スーパーコンピュータ「不老」
～大規模コールドストレージ導入による データサイエンス支援を指向したスパコン～

名古屋大学　情報基盤センター
片桐孝洋

Gfarmシンポジウム2020
2020年10月9日(金) 14:10~14:50　AP秋葉原O+Pルーム＆オンライン開催
〒110-0006　東京都台東区秋葉原1-1　秋葉原ビジネスセンター

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. 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）
  2. **AI/機械学習**研究を加速するGPU（Type Ⅱ サブシステム）
  3. 100年データ保存可能な最大6PBの光ディスクストレージ
  4. 充実した可視化システム
  5. 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

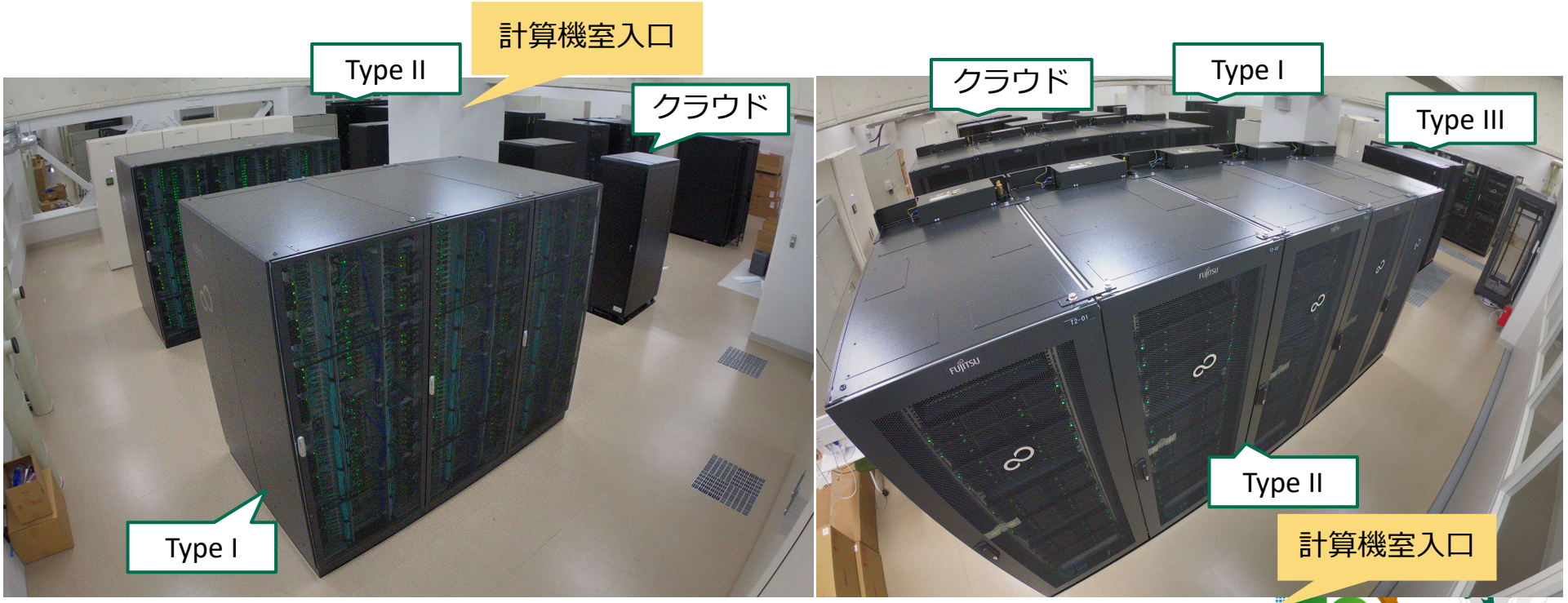
導入の背景

- **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

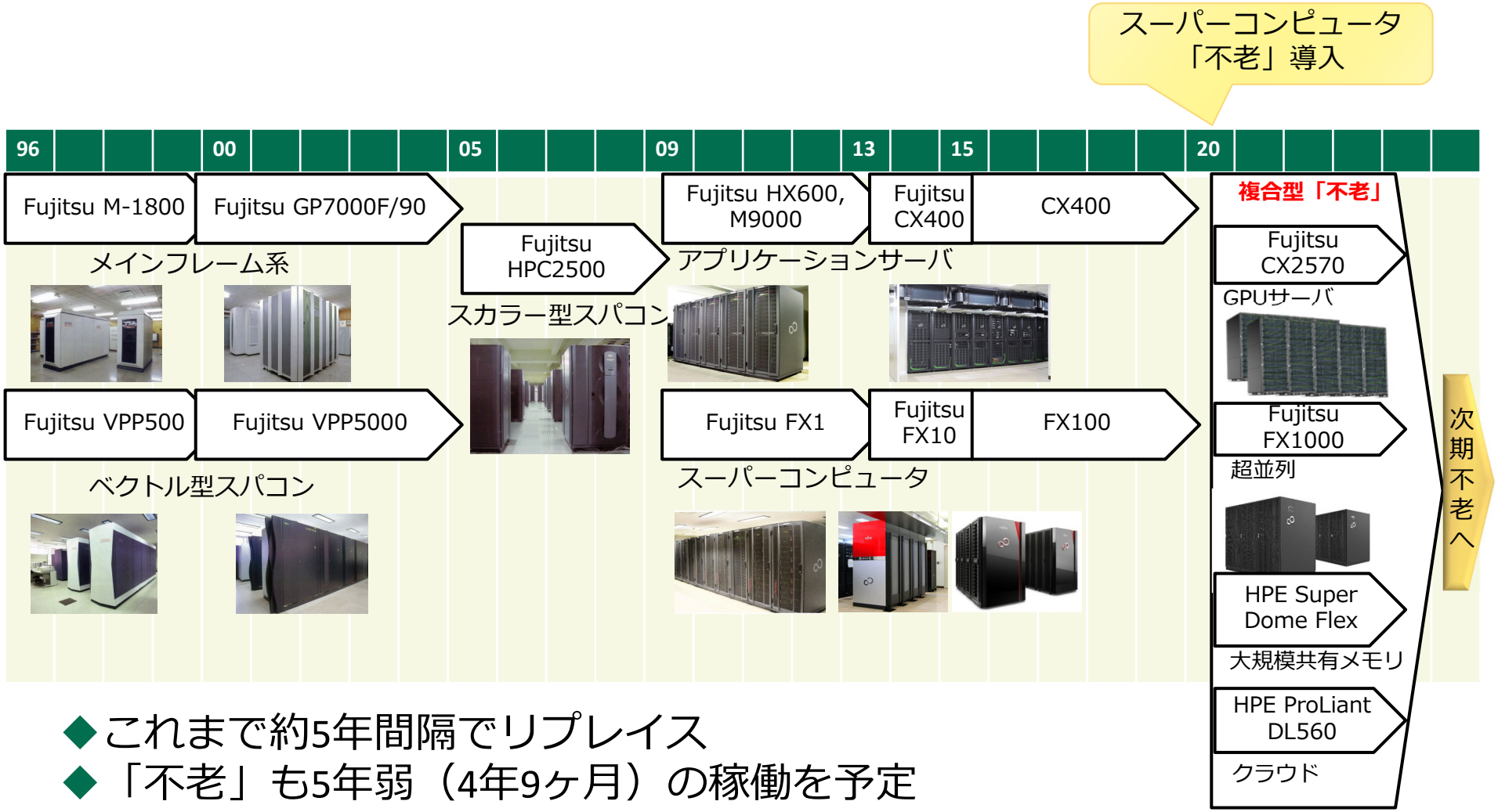
- 従来のスパコンでは**明らかな能力不足**

設置状況

- 2020年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



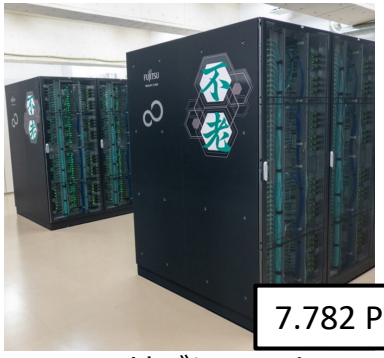
名古屋大学情報基盤センターのスパコンの歴史



- 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

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）



Type Iサブシステム
FUJITSU Supercomputer FX1000
「富岳」型

Type IIサブシステム
FUJITSU Server PRIMERGY CX2570 M5
GPUスパコン

Type IIIサブシステム
HPE Superdome Flex
大容量メモリ・可視化

クラウドシステム
HPE ProLiant DL560
バッチ＆インタラクティブ

ホットストレージ
FUJITSU PRIMGERY RX2540 M5
FUJITSU ETERNUS AF250 S2
DDN SFA18KE
DDN SS9012

コールドストレージ
SONY PetaSite 拡張型 Library
↓ 2020年2月更新
SONY PetaSite 拡張型 Library

484 TB → **6 PB**

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

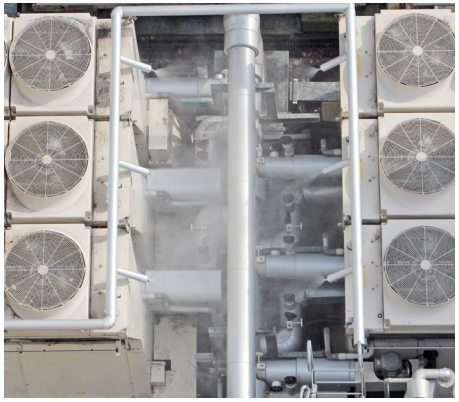
最大消費電力

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

湧水を用いた冷却

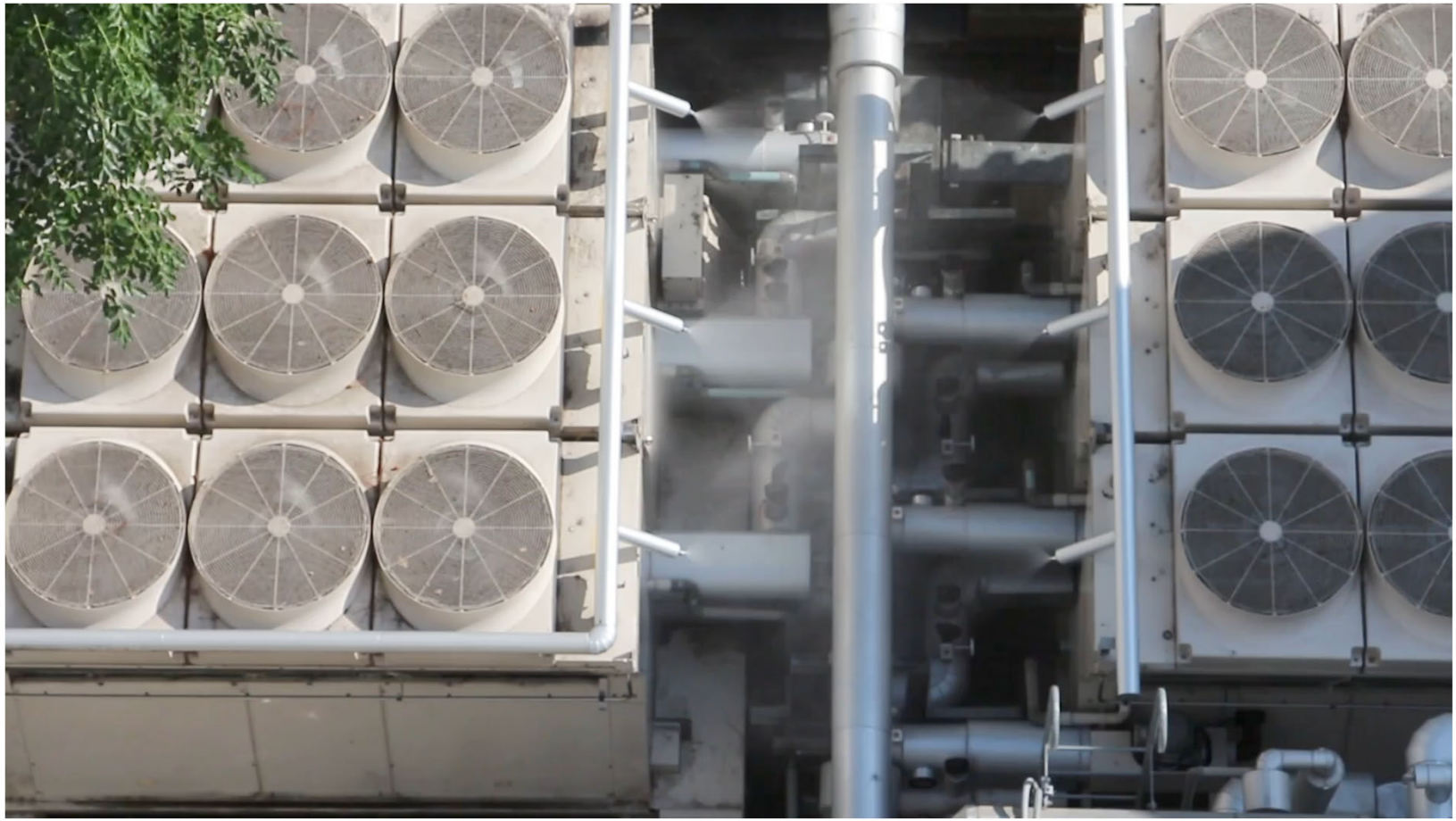
- 地下の湧水を活用したら総合評価時加点
- 屋外チラーに散水して冷却



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
  - 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

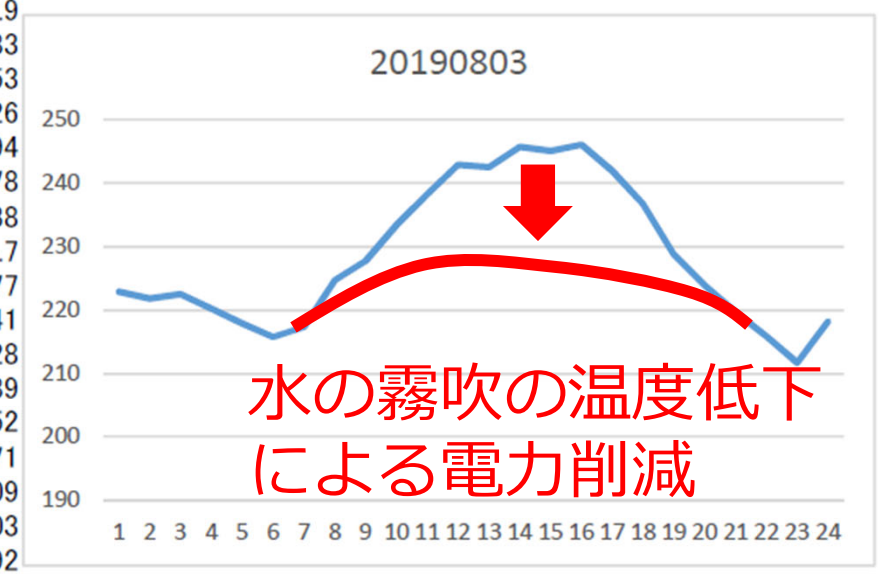


名古屋大学 NAGOYA UNIVERSITY

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)

2019年08月03日	222.9
2019年08月03日	221.83
2019年08月03日	222.53
2019年08月03日	220.26
2019年08月03日	217.94
2019年08月03日	215.78
2019年08月03日	217.38
2019年08月03日	224.7
2019年08月03日	227.77
2019年08月03日	233.41
2019年08月03日	238.28
2019年08月03日	242.89
2019年08月03日	242.52
2019年08月03日	245.71
2019年08月03日	245.09
2019年08月03日	246.03
2019年08月03日	241.92
2019年08月03日	236.67
2019年08月03日	228.74
2019年08月03日	223.93
2019年08月03日	219.69
2019年08月03日	215.88
2019年08月03日	211.76
2019年08月03日	218.28

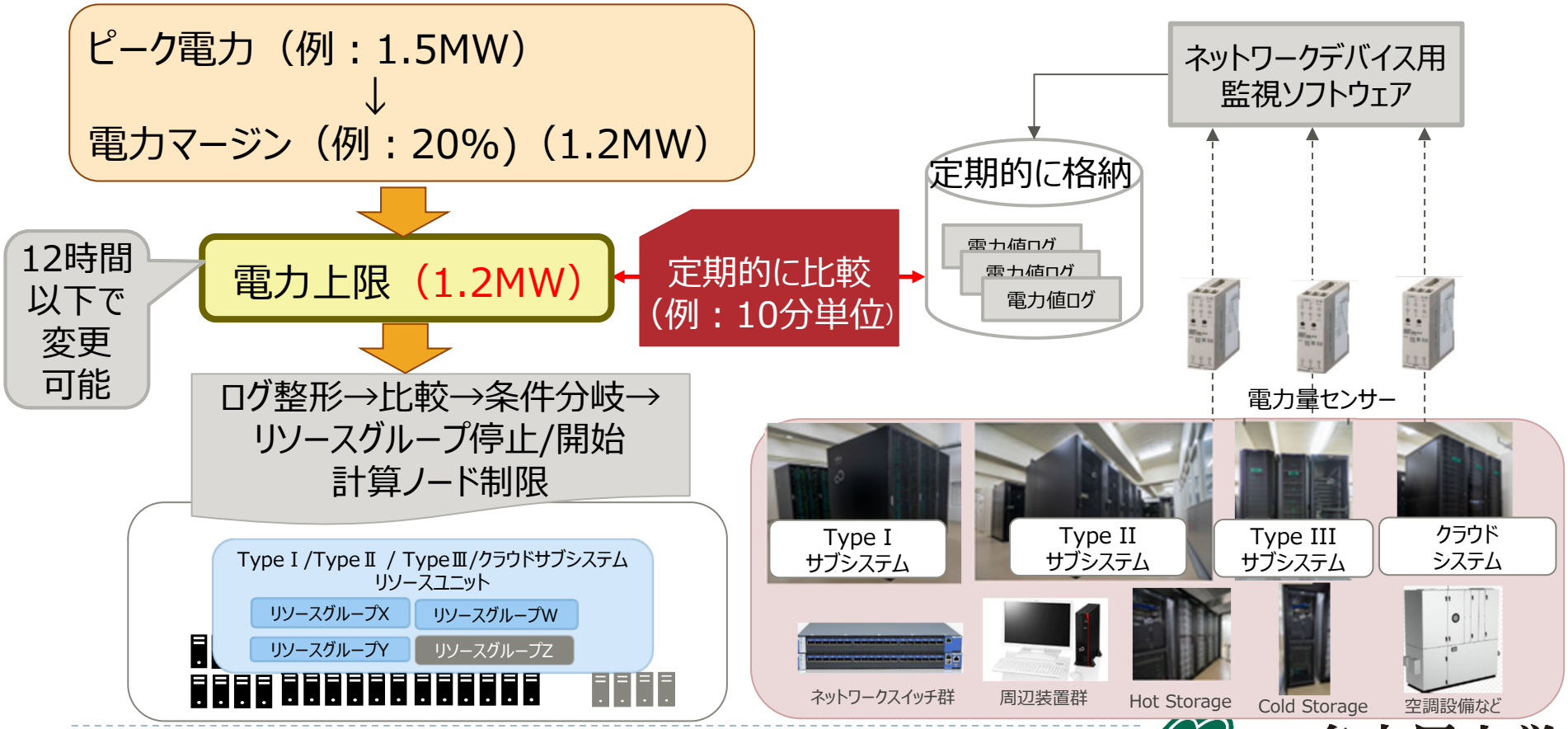


年間数百万円程度の電気代削減を予想

名古屋大学 NAGOYA UNIVERSITY

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

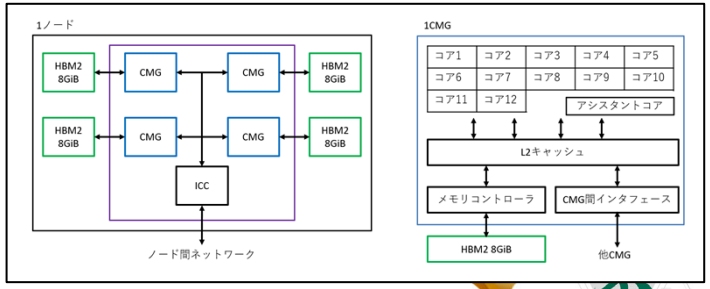


各サブシステムの仕様と特徴：Type I サブシステム

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, 4ソケット相当
	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

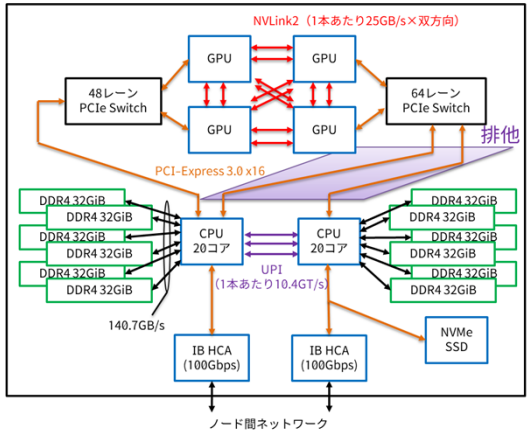


各サブシステムの仕様と特徴：Type II サブシステム

機種名		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



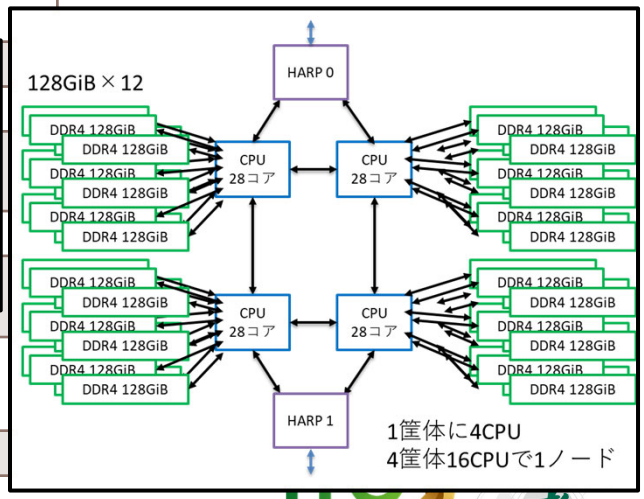
名古屋大学 NAGOYA UNIVERSITY

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

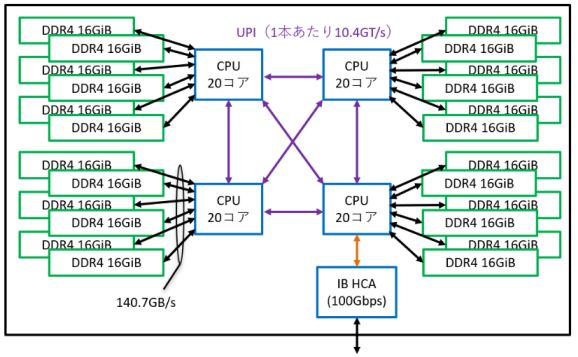


各サブシステムの仕様と特徴：クラウドシステム

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



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 &tiems; 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拡張型 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

- **TypeI～TypeIII、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

ベンチマーク結果

スーパーコンピュータ「不老」ベンチマーク結果

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）** **(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）** **(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] TypeⅢサブシステム	**9.79倍** 12.6[秒](UV2000)

大島聡史准教授提供

各種ベンチマークによる性能評価

- FX1000、Cascade Lake、V100それぞれの性能自体は既知の部分が多いため、同じ問題を各サブシステムで実行した場合の性能など、「不老」ならではの比較結果を紹介する
- 主要ベンチマーク
  - Stream, HPL, HPCG
- 通信性能
  - OSU Micro-Benchmarks
- ストレージ
  - 自作の単純なテストプログラム

コンパイラなど（特に断りのない限り）
- Type Iサブシステム：TCS 1.2.26
- Type IIサブシステム：Intel 2019.5.281, CUDA 10.2, PGI 20.4
- クラウドシステム：Intel 2019.5.281

Type IIのPCIe/NUMA構成は先に示したとおり
（「片寄せ」で「SNC無効」）

主要ベンチマーク性能：

大島聡史准教授提供

Stream Triad (N=20,000,000)

- A64FXのHBM2の速さが良くわかる結果
- 4CPUのクラウドシステムは2CPUのType IIの倍は出ない

- Type Iサブシステム
  - 1ノード：826 GB/s
    - コンパイル時オプション -Kfast,openmp,zfill
    - 実行時環境変数 XOS_MMM_L_PAGING_POLICY=demand:demand:demand
    - 特にzfillとdemand設定が大きく影響、CとFortranの差はなし
- Type IIサブシステム
  - 1ノード 2CPU：202 GB/s　(-xHost -qopenmp -qopt-zmm-usage=high)
    - -qopt-streaming-stores は
      変更しても向上せず（デフォルトはauto）
  - 1ノード 1GPU：777 GB/s
    ※streamベンチコードにOpenACC指示文を入れて測定
- クラウドシステム
  - 1ノード 4CPU：338 GB/s　（〃）

主要ベンチマーク性能：

大島聡史准教授提供

HPL

- GPUの性能の高さと
  A64FXの効率の良さがよくわかる結果

- Type Iサブシステム
  - 1ノード：2.4849 TFLOPS： 2.4849/3.3792=73.5%
  - 全系：6.6178 PFLOPS： 6.6178/7.782=85.0%
    - TOP500 2020年6月版で世界36位・国内5位（「富岳」、ABCI、OFP、TSUBAME3.0の次）
    - メモリサイズが小さめ＝大きな問題を解けないため1ノードの効率は控えめ、
      大規模では高効率
- Type IIサブシステム（GPU利用）
  - 1ノード：24.440 TFLOPS： 24.440/33.888=72.1%
  - 全系：4.880 PFLOPS（測定時期の都合でTOP500未登録、〃51位相当）：
    4.880/7.489=65.2%
- クラウドシステム
  - 1ノード：3.25 TFLOPS：3.250/5.376=60.5%
  - Type I, II は富士通による最適化・測定
  - クラウドはIntelコンパイラ2019.5.281（のMKL）に
    付属のmp_linpack

主要ベンチマーク性能：

大島聡史准教授提供

HPCG

- HPL同様に、GPUの性能の高さとA64FXの効率の良さがよくわかる結果

- Type Iサブシステム
  - 1ノード：105.956 GFLOPS： 105.956/3379.2=3.14%
  - 全系：230.594 TFLOPS： 230.594/7782=2.96%
    - HPCG 2020年6月版で世界16位・国内4位、「不老」以上の性能で2%以上は「富岳」の2.6%のみ
    - 富士通による最適化・測定
- Type IIサブシステム（GPU利用）
  - 1ノード：550.6 GFLOPS： 550.6/33888=1.62%
  - 100ノード：48.2544 TFLOPS： 48.2544/33888=1.42%
    - 全系では100 TFLOPS程度か？
    - HPCG Webサイトにて配布されているHPCG3.1バイナリ(2019.12.05版)で測定、60秒試行
- クラウドシステム
  - 1ノード 1CPU（1MPIプロセス）：16.6079 GFLOPS： 16.6079/1344=1.24%
  - 1ノード 4CPU（4MPIプロセス）：61.7766 GFLOPS： 61.7766/5376=1.15%
    - Intelコンパイラ2019.5.281（のMKL）に付属のベンチマークにて測定、60秒試行

通信性能比較：
OSU Micro-Benchmarks

測定項目

- Type IとType IIの通信性能比較
  - latency
  - allreduce, alltoall
  - 今回はCPU間通信のみ
  - TypeIのノード割り当てポリシー：
    mesh と torus
  - Type IIのMPIの違い：
    Intel MPI と OpenMPI

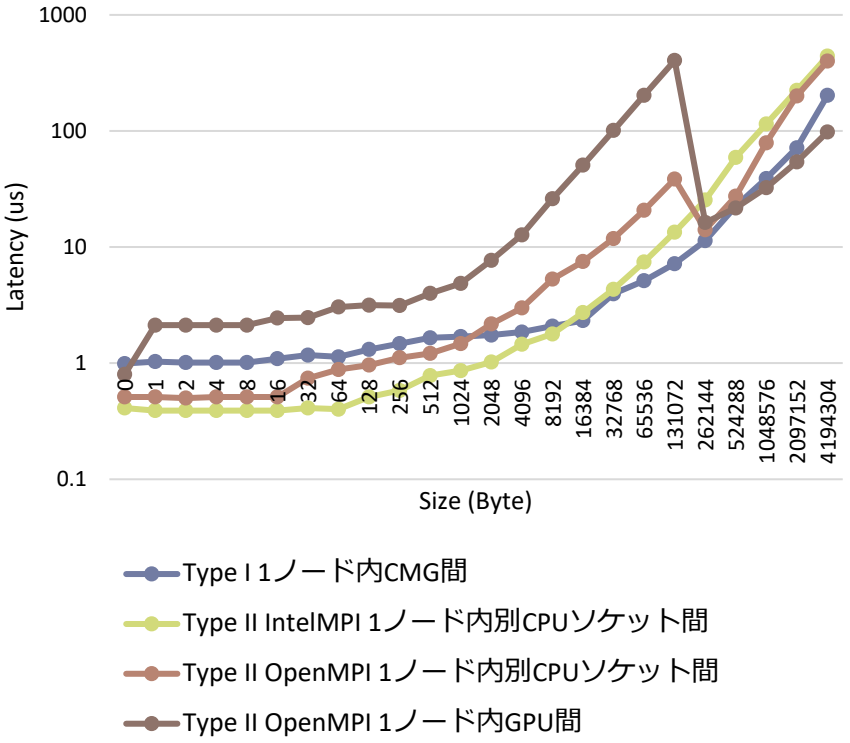
通信性能比較：

大島聡史准教授提供

OSU Micro-Benchmarks：latency

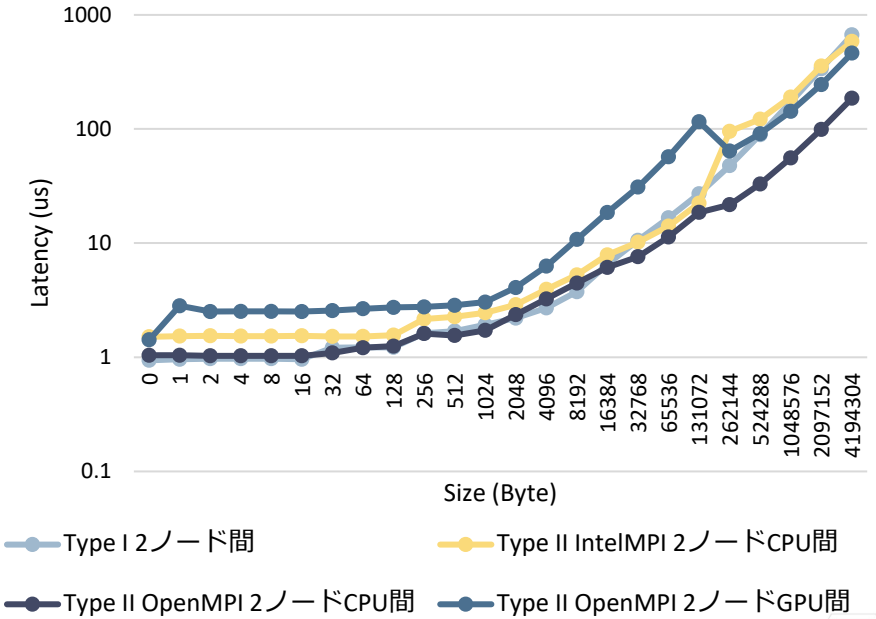
●ノード内

- サイズ小
  - Type IIのCPUソケット間（特にOpenMPI） が速い
- サイズ大
  - Type IのCMG間が速い、TypeIIの GPU間も速い



●ノード間

- Type IIのOpenMPIが速い
- Type Iもサイズ小では速いが、サイズ大はやや遅い
- GPU間は他より遅い



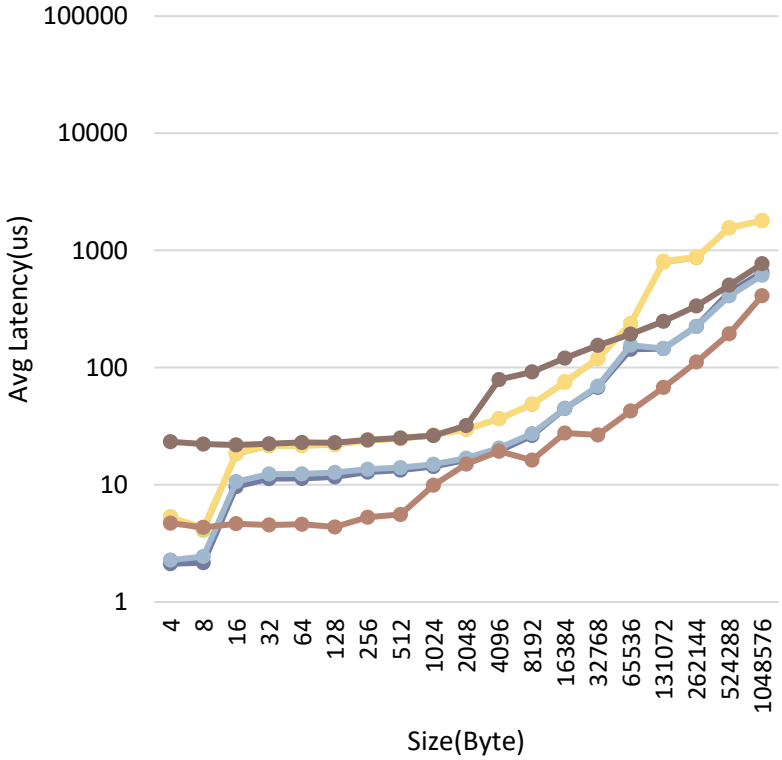
名古屋大学 NAGOYA UNIVERSITY

通信性能比較：

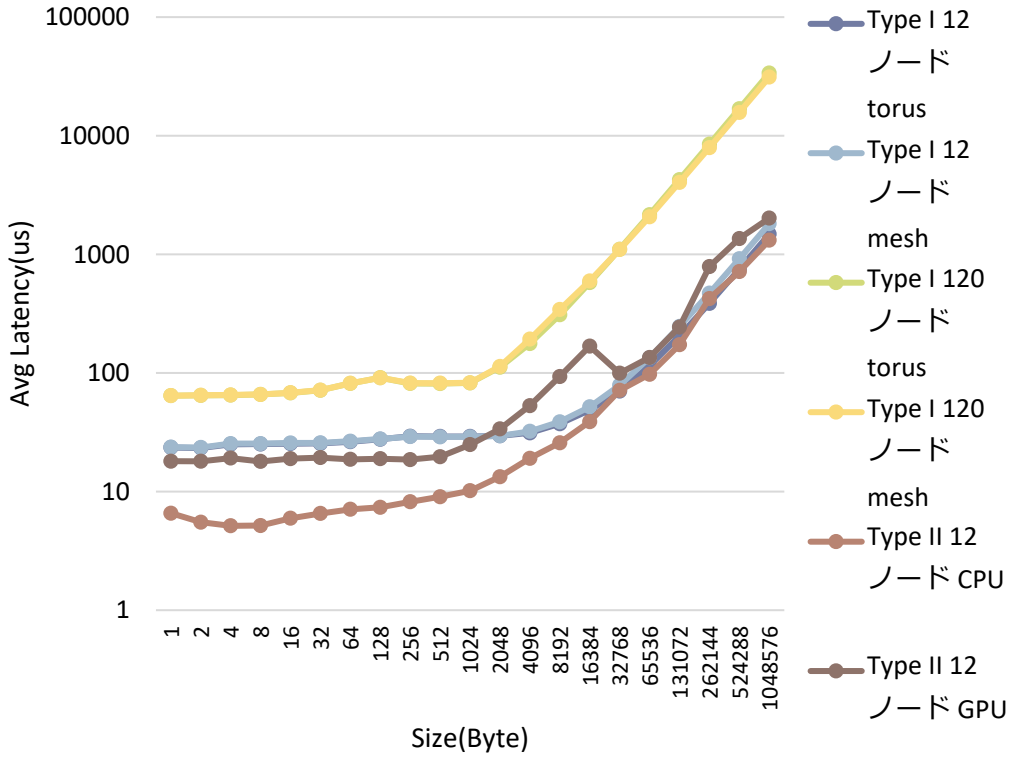
大島聡史准教授提供

OSU Micro-Benchmarks：allreduce, alltoall

● allreduce



● alltoall



- 全体的にはType IIの方が低レイテンシの傾向
- Type Iについてはより多くのノードでも実験するべきであったか

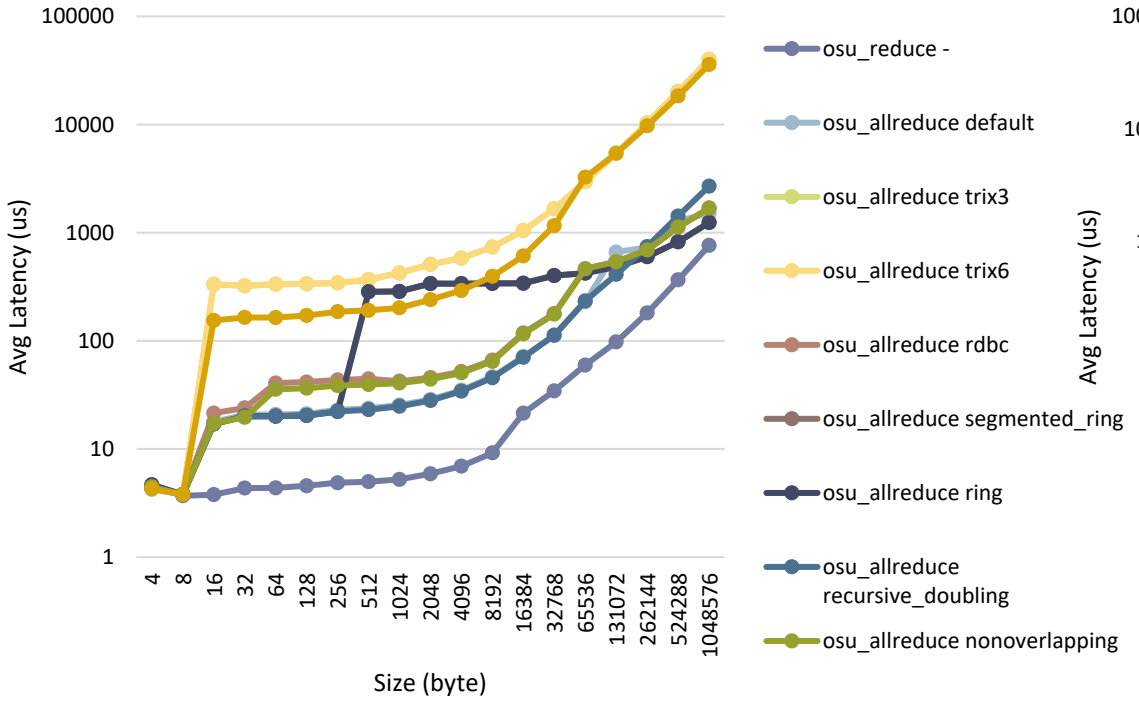
- GPU to GPUの通信性能はパラメタ等精査中……

名古屋大学 NAGOYA UNIVERSITY

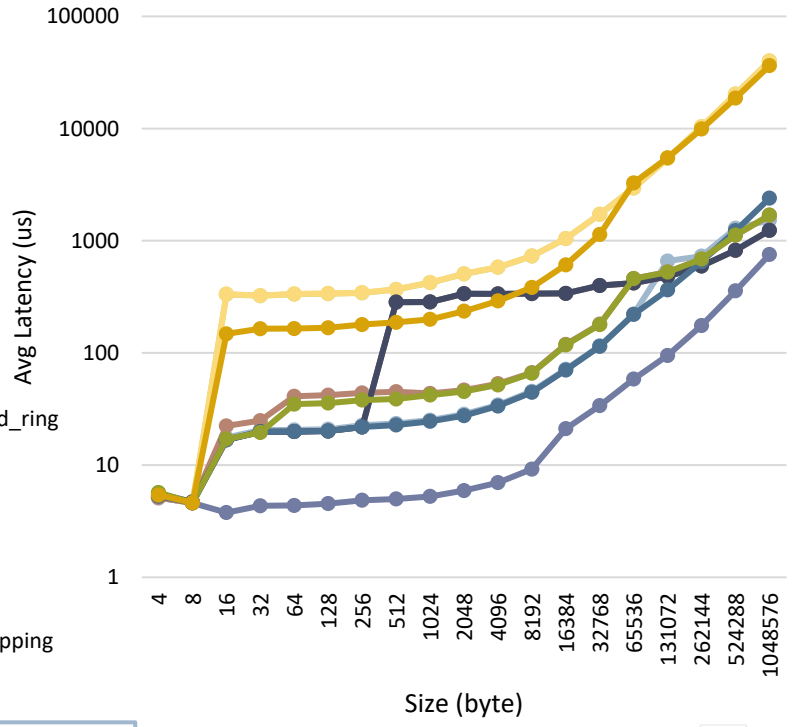
Type IのallreduceアルゴリズムTuning：96ノード（Type Ⅰ）

大島聡史准教授提供

● 96nodes:　torus



● 96nodes:　mesh

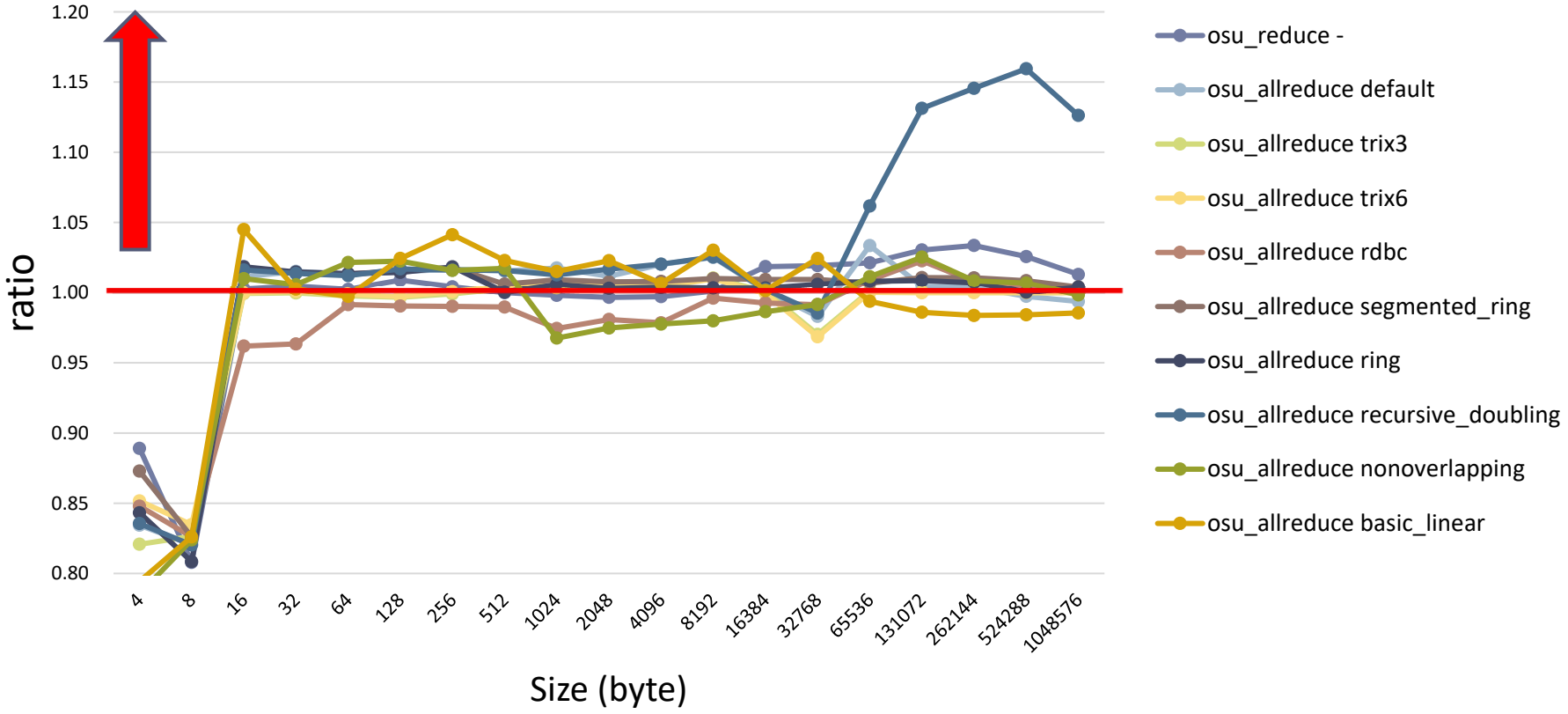


- デフォルト（未指定）でほぼ最速、ただし128KBが若干遅い

Type IのallreduceアルゴリズムTorus 対 mesh（Type Ⅰ）

meshが高速　　96nodes, torus/mesh

大島聡史准教授提供

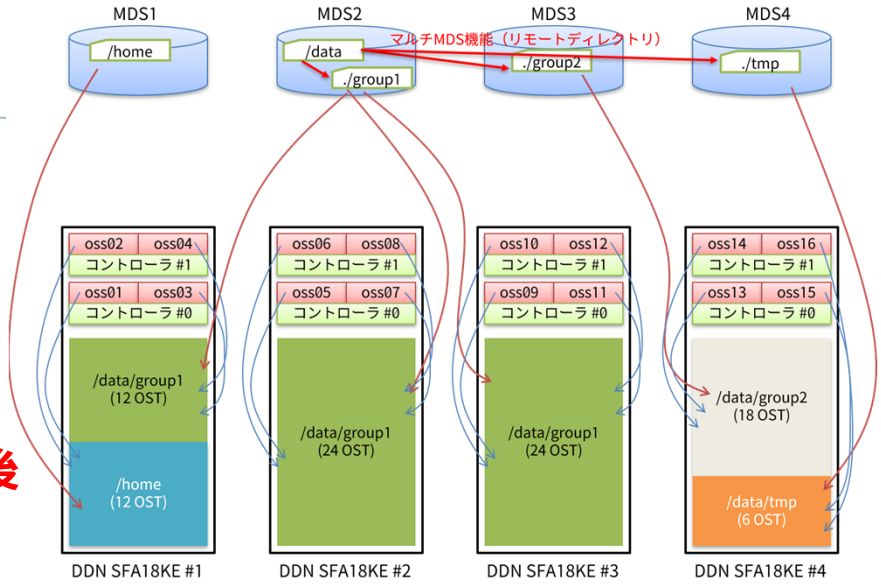


- 4Byte, 8Byteはtorusの方が明らかに高速、16byte以上ではあまり変わらない（±5%）
- recursive_doublingだけはサイズが大きな時にmeshが高速

名古屋大学 NAGOYA UNIVERSITY

ストレージ性能：IOR（運用前性能）

大島聡史准教授提供



- IORベンチマーク性能
  - 一般的なバッチジョブで利用が推奨されている/data/group1に対してType Iサブシステムから読み書き
    - **File Per ProcessでR/Wともに100GB/s前後**
    - **Single Shared FileでR/Wともに80GB/s強**

		File Per Process 性能		Single Shared File 性能	
操作対象	OST	write 平均値	read 平均値	write 平均値	read 平均値
ディレクトリ	クライアント数	[MiB/sec]	[MiB/sec]	[MiB/sec]	[MiB/sec]
/data/group1	576	93130.36	106612.15	84112.94	83286.16
/data/group2	288	33770.22	38685.99	32984.10	36369.85
/home	144	17928.59	22037.02		
/data/tmp	144	12697.28	13242.41		
		write+read 合計値	169052.000	write+read 合計値	118381.525

名古屋大学 NAGOYA UNIVERSITY

ストレージ性能：
自作のテストプログラム

大島聡史准教授提供

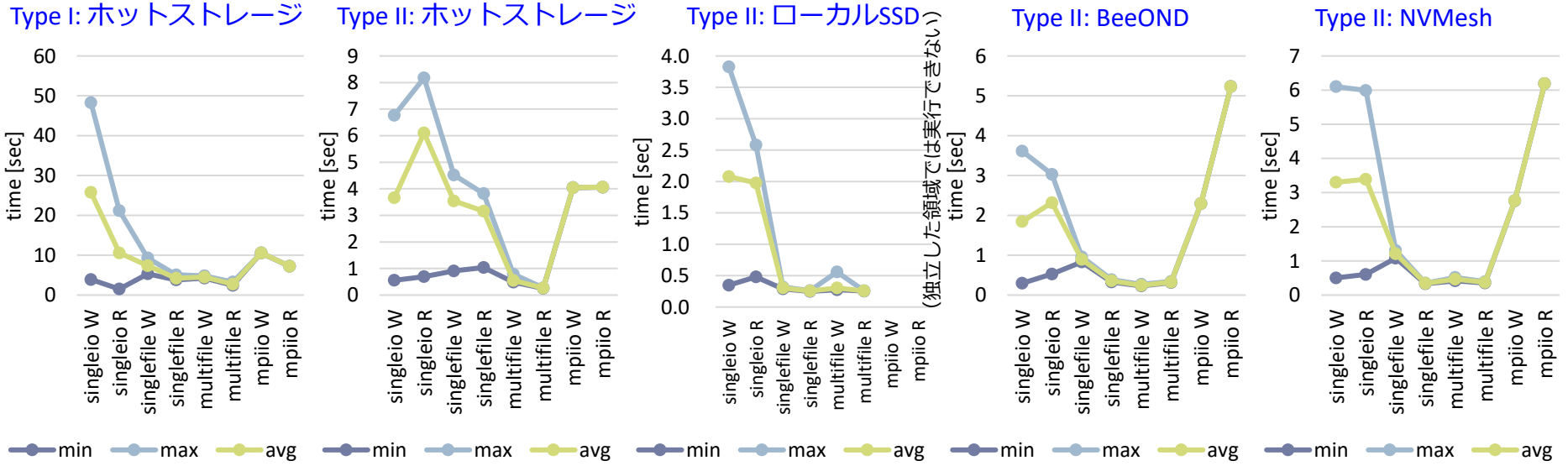
- 実際の使い方を想定したテストプログラムを作成し運用中に測定
- プログラム内容（いずれもファイルのopen/close時間を含む）
  1. **singleio**：マスタープロセスのみがファイル操作、配付と集約はMPI通信
     - ノード数が増えた場合に全プロセス分を持てないことを想定し、逐次的に1対1通信
  2. **singlefile**：全プロセスがfread/fwriteで1ファイルを読み書き
  3. **multifile**：全プロセスが個別の1ファイルをfread/fwrite
  4. **mpiio**：MPI-IOで1ファイルを読み書き（MPI_File_set_view, 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）
  - {1M,10M,100M,1G}elements/processで測定、今回は100M版を提示（10M以上は同様の傾向）
  - データ型は全てdouble型（8byte）
    - 100M elements/processはプロセス当たり800MB、1秒で終了すれば9600MB/sec相当
- **githubに公開してあるため興味がある人はご自由にお試しください**
  - **https://github.com/exthnet/iotest**

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

大島聡史准教授提供

一発測定、 min, max, avgは12ノード中のmin, max, avg

※1秒=9600MB/sec



- multifileが良い性能：max値（R/W）を比較すると、左から順に
  W/R = 4.82/3.27, 0.79/0.28, 0.56/0.26, 0.27/0.34, 0.52/0.39
  - 実はType IがType IIよりかなり遅い、ただしホットストレージは計算ノードとストレージ両方の負荷の影響を受けるため測定タイミングの影響もあるかもしれない？
- mpiioの性能が今ひとつ、特にBeeOND・NVMeshが遅い
- 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。

名古屋大学 NAGOYA UNIVERSITY

スーパーコンピュータ「不老」アプリケーション

名古屋大学
NAGOYA UNIVERSITY

スーパーコンピュータ「不老」で動かすアプリケーション例

■ 大規模数値計算（Type Ⅰサブシステム）

- 名大 坪木和久 教授：スーパー台風解析（雲解像モデル CReSS）
- 名大 渡邉智彦 教授：プラズマシミュレーションGKV

■ AI／GPUコンピューティング（TypeⅡサブシステム）

- 名大 森健策 教授 （GPU大規模AI計算）：AIによる医用画像診断支援技術
- 名工大 本谷秀堅 教授（規模GPU計算）：医用画像処理 LDDMM

■ 大規模数値計算＋AI（Type Ⅰサブシステム＋TypeⅡサブシステム）

- 立教大 望月祐志 教授：COVID-19解析、フラグメント分子軌道計算（ABINIT-MP） 理研R-CCS：新型コロナウイルス対策を目的としたスーパーコンピュータ「富岳」の優先的な試行的利用採択課題
- 山梨大 相馬一義 准教授：機械学習による気象予測

■ 可視化・ストレージ（TypeⅢサブシステム）

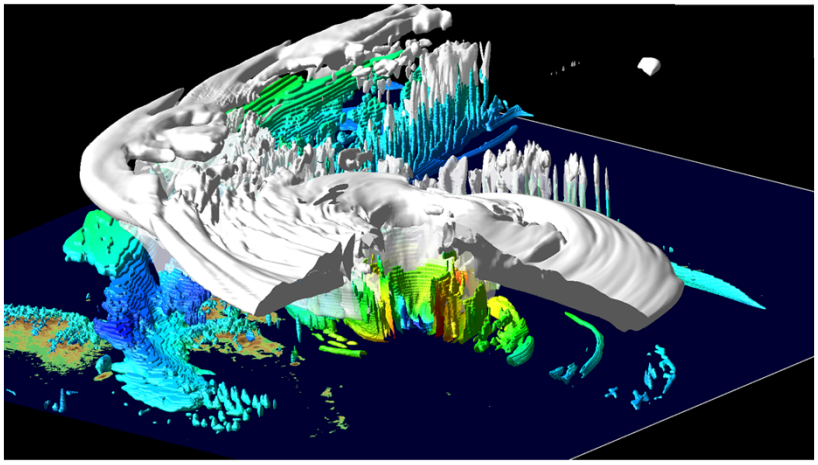
- 名大 高橋一郎 特任主任技師：
  可視化ツール VisPlus、コールドストレージ操作ツールODAPLUS

■ 大規模数値計算（Type I サブシステム）アプリケーション

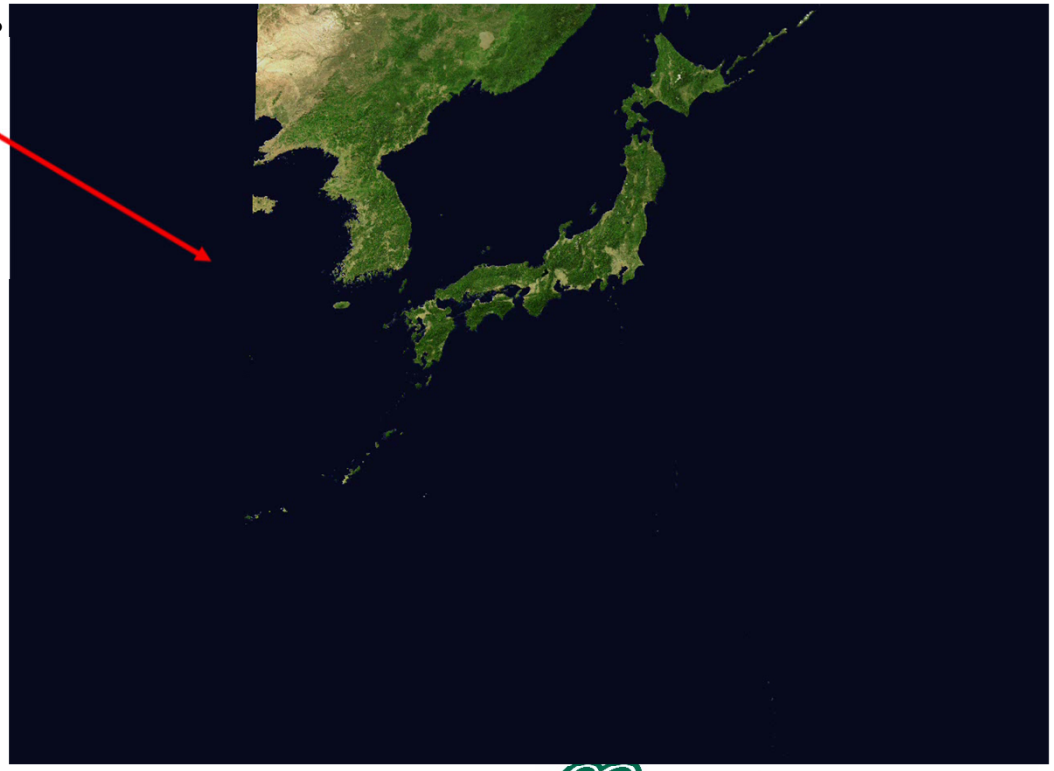
スーパー台風のメカニズム解析

（坪木和久教授提供）

- **名古屋大学 坪木和久教授 開発による 雲解像モデルCReSS**
  (Cloud Resolving Storm Simulator)によるシミュレーション
- **未来の台風（2076年09月に発生）**に伴う雲を立体的表示
  この台風は太平洋上を北上し、日本に上陸する直前でも中心気圧880 hPa以下を維持。
  台風が太平洋上にあるとき、中心気圧870～860 hPa、最大地上風速70～80 m/sを4日間維持し、
  ほぼその強度のまま関東地方に上陸する。



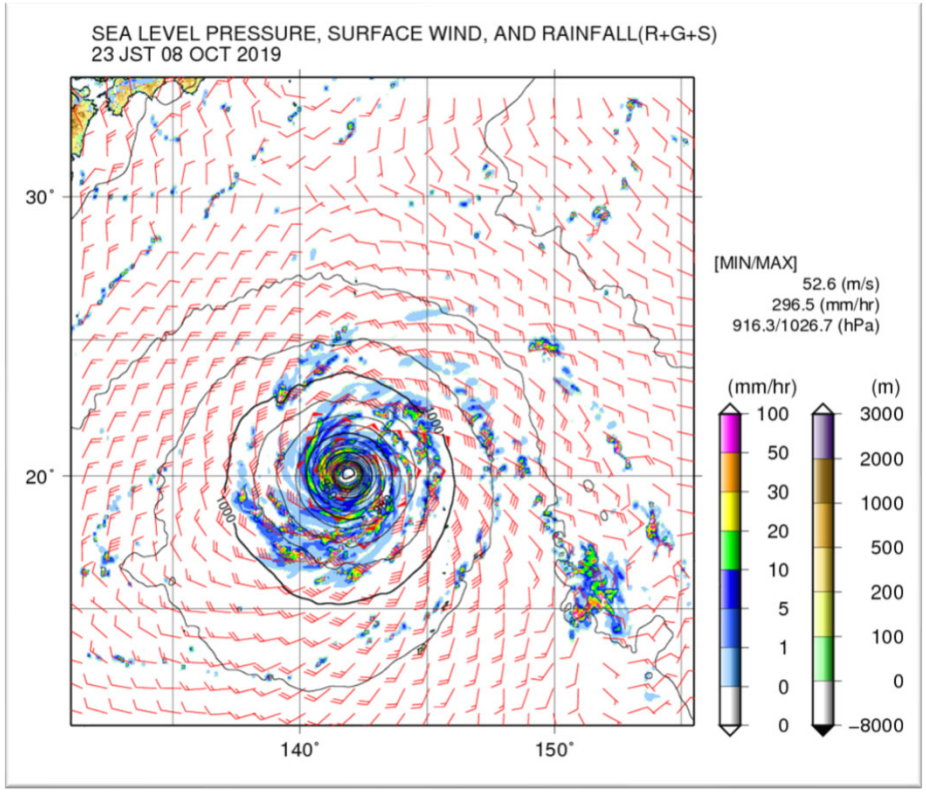
伊勢湾台風のシミュレーション結果

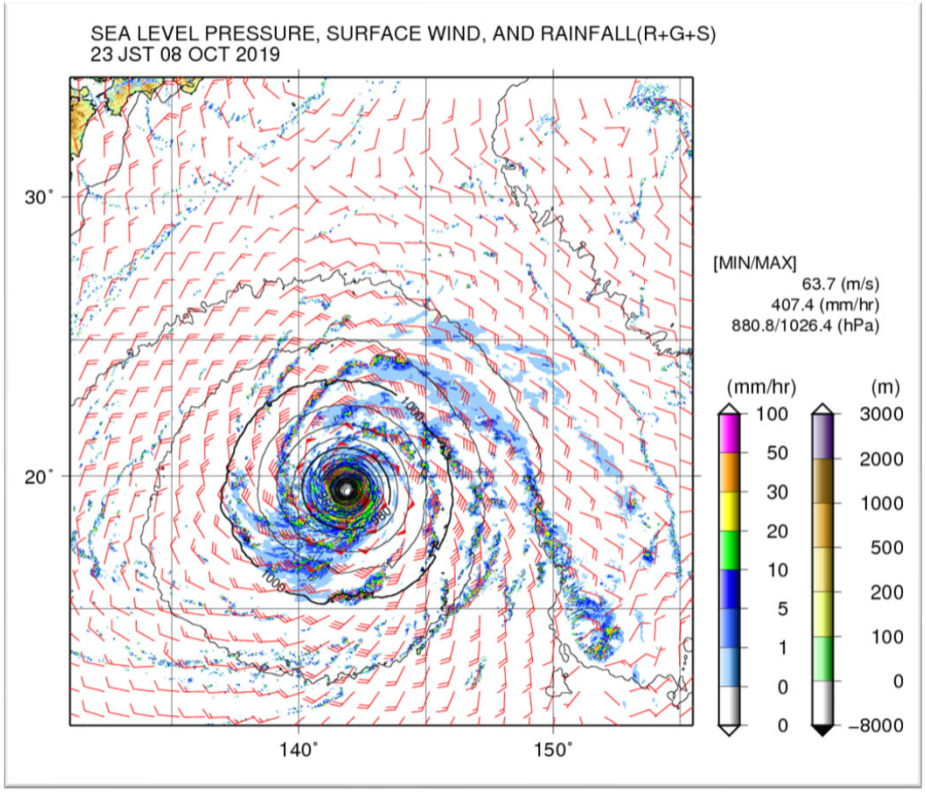


台風のメカニズム解析
：スーパーコンピュータ「不老」でのターゲット

- 旧システム（FX100）
  - 水平間隔 2 km 格子、東西・南北それぞれ1000～2000格子、
    鉛直は地上から上空約20 km まで 100層、
    積分時間数日程度の計算
  - 数日の実行時間を必要
- スーパーコンピュータ「不老」によって
  - より細かく短時間に
  - 水平1 kmあるいは 500 m 、かつより広い領域で、
    台風の急発達のメカニズムや詳細な構造が明らかになる
  - 数100 m 格子でより詳細な地形を考慮した
    シミュレーションにより豪雨の詳細な構造を明らかにし、
    豪雨に伴う被害の詳細な予測が可能に

台風のメカニズム解析
：スーパーコンピュータ「不老」での計算結果





従来：水平格子間隔約2.5km

水平格子間隔約1.0km

スーパーコンピュータ「不老」で計算

1600ノード、24時間程度
現在計算中・結果検証中

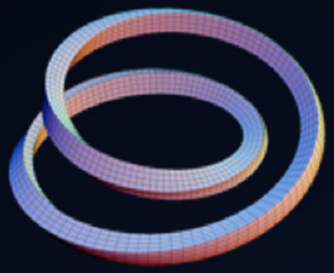
核融合プラズマ乱流コードGKVの概要

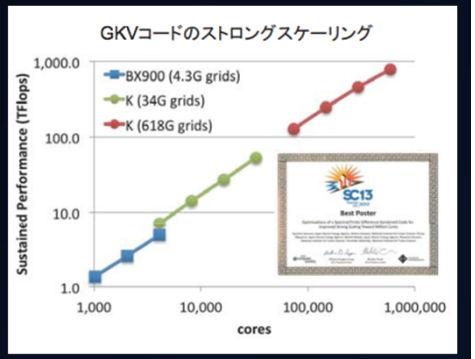
GyroKinetic Vlasov code

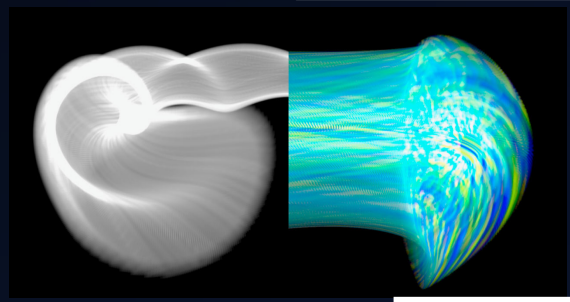
- ジャイロ運動論モデルに基づく核融合炉心プラズマ乱流の第一原理シミュレーション
- 5次元位相空間上の移流・拡散
  - MPI/OpenMPハイブリッド並列
  - FFTスペクトル法(x, y)+差分法$(z, v_{\parallel}, \mu)$
  - 時間積分：陽的ルンゲクッタ法+陰解法衝突項
- フラックスチューブ配位による局所乱流の高精度・高解像度計算
  - 電子・イオン系マルチスケール乱流
  - 多粒子種プラズマ乱流

開発者：
Tomo-Hiko Watanabe (Nagoya Univ.)
Shinya Maeyama (Nagoya Univ.)
Masanori Nunami (NIFS)
Motoki Nakata (NIFS)
Akihiro Ishizawa (Kyoto Univ.)
Yuuichi Asahi (JAEA)

Flux tube







スーパーコンピュータ「不老」Type I サブシステム全系（2304ノード（11万コア））ベンチマーク

スーパーコンピュータ「富岳」想定規模：格子点数：2048*2048*48*96*48*3=1.3×10^12、
MPI並列数：16*4*8*6*3=9216 = (2304ノード×4MPI / ノード)、OpenMP並列数: 12

本研究の一部は，文部科学省「富岳」成果創出加速プログラム「核燃焼プラズマ閉じ込め物理の開拓」の一環として実施したものです。

GKVフルアプリ結果（その１）

- スーパーコンピュータ「不老」Type I サブシステム（富岳型ノード）
  - 288ノードジョブ（13,824コア）
  - 理論性能：973 TFLOPS
- 演算効率：6.78%（ 66.06 TFLOPS ）

- 問題サイズ
  - 512 * 256 * 48 * 96 * 48 * 3 = 8.7×10^{10} 格子点
  - 2 * 4 * 8 * 6 * 3 = 1,152 MPI (= 288node * 4MPI/node)
- ラージページ指定効果
  - export XOS_MMM_L_PAGING_POLICY=demand:demand:demand
  - 適用前：77.66 [秒] → 適用後：75.44 [秒] （2.9%高速化）

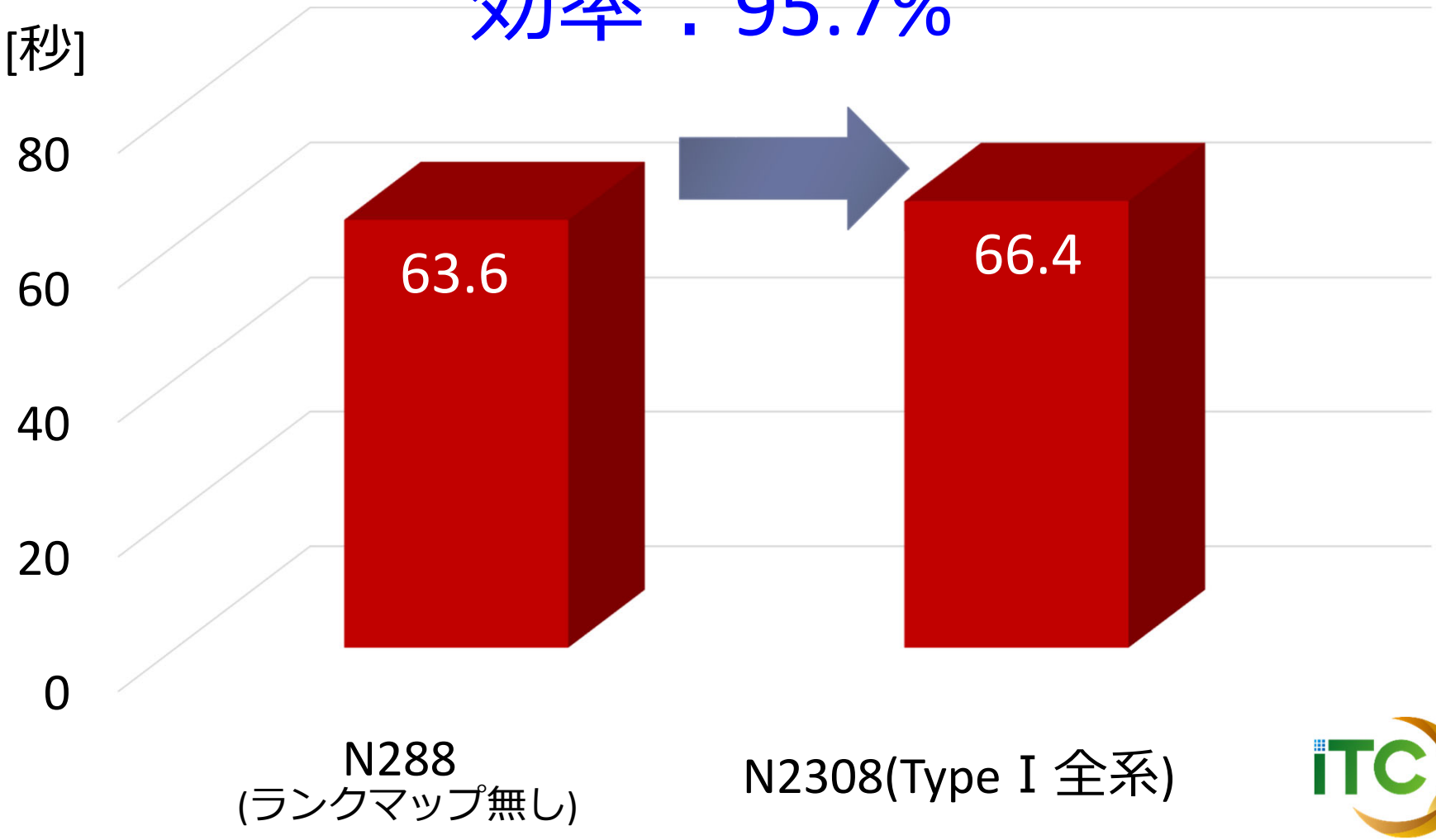
GKVフルアプリ結果（その２）

- スーパーコンピュータ「不老」
  Type I サブシステム（富岳型ノード）
  - 全系ジョブ
  - 2304ノードジョブ（110,592コア）
  - 理論性能：7.782 PFLOPS
- **演算効率：6.67%（519 TFLOPS）**
  - ランクマップなし：76.7[秒] → あり：66.4[秒] (15.5%高速化)

- 問題サイズ
  - 1024 * 1024 * 48 * 96 * 48 * 3 = 7.0×10^{11} 格子点
  - 16 * 4 * 8 * 6 * 3 = 9,216 MPI (= 2,304 node * 4MPI/node)

GKVフルアプリ
弱スケーリング結果

効率：95.7%



GKVフルアプリ
ランクマップの効果（２３０４ノード）

- ランクマップなし

MPI	%	Communication (s)	Start	End		
847695	1.1740	84764.0938	--	--	Application	
847672	1.1739	84761.7969	976	1032	gkv_bndry.bndry_zv_sendrecv_	

- ランクマップあり

MPI	%	Communication (s)	Start	End		
472885	0.7353	47284.8438	--	--	Application	
472847	0.7353	47281.0430	976	1032	gkv_bndry.bndry_zv_sendrecv_	

■ AI／GPUコンピューティング（TypeⅡサブシステム）アプリケーション

医用画像処理　名工大　本谷秀堅　教授

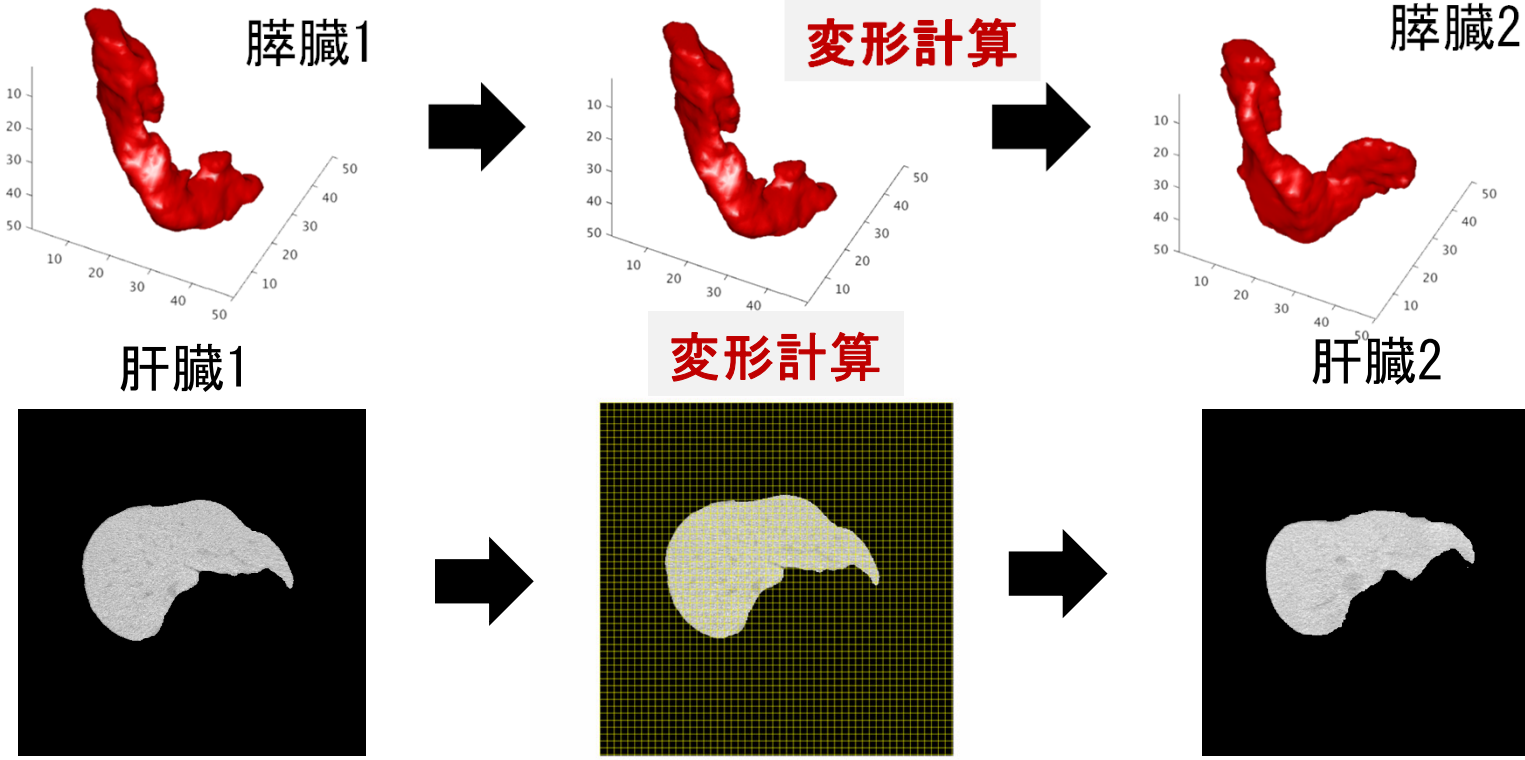
人の臓器形状の滑らかな補間のための高速計算

- 人工知能の学習には多数のデータが必要
- 人の臓器の形のサンプルを多数揃えるのは大変

大量データ/大規模計算
→スパコンで高速化

形の違う臓器二つの間を自然に**高速変形**させることで学習データを多数獲得

→病気の自動診断システム開発へ

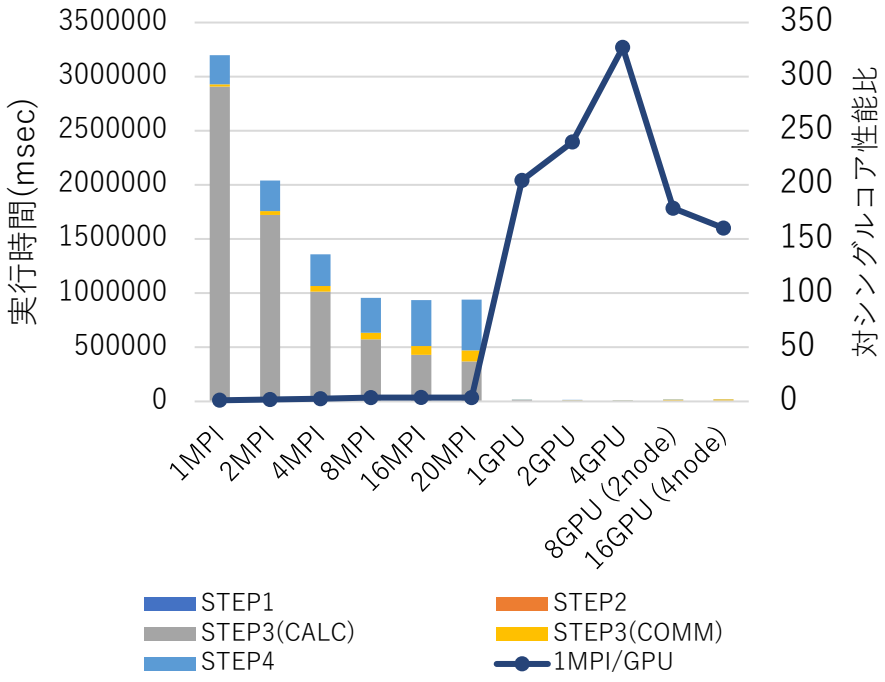


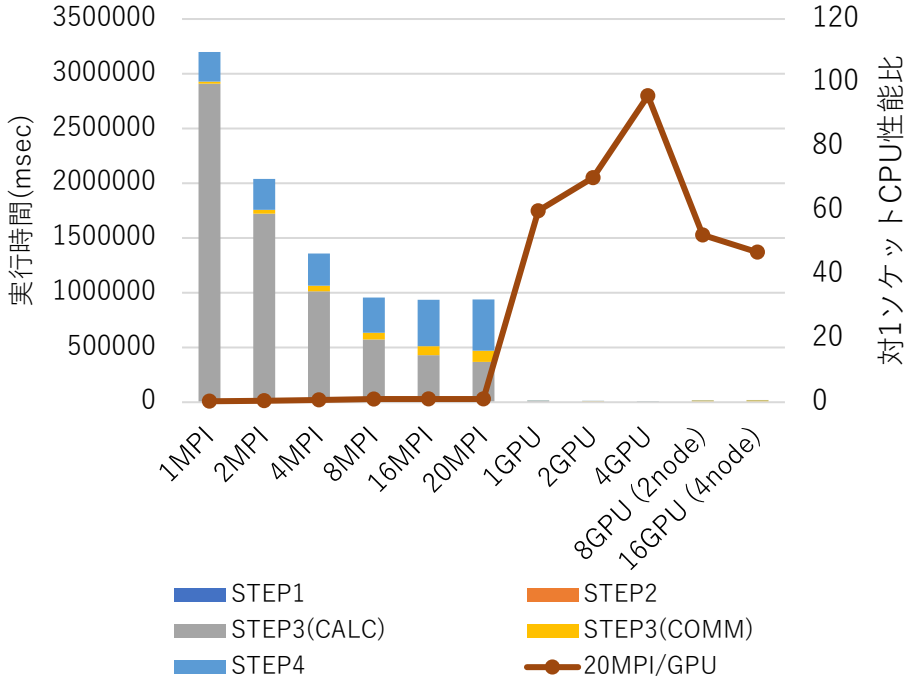
（本谷秀堅教授提供）

医用画像処理　名工大　本谷秀堅　教授

スーパーコンピュータ「不老」TypeⅡサブシステム（GPU）

- 大変形微分同相写像（Large Deformation Diffeomorphic Metric Mapping,LDDMM)法によるプログラム（本谷研究室開発）
- プログラムをGPU化（OpenACC）





- CPUシングルコアに対し：326倍
- CPU（1ソケット）に対し：96倍

■ 大規模数値計算＋AI（Type Ⅰ サブシステム＋Type Ⅱ サブシステム）

ABINIT-MP

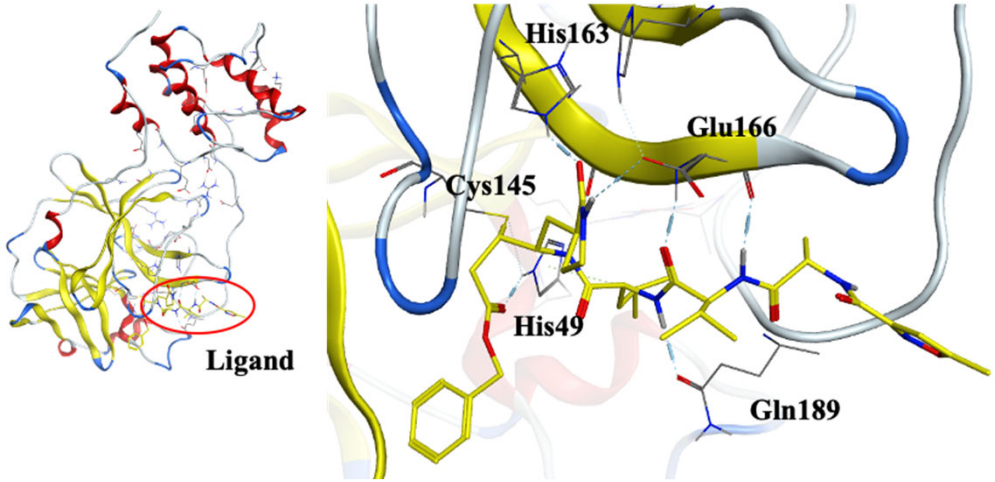
（望月祐志教授提供）

コロナウイルスのメインプロテアーゼとN3阻害剤の複合構造に関するフラグメント分子軌道計算

畑田崚, 奥脇弘次, ○望月祐志（立教大学）, 福澤薫 （星薬科大学）
古明地勇人（産業技術総合研究所）, 沖山佳生（国立医薬品食品衛生研究所）,
田中成典（神戸大学院）

●SARS-CoV-2 メインプロテアーゼとN3阻害剤の結晶構造

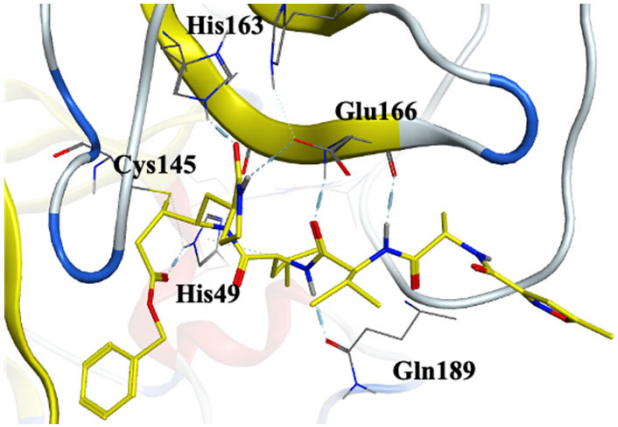
- 2020年2月に、Liuらによって SARS-CoV-2のメインプロテアーゼと、阻害能を有する化合物の複合体の結晶構造が発表された (PDB ID: 6LU7)



本研究では、フラグメント分子軌道法と6LU7結晶構造を用いてメインプロテアーゼ-N3阻害剤間の相互作用解析を行った

●研究内容はChemRxivで2020/3/17に公開 (https://bit.ly/3f67nlc)
●アメリカ化学会の専門誌に採録：Fragment molecular orbital based interaction analyses on COVID-19 main protease - inhibitor N3 complex (PDB ID:6LU7), J. Chem. Inf. Model. 2020, June 15, 2020 (https://doi.org/10.1021/acs.jcim.0c00283)

フラグメント分子軌道法を用いた相互作用解析



- **解析方法**

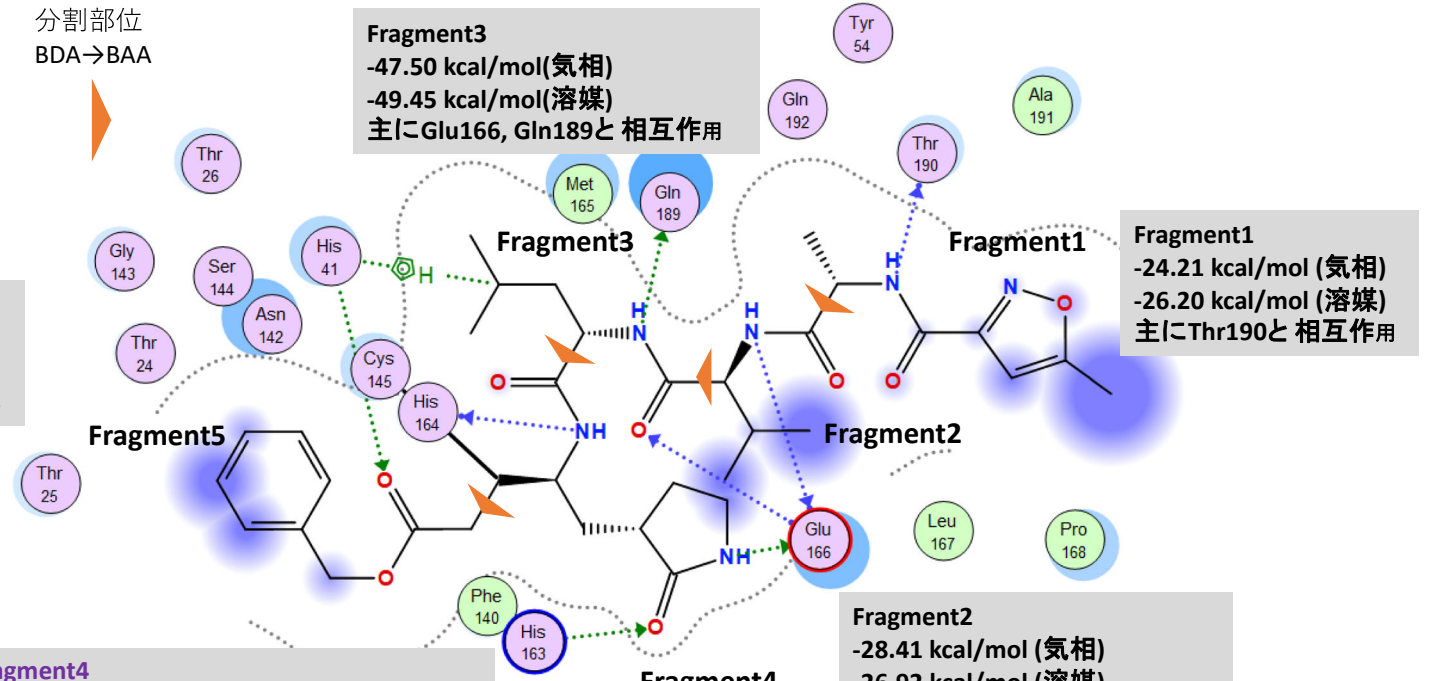
N3阻害剤を5つの部位に、
タンパク質をアミノ酸単位にフラグメント分割し、
フラグメント間の相互作用エネルギー(IFIE)を算出
気相条件・溶媒条件下で各々MP2/6-31G*レベルで計算

- **計算環境： ABINIT-MPプログラム @ 名大FX-100**

気相条件 : 128ノード（16スレッド - 256プロセス） 1.4時間
溶媒条件 : 192ノード（ 16スレッド - 384プロセス） 30.5時間

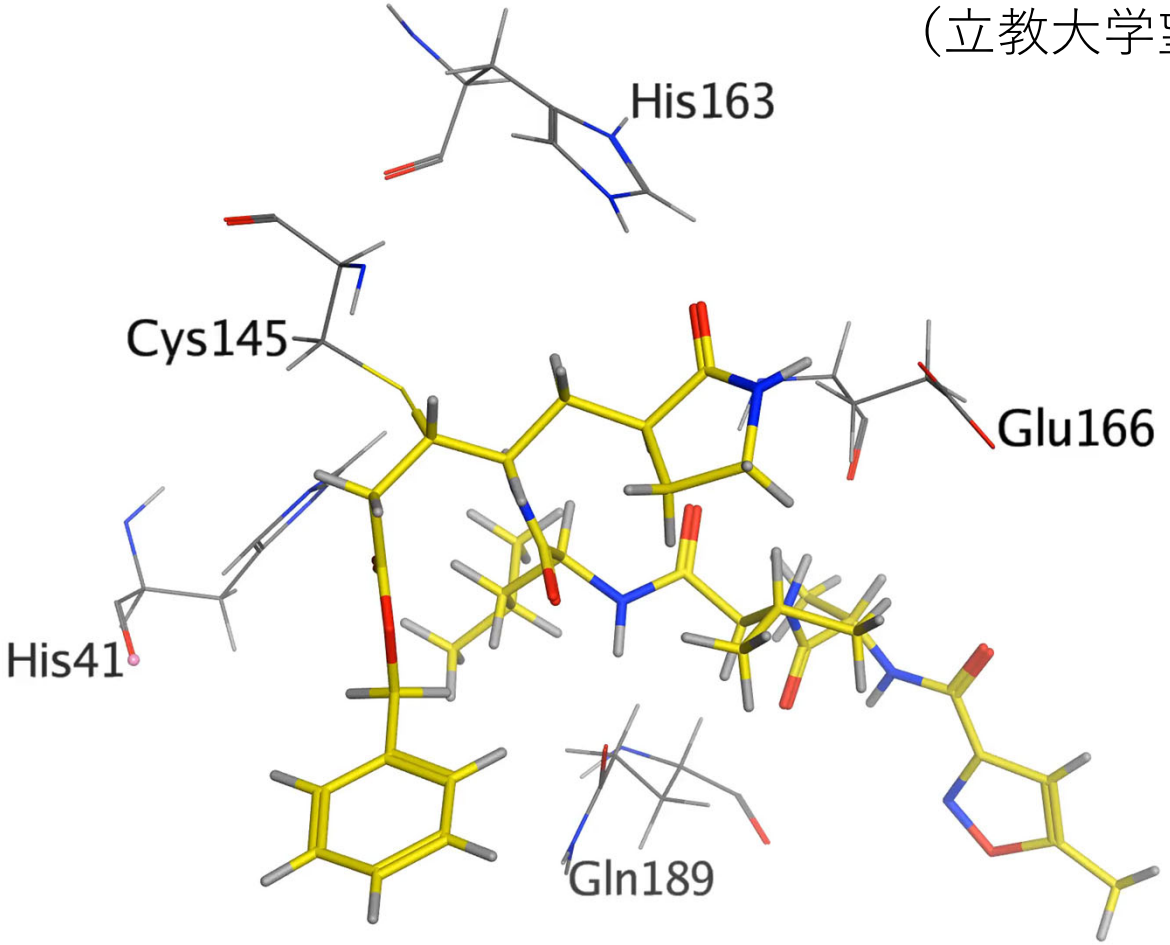
- **解析結果**

分割部位
BDA→BAA



新型コロナウイルスのメインプロテアーゼと結合した阻害剤N3の重要部位の構造ゆらぎ

（立教大学望月研究室提供）



スーパーコンピュータ「不老」で以下の研究開発を予定：

- **Type Ⅰサブシステム**：SIMD化強化等のコードチューニング
- **Type Ⅱサブシステム**：AI処理連携

雲解像モデルを用いたアジアメガシティの都市気象シミュレーション

（相馬一義准教授提供）

相馬一義1，　倉上健 2，宮本崇1，古屋貴彦1，馬籠純1，石平博1，坪木和久3，草野完也3，田中賢治4
1 山梨大学大学院総合研究部，2 日本工営株式会社　中央研究所
3 名古屋大学宇宙地球環境研究所，4 京都大学防災研究所

近年短時間強雨が増加し、被害が多発

短時間強雨をより早い段階で予測することが必要

→数値気象モデルを用いた降水予測が有効な手段

洪水・土砂災害予測等、減災へ活用

→降水予測結果を定量的に用いる必要がある

https://www.zenchin.com/news/27-6.php

https://www.jma.go.jp/jma/kishou/books/hakusho/2017/index4.html

※数値気象モデル　物理式を用いて三次元の風の動きや降水量などを予測するモデル

（相馬一義
准教授
提供）

複数のショートカット接続を含む深層畳み込みニューラルネットワーク

計算設定2

入力データ	**降水強度 +地上鉛直風速**
入力データ数	学習：9222 検証:1000
入力層のノード数	40×40
出力層のノード数	40×40
損失関数	MSE
エポック数	500
使用するネットワーク構造	U-Net

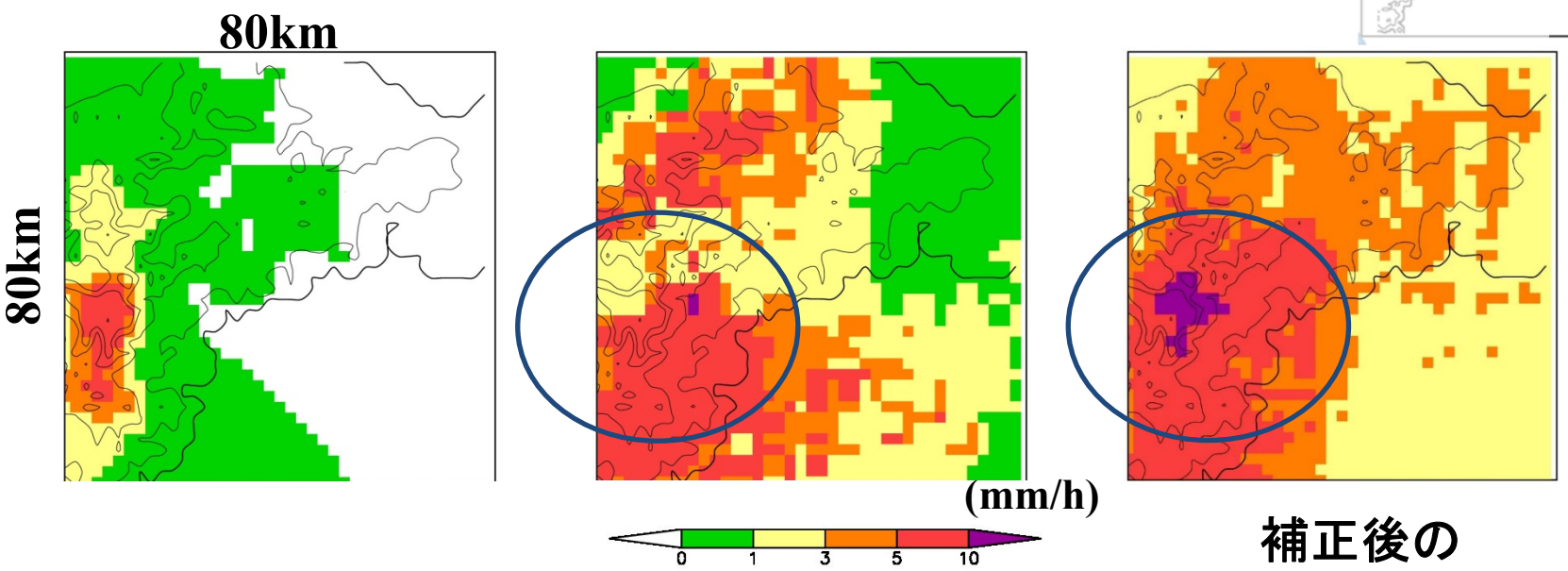
倉上ら，土木学会論文集G，2020

検証過程

定性的評価(計算設定2)

(相馬一義
准教授
提供)

空間スケールの大きい(台風等による)降水
補正によって減災上重要な局所的に降水強度が
大きな領域(10mm/hr)を表現可能になった



数値気象モデルの
降水強度出力

気象庁解析雨量データ

補正後の
数値気象モデルの
降水強度出力

(2001年8月21日1時)のデータ

倉上ら，土木学会論文集G，2020

■ 可視化・ストレージ（TypeⅢサブシステム）センター独自開発ソフトウェア

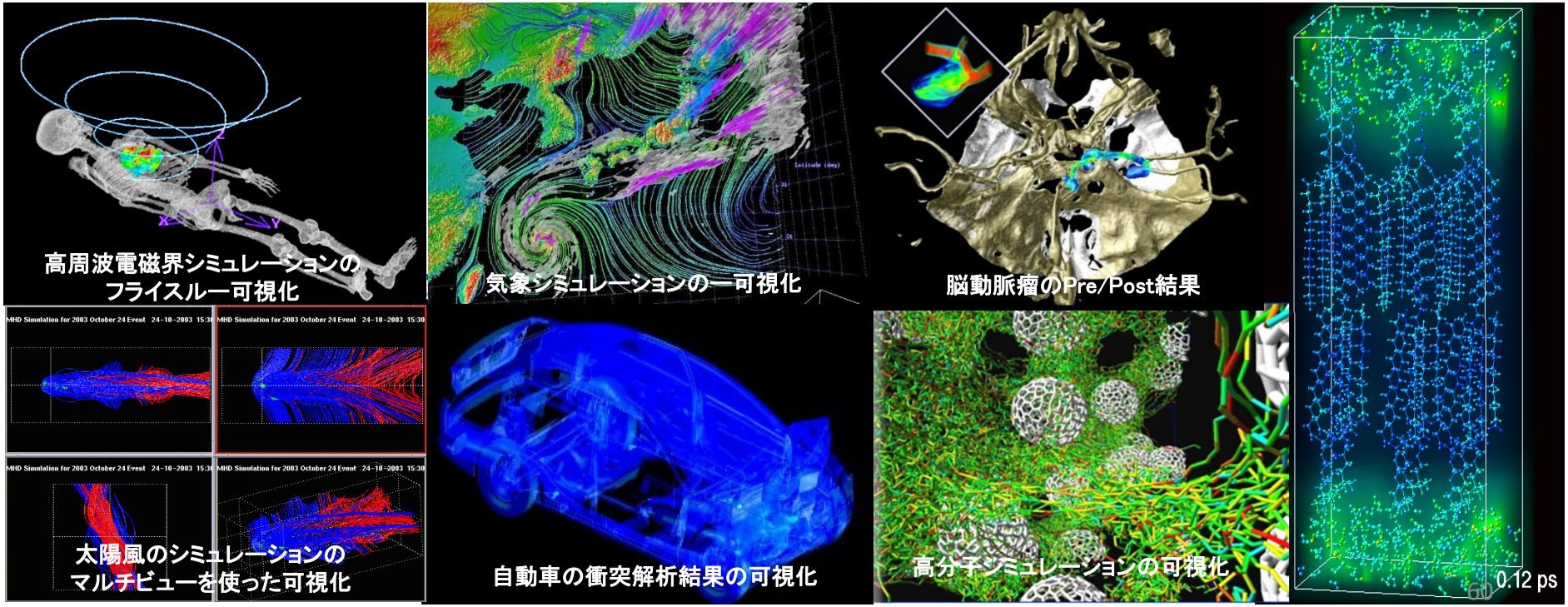
プログラム名： VisPlus

開発者： 名古屋大学　高橋一郎

概要　： 多くの大学や研究機関で利用されている可視化アプリケーション開発ツール”*AVS Express*”を使って開発した可視化アプリケーションプログラムとユーティリティプログラムの集合体である。

　可視化アプリケーションプログラムは、AVSのNetworkエディタを使ってカストマイズしてオープンプラットホームで利用することができる。
また、リモート可視化、ローカル可視化にも対応している。

動作環境： Linux, Windows, Mac

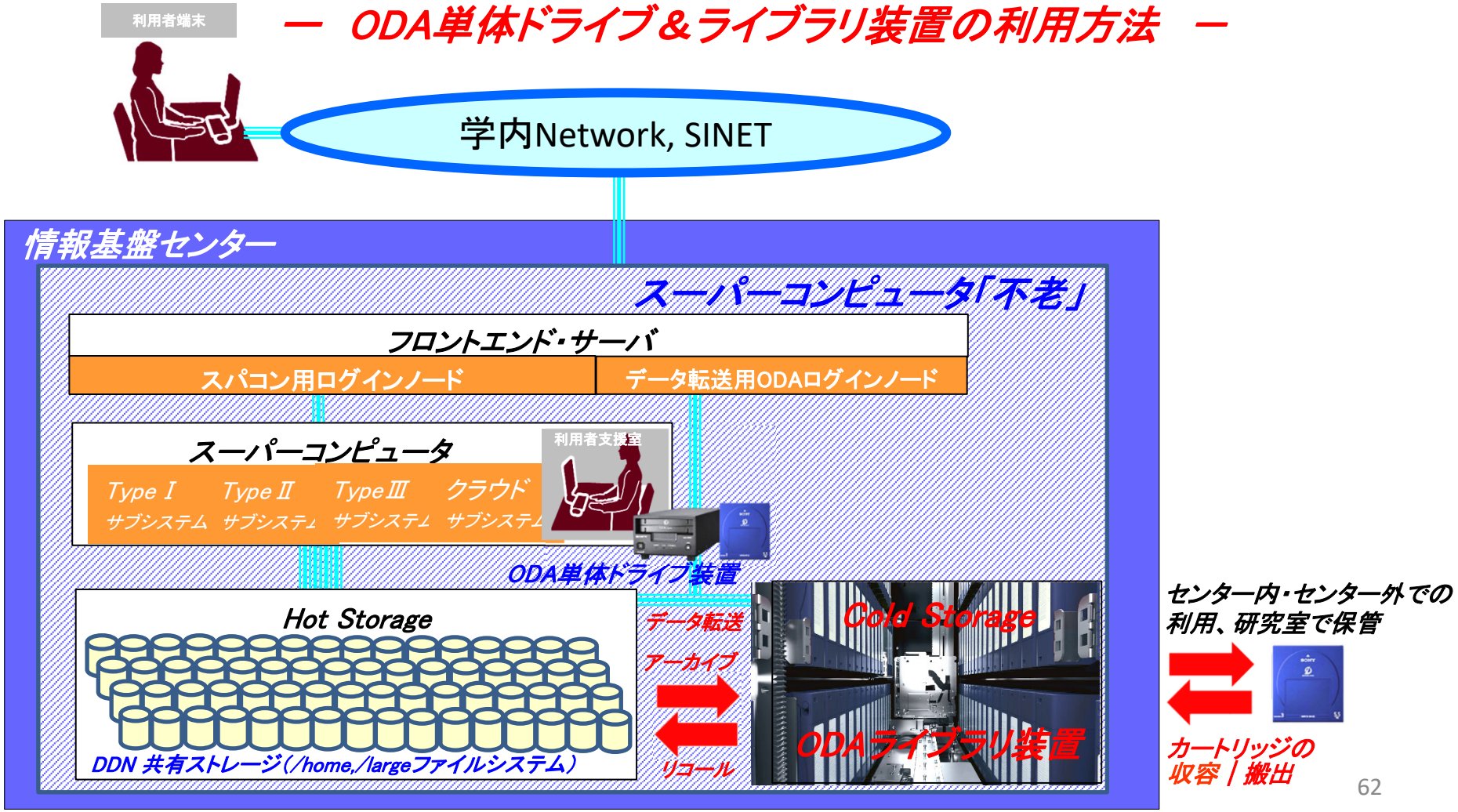


コールドストレージへの展開

プログラム名： ODAPLUS

開発者： Sony, 高橋一郎

概要　： Sony社製コールドストレージ（光Diskライブラリ ODA)システムの管理・運用を行うプログラム(現在開発中）



※以降の内容は開発中のため、サービス体系、機能、コマンド等の変更がありえます。詳細は、名古屋大学情報基盤センターのHPでご確認ください。

高橋一郎　特任主任技師提供

オプティカルディスク・アーカイブとは

オプティカルディスク・アーカイブ(ODAと呼ぶ)は、
デジタルデータの長期保存(アーカイブ)を目的とした、
大容量光ディスクストレージシステムです。

データを格納するカートリッジに複数枚の光ディスクを
格納し、1つのボリュームとして光ディスクを大容量に
利用することができます。

① **Archival Disc(業務用次世代光ディスク)**

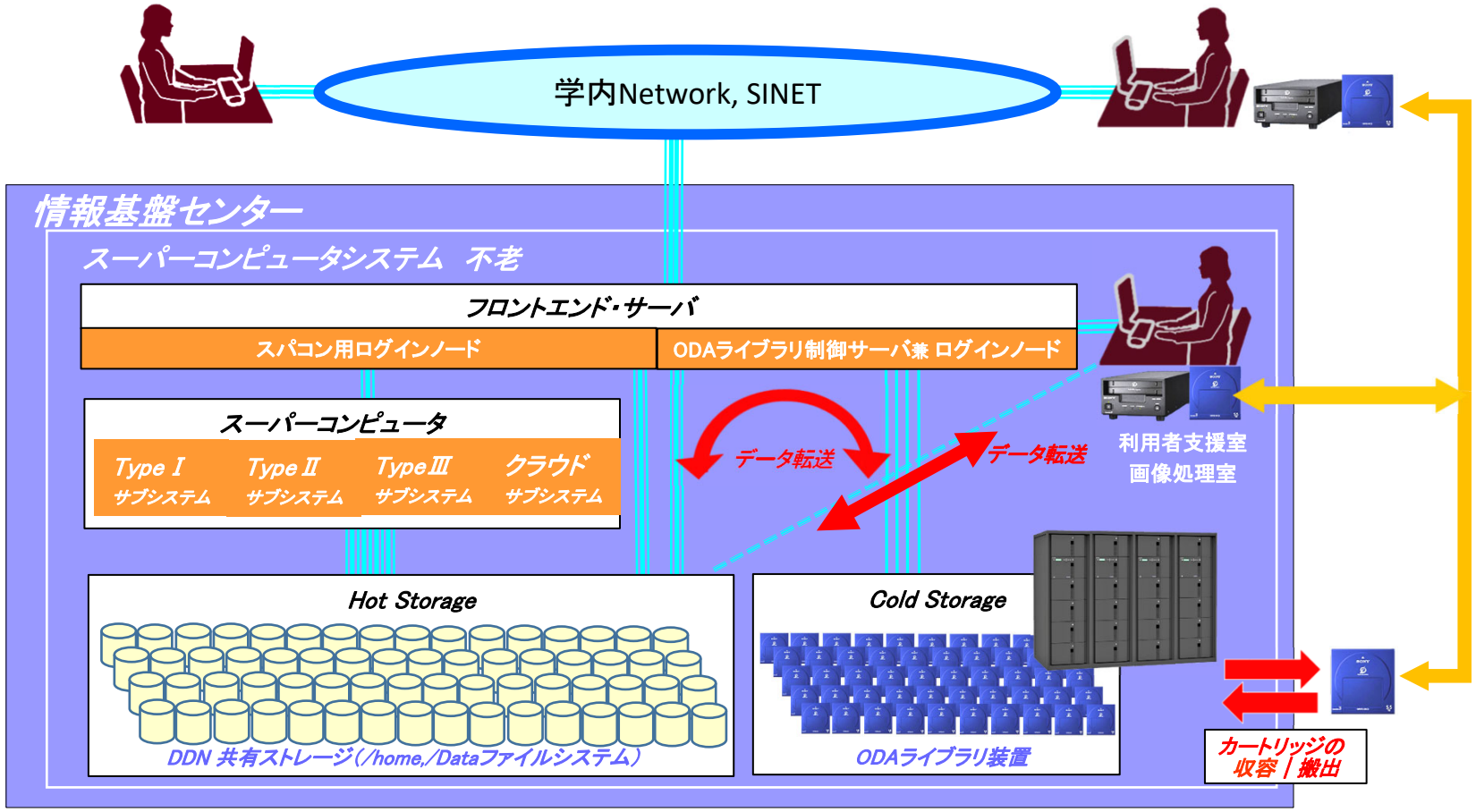
② **ODAドライブユニット**
PCやWSに、USB接続して利用する。

③ **PetaSite拡張型ライブラリー**

資料) オプティカルディスク・アーカイブのご紹介
https://www.youtube.com/watch?v=TAYQ_FiJrAk
Introduction of PetaSite Library Systems
https://www.youtube.com/watch?v=pPbt0sARido
オプティカルディスク・アーカイブ・カートリッジ耐久性実験
https://www.youtube.com/watch?v=UnIrreBvndg
https://www.youtube.com/watch?v=nGzrGgDMupw

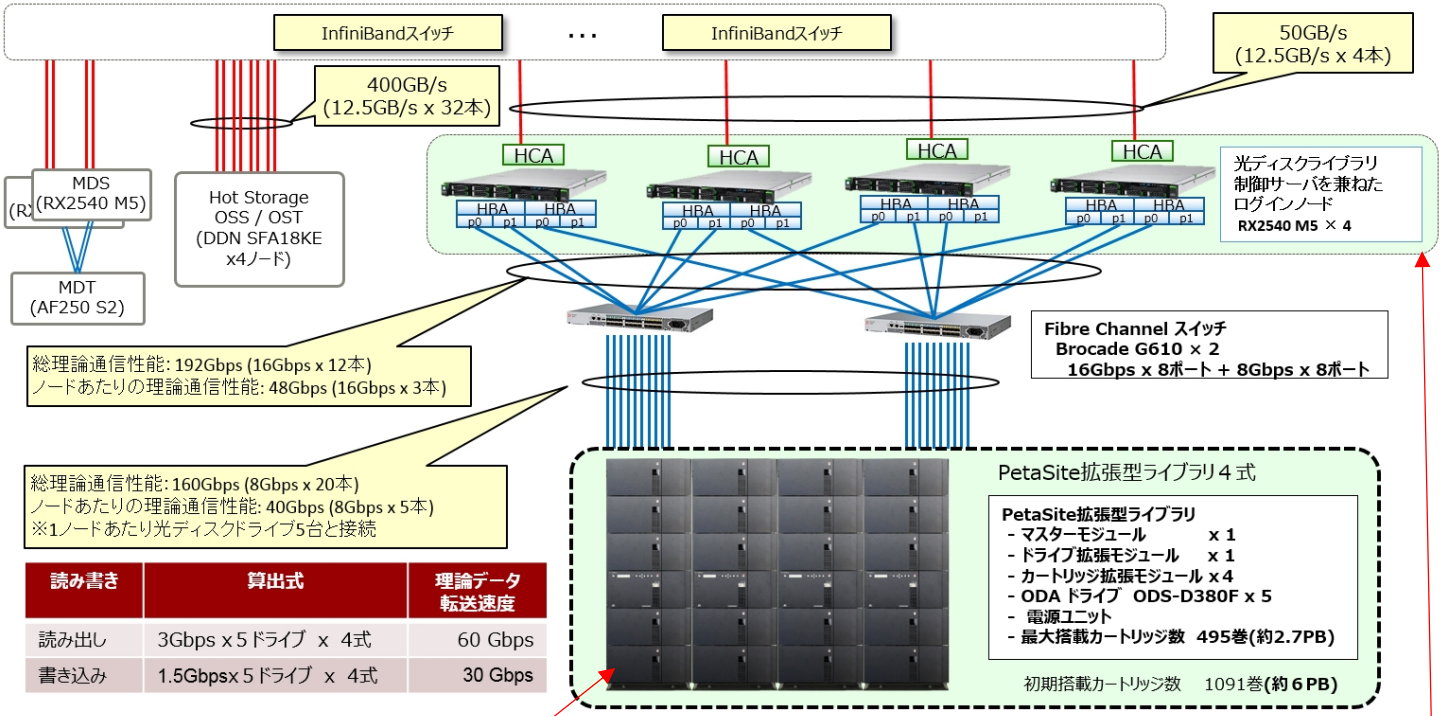
高橋一郎　特任主任技師提供

ODAカートリッジとデータの流れ



高橋一郎　特任主任技師提供

コールドストレージシステム構成（Phase2）



読み書き	算出式	理論データ転送速度
読み出し	3Gbps x 5ドライブ x 4式	60 Gbps
書き込み	1.5Gbpsx 5ドライブ x 4式	30 Gbps

ODAライブラリ装置(PetaSite拡張型ライブラリ)は、ODAカートリッジを格納するキャビネット、ドライブ、キャビネット内のODAカートリッジを識別してドライブにセットするロボット機構、ODAカートリッジをライブラリ内に収容および搬出する機構から成る。
このODAライブラリ装置は、Fibre Channelでフロントエンドサーバを兼ねているODAライブラリ制御サーバに接続されている。

高橋一郎　特任主任技師提供

単体ドライブを利用する方法

情報基盤センター内の利用者支援室と画像処理室には、端末としての
利用やファイル転送を行うためにODA単体ドライブがUSB接続されている。
Linux環境とWindows環境で利用できる。

Windows環境では、ベンダー提供のWindows Explorer相当のGUIを使った
ファイル操が行えるソフトウエアが利用できる。

Linux環境では、ワークステーションにスパコンのHot Storageのファイルシステムが
富士通のFEFSでマウントされているため、Hot Storageとの高速ファイル転送に利用できる。

高橋一郎　特任主任技師提供

ODAライブラリ装置を利用する方法

① ターミナルソフト（PuTTYやTeraTerm等）を使って、公開鍵認証で「指定されたODAのフロントエンドサーバ」にログインします。

② ODAライブラリ内のカートリッジへのアクセスは、センターが提供する次の2種類のバッチジョブ投入コマンドを使ってジョブを投入して、ファイル操作を行います。

- インターラクティブ・ジョブで実行(@isubコマンド)
- バッチジョブで実行(@qsubコマンド)

高橋一郎　特任主任技師提供

1）バッチジョブ

まず、ファイル操作を行うコマンド列を記載したテキストファイル(バッチファイル)を作成します。そして、そのファイルの名前をジョブ投入コマンドのオペランドに指定してバッチジョブを投入してカートリッジへのアクセスを行います。

2）インターラクティブ・バッチジョブ

ジョブ投入後、一定時間内に対話型でコマンドを入力してカートリッジへのアクセス※1を行います。

※1） センター提供のodaallocコマンドを使って、使用するカートリッジをドライブに取付けてマウントします。そして、Hot Storage間のファイルコピーやファイル操作を行います。最後に、odafreeコマンドを使って、**アンマウント後カートリッジをキャビネットに戻します。**

高橋一郎　特任主任技師提供

ODAライブラリ装置およびカートリッジの利用上の留意点

① 記録メディアは、Write Onceです。

② ODAサーバー当りのドライブユニットの数は、5ドライブです。
サーバーのドライブ数が少ないため、カートリッジへのアクセスは、
バッチジョブを使って利用します。

③ 書き出し操作は、シーケンシャル記録のみ。
同一メディアに対しての多重アクセス（メディア内のコピーを含む）は行えない。

④ カートリッジ当りの記憶容量は、次のとおりです。
記録容量　　：　5TB 程度
作成可能ファイル数：　80万ファイル程度　　←　iノード制限値
ディレクトリおよびファイル階層：　最大64 階層まで（ルートディレクトリは、1階層とカウント）

⑤ 使用できるLinuxコマンドに制限があります。

高橋一郎　特任主任技師提供

⑥ ファイル名およびディレクトリ名の命名規約
- Unicode 2.0 で表現可能な1 文字以上、最大127 文字です。
- 同一ディレクトリ内に、大文字小文字でファイル名を重複させることはできません。
- 次の文字は、使用できません。
  “　＊　／　：　＜　＞　？　｜　￥(back slash or Yen)　(DEL)
- .(dot)で始まる文字列は利用できません。

⑦ ODAのRead/Write性能は、ファイルサイズに依存します。
ファイルサイズが大きいほど高速になります。ファイルサイズは、50MB以上を推奨します。

ファイルサイズの小さなデータを大量に取り扱う場合は、iノード不足になったり
カートリッジへのReadおよびWrite処理に時間がかかります。
このような場合は、複数ファイルをtarコマンドを使ってアーカイブファイルに
取りまとめてからカートリッジに記録してください。
※ tarコマンドを使ったアーカイブ＆圧縮｜解凍コマンドやカートリッジ内の
ファイル情報を管理するソフトウエアは、ご提供いたします。

⑧ マルチボリュームは利用できません。事前にLinuxのsplitコマンドでファイルを
分割してから、カートリッジにデータを記録してください。

高橋一郎　特任主任技師提供

コールドストレージシステム利用申請区分と利用負担金

利用申請書 (区分)	ファイル利用負担経費 (年間)	内容	備　　考
新規申請	管理費：1口　1万円 ファイル負担経費(光ディスクの生ディスク代金相当)：1口　19万円	カートリッジはセンター提供 1口：10カートリッジ (50TB相当)	管理費は初年度のみ 利用終了にカートリッジを利用者に返却
継続申請	管理費：1口　1万円	1口：10カートリッジ	次年度以降
変更申請	な　し	利用メンバーの変更 カートリッジの持ち出し利用 カートリッジの持ち込み利用 利用停止	
カートリッジ 持込み申請	管理費：1口　1万円	カートリッジは利用者持込み 1口：10カートリッジ（上限）	

高橋一郎　特任主任技師提供

ODAライブラリ装置の「新規申請」方法

①　「コールドストレージ新規利用申請書」で利用申請を行います。
年度単位の申請で、1口：10ボリューム（50TB）（初年度）年間20万円、（次年度以降）年間1万円です。

→

1口の申請で10ボリュームが作成され、ボリューム通番とログインノードが
利用者に通知されます。
ボリューム通番とは、ライブラリ内でカートリッジを識別して取扱うための名前です。
ボリューム通番の形式は、登録番号に3桁の数字が付きます。
このボリューム通番は、メディアには記録されません。

ボリューム通番の形式：　登録番号 ＋ 3桁の通番
例：　x49999aの場合：　x49999a001, …　x49999a009

※　ライブラリ内のカートリッジの利用は、利用者制限がかかっています。
「新規利用申請書」の利用者記入欄に、利用を許すメンバーを10名まで申請できます。
ODAログインノードの利用者制限にも利用する予定です。

高橋一郎　特任主任技師提供

ODAライブラリ装置の「継続利用」申請方法

継続して利用する場合は、毎年、「継続利用申請」が必要です。
年度単位の申請で、カートリッジの有無に関係なく、
１口：10カートリッジの管理費：年間１万円　が必要です。

ODAライブラリ装置の「持込み利用」申請方法

カートリッジを持込んで利用する場合※1、「持込み利用申請書」を使って利用申請が必要です。
年度単位の申請で、カートリッジの有無に関係なく、
１口：10カートリッジの管理費：年間１万円　が必要です。
収納するカートリッジの本数を記載し、申請書と一緒に持ち込みカートリッジを提出してください。
カートリッジに順序等がある場合は、記入欄に記載してください。
→　ライブラリに収納後、ボリューム通番とログインノードをお知らせします。

※1）　新規・継続・持ち込み申請を除く。

高橋一郎　特任主任技師提供

ODAライブラリ装置の「利用変更」申請方法

コールドストレージサービス「変更申請書」では、次の申請ができます。

① 利用メンバーの変更申請

② 利用停止申請
　カートリッジの必要・不要を申請書に記載してください。
　必要の場合は、カートリッジを搬出してお渡しします。

③ ライブラリ内のカートリッジの搬出申請
　搬出するボリューム通番を記載してください。
　→　カートリッジを搬出してお渡しします。

④ ライブラリ内にカートリッジの収納申請
　収納するカートリッジの本数を記載し、申請書と一緒にカートリッジを提出してください。
　カートリッジに順序等がある場合は、記入欄に記載してください。
　→　ライブラリに収納後、ボリューム通番をお知らせします。

高橋一郎　特任主任技師提供

ODAライブラリ装置とODAライブラリ制御サーバの諸元

	Phase1	Phase2
ODA制御サーバ数	1	4
ODAライブラリ数	1	4
機種名	PetaSite拡張型ライブラリシステム	
総物理容量	484TB	10.89PB　(2.72PB×4ライブラリ)
総スロット数	88巻	1980巻　(495×4ライブラリ)
総ドライブ数	5	20　(5×4ライブラリ)
ロボット機構	1	4　(1×4ライブラリ)

ODAライブラリ装置の諸元

項　目	仕　様
CPU	Intel Xeon Gold6248(2.5GHz/20コア)×2 理論演算性能　3.2TFLOPS
主記憶	384GiB　(32GiB DDR4×12)
内蔵ストレージ	2TB SSD, 19.2 SSD
インターコネクト	InfiniBand EDR×1
SANインターフェース	FC16Gbps×4
ネットワークインターフェース	10GTwinax×2,　1000BASE-T×2

ODAライブラリ制御サーバ兼ODAフロントエンドサーバの諸元

Phase1は1台、Phase2は4台で構成

高橋一郎　特任主任技師提供

光ディスクドライブの仕様

- ODAライブラリ装置の光ディスクドライブユニットとして「ODS-D380F」
- ODA単体ドライブの光ディスクドライブユニットとして「ODS-D380U」

ODS-D380F

ODS-D380U

仕様		ODS-D380F	ODS-D380U
入出力インターフェース		Fibre Channel 8Gbps	SuperSpeed USB 10 Gbps (USB 3.2 / USB3.0互換
タイプ	読み出し再生	3Gbps	
	ベリファイ記録	1.5Gbps	
電源		DC 19.5V	
消費電力		約115W	約105W
動作温度		5℃ to 40℃	
動作湿度		20% to 90%	
質量		約4.9Kg	約4.8kg
外形寸法		146x94.2x401.8mm	146x95.5x414.4mm
その他		各ドライブは5.5TBの光ディスクに対応しており、双方で記録したデータは、双方で読み取ることが可能（完全互換）	

高橋一郎　特任主任技師提供

ODAライブラリ操作コマンドの利用方法

コマンド名	機　　能
odaalloc	ライブラリ内のカートリッジのマウント処理を行う。
odafree	ライブラリ内のカートリッジのアンマウント処理とDBのファイル情報を更新を行う。
@isub	ライブラリ内のカートリッジにアクセスするインターラクティブ・ジョブを投入する。
@qsub	ライブラリ内のカートリッジにアクセスするバッチジョブを投入する。
qstat	ジョブの状態を表示する（ジョブスケジューラ PBSproのコマンド ）。
qdel	ジョブをキャンセルする（ジョブスケジューラ PBSproのコマンド ）。
@vlist	利用可能なライブラリ内のカートリッジの一覧情報を表示する。
@flist	指定したカートリッジのDBに記録しているファイル情報を表示する。　※ライブラリー内のカートリッジに限る。
@find	利用可能なカートリッジに記録しているファイル情報を、DBを参照してキーワードやワイルドカードで検索する。
@update	指定したカートリッジの最新のファイル情報で、DBのファイル情報を更新を行う。
@tree	@lsコマンドで格納したファイル情報をもとに、ディレクトリ構造のツリー表示を行う。
@ls	指定したディレクトリ配下のディレクトリとファイル情報を、利用者が指定するファイルに格納する。 DBとは別に、利用者が独自にファイル情報を管理することができます。　Hotストレージでも利用できます。
grep	Linuxのgrepコマンド。@lsコマンドで格納したファイル情報を、キーワードやワイルドカードを使って検索する。

【例題】
ボリューム通番：x49999a001に、saveというディレクトリを作成して、/home上のdata01ディレクトリ全体をsaveにコピーする手順を以下に示します。

【@qsubコマンドを使ったバッチジョブの利用方法】

```
[oda01~]$ @qsub  ex1.bat
```

@qsubコマンドのオペランドにファイル操作を記載したバッチファイルの名前を指定してジョブを投入します。
※@subコマンドで投入するジョブの最大実行時間は、24Hです。

【@isubコマンドを使ったバッチジョブの利用方法】

```
[oda01~]$ @isub
*****************************************
*  interactive job submiied.            *
*****************************************
qsub: waiting for job 60.oda01 to start
qsub: job 60.oda01 ready
[oda01~]$
[oda01~]$ odaalloc  x49999a001
[oda01~]$ cd  /oda/x49999a/x49999a001
[oda01~]$ mkdir  save
[oda01~]$ cp -ar  /home/x49999a/data01  save/.
[oda01~]$ ls -al  save
[oda01~]$ odafree  x49999a001
[oda01~]$ exit
```

@isubコマンドのジョブが正常に実行されると、
qsub: ..... to start、qsub:..... readyメッセージの後に、
＄プロンプトが表示され、コマンドが入力できる状態になります。
対話形式でコマンドを入力して、ファイル操作を行います。最後に、exitコマンドを入力してジョブを終了します。
※@isubコマンドで投入するジョブの最大実行時間は、60分です。

バッチファイル　　ファイル名：ex1.bat

```
odaalloc  x49999a001                    ← ボリュームのマウント
cd  /oda/x49999a/x49999a001             ← ディレクトリの移動
mkdir  save                             ← ディレクトリ作成
cp -ar  /home/x49999a/data01  save/.    ← ファイルのコピー
ls -al  save                            ← lsコマンド
cd  .                                   ← カレントディレクトリをボリューム外に移動
odafree  x49999a001                     ← ボリュームのアンマウント
```

※1）計算依頼したのジョブの操作コマンド

qstatコマンド：　ジョブの状態を表示する。
qdelコマンド：　指定したジョブIDのジョブを削除する。

高橋一郎　特任主任技師提供

1）odaallocコマンド

コマンドの形式：
odaalloc　ボリューム名

マウントポイントの形式：
/oda/ログインID/ボリューム名

機能：

odaallocコマンドは、マウントポイントを作成して指定したボリュームをマウントします。

次の一連の処理が順に実行され、取付けたドライブ名とマウントポイントが表示されます。

- 指定したカートリッジの所有者の認証
- ドライブの状態確認
- キャビネットからドライブにカートリッジを搬送
- マウントポイントの作成
- カートリッジのマウント処理

【使用例】
[oda01~]$ odaalloc　x49999a001

allocate start -> check -> move -> drive -> mount -> termination
oda　drive: /dev/sdh　… メディアを取付けたドライブ名
mount point: /oda/x49999a/x49999a001 … マウントポイント

【使用例】

ログインIDが「x49999a」、マウントするカートリッジのボリューム名「x49999a009」の場合のマウントポイント。

/oda/x49999a/x49999a009a

高橋一郎　特任主任技師提供

2) odafreeコマンド

コマンドの形式：
odafree　ボリューム通番

機能：
odafreeコマンドは、指定したカートリッジのアンマウントを行います。

次の一連の処理が順に実行されます。
- 指定したメディアの所有者の認証
- ドライブの状態確認
- アンマウント処理（カートリッジの記録情報をDBに反映）
- ドライブの開放
- カートリッジをキャビネットに搬送
- マウントポイントの消去

odafreeコマンドを実行する前に、必ず、マウントポイント配下のファイルは使用していないことを確認してください。
使用しているとBusy状態となり、エラーが発生します。
その場合は、原因を対処した後に、再度，odafreeコマンドを入力してください。

【使用例】
```
[oda01~]$ odafree  x49999a001
free start-> unmount-> eject-> mount point delete->
termination
```

高橋一郎　特任主任技師提供

3) @vlistコマンド

コマンドの形式：
@vlist

機能：
割当てられているライブラリ内のカートリッジの一覧と空きスペース量を表示します。

【使用例】
[oda01~]$ @vlist

Label	Barcode	FreeSize(byte)	iFree
w49000a001		5204767932416	799998
w49001a001		4898986655744	799998
w49001a002		4346016694272	796380
w49001a003		4895680233472	799998

高橋一郎　特任主任技師提供

4) @flistコマンド

コマンドの形式：
@flist　ボリューム名

機能：
指定したカートリッジのファイル情報を、ODAライブラリのDBを参照して表示します。

【使用例】

```
[oda01~]$ @flist  w49001a001
          FilePath
          /odatest/500mb
          /odatest/100mb
          /odatest1/01x10000
          /odatest1/1000x5
          /odatest2/1mb
          /odatest2/10mb
          /odatest3/50mb
```

高橋一郎　特任主任技師提供

5) @driveコマンド

コマンドの形式：
@drive

機能：
ODAライブラリの空きドライブ数と、ドライブの利用状況を表示します。

【使用例】

```
[@oda01 ~]$ @drive
                空きドライブ数                メディアの固体番号          所有者      ボリューム名
Empty Drives Count: 4
 DriveID  Filesystem      Mounted on        MediaSN          Owner        Label
       1  /dev/sdh                       51AA2013WR70416    x49999a     x49999a001
       2  /dev/sdi
       3  /dev/sdj
       4  /dev/sdk
       5  /dev/sdg
```

高橋一郎　特任主任技師提供

6) @findコマンド

コマンドの形式：
@find　キーワード

機能：

利用可能なカートリッジに記録しているファイル情報を、DBを参照して
キーワードやワイルドカードで検索する。

```
[@oda01 ~]$ @find  *xy*

Label           Barcode   FilePath
w40001a001                /test999/Demo/aaa/rxy02.dat
w40001a002                /test999/Demo/bbb/rxy02.001.dat
w40001a003                /test999/Demo/ccc/rxy02.002.data
```

高橋一郎　特任主任技師提供

7) @updateコマンド

コマンドの形式：
@update　ボリューム名

機能：

マウント状態の指定したカートリッジの現在のファイル情報をDBに反映する。

【使用例】
[oda01~]$ @update　w49001a001

高橋一郎　特任主任技師提供

8) @lsコマンド

コマンドの形式：
　@ls　ディレクトリ名　　[出力ファイル名]

機能：
@lsコマンドは、指定したディレクトリ配下のすべての「ディレクトリ情報」と「ファイル情報」を、フルパスの形式で出力ファイルに保存します。
gzip形式で圧縮保存する場合は、出力ファイル名のサフィックスに「.gz」を付けます。
このコマンドは、ODAライブラリシステムのDBとは別に、利用者が独自にファイル情報を管理することができます。 Hotストレージでも利用できます。
次のLinuxコマンドを使って検索および表示を行うことができます。

cat,less,more	zcat,zless,zmore	→	内容を表示
find, grep	zgrep	→	パスを検索
diff, cmp	zdiff,zcmp	→	比較
@tree		→	ディレクトリ構造の表示

【使用例】

```
[@oda01 ~]$ @ls  /oda/x49999a/x49999a009   009.20200901.gz

[@oda01 ~]$ @ls  /oda/x49999a/x49999a009   009.20200901.gz| @tree -d
```

高橋一郎　特任主任技師提供

9) @treeコマンド

コマンドの形式：
@tree　@lsコマンドの出力ファイル名　[-d]

機能：

@treeコマンドは、@lsコマンド又はodafilelistコマンドで作成したファイルを、tree構造で表示します。「-d」オプションを指定すると、ディレクトリ情報のみ表示します。

【使用例】

[oda01~]$ @tree　out.gz

おわりに

おわりに

- 名古屋大学情報基盤センターで7月1日から
  正式サービス開始のスーパーコンピュータ「不老」
  - **数値シミュレーションとデータサイエンス（主に機械学習）を融合可能な設計と運用**
    1．TypeIサブシステム（「富岳」型ノード）⇒超並列処理用、国策スパコン連結
    2．TypeⅡサブシステム　⇒機械学習、高速ローカルディスク
    3．TypeⅢサブシステム　⇒48TBの大規模共有メモリによる
       プレ／ポスト／可視化処理
    4．クラウド　⇒PCクラスタ利用、時刻指定予約
    5．可視化システム　⇒詳細可視化ディスプレイ連結、リモート可視化
    6．大規模共有ホットストレージ
       ⇒１～５に連結、シームレスなデータ移動
    7．コールドストレージ　⇒データを100年保存可能、簡便な利用コマンドをセンターで開発、光ディスク持ち込みも可能、（業界初）サービス終了時に光ディスクを持ち帰り可能

謝辞

- ベンチマークの取得・資料提供に関して
  以下の皆様のご協力ありがとうございました。

  - 名古屋大学情報基盤センター　大島聡史　准教授
  - 名古屋大学情報基盤センター　永井亨　助教
  - 富士通社　SE各位

参加募集

- イベント開催予定：詳しくはWebで

  http://www.icts.nagoya-u.ac.jp/ja/sc/

  - 「富岳」ノード利用：第4回 スーパーコンピュータ「不老」利用型講習会 ライブラリ利用講習会（初級）
    2020年10月19日
  - GPUノード利用：スーパーコンピュータ「不老」
    Type IIサブシステム利用 Optuna利用講習会
    2020年10月23日
- その他、TypeⅡのGPU (V100)を活用した講習会を多数実施予定です。