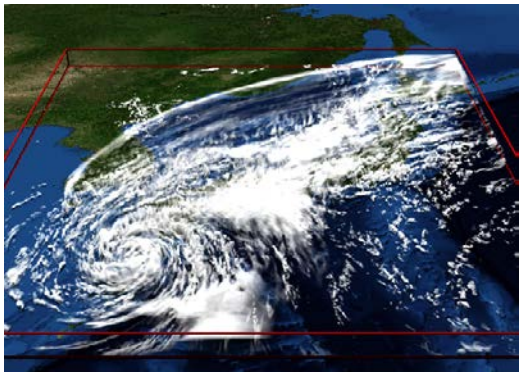
TSUBAME3.0冷却システムの紹介

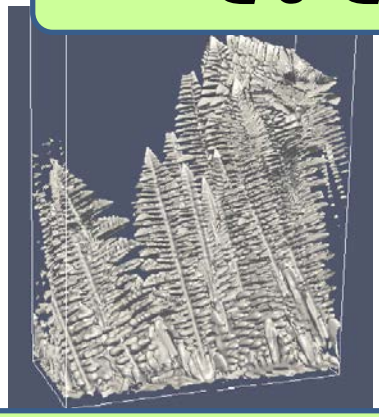
2018年4月19日（木）

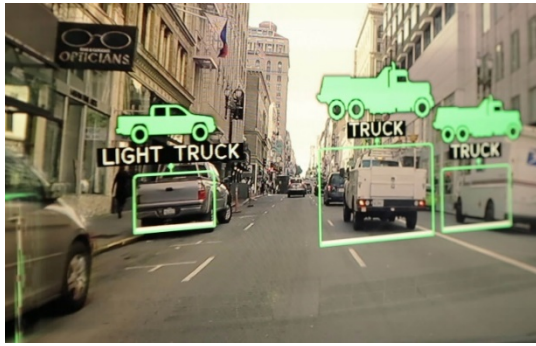
東京工業大学
学術国際情報センター
遠藤敏夫

AI時代・科学技術・ものづくりを支えるスーパーコンピュータ

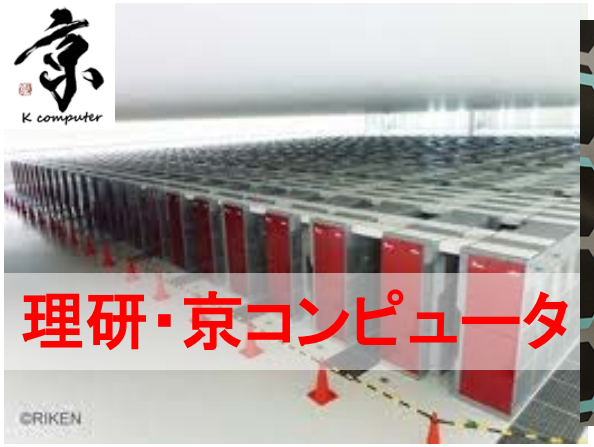
さまざまな計算







さまざまなスーパーコンピュータ

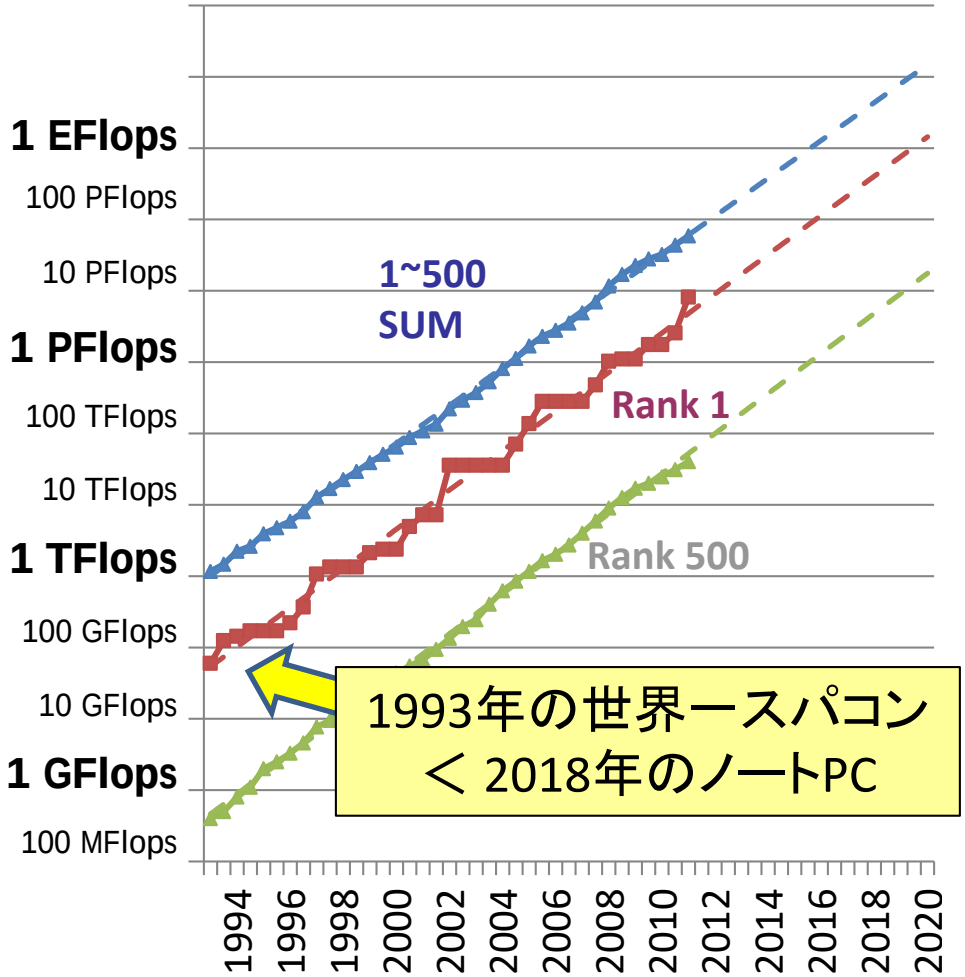


理研・京コンピュータ

東工大・TSUBAME3.0

中国・天河二号

スパコン性能の進化



1993年の世界一スパコン
＜ 2018年のノートPC

※Flops: 1秒間に行える浮動小数演算の回数
10PFlopsなら、1秒間に一京回の演算。
スパコン性能を表す代表的な単位

スパコンの性能は**指数関数的**に伸びてきた

- 20年で性能500,000倍！

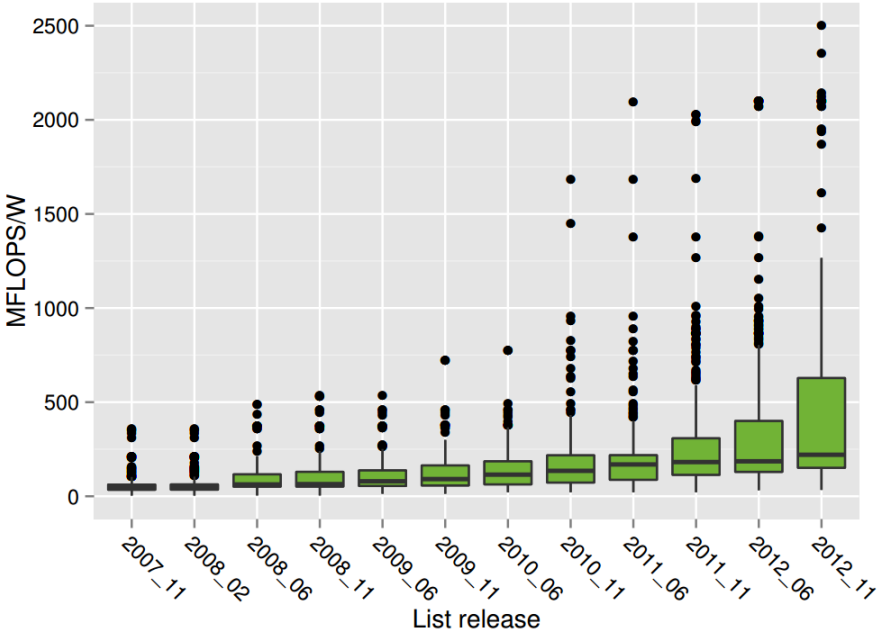
今後も同じように伸びるか？

消費電力が最大の壁に！

現在のトップスパコンは
すでに数十MW（メガワット）

将来のスパコンは「電力あたりの性能」で決まる

省エネスパコンランキング
Green500におけるFlops/Wの推移
[Wu Feng et al, IGCC13]



- 現実的なスパコンセンターの電力の限界は数十MW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

→ *2022年ごろに達成できるか？*

一口にスパコンの電力と言っても

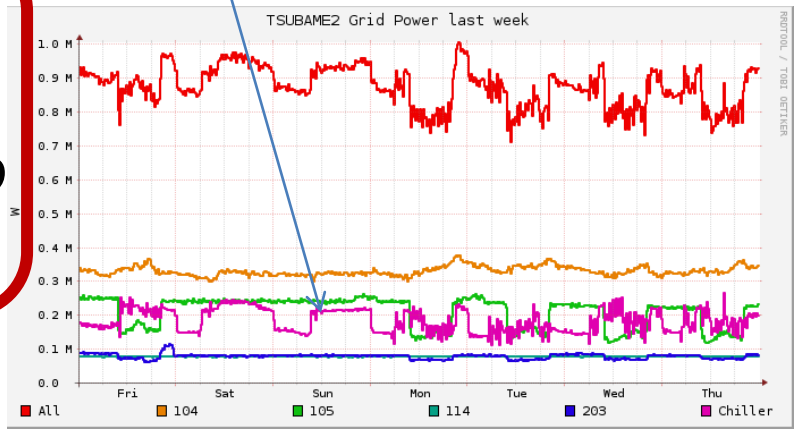
- 計算機システムそのものの電力
- 冷却などの設備電力

スパコンをよりグリーンにするには

- 計算機システム電力の削減
  - プロセッサの性能÷電力比の向上を利用
    - ムーアの法則
  - メニーコア技術に基づくGPGPUの活用
  - 通信ボトルネックを解消するシステムデザイン
  - システムを効率活用するためのソフトウェア最適化技術

- 冷却システム電力の削減
  - 空冷 → 液冷の流れ
  - 冷たい冷却水(10℃以下)をどうやって排除するか？

TSUBAME2(2010-2017)では、
チラーが25%電力を占めていた！

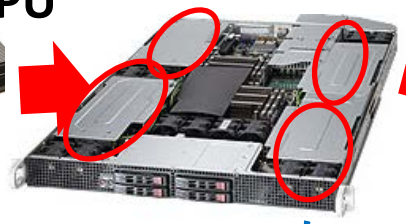


TSUBAME-KFC: ウルトラグリーン・スパコン試作機

油浸冷却＋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

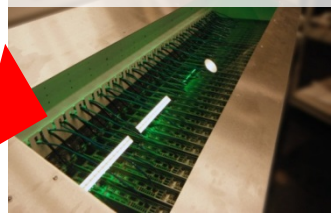
4GPU搭載計算サーバ群

K20X GPU



NEC LX104Re-1G改×40台

液浸サーバラック
熱はプロセッサチップから油へ



熱交換器
熱は油から水へ

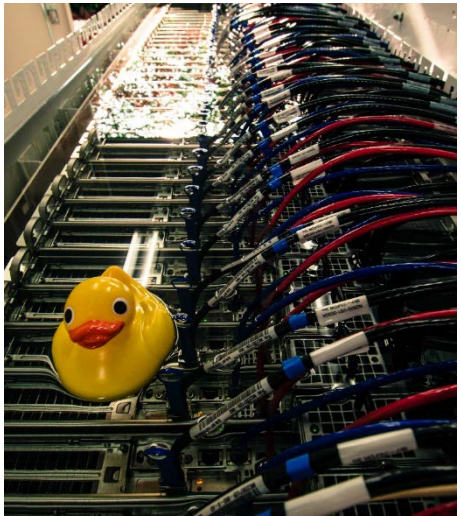
蒸散熱
自然大気中へ

合計理論性能
217TFlops（倍精度）
645TFlops（単精度）

設計時目標

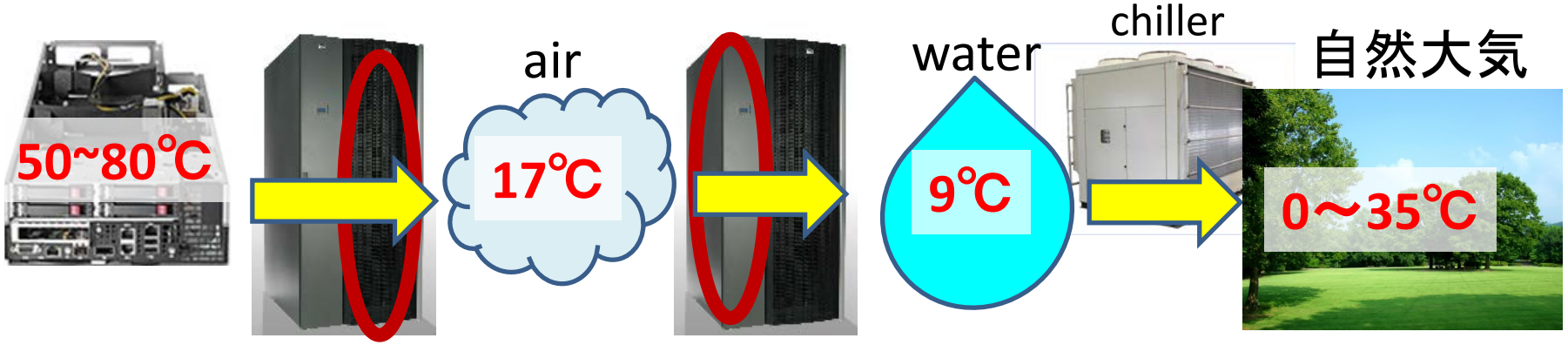
- 世界トップクラスの電力性能比, 3GFlops/Watt以上
- 次世代の超省電力スパコン技術の実証実験

TSUBAME-KFC

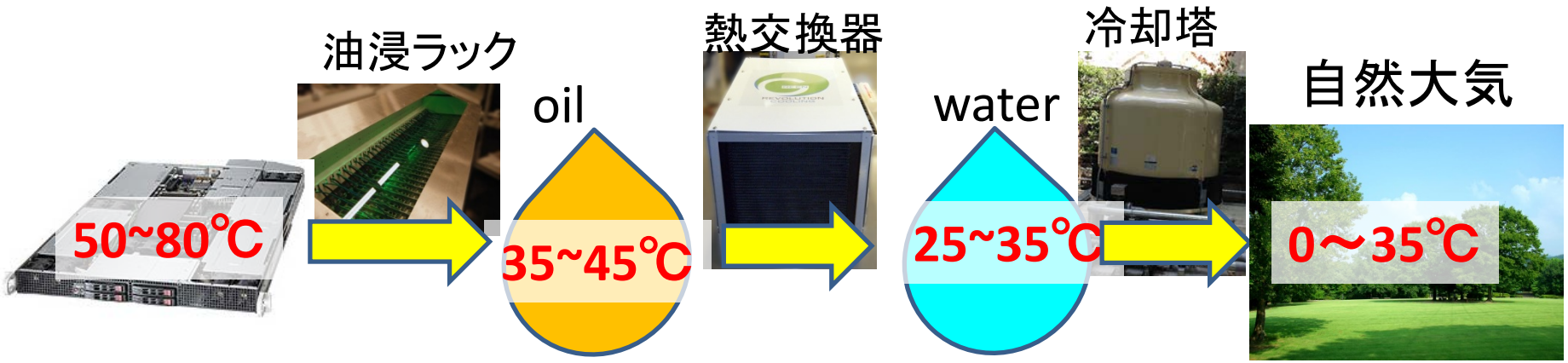


なぜ油浸冷却は効率的?

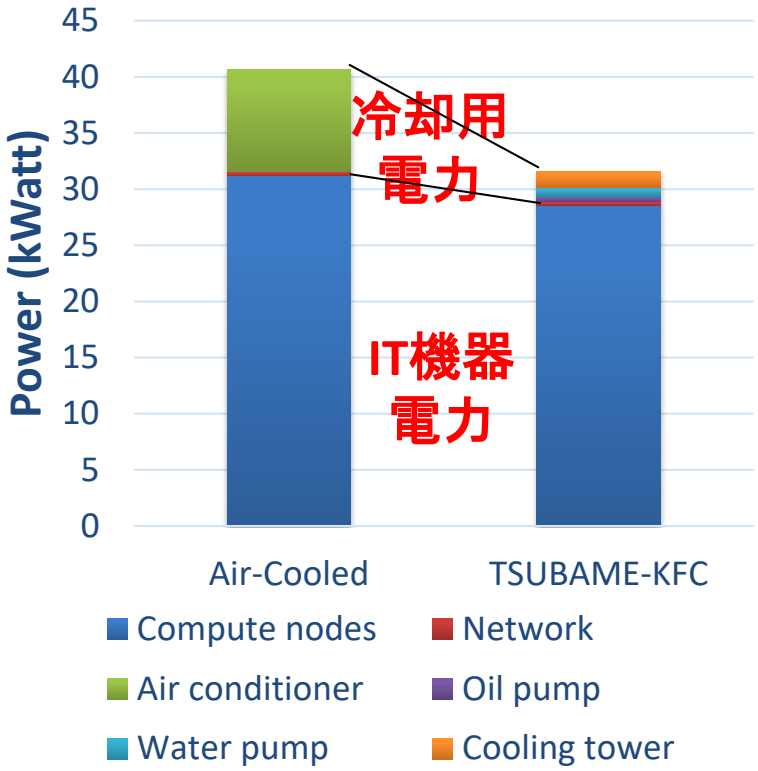
一般的な冷却 (TSUBAME2の場合)



TSUBAME-KFC



TSUBAME-KFCの電力評価



同種の計算ノードの空冷の場合と比較
⇒**液浸**により、
- **冷却用電力は70%減！**
- **計算機電力も8%減**

計算機電力減少の理由:
- サーバ内のファンの除去
- チップ温度低下によるリーク電流削減

空冷ではPUE=1.29 (TSUBAME2なみ)と仮定

$$PUE = \frac{(\text{IT機器電力}+\text{冷却等電力})}{\text{IT機器電力}}$$

KFCの**PUE = 1.09**
(設備部分には冷却のみ算入)

TSUBAME-KFCでの実験を経て

油浸冷却の利点と課題

利点

- [省エネ] 系内に10℃以下の箇所がない。冷却水ループ・油ループとも30℃以上
- [省エネ] 計算機内のファンが不要・チップ温度の低下
- [汎用性] 通常のラックマウントサーバを利用可能

課題

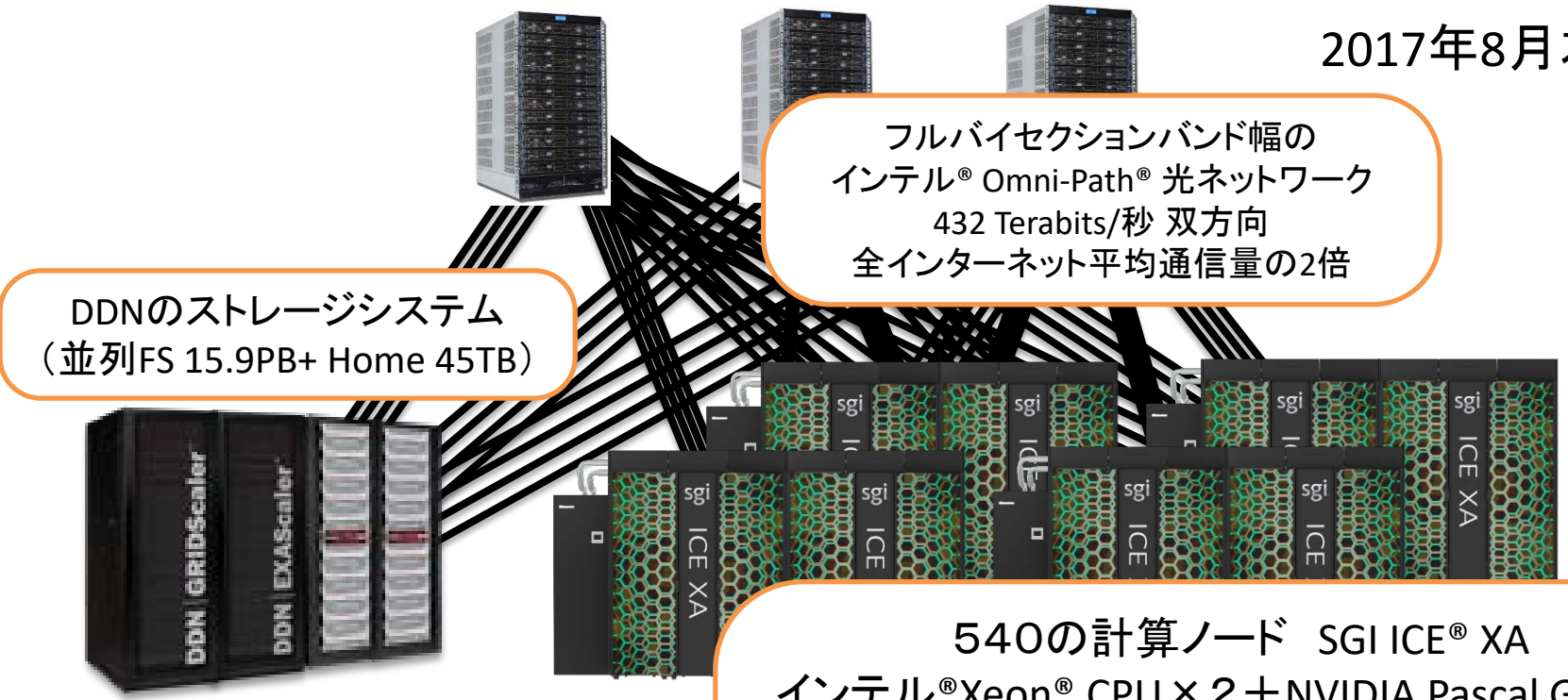
- [設置環境] ラックあたり、2倍近い広さ・油のために+1トン必要
- [メンテナンス] 計算機パーツ交換のためには「油揚げ」必要
- [条例] 数十トンの油の設置のためには消防法からの制約も

→ TSUBAME3.0冷却システムでは上記を考慮、2017年に稼働開始

東工大TSUBAME 3.0 のシステム概要

2000枚以上の最新P100 GPU・1000基以上のCPUにより、倍精度12ペタフロップス

Integrated by
Herlett Packard (HPE)
2017年8月本稼働



フルバイセクションバンド幅の
インテル® Omni-Path® 光ネットワーク
432 Terabits/秒 双方向
全インターネット平均通信量の2倍

DDNのストレージシステム
（並列FS 15.9PB+ Home 45TB）

540の計算ノード　SGI ICE® XA
インテル®Xeon® CPU×２＋NVIDIA Pascal GPU×４
256GBメモリ、2TBのNVMe対応インテル®SSD
47.2 AI-Petaflops, 12.1 Petaflops

システム消費電力
- 約500kW (平均運用時)
- 約800kW (ピーク時)

ラックあたり、50～60kW！！

TSUBAME3.0の冷却システムの方針

- 主要熱源である、GPU/CPUは直接液冷とする
  - 水パイプを計算機中に通す
  →液冷方式としては、液浸よりもメジャー
  ただし、水温を高くする方針はKFCから受け継ぐ
- 汎用性についてはどうか？
  - スパコンにおいては、サーバの物理的な種類は少ない。TSUBAME3.0においては、540台が均一
  → HPの協力のもと、新規ノード設計・水冷パイプ設計
- GPU/CPU以外の要素(メモリ/SSDなど)は空冷
  - この空気についても、上記冷却水との熱交換

- 冷却水ループは25℃以上(通常32℃)であり、KFC同様に超省エネをねらう

高効率なTSUBAME3.0を支える設備

計100tを支える強化された床部

省スペース化

電気配線・冷却水管天井配置を基本とし、設備レイアウトの自由度・床耐荷重を最大化

32度の高温冷却水による冷却

低損失化

420Vの高電圧で配電することにより、配電エネルギーの点損失化と配線コストを低減

床部の強化により高密度なシステムを導入可能に。
1t/m^{2}以上を実現

高密度化

天井部に設置された配管・配線

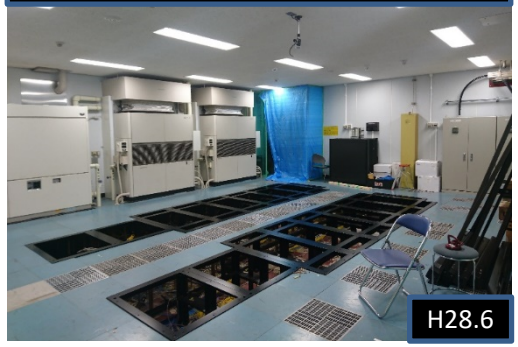
冷却塔によるフリークーリング技術を利用した、高効率な冷却技術の適用

省エネルギー化

420V 3相4線式 配電設備

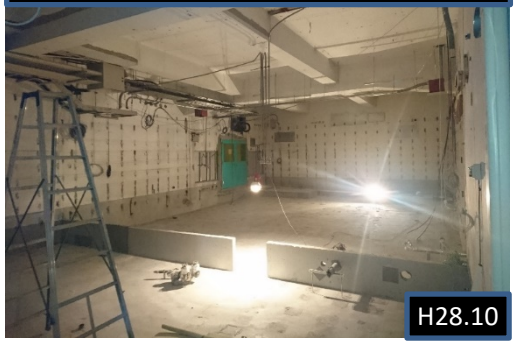
サーバルーム改修の様子

ストレージ撤去後の様子



壁・天井の撤去

撤去後の様子

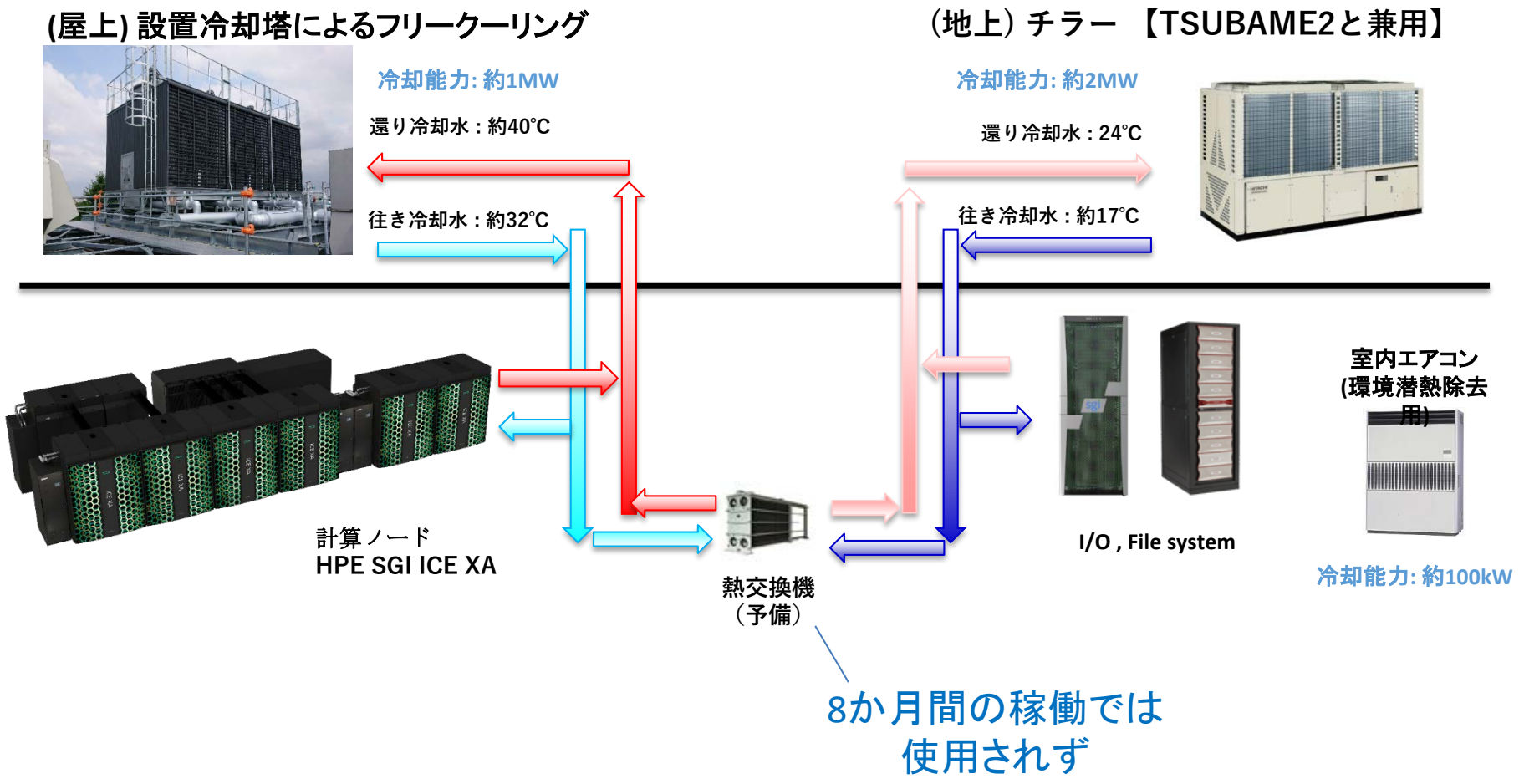


床の撤去

床が撤去され地下室が露出

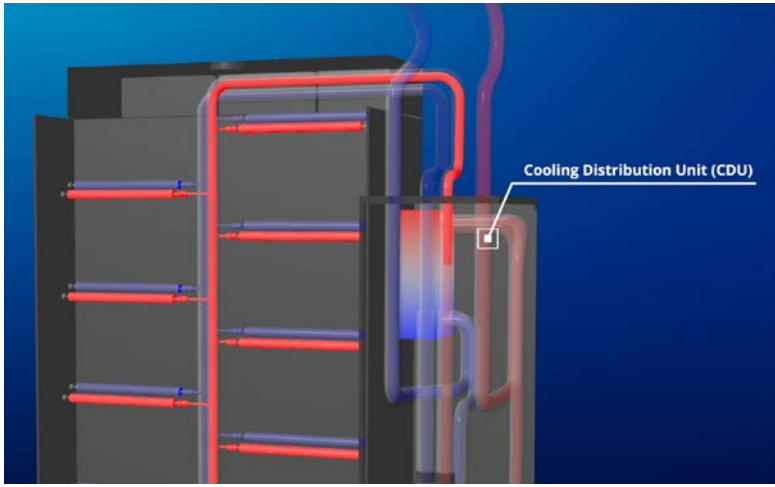
地下室の床の基礎 再構築

TSUBAME3.0の冷却システム

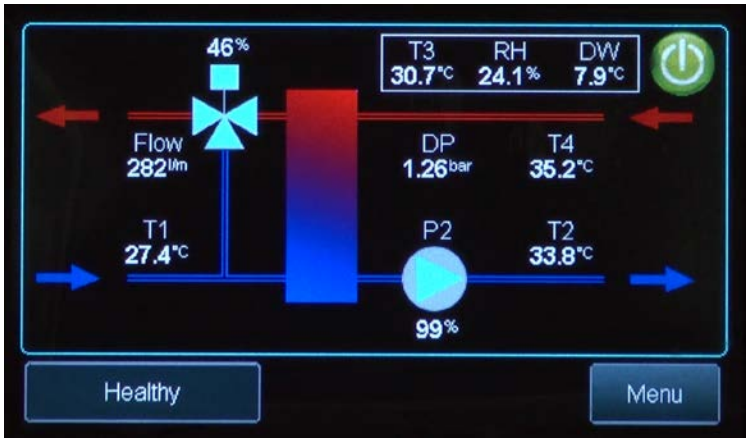


マシン室の冷却水

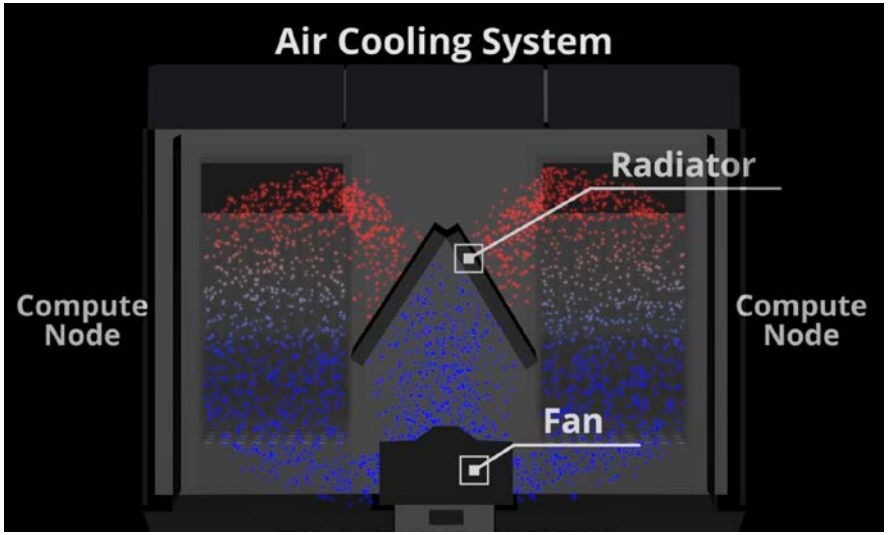
天井に配管、各CDUへ



CDUから計算ラックへ

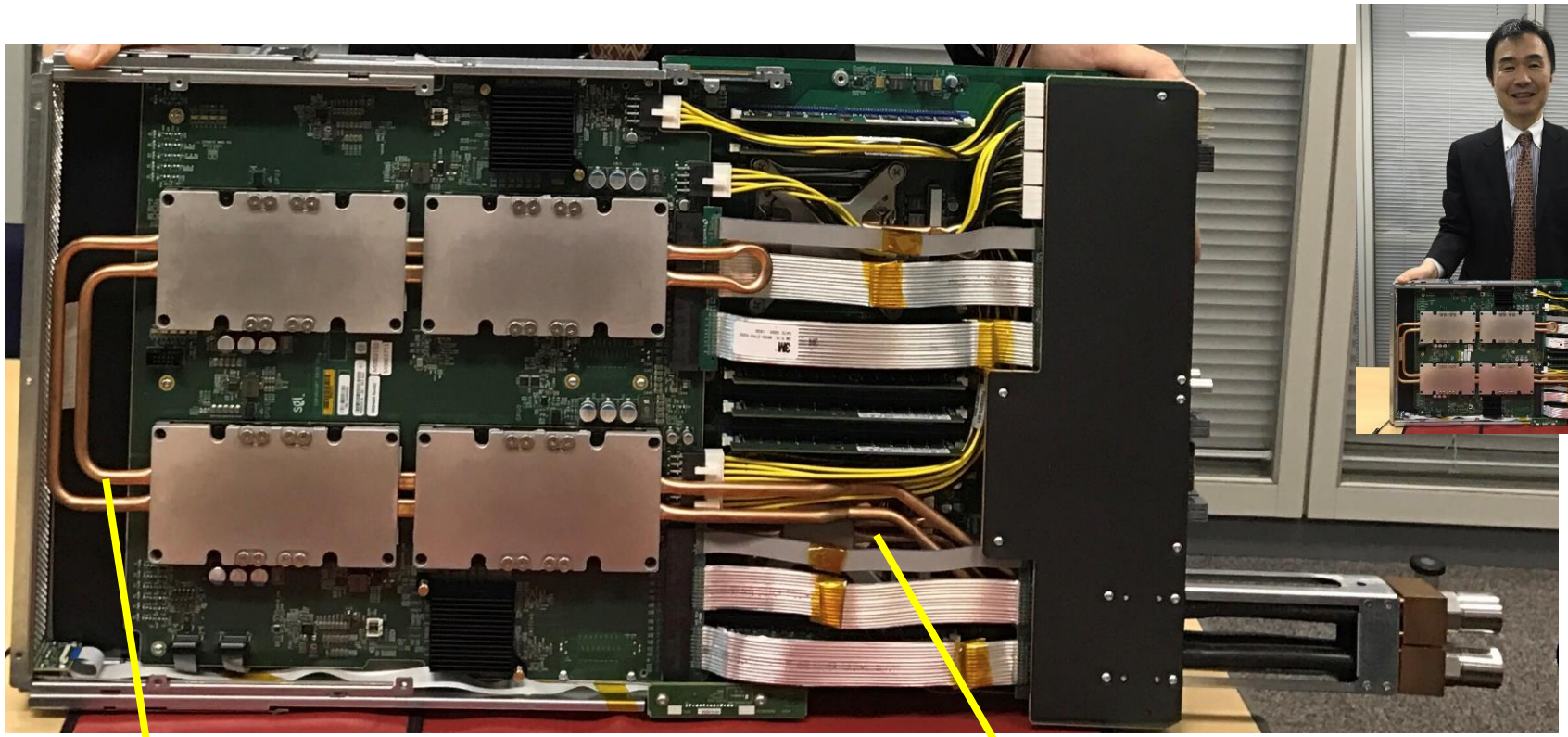


冷却水モニターの様子
いずれの箇所も25℃以上



暖かい冷却水を空冷(メモリ等)にも利用

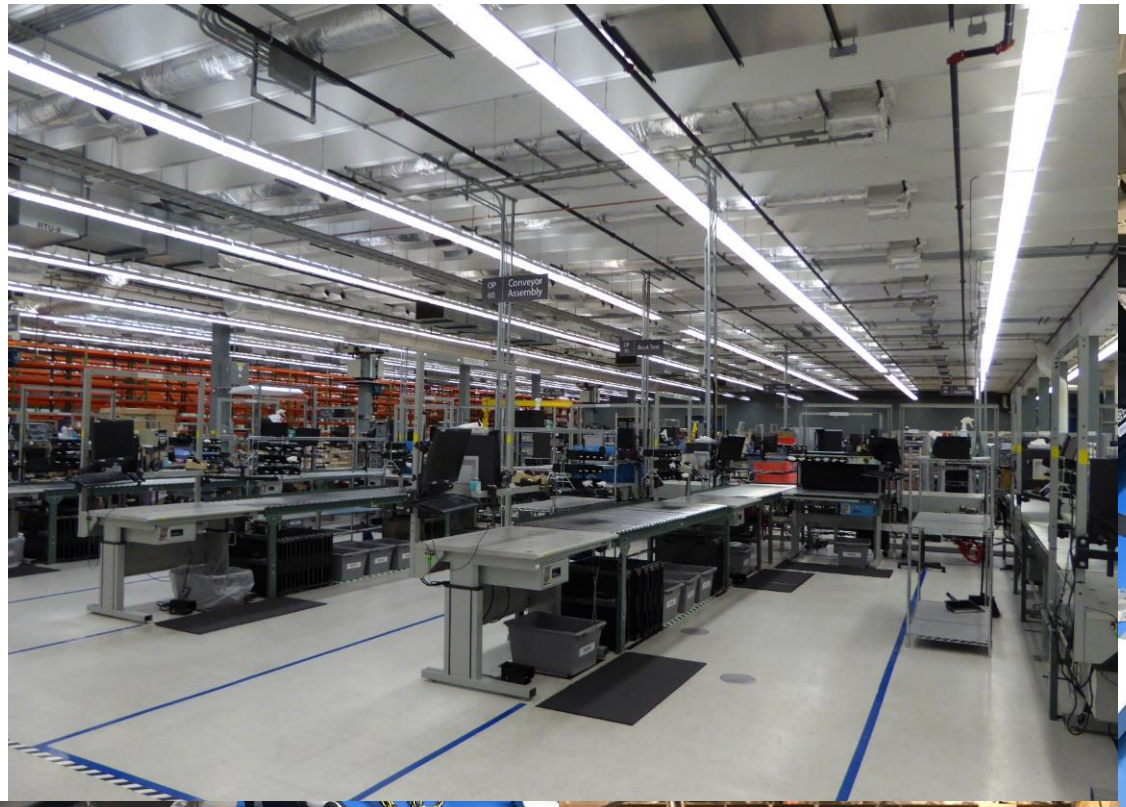
新規設計されたTSUBAME3.0水冷ノード
HPによるさらなる改良後、製品化予定

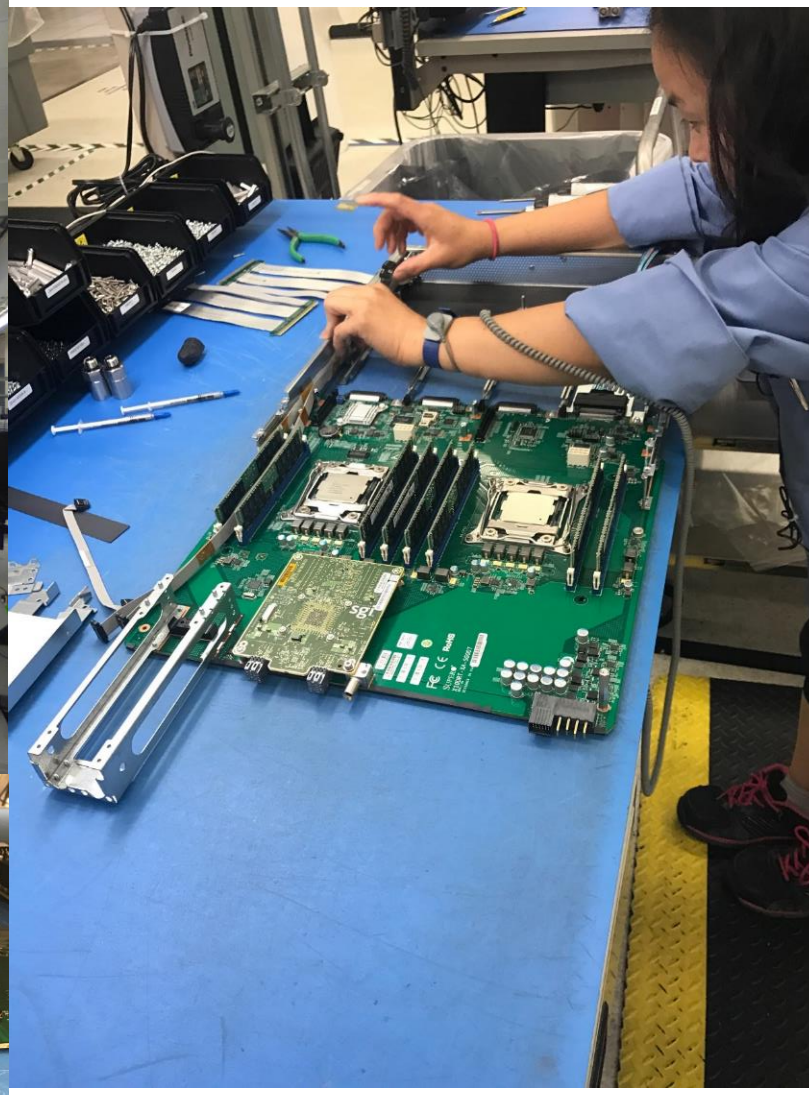


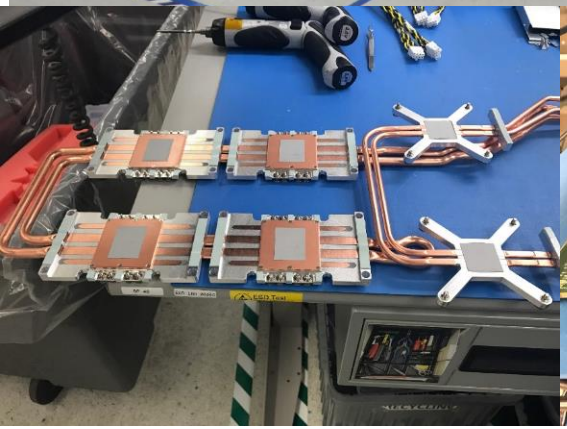
4GPUを通過する
冷却水パイプ

2GPUを通過する
冷却水パイプ(下層)

米国HPE向上での組み立ての様子







おわりに

- TSUBAME2, TSUBAME-KFCでの経験を元に、新しいスパコンTSUBAME3.0の冷却システムを実現した
- 平均500kW、ピーク800kWのスパコンの冷却を省エネに
  - ラックあたり、平均33kW, ピーク53kW程度