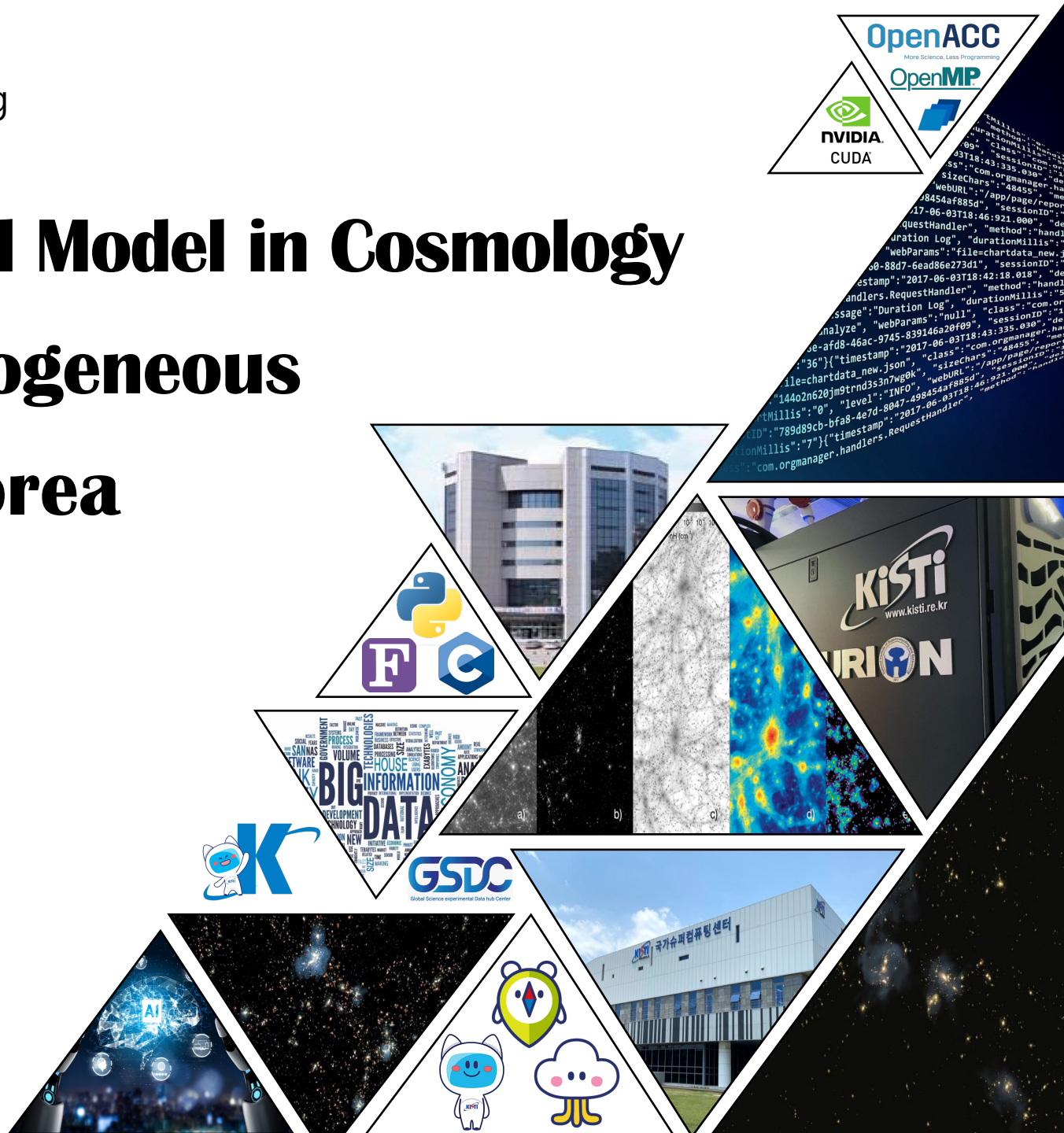
The 10th East-Asia Numerical Astrophysics Meeting

A State-of-the-Art Numerical Model in Cosmology for Next-Generation Heterogeneous Computing Systems in Korea

Yonghwi Kim (金龍輝)

Center for Advanced Scientific Computing,
Division of National Supercomputing R&D

KiSTi
www.kisti.re.kr
Korea Institute of
Science and Technology Information



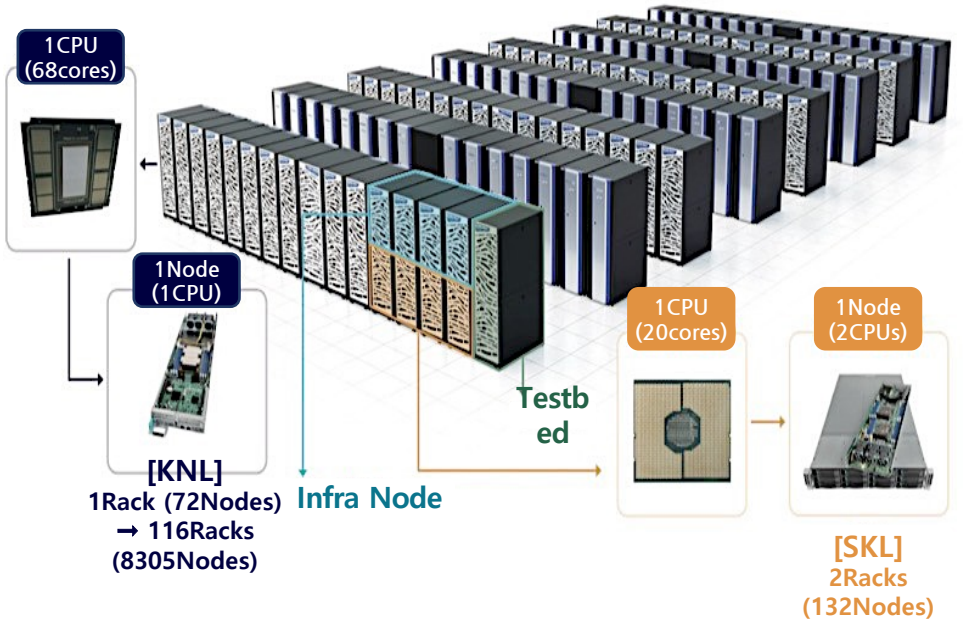
KiSTi
Korea Institute of
Science and Technology Information
www.kisti.re.kr

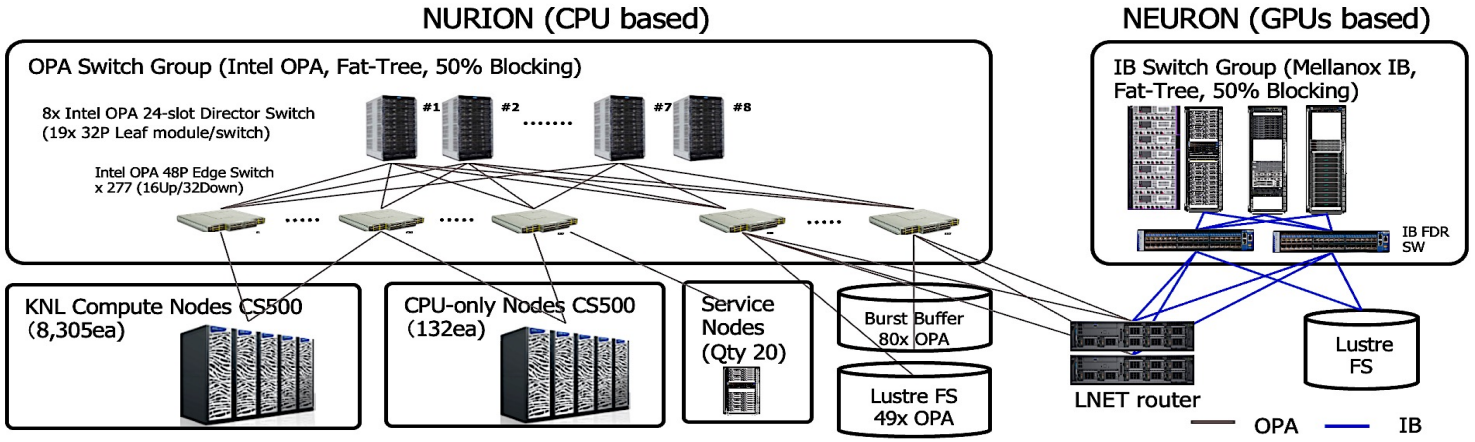
Contents

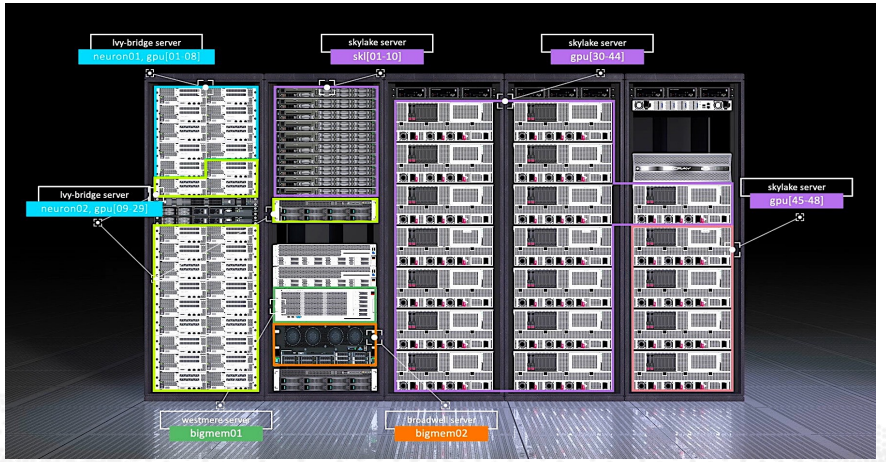
Leading R&D Transformation
with AI and HPC

National Flagship Supercomputer – KISTI-5

- **"NURION" CPU System** (Top500 #11 in 2018)
  - **25.7 Pflops** / 33.88 PB storage resource
  - **8,305** KNL Compute Nodes + **132** Xeon Skylake CPU Nodes
- **"NEURON" CPU-GPU Integrated System**
  - Focused on AI/DL computing and GPU-accelerated applications (official service in July 2019)
  - 140 V100s + 120 A100s + 40 H200s + 2 GH200s (BMT servers)
  - Total **6.4PFlops** & to be expanded to **+8 PFlops (FP64)** in 2025







Next Generation Supercomputer – KISTI-6

Leading R&D Transformation with AI and HPC

Courtesy by HPE

Total Peak
+600 PF (FP64)

** 33.6EF(FP8), 16.8EF(FP16)*

Computing Node

HPE Cray EX4000 System

GPU R_{peak} 588.28 PF (**2,084 nodes x 4 GH200s**)

CPU R_{peak} 15.7 PF (802 nodes, AMD Turin)

SlingShot-400 High Performance Interconnect

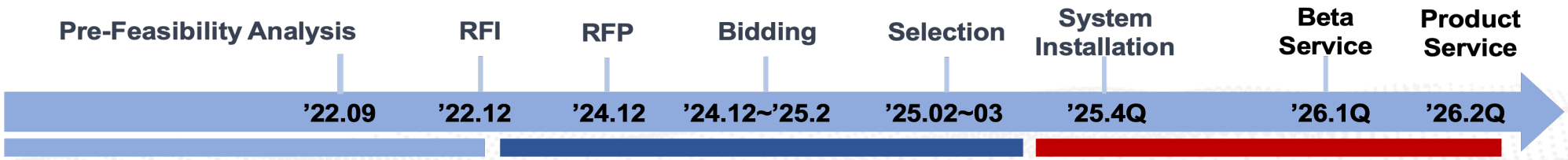
Storage System

HPE Cray Supercomputing Storage System

Over 200 PB Usable Capacity

Flash >9.0 TB/s (r) and >6.5 TB/s (w)

☑ Expected Timeline of KISTI-6 Supercomputer



Rapid Change of GPU Architecture

- **Nvidia** GPU

2022	2023	2024	2025	2026	2027
Hopper (H100 GPU) - FP8 Transformer Engine - HBM3	Hopper+ (H200 GPU) - HBM3e Grace CPU Launch - 72-core Arm CPU - CPU-GPU coherence GH Superchip	Blackwell (B100) - 8x HBM3e - 20 PFLOPS FP4 GB Superchip	Blackwell Ultra(B100-U) - Enhanced performance - Likely Grace refresh	Rubin (R100) - 8x HBM4 - 50 PFLOPS FP4 Vera Arm CPU	Rubin Ultra (R100-U) - 12x HBM4 - 100 PFLOPS FP4 Kyber Rack + Vera

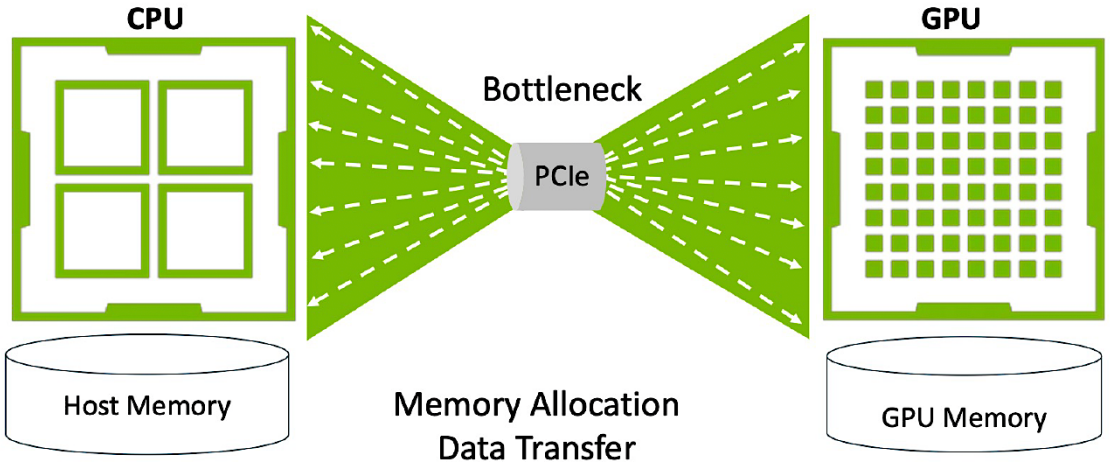
- **AMD** GPU

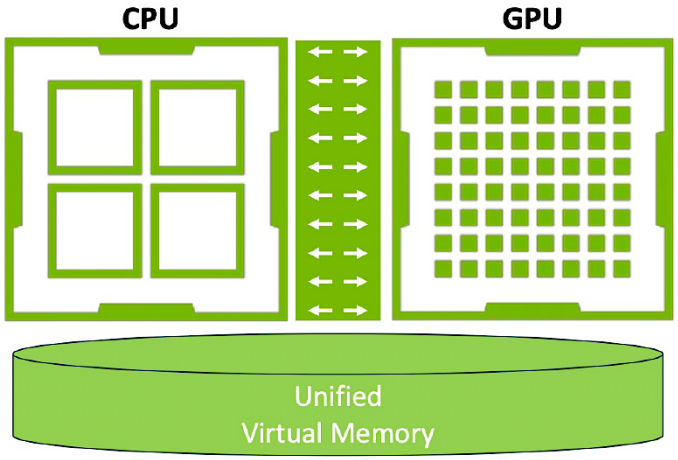
2022	2023	2024	2025	2026
MI200 (CDNA 2) MI250X: HPC/AI - Based on 6nm	MI300X (GPU) - CDNA 3 - HBM3	MI325X - HBM3e MI300A: CPU+GPU (Zen 4 + CDNA 3) - 128GB HBM3	MI350 (Expected) - CDNA 4 Architecture - Advanced AI / HPC capabilities	MI400 APU (Expected) - Zen 5 + CDNA 4 - Next-gen HBM4 - Higher efficiency & coherence

* APU (Accelerated Processing Unit) integrates a CPU and GPU onto a single die.

Memory Bottleneck

- Accelerators (GPUs) and CPUs are separate devices, each with their own dedicated memory.
- Accelerators (GPUs) and CPUs are integrated into a unified System-on-Chip (SoC).





❑ PCIe bottlenecks CPU-GPU communication
❑ GPU can't access CPU memory directly
❑ Complex Programmability
❑ Low CPU Performance per Watt

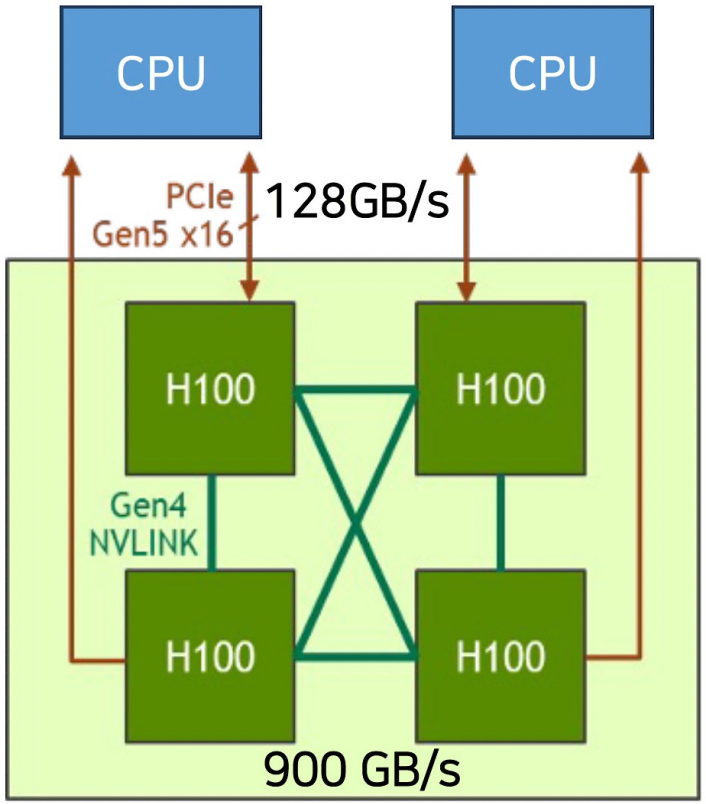
Legacy PCIe Architecture

We need a new chip architecture:

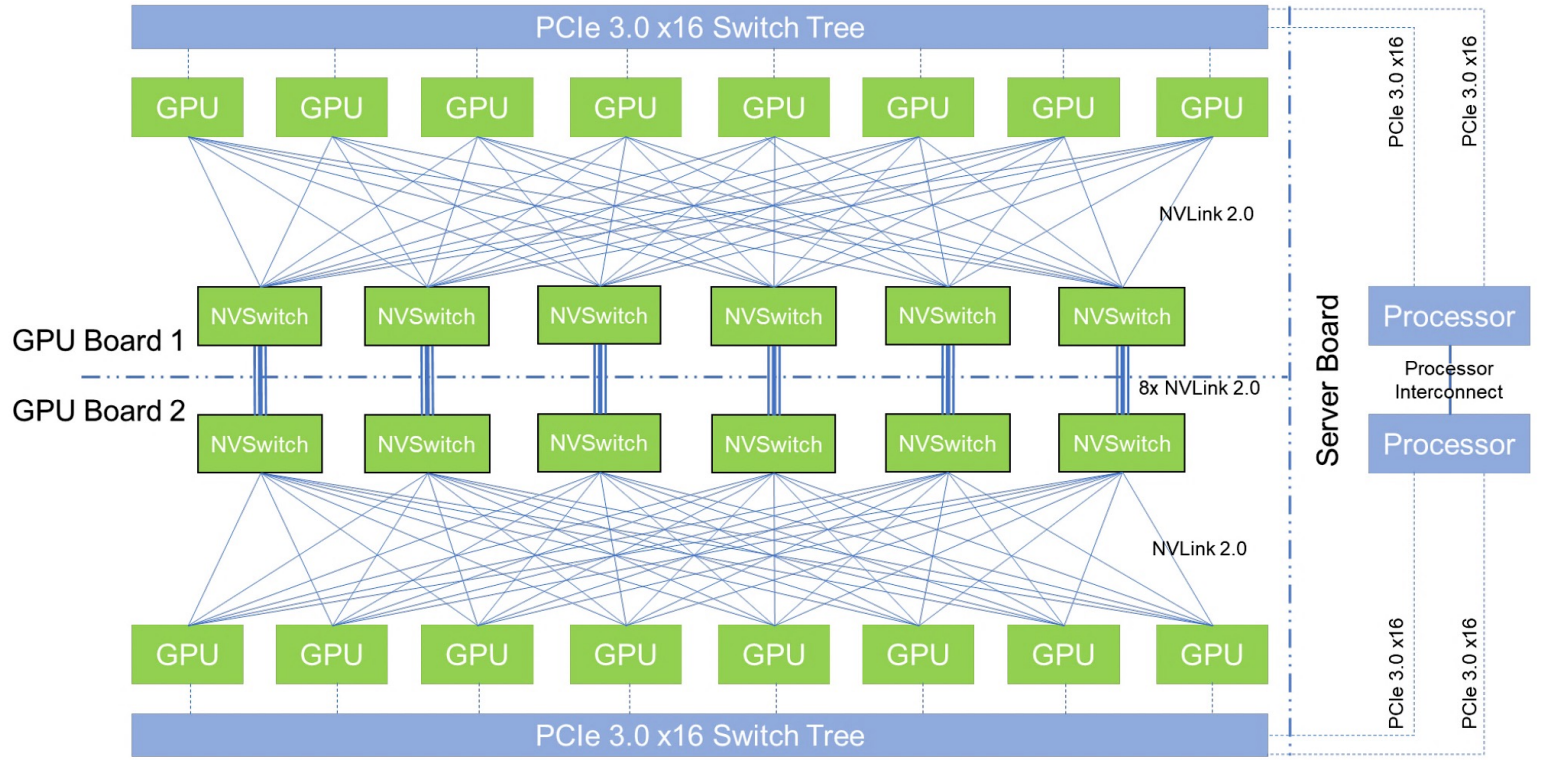
✓ Accelerate CPU-GPU data transfers
✓ Coherent CPU and GPU cache memory
✓ Simplify Programmability
✓ Deliver more performance per Watt

Network Bottleneck

- Traditional CPU-GPU data exchange bottlenecks: Affecting network throughput as well



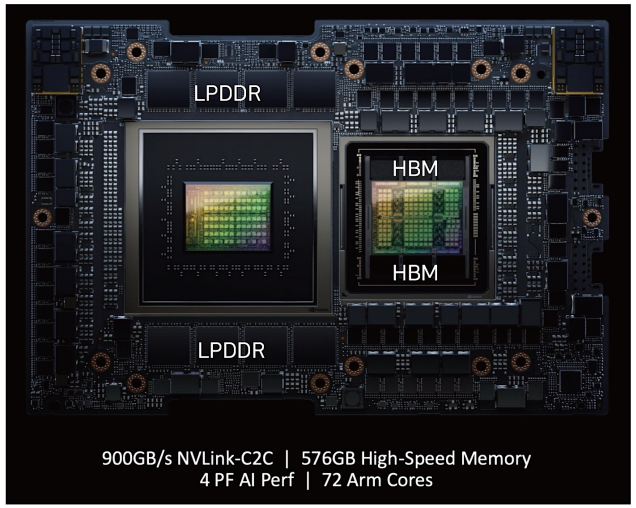
https://developer.nvidia.com/



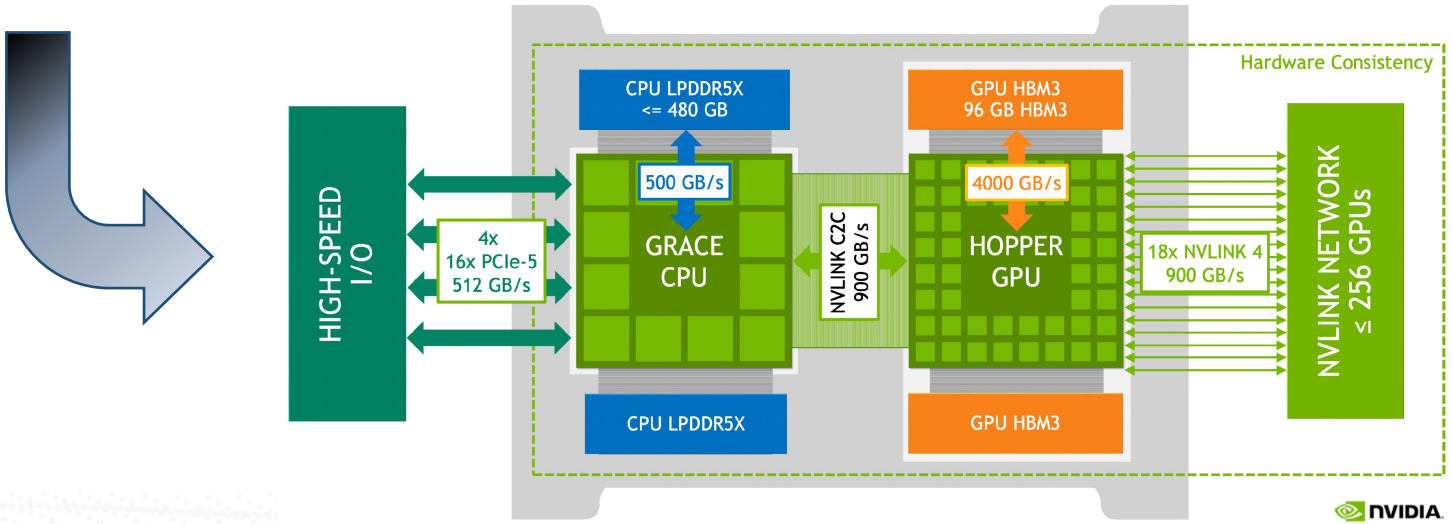
https://techjunction.co/encyclopedia/nvswitch/

GPU-CPU Integrated Architecture

- Grace-Hopper Superchip (**NVIDIA**)

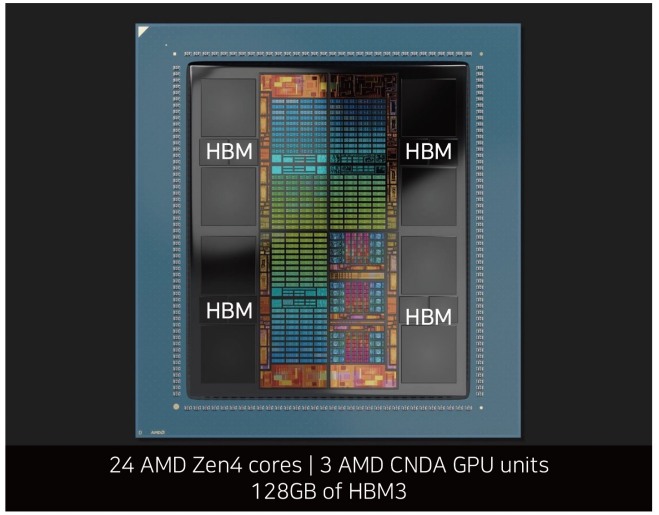


NVIDIA



NVIDIA

- MI300A APU (**AMD**)

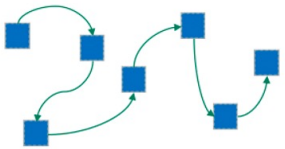


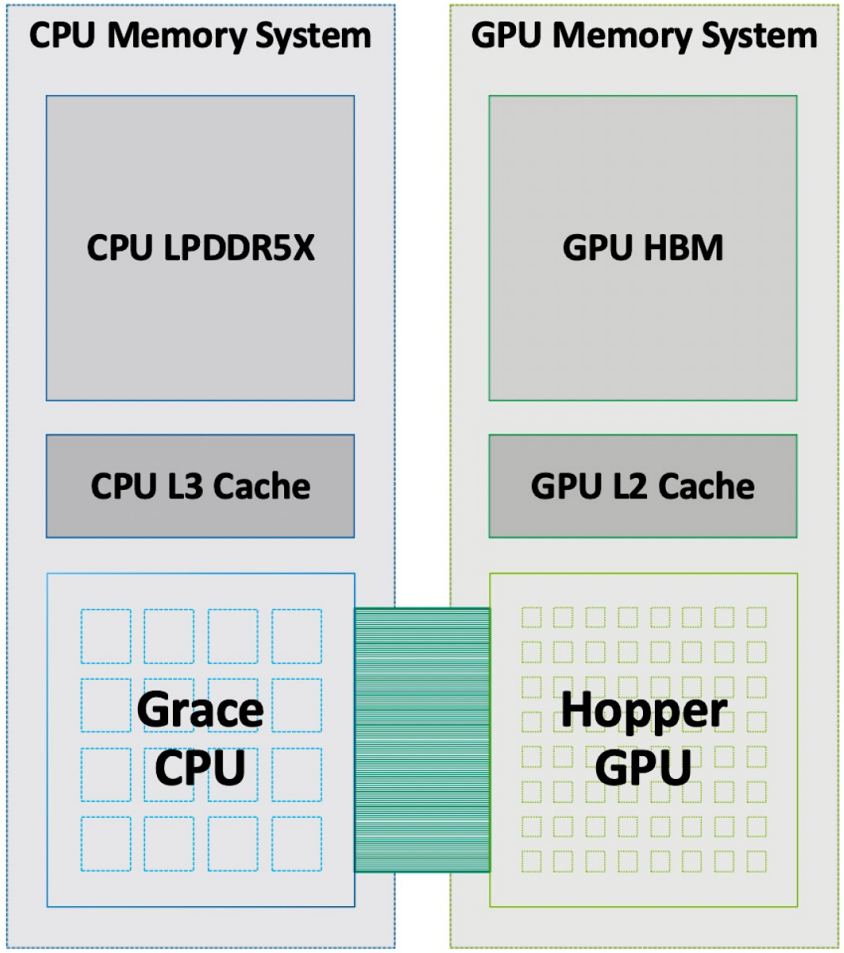
AMD

Two Memory Systems

CPU memory system is optimized for **low latency** and **deep cache hierarchy**

Run **latency-sensitive** code on the CPU, e.g., a linked list



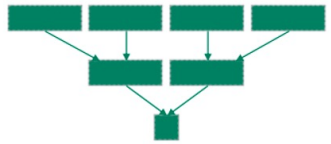


NVLink-C2C

CPU and GPU each have **full coherent access** to memory

GPU memory system is optimized for **high throughput** and **high bandwidth cache**

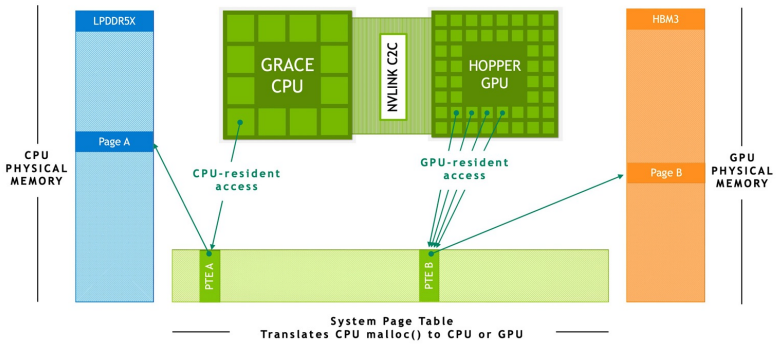
Data- and math-intensive code on the GPU, e.g., vector reduction



GH Memory Model

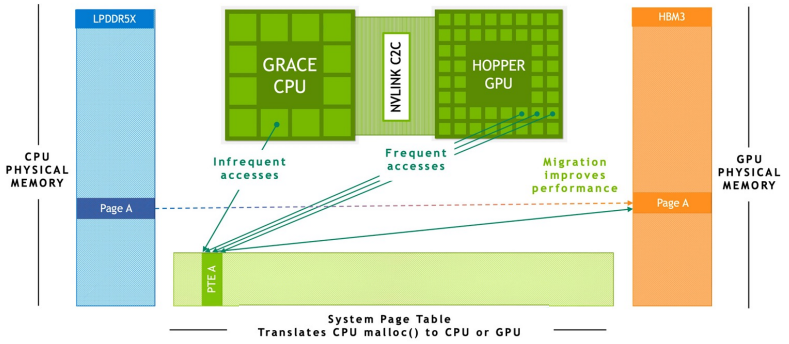
➢ Model (I)

: Local memory pool accesses at link speeds



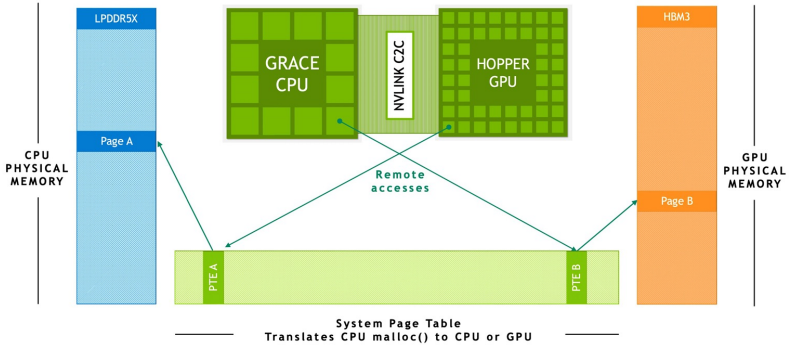
➢ Model (II)

: Simplified Unified Memory via Address Translation Service (ATS) & automatic migrations



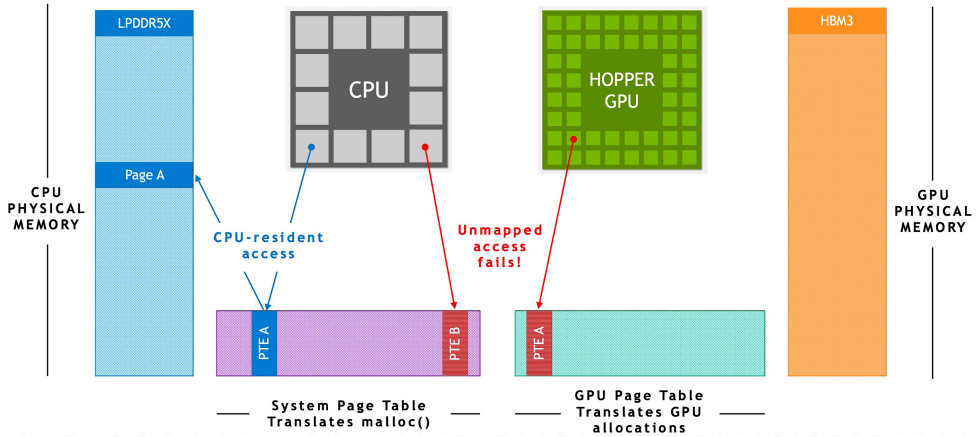
➢ Model (III)

: C2C enabling remote accesses



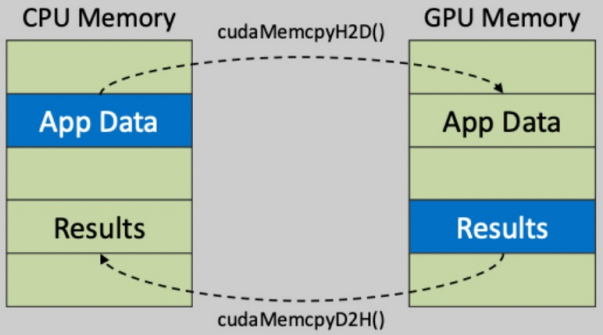
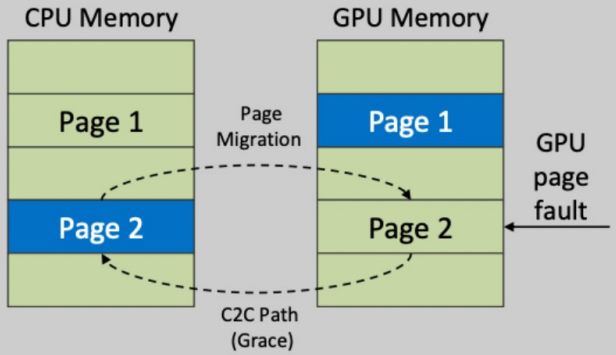
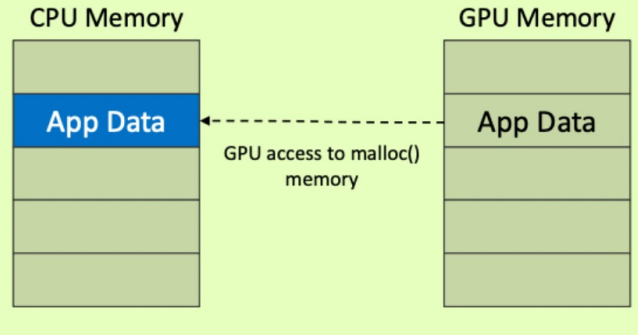
※ Previous Memory Model

: CPU and GPU have separate memories and page tables.



Grace-Hopper Programming Model

- Full CUDA support with additional Grace memory extensions

	Explicit Copy	Managed Memory	System Allocated
	Application explicitly moves data between CPU & GPU as needed	CPU and GPU can access memory on-demand and data migrated locally for higher BW access	GPU can access memory allocated from `malloc()`, `mmap()`, etc.
			
HGX	~60 GB/s PCIe Gen5 transfers (H2D/D2H)	Requires migration to GPU	Access possible with explicit call to `cudaHostRegister()` at PCIe speeds Requires HMM patch in Linux Kernel
G+H	7x faster transfers, up to 450 GB/s (NVLink C2C)	Migrations not required and faster migrations when they happen at NVLink C2C speed	`cudaHostRegister()` not needed; access at NVLink C2C speeds

Advanced Utilization of HPC Program

- Solving scientific problems & Creating innovative technologies
  - Through support for large-scale group research
  - Based on ultra-large data & simulation
  - Using high-performance computing

DARWIN project

: $3.5 million (2022~2027, 4.5 years)

- Major strategic fields
  - Material science
  - Life science
  - ICT
  - Meteorology
  - Self-driving
  - **Astronomy**
  - Nuclear fusion
  - Manufacturing technology
  - Natural disaster
  - National defense

한국천문연구원 Korea Astronomy & Space Science Institute
KISTI 한국과학기술정보연구원 Korea Institute of Science and Technology Information www.kisti.re.kr
KIAS 고등과학원 KOREA INSTITUTE FOR ADVANCED STUDY
서울대학교 SEOUL NATIONAL UNIVERSITY
연세대학교 YONSEI UNIVERSITY
경희대학교 KYUNG HEE UNIVERSITY
충남대학교 CHUNGNAM NATIONAL UNIVERSITY

- What is '**DARWIN**'?
  - **DA**zzling **R**ealization of d**W**arf galaxies **I**n the **N**ext generation of cosmological HD simulations
  - A high-resolution and large volume simulation model to understand the next generation of dwarf galaxy observational research.

3 Steps of DARWIN Simulations

Horizon Run 5
2019, 2021

Nurion (KISTI 5th)
- ❑ # of cores~ 1.7×10^5
- ❑ CPU time ~ 3.8×10^8
- ❑ RAMSES-**OMP**
- ➢ Δx = 1 kpc
- ➢ $L^3 = 1.0\times0.1\times0.1$ Gpc^3

DARWIN-1
2024 on going

Nurion (KISTI 5th)
- ❑ # of cores ~ 10^5
- ❑ CPU time ~ 2.2×10^8
- ❑ RAMSES-**DARWIN-β**
- ➢ Δx = 0.5 kpc
- ➢ $L^3 = 65^3$ Mpc^3

DARWIN-2
2024 ~ 2025

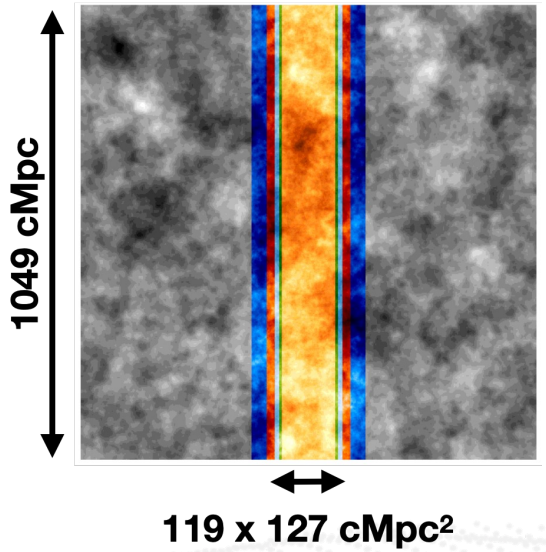
KISTI 5th/6th (x20 R_{peak})
- ❑ RAMSES-**DARWIN**
- ➢ Δx = 62.5~125 pc
- ➢ $L^3 = 30^3$ Mpc^3
- ➢ Additional physics

~500×t_{calc}(DARWIN-1)

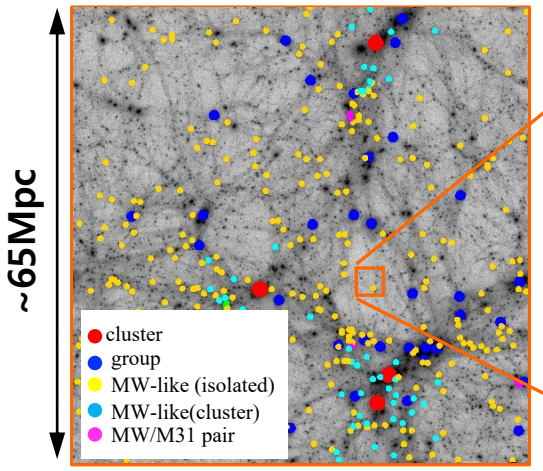
DARWIN-3
2025 ~ 2027

Upcoming **KISTI 6th**
- ❑ RAMSES-**DARWIN**
- ➢ Δx = 4~8 pc
- ➢ $L^3 = (3\sim5)^3$ Mpc^3
- ➢ Additional sub-physics

~500×t_{calc}(DARWIN-2)

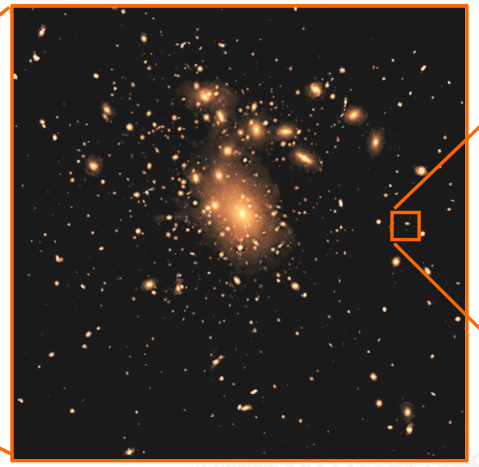


Δx=500pc



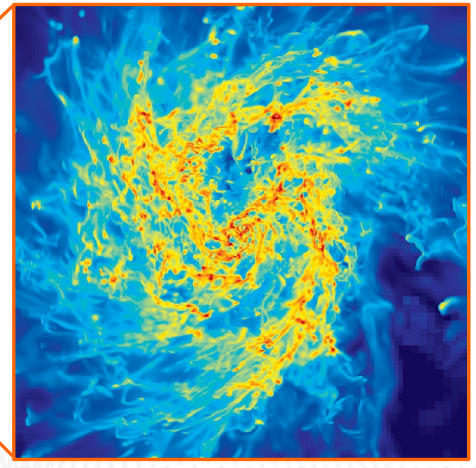
DARWIN-1

Δx=125pc, 62.25pc



DARWIN-2

Δx=4-8 pc



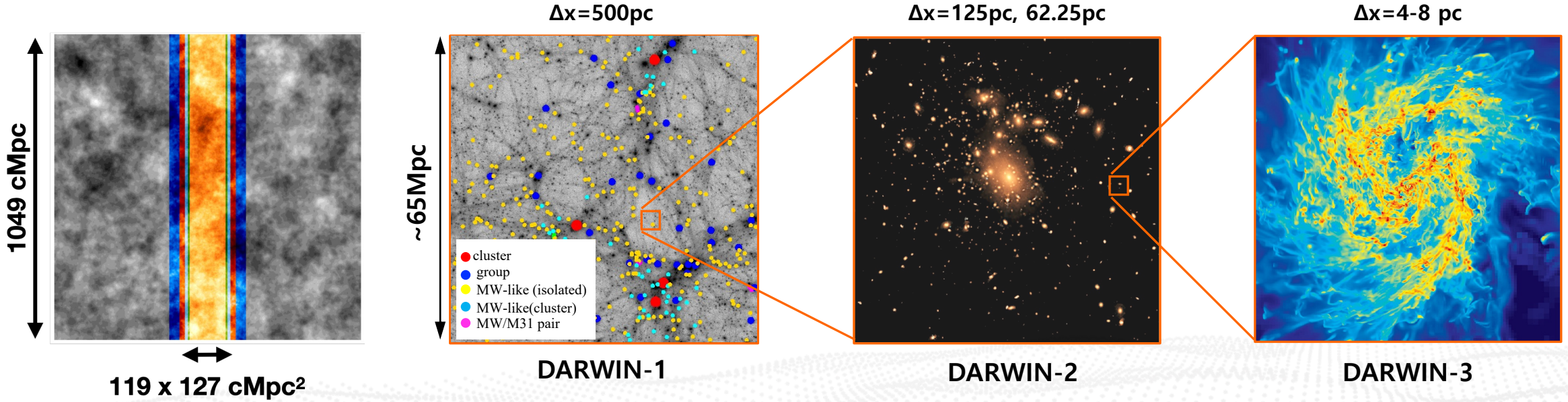
DARWIN-3

3 Steps of DARWIN Simulations

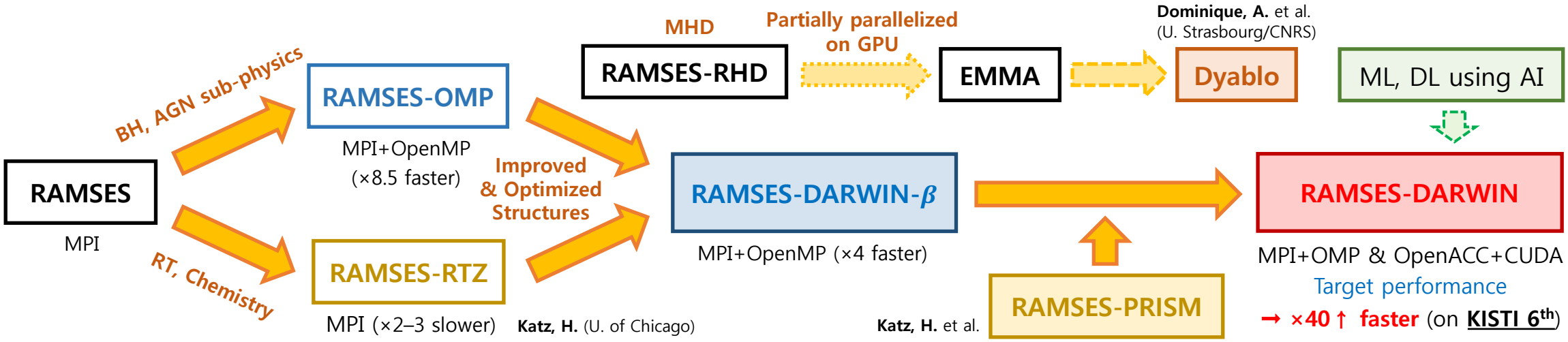
- Significance of **DARWIN** project
  - Theoretical model to interpret next-generation observations
  - High resolution (detailed features) + Precise baryonic physics + Cosmological volume
    : challenging task in galaxy formation simulation ➔ The next-generation of the cosmological HD simulations

⇨ Large volume and the world's highest level of resolution: 3-step multi-resolution with AI algorithms

⇨ Developing code for the CPU-GPU **heterogeneous HPC** system: **Exa-scale** Supercomputer



Code Development History



- Code Structure
  - `patch/RTZ & PRISM` ⇦ Original RAMSES + SF, stellar winds, SNe feedbacks (Ia, II) + RT(Z)
  - `patch/DARWIN-β` ⇦ + Sink (Stars & BHs) + AGN + OpenMP
  - `patch/DARWIN v1.0` ⇦ + **new** SF/SNe + Chemistry (>12 elements) + partially GPU optimized
  - `patch/DARWIN v2.0` ⇦ + fully supported **OpenACC** / **CUDA**

Benchmark Tests

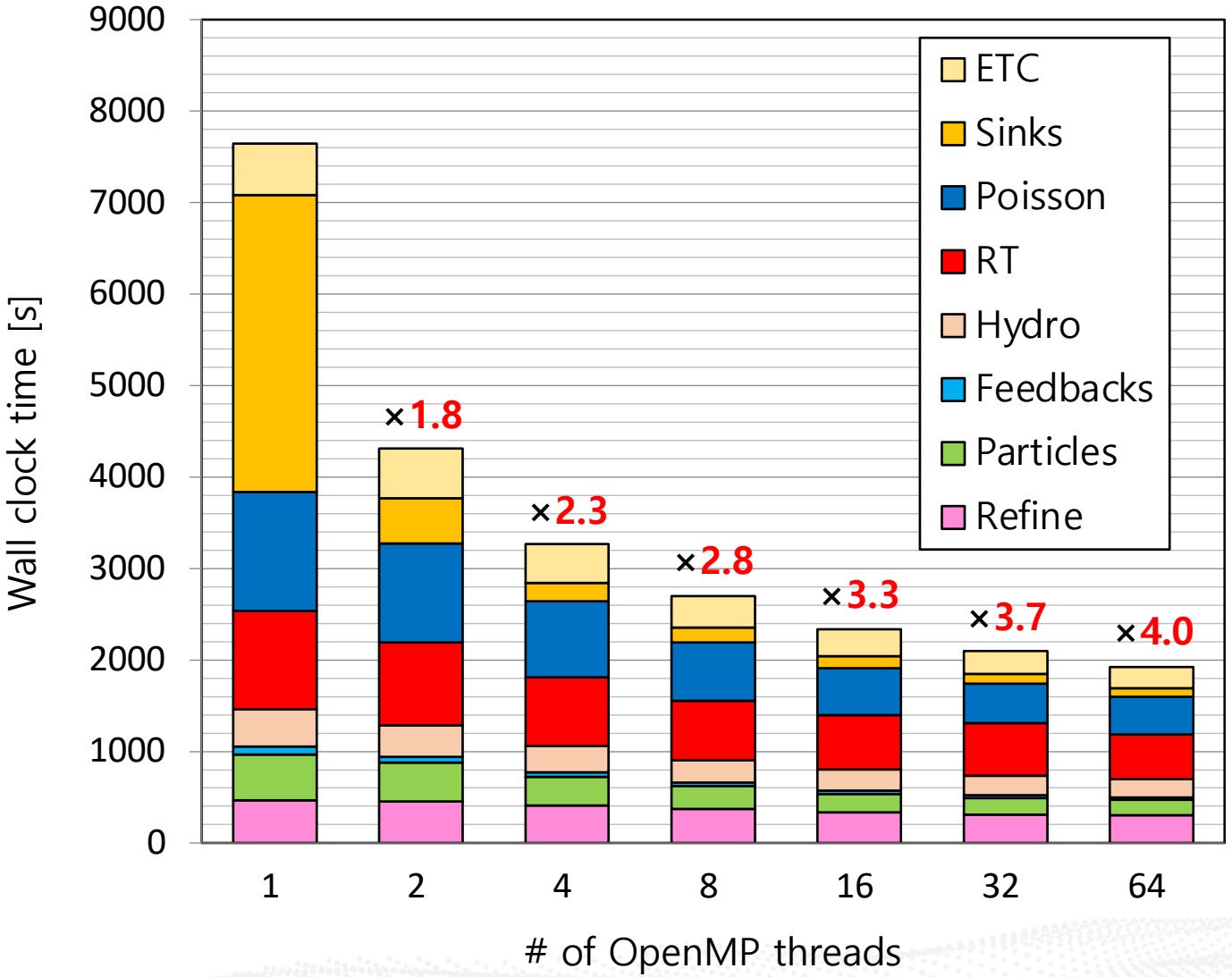
- Test systems
  - KISTI **NURION** (**25.7 Pflops**)
    - #11 in the world in 2018, #75 as of 2024
    - 8305 many-core CPU nodes (KNL)
    + 132 general purpose CPU nodes (Skylake)
    - 100Gbps Omni-path Architecture
    - 800GB Parallel Filesystem
    - 20PB Lustre filesystem
  - KISTI **NEURON**
    - GPU-based sub-system
    - > 250 V100 & **A100** GPUs
    - 10 CPU-only Skylake nodes
    + **H100** & **GH200** test servers (6 nodes)

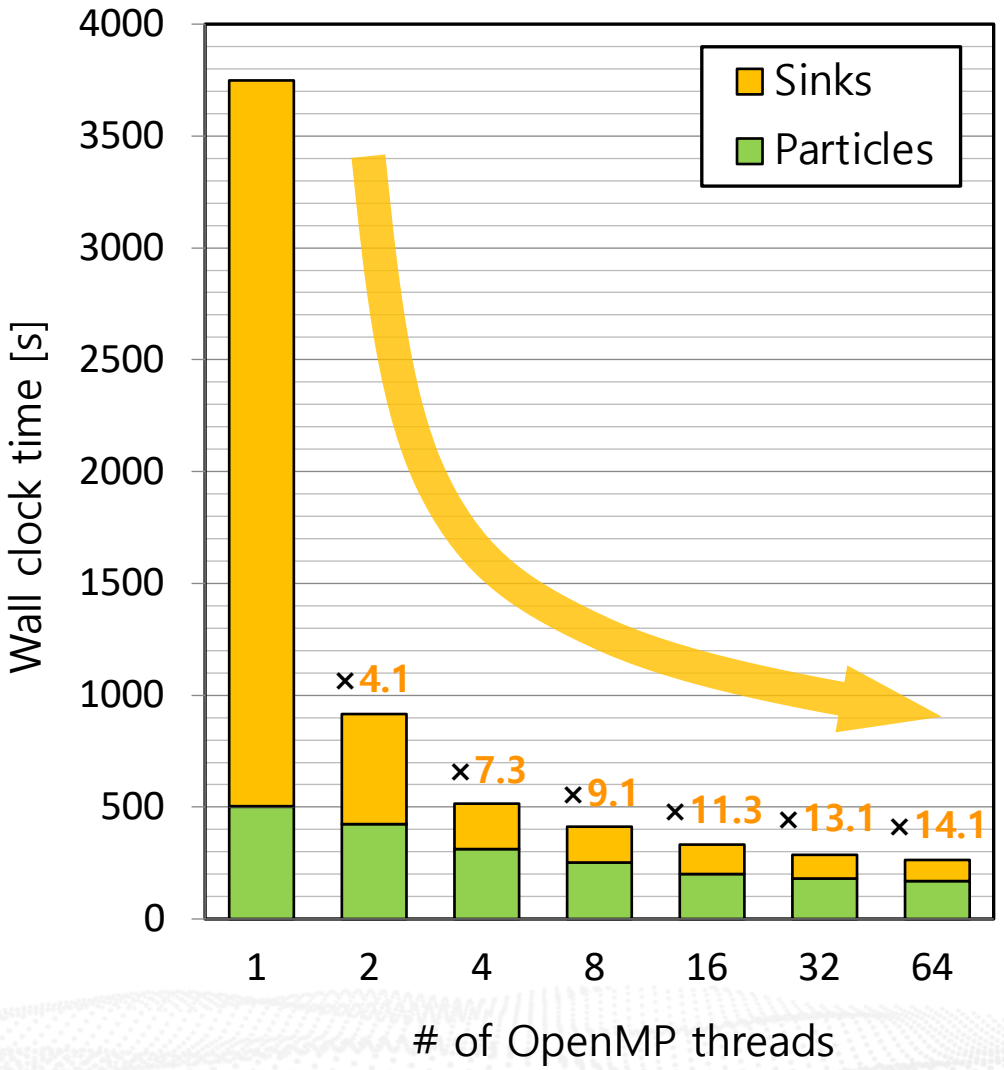
- BMT case in astrophysical environments
  - A galaxy cluster region
  - IC – 32^3 Mpc^3 with 512^3 grids
  - AMR levels = 9 – 16
  - # of MPI tasks = 1024

+ **EuroHPC/FZJ(JSC)** : **JEDI** (GH200)

Site:	EuroHPC/FZJ		
Nodes	48		
GH200/node	4 ×	NVIDIA GH200 Grace-Hopper Superchip	
	Processor/GH 200:	1 ×	NVIDIA Grace (Arm Neoverse-V2), 72 cores, 120 GB LPDDR5X, 384 GB/s
	Accerator/GH 200:	1 ×	NVIDIA HOPPER H100, 96 GB HBM3, 4 TB/S
Interconnect:	Quad-Rail NVIDIA InfiniBand NDR200		
	CPU-GPU	900 GB/S	
	GPU-GPU(intra node)	300 GB/S	
	CPU-CPU(inter node)	200 GB/S	

Performance – OMP Scalability

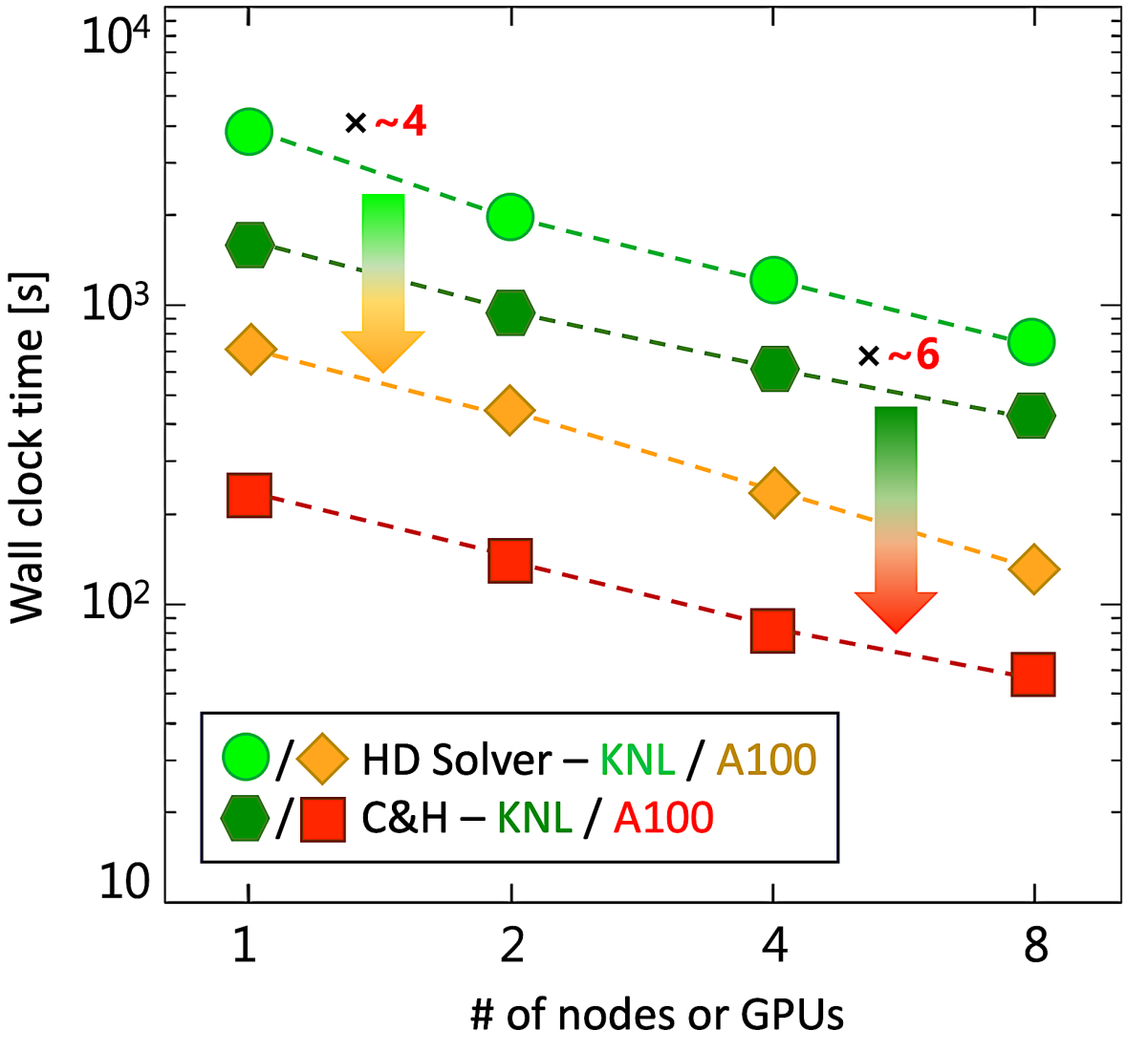




GPU Test I – RAMSES-DARWIN

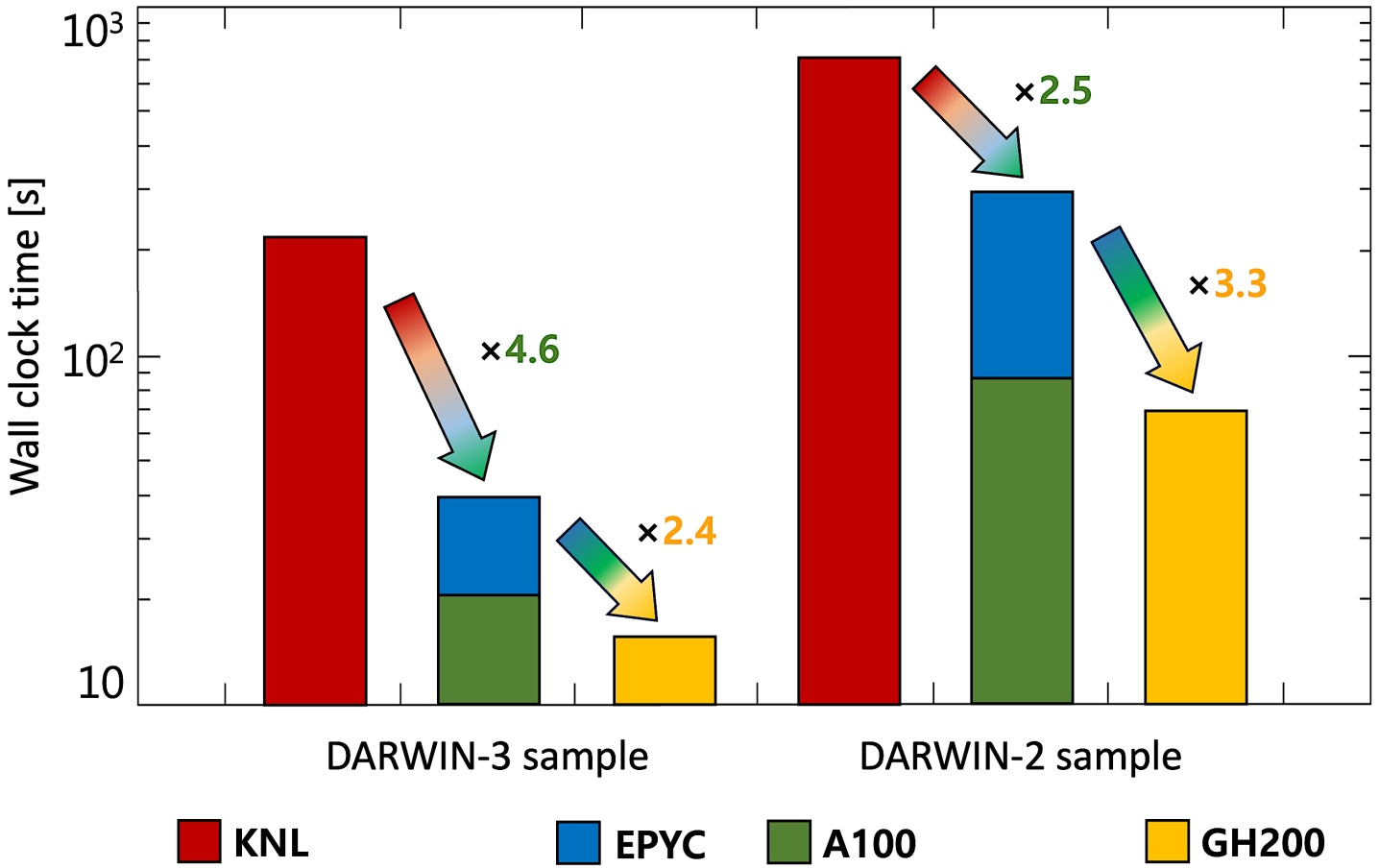
- Physics parts : OpenMP ⇨ OpenACC
  - HD solver (Godunov & interpolation)
  - Cooling & Heating

- Test systems
  - CPU server : KISTI Nurion KNL
    - Intel Xeon Phi 7250 / 1 node / 68 cores
    - Total 25.3 Pflops / 3.0 Tflops per node
  - GPU server : KISTI Neuron A100
    - AMD EPYC 2.8G 32×2 cores (5.6 Tflops)
    - GPU: 4 × NVIDIA A100 (4 × 9.7 Tflops FP64)

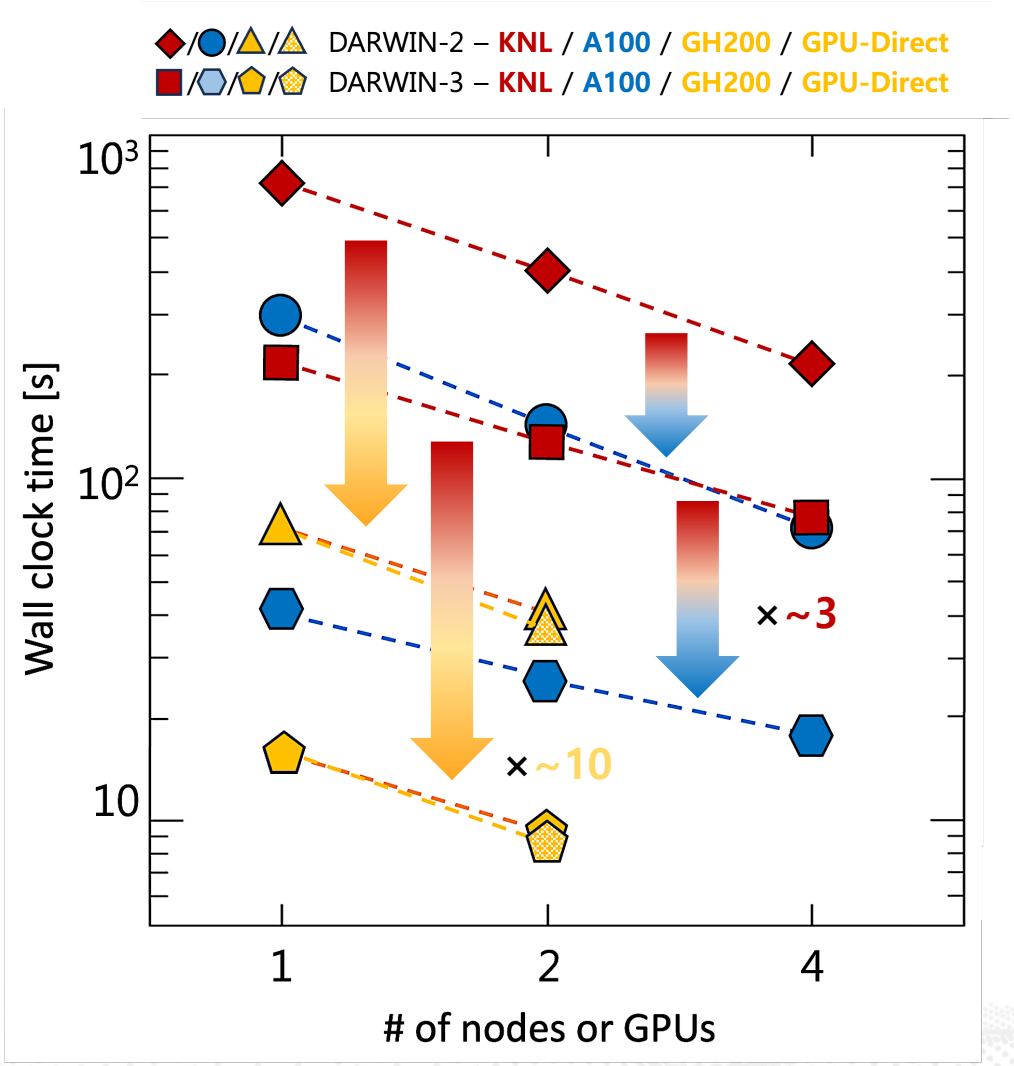


GPU Test II – Performance

- Test environment: KNL, A100 (with AMD EPYC 32×2 cores), GH200
  - `nvidia_hpc_sdk/24.1; cuda/12.3; gcc/11.4.1`

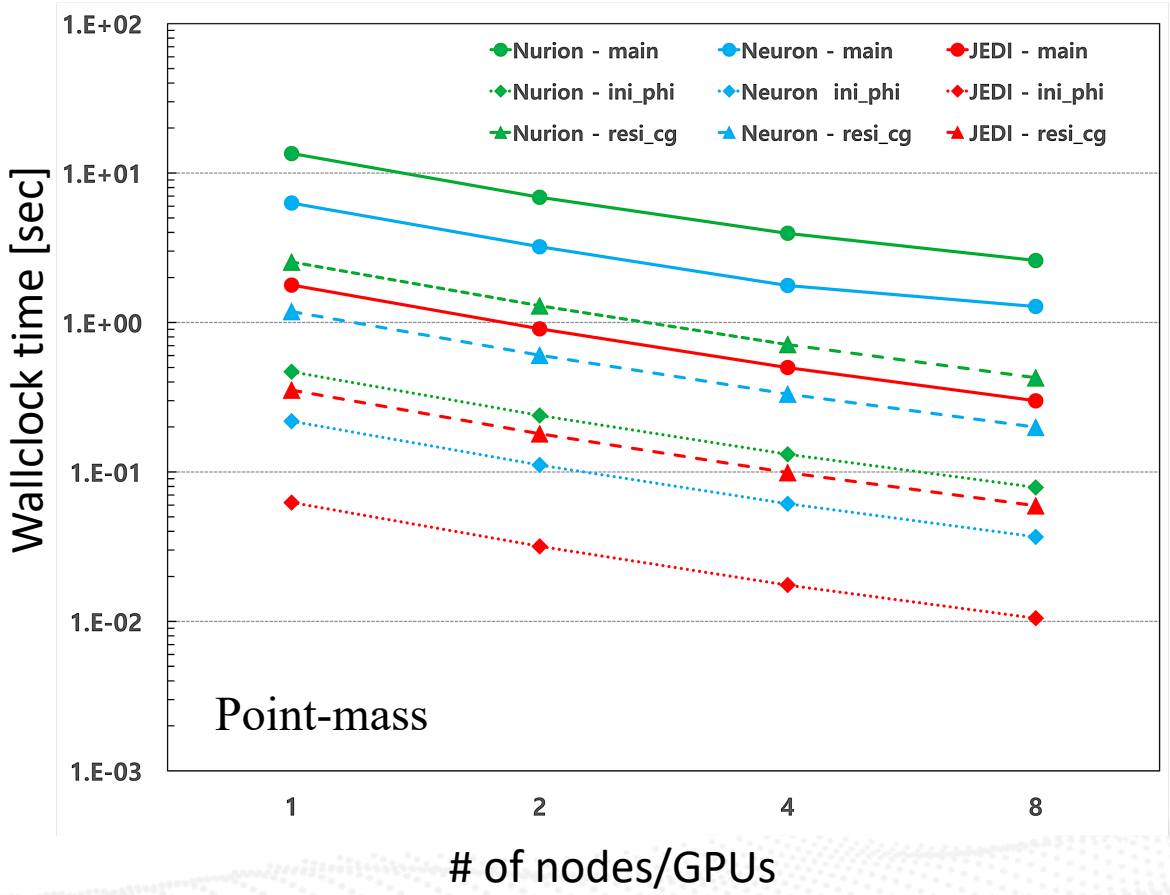


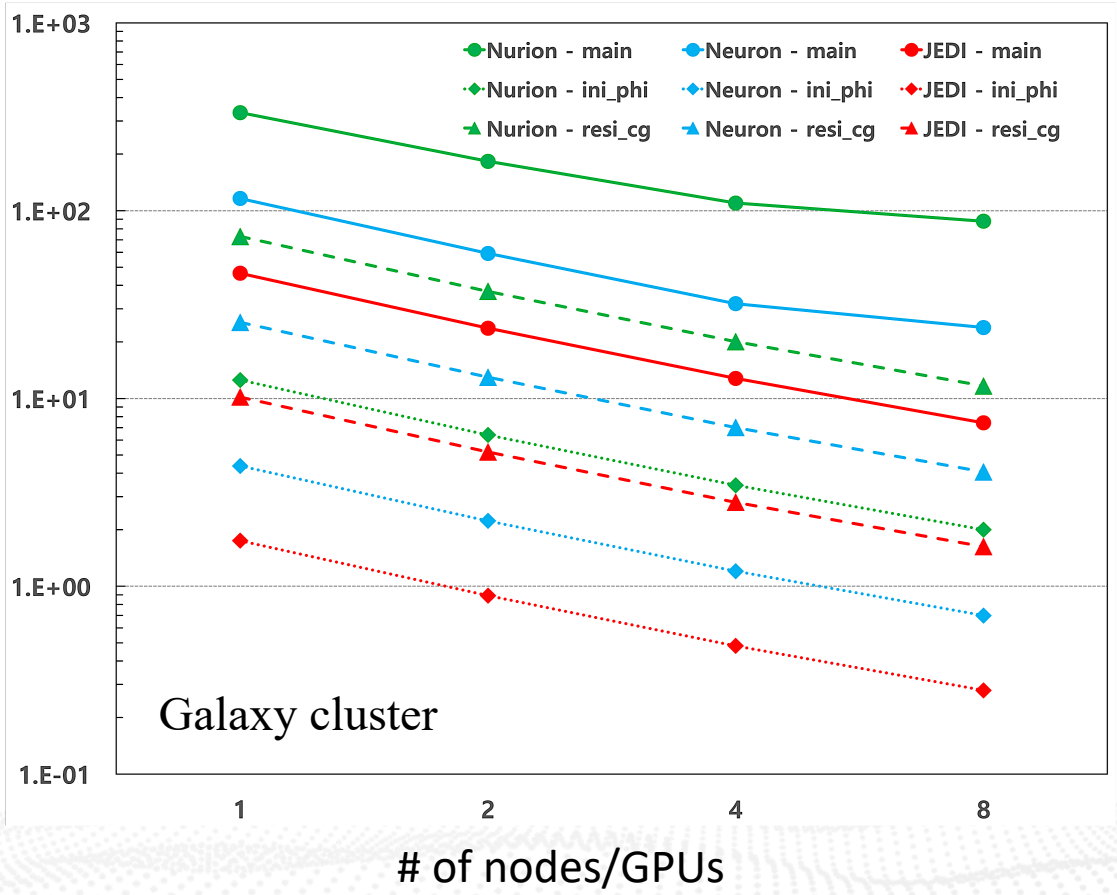
CUDA-Aware-MPI



GPU Test III – Multi GPU Performance

- BMT cases in astrophysical environments
  - Single galaxy (point mass) at the center of a cube with 128^3 grids, AMR level varies from 2^7 to 2^{12}
  - A galaxy cluster in 30^3Mpc cube with 512^3 grids, AMR level varies from 2^9 to 2^{16}





Summary

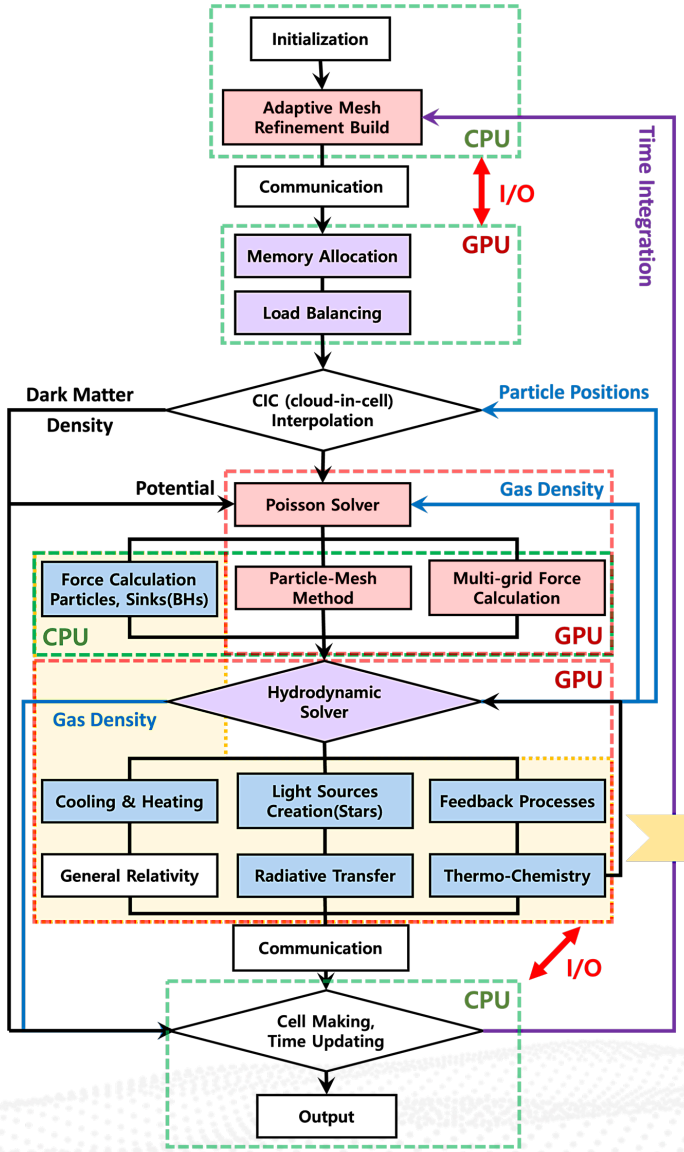
- In 2025,
  - Dynamics parts
    - Poisson solver: cg, mg ⇨ CUDA

    ※ I/O & Load balancing ⇨ OpenACC
  - Physics parts
    - Hydro solver ⇨ OpenACC
- Until mid of 2026,
  - Dynamics parts
    - Mesh structures (AMR build)

      ⇨ CUDA (partially OpenACC)
    - PM ⇨ CUDA (partially OpenACC)
    - I/O & Load balancing ⇨ CUDA
  - Physics parts
    - N-body / Tree / RT, Chemistry & etc.

      ⇨ OpenACC (partially CUDA)

    ※ All hydrodynamic parts ⇨ CUDA



Key Parts in Sub-physics Modules

DARWIN-2

- Force calculation of **Sinks** as **BHs**
- Light Sources Creation (Pop-III, Pop-II old stars)
  - Metal-poor, seeds of BHs in early universe
- **AGN** activities with Super-massive BHs growth
  - Mechanical Jet and thermal feedbacks

Force Calculation of Sinks as **BHs** | Light Sources Creation - **Old Stars** | **AGN Feedback** Processes | **Cooling & Heating**

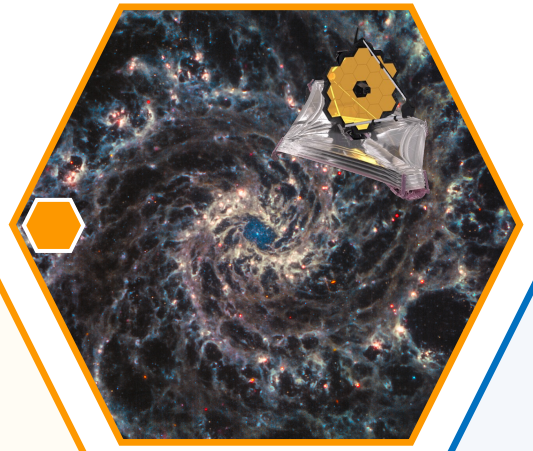
DARWIN-3

- Force calculation of **Sinks** as newly born **STARS**
- Light Sources Creation (Pop-II, Pop-I young stars)
  - Metal-rich (Cooling & Heating)
  - ➔ **Thermo-Chemistry**
- Supernovae(SNe) feedbacks
  - Mechanical and Thermal feedbacks

Force Calculation of Sinks as **STARs** | Light Sources Creation - **Young Stars** | **Radiative Transfer** | **SNe Feedback** Processes | **Thermo-Chemistry**

CUDA
Hybrid (CUDA+OpenACC)
OpenACC (partially CUDA)

Our Steps Towards to the Future



High Resolution Observations

Large-scale Cosmological Model

➔ Big data

Next-generation of HPCs

Upgraded Analysis Methods

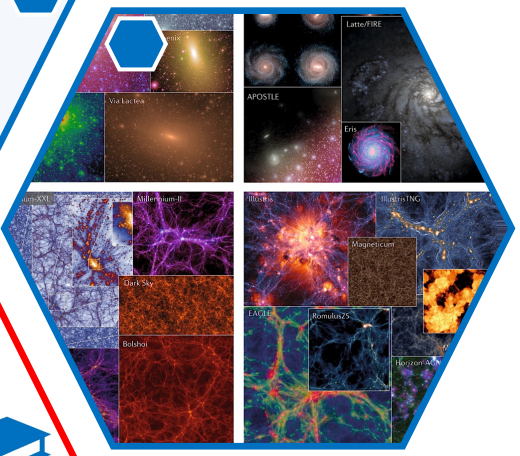
Parallelization

Internal and External Collaborations

Various Scale

Various Methodology

Various Research Projects



Thank you