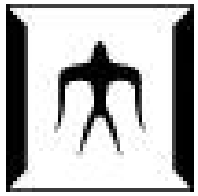
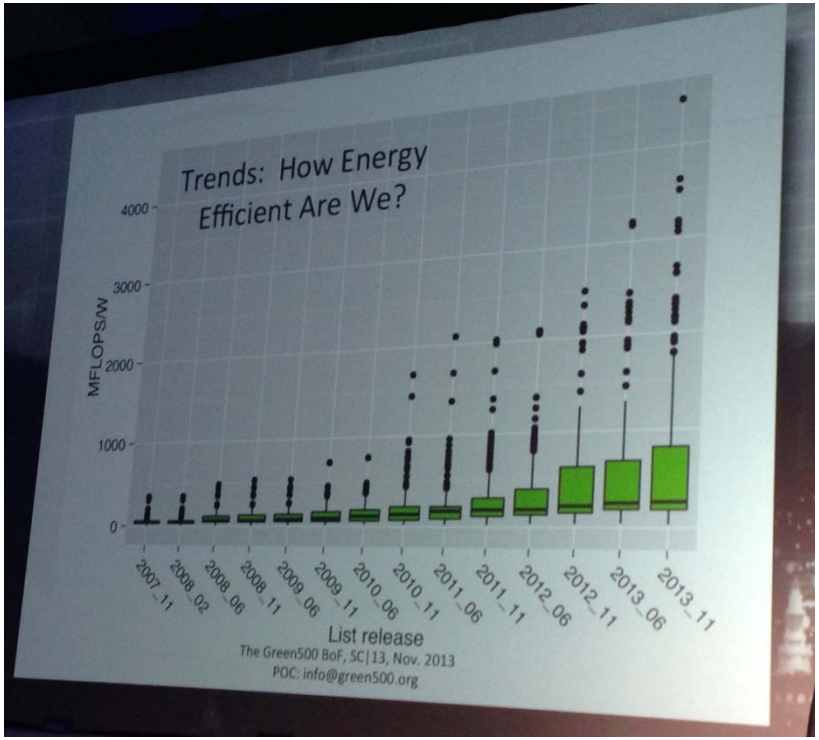
TSUBAME-KFC:
液浸冷却を用いた
ウルトラグリーンスパコン研究設備

遠藤敏夫、額田彰、松岡聡

東京工業大学学術国際情報センター



現在～将来のスパコンは電力あたり性能で決まる



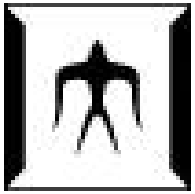
From Wu Feng's presentation
@Green500 SC13 BoF

- 現実的なスパコンセンターの電力の限界は20MW程度とされる
- Exaflopsのシステムを実現するには、50GFlops/W を実現する技術は不可欠

- Exaflops 2020年ごろ
- 冷却などの設備電力も考慮する必要

3年前のTSUBAME2.0

TSUBAME 2.0は0.96GFlops/Wを実現



- 2010/11Green500にて、2位(事実上3位)
- Greenest Production Supercomputer 賞

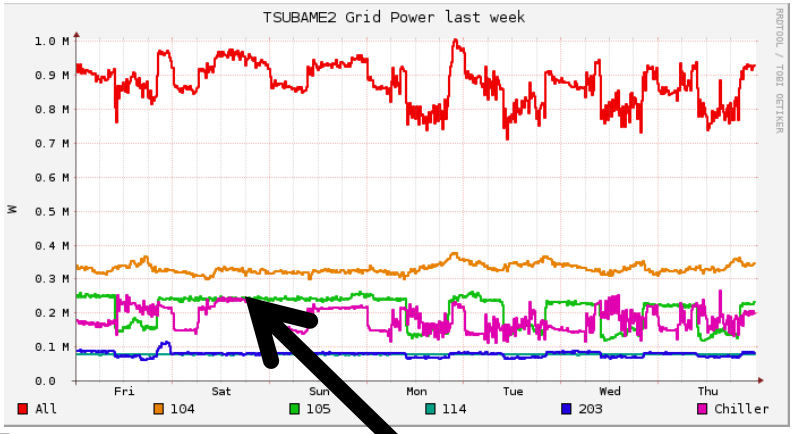
2015~16のTSUBAME3.0やその後へ向け
さらにグリーンにする必要性!!

さらにグリーンにするアプローチ

- 計算機電力の削減
  - プロセッサのプロセス縮小
  - スループット重視コア・アクセラレータ活用アーキテクチャ・およびソフトウェア技術

- 冷却設備電力の削減
  - 空冷より液冷のほうが有利
    - 高い比熱・熱容量
  - 「冷たすぎる」冷媒生成の除去

→ 本研究では液浸冷却に着目し、テストベッド***TSUBAME-KFC***を構築



TSUBAME2では
チラー電力が全体電力の25%！

今回の成果

- 2015年度末稼働予定のTSUBAME3.0のプロトタイプである**TSUBAME-KFC**が11月、SC13国際会議にて発表されたスパコンの**電力効率ランキングGreen500, Green Graph 500の両方において世界一位となり、世界初の二冠を達成**
- **両リストとも日本のスパコンが一位になるのは初めて**
- Green500において達成した電力効率は4.508GigaFlops/Wと、前回の一位から5割近く向上、今回の二位も24%引き離す。

TSUBAME-KFC

KFC: Kepler Fluid Cooling

= (液浸冷却技術

+ 外気冷却技術

+アクセラレータ付高密度ノード)

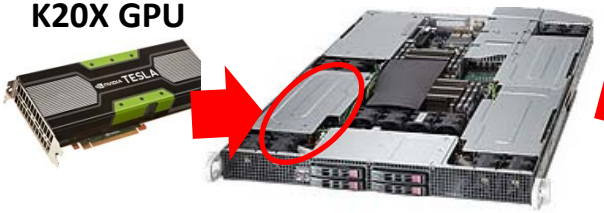
を20フィートコンテナ中に

TSUBAME-KFC: ウルトラグリーン・スパコンテストベッド

液浸冷却＋大気冷却＋高密度スパコン技術を統合した
コンテナ型研究設備 → TSUBAME3.0プロトタイプ

実証実験用計算サーバ群

K20X GPU

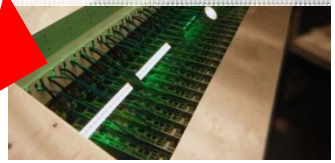


NEC LX104Re-1G改 ×40台

サーバ1台あたり
- Intel IvyBridge 2.1GHz 6core×2
- NVIDIA Tesla K20X GPU ×4
- DDR3メモリ 64GB, SSD 120GB
- 4x FDR InfiniBand 56Gbps

合計理論性能
210TFlops（倍精度）
630TFlops（単精度）

GRC製液浸サーバラック
プロセッサチップ 60~80℃
⇒ 冷媒油 35~45℃



熱交換器
冷媒油 35~45℃
⇒ 水 25~35℃

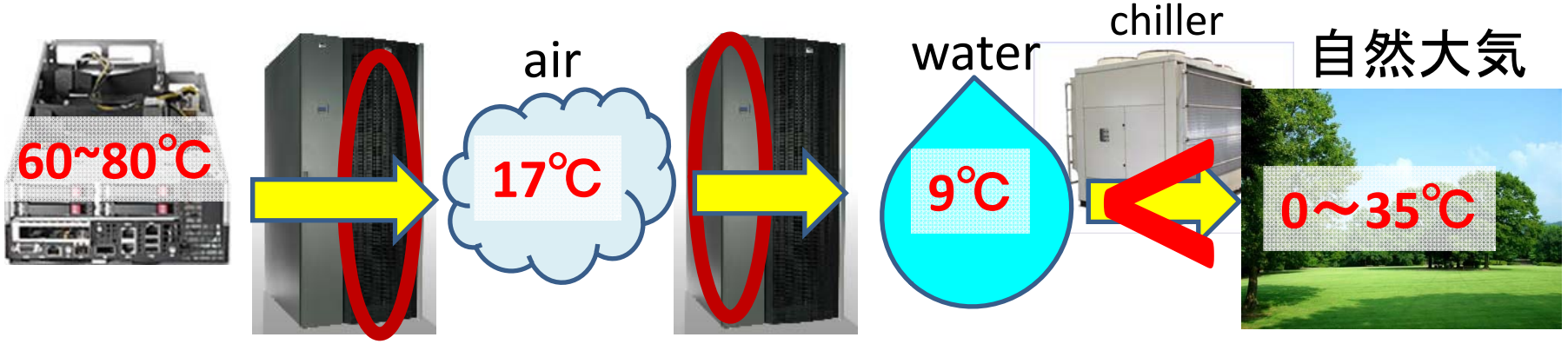
蒸散熱
自然大気中へ

ねらい

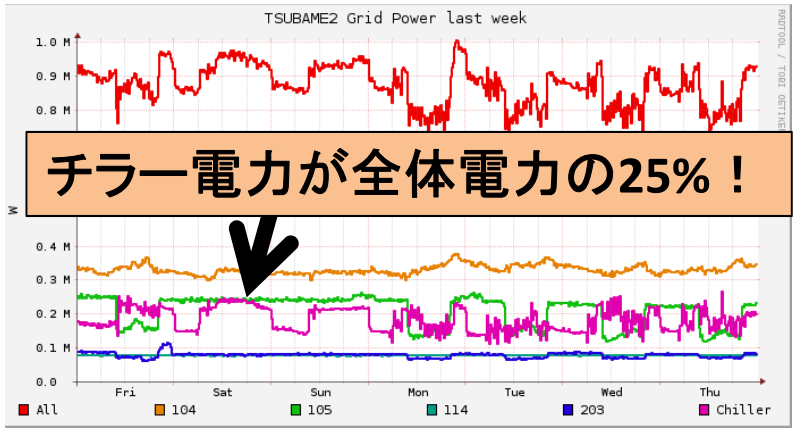
- 世界トップクラスの電力性能比, 3GFlops/Watt以上
- 平均PUE 1.05 (←この点は未実現・将来課題)
- ULP-HPC技術の実証実験

なぜ効率的な冷却と期待される？

一般的な冷却 (TSUBAME2の場合)

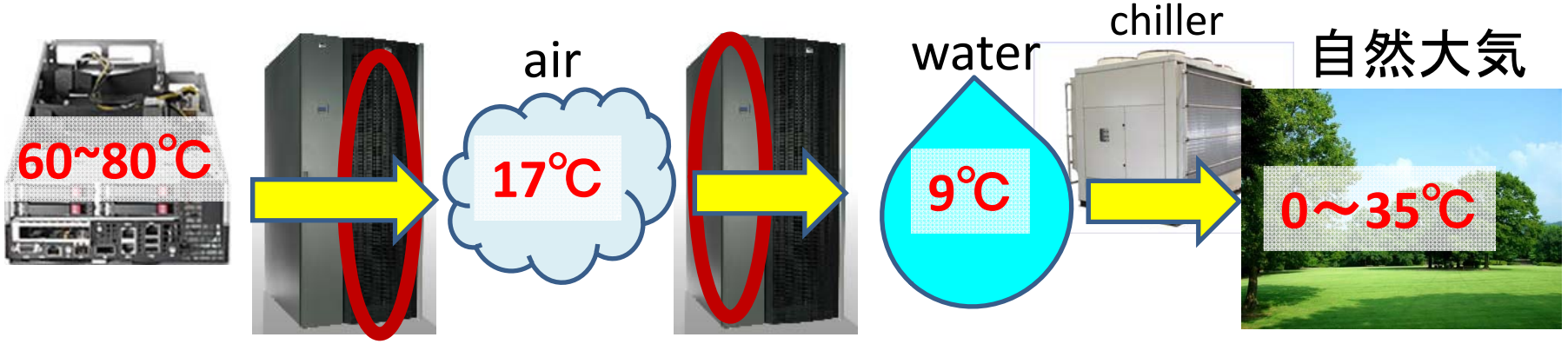


- 外気温より低温の冷媒水を作るためのチラーが電力を食う
  - 冷蔵庫・クーラーと同様にコンプレッサーなどを使うため

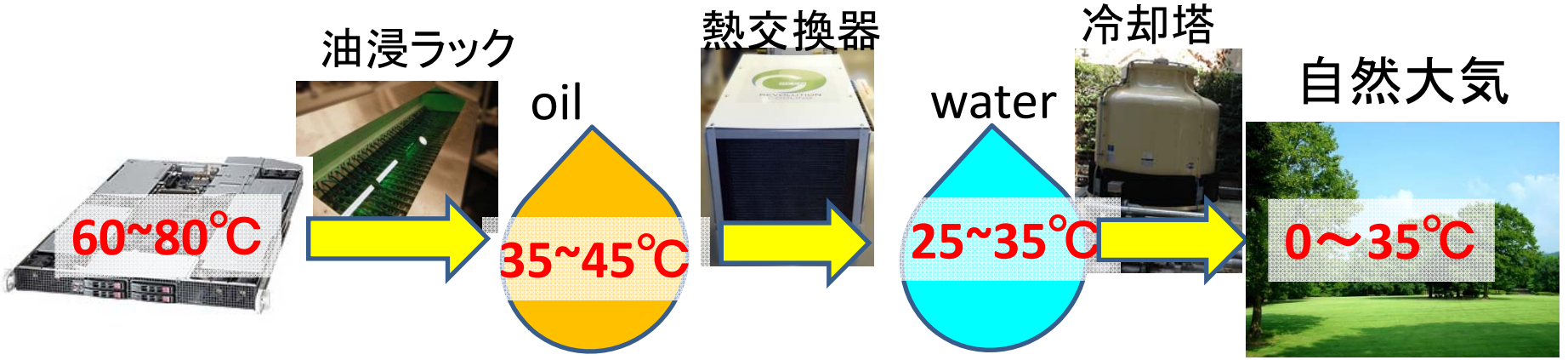


なぜ効率的な冷却と期待される？

一般的な冷却 (TSUBAME2の場合)

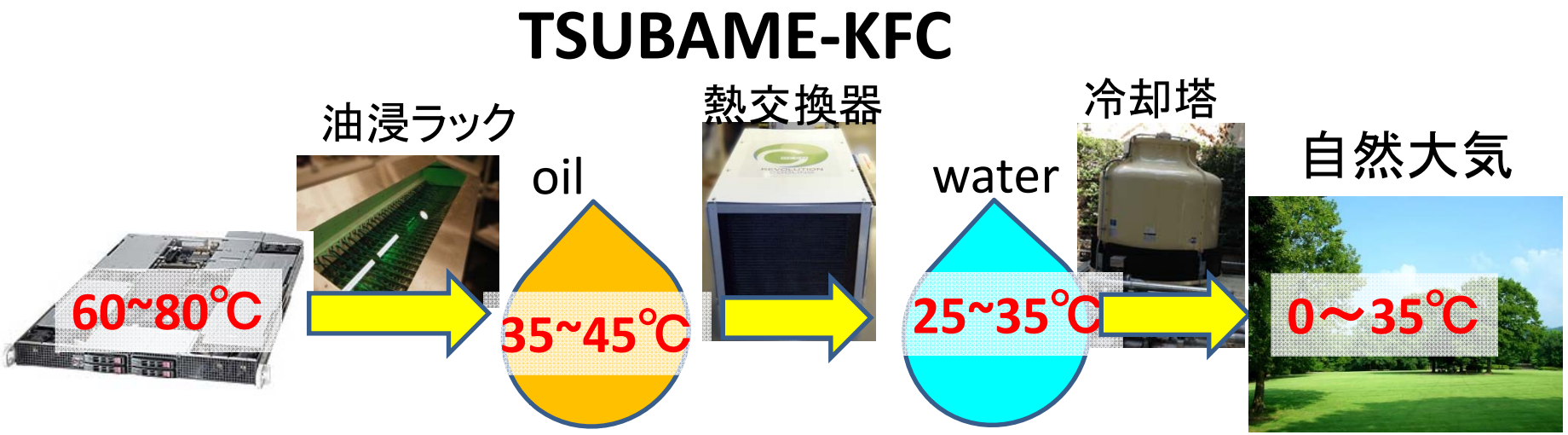


TSUBAME-KFC



なぜ効率的な冷却と期待される？

- KFCでは高温部→低温部に熱が流れる
  - 液体の比熱 > 空気の比熱のため有利
  - 原則的に、冷媒を動かすための電力のみ(ポンプ)
  - 真夏にどうなるかの評価は将来課題



東京において自然冷却可能な時期

外気湿球温度(℃)

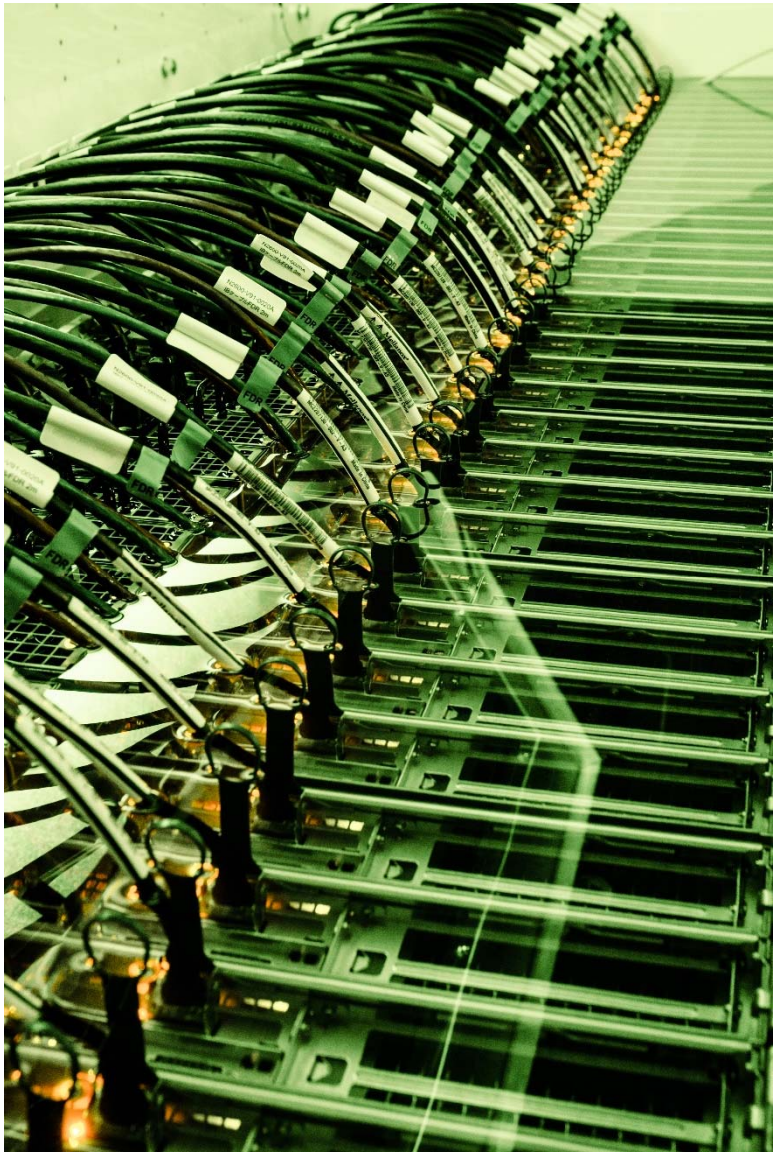
	東京		札幌	
	最高	平均	最高	平均
1月	15.0	4.0	4.2	−4.5
2月	15.6	4.2	5.6	−3.0
3月	17.2	7.3	8.6	−0.2
4月	21.7	11.7	16.4	5.5
5月	22.8	16.1	17.7	9.9
6月	25.2	19.4	21.8	14.1
7月	26.9	23.0	24.5	19.1
8月	27.4	23.5	24.1	18.8
9月	26.0	20.3	23.5	16.0
10月	24.3	15.8	20.1	10.1
11月	20.0	10.8	14.4	3.2
12月	17.5	6.7	8.1	−1.6

冷却塔の性質より、
冷却水温度≧外気湿球温度

- 青: 問題なし
- 黄: 冷却可能見込み
- 赤: 冷媒高温時の調査必要

独SuperMUCスパコンの温液冷却の成果を見ると、赤の時期ですらokな見込み

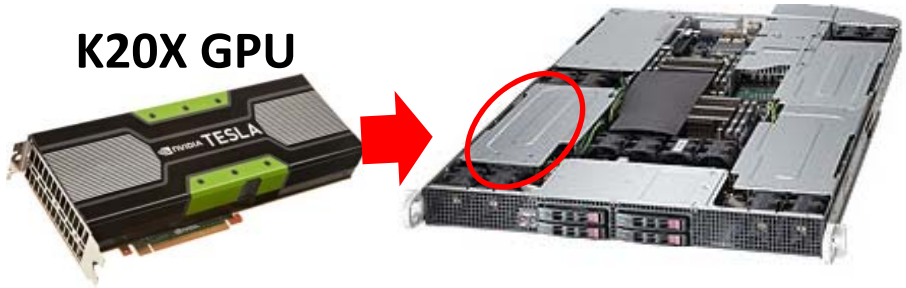
KFC計算ノード



NEC LX 1U-4GPU Server, 104Re-1G

(SUPERMICRO OEM)

- 2X Intel Xeon E5-2620 v2 Processor (Ivy Bridge EP, 2.1GHz, 6 core)
- **4X** NVIDIA Tesla K20X **GPU**
- 1X Mellanox FDR InfiniBand HCA
- 1X 120GB SATA SSD



Peak Performance (DP)

Single Node	5.26 TFLOPS
System (40 nodes)	210.61 TFLOPS

CentOS 6.4 64bit Linux
Intel Compiler, GCC
CUDA 5.5
OpenMPI 1.7.2

冷媒油の選定

GRC社標準の冷媒が，日本では第四類危険物に相当すると判明
⇒　検討の結果，
ExxonMobil SpectraSyn Polyalphaolefins (PAO)
を選定

	4	6	8
40℃動粘度	19 cSt	31 cSt	48 cSt
Specific Gravity@15.6C	0.820	0.827	0.833
Flash point (Open Cup)	220 C	246 C	260 C
Pour point	-66 C	-57 C	-48 C

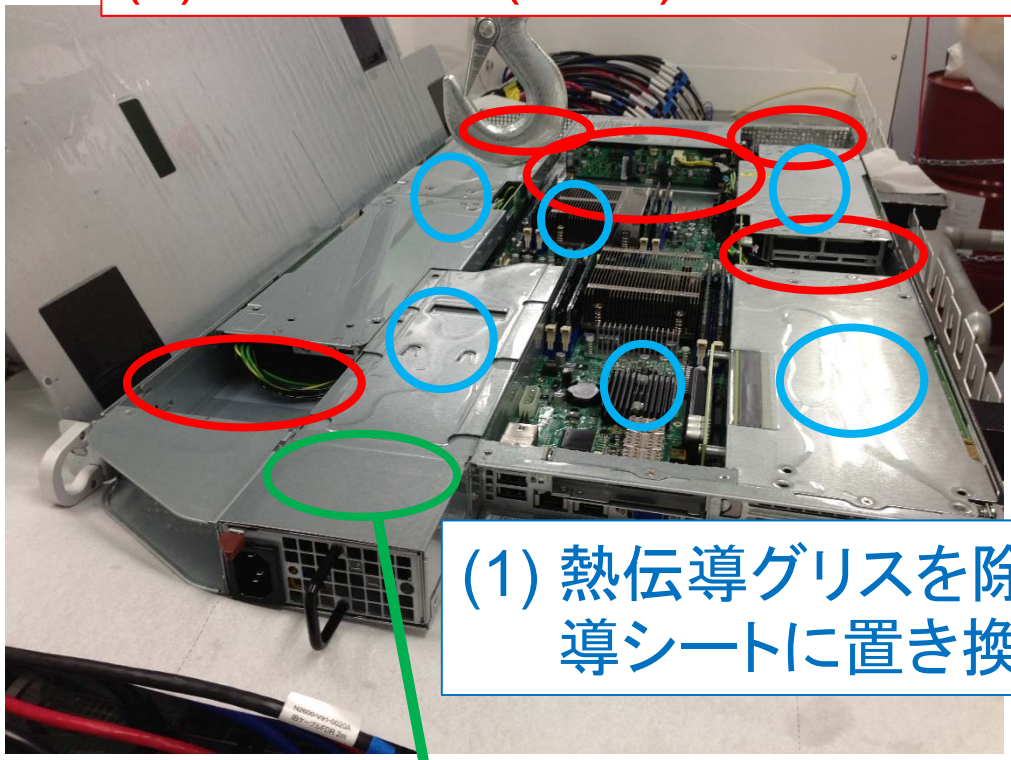
田園調布消防署

消防法における危険物該当外である，引火点が250℃超の油を選定

消防署との協議により，危険物の安全規定を考慮
- 油槽の周りの間隔，コンテナ・扉の材質など

計算ノードの改造

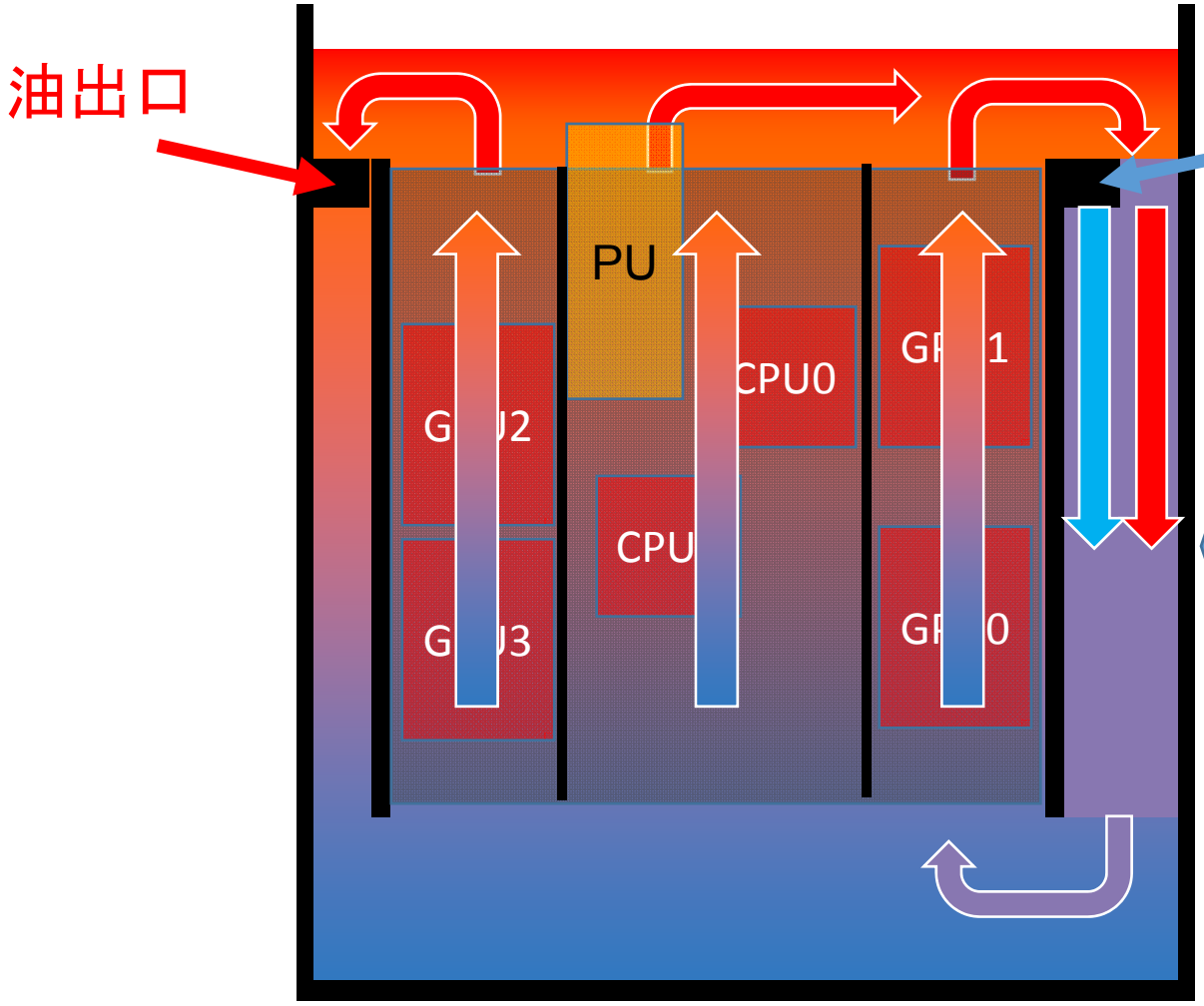
(2) 冷却ファン(12個)を除去

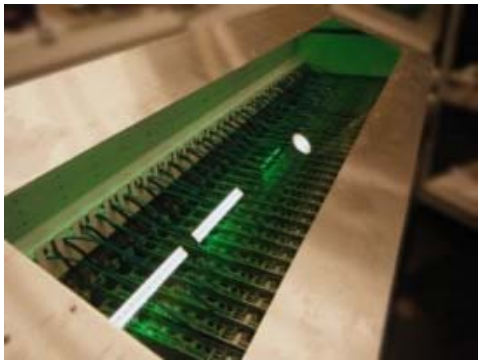


(1) 熱伝導グリスを除去，熱伝導シートに置き換え

(3) ファームウェアを変更し，冷却ファンが除去・停止しても稼働可能に

Green Revolution Cooling社 CarnotJet システム





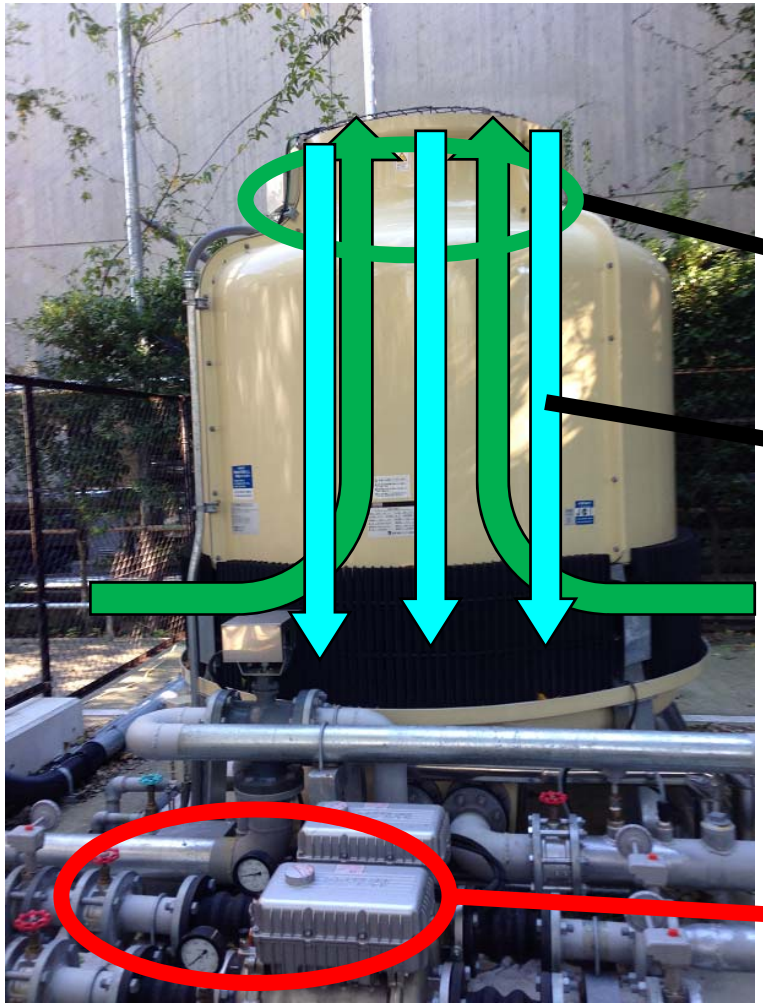
油⇔水の熱交換器

チューブ型熱交換器×３

冷媒油ポンプ×２

ポンプの流速は、油温・水温に従って
インテリジェントに調整

コンテナ外冷却塔



ファンあり：大気を下から上へ吹上

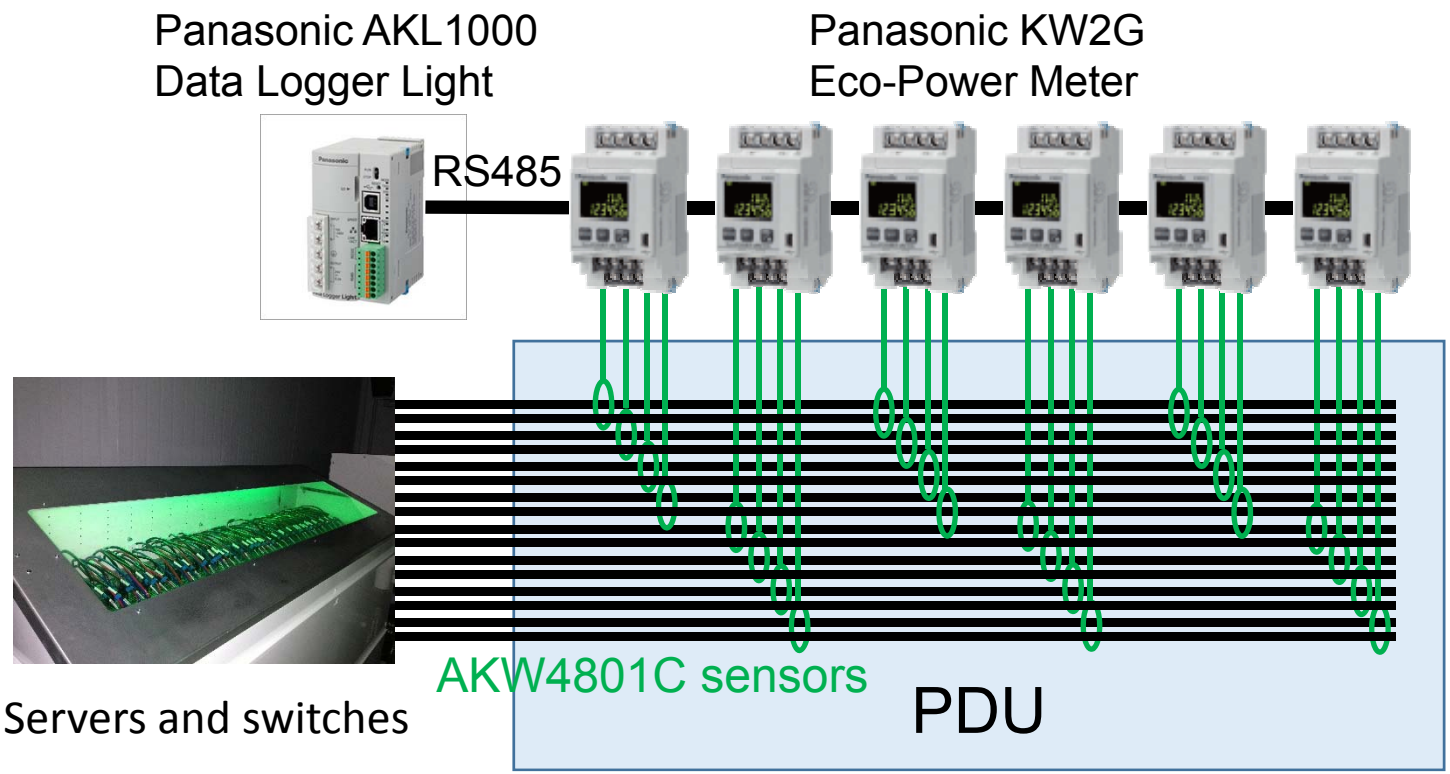
冷媒水は上から下へ

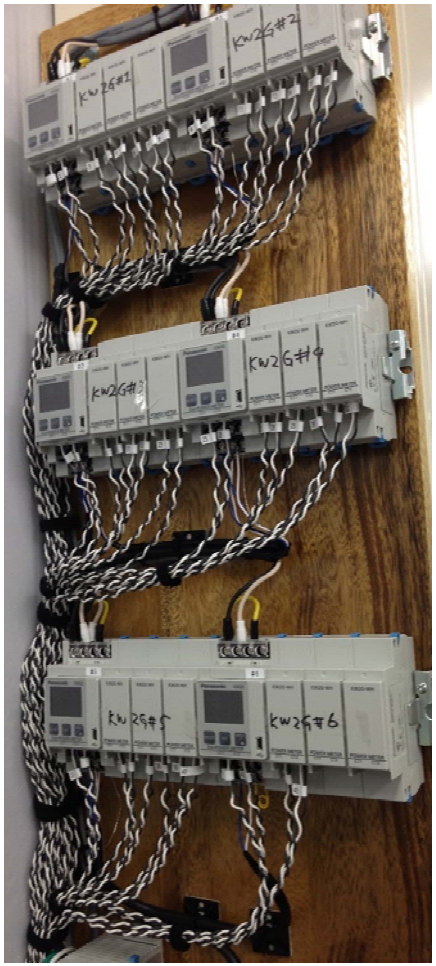
冷媒水用ポンプ×２

電力測定システム

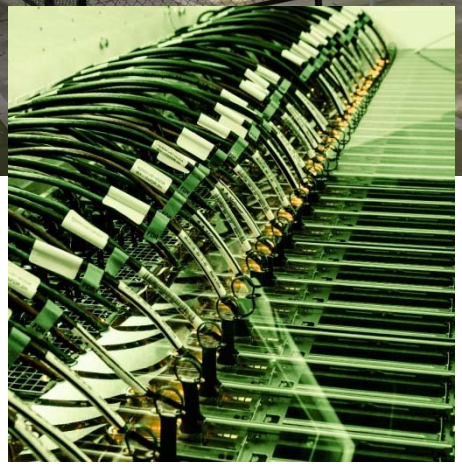
TSUBAME-KFCでは，毎秒毎に
- 各計算ノード
- ネットワークスイッチ

の電力を記録





TSUBAME-KFC外観



2013年9月インストール完了

電力性能評価指標

PUE (Power Usage Effectiveness)

$$PUE = \frac{(\text{IT機器に使う電力}+\text{冷却等電力})}{\text{IT機器に使う電力}}$$

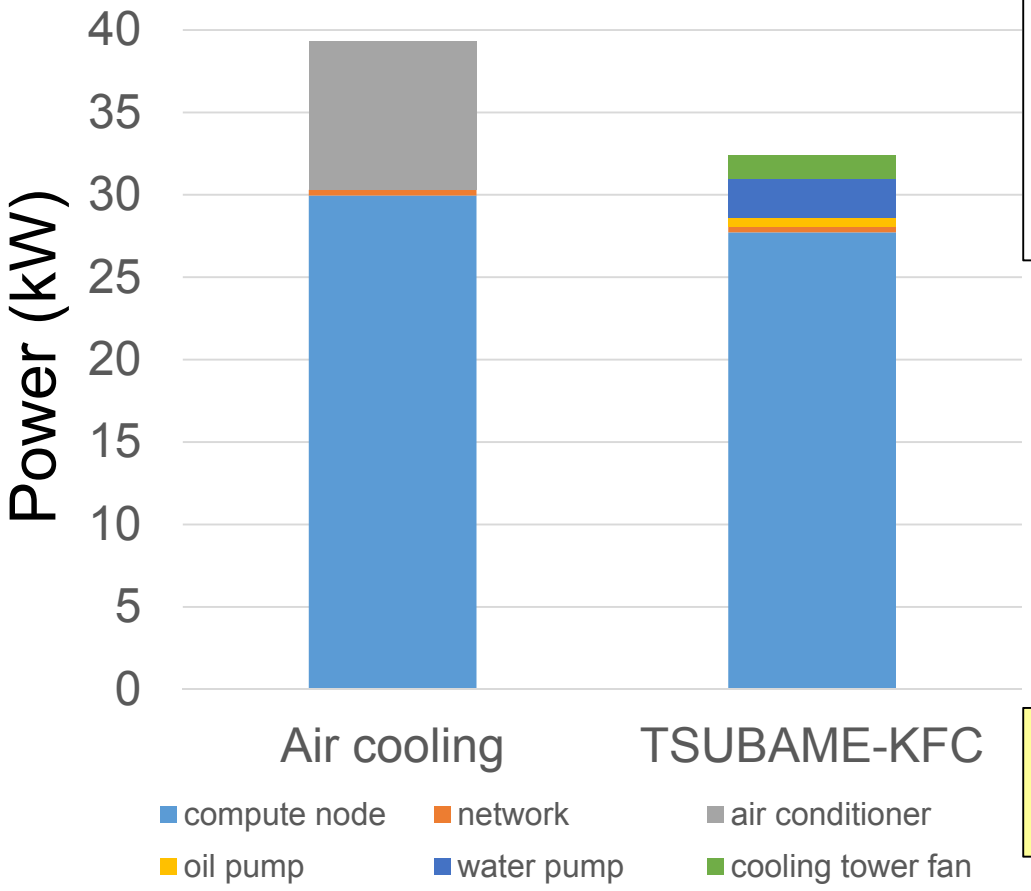
- 1が理想、2以上だとへぼいセンターと言われる
- TSUBAME2は年間平均1.3 ⇒ KFCで1に近づける！
- IT機器の効率性は入らない指標

Green500ランキングの指標

$$\text{効率(Flops/W)} = \frac{\text{Linpack性能 (Flops)}}{\text{Linpack時IT機器電力(W)}}$$

- Linpack時の効率を考慮
- 冷却等電力は入らない ⇒ KFCであまり有利でない指標
- 分母に冷却電力も含めれば、まあまあよい指標か
  2020年に50GFlops/Wめざす！

TSUBAME-KFCのPUE評価



空冷ではPUE=1.3と仮定

KFCの**PUE = 1.15**

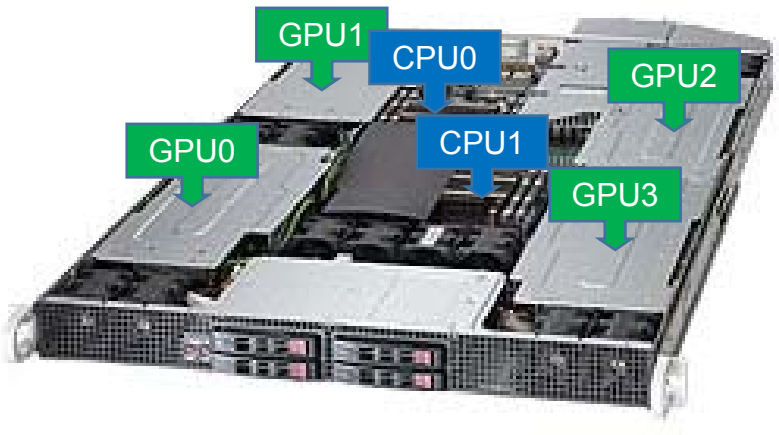
- GPU DGEMM時
- 空冷時ノード電力を基準にすると**1.068**

油ポンプ (60%)	0.53 kW
水ポンプ	2.40 kW
冷却塔ファン	1.40 kW
冷却電力合計	4.33 kW

水ポンプの電力が想定より大きく，PUEがやや悪化 ⇒ 今後の課題

計算ノード内温度と電力

上: GPU上でDGEMM(行列積)実行時
下: アイドル時



プロセッサ温度はIPMIで取得

油温の低下によりプロセッサ温度低下

	Air 26 deg. C	Oil 28 deg. C	Oil 19 deg. C
CPU0	50 (43)	40 (36)	31 (29)
CPU1			
GPU0	52 (33)	47 (29)	42 (20)
GPU1	59 (35)	46 (27)	43 (18)
GPU2	57 (48)	40 (27)	33 (18)
GPU3			42 (18)
Node Power	749W (228W)	693W (160W)	691W (160W)

28℃の油は
26℃の空気より冷える

~8% 電力減！

空冷⇒液浸で8%電力減
- ノード内ファンの除去
- リーク電流減少

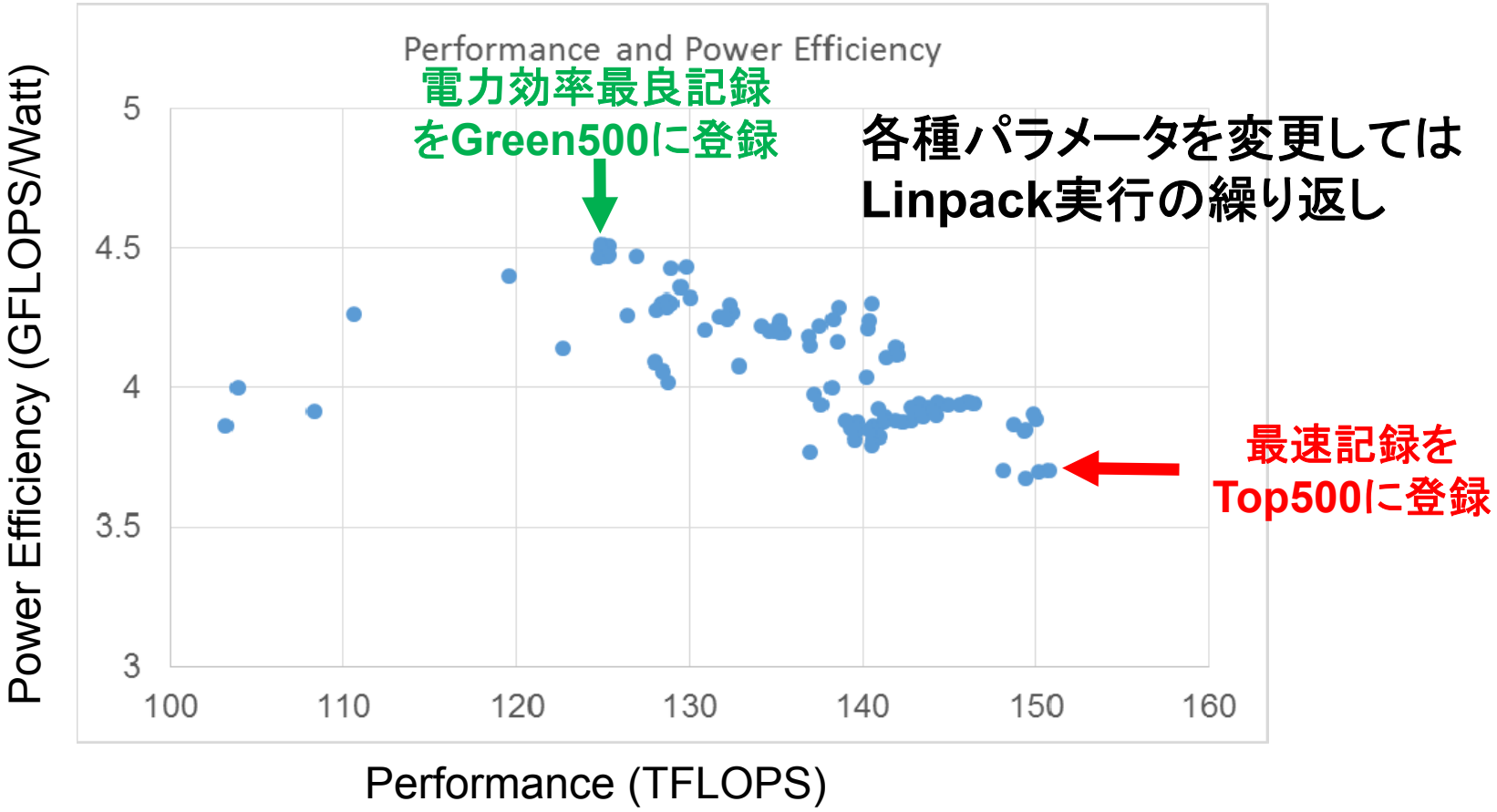
外気環境のシステムへの影響

	雨天 Oct. 29th 17pm	曇天 Oct. 30th 17pm	晴天 Oct. 31th 17pm
外気温	14.8 C	19.7 C	19.8 C
外気露点温度	15.2 CDP	15.9 CDP	11.7 CDP
湿度	99%	75%	56%
冷媒水温	14.8 C	16.8 C	14.9 C
油槽上部温度 (2センサー)	25.7 / 28.0 C	27.0 / 29.4 C	25.4 / 27.4 C
冷媒油温(out)	24.2 C	23.3 C	23.5 C
熱交換 (in)	18.0 C	19.3 C	17.8 C
熱交換 (out)	18.9 C	19.9 C	18.5 C
熱交換器電力 (主に油ポンプ)	572W	566W	555W

Top500とGreen500ランキング

(www.top500.org, www.green500.org)

- Top500: Linpackベンチマークの速度性能(Flops)でランク
- Green500:ワットあたりのLinpack速度性能(Flops/Watt)でランク
  - 速度性能がTop500 500位以上であることが出場条件



KFCがGreen500で有利である理由

計算ノードデザインによる利点

- GPU:CPU比が4:2 (TSUBAME2.5では3:2)
- 省電力Ivy Bridge CPU (TSUBAME2.5ではWestmere)
- 冷却方法の影響：ノード内ファンの除去，チップ温度低下

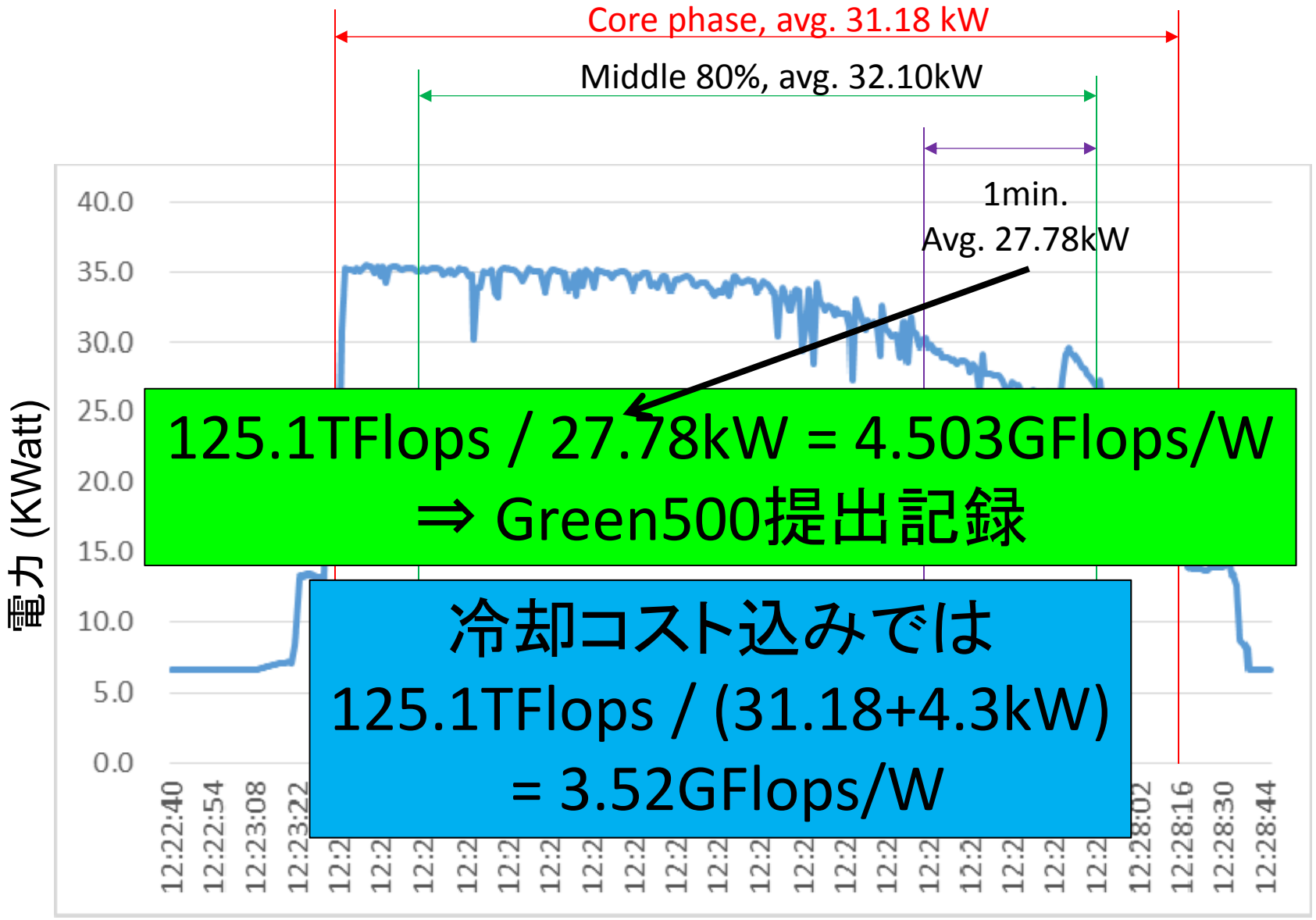
ソフトウェア・チューニングによる利点

- Linpackソフトウェア
  - 今回はNVIDIA提供のバージョンが最良（遠藤版は勝てず）
  - 行列サイズはGPUメモリに収まる範囲
- GPUクロック周波数・電圧のチューニング
  - K20Xで選択可能な周波数 (MHz):

    614 (best), 640, 666, 705, 732 (default), 758, 784

    ⇒ クロック・電圧を落とすほうが電力効率良
- Linpackパラメータのチューニング
  - 主にブロックサイズ (NB)，プロセスグリッド (P&Q)

Linpack中の電力推移とGreen500提出記録



2013/11 Green500ランキング

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss Centre	nterconnect ,	1,753.66
5	3,130.95	ROME Ardenn	00GHz,	81.41
6	3,068.71	GSIC C Techno	6C 2.930GHz,	922.54
7	2,702.16	Univers	finiband	53.62
8	2,629.10	Max-Pl	Infiniband,	269.94
9	2,629.10	Financ	Infiniband,	55.62
10	2,358.69	CSIRO	.000GHz,	71.01

おわりに

TSUBAME-KFCは4.5GFlops/Wで、Green500 世界一

- 国内スパコンとしては初
- 二位と24%差
- 冷却コスト込み3.5GF/W　⇒ 50GF/Wへ向け邁進

日本電気、NVIDIA、Green Revolution Cooling、
Super Micro、Mellanox、東工大関連部署
をはじめとする皆様に深く感謝します