AMDが支えるAI for Scienceの最前線

2026年7月2日

日本AMD株式会社
大原久樹 (Hisaki.Ohara@amd.com)

AMD
together we advance_

AMD INSTINCT

Strategic Pillars

Performance & TCO Leadership

Leadership products on an annual cadence with breakthrough economics

Ease of Migration

Drop-in compatible with existing infrastructure for hardware and software

Commitment to Openness

Investment & participation in open standards across the entire ecosystem

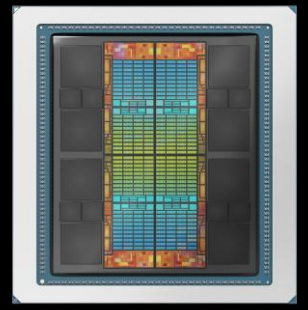
Customer Focused

Roadmap and support structure geared toward customer success

Extending Leadership from GPU to Rack-Scale Infrastructure

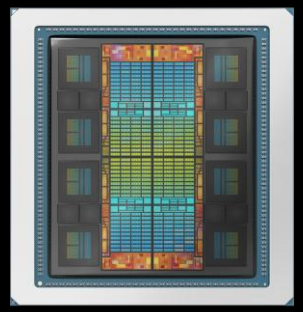
AMD INSTINCT

AMD Instinct™ MI300A/X



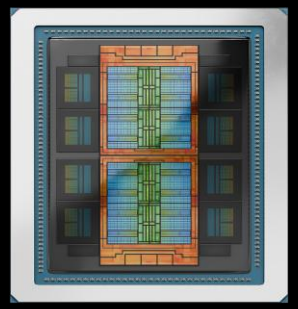
2023

AMD Instinct™ MI325X



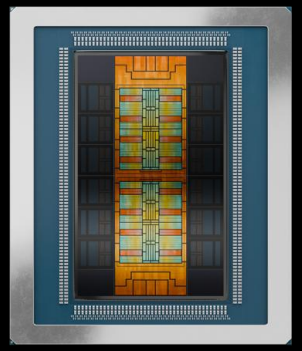
2024

AMD Instinct™ MI350 Series



2025

AMD Instinct™ MI400 Series



2026

AMD Instinct™ MI500 Series

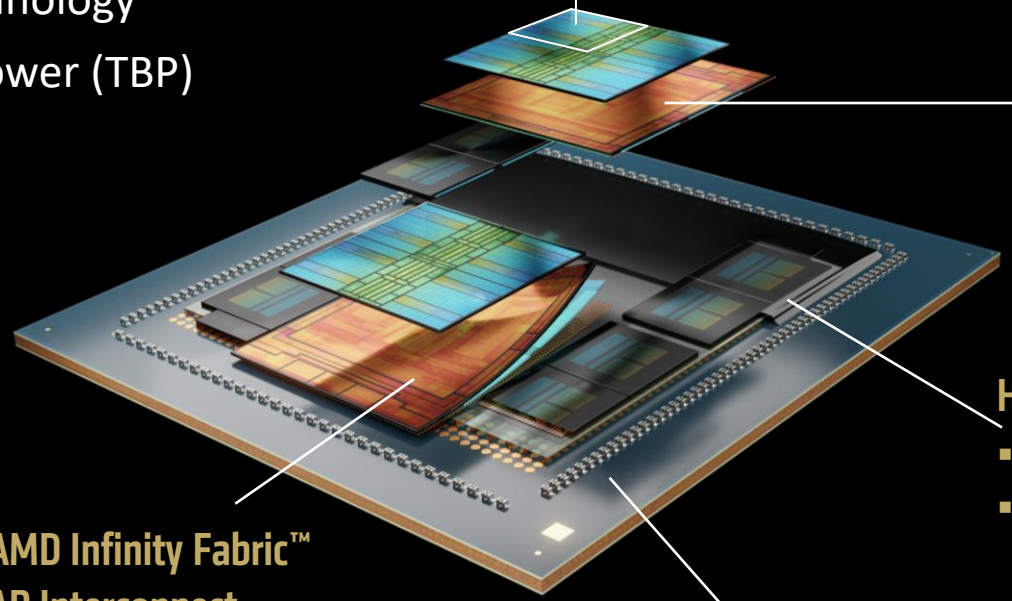
2027

Roadmap subject to change

AMD Instinct™ MI350 Series GPU

State-of-the-Art Construction

- 185 Billion transistors
- 3D Multi-Chiplet (2 chiplet types + HBM3E)
- Heterogenous process technologies
- Proven COWOS-S packaging technology
- Up to 1,400 watts Total Board Power (TBP)



Accelerator Complex Die (XCD)

- Latest TSMC 3nm (N3P) Technology
- Latest process for compute perf/watt optimization
- Max 8 per package (Max 4 per IOD)

AMD I/O Base Die (IOD)

- TSMC 6nm FinFET (N6) Technology
- Mature process for optimal yield & cost
- Max 2 per package

HBM3E

- Max 4 stacks per IOD (Max 8 per package)
- Up to 288 GB of HBM memory

AMD Infinity Fabric™ AP Interconnect

Package

AMD Instinct™ MI350 Series Compute Units (CU)

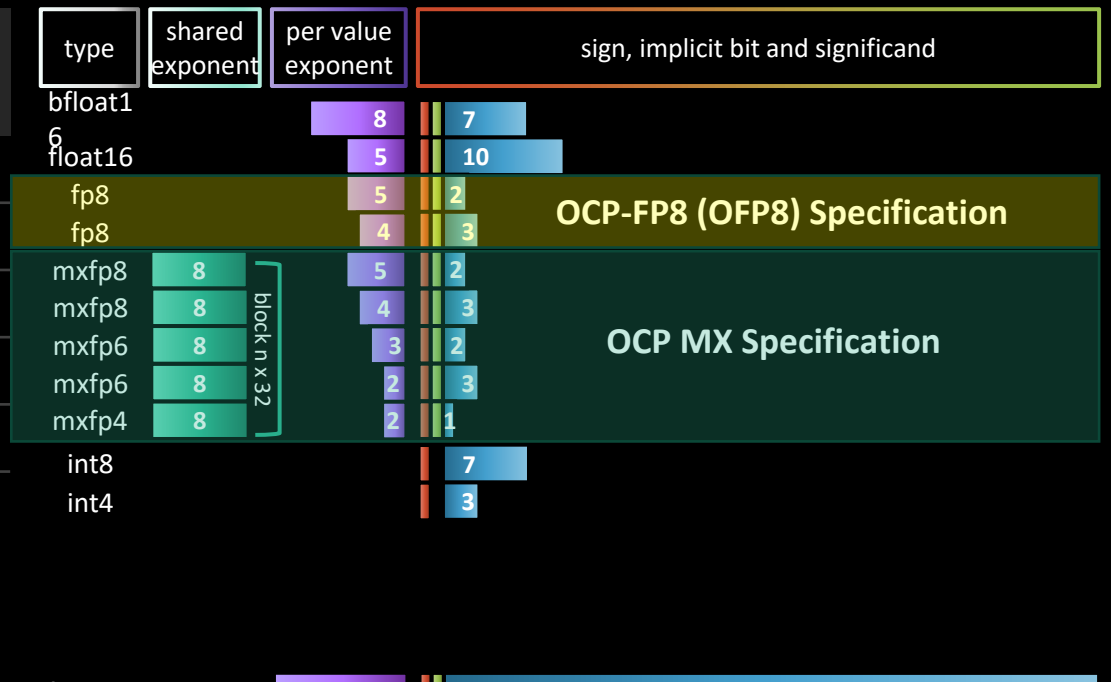
Supported Data Formats & Generational Performance Improvements

AI Lower Precision Data Types

Computation	MI350 Series (FLOPS/clock/CU)	MI355X (Peak Theoretical)	MI355X Peak Speedup‡ (Compared to MI325X)
Vector FP16+	256	157.3 TFLOPs	~1.0x
Matrix FP16/BF16	4096	2.5 PFLOPs	1.9x
Matrix FP8	8192	5.0 PFLOPs✧	1.9x
Matrix INT8/INT4	8192	5.0 PFLOPs	1.9x
Matrix MXFP6/MXFP4	16834	10 PFLOPs	New to MI350*

High Performance Computing (HPC)

Vector FP64	128	78.6 TFLOPs	~1.0x
Matrix FP64	128	78.6 TFLOPs	~0.5x
Vector FP32+	256	157.3 TFLOPs	~1.0x
Matrix FP32	256	157.3 TFLOPs	~1.0x



Peak theoretical performance without Sparsity

‡ Compared to MI300X * AMD Instinct™ MI300X GPUs do not support MXFP6, MXFP4

+ Refers to non-packed vector instructions on AMD Instinct™ MI300X and MI350X Series GPUs

✧ AMD Instinct MI350 Series GPUs support both OFP8 format as well as OCP MX formats

◊ AMD Instinct™ MI350 Series GPUs support TF32 through software emulation

AMD
together we advance_

AMD INSTINCT MI350P

Integrate Into Existing Data Center Infrastructure

Leadership Costs and Exceptional AI Performance

Open Ecosystem and Enterprise Ready AI Software

AMD Instinct™ MI350P PCIe® Card

Deploy and Scale in Existing Data Center Infrastructure

Built for Existing Enterprise Infrastructure

10.5” FHFL Dual-Slot Design
600W Passive Air-cooled
450W Power Cap Configurable
PCIe Gen 5 Host Interface

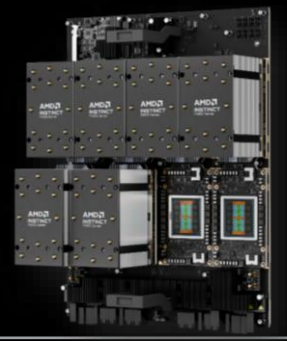
Leadership GPU Architecture

AMD CDNA 4™ Architecture
128 Compute Units
Support for MXFP8, MXFP6, MXFP4
144GB HBM3E Capacity, up to 4.0 TB/s
Dedicated Video and JPEG HW Decode

Open Enterprise Ready Software

Enterprise AI Reference Stack with Inference Microservices
GPU Independence with Cross Platform Interoperability
License-Free, Open Source, Standards based Software
Broad Ecosystem to support Custom Software Stacks

AMD AI Solution Portfolio – Air Cooled



	EPYC™ CPU	Radeon AI PRO™	AMD Instinct™ PCIe®	AMD Instinct™ UBB8
Positioning	Mixed Workload General purpose + AI Batch Inference	Mixed Workload Edge AI + Graphics	Low-latency, Memory-intensive Compute	High- throughput AI at Scale for Inference + Training
Use cases	Inference: Classical ML, RecSys, SLM/RAG, NLP	SLM Inference / RAG / MCP / Txt to Image / Video AI	SLM / MLM/ LLM Inference / RAG	Large-scale LLM Inference & Training
Model Size* (Max)	~ 30B per CPU	~ 40-50B per GPU	~ 200-250B per GPU	~up to 500B per GPU
AMD offering	AMD EPYC 5th Gen	R9700S | R9600D	MI350P[1]	MI350X | MI355X
GPUs/node	0	Up to 8GPUs per node	Up to 8 GPUs per node	8 GPU per node
GPU-to-GPU Interconnect	N/A	N/A	N/A	Fully-connected 8-GPU Infinity Fabric
Memory Capacity with Peak Bandwidth	614 GB/s per CPU 6400 MT/s	32G GDDR6 640 GB/s	144GB HBM3e 4 TB/s	2.3TB HBM3e 8TB/s per GPU

*Model size is estimated based on FP4/MXFP4. Estimated by AI

Introducing

AMD Instinct™ MI400 Series GPUs

Leadership Performance, Open Standards, Gigawatt Scale

AMD
together we advance_

AMD Instinct™ MI400 Series GPUs

AMD Instinct™ MI455X GPU

AMD Instinct™ MI430X GPU

AMD CDNA 5

HBM4
Leadership capacity

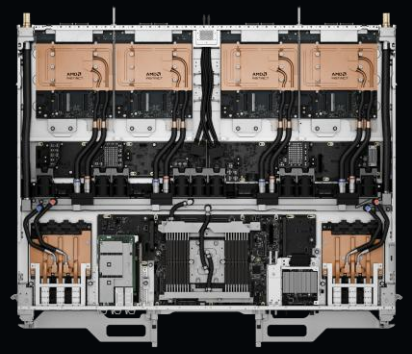
AMD Helios Rackscale Solution

Rack Performance is Powered by AMD Instinct™ MI455X GPUs

Position the AMD Helios rackscale solution from the top down—rack to tray to EAM to die—with the Instinct MI455X GPU as the hardware foundation of the system.

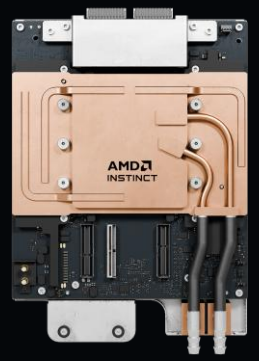
AMD Helios Rackscale Solution

18 trays / 72 GPUs



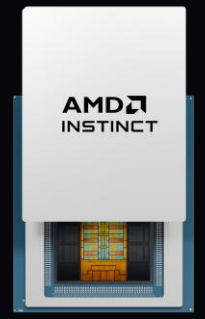
4-GPU compute tray

Repeatable Building Block



AMD Instinct MI455X EAM

Module-Level View



AMD Instinct MI455X die

Silicon/Package

Per-GPU Capability

Up to 40 PFLOPs
Peak 4-bit Precision (OCP MXFP4)

Up to 20 PFLOPs
Peak 8-bit Precision (OCP MXFP6/MXFP8)

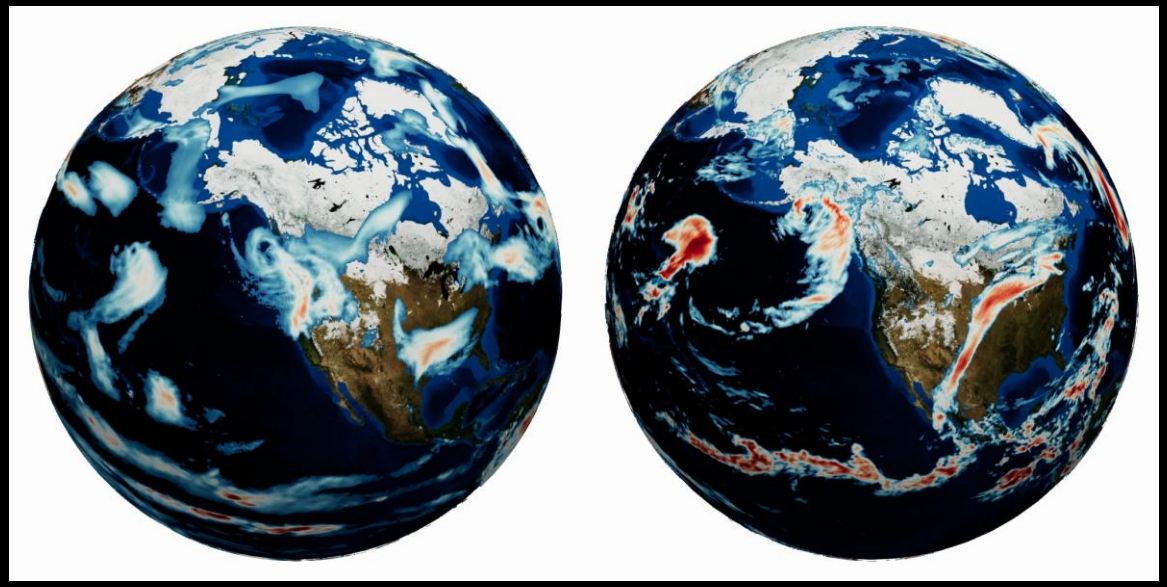
Up to 432 GB
HBM4 Memory

Up to 19.6 TB/s
Memory Bandwidth

Up to 600 GB/s
Scale out Bandwidth / per GPU

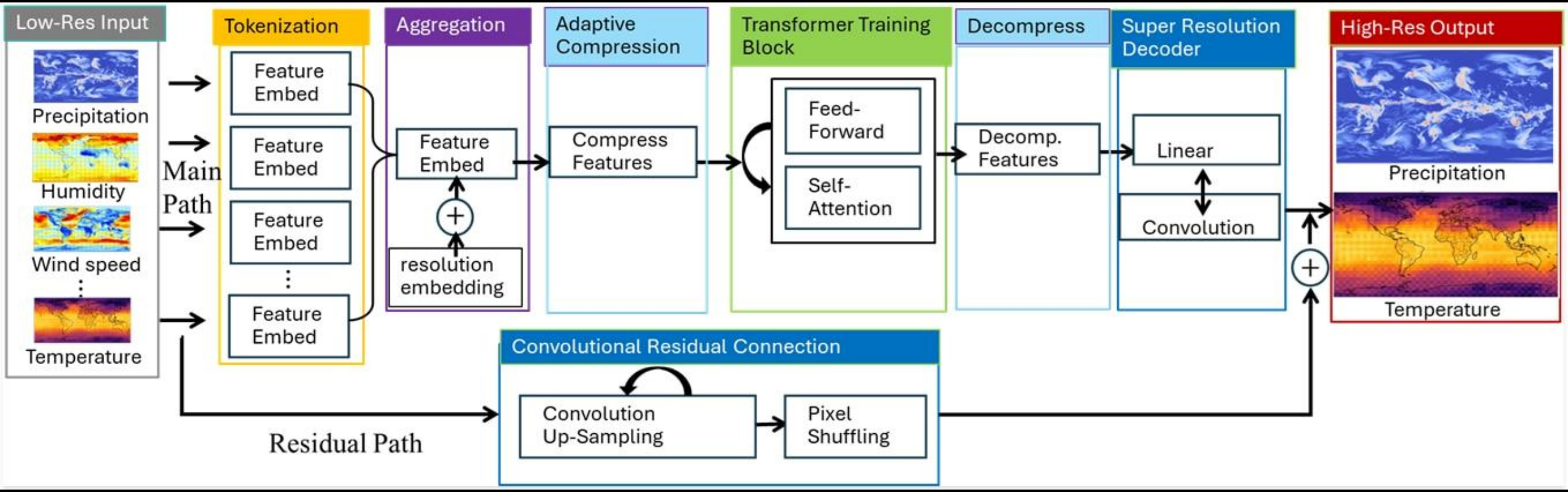
[ORNL] ORBIT-2: Advancing Earth System Intelligence at Exascale

- 全球の気象予測は進化しているが、10–30 km級の解像度では局地的な豪雨・高温・地形影響を十分に表現しにくい
- 一方、災害対応・インフラ計画には、1–5 km級、あるいはそれ以下の空間解像度が重要
- ORBIT-2の目的は、低解像度の気象場から高解像度の気象変数を推定するdownscalingを実現すること



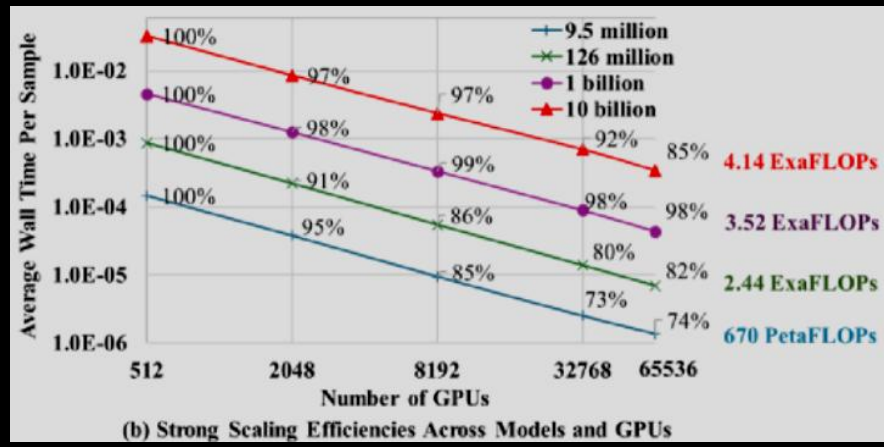
左：28km解像度の入力イメージ　右: 7km解像度にdownscaleされた予測結果

ORBIT-2: Model



- Reslim Architecture
- TILES Algorithm: provide efficient parallelism across GPUs at training

ORBIT-2: Exascale training and Inference



- Training
  - Frontier 上で、AMD EPYC CPU + AMD Instinct MI250X GPUを使用
  - PyTorch + ROCm 6.3.1、FlashAttention、FSDP、Tensor Model Parallelism
  - 最大10B parameters、65,536 GPUs、4.1 EFLOPS (BF16)、
    74–98% strong scaling efficiencyを達成

- Inference
  - 元々、”days or weeks on a large supercomputer” → 0.55 sec on 10B-parameter model
  - ORBIT-2 based Weather and Climate Downscaling and Downscaled Global Forecasts on AMD Instinct
    - https://rocm.blogs.amd.com/artificial-intelligence/orbit2-inference/README.html#installation-setup-for-orbit-2-inference
    - https://github.com/silogen/ai-samples/tree/main/ai4sciences/orbit2-inference

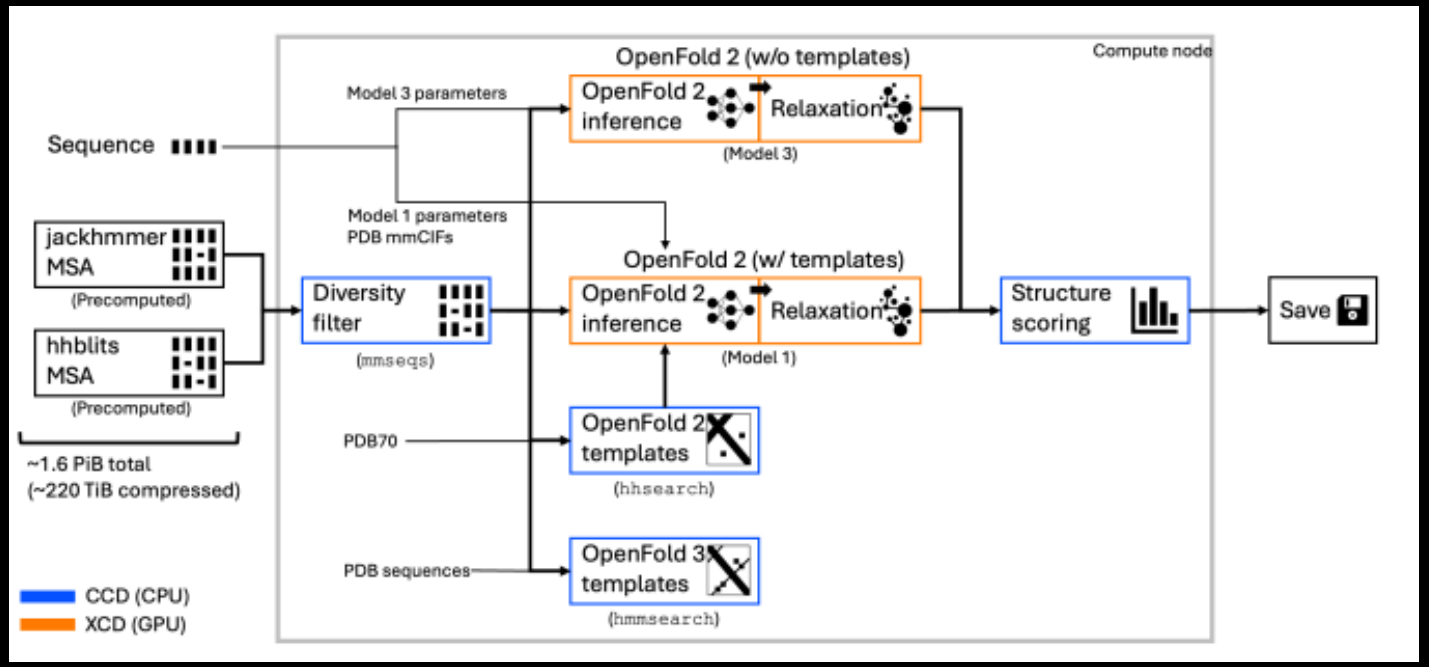
- 成果
  - 全球スケールで高解像度の 降水量・気温 を推定
  - GenCastなどの予測モデルと組み合わせることで、高解像度な全球気象予測への応用の可能性を示した
  - SC25 Best Paper Award / Gordon Bell Prize finalist

- References
  - https://www.amd.com/en/blogs/2025/earth-system-modeling-with-orbit-2.html
  - https://rocm.blogs.amd.com/artificial-intelligence/orbit2-inference/README.html

[LLNL] MI300Aを用いたOpenFold用蒸留データの生成

- 実験構造データだけでは、OpenFold3のような大規模モデルの学習には不足
- OpenFold3の学習に必要な蒸留データをOpenFold2の推論で大規模生成
  - AlphaFoldはライセンス上、蒸留データ生成への利用に制約

- ElmerFold
  - CPU: 前処理・IO
  - GPU: OpenFold2の推論



ElMerFold

- 推論基盤
  - El Capitan 11,000 ノード、44,000 MI300A APU

- 性能
  - ElMerFold processes ~41 million proteins at 2,378 structures/s, achieving a 16.3× improvement over prior approaches and reaching 969 PFLOP/s FP32 inference performance.
  - 72.9% parallel efficiency

- 最適化手法
  - Triton + DaCe

- **Triton** provides a portable, high-level (Python) approach for writing efficient GPU kernels
  - FlashAttention-style kernels adapted to EvoFormer
- **DaCe** allows lower-level control of memory operations to improve performance

Pointwise attention kernel (#residues = 256)

	Time	Mem. read BW	Mem. write BW
Original	2.30 ms	—	—
Triton	1.12 ms	260.79 GB/s	14.49 GB/s
DaCe	167.76 µs	3,630.00 GB/s	178.84 GB/s


AMD