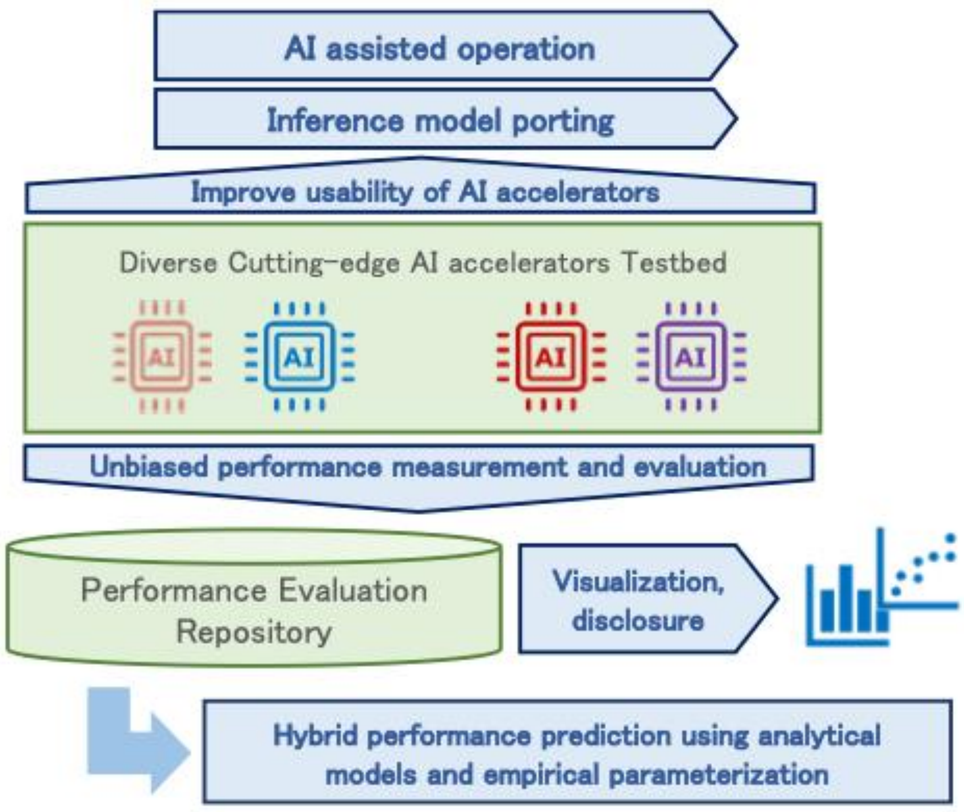
AIST
Create the Future, Collaborate Together

NAPA: Next-generation AI Platform Advancement

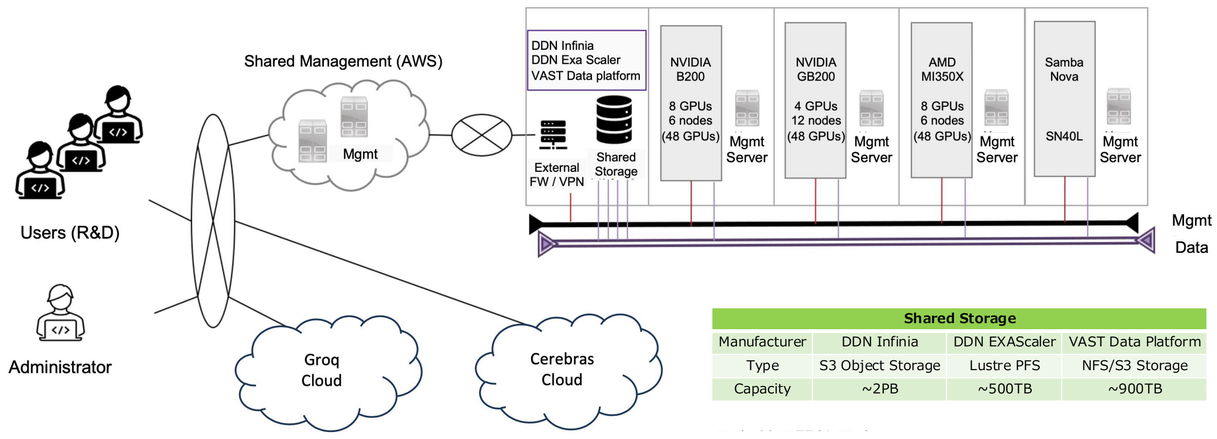
Overview

- Aims to continuously evaluate diverse and cutting-edge AI accelerator chips and developing technologies to realize highly efficient use of AI computing resources with those accelerators, for next generation AI platforms.
- Bridge the gap between AI Research and Industry: Open Innovation Platform to accelerate joint academic-industry R&D for AI
- 2025-2030, $180M+, New Energy and Industrial Technology Development Organization (NEDO)



Work Packages

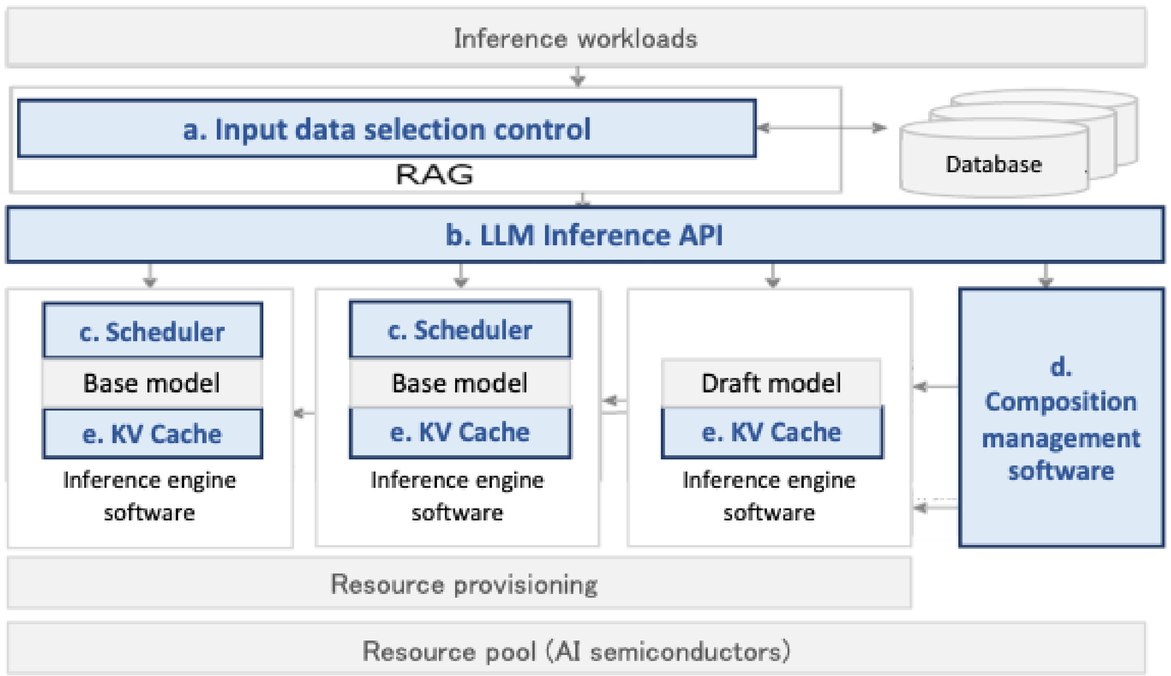
- WP1: Building and evaluating an AI accelerator testbed
  - Introduce multi-vendor-based AI accelerators targeting large-scale training and inference at data centers
  - Neutral and fair evaluation of AI accelerators using benchmarks, real workloads, and testbed operational data
  - Improve portability of AI application across different accelerators and operation of computing resources
- WP2-1: Improving resource utilization based on accurate performance prediction
  - Develop highly-accurate execution performance prediction technology for both training and inference
  - Develop a system to continuously evaluate performance of training/inference processing and creating open performance prediction datasets
- WP2-2: Developing technologies of highly efficient, highly-performant inference system
  - Input data selection control, LLM inference API with speculative decoding, configuration management software, advance scheduling technology, KV cache offload design guidelines



Shared Storage			
Manufacturer	DDN Infinia	DDN EXAScaler	VAST Data Platform
Type	S3 Object Storage	Lustre PFS	NFS/S3 Storage
Capacity	~2PB	~500TB	~900TB

	B200	GB200 NVL4	MI350X
Manufacturer	NVIDIA	NVIDIA	AMD
# GPUs/nodes	8	4	8
# of nodes	6	12	6
Integrator	SuperMicro	SuperMicro	SuperMicro
Cooling Method	Air	Liquid	Air
GPU memory	180GB×8 GPU 1.44TB / node	192GB×4 GPU 0.768TB / node	288GB×8 GPU 2,304TB / node
Memory capacity	2.0TB / node (64GB×32 DIMM)	960GB / node (ECC LPDDR5X Onboard)	3.0TB / node (128GB×24 DIMM)
NVMe Storage	307TB / node (30.7TB×10 NVMe)	30.72TB / node (7.68TB×4 NVMe)	245.6TB / node (30.7TB×8 NVMe)
NIC for RoCE	400G×8	400G×4	400G×8
Storage NIC	200G	200G	200G

	SN40L
RDU maker	Sambanova
RDUs/node	8
# of nodes	2
Integrator	Sambanova
Cooling	Air cooling
RDU memory	SRAM: 520MB x 16 RDU HBM3: 64GB x 16 RDU DDR: 768GB x 16 RDU
Memory	2.0TB / node
NVMe storage	6 x 7.6TB /node
NIC	100Gbps x 2 / node



This presentation is based on results obtained from the project, “Research and Development Project of the Enhanced Infrastructures for Post-5G Information and Communication Systems” (JPNP25013), commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

AIST SC Website