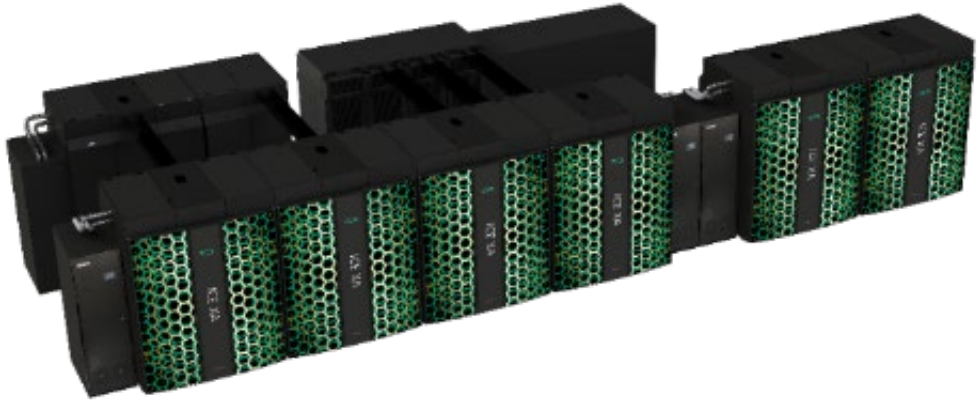
TSUBAMEスパコンの過去、現在、未来

東京工業大学
遠藤　敏夫

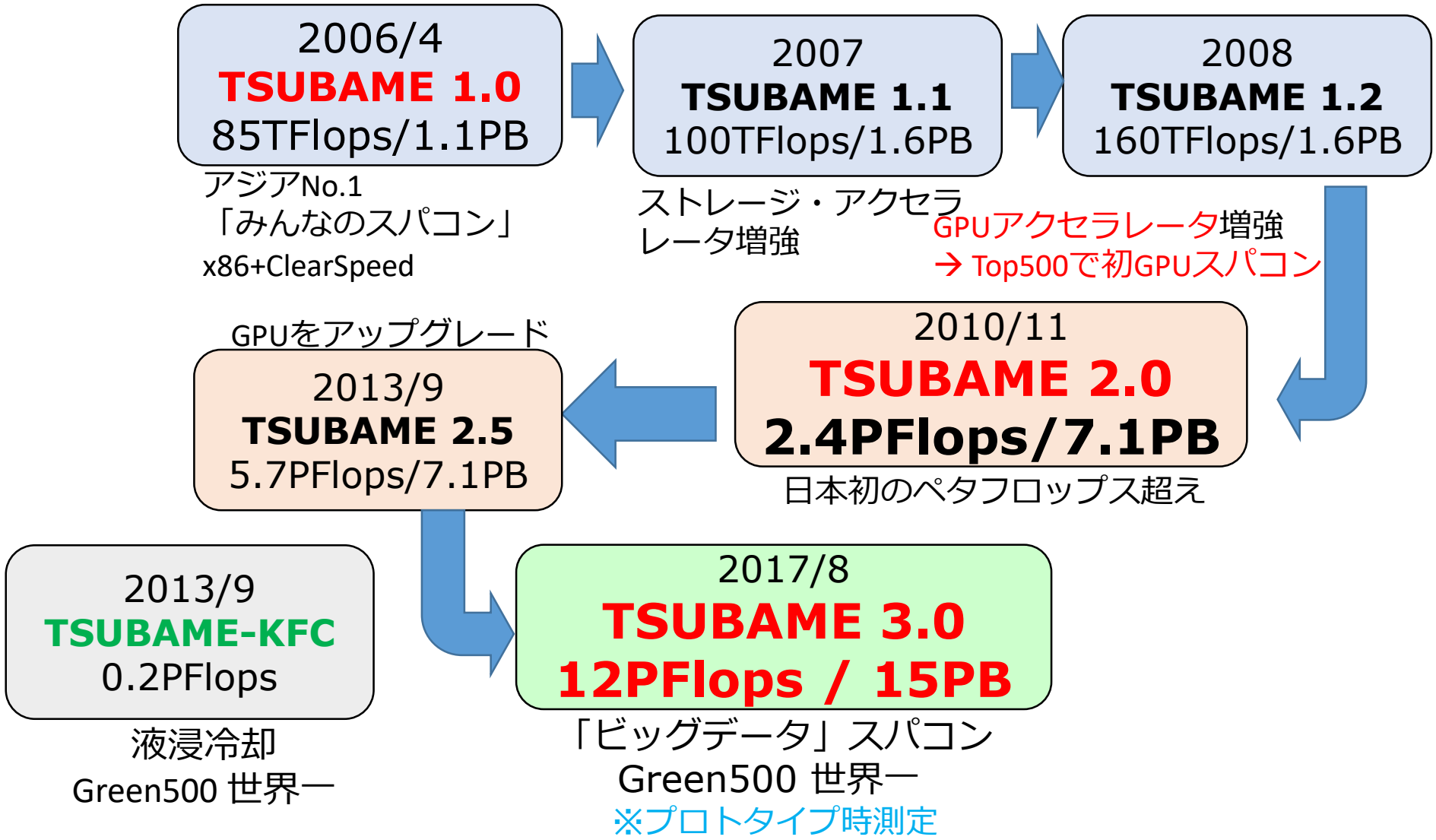
2021年12月9日（木）
PCクラスタシンポジウム

東工大TSUBAMEスパコンの概要

- 東京工業大学学術国際情報センターで運用するスーパーコンピュータ ➔ 現在はTSUBAME3.0
- 当初からのキャッチフレーズは「みんなのスパコン」
- 多分野・学内外（学生も）・産業利用含むユーザ
  - アクティブユーザ1000～1500人
- GPU等のアクセラレータ採用により計算能力向上
- シミュレーション等に加え機械学習・深層学習の利用者増



東工大TSUBAMEスパコンの歴史



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GPUのリーディングスパコンとして世界的にも知られ、次期TSUBAME4.0へも注目

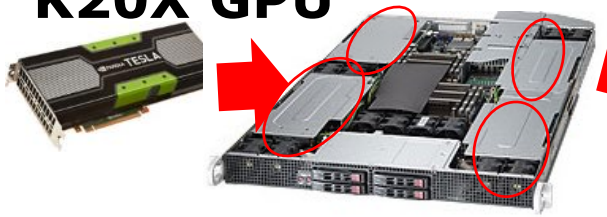
- アジアNo.1 スパコン認定 (2006)
- 世界初大規模GPUスパコン (2008)
- Top500:演算性能世界4位 (2010)
- ACMゴードンベル賞 (2011)
  - Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer
- 文部科学大臣表彰科学技術賞(開発部門)(2012)
  - 運用世界一グリーンペタスパコンの開発
- Green500:省エネ性能世界一 (2017)

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

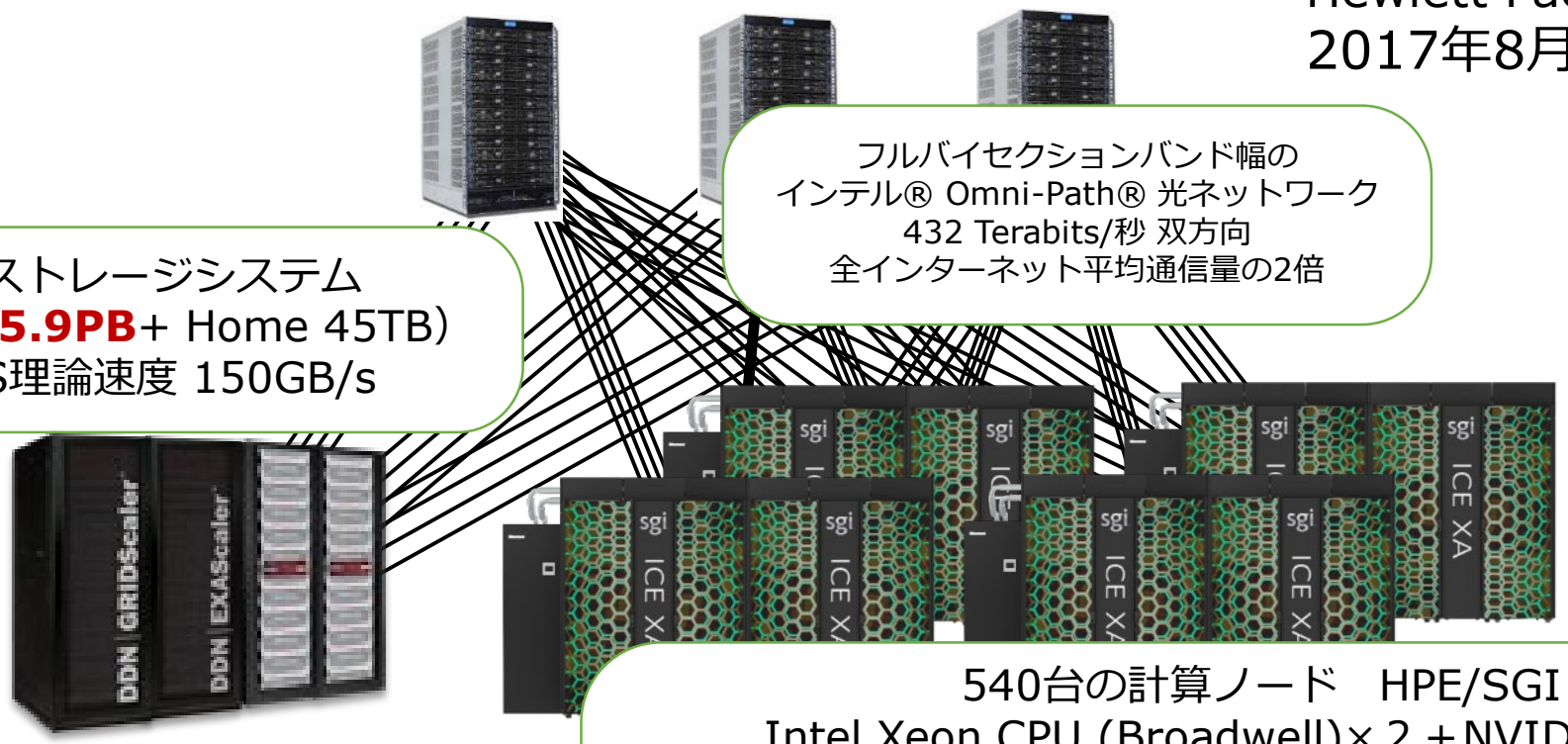
- 世界トップクラスの電力性能比, 3GFlops/Watt以上
- 次世代の超省電力スパコン技術の実証実験

TSUBAME 3.0 のシステム概要

Integrated by
Hewlett Packard (HPE)
2017年8月～~~2022~~年7月
2023年

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ストレージシステム
（並列FS **15.9PB**+ Home 45TB）
並列FS理論速度 150GB/s



540台の計算ノード　HPE/SGI ICE XA
Intel Xeon CPU (Broadwell)×２＋NVIDIA P100 GPU×４
256GBメモリ、２TBのNVMe対応Intel SSD
1台あたり計算能力 22 Tflops (FP64), 87 Tflops (FP16)

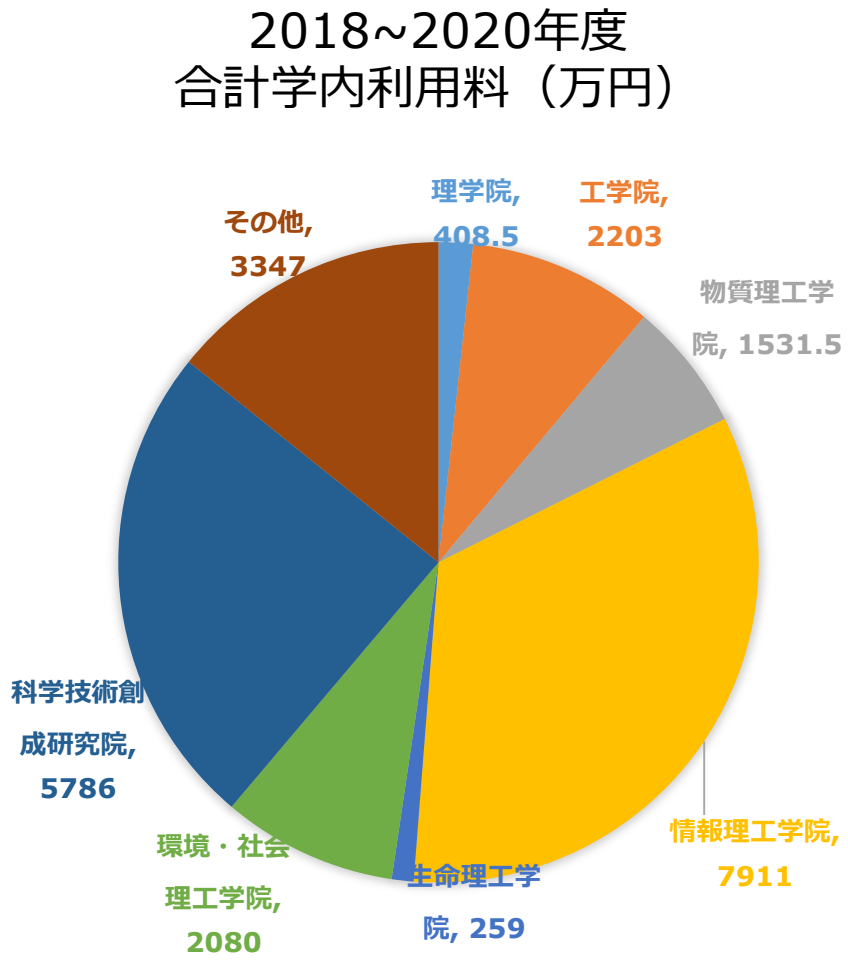
システム消費電力
- 約600kW (平均運用時)

➔ **合計計算能力 12.1 PFlops (FP64), 47.2 PFlops (FP16)**

多種シミュレーション

深層学習

広分野で利用されるTSUBAME



東工大の6学院+研究院のすべてから有償利用されている

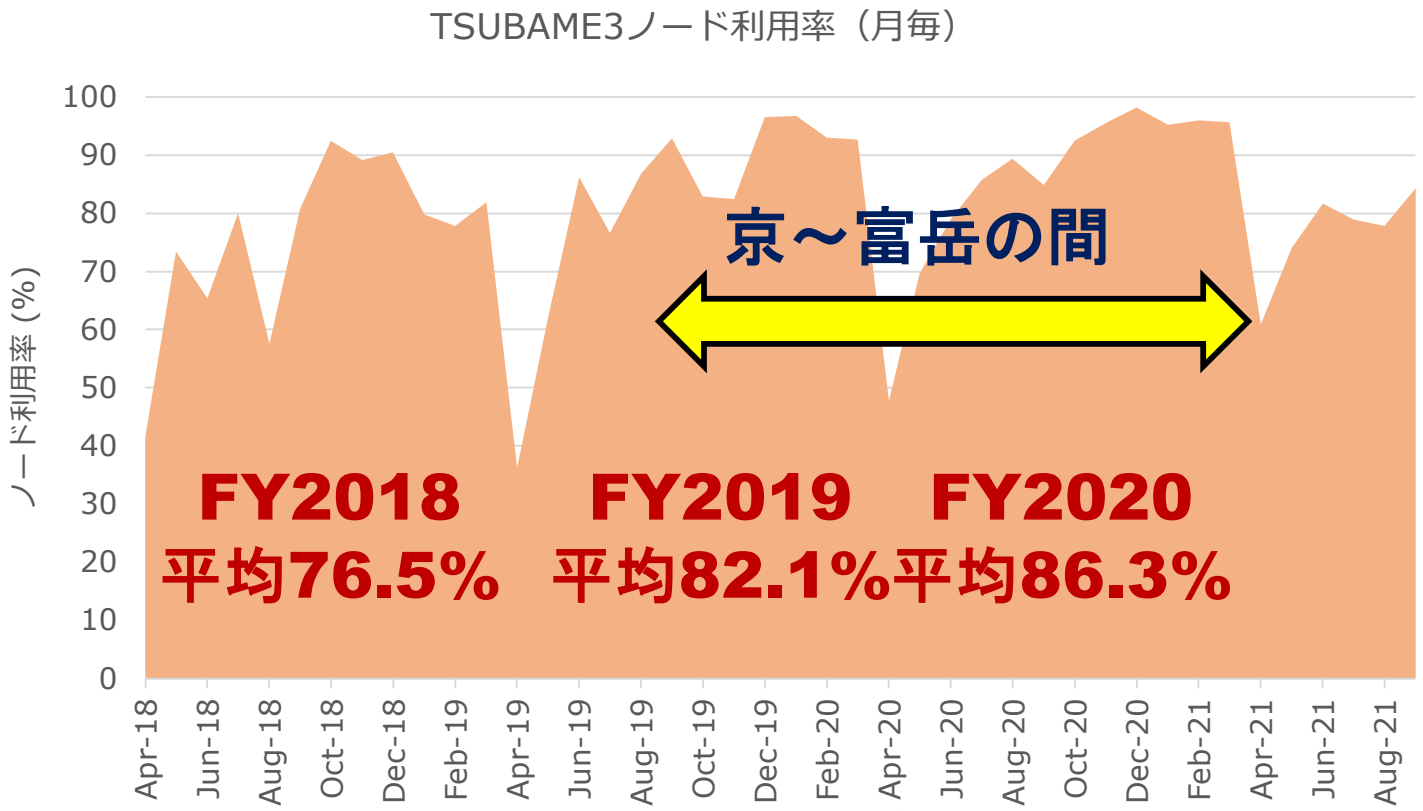
東工大内の利用例

- 創薬の大規模分子シミュレーション(COVID-19の阻害剤など)
- 創薬手法開発と治療薬探索
- 素粒子の量子ダイナミクス研究
- 高イオン伝導性固体電解質の探索
- 深層学習による自然言語処理
- 深層学習による音声・動画認識

：

高いTSUBAMEの利用率

アカウント数
約5600
うち、アクティブ
約1200



※ノード単位の利用率なので、コア単位ではもう少し空きあり

- 高い利用率 ➔ 運用効率はよいが、ユーザにとっては長い待ち時間
  - 特に2020年度後半は>95%
- 「次の」計算資源も重要だが、運用による影響軽減も必要

TSUBAME3.0の運用

- OS: SLES12 + ジョブスケジューラ: Univa Grid Engine

- PyTorch, TensorFlowなどをプレインストール、しかしバージョン進展速い

➔ Singularityコンテナにより、ユーザの希望する環境を利用可能

- TSUBAMEノード(28CPUコア+4GPU)はファット

➔ ノードを動的分割してスケジュール対象

- ほかにも、利用を容易にする仕組みを中途追加

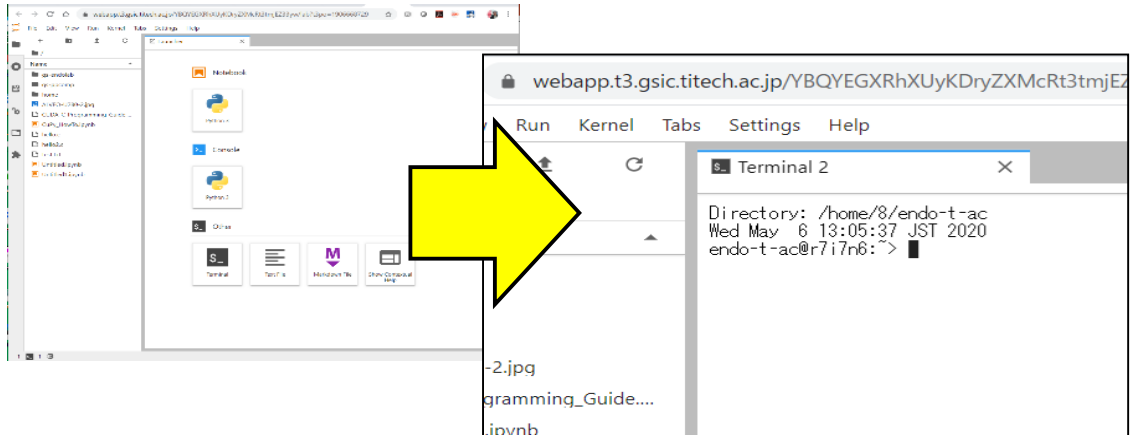
2つの運用改善の合わせ技で「みんなのスパコン」強化へ

[改善1] インタラクティブ専用キュー

- TSUBAMEの540台の計算ノードから4台(**<1%**)をインタラクティブジョブ専用に切り出し、別キューとして提供
  - ¼ノード(7コア・1GPU・64GiB実メモリ)を最大7人が共有
    - 同時実行 4 x 4 x 7 = 112ジョブ, 実習系講義が1クラス分収まるレベル

[改善2] Webブラウザ利用

- TSUBAME利用者ポータルでユーザ認証ができる ➔ ノードログインに使ってもよいことに
- Jupyter Lab (コマンドプロンプトあり), Code Serverを提供中 ➔ SSHを知らなくてもGPUを使ったデータ分析可能



TSUBAME4.0:

データ・科学・AI融合のための「もっと」みんなのスパコン

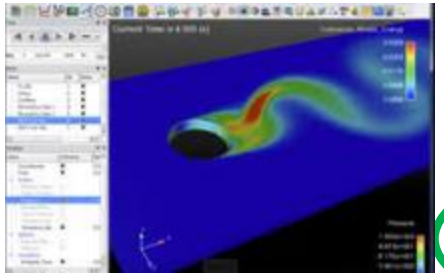
対話的データ解析

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

深層習学との融合による
シミュレーション革新
Ex) AlphaFold2

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

ビッグデータ解析

計算科学・
シミュレーション

AI・深層学習

ドライブレコーダデータ
に基づく深層学習

TSUBAME4.0
(イメージ図)

2023/8～2029/7 運用予定

TSUBAME4.0の特徴

- **現行TSUBAME3と比べ、約6倍の演算速度** (倍精度70～90PFlops程度)
  - ➔ 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等の大幅利用による加速**
  - TSUBAMEシリーズではGPU等の利用により、演算速度効率が数倍に
  - ➔ 投資あたりの研究成果の増大

一方、「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

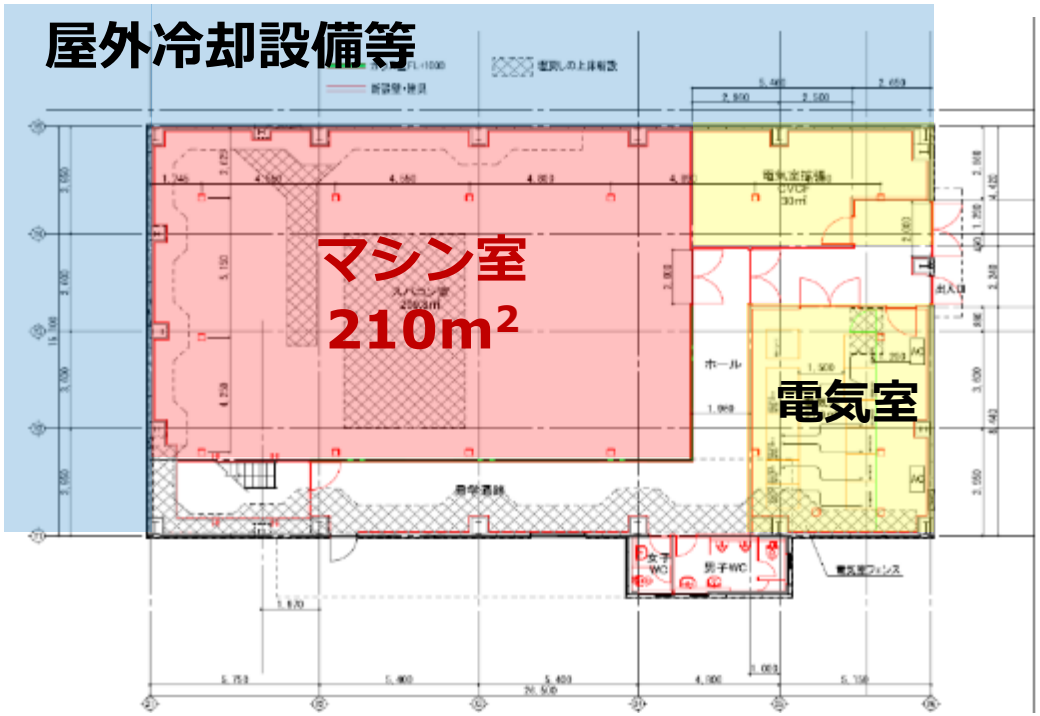
データ・計算科学・AIを中心とした研究成果創出の支援を大幅強化

ストレージの階層構造？ユーザにどのようなビューを見せる？

TSUBAME4.0は東工大すずかけ台キャンパスへ

TSUBAME1~3は大岡山キャンパス・GSIC情報棟に設置

TSUBAME4: **すずかけ台キャンパスG4A棟**(旧MHD棟)
を**新データセンター**として改修後に設置予定



ありがとうございました

← スライド表紙はすずかけ台キャンパスの画像でした
（建物は別）