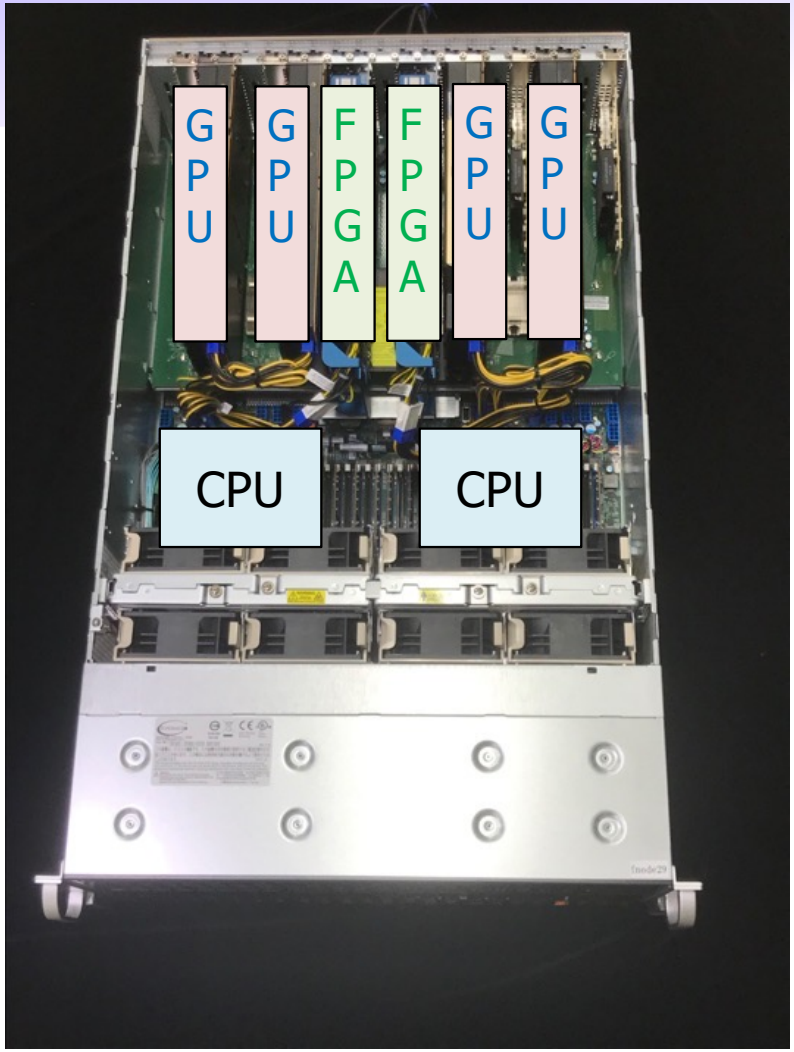
# 筑波大学における
# ビッグメモリスーパーコンピュータの導入

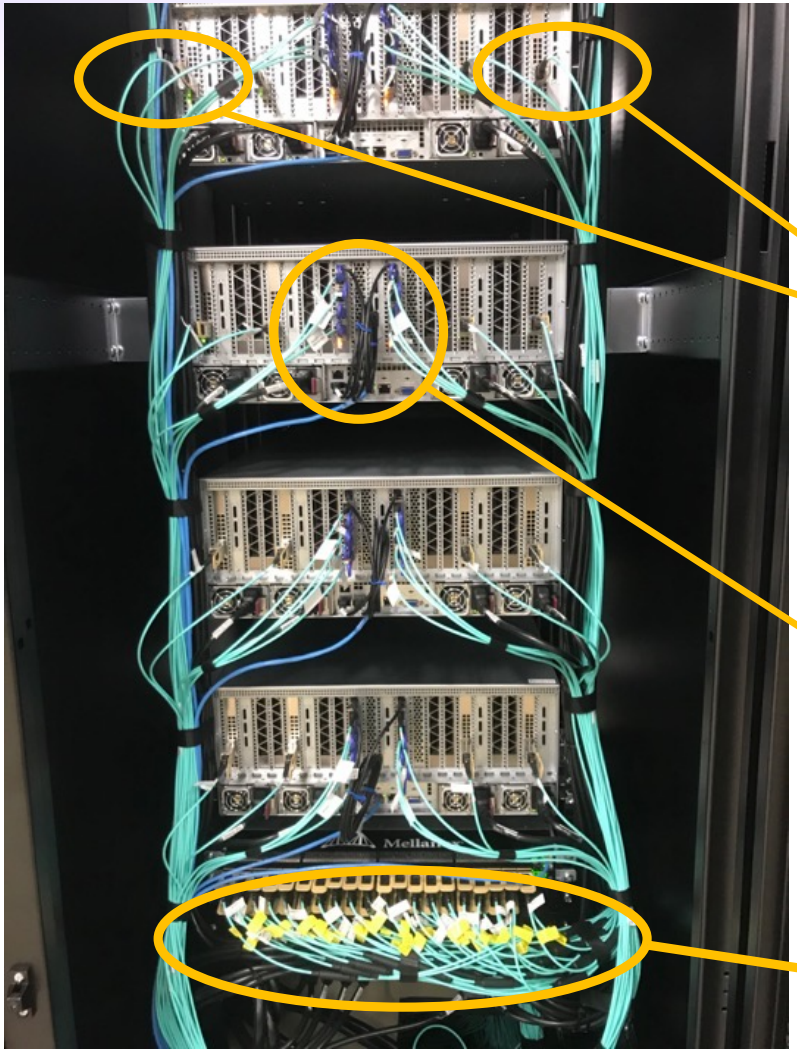
朴泰祐（＆建部修見）
筑波大学計算科学研究センター・センター長
taisuke@ccs.tsukuba.ac.jp (& tatebe@cs.tsukuba.ac.jp)

# 筑波大CCS（Center for Computational Sciences)で稼働中の GPU+FPGA多重複合演算加速スーパーコンピュータCygnus




全81ノード全てにNVIDIA Tesla V100 GPU x 4基搭載
うち32ノードには加えてIntel Stratix10 FPGAボード (BittWare 520N)
総ピーク性能 2.43PFLOPS (DP) （GPU+CPU）



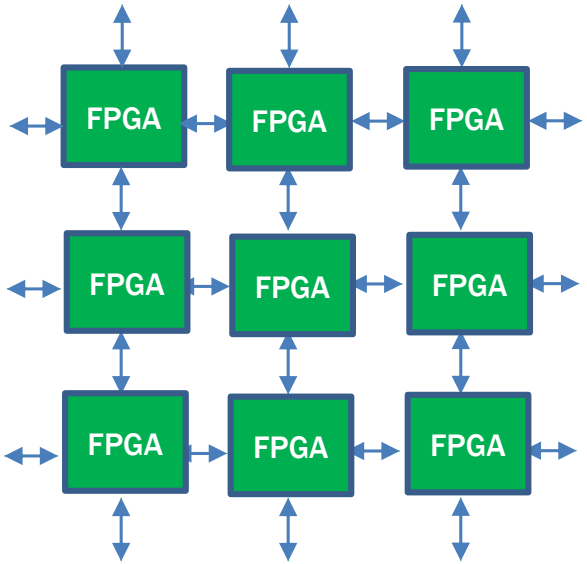
IB HDR100 x4
⇨ HDR200 x2

100Gbps x4
FPGA optical
network

IB HDR200
switch (for
full-bisection
Fat-Tree)

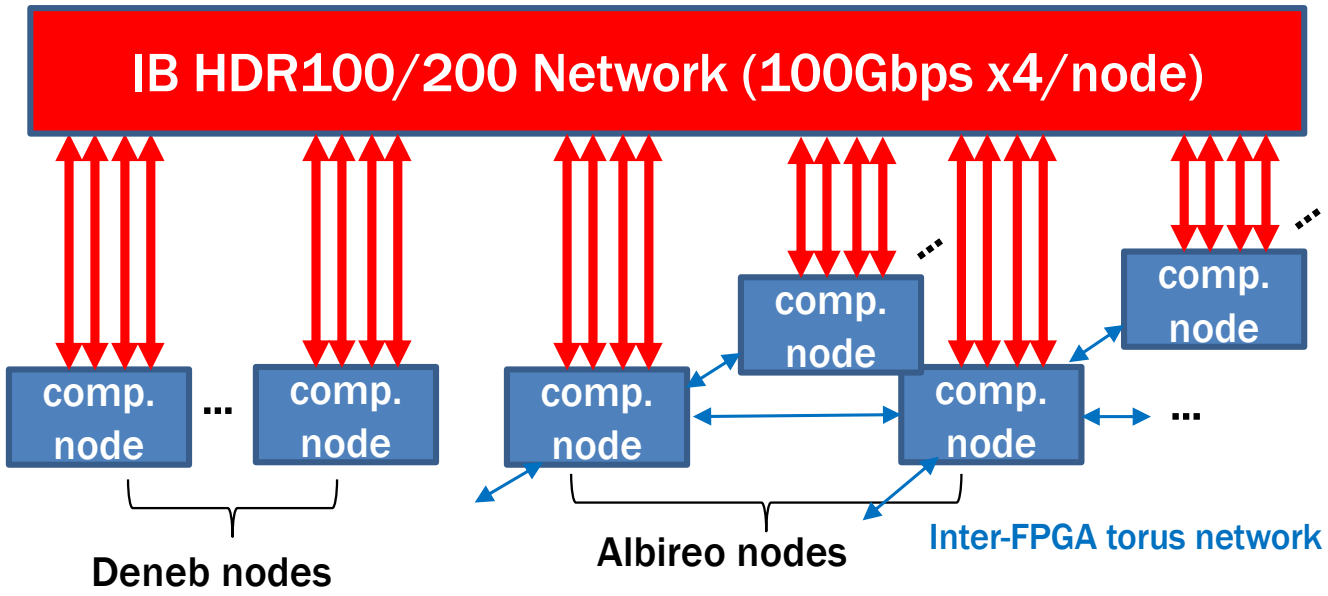
# Two types of interconnection network

## Inter-FPGA direct network (only for Albireo nodes)




64 of FPGAs on Albireo nodes (2 FPGAS/node) are connected by 8x8 2D torus network without switch

## InfiniBand HDR100/200 network for parallel processing communication and shared file system access from all nodes




For all computation nodes (Albireo and Deneb) are connected by full-bisection Fat Tree network with 4 channels of InfiniBand HDR100 (combined to HDR200 switch) for parallel processing communication such as MPI, and also used to access to Lustre shared file system.

# Cygnusが目指しているもの

- GPUのみによる演算加速をさらに強力にするため（特にstrong scaling）FPGAによる演算加速も取り入れる
  - 単純な大規模SIMDだけで不足する部分的な並列性欠如，条件分岐による効率低下，頻繁なノード間通信に対応
  - Multi-Physicsアプリケーションを中心とする適材適所的な演算加速器の適用
- 計算性能＋ノード間通信重視
  - 現在の流れ： Big Data + Extreme Computing
  - Big DataについてもGPUの有効活用（＋FPGAによるさらなる高速化）で対応
    ⇒より大規模な計算への対応は？
- 大容量メモリを用いた新しいソリューションによりBig Data処理を加速

# 次の一手：Big Memory Supercomputer ～ Cygnus-BD（仮称）

- 最新技術である不揮発性メモリを用いた新たなプラットフォーム
  - CPU core当たりのメモリ容量はどんどん減少している
  - PMEM (Persistent Memory)を用いた，一桁近く容量の増えたユーザメモリにより，メモリ容量ボトルネックとなる計算科学問題に対応
- PMを temporal file system として利用
  - PMをIOデバイスとしてアクセスし，ノード分散高速ファイルシステムを構築
  - 大規模計算とAI処理のステージングを高速処理
- 最新GPUを搭載
  - traditional HPC + AI/BigData処理
  - AI for HPC
- その他のPMを中心とする研究とアプリケーション実行を推進

⇒ Cygnus-BD (Cygnus for Big Data)

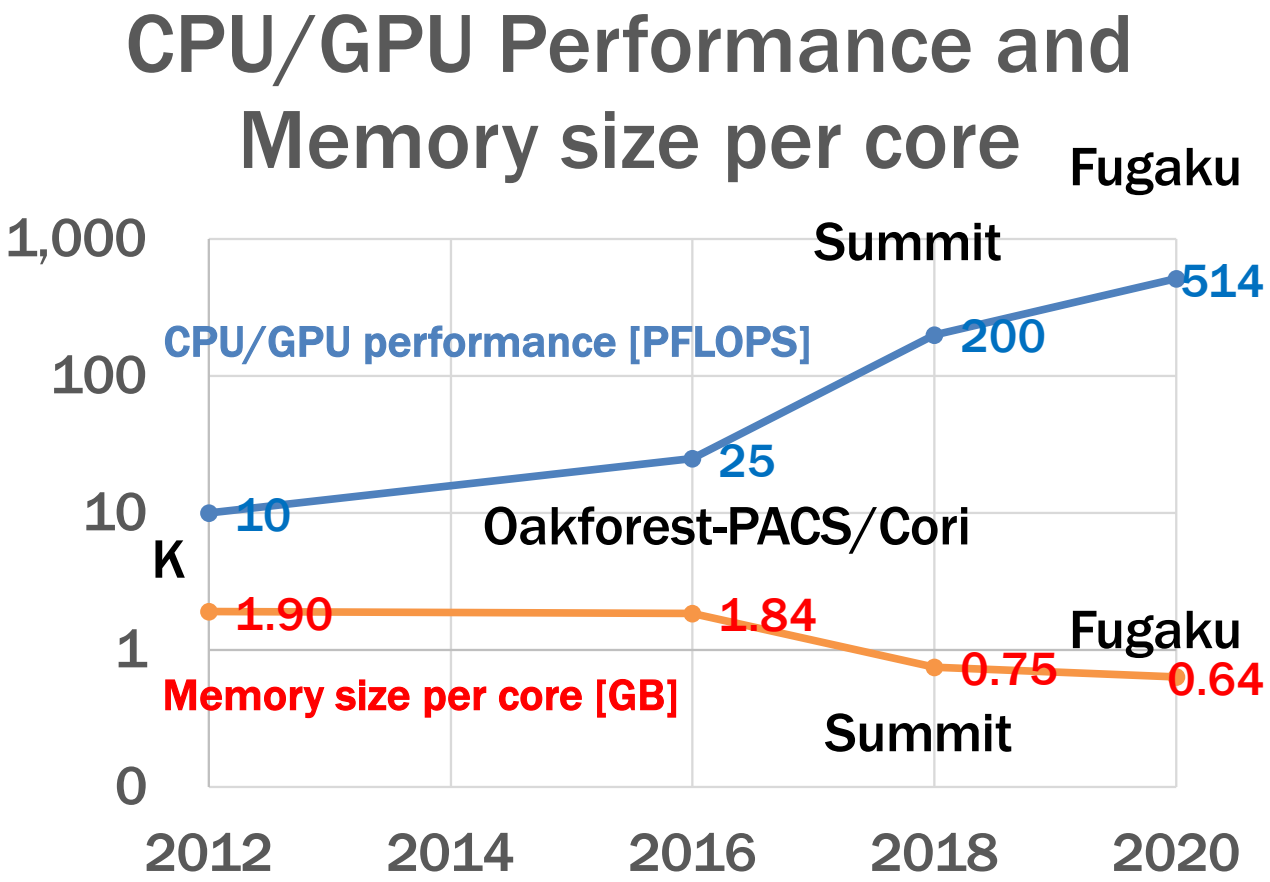
⇒ 従来のCygnusの拡張として考える（旧Cygnus = Cygnus-EC）

# 筑波大CCSのスーパーコンピュータ運用方針

- 筑波大CCSは今後もJCAHPCの下で東大ITCとの連携を継続
  - 両センターは次期OFPシステム (OFP-II)の開発・導入・運用を継続する
- 従来のCOMA, Cygnus(-EC) のようにJCAHPCシステムと平行して独自システムの開発・運用を継続
  - Cygnus-ECの運用継続（～2024年度）
  - Cygnus-BDの導入（2022年11月～）
- ２種類のスーパーコンピュータの棲み分け
  - JCAHPCのOFP系では広範囲のアプリケーションに対し，使いやすいシステムとして幅広いユーザ層を支える ⇨ OFP2（仮称）
  - 筑波大独自システム（ex. Cygnus）ではやや「尖った」ユーザも視野に入れ，高性能・省電力な Advanced Accelerated Supercomputing 環境を提供していく

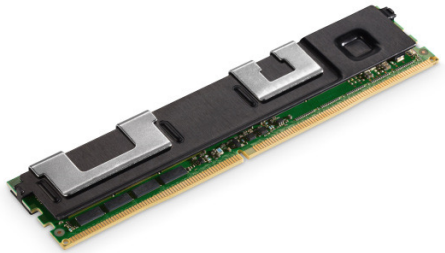
# なぜ Big Memory が必要なのか？

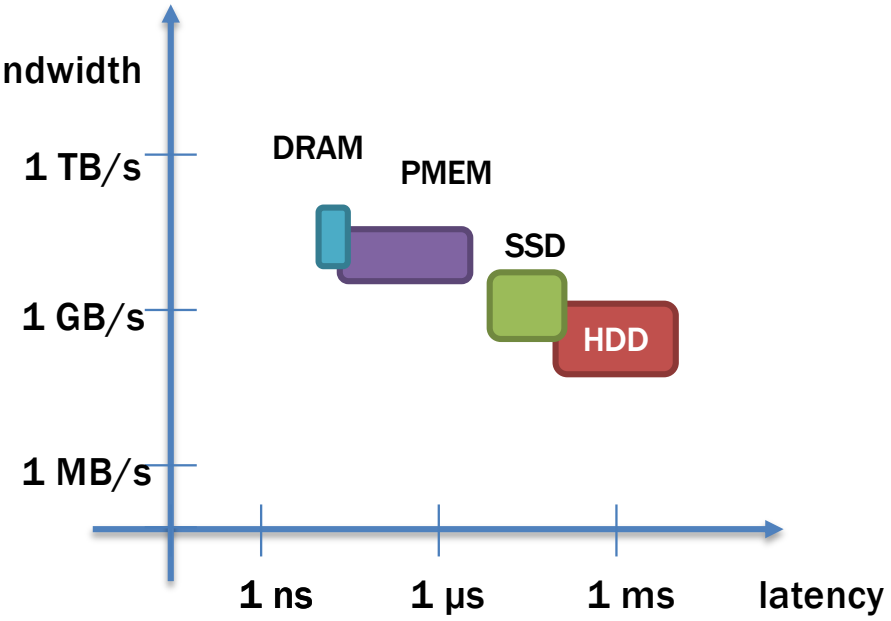
- CPUの性能向上50倍に対し、メモリサイズは3.8倍にしか増えていない
- データ科学・AIによる科学に大きな影響
  - メモリサイズとストレージ性能が重要
- Persistent Memory の導入
  - Memory mode: 大容量メモリ
  - AppDirect mode: 高速ストレージ



# Persistent Memory（永続性メモリ）

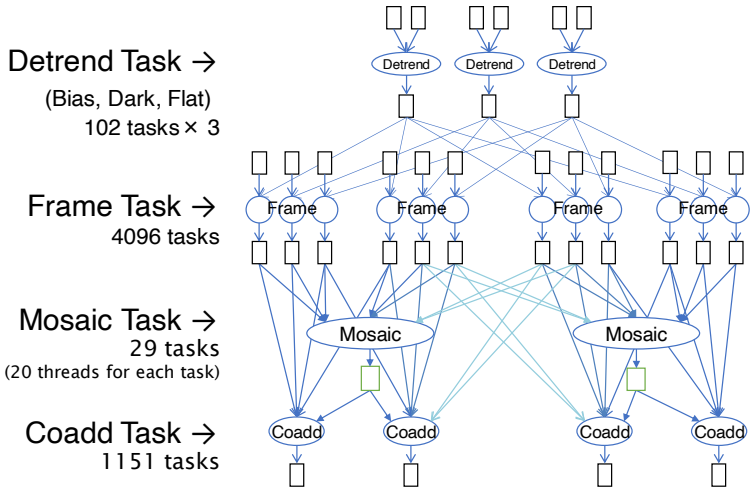
- 容量/価格がDRAMより遥かに良い
- 最小レイテンシ ~60 ns (DRAMより大きくは劣っていない)
- バンド幅はDRAMの半分程度
- Memory mode
  - 性能をそれほど低下させず大容量メモリが利用可能
  - DRAMを last level cache に使える
- App direct mode
  - Byte-accessible な永続メモリとして使え，I/O性能を飛躍的に高められる

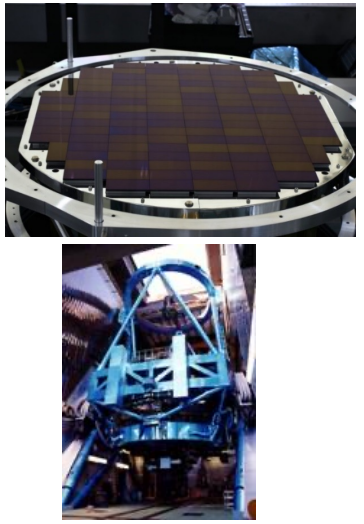




# 大規模データに基づく計算・データ科学：宇宙物理

- すばる望遠鏡の観測データの解析



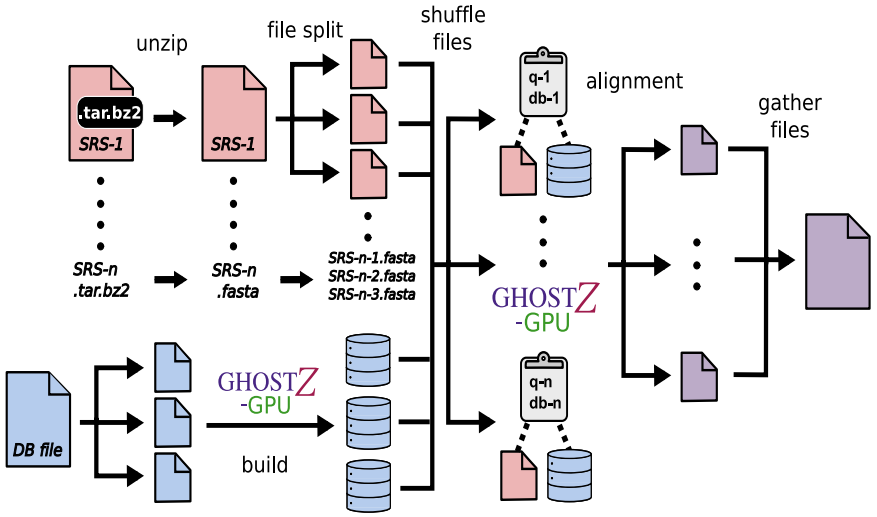



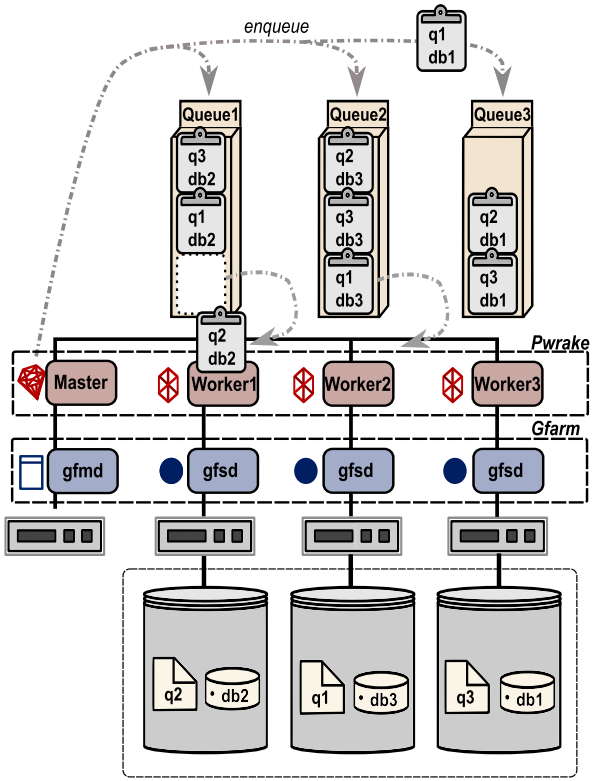
© NAOJ

- 現在はファイルを介したステージングによる処理
  ⇒ 大規模メモリによりオンラインのままで処理が可能に（**in situ**）

# 大規模データに基づく計算・データ科学：遺伝子情報処理

- 環境遺伝情報 ⇒ 特定遺伝情報解析
- メモリ不足 ⇒ 問い合わせベースのファイルI/O処理
- 大容量メモリによる大幅な性能改善を期待



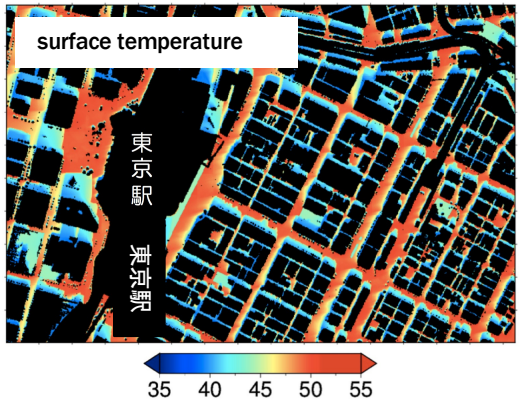



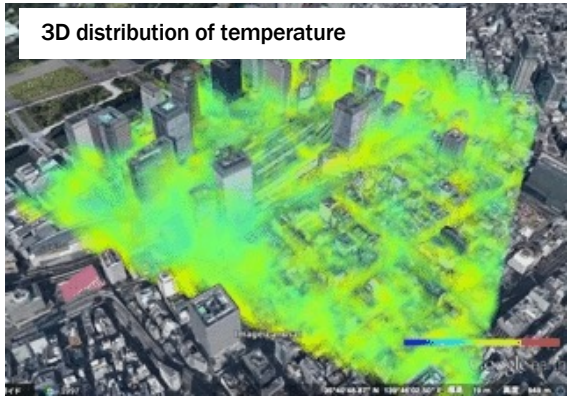
# 大規模データに基づく計算・データ科学：気象シミュレーション

## マルチフィジックス超高解像度都市気象シミュレーション

- **City-LES: Large Edy Simulation** による都市気象シミュレーション，力学課程，建物，太陽光輻射，地表面の植生などを含む風速・気温・気圧などの推測

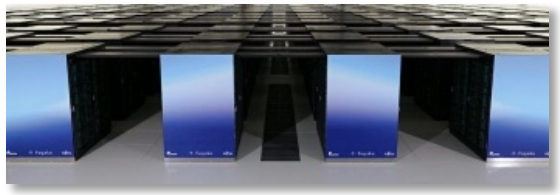
**TOKYO2020 model around Tokyo Station, 5m grid**



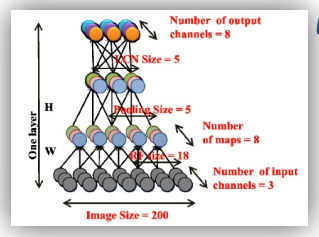




- 高解像度で時間発展シミュレーションを実施 ⇨ データ量が膨大
- **GPU化により最大17倍の速度向上 ⇨ strong scaling による通信を避けたい**

# Big Data & AI for HPC in Bio-Science & Drug Discovery



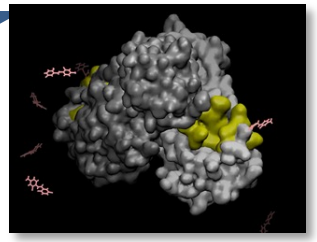
## Data Science & Machine Learning

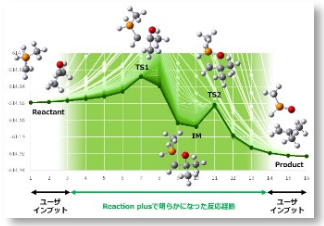



high speed
high accuracy

data generation

## Simulation (MD, docking)






minimum real data by medical research

## Big Data (medical experiment and database)

(slide by Yasushi Okuno, Kyoto U.)

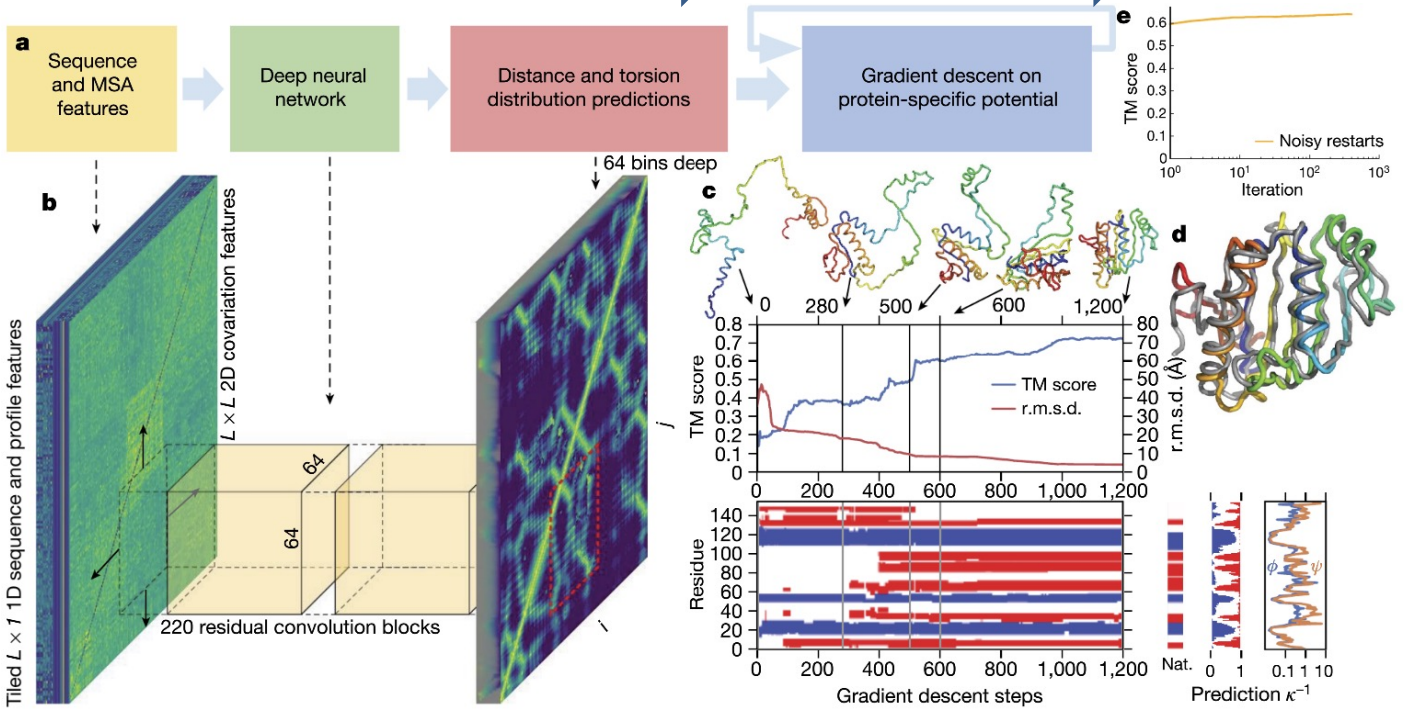
# Alpha-Fold: Protein structure prediction by AI with

**Protein structure prediction by ultra high-performance MD → AI replaces**

deep learning

molecular simulation

amino acid array



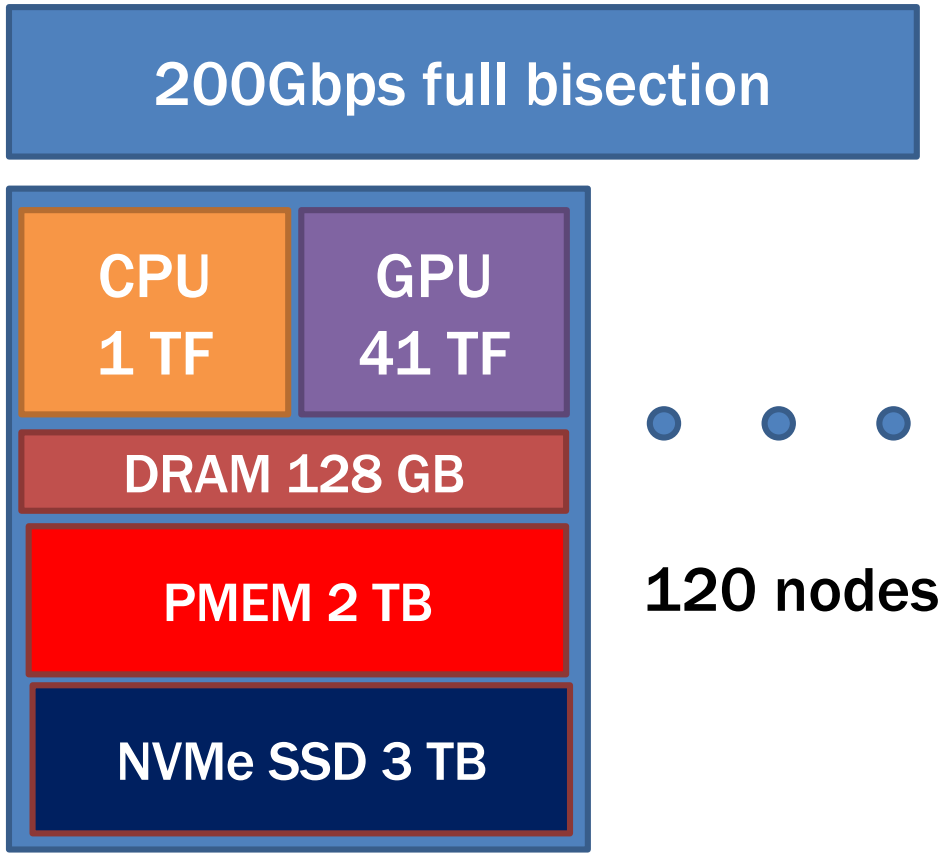

protein structure

(slide by Yasushi Okuno, Kyoto U.)

Backpropagation

# Cygnus-BD（仮称）: 世界初の PMEM + Xeon-next + H100

- 2022年11月稼働開始
- ピーク性能
  - 120 nodes, 6 PFlops, 240 TiB
- ノード構成
  - NVIDIA Tesla H100 PCIe
  - Intel next gen Xeon
  - Intel next gen Optane
- ノード性能
  - 2 TFlops (CPU), 48 TFlops (GPU with tensor core)
  - 128 GB DRAM, 2 TiB PMEM
  - 6.4 TB NVMe SSD
- 相互結合網
  - 200 Gbps full bisecion (NVIDIA Quantum-2 IB)
- 並列ファイルシステム (DDN)
  - 7.1 PByte, 40 GB/s



# Cygnus-EC (現在のCygnus) とCygnus-BDの比較

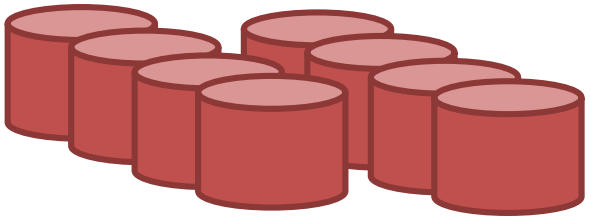
| | Cygnus-EC (2019) | Cygnus-BD (2022) |
|---|---|---|
| PFLOPS (DP) | 2.3 | 6.0 (2.6x) |
| PMEM (TiB) | 0 | 240 |
| CPU/nd (TFLOPS) | 0.16 | ? (???) |
| GPU/nd (TFLOPS) | 7 | 48 (6.8x) |
| FPGA/nd (SP TF) | 1.2 | 0 |
| DRAM/nd (GiB) | 192 | 128 (0.67x) |
| PMEM/nd (TiB) | 0 | 2 |
| Storage (PB) | 2.4 | 7.1 (3x) |

System Integration: NEC

(subject to change)

# Ad hoc file systemの研究

- Temporal parallel file system using node-local storage
- Fill the performance gap between CPU/GPU and storage



- We are developing CHFS (Consistent Hash File System) ad hoc file system to utilize persistent memory
  - No metadata server, no sequential processing for performance and scalability

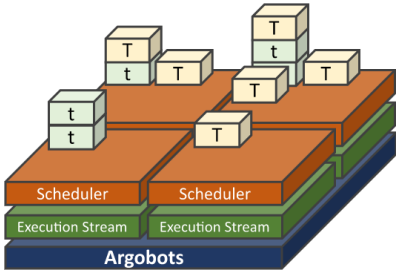
* O. Tatebe, et. al., **"CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory", HPC Asia 2022**

# Design goal of CHFS

- Utilizing persistent memory performance
  - In-memory persistent key-value store (not block-based file system)
- Reduce metadata overhead and achieve scalable performance improvement
  - No dedicated metadata server
  - No sequential execution
  - Based on highly parallel distributed key-value store without any central data structure
- Improve single-shared-file performance
  - File is divided into fixed-size chunks to distribute a single file among servers

# Implementation of CHFS

- Mochi-Margo [JCST 2020]
  - https://mochi.readthedocs.io/en/latest/
  - Communication library using Mercury and Argobots
- Mercury [Cluster 2013]
  - Async RPC, RDMA communication library
  - libfabric, CCI, shared memory plug-ins
- Argobots [IEEE TPDS 2018]
  - Light-weight thread library
- pmemkv
  - cmap – concurrent hash map

# まとめ

- 筑波大CCSの次の一手として，HPC/BD/AIの総合的な高速演算・大容量データ処理のためPMEMを最大限に活用する新スーパーコンピュータCygnus-BD（仮称）を導入
  ⇨ 2022年11月稼働開始予定
- これまでHPC技術はデータ科学・AIに大きく貢献してきたが，今後はAI技術に支えられた効率的なHPCを実現し，エコサイクルを構築すべき
- ノード当たりのCPU性能とメモリ容量のギャップは年々広がっており，データ科学にとって大きな障害となりつつある
  ⇨ 性能向上のための並列化より大規模データのため並列化せざるを得ない
- PMEMの活用により，大容量メモリと超高速並列分散 ad hoc ファイルシステムを両立させ，新しい時代のデータ科学に備える
- 年度後半に筑波大学学際共同利用などに投入予定，HPCIも視野に