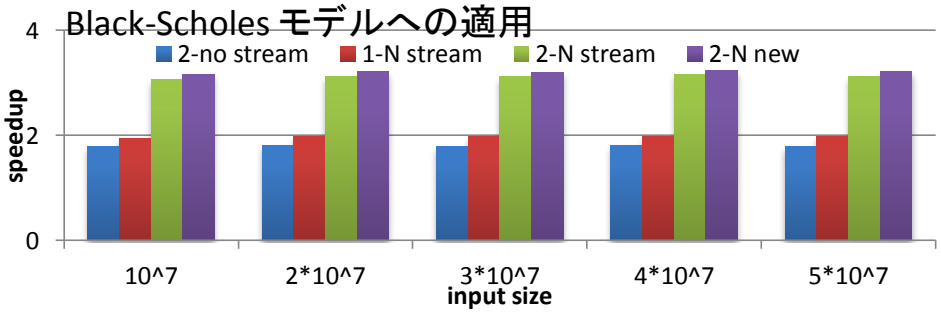
対称 GPU のための最適解：スループット凸包



漸近最適なタスク分割とスケジューリング



Black-Scholes モデルへの適用



GPU の実行時間・エネルギー消費推定モデル

CUDA コードを Ocelot により分析

↓

メモリ並列度 (MPD)，計算並列度 (CPD)

↓

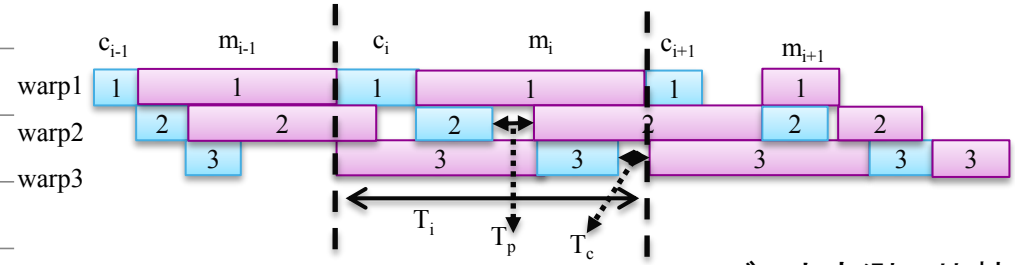
実行時間・エネルギー推定モデル

実行時間推定モデル

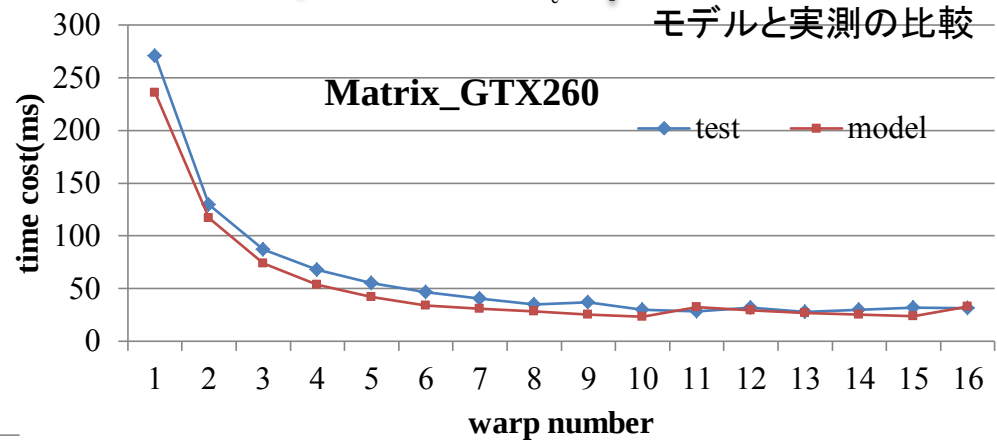
$$T_i = \lceil (N_{act} * c_i)/CPD \rceil + T_p + T_c$$

$$T_p = \max\{m_{i-1} * \lfloor (N_{act}/MPD) - 1 \rfloor - (N_{act} - 1) * c_i, 0\}$$

$$T_c = \max\{m_i - ((N_{act} - 1) * c_i + T_p), 0\}$$



モデルと実測の比較



推定精度(%)	Matrix	Linear filter	Black-scholes
GTX260 time	90.7	91	88.4
C2050 time	84.4	89.5	90.1
C2050 energy	85.8	91.9	88.9

OMPCUDA: GPU向けOpenMP処理系(電通大G)

CPU向け並列化プログラミング環境でGPUを利用可能に

アプリケーションプログラマが
記述したOpenMPプログラム

```
float gdata[N];
int main(){
  ......;
#pragma omp parallel for
{
  ...... = gdata[...];
  gdata[...] = ......;
  ......;
}
  ......;
  return 0;
}
```

プログラム変換機構

GPUコード

```
__device__ float gdata[N];
__global__ void
 _ompcuda_func_1(float *arg0, float *arg1){
  gdata = arg0;
  ...... = gdata[...];
  gdata[...] = ......;
  ......;
}
```

実行時ライブラリ呼び出し

CUDAコンパイラへ

CPUコード

```
float gdata[N];
int _ompcuda_main
  ......;
  cuMemcpyHtoD(..., &gdata[i], ...);
  cuParamSeti(...);
  cuLaunchGrid(_ompcuda_func_1, ...);
  cuMemcpyDtoH(&gdata[i], ...);
  ......;
  return 0;
}

int main(){
  _ompcuda_init();
  _ompcuda_main();
}
```

CUDA実行時ライブラリ
呼び出し

backendコンパイラへ

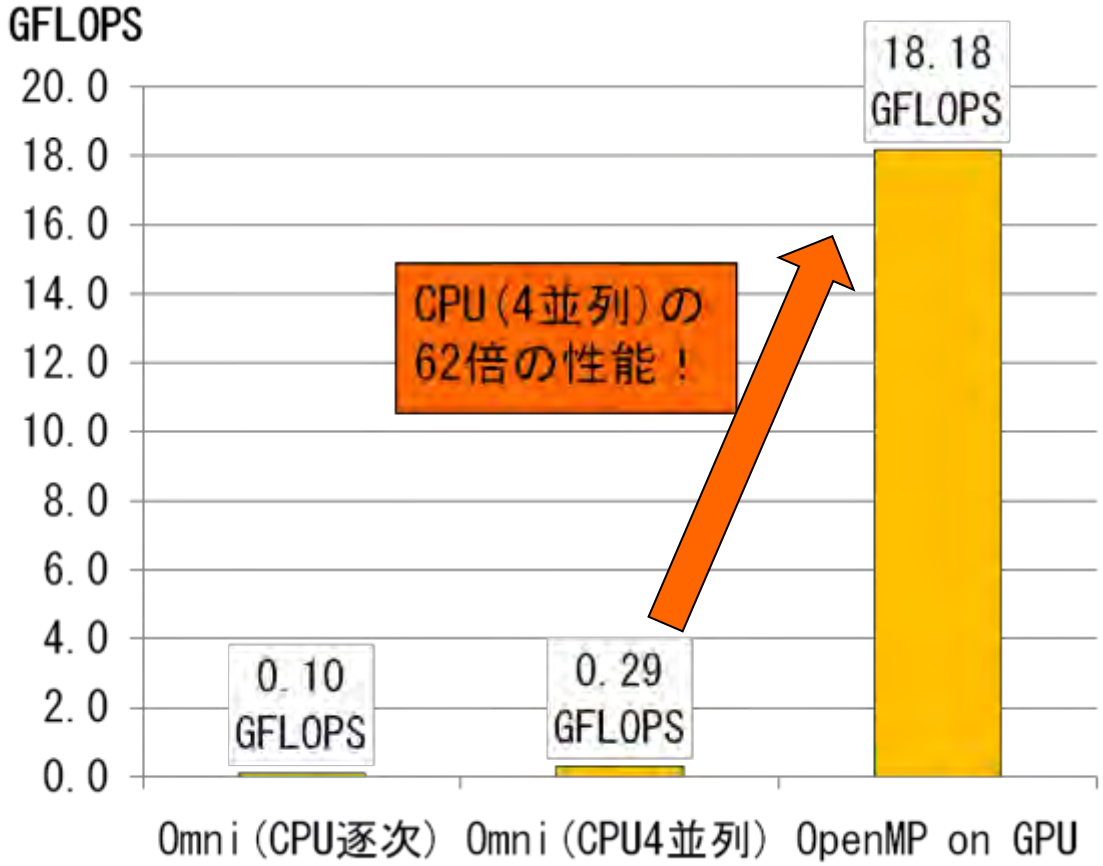
実行時ライブラリ呼び出し

- OpenMP for CUDA
  - CUDA対応GPU向けOpenMP処理系
  - GPU上にある多数の演算器を利用し
    高い並列度を活かす
  - #pragma omp parallel for および reduction に対応
  - GlobalMemoryと多数の演算器を用いた
    高い並列度の並列処理

性能評価
同一ソースコードによる実行性能比較

アプリケーションソースコード
(行列演算2重ループ)

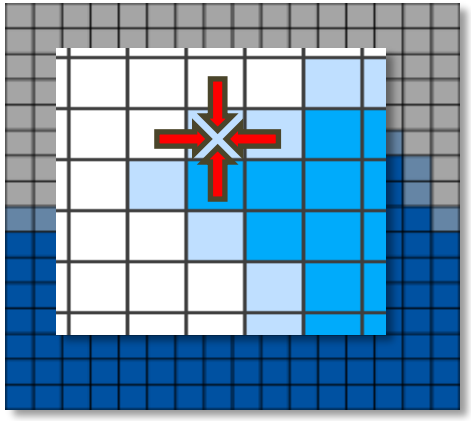
```
#define N 1024
float a[N*N], b[N*N], c[N*N];

#pragma omp parallel for
private(j)
  for(i=0; i<N*N; i++){
    float tmp = 0.0f;
    for(j=0; j<N; j++){
      tmp += a[(i/N)*N+j] *
                b[j*N+(i%N)];
    }
    c[i] = tmp;
  }
```



提案手法による実行は行列積において同一ソースコードを用いて
CPU(逐次実行)の**176倍**、CPU(4並列実行)の**62倍**の性能を達成

CUDAによる共有メモリを用いない実装では 22.2GFLOPS、
共有メモリを用いる実装では 83.4GFLOPS を達成可能

並列GPUコード自動生成による ステンシル計算のスケーラビリティ [SIAM PP 11]



計算内容は簡潔でも,
- 領域分割
- MPI+CUDA通信
- 通信オーバラップ

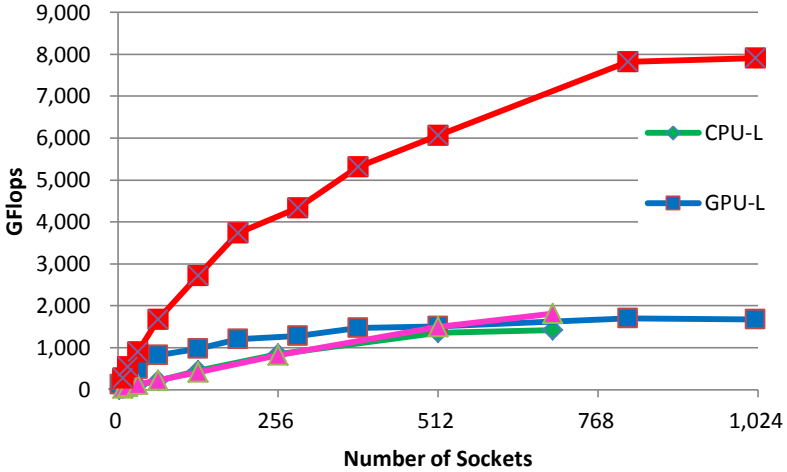
のためコードは煩雑
→ 自動コード生成システムを構築中

ユーザは一点の計算を表現する関数を記述

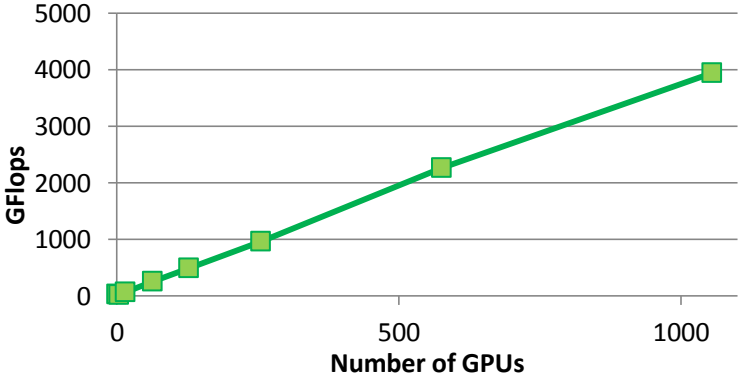
```
__stencil__ void average(int x, int y, int z, grid3d_real
g) {
    if (x == 0 || y == 0 || z == 0) return;
    float ret = psGridGet(g,0,0,0)
       + psGridGet(g,-1,0,0) + psGridGet(g,1,0,0)
       + psGridGet(g,0,-1,0) + psGridGet(g,0,1,0)
       + psGridGet(g,0,0,-1) + psGridGet(g,0,0,1);
   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

GPUクラスタ用高性能並列コードを自動生成

TSUBAME2.0 1024GPUで**姫野ベンチ8TF; 世界記録達成！**

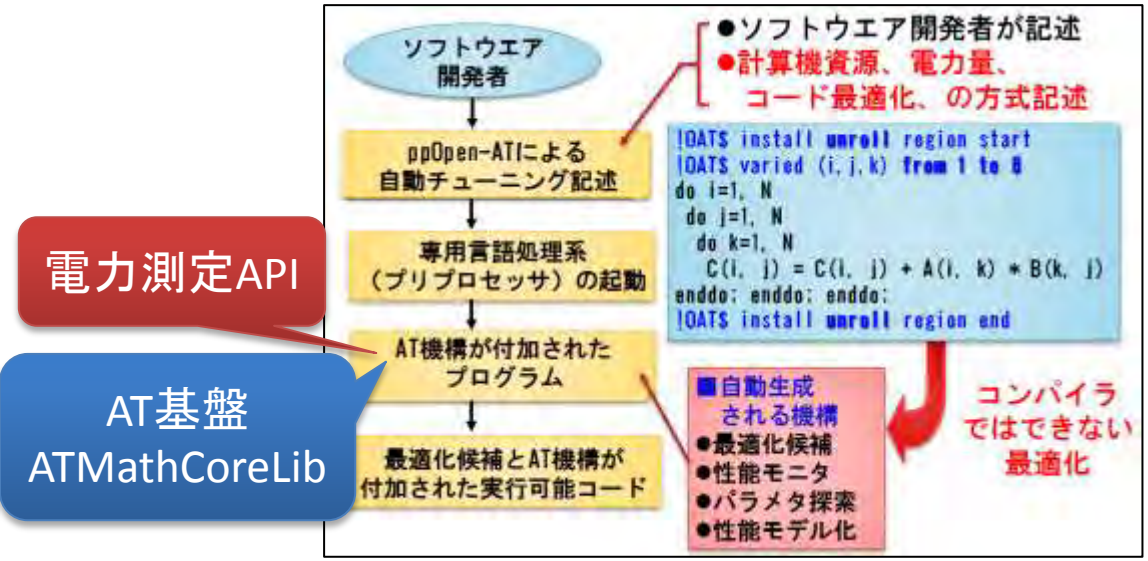


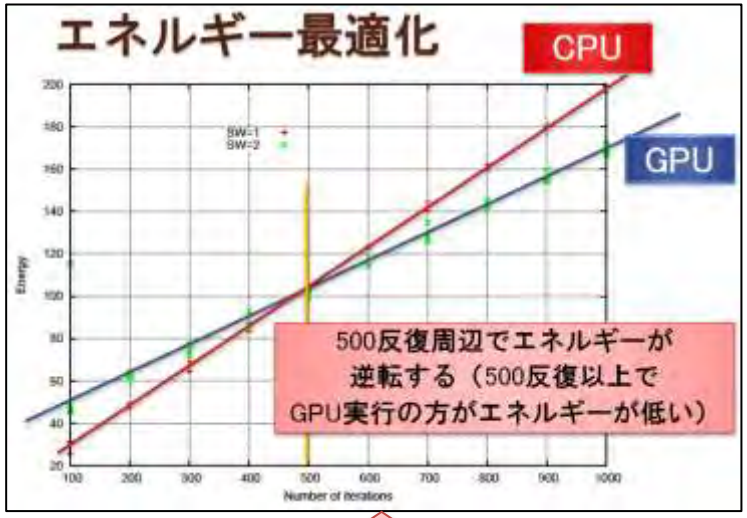
東大地震コードの弱スケーラビリティ



ppOpen-ATの省電力拡張

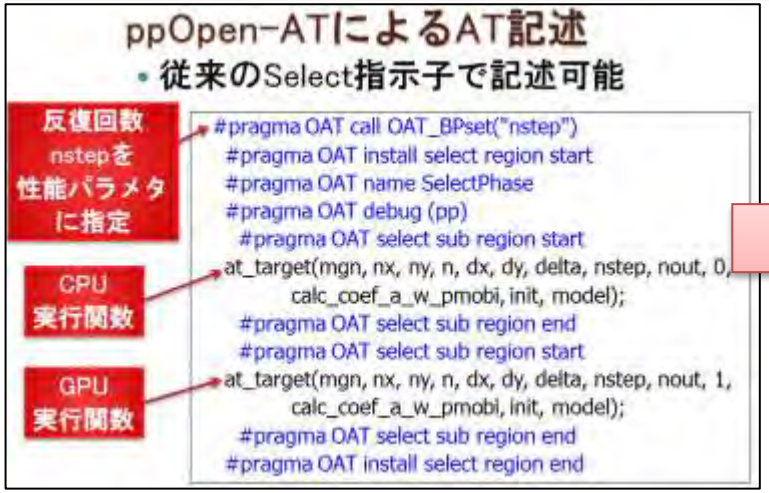
- AT言語ppOpen-ATに省電力化AT機能を実装。任意箇所のエネルギーを最適化可能。





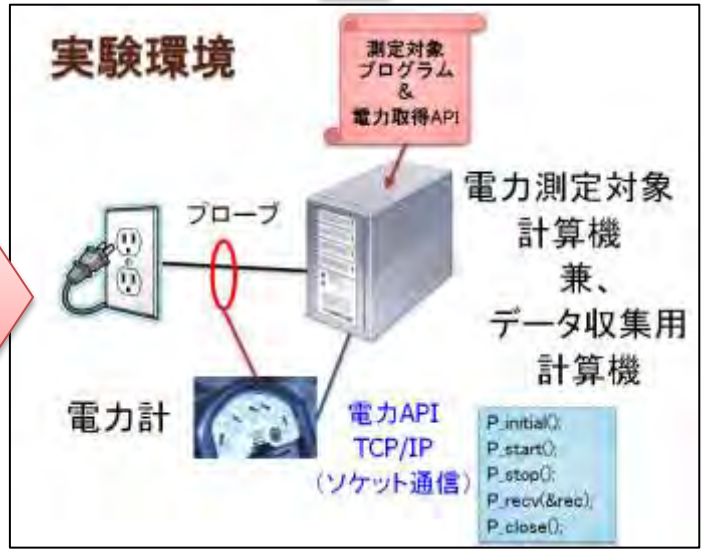
例）フェーズフィールドの反復回数ごとのCPU、GPU切り替えによるエネルギー最適化

自動最適化



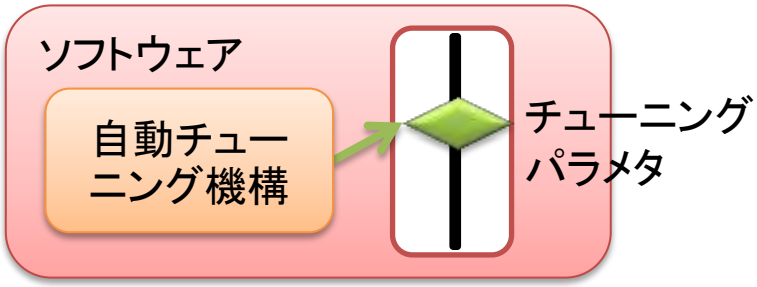
自動生成

エネルギー最適化のAT機能を含むコード



自動チューニングのための数理手法の研究と自動チューニング数理ライブラリ ATMathCoreLib

自動チューニング
ソフトウェアに組み込まれた可変性を自動的に調整して，できるだけ高性能・省電力を実現する仕組み



提案手法
- オンライン自動チューニング
- ワンステップ近似
- ワンステップ近似の高速化
  - ランダムサブセット法
  - 確率的候補選択法
- 並列処理の自動チューニング
  - 並列実験
  - 並列試行
- AT メーター
- 4DAC

ライブラリ ATMathCoreLib として公開中

◆ **オンライン自動チューニング**

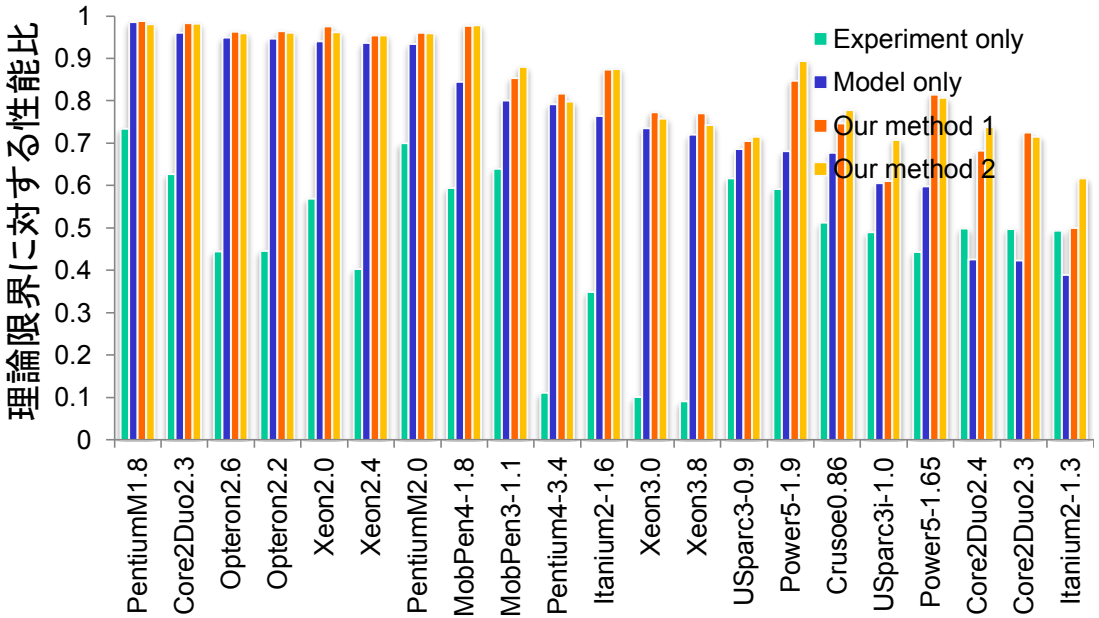
$$w_i^{(k)} = \mu_i^{(k)} + (k-1)E\left(\min\left\{\mu_i^{(k-1)}, \mu_{\min'}^{(k-1)}\right\}\right)$$

◆ **オフライン自動チューニング**

$$w_t = \mu_t + \kappa E_y\left(\min\left\{\mu_t^{post}(y), \mu_{\min'}\right\}\right)$$

◆ **並列自動チューニング**

$$W_n = \max\{t_{P+1},\ldots,t_{P+n}\} + (K-2)\min\{t_{P+1},\ldots,t_{P+n},t_{opt}\}$$



行列積の外・中アンローリング：n = 1 から 256/512 の平均

国内外の自動チューニング研究をリード

iWAPT (international Workshop on Automatic Performance Tuning) を主催

	開催地		開催地
iWAPT 2006	東京	iWAPT 2010	Berkeley
iWAPT 2007	東京	iWAPT 2011	Singapore
iWAPT 2008	筑波	iWAPT 2012	神戸
iWAPT 2009	東京	iWAPT 2013	Barcelona

編集した書籍，学会誌特集号

- 「**情報処理**」2009 年 6 月
- Springer "**Software Automatic Tuning**" 2010 年 9 月
- 「**応用数理**」2010 年 9 月・12 月

様々な学会でセッション実施

- 情報処理学会 HPC 研究会
- 日本応用数理学会 年会
- 日本計算工学会 講演会
- SIAM Parallel Processing (PP)
- SIAM Comp. Sci. and Eng. (CSE)

招待講演

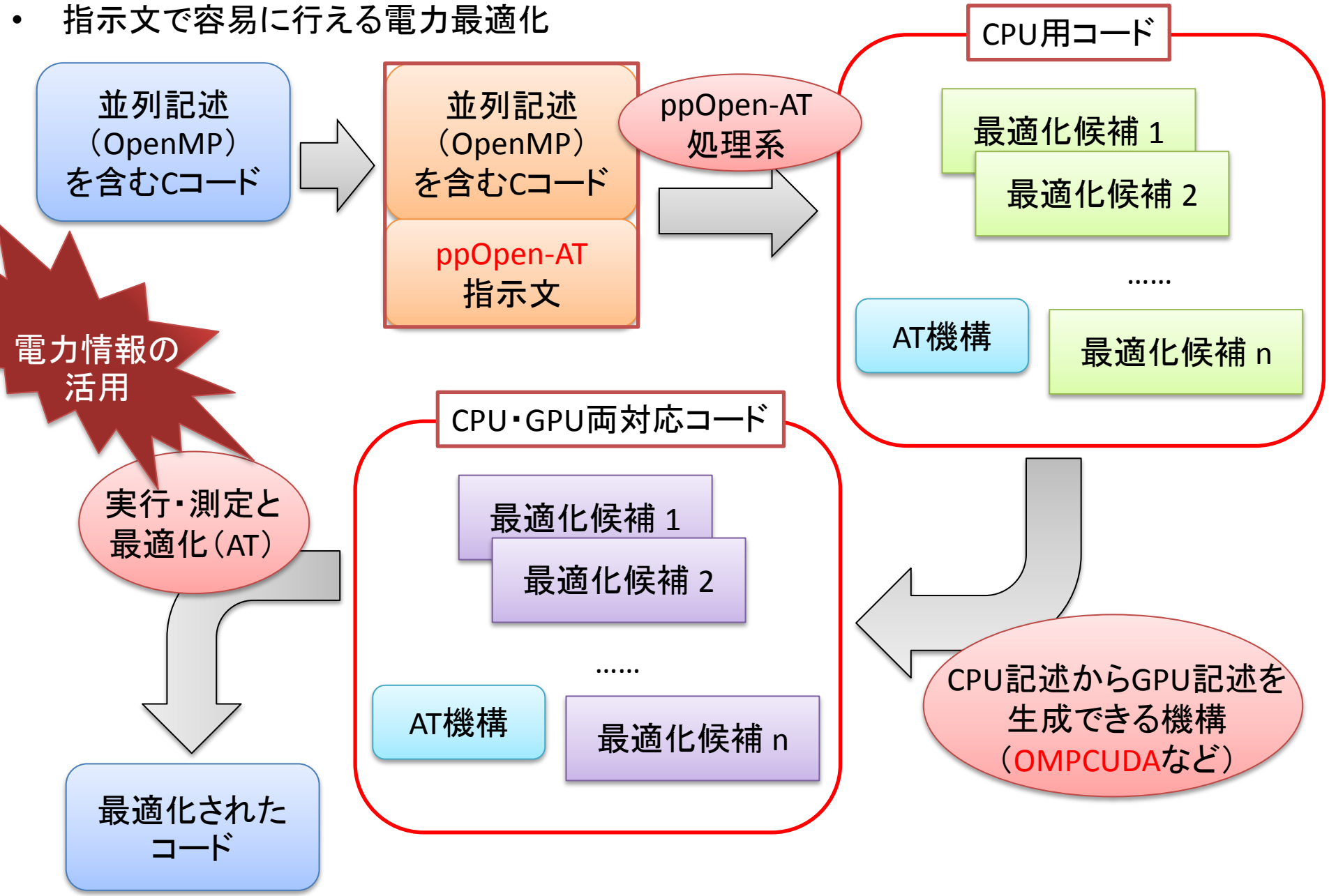
- 2009 ITBL シンポ（須田）
- 2009 数値解析シンポ（片桐）
- 2009 Workshop on Libraries and Autotuning for Petascale Applications（片桐）
- 2011 HLPP 2011（片桐）

その他関連するもの

- 科学技術分野の文部科学大臣表彰 若手科学者賞（片桐）
- SIAM News 2008/06 号で紹介
- 文科省 科学技術政策研究所「科学技術動向」2009/11 紹介

ppOpen-ATとOMPCUDAの連携

- 指示文で容易に行える電力最適化



省電力インターコネクト
（イーサネット、InfiniBand)

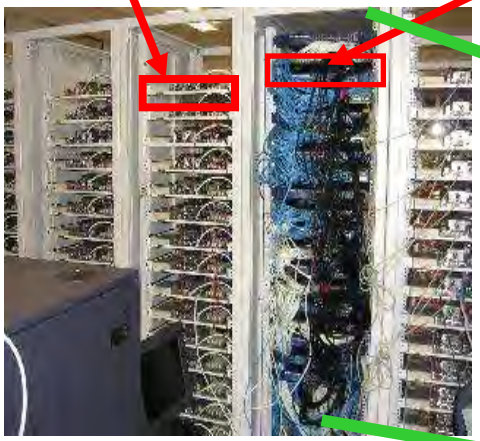
同志社大学SuperNovaクラスタ
（225ホスト）

Before

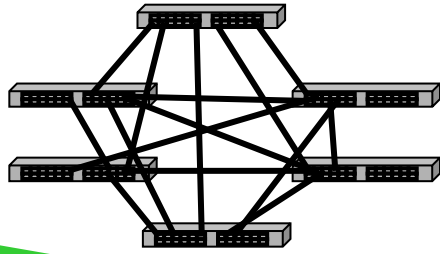
巨大なモンスタースイッチ
（312ポート@1）

PC

スイッチ



After(ULP-HPC)



小規模スイッチ
（48ポート@8)

導入例（2009時点）

<u>消費電力</u>：小規模スイッチ（～87%減）×マルチパストポロジ(6割増）×On/Off　リンク制御（～37%減）≒8分の1

<u>導入コスト</u>：小規模スイッチ（～98%減) ×マルチパストポロジ(6割増）×On/Off　リンク制御　≒25分の1

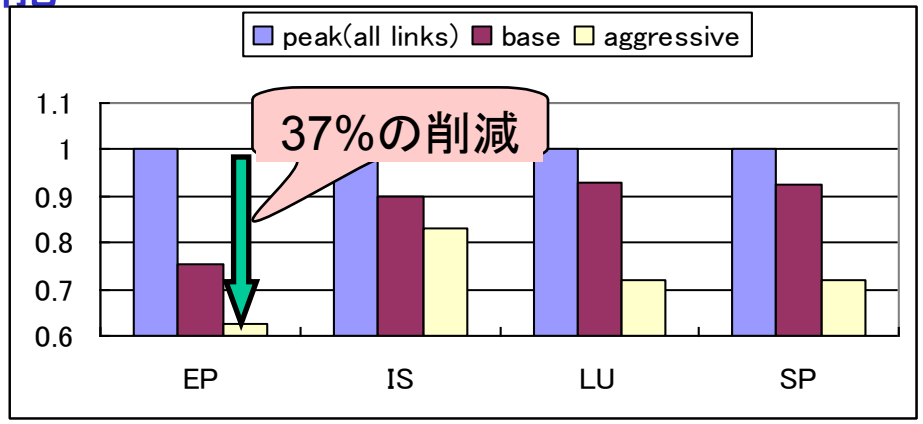
On/Off リンクレギュレーション

マルチパスイーサネット、InfiniBand において
On/Off＆マルチスピードリンクレギュレーション法を実現

- トラフィック負荷によらず、スイッチの消費電力はほぼ一定
- スイッチ消費電力の中でポートの消費電力が支配的
  - 並列アプリケーションの通信パターンの解析結果より、利用率の低いリンクをオフ/減速することで電力減
  - 外部操作で電力削減が可能

表： スイッチ単体の消費電力の内訳(単位 W)

スイッチの型番	All except ports	GbE port	Peak（ポートの割合）
SFS7k(IB)	21.4	0.9	43(50%)
PC62(GbE)	56.8	2.1	155 (63%)
SF-420(GbE)	32.6	1.0	55.4(41%)



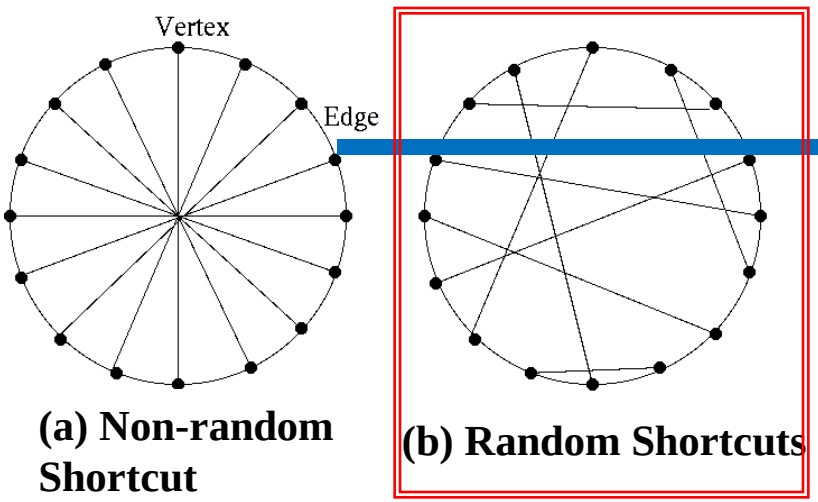
NWの電力削減効果
（NAS並列ベンチマーク, Class C）

ランダムトポロジを利用して スケーラブルな省電力インターコネクトへ

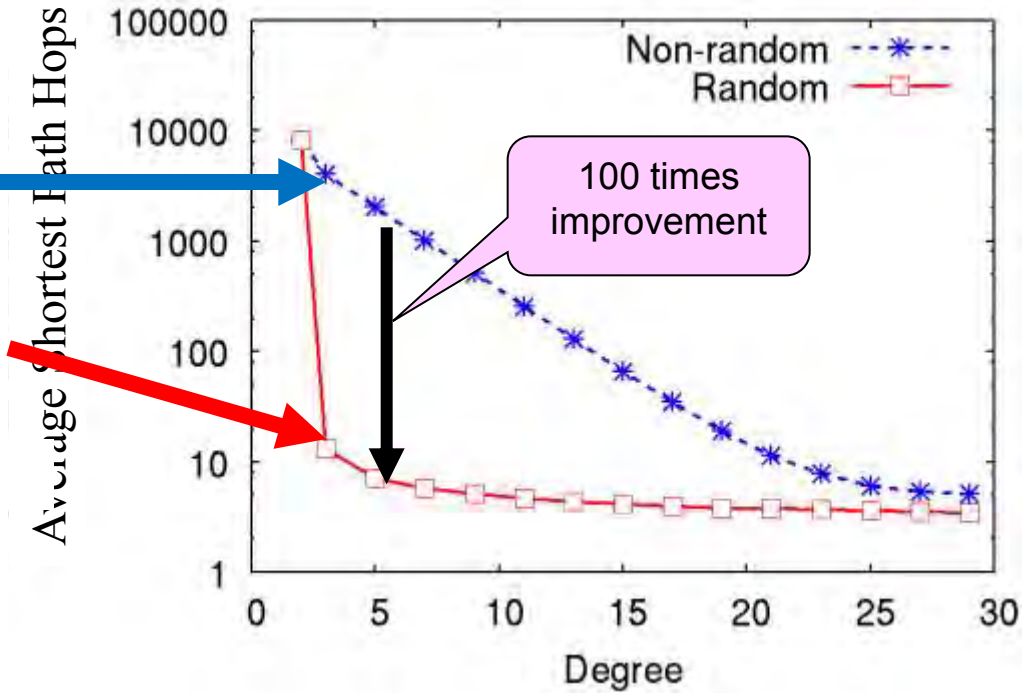
エクサ級スパコンではインターコネクトの低遅延化（システム最長1us)と省電力化が必要

- トポロジの課題：最長，平均ホップ数の激減必→ランダムなトポロジの利用
- 省電力の課題：スケーラブルなリンクの動的On/Off制御

開発並列アプリの通信パターン
↓
On/Offリンク省電力制御
↓
ネットワーク再構成



Ring + Shortcuts



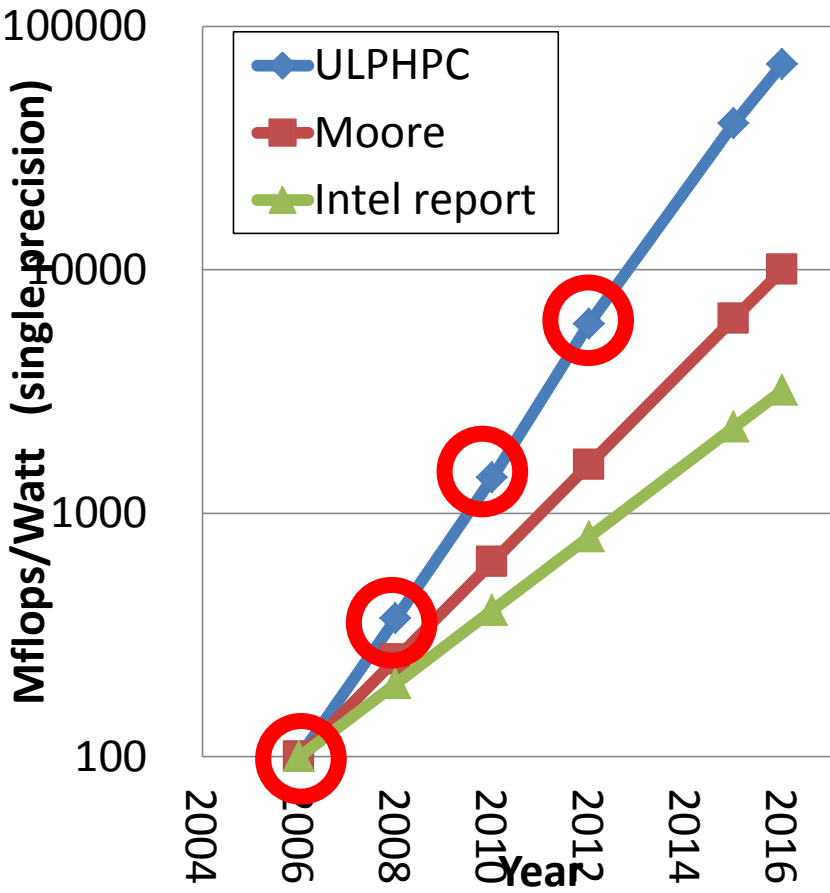
32,768ノードにおけるホップ数

11月最終報告会デモについて

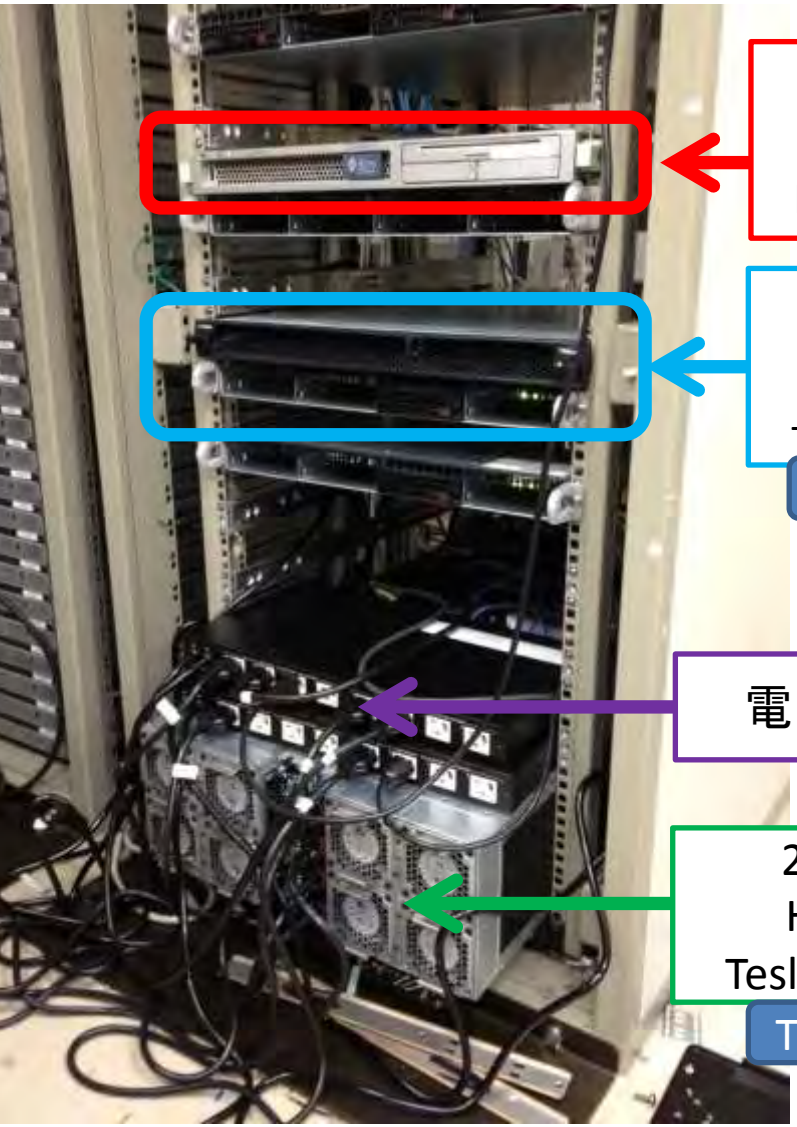
- 本チームの数値目標がfeasibleであり，
- グループ間の技術統合により可能である

ことを示す

- 世代の異なる複数の計算機を並べ，アプリケーションを実行
  - 各電力性能比が，10年約1000倍ラインに乗っていることを示す



計測用システム (デモ展示)



TSUBAME 1.0同等

2006年マシン
SunFire X4100
Dual-Core Opteron

2008年マシン
Tesla S1070(GT200)
+ Supermicroサーバ

TSUBAME 1.2同等

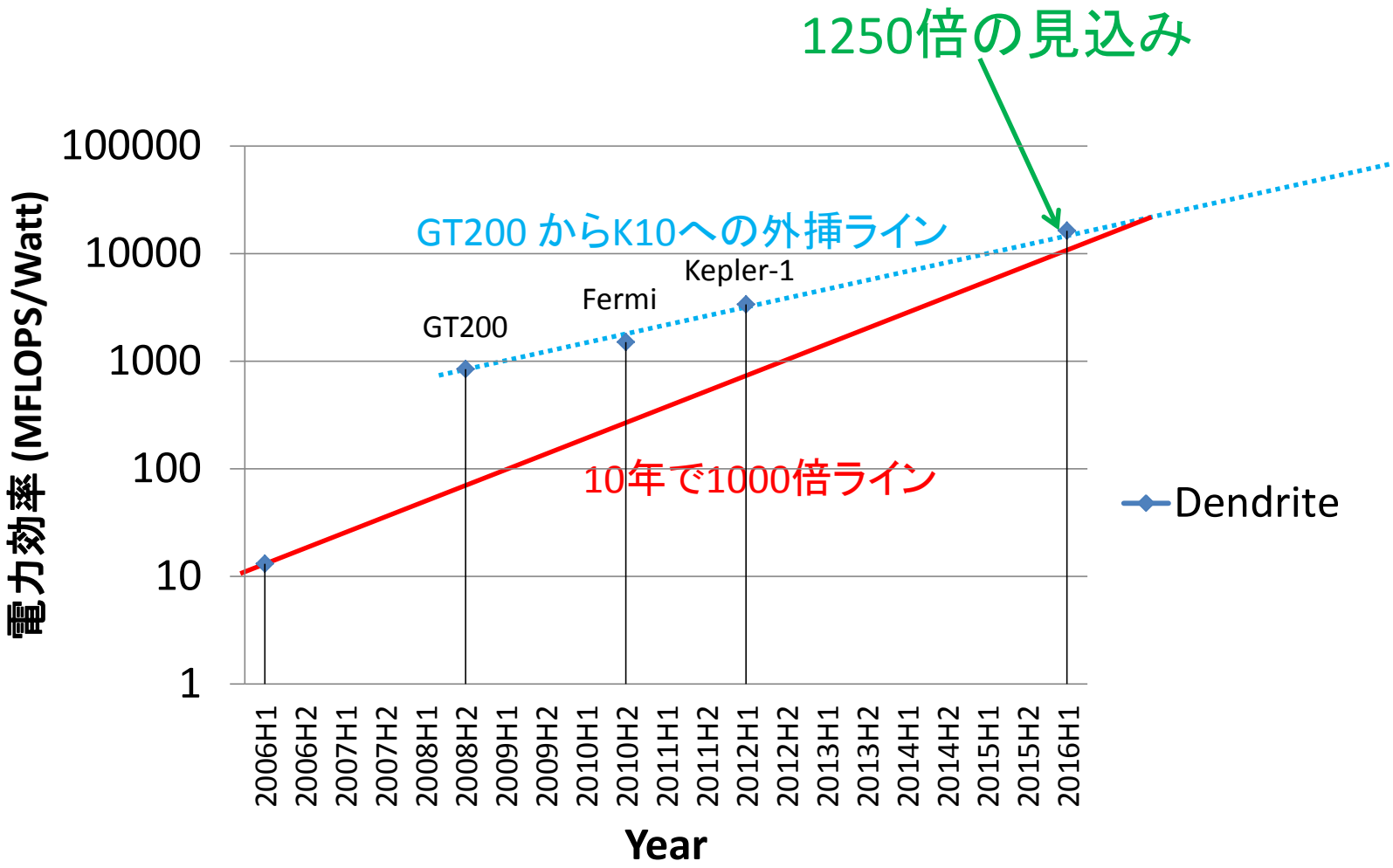
電力モニタ装置

2010年マシン
HP SL390s G7
Tesla M2050(Fermi)

TSUBAME 2.0同等

2012年マシン
Tesla K10(Kepler) x4
油浸冷却

凝固成長シミュレーションにおける各世代マシンでの電力効率の推移



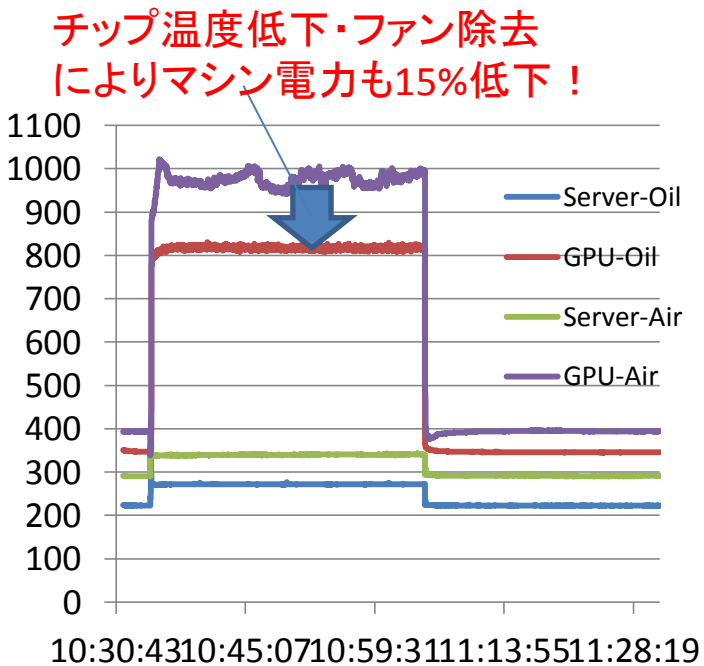
GT200世代でのGPU化によって60倍以上の電力効率向上。

先進的冷却技術の適用による HPCの更なる省電力化

- 計算システムの省電力化だけでなく冷却システムの省電力化が必須
- 油浸冷却技術を導入・評価
  - 東工大GSIC概算要求「スパコン・クラウド情報基盤におけるウルトラグリーン化技術の研究推進」と共同
  - PUE<1.05目標

米国Green Revolution Cooling社製の油槽

自作油槽に4 Kepler GPUマシンを液浸した様子



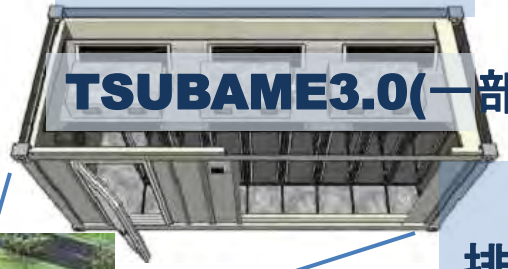
概算要求: スパコン・クラウド情報基盤における ウルトラグリーン化技術の研究推進

(3) JST ULPHPC 基礎研究の適用によるスパコン用超省電力ミドルウェア・アプリ等

(2) TSUBAME2.0 (2010) ~2.4PFlops
世界一の性能および性能電力比
ラップトップの数倍

コンテナ型スパコンセンター(概念図)

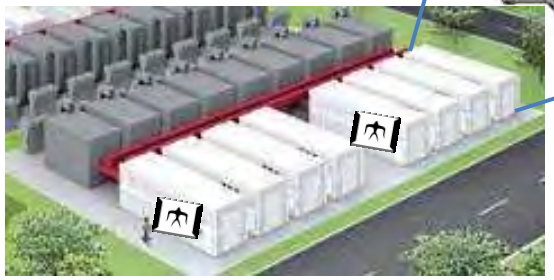
高密度コンテナ



TSUBAME3.0(一部)

冷却用CO_2排出の大幅削減

Nvidia Tesla K20 (2013年度前半)追加
TSUBAME2.5 単精度 ~10PF
消費電力の削減



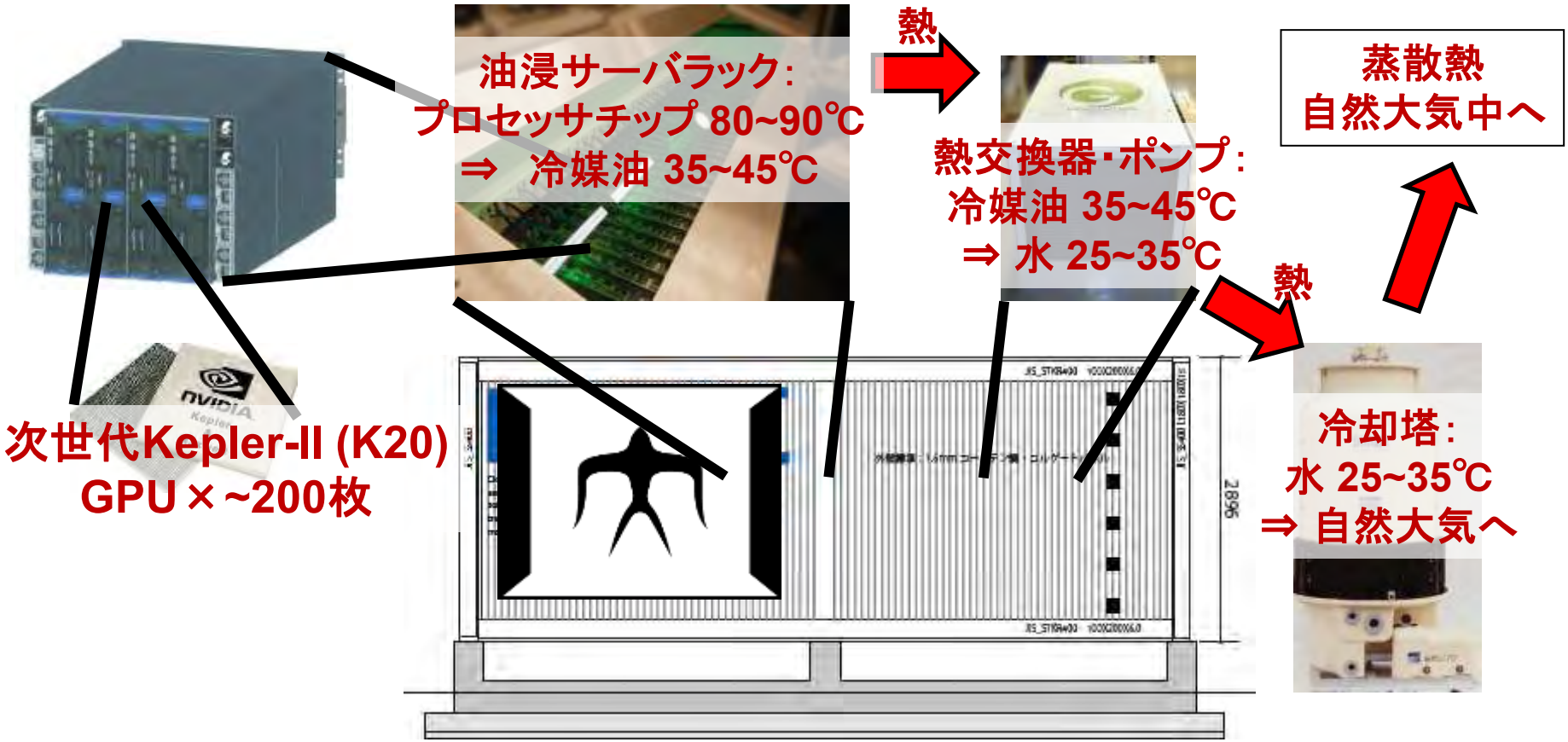
無人・リモート自律運用技術

-最終成果-
2015年 TSUBAME3.0 倍精度25~30PFlops---TSUBAME1.0比で

(1)先進冷却PUE < 1.1
年2/3以上は冷却電気ほぼゼロ

2013/2 TSUBAME-KFC実験設備

計算機電力70~100kW規模の、コンテナ型研究設備
油浸冷却＋大気冷却＋高密度スパコンの統合

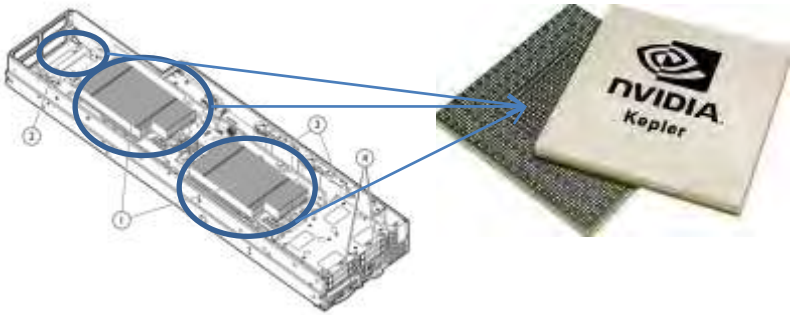


目標:

- 現在の世界トップ(2012年BG/Q , 2.1GFlops/Watt)以上の電力性能比
- 年間のほとんどで、事実上PUE<1.0

TSUBAME2.0から2.5への中間アップグレード (2013年7月)

- TSUBAME2.0のGPU(Fermi 2050)を全部または部分的に最新のアクセラレータに交換

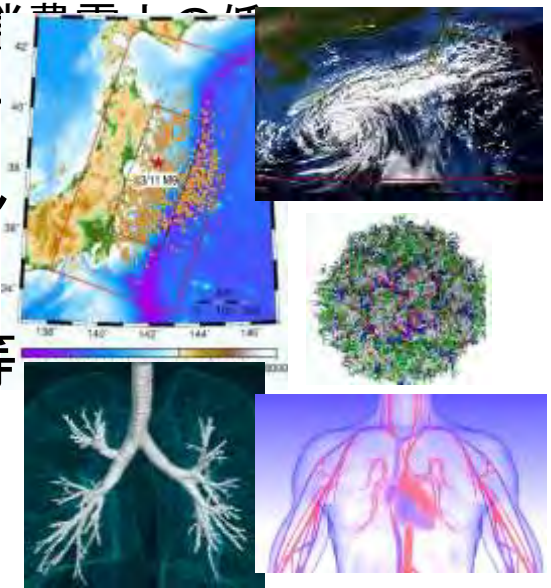


TSUBAME2.0計算ノード
Fermi GPU 3機 **x 1408**台 **= 4224**機

性能向上:
4.8PF/2.4PF→12PF以上/4PF以上へ
(単位:単精度/倍精度ペタフロップス**)**
c.f. 京 **11.2/11.2, T2.5**が「単精度日本一」
実アプリの向上は更に見込める
信頼性の向上・消費電力の低
時期:**2013**年**7**

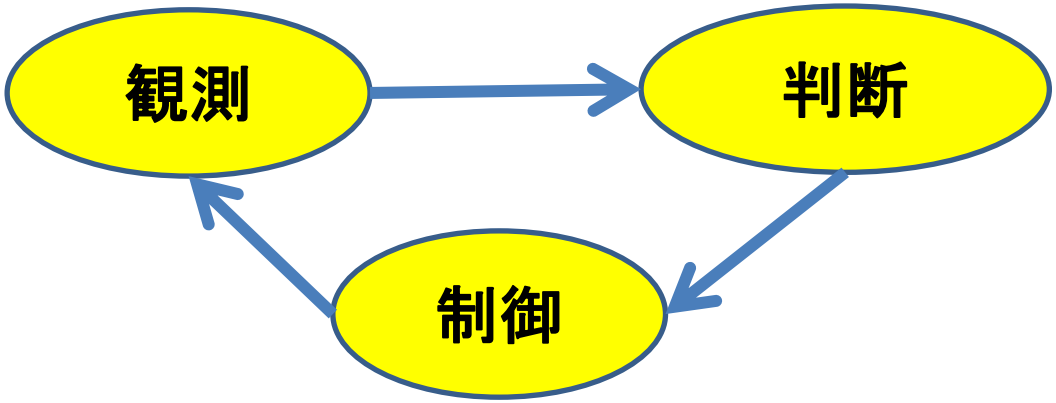
学内**150**研究グループ
＋
全国基盤センターとしての
性能向上が急務
他機関の同等の代替資源は基本的に日本では存在しない

重要な社会貢献を行う
学内外アプリケーション
にもフォーカス
災害・環境・健康医療等
東工大の貢献



Future Work: より合理的な電力最適化に向けて

システム全体レベルの電力キャップに向けて
フィードバック制御システムが必要



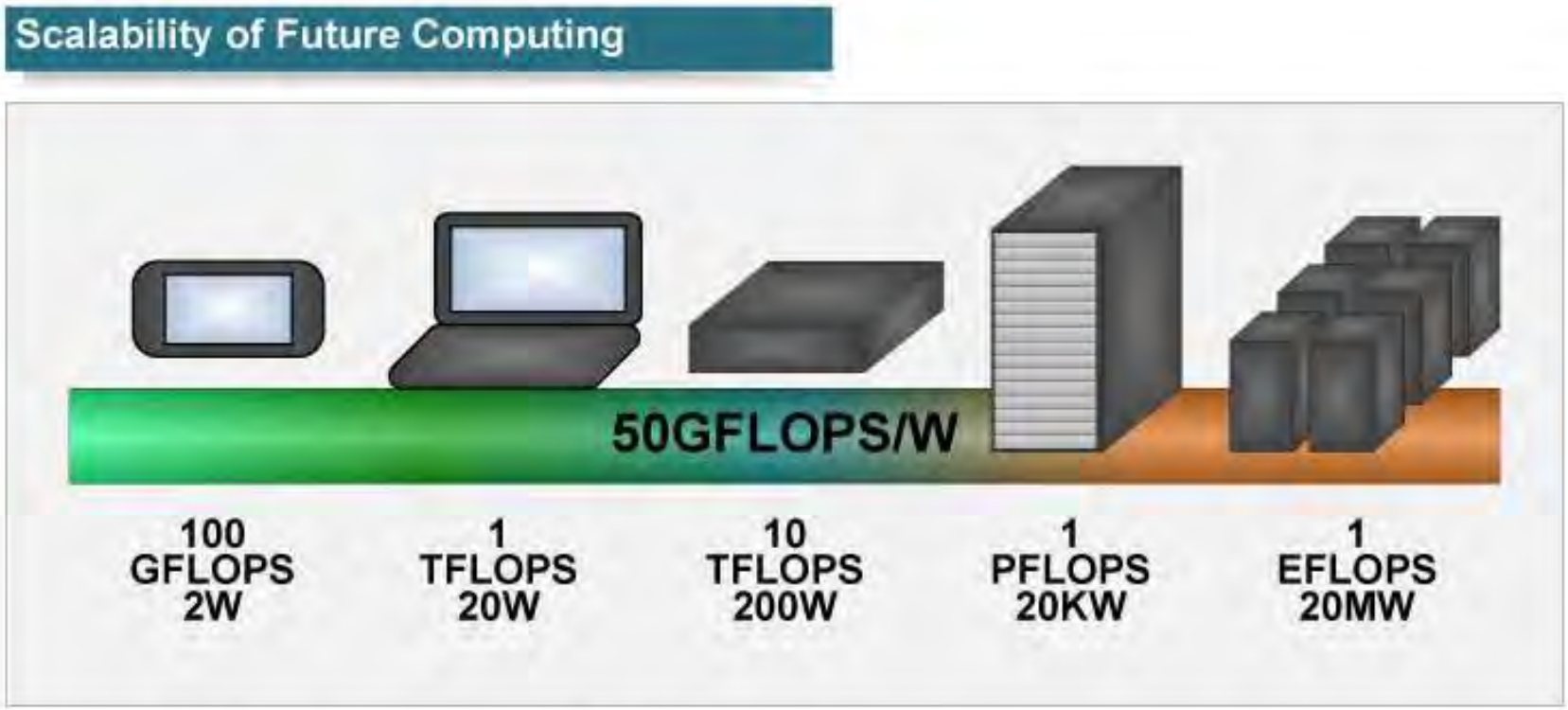
観測における課題:

- 電力測定の空間・時間的粒度問題

判断における課題:

- 電力バジェットの割り当て手法
  - 多数ノード間，多数アプリ間
  - CPUジョブとGPUジョブ間の公平性とは何か

2020年
エクサフロップススパコンへ向けて

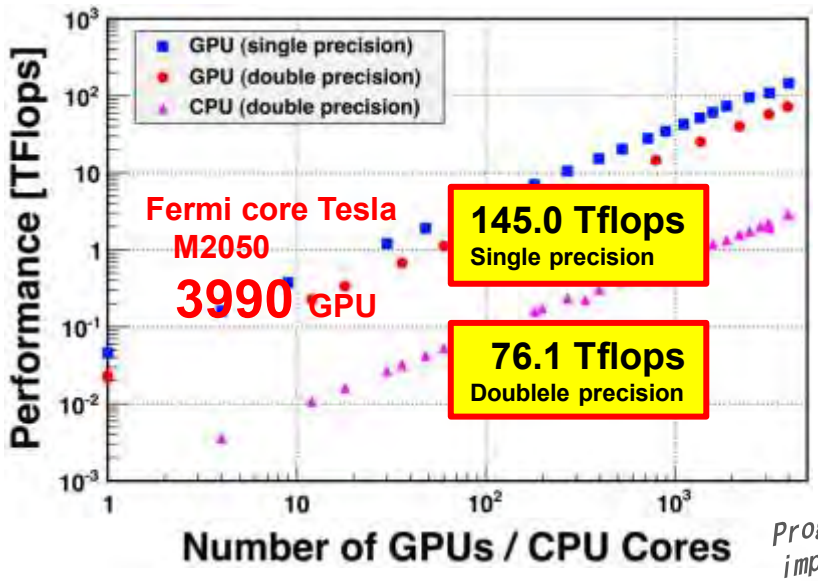


Machine	Power (incl. cooling)	Linpack Perf (PF)	Linpack MFLOPs/ W	Factor	Total Mem BW TB/s (STREAM)	Mem BW MByte/S / W	Factor
Earth Simulator 1	10MW	0.036	3.6	13,400	160	16	312
Tsubame1.0 (2006Q1)	1.8MW	0.038	21	2,368	13	7.2	694
ORNL Jaguar (XT5. 2009Q4)	~9MW	1.76	196	256	432	48	104
Tsubame2.0 (2010Q4)	**1.8MW**	**1.2**	**667**	**75**	**550**	**305**	**16**
K Computer (2011Q2)	~16MW	10	625	80	3800	236	21
BlueGene/Q (2012Q1)	~12MW?	17	~1417	~35.3	~3000	~245	20
(TSUBAME2.5 (2013Q3))	**1.8MW**	**~4**	**~2000**	**~22.5**	**~1000**	**~580**	**8.6**
Tsubame3.0 (2015Q3)	**1.8MW**	**~20**	**~11,000**	**~4.6**	**~4000**	**~2200**	**2.3**
EXA (2020)	20MW	1000	50,000	1	100K	5000	1

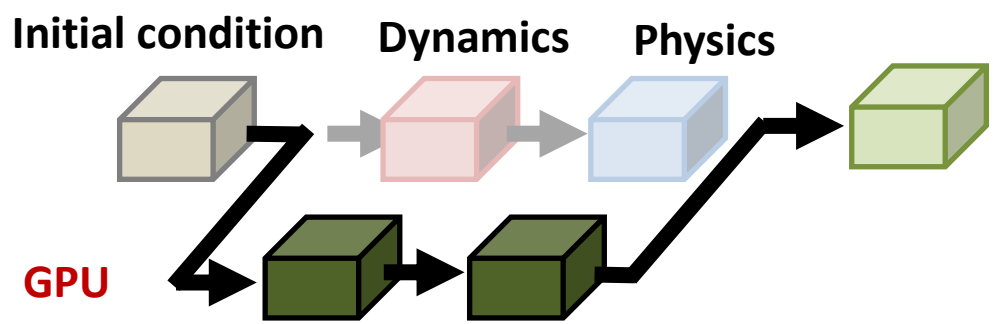
x31.6 ***x42*** ***~x16*** ***~x6.6***

現業用途の気象予報コードのGPUコンピューティングによる低消費電力化

気象庁数値予報課との共同研究



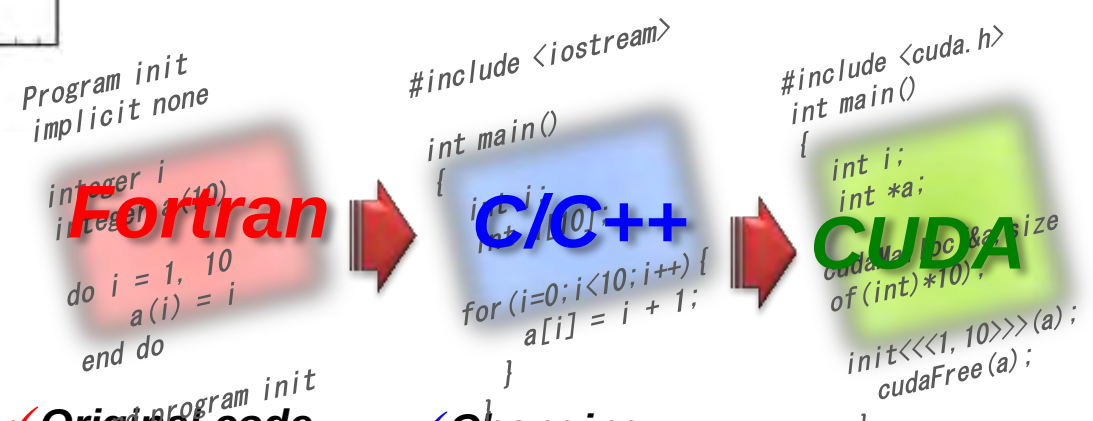
次世代気象予報コード ASUCA の
Full-GPU Implementation



世界中の気象庁、気象研究者に強い影響

スイス気象庁も気象予報コード COSMO をフルGPU化するプロジェクトを開始

SC'10 Best Student Paper Finalist



✓ *Original code at JMA*

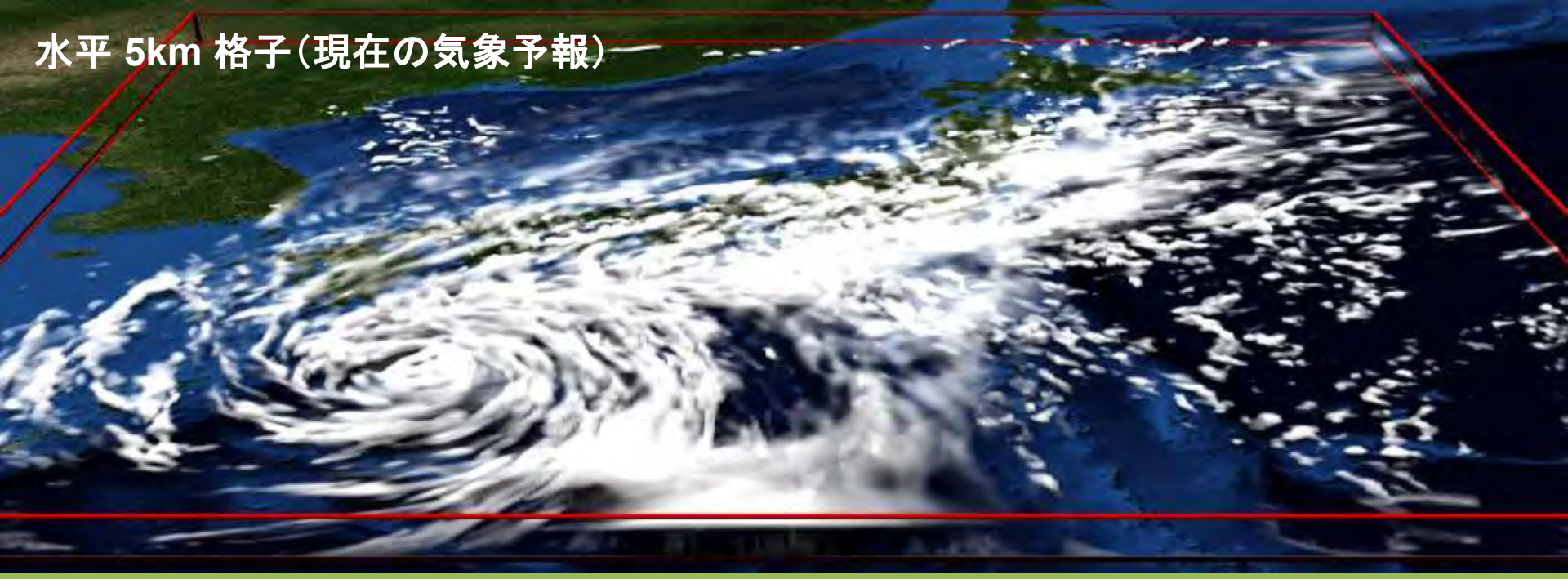
z,x,y (k,i,j)-ordering

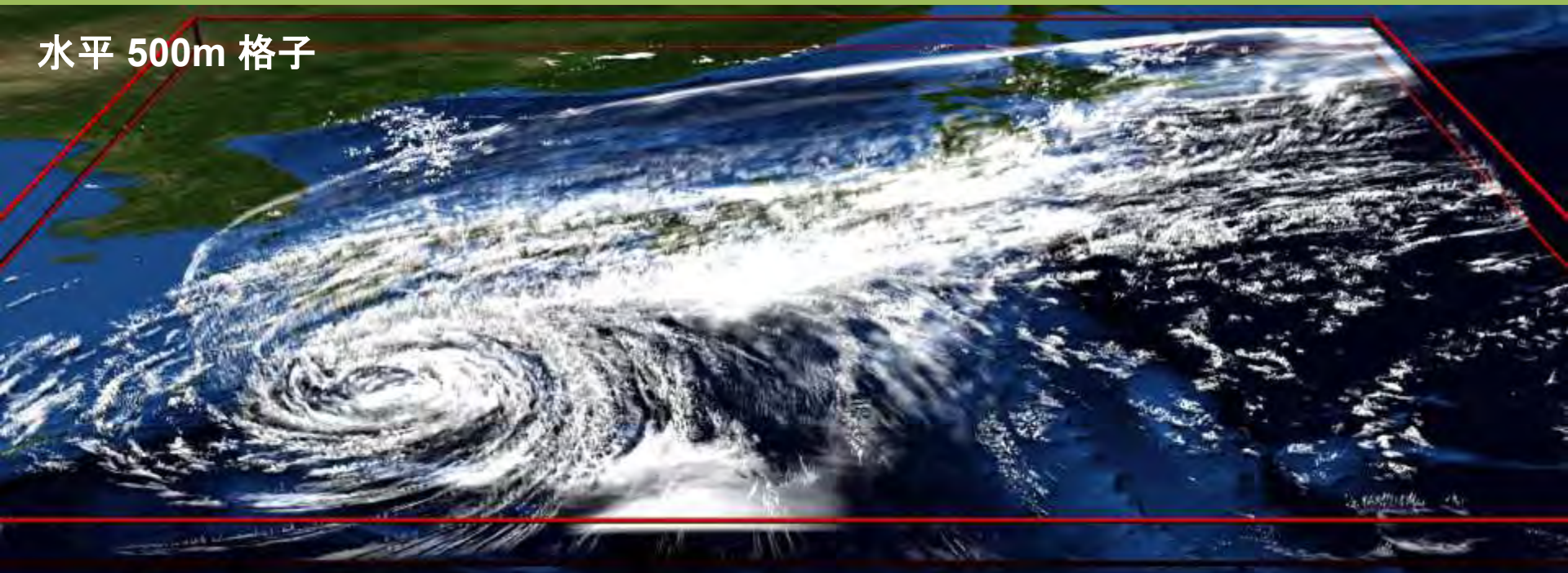
✓ *Changing array order*

x,z,y (i,k,j)-ordering

✓ *GPU code*

x,z,y (i,k,j)-ordering


水平 5km 格子（現在の気象予報）


水平 500m 格子