AIとHPC融合に向けた TSUBAME4.0スーパーコンピュータ

東京工業大学
遠藤 敏夫

2023/6/22 PCクラスタワークショップ in 大阪2023

東工大TSUBAMEスパコンシリーズ

TSUBAME1.0～1.2
(2006～2010)

TSUBAME2.0/2.5
(2010～2017)

TSUBAME3.0
(2017～)

TSUBAMEシリーズは、世界に先駆けたGPUの採用など、先進的な取り組みを行ってきた

- アジアNo.1 スパコン認定 (2006)
- 世界初大規模GPUスパコン (2008)
- Top500:演算性能世界4位 (2010)
- ACMゴードンベル賞 (2011)
  - Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer
- 文部科学大臣表彰科学技術賞(開発部門)(2012)
  - 運用世界一グリーンペタスパコンの開発
- Green500:省エネ性能世界一 (2017)

東工大TSUBAMEの歴史

2006/4
TSUBAME 1.0
85TFlops/1.1PB

アジアNo.1
「みんなのスパコン」
AMD+ClearSpeed

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アクセラレータ増強
→ Top500で初GPUスパコン

2010/11
TSUBAME 2.0
2.4PFlops/7.1PB

日本初のペタフロップス超え

GPUをアップグレード

2013/9
TSUBAME 2.5
5.7PFlops/7.1PB

2013/9
TSUBAME-KFC
0.2PFlops

液浸冷却による
小型実験スパコン
Green世界一

2017/8
TSUBAME 3.0
12PFlops / 15PB

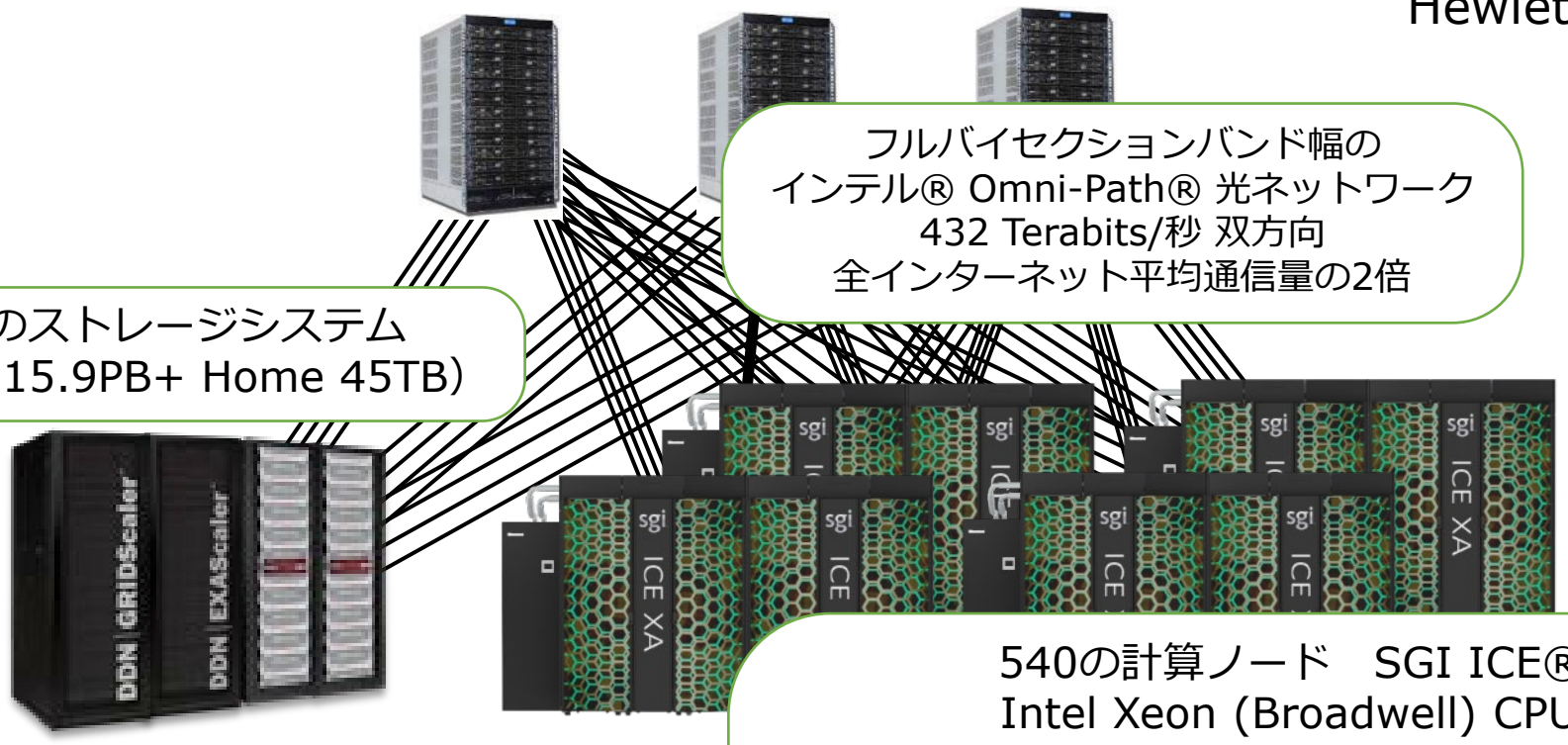
省エネ型「ビッグデータ」スパコン
Green世界一

2024/4 予定
TSUBAME 4.0
66.8PFlops/44PB

現行TSUBAME 3.0 のシステム概要

2017/8~

Integrated by
Hewlett Packard (HPE)



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
Intel Xeon (Broadwell) CPU×２
＋NVIDIA P100 GPU×４
256GBメモリ、２TBのNVMe対応インテル®SSD
12PFlops (倍精度), 47PFlops (FP16)

- ユーザには学内・学外研究者・産業利用を含み、アクティブユーザ数1,400
- 東工大では200近くの研究室が利用
- 深層学習ユーザが大幅増加

TSUBAME3.0の特徴

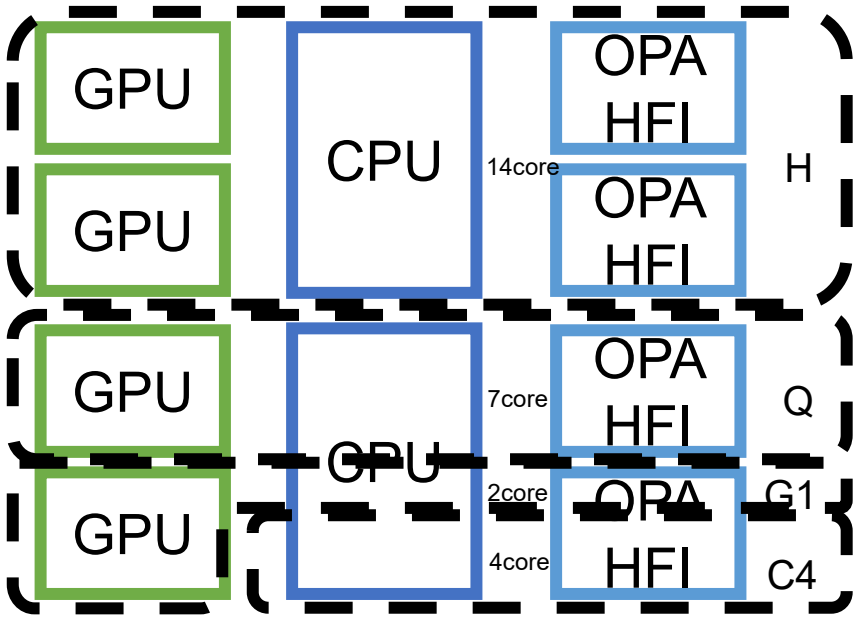
AI/機械学習を強力にサポートするためのTSUBAME3.0の方策：計算科学・シミュレーションとの共存

- ハードウェア面 ➔ **GPU!!**
- ソフトウェア・運用面
  - 機械学習/深層学習フレームワークの提供 (TensorFlow, PyTorch, Chainer…)
  - ↑バージョンアップが早い。ユーザが必要なバージョンを入れられるよう、Singularityコンテナの用意
  - ばらばらな**ジョブの粒度**への対応：>100GPUのMPIジョブもあれば、1コア+1GPUで十分な場合、Pythonスクリプトが動けばよい場合・・・
  - **インタラクティブ利用**需要への対応：バッチスケジューラだけでなく、Webベース利用(2020～)

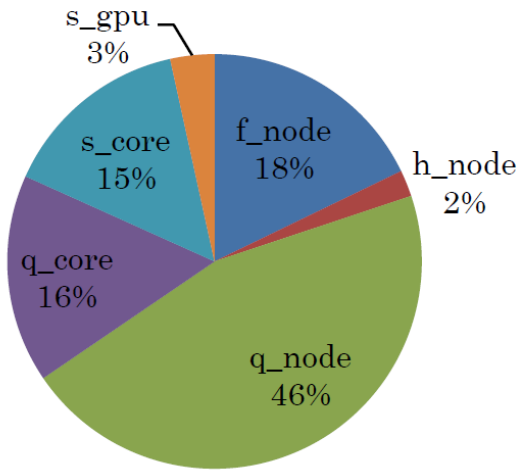
CPU・GPU資源を有効活用しつつジョブ粒度対応

TSUBAME3.0は「ファットノード」：28 core + 4 GPU + 256GB memory

- 「小さい」ジョブにノード丸ごとはもったいない ➔ **ノード分割**の導入
- TSUBAME3上で、複数の「インスタンスタイプ」を定義した
  - GSIC野村准教授(現在)が設計



- f_node: 分割しないノード丸ごと
- h_node: 14 core + 2GPU
- q_node: 7 core + 1GPU
- s_core: 1 Core
- q_core: 4 Core
- s_gpu: 2 Core + 1 GPU



ジョブ件数の統計(2023/3):
>80%が分割タイプ指定

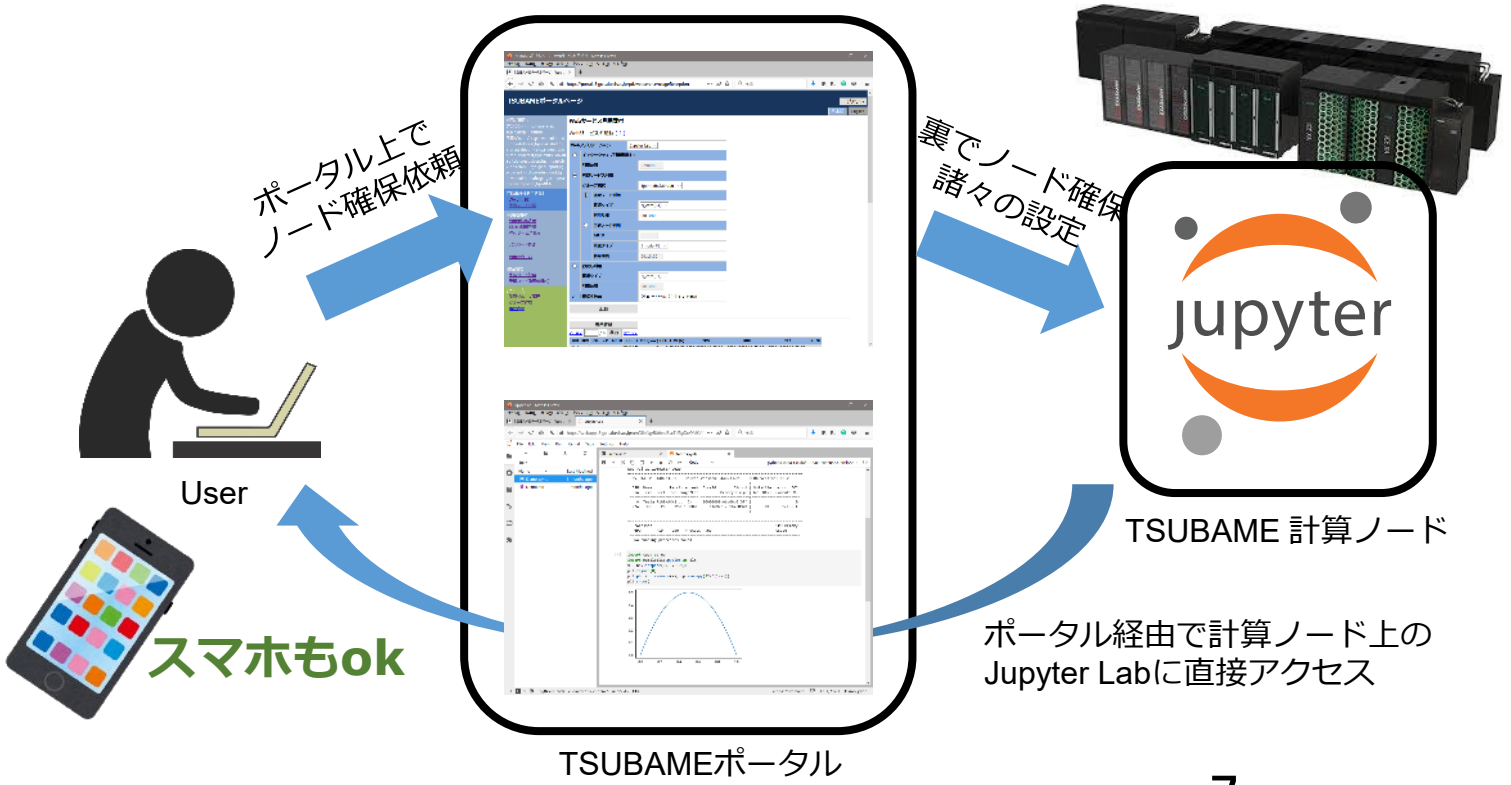
ユーザは、ジョブ投入時に「タイプ」×「インスタンス数」を指定
➔ ユーザは日常的にタイプを選んでいる

2020~

TSUBAME + シングルサインオン+ Jupyter = みんなのビッグデータ・AI/ML基盤 [野村 SWoPP2020]

JupyterNoteBook(当時)をはじめ、Web上での計算資源利用が急激に普及

➔ TSUBAME上にWebインタフェースを実装: JupyterLab / CodeServer / noVNC
- ➔ ブラウザだけでGPUを含めTSUBAMEを直接**インタラクティブ利用**できる
- ➔ 注：OpenOnDemandがメジャーになる少し前だった



この図のどこにもSSH・公開鍵は出てこない→ 覚える必要がない

2022年度に約240名の利用実績
うち1/3 (推定)は、Webのみ利用
➔ スパコンユーザの新規発掘

次期TSUBAMEにおいて、計算性能・使いやすさの両方をさらに向上

7

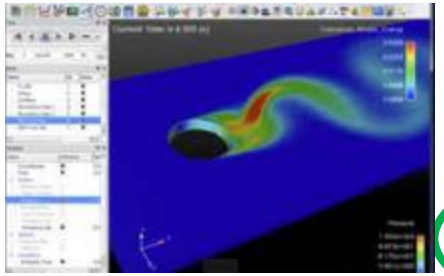
TSUBAME4.0:

データ・計算科学・AI融合のための「もっと」みんなのスパコン により、コンバージェンス・サイエンスの中核インフラへ

対話的データ解析

深層学習との融合による
シミュレーション革新

SNSのフォロー関係解析



シミュレーションと
リアルタイム可視化

深層学習による
画像等認識

HPCI・JHPCNの計算
資源として参加

SINET

TSUBAME4.0
2024/4運用開始予定

大学統合後の
東京科学大学(仮称, 2024秋～)
においても重要な研究教育インフラ

TSUBAME4.0の特徴

- **現行TSUBAME3と比べ、5倍以上の演算速度** ※倍精度・行列演算にて
  - ➔ AI・シミュレーションにおいてさらに増大する計算量への対応
  - ➔ 混雑の緩和へ
- **対話的利用・コンテナ技術の拡充**
  - ➔ビッグデータ解析や可視化を容易化、研究のPDCAを加速
  - ➔ 各ユーザの欲しいソフトウェア環境を迅速に準備
  - ➔待ち時間を短縮するスケジューリングにより、ライトユーザへも恩恵

- **GPUの大幅利用による加速**
  - TSUBAMEシリーズではGPUの利用により、演算速度効率が数倍に
  - ➔ 投資あたりの研究成果の増大

「GPU利用には工夫が必要」という課題に対しては、以下の対応

- 東工大の長年のGPUに関する教育・研究コミュニティの実績
- 深層学習分野ではGPUがデファクトスタンダードになっており、急速に整備

データ・計算科学・AIを中心とした研究成果創出の支援を大幅強化

仕様書時点のTSUBAME4.0の性能

- 総演算性能
  - 倍精度: ≧60PFlops
  - 半精度: ≧ 240PFlops

  } *TSUBAME3.0の5倍以上*
- 計算ノード(少なくとも一部)は、x86互換CPU + CUDA対応GPU
  - 現行機との連続性
  - 単体GPUは、少なくとも以下の性能
    - 倍精度 ≧36TFlops　　※A100は要件を満たさない
    - メモリ容量 ≧64GB
- ストレージ
  - 共有ストレージ ≧38PB
  - 一部、SSDベースの高速共有ストレージ

2023/3/22
日本ヒューレット・パッカードが落札

Hewlett Packard Enterprise

2023/5/18 プレスリリース

TSUBAME4.0性能概略

	TSUBAME3.0	TSUBAME4.0
総演算性能		
• 倍精度演算	12PFlops	**66.8PFlops** (行列演算) **34.7PFlops** (汎用演算)
• 深層学習	47PFlops	**952PFlops (FP16)**
計算ノード数	540ノード (均一)	**HPE Cray XD6500シリーズ** **240ノード (均一)** (縮退時138ノード)
総GPU数	2160個	**960個**
共有ストレージ	DDN社 SFA	**HPE ClustreStor E1000**
• 容量	16PByte	**44PByte** **+ AllFlash 327TByte**

11

TSUBAME4.0ノード構成

	TSUBAME3.0	TSUBAME4.0
CPU	Intel Xeon 2680v4 ×2	**AMD EPYC 9654 ×2**
• 周波数、コア数	2.4GHz, 28コア (=14×2)	**2.4GHz, 192コア (=96×2)**
メインメモリ	DDR3-2400 4ch×2	**DDR5-4800 12ch×2**
• 容量	256GiB	**768GiB**
ネットワーク	OmniPath 100Gbps×4	**InfiniBand NDR 200Gbps×4**
OS	SUSE Linux Enterprise 12	**RedHat Enterprise Linux 8**
GPU	NVIDIA P100 SXM×4	**NVIDIA H100 SXM5 94GB HBM2e ×4**
以下、1GPUあたり		
• 演算性能(倍精度)	5.3TFlops	**66.9TFlops (行列), 33.4TFlops(汎用)**
• メモリ容量	16GB	**94GB**
• メモリ速度	0.73TB/s	**2.39TB/s**

※通常のH100製品(80GB, 3.3TB/s)
とメモリ性能が異なる

搭載予定GPU

NVIDIA H100 SXM5 94GB HBM2e

- ノードあたり4基×240ノード = 960基

	H100 94GBモデル (TSUBAME4)		H100 通常モデル	
• 演算性能(倍精度)	**66.9TFlops (行列), 33.4TFlops(汎用)**	＝	**66.9TFlops (行列), 33.4TFlops(汎用)**	
• 演算性能(FP16)	**990TFlops (行列)**	＝	**990TFlops (行列)**	←疎性算入せず
• メモリ容量	**94GB**	＞	**80GB**	
• メモリ速度	**2.39TB/s**	＜	**3.35TB/s**	
	HBM2e (0.4TB/s?)× 6		HBM3 (0.55TB/s?) × 5	

各種LLMやAlphaFoldなどの大規模モデル学習・推論においてGPUメモリ容量制限の緩和は重要

⇔ メモリ速度低下は残念だが、演算時間の延長でカバーできる

計算ノードのソフトウェアと利用イメージ

>1,000人のユーザが計算資源を共有

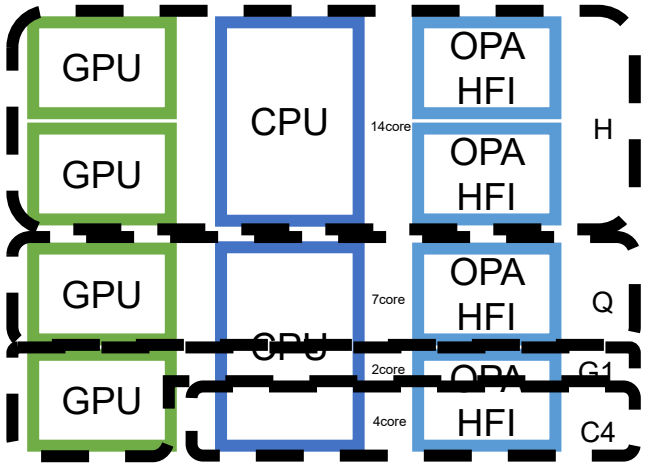
- 計算ノードとジョブの対応を、基本的にはジョブスケジューラが管理する
  - Altair Grid Engine
  - ノード論理分割と組み合わせ
- 計算ノードOSはRedhat Enterprise Linux 8
  - ユーザはコンテナソフトウェアApptainer (旧Singularity)を利用可能
- SSHログインに加え、Webブラウザ利用可能
  - OpenOnDemand検討中

ノード構成バランス

TSUBAME3に比べて、ノード数44%

ノードあたりでは：

- GPU数: 4 ➔ 4 (GPUあたり性能は>12倍)
- CPUコア数: 28コア ➔ 192コア
- 主メモリ: 256GiB ➔ 768GiB
  - コアあたりメモリは 9.1GiB ➔ 4GiB で減少してしまう



ノード論理分割の重要度は増す一方

TSUBAME3の28コアに比べ、TSUBAME4の192コアノードでは、分割制度をより細やかにする必要

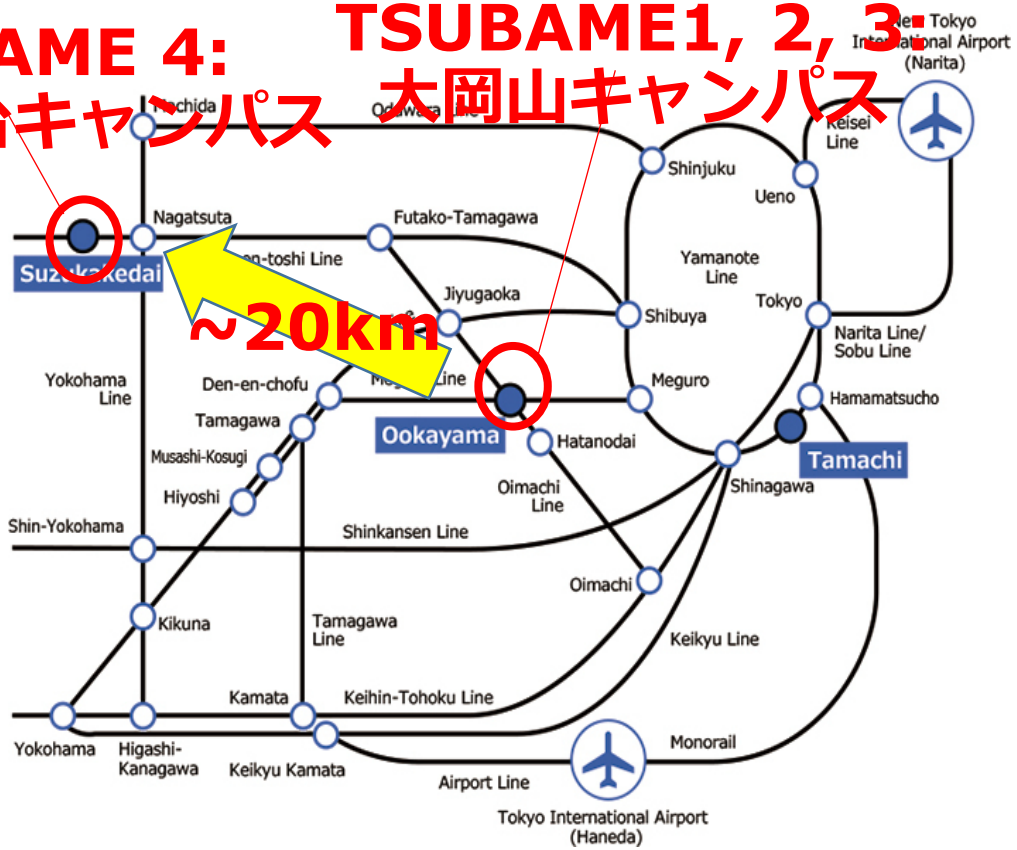
- GPU数も44%になり、「1GPU」がより貴重。システム側でのGPU分割(MIG機能)検討中

このような点含めて鋭意詳細運用設計中

2024/4 TSUBAME4は東工大すずかけ台キャンパスで稼働

TSUBAME 4:
すずかけ台キャンパス

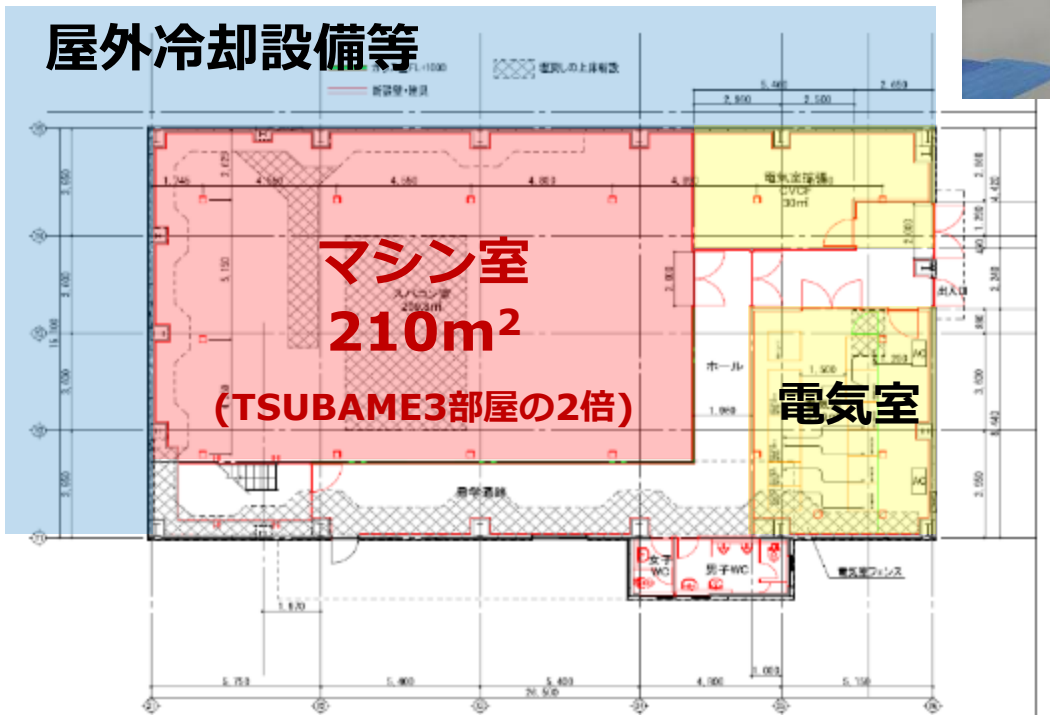
TSUBAME1, 2, 3:
大岡山キャンパス

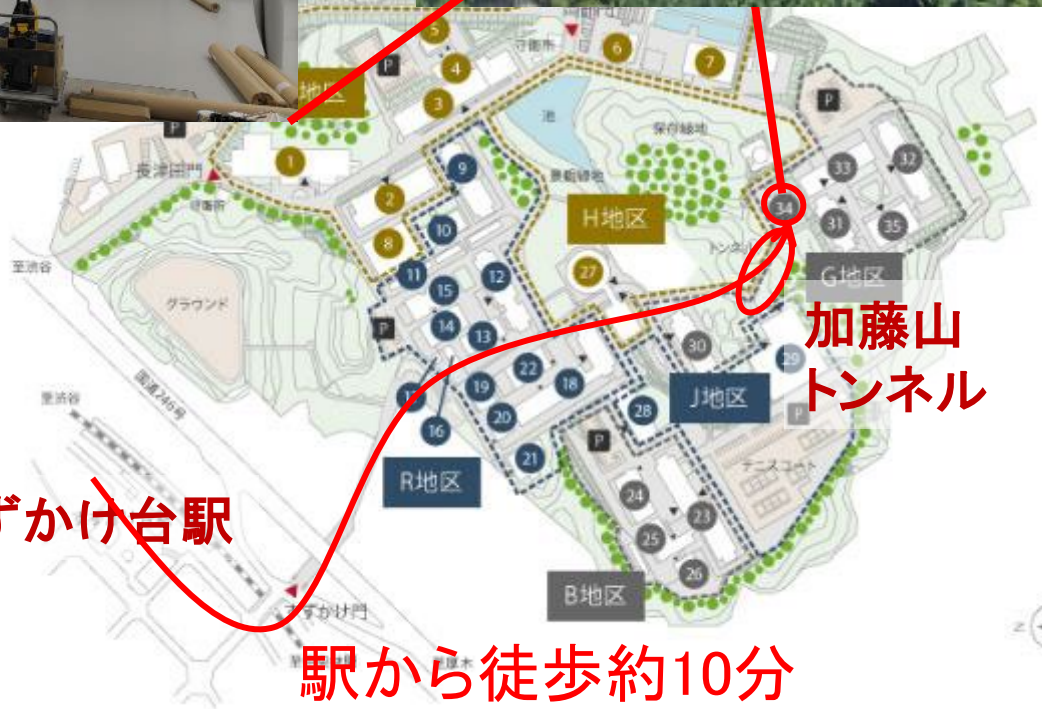


すずかけ台
(横浜市緑区)

2024/4 TSUBAME4は東工大すずかけ台キャンパスで稼働

すずかけ台キャンパスG4A棟(旧MHD棟)を**新データセンター**として改修済





Tokyo Tech

ありがとうございました