東京工業大学 学術国際情報センター

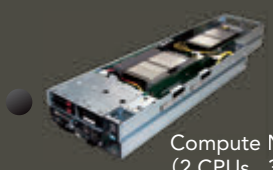
HARDWARE SOFTWARE SPECIFICATIONS

Global Scientific Information and Computing Center

TSUBAME 2.5

ハードウェア・ソフトウェア スペシフィケーション（仕様）

- GPUの大量搭載による高性能計算ノード
- 高速ネットワークによる内部接続
- 高速・高信頼性ストレージ
- 低消費電力・グリーン運用
- システム・アプリケーション・ソフトウェア

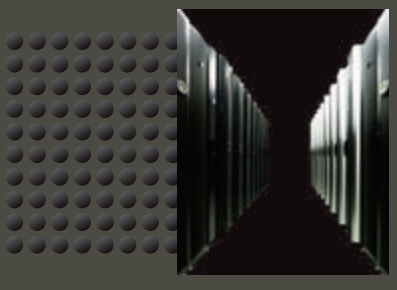


Compute Node
(2 CPUs , 3GPUs)

4.08 TFLOPS.
58.0 GB（CPU）+ 18 GB（GPU）

Rack（30 nodes）

122 TFLOPS
2.28 TB



System（58 Racks）

1442nodes
2952 CPU sockets :
224.7 TFLOPS
※Turbo boost

4360GPUs:
5.562 PFLOPS

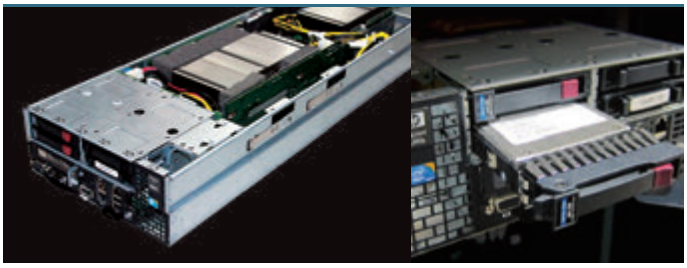
Total:
5.787 PFLOPS

Memory:
116 TB

■ GPUの大量搭載による高性能計算ノード

計算ノードはThin、Medium、Fatの3種類のノードから構成されています。演算性能の殆どを占めるThinノードは17/2インチ幅、高さ2Uのサイズにに CPUを2個、Keplerコアの GPUを3個搭載するコンパクトな設計になっています。さらにQDR InfiniBand HCAを2つ接続しつつPCI Express Bus の通信帯域を確保しています。電源ユニットも3+1に多重化され、高信頼性も兼ね備えています。

Thinノード　1408ノード



HP ProLiant SL390s G7

CPU : Intel Xeon X5670 (Westmere-EP, 2.93GHz, 3.196GHz@Turbo boost)
×2 ソケット　ソケットあたり 6 コア, ノード内合計 12 コア
GPU : NVIDIA Tesla K20X (GK110) ×3, GPU1 個あたり 1.31TFLOPS, VRAM 6GB
Memory : 58GB DDR3 1333MHz　一部 103GB
SSD : ノードあたり 120GB (60GB × 2)　一部 240GB (120GB × 2)
Network: 4X QDR InfiniBand × 2

Mediumノード　24ノード

HP ProLiant DL580 G7

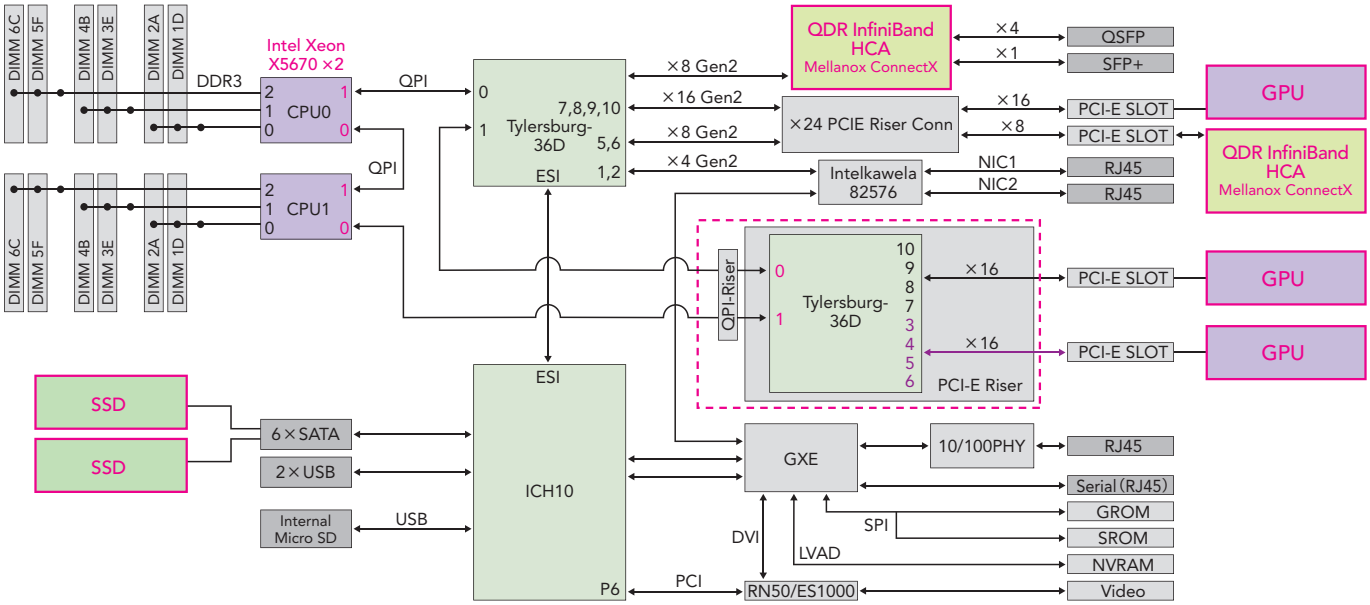
CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz × 4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
(NVIDIA Tesla C1060 × 4) or
NextIO vCORE Express 2070
(NVIDIA Tesla M2070 × 4)
Memory:137 GB (DDR3 1066MHz)
SSD: 120GB × 4 (480GB/node)
Network: 4X QDR InfiniBand

Fatノード　10ノード

HP ProLiant DL580 G7

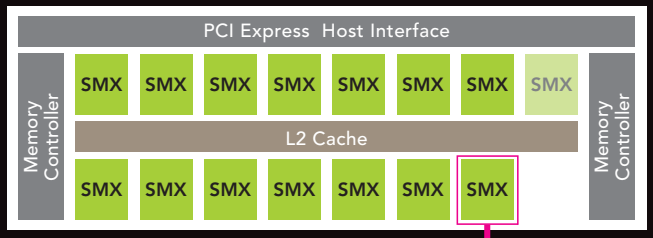
CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz × 4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
(NVIDIA Tesla C1060 × 4)
Memory: 274 GB (8 nodes),
548 GB (2 nodes)
DDR3 1066MHz
SSD: 120GB × 5 (600GB/node)
Network: 4X QDR InfiniBand

Thinノード　ブロック図



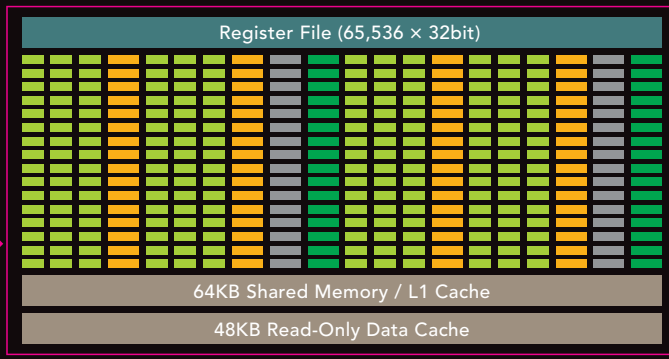
GPUの詳細

K20X Architecture (Kepler GK110 Core)



・ピーク性能：1.31 TFLOPS（倍精度）
3.95 TFLOPS（単精度）
・シェーダクロック：732 MHz
・CUDA コア（SP）数：2,688 cores
・Streaming Multiprocessor（SMX）：14 SMX
・ライタブル L2 キャッシュ：1.5MB
・メモリ帯域：250GB/s
・メモリクロック：2.6GHz（GDDR5）
・ECC メモリ：内部及び外部メモリ
・オンボードメモリ：6GB

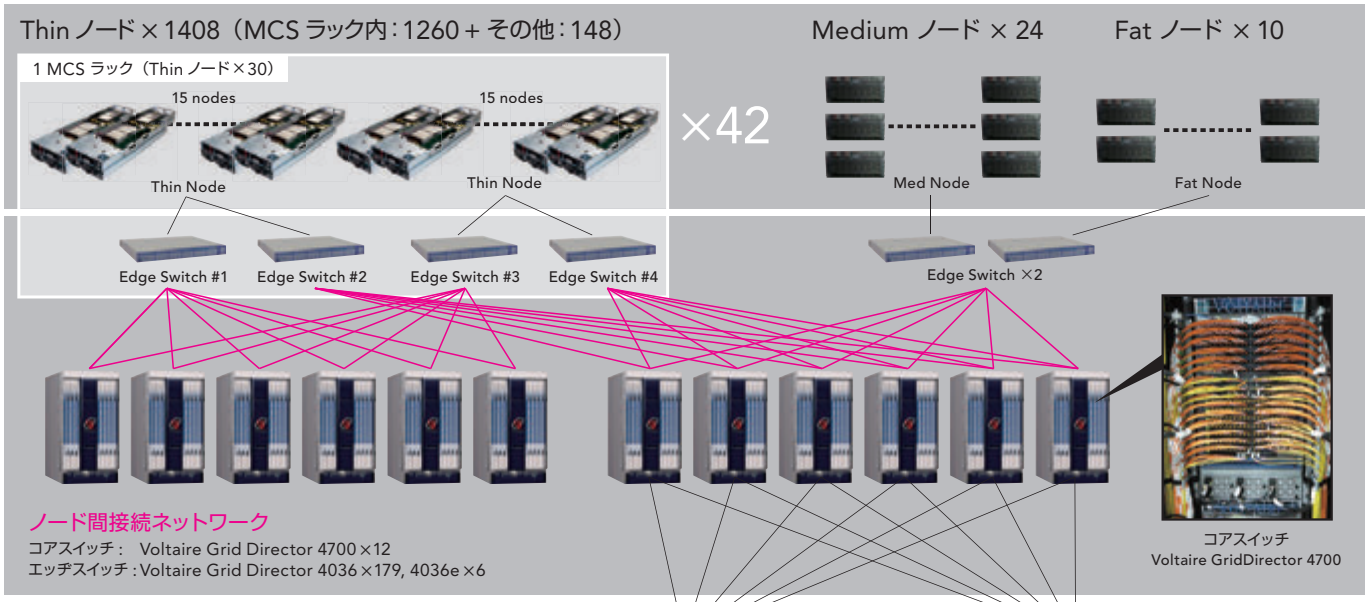
SMX の詳細

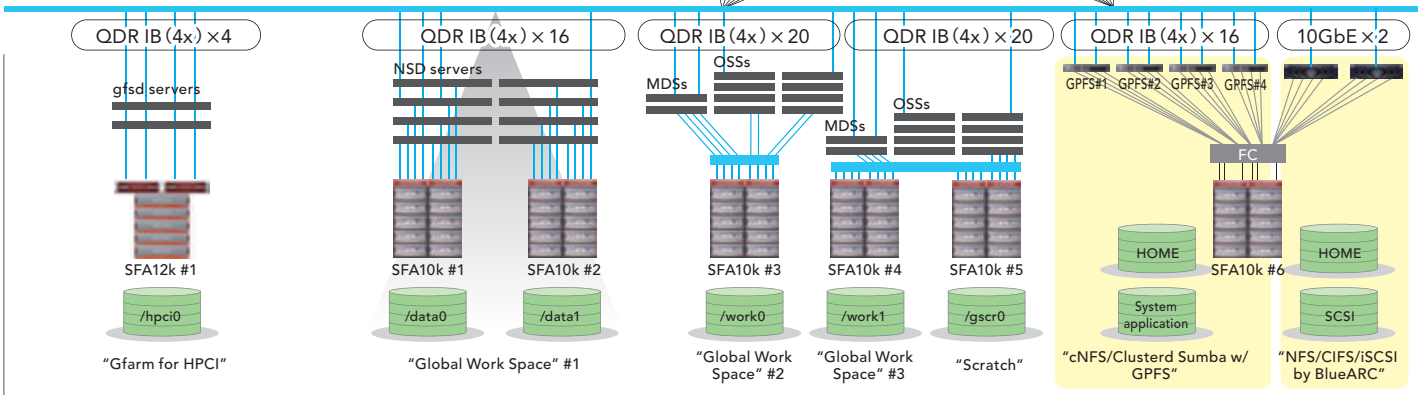


・CUDA コア（SP）/ SMX：192 cores
・DP ユニット / SMX：64　・SFU / SMX：32
・WARP スケジューラ / SMX：4 units
・シェアードメモリ / SMX：16KB or 32KB or 48KB
・ライタブル L1 キャッシュ / SMX：48KB or 32KB or 16KB
・リードオンリー・データキャッシュ / SMX：48KB

■ 高速ネットワークによる内部接続

TSUBAME2.5ネットワークはDual-Rail QDR InfiniBandをベースに構成され、計算ノード間はFat-Tree型のインターコネクションにより
フル・バイセクションバンド幅として200Tbpsを達成しています。計算ノード間End-to-Endの遅延もマイクロ秒オーダーと非常に小さく高速であり、
高信頼ストレージとも高速に接続しています。このネットワークは総計100Km, 3000本あまりの光ファイバーケーブルにより支えられています。



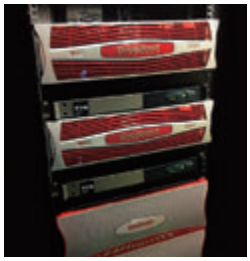


HPCI

Gfarm: ～ 600TB

gsfd server:
HP ProLiant DL360p Gen8 × 2
Intel XeonE5 2640 × 2,
64 GB Mem,
QDR IB (4x) × 2

Storage:
DDN SFA12k × 1
4TB SAS HDD × 155 disks



GPFS with HSM

GPFS 2.4 PB

NSD server: HP ProLiant DL380 G6 × 4
Intel Westmere EP × 2,
48GB Mem, QDR IB (4x) × 2

HP ProLiant DL360 G6 × 4
Intel Westmere EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA10k × 2
2TB SATA HDD × 1180 disks
600GB SAS HDD × 20 disks

Lustre

Lustre 3.6 PB

MDS : HP ProLiant DL360 G6 × 4
Intel Westmere-EP × 2,
48GB Mem, QDR IB (4x) × 2

OSS : HP ProLiant DL360 G6 × 16
Intel Westmere-EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA 10k × 3,
2TB SATA HDD × 1770 disks,
600GB SAS HDD × 30 disks

Home

Home 1.2 PB

cNFS (GridScaler)/Clustered Samba w/ GPFS:
HP ProLiant DL380 G6 × 4
Intel Westmere EP × 2,
48GB Mem, QDR IB (4x) × 2

NFS/CIFS/iSCSI:
BlueArc Mercury 100 × 2
10Gbps × 2

Storage: DDN SFA 10k × 1
2TB SATA HDD × 600 disks

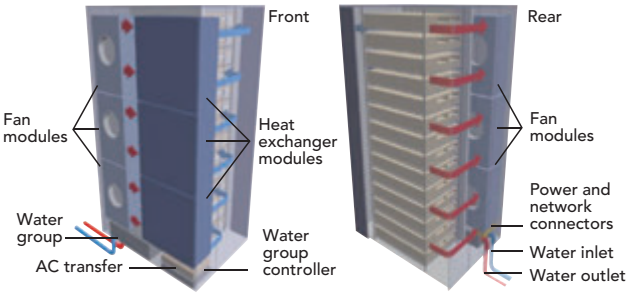
■ 高速・高信頼性ストレージ

TSUBAME2.5は、各計算ノードに備えたスクラッチ出力用の合算約190TBのSSD、高速なI/Oを行うための5.9PBのLustre、
GPFSなどの並列ファイルシステム領域、クラウドサービス用の1.2PBのホーム領域、GPFS並列ファイルシステムと連動し
階層型ストレージを構成する4PB超のテープライブラリなど、使用目的に応じて多様な計11PBもの莫大なストレージ領域を提供します。

■ 低消費電力·グリーン運用

Linpackベンチマーク電力性能：3068.71（MFLOPS/W）
システム機器ピーク消費電力：　1620（KW）
システム機器平均消費電力*：　698（KW）
システム機器アイドル消費電力：470（KW）
年間平均PUE：　1.285
「*」平均消費電力はTSUBAME 2.0 実績の年間平均

冷却：Modular Cooling System

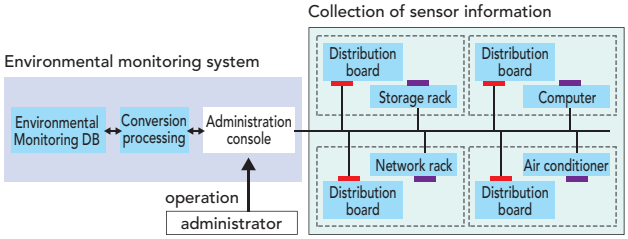


ラック内に熱交換システムを内蔵した密閉型水冷システムにより、通常のデータセンターに比べ高密度な世界トップクラス（ラックあたり最大35KW）の冷却が可能です。サーバの吸入口に均質な冷却風を提供し、ドア開閉は自動化·加湿不要となっています。完全自動温度制御による最適な消費電力点の制御を行い、95% から97% の熱を水冷で除去することが可能です。また、ポリカーボネート製のドアは大幅なノイズ削減にも貢献しています。

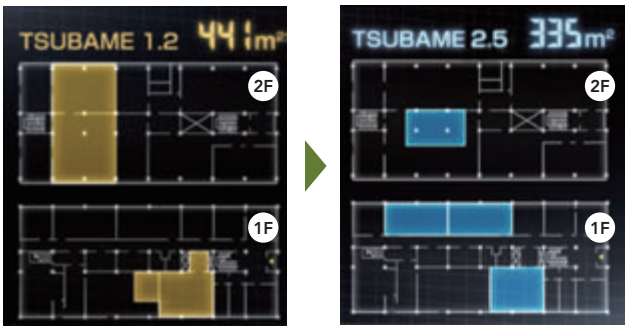
- 空調機器ピーク消費電力：460（KW）
- 空調機器平均消費電力*：204（KW）

グリーン運用：環境モニタリング

計算機ルームだけではなく、計算ノード、ラック毎の温度、消費電力などをリアルタイムで監視しています。



狭い設置面積



TSUBAME 1.2 から性能が70倍以上向上したのに設置面積は逆に狭くなっています。

■ System Software　WindowsとLinuxを動的に切り替える“Dynamic provisioning”

ジョブ管理システムとクラスタ管理システムを連携させてユーザ環境を管理し、ノードプールから計算リソースを取り出して足りない部分に配分します。Linux用とWindows用のバッチスケジューラにより計算ノードを管理し、ノードの動的な増加·削減に対応しています。仮想マシンの実行をサポートし、それらもジョブスケジューリングの対象として管理します。

OS	SUSE Linux Enterprise Server 11 SP1 Windows HPC Server 2008 R2
バッチシステム	PBS Professional

■ ISV（commercial） Software

（「*」は GPU 対応または一部対応）（2013年11月現在）

Compilers, Debuggers and Libraries

Intel Compiler（C/C++/Fortran）
PGI Compiler*
（C/C++/Fortran, OpenACC, CUDA Fortran）
Total View Debugger*
CAPS Compiler*（HMPP, OpenACC）
CULA*（Numerical Libraries for CUDA）

Applications

ANSYS Fluent*, Workbench*
MSC Nastran*
LS-DYNA
Gaussian, Gauss View
Molpro
Scigress
MATLAB*
AVS/Express, AVS/Express PCE
ABAQUS*, ABAQUS CAE
Patran
CST STUDIO SUITE*（MW-Studio*）
AMBER*
Materials Studio, Discovery Studio
Mathematica*
Maple*
EnSight

：全てのユーザ使用可能ライセンス　：学内ユーザのみ使用可能ライセンス　：産業利用のユーザのみ使用可能ライセンス

発行:東京工業大学 学術国際情報センター
〒152-8550 東京都目黒区大岡山 2-12-1　電話:03-5734-2087　FAX:03-5734-3198　E-mail:tsubame@gsic.titech.ac.jp

http://www.gsic.titech.ac.jp/

201311