

6月23日(木)-24日(金)

PCクラスタワークショップ in 神戸2022 「クラウドとHPC」

@理化学研究所 計算科学研究センター 1階セミナー室 (神戸市)

6月24日(金) 10:40-11:00

大島聡史

(名古屋大学 情報基盤センター 大規模計算支援環境研究部門 准教授)

スーパーコンピュータ「不老」の “クラウド的な”利用の状況について



自己紹介

- 栃木県塩谷郡（那須と日光の間ぐらい） 出身
- 40歳（文科省基準ではもう若手ではないそうです……）
- 経歴
 - 2000年4月-2009年3月 電気通信大学 電気通信学部 → 大学院情報システム学研究科
 - 2009年4月-2009年9月 電気通信大学 大学院情報システム学研究科 産学官連携研究員
 - 2009年10月-2017年3月 東京大学 情報基盤センター 助教
 - 2014年4月-2016年3月 兼業 文部科学省 研究振興局 参事官（情報担当）付
計算科学技術推進室 技術参与
 - 2017年4月-2019年6月 九州大学 情報基盤研究開発センター 助教
 - 2019年7月～（現職） 名古屋大学 情報基盤センター 准教授
- スパコンの調達・運用に関わりながら最新の計算機環境の活用について研究
 - GPUコンピューティング・アプリケーションのGPU化
 - 高性能な並列数値計算の実装法：行列計算、疎行列ソルバー、ライブラリ
 - プログラミング環境：言語やライブラリ

今日お話しすること

• プログラム掲載情報

10:40~11:00 「スーパーコンピュータ「不老」の"クラウド的な"利用の状況について」

大島聡史（名古屋大学）

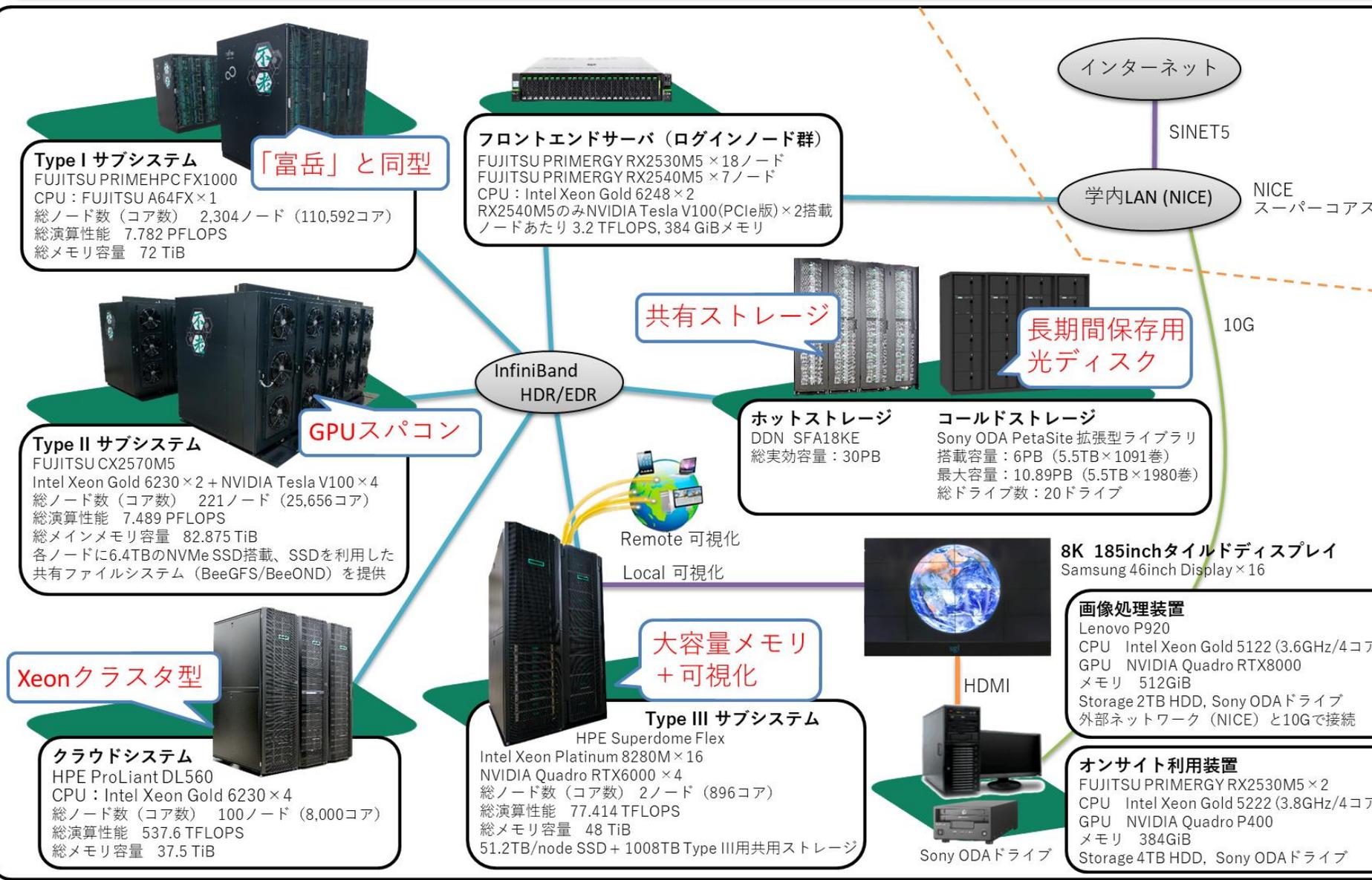
▼ 講演概要を閉じる

スーパーコンピュータ「不老」の運用開始から丸2年が経過しようとしています。「不老」は従来のシステムと比べて多様なユーザに利用してもらうべく、多様なサブシステムによる構成、GPUの本格的な導入、機械学習ユーザ向けの情報提供、時刻指定ジョブの導入、コールドストレージの採用など、仕様や運用に様々な工夫を採り入れてきました。本発表ではその狙いと現状について紹介します。

- この2年、どのようなことを考えて制度設計や運用をしてきたか
- 実際の利用状況はどうか
- （なるべく）「クラウド」に関連付けてお話しします

特徴の異なる複数の計算サブシステム + 複数種類のストレージ

「不老」の概要：全体構成図



各計算サブシステムの主な特徴

Type I サブシステム

- 「富岳」同型 (FX1000、先にサービス運用を開始)
- 最大のノード数とFLOPS

Type II サブシステム

- 各ノードにV100×4搭載 (名大初のGPUスパコン)

Type III サブシステム

- 大容量メモリ
- 可視化システムと連携

クラウドシステム

- Xeon クラスタ型
- 時刻指定実行も可能

Type I サブシステム
FUJITSU PRIMEHPC FX1000
CPU : FUJITSU A64FX×1
総ノード数 (コア数) 2,304ノード (110,592コア)
総演算性能 7.782 PFLOPS
総メモリ容量 72 TiB

「富岳」と同型

フロントエンドサーバ (ログインノード群)
FUJITSU PRIMERGY RX2530M5 × 18ノード
FUJITSU PRIMERGY RX2540M5 × 7ノード
CPU : Intel Xeon Gold 6248 × 2
RX2540M5のみNVIDIA Tesla V100 (PCIe版) × 2搭載
ノードあたり 3.2 TFLOPS, 384 GiBメモリ

Type II サブシステム
FUJITSU CX2570M5
Intel Xeon Gold 6230 × 2 + NVIDIA Tesla V100 × 4
総ノード数 (コア数) 221ノード (25,656コア)
総演算性能 7.489 PFLOPS
総メインメモリ容量 82.875 TiB
各ノードに6.4TBのNVMe SSD搭載、SSDを利用した共有ファイルシステム (BeeGFS/BeeOND) を提供

GPUスパコン

共有ストレージ

ホットストレージ
DDN SFA18KE
総実効容量 : 30PB

コールドストレージ
Sony ODA PetaSite 拡張型ライブラリ
搭載容量 : 6PB (5.5TB×1091巻)
最大容量 : 10.89PB (5.5TB×1980巻)
総ドライブ数 : 20ドライブ

長期間保存用光ディスク

Xeon クラスタ型

クラウドシステム
HPE ProLiant DL560
CPU : Intel Xeon Gold 6230 × 4
総ノード数 (コア数) 100ノード (8,000コア)
総演算性能 537.6 TFLOPS
総メモリ容量 37.5 TiB

Type III サブシステム
HPE Superdome Flex
Intel Xeon Platinum 8280M × 16
NVIDIA Quadro RTX6000 × 4
総ノード数 (コア数) 2ノード (896コア)
総演算性能 77.414 TFLOPS
総メモリ容量 48 TiB
51.2TB/node SSD + 1008TB Type III用共用ストレージ

大容量メモリ + 可視化

8K 185inchタイルドディスプレイ
Samsung 46inch Display × 16

画像処理装置
Lenovo P920
CPU Intel Xeon Gold 5122 (3.6GHz/4コア)
GPU NVIDIA Quadro RTX8000
メモリ 512GiB
Storage 2TB HDD, Sony ODAドライブ
外部ネットワーク (NICE) と10Gで接続

オンサイト利用装置
FUJITSU PRIMERGY RX2530M5 × 2
CPU Intel Xeon Gold 5222 (3.8GHz/4コア)
GPU NVIDIA Quadro P400
メモリ 384GiB
Storage 4TB HDD, Sony ODAドライブ

Sony ODAドライブ

「不老」導入にあたって特に考えたこと

- 維持したこと
 - 国産フラッグシップ互換機※
+ x86系 + 大容量メモリ・可視化
を軸としたマルチサブシステム構成
 - サブシステム共通のストレージ・共通のSSH
公開鍵認証
 - サブシステム共通の利用ポイント制度
- 新しく考えたこと
 - GPU (Type II)
 - SSD (Type II)
 - 時刻指定実行 (クラウド)
 - ノード準占有 (Type II、クラウド)
 - コールドストレージ
 - 縮退運転しやすい体制
 - 湧水の活用

※調達の様式的にはフラッグシップ互換機を必須としたり優遇したりしたわけではなく、候補としてありえるようにした

各計算サブシステムの利用状況

- (時期によって変動は大きいが……)
 - Type Iサブシステム
 - ノード稼働率は半分程度
 - 多ノードのジョブが流れやすい点は便利
 - Type IIサブシステム
 - 統計としてのノード稼働率は半分程度だが、利用状況の波が大きい
 - 計算ノードがほとんど埋まっているようなこともしばしば
 - Type IIIサブシステム
 - 特定の利用者に継続して多く利用されている
 - クラウドシステム
 - ノード稼働率が高く混んでいることが多い (通年で見ても7割を超える)
 - 多くの一般ユーザにとってはXeon CPUが使いやすいうえに少量使う分には安価なためか
- ※ 残念ながらシステム全体でのアプリ単位の統計やユーザの分野別内訳などの情報はないため、どのようなユーザに使われているかなどの詳しい情報はありません

「クラウドシステム」？

- 「不老」のクラウドシステムは
パブリッククラウドなどにつながったシステムではない
- 特徴
 - Xeon Goldを4ソケット搭載
 - Type IIと同じCPU（ただしType IIは水冷、クラウドは空冷）
 - ノードあたりメモリ量もType II同様（ソケット単位では半量、ソケットあたりバンド幅は同じ）
 - 時刻指定実行が可能
 - UNCAIによるWebシステムからノードの予約ができる（クラウド的な利用もできる）
 - 利用ポイントはバッチ実行と同様
 - 優遇もデメリットもないが、半ソケットからの予約実行が可能（小規模利用にも適する）

Xeon クラスタ型



クラウドシステム

HPE ProLiant DL560

CPU : Intel Xeon Gold 6230 × 4

総ノード数 (コア数) 100ノード (8,000コア)

総演算性能 537.6 TFLOPS

総メモリ容量 37.5 TiB

UNCAI

富士通製のWeb予約システム

- 指定した時刻になると仮想マシンが起動する
- 起動後はログインノード経由でssh接続してインタラクティブ利用
- 起動時に自動で実行されるスクリプトを追加指定可能（「時刻指定バッチスクリプト」）
- 使った分（仮想マシンが動いた時間）だけ利用ポイントが減算される
- 0.5ソケット分から利用が可能
 - バッチジョブ実行より小さい単位で使える
- 複数台の仮想マシンの予約やMPI実行も可能
- センタースパコンでUNCAIを導入しているのは名大と九大のみ？

予約表

日表示 ○ リソース検索 ○ 予約一覧

現在日時を表示 2022年6月10日(金)

仮想	08	09	10	11	12	13	14	15	16	17
VS	40	40	40	40	40	40	40	40	40	40
VM	20	20								
VL	10	10								
VX	5	5								

凡例:
 上段 空き (予約作成可能) 予約済み (予約不可)
 下段 操作ユーザの予約 環境作成

予約表について

- ・予約表中の数字は、予約可能数を表しています。
- ・各リソースの仕様は、informationアイコンにマウ

予約作成

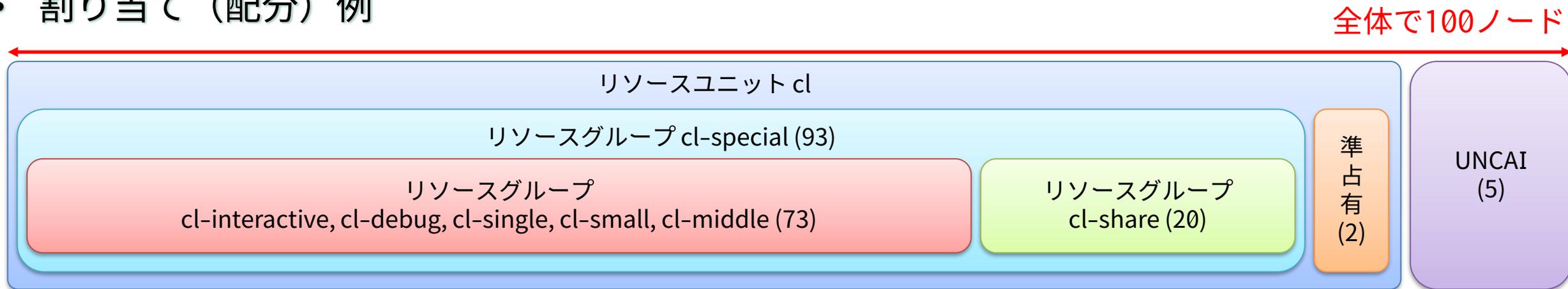
予約者: []
 所属グループ: []
 メールアドレス: []
 テンプレート: VL(Virtual,40cores,180GB Mem) CentOS7.7 台
 時刻指定バッチスクリプト: ~/.batch/指定なし(バッチ実行しない) MPI実行

利用期間
 開始: 2022年6月10日(金) 13:30
 終了: 2022年6月10日(金) 17:30

閉じる 予約作成 予約変更 予約削除

バッチ実行と時刻指定実行（UNCAI）の共存（ノードの分配）

- クラウドシステムは合計100ノード
- 100ノードをバッチ実行用と時刻指定実行用に明確に分けて運用
- 課題：需要にあった分け方をしないとノード利用率の低下や待ち時間の増加を招く
 - バッチジョブを実行しすぎるとUNCAIで予約できるノードがなくなってしまう
 - UNCAIに確保しすぎると遊休ノードの増加やバッチ実行の待ちが増えてしまう
- 対応：利用状況を見て割合を調整することにした
 - 利用実績のあるバッチをメインにし、UNCAIのノード数は様子を見ながら調整
- 割り当て（配分）例



UNCAIの利用状況

- 全期間を通してほとんど使われていない
 - ほとんどのノードをバッチジョブ実行用にしており、バッチジョブのノード占有率は高い
- 何故？
 - 時刻指定実行の存在が伝わっていない？ (UNCAIがあまり使われていない明確な理由がわかっているわけではない)
 - 利用手引書や講習会などで伝えてはいるのだが
 - そもそもバッチ実行しか求められていない？
 - 時刻を指定して実行できることにメリットを感じる人が少ない？
 - UNCAI用に数ノード確保してあるため、バッチ利用がいつぱいでも実行できるのだが……
 - 現状あまり使われていないため確保数は最低限
 - UNCAIを使うことに抵抗がある？
 - めんどくさい？難しい？
 - 特定アプリをもっと簡単に使えるようにしたら需要がある？
 - Jupyter Notebookの起動と接続がもっと何も考えずに行える、とか？
 - 現時点でも簡単に起動できるが、SSHポート転送などのちょっとした作業は必要
 - ユーザの需要とマッチしていない？

パブリッククラウドとは使い勝手が違うのが難点？

- (大手のパブリッククラウドをほんの少ししか使ったことがない私が言うのもどうかと思いますが) パブリッククラウドのWebインターフェイスも複雑で使いやすいとは言い難いような……
- 時刻指定実行 (UNCAI) にはもっと需要がありそうな気がするのだが……何が足りない？
 - 紹介動画？ (具体的に手順などを説明した動画)
 - (パブリッククラウドには必ずある気がする。不要なレベルで簡単なハズなのだが。)
 - 外部ツールとの連携？
 - (Webインターフェイスではなく3rd partyのツールやWeb APIから使う需要の方が大きい？)
 - 動かすアプリ？
 - (時刻指定して動かしたいものがない？外部のシステム等ともっと連携しやすくなれば使う？)
 - 外部のパブリッククラウドなどとの連携機能？
 - (パブリッククラウドとつながっていることが重要？そのあたりはこのあと伊達先生と南里先生が話してくれる？)
 - 実は需要があまりない？
 - もっと「あの環境と同じように使える感」が必要？

「不老」クラウドシステムの現状：まとめ

- 「不老」のクラウドシステムは“クラウド的な利用”にも対応した計算サブシステム
- 現状では“クラウド的な利用”であるUNCAIの利用はあまり行われておらず、従来型のバッチジョブ実行で混雑している
 - スパコンとして悪いことではないが、新しい使い方が有効に使われていない点は少し残念
- 今後“クラウド的な利用”が進むとしたら、特に適した用途（アプリ）が出てきた場合だろうか？
- （ポスト「不老」ではどうしよう……？ ※まだ具体的には考えていません）

Type IIサブシステムの利用状況

GPUスパコン



Type II サブシステム

FUJITSU CX2570M5

Intel Xeon Gold 6230×2 + NVIDIA Tesla V100×4

総ノード数 (コア数) 221ノード (25,656コア)

総演算性能 7.489 PFLOPS

総メインメモリ容量 82.875 TiB

各ノードに6.4TBのNVMe SSD搭載、SSDを利用した共有ファイルシステム (BeeGFS/BeeOND) を提供

- Type IIサブシステムは x86 (Xeon) + GPU (V100) + SSD を搭載した高性能サブシステム
- 他のサブシステムに比べてより先進的な利用・多様な利用を意識
 - 古典的なスパコン利用はType Iサブシステムとクラウドシステムが引き受ける
 - Type IIサブシステムはより現代的な計算資源利用を担当 というイメージ
- **Type IIサブシステムは名大センター初のGPUスパコン**
 - GPUクラウド (パブリッククラウド) と直接「戦う」ことになるシステムとも言える
 - 正直言って、十分使ってもらえるかはやや不安だった
 - 統計としてみると利用率が低い月もあった
 - 自分が利用に困るほど混んでいることもたびたびあり、ある程度しっかり使えてもらっている感触
 - 若干ヘビーユーザがいる模様 (HPCI関係?)
 - 機械学習ユーザが増えてくれた?
 - AlphaFoldユーザの影響?

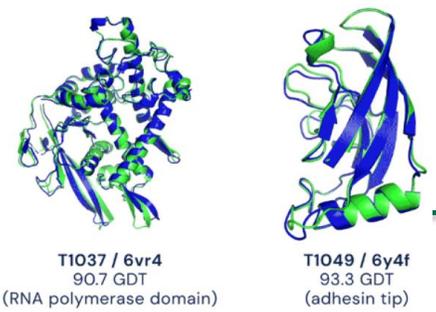
需要・状況に合わせた対応

- HPCユーザはCUDA, HPC SDK, CUDA-Aware MPIあたりさえ整備されていれば十分使える
 - やや暴論？、だが恐らく事実？
- 機械学習ユーザはGPUのヘビーユーザになる可能性があるがスパコンに慣れていない
 - 需要・要望にあわせてHPCユーザよりも手厚いサポートが必要
 - 実際に行ったこと
 - コンテナ (Singularity) 環境の整備
 - ドキュメントの充実 (ベンダーによる利用手引書に加えて幾つかの資料を作成し公開)
 - 講習会を多めに開催 (MPI, OpenMP, OpenACCよりも高い頻度でハンズオン講習会を実施)
 - NVIDIA社と協力してハッカソンイベント等の実施

コンテナ環境の整備

- (管理者権限の都合で運用しにくいDockerではなく) Singularityを導入
 - (今はDockerもroot権限を持たないユーザでも使えるとかなんとか……ちゃんと調べてません)
- スパコンの課題の1つであるソフトウェア環境整備に良い影響
 - 最新ソフトウェアの速やかな提供が難しい・共有環境のためユーザが好き勝手にソフトウェアをインストールしにくい
 - 動作確認が取れていないソフトウェアを提供できない
 - 他のソフトウェアとの競合の可能性に時間がかかる・競合すると導入できない
 - コンテナを持ち込むことである程度は解消 (軽減)
- OSSのツール・フレームワークをとっかえひっかえする (特定バージョンの組み合わせに縛られる) 機械学習ユーザの利用に対応できる
 - 機械学習講習会などでコンテナやPython仮想環境の使い方を紹介し使ってもらっている

AlphaFold (AlphaFold2)の導入



- AlphaFold (AlphaFold2)
 - タンパク質の構造予測に有用なソフトウェア
 - 森脇先生（東大）・望月先生（立教大）らの協力により「不老」Type IIサブシステムにAlphaFold2を導入。利用手順を公開し、機械学習講習会の中でも具体的な利用方法を紹介。
 - Type IIサブシステムではGPUによる高速化以上にSSDによる高速化が劇的
 - 大容量のDBを参照する負荷が非常に大きい
 - HDD(共有ストレージ)では11時間かかった処理がSSDを使うことで10分に短縮された例もあり
 - NVMEESH上にDBファイルを置き、誰でも参照できるようにしてある
- 現在のType IIサブシステムの利用ポイントの算出には占有するGPUの数が使われている
- NVMEESHを使えるリソースグループ（キュー）の最小単位は1ノード=4GPU
- AlphaFoldは4GPUを活用できる作りになっていないためポイントの消費効率が悪い
- →1ノードを4ユーザで共有（1ジョブ=1GPU）しNVMEESHも触れる共有キューを準備中

画像引用元

<https://github.com/deepmind/alphafold>

Type IIサブシステムの現状と課題や展望：現状

- 現状
 - 大規模なGPU計算環境を（パブリッククラウドよりも）安価にかつ安定して提供している
 - コンテナも活用することで、またPython関係の開発環境をユーザレベルで比較的容易に整備できるため、利用者の権限がやや制限された共有スパコン環境でもユーザが比較的自由に計算環境を整えて利用できていると思われる
 - 大量のGPU資源が欲しいが資金を潤沢に持っていない・計算機の管理をしたくない利用者の需要をある程度受け止められている
 - 当初不安したよりも多くのGPU利用に恵まれている

Type IIサブシステムの現状と課題や展望：課題と展望

• 課題と展望

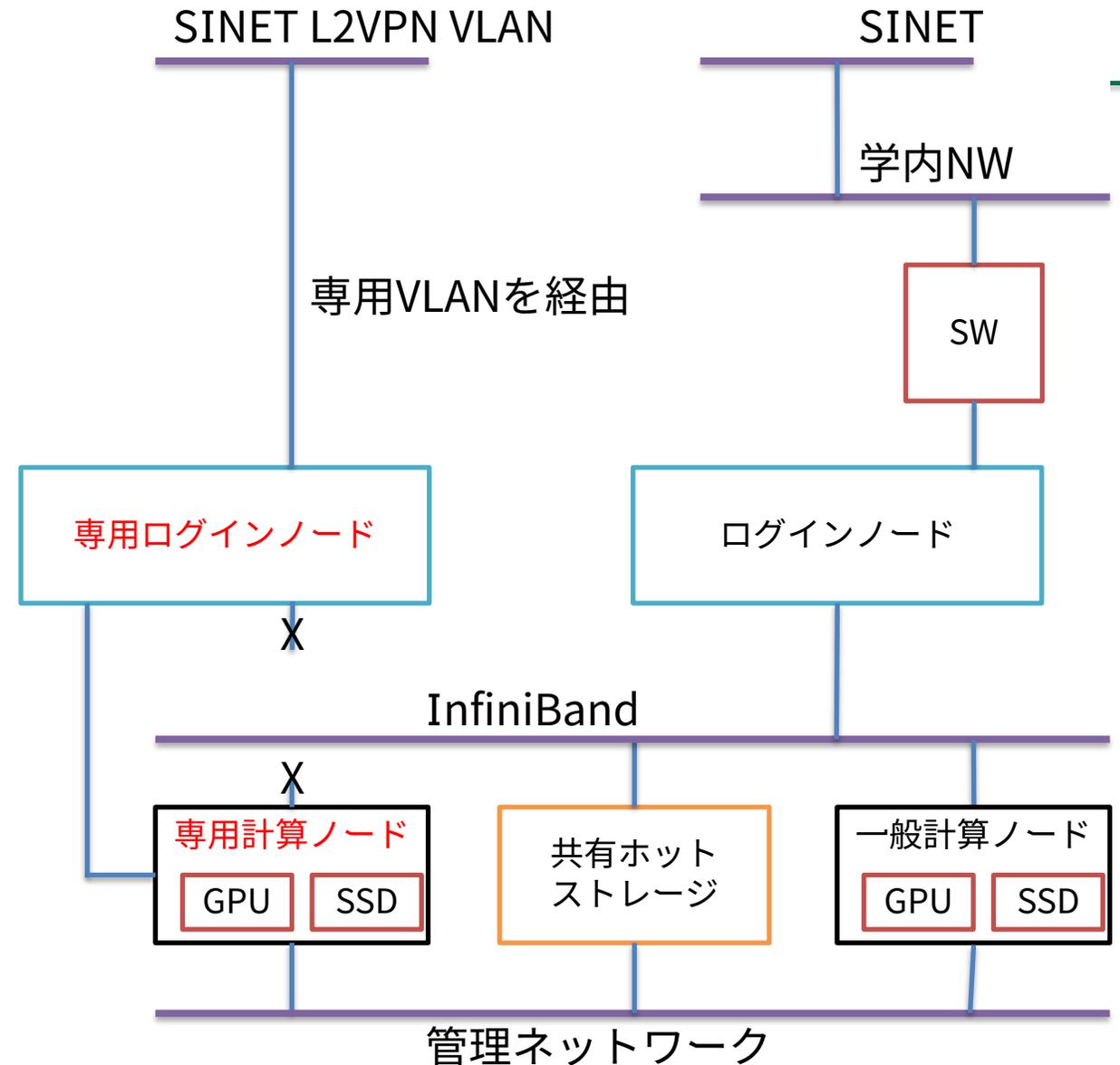
- A100が普及し始めたことやH100の普及が始まることによりV100の魅力が低下する問題
 - 5～6年程度使い続けねばならないセンタースパコンの宿命のようなもの
 - 電気代分をユーザに支払ってもらっているセンターとしては、HWが古くなってきたからと言って安易に値下げできないという事情がある
 - それどころか値上げ待ったなしの世界情勢・電気料金高騰……
 - パブリッククラウドのGPUインスタンスの有利なところ？
- ブラウザからの利用など従来のスパコンではあまり使われていない使い方への対応も必要？
 - Jupyter Notebookの利用方法は提供済
 - AlphaFoldもブラウザからの利用が求められている？
 - 多くのAlphaFoldユーザはGoogle Colaboratory (Colab)上のWebインターフェイス付き無料AlphaFoldを使っているらしい。スパコンでも真似できれば需要が増える？
- 今後も新しい使い方継続して対応していくことはとても重要

スパコン外部とのさらなる連携について

- 「システムをまたいだスパコンの利用」の需要が高まりつつある
 - シミュレーションと機械学習の連携など得意とするシステムが異なる計算を複数組み合わせる使い方
 - ← スパコン外部との連携もスパコン内の複数システムの連携も考えられる
 - センサデータなどを参照して行う使い方
 - ← 主にスパコン内部とスパコン外部との連携
- こうした新しい利用形態は今後さらに広がっていくだろう
- 初めから想定していなかった使い方への対応は容易ではない
 - システムの設定変更が必要となる可能性
 - ハードウェアの追加が必要となる可能性
- 科学的な価値の可能性 ⇔ 追加金額・手間
- 人的資源も予算も限られるが、可能な範囲で様々な利用に対応する必要がある

最近行っている取り組み

- 実験中：外部からセキュアに計算ノードを利用するためのテスト
 - 専用ログインノードとVPN・VLANを利用し、なるべくダイレクトに
(共有ストレージに一切のデータを置かずに)計算ノードのGPUを使う
 - 医用画像などデータの管理に厳しい分野で有効に活用できる可能性がある



コールドストレージ

- 「不老」はHDD RAIDのホットストレージに加えてSONYのOptical Disc Archiveを用いたコールドストレージシステムを提供
 - 保存寿命100年以上の追記型光ディスク
 - 1カートリッジあたり5.5TB
 - 総搭載容量6PB（半分くらい販売済）
 - システム総搭載可能容量10.89PB
 - ディスクの持ち帰りが可能
 - USB外付け型ドライブの貸し出しも可能
- 重要データの保存に好適
 - 外部システムからネットワークを通じたコールドストレージへの書き込みも対応可能
 - （オンザフライではなく一時保存して書き込み）
 - 消費電力上のメリットもあり名大では重要と考えているが、コールドストレージを備えたスパコンは少ない……
 - 従来からのスパコンのストレージやクラウドのストレージとは別の有用性があると思うのだが？



全体のまとめ

- 約2年間利用してきた「不老」について、「クラウド」のキーワードを絡めて現在の状況を報告した
- クラウドシステムとType IIサブシステムの状況を中心に報告
 - クラウドシステム
 - パブリッククラウド連携などの仕組みはない
 - WebシステムUNCAIからのクラウド的利用とバッチジョブ実行のハイブリッド運用だが、バッチジョブ実行による利用がほとんどを占めている
 - クラウド的な利用はHPCユーザには人気がない？適切な利用方法が共有できていない？もっとクラウド的に使えるものが望まれる？
 - Type IIサブシステム
 - x86 + GPU + SSD 環境として有効に活用されている
 - コンテナを用いた機械学習やGPU+SSDフル活用のAlphaFoldなど新しい使われ方の需要が大
- HWの更新はできないが、今後もユーザの需要・意見を反映してソフトウェア環境など使い勝手を改善し、「老けない」システムを提供していきたい