GW-scale Open AIDC Framework White Paper (Draft)

—— An Open, Green, and Intelligent System Architecture for GW (Gigawatt)-Scale AI Data Centers

Version Information:
Version: v0.8.3 (Draft)
Start date: December 9, 2025
Revision date: June 16, 2026
Compiled by: GW-scale Open AIDC Working Group (OCP China Community / OCTC)
Project coordinator:

Founding members of the Working Group
CESA (China Electronics Standardization Association), Baidu, Inspur Information, Kuaishou, China Mobile, VNET, AVIC Optoelectronics, Guangdong Connector Association, Luxshare Tech, EVE Energy, Anlan Wanjin ……

Newly joined members

Authors and contributors
……

Technical review: OCP, OCTC, ……

Table of Contents

[In Word, right-click and select “Update Field” to generate the table of contents]

Foreword

Artificial Intelligence (AI) technology is evolving at an unprecedented pace. New application paradigms such as large models, generative AI (AIGC), and agentic AI continue to push compute demand toward ever larger scale, higher density, and higher energy efficiency. Worldwide, computing infrastructure is accelerating from the traditional data center model centered on servers and clusters toward a new generation of AI Data Centers (AIDC) characterized by system-level co-design.

Against this backdrop, the Open Compute Project (OCP) community has put forward

the “Open Systems for AI” initiative, emphasizing open standards, a systems perspective, and cross-layer co-design to drive an overall leap in the performance, energy efficiency, scalability, and sustainability of AI infrastructure. This initiative provides an important methodological foundation and open collaboration framework for the development of AI infrastructure worldwide.

At the same time, China is at a critical stage of scaling up its compute infrastructure. On one hand, new demands—large-model training and inference, industrial intelligence upgrades, and agent applications—are driving the continued construction of ultra-large-scale AI data centers. On the other hand, China has formed unique advantages in renewable energy supply, HVDC transmission, cross-regional compute deployment, manufacturing supply chains, and engineering organization capability, providing fertile ground for exploring this new form of gigawatt (GW)-scale AI data center.

In this context, under the joint guidance of the OCP China Community and the Open Compute Technology Committee (OCTC), the GW Scale Open AIDC Working Group was formally established, with the goal of systematically reviewing and building—oriented to China’s conditions and engineering practice—a set of open, green, intelligent, and replicable reference architectures and practical frameworks for GW-scale AI data centers.

This white paper is not a manual for a specific product or a single technical solution, but a system-level reference framework. Its objectives are to:

distill the common, cross-domain coordination problems of GW-scale AI data centers in energy, compute, and networking;

clarify system-level design principles, the layered architecture, and key interface directions;

provide industry, operators, equipment vendors, and standards bodies with an alignable, co-buildable reference foundation;

promote two-way interconnection and collaborative evolution between China’s practice and the global OCP system.

Through this white paper, the GW Scale Open AIDC Working Group hopes to build industry consensus, lower the cost of system-level innovation, foster the development of an open ecosystem, and provide solid support for the sustainable evolution of next-generation AI infrastructure.

1. Industry Background

The training scale and compute demand of AI models are growing exponentially, driving simultaneous leaps in compute density, network bandwidth, energy efficiency, and green compliance:

- Training cluster scale: from 10,000 cards → 100,000 cards
- High-speed interconnect: from 400G → 800G → Tb/s
- Cooling method: from air cooling → cold plate → immersion
- Data center power: from a few MW → hundreds of MW → toward the GW scale

Traditional data center architectures will be unable to meet the energy-efficiency and scaling needs of future large models, AGI, and the agent economy. Worldwide, a new form of data infrastructure centered on “energy–compute integration” is emerging. The design of AI computing architecture must evolve from the “chip level” and “server level” up to the “rack level” and “system level,” as shown below.

2. China’s Distinctive Conditions

China possesses a set of unique conditions for exploring the ultra-large-scale AI data center model:

(1) Coordinated advancement of energy structure and compute deployment

Building on the “Eastern Data, Western Computing” program, China has

accumulated system-level engineering experience in cross-regional compute deployment, power allocation, and network coordination. The northwest and north of China have concentrated, contiguous wind and solar resources, providing a sustainable, low-carbon energy base for ultra-large-scale AI data centers; meanwhile, the HVDC transmission and grid-dispatch systems create real conditions for cross-domain coordination of compute and energy.

(2) Advantages in manufacturing and the key-component supply chain

China has formed a relatively complete and internationally competitive industrial chain in key areas such as complete servers, liquid cooling systems, connectors, optical modules, power supplies, and power devices. This advantage provides important support for high power density, high-bandwidth interconnect, and rapid large-scale delivery.

(3) Strongly application-demand driven

China's AI applications show pronounced diversity, spanning both internet/cloud-computing scenarios and the intelligence needs of manufacturing, energy, transportation, finance, and other industries. This pattern of multiple scenarios developing in parallel requires compute infrastructure to be highly general-purpose, scalable, and capable of long-term evolution.

(4) Driven by policy and demonstration projects

Governments at all levels regard AI and compute infrastructure as a key strategic direction, accelerating the construction of new AI data centers through policy guidance, demonstration projects, and industrial coordination, providing a real-world testbed for system-level innovation.

Taken together, these factors mean that, in advancing the evolution of AI infrastructure, China needs not only "larger data centers" but a systematic architectural methodology oriented to the GW (gigawatt) scale.

3. A Paradigm Shift

A GW-scale AI data center is not a linear scaling-up of a traditional data center, but a fundamental shift of the system paradigm, whose core characteristics are:

(1) Larger design scale

Architectural design rises from the server and rack levels to the campus and even regional levels.

(2) Cross-domain coordination becomes a precondition

Energy, compute, and network systems are no longer independent subsystems; they must be co-designed and jointly optimized at the planning stage.

(3) System stability and operability take priority

At ultra-large scale, reliability, maintainability, and lifecycle cost are often more decisive than single-point performance.

(4) Greater importance of openness and standardization

As system complexity rises, closed, custom solutions are hard to sustain over the long term; open interfaces and standards coordination become key to reducing systemic risk.

Therefore, building a GW-scale AI data center is in essence a systems-engineering problem, requiring holistic, architecture-level coordination of technology, energy, networking, operations, and ecosystem.

4. Vision and Goals

Based on the above background, the GW Scale Open AIDC Working Group proposes the following overall vision:

(1) Build an open reference architecture for the GW scale

Through systematic layered design, clarify key capability boundaries and interface directions and reduce the cost of cross-stakeholder collaboration. Use an open system to advance the standardization and normalization of China's compute infrastructure, and promote the re-globalization of the supply chain.

(2) Promote a green development model that coordinates energy and compute:

Treat energy efficiency and sustainability as core system-level constraints, achieve coordinated evolution of compute growth and green goals, and advance the sustainable development of compute power.

(3) Support the long-term evolution of diverse AI application scenarios

Meet the differentiated demands placed on compute systems by large models, agents, industrial intelligence, and other application types.

(4) Foster the joint construction of China's practice and the global standards system

On the basis of aligning with OCP's global methodology, summarize China's engineering experience and help form systematic solutions that can be understood and adopted internationally—that is, the ability to both "bring in" and "go global."

Chapter 1 Infrastructure Layer

1.1 Building and Space Planning

The building and space planning of a GW-scale AI data center (AIDC) is the campus-level civil design and coordination oriented to gigawatt-scale total power, high-power-density clusters, and energy–compute integration. It focuses on five dimensions—overall site layout, building unit space, modular adaptation, fire/security, and long-term evolution—providing a baseline for civil/space design of cross-regional, large-scale, standardized compute infrastructure.

1.1.1 Core Planning Principles

The building and space planning of a GW-scale AI data center breaks the design logic of traditional MW-scale data centers. Taking system design as the center and the campus-level AIDC as the scale, it establishes six core principles, all focused on building-space form, layout, scale, and compliance.

1.1.1.1 Policy-Adaptation Principle

Fully align with national new-type data center construction, the "Eastern Data, Western Computing" compute-hub layout, "compute–power coordination" policy requirements, the dual-carbon goals, and local compute-infrastructure policies. Building-space planning must satisfy hard policy requirements such as green-power consumption, HVDC supply access, liquid-cooling deployment, and energy saving. Core metrics such as building grade, fire protection, seismic resistance, and evacuation strictly follow the national standard for Grade-A data centers, ensuring that the GW-scale campus complies with national and regional compute-infrastructure requirements from the planning stage and serves as the spatial carrier for the national compute-network strategy.

1.1.1.2 Safety-First Principle

Taking Grade-A data center reliability as the bottom line for building-space design, build a full-dimension safe-space system across overall site layout, building structure, functional zoning, and fire evacuation. For the four high-risk areas of a GW-scale campus—the 220kV HVDC supply area, the large-capacity energy-storage area, the liquid-cooling pipe-network area, and the high-density compute-cluster area—use civil measures such as physical isolation, spatial separation, load reinforcement, and leak-/flood-prevention space reservations to build a solid spatial-safety foundation for the

ultra-large campus.

1.1.1.3 High-Power-Density Adaptation Principle

Building space fully adapts to the high-power-density characteristics of a GW-scale AI data center—single rack above 140kW, campus total power ≥1000MW—focusing on four civil metrics: structural load, indoor clear height, column-grid scale, and pipe/cable space. By enlarging the building-space scale, strengthening structural bearing capacity, and reserving vertical pipe/cable space, it meets the installation and O&M space needs of wide-body racks, liquid-cooling CDU units, and rack-level backup-power equipment.

1.1.1.4 Network-Topology Adaptation and Evolution Principle

The overall site plan and building-unit layout are synchronized with the high-speed network topology of the AI data center. The core compute area, network-switching area, and storage area adopt close-proximity physical concentration, reducing network-interconnect loss by optimizing spatial distance. Dedicated space and vertical pipe/cable channels are reserved inside buildings for fiber distribution, high-speed interconnect racks, and cross-cluster links, supporting spatial evolution for generational bandwidth upgrades without later civil renovation.

1.1.1.5 Elastic and Flexible Design and Expansion Principle

Adopt a two-dimensional elastic design of campus-level overall planning plus building-level modular partitioning. Building space is divided by Pod compute-cluster units, supporting phased deployment and on-demand expansion. The building structure, power distribution, cooling pipe network, and fire systems all use pooled, reserved designs, so the spatial layout can flexibly adapt to dynamic training/inference cluster adjustments and air-/liquid-cooling hybrid deployment, avoiding the space waste of one-time construction and improving the space utilization and operational flexibility of the GW-scale campus.

1.1.1.6 Green, Intensive, and Efficient Principle

With lowering PUE and WUE as the core goal of building-space planning, the overall site layout prioritizes the use of natural cooling sources and optimizes the heat-transfer spatial path between the cooling system and the compute area. The building envelope adopts high-efficiency thermal insulation to reduce cooling energy loss; functional zoning follows an intensive, compact principle, strictly controlling the area of auxiliary supporting zones to improve land and building-space utilization. Differentiated spatial design is applied to water-scarce western regions and water-rich southern regions, achieving the dual goals of green/low-carbon and spatial efficiency.

1.1.2 Site Selection and Site Planning

Site planning is the prerequisite for the building space of a GW-scale AI data center, mainly addressing four civil issues: campus siting conditions, overall site layout, vertical design, and pipe-network coordination.

1.1.2.1 Basic Site-Selection Requirements

1. Core-resource adaptation: Prioritize siting at national compute-hub nodes of "Eastern Data, Western Computing." The site must have campus-level power-access conditions (specific access capacity and voltage levels are given in Section 1.2) and be close to renewable-energy-rich areas; ensure communication-trunk reachability to meet the network-transport needs of

cross-regional compute dispatch. Water-rich areas should have stable industrial-water quotas; water-scarce/no-water areas should reserve space for water-free cooling systems.

2. Environmental safety isolation: The site should be far from dust, harmful gases, corrosive and flammable/explosive facilities, and avoid geological-hazard-prone areas, ensuring structural safety of the building site at the source.

3. Interference avoidance: The site should be far from strong vibration sources, strong noise sources, and strong electromagnetic-interference areas, and keep a compliant safe distance from civil buildings and public facilities, meeting the quiet-operation spatial-environment requirements of the AI data center.

1.1.2.2 Basic Site-Planning Requirements

1. Building layout

The building layout may use the “main computer-room building + office support building” model (hereinafter the traditional model) or the building-cluster model. The building-cluster model splits the computer rooms and power/utility rooms of the main building in the traditional model into separate buildings. When using this model, the overall plan should adopt a centripetal building-cluster + physical-isolation layout—the computer-room building at the center of the campus, with power buildings around the IT buildings—forming a standardized civil/space structure, divided into two typical layouts:

1) High-rise single-building centripetal layout: a single high-rise computer-room building serves as the geometric center of the campus, with power buildings arranged symmetrically around it, forming a “one-core, four-auxiliary” symmetric structure. In this layout, the pipe/cable paths between the power buildings and the core computer-room building are equal in length and the spatial distance is shortest, greatly reducing the civil routing length of power, cooling, and communication lines.

2) Multi-story cluster centripetal layout: 2–3 multi-story computer-room buildings form the campus core cluster, surrounded by power buildings, forming a “multi-core, multi-auxiliary” expandable structure. This layout supports phased construction—computer-room and power-building clusters can be added as compute demand grows without changing the campus trunk pipe routing—adapting to the elastic-expansion principle of a GW-scale campus while keeping the spatial distances of each computer-room and power building consistent for standardized O&M and pipe/cable layout.

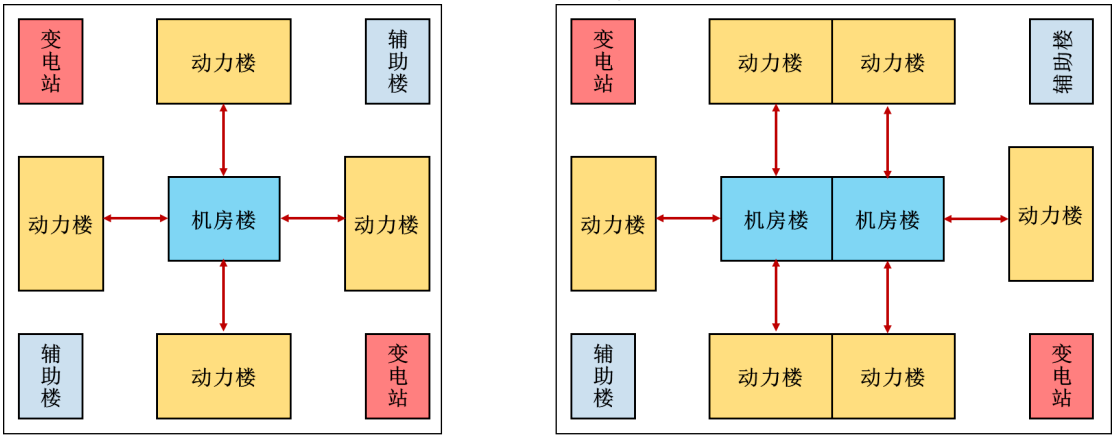


Figure 1.1 Schematic of the centripetal building-cluster layout

2. Entrances/exits and circulation

The campus sets independent pedestrian, vehicle, and logistics entrances/exits, achieving three-stream separation. The main entrance reserves turning and turn-around

space for large prefabricated modules and heavy racks; fire lanes are ≥4m wide with a turning radius of ≥15m for freight, forming a continuous fire loop around the campus.

3. Site vertical design

Site drainage slope ≥0.5% to eliminate ponding; the indoor-outdoor level difference is reasonably set based on terrain and the 100-year flood level; in typhoon areas, increase the indoor-outdoor level difference and add backflow-prevention measures.

4. Integrated pipe network

Coordinate HVDC supply, liquid-cooling/fluorinated-fluid circulation, communication fiber, fire protection, and waste-heat-recovery pipelines, laid centrally in an integrated utility tunnel. The pipe-network routing matches the building layout and network topology, reserving expansion interfaces to avoid later renovation damaging the civil structure.

1.1.3 Core Building-Unit Design

The core building unit is the key carrier of compute hosting, energy assurance, and cooling operation of a GW-scale AI data center; its design directly determines campus operational efficiency, safety/reliability, and long-term evolution capability. This subsection focuses on five core civil dimensions—functional zoning, spatial scale, structure, envelope, and finishing—and, combining national standards with GW-scale campus practice, refines design requirements and clarifies their basis, providing dedicated, compliant, and deployable building space for GW-scale compute, energy, cooling, and O&M systems.

1.1.3.1 Functional Zoning

In the traditional building-layout model, the main computer-room building should adopt a centripetal single-building layout of “main computer room in the center, support facilities around the periphery.” It is divided by function into four core zones, each spatially independent with clear boundaries. The refined design requirements and basis for each zone are as follows:

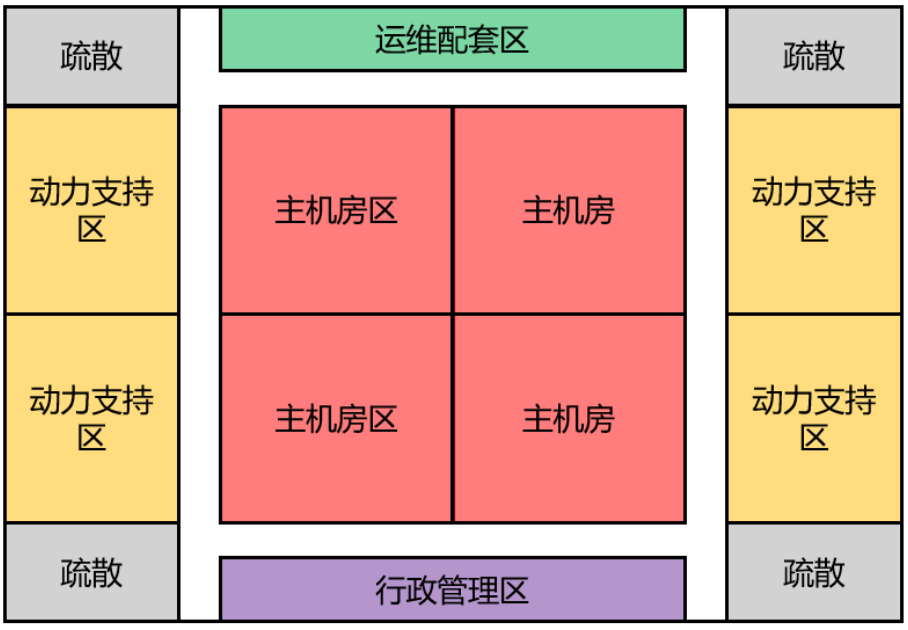


Figure 1.2 Schematic of the centripetal single-building layout

1. **Core main-computer-room zone: as the core space of the compute building, accounting for 20%–30% of the total floor area of a single computer-room building, the main computer room is the core hosting area for high-density rack clusters. Its layout should be close to the power-support zone to shorten pipe/cable**

transmission paths; the space should be regular to accommodate batch deployment of wide-body racks and whole-rack super nodes, reserving rack-row spacing and O&M operating space.

2. **Power-support zone: may use a separate building or an independent partition within the computer-room building. When a separate power building is used, it must keep a fire-separation distance from the compute building. When inside the building, the support zone accounts for 40%–50% of the total floor area of a single building; as single-rack power grows, the support-zone area expands accordingly. Diesel generator sets may use outdoor containerized form.**

3. **O&M support zone: accounts for 8%–10% of the total floor area of a single building. It meets the space needs for daily O&M, monitoring, equipment commissioning and fault handling, and spare-parts storage.**

4. **Administrative-management zone: set as needed, accounting for 5%–10% of the total floor area of a single building. The office area must be physically isolated from the core compute zone and energy-assurance zone to avoid interference.**

1.1.3.2 Building Design

1. Unit scale: a single core computer-room building of a GW-scale campus has a floor area ≥20,000 m², modularly partitioned by Pod compute units, with a single Pod adapted to a 50MW–100MW power load. The number of floors of a single building should be controlled at 3–5; too many floors increase vertical-transport cost and cooling-pipe loss, while too few occupy more land. Building length should not exceed 120m and width should not exceed 60m, facilitating fire evacuation and equipment O&M while suiting the transport and installation of modular prefabricated components, in line with the green/intensive/efficient principle.

2. Floor-plan layout: when the core compute area uses non-immersion cooling, it should adopt a row-based layout with closed hot/cold aisles, with racks arranged face-to-face and back-to-back to form clear hot/cold aisles, where the cold-aisle width ≥1.5m and the hot-aisle width ≥1.8m for cooling circulation and heat removal; liquid-cooling racks and CDU units should be placed nearby to shorten the secondary-side pipe distance and reduce cooling loss and pipe resistance. Liquid-cooling rooms and CDU equipment areas use locally raised floors with sloped drainage and supporting flood-diversion space. Dedicated pipe/cable shafts should be reserved inside the building for centralized laying of power, cooling, and network lines, avoiding exposed lines that hinder O&M and space use; shaft size ≥1.2m × 1.2m for installation and maintenance.

3. Floor height and column grid: a floor height of 7.2m per level is recommended, reserving vertical installation space for liquid-cooling racks (≈2.5m high), overhead cable trays (≈2.9m high), and CDU equipment (≈1.8m high), and reserving 10%–15% vertical redundancy for later equipment upgrades. The column-grid spacing should match the rack-arrangement scheme to ensure a reasonable number of racks per row and sufficient O&M space.

4. Envelope: exterior walls should use Grade-A fire-resistant insulation, preferring high-efficiency materials to reduce cooling energy. Liquid-cooling areas should add a secondary anti-seepage layer to prevent coolant leaks from damaging the building structure. Doors and windows should use high-airtightness, dust-proof, and sound-insulating designs to reduce heat conduction and external interference and meet computer-room cleanliness requirements.

5. Interior finishing: all areas should use Grade-A non-combustible, anti-static, dust-proof, and moisture-proof finishing materials to eliminate fire hazards and dust pollution.

Floors should use anti-static flooring; liquid-cooling areas add anti-slip and

leak-diversion channels connected to a collection sump for easy collection and handling of leaked media. Walls should use fire-, moisture-, and dust-proof latex paint or color-steel panels in light colors to enhance indoor lighting.

1.1.3.3 Structural Design

1. Structural design standards:

For new GW-scale campus building units, the building structure safety grade should be Grade 1; seismic fortification category: not lower than Category B (key fortification) for new buildings and not lower than Category C (standard fortification) for retrofits; the structural design standard of supporting buildings/structures should be consistent with the main computer-room building.

2. Reasonable structural layout:

1) Neither new GW-scale AI data centers nor their supporting building units should use single-span structures or structures dominated by partial single-span frames;

2) Given the large load and high clear-height needs of GW-scale AI data centers, the structural column grid can adopt different spans in the X and Y directions based on rack-layout characteristics—maximizing span in one direction to fit the most rack rows, and appropriately reducing the other; secondary-beam spacing can be densified in the long-span direction, while the short-span direction serves as the main load-bearing beams with larger frame-beam section height; optimal span modules include 9.6m × 6.0m (short direction) and 12.0m × 7.2m (short direction);

3) Structural-joint layout: structural joints must not be crossed within the main computer room; the overall building should, through appropriate measures, avoid structural joints as far as possible;

4) Given the GW-scale computer room's features of heavy load, large span, high floor height, and large investment, the structure should consider seismic isolation/damping design—for the whole building, and also using isolation pads for heavy single racks, or overall isolation at a raised floor layer; at least one of these isolation/damping measures should be adopted;

5) Relatively heavy equipment should be placed on lower floors and relatively light equipment on the top floor;

6) For the roof structure, engineering experience shows that structural sloping is favorable, freeing up load margin to allocate to large rooftop equipment;

3. Load values:

1) For structural calculation of new GW-scale AI data centers, the standard live-load values of floor slabs in major functional rooms should not be less than the following:

Liquid-cooling CDU (main computer room; single-rack weight 2 t) ≥16 kN/m²;

Liquid-cooling CDU (main computer room; single-rack weight 3 t) ≥20 kN/m²;

Diesel generator sets (generally built separately outdoors), floor live load ≥24 kN/m²;

UPS room 10 kN/m²; gas-cylinder room 8 kN/m²;

Battery room / energy-storage equipment area floor live load: 24 kN/m²;

Power distribution room: 16 kN/m²;

HVAC area: 8 kN/m²; equipment transport corridor: 12 kN/m²;

Computer-room and diesel-generator areas should consider a dynamic factor of 1.05;

Suspended load under the slab in computer-room areas: 2.0–4.0 kN/m²;

Large rooftop equipment (dry coolers, fluorine pumps, cooling towers, etc.) should be included in structural calculation by actual weight;

2) For new GW-scale AI data centers, the load-combination, frequent-value, and quasi-permanent-value coefficients of live loads may be taken from Appendix 1 of the "Code for Design of Data Centers"; when calculating seismic action, the combination coefficients of the above variable loads are as follows:

Main computer room / HVAC area / substation: 0.9;

Battery room / UPS room / pipe-cable suspension / diesel generator: 0.9–1.0

Equipment transport corridor: 0.7

3) Given the high load level of GW-scale AI data centers but the spacing requirements of rack and equipment layouts, when calculating components at different structural locations, reference can be made to Table 8.2.2 of the “Code for Design of Telecommunication Buildings”: for rooms with loads greater than 8 kN/m², the live-load reduction factor for main frame beams, columns, walls, and foundation components may be 0.75; for secondary-beam components it may be 0.85; slab calculation is not further reduced;

4) For retrofit projects of existing buildings, the above live-load values may not apply. In design, loads may be input as local slab-surface loads based on the actual weight and specific positioning of equipment; but considering maintenance/installation and liquid-leakage risk, a small uniformly distributed live load should be appropriately superimposed;

5) Since the design working life of a computer-room building is generally longer than the lifecycle of the equipment, and as technology iteration in the data industry accelerates, it is recommended where feasible to appropriately relax the live-load values for floors at the design stage, taking the upper-limit standard as far as possible, so that the building can be flexibly re-laid out within its working life;

4. Material selection:

Given the high load level of GW-scale AI data centers, when using reinforced-concrete structures, the main load-bearing members should use HRB500/HRBF500 grade rebar; when using steel structures, high-strength steel of Q355 or above should be used;

5. Structural design and detailing requirements:

1) When the computer room requires a raised platform, the design standard and load values of the raised structure should be consistent with the main structure; the raised (steel-frame) structure also requires ultimate-limit-state and serviceability-limit-state checks;

2) When major equipment transfers (concentrated) forces to the main slab through the raised structure, the punching-shear capacity of the slab should be additionally checked;

3) Structural calculation should fully consider the effect of the raised-platform detailing on the main structure—i.e., the seismic mass point is actually located at the top of the raised platform, so the stirrup-densification zone of the main frame columns should include the raised height; where necessary, the raised structure and the main structure should be modeled together and calculated as a whole;

4) Seismic supports/hangers for pipes and cables should be sized and spaced based on calculation;

5) Given the high floor height of GW-scale AI data centers, stairs are generally at least 3 or 4 flights; in structural calculation, special attention should be paid to taking the stair self-weight and stair live load as double or triple according to the number of flights;

6) If a GW-scale AI data center requires whole-machine delivery, lifts must be arranged in the transport-corridor area of the computer room, and structural members related to the lifting equipment should consider a dynamic factor of 1.05–1.2;

7) Per substation design requirements, areas with transformers should have lifting openings and hooks; hooks should be arranged on structural beams, and the hook beams should include the lifting-condition load; the lifting-opening size should fully consider operating buffer space;

8) When the computer room is of steel-structure type, fire-protection and anti-corrosion design should be carried out per the corresponding fire-protection standards;

1.2 Power Supply and Backup System

1.2.1 Power-Demand Characteristics of Gigawatt-Scale AI Data Centers

1.2.1.1 Power Demand of AI Training and High-Density Power Racks

The GW-scale AI data center has become the core construction form of ultra-large-scale compute infrastructure. Its power demand differs fundamentally from that of traditional MW-scale data centers, mainly in ultra-high single-rack power, dynamic load characteristics, gigawatt-scale supply capacity, and fine-grained redundancy needs—these are the core origin and underlying basis of power-distribution architecture design.

The load-rate fluctuation of an AI data center is markedly higher than that of a traditional data center: when a training job starts, the load can rise rapidly from a low trough to a high peak, and fall rapidly after the job ends. When large numbers of racks start simultaneously or training jobs enter peak phases in sync, the supply system experiences large-scale, high-amplitude instantaneous power surges. The power-distribution system must therefore have very strong dynamic-load adaptability, fast voltage-regulation capability, and load-balancing dispatch capability to effectively damp instantaneous shocks at the architectural level.

1.2.1.2 Overall Planning of the Power Supply and Backup System

Power-capacity planning of a GW-scale AI data center must adapt to the rapid-growth characteristics of AI compute, following the core principles of “advance planning, phased construction, flexible expansion, redundant reliability.” The campus’s external-power access capacity should meet full supply of the core load, reserve expansion headroom, and adapt to the stepwise growth of compute demand over the next 3–5 years.

As shown in Figure 1.3, the power supply and backup system of a GW-scale AI data center is organized from high to low voltage levels and divided into sections by on-site deployment location, complete power chain, and system function, explaining the system architecture, configuration features, and operating logic block by block. The campus trunk-interconnect layer is the main artery of power transmission, handling the transformation from 220kV to 10kV or 35kV; the building-branch-interconnect layer is the branch transmission network linking the campus trunk ring to each computer-room building, handling distribution from the campus central substation to the distribution rooms in each building, using 10kV or 35kV, then converting AC to 800V DC via an energy router or solid-state transformer (SST); the in-room distribution-interconnect layer is the trunk distribution network inside the building, handling power delivery from the building distribution room to each rack zone, using the 800V DC voltage level. In-rack power supply is detailed in later chapters.

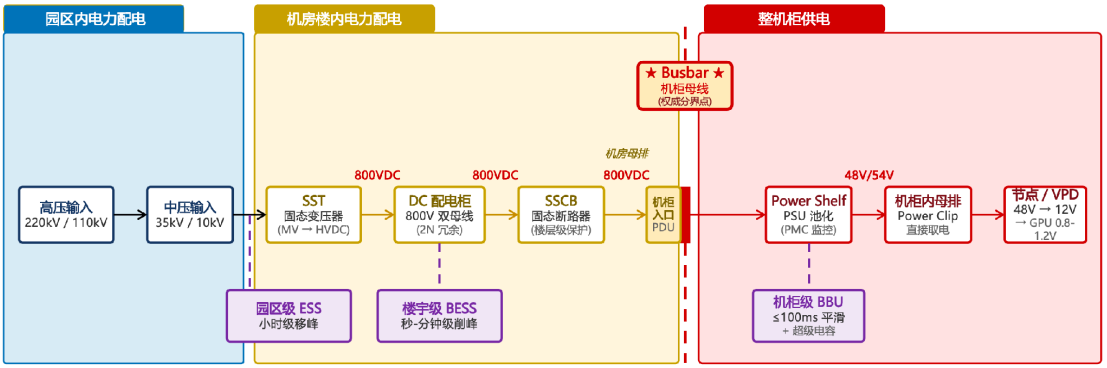


Figure 1.3 Schematic of the power supply and backup system stages

1.2.2 Campus-Level High-Voltage Supply System

The GW-scale AI data center campus builds an integrated overall architecture of source–grid–load–storage coordination, achieving reliable supply, efficient transmission, flexible dispatch, and low-carbon consumption of electricity. Based on a multi-source coordinated supply system, it gradually raises the share of new energy such as wind, solar, hydro, and nuclear, and deploys energy storage to increase the green-power consumption ratio and reduce carbon emissions.

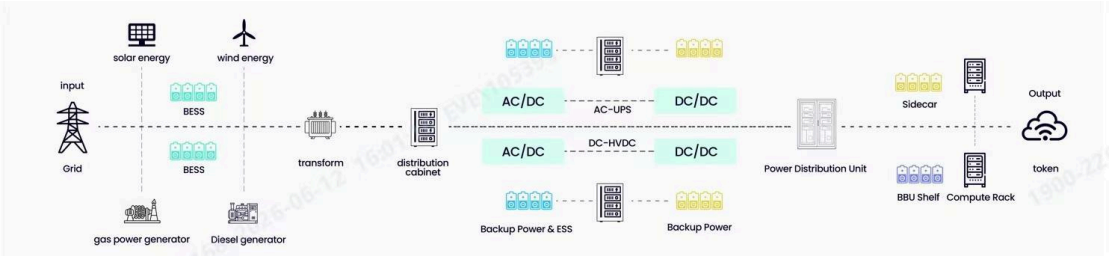


Figure 1.4 Schematic of the multi-level energy-storage coordinated layout

As shown in Figure 1.4, on the storage side a three-level “campus–building–rack” coordinated energy-storage layout is built: the campus level deploys long-duration storage handling grid peak-shaving, long-duration backup, and virtual-power-plant operation; the building level deploys short-duration storage handling peak-valley arbitrage, power-quality management, and short-duration backup; the rack-level BBU can handle short-duration energy buffering, transient power support, and backup-power transition support—forming an all-time, all-scenario energy-storage assurance and dispatch system.

Through the campus-level energy-management system, data interconnection and coordinated dispatch are achieved—optimal power allocation and maximum energy efficiency under normal conditions, uninterrupted load assurance under grid-fault conditions, and maximum renewable-energy consumption and power-trading operation under low-carbon-operation scenarios—fully adapting to the high-capacity, high-reliability, low-carbon operation needs of a GW-scale AI data center.

1.2.3 Campus Medium-Voltage Distribution System

1.2.3.1 35kV vs 10kV

GW-scale AI data centers cover a large area, with dispersed computer-room buildings and large single-building load capacity. The medium-voltage distribution system must balance three core needs: low loss in high-capacity long-distance transmission, flexible dispatch of multiple load zones, and rapid isolation under fault scenarios. Although 10kV distribution still dominates at present, 35kV—as a mainstream medium-voltage level in China—offers differentiated advantages. The core advantages of 35kV are concentrated in three dimensions—high capacity, low loss, and long-distance transmission: at the same current-carrying capacity, the transmission capacity of 35kV is more than three times that of 10kV, greatly saving investment in cables, utility tunnels, and switchgear.

1.2.3.2 Medium-/Low-Voltage Supply Architecture

The GW-scale AI data center campus uses a two-level layout of “campus central main substation + distributed annex substations at each building.” The medium-voltage side may use a double-busbar sectionalized scheme, with the two busbars fed from different sections of the campus 35kV/10kV and backing each other up; the low-voltage DC side uses a double-busbar connection, with each DC busbar segment fed by an independent rectifier group, forming a dual-path redundant DC supply that directly powers the building.

1.2.3.3 Storage Replacing Part of the Diesel-Generator Scheme

For the backup-power requirements of a GW-scale AI data center, a backup architecture of “long-duration storage as the main, retaining some diesel generators for emergency backup” is adopted. The campus core IT load is provided with 2–8 hours of storage to handle priority backup during external-power outages. Diesel generators with 50% of the core-load capacity are configured, started only in extremely low-probability scenarios such as prolonged grid faults or extreme storage-system failures.

Energy storage adopts prefabricated-container modular form. In the market, large-capacity batteries are trending toward single-cell capacity greater than 600Ah; on a high-energy-density basis, ultra-simple integration can be achieved, improving system reliability, safety, and economy—an ideal storage route for GW-scale compute centers.

1.2.4 In-Building Power Distribution System

1.2.4.1 Evolution Path of the Supply Architecture

The traditional 380V low-voltage AC supply architecture has a long power chain and many conversion stages, possibly facing efficiency bottlenecks under high-power scenarios. The industry is gradually introducing 240V/336V DC schemes, directly powering servers via DC busbars and enabling direct battery connection.

By further raising the output voltage level—e.g., 800VDC or ±400VDC—the data center meets higher power demand while effectively reducing transmission current and line loss and improving system efficiency. Meanwhile, the HVDC architecture has a natural advantage of directly coupling with DC energy systems such as PV and storage, restructuring the traditional supply model at the system level and becoming an important evolution direction for data centers oriented to high compute density, high energy efficiency, and green sustainability.

1.2.4.2 Full-DC Supply Architecture

A currently promising 800VDC or ±400VDC supply architecture uses an energy router or solid-state transformer (SST) to directly convert 10kV medium-voltage AC into stable, controllable 800VDC or ±400VDC, then provides servers with 48V or lower via high-efficiency DC-DC modules. This architecture significantly shortens the power chain and reduces the number of AC/DC conversion stages. In terms of efficiency, traditional UPS and HVDC overall efficiency is typically 95%–97.5%, whereas the SST-based 800VDC or ±400VDC architecture has measured efficiency above 98%, with significant annualized energy savings—an end-to-end system efficiency improvement of about 3%–5%.

In terms of materials, the high-voltage, low-current full-DC scheme uses about one-third the copper of a traditional AC scheme; in terms of power density and space use, reducing the number of supply devices cuts footprint by 40%–50%; in terms of construction difficulty, modular, prefabricated systems greatly shorten the build period; in terms of energy, new energy and storage can be connected in DC form, reducing unnecessary AC/DC conversion stages.

1.2.4.3 HVDC Storage Configuration

HVDC-based storage must provide 5–15 minutes of short-duration backup to support the transition of backup-power switching. As campus scale grows, the imbalance of the computer-room gray/white space requires higher-energy-density storage equipment, which must also be highly reliable, highly safe, and feature intelligent monitoring. Lead-acid batteries can no longer meet current needs; the lithium-iron-phosphate route—thanks to large-scale application in the storage industry, especially the development of stacked-cell technology—offers high-rate discharge, low temperature rise, high safety, and real-time state monitoring, making it the main technical direction for large-scale data-center backup.

1.2.5 Operation Protection and Control System

1.2.5.1 Multi-Level Protection Strategy

The core difficulty of protecting a full-DC supply system is that DC current has no natural zero crossing, the arc is hard to extinguish during a short-circuit fault, and the fault current rises extremely fast. The solid-state circuit breaker, with its microsecond-level breaking speed and arc-free breaking, becomes the core protection device of a full-DC supply system. Combined with solid-state breakers, a multi-level full-chain protection strategy of “campus–building–rack–device” is built to achieve rapid fault isolation and minimize impact.

From device level to campus level, protection-action times are set with a progressively increasing time gradient, ensuring lower-level protection acts first and upper-level protection acts only as backup when the lower level fails to act; each protection level has fault-direction discrimination, acting only on faults within its protection zone and not on through-faults outside it. Through the campus SCADA system, protection devices at all levels achieve data interconnection and coordinated linkage; when a fault occurs, the fault point can be precisely located and commands issued simultaneously to achieve precise isolation and automatic load transfer, further improving protection-system reliability.

1.2.5.2 Intelligent Control System of the AI Data Center

Through a full-lifecycle digital management system for power distribution, combined with the campus-level energy-management system, the transition from “passive O&M” to “proactive and predictive O&M” is achieved. It accurately reconstructs the full-chain supply architecture, achieving real-time mapping and two-way interaction between the physical system and the digital model. Core functions and applications include full-system real-time visual mapping, multi-scenario simulation and scheme validation, remote control and unattended O&M, integrated energy management, and power-trading operation—comprehensively improving system reliability, O&M efficiency, and energy efficiency, while also achieving additional economic functions such as maximum renewable-energy consumption and power trading under low-carbon operation.

1.2.6 Conclusions and Outlook

Combining current industry technology trends, development mainlines, technical gaps, and engineering-deployment needs, the core technology trends of the GW-scale AI data center power supply and backup system can be summarized into six directions, covering architecture upgrade, safety protection, storage backup, standards building, intelligent O&M, and low-carbon transformation.

1. High-voltage full-DC efficient supply

800V full-DC supply fully replaces traditional AC UPS supply, centered on the energy-router/SST, simplifying the power chain and improving system efficiency and power density to meet MW-scale single-rack supply needs. Going forward, the application of wide-bandgap semiconductor devices in supply rectification and protection will be explored, the 35kV HVAC-to-800VDC direct-conversion technology will be validated at scale, and the multi-level full-chain supply architecture across campus, rack, and chip will be continuously optimized to comprehensively improve transmission efficiency.

2. Fine-grained safety protection of HVDC

DC protection technology is iterating toward microsecond-level, full-chain, and intelligent directions; solid-state breakers will achieve full-chain fine-grained protection, DC short-circuit arc extinguishing, and multi-level protection coordination. Going forward, focusing on 800V-and-above HVDC scenarios, key research will target rapid short-circuit fault detection, arc-free breaking, and precise ground-fault location, optimize multi-level intelligent protection coordination, and improve the HVDC supply safety-protection system and industry technical standards.

3. Multi-functional storage-backup integration

Storage systems have iterated from single-purpose emergency backup to a multi-function fusion system of “backup + peak-shaving + power-quality management + virtual power plant + green-power consumption,” with the “long-duration storage + backup power” model achieving low-carbon, high-economy backup. Going forward, large-scale application research on hybrid storage systems will be carried out, capacity configuration and coordinated dispatch optimized, and a multi-level storage safety-protection system built to improve intrinsic safety.

4. Standardization of the power-distribution interconnect system

For the full-DC supply architecture, the industry is accelerating the building of unified interconnect standards, regulating the parameters and design of core components such as device interfaces, transmission lines, and connectors, to achieve full-industry-chain interoperability and compatibility and lower engineering-application cost. Standardization research on full-DC power-distribution systems has begun; going forward, unified technical parameters, test standards, and design specs for 800V DC power sources, busways, connectors, and other devices will be formulated to solve the pain points of inconsistent device interfaces and poor compatibility and promote large-scale deployment of full-DC technology.

5. All-scenario intelligent O&M and control

Relying on digital twin, AI large models, and IoT, the industry is gradually building a full-lifecycle digital management system for power distribution, achieving visual monitoring, predictive O&M, and intelligent coordinated dispatch. Going forward, key breakthroughs will target real-time closed-loop control between the digital twin and the physical system, develop AI-large-model-driven intelligent dispatch algorithms to maximize operating efficiency and economy, and deepen research on intelligent fault diagnosis and predictive maintenance to help achieve zero-fault operation.

6. Building a zero-carbon power-distribution system

Deep integration of the power-distribution system with renewable energy is a core trend. Through source–grid–load–storage, direct green-power connection, energy-efficient equipment, and clean backup, system loss and carbon emissions are continuously reduced to build a zero-carbon power-distribution system. Going forward, the deep coordination of new energy and compute centers will be continuously optimized, green-power consumption and coordinated-control capability improved, and a full-chain zero-carbon-emission power-distribution architecture built to drive the full transformation of GW-scale compute centers into zero-carbon compute facilities.

1.3 Cooling System of the Gigawatt-Scale AI Data Center

The cooling system of an AI data center is key infrastructure for ensuring stable IT-equipment operation, controlling energy cost, and improving core competitiveness. Its core goal is to efficiently transfer the large amount of heat generated during server operation to the outside, keeping core server components (CPU, GPU, etc.) in a safe, stable operating range while achieving optimal energy efficiency and controllable cost. As AI compute demand explodes, single-rack power density has soared from a traditional 8–15kW to 30–60kW; with the deployment of architectures such as NVIDIA GB300, single-rack cooling power has exceeded 150kW, and full-liquid-cooling architectures such as Rubin Ultra push single-rack power to 600kW. Data-center energy use is also moving from the MW scale toward the GW scale. Against this backdrop, the cooling architecture for GW-scale AI data centers has iterated from traditional air cooling toward a liquid-cooling-dominated, multi-technology-fusion direction.

1.3.1 Cooling System Architecture and Core Equipment

Current mainstream cooling architectures in the AI data center industry fall into three technical paths:

- Forced air cooling: cooling servers via fans, suited to 12–15kW single-rack power density; advantages are low cost and simple O&M, disadvantages are high noise and high energy use.
- Cold-plate liquid cooling: centered on a cold plate attached to core heat-generating components such as CPU and GPU, directly absorbing the heat via circulating coolant, exchanging heat through the CDU, and finally transferring it to the outside atmosphere via cooling-source equipment. This method has good compatibility, balances cooling efficiency and O&M difficulty, and is currently the preferred solution for high-power-density data centers.
- Immersion liquid cooling: fully immersing servers in insulating coolant, with the highest cooling efficiency, no local hotspots, and low noise. As coolant localization matures and process optimization, supply-chain completeness, and commercial application advance, the economics of single-phase immersion cooling are gradually emerging. The commercial deployment of immersion cooling is now advancing rapidly—on the eve of large-scale deployment.

From the comprehensive comparison above, cold-plate liquid cooling solves the pain points of insufficient cooling efficiency and high energy use of traditional air cooling, while immersion cooling has not yet reached large-scale deployment, with industrialization and scaling not yet formed. Based on the mainstream positioning and broad application value of cold-plate liquid cooling, the following fully dissects its system architecture, operating principle, core equipment, and engineering validation as a reference for practical application and deployment.

1.3.1.1 System Architecture

1. Primary-side system architecture

The primary-side liquid-cooling architecture should be selected based on the site's meteorological conditions (including dry-bulb and wet-bulb temperatures), water resources, and the inlet-coolant temperature of the liquid-cooled servers. Specific selection methods are shown in the table below.

Primary-side liquid-cooling system architecture				
Climate conditions	With a cold season: low ambient temperature, can satisfy year-round free cooling		With a warm season: high ambient temperature, requires mechanical-cooling supplement	
Water-source conditions	Water-rich area	Water-scarce area	Water-rich area	Water-scarce area
System architecture	Open cooling tower + plate exchanger, closed cooling tower (Figure 1.5)	Dry cooler (Figure 1.6), air-cooled condenser (Figure 1.7)	Chiller + cooling tower + plate exchanger (Figure 1.8)	Air-cooled chiller (Figure 1.9)
Economic comparison				
Cost	Low	Relatively low	Relatively high	High
Energy consumption	Low	Relatively low	Relatively high	High

O&M difficulty	Relatively low	Low	High	Relatively high

Table 1.1 Primary-side liquid-cooling system architecture

The following are schematics of several architectures:

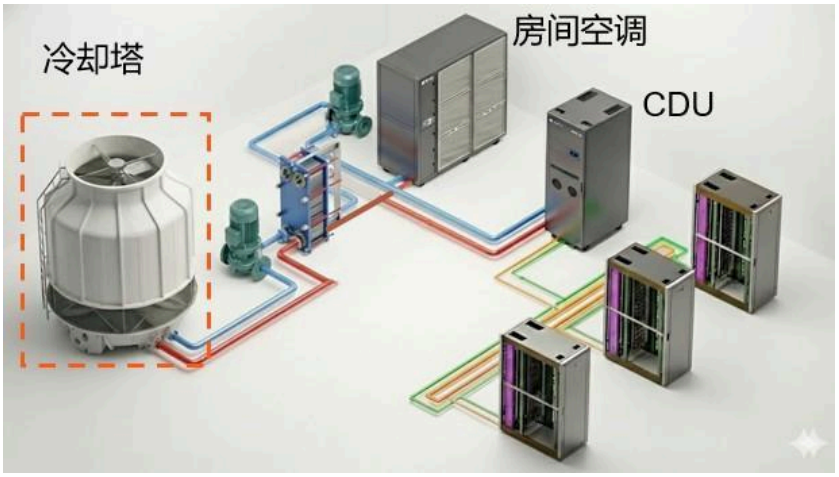


Figure 1.5 Cooling-source scheme: open cooling tower + plate heat exchanger, or closed cooling tower

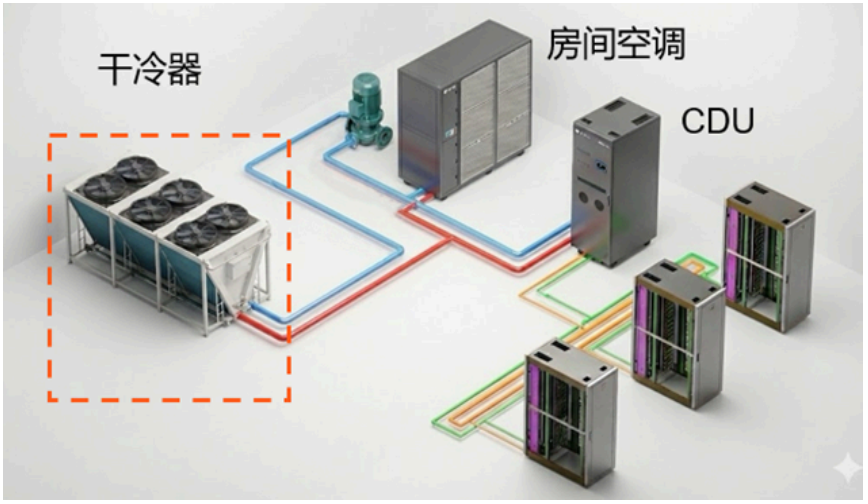


Figure 1.6 Cooling-source scheme: dry cooler

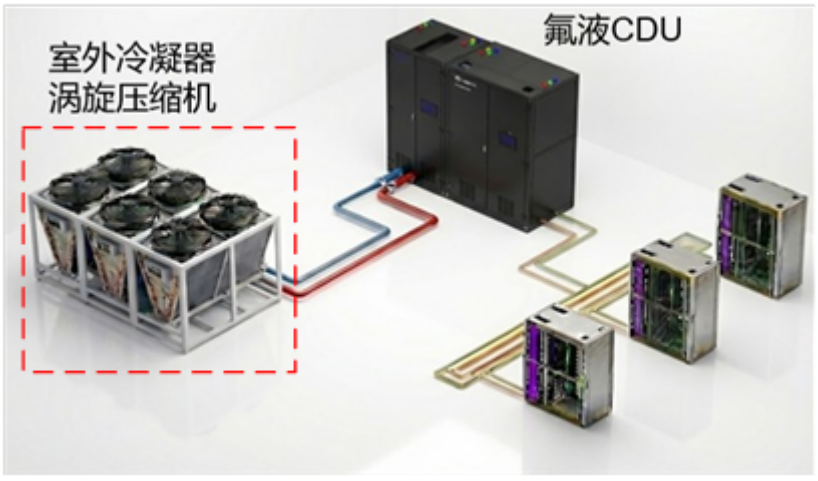


Figure 1.7 Cooling-source scheme: air-cooled condenser

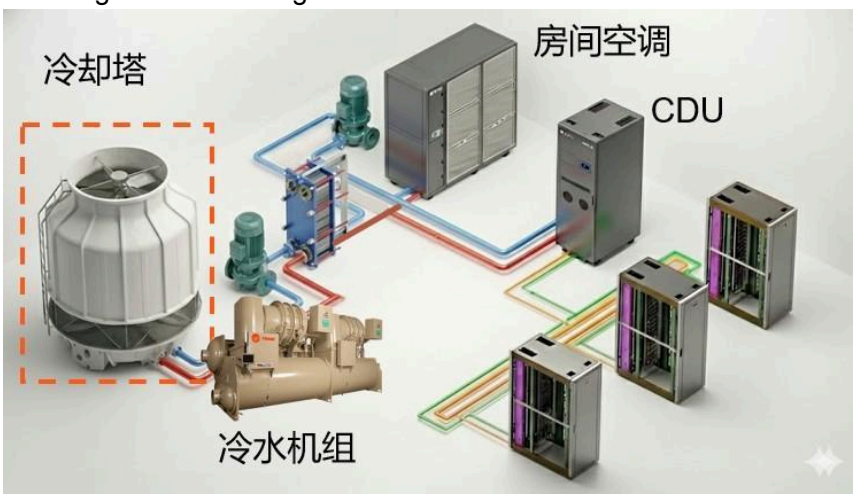


Figure 1.8 Cooling-source scheme: water-cooled chiller

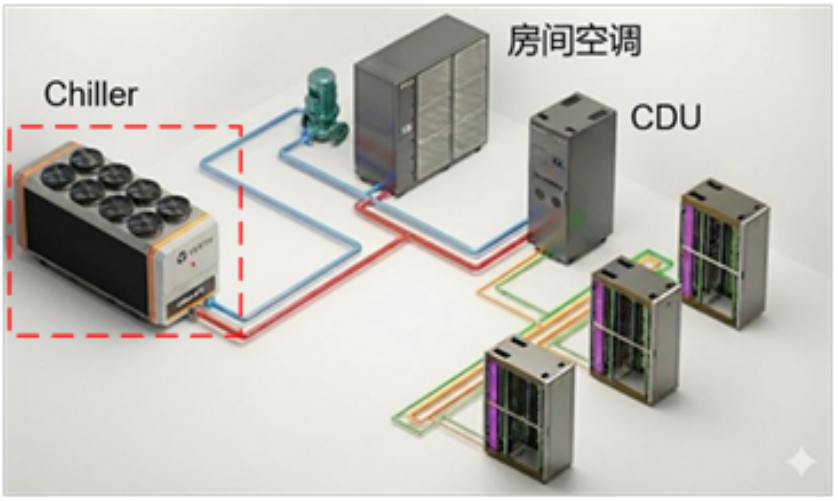


Figure 1.9 Cooling-source scheme: air-cooled chiller

2. Secondary-side system architecture

- **Centralized liquid-cooling system architecture**

The centralized liquid-cooling system is characterized by centralized CDU control and unified pipe-network distribution. It commonly uses 2N or N+X redundancy to suit the reliability-level needs of different compute scenarios, providing redundancy assurance for stable operation of high-density compute equipment.

2N architecture (below): a dual-redundant, dual-bus architecture whose core is deploying two fully independent, equal CDU sets each with 100% load capacity, fully independent in structure, electrical, and control while backing each other up. Under normal conditions, the two CDU sets work in parallel, jointly carrying 100% of the system load; if either CDU fails, the other automatically switches in without interruption via preset logic to carry the full system load, ensuring stable operation of compute equipment. The 2N architecture offers convenient O&M: one CDU set can be powered down for maintenance without interrupting the other.

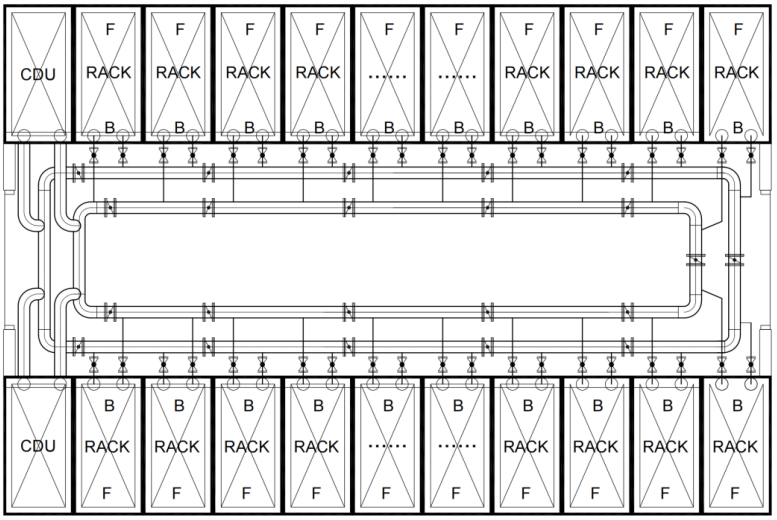


Figure 1.10 Centralized liquid-cooling system, 2N architecture

N+X architecture (below): a modular redundant architecture achieving load balancing and automatic fault switching via intelligent group control; N: the minimum number of CDUs needed to meet the total system load; X: the number of redundant standby CDUs. All CDUs are connected in parallel to the same supply/return pipe network, jointly participating in cooling-capacity distribution and intelligent control. Under normal conditions, N CDUs run at full load and X CDUs are in hot standby; when a CDU fails, a standby CDU quickly switches in automatically to keep compute equipment running safely. Compared with the 2N architecture, the N+X architecture offers better cost and linear-expansion friendliness.

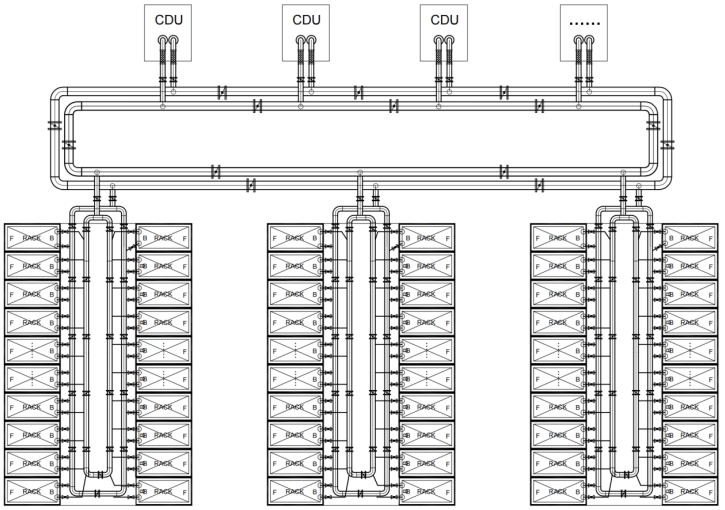


Figure 1.11 Centralized liquid-cooling system, N+X architecture

The 2N architecture focuses on the zero-interruption needs of core compute scenarios, providing the highest reliability; the N+X architecture balances reliability and cost-effectiveness. Both can meet standardized, modular, high-reliability design requirements, providing solid technical support for the green, low-carbon, efficient, and stable operation of high-density AI data centers.

- **Distributed liquid-cooling system architecture**

The distributed liquid-cooling system is characterized by placing the CDU inside the rack and achieving rack-level precise cooling via an independent closed loop. It features flexible deployment, on-demand expansion, the ability to ship with whole-rack servers,

and good isolation—a single-rack fault does not affect the whole. It suits retrofit rooms of high-density traditional data centers and edge-computing nodes with high space requirements. For more details, see Section 2.1 AI Super Node.

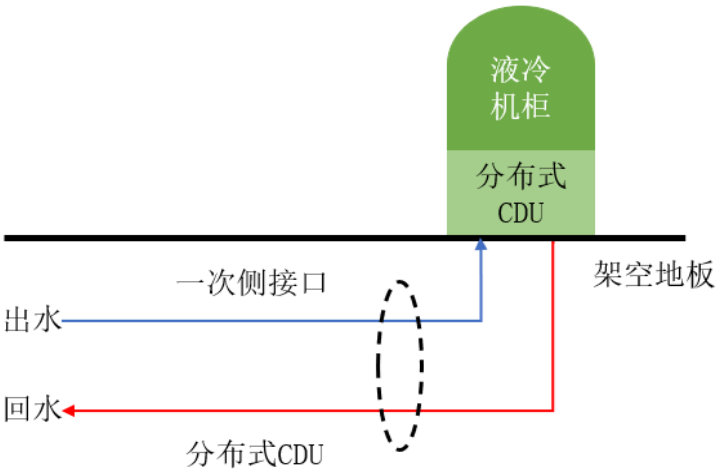


Figure 1.12 Distributed liquid-cooling system

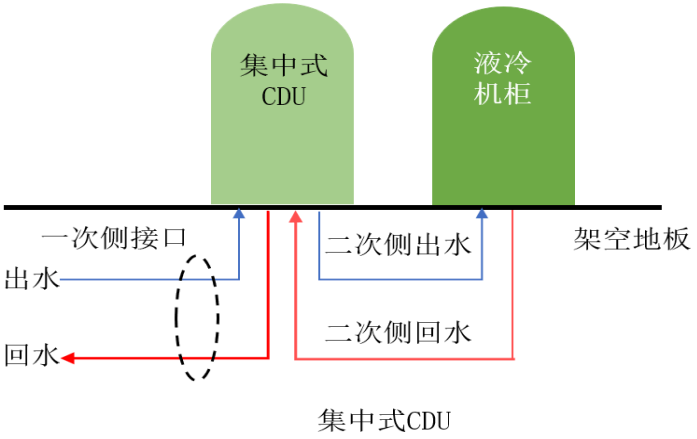


Figure 1.13 Centralized liquid-cooling system

A typical liquid-cooled whole-rack layout is shown below, including compute nodes, switch nodes, power shelves, blind-mate manifolds, cable cartridges, etc.

1.

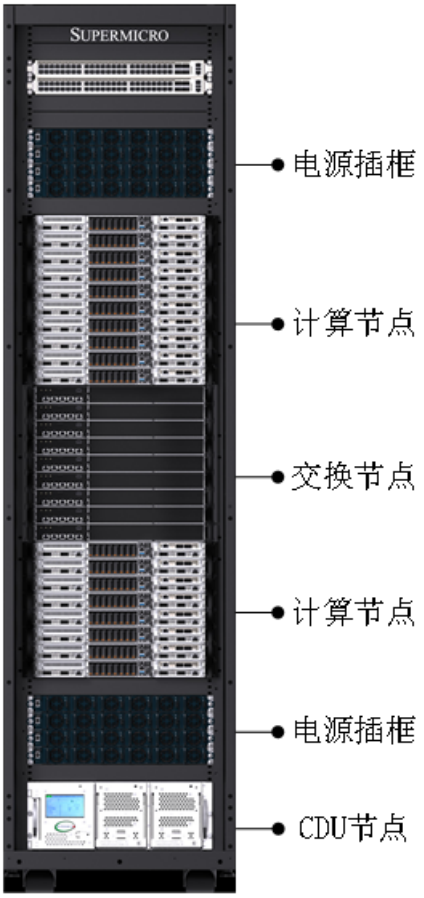


Figure 1.14 Whole-rack architecture

1.3.1.2 Control System

1. Composition and architecture of the cooling-source automation system

The cooling-source automation system mainly monitors and controls the primary-side circulation equipment and cooling towers. It comprises a cooling-source automation controller, pump flowmeters, pressure sensors, temperature sensors, motorized valves, cooling-tower sensors, CDU monitoring, other sensors, and a display. The system architecture is as follows:

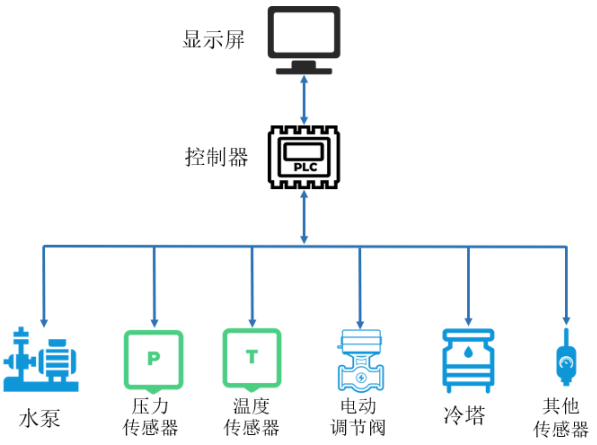


Figure 1.15 Cooling-source automation system

Two schemes—PLC and DDC—are commonly used for the cooling-source

automation system. PLC is stable and reliable, with flexible function development (e.g., hot standby, scheduled cycling), a wide selection, easy expansion, and industrial-grade adaptability to complex environments. DDC is better suited to building HVAC control and is relatively easy to develop and commission.

2. Secondary-side system control architecture

- **Control architecture of the 2N system**

In the 2N architecture, each ring network contains 2 CDUs, for a total of N ring networks. The schematic is below:

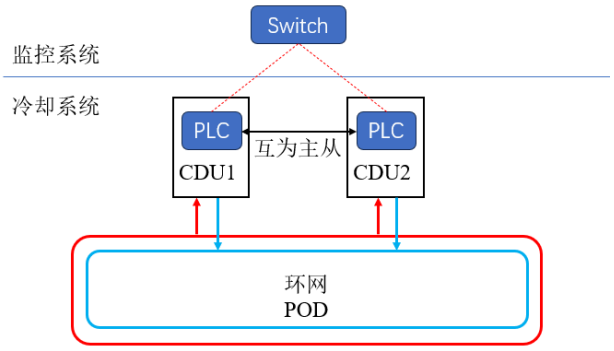


Figure 1.16 2N control architecture

Each CDU in the 2N system has an independent PLC controller, with the CDUs in a master–slave relationship. Operating modes include group-control mode and standalone mode: in group-control mode, the CDU receives command parameters from the group-control host; a CDU can leave group control and run standalone, executing control logic via its own PLC. Each CDU controller communicates with the host monitoring system and uploads operating data.

- **Control architecture of the N+X system**

The N+X architecture suits systems where a single ring network has 3 or more CDUs—N CDUs running and X CDUs in cold standby. The schematic is below:

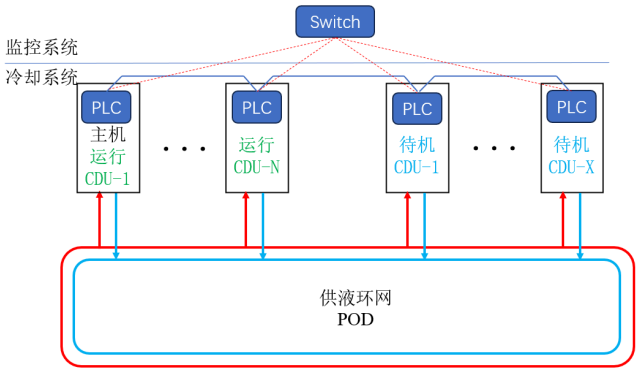


Figure 1.17 N+X control architecture

In the N+X system, one CDU must serve as the group-control host; in group-control mode, the system executes the host's commands, and a CDU can leave group control and run standalone via its own PLC.

Each CDU in the N+X system has an independent PLC controller; the controllers are peers with no single one-to-many master controller, and any CDU controller can serve as the system group-control host. The CDU controllers communicate over a control network to synchronize key data, achieving fault switching and manual master/standby setting. Any two PLCs in the automation network can communicate with each other. Each CDU controller communicates with the monitoring system and uploads operating data.

Host setting and switching of the N+X system: the N+X host can be manually designated, and at any moment there is one and only one host in the system. In special

cases—e.g., the host PLC loses communication with all other PLCs, or the host PLC loses power or fails—the remaining online CDUs can automatically elect a host for group control.

- **Group-control system control logic**

The group-control host uniformly schedules the start/stop, fault switching, and rotation of all CDU units in the system, as well as the adjustment of the secondary-side pumps and secondary-side two-way valves inside the CDUs.

Cold/hot standby: in group-control cold-standby mode, standby CDUs stop and run in rotation at a fixed interval to balance unit running time and extend service life; in group-control hot-standby mode, CDUs in the system run together and share the heat-exchange load equally.

Data synchronization: the control mode and parameters of CDUs in the same system are kept consistent; changing a parameter on any CDU synchronizes the data to the entire system.

Fault switching: CDU fault levels are managed hierarchically; when a fault affecting unit operation occurs, the standby unit must be started and the faulty unit stopped, while keeping stable flow during switching.

Reliability design: key sensors inside the system are backed up to prevent system faults caused by sensor failure. In addition, to keep the CDU from losing heat-exchange capacity due to controller failure, when the controller fails the CDU pump keeps running in its pre-fault state.

1.3.1.3 Heat-Exchange Equipment

1. Centralized CDU key points and trends

Centralized CDUs are divided into in-row and non-in-row types by their physical layout in the computer room. In-row CDUs are placed in line with server racks, generally matching rack height and depth, providing cooling for that row of racks. Non-in-row CDUs are placed outside the server-rack rows in a specific area of the room, generally without strict size requirements; multiple CDUs are usually set to jointly cool the racks of the whole room.

Figure 1.18 Centralized CDU

Design key points

Heat exchanger: when the cold plate is copper, a copper-brazed heat exchanger should be used; when the cold plate is aluminum, an all-stainless-steel brazed or removable heat exchanger should be used;

Pump: the pump mechanical seal should be SiC-on-SiC or SiC-on-tungsten-carbide,

not graphite; wetted material should be 304 stainless steel or above; for system reliability, the pump set needs 2N redundancy and online-replacement capability;

Pipe flow velocity: the internal pipe flow velocity on both primary and secondary sides of the CDU should be ≤3 m/s;

Filtration: primary-side filtration precision should be ≥50 mesh, secondary-side ≥270 mesh; both filters should support online cleaning;

Auto-venting: automatic vent valves should be set at the highest point and local high points of the CDU;

Pressure-relief protection: an automatic relief valve must be configured inside the CDU, able to relieve automatically when system pressure exceeds 5 bar;

Anti-condensation: the CDU must have anti-condensation control to avoid condensation risk in the secondary-side system;

Wetted materials: all wetted materials must meet compatibility requirements; pipes recommended to be 304 stainless steel or above, with matching stainless-steel solder; gaskets and hoses should use EPDM, PTFE, or PVDF. Bronze, carbon steel, brass with >15% zinc, high-lead brass, aluminum and aluminum alloys, PTFE thread tape, thread sealant, etc. are prohibited;

Development trends

High heat-flux density: single-CDU power keeps growing; 2MW single-CDU cooling capacity already exists, and 3MW or higher will appear to meet the high-density needs of AI data centers and supercomputing clusters; multi-module parallel connection supports linear expansion and adapts to phased construction;

Full-chain intelligence: room-level AI scheduling can dynamically optimize supply temperature and flow based on outdoor temperature/humidity and IT load, saving 20%–30% energy; full-chain CDU+piping+rack simulation gives early warning of clogging, leakage, and hydraulic imbalance; prediction of pump life, heat-exchanger fouling, and pipe-corrosion trends;

Higher temperatures + waste-heat recovery: raising supply-water temperature to 40–55°C enables year-round free cooling, lowering PUE to 1.05–1.1; using coolant waste heat for campus heating, domestic hot water, and dehumidification raises overall energy utilization to ≥80%;

Low energy use, long life: high-efficiency magnetic-levitation pumps + variable-frequency control cut transport energy by 40%; corrosion-resistant new materials (titanium alloy, special plastics) extend life to over 15 years.

2. Secondary-side pipe-network key points and trends

The secondary-side pipe network is usually factory-prefabricated: pipes are grouped, with most pipe-module welding, cleaning, and tightness testing done in the factory, then transported under nitrogen pressure to the site for assembly. Only a small amount of pipe installation is done on site, commonly using “hot-work-free” non-welded connections such as clamps and flanges.

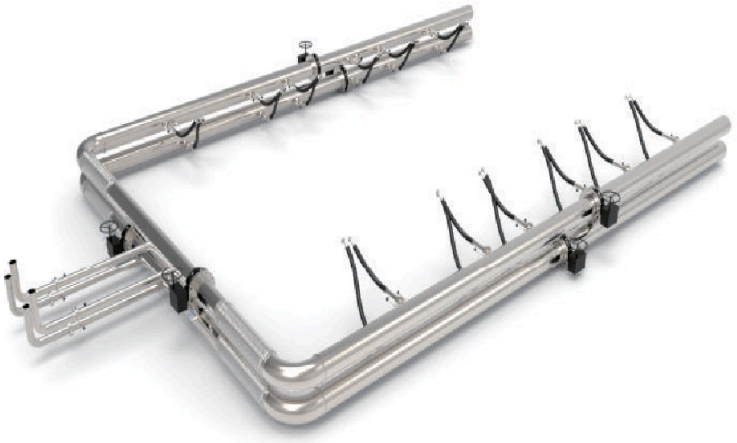


Figure 1.19 Secondary-side engineering pipe network

Design key points

Material selection: liquid-cooling CDUs typically run between 15°C and 45°C, with the working fluid mostly pure water or inhibited glycol. Pipe-material selection must offer corrosion resistance, mechanical strength, machinability, and chemical compatibility with the coolant. In practice, 304L and 316L are the most widely used; the main difference is nickel (Ni) content and the addition of molybdenum (Mo). Balancing overall performance and cost, 304L is preferred for secondary-side piping.

Flexible connection: in pipe design and installation, the effect of thermal expansion and contraction must be considered, allowing expansion/contraction without imposing force on connected pipes or equipment, and meeting pressure-tightness tests under both high- and low-temperature conditions. The tightness-test pressure should be higher than the design pressure (e.g., 1.5× design pressure);

Maintainability: the pipe network should be designed to be removable and flushable, with inspection ports reserved at key nodes;

Cleanliness requirements: before commissioning, the whole system must be circulation-flushed, and coolant samples tested for dielectric constant, volume resistivity, water content, etc., with the liquid free of impurities and the surface free of oil;

Drip trays: drip trays must be set at the bottom of the secondary-side pipe network, usually fully welded around the perimeter. Matching different pipe segments, the drip trays must also be prefabricated in segments and designed with a slope for directional drainage;

Development trends

As AI data center liquid-cooling systems demand higher reliability and stability, the liquid-cooling pipe network is evolving from a “passive heat-dissipation channel” to “intelligent thermal-management infrastructure,” for example:

Distributed sensor network: deploy pressure, temperature, flow, and leak-detection sensors on each branch for real-time sampling and monitoring, detecting operating faults and risks in time;

Digital-twin integration: feed the sensor data from the liquid-cooling pipe network into the AI data center’s digital-twin platform for fault prediction and heat-flow simulation optimization.

1.3.2 Room-Level Cooling System Engineering Validation

1. Tightness test

Before a room-level liquid-cooling system is put into operation, an overall secondary-side tightness test is performed to ensure no seepage or leakage in the whole

secondary-side system, usually by gas or liquid testing.

2. Liquid-supply characteristic test

Used to verify whether the relationship between secondary-side flow resistance and flow rate meets technical requirements; simulation tests are run based on the liquid-cooling system's flow and pressure differential, also probing the upper and lower flow limits of the whole secondary-side system, providing data support for later operation.

3. CDU fault-switching and switching-time test

Simulates the unit-switching stabilization time when a CDU raises a fault alarm, and records the flow fluctuation during switching, providing data support for later operation.

4. Heat-exchange capacity test

Simulates and tests whether the CDU's heat-exchange capacity and the cooling tower's heat-rejection capacity can reach the design values.

1.3.3 Future Outlook

As AI large models iterate at an astonishing pace and compute demand climbs exponentially, server power and cooling needs rise accordingly. Under the dual squeeze of compute density and energy policy, liquid cooling has gradually shifted from an "option" to a "must," becoming a rigid standard. In the future, the cooling system of gigawatt-scale AI data centers will evolve mainly around high-density heat dissipation, parallel cold-plate + immersion dual tracks, extreme PUE, and full-stack intelligence.

1. High-density heat dissipation: single-chip power is leaping from 200W to a future 3000W (NVIDIA GB300/Rubin), and single-rack power is rising to over 150kW, even to the MW level. The heat-flux-density demand on liquid-cooling systems keeps rising, as does the heat-flux-density requirement on heat-exchange components directly involved with servers, such as CDUs and cold plates. Rack-mounted CDU heat-flux density of 4U–150kW has become mainstream, centralized-CDU heat-exchange power has exceeded 2MW, and liquid-cooling cold plates achieve heat-flux density over 150 W/cm² via manifold-microchannel and similar structures. Gigawatt-scale AI data centers are creating strong demand for various new technologies, materials, and architectures.

- New technology: a two-phase liquid-cooling system based on the physical property of latent heat of vaporization through working-fluid phase change (boiling/condensation) further improves cooling efficiency—about 1.3–1.5× that of current single-phase liquid cooling;
- New material: cold plates combining ultra-high-thermal-conductivity materials such as diamond with manifold-microchannel and similar techniques achieve heat-flux density exceeding 1000 W/cm², but due to fabrication-process issues this remains at the pre-research stage;

2. Parallel cold-plate + immersion dual tracks: because cold-plate liquid cooling is mature, stable, reliable, and cost-controllable, it currently dominates. But as heat-flux density rises, immersion cooling is gradually developing. For some time to come, the two mainstream architectures—cold-plate and immersion—will coexist. As immersion cooling continues to iterate and optimize, its standards system matures, device interoperability becomes fully compatible, hardware architecture is deeply decoupled, and commercial scale keeps expanding, immersion cooling will become an indispensable mainstream cooling solution for high-density AI data centers, supporting large-scale deployment of GW-scale AI data centers, lowering the per-unit-compute token cost, and helping realize inclusive compute.

3. Full-stack intelligence: to improve the stability, reliability, and PUE of liquid-cooling systems, intelligent and fine-grained management is imperative. From full-parameter sensing of the CDU, building a CDU health-management system to predict CDU faults and parts maintenance, to introducing an AI group-control system for adaptive coolant-temperature regulation, load

forecasting, and fault self-healing, the liquid-cooling system can gain a smarter "brain" to achieve true "predictive cooling."

4. Extreme PUE: combining the liquid-cooling system with waste-heat recovery, using coolant waste heat for campus heating, domestic hot water, dehumidification, and other scenarios achieves comprehensive energy utilization and lowers the overall PUE of the AI data center campus.

Chapter 2 IT Facility Layer

2.1 AI Super Node

An AI super node is essentially an integrated super-server. Relying on high-speed buses such as UALink, Ethernet, and CLink, it vertically scales up AI accelerator modules such as GPUs; leveraging frontier technologies such as architecture decoupling and liquid cooling, it achieves extreme high-density integration of AI accelerator modules. By shortening the high-speed signal-transmission path between modules, it effectively increases interconnect bandwidth and reduces communication latency, fully adapting to scenarios such as MoE, PD separation, agents, long-sequence input/output, and multi-turn interaction.

The development of AI large models continuously drives substantial increases in the performance and power of AI accelerator modules, while the supply and cooling infrastructure of existing data centers varies widely. To adapt to the carrying capacity of existing facilities, the distributed super-node form has emerged, forming a complementary layout with the whole-rack super node to jointly ensure efficient, stable operation of upper-layer AI services.

The whole-rack super node uses the rack as the unit, suited to high-power-density, liquid-cooled data centers. The distributed super node uses the chassis as the unit and can be directly racked into already-deployed racks, suited to data center environments with limited power/cooling conditions or those built incrementally.

2.1.1 Whole-Rack Super Node

The novel whole-rack super-node architecture achieves high-density integration, high-speed interconnect, liquid cooling, and centralized power supply. It interconnects GPU chips via high-speed buses and switch chips, achieving interconnection of GPU chips at different scales—32, 64, 128 cards—within a single rack or across racks, overcoming the in-rack challenges of compute, memory, storage, power, and cooling, and solving the problems of resource pooling and integrated design.

2.1.1.1 Compute Node

The compute node is the core carrier of the CPU, GPU, memory, and in-rack high-speed Scale-up interconnect interface—the core compute carrier of the whole-rack super node. The extreme pursuit of performance and compute density brings dual challenges of high-power supply and cooling. In current mainstream industry compute-node solutions, centralized power supply and full-liquid-cooling cold-plate cooling are commonly used.

Taking the super node of Baidu, a leading Chinese internet vendor, as an example, its compute node integrates three functional modules—a CPU compute board, a PCIe Switch board, and a GPU board—in a 1:2:4 resource ratio, plus 4 smart NICs for expansion. The compute node has a 1U-height design, 21 inches wide and 1.2 m deep; the NICs, NVMe-SSDs, M.2, and other components use a front layout, while the Scale-up high-density connectors are placed at the rear.

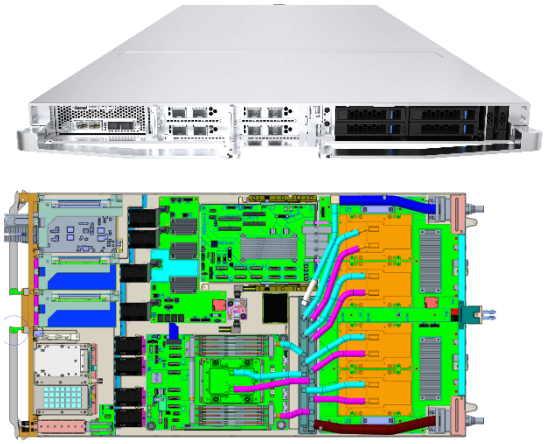


Figure 2.1 Baidu whole-rack super-node compute-node design example

Different AI application scenarios have markedly different CPU-to-GPU resource-ratio needs, leading to very different super-node architecture designs.

In AI training, the CPU mainly handles data loading, task dispatch, and coordinated scheduling of GPU compute, emphasizing low-latency, high-bandwidth interconnect rather than simply stacking cores for stronger compute. NVIDIA GB200 NVL72 uses 36 Grace CPUs of 72 cores each + 72 Blackwell GPUs, a 1:2 CPU-to-GPU ratio; the Vera Rubin super-node overall architecture continues the 1:2 CPU-to-GPU ratio (CPU core count raised to 88), representing the current mainstream architecture for training-scenario super nodes.

In AI inference, especially complex applications such as AI agents, as task complexity rises and multi-model scheduling increases, the CPU's load rises significantly in task orchestration, memory management, scheduling, and transferring data to GPU accelerators. A tightly coupled architecture with a fixed CPU-to-GPU ratio is prone to resource mismatch and compute waste, whereas a CPU-GPU decoupled, pooled architecture can adapt to a wider range of applications and significantly improve resource utilization, becoming an important direction for inference clusters. Alibaba Panjiu AI Infra 2.0 128 is a typical example of such an architecture. At GTC 2026, NVIDIA introduced the Groq 3 LPX LPU whole-rack solution, further enriching the compute-resource types of inference-scenario super nodes with more flexible ratios.

2.1.1.2 Rack

The rack is the physical carrier and structural core of the whole-rack super-node system, used to deploy compute nodes, power shelves, distributed CDUs, network nodes, the Scale-up high-speed network, the Scale-out interconnect network, etc., and to provide reliable fixing and structural support for the power busbars and liquid-cooling pipes, achieving precise interconnection and coordination among functional modules in the rack—the core physical framework by which the super node achieves compute aggregation. The marked improvements in structural strength, machining precision, and coupling fit with internal equipment are the core differences between a whole-rack super-node rack and a general-purpose ICT rack.

The rack spec of a whole-rack super node must match the overall super-node architecture, and the industry currently has multiple differentiated schemes. Among them, racks based on the directionally optimized OCP Open Rack V3.0 (ORV3), driven by NVIDIA, have a relatively mature industrial ecosystem; more scalable wide-body racks have also been launched, with OCP Open Rack Wide (ORW) being a widely applied

industry example, shown by Meta, AMD, and others at the 2025 OCP Global Summit. But Meta also noted in its technical sharing that larger racks mean higher design and manufacturing difficulty, higher transport cost, and greater O&M challenges, so the future direction may be multiple decoupled low-density ORV3 racks interconnected optically for higher scalability.

The super-node rack of Baidu, a leading Chinese internet vendor, inherits the Scorpio whole-rack server architecture concept, with a rack height of 46U, external dimensions of 2300mm (±2) × 600mm (0, −3) × 1200mm (±2), and a unit RU height of 46.5mm; the rack rated load is no less than 1600kg, supporting water, power, and network blind-mate interfaces for node blind-mate O&M. The liquid-cooling manifold uses a top/bottom inlet/return design, supporting whole-rack integrated delivery and significantly improving delivery and O&M efficiency.

Figure 2.2 Baidu whole-rack super-node rack design example

2.1.1.3 Whole-Rack Power Supply

At the power-access stage, North American data centers traditionally use 480Y/277VAC and 208VAC as the main input formats, while Chinese data centers commonly use 380Y/220VAC, 240VDC, and 336VDC. As AI super-node power density rises rapidly, higher-voltage DC input schemes such as ±400VDC and 800VDC will gradually mature and become widespread.

1. Power-shelf supply

The Power Shelf is the supply-conversion hub of the whole-rack super-node system. It uses a pooled architecture to centrally deploy the rack's power supply units (PSUs) and is equipped with a Power Management Controller (PMC) to monitor the operating status and health of PSUs and other supply components in real time. The Power Shelf connects to the input power via a PDU, steps down through the PSUs to 54VDC or 48VDC, and outputs to the rack busbar to power equipment uniformly. A directionally optimized Power Shelf can also provide several standard supply interfaces to power devices with other rated-voltage inputs.

As single-rack power rises rapidly, improving power density and supply-conversion efficiency has become the core design goal of the PSU. Applying new materials and

technologies—third-generation semiconductors (e.g., GaN devices), high-frequency low-loss magnetics, advanced power-conversion topologies (e.g., flying-capacitor four-level FC4L topology)—can effectively shrink device size, reduce hysteresis and eddy-current losses, lower device voltage stress, and significantly improve the device figure of merit, ultimately achieving higher power density and conversion efficiency.

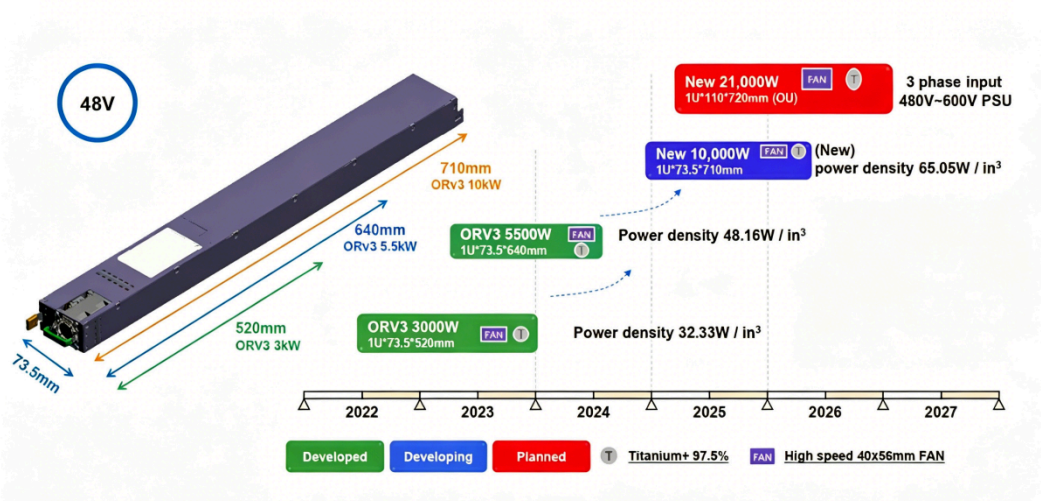


Figure 2.3 PSU power and size roadmap

An in-rack integrated Power Shelf squeezes the deployment space of compute and switch nodes. When single-rack power exceeds 200kW, continuing to use the built-in Power Shelf architecture would significantly reduce compute-deployment density and even affect the operating efficiency of the whole compute cluster. As single-rack power rises further, the model of co-locating supply units with compute and network equipment becomes unsustainable, and the power supply will have to be separated from the service rack, with the Sidecar standalone power-rack scheme gradually becoming mainstream for AI data centers.

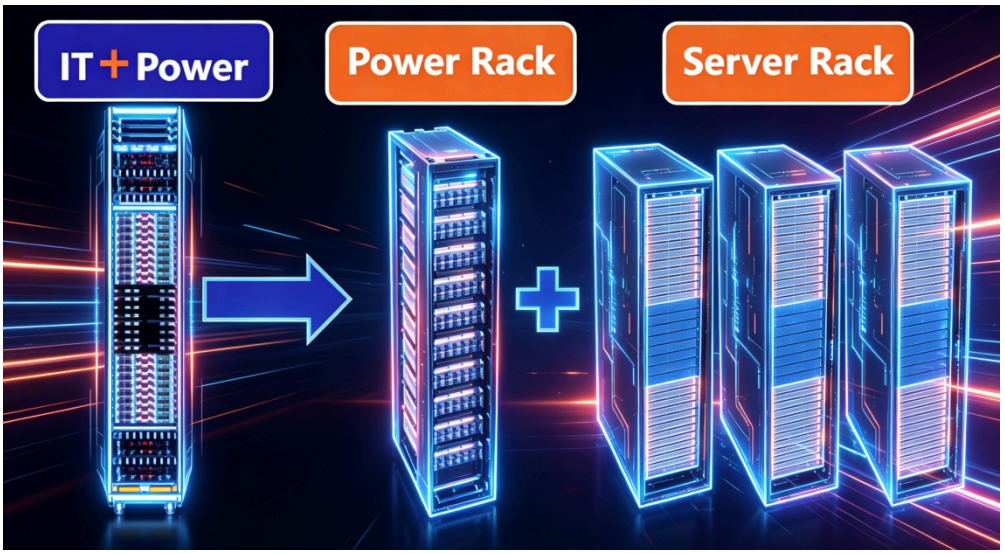


Figure 2.4 Whole-rack super-node power architecture shifting from co-rack to the Sidecar standalone power-rack scheme

2. Compute-node power supply

In current industry practice, compute and network nodes usually draw 54VDC or 48VDC directly from the rack busbar via a Power clip to power the GPU modules, while powering the in-node CPU and other components requires further step-down conversion to 12VDC. With the application of ±400VDC and 800VDC input schemes, the busbar

voltage will also be upgraded to ±400VDC or 800VDC, requiring DC conversion modules inside the compute and network nodes for step-down to output 54VDC, 48VDC, 12VDC, etc. to power the GPU modules and other in-node components. As chip power for GPUs, CPUs, etc. keeps climbing (single-GPU power has reached 2kW), supply-path design challenges grow, and more efficient new supply architectures such as vertical power delivery have emerged.

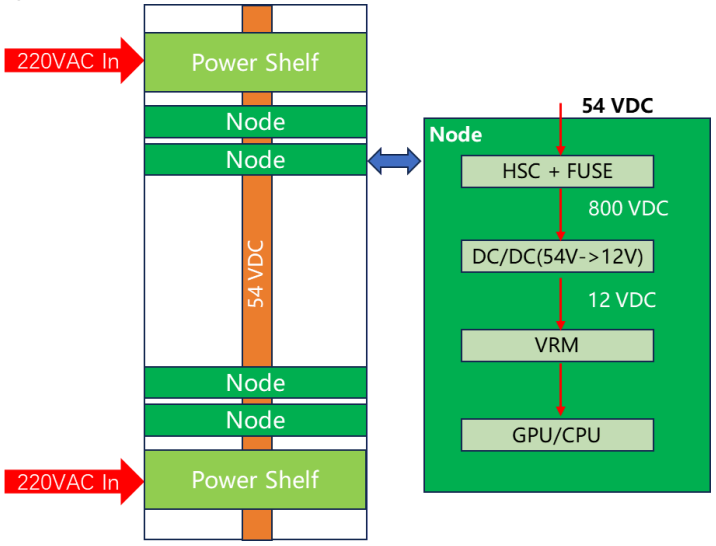


Figure 2.5 Current 220VAC whole-rack power topology

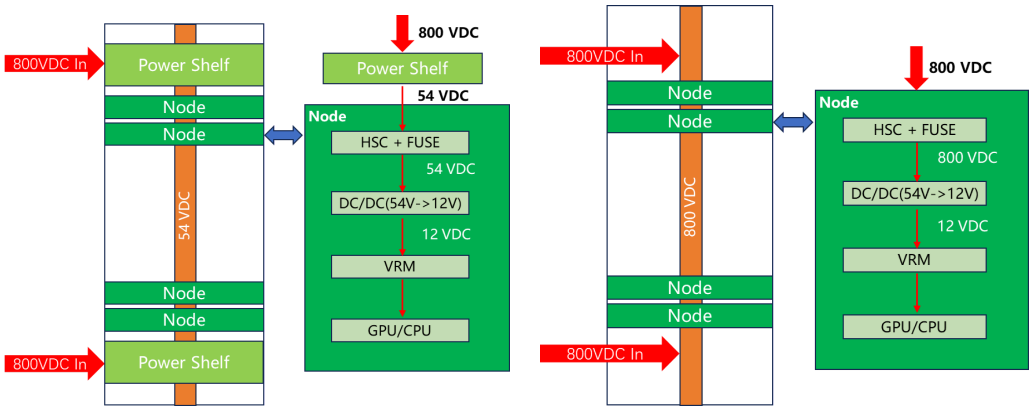


Figure 2.6 Future 800VDC power topology

2.1.1.4 Whole-Rack Cooling

At the whole-rack system level, rack-level cooling is moving from air cooling to air-liquid hybrid cooling and ultimately to full liquid cooling. As compute, transport, and storage capacity and power of compute, network, and storage nodes increase, maintaining excellent performance and system stability/reliability is a technical challenge the super node must overcome, and liquid cooling offers a better cooling solution for high-power-density scenarios.

To deploy extreme compute density, large cloud service providers (CSPs) and next-generation cloud data center users commonly adopt liquid-cooled whole-rack super-node solutions. For small/medium users with lower compute-density requirements or data centers not yet liquid-cooling-ready, air-cooling or air-liquid hybrid cooling can be selected.

Current liquid-cooled whole-rack super nodes commonly use the cold-plate scheme; the in-rack liquid-cooling components mainly include the manifold, cold plates, quick connectors, and the in-rack CDU.

The in-rack manifold is the piping system connecting the CDU to the node cold

plates. Based on the connection form of the quick connectors between manifold and cold plate, it is divided into manual-insert and blind-mate manifolds; blind-mate manifolds are usually equipped with anti-splash components and diversion grooves. For reliable assembly, the flatness of the manifold connector mounting surface must not exceed 0.1mm. To keep in-rack node cooling balanced, the manifold branch-flow deviation must not exceed 10%.

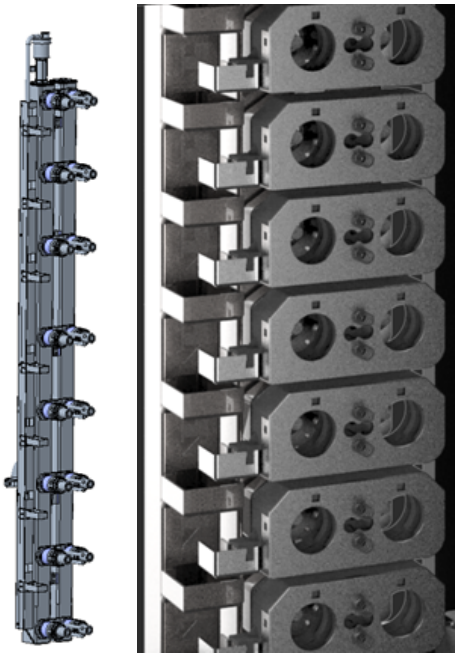


Figure 2.7 Schematic of the manifold and anti-splash components

Cold plates are located inside compute and switch nodes; their efficient design and fabrication process are crucial to the long-term reliable operation of high-power components such as GPU, CPU, and ASIC and to system energy saving. At the design level: optimize the fin structure to increase heat-dissipation area and enhance fluid turbulence; use a jet-cavity design to achieve direct impingement cooling of hotspot regions and improve local heat-dissipation intensity; introduce two-phase cold-plate technology and high-efficiency evaporative heat-exchange structures, using liquid phase change to absorb large latent heat and further strengthen heat dissipation. At the fabrication and inspection level: perform supplementary machining after welding to ensure flatness; perform 100% ultrasonic non-destructive testing on welds to eliminate defects; use nickel plating or passivation for long-term protection.

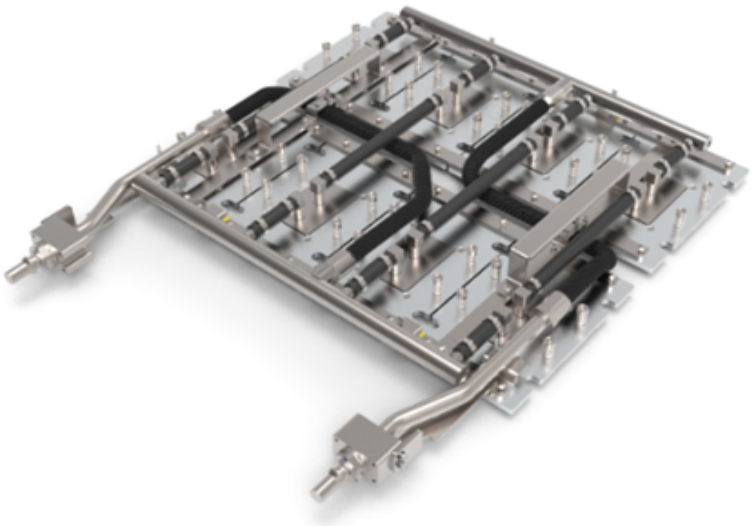


Figure 2.8 GPU cold-plate assembly

By mounting location, quick connectors are divided into in-rack mounting (on the manifold side) and in-server mounting (on the cold-plate side). By mating operation, they are divided into manual-insert and blind-mate quick connectors. Blind-mate quick connectors are widely used in super-node systems. To achieve reliable connection, blind-mate quick connectors must meet stringent radial, axial, and angular float requirements: radial float ≥±1.0mm, axial float ≥1.0mm, and with a floating module, angular float ≥±2.5°.

Figure 2.9 Liquid-cooling quick connector

Based on the overall CDU layout, whole-rack super-node liquid-cooling schemes fall into distributed cooling and centralized cooling architectures: the distributed scheme integrates the CDU and IT equipment in the same rack, better suited to small/medium deployments not pursuing extreme compute density; large CSPs and Neo cloud users commonly adopt the centralized scheme with standalone CDU racks, helping achieve higher compute density, simplify liquid-cooling piping, and greatly improve O&M efficiency.

Figure 2.10 Distributed-cooling whole-rack super node

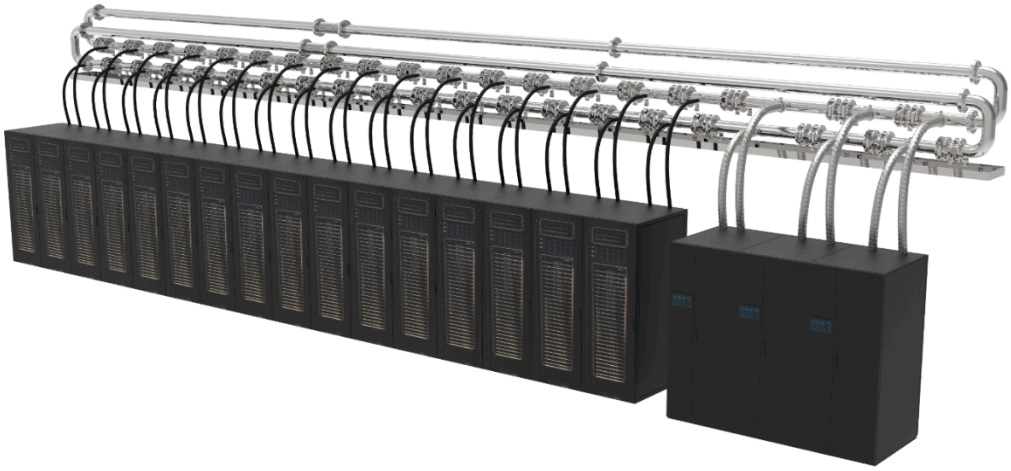


Figure 2.11 Centralized-cooling whole-rack super-node cluster

2.1.1.5 Scale-up Interconnect

Relying on a dedicated Scale-up interconnect protocol, the super node can build a globally unified address space, abstracting the memory and VRAM of all AI accelerator chips in the system into a unified storage resource pool; compute units can directly access other chips' storage resources with memory semantics, eliminating complex data copies and reducing end-to-end latency to the sub-microsecond level.

Super-node Scale-up interconnect must meet core needs such as high bandwidth, low latency, high reliability, and high density, and can be divided into two technical routes: copper interconnect and optical interconnect.

1. The copper-interconnect route uses copper conductors as the transmission medium, offering high reliability, low latency, a mature industrial chain, and controllable cost, currently dominating. Cable-tray interconnect and orthogonal board-to-board interconnect are the two core architectures.

- **Cable-tray interconnect**

Through continuous optimization of connectors and copper-cable processes, cable tray greatly improves link-loss margin, extending interconnect length to 2m. Its blind-mate self-guiding floating mechanism eliminates the high-speed signal degradation caused by incomplete mating in traditional hard connections, while supporting flexible interconnect-form conversion such as wire-to-board and wire-to-wire, offering high design flexibility and reliability. This scheme especially suits the architecture of horizontally laid out, vertically stacked compute and network nodes and is currently the most widely used in-rack interconnect scheme for whole-rack super nodes.

To ensure long-term reliable connection of the cable tray, the requirements on the floating mechanism and mating force must be quantified. Taking Baidu's super node as an example, its cable-tray quantitative metrics are:

i. Panel-side X/Y-axis float: within the specified pressure range, achieve ±3.0mm self-correcting guidance, with no jamming, locking, or rebound failure throughout;
ii. Panel-side Z-axis travel float (spring-compression direction): within the specified pressure range, achieve a maximum 3.5mm spring compression under node over-pressure, with no jamming, locking, or rebound failure throughout;
iii. Spring-force performance: the spring is the core component for X/Y/Z three-axis float and must stably meet elastic-performance requirements across different design pressure ranges.

- **Orthogonal board-to-board interconnect:**

Orthogonal board-to-board interconnect breaks the structural constraint of horizontally laid out, vertically stacked in-rack nodes, enabling front-to-back vertical mating of compute and network nodes (both compute and network nodes use a vertical layout for mating), further improving the compute-integration density of the super-node system.

Orthogonal midplane interconnect: suited to small-scale short-distance interconnect scenarios, typically used inside a GPU Tray and between adjacent GPU Trays and Switch Trays (interconnect distance usually 0.5m–0.7m), offering simple architecture, high integration, low cost, and easy maintenance, and widely used in distributed super-node systems.

Orthogonal Direct interconnect: relying on low-loss PCB materials, high-speed-connector material upgrades and process iteration, plus architectural optimization, it completely eliminates mid-backplane signal loss and greatly extends high-speed signal transmission distance. Its high-speed channels are of two types—PCB and copper cable: copper-cable interconnect distance reaches 1m–1.3m at 112G PAM4 and 0.4m–0.8m at 224G PAM4; PCB-trace interconnect distance is slightly lower than copper cable. Orthogonal Direct especially suits ultra-high-integration-density whole-rack super-node architectures and is an important technical route for next-generation Scale-up high-speed interconnect.

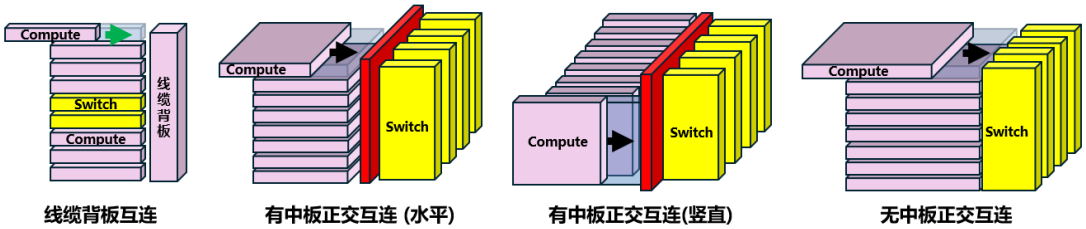


Figure 2.12 Various orthogonal board-to-board interconnect schemes

As super-node signal-transmission rates keep rising, to solve the in-chassis interconnect and transmission bottleneck of signals above 112Gbps, NPC and CPC

become the optimal solutions:

NPC (Near Package Copper): oriented to rack-level long-distance interconnect, centered on loss optimization and distance extension, supporting 112G/224G rates with a transmission distance of 1.5m–2m;

CPC (Co-Packaged Copper): oriented to package-level ultra-high-density interconnect, centered on rate improvement and energy reduction, bringing 224G/448G signals inside the package with a transmission distance of 1m.

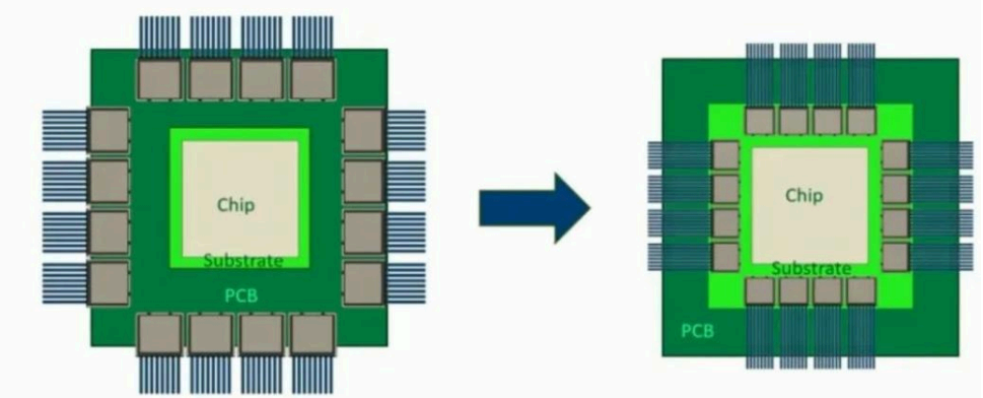


Figure 2.13 NPC and CPC interconnect schemes

2. The optical-interconnect route uses optical fiber as the transmission medium, using optical signals for long-distance, ultra-high-bandwidth data transmission—an important technical route to break the in-rack and inter-rack interconnect bottleneck of super nodes.

LPO (Linear Pluggable Optics) keeps the traditional pluggable-module form; by removing the built-in DSP and shifting complex functions such as equalization and clock recovery to the SerDes (serializer/deserializer) chip on the ASIC side, it becomes an important transitional form of a short-distance, low-power pluggable solution.

Current hot technologies in super-node optical interconnect are NPO (Near Packaged Optics) and CPO (Co-packaged Optics). Both remove the built-in DSP from the optical module and integrate it closer to the ASIC, significantly reducing optical-module power (single 400G-port power from about 8W to 3W), lowering single-end latency by about 100ns, and significantly improving port density and system reliability. The NPO scheme, being more mature and open industrially, currently gains more support from system vendors and users.

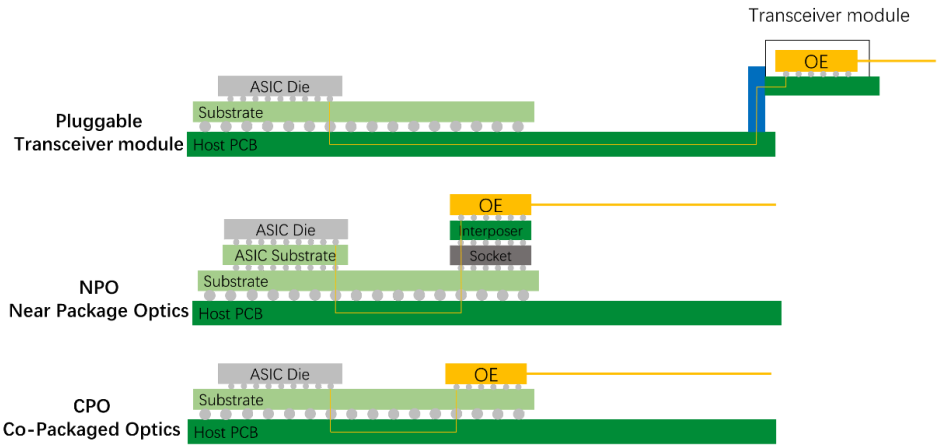


Figure 2.14 Link diagrams of pluggable optical module, near-packaged optics, and co-packaged optics

Since the second half of 2025, Chinese vendors have intensively explored NPO technology and product layouts: Tencent led the ETX Ultra 512-card super-node solution,

using NPO modules to achieve full-optical interconnect inside the super node; Alibaba Cloud first released an NPO module and a 102.4T NPO Switch concept prototype at the September 2025 Apsara Conference, and was the first to apply 3.2T NPO technology in its new-generation domestic four-chip switch.

2.1.2 Distributed Super Node

The distributed super node is a system form that efficiently aggregates small-scale compute chips within a single chassis through a tightly coupled architecture and high-speed protocol coordination. It has a smaller unit scale and finer compute granularity and can scale flexibly to business needs, suited to small/medium deployments or existing data centers with limited power density.

2.1.2.1 Modules

1. AI accelerator module

In physical form, the distributed super-node AI accelerator module can be divided into two types: PCIe add-in-card and board-mounted integrated.

PCIe add-in-card AI accelerator module: usually uses PCIe as the card-to-card interconnect bus, with high standardization but relatively limited cooling and power ceilings, suited to compute scenarios with moderate per-module compute requirements.

Board-mounted integrated AI accelerator module: uses open-standard or vendor-defined interfaces and protocols, breaking the interface and size limits of the PCIe add-in-card module, supporting higher power input and system cooling with stronger AI compute—currently the mainstream form of AI accelerator module.

Figure 2.15 Board-mounted integrated AI accelerator module design example

Taking Rubin/Rubin Ultra as the technical reference, for training and high-density-inference GPU integrated modules, power will rise further to 3600W in 2027, with significantly higher volume, heat-flux density, and supply demand, posing systemic challenges to cold-plate/immersion cooling, supply regulation, rack cooling, and structural design.

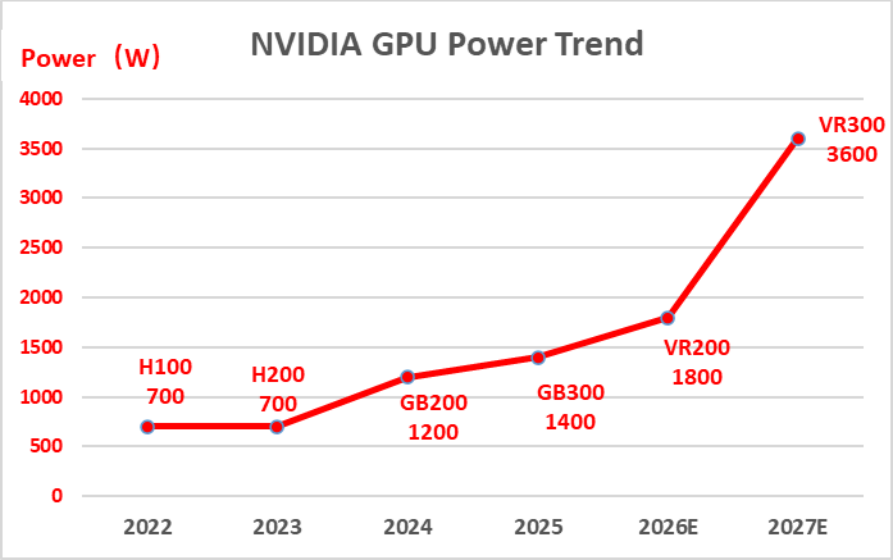


Figure 2.16 NVIDIA GPU power trend

2. AI carrier-board module

The AI carrier-board module is the core substrate unit carrying AI accelerator modules such as GPUs. A single carrier board can mount multiple AI accelerator modules interconnected via bus protocols, forming a basic AI-acceleration resource-pool unit. The carrier board provides external high-speed buses and expansion network interfaces, supporting multi-board Scale-up or Scale-out expansion.

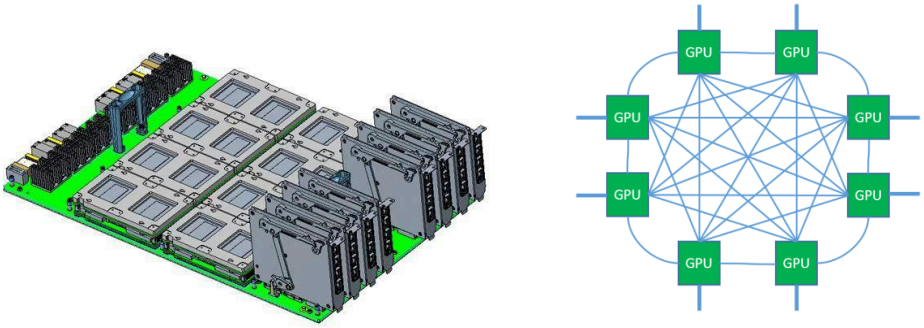


Figure 2.17 AI accelerator carrier-board module with expansion cards and its bus topology

As the power, size, and bus rate of AI accelerator modules such as GPUs grow rapidly, the design of AI carrier-board modules also faces more complex architecture, topology, and engineering-implementation challenges.

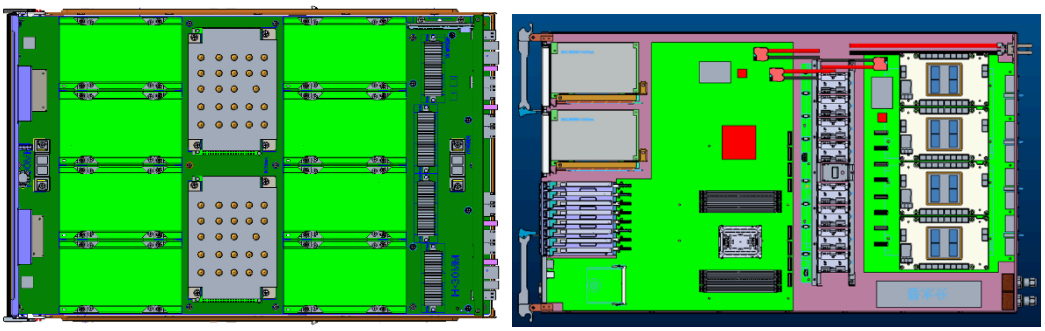


Figure 2.18 Design comparison of 8×GPU and 4×GPU carrier boards

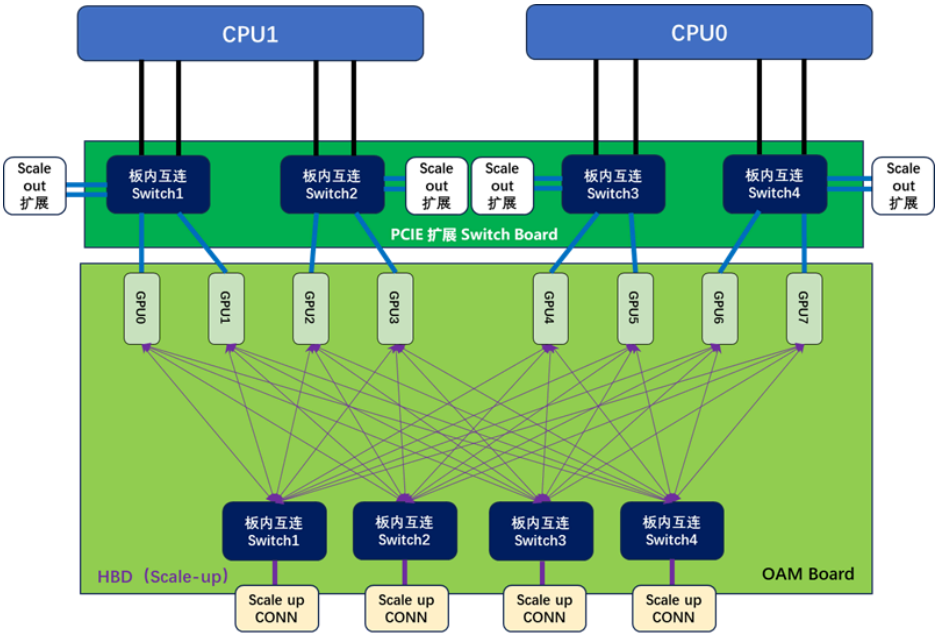


Figure 2.19 8×GPU carrier-board system topology

2.1.2.2 Topology

1. Cluster-layer topology

At the cluster layer, distributed super nodes are networked via IB or Ethernet protocols in network architectures such as fat-tree, allowing flexible configuration of NICs and the data center network topology.

A typical cluster topology is a three-tier CLOS/Fat-Tree.

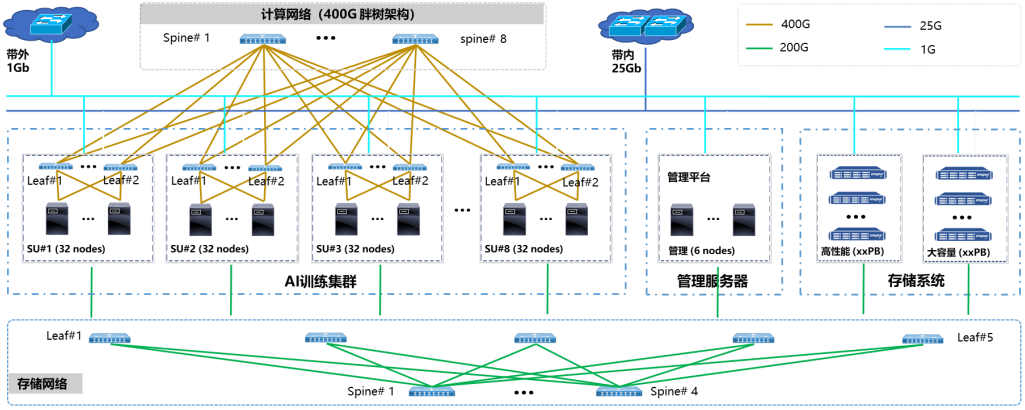


Figure 2.20 Example of a typical AI cluster networking topology

2. Rack-layer topology

There are two mainstream expansion schemes at the rack layer:

- Integrating compute, storage, and accelerator nodes within a rack—more flexible deployment, suited to small/medium-scale cluster scenarios;
- Integrating only a single type of node within a rack, then expanding the cluster by rack as the unit—suited to large-scale cluster scenarios.

For the scheme integrating multiple node types within a single rack, a bus architecture should be used to achieve interoperable access among various CPU compute nodes, GPU nodes (typical scenarios such as PD separation), and storage nodes, using one or more open-standard interconnect protocols including Ethernet, OISA, UALink, and CLink.

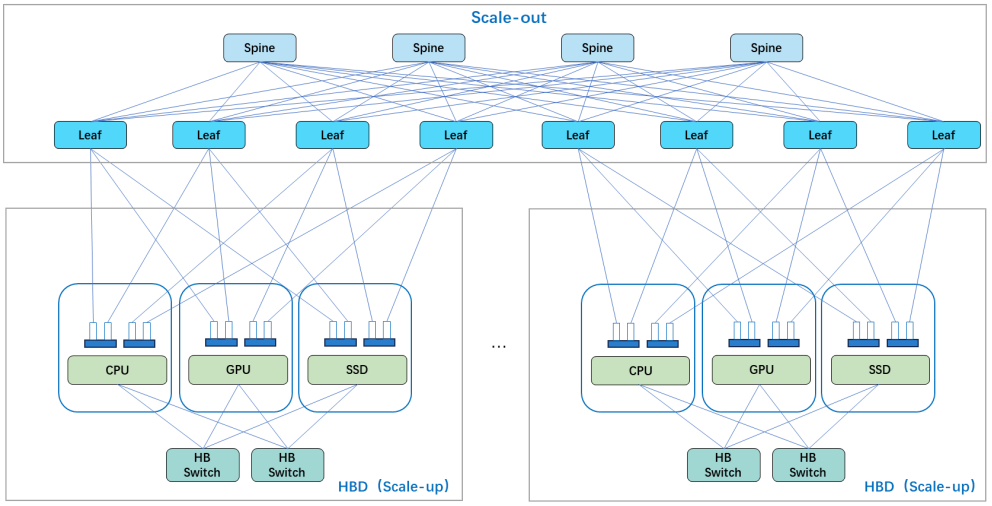


Figure 2.21 Cluster topology integrating compute, storage, and AI accelerator nodes within a single rack

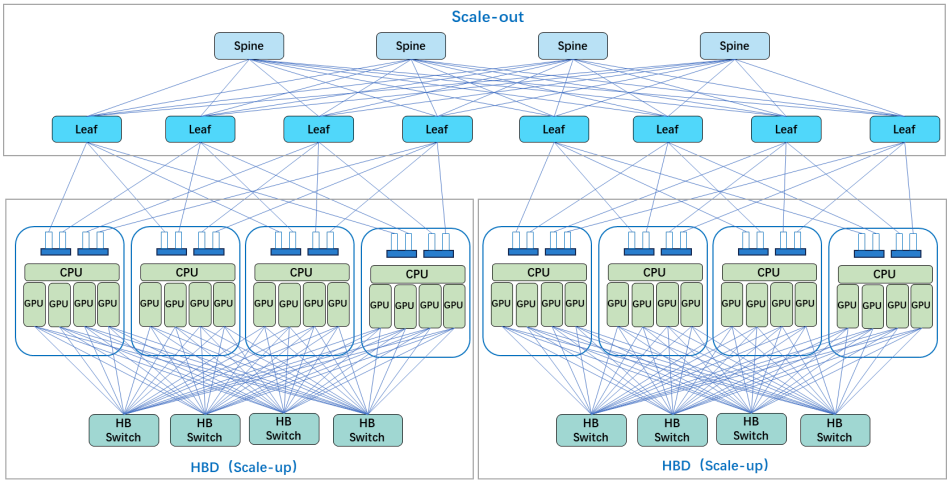


Figure 2.22 Cluster topology integrating only AI accelerator nodes within a single rack

3. Node-layer topology

It includes multiple physical topologies, supporting more flexible CPU, memory, and GPU ratios.

- Traditional CEM standard card: the CPU connects directly to GPUs via the PCIe bus, and GPUs interconnect within the node via a PCIe switch. This scheme has a simple topology, high reliability, and high flexibility, but limited intra- and inter-node interconnect bandwidth;
- OAM card: dominated by direct-connect topologies under the OCP UBB spec (Full-mesh, Hybrid Cubic). In the future, an on-board switch chip will be added for programming friendliness and further performance gains;

As in the following NVIDIA BFX NVL16 topology example, the GPU connects upward directly to the CPU and downward via a Switch for card-to-card interconnect, effectively combining the flexibility of the CEM standard card with the high-interconnect-bandwidth performance advantage of the NVSwitch

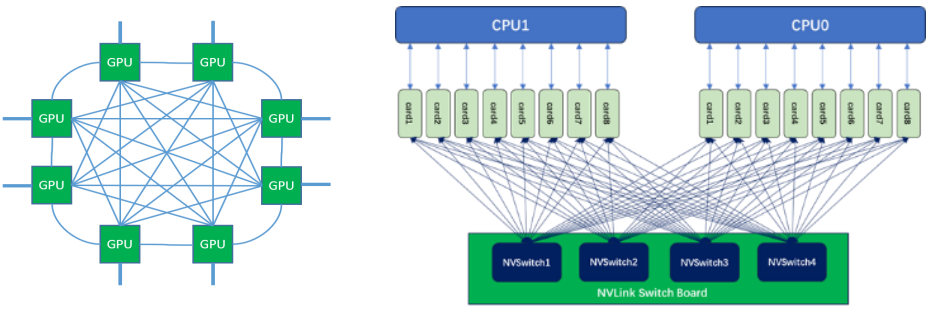


Figure 2.23 Full-mesh interconnect topology and NVL16 interconnect topology

2.1.2.3 Power Supply

The distributed super node integrates power modules. As GPU-module power and deployment density rise, single-system power has reached tens of kW and single-PSU output has exceeded 5kW. To match the internal multi-carrier-board vertically stacked architecture, the system commonly uses a busbar unified-supply scheme.

2.1.2.4 Cooling

Thanks to the small granularity and flexible networking/deployment of the distributed super node, a certain number of distributed super nodes and a distributed CDU can be integrated in the same rack based on infrastructure conditions, forming a standard rack unit, and then the cluster can be expanded on the basis of standard rack units.

When the user's deployment scale is large, a standalone CDU-rack scheme can be used for intensive deployment.

2.1.2.5 IO Expansion Card

To further improve the expansion flexibility of distributed super nodes and GPU carrier-board modules, a standard Scale-up expansion card can be developed. The Scale-up expansion card often needs to integrate functional modules such as signal recovery and signal amplification to achieve reliable, efficient interconnect of GPU modules between distributed super nodes or between carrier-board modules.

2.2 High-Speed Interconnect Physical Layer and Connectors

2.2.1 High-Speed Interconnect Network of the Gigawatt-Scale AI Data Center

A super node is a unified compute unit composed of multiple GPU cards interconnected point-to-point; its core is achieving high-bandwidth, low-latency intra-node communication via Scale-up technology to solve the communication bottleneck of AI computing. With Scale-up optimizing multi-card coordination within a single node and Scale-out supporting expansion between super nodes, flexible networking of 10,000-card-scale clusters is achieved. Scale-up high-speed interconnect hardware applications include: compute-node high-speed interconnect, switch-node high-speed interconnect, in-rack node-to-node high-speed interconnect, and rack-to-rack high-speed interconnect.

2.2.1.1 Super-Node High-Speed Interconnect Hardware Architecture

The larger-scale XPU compute integrated by super nodes, together with the high-bandwidth, low-loss goals driven by explosive compute demand, jointly drive the transformation of high-speed interconnect hardware architecture. Super nodes must comprehensively consider compute capacity, efficiency, cooling, and scalability, with the architecture gradually evolving toward cable-tray, orthogonal-midplane, orthogonal-direct, and NPO/CPO architectures.

Cable-tray architecture: nodes slide in horizontally from the front panel, with the ends interconnecting to the backplane. It is currently the most widely used scheme in the industry, e.g., NVIDIA NVL72–144, ByteDance Dayu, Baidu Tianchi.

Orthogonal-midplane architecture: compute and switch nodes are loaded vertically or horizontally from the front/rear panels and interconnect via a mid-backplane, e.g., NVIDIA Rubin Ultra NV144, Sugon Scale X640.

Orthogonal-direct architecture: compute and switch nodes mate directly and vertically at the front/rear panels, e.g., Huawei CloudMatrix 384, Alibaba Panjiu AL128, Baidu Tianchi.

2.2.1.2 Compute-Node High-Speed Interconnect

The compute node is the core compute carrier of the whole-rack super node, with multiple XPU cards interconnected within a single node (4 to 16 cards per node is common in the industry). As AI large models pursue extreme performance and compute density, copper-based high-speed interconnect faces challenges such as signal attenuation and rising power. The industry has proposed technical directions such as high-density large-channel cable assemblies, near-package copper (NPC), and near-packaged optics (NPO), while driving rate upgrades of standard (e.g., PCIe protocol) high-speed cables, jointly working toward core goals such as signal-loss optimization and transmission-distance extension.

2.2.1.3 Switch-Node High-Speed Interconnect

The switch node is the core interconnect hub of the whole-rack super node. As communication needs explode for inter-node-group XPU card interconnect (16–640 cards per super node), large numbers of compute units and switch chips must achieve high-density interconnect in limited space.

In gigawatt-scale AI data centers, the extreme pursuit of communication latency and network bandwidth by AI large models has driven rapid development of Ethernet switch chips: rates rising from 56Gbps/lane toward 224Gbps/lane & 448Gbps/lane. Under this high-speed network evolution, high-speed interconnect faces multiple challenges such as signal attenuation, rising power, and limited rear-window panel space. To this end, the industry has proposed a series of technical paths—high-density large-channel cable assemblies, near-package copper (NPC), near-packaged optics (NPO), co-packaged copper (CPC), and co-packaged optics (CPO)—and, through rapid practice and iteration, drives the coordinated evolution of RNICs and new interconnect schemes to jointly ensure stable, efficient operation of large-scale AI networks under high-throughput scenarios.

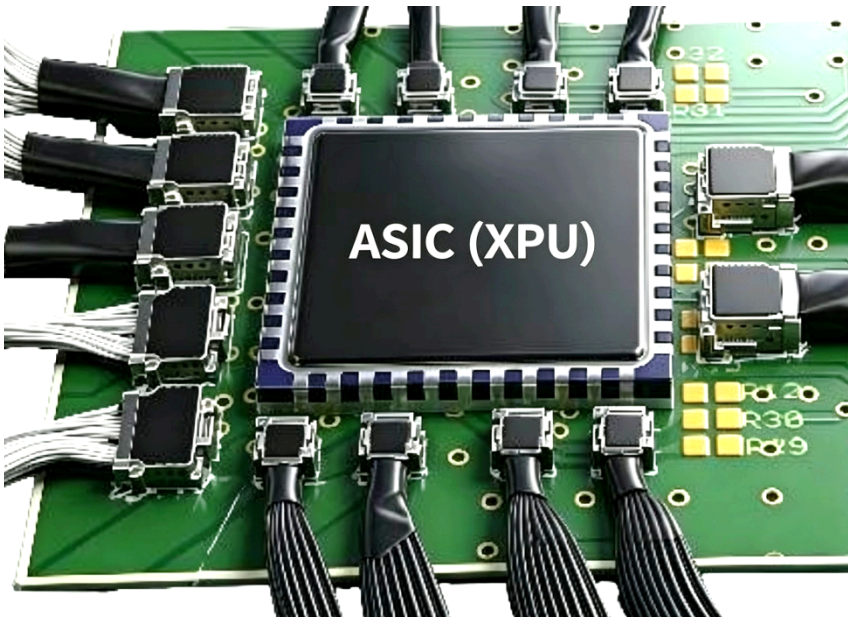


Figure 2.17 Schematic of an NPC product application for node high-speed interconnect

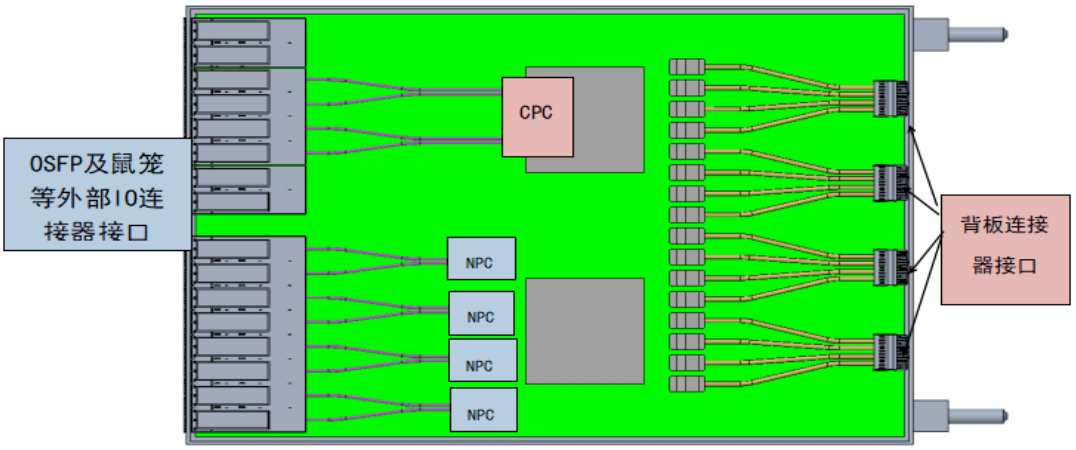


Figure 2.18 Schematic of switch-node high-speed interconnect application

2.2.1.4 In-Rack Node-to-Node High-Speed Interconnect

In the application scenario of gigawatt-scale AI data center racks, the whole rack is evolving from the traditional PCB-backplane architecture toward cable-tray, orthogonal-midplane, orthogonal-direct, and similar architectures.

1. Cable-tray architecture

The cable-tray architecture uses high-speed copper-cable modules to replace the traditional PCB backplane, building direct connections between compute and switch nodes to form a flexible mesh topology. This scheme breaks the physical-size limit of the backplane, supports high-density, multi-direction expansion, and offers advantages such as low cost, high reliability ensured by floating mechanisms, and multi-XPU compatibility, with bandwidth upgrades via rapid cable replacement. Its open layout also conveniently adapts to various cooling needs such as air cooling (optimized air ducts) and liquid cooling (with cold plates and manifolds), achieving efficient cross-node interconnect deployment inside the rack.

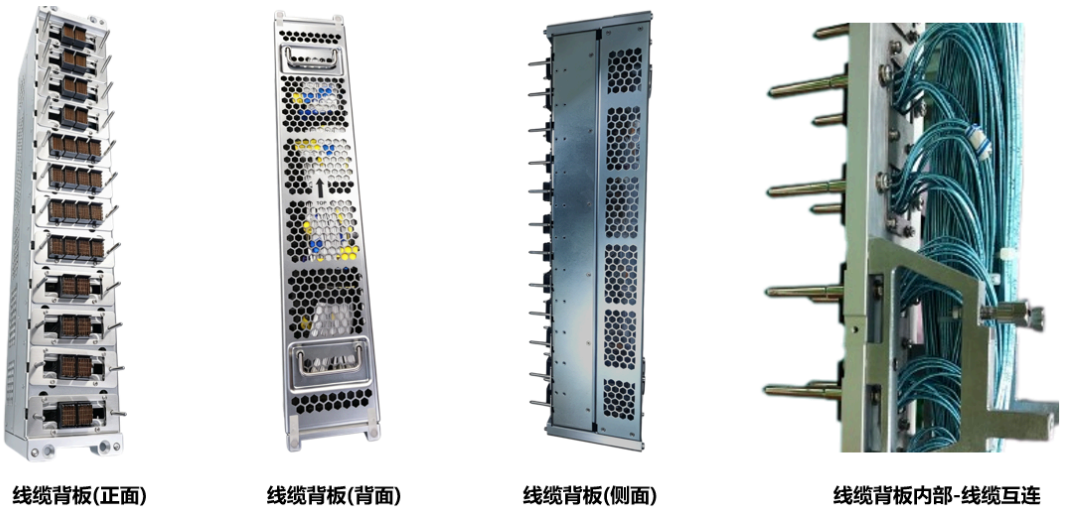


Figure 2.19 Schematic of a cable-tray product

2. Orthogonal-midplane architecture

The orthogonal-midplane architecture plugs compute and switch nodes onto a common mid-backplane at a 90° vertical orthogonal angle, building a non-blocking rectangular topology. With extremely short signal links, it achieves ultra-low latency and excellent signal integrity; with a stable connector design, the connector can support both PCB board-level schemes and cable schemes, supporting front/rear blind-mate maintenance and efficient straight-through airflow cooling, while offering high density, high reliability, and good cross-generation compatibility—though it correspondingly introduces new challenges in architecture and hardware design.

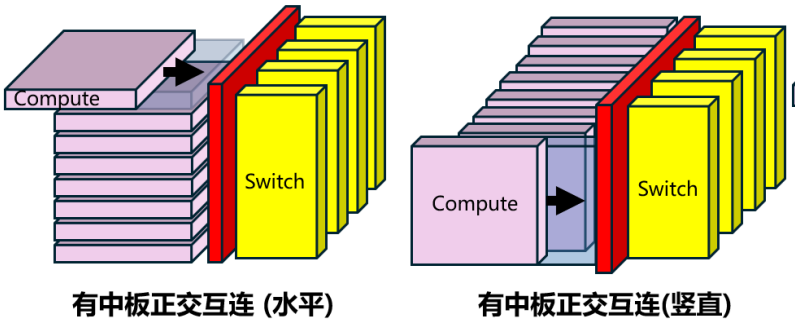


Figure 2.20 Schematic of the orthogonal-midplane architecture

3. Orthogonal-direct architecture

The orthogonal-direct architecture eliminates the common mid-backplane, using a 90° vertical orthogonal direct-connect hardware structure between compute and switch nodes. This design breaks the line-rate bottleneck of the traditional backplane, achieving ultra-large switching capacity and high scalability while maintaining extremely low signal attenuation. However, this highly integrated direct-connect mode also poses greater engineering challenges to mechanism precision, interconnect stability, cooling/liquid-cooling planning, power supply, and hardware design.

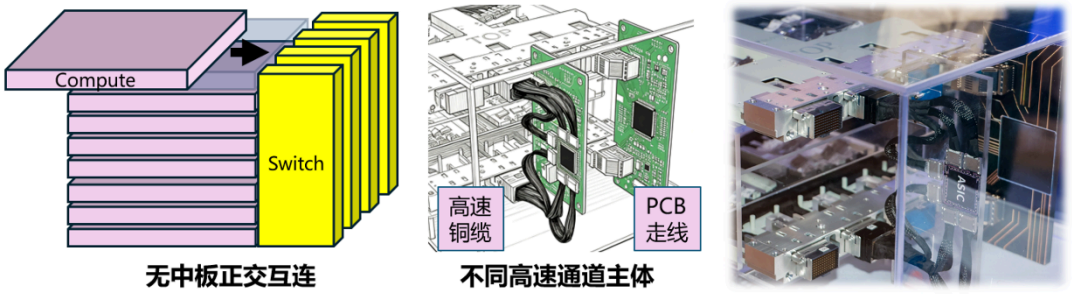


Figure 2.21 Direct orthogonal + PCB architecture interconnect scheme

2.2.1.5 Rack-to-Rack Interconnect

Inter-device interconnect schemes fall mainly into two technical routes: copper and optical. Copper transmission schemes are mainly DAC (Direct Attach Cable), ACC (Active Copper Cable), AEC (Active Electrical Cable), and backplane-connector direct-attach copper cables. As data center bandwidth evolves from 100G/400G toward 800G/1.6T, copper interconnect is upgrading from traditional DAC to ACC and AEC with stronger signal compensation. ACC suits medium-distance scenarios such as intra-rack, while AEC better suits long-distance interconnect needs such as cross-rack. In optical interconnect, AOC and pluggable optical modules are the core schemes, with their evolution focused on system integration with higher bandwidth, lower power, smaller size, and better cost.

2.2.2 Interconnect Technology Implementation Paths

For gigawatt-scale AI data centers, high-speed interconnect faces enormous challenges. SerDes technology has evolved from 28Gbps to 224Gbps; 56G/112Gbps is now widely used and is strongly driving network-architecture innovation. Therefore, systematically reviewing the interconnect technology paths across all rate segments is key to ensuring smooth system upgrades.

2.2.2.1 Termination Forms: Press-fit/SMT/BGA/LGA

To meet rising transmission rates, connector termination forms have evolved: the finished hole of Press-fit (press-fit pins) keeps shrinking, gradually toward 0.25 mm; the pitch of SMT, BGA, and LGA continues to densify toward <0.6 mm to solve impedance-consistency issues under high-bandwidth scenarios.

Table 2.1 Comparison of termination-form application information

Termination form	Typical spec range	Advantages	Disadvantages
Press-fit	56Gbps recommended finished hole diameter: 0.31–0.45 mm	High mechanical reliability, easy insertion/removal, convenient repair and rework;	Hole diameter is limited by PCB stack-up thickness, stamping, and other processes, with limited spatial density;
	112Gbps recommended finished hole diameter: 0.29–0.36 mm		
	224Gbps recommended finished hole diameter: 0.25–0.36 mm		
SMT	Terminal pitch: 0.5–0.6 mm	Higher density and miniaturization, better than Press-fit	Difficult repair and rework;
BGA	Terminal pitch: 0.4–0.6 mm	Higher density and miniaturization	Difficult repair and rework;
LGA	Terminal pitch: 0.4–0.6 mm	Higher density and miniaturization	Slightly lower mechanical reliability than Press-fit;

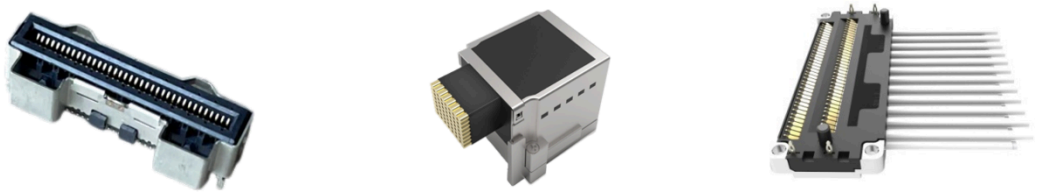
2.2.2.2 High-Speed Connectors

The main interconnect form between a single node's end panel and the mating node is the high-speed backplane connector. As interconnect applications diverge, one can choose board-level or cable types and extend different series of 56Gbps–224Gbps high-speed connectors. [Sections 2.2.2.2–2.2.2.4 — detailed connector-selection recommendation panorama — refer to the product table in the "Gigawatt-Scale AI Data Center High-Speed Interconnect Technology White Paper"]

Figure 2.22 Schematic of a high-speed backplane connector and cable assembly product

2.2.2.3 Internal High-Speed I/O

Internal high-speed I/O achieves high-speed interconnect within a single node. As interconnect protocols diverge, one can choose standard connectors (MCIO, Gen-Z, etc.) or custom connectors (NPC/CPC) and adapt to the high-speed characteristics of different protocols, such as PCIe 5.0–PCIe 7.0 (per-lane 32GT/s–128GT/s), with SerDes rates covering 56Gbps–224Gbps.



高速I/O连接器（标准类型） 高速I/O连接器— NPC（定制类型） 某 LGA 形态的 NPC（定制类型）

Figure 2.23 Schematic of a high-speed I/O connector and custom connector product

2.2.2.4 External High-Speed I/O

External high-speed I/O achieves high-speed interconnect across nodes and racks. As interconnect protocols diverge, one can choose standard cables (QSFP, OSFP-XD, etc.) or custom backplane-connector cables and adapt to the high-speed characteristics of different protocols. In terms of product cable length, passive DAC cables are mostly 0.5–2.0 m, active ACC cables mostly 2.0–4.0 m, and active AEC cables mostly 4.0–7.0 m.

Figure 2.24 Schematic of an external high-speed I/O cable product

2.2.2.5 High-Speed PCB Materials

The loss requirements for high-speed PCB materials, by signal-link needs at different rates, are as follows:

Table 2.2 High-speed PCB material application information

Signal rate	Dielectric loss (Df)	Non-domestic materials	Domestic materials
56Gbps	0.003–0.005	MEGTRON6 (G), EM891, TU-883, IT-968, etc.	NY6300S, H360Z, S6, etc.
112Gbps	0.002–0.003	MEGTRON7 (N), EM890K, TU-933, IT988GSE, etc.	NY-P2, S6NX, etc.
112Gbps	0.0015–0.002	MEGTRON8 (N), EM892K, TU943SN, IT998G, etc.	NY-P4N, S9N, etc.
112Gbps 224Gbps	0.0015–0.002	MEGTRON8 (N), EM892K, TU943SN, IT998G, etc.	NY-P4N, S9N, etc.
224Gbps	0.0001–0.0015	MEGTRONS (U), EM892K2, IT998GSE, etc.	NY-P4, S9N2, etc.
224Gbps	0.00005–0.0001	MEGTRON9 (U/Q), EM896K2/K3, IT999GSE2/3, etc.	NY-P5N2/Q, S10GN2/Q, etc.

2.2.2.6 High-Speed Raw Cable

Higher-rate high-speed raw cables require smaller size and better SDD21 loss. To this end, designs often use insulation with lower dielectric constant (Dk) and loss factor (Df), or reduce loss by enhancing coupling between conductors.

Solid insulation cannot reach below 2.0 due to material Dk limits; conventional PE/PP/FEP materials have Dk values of 2.03–2.3. To reduce loss, the insulation design of high-speed raw cables is optimized from single-layer insulation to double-layer insulation, foamed insulation, or twin-core co-extrusion structures.

Typical SI metrics are also increasingly stringent, for example: center impedance matching the connector, with tolerance tightened from ±5Ω to ±2Ω; intra-pair skew reduced from 10 ps/m to 2 ps/m; a smooth SDD21 curve with bandwidth rising to 40 GHz, 67 GHz, even 110 GHz without glitches.

The table below provides only the SI performance of a typical ground-wire-free high-speed copper cable; in practice, different structures such as dual-ground-wire or single-ground-wire can be selected as needed.

Table 2.3 High-speed Raw Cable application information

AWG	Conductor OD	Insulation structure	Ground wire	Wrap-tape size (mm)	Impedance /ohm	112Gbps	224Gbps	448Gbps	
						SDD21 @26.56GHz	SDD21 @53.12GHz	SDD21 @84GHz	SDD21 @110GHz
32	0.203SC	FEP twin-core	None	1.55×0.89	92	7.9–8.4 dB	11.5–12.0 dB	16.0 dB	19.5 dB

		co-extrusion							
30	0.285SC	FEP+FEP	None	1.70×1.20	92	6.1 dB	8.8 dB	12.3 dB	14.5 dB
28	0.35SC	Foam FEP+PE	None	1.99×1.18	92	5.4–5.6 dB	8.3 dB	10.2 dB	11.9 dB
26	0.394SC	Foam FEP+PE	None	2.10×1.14	92	4.7–5.1 dB	8.0 dB	9.7 dB	11.3 dB
25	0.455SC	Foam FEP+PE	None	2.55×1.32	92	4.1–4.4 dB	7.3–7.7 dB	8.4 dB	9.8 dB

2.2.3 Outlook on Next-Generation Interconnect Technology Paths

As AI large models and 10,000-card clusters explode, gigawatt-scale AI data center interconnect faces rigid demands for high bandwidth, low latency, and low cost, driving per-lane rates toward 448Gbps. Current interconnect technology is breaking through in parallel along the copper and optical paths.

Copper interconnect is undergoing system-level innovation: from coding/modulation (PAM4/6/8) to architecture (evolving toward various orthogonal forms) to product-structure upgrades. The core focus, CPC (co-packaged copper), through integrated “chip–connector–copper-cable” design, offers multiple advantages over traditional schemes—lower link loss, lower power, lower system cost, and both high density and maintainability. Meanwhile, the evolution of PCB and Raw Cable, with the introduction of new materials such as high-frequency board materials and foamed insulation, jointly helps copper interconnect keep breaking the distance boundary and overcome the high-speed-interconnect challenges at 110GHz and above.

Optical interconnect evolves from pluggable optical modules toward NPO/CPO. CPO co-packages the optical engine with the electrical chip so that the electrical signal travels only a few millimeters, greatly improving bandwidth density and reducing bit-error rate—key to breaking the panel IO-density limit. However, constrained by the industry ecosystem, it is still in the exploratory stage, the overall industrial chain is not yet complete, and upstream/downstream partners must jointly mature its productization and deployment.

In the future, “optical–copper coordination” will be the long-term landscape: copper interconnect holding the short-distance in-rack scenarios and optical interconnect handling long-distance cross-rack communication, complementing each other to build a high-efficiency compute foundation.

2.3 AI Network and Switching System

2.3.1 Scale Expansion

AI compute Scale-Out networks are evolving toward larger-scale cluster architectures, driven mainly by two factors.

1. Technology-demand driven: advanced large models represented by GPT-5.5, Claude, and DeepSeek v4 have verified that a model’s intelligence is directly and positively correlated with its parameter scale, training-data volume, and total compute. To pursue stronger reasoning, generalization, and multimodal capabilities, ultra-large clusters of more GPUs must be built, placing extremely high demands on the underlying network’s Scale-Out capability.

2. Economic-efficiency driven: in fierce global AI competition, shortening the model-training cycle has become a core competitive advantage. Larger GPU clusters can

significantly accelerate training jobs, and the resulting efficiency gains directly lower the operating cost per unit of compute, achieving dual optimization of efficiency and cost.

2.3.1.1 Rapid Evolution of Switch-Chip Bandwidth Drives a Generational Leap in Compute Networks

The coordinated development of AI data center switch chips and networking architectures is the cornerstone supporting the evolution of 10,000-card clusters. Its evolution clearly reflects the shift from “using architectural complexity to compensate for insufficient bandwidth” to “using chip performance to drive architectural simplification.”

Early architecture-compensation stage: limited by single-chip switching bandwidth, the industry had to use complex architectures such as chassis-box networking and three-tier box-box networking to meet early 10,000-card networking needs.

Current chip-driven simplification stage: as switch-chip performance rises, single-chip bandwidth has moved from 12.8T toward the 51.2T and even 102.4T era. This makes simple, flat two-tier box-box networking mainstream, significantly reducing management complexity and cost.

The core driver of this shift comes from continuous process and architecture breakthroughs by chip suppliers represented by Broadcom. By adopting advanced processes such as 5nm and 3nm and iteratively releasing 51.2T to 102.4T chip series, they provide higher integration and bandwidth, laying a solid hardware foundation for building simple, efficient compute networks.

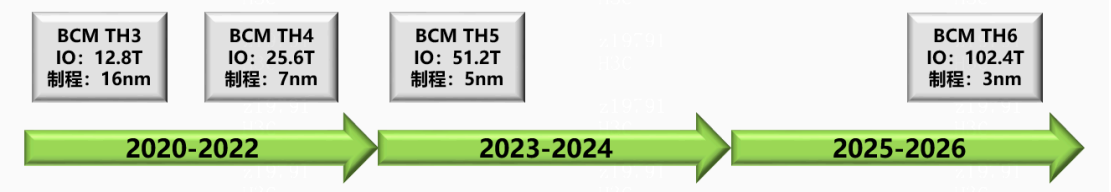


Figure 2.25 Broadcom compute-chip roadmap

To match the rigid demand for switch-chip capacity evolving from 51.2T toward 102.4T and beyond, the per-lane interface rate and signal-modulation technology are undergoing a key generational leap.

Port-rate and SerDes technology trends: ports are still mainly 400G at present, with 800G about to surge. Affected by chip process and ecosystem maturity, technology platforms based on 112G SerDes will persist domestically for a considerable time. As 224G SerDes gradually lands, NPO (near-packaged optics) modules are expected to deploy at scale from 2027, while CPO (co-packaged optics) switches are expected to achieve large-scale commercial use after 2028.

Coordinated evolution of SerDes and modulation: per-lane SerDes rates are migrating rapidly from 56Gbps to 112Gbps, with the Tomahawk5 chip fully supporting 112Gbps. The next-generation Tomahawk6 will further introduce 224Gbps SerDes. In modulation, PAM4 has become the mainstream standard for high-speed interconnect, with transmission efficiency twice that of traditional NRZ. In future 102.4T-and-above chips, PAM4 will continue to combine with 112G/224G SerDes to build a high-bandwidth, high-efficiency physical-layer foundation.

2.3.1.2 Elastic Compute Network Architecture Supporting Ultra-Large-Scale Expansion

The Scale-Out network of an AI compute cluster aims to achieve efficient, lossless interconnect of GPU servers. Its networking architecture must scale elastically with cluster size, mainly in the following modes.

1. Single-machine direct connect: suited to small-scale inference

Scenario: suited to small-scale inference or development/testing of tens of GPU cards.

Scheme: connect all GPU cards directly to a single high-performance switch for the

simplest non-blocking network. For example, H3C's S9827-128DH series switch based on the Tomahawk5 chip can support up to 128 400G ports per unit, meeting hundred-card cluster needs.

Value: easily supports hundred-card clusters, with quick deployment and simple O&M.

2. Two-tier CLOS architecture: supporting medium/large training clusters

Scenario: supports medium/large training clusters of hundreds to tens of thousands of GPU cards—currently the mainstream scheme.

Scheme: uses the classic two-tier flat Spine-Leaf architecture. By choosing Spine/Leaf switches of different bandwidths (e.g., 12.8T to 102.4T) and linearly increasing the number of devices, flexible, linear expansion from 512 cards to 16K cards is achieved while maintaining constant low latency and high bandwidth.

Value: a three-hop non-blocking network with low latency, simpler design, deployment, and O&M, and easy fault location.

Table 2.4 Scale of the two-tier CLOS Scale-Out network architecture

	400G card access scale	400G card access scale
Spine: 12.8T, 32×400G Leaf: 12.8T, 32×400G	Spine: 16 units Leaf: 32 units	512
Spine: 25.6T, 64×400G Leaf: 25.6T, 64×400G	Spine: 32 units Leaf: 64 units	2K
Spine: 51.2T, 128×400G Leaf: 51.2T, 128×400G	Spine: 64 units Leaf: 128 units	8K
Spine: 102.4T, 128×800G Leaf: 102.4T, 128×800G, one-to-two access	Spine: 64 units Leaf: 128 units	16K

Note: in the 102.4T scheme, Leaf switches use 800G ports connected to 400G GPU cards via breakout cables (e.g., one-to-two), doubling port density.

3. Three-tier CLOS architecture: building 100,000-card ultra-large-scale clusters

Scenario: used to build a single ultra-large-scale model-training cluster of tens of thousands to 100,000 cards.

Scheme: introduce a Core layer on top of the two-tier architecture, forming a Core-Spine-Leaf three-tier architecture. This architecture scales horizontally based on POD (Performance-Optimized Module) units, each POD being a two-tier CLOS internally. By stacking multiple PODs interconnected via the Core layer, near-unlimited linear scalability is achieved—the infrastructure for training next-generation trillion-parameter models.

Value: an ultra-large-scale expansion scheme that two-tier CLOS cannot replace. But it also has issues such as more hops, increased latency, complex design/deployment/O&M, and high cost.

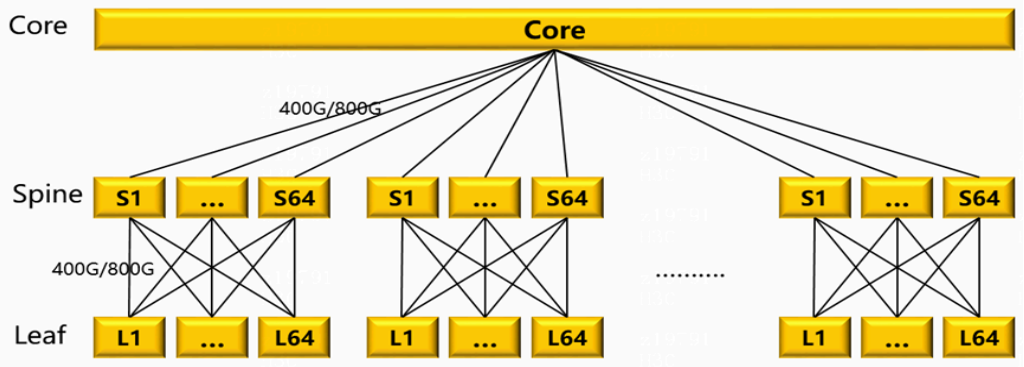


Figure 2.6 Three-tier CLOS Scale-Out network architecture

4. Multi-plane networking: a heterogeneous expansion scheme for high-performance NICs

Scenario: with high-performance smart NICs supporting port-aggregation splitting (e.g., NVIDIA CX8 800G, currently supporting 2×400G), use 2 planes—two independent two-tier CLOS networks—to provide an ultra-large cluster of tens of thousands of cards.

Scheme: connect the multiple NIC ports of a single server separately into multiple independent two-tier Spine-Leaf network planes. Each plane provides a complete network path for the server, achieving cross-plane bandwidth aggregation and fault isolation. Based on the Tomahawk6 chip, a single plane can provide 32K 400G interfaces.

Value: using a simple, low-cost two-tier CLOS, an ultra-large cluster of tens of thousands of cards can still be built.

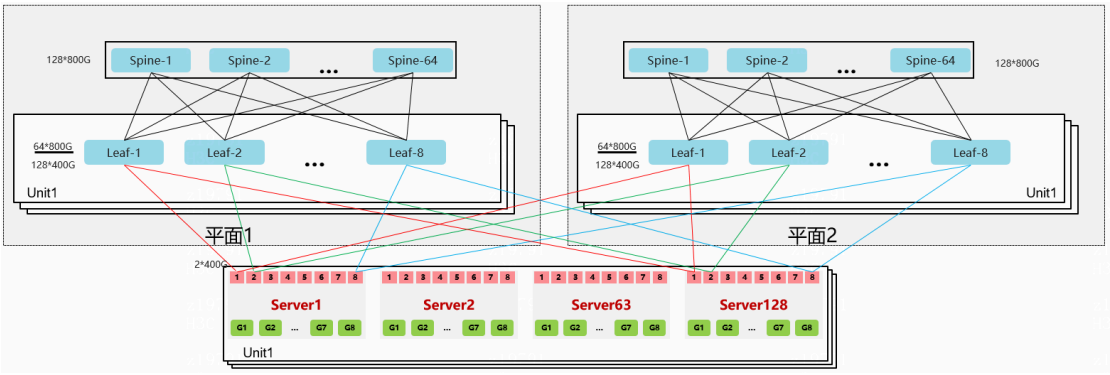


Figure 2.27 Dual-plane Scale-Out network architecture

2.3.2 Communication Efficiency

The traffic model of compute networks is dominated by a few fixed collective-communication operations such as All-Reduce and All-to-All. Its traffic has the typical characteristics of low entropy (few flows, single feature), burstiness, and long flows, which easily cause instantaneous congestion in the switching network.

Since GPU training follows a strict synchronous paradigm, all nodes must complete collective communication before entering the next round of computation, prone to a “barrel effect.” Therefore, end-to-end latency jitter caused by congestion directly slows the overall training job, becoming the core performance bottleneck of compute networks.

To overcome this, the industry focuses on improving network-communication efficiency and proposes two technical paths—host-network coordinated scheduling and pure-network-side scheduling—to achieve finer load balancing and congestion control.

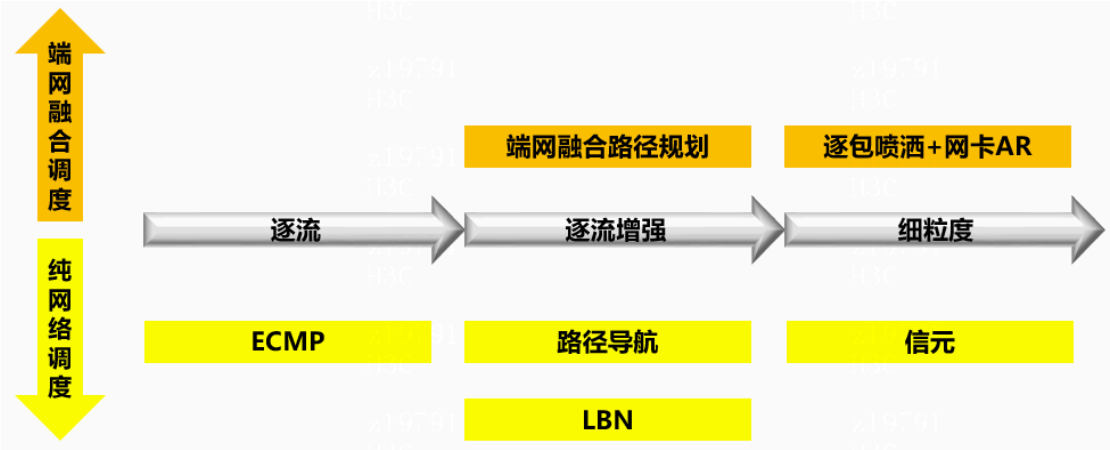


Figure 2.28 Scale-Out network load-balancing technology system

2.3.2.1 Pure-Network-Scheduling Load-Balancing Technologies

These are technical schemes in which traffic-optimization scheduling is achieved

autonomously entirely by network devices (mainly switches), without any cooperation or awareness from servers or NICs.

Table 2.5 Pure-network-scheduling load-balancing technologies

Item	Description
ECMP	Switch N-tuple hash algorithm, currently the most widely used in the industry; but due to low-entropy characteristics, load-balancing may be poor, which can be tuned via multiple QPs, adjusting the hash algorithm, etc.
Path navigation	The controller collects switch information, identifies congestion points and contributing flows, and re-plans to move congested flows to lightly loaded paths; only suitable for fixed-communication scenarios, with poor effect for unpredictable traffic such as MoE.
LBN	Ingress-port-based hashing ensures even Leaf uplink traffic; under large-scale networking, congestion probability is high when multiple Leaves go downstream through the same Spine—suited to small-scale networking where the host side does not support multiple QPs.
Cell	Cell-switching technology splits traffic into cells at the Leaf, evenly sprays them to all Spines, then reassembles, achieving load balancing; naturally heterogeneous-decoupled, requires no NIC support, and adapts to arbitrary traffic.

2.3.2.2 Host-Network-Fusion Scheduling Load-Balancing Technologies

A technical architecture that achieves global traffic optimization through information exchange and joint decision-making between the network side (controller, switches) and the compute side (smart NICs, collective-communication libraries).

Table 2.9 Host-network-fusion scheduling load-balancing technologies

Item	Description
Host-network-fusion path planning	The controller obtains GPU communication relationships via the CCL and pre-plans non-overlapping network paths to avoid congestion; suited to training scenarios with fixed communication patterns, with poor effect for unpredictable traffic such as MoE.
Per-packet spraying + NIC AR (Adaptive Routing)	Based on packet-level even spraying, paired with receive-end NIC AR out-of-order reordering, load-balancing performance is excellent; currently mainly relies on NVIDIA's closed ecosystem, and is expected to become widespread as domestic and Broadcom NICs support AR in the future.

2.3.3 Reliability and Maintenance

2.3.3.1 Switch Node-Level Reliability and Maintenance

Compute networks usually consist of tens to hundreds of box-architecture switches; their hardware design must ensure key parts (e.g., power supplies, fans) support M+N redundancy and have fault monitoring and O&M capability for core components such as CPU and memory.

2.3.3.2 Network Architecture-Level Reliability and Maintenance

Intra-network reliability: core reliance on ECMP for rapid self-healing. Taking the H3C S9827-128DH as an example, 64-path ECMP groups can be formed between Spine and Leaf; when a device or link fails, traffic automatically switches to other paths for

lossless self-healing. Combined with data-plane fast-convergence technology, fault-switching packet loss can be controlled to the millisecond level, ensuring business continuity.

Network-edge reliability: a multi-plane redundant architecture improves access reliability. Server NICs are dual-homed to Leaves on 2 planes; normal traffic is forwarded in isolation across the dual planes, and when an access link or device fails, host-side CX8 & NCCL escape provides fault redundancy, avoiding single-point failure at the network edge.

2.3.3.3 Link-Level Reliability and Maintenance

Compute networks rely on tens of thousands of optical modules interconnected, whose transmission stability directly affects the continuity of AI training and inference jobs. To ensure reliability, the industry takes two-level measures—pre-training and during-training.

Pre-training PRBS stress test: before networking is complete and services go live, PRBS (pseudo-random binary sequence) bit streams are commonly used to test all optical modules at scale. This screens out underperforming modules in advance, avoiding service interruptions caused by hardware-quality issues at the source.

During-training optical-link monitoring: after services run, optical modules undergo probabilistic performance degradation over time and environmental factors, possibly causing port faults or link flaps. To this end, the industry commonly monitors key metrics such as SNR, FEC counts, symbol error rate (SER), and bit error rate (BER) in real time. By analyzing the degradation trends of these metrics, potentially faulty modules can be warned and replaced in advance, achieving predictive maintenance, ensuring transmission quality and significantly reducing the risk of AI-job interruption.

Through the combined strategy of “strict pre-deployment screening” and “continuous in-operation monitoring,” compute networks can build a highly reliable optical-interconnect layer, providing a stable foundation for large-scale AI services.

2.3.3.4 Service-Level Reliability and Maintenance

Compute networks place higher demands on congestion monitoring; traditional switches’ second-level port statistics struggle to meet these needs.

Limits of traditional monitoring: switches can detect network congestion and indicate performance bottlenecks via ECN marking, PFC packet rate, and other statistics. But such monitoring is coarse-grained, unable to capture millisecond-to-hundred-millisecond instantaneous micro-bursts or locate the specific business flows causing congestion, making bursty congestion hard to observe and present effectively.

Micro-burst monitoring capability: to solve this, compute switches must count port buffer usage and forwarding rate at millisecond granularity. This lets the system accurately capture and report instantaneous traffic peaks when micro-bursts occur, avoiding their being “averaged out” by second-level means, achieving effective burst monitoring and problem location.

Flow-level monitoring and tuning: to further locate the congestion root cause, flow-level monitoring must be enabled. By building flow tables for business flows and counting the forwarding latency of each flow, the “congestion-contributing flows” with significantly increased latency can be precisely identified. Combined with the network controller, these flows can be dynamically redirected to lightly loaded paths, achieving real-time-state-based load-balancing tuning to optimize network performance.

In summary, by combining millisecond-level port monitoring with flow-level latency analysis, compute networks can achieve closed-loop management from “detecting congestion” to “locating the root cause” to “dynamic tuning,” effectively addressing the high-burst, low-latency AI-traffic challenge.

2.3.4 Future Outlook

2.3.4.1 Network-Management Software Supports Compute Services

Compute-network construction and O&M is a systems-engineering effort spanning multiple service layers, requiring deep integration and coordination of the network with switches, NICs, GPUs, and other devices. Network-management software, centered on automation, systematization, and intelligence, provides three key capabilities.

Host-network coordination, rapid bring-up: through global visualization and automated deployment, achieve unified management of devices, GPUs, and NICs, support one-click policy delivery and health stress testing, and greatly improve large-scale networking delivery efficiency.

Global control, network tuning: for compute traffic's low-entropy, bursty, elephant-flow characteristics, sense congestion at the service level and dynamically tune to achieve dynamic load balancing and global path optimization, improving network throughput and training efficiency.

Intelligent O&M, training assurance: build a full-cycle assurance system of pre-training inspection, during-training real-time monitoring (path, congestion, hashing), and post-training fault demarcation and traceback, achieving rapid sensing, location, and optimization.

This forms a closed loop of "rapid delivery – dynamic tuning – intelligent O&M," providing stable, efficient, and visible network support for compute services.

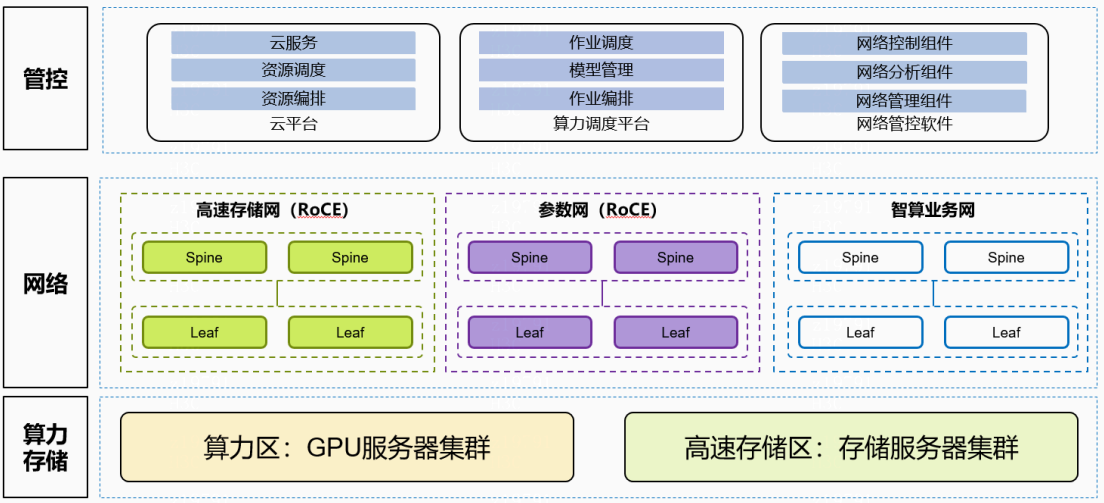


Figure 2.29 Compute-network management software architecture

2.3.4.2 An Open, Decoupled, and Healthy Ecosystem

Compute networks mainly consist of switches, NICs, and GPUs. Taking a typical inter-GPU RDMA Write operation as an example, the DMA process is as follows.

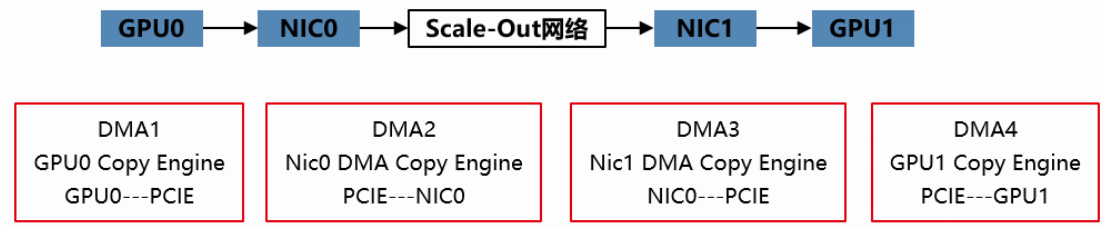


Figure 2.30 Participants in the RDMA flow from a hardware perspective

Compute-network construction often faces closed-ecosystem and integration challenges. The current market is mainly supplied with independent components such as

switches, NICs, and GPUs by different vendors, requiring system integration to build a complete solution. However, some large vendors, leveraging full-stack product capabilities, tend to promote closed “all-in-one” solutions, objectively limiting customers’ opportunity to choose the industry’s best technology and product combinations and bringing problems such as a single-source supply chain and uncompetitive cost.

To address this challenge, DDC (Distributed Disaggregated Chassis / dynamic multi-connection) technology emerged. Its core value is decoupling the network from the upper-layer AI framework and heterogeneous compute. By fully integrating functions such as cell spraying, packet reassembly, and reordering into the network side, regardless of the brand or model of GPU and NIC a customer uses—even if it lacks advanced out-of-order processing—a stable, high-performance transmission experience can be obtained via the DDC network, truly achieving “heterogeneous compatibility, performance worry-free.” This technology not only breaks ecosystem lock-in but also provides key technical support for customers to build an open, efficient, and autonomous compute network.

2.4 AI Storage Technology and System

2.4.1 Overview

For AI storage infrastructure in large data centers, the capability boundary of storage nodes—which handle training hot data, inference caching, vector-index file persistence and high-speed reads, checkpoints, parallel file services, and metadata services—is mainly determined by: first, the PCIe-lane scale, memory-channel scale, and NUMA topology provided by the CPU; second, the front-end NVMe media density and local flash-pool capability; third, the access mode of 400G and emerging 800G-class RDMA networks.

Based on current large-data-center deployment practice, high-density all-flash nodes have become an important representative form of high-performance AI storage servers; such nodes usually provide high-bandwidth data services for the hot-data tier, with the design goal of achieving systematic matching among CPU, SSD, memory, and high-speed RDMA NICs.

2.4.2 AI Storage Technology

2.4.2.1 High-Speed Inference Storage Architecture

The high-speed inference storage architecture can be summarized as four tiers: GPU VRAM, CPU memory, local SSD, and network SSD. GPU VRAM carries the hottest KV Cache and the current request context; CPU memory serves as the active cache and swap-in/swap-out buffer; local NVMe SSD provides low-latency capacity expansion; network SSD or a remote flash pool is used for cross-node sharing, elastic expansion, and warm/cold data.

The significance of this architecture is placing the inference context in tiers by access heat, so the GPU need not frequently wait for data due to long context, multi-turn dialogue, or agentic-AI tasks.

Figure 2.31 Schematic of the four-tier high-speed inference storage architecture

The NVIDIA Inference Context Memory Storage Platform (CMX) is a dedicated "inference-context memory storage platform" released by NVIDIA at CES in early 2026 for long-context LLM / multi-turn dialogue / agentic AI inference. Its core uses the BlueField-4 DPU for hardware acceleration, inserting a Tier-3.5 (see table) Pod-shared context layer between GPU HBM and traditional storage, specifically solving the problems of KV-cache VRAM overflow, high latency, and high power.

Table 2.10 NVIDIA divides the entire AI memory/storage hierarchy into G1–G4, with CMX inserted in the middle:

Tier	Location	Media	Latency	Use
1	GPU local	HBM3e	~100ns	Model weights, hottest KV
2	CPU near-memory	Vera CPU + LPDDR5X	~1μs	Overflow KV, medium heat
3.5 (CMX, Tier 3.5)	Pod-shared	BlueField-4 + Flash	~5μs	Long-context KV, multi-turn sharing, agent memory
4	Cluster / cloud	Traditional NAS/S3	~ms	Cold data, persistence

2.4.3 AI Storage Node

2.4.3.1 CPU

In storage-server architecture, the core value of the x86 CPU lies mainly in a mature I/O ecosystem, stable enterprise-grade RAS features, and long-term adaptation to mainstream storage software stacks. For AI storage nodes, CPU selection must consider: how a single server supports pass-through access for large numbers of NVMe SSDs; and how to attach one or more high-speed RDMA NICs.

Platform planning first looks at the number of PCIe lanes. Enterprise NVMe SSDs commonly use a PCIe x4 interface; 400G/800G-class high-speed NICs usually use PCIe x16 slots, plus lane occupation by boot drives, BMC, on-board controllers, etc. In terms of PCIe 5.0 lanes the CPU can directly provide (CPU native direct-connect, not the final motherboard-available count): Intel Xeon 6 (Granite Rapids, 6th-gen Xeon) up to about

96 per socket; AMD EPYC 9005 (Turin) up to 128 per socket, 160 in a dual-socket config. Note that a dual-socket system is not a simple doubling—some lanes between the two CPUs are used for UPI/Infinity Fabric interconnect, and actual usable lanes are also taken by boot drives, BMC, etc. Thus different platforms differ markedly in single-socket direct-connect capability, dual-socket expansion, and whether an extra PCIe Switch is needed.

When a node deploys multiple high-speed RDMA ports, the affinity among CPU, NIC, and SSD is especially critical. NICs should be in the same NUMA domain as the NVMe pool carrying the main I/O to reduce cross-socket data movement; with a dual-socket architecture, the impact of cross-socket paths on bandwidth efficiency and tail-latency stability must be fully evaluated.

2.4.3.2 Memory

The following two scenarios significantly increase memory-capacity configuration.

1. When KV Cache uses memory caching, its capacity is approximately linear with the number of tokens, model layers, KV heads, per-head dimension, data type, and number of concurrent sessions. In long-context or high-concurrency inference, cache demand easily exceeds the single-card VRAM limit. To this end, the industry commonly uses a four-tier offload strategy: GPU VRAM → CPU memory → local SSD → network SSD. CPU memory not only serves as an intermediate buffer for swapped-out KV blocks but also handles prefetch and asynchronous swap-in/swap-out scheduling. In capacity planning, memory is generally configured at 3–6× the GPU HBM; taking 4 H100s (320 GB HBM total) as an example, 1–2 TB DRAM is recommended, dedicated to storing inactive KV blocks swapped out from GPU VRAM.

2. The metadata-service layer of a parallel file system requires large amounts of memory, with memory capacity closely related to the metadata hot set composed of inodes, dentries, stripe mappings, and client lock tables. The memory footprint of the client lock table usually grows approximately linearly with the number of concurrent clients; under high lock contention, due to frequent lock callbacks and lock escalation, growth may exceed linear, so extra capacity should be reserved in planning.

2.4.3.3 NVMe SSD

At the local persistent-media layer, the NVMe SSD is a core part of the AI storage server. The mainstream direction is enterprise NVMe SSDs, with common forms including U.2/U.3 and EDSFF specs such as E1.S and E3.S. Scenarios such as AI training, inference, checkpoints, vector indexing, and high-concurrency metadata access have clear requirements for low latency, high-concurrency random read/write, and stable sequential throughput, so SSD selection should comprehensively evaluate steady-state performance, write consistency, power-loss protection, thermal management, and long-term operability.

By interface generation, PCIe 5.0 x4 has become the mainstream foundation for high-performance enterprise NVMe SSDs; its single-drive sequential read/write, random IOPS, and queue concurrency are markedly improved over PCIe 4.0, better handling loads such as training hot data, checkpoint writes, vector-index loading, and inference-cache expansion. For AI storage nodes, the significance of Gen5 SSDs is not only faster single-drive speed but also that fewer drives can form higher node throughput, reducing structural pressure among drive bays, power, PCIe Switch, and network egress.

PCIe 6.0 raises the link rate to 64 GT/s and introduces PAM4, FLIT encoding, and stronger error correction, providing a higher host-side bandwidth ceiling for next-generation Gen6 NVMe SSDs. As Gen6 SSDs, PCIe 6.0 Switches, CXL-related memory expansion, and 800G/1.6T networks gradually mature, AI storage nodes can provide higher aggregate bandwidth with fewer PCIe lanes, fewer drive bays, or fewer access-tier hops, of direct value to long-context inference, KV-cache tiering, real-time feature reads, and multi-tenant vector retrieval.

Longer term, the PCIe 7.0 spec has evolved toward a 128 GT/s link rate, and future

higher-generation SSDs will push the local flash layer to keep approaching the access capability of the memory layer. For AI storage, the value of advanced SSD technology is not simply chasing peak benchmarks, but shortening the time GPUs wait for data, improving recovery speed after cache hits, lowering the expansion cost of the hot-data tier, and providing a more reliable media foundation for storage-decoupled GPU nodes (few/no local drives, storage pooled at the rack level), network context caching, and rack-level flash pools.

In terms of media route, the applicable positions of TLC and QLC should be distinguished by workload type. TLC suits the training-hot-data tier, checkpoint tier, metadata tier, and mixed read/write tier; its steady-state write capability and endurance suit long-term heavy-load scenarios. QLC suits read-heavy, write-light, capacity-first AI scenarios, such as the inference-cache expansion tier, vector-database main-data tier, RAG raw-index tier, and read-only-dataset tier.

NVMe SSD error diagnosis and early fault prediction can achieve device monitoring via standardized management capabilities. The NVMe-MI spec requires unified interfaces for NVMe-device discovery, monitoring, configuration, and update on both in-band and out-of-band paths, facilitating predictive maintenance of large-scale nodes. On the other hand, FDP (Flexible Data Placement) lets the host provide finer-grained write-placement hints to the SSD, reducing garbage collection and write amplification and further improving media life, performance consistency, and space efficiency.

In terms of form and power, new interfaces such as E1.S and E3.S will replace traditional drive-bay descriptions. SNIA's public description of EDSFF emphasizes its better adaptability in flash density, power capability, thermal design, and serviceability compared with traditional forms, helping accommodate more high-performance flash in a single node. In single-drive power, the focus is on lower write amplification, more reasonable airflow, and a more stable thermal-control curve.

Table 2.11 Applicable positions of TLC and QLC in AI storage

Media type	**More suitable workloads**	**Main advantages**	**Risks to control**
TLC	Training hot data, checkpoints, metadata tier, mixed read/write cache tier	Stronger steady-state writes, higher endurance, easier tail-latency control	Higher cost; should be prioritized for tiers with the highest performance-consistency needs
QLC	Inference corpora, vector-database body, read-heavy/write-light datasets, capacity-expansion tier	Higher capacity efficiency, lower unit cost	Sustained rewrites and high-frequency random writes expose steady-state-performance and lifespan limits faster

2.4.3.4 RDMA NIC

In AI storage servers, the high-speed RDMA NIC is a core component as important as the CPU and NVMe SSD. As local NVMe-pool aggregation keeps improving, 400G/800G-class RDMA network interfaces have become a baseline direction for high-performance AI storage nodes. New-generation high-speed NICs place higher demands on PCIe 5.0 or PCIe 6.0 motherboard and CPU platforms, so network planning must be designed together with the host platform.

GPU Direct Storage (GDS) is a storage-I/O acceleration technology proposed by NVIDIA. It establishes a direct DMA data path between GPU VRAM and storage devices, bypassing the bounce buffer in CPU memory, thereby reducing CPU load and easing the system bandwidth bottleneck. In local-storage scenarios, GDS directly opens the DMA

path between GPU VRAM and NVMe SSD; combined with RDMA NICs, AI storage nodes further shorten the GPU data path and reduce CPU involvement at the storage node.

Before introducing EBOF, JBOF, and storage-decoupled architectures, the boundaries of these terms must be clarified. JBOF (Just a Bunch of Flash) usually refers to an external flash enclosure of multiple NVMe SSDs providing shared block access via NVMe-oF; EBOF (Ethernet Bunch of Flash) can be understood as a JBOF form carried over Ethernet/RoCE; storage decoupling designs GPU compute nodes with no local persistent data drives, with the compute side accessing a remote flash pool or parallel file system over a high-speed network. All three point to the resource-pooling direction of "compute–flash decoupling." EBOF/JBOF/Diskless, parallel file systems, and distributed storage all rely heavily on high-speed RDMA networks to ensure sustained bandwidth and low latency for concurrent multi-client access. In inference systems, whether remote network-storage tiering of KV Cache or context sharing and migration, all rely on high-speed-network support.

2.4.3.5 DPU NIC

The DPU-based high-speed inference storage architecture is NVIDIA's Inference Context Memory Storage Platform. Driven by the NVIDIA BlueField-4 DPU, it targets long-context, multi-turn, and agentic AI inference, aiming to expand GPU context capacity at the rack and cross-node-cluster level and support high-bandwidth KV-Cache sharing. The network base uses NVIDIA Spectrum-X Ethernet, providing RDMA-based high-speed KV-Cache access; at the software level, it works with NVIDIA Dynamo, using the asynchronous transfer library NIXL to enable seamless KV-Cache flow among HBM, CPU memory, and off-chip storage. The emergence of this platform shows that the DPU's role in AI storage is evolving from traditional protocol offload further toward context storage, cache routing, and an efficient data-sharing layer.

2.4.3.6 Storage Drive Expansion

Storage drive expansion can be understood, beyond "drive-enclosure expansion," as "flash resource-pool expansion." JBOF places large numbers of NVMe SSDs in an independent flash enclosure exposing shared access via NVMe-oF; EBOF emphasizes that the flash enclosure connects to the compute network mainly via Ethernet and RoCE. The core value of both is decoupling flash resources from compute nodes, managing them uniformly as an independent resource pool, and providing composable flash-access services to upper-layer nodes via NVMe-oF over RoCE or NVMe-oF over TCP. This architecture enables finer-grained fault-domain isolation during expansion, maintenance, firmware upgrade, and media replacement, and better suits managing flash as a resource pool in large data centers.

The diskless architecture further advances this resource-pooling idea: GPU compute nodes no longer have local drives for business-data persistence, instead placing models, datasets, vector indexes, checkpoints, or context caches in a remote flash pool, parallel file system, or object/file-converged storage. Some implementations still keep a small amount of local NVMe for system boot, logs, or temporary cache, but the core data path relies on high-speed-network access to remote storage. This approach especially suits AI clusters that need elastic expansion, rapid replacement of stateless compute nodes, unified management of flash capacity, and splitting fault domains by role such as training, inference, and retrieval. Combining JBOF/EBOF with the diskless architecture forms a multi-tier AI storage system from local cache to remote flash pool.

Chapter 3 Network and High-Speed Interconnect Layer

The interconnect system of a compute cluster can be divided into three levels: Scale-up (vertical scaling), Scale-out (horizontal scaling), and Scale-across (long-distance scaling). Scale-up focuses on integrating resources within a single logical compute node, achieving unified addressing among multiple GPUs and breaking the bandwidth and latency bottlenecks of the traditional bus. Scale-out relies on a lossless network for inter-node communication, focusing on cluster-scale expansion. Scale-across further extends Scale-out clusters distributed across regions, forming a cross-regional compute pool, commonly used in ultra-large AI data centers with more than 10,000 cards. The three interconnect networks form a complementary hybrid compute architecture, greatly improving node-cluster compute density and parallel-computing efficiency.

3.1 Scale-Up Technology Analysis

3.1.1 Research Background and Significance of Scale-up

Current large models place stringent demands on the compute performance, storage capacity, and data-interaction capability of the underlying compute infrastructure. Compute demand is not merely stacking chip counts, but a rigid requirement for low latency, strong consistency, and high compute density. The scalability of GPU compute and storage space directly determines the efficiency of model training, inference, and multi-node collaborative computing. Traditional interconnect architectures generally have prominent bandwidth bottlenecks, high communication latency, and limited scalability, hard to adapt to the massive high-frequency data interaction of model training/inference. Against this industry backdrop, Scale-up vertical scaling has become a key technical path to improving compute.

In the early stage of compute-cluster infrastructure, the vertical-scaling architecture commonly used the PCIe bus for interconnect. As compute and interconnect technology iterated, the traditional architecture gradually evolved toward high-performance super-node architectures. The mainstream Scale-up interconnect protocols and technology systems in current super-node scenarios mainly include NVLink, OISA, UALink, SUE, ESUN, CLINK, Co-Fabric, etc.

3.1.2 Card-to-Card High-Speed Interconnect Protocols and Technical Standards

3.1.2.1 NVLink Protocol

NVLink is a high-speed interconnect technology proposed by NVIDIA in 2014. Unlike the traditional architecture where GPUs must forward data via the CPU, NVLink uses a GPU direct-connect design, working with dedicated NVSwitch chips to let multiple GPUs exchange data directly, greatly improving communication efficiency. NVLink supports memory coherence and unified memory addressing, integrating the local VRAM of multiple GPUs into a large unified memory pool. It also integrates hardware-acceleration protocols, optimizing collective operations such as All-Reduce, reducing multi-GPU communication latency and improving data-interaction efficiency and stability.

As the benchmark scheme of the Scale-up architecture, NVLink has gone through multiple iterations with continuous leap-style performance gains. From the first-generation NVLink 1.0 at 160GB/s bidirectional, to the currently commercial NVLink 5.0 at 1.8TB/s bidirectional—a more-than-10× improvement; the next-generation NVLink 6.0 is expected to be commercially available in 2026, further raising bidirectional

bandwidth to 3.6TB/s. At the system level, NVIDIA's Blackwell Ultra NVL72 rack achieves an all-to-all topology of 72 GPUs, with total whole-rack communication bandwidth up to 130TB/s.

Despite its outstanding performance, NVLink has clear shortcomings, the most central being a highly closed ecosystem, strong vendor lock-in, and poor compatibility. Although NVLink Fusion, an open interconnect scheme, was introduced in 2025, vendors face strict licensing-access thresholds and it is not yet widely deployed. Once adopted, users are deeply locked into NVIDIA's hardware ecosystem. Second, it is costly: NVLink requires dedicated NVSwitch chips, proprietary backplanes, and power/cooling designs, with stringent whole-machine requirements and deployment/O&M costs higher than other general-purpose schemes. These limits pose high deployment and adaptation thresholds in construction.

3.1.2.2 Omni-directional Intelligent Sensing Interconnect (OISA)

In 2024, China Mobile proposed the Omni-directional Intelligent Sensing Express Architecture (OISA) for Scale-up scenarios in AI data centers, and released the updated OISA 2.0 protocol in 2025. It supports efficient interconnect of 1024 GPUs within a super node, with card-to-card bandwidth exceeding the TB/s level and latency reduced to hundreds of nanoseconds, strongly supporting data-intensive AI applications such as large-model training and inference.

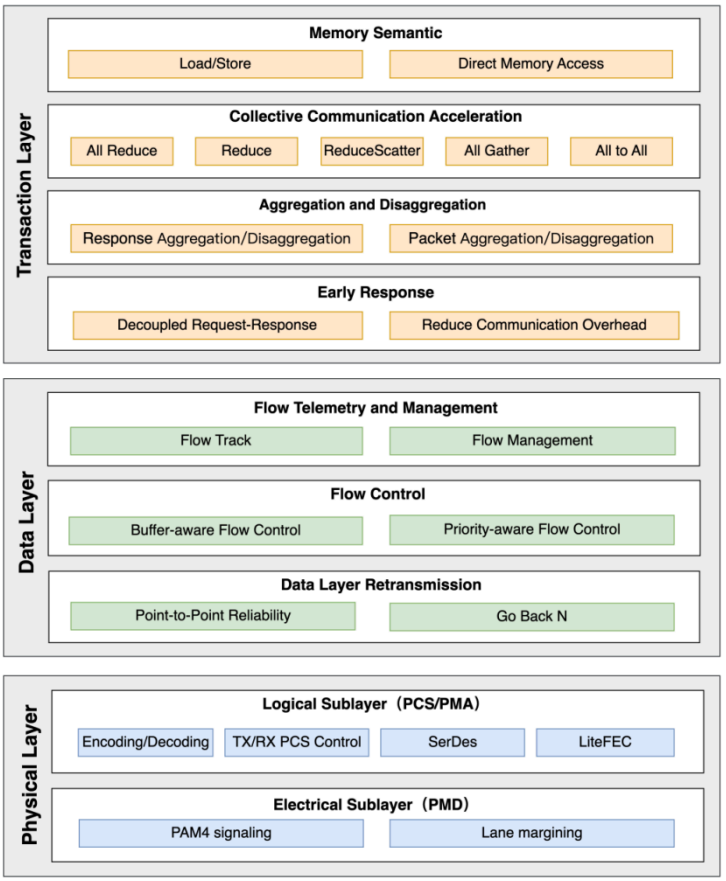


Figure 3.1-1 OISA protocol architecture

As shown in Figure 3.1-1, the OISA protocol stack consists of the transaction layer (TL), data layer (DL), and physical layer (PL). The transaction layer, at the top of the stack, interacts with the GPU's on-chip network, receiving transaction-related signals from the GPU, encapsulating them into transaction-layer packets (TLP), and forwarding them to the data layer. The data layer performs flow control and intelligent traffic sensing to ensure transmission stability. The physical layer uses an Ethernet-based

implementation, leveraging mature Ethernet physical-layer technology (e.g., SerDes) to support high-speed transmission between GPU devices.

The OISA 2.0 protocol has several features: first, support for native memory semantics, enabling AI chips to access data across nodes without obstacles; second, an innovative TLP packet-reconstruction technology that aggregates tiny compute transactions to improve bandwidth-utilization efficiency; third, intelligent sensing, monitoring the interconnect status of global GPUs and switch chips in real time to support dynamic routing, congestion control, and intelligent O&M; fourth, collective-communication hardware acceleration, offloading complex operations such as collective communication in AI training to switch chips via dedicated hardware engines, bringing significant performance gains to scenarios such as large-model training.

By building a unified, open technology platform, OISA provides a new domestic solution for super-node card-to-card interconnect and is a representative domestic Scale-up scheme. This will effectively stimulate industry-chain innovation, drive technology upgrades and iterative innovation of related products, and build a new compute-industry ecosystem of positive interaction and continuous evolution.

3.1.2.3 SUE Technology

To achieve efficient transmission among XPUs at large scale, Broadcom released the SUE (Scale-up Ethernet) protocol in April 2025; Figure 3.1-2 shows the SUE protocol-stack architecture. Besides the standard mode, SUE offers a streamlined mode called SUE Lite, which streamlines the packet format and header length, keeping only the header used for data forwarding to reduce hardware overhead and transmission latency.

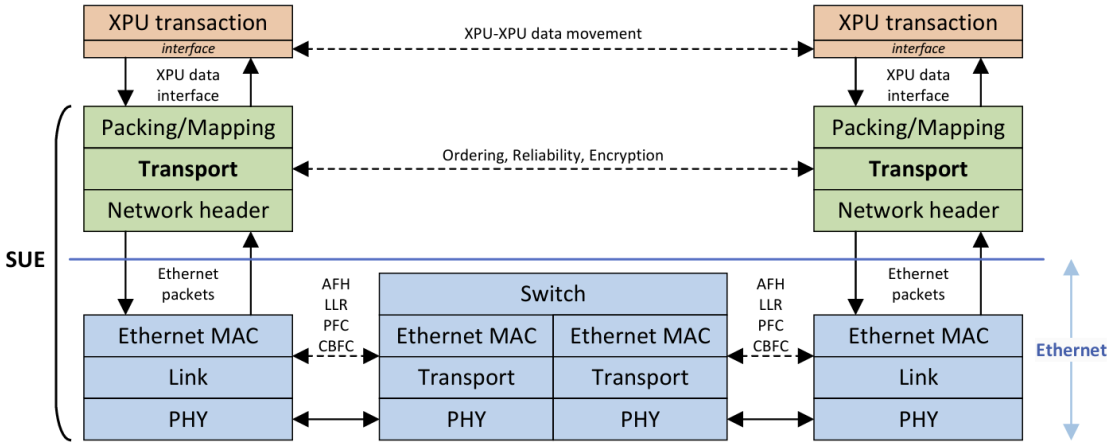


Figure 3.1-2 SUE protocol architecture

For transmission reliability, SUE supports credit-based flow control (CBFC), priority-based flow control (PFC), and link-layer retry (LLR), building an end-to-end lossless transmission system. The three work together to significantly reduce latency jitter and packet loss, meeting the stringent transmission-stability requirements of large-model training and inference.

In June 2025, the Open Compute Project (OCP) announced the establishment of the SUE-T (Scale-up Ethernet Transport) working group to improve the protocol system of Scale-up Ethernet interconnect. SUE-T will carry on and continue advancing part of the core transport-layer work of the original SUE, focusing on the host-side transport-layer protocol to decouple the transport layer from the switching layer.

At the hardware-ecosystem level, SUE adapts to Broadcom's Tomahawk Ultra and Tomahawk 6 series switch chips, relying on their forwarding performance to provide ample bandwidth and low latency for large-scale Scale-up clusters. SUE's deep coordination with switch chips makes the Ethernet-based Scale-up architecture a core support for AI super nodes to move from technical scheme to commercial deployment.

3.1.2.4 UALink Protocol

In October 2024, to break the monopoly of proprietary protocols and let accelerators

and switch chips from different vendors work together efficiently, nine companies—AMD, AWS, Astera Labs, Cisco, Google, HP, Intel, Meta, and Microsoft—announced the founding of the UALink (Ultra Accelerator Link) Consortium. The UALink protocol is the core technology led by the consortium, with version 1.0 officially released in April 2025 and version 2.0 updated in April 2026. The protocol is open to consortium members. Its architecture, shown in Figure 3.1-3, comprises, from top to bottom, the protocol layer, transaction layer, data-link layer, and physical layer.

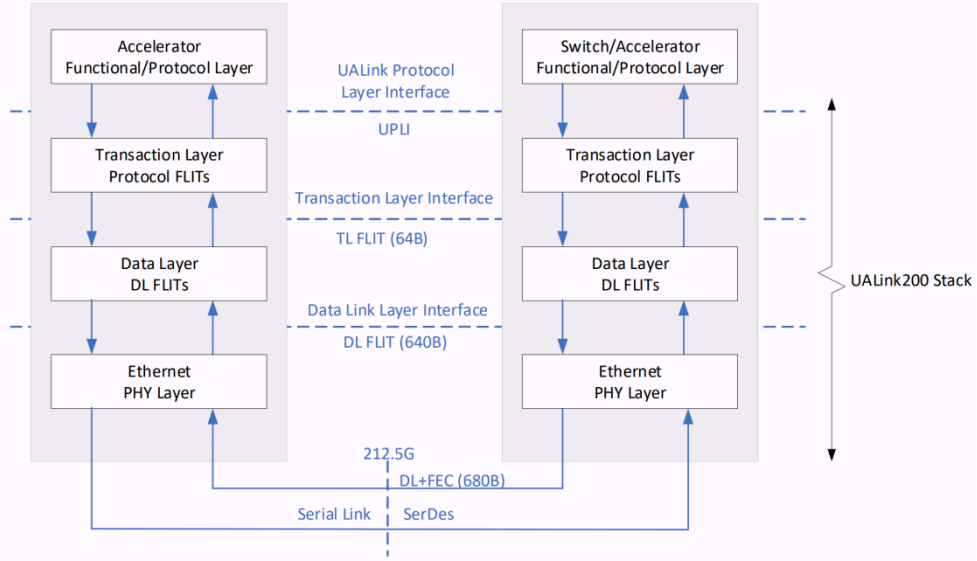


Figure 3.1-3 UALink protocol architecture

The UALink protocol layer, also called UPLI (UALink Protocol Level Interface), is the interaction interface between the protocol and the accelerator. Its transaction layer breaks protocol-layer instructions into standard transaction-layer transfer units, TL Flits (64 Byte), and passes them to the data-link layer. Its data-link layer adds CRC and a frame header to each Flit, encapsulating them into data-link-layer transfer units, DL Flits (640 Byte), and passing them to the physical layer; it also has link-retry and credit-based flow-control mechanisms to ensure reliability. Its physical layer is based on IEEE 802.3dj with partial modifications, letting the protocol directly reuse mature Ethernet SerDes, cables, and other supply-chain resources, lowering hardware mass-production and deployment cost.

UALink integrates the UALinkSec security mechanism, supporting end-to-end data encryption and trusted-execution-environment adaptation, achieving tenant-traffic isolation to meet the security-compliance needs of cloud-native multi-tenant scenarios. However, UALink does not support a coherence protocol; data coherence must be designed and implemented by vendors themselves, e.g., via software or other means. In industry promotion, the first products supporting the UALink spec are expected to be commercially available in 2026–2027, becoming a core support for opening up AI compute infrastructure.

3.1.2.5 Other Card-to-Card Interconnect Protocols

PCIe, with its mature industrial chain, controllable cost, and strong ecosystem compatibility, became the technical route for rapid deployment in the early stage of compute infrastructure. Its typical architecture uses a full-interconnect topology within a single node, with inter-node multi-machine communication via PCIe Switch, solving trillion-parameter large models' core needs for large VRAM and low-latency communication. But limited by the protocol-stack architecture and transmission characteristics, native point-to-point direct connection between nodes cannot be achieved, affecting bandwidth utilization.

Inspur Information deeply modified the PCIe protocol and developed the Co-Fabric interconnect protocol to build Scale-up compute systems of 32, 64, 128 cards and larger.

On top of the general bus protocol, Co-Fabric adds an Agent Layer and an Adapter Layer, breaking the expansion limits of native bus protocols and supporting full-switched non-blocking interconnect of more than a hundred cards. Through local virtual devices, it achieves transparent mapping of remote GPU VRAM and builds a globally unified address space. Through efficient data forwarding within the switching domain, it compresses communication latency to the hundred-nanosecond level. Taking a 64-card super node as an example, the DeepSeek R1 large-model token-generation speed (TPOT) reaches 8.9ms, and the DeepSeek V3 large-model single-card throughput reaches 631.4 tokens/s.

In September 2025, led by ODCC and headed by CAICT and Tencent, the ETH-X interconnect protocol white paper was released. The spec's transaction layer defines the implementation of PAXI (Peer-to-Peer AXI), extending AXI bus transactions to remote nodes over Ethernet; the data-link layer defines the frame structure and retransmission mechanism; the physical layer is compatible with the PMD and PCS definitions of standard Ethernet IEEE 802.3. ETH-X aims to provide Scale-up interconnect capability based on the Ethernet ecosystem.

At the October 2025 OCP Summit, 12 tech giants including AMD, Arista, and Broadcom jointly launched the ESUN (Ethernet for Scale-Up Networking) alliance, focusing on standardizing the lower network layers in accelerators and switches, coordinating with the transport-layer-focused SUE-T to jointly drive large-scale application of open Ethernet in super nodes and break the monopoly of proprietary protocols.

In November 2025, under the guidance of the Ministry of Industry and Information Technology, the China Electronics Standardization Institute launched the Compute Interconnect Bus protocol (CLink) joint initiative. Through standardization to lead compute-technology innovation, it jointly drives the continuous iteration of the compute-interconnect-bus protocol-standard cluster. CLink covers Scale-up overall architecture design, communication semantics, flow control, and other core content, providing unified technical specs for large-scale compute clusters and improving large-scale-deployment efficiency.

In addition, there are: Alibaba Cloud's Scale-up open ecosystem ALS (ALink System, accelerator interconnect system); ByteDance's Ethernet-based EthLink, supporting Load/Store and RDMA semantics; Huawei has published UB (Unified Bus) 2.0, an interconnect protocol for super nodes, enabling peer-to-peer direct communication among heterogeneous processing units such as CPU, GPU, NPU, DPU, and Switch; the high-throughput Ethernet ETH+ launched by Alibaba Cloud and the Institute of Computing Technology, whose protocol standard supports open-decoupling ideas; ACClink, self-developed by Tanwei Xinlian, with related products expected to complete technical validation by the end of 2026; OLink, independently developed by ZTE, compatible with the RDMA standard; and HSL, a fully open interconnect-bus protocol from Hygon, reducing the coordination latency and adaptation cost between GPU and CPU.

3.1.3 Future Development Trends and Application Potential

Scale-up technology will continue to optimize along three directions: physical layer, protocol stack, and boundary expansion. As the underlying core support, the 448Gbps per-lane physical-layer technology based on the OIF CEI-448G framework completed its standard-framework release in November 2025; it can double link bandwidth without changing port density, greatly improving single-rack compute density, but also places extremely high demands on the signal integrity of PCB materials and connectors, requiring low-loss materials and optical-electrical fusion schemes for stable deployment. To carry the physical-layer evolution, optical interconnect will become the core scheme to break the limits of electrical interconnect; technologies such as near-packaged optics

(NPO) and co-packaged optics (CPO) will gradually be commercialized—see Section 2.2 of this book for details.

In the protocol stack, in-network computing has achieved initial deployment in the large-model training clusters of leading vendors, realizing standardized hardware offload of collective-communication operators to reduce host-side resource consumption. But it also faces issues such as limited operator programmability, poor cross-vendor compatibility, and rising chip-design complexity and power; mismatched coordination of flow control and compute scheduling may also cause new congestion risks. This technology will keep iterating and co-building the ecosystem.

Advancing in parallel is the extension of architectural boundaries: Scale-up will extend from a single rack to a two-tier cross-rack network, breaking the single-rack compute ceiling and simplifying the networking complexity of large clusters. But due to increased hops causing higher latency, multi-path load balancing, adaptive routing, packet reordering, broadcast-storm risk from an enlarged broadcast domain, and the significantly enlarged impact scope of single-point failures, the cluster scale and interconnect performance must be balanced. Overall, Scale-up technology will keep evolving toward higher bandwidth, lower latency, and better adaptability.

3.2 Scale-Out Network Technology

3.2.1 Scale-Out Network Requirements and Goals

3.2.1.1 Scale-Out Requirements

The Scale-out network of an AI data center must simultaneously support cross-node communication for large-scale distributed training and inference, with core needs of ultra-high bandwidth, ultra-low latency, and lossless transmission. Training tasks mainly generate Scale-out traffic from data-parallel gradient All-Reduce and pipeline-parallel cross-node passing (tightly coupled communication such as tensor parallelism is mainly carried by the Scale-up domain); inference tasks, with the rise of PD separation, MoE expert parallelism (All-to-All), and cross-node KV-cache sharing, place increasingly strong bandwidth and latency demands on Scale-out.

1. High-bandwidth requirement

When network bandwidth is insufficient, the wait time caused by data-parallel communication can account for 55–85% of the model-training iteration time, becoming the main performance bottleneck. To complete training synchronization within milliseconds or microseconds, each compute node needs extremely high network egress bandwidth, e.g., 400Gbps, 800Gbps, or even 1.6Tbps.

2. Ultra-low-latency requirement

Collective communication often requires all worker nodes to wait for the slowest communication to finish before the next iteration, so network tail latency directly determines the progress of the whole training job. In addition, inference PD separation and KV-cache transfer are also highly sensitive to end-to-end latency and jitter. Thus the Scale-out network must provide microsecond-level end-to-end latency with minimal jitter.

3. Lossless-transmission requirement

Packet loss triggers excessive retransmissions, sharply degrading network performance. Therefore, the Scale-out network is based on the high-performance communication protocol RDMA to achieve lossless transmission of work traffic.

3.2.1.2 Goals of the Scale-Out Network

The goal of the Scale-out network is to maximize the overall communication performance of distributed training and minimize the idle wait time of XPUs (GPU, TPU, LPU, NPU, etc.) caused by the network. Its design metrics include: 1) supporting bandwidth up to 100 GB/s, and 2) achieving end-to-end round-trip latency below 10 microseconds within a range of hundreds of meters.

3.2.2 Network Operating System SONiC

SONiC (Software for Open Networking in the Cloud) is a Linux-based open-source network operating system initiated and open-sourced by Microsoft in 2015, currently a sub-project of the OCP networking domain. SONiC aims to provide a highly scalable network-operation solution for data centers through software-hardware decoupling; it supports multiple ASIC chips and CPU architectures, breaking the hardware lock-in of traditional operating systems. SONiC has strong ecosystem support, covering chip vendors such as Broadcom and Mellanox and equipment vendors such as H3C, Dell, and Arista, and has been deployed in production by hundreds of large organizations including Baidu, Tencent, and LinkedIn.

SONiC's core architecture is built on the Switch Abstraction Interface (SAI), using Docker container technology to encapsulate network function modules (such as BGP) as independent processes, achieving full decoupling among hardware, software, and protocol modules. It allows operators to share the same software stack across multiple vendors' hardware and supports rapidly adding new components or third-party software.

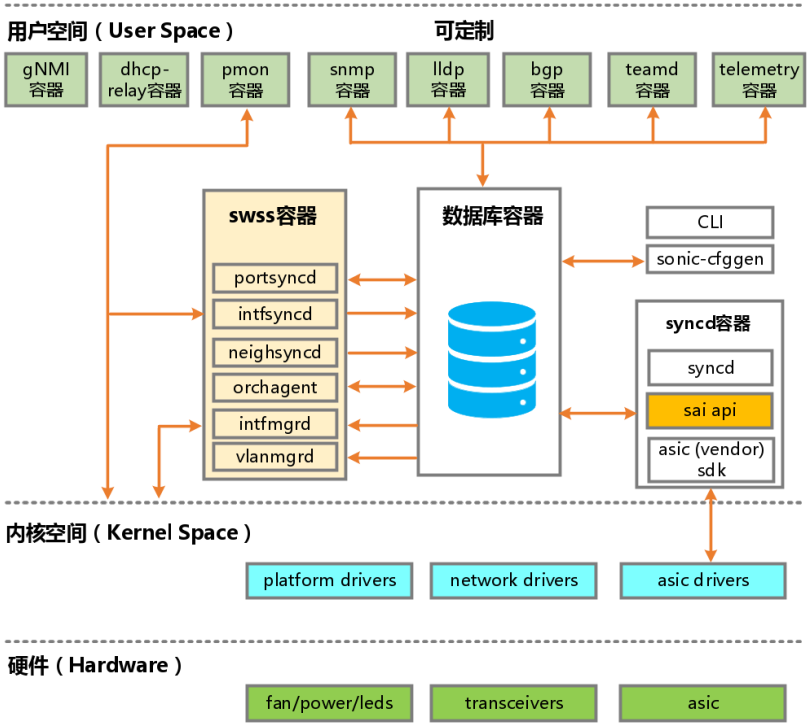


Figure 3.4 SONiC system architecture

In Scale-out networks, SONiC's standardized interfaces and containerized architecture fit the needs of building large-scale AIDC clusters, handling high-speed interconnect and dynamic configuration among many nodes, supporting high-throughput scenarios such as RDMA, and achieving linear expansion of hardware nodes without changing software. As the commercial ecosystem matures, SONiC is expected to become the mainstream network operating system for AIDCs, with a status similar to Linux in the server domain.

3.2.3 Scale-Out Protocols and Software Stack

3.2.3.1 Communication Protocols

AIDC Scale-out networks mainly use communication protocols such as InfiniBand, RoCEv2 (RDMA over Converged Ethernet v2), and Ultra Ethernet (UE), plus various application protocols.

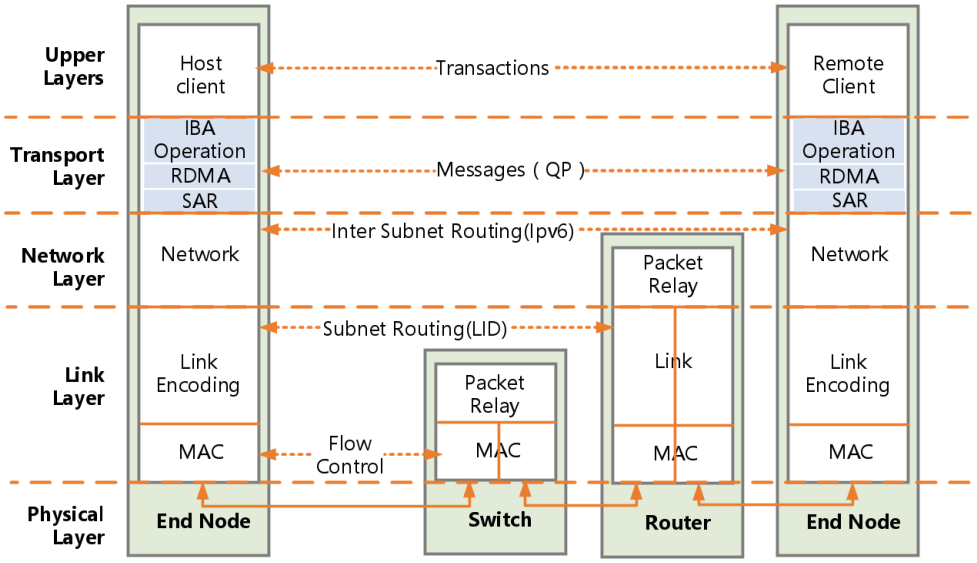


Figure 3.5 InfiniBand protocol stack

1. InfiniBand RDMA

InfiniBand RDMA is a network interconnect protocol designed for high-performance computing and AI clusters, allowing one computer's NIC to directly read/write another computer's memory, bypassing the OS kernel and CPU protocol processing to reduce latency. The InfiniBand link layer uses a credit-based flow-control protocol for lossless transmission and guarantees packet order within a single virtual-channel path. RDMA operations are executed by the RDMA-NIC hardware under the control of the InfiniBand Verbs interface, providing low-latency, high-bandwidth data transfer. Moreover, most key data-transfer operations of the protocol stack are offloaded to hardware for acceleration. At present, despite competition from Ethernet schemes such as RoCEv2 and UEC, InfiniBand RDMA still holds an advantage in AI scenarios with extreme performance requirements.

2. RoCEv2

In AI data center Scale-out networks, RoCEv2 is currently the main standard among Ethernet schemes. RoCEv2 is an implementation of RDMA running over Ethernet and supporting Layer-3 IP routing, using UDP at the transport layer and supporting collective-communication libraries and RDMA APIs at the application layer. RoCEv2's biggest advantage over RoCEv1 is cross-subnet routing, enabling larger-range interconnect expansion in Scale-out environments.

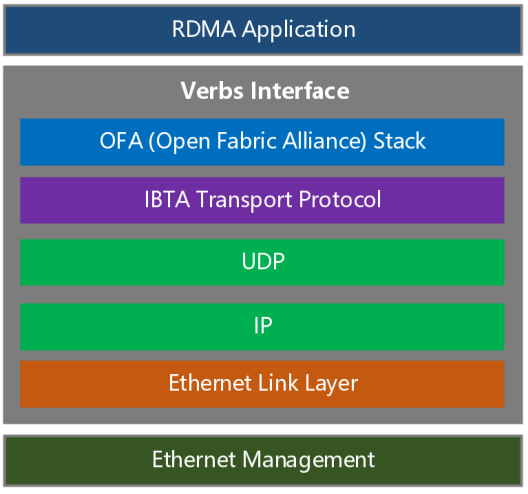


Figure 3.6 RoCEv2 protocol stack

RoCEv2 is based on the connectionless UDP transport protocol and lacks TCP-like reliability and congestion control, so it usually needs to combine PFC, ECN, and DCQCN to ensure reliable communication and high-performance transmission in large-scale XPU clusters. RoCEv2 fits the existing IP/Ethernet-based protocol stack of data centers, avoiding the extra cost and complexity of deploying a dedicated InfiniBand network.

3. UE

Ultra Ethernet (UE) defines a set of Ethernet-based standards and specs to meet the high-bandwidth, low-latency communication needs of HPC and AI applications. In particular, UE is the Scale-out network recommended by OCP Open Systems for AI.

The UE software layer defines APIs and specs, with its core being support for the libfabric (v2.0) interface.

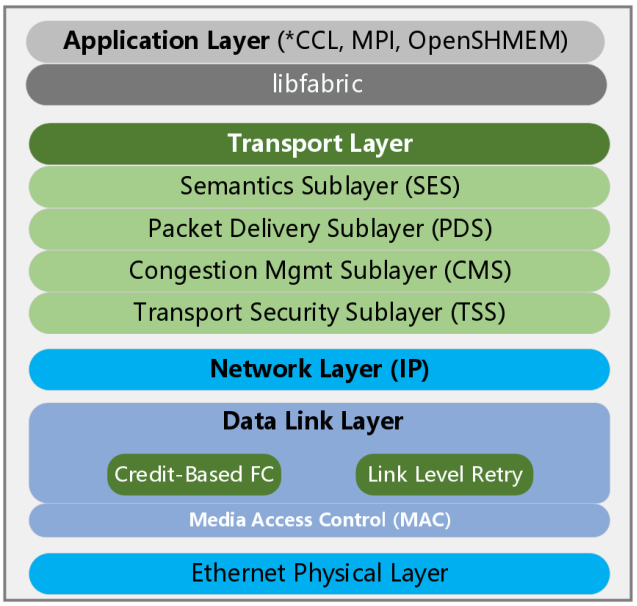


Figure 3.7 UE protocol stack

As shown above, the transport layer is the innovative core of UE, supporting RDMA and aiming for ultra-large-scale expansion to millions of endpoints, composed of the semantic sub-layer, packet-delivery sub-layer, congestion-management sub-layer, etc.

• The semantic sub-layer (SES) maps the libfabric API to network operations (e.g., RDMA read/write), supporting deferred send and message-rendezvous protocols to optimize the offload efficiency of AI collective-communication libraries;

• The packet-delivery sub-layer (PDS) handles reliability and ordering, achieving connection management via the ephemeral Packet Delivery Context (PDC), supporting zero-RTT context establishment—i.e., the first arriving packet can establish the reliability context;

• The congestion-management sub-layer (CMS) provides end-to-end congestion control, supporting sender- and receiver-side congestion control. The key feature of CMS is packet spraying, which distributes a single flow's packets across all available paths by changing the entropy value (EV), eliminating network hotspots.

The UE network layer uses standard IPv4 and IPv6 to ensure compatibility with existing data center routing architectures and management tools. Packet Trimming (PT) is a key optional feature of the network layer. When switch buffers congest, PT trims the packet payload and forwards only the header, serving as an extremely fast explicit packet-loss signal, letting the receiver immediately sense it and request retransmission, greatly reducing tail latency.

The UE link-layer protocol maintains compatibility with the IEEE 802.3 standard while introducing optional performance-enhancing features, including the link-layer retry

(LLR) protocol and credit-based flow control (CBFC). The Link Layer Discovery Protocol (LLDP) is used to negotiate with the communication peer to enable LLR and CBFC.

The table below summarizes the comparison of InfiniBand, RoCEv2, and UE.

Table 3.1 Comparison of InfiniBand, RoCEv2, and UE

	InfiniBand	**RoCEv2**	**UE**
Current stage	Mature / high-end market monopoly	Mature / large-scale deployment	Introduction / standardization
Core advantages	Extreme performance, full-stack optimization	Openness, cost, ecosystem, fast iteration	Open architecture, future-oriented design
Core disadvantages	Closed, expensive, slow iteration	O&M complexity from an open ecosystem	Not yet deployed at scale, with uncertainty
Future development	Market may gradually be taken by RoCE	A long-time de-facto market mainstream	An important future investment and technology bet

4. Application-layer communication protocols

The core role of application-layer protocols is to abstract the underlying RDMA or network hardware, implement collective-communication operations (All-Reduce, etc.), and provide high-performance, scalable communication interfaces. For example, NCCL is used for automatic topology awareness and efficient collective communication. In addition, the RDMA network can be operated directly via the RDMA API for custom libraries, storage, and network-optimization functions.

3.2.3.2 Network Congestion Control

1. Adaptive routing

Adaptive routing is an end-to-end fine-grained load-balancing protocol jointly implemented by switches and NICs, aimed at solving common issues such as elephant-flow collisions and network congestion in AI training. The Scale-out adaptive-routing protocol supports per-packet multi-path transmission, forwarding packets to available ports simultaneously based on egress-queue availability, achieving packet spraying and receive-end reordering, reducing link congestion and greatly improving effective network bandwidth utilization.

2. Closed-loop control

For many-to-one (Incast) congestion caused by multiple senders sending to a single receiver simultaneously, the Scale-out network combines congestion-control mechanisms such as PFC, ECN, DCQCN, and packet trimming, controlling the data-injection rate via feedback signals to achieve closed-loop congestion control in Incast scenarios.

3. Network telemetry

AIDC Scale-out routing and control protocols rely on measuring network state. Switches detect congestion hotspots in real time via telemetry and feed the information back to NICs. NICs perform congestion control with microsecond-level latency, finely regulating the traffic sources causing congestion to avoid network congestion. In addition, real-time telemetry ensures performance isolation in multi-tenant environments.

3.2.4 Performance Evaluation and Tuning

3.2.4.1 Scale-Out Network Performance Metrics

Scale-out network performance metrics include: 1) bandwidth and latency in RDMA bisection scenarios; 2) large-scale All-Reduce and All-to-All bandwidth; 3) effective bandwidth utilization; 4) scaling efficiency of distributed training jobs, i.e., the improvement ratio of job-completion time when XPU cards scale from N to 2N; 5)

network resilience, i.e., network performance when some links are down; 6) performance before and after multi-tenant network virtualization.

3.2.4.2 Performance-Tuning Techniques

Scale-out network performance-tuning techniques include traffic regulation, topology-aware routing, end-to-end coordination, and RDMA fault recovery. Traffic tuning dynamically adjusts ECN thresholds, PFC limits, etc. for different traffic patterns; topology-aware routing uses routing mechanisms based on the job communication matrix and topology information (e.g., multi-path traffic engineering) to improve load-balancing protocols; the end-to-end coordination mechanism performs full-stack joint tuning via software communication libraries, NIC drivers, and switch OS; the RDMA fault-recovery protocol provides fault detection and recovery, reducing the impact time of faults on compute jobs via backup multi-paths.

3.2.5 Industry Frontier Cases and Trends

3.2.5.1 AIDC Scale-Out Network Practice

1. InfiniBand-based Scale-out network

NVIDIA DGX SuperPOD uses InfiniBand to build large-scale distributed AI clusters and achieves a Scale-out network through modular design, adopting adaptive routing, multi-path traffic-sharing mechanisms, and hardware-based network self-healing and error-detection to ensure high-performance RDMA communication.

2. RoCEv2-based Scale-out network

Azure and Meta use RoCEv2 + custom Ethernet to achieve large-scale AI Scale-out networks with near-InfiniBand performance but more open and lower cost.

3. UEC-based Scale-out network

Broadcom's Scale-out network scheme is based on UEC, using enhanced RDMA, multi-path, and programmable congestion control to upgrade Ethernet into a high-performance interconnect system for million-node AI clusters.

4. Heterogeneous interconnect and customized schemes

Google's OCS Scale-out network essentially replaces traditional multi-tier electrical-switching networks with "reconfigurable optical paths + SDN control," achieving low-latency, high-bandwidth, low-cost interconnect for AI clusters. OCS is pure physical-layer switching, performing no protocol parsing or electrical processing of packets, so it is transparent to encapsulation protocols (such as RoCE).

5. VNET Scale-out case

VNET's AIDC uses a two-tier non-blocking Spine-Leaf architecture to build a 1:1 non-oversubscribed Scale-out network. The whole network is based on RoCEv2 lossless Ethernet, with thousand-card commercial clusters deployed at national compute hubs and bases such as Ulanqab, Zhangjiakou, the Yangtze River Delta, and the Greater Bay Area, smoothly expandable to the 10,000-card scale, with end-to-end communication latency stably controlled within 5μs and peak network utilization above 90%.

6. Purple Mountain Laboratories Scale-out case

Purple Mountain Laboratories built an ultra-wide lossless compute-network, based on a self-developed 51.2T Ethernet white-box switch running the UniNOS network-device operating system, using a two-tier Spine-Leaf architecture to build an end-to-end 400G Scale-out ultra-wide RoCEv2 network, deploying PFC and ECN for lossless transmission while using adaptive routing and per-packet load-sharing, with link-bandwidth utilization no less than 95%.

3.2.5.2 Future Trend Outlook

InfiniBand and RoCEv2 are still typical high-performance Scale-out network protocols, while protocols such as UE and MRC (Multipath Reliable Connection) iterate rapidly and may become mainstream in the future. Meanwhile, RDMA-related protocols deployed over OCS are an important evolution direction for AIDC Scale-out networks. In addition, Scale-up and Scale-out protocols show a trend of coordinated evolution, e.g.,

ESUN and UEC collaborating to align open standards.

3.3 Scale-Across Network Technology

3.3.1 AI Data Center Interconnect Requirements

3.3.1.1 Compute Expansion

The Scale-across network extends AI fabrics beyond a single AIDC, with connection distances spanning tens of kilometers. Via the Scale-across network, operators can interconnect multiple AI data centers to form ultra-large-scale AI infrastructure spanning multiple sites.

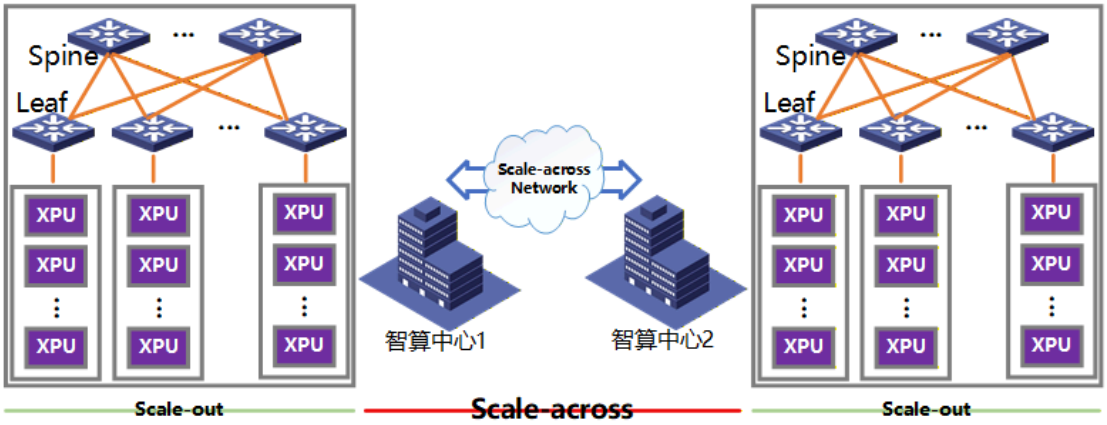


Figure 3.8 Schematic of the Scale-across network

Scale-across compute expansion connects the compute of multiple data centers in different geographic locations via the Scale-across network so that they logically appear as a unified, schedulable ultra-large compute fabric. Its need mainly arises from the physical and engineering limits a single data center faces when carrying trillion-parameter-scale model training. In addition, as AI business focus shifts to inference, compute needs to be distributed in metropolitan areas closer to users; via Scale-across technology, inference nodes within multiple metro areas can be organized into a unified resource pool for distance-first scheduling and dynamic load allocation.

3.3.1.2 Physical Limits

Besides compute-expansion needs, the need to connect multiple data centers via a Scale-across network also arises from the physical limits of a single data center, including power, cooling, and floor-space limits.

1. Power-supply limit

Single-rack power is evolving from the current ~150kW toward 200–500kW (or higher) (see 1.3, 2.1); a single AIDC struggles to support the enormous power and infrastructure consumption needed for millions of XPUs.

2. Cooling challenge

Traditional air cooling can no longer meet the cooling needs of ultra-high-power racks, requiring liquid-cooling systems, which place higher demands on a single building in physical space and engineering implementation.

3. Floor-space limit

The physical space of a single building or campus is limited, hard to keep expanding to accommodate million-scale XPU clusters.

3.3.2 Lossless RDMA

3.3.2.1 Distance-Adaptive Load Balancing

Compared with a single-data-center network, the core challenge of the Scale-across

network is the marked difference in link distances, causing obvious latency differences across RDMA paths; load also easily concentrates on some links, and traditional static routing struggles to fully utilize the whole network's resources, so an efficient dynamic load-balancing mechanism is needed to improve performance and stability.

The Distance-Adaptive Load Balancing (DALB) protocol, through dynamic path-weight computation and distance- and load-based adaptive traffic allocation, can significantly improve the performance and reliability of ultra-large Scale-across networks. DALB adaptively allocates traffic based on path distance and current network load, and through dynamic monitoring of network state and periodic weight adjustment, balances hotspot links and fully utilizes multi-path resources, enhancing the stability of InfiniBand RDMA, RoCEv2, and UE RDMA protocols in Scale-across networks. When a link is abnormal, DALB can automatically lower its weight or switch traffic to other paths, ensuring communication continuity and overall network reliability.

3.3.2.2 End-to-End Congestion Control

Scale-across networks involve cross-rack, cross-Pod, or cross-data-center communication, with complex network topology and significant differences in link latency and bandwidth. The end-to-end congestion-control protocol (E2E CC) monitors network congestion between the two communicating parties and adaptively adjusts the sending rate to avoid link overload. The protocol usually relies on the explicit congestion notification (ECN) mechanism: when a network device detects an excessively long queue or potential congestion, it marks packets via ECN, and the sender reduces its rate accordingly, preventing packet loss and latency accumulation.

In large-scale, heterogeneous, multi-path network environments, the end-to-end congestion-control protocol can be combined with the multi-path traffic-engineering (MPTE) protocol, spreading traffic across multiple links and dynamically adjusting the allocation ratio based on each path's congestion state so that low-congestion paths carry more traffic. By combining end-to-end congestion control and ECN notification with multi-path traffic scheduling, the Scale-across network can sense and respond to congestion in advance, achieving RDMA lossless transmission.

3.3.3 Wide-Area Deterministic Interconnect

3.3.3.1 Precise Latency Management

Latency management in wide-area Scale-across networks is achieved via mechanisms such as network-resource reservation, multi-path traffic allocation, load-balancing weight optimization, fast congestion notification, and elastic bandwidth allocation, relying on close coordination of compute and network scheduling. The network scheduler reserves bandwidth resources for each communication path in advance, while combining multi-path traffic engineering (MPTE) and backup-path strategies for fine traffic control. When a link is about to congest, network nodes can identify the affected link and notify the sending node, which dynamically adjusts its sending rate accordingly to suppress latency fluctuation while maintaining high throughput and end-to-end lossless performance. In addition, elastic bandwidth and multi-tenant scheduling strategies let network resources grow or shrink dynamically with compute-task demand.

3.3.3.2 Network State Awareness

In Scale-across networks, network-state awareness (e.g., in-band telemetry) is a core performance-assurance mechanism for ultra-large distributed compute systems. The protocol incorporates the physical distance of Scale-across links and the resulting propagation delay into the flow-control model, so each link's data-injection rate can be optimized based on its propagation characteristics to minimize tail latency; it can also automatically identify the network topology and intelligently allocate traffic among multiple feasible paths. In-band telemetry provides the network with sub-microsecond real-time monitoring, feeding congestion signals directly back to the sender so the NIC can quickly complete rate adjustment, and provides end-to-end health monitoring and

troubleshooting.

3.3.4 Industry Frontier Schemes and Trends

3.3.4.1 Long-Distance Cross-AI-Data-Center Training Schemes

As shown below, the industry deploys XPUs in multiple data centers and interconnects them via DCI devices to form a Scale-across network for cross-DC collaborative training.

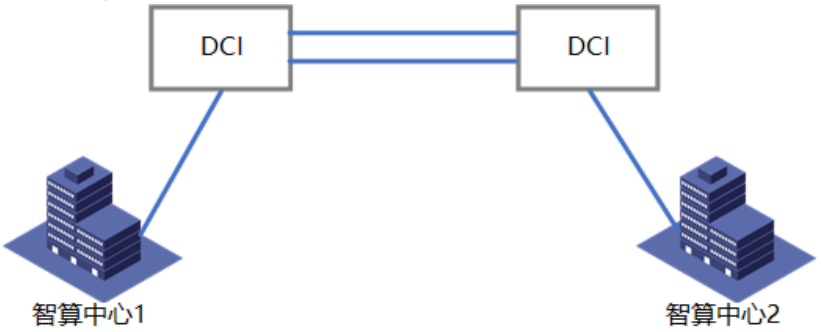


Figure 3.9 Schematic of long-distance AI data center interconnect

North America has completed long-distance cross-DC collaborative-training deployment; limited by technology, distances are currently within 100km. Specifically, the challenges and solutions for the 100km long-distance cross-data-center training scenario are shown in the table below.

Table 3.2 Challenges and solutions for long-distance cross-data-center training

100km-scenario technical challenge	Description	Solution
Long-distance congestion-notification efficiency	Devices in the near-end AI data center mark ECN; ECN marking – CNP congestion notification – sender rate reduction takes 1ms in total	Near-end DCI devices support FastCNP, proxying CNP replies to achieve microsecond-level sender rate reduction
Buffer backlog of the long-distance congestion proxy	Devices in the far-end AI data center congest; CNP notification – sender rate reduction takes 1ms total, during which the far end triggers PFC, and the far-end DCI must wait 1ms to release the buffer, requiring 50MB buffer per port, prone to overflow and packet loss	DCI devices use models with large HBM buffers, e.g., Broadcom's DUNX-series chips, or router NP chips

3.3.4.2 Scale-Across Network Practice

1. NVIDIA Scale-across case

NVIDIA launched Spectrum-XGS Ethernet, optimized specifically for Scale-across scenarios, enabling geographically distributed facilities to operate as a single "Giga-Scale" AI factory. In an NCCL All-Reduce collective-communication test between data centers 10 km apart, Spectrum-XGS performed 1.9× better than traditional Ethernet.

The NVIDIA Scale-across scheme supports a distance-adaptive congestion-control protocol, introducing algorithms aware of inter-site distance that dynamically optimize flow control to minimize the tail latency of long-distance transmission. Its topology-aware load balancing automatically adjusts traffic distribution across Scale-across fabrics, eliminating network hotspots and maximizing throughput. Through control logic based on global telemetry, it ensures deterministic latency and jitter QoS.

2. Broadcom Scale-across case

In the Scale-across domain, Broadcom proposed a scheme based on open Ethernet standards, aiming to integrate scattered AI clusters into an ultra-large compute pool supporting million-XPU nodes through city- or regional-level long-distance interconnect.

Broadcom's scheme combines switch-side drop congestion notification (DCN), priority flow control (PFC), and explicit congestion notification (ECN), with end-to-end traffic visibility, to solve the oversubscription challenge where Scale-across link bandwidth is usually smaller than intra-cluster bandwidth. At the Scale-across level, it uses deep-buffer technology to absorb the latency and micro-burst congestion of long-distance transmission, leveraging GB-level buffering to keep RDMA traffic lossless over long distances.

3. VNET Scale-across case

VNET builds AIDC clusters with a three-tier CLOS-like architecture (Core-Spine-Leaf) within cities such as Beijing and Suzhou: a Hub-Spoke-Edge topology. The Hub is the core hub AIDC, deployed in suburban areas with ample power and accessible green power, carrying 10,000-card-scale GPU training/inference clusters; the Spoke is a medium radiating AIDC, the relay hub of metro compute, evenly deployed at core locations of each city district; the Edge is an edge AIDC, pushed down to core business districts and industrial parks, deployed close to users and data sources to provide sub-millisecond low-latency edge-intelligence services. The underlying layer rebuilds the metro compute bus via an all-optical switching network, achieving one-hop direct all-optical interconnect across all Hub-Spoke-Edge nodes, with end-to-end optical-layer latency stably controlled within 0.5 ms, building a non-blocking AI Fabric within the city.

VNET proposed the Universal Cross Connection (UCC) protocol, based on standard Ethernet and RDMA, deeply optimizing and lightening the network and transport protocols, with a built-in endogenous network-security mechanism based on a public/private-key system, achieving trusted, manageable, and controllable connections within the metro and ensuring high-speed, secure cross-node compute scheduling.

4. Purple Mountain Laboratories Scale-across case

Purple Mountain Laboratories built a multi-compute-center cross-WAN-interconnect RDMA network based on the CENI network, deploying RDMA gateway devices at compute-center egresses. The gateways use a self-developed 12.8T programmable white-box switch running UniNOS, using 400G ports for interconnect. FastCNP fast-congestion-control technology is deployed on the gateways to achieve fast congestion-control feedback, solving the host-side RDMA-NIC flow-control problem caused by wide-area long-distance latency; an elephant-flow load-balancing technology is also deployed on the gateways, achieving lossless splitting of RDMA traffic based on 5-tuple + QPID, with dynamic orchestration via SRv6 multi-path policies for precise flow splitting and optimal path matching, improving wide-area link utilization. The CENI network provides 400G interconnect for deterministic transmission. Ultimately, in long-distance transmission over 100km, network throughput stays above 90%.

3.3.4.3 Future Trend Outlook

The AIDC Scale-across network is in a burgeoning development stage; it is not only the logical extension of Scale-out but also a key path to break the physical limits of a single data center and achieve the continuous evolution of trillion-parameter large models and multi-agent systems.

Driven by organizations such as the Ultra Ethernet Consortium (UEC), Scale-across network protocols are more likely to be built on Ethernet standards. Meanwhile, an open ecosystem is also an important direction for the evolution of Scale-across protocols.

Chapter 4 System Software and Resource Management Layer

4.1 AI System Software

4.2 Power and Environment Monitoring

4.3 Intelligent Autonomous O&M

4.4 Unified O&M Management

4.4.1 Current Status and Analysis of AIDC O&M

Currently, the maximum single-rack power density of AI super nodes is about 150kW and evolving toward 200–500kW. Liquid cooling, RoCE lossless networks, and large-scale server clusters have become standard, with business mainly large-model training, inference, and compute leasing. Demands for equipment uptime, rapid fault handling, and remote control far exceed those of traditional IDC, and AIDC O&M is at a critical period of transformation from traditional IDC hosting to high-density, high-reliability, intelligent O&M.

1) O&M model: gradually shifting from manual on-site inspection to a combination of remote centralized O&M + automated O&M + AI predictive O&M, with on-site staff only handling hardware replacement and emergency repair.
2) Core demands: the compute cluster must not be interrupted; a single server/card fault needs minute-level location, isolation, and recovery; massive servers (tens of thousands / hundreds of thousands) need unified management.
3) O&M tiering: the industry generally divides into in-band O&M (operating system, business, cluster software) and out-of-band O&M (BMC) (hardware bottom layer, power-on, BIOS, health status); the two complement each other, and BMC has become the foundation of AIDC O&M.

4.4.2 Whole-Rack Super-Node Management Architecture

Super-node system-management software has clear hierarchical control and integrated characteristics, with a technical architecture covering, from top to bottom, four levels: rack-level control, node-level management, network management, and centralized O&M control.

1) Rack-level management extends the control scope to the rack-level micro-environment, achieving power management, distributed-CDU cooling regulation, and unified monitoring of environmental sensors such as temperature and humidity.
2) Node-level out-of-band management relies on the board management controller, based on standard protocols such as IPMI and Redfish, achieving deep monitoring and control of server hardware status.
3) The network manager, for high-performance interconnect networks, provides topology auto-discovery, health inspection, and visual O&M, ensuring stable, efficient operation of the data center network.

4) Finally, through the centralized O&M console, global management capabilities are aggregated, providing operators a global view and supporting unified cross-domain policy delivery and automated O&M.

Super-node system-management software uses unified modeling and telemetry, building a unified hardware model based on protocol standards such as Redfish, abstracting all components in the super node—CPU, accelerator cards, memory, NICs, and PCIe/CXL devices—into a standardized data model, and performing fine-grained telemetry-data collection based on this model.

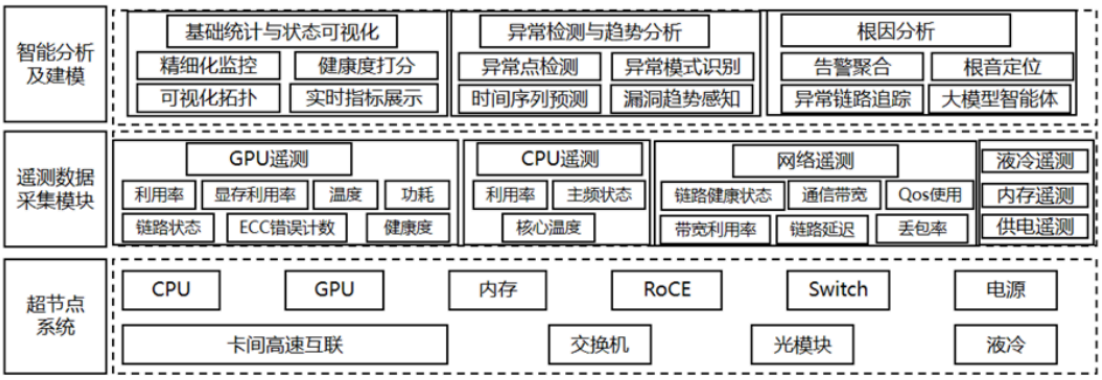


Figure 4.1 Whole-rack super-node management function framework

4.4.3 Super-Node Rack-Level Management

The whole rack manages the Power Shelf, PSU (Power Supply Unit), and CDU (Coolant Distribution Unit) system via the PMC (Power Management Controller) management module.

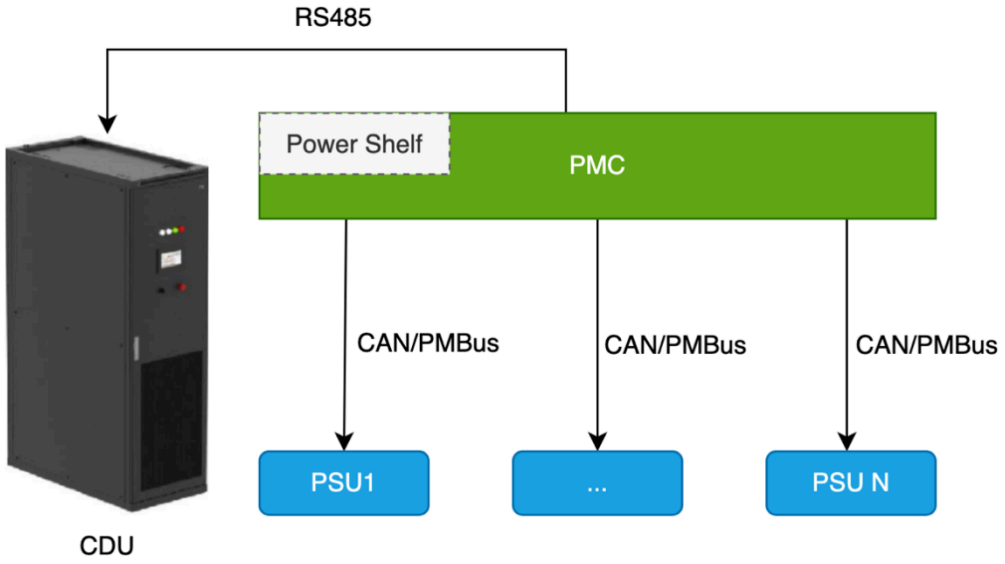


Figure 4.2 Rack-level management block diagram

4.4.3.1 Power-System Management

The Power Shelf must support sensor monitoring of core voltage, current, input power, output power, ambient temperature, Shelf output busbar clip temperature, etc., as well as monitoring of the total number of PSUs in the Power Shelf, supporting Power Shelf health-status logging and alarms.

For each PSU in the Power Shelf, PSU assets support query functions, including firmware, manufacturer, serial number, FRU, etc.; PSU health monitoring must support voltage, current, temperature, and power monitoring and log alarms, support PSU

black-box log download, and support service-transparent PSU firmware upgrade/downgrade.

4.4.3.2 Liquid-Cooling System Management

The whole-rack super-node leak-detection sources must cover the whole-rack manifold inlet/return quick connectors, flanges, and node core components (CPU/GPU, etc.), supporting leak monitoring and log alarms.

For the scenario where a distributed CDU independently supplies the whole rack, CDU asset acquisition must be supported, including manufacturer, product name, serial number, firmware version, liquid level, ambient temperature, liquid temperature, fan, and power management, supporting fault monitoring and log alarms for core components such as leakage, temperature, fan, and power, with logs retained through power loss. CDU management must support remote and local control of CDU operating parameters, support remote CDU firmware upgrade, and meet Modbus RTU/TCP for the external interface protocol.

4.4.3.3 Rack Management Protocol

The rack PMC (Power Management Controller) connects to the upper-layer O&M management platform via Redfish, IPMI, and SMASH CLP protocols to monitor rack health status and attributes.

4.4.4 Super-Node Node-Level Management

The node-level BMC achieves observability and control of the super-node infrastructure by uniformly orchestrating the compute Tray, Switch Tray, and power and cooling subsystems.

Node-level management mainly includes:

Core-component status monitoring and lifecycle management

PCIe Switch management and topology configuration

Power and cooling control

Figure 4.3 Node-level management block diagram

4.4.4.1 Core-Component Management

1) Motherboard management

The BMC handles unified monitoring and management of basic motherboard resources; the system continuously collects board-level telemetry via PMBus, I2C, GPIO, ADC, and other interfaces, and uses an event-driven mechanism for rapid detection and reporting of abnormal states. It mainly includes:

- Voltage, current, VR, and power monitoring
- Power-on sequence detection and alarms
- FRU management, supporting modification of major fields
- Firmware lifecycle management, supporting firmware upgrade and version management for BMC, BIOS, CPLD, PSU, GPU FW, etc., where BMC upgrade does not affect stable system operation
- BIOS configuration management, supporting out-of-band BMC modification of BIOS Setup options

2) CPU management

CPU management mainly covers CPU lifecycle management, operating-status monitoring, and fault detection, including:

- CPU temperature and power monitoring
- CPU asset-information management
- CPU presence, fault detection, and alarms

3) GPU management

The GPU is the core compute resource in the AIDC super node, and the node-level BMC must have GPU identification, status monitoring, and fault management.

GPU management capabilities include:

- GPU asset management
- GPU temperature and power monitoring
- GPU presence detection, utilization and VRAM-utilization monitoring, VRAM recoverable errors, unrecoverable errors, and ECC Mode monitoring

4) Smart-NIC management

As AI data centers gradually evolve toward intelligent networking, smart NICs have become a key component in super nodes. The BMC can achieve unified DPU management via MCTP, NC-SI, PCIe VDM, etc.

The node-level management system must support:

- Smart-NIC asset management
- Smart-NIC temperature and power monitoring
- Smart-NIC presence detection, port-connection status, negotiated rate, and other monitoring

5) Memory management

AI training scenarios place extremely high demands on memory stability; the BMC must continuously monitor:

- Memory temperature monitoring
- Memory asset-information monitoring, including manufacturer, serial number, memory capacity, bit width, maximum frequency, current frequency, etc.
- Memory presence detection, recoverable and unrecoverable error monitoring

6) PCIe Switch management

In AI super-node scenarios, the stability of the PCIe Fabric directly affects inter-GPU communication performance and overall compute efficiency. The BMC must continuously monitor PCIe link health and detect and alarm in real time on events such as link speed reduction, link jitter, error recovery, and abnormal interruption, ensuring high availability and stable operation of the PCIe Fabric.

PCIe Switch management mainly includes:

- Link Width/Speed status monitoring

- Port-traffic detection
- AER error collection
- Switch firmware management

In addition, the system should support dynamic reconstruction and state synchronization of the PCIe topology. In scenarios such as GPU Reset, PCIe Recovery, link anomalies, or topology changes, it can automatically rediscover and update topology information, ensuring the BMC-side topology view stays consistent with the actual hardware state.

PCIe Switch topology management mainly includes the following device types:

- PCIe Switch
- GPU Endpoint
- RNIC Endpoint
- PHY Retimer
- NVMe

Figure 4.4 PCIe SW topology management

4.4.4.2 Power Management

The node BMC can perform power control on the host server, including power-on, forced shutdown, soft shutdown, reset/restart, and forced restart; for smart-card configuration, it must support smart-card power on/off, described as follows.

1) Power-on: equivalent to a short press of the Power Button when the server is off, powering on each component in the designed sequence.

2) Soft shutdown: equivalent to a short press of the Power Button when powered on; the OS must be correctly configured to set the short press to direct shutdown for this to take effect. Some OSes pop up a confirmation upon receiving the soft-shutdown signal, requiring user confirmation in the OS before actual shutdown.

3) Forced shutdown: equivalent to a long press (over 5s) of the Power Button to force the server off, with the CPU/GPU processors powering down simultaneously.

4) Reset/restart: the system restarts immediately, simulating a short press of the Power Reset Button.

5) Forced restart: shut down then power on—equivalent to a long press of the Power Button to shut down, then waiting 10s and short-pressing the Power Button to

power on.

6) Smart-card power on/off: the smart card carries bare-metal and VM services; when the BMC detects the smart card present, normal power operations require the smart card not to power down. When the smart card needs to be powered on/off, an additional instruction must be provided to ensure new smart-card features take effect.

4.4.4.3 Cooling Management

For node-component cooling, apart from the liquid-cooled parts, the remaining fan-cooled components rely on the BMC adjusting fan speed to maintain normal operating temperature, supporting two speed-control modes—manual and automatic:

- Manual fan control: set the fan-control mode to manual to customize fan speed
- Automatic fan control: set the fan-control mode to automatic for intelligent automatic speed regulation

For cooling-anomaly handling, the following scenarios must be covered:

- BMC hang: ramp up to safe speed from the anomaly scenario;
- During BMC restart or IPMI-service anomaly: keep safe speed until the BMC finishes booting and the cooling module works normally;
- During firmware update, ensure no cooling risk (if the cooling module cannot regulate speed during the upgrade, ramp up to safe speed at the start of file upload);
- Cooling-module anomaly: keep safe speed until normal speed regulation is restored;

When core components (CPU, GPU, smart NIC, etc.) have abnormal temperature readings (0 value or unreadable) for 3 or more consecutive reads, immediately perform anomaly speed regulation to safe speed.

4.4.4.4 Log Diagnostics

The BMC must have one-click log-collection capability, automatically completing multi-source log collection, packaging, upload, or export, facilitating rapid problem location and O&M analysis.

Log-collection content should include:

- System Event Log (SEL)
  Records hardware-sensor fault events of the motherboard, CPU, GPU, memory, etc.
- Audit/operation logs
  Records non-query operations performed on the device.
- HOST logs
  Records BIOS and OS boot and runtime SOL logs.
- BMC logs
  Records BMC runtime logs, including dmesg, trace log, internal-module anomaly logs, etc.
- Device comprehensive diagnostic data
  Records in-system component asset information, firmware versions, historical sensor data, BIOS option data, cooling speed-regulation logs, CPU key status-register information, the last screen before a crash, etc.

4.4.4.5 Alarm Settings

The BMC must be able to configure alarm logs, including syslog alarm settings, SNMP settings, and email alarm settings, with the following requirements:

1) SNMP settings

Forward alarm logs externally via SNMP Trap, supporting configuration of host identifier, Trap version, alarm level, target-device IP, etc.

SNMP Get/Set is used to read device logs, supporting configuration of SNMP V1/V2C/V3, community name, etc.

2) Syslog alarm settings

Forward alarm logs externally via syslog remote forwarding, supporting configuration

of target-device IP, alarm level, transport protocol, etc.

3) Email alarm settings

Forward alarm logs externally via email, supporting SMTP server address, port, alarm level, etc.

4.4.4.6 Remote Access

The BMC has remote-access and server-operation capabilities, including graphical access, command-line access, and remote media, without a local monitor.

1) Graphical access

Remotely view and operate the host screen via HTML5 KVM and VNC; VNC supports password change.

2) Command-line access

Perform input/output operations by redirecting the server's serial port to the SOL console.

3) Virtual media

Supports remote mounting of ISO image files, CD/DVD, HDD, and folder image mounting for the host server's remote-media use. Virtual media is disabled by default; the virtual USB NIC device is required to be automatically disabled in the OS, and virtual media can be opened via IPMI/Redfish instructions or the web.

4.4.4.7 Users and Security

The BMC is an important management entry point for AI infrastructure. The node-level management system must build a complete system of identity authentication, permission control, network-access protection, and certificate security to ensure the security and reliability of the super-node platform.

1) User and permission management

The node BMC supports multi-level user management and Role-Based Access Control (RBAC), assigning differentiated management permissions to different user roles, effectively reducing the risk of misoperation and achieving secure isolation and fine-grained access control of system resources.

The node-level management system supports a typical three-level permission model, with permissions divided as follows:

Function module	Administrator	Operator	General user
User configuration	✓	×	×
Configure own account	✓	✓	✓
General configuration	✓	✓	×
Power control	✓	✓	×
Remote media	✓	✓	×
Remote KVM	✓	✓	×
Security configuration	✓	×	×
Debug diagnostics	✓	×	×
Query functions	✓	✓	✓

2) Password security policy

To improve system security, the BMC supports password-complexity policy configuration and enables strong-password requirements by default.

The password policy supports the following:

Minimum password-length limit

Password-complexity verification

Password-validity-period management

Historical-password repetition detection

Login-failure lockout mechanism

3) System firewall

The node BMC supports a built-in system firewall to limit illegal access and network-attack risk; administrators can restrict access to the management interface from specific sources based on the data center network plan, improving management-network security.

The system supports the following firewall-rule types:

IP-address firewall rules, supporting access control based on IPv4/IPv6 black/white lists.

MAC-address firewall rules, supporting access control based on MAC-address black/white lists.

Port firewall rules, supporting access restriction for management-service ports.

4) SSL certificate management

To ensure management-interface communication security, the node-level BMC supports SSL/TLS certificate management for secure communication such as HTTPS.

The system supports generating a CSR (Certificate Signing Request) file via the web interface and supports certificate import and update.

This effectively protects management-data transmission, preventing sensitive-information leakage and man-in-the-middle attacks.

4.4.4.8 BMC Network Management

The node BMC supports flexible network configuration and management to meet out-of-band management needs in different data center network deployments, supporting IPv4 and IPv6 dual-stack protocols and dynamic switching between static IP and DHCP modes.

To improve the reliability and availability of the node-level management network, the BMC network service has a Self-Recovery Mechanism. Through network self-recovery, the node-level BMC can effectively improve the stability and continuous operation of the management network and reduce the impact of network anomalies on remote O&M and cluster management. The network-recovery scheme is as follows:

1) When the system detects an abnormal loss of the IPv4 or IPv6 DHCP address, the network service automatically sends periodic DHCP/DHCPv6 requests to reacquire a valid network address, ensuring the out-of-band management network recovers promptly.

2) When the IPv4/IPv6 network-service process exits abnormally, crashes, or is lost, the system can automatically perform service restart and network recovery.

4.4.5 Development Trend Outlook

In the future, the BMC will gradually evolve from a traditional out-of-band management module into an infrastructure control node integrating standardized interfaces, hardware observability, automated O&M execution, secure and trusted control, and scenario coordination. Driven by new compute scenarios such as AIDC, liquid cooling, and GPU clusters, the BMC's capability boundary will keep expanding, becoming a key part of the O&M system at the server, rack, and even cluster level—the BMC is moving from an auxiliary management component toward a core control node of intelligent compute infrastructure.

The main trends can be seen in five directions:

Standardization: moving from IPMI to Redfish dominance, unifying interfaces across all models and lowering the difficulty of managing multiple devices.

Intelligence: moving from "monitoring" to "diagnosis + decision + linkage"; the BMC is no longer just "collecting data" but gradually gaining hardware autonomous-O&M capability.

Security: from a management entry point to a security boundary; future security requirements for the BMC will keep rising, and the BMC will be not just an "O&M entry" but an important part of the server hardware chain of trust.

Scenario specialization: deeply evolving toward AIDC/liquid-cooling/GPU clusters, the BMC's responsibilities will clearly expand, moving from a

“single-machine view” to a “node + rack + cluster” coordinated view.

Platformization: from single-point management to large-scale orchestration; the future value of the BMC lies not in clicking into a single machine’s page, but in whether it can be used in a batch, automated, and platform-based way.