

「不老」 システム概要紹介

名古屋大学 情報基盤センター
片桐孝洋

第1回 スーパーコンピュータ「不老」ユーザ会
2020年8月31日（月） 14:05-14:25
オンライン開催（ZOOM）

まとめ

▶ スパコン「富岳」ベース

- ▶ 2020年6月のTOP500で世界一 (415PFLOPS) のスーパーコンピュータ「富岳」の同型機を
正式運用開始 (2020年7月1日～、Type I サブシステム)

▶ スーパーコンピュータ「不老」の特徴

1. **数値計算とAIを融合するスパコン**
2. **敷居が低い** (研究目的等に問題無なら有資格者が誰でも利用可)
3. **AI/機械学習**研究を加速するGPU (Type II サブシステム)
4. **100年データ保存可能**な最大6PBの光ディスクストレージ
5. 充実した可視化システム
6. **湧水を用いたエコ冷却**、夏季電力を制御するシステム



まとめ

- ▶ 異常気象解析、津波シミュレーション、
遺伝子解析、医療診断支援、自動運転から
宇宙の仕組みの解明まで**幅広い研究**をサポート
- ▶ スーパーコンピュータ「不老」特有のアプリケーション例
 - ▶ 台風のメカニズム解析
 - ▶ 医用画像処理
 - ▶ 自動運転
 - ▶ プラズマシミュレーション
 - ▶ 高精細可視化

システム紹介



名古屋大学
NAGOYA UNIVERSITY

スーパーコンピュータ「不老」の役割

1. 全国共同利用・共同研究拠点として学内外へ計算資源提供

- ▶ **全国共同利用・共同研究拠点**として国が位置づけ
 - ▶ 全国の研究者の世界トップレベル研究を強力に支援

スーパー
コンピュータ
「富岳」

簡便な
移行支援

世界トップレベル
研究の支援

HPCI (High Performance
Computing Infrastructure) 利用者

名大拠点利用者

JHPCN利用者

名大「不老」
Type I システム
(富岳型ノード)



高性能化／
特殊処理対応／
長時間実行支援

世界トップレベル
研究成果創出

2. ものづくり企業支援 (地域イノベーションコア形成)

- ▶ **産業利用制度 (公開、非公開)**
- ▶ 計算機利用型講習会による並列処理・大規模計算普及 (地域特有の中小企業支援)

3. 新しい計算需要に向けたサービス開拓

- ▶ **データサイエンス (ビッグデータ)**、AI基盤の提供による新サービスの開拓

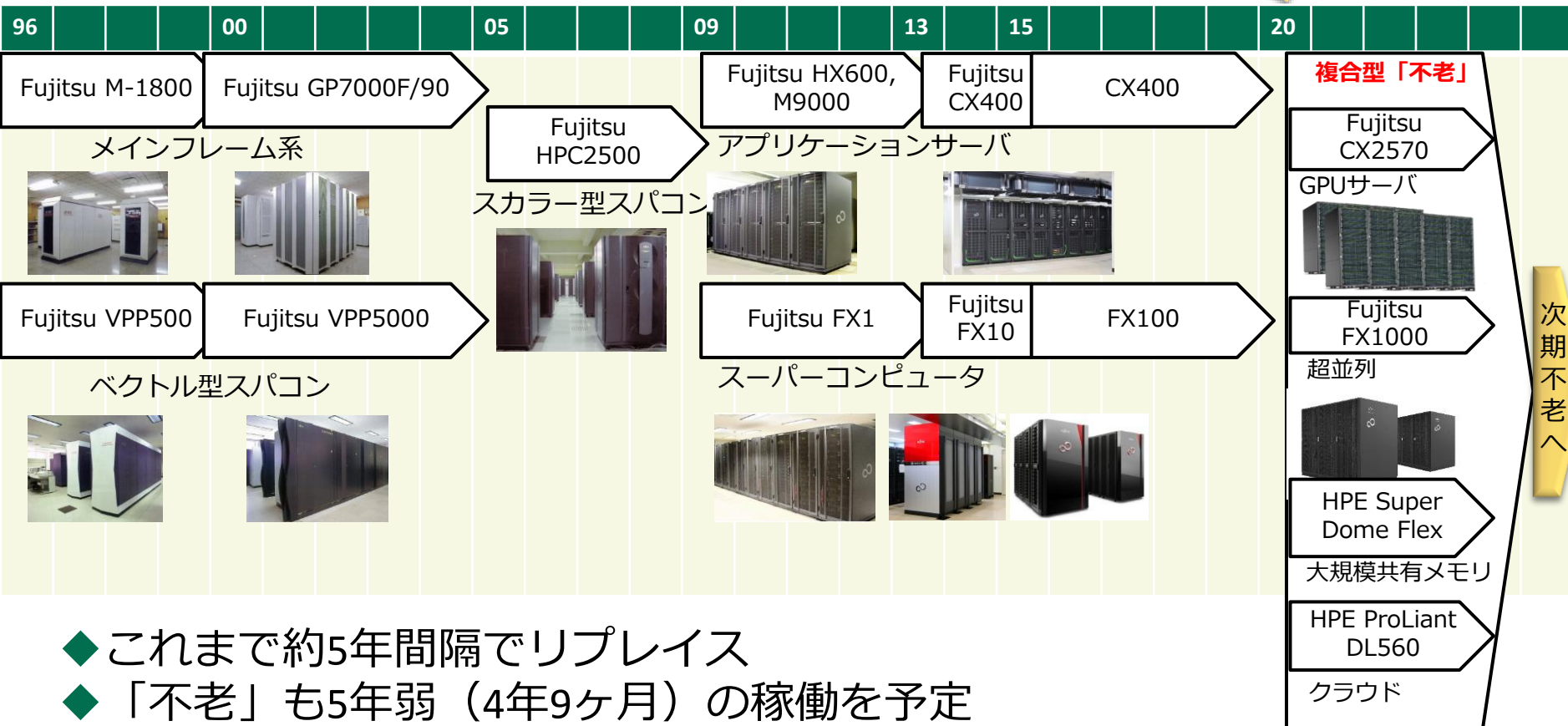
4. 指定国立大学として重要な役割

- ▶ **数理・データ科学教育**
- ▶ 人材育成・研究力強化・社会との連携

- ・数理データ科学分野の人材育成
- ・AI技術基盤提供 (大規模AI計算基盤、大規模ストレージ、サイバーフィジカル)
- ・技術コンサルティング・・・等による社会貢献

名古屋大学情報基盤センターの スパコンの歴史

スーパーコンピュータ
「不老」導入



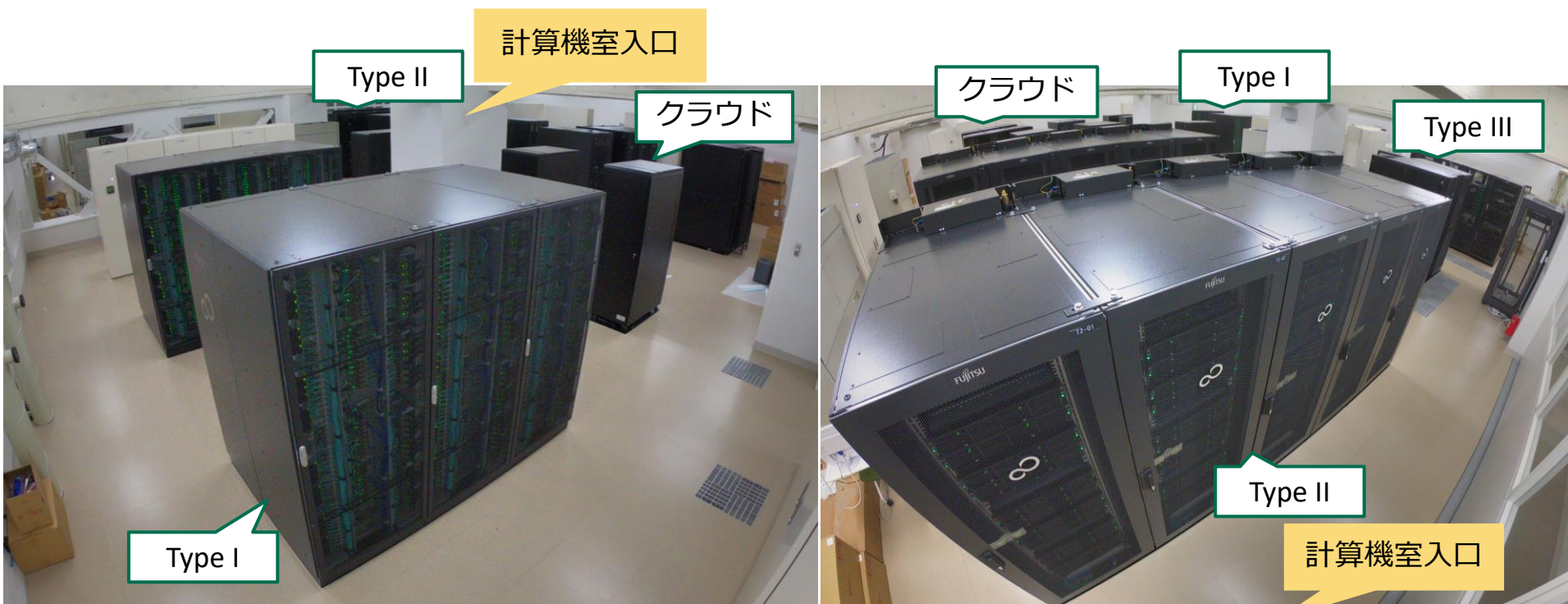
- ◆これまで約5年間隔でリプレイス
- ◆「不老」も5年弱（4年9ヶ月）の稼働を予定

導入の背景

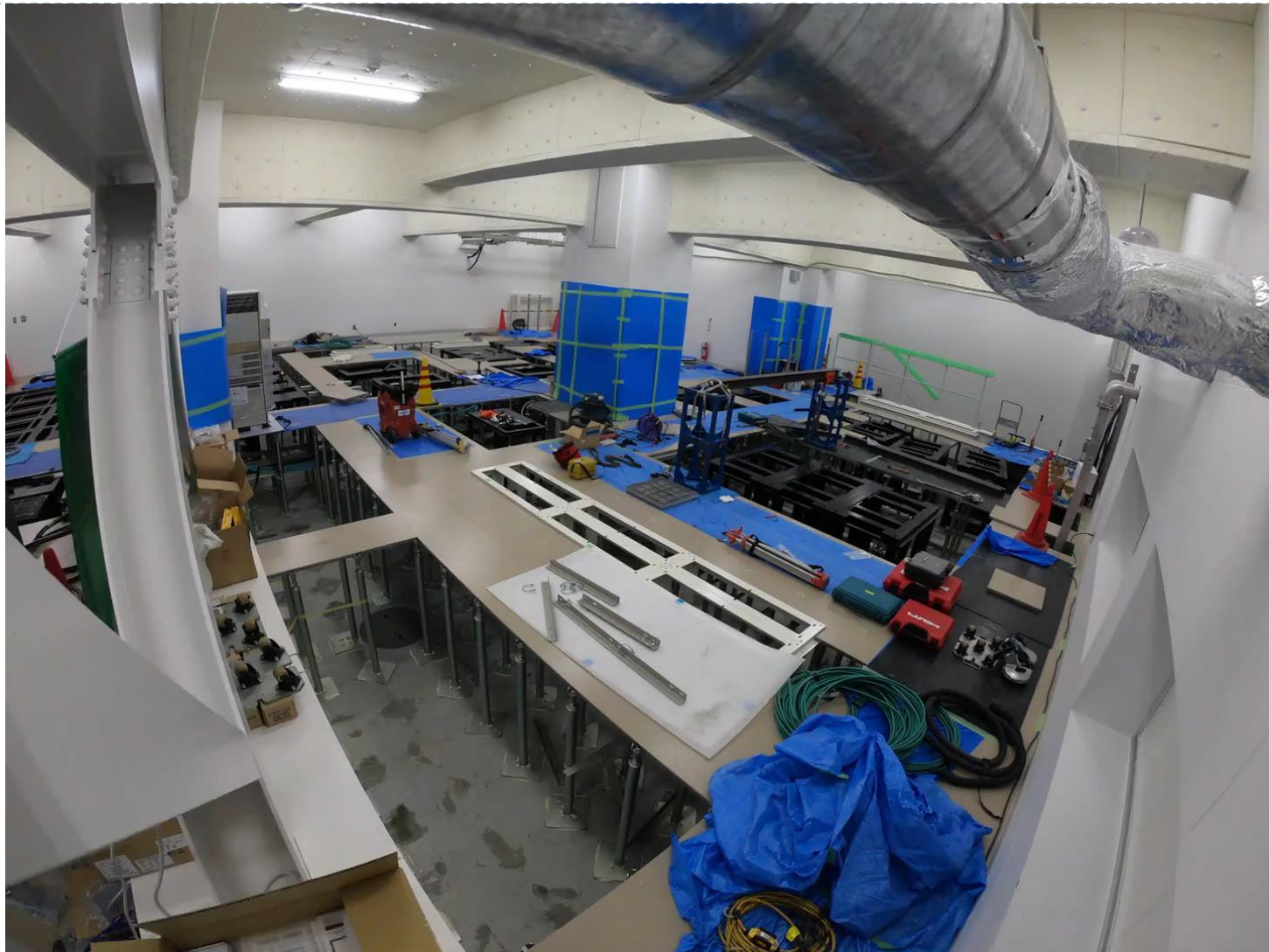
- ▶ **研究のデジタル化 (デジタルサイエンス)**
 - ▶ コンピューティングを活用した研究の広まり
- ▶ **AI/機械学習研究の増大**
 - ▶ 自動運転、医療、創薬
- ▶ **シミュレーション研究の増加**
 - ▶ 異常気象、津波など国民の安全に密接にかかわる現象
 - ▶ 生命・宇宙などの基礎科学
- ▶ **データの爆発的増大**
 - ▶ 元データ、解析結果、AI学習結果など
- ▶ 従来のスパコンでは**明らかな能力不足**

設置状況

- ▶ 7月1日、スーパーコンピュータ「不老」が稼働開始しました。現在も順調に稼働中です。
- ▶ 名古屋大学 情報基盤センター 本館地下1階の様子



スーパーコンピュータ「不老」 Type I サブシステム設置動画



スーパーコンピュータ「不老」 Type II サブシステム設置動画

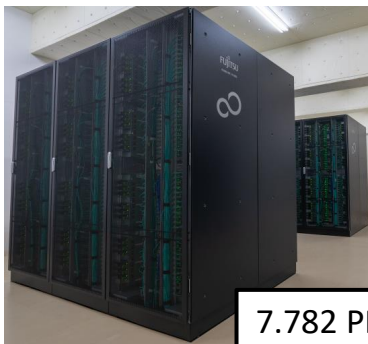


スーパーコンピュータ「不老」 全体図



実際に入ったもの（主な構成要素）

Type I, II, III, クラウドの合計で15.886PFLOPS
(旧システムの約4倍)



7.782 PF

Type Iサブシステム

FUJITSU Supercomputer FX1000

「富岳」型



7.489 PF

Type IIサブシステム

FUJITSU Server PRIMERGY CX2570 M5

GPUスパコン



77.414 TF

Type IIIサブシステム

HPE Superdome Flex

大容量メモリ・可視化



537.6 TF

クラウドシステム

HPE ProLiant DL560

バッチ&インタラクティブ



30 PB

ホットストレージ

FUJITSU PRIMERGY RX2540 M5

FUJITSU ETERNUS AF250 S2

DDN SFA18KE

DDN SS9012



484 TB → 6 PB

コールドストレージ

SONY PetaSite Library

↓2020年2月更新

SONY PetaSite 拡張型 Library



性能諸元（主要サブシステム群）

		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	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

消費電力・省電力対策

▶ 最大消費電力

サブシステム名	消費電力
TypeIサブシステム	628.1kVA
TypeIIサブシステム	393.5kVA
TypeIIIサブシステム	21.6kVA
クラウドシステム	93.0kVA
ストレージ	49.9kVA
フロントエンド	19.6kVA
運用管理システム他	52.3kVA
冷却設備	641.9kVA
合 計	1,899.9kVA

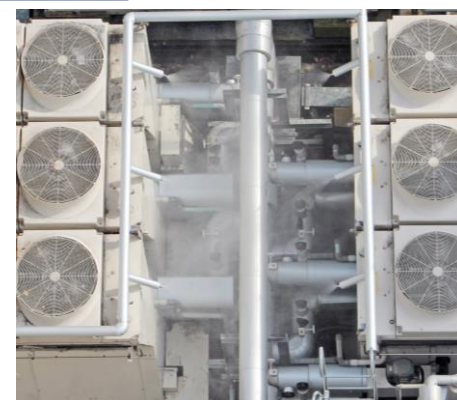
▶ 電力可視化



PUE < 1.4?

▶ 湧水を用いた冷却

- ▶ 地下の湧水を活用したら総合評価時加点
- ▶ 屋外チラーに散水して冷却

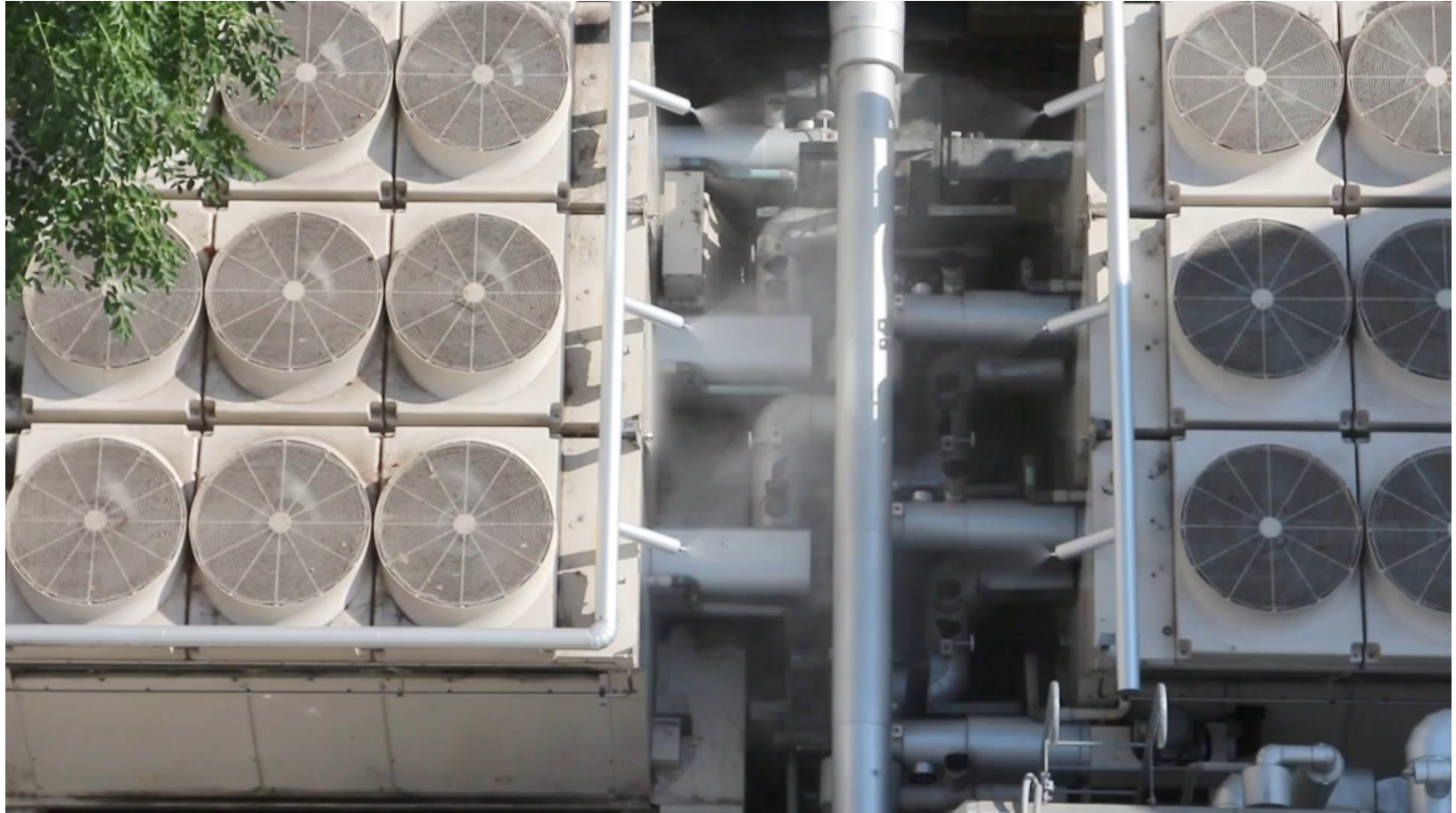


湧水による冷却システム

- ▶ 情報基盤センターの地下は夏季でも18℃程度の湧き水が毎分30L程度湧く
- ▶ この湧き水は、地下からポンプで吸い上げて雨水扱いで捨てていた
- ▶ 今回の仕様で、湧き水を冷熱源として使用する場合は加点点
 - ▶ 冷却水としての利用許可・水質検査済み

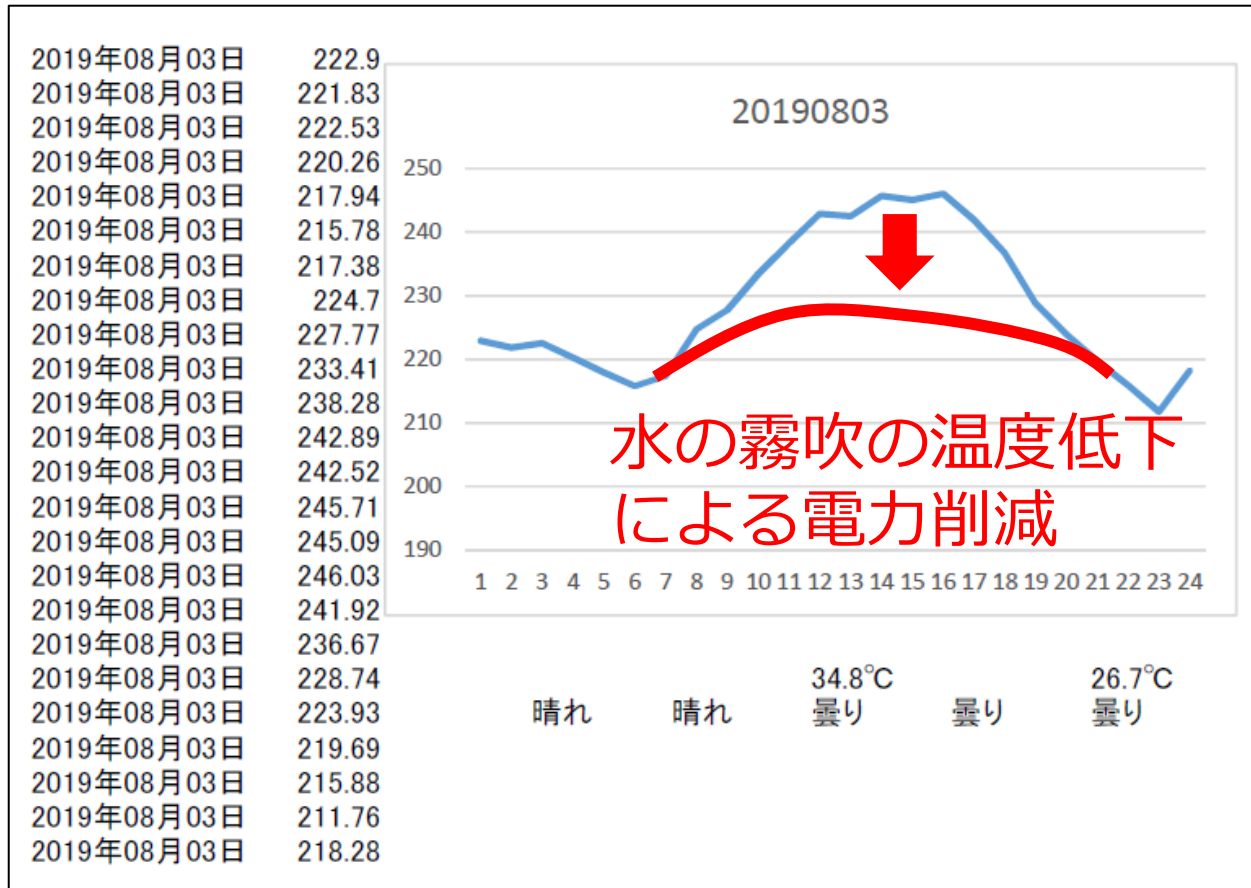


湧水による冷却システム



湧水による冷却システム

- ▶ 気温の高い
4月から11月
の間で利用
- ▶ 夏季の1日
(2019年8月3日)
の(旧) FX100
システムの
水冷チラーの
電気使用量
(KW)



年間数百万円程度の電気代削減を予想



使用最大電力の動的制御機構

- 監視ソフトウェアから一定時間毎に電力値を取得
- 出力された電力値と、あらかじめ規定したシステム全体の使用最大電力の上限値を比較し、最大電力の上限を超えないよう、計算ノードやジョブ実行可能範囲を制限

ピーク電力（例：1.5MW）



電力マージン（例：20%）（1.2MW）

12時間
以下で
変更
可能

電力上限（1.2MW）

定期的と比較
（例：10分単位）

ログ整形→比較→条件分岐→
リソースグループ停止/開始
計算ノード制限

Type I / Type II / Type III / クラウドサブシステム
リソースユニット

リソースグループX リソースグループW

リソースグループY リソースグループZ



定期的に格納

電力値ログ
電力値ログ
電力値ログ

ネットワークデバイス用
監視ソフトウェア



電力量センサー

Type I
サブシステム

Type II
サブシステム

Type III
サブシステム

クラウド
システム



ネットワークスイッチ群



周辺装置群



Hot Storage



Cold Storage



空調設備など

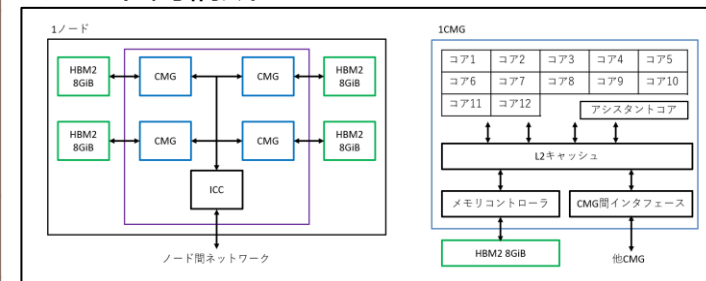
各サブシステムの仕様と特徴： Type I サブシステム



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ソケット
	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

- 世界初正式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



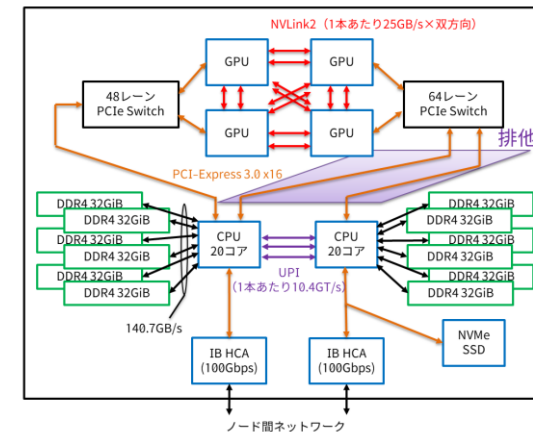
各サブシステムの仕様と特徴： Type II サブシステム



- 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

機種名	FUJITSU Server PRIMERGY CX2570 M5
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

ノード内構成



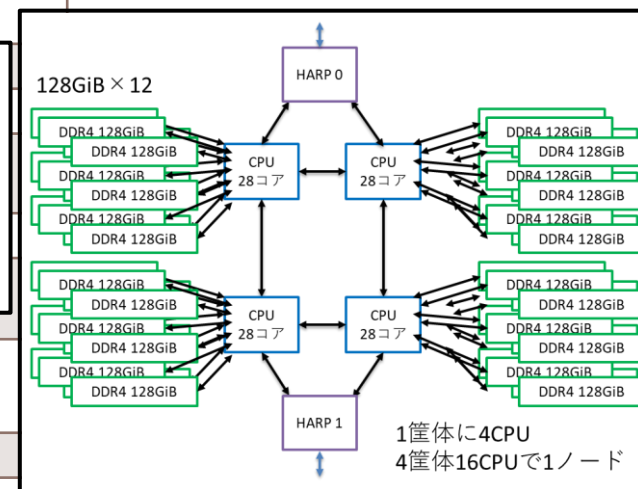
各サブシステムの仕様と特徴： Type III サブシステム



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 2419.2 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

ノード内構成



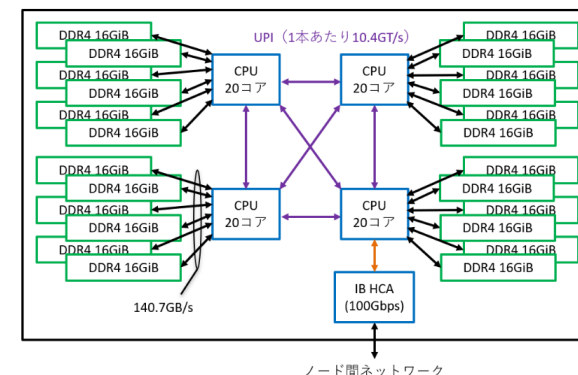
各サブシステムの仕様と特徴： クラウドシステム



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

ノード内構成



ノード間ネットワーク

ホットストレージの仕様と特徴

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

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



コールドストレージの仕様と特徴

フェーズ1: 2020年7月1日より稼働開始

機種名	PetaSite Library
総スロット数 (最大搭載可能カートリッジ数)	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
総スロット数 (最大搭載可能カートリッジ数)	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

その他の構成要素

- ▶ フロントエンドシステム（ログインノード群）
 - ▶ 合計25ノード
 - ▶ Type I用も含めて全てXeon Gold 6248(Cascade Lake)×2、一部にTesla V100(PCIe)搭載
- ▶ オンサイト利用装置・画像処理装置
 - ▶ センター内の利用者支援室・可視化室に設置、訪問者が利用できる機器
 - ▶ SINETやIBでシステムに接続、持ち込みUSB機器（ハードディスク）を利用してデータの出し入れが可能
 - ▶ 画像処理装置はSINET(10G)と可視化設備に接続
 - ▶ オンサイト利用装置はコールドストレージ単体ディスクドライブを装備

運用形態

- ▶ **Type I～Type III、クラウドの4サブシステムは1つの申込で利用可能、かつ、共有ファイルシステム（ホットストレージ）で連結 ⇒ シームレスなデータ移動**
- ▶ Type I, IIサブシステムは完全にバッチ処理運用
- ▶ Type IIIサブシステムはノード毎に別の運用形態
 - ▶ 1ノードはバッチ処理運用
 - ▶ 1ノードは可視化室の機器に接続
 - ▶ 可視化室で直接操作、SSH接続、NICE DCVによるリモートデスクトップ接続
- ▶ クラウドシステム
 - ▶ 一部ノードはバッチ処理運用、一部ノードはUNCAIによる時刻指定利用
 - ▶ 利用状況にあわせて割合を調整していく予定
- ▶ コールドストレージ
 - ▶ 専用ログインノードから専用コマンドで操作、ホットストレージとのデータコピーが可能
- ▶ バッチ処理システム運用上の工夫
 - ▶ **ノード共有（1/4ノード）キュー、優先キュー（消費係数2倍）、インタラクティブキュー、節電時の縮退運用時には止まるextraキュー**

利用制度・課金制度の特徴

▶ 課金制度の特徴

- ▶ 前払い、システム共通の利用ポイント制度
 - ▶ 購入した利用ポイントを全システムで利用できる、消費係数がシステムごとに異なる
 - ▶ **初期費用は1ユーザ10,000円** ※登録料という扱いだが利用ポイントに変換される
 - ▶ 一度に50万円以上購入すると1.25倍のポイントになる
- ▶ 優先キュー：ポイント消費2倍で投げられるキュー
 - ▶ 旧システム運用後半は優先キューすら混む事態
- ▶ ログインノードの利用も課金対象、ストレージは一定量を超えたら課金対象

▶ 「不老」の利用制度

- ▶ 基本的には従来の制度を継承、いくつかの新制度を導入
- ▶ **グループ利用**：20アカウントまでで1グループ、グループ内のポイント融通が可能
- ▶ **準占有制度**：1時間以内の実行を保証、空き時間はdebugキューで活用
- ▶ **クラウドノード予約利用**：Web（UNCAI）で予約を行い実行時間帯を確定させての利用
 - ▶ 自動的にバッチが起動する**時刻指定バッチジョブ**と
sshでログインして利用する**時刻指定インタラクティブ実行**



参加募集

- ▶ スーパーコンピュータ「不老」TypeIサブシステム（「富岳」型ノード）
を無料で利用できる、プログラミング講習会を開始
 - ▶ 完全遠隔講習会
 - ▶ 企業の方も参加できます
 - ▶ 登録上限（新規ユーザ10名程度）があります。早めにご登録ください。
- 1. **第1回スーパーコンピュータ「不老」利用型講習会 MPI（初級）**
2020年9月28日（月）10：00～17：30
- 2. **第2回スーパーコンピュータ「不老」利用型講習会 OpenMP（初級）**
2020年9月30日（水）10：00～17：30
- 3. **第3回スーパーコンピュータ「不老」利用型講習会 OpenACC（初級）**
2020年10月7日（水）10：00～17：30
 - ▶ Type II サブシステム（GPU）

申込：

<https://www2.itc.nagoya-u.ac.jp/cgi-bin/kousyu/csview2.cgi>