東工大
TSUBAME3.0の概要

東京工業大学　学術国際情報センター　教授
産業技術総合研究所 人工知能研究センター
特定フェロー

松岡　聡

2017/2/17 記者会見用資料

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

- 大規模なGPU採用による高性能と低電力の両立
- 最小の設置面積(200m2程度)、高いコストパフォーマンス
- 高性能にマッチした光ネットワーク、SSDストレージ

各種基礎研究がベース
メーカーと新規共同開発

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS
32nm

GPUs(Tesla M2050)
515 GFLOPS
3 GB 40nm

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB
>400GB/s Mem BW
80Gbps NW BW
~1KW max

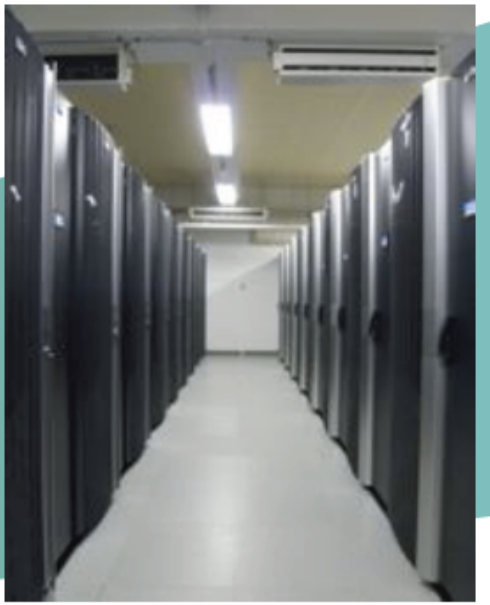
Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB
>1.6TB/s Mem BW

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB
>12TB/s Mem BW
35KW Max

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



2.4 PFLOPS
80 TB
>600TB/s Mem BW
220Tbps NW
Bisecion BW
1.4MW Max

Integrated by NEC Corporation

TSUBAME2.0のアプリケーションの受賞


SC11

ACM Gordon Bell Prize

Special Achievements in Scalability and Time-to-Solution

Takashi Shimokawabe, Takayuki Aoki,
Tomohiro Takaki, Akinori Yamanaka,
Akira Nukada, Toshio Endo,
Naoya Maruyama, Satoshi Matsuoka

Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

IEEE COMPUTER SOCIETY

ACM Gordon Bell Prize 2011
ACM ゴードンベル賞
2.0ペタフロップス達成
(京コンピュータと同時受賞)

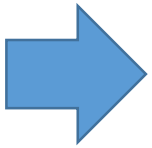
Special Achievements in Scalability and Time-to-Solution

“Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer”

TSUBAME2.0⇒2.5 計算ノードの進化(2013/9)

- 全4224GPUを最新のKepler GPUにほぼ運用中断なく交換
- 幾つかの技術上・運用上の問題をメーカーと共同で克服
- 低コスト・短期間でマシンの能力を2-3倍に向上に成功

NVIDIA Fermi
M2050
1039/515GFlops
3GBメモリ



NVIDIA Kepler
K20X
3950/1310GFlops
6GBメモリ

TSUBAMEと京との比較(1)

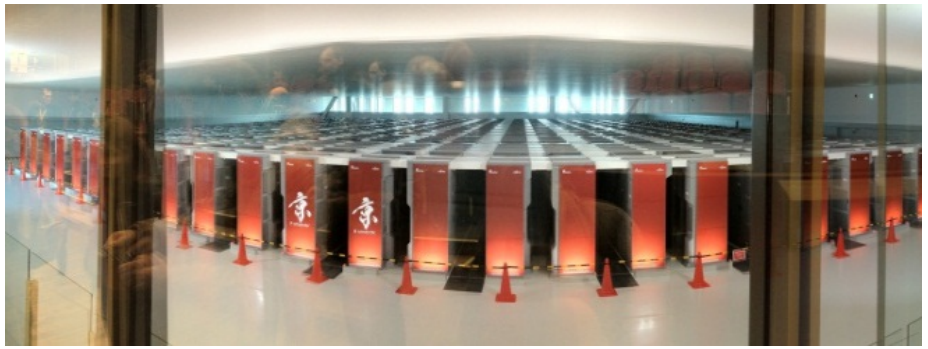
東京工業大学
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

TSUBAME2.0(2010)
→ TSUBAME2.5(2013)
単精度17.1 Petaflops(最速)
倍精度5.76 Petaflops
約50億円/6年 (電気代等含)

性能≒
コストくく

独立行政法人理化学研究所
計算科学研究機構
RIKEN Advanced Institute for Computational Science

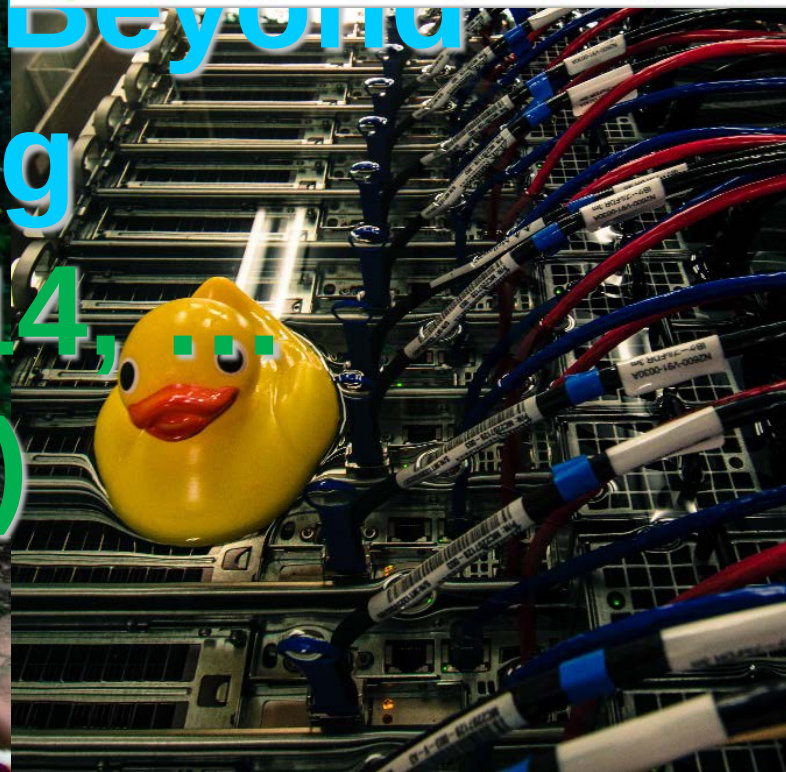


京コンピュータ (2011)

単精度11.4Petaflops
倍精度11.4Petaflops(最速)

約1500億円?/6年
(電気代等含)

TSUBAME-KFC
Towards TSUBAME3.0 and Beyond
Oil-Immersive Cooling
#1 Green 500 SC13, ISC14, …
(Paper @ ICPADS14)



TSUBAME-KFC: ウルトラグリーン・スパコン研究設備
(文部科学省概算要求・2011-2015・約2億円)

液浸冷却＋高温大気冷却＋高密度実装＋電力制御のスパコン技術を統合
TSUBAME3.0のプロトタイプ

高密度実装・油浸冷却
210TFlops (倍精度)
630TFlops (単精度)
1ラック

2015アップグレード
500TFlops (倍精度)
機械学習等1.5PFlops (単精度)
世界最高性能密度
(やはり1ラック=>7ラックで京相当)

高温冷却系
冷媒油 35~45℃
⇒ 水 25~35℃
(TSUBAME2は7~17℃)

冷却塔:
水 25~35℃
⇒ 自然大気へ

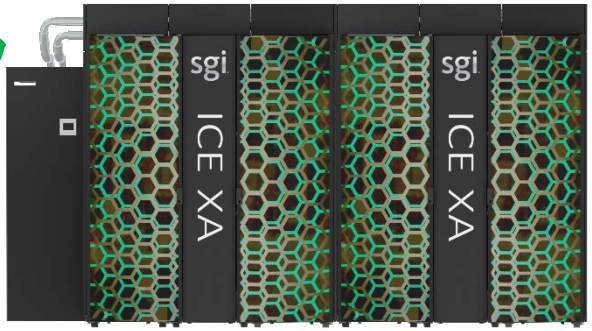
コンテナ型研究設備
20フィートコンテナ(16m²)
無人自動制御運転

将来はエネルギー回収も

2013年11月/2014年6月
Green500ランキング
世界一位(日本初)

2017年 リーディングスパコンTSUBAME3.0へ
「最先端の技術チャレンジに挑むスパコン」

1. HPCIリーディングスパコン：TSUBAME2と合算し京の二倍の性能：15-20ペタフロップス、4-5ペタバイト/秒メモリバンド幅、ペタビット級光ネットワーク
2. ビッグデータスパコン：大規模シリコンストレージによりマルチテラバイト/秒のI/O、機械学習・AI含むビッグデータ用ミドルウェアやアプリ
3. グリーンスパコン10ギガフロップス/W以上の性能電力性能・グリーン・PUE < 1.05
4. クラウドスパコン：種々のクラウドサービスと高速性の両立
5. 「見える化」スパコン：超複雑なマシンの状態が「見える」

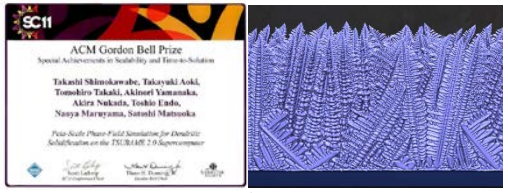


2017年 TSUBAME3.0
12.1ペタフロップス
グリーン・ビッグデータ
HPCIリーディングスパコン

2013年TSUBAME2.5
アップグレード(補正予算)
5.7ペタフロップス

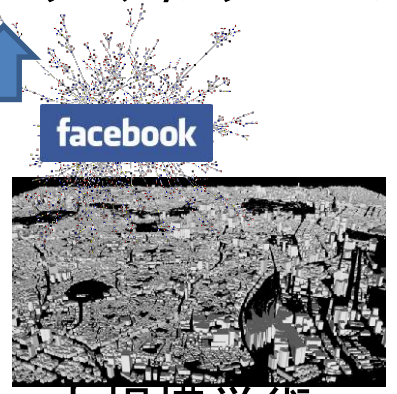
2010年 TSUBAME2.0
2.4ペタフロップス・世界4位
運用スパコングリーン世界一

2006年 TSUBAME1.0
80テラフロップス・アジア一位
世界7位



2011年ACMゴードンベル賞

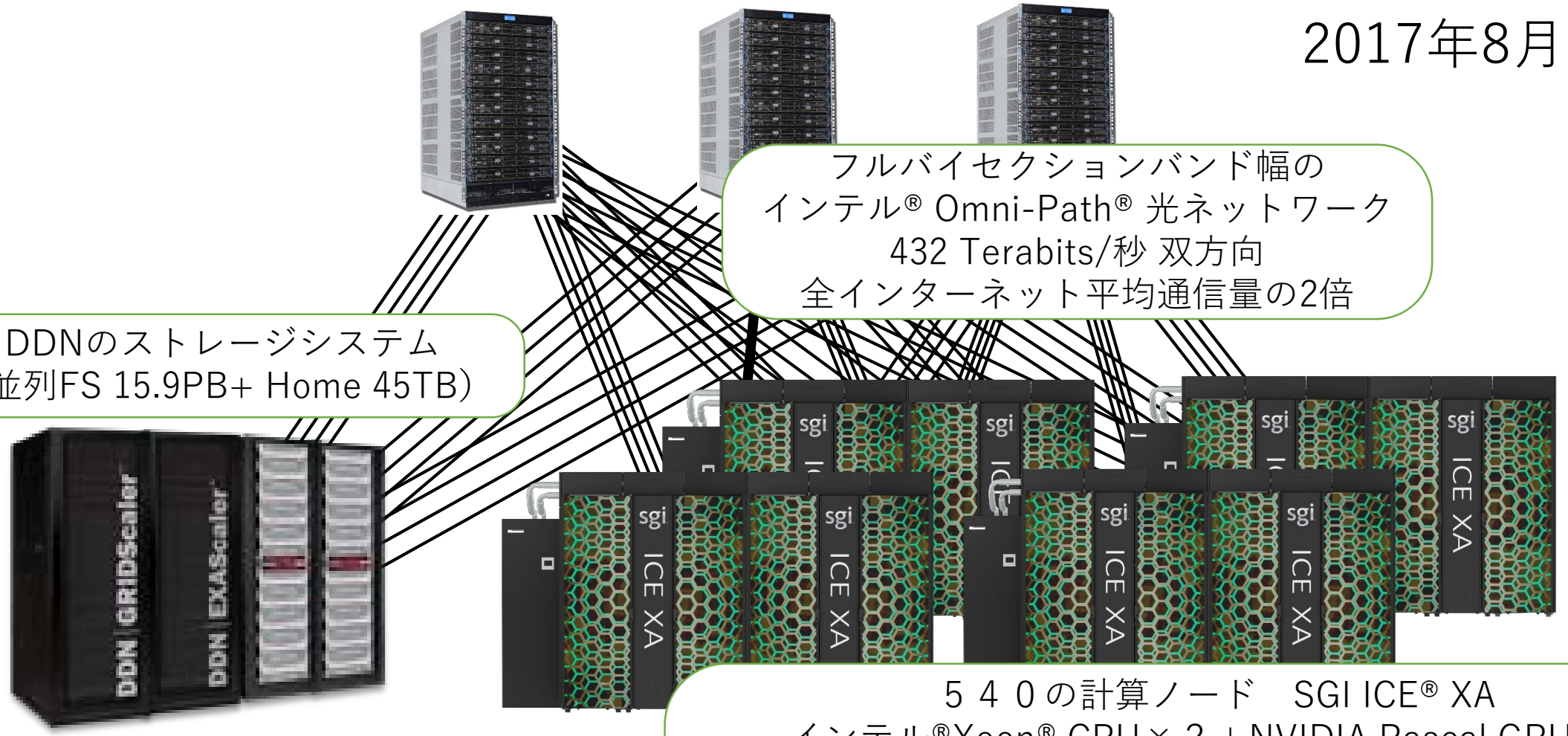
2013年TSUBAME-KFC
スパコングリーン世界一



大規模学術
及び
産業利用アプリ

TSUBAME 3.0 のシステム概要

2017年8月本稼働



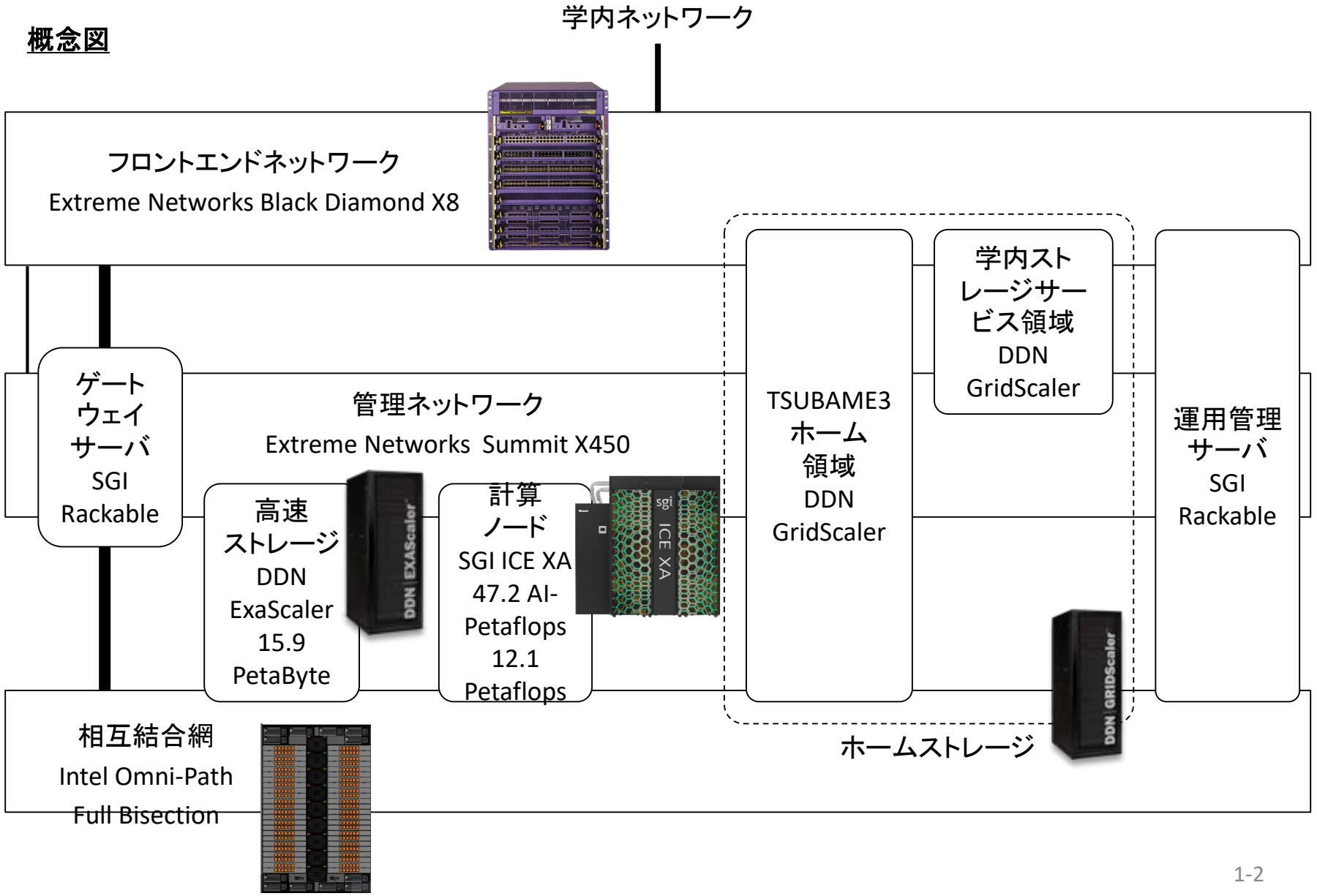
フルバイセクションバンド幅の
インテル® Omni-Path® 光ネットワーク
432 Terabits/秒 双方向
全インターネット平均通信量の2倍

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

TSUBAME 3.0 全体構成

概念図



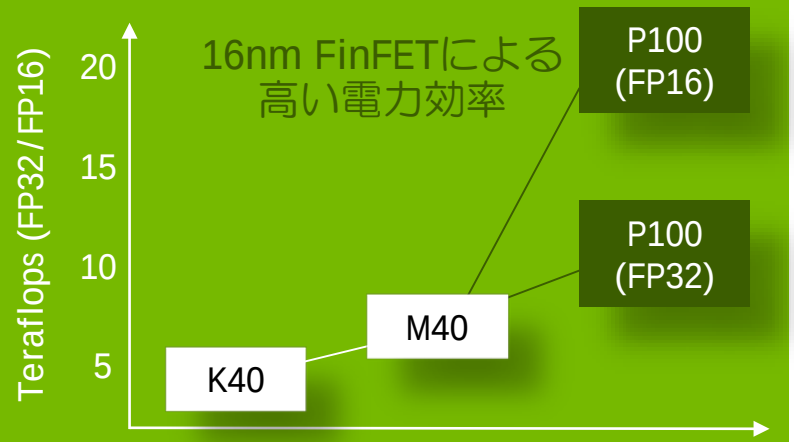
NVIDIA TESLA P100 “PASCAL” GPU

5つの技術革新

TSUBAME3では2160機採用

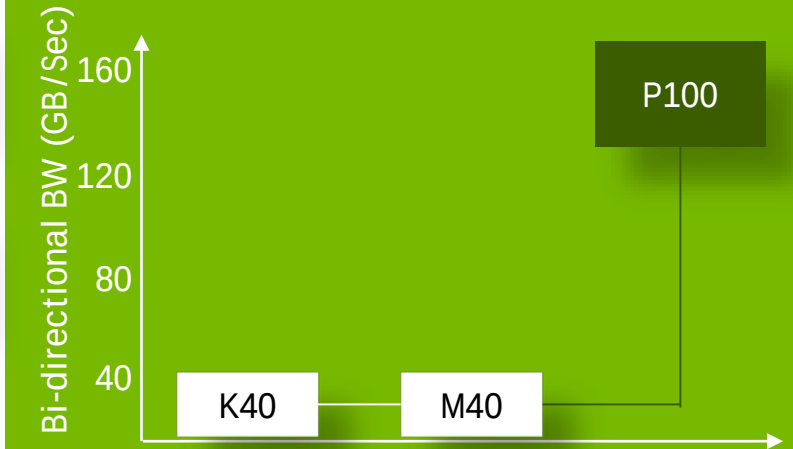
HPC&AI高性能演算性能

21 Teraflops of FP16 深層学習性能
5.3 Teraflops FP64 HPC



NVLINKによる高速GPU間通信

PCI-e 3 比で5x GPU-GPU通信バンド幅

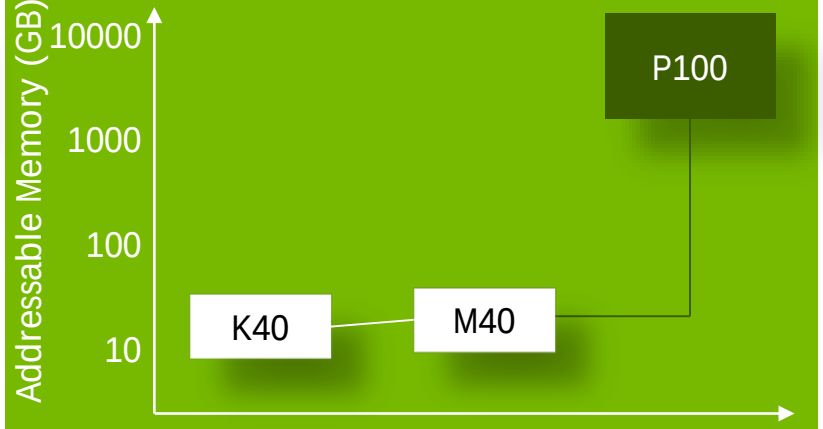


HBM2 2.5次元積層メモリ

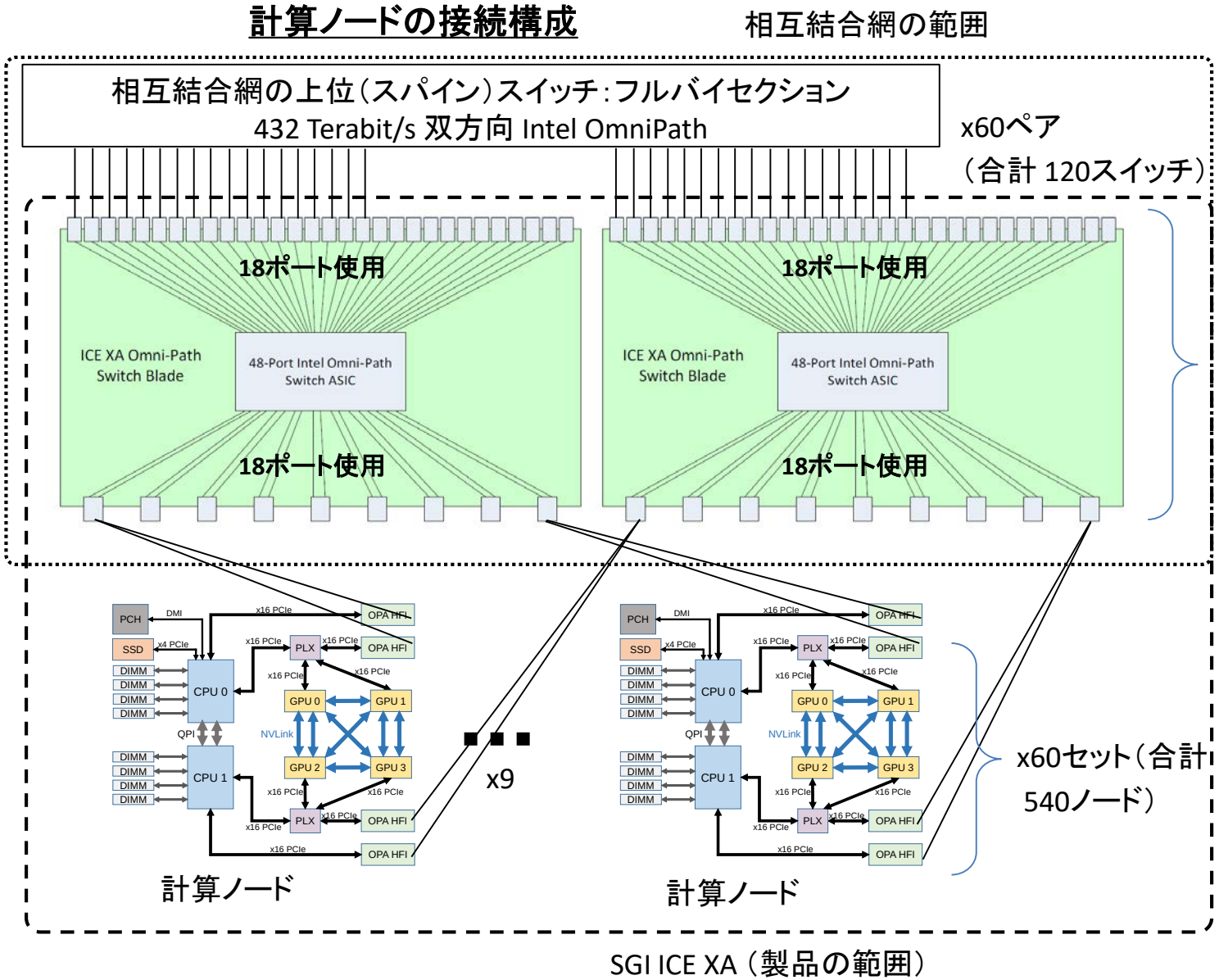
大規模HPC&データのワークロードにて
3倍のメモリバンド幅

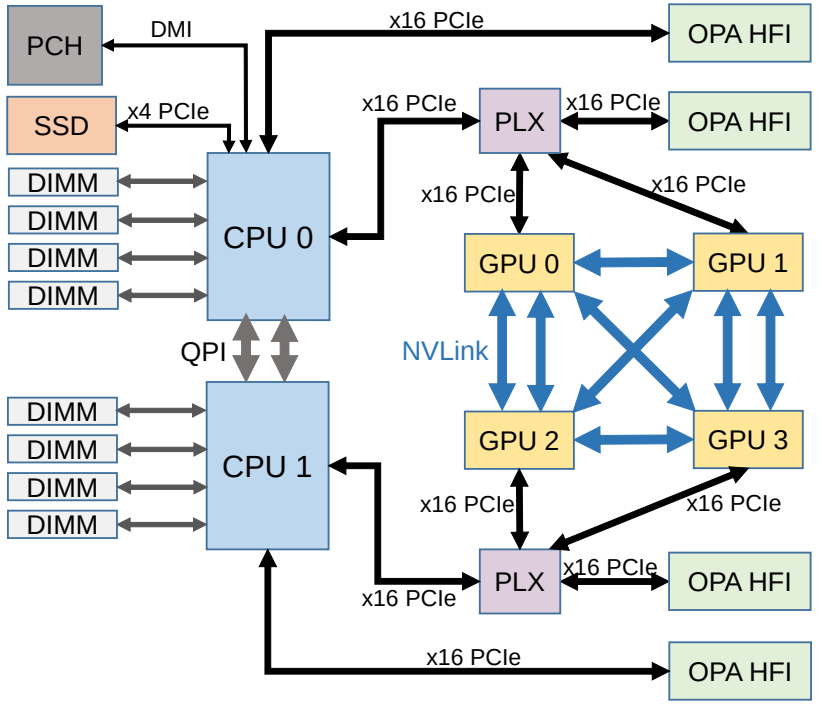
ハードウェアによるCPUとのメモリ共有

GPUのメモリ制限を撤廃



TSUBAME3.0 計算ノード（Ⅱ）SGI ICE-XA ベース、ノードはSGI社と東工大GSICの共同設計によるTSUBAME3.0用新設計





超高性能な”Fat Node”構成(TSUBAMEの伝統)

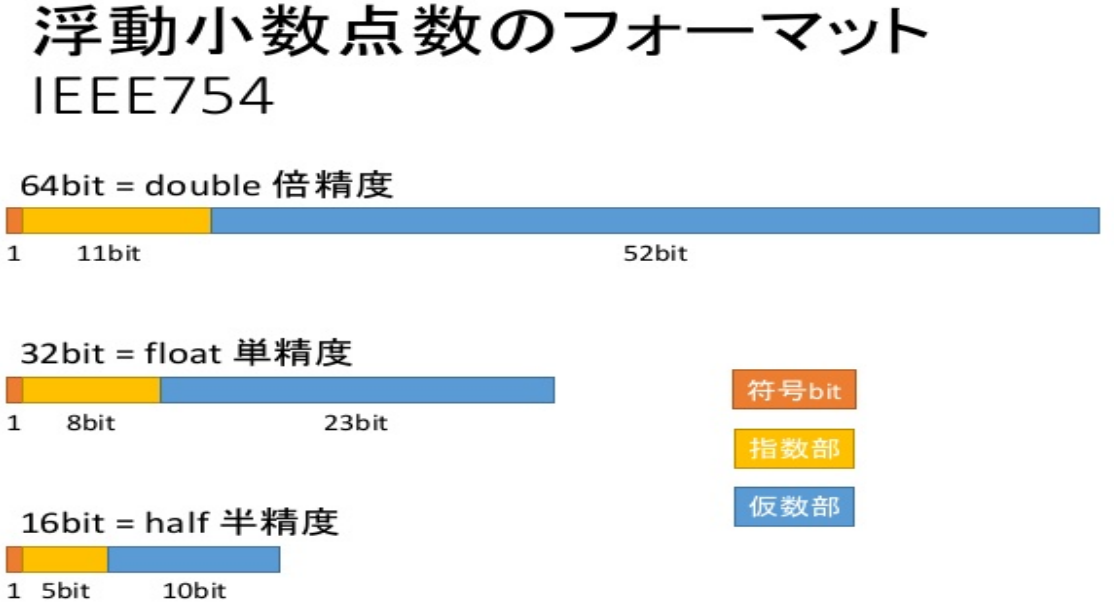
- 4 SXM2(NVLink) NVIDIA Pascal P100 GPU
- 高ネットワークバンド幅 – Intel Omnipath ネットワーク 100GBps x 4 = 400Gbps
- 高 I/O バンド幅- Intel 2 TeraByte NVMe
  - システム合計容量1ペタバイト以上、2Terabyte/秒の合算バンド幅
- 高密度・高温水冷ブレード – 1ラックあたり 36 ブレード = 144 GPU + 72 CPU、50-60KW
  - 一般データセンターの10-20倍

浮動小数点：倍精度、単精度、半精度

(Floating Point Precisions: Double, Single, Half)

- 科学技術や機械学習で使われる「実数」の演算
  - 例小数：3.1415926…、小さい数字：.000000…0001、大きい数字：999999…999999999
- その表現法:IEEE 754 浮動小数点規格
  - 倍精度 (double precision):8バイト、10進で約16ケタ
  - 単精度 (single precision): 4バイト、10進で約7ケタ
  - 半精度(half precision): 2バイト、10進で約3.3ケタ

- 科学技術計算では、従来は倍精度が重視されてきたが、近年は単精度を利用しての高速化が増加
- 機械学習/AIでは、単精度が主流で、近年は半精度が用いられ始めている



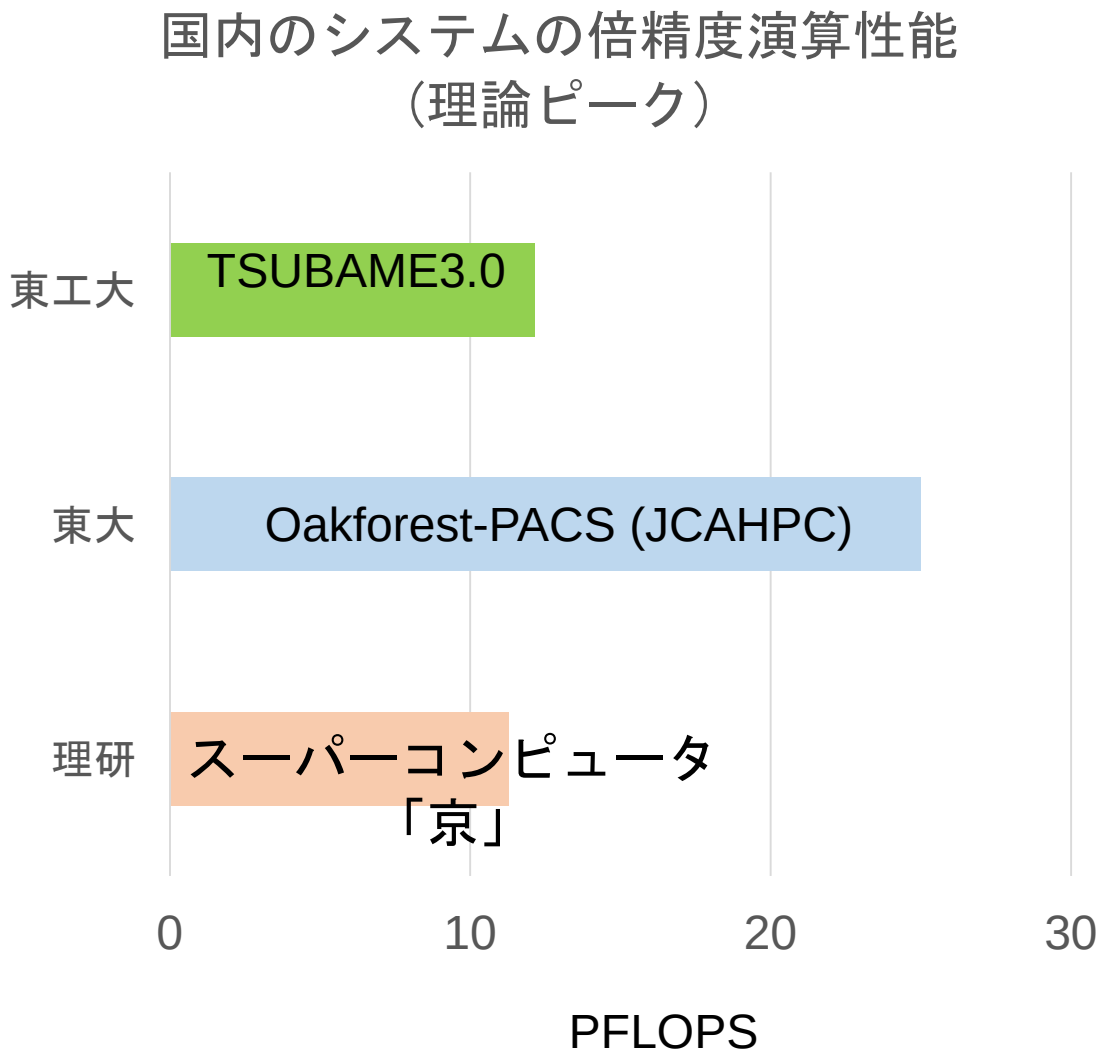
TSUBAME3諸元:

計算ノードの構成

CPU	Intel Xeon E5-2680 V4 (14 core) ×2
GPU	NVIDIA TESLA P100 (16GB, SXM2) ×4
Memory	DDR4-2400 DIMM 256GB
Network	Intel Omni-Path HFI 100Gbps ×4
SSD	Intel 2TB, NVMe

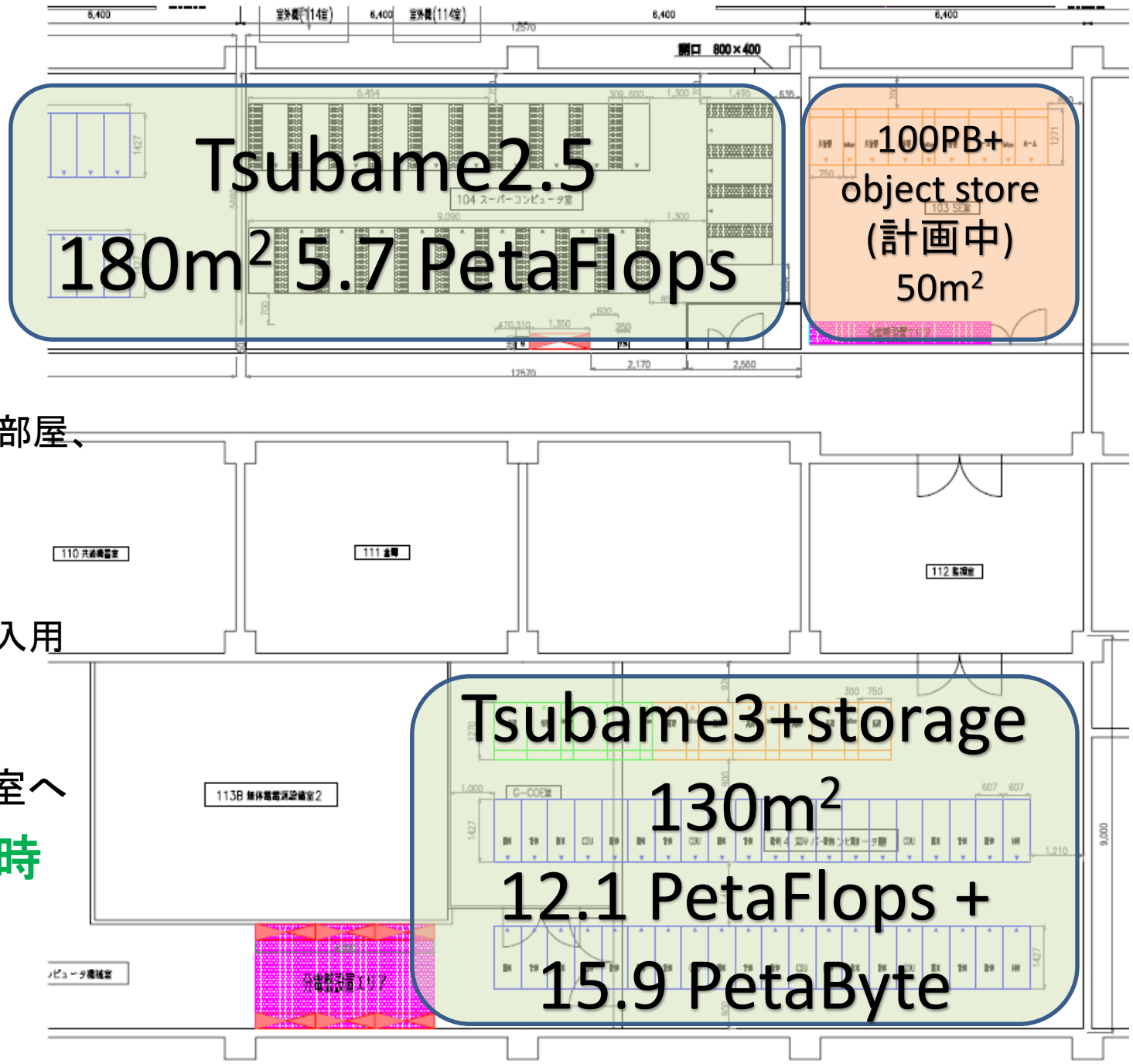
システム総計

倍精度演算性能	12.15 PFLOPS
メモリバンド幅	1.66 PByte/s
メモリ容量	168.75 TiB
インジェクション及びバイセクションバンド幅 双方向	216.0 Tbps 432.0 Tbps
NVMeローカルストレージ容量 ローカルストレージバンド幅	1080 TByte 最大2 TByte/s

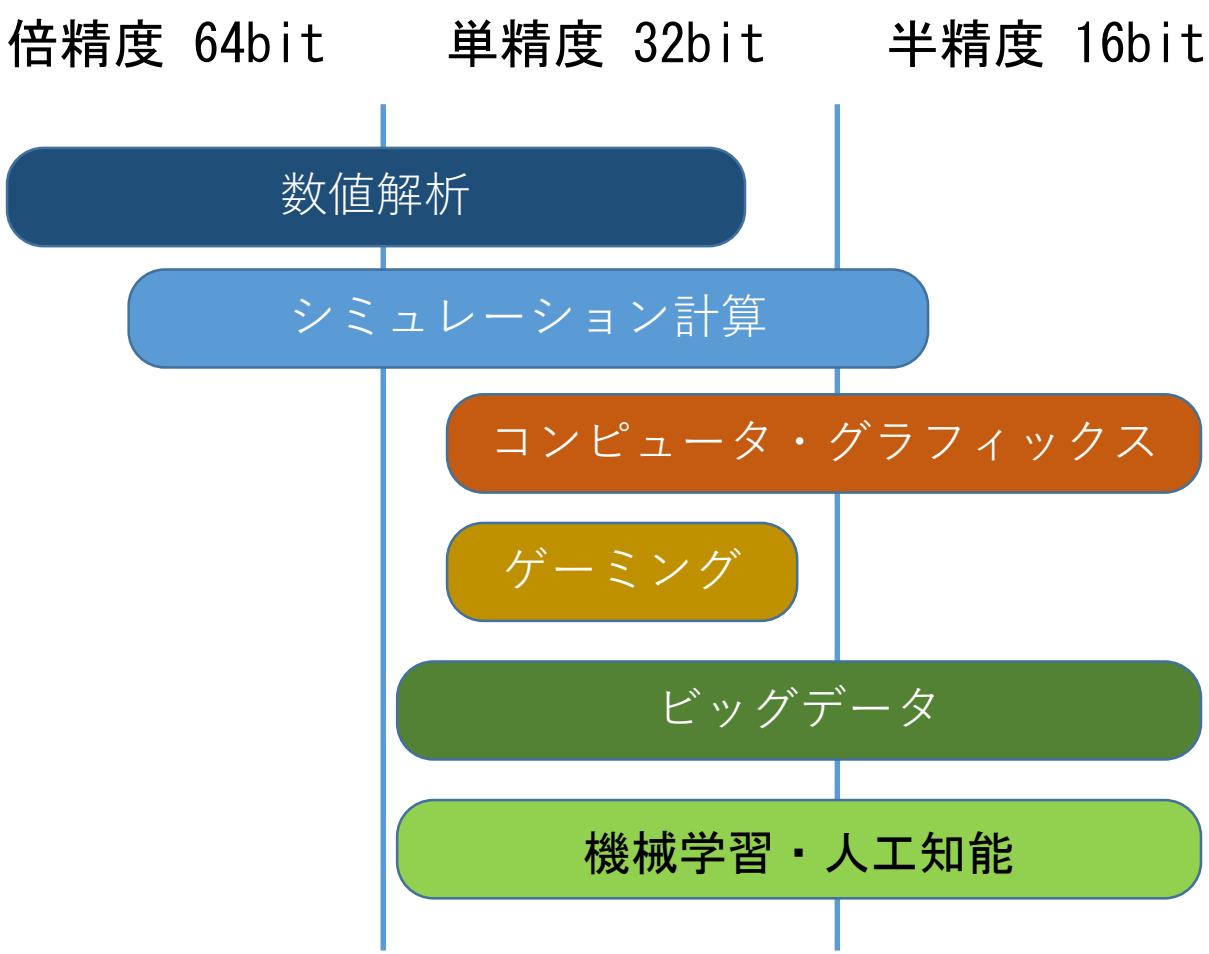


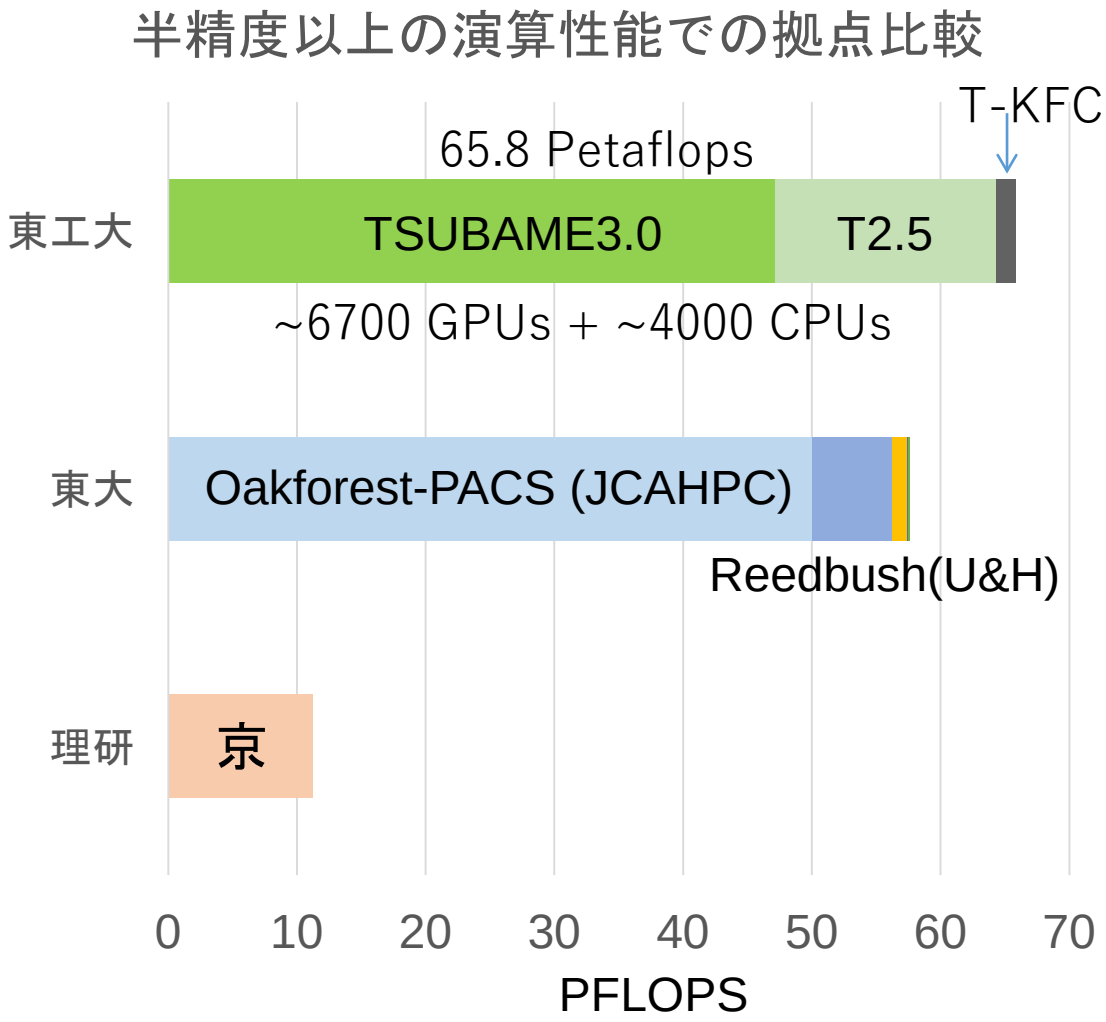
TSUBAME2&3同時運用フロアプラン

- 以前TSUBAME2のストレージと一部GSICの実験機器等がある114を改修、130m2以上のTSUBAME3新スパコン室へ
  - 撤去：真中の非構造壁、旧GCoEサーバー小部屋、高床構造、空調機など
  - 移動：TSUBAME2ストレージ(2Fに一時移動)
  - 強化：床耐荷重1m2あたり1t
  - 新設：GSIC搬入口近辺へ外気の遮断壁と搬入用ドア
- TSUBAME3の計算ノード、ネットワーク、ストレージ、管理ノード等全てが新スパコン室へ
- **研究成果による電力削減により、同時運用でも電気代はほぼH26のTSUBAME2の電気代並み**



浮動小数演算の精度





TSUBAME3+2.5+KFCの合算で、機械学習/AIの性能はスパコン・クラウドすべて含め日本トップ

FP16 on Tesla P100

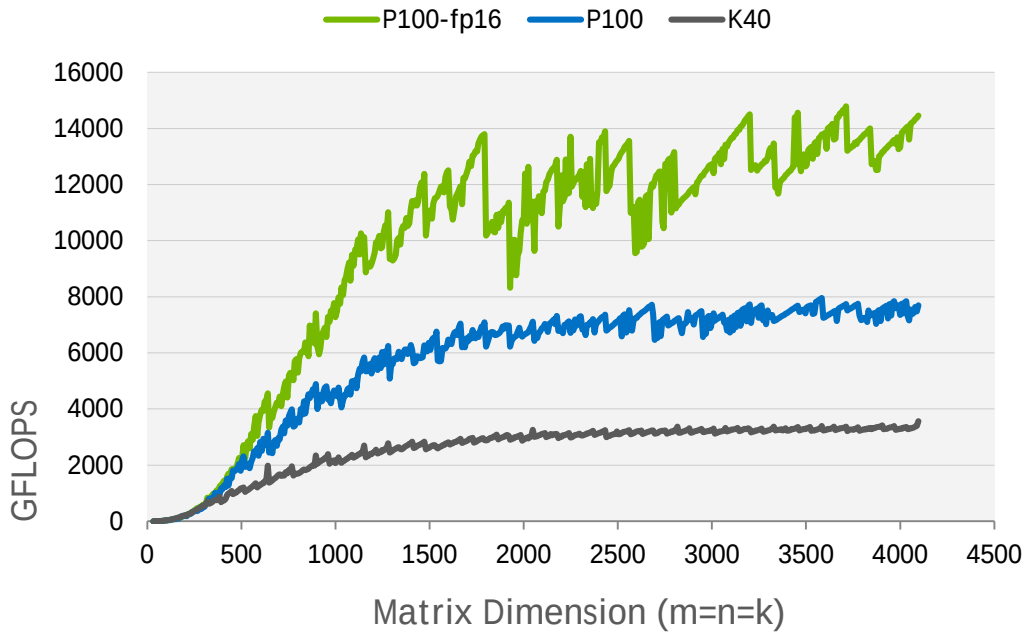
Half precision support for new type of applications

Tesla P100 supports FP16

Tesla P100 supports new FP16 compute capability, for applications that can take advantage of the new half precision, they can experience up to 2x speedup over using FP32

Storing FP16 data compared to higher precision FP32 or FP64 reduces memory usage of the neural network, allowing training and deployment of larger networks, and FP16 data transfers take less time than FP32 or FP64 transfers

> 4x Faster GEMM Performance with FP16 on Tesla P100



- Comparing GEMM performance on K40m (FP32) and P100 (FP32 and FP16)
- cuBLAS 8 on P100, Base clocks (r361)
- cuBLAS 8 on P40, Base clocks (r367)
- cuBLAS 7.5 on K40m, Base clocks, ECC ON (r352)
- Input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3@ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory
- m=n=k=4096

計算ノードあたりの性能比較（理論値）

指標	TSUBAME2.5(2013)	TSUBAME3.0(2017)	倍率
CPUのコア数×動作周波数（GHz）	35.16	72.8	2.07
CPUのメモリ容量（GB）	54	256	4.74
CPUのメモリバンド幅（GB/s）	64	153.6	2.40
GPUのCUDAコアの数	8,064	14,336	1.78
GPUのFP64（TFLOPS）	3.93	21.2	5.39
GPUのFP32（TFLOPS）	11.85	42.4	3.58
GPUのFP16（TFLOPS）	11.85	84.8	7.16
GPUのメモリ容量（GB）	18	64	3.56
GPUのメモリバンド幅（GB/s）	750	2928	3.90
SSDの容量（GB）	120	2000	16.67
SSD READ（MB/s）	550	2700	4.91
SSD WRITE（MB/s）	500	1800	3.60
ネットワーク転送速度（Gbps）	80	400	5.00

計算ノードあたりの性能比較（ベンチマーク）

指標	TSUBAME2.5	TSUBAME3.0	倍率
[CPU] SPEC CINT2006_rate base	330	1230	3.73
[CPU] SPEC CFP2006_rate base	235	899	3.83
[GPU] HPL（TFLOPS）	2.02	15.00	7.42
[GPU] HPCG（GFLOPS）	61.19	339.59	5.55

TSUBAME3/2.5でサポートされる エヌビディア ディープラーニング プラットフォーム

アプリケーション

Image Classification　Object Detection

コンピュータ ビジョン

Voice Recognition　Translation

会話と音

Recommendation Engines　Sentiment Analysis

振る舞い

フレームワーク

Caffe　Chainer　DL4J Deeplearning4j　Mocha.jl julia　KERAS　Microsoft CNTK　MatConvNet　mxnet　MINERVA　Purine　OpenDeep　Pylearn2　TensorFlow　theano　torch

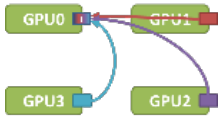
ディープラーニングSDK

cuDNN　GIE

ディープラーニング

cuBLAS　cuSPARSE　cuFFT

数学ライブラリ



NCCL

マルチ GPU 間通信

GPU プラットフォーム

amazon web services　IBM SOFTLAYER　Microsoft Azure　阿里云计算 Alibaba Cloud Computing

クラウド

Tesla P100　Tesla K80/M40/M4　Jetson TX1　DRIVEPX2

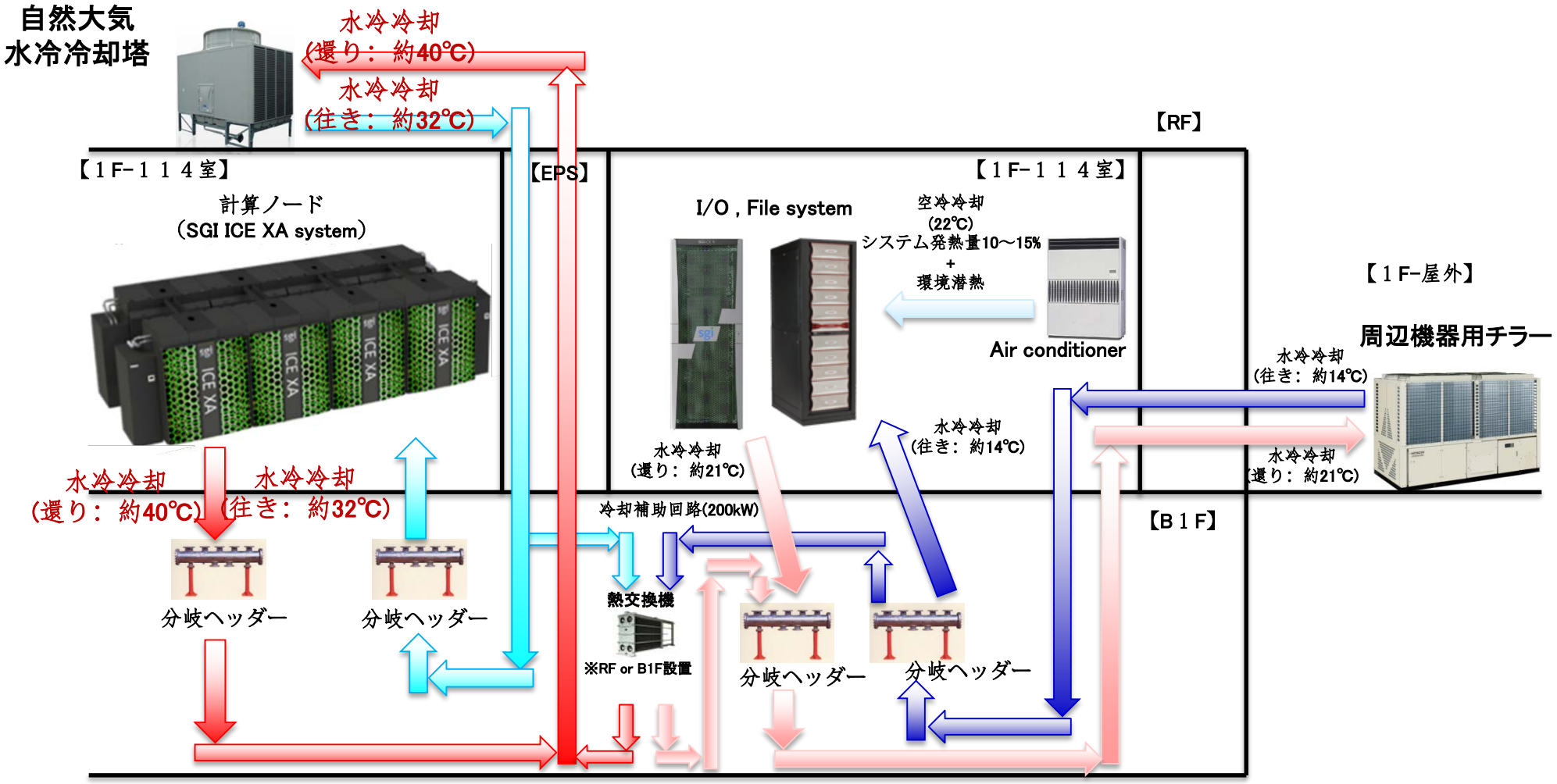
GPU

DGX-1

サーバー

NVIDIA

TSUBAME3.0 冷却システム系統図
32度の自然大気冷却水による高効率高温冷却



TSUBAME3.0 世界トップクラスの冷却効率

PUEは冷却のオーバーヘッドを表す。1.0に近いほど良い
PUE = {（計算ノードの消費電力）＋（計算ノードの冷却に必要なすべての設備消費電力）}
／（計算ノードの消費電力）

- **通常のデータセンター：PUE = 2～3（冷却の方がマシンより電気を食っている）**
- **最新の空冷データセンター、TSUBAME1：PUE = 1.4～1.6**
- **TSUBAME2.0：PUE = 1.28**
- **TSUBAME-KFC：PUE = 1.09**

TSUBAME 3.0 PUE 予測 (900KW消費仮定) 2013～2015年の天候データを元に計算

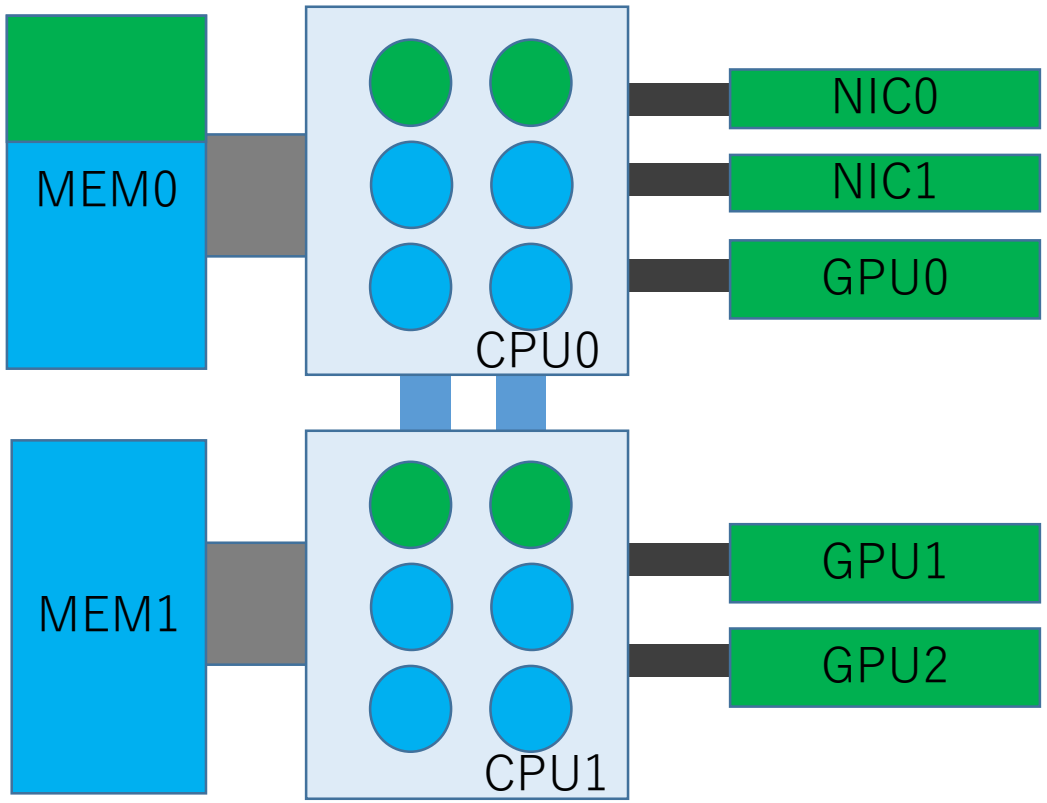
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	年間平均
冷却設備平均消費電力[kW]	28. 83	28. 83	28. 83	28. 83	29. 21	30. 06	30. 06	32. 00	30. 06	29. 21	28. 83	28. 83	29. 465
PUE	1. 032	1. 032	1. 032	1. 032	1. 032	1. 033	1. 033	1. 036	1. 033	1. 032	1. 032	1. 032	1. 033

TSUBAME3.0の計算ノードの年間PUE平均値は『1．033』
世界トップクラス

クラウド機能
複数のジョブによる計算ノードの共有

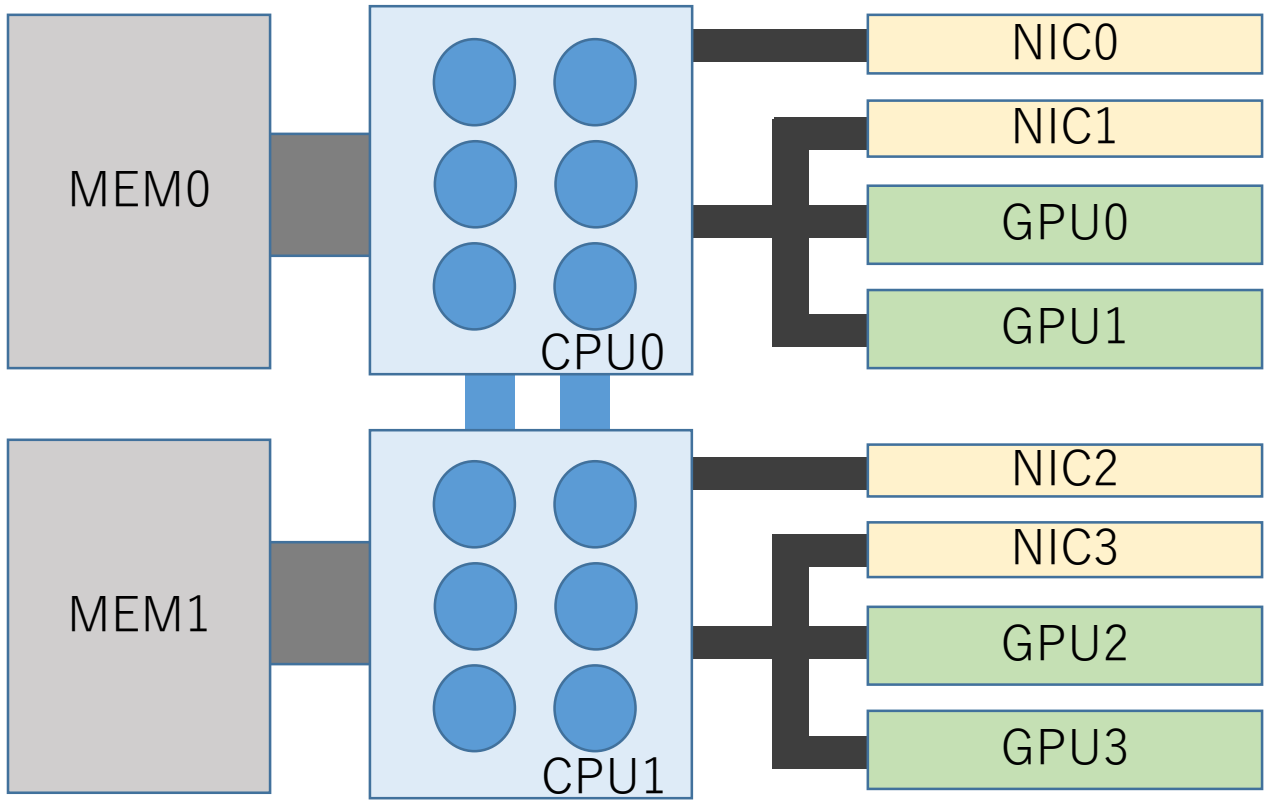
TSUBAME2.5は固定分割

Gキュー（4コア＋3GPU）とVキュー（８コア）



ジョブの種類	状況
多数のGPUを使うジョブ	Gキューへ
CPUだけ使用するジョブ	Vキューへ
GPUを１，２個使うジョブ	GPUが余る
メモリ使用量が少ないジョブ	メモリが余る

TSUBAME3.0の細粒度なリソース割り当て

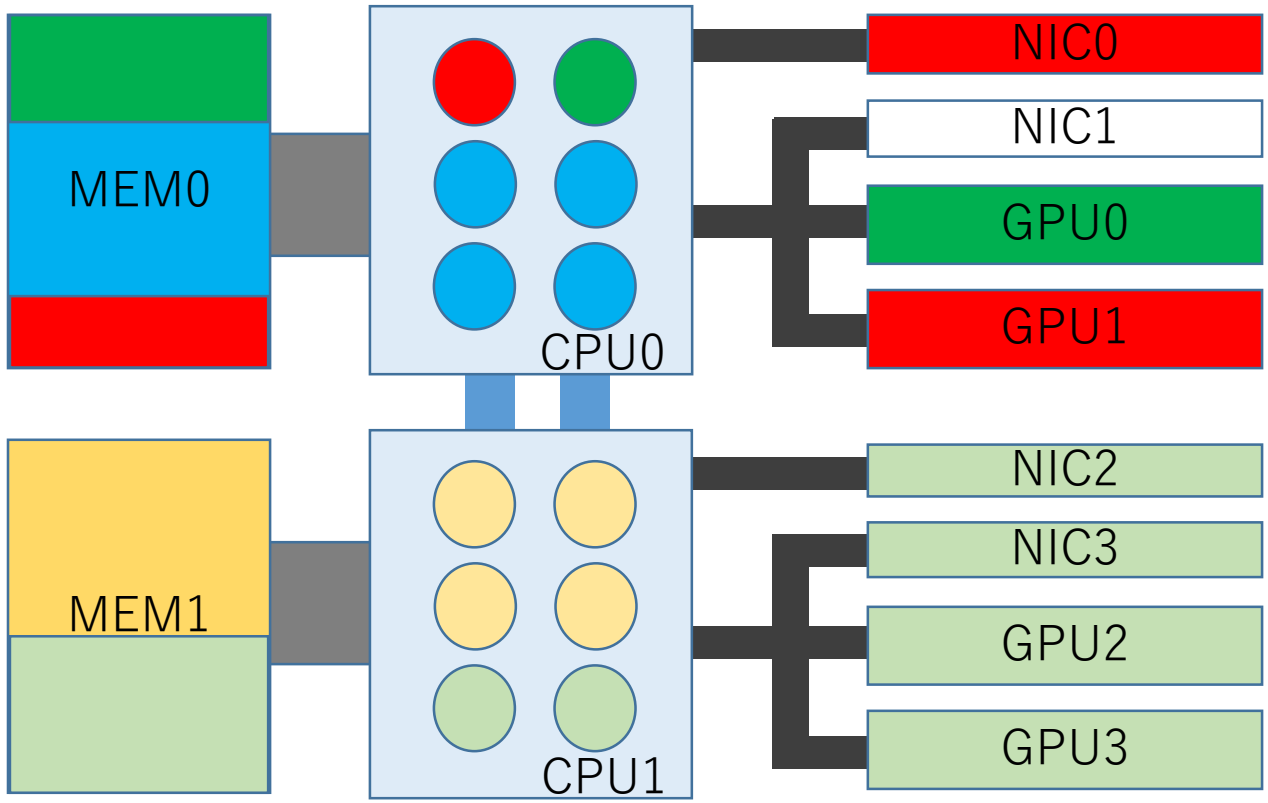


図ではCPUのコア数は少なく記載

Dockerにより計算ノードのリソース（CPUコア、GPU、NIC、メモリ）を分割。

ジョブスケジューラと連携しジョブが必要なリソースを割り当て、ノード内のリソースを無駄なく活用。

TSUBAME3.0の細粒度なリソース割り当て



ジョブ	割り当てリソース
1	CPU2コア、NIC0、GPU1、32GBメモリ
2	CPU8コア、64GBメモリ
3	CPU4コア、GPU0、16GBメモリ
4	CPU8コア、80GBメモリ
5	CPU4コア、NIC2&3、GPU2&3、48GBメモリ

図ではCPUのコア数は少なく記載

＜OILの目的＞

個人から社会の
挙動を表す
実社会ビッグデータ

企業・自治体保有のビッグデータ

- 生体データ：ヘルスケア，製薬
- 個人行動データ：eコマース，通信キャリア，交通
- 集団行動データ：電機メーカ，流通，交通

学術データ

- 研究・実験データ：大学・研究機関・医療機関

オープンデータ

- 地域活動データ：地域経済分析システム RESAS
- 高齢者データ：高齢者健康調査サンプル JAGES

…

ビッグデータを高度に解析するためのハード・ソフトウェアの研究開発

橋渡し

企業との連携拠点の形成

【東工大の強み】ハードウエア構築技術

- 世界トップクラスの大規模スーパーコンピュータ構築技術，TSUBAME KFC（Kepler Fluid Cooling）などの省電力計算機技術と豊富な運用実績
- 高速深層学習基盤の構築技術，大規模シミュレーション技術，統計物理に基づくモデリング技術，生命情報解析技術

【産総研の強み】ビッグデータ活用ソフトウェア開発

- 機械学習などデータ処理用計算機と高性能計算機をつなぐシステム連携技術
- 大規模環境計測データの解析技術，サービス設計技術，確率モデリング技術，多次元データ分析と可視化

産総研・東工大 OIL 実社会ビッグデータの活用基盤の構築

1. ビッグデータ処理オープンプラットフォームの確立
   - データ処理環境：ビッグデータ処理システム（TSUBAME／GPGPU計算機／次世代計算機）
2. ビッグデータを活用するデータ処理技術の開発
   - データ処理：確率モデリング／データ可視化／データマイニング／シミュレーション

産総研が産業界との連携・開発技術の実用化を主導

企業等

構築したビッグデータ活用基盤の利用を産業界に広く提供し、保有する実社会ビッグデータからの価値創造

- 自社サービスの向上に利用：ヘルスケア，eコマース
- 新サービスの創出に利用：プロバイダ，シンクタンク
- 他企業へ売却：銀行・証券・保険，不動産，製造業

産総研・東工大 実社会ビッグデータ活用オープンイノベーションラボラトリ

ラボ長
東工大 松岡聡 教授
（クロスアポイントメントにより現 産総研 特定フェロー）

副ラボ長
産総研 小川宏高 研究チーム長

運営連絡会
- 産総研 担当領域長
- 東工大 担当学院長
- ラボ長、副ラボ長
- 各機関の事務担当者，知財担当者

東工大 大岡山キャンパス（設置場所）

産総研 つくば第１（事務支援）

実社会HPC班

研究課題１：
ビッグデータ処理オープンプラットフォームの確立

（班長： 産総研 高野了成 研究グループ長）

- 東工大
  - 松岡聡 教授
  - 横田理央 准教授、遠藤敏夫 准教授
- 産総研
  - 高野了成 研究グループ長
  - 小川宏高 研究チーム長
  - 須崎有康主任研究員、広渕崇宏主任研究員
  - 谷村勇輔主任研究員、佐藤仁研究員
- その他（予定）
  - ポスドク　2名、テクニカルスタッフ 4名
  - RA 10名（博士 4名、修士 6名）、
  - 技術研修生 10名

BD/AI研究資源
東工大TSUBAME3/2.5
産総研ABCI/AAIC
⇨ **200 AI-FLOPS 以上**

データ活用班

研究課題「ビッグデータを活用するデータ処理技術の開発」
（班長：産総研 小川宏高 研究チーム長）

研究課題２-1：深層学習処理基盤を用いた大規模環境計測データの解析と応用
- 東工大 篠田浩一 教授、村田剛志 准教授
- 産総研 小川宏高 研究チーム長、李時旭 主任研究員、中村良介 研究チーム長
- その他（予定） ポスドク　2名、テクニカルスタッフ 4名、技術研修生 25名

研究課題２-2：実社会ビッグデータ分析とエージェントモデリングの融合技術研究
- 東工大 寺野隆雄 教授、小野功 准教授、高安美佐子 准教授、出口弘 教授
- 産総研 本村陽一 首席研究員、櫻井 瑛一 研究員
- その他（予定） ポスドク　3名、テクニカルスタッフ 3名、
  RA 2名（博士1名、修士1名）、技術研修生 2名 予定

研究課題２-３：超多次元データからの弁別特徴発見システム開発と実データの応用
- 東工大 秋山泰 教授、生命理工学研究科 山田拓司 准教授、
  石田貴士 准教授、大上雅史 助教、関嶋政和 研究ユニットリーダー
- 産総研 瀬々潤 研究チーム長、永田賢二 主任研究員、富井健太郎 研究チーム長
- その他（予定） ポスドク　4名、
  RA 16名（博士1名、修士15名）、技術研修生 12名、非常勤研究者（他予算）9名

研究課題１「ビッグデータ処理オープンプラットフォームの確立」

【研究内容】
産総研が整備するAIクラウド（ABCI）と東工大が有するTSUBAMEなどのスパコン（TSUBAME2.5/3.0）を高速に相互運用するための技術を導入し課題2で開発するビッグデータ解析汎用ツールを搭載する。企業ユーザ利用を促進するためにオープンプラットフォームを形成する。技術仕様やソースコードを公開するオープンプラットフォームとして開発することで、外部からの開発者やサードパーティなどの参入障壁を解消する。

研究項目1
高速化・低消費電力化に向けた
ミドルウェアの研究

・ABCIやTSUBAME3.0のアーキテクチャに基づいて
ビッグデータ処理に特化した運用技術を最適化するミドルウェアを開発し、高速化・低消費電力化を実現する。

開発項目2
効率的かつ簡便なビッグデータ処理を
支援するツールの開発

・開発したミドルウェアを利用可能な環境を簡易に構築するために、必要なソフトウェア群を自動的にインストールするソフトウェアスタックを開発する。
・複数のプロセスを分離可能なコンテナ技術を応用して、
安定した実行環境を提供する。

橋渡し

【連携が想定される企業】
高速稼働するビッグデータ処理プラットフォームの提供（D社等）
開発した設計・運用技術のデータセンター運営企業への技術移転（S社等）

研究課題２ ビッグデータを活用するデータ処理技術の開発

研究課題２-１：「深層学習処理基盤を用いた大規模環境計測データの解析と応用」

【研究内容】
連続的なデータとして高精度センサ（ドライブレコーダ、監視カメラ、航空機・人工衛星等）から得られる異種・大量な環境計測データからヒトやモノの動きをリアルタイムで状況認識・異常検知の解析を行う手法を研究し汎用ツール・ライブラリを設計・実装・評価する。

橋渡し

【連携が想定される企業】
商業施設での人流解析シミレーション（P社等）
自動車メーカー、自動運転技術開発、
ドローンメーカー、衛星画像処理（N社等）

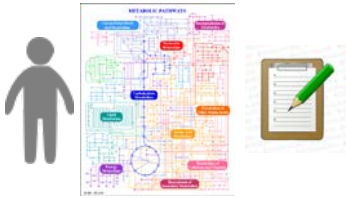
研究課題２-２：「実社会ビッグデータ分析とエージェントモデリングの融合技術研究」

【研究内容】
社会システムのデータとして数十万人の群衆の挙動などのビッグデータを取り扱うエージェントシュレーションを高速かつ高性能に実現するための手法を研究し汎用ツール・ライブラリを設計・実装・評価する。

橋渡し

【連携が想定される企業】
工員の動きを含む大規模工場操業の効率化（A社等）
新サービスや製品設計支援の有効性の事前評価
（H社、T社等）

研究課題２-３：「超多次元データからの弁別特徴発見システム開発と実データの応用」



【研究内容】
データの自由度（特徴量の次元）が高い超多次元データを２群または多群に弁別するため、特徴量を発見し仮説検定を行うまでの作業を自動化する手法を研究し汎用ツール・ライブラリを設計・実装・評価する。

橋渡し

【連携が想定される企業】
創薬関連・製薬企業
（T社、A社、E社、D社、F社等）
健康食品産業