AXIES2020 2020.12.09-10@オンライン
2020.12.11 9:00-10:30 HPCテクノロジー1
9:00-9:18 FA1-1

大島 聡史, 永井 亨, 片桐 孝洋（名古屋大学）

スーパーコンピュータ「不老」のシステム構成と性能

名古屋大学 NAGOYA UNIVERSITY
ITC
不老

AXIES2020：スーパーコンピュータ「不老」に関する発表について

- FA1-1 スーパーコンピュータ「不老」のシステム構成と性能
  - 全体的なシステム構成と性能について

- FA1-2 スーパーコンピュータ「不老」のサービスとエコシステム
  - 特徴的な利用制度と消費電力抑制について

- FA1-3 スーパーコンピュータ「不老」における光ディスクライブラリを用いたコールドストレージシステムの構築
  - コールドストレージの設計と運用について

名大情報基盤センターのスパコンの歴史

2020年7月1日
スーパーコンピュータ「不老」導入
（運用開始）

96 00 05 09 13 15 20

Fujitsu M-1800 → Fujitsu GP7000F/90

メインフレーム系


Fujitsu VPP500 → Fujitsu VPP5000

ベクトル型スパコン


Fujitsu HPC2500

スカラー型スパコン

Fujitsu HX600, M9000 → Fujitsu CX400 → CX400

アプリケーションサーバ


Fujitsu FX1 → Fujitsu FX10 → FX100

スーパーコンピュータ



複合型「不老」

Fujitsu CX2570

GPUサーバ

Fujitsu FX1000

超並列

HPE Super Dome Flex

大規模共有メモリ

HPE ProLiant DL560

クラウド

次期不老へ

- これまで約5年間隔でリプレイス
- 「不老」は2026年3月末まで（5年9ヶ月）運用の予定

スーパーコンピュータ「不老」　導入の背景と狙い

- 研究のデジタル化（デジタルサイエンス）
  - 従来よりさらに多くの分野で計算機が必要に
- 大規模数値シミュレーション研究に対する需要の増大
  - 異常気象や津波など安心・安全のための研究の増加
  - 生命や宇宙など基礎科学分野の要求も引き続き大きい
- AI・機械学習に関する研究の増大
  - 自動運転、医療、創薬、etc.　⇒　GPUへの期待
- データの爆発的増大
  - 計測データ、解析結果、AI学習結果など
  - 高速アクセス　＋　長期保存

- ◆ 大規模かつ複合型のスパコンを導入して様々な要求に応える
- ◆ 従来型の大規模共有ストレージに加えて長期データ保存向けのコールドストレージを搭載

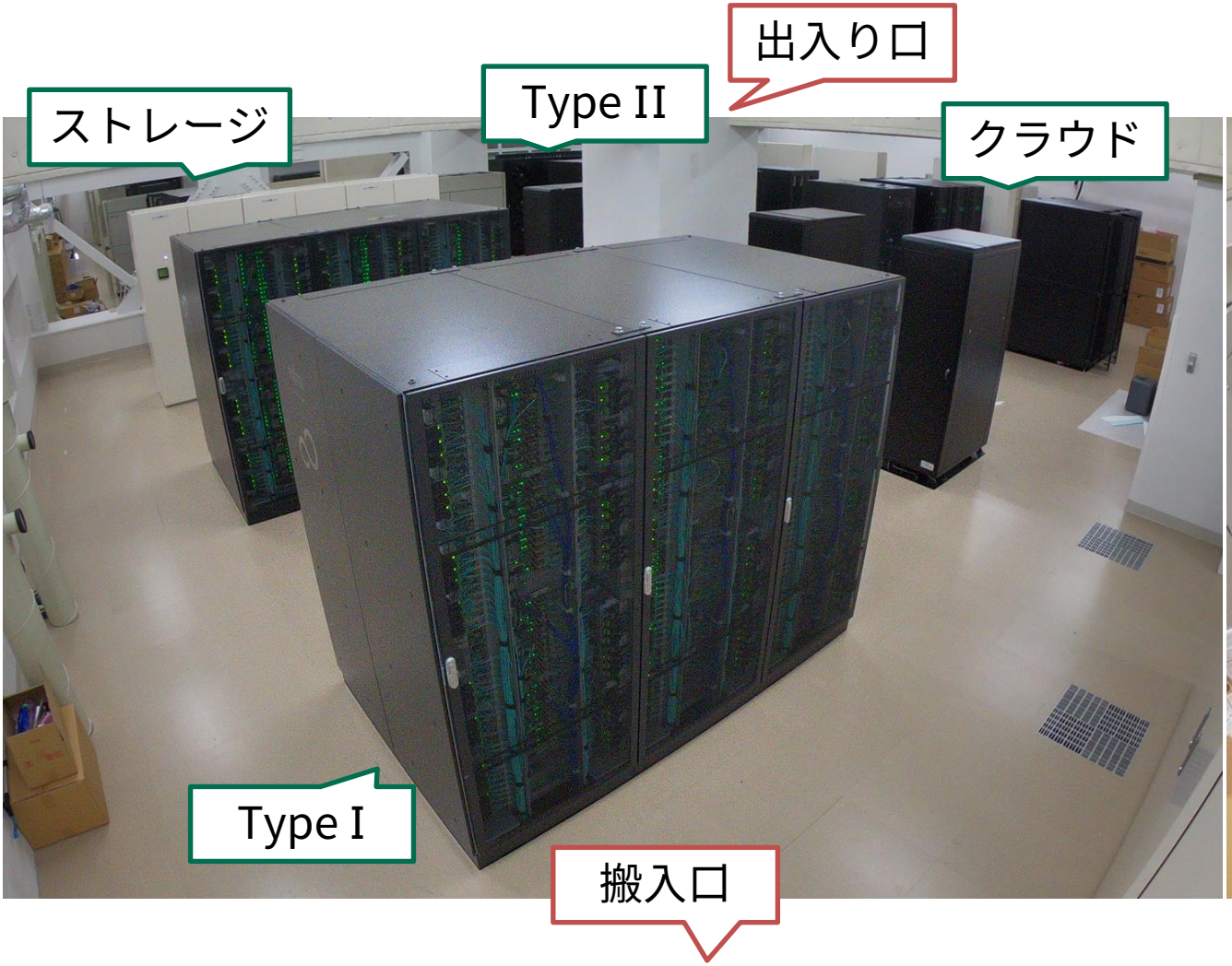
愛称とロゴについて

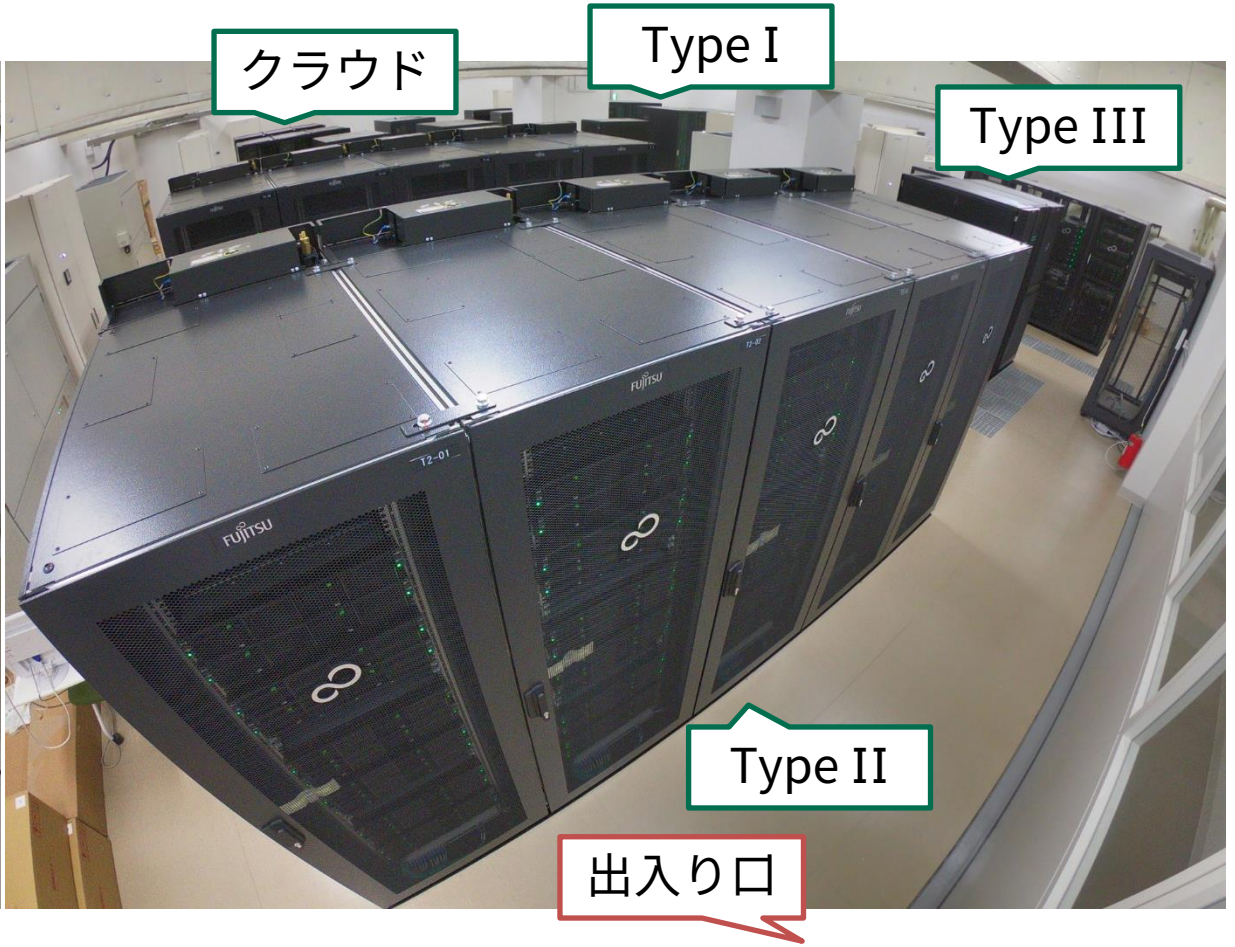
- 利用者により身近に感じてもらえるシステムとなることを目指して、本センターとしては初めて公募を行い愛称を付けた
  - 愛称を付けること自体が初めて

- 愛称：スーパーコンピュータ「不老」
  - 商標等の都合により、「不老」ではなく
    スーパーコンピュータ「不老」が正式な愛称

- 採用されたのは名古屋市内の高校生による案
- 由来は名古屋大学の所在地「愛知県名古屋市千種区不老町」
  - このシステムによって人類が得た恩恵がその後末永く
    人類の文明の中に生き続けることを願って命名。
  - なお英語名称はFlow。そのため不老も「フロウ」ではなく
    「フロー」と読むことになっている。

設置状況（名古屋大学 情報基盤センター　本館地下１階）

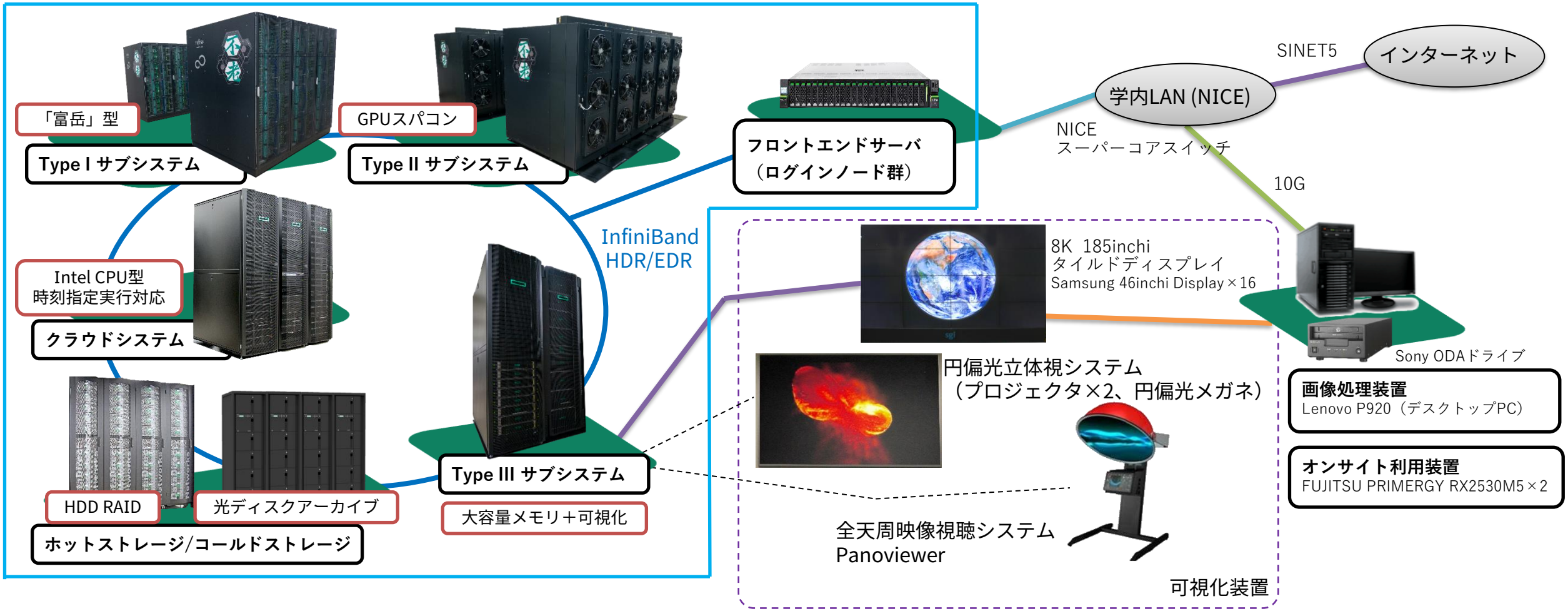




システム構成（全体の概要）

- 特徴の異なる4つの計算サブシステムと
  2つのストレージを中心とした複合型のシステム

総理論演算性能 15.886PFLOPS、総主記憶容量240.375TiB
Hot Storage 30PB、Cold Storage 6PB

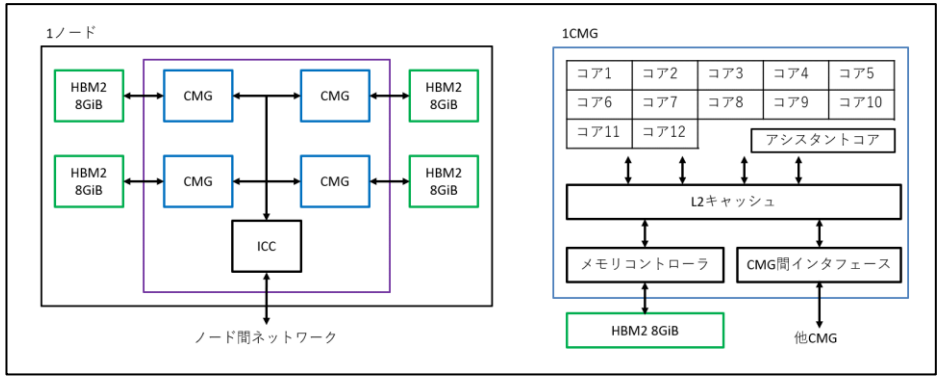


スーパーコンピュータ「不老」Type I サブシステム

機種名		FUJITSU Supercomputer PRIMEHPC FX1000
計算ノード	CPU	A64FX (Armv8.2-A + SVE), 48コア+2アシスタントコア (I/O兼計算ノードは48コア+ 4アシスタントコア), 2.2GHz, 1プロセッサ
	メインメモリ	HBM2, 32GiB
	理論演算性能	倍精度 3.3792 TFLOPS, 単精度 6.7584 TFLOPS, 半精度 13.5168 TFLOPS
	メモリバンド幅	1,024 GB/s (1CMG=12コアあたり256 GB/s, 1CPU=4CMG)
ノード数、総コア数		2,304ノード, 110,592コア (+4,800アシスタントコア)
総理論演算性能		7.782 PFLOPS
総メモリ容量		72 TiB
ノード間インターコネクト		TofuインターコネクトD 各ノードは周囲の隣接ノードへ同時に合計 40.8 GB/s × 双方向 で通信可能（1リンク当たり 6.8 GB/s × 双方向, 6 リンク同時通信可能）
ユーザ用 ローカルストレージ		なし
冷却方式		水冷

- 世界初運用の スーパーコンピュータ「富岳」型 システム
- 自己開発のMPIプログラム向き
- 超並列処理用
- AIツールも提供

ノード内構成

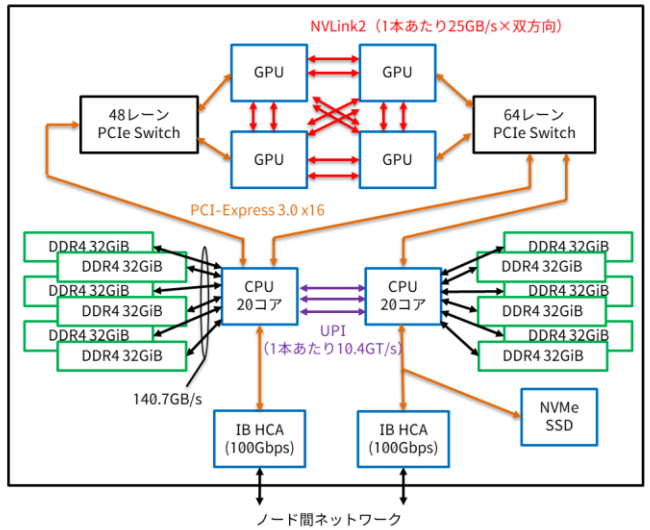


スーパーコンピュータ「不老」Type II サブシステム

種名		FUJITSU Server PRIMERGY CX2570 M5
計算ノード	CPU	Intel Xeon Gold 6230, 20コア, 2.10 – 3.90 GHz × 2 ソケット
	GPU	NVIDIA Tesla V100 (Volta) SXM2, 2,560 FP64コア, up to 1,530 MHz × 4ソケット
	メモリ	メインメモリ(DDR4 2933 MHz)：384 GiB (32 GiB × 6 枚 × 2 ソケット) デバイスメモリ(HBM2)：32 GiB × 4 ソケット
	理論演算性能	倍精度 33.888 TFLOPS　(CPU 1.344 TFLOPS × 2 ソケット, GPU 7.8 TFLOPS × 4 ソケット)
	メモリバンド幅	メインメモリ 281.5 GB/s (23.464 GB/s × 6 枚 × 2 ソケット) デバイスメモリ 900 GB/s × 4 ソケット
	GPU間接続	NVLINK2 (1GPUから他の3GPUに対してそれぞれ50GB/s×双方向)
	CPU-GPU間接続	PCI-Express 3.0 (x16)
ノード数、総コア数		221ノード、8,840 CPUコア + 2,263,040 FP64 GPUコア
総理論演算性能		7.489 PFLOPS (CPU 0.594 PFLOPS, GPU 6.895 PFLOPS)
総メモリ容量		メインメモリ 82.875 TiB、デバイスメモリ 28.288 TiB
ノード間インターコネクト		InfiniBand EDR 100 Gbps × 2, 200 Gbps
ユーザ用 ローカルストレージ		NVMe SSD 6.4TB, 一部ノードにて BeeGFS/BeeOND/NVMesh (ローカルストレージを使用した共有ファイルシステム) を提供
冷却方式		水冷

- データサイエンス研究、機械学習用のGPUクラスタ型
- 最新GPU (Volta) 4台／ノード
- 充実したAIツール
- 高速SSDローカルディスク

ノード内構成

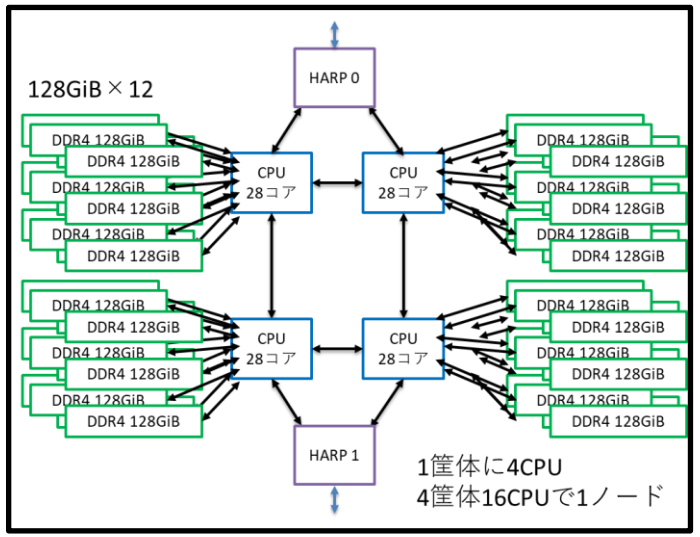


スーパーコンピュータ「不老」Type III サブシステム

機種名		HPE Superdome Flex
計算ノード	CPU	Intel Xeon Platinum 8280M, 28コア, 2.70 – 4.00 GHz × 16 ソケット
	GPU	NVIDIA Quadro RTX6000 × 4
	メモリ	メインメモリ(DDR4 2933 MHz)：24 TiB (128 GiB × 12枚 × 16 ソケット) デバイスメモリ(GDDR6)：24 GiB × 4
	理論演算性能	倍精度 38.7072 TFLOPS (CPU 2.4192 TFLOPS × 16 ソケット)
	メモリバンド幅	メインメモリ 2252.544 GB/s (23.464 GB/s × 12枚(6チャネル) × 16ソケット)
	CPU-GPU間接続	PCI-Express 3.0 (x16)
ノード数		2
総理論演算性能		77.414 TFLOPS (38.7072 TFLOPS × 2 ノード)
総メインメモリ容量		48 TiB
ノード間インターコネクト		InfiniBand EDR 100 Gbps
ユーザ用 ローカルストレージ		一方のノードに102.4 TB SSD、もう一方のノードに1008 TB 共有ストレージを接続
冷却方式		空冷

- 大規模共有メモリ（24TiB）
- プリポスト処理用、可視化処理用
- NICE DCVを用いたリモート可視化
- 1ノードをバッチ処理、1ノードを会話型処理に利用

ノード内構成

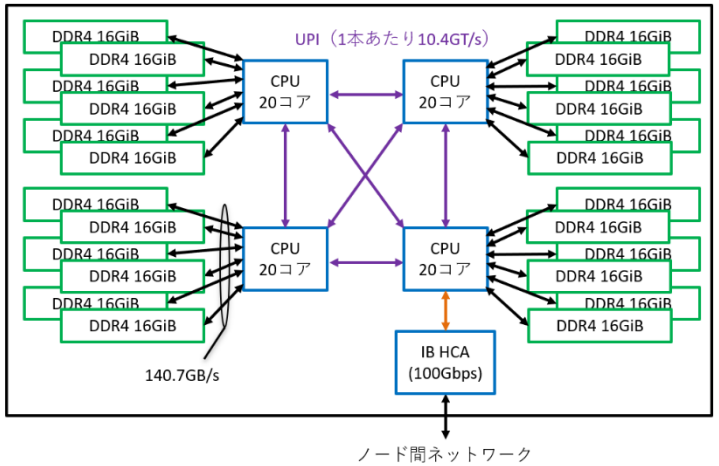


スーパーコンピュータ「不老」クラウドシステム

機種名		HPE ProLiant DL560
計算ノード	CPU	Intel Xeon Gold 6230, 20コア, 2.10 – 3.90 GHz × 4 ソケット
	メモリ	メインメモリ(DDR4 2933 MHz) 384 GiB (16 GiB × 6 枚 × 4 ソケット)
	理論演算性能	倍精度 5.376 TFLOPS (1.344 TFLOPS × 4 ソケット)
	メモリバンド幅	メインメモリ 563.136 GB/s (23.464 GB/s × 6枚 × 4 ソケット)
ノード数		100
総理論演算性能		537.6 TFLOPS (5.376 TFLOPS × 100 ノード)
総メインメモリ容量		37.5 TiB
ノード間インターコネクト		InfiniBand EDR 100 Gbps
ユーザ用ローカルストレージ		なし
冷却方式		空冷

- 研究室クラスタから移行しやすいIntel CPU搭載システム
- 高いノードあたりCPU性能（4ソケット）
- 時刻を指定してのバッチジョブ・インタラクティブ利用が可能（UNCAI）

ノード内構成



スーパーコンピュータ「不老」ホットストレージ

メタデータサーバ(MDS)	
機種名	FUJITSU PRIMERGY RX2540 M5
CPU	Intel Xeon Gold 5222 (3.80GHz, 4コア) × 2
メインメモリ	DDR4 192 GiB
HDD	SAS 900 GB 10krpm × 2 (RAID1)
Interconnect	InfiniBand EDR × 2
SAN	FibreChannel 32 Gbps × 2
OS	RedHat Enterprise Linux
ノード数	4台
メタデータストレージサーバ(MDT)	
機種名	FUJITSU ETERNUS AF250 S2
SSD	RAID1+0 [4D+4M] × 2 + 2HS RAID1+0 [3D+3M] × 1 + 2HS
ノード数	1台

データストレージ(OSS/OST)	
機種名	DDN SFA18KE × 1台 DDN SS9012 × 10 台
HDD	NL-SAS 14TB 7.2krpm × 730、RAID6 [8D+2P] 30 Device × 24 DCR Pool + 10HS
Interconnect	InfiniBand EDR × 8
搭載セット数	4
総容量	
物理容量	40.32 PB (Global Spareを除く)
実効容量	約30.44PB

- 大容量：30.44 PB （実効容量）
- 超高速アクセス性能：384 GB/s

スーパーコンピュータ「不老」コールドストレージ

フェーズ1: 2020年7月1日より稼働開始

機種名	PetaSite 拡張型 Library
カートリッジ数（総カートリッジ数 /最大搭載可能カートリッジ数)	88巻 / 88巻
総物理容量 / 最大搭載可能容量	484 TB / 484 TB
総ドライブ数	6
ODAサーバ数	1

フェーズ2: 2021年2月1日より稼働開始予定

機種名	PetaSite 拡張型 Library
カートリッジ数（総カートリッジ数 /最大搭載可能カートリッジ数)	1,092巻 / 1,980巻
総物理容量 / 最大搭載可能容量	6 PB / 10.89 PB
総ドライブ数	20
ODAサーバ数	4

- １度書き込み（追記）のみの光ディスクストレージ
- 実験データ等の長期データ保存用
- 理論上100年データ保持可能
- 水にぬれても読み出せる
- サービス終了後ユーザに光ディスクを返却

性能諸元（主な計算サブシステム群）

		「富岳」型	GPU	大容量メモリ	4ソケットCPU＋時刻指定
		Type I	**Type II**	**Type III**	**クラウド**
ノードあたり	CPU	A64FX ×1 (Armv8.2-A + SVE) 48+2コア、2.2GHz	Xeon Gold 6230×2 (Cascade Lake) 20コア、2.10-3.90 GHz	Xeon Platinum 8280M×16 (Cascade Lake) 28コア、2.70-4.00 GHz	Xeon Gold 6230×4 (Cascade Lake) 20コア、2.10-3.90 GHz
	メインメモリ	HBM2, 32GB	DDR4, 384GB	DDR4, 24TB	DDR4, 384GB
	GPU	-	Tesla V100×4 (Volta) HBM2, 32GB	Quadro RTX6000×4 (Turing) GDDR6, 24GB	-
	理論性能（倍精度浮動小数点演算性能とメモリバンド幅）	3.3792 TFLOPS(DP) 1,024 GB/s	・CPU 1.344 TFLOPS(DP)×2 140.784 GB/s×2 ・GPU 7.8 TFLOPS(DP)×4 900 GB/s×4	・CPU 2.4192 TFLOPS(DP)×16 140.784 GB/s×16	1.344 TFLOPS(DP)×4 140.784 GB/s×4
ノード数		2,304	221	2	100
ノード間接続		TofuインターコネクトD	InfiniBand EDR ×2	InfiniBand EDR	InfiniBand EDR
総理論性能		7.782 PFLOPS(DP) 2.359 PB/s	7.489 PFLOPS(DP) 857.8 TB/s	77.414 TFLOPS(DP) 2.253 TB/s	537.6 TFLOPS(DP) 56.314 TB/s
冷却方式		水冷	水冷	空冷	空冷
その他			SSD搭載	ローカルストレージあり	一部ノードは時刻指定実行に対応

「不老」と旧システムとのプログラム実行性能比較
主にFX100(2015.09.01稼働開始)とFX1000との性能比較

ベンチマーク名	性能	旧システムからの速度向上
TOP500（HPL） 連立一次方程式の求解	6.617 PFLOPS (世界 36位)（国内5位）→41位 (2020年6月：Type I サブシステム)	2.27 倍 2.910 PFLOPS (FX100)
HPCG 産業利用で多い疎行列反復解法	0.231 PFLOPS (世界 16位)（国内4位）→21位 (2020年6月：Type I サブシステム)	2.65 倍 0.087 PFLOPS (FX100)
GKVカーネルベンチ 名大独自開発のプラズマ シミュレーションベンチマーク	0.258 [秒] (kernel2_intgrl) Type I サブシステム	9.16 倍 2.36[秒](FX100)
Modylas 分子動力学ソフトウェア	20.72 [秒] Type I サブシステム	2.99 倍 61.9[秒] (FX100)
VOLR ボリュームレンダリング	1.29 [秒] Type III サブシステム	9.79 倍 12.6[秒](UV2000)
HPL-AI 人工知能で必要な演算	30.1 PFLOPS（96.6%） (世界 5位 ）（国内2位） (2020年11月：Type I サブシステム)	--倍

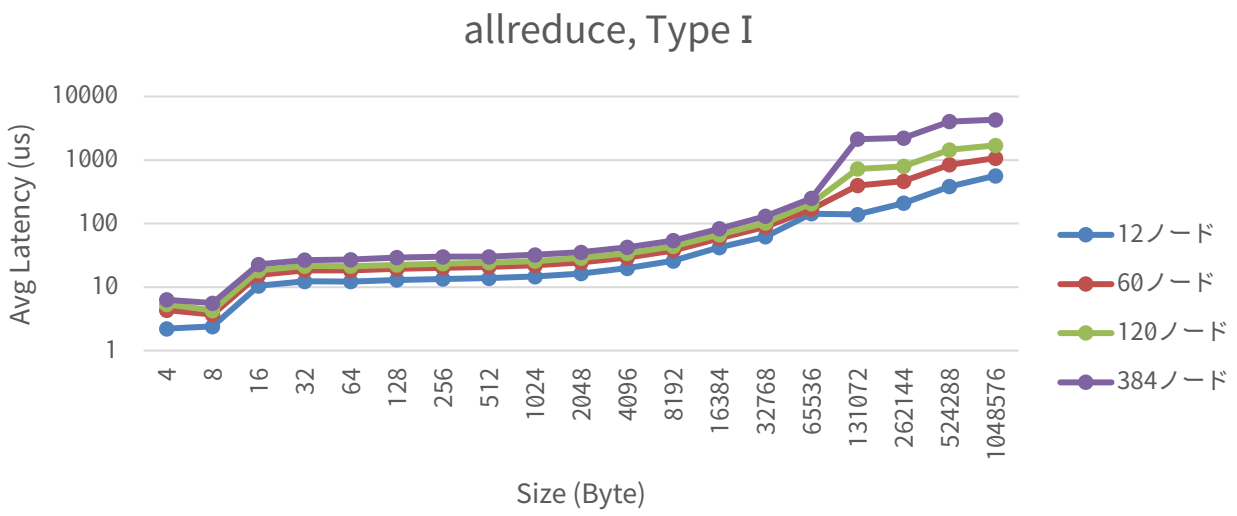
主要ベンチマーク*にて得られた「不老」各サブシステムの性能

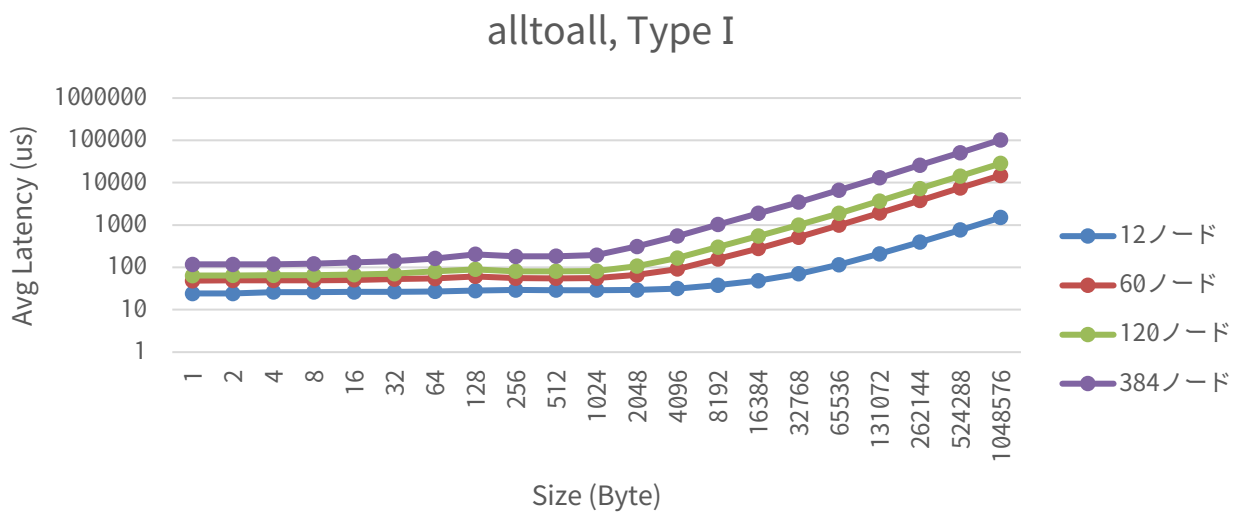
* STREAM以外は「富岳」が2連続で4冠となった4つのベンチマーク（残りはGraph500、「不老」では未測定）

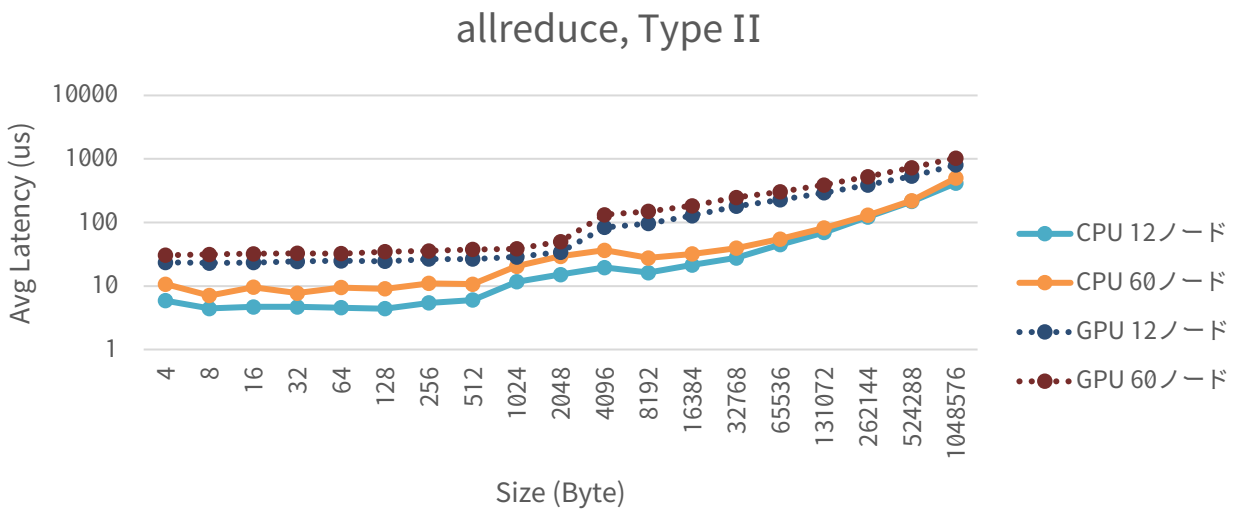
- メモリ性能：STREAM Benchmark
- 密行列演算性能：High Performance Linpack (HPL) Benchmark、TOP500
- 疎行列演算性能：High Performance Conjugate Gradients (HPCG) Benchmark
- HPL-AI：HPLの低精度版

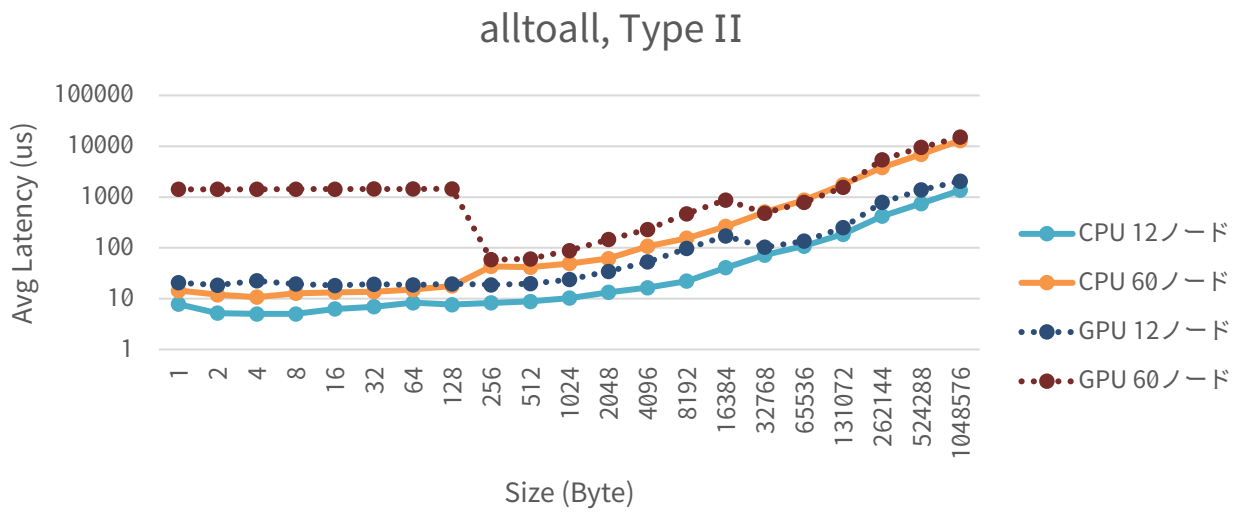
	Type I サブシステム	Type II サブシステム	クラウドシステム
STREAM	826 GB/s	2CPU　202 GB/s 1GPU　777 GB/s	4CPU　338 GB/s
HPL	1ノード　2.49 TFLOPS 全ノード　6.62 PFLOPS 世界36位→41位	1ノード　24.44 TFLOPS 全ノード　4.49 PFLOPS 世界58位	1ノード　3.25 TFLOPS
HPCG	1ノード　105.96 GFLOPS 全ノード　230.59 TFLOPS 世界16位→21位	1ノード　550.60 GFLOPS 全ノード　97.16 TFLOPS 世界34位	1ノード　61.78 GFLOPS
HPL-AI	全ノード　30.10 PFLOPS 世界5位		

ネットワーク性能（OSU Micro-benchmarks）









Type IIはOpenMPI(HPCX2.6)を利用、GPUの通信性能改善については継続して調査中

ストレージ性能（自作のテストプログラム）

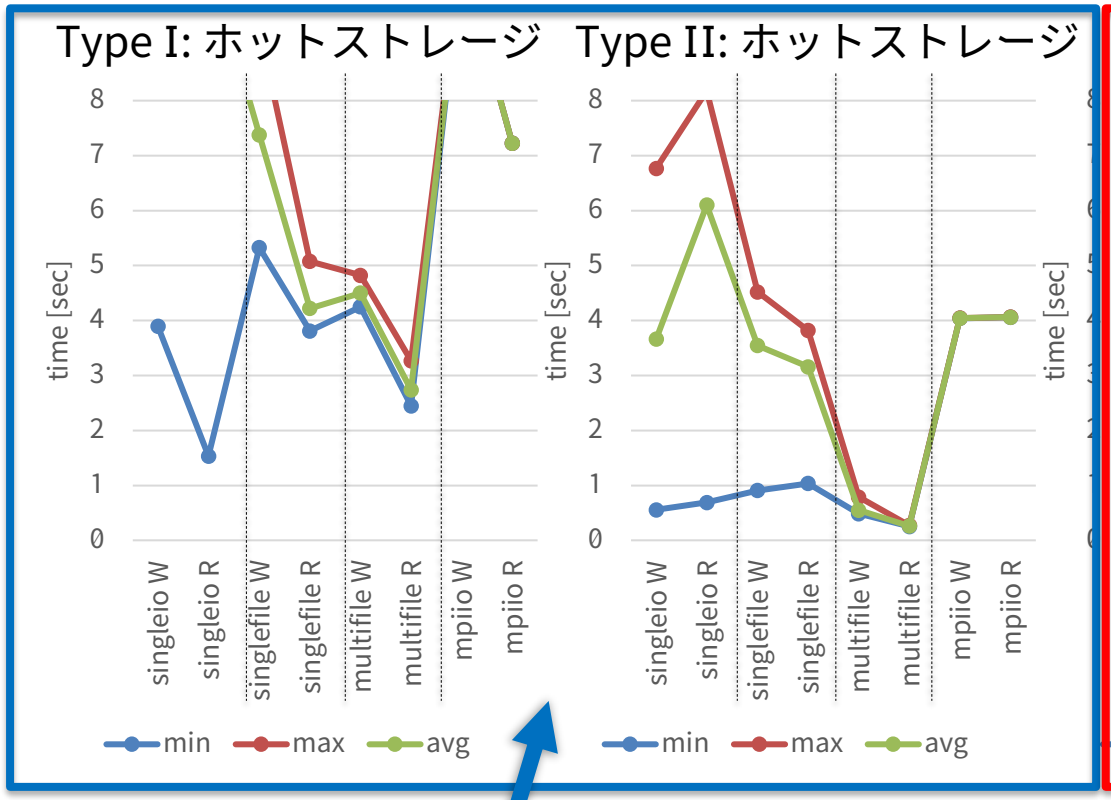
- 実際の使い方に近いシンプルなテストプログラムを作成し、運用中に測定してみた

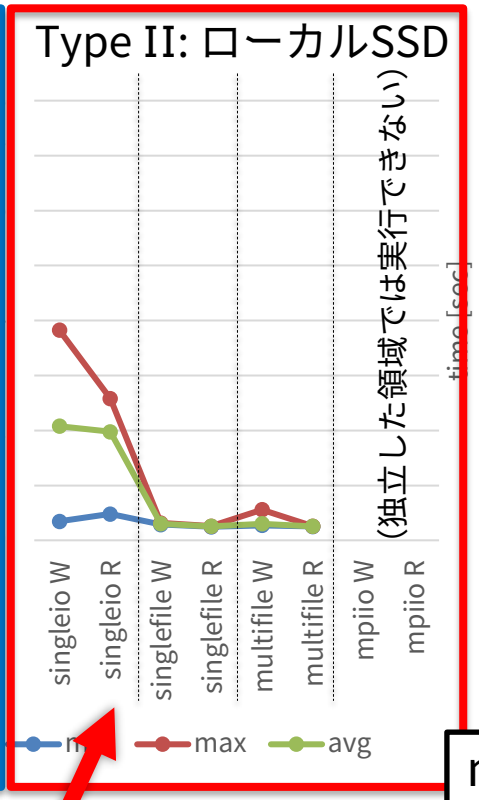
- プログラム内容（いずれもファイルのopen/close時間を含む）
  1. singleio：マスタープロセスのみがファイル操作、配付と集約はMPI通信
     - ノード数が増えた場合に全プロセス分を持てないことを想定し、逐次的に1対1通信
  2. singlefile：全プロセスがfread/fwriteで1ファイルを読み書き
  3. multifile：全プロセスが個別の1ファイルをfread/fwrite
  4. mpiio：MPI-IOで1ファイルを読み書き（MPI_File_set_view, MPI_File_write/read_all)

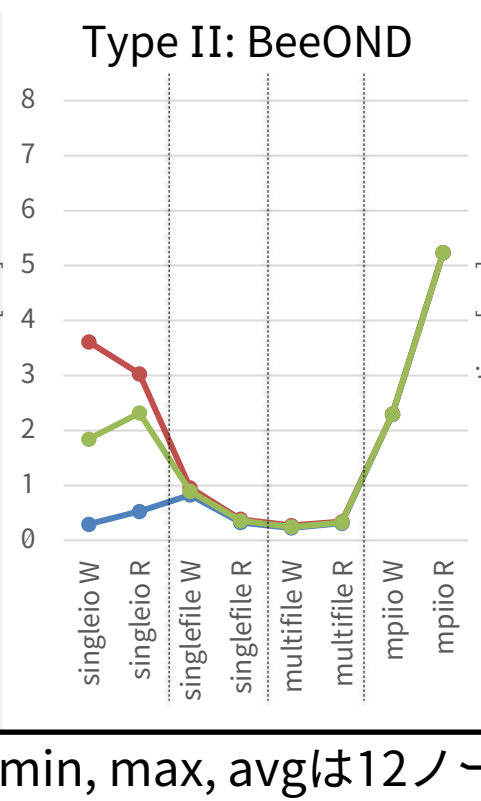
- 問題設定（プロセス数と容量）
  - 12プロセス、1プロセス/1ノード（Tofu-Dを考慮（12:mesh）、比較のためType IIも12ノード）
  - 1プロセスあたり100M要素を読み書き、データ型は全てdouble型（8byte）
    - プロセスあたり800MB、1秒で終了すれば9600MB/sec相当

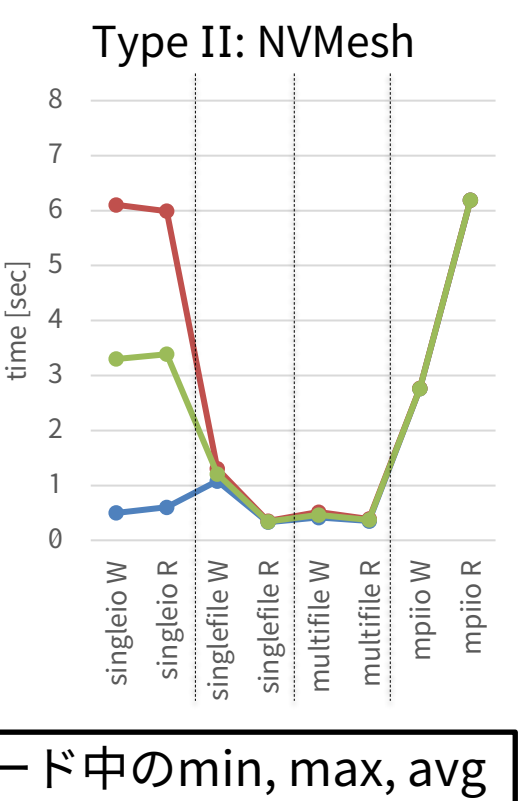
※1秒=9600MB/sec

テスト結果：12ノード、1プロセス/1ノード、プロセスあたり100M要素









min, max, avgは12ノード中のmin, max, avg

- ホットストレージの性能は計算ノードとストレージ両方の利用状況に影響を受けるためばらつきやすい
- SSDはノード外の影響がないため安定して高速
- mpiioの性能は今ひとつ、特にBeeOND・NVMeshが遅い
  - 東工大TSUBAME3でも同様にBeeONDのmpiioは遅いようで、使い方の想定があっていないと思われる。
- SSDによる劇的な性能向上は観測できていないが、singlefileでも高性能、他のジョブの影響も受けない利点有。

まとめ

- 2020年7月1日に運用を開始したスーパーコンピュータ「不老」の概要と性能を紹介した

- 「不老」の狙い・特徴
  - 不足してきた計算性能を補うだけではなく、新たな需要に応えるために、特徴の異なる4種類の計算サブシステムと2種類のストレージから構成される「複合型」のスパコン
  - 電力や運用についてはこの次の発表を参照

- 「不老」の性能
  - Type Iサブシステムは旧システム（FX100）と比べてHPLやHPCGで2倍以上の性能
  - アプリケーション性能では10倍近い性能向上（GKV、VOLR）
  - 単体ノード性能ではGPUを備えるType IIサブシステムも非常に高い性能を発揮