开放计算标准工作委员会
Open Compute Technology Committee

OCTC AB03—2026

OCP China Community ｜ 开放计算中国社区

开放计算标准工作委员会 OCTC（中国电子工业标准化技术协会）

吉瓦（GW）级开放智算中心框架技术报告

开放、高效、绿色、智能的吉瓦级智算中心体系架构

GW-scale Open AIDC Blueprint White Paper

An Open, Efficient, Green, and Intelligent Architecture for GW-Scale AI Data Centers

编制单位：GW-scale Open AIDC 工作组（OCP China Community / OCTC）
指导单位：OCP 中国社区 · 开放计算标准工作委员会（OCTC）
发布日期：2026 年 7 月

贡献者与编制信息

编制单位

本技术报告由 GW-scale Open AIDC 工作组在 OCP 中国社区（OCP China Community）与开放计算标准工作委员会（OCTC）的共同指导下编制完成。工作组面向中国国情与工程实践，将系统梳理并构建一套开放、高效、绿色、智能、可规模复制的吉瓦（GW）级智算中心参考架构与实践框架。

主要编写者（单位）

OCP 中国社区

中电标协开放计算标准工作委员会（OCTC）

百度在线网络技术（北京）有限公司

浪潮电子信息产业股份有限公司

中航光电科技股份有限公司

世纪互联集团有限公司

中国移动通信有限公司研究院

中国电子技术标准化研究院

新华三技术有限公司

北京极客天成科技有限公司

江苏致网科技有限公司

广东省连接器协会

中国建筑标准设计研究院有限公司

惠州亿纬锂能股份有限公司

东莞立讯技术股份有限公司

庆虹电子（苏州）有限公司

江苏安澜万锦电子股份有限公司

行吟信息科技（上海）有限公司

主要编写者与其他贡献者

本技术报告各章节由工作组成员单位的 AIDC、超节点、连接器、供配电、冷却、网络、存储、互连协议，管理运维和建筑设计等领域专家共同撰写、审阅与校订。主要编写者如下：叶毓睿、武正辉、张斌、张群、梁宇彤、袁俊峰、陈宣豪、高小淇、刘石磊、蒋正顺、李红哲、邹胜亮、雷雄、解爽、齐韬然、王佳森、司昌亮、吴晓晖、席晓光、任晓磐、罗振、张军萍、李恒聿等。

在此也感谢何永占、陈海、俞雄杰、李典林、鲍益、秦晟煜、秦世娟、房思源、路鑫、卜莹琼、高建国、王兴隆、刘升福、张晓光、汪硕、王志强、赵伟、张政、庞斌、郝玉涛、

汪超、陈玥、赵静涛、邓昊昆、李波涛、杨海忠、苗睿、张洋、张先国、管峥朝等为技术报告做出了贡献。

技术审阅

OCP 中国社区、开放计算标准工作委员会（OCTC）等相关专家组。

特别致谢

感谢 Open Compute Project（OCP）Open Systems for AI 战略倡议提供的方法论基础与开放协作框架，为本技术报告的系统级设计视角提供了重要参考。

执行摘要

面向新一代 AI 基础设施的系统级愿景与中国实践

人工智能正以前所未有的速度演进。大模型、生成式人工智能（AIGC）与智能体（Agentic AI）等新型应用持续推动算力需求向更高规模、更高密度与更高能效跃迁。计算基础设施正从以服务器和集群为中心的传统数据中心形态，加速迈向以系统级协同设计为特征的新一代人工智能数据中心（AI Data Center，AIDC）。在这一进程中，数据中心功率正从数兆瓦（MW）向数百兆瓦乃至吉瓦（GW）级跨越，传统架构已难以满足未来大模型、通用人工智能与智能体经济在能效与可扩展性方面的需求。

Open Compute Project（OCP）社区提出的 Open Systems for AI 倡议，强调通过开放标准、系统视角与跨层协同设计，推动 AI 基础设施在性能、能效、可扩展性与可持续性方面的整体跃迁，为全球 AI 基础设施的发展提供了重要的方法论基础与开放协作框架。与此同时，中国正处于算力基础设施规模化发展的关键阶段：依托“东数西算”工程，在新能源供给、高压直流输电、跨区域算力布局、制造业供应链与工程组织能力等方面形成了独特优势，为探索吉瓦级智算中心这一新形态提供了现实土壤。

吉瓦级智算中心并非传统数据中心在规模上的线性放大，而是一种系统范式的根本转变：设计尺度从服务器级、机柜级上升至园区级乃至区域级；能源、计算与网络不再是相互独立的子系统，而需在规划阶段进行协同设计与联合优化；在超大规模条件下，可靠性、可维护性与全生命周期成本往往比单点性能更具决定性意义；系统复杂度的提升，使开放接口与标准协同成为降低系统性风险的关键手段。因此，吉瓦级智算中心的建设本质上是一个系统工程问题。

本技术报告并非具体产品或单一技术方案的说明文档，而是一个系统级参考框架。全文按四个相互依存的层次组织：基础设施层（建筑与空间规划、供备电系统、冷却系统）、IT 设施层（AI 超节点、高速互连物理层与连接器、AI 网络与交换系统、AI 存储）、网络与高速互连层（Scale-Up、Scale-Out 与 Scale-Across）、以及系统软件与资源管理层（AI 系统软件、动环监控、智能自治运维与统一运维管理），由于能力、时间所限，当前版本的这份报告，。各层均围绕开放、高效、绿色、智能四大主线，明确系统级设计原则、分层架构与关键接口方向。

通过本技术报告，GW-scale Open AIDC 工作组期望提炼吉瓦级智算中心在能源、计算、存储、网络等跨域协同层面的共性问题，为产业界、运营方、设备厂商及标准组织提供可对齐、可共建的参考基础，推动中国实践与 OCP 全球体系之间的双向互通与协同演进——**既能“引进来”，也能“走出去”**，从而凝聚产业共识、降低系统级创新成本、促进开放生态发展，为新一代 AI 基础设施的可持续演进提供坚实支撑。

目录

引言（Introduction）

人工智能（Artificial Intelligence，AI）技术正以前所未有的速度演进，大模型、生成式人工智能（AIGC）以及智能体（Agentic AI）等新型应用形态持续推动算力需求向更高规模、更高密度和更高能效方向跃迁。全球范围内，计算基础设施正从以服务器和集群为中心的传统数据中心形态，加速迈向以系统级协同设计为特征的新一代人工智能数据中心（AI Data Center，AIDC）。

在这一背景下，Open Compute Project（OCP）社区提出 Open Systems for AI 倡议，强调通过开放标准、系统视角与跨层协同设计，推动 AI 基础设施在性能、能效、可扩展性和可持续性方面的整体跃迁。这一倡议为全球 AI 基础设施的发展提供了重要的方法论基础和开放协作框架。

与此同时，中国正处于算力基础设施规模化发展的关键阶段。一方面，大模型训练与推理、产业智能化升级、智能体应用等新需求，正在推动超大规模智算中心持续建设；另一方面，中国在新能源供给、高压直流输电、跨区域算力布局、制造业供应链和工程组织能力等方面形成了独特优势，为探索吉瓦（GW）级智算中心这一新形态提供了现实土壤。

在此背景下，在 OCP 中国社区（OCP China Community）与开放计算标准工作委员会（OCTC）的共同指导下，GW-scale Open AIDC 工作组正式成立，旨在面向中国国情与工程实践，系统梳理并构建一套开放、高效、绿色、智能、可规模复制的吉瓦级智算中心参考架构与实践框架。

本技术报告并非具体产品或单一技术方案的说明文档，而是一个系统级的参考框架，目标在于：

提炼吉瓦级智算中心在能源、计算、存储、网络等跨域协同层面的共性问题；

明确系统级设计原则、分层架构与关键接口方向；

为产业界、运营方、设备厂商及标准组织提供可对齐、可共建的参考基础；

推动中国实践与 OCP 全球体系之间的双向互通与协同演进。

通过本技术报告，GW-scale Open AIDC 工作组期望凝聚产业共识，降低系统级创新成本，促进开放生态发展，为新一代 AI 基础设施的可持续演进提供坚实支撑。

1.行业背景

AI 模型的训练规模和计算需求呈指数级增长，推动算力密度、网络带宽、能效、绿色合规等需求同步跃升：

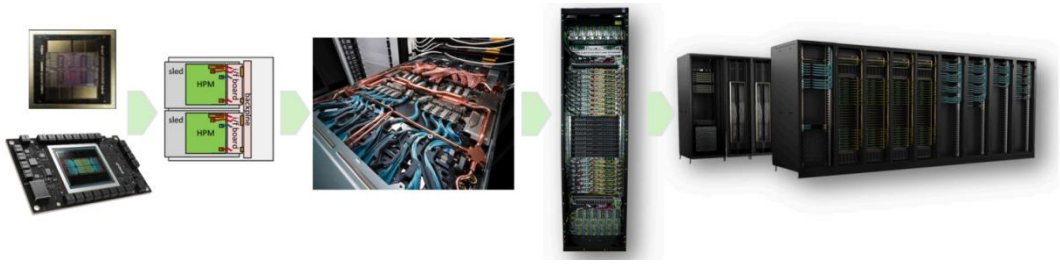
(1) 训练集群规模从万卡→十万卡→百万卡

(2) 高速互连从 400G→800G→1.6T，甚至更高速率

(3) 冷却方式从风冷→冷板→浸没式

(4) 数据中心功率从数 MW→数百 MW→向吉瓦级跨越

传统数据中心架构将无法满足未来大模型、AGI 及智能体经济的能效与扩展需求，全球范围内正在出现以“能源—算力一体化”为核心的新型数据基础设施形态。AI 计算架构的设计必须从“芯片级”、“服务器级”提升到“机柜级”（Rack-level）、“系统级”（System-level）方向演进，如下图所示。



引言 - 图 1

2. 智算中心的中国特色

中国在探索超大规模智算中心形态方面，具备一系列独特条件：

（1）能源结构与算力布局协同推进

依托“东数西算”工程，中国已在跨区域算力布局、电力调配和网络协同方面积累了系统级工程经验。西北、华北等地区具备集中连片的风电、光伏等新能源资源，为超大规模智算中心提供了可持续、低碳的能源基础；同时，高压直流输电与电网调度体系，为算力与能源的跨域协同创造了现实条件。

（2）制造业与关键部件供应链优势

中国在服务器整机、液冷系统、连接器、光模块、电源与功率器件等关键领域，已形成较为完整且具备国际竞争力的产业链体系。这一优势为高功率密度、高带宽互连、快速规模化交付提供了重要支撑。

（3）应用需求驱动显著

中国 AI 应用呈现出显著的多样性特征，既包括互联网和云计算场景，也涵盖制造、能源、交通、金融等产业智能化需求。这种多场景并行发展的格局，使得智算基础设施必须具备高度通用性、可扩展性与长期演进能力。

（4）政策与示范工程牵引

各级政府将人工智能与算力基础设施视为重要战略方向，通过政策引导、示范工程和产业协同，加速新型智算中心建设，为系统级创新提供了现实试验场。2026 年开始，中国出台了“算电协同”相关的政策。

上述因素共同决定了，中国在推进 AI 基础设施演进过程中，不仅需要“更大的数据中心”，更需要一套面向吉瓦级规模的系统性架构方法。

3. 范式转变

吉瓦级智算中心并非传统数据中心在规模上的线性放大，而是一种系统范式的根本转变，其核心特征体现在：

（1）设计尺度提升

架构设计从服务器级、机柜级，上升至园区级乃至区域级。

（2）跨域协同成为前提条件

能源系统、计算系统与网络系统不再是相互独立的子系统，而是需要在规划阶段进行协同设计与联合优化。

（3）系统稳定性与可运维性优先

在超大规模条件下，可靠性、可维护性与生命周期成本，往往比单点性能更具决定性意义。

（4）开放性与标准化的重要性提升

系统复杂度的提升，使得封闭、定制化方案难以长期维持，开放接口与标准协同成为降低系统性风险的关键手段。

因此，吉瓦级智算中心的建设，本质上是一个系统工程问题，需要在架构层面对技术、能源、网络、运维和生态进行整体统筹。

4. 愿景目标

基于上述背景，GW-scale Open AIDC 工作组提出如下总体愿景：

（1）构建面向吉瓦级规模的开放参考架构

通过系统化分层设计，明确关键能力边界与接口方向，降低跨主体协作成本。以开放体系推动中国智算基础设施的标准化和规范化；并推动供应链再全球化。

（2）推动能源与算力协同的绿色发展模式：

将能效与可持续性作为系统级设计的核心约束，实现算力增长与绿色目标的协同演进，推动智算力的可持续发展。

（3）支撑多样化 AI 应用场景长期演进

满足大模型、智能体及产业智能化等多类应用对算力系统的差异化需求。

（4）促进中国实践与全球标准体系的协同共建

在对齐 OCP 全球方法论的基础上，总结中国工程经验，推动形成可被国际理解和采纳的系统性方案。也即，既能“引进来”，也能“走出去”。

第一章 基础设施层

1.1 建筑与空间规划

吉瓦级智算中心（AIDC）建筑与空间规划是面向**吉瓦级总功率、高功率密度集群、能源-算力一体化**的园区级土建设计与统筹，核心聚焦总平面布局、建筑单体空间、模块化适配、消防安防及远期演进五大维度，为跨区域、规模化、标准化智算基础设施提供土建空间设计的基准。

1.1.1 规划核心原则

吉瓦级智算中心建筑与空间规划突破传统 MW 级数据中心设计逻辑，以系统设计为中心，以**园区级 AIDC 为尺度，**确立六大核心原则，所有原则均聚焦建筑空间形态、布局、尺度与合规性。

1.1.1.1 政策适配原则

全面对接国家新型数据中心建设、“东数西算”算力枢纽布局、“算电协同”政策要求、双碳目标及地方算力基础设施专项政策，建筑空间规划满足绿电消纳、高压直流供电接入、液冷系统落地、节能降耗等政策硬性要求。建筑等级、防火、抗震、疏散等核心指标严格遵循 A 级数据中心国标规范，确保吉瓦级园区从规划阶段即符合全国及区域算力基础设施建设的合规要求，适配国家算力网络战略落地的空间载体需求。

1.1.1.2 安全优先原则

以 A 级数据中心可靠性为建筑空间设计底线，从总平面布局、单体结构、功能分区、消防疏散等维度构建全维度安全空间体系。针对吉瓦级园区 **220kV 高压供电区、大容量储能区、液冷管网区、高密算力集群区**四大高风险区域，通过物理隔离、空间分隔、荷载强化、防泄漏/防淹空间预留等土建手段，筑牢超大规模园区的空间安全基底。

1.1.1.3 高功率密度适配原则

建筑空间全维度适配吉瓦级智算中心**单机架 140kW 以上、园区总功率≥1000MW** 的高功率密度特征，核心聚焦结构荷载、室内净高、柱网尺度、管线空间四大土建指标。通过放大建筑空间尺度、强化结构承载能力、预留竖向管线空间，满足宽体机柜、液冷 CDU 单元、机柜级备电设备的安装与运维空间需求。

1.1.1.4 网络拓扑架构的适配和演进原则

建筑总平面与单体空间布局同步匹配智算中心高速网络拓扑架构，核心算力区、网络交换区、存储区采用**近距离物理集中布局**，通过优化空间距离降低网络互连损耗。建筑内部预留光纤配线、高速互连机柜、跨集群链路的专用空间与竖向管线通道，支持网络带宽代际升级的空间演进，无需后期土建改造即可适配网络架构迭代。

1.1.1.5 弹性灵活设计及扩展原则

采用**园区级整体规划+单体模块化分割**的双维度弹性设计思路，建筑空间按 Pod 算力集

群单元划分，支持分期部署、按需扩容。建筑结构、供配电管线、冷却管网、消防系统均采用**池化预留设计**，空间布局可灵活适配训练/推理集群动态调整、风冷/液冷混合部署的空间切换需求，避免一次性建设造成的空间浪费，提升吉瓦级园区的空间利用率与运营灵活性。

1.1.1.6 绿色集约高效原则

以降低 PUE、WUE 为建筑空间规划核心目标，总平面布局优先利用自然冷源，优化冷却系统与算力区的热量传输空间路径。建筑围护结构采用高效隔热保温设计，减少制冷能耗损耗；功能分区遵循**集约紧凑**原则，严格控制辅助配套区面积，提升土地与建筑空间利用效率。结合西部无水区域、南方有水区域的地理特征做差异化空间设计，实现绿色低碳与空间高效的双重目标。

1.1.2 选址与场地规划

该规划是吉瓦级智算中心建筑空间的前置基础，核心解决**园区落地条件、总平面布局、竖向设计、管网统筹**四大土建问题。

1.1.2.1 选址基本要求

1.核心资源适配：优先选址于“东数西算”国家算力枢纽节点，场地具备园区级电力接入条件,具体接入容量与电压等级见 1.2 篇章；靠近新能源富集区域；保障通信干线通达，满足跨区域算力调度的网络空间传输需求。有水区域具备稳定工业用水指标，无水/少水区域预留去水化冷却系统安装空间。

2.环境安全隔离：场地远离粉尘、有害气体、腐蚀性及易燃易爆场所，避开地质灾害隐患区，从源头保障建筑场地结构安全。

3.干扰规避：场地远离强振源、强噪声源与强电磁场干扰区，与民用建筑、公共设施保持合规安全距离，满足智算中心安静运营的空间环境要求。

1.1.2.2 场地规划基本要求

1.建筑布局

建筑布局可采用主机房楼+办公配套楼模式（下称传统模式）或建筑集群模式。建筑集群模式是指将传统模式中主机房楼的机房与动力等功能房间拆分了独立楼栋，以建筑集群模式布局，当采用该模式布局时，总平面宜采用向心式建筑集群+物理隔离模式，即机房楼位于园区中间，动力楼置于 IT 楼四周，从而形成标准化的土建空间结构，分为两种典型布局模式：

1）高层单体向心布局：以单栋高层机房楼为园区几何中心，四周对称布置动力楼，形成“一核四辅”的对称结构。该布局下，动力楼与核心机房楼的管线路径均等、空间距离最短，可大幅降低电力、冷却、通信管线的土建敷设长度。

2）多层组团向心布局：以 2-3 栋多层机房楼为园区核心组团，周边环绕布置动力楼，形成“多核多辅”的扩展结构。该布局支持分期建设，可根据算力需求新增机房楼与动力楼组团，无需调整园区主干管线走向，适配吉瓦级园区的弹性扩展原则，同时保障各机房楼与动力楼的空间距离一致，实现运维与管线布局的标准化。

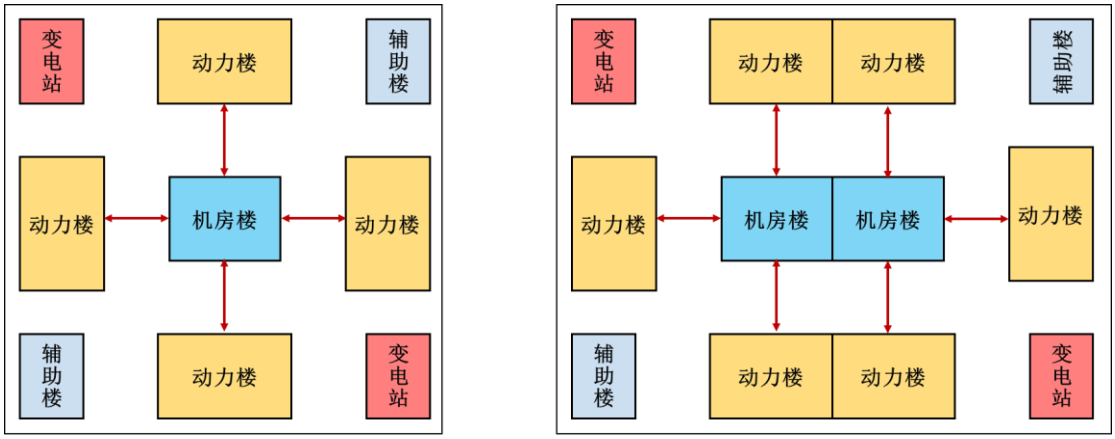


图 1.1 向心式建筑集群布局示意图

2. 出入口及流线

园区设置独立人流、车流、物流出入口，实现三分流管控。主入口预留大型预制模块、重型机柜运输转弯与回车空间；消防车道宽度≥4m、转弯半径兼顾货运按≥15m，沿园区形成贯通式消防环路。

3. 场地竖向

场地排水坡度≥0.5%，杜绝积水；通过地势、百年洪水位等信息合理确定室内外高差，台风地区适当增加室内外高差及防水倒灌措施。

4. 综合管网

统筹高压供电、液冷/氟循环、通信光缆、消防、余热回收管线，采用综合管廊集中敷设。管网走向与建筑布局、网络拓扑匹配，预留扩容接口，避免后期改造破坏土建结构。

1.1.3 核心建筑单体设计

核心建筑单体是吉瓦级智算中心算力承载、能源保障、冷却运行的核心载体，其设计直接决定园区运营效率、安全可靠性与长期演进能力。本小节聚焦**功能分区、空间尺度、结构、围护、装修**五大核心土建维度，结合国标规范与吉瓦级园区实践，细化设计要求、明确设计依据，为吉瓦级算力、能源、冷却、运维系统提供专属、合规、可落地的建筑空间支撑。

1.1.3.1 功能分区

主机房楼宜采用**“主机房居中、配套围绕周边”**向心式单体布局模式。按功能划分为四大核心区域，各区域空间独立、边界清晰，各区域细化设计要求及依据如下：

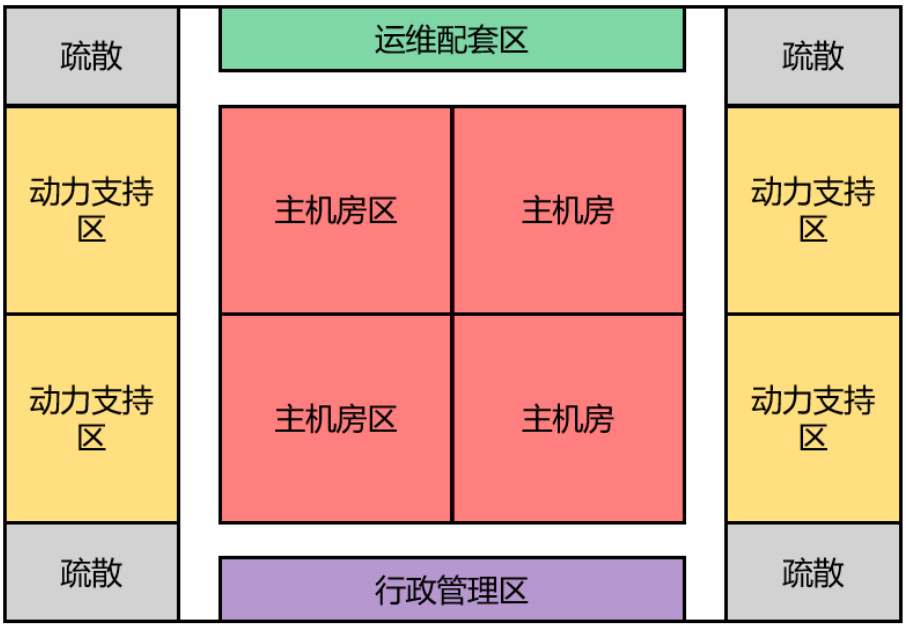


图 1.2 向心式单体布局示意图

1.核心主机房区：作为算力建筑的核心空间，占单栋机房楼总面积的 20%-30%，主机房是高密机柜集群的核心承载区域，布局需考虑靠近动力支持区，缩短管线传输路径；空间需规整，适配宽体机柜、整机柜超节点的批量部署，预留机柜排列间距与运维操作空间。

2.动力支持区：可采用独立建筑或机房楼内独立分区布局，采用独立动力楼时需与算力楼保持防火间距。在建筑内时，支持区面积占单栋建筑总面积的 40%-50%，随单机柜功率扩大，后期支持区面积相应扩大，柴油发电机组可采用室外集装箱式。

3.运维配套区：面积占单栋建筑总面积的 8%-10%。满足日常运维、监控、设备调试与故障处置、备品备件存储的空间需求。

4.行政管理区：按需求设置，面积占单栋建筑总面积的 5%-10%，办公区域需与核心算力区、能源保障区物理隔离，避免干扰。

1.1.3.2 建筑设计

1.单体规模：吉瓦级园区单栋核心机房楼建筑面积≥20000 ㎡，按 Pod 算力单元模块化分割，单 Pod 空间适配 50MW－100MW 功率负荷部署。单栋建筑层数宜控制在 3-5 层，层数过多会增加垂直运输成本与冷却管网损耗，过低则占用土地资源，建筑长度不宜超过 120m，宽度不宜超过 60m，便于消防疏散与设备运维，同时适配模块化预制构件的运输与安装，符合绿色集约高效原则。

2.平面布局：核心算力区当采用非浸没式制冷方式时，宜采用行列式+冷热通道封闭布局，机柜面对面、背对背规整布置，形成明确的冷热通道，其中冷通道宽度≥1.5m，热通道宽度≥1.8m，便于冷量循环与热量排出；液冷机柜与 CDU 单元宜就近布局，缩短二次侧管路空间距离，减少冷量损耗与管路阻力，液冷机房、CDU 设备区采用局部抬高找坡排水，配套防淹导流空间。建筑内部应预留专用管线井，集中敷设供电、冷却、网络管线，避免管线外露影响运维与空间利用，管线井尺寸≥1.2m×1.2m，便于管线安装与检修。

3.层高与柱网：建议智算中心每层层高 7.2m，预留液冷机柜（高度约 2.5m）、上走线

桥架（高度约 2.9m）、CDU 设备（高度约 2m）的竖向安装空间，同时建议预留 10%-15%的竖向冗余空间，适配后期设备升级需求。柱网间距选择需结合机柜排列方式，确保每排机柜数量合理、运维空间充足。

4. 围护结构：外墙应采用 A 级防火保温材料，优先选用高效保温材料，降低制冷能耗。液冷区域应增设二次防渗漏构造层，防止冷却介质泄漏损坏建筑结构。门窗应采用高气密性、防尘、隔音设计，减少热量传导与外界干扰，满足机房洁净度要求。

5. 室内装修：全域应采用 A 级不燃性、防静电、防尘、防潮的装修材料，杜绝火灾隐患与粉尘污染。

地面应采用防静电地板材料，液冷区增设防滑与漏液导流槽，导流槽连接集液池，便于泄漏介质收集与处理。墙面应采用防火、防潮、防尘的乳胶漆或彩钢板，颜色宜选用浅色系，可增强室内采光。

1.1.3.3 结构设计

1. 结构设计标准：

对于新建吉瓦级园区建筑单体，建筑结构的安全等级宜按：一级；抗震设防类别：新建建筑不低于乙类（重点设防类）、对于改造建筑不低于丙级（标准设防类）；配套建筑（构筑）物结构设计标准应与主机房一致；

2. 合理的结构布局：

1）新建吉瓦级智算中心和其配套建筑单体，均不应采用单跨结构或局部单跨框架为主的结构型式；

2）鉴于吉瓦级智算中心荷载需求较大，净高需求也较高，结构柱网可以根据机柜布置特点，采用 X\Y 向取不同跨度的方式制定方案，一个方向尽量加大跨度，满足最大布设机柜列数，另一个方向则宜适当减小；长跨方向可加密次梁间距，短跨方向作为主受力梁，宜加大框架梁截面高度；最优跨度模数可以选择 9.6mx6.0m（短向）；12.0mx7.2m（短向）等；

3）结构缝布局：主机房内严禁穿越结构缝；整体建筑宜通过采取措施，尽量避免结构缝设置；

4）根据吉瓦级机房荷载大、跨度大、层高大、投资大的特点，结构应考虑采用减隔震设计；减隔震即针对建筑整体，还需对质量较大的单机柜采用隔震垫等措施；或在架空层进行整体隔震；以上不同系统的减隔震措施至少应考虑采用一种；

5）相对重型的设备宜布置在建筑下层，相对轻型设备置于建筑顶层；

6）对于屋面结构，根据工程经验，采用结构找坡的方式较为有利，可以优化出荷载裕度贴补到屋顶大设备上；

3. 荷载取值：

1）新建吉瓦级智算中心进行结构计算时，各主要功能房间的楼板活荷载标准值不宜小于以下标准：

液冷 CDU（主机房；单机柜重量 2 吨时）≥16kN/㎡；

液冷 CDU（主机房；单机柜重量 3 吨时）≥20kN/㎡；

柴发机组（一般设于室外单独建设）楼面活荷载≥24kN/㎡；

UPS 室 10kN/㎡； 钢瓶间 8kN/㎡；

电池室、储能设备区楼面活荷载：24kN/㎡；

配电室：16kN/㎡；

空调区：8kN/㎡；设备运输通道:12kN/㎡；

机房区、柴发区宜考虑动力系数 1.05；

机房区域板底吊挂荷载 2.0~4.0kN/㎡；

屋面大型设备（干冷器、氟泵、冷却塔等）宜按实际重量参与结构计算；

2）新建吉瓦级智算中心活荷载组合值系数、频遇值系数、准永久值系数可以按《数据中心设计规范》附录一计取；计算地震作用时，以上可变荷载的组合值系数按下：

主机房区/空调区/变配电： 0.9；

电池室/UPS 室/管线吊挂/柴发：0.9~1.0

设备运输通道：0.7

3）鉴于吉瓦级智算中心荷载水平较大，但机柜和设备等平面布置均有一定的间距要求，针对结构不同部位构件计算时，可以参考《通信建筑工程设计规范》表 8.2.2：大于 8KN/㎡荷载的房间，在对主框架梁、柱、墙及基础构件其活荷载折减系数可以取 0.75；在对于次梁构件其活荷载折减系数可以取 0.85；楼板计算不再折减；

4）对于既有建筑改造工程，可以不受上述活荷载取值限制。设计时，可以按设备的实际重量和具体定位，以板面局部荷载的方式输入计算；但考虑检修安装和液体渗漏风险，应适当叠加计入少量均布活荷载；

5）鉴于机房建筑的设计工作年限一般会大于设备的寿命周期，随着数据产业技术更新迭代加速，在可行的情况下，建议在设计时对楼层使用活荷载取值适当放宽，尽量计取上限标准，以便于建筑在工作年限内可以灵活改动布置；

4.材料选用：

鉴于吉瓦级智算中心荷载水平较大，当采用钢筋混凝土结构时，结构主要承载构件宜采用 HRB500\HRBF500 级钢筋；当采用钢结构时，宜采用 Q355 及以上的高强钢材；

5.结构设计及构造要求：

1）当机房有架空需求时，架空结构的设计标准和荷载取值应与主体结构保持一致；架空（钢架）结构也需要进行结构承载力极限状态验算和正常使用极限状态验算；

2）重大设备通过架空结构作用到主体结构楼板时（集中力），宜补充核算楼板的抗冲切承载力；

3）结构计算应充分考虑架空层构造对主体结构的影响，即：地震作用的质点实际位于架空层顶面，从而建筑主体框架柱的箍筋加密区定义就应计入架空高度；必要时，架空结构与主体结构共同建模，整体计算；

4）管线抗震支吊架应通过计算确定构件尺寸和布置间距；

5）鉴于吉瓦级智算中心结构层高较高，楼梯的布置一般至少为 3 跑或 4 跑以上，结构计算时，应特别注意楼梯自重及楼梯活荷载应按相应梯跑数量记取为双倍或 3 倍；

6）如吉瓦级智算中心有整机交付需求时，机房内运输通道区域需布置升降机，与升降设备相关的结构构件应考虑动力系数 1.05~1.2；

7）根据变配电室设计要求，设置有变压器的区域应设吊装口及吊钩，吊钩应布置在结构梁上，吊钩梁应计入吊装工况荷载；吊装口尺寸应充分考虑操作缓冲空间；

8）机房为钢结构型式时，应按相应防火标准，进行防火和防腐设计。

1.2 供备电系统

1.2.1 吉瓦级智算中心电力需求特征

1.2.1.1 AI 训练与高密度功率机柜的电力需求

吉瓦级智算中心已成为超大规模算力基础设施的核心建设形态，其电力需求与传统兆

瓦级数据中心存在本质差异，核心体现为单机柜功率超高化、负载特性动态化、供电规模吉瓦化、冗余需求精细化，是供配电架构设计的核心原点与底层依据。

智算中心的负载率波动明显高于传统数据中心，训练任务启动时负载可从低峰快速攀升至高峰，任务结束后负载快速回落，当大规模机柜同时启动或训练任务同步进入峰值阶段时，会对供电系统造成大规模、高幅度的瞬时功率冲击，因此要求供配电系统具备极强的动态负载适配能力、快速调压能力与负荷均衡调度能力，从架构底层实现对瞬时冲击的有效平抑。

1.2.1.2 供备电系统整体规划

吉瓦级智算中心的电力容量规划需适配 AI 算力的高速增长特性，遵循“超前规划、分期建设、灵活扩容、冗余可靠”的核心原则。智算中心园区配套外电接入容量应满足核心负载全量供电需求，预留扩容空间，适配未来 3-5 年内算力需求的阶梯式增长。

如图 1.3 所示，吉瓦级智算中心的供备电系统按照电压等级从高到低依次梳理，并结合现场部署区位、完整供电链路、系统功能定位进行章节划分，分板块逐一阐述系统架构、配置特点与运行逻辑。其中，园区内主干互连层是吉瓦级智算中心电力传输的主动脉，承担着从 220kV 到 10kV 或 35kV 的电力变换；楼宇分支互连层是连接园区主干环网与各机房楼的分支传输网络，承担着从园区集中变电站到各机房楼内配电室的配电功能，采用 10kV 或 35kV 电压等级，再通过能量路由器或固态变压器 SST 将交流转化为 800V 直流；机房内配电互连层是机房楼内的主干配电网络，承担着从机房楼配电室到各机柜分区电力配送功能，采用 800V 直流电压等级。机柜内供电将在后续章节详细阐述。

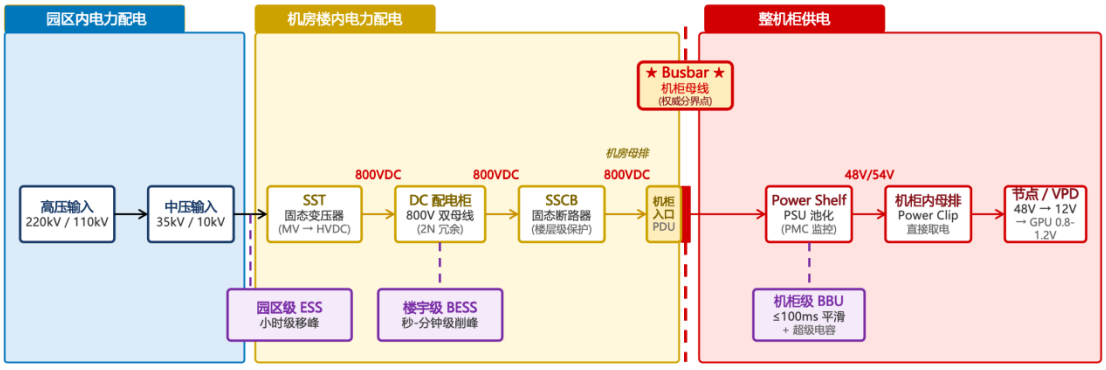


图 1.3 供备电系统环节示意图

1.2.2 园区级高压供电系统

吉瓦级智算中心园区通过构建源网荷储协同联动的一体化总体架构，实现电力的可靠供应、高效传输、灵活调度与低碳消纳。基于多电源协同供给体系，逐步提升风光水核等新型能源占比，并配置储能提升绿电消纳比例，降低碳排放。

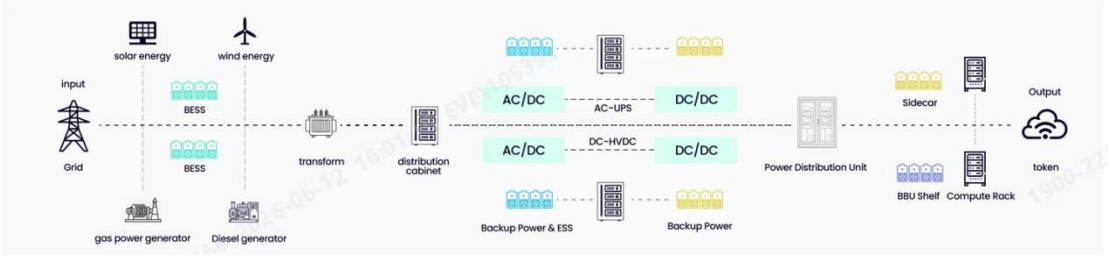


图 1.4 多级储能协同布局示意图

如图 1.4 所示，储侧构建“园区级-楼宇级-机柜级”三级储能协同布局体系，园区级配置长时储能系统，承担电网调峰、长时间备电、虚拟电厂运营核心功能；楼宇级配置短时储能系统，承担峰谷套利、电能质量治理、短时备电功能；机柜级 BBU 可承担短时能量缓冲、瞬态功率支撑和备用电源过渡支撑等功能，形成全时段、全场景的储能保障与调度体系。

通过园区级能源管理系统实现数据互通、协同调度，在正常工况下实现电力的最优分配与能效最大化，在电网故障工况下实现负荷的不间断保障，在低碳运营场景下实现可再生能源的最大化消纳和电力交易运营，全面适配吉瓦级智算中心的大容量、高可靠、低碳化运营需求。

1.2.3 园区中压配电系统

1.2.3.1 35kV vs 10kV

吉瓦级智算中心占地面积广、机房楼分布分散、单机房楼负载容量大，中压配电系统需兼顾大容量远距离传输的低损耗、多负载分区的灵活调度、故障场景的快速隔离三大核心需求。虽然目前仍以 10kV 配电为主，但 35kV 作为国内中压配电的主流电压等级，呈现出差异化的优势。35kV 电压等级核心优势集中体现在大容量、低损耗、远距离传输三大维度，在相同载流量下，35kV 电压等级的传输容量是 10kV 的 3 倍以上，大幅节省电缆、管廊、开关设备的投资成本。

1.2.3.2 中低压供电架构

吉瓦级智算中心园区采用“园区集中式主变配电站+机房楼分布式附设配电室”的两级布局模式。中压侧可采用双母线分段接线方案，双母线分别来自园区 35kV/10kV 的不同段，互为备用，低压直流侧采用双母线接线，每段直流母线分别由独立的整流器组供电，形成双路冗余直流供电架构，直接为机房楼供电。

1.2.3.3 储能替代部分柴发方案

针对吉瓦级智算中心的备电要求，采用“长时储能为主、保留部分柴发应急备用”的备电架构，为园区核心 IT 负载配置 2-8 小时储能系统，承担外电中断场景下的优先备电功能。配置 50%核心负载容量的柴油发电机，仅在长时间电网故障、储能系统极端故障等极小概率场景下启动。

储能采用预制舱式模块化形态，市场中大容量电池在朝着单体容量大于 600Ah 方向发展，在高能量密度基础上，可以做到极简集成，提高系统可靠性、安全性，经济性，是吉瓦级算力中心的理想储能路线。

1.2.4 机房楼内配电系统

1.2.4.1 供电架构发展路径

传统 380V 低压交流供电架构供电链路较长，能量转换级数多，在高功率场景下或面临效率瓶颈，行业逐步引入 240V/336V 直流方案，通过直流母线直接为服务器供电并实现电池直挂。

通过进一步提升输出电压等级，如 800VDC 或±400 VDC，数据中心在满足更高功率需求的同时有效降低传输电流与线路损耗，提升了系统效率。同时，高压直流架构还具备与光伏、储能等直流能源系统直接耦合的天然优势，使其从系统层面对传统供电模式进行重构，成为面向高算力密度、高能效与绿色可持续发展数据中心的重要演进方向。

1.2.4.2 全直流供电架构

当前较具潜力的 800VDC 或±400VDC 供电架构是通过能量路由器或固态变压器 SST，将 10kV 中压交流直接转换为稳定可控的 800VDC 或±400VDC，再通过高效 DC-DC 模块为服务器提供 48V 或更低电压。这一架构显著缩短供电链路，减少交直流转换级数，在效率方面，传统 UPS 和 HVDC 整体效率通常在 95%-97.5%之间，而采用 SST 的 800VDC 或±400VDC 架构，实测效率可达 98%以上，年化节能效果显著，系统端到端效率提升约 3%-5%。

在材料方面，高电压、低电流的全直流供电方案用铜量约为传统交流方案的三分之一；在功率密度与空间利用方面，通过减少供电设备数量，占地面积减少 40%-50%；在建设难度方面，模块化、预制化系统，大幅缩短建设周期；在能源方面，新能源和储能可以直流形式接入，减少不必要的交直流转换环节。

1.2.4.3 HVDC 储能配置

基于 HVDC 的储能需满足 5－15 分钟（6C－10C 放电倍率）短时备电需求，用于支撑备用电源切换过渡和负载瞬态功率波动。随着园区规模和单机房楼功率密度持续提升，储能系统需要在有限部署空间内提供更高备电容量，因此对高能量密度、高倍率、高可靠、高安全及智能监测能力提出更高要求。

铅酸电池已难以满足当前需求，磷酸铁锂技术路线凭借储能行业大规模应用基础，特别是叠片技术的发展，具备高倍率放电、低温升、高安全、状态实时监测等优势，是数据中心大规模短时备电的主要技术方向。

1.2.5 运行保护与控制系统

1.2.5.1 多层级保护策略

全直流供电系统的保护核心难点，在于直流电流无自然过零点，短路故障时电弧不易熄灭，且故障电流上升速度极快。固态断路器凭借微秒级的分断速度、无电弧分断的核心优势，成为全直流供电系统保护的核心设备，结合固态断路器构建“园区级-楼宇级-机柜级-设备级”的多层级全链路保护策略，实现故障的快速隔离与最小化影响。

从设备级到园区级，保护动作时间设置逐级递增的时间梯度，确保下级保护优先动作，只有当下级保护拒动时，上级保护才作为后备保护动作；各级保护均配置故障方向判别功能，仅对保护范围内的故障动作，对保护范围外的穿越性故障不动作。通过园区 SCADA 系

统，实现各级保护装置的数据互通与协同联动，故障发生时，可精准定位故障点，同步下发指令，实现故障的精准隔离与负荷自动转移，进一步提升保护系统的可靠性。

1.2.5.2 智算中心智能控制系统

通过供配电系统全生命周期数字化管理体系，与园区级能源管理系统结合，实现从“被动运维”向“主动运维、预测性运维”的转型，精准还原全链路供电架构，实现物理系统与数字模型的实时映射、双向互动，核心功能与应用包括全系统实时可视化映射、多场景模拟仿真与方案验证、远程操控与无人值守运维、综合能源管理及电力交易运营等，全面提升系统运行的可靠性、运维效率与能效水平，同时在低碳运营场景下实现可再生能源的最大化消纳和电力交易等额外经济性功能。

1.2.6 结论与展望

结合现有行业技术趋势、发展主线及技术短板、工程落地需求，可将吉瓦级智算中心供备电系统的核心技术趋势总结为六大方向，覆盖架构升级、安全保护、储能备电、标准建设、智能运维、低碳转型 。

1.高压全直流高效供电

800V 全直流供电全面替代传统交流 UPS 供电，以能量路由器 SST 为核心，简化供电链路、提升系统效率与功率密度，适配兆瓦级单机柜供电需求。后续将探索宽禁带半导体器件在供电整流、保护环节的应用，规模化验证 35kV 高压交流直转 800VDC 技术，持续优化园区、机柜、芯片多级全链路供电架构，全方位提升供电传输效率。

2.高压直流精细化安全保护

直流保护技术正向微秒级、全链路、智能化方向迭代，固态断路器将实现全链路精细化保护、直流短路电弧熄灭、多级保护协同等核心功能。后续将聚焦 800V 及以上高压直流场景，重点研究短路故障快速检测、无电弧分断、接地故障精准定位技术，优化多级保护智能协同策略，完善高压直流供电安全防护体系与行业技术标准。

3.多功能储能备电集成

储能系统已从单一应急备电，迭代为“备电+调峰+电能质量治理+虚拟电厂+绿电消纳”多功能融合体系，“长时储能+备用电源”模式实现备电系统低碳化、高经济性。后续将开展混合储能系统规模化应用研究，优化容量配置与协同调度策略，搭建储能多层级安全防护体系，提升备电系统本质安全水平。

4.供配电互连系统标准化

针对全直流供电架构，行业正加速推进统一互连标准建设，规范设备接口、传输线路、连接器等核心部件参数与设计，实现全产业链设备互联互通、兼容适配，降低工程应用成本。目前已开展全直流供配电系统标准化研究，后续将同步制定 800V 直流电源、母线槽、连接器等设备的统一技术参数、测试标准与设计规范，解决行业设备接口不统一、兼容性差的痛点，推动全直流技术规模化落地。

5.全场景智能运维调控

行业依托数字孪生、AI 大模型、物联网技术，逐步搭建供配电系统全生命周期数字化管理体系，实现可视化监控、预测性运维与智能协同调度。后续将重点突破数字孪生与物理系统的实时闭环控制技术，研发 AI 大模型驱动的智能调度算法，最大化系统运行效率与经济效益；深化故障智能诊断、预测性维护技术研究，助力供配电系统实现零故障运行目标。

6. 打造零碳化供配电系统

供配电系统与可再生能源深度融合是核心发展趋势，通过源网荷储、绿电直连、节能高效设备、清洁备电等，持续降低系统损耗与碳排放，打造零碳供配电体系。后续将持续优化新能源与算力中心的深度协同，提升绿电消纳能力和协同控制能力，构建全链路零碳排放供配电体系，推动吉瓦级智算中心全面向零碳算力设施转型。

1.3 吉瓦级智算中心冷却系统

智算中心冷却系统是保障 IT 设备稳定运行、控制能耗成本、提升核心竞争力的关键基础设施，其核心目标是将服务器运行过程中产生的大量热量高效转移至外界，维持服务器核心器件（CPU、GPU 等）处于安全稳定的运行区间，同时实现能效最优、成本可控。随着 AI 算力需求爆发，单机柜功率密度从传统 8-15kW 飙升至 30-60kW，随着英伟达 GB300 等架构的落地，单机柜散热功耗突破了 150kW，甚至 Rubin Ultra 等全液冷架构，使单机柜功耗达到 600kW，智算中心能耗也从兆瓦（MW）级逐渐向吉瓦（GW）级迈进，基于此背景下，针对吉瓦级智算中心的冷却散热系统架构已从传统风冷向液冷主导、多技术融合的方向迭代。

1.3.1 冷却系统架构与核心设备

当前智算中心行业主流散热架构分为以下三大技术路径：

a) 强制风冷架构：通过风扇对服务器进行散热，适配单机柜 12-50kW 功率密度，优势是成本低、运维简单，劣势是噪音大、能耗高。

b) 冷板式液冷架构：核心为贴合 CPU、GPU 等核心发热器件的冷板，通过循环冷却液直接吸收核心器件产生的热量，经过 CDU 换热，最终通过冷源设备传导至外部大气。该种散热方式兼容性好，兼顾散热效率与运维难度，是目前高功率密度数据中心的首选方案。

c) 浸没式液冷架构：将服务器全浸没于绝缘冷却液中，散热效率最高，无局部热点问题，噪音低。随着冷却液国产化技术成熟，系统工艺优化、供应链健全与商业化应用，单相浸没液冷的经济性正在逐步凸显。目前浸没式液冷的商业化部署节奏正在快步前进，处于规模化部署的前夜。

从上述各类架构的综合对比来看，冷板式液冷系统解决了传统风冷散热效率不足、能耗偏高的痛点，而浸没式液冷尚未达到规模化部署，产业化及规模化尚未形成。基于冷板式液冷系统的主流定位与广泛应用价值，接下来将全面拆解其系统架构、运行原理、核心设备及工程验证，为实际应用与落地提供参考。

1.3.1.1 系统架构

1.一次侧系统架构

一次侧液冷系统架构宜根据项目地所在的气象条件（包括干球温度、湿球温度）、水资源情况、液冷服务器的进液温度等条件进行选择。具体选择方式见下表。

表 1.1 一次侧液冷系统架构

一次侧液冷系统架构				
气候条件	有寒冷季节：自然温度低，可以满足全年自然冷源制冷		有温热季节：自然温度高，需要机械制冷补充	
水源情况	有水地区	缺水地区	有水地区	缺水地区

系统架构	开式冷却塔+板换、闭式冷却塔（如图 1.5）	干冷器（如图 1.6）、风冷冷凝器（如图 1.7）	冷水机组+冷却塔+板换（如图 1.8）	风冷冷水机组（如图 1.9）
经济性对比				
成本	低	较低	较高	高
能耗	低	较低	较高	高
运维难度	较低	低	高	较高

以下是几种架构的示意图：

图 1.5 开式冷却塔+板换或闭式冷却塔冷源方案

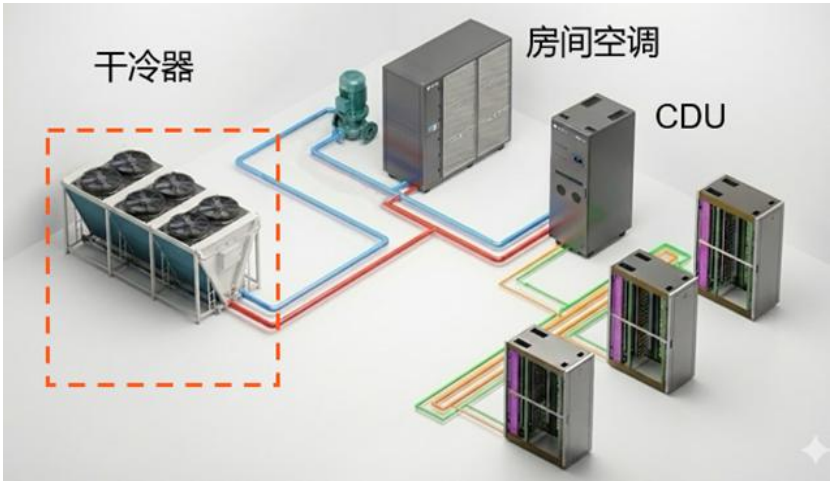


图 1.6 干冷器冷源方案

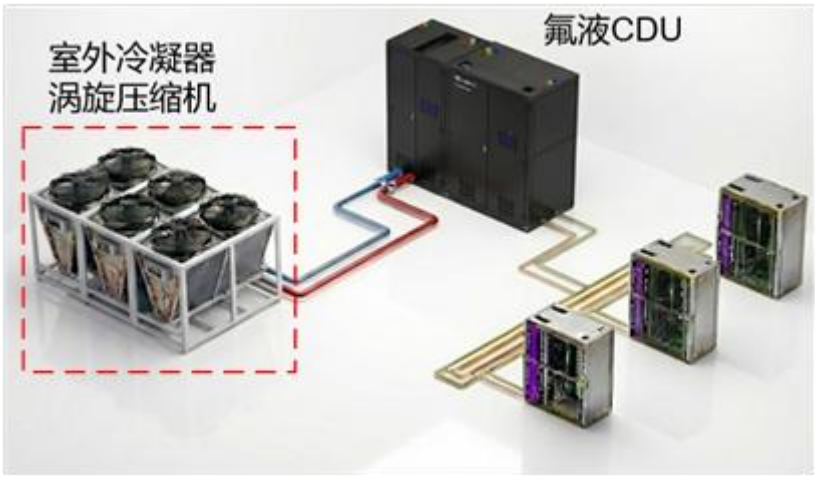


图 1.7 风冷冷凝器冷源方案

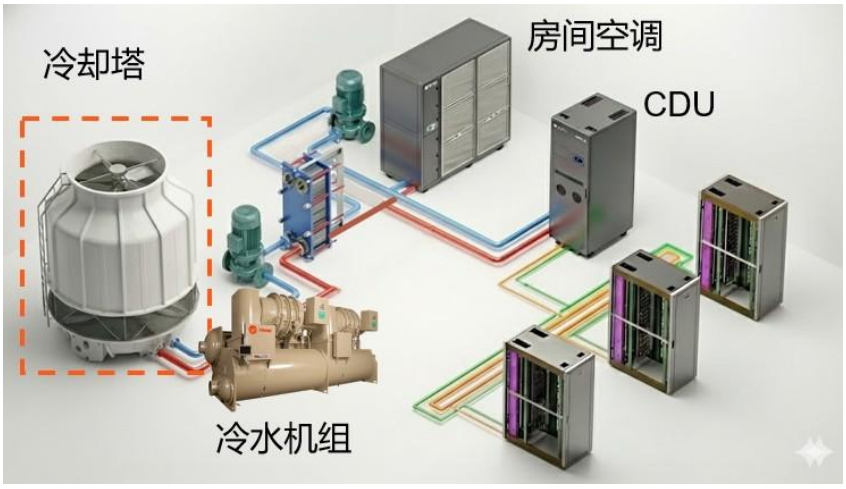


图 1.8 水冷冷水机组冷源方案

图 1.9 风冷冷水机组冷源方案

2. 二次侧系统架构

1） 集中式液冷系统架构

集中式液冷系统以 CDU 集中管控、管网统一输配为核心特征，主流常采用 2N、N+X 冗余设计，可适配不同算力场景的可靠性等级需求，为高密度算力设备的稳定运行提供冗余保障。

2N 架构（如下图）：双冗余、双总线架构，其核心是部署两套完全独立、对等且具备 100%负载能力的 CDU，两套系统在结构、电气及控制上完全独立又互为备份。常规工况下，两套 CDU 并行协同，共同承担 100%的系统负载；任意一套 CDU 发生故障，另一套可通过预设逻辑实现无中断自动切换，承担系统全部负载，确保算力设备稳定运行。2N 架构具备便捷的运维性，无需中断其中一套 CDU 的正常运行，即可对另一套 CDU 进行断电维护。

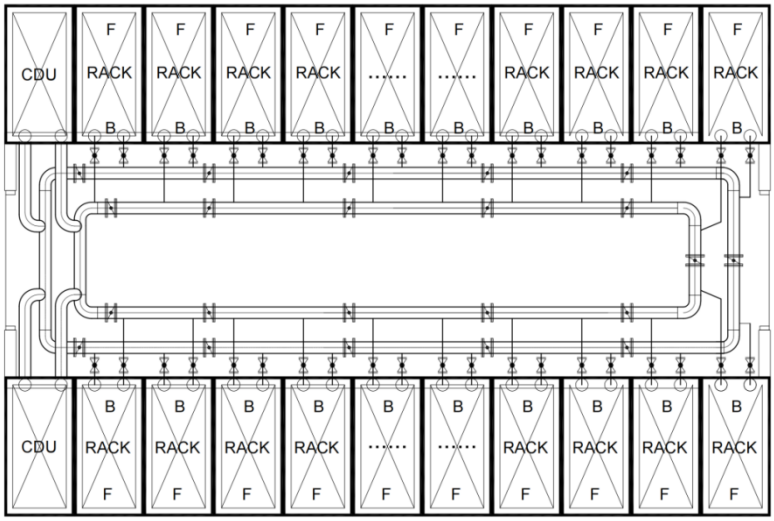


图 1.10 集中式液冷系统 2N 架构

N+X 架构（如下图）：模块化冗余架构，通过智能群控实现负载均衡与故障自动切换；N：满足系统总负载所需的最少 CDU 数量；X：冗余备用 CDU 数量。CDU 全部并联接入同一套供回液管网，共同参与系统冷量分配及智能控制。常规工况下，N 台 CDU 满载运行，X 台 CDU 处于热备状态；CDU 发生故障时，备用 CDU 可快速自动切入，保证算力设备安全运行。相较于 2N 架构，N+X 架构具备成本更优、线性扩容友好的优势。

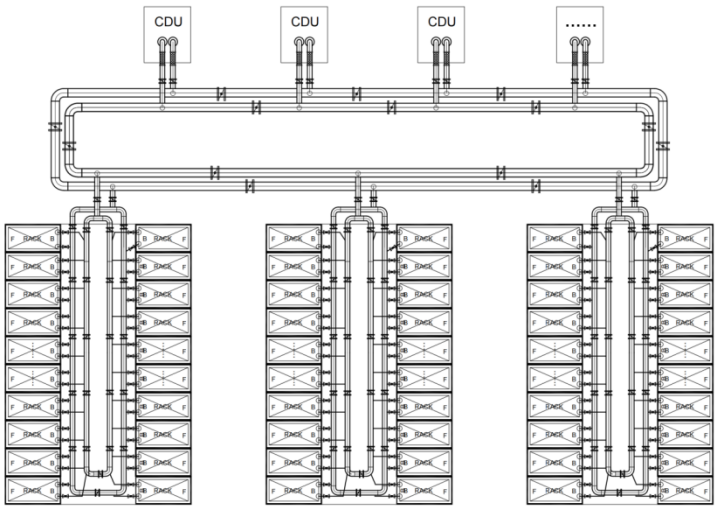


图 1.11 集中式液冷系统 N+X 架构

2N 架构聚焦核心算力场景的零中断需求，提供最高等级的可靠性保障；N+X 架构兼顾可靠性与性价比，两种架构均可实现标准化、模块化、高可靠的设计要求，为高密度智算中心的绿色低碳、高效稳定运行提供坚实的技术支撑。

2） 分布式液冷系统架构

分布式液冷系统是以 CDU 布置在机柜内部，通过独立的闭环回路，实现机柜级的精准散热为核心特征，特点为部署灵活，可按需扩展，可随服务器整机柜部署出货，且单柜故障不波及全局，隔离性较好。适用于高密度传统数据中心的改造机房以及空间要求较高的边缘计算节点。更多详情，参见 2.1 节 AI 超节点。

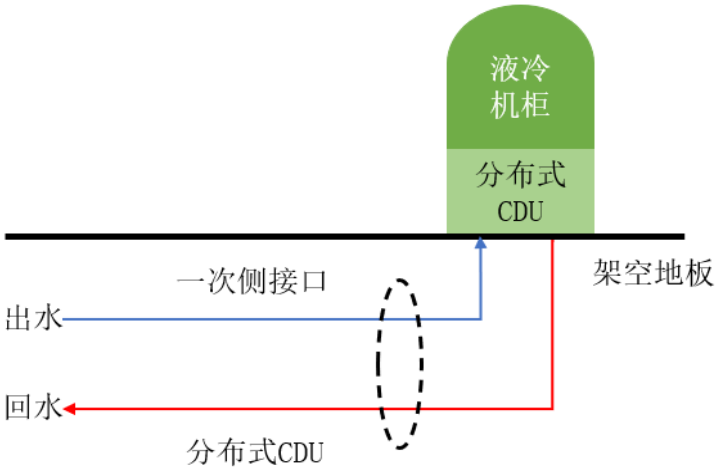


图 1.12 分布式液冷系统

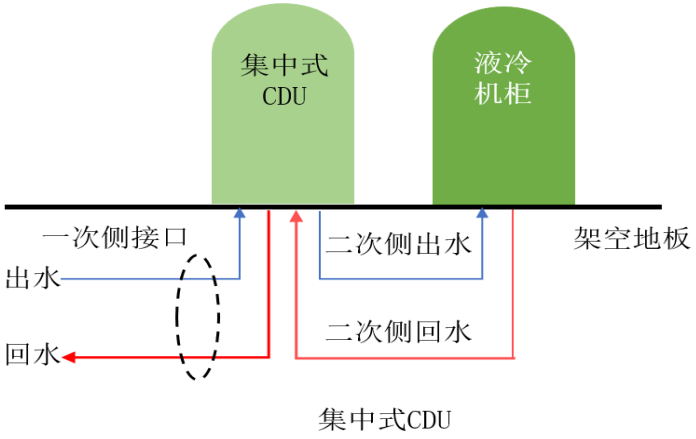


图 1.13 集中式液冷系统

典型的液冷整机柜布局如下图所示，包括计算节点、交换节点、电源插框、盲插 Manifold、Cable Cartridge 等。

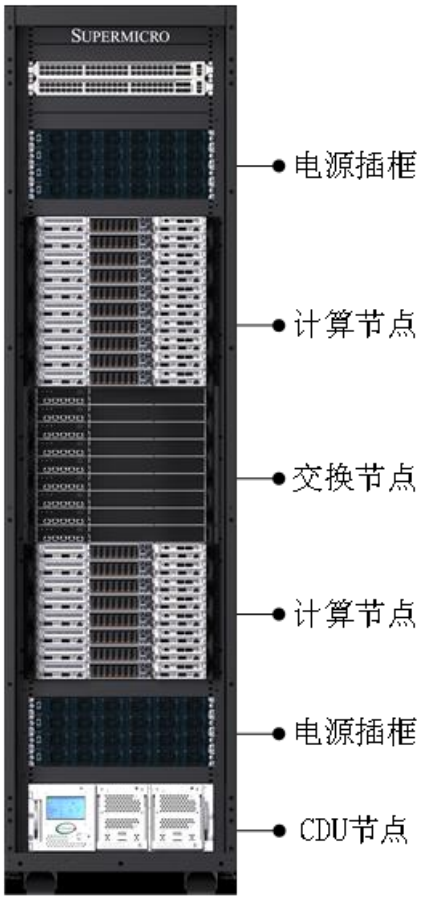


图 1.14 整机柜架构

1.3.1.2 控制系统

1.冷源自控系统的组成及架构

冷源自控系统主要负责一次侧循环设备和冷塔的监控，由冷源自控控制器、水泵流量计、压力传感器、温度传感器、电动阀、冷却塔传感器、CDU 监控、其他传感器及显示屏组成，系统架构如下：

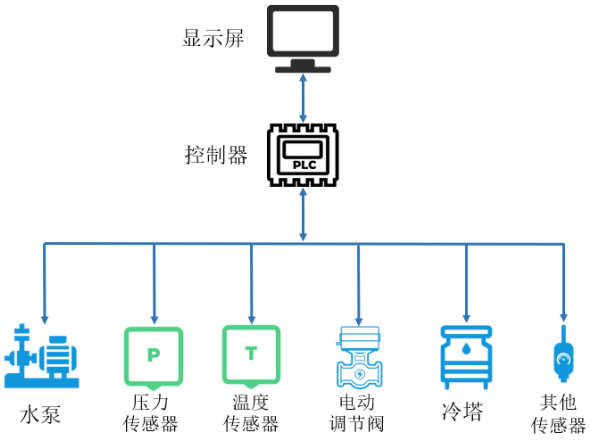


图 1.15 冷源自控系统

冷源自控系统目前常用的有 PLC 和 DDC 两种方案。PLC 性能稳定可靠，功能开发灵活（如热备、定时轮循）、选型种类较多、扩充方便，工业级可以适应复杂环境。DDC 更适合大楼暖通系统的控制，开发调试较为容易。

2. 二次侧系统控制架构

1）2N 系统的控制架构

2N 系统架构为同一个环网内的 CDU 数量为 2 个，共 N 个环网的系统架构，示意图如下：

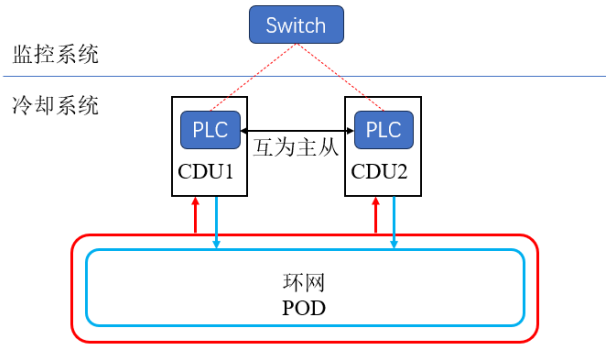


图 1.16 2N 控制架构

2N 系统的每台 CDU 内通常配置独立的控制器（如 PLC 或嵌入式单片机），CDU 互为主从关系，机组运行模式包括群控模式和单机模式，当 CDU 处于群控模式时，接收群控主机的命令参数，CDU 可脱离群控处于单机模式，由自身 PLC 执行控制逻辑。各 CDU 的控制器与上位机监控系统通讯，上传运行数据。

2）N+X 系统的控制架构

N+X 系统架构适用于同一环网内的 CDU 数量为 3 台及以上的系统，N 台 CDU 运行，X 台 CDU 待机冷备，示意图如下：

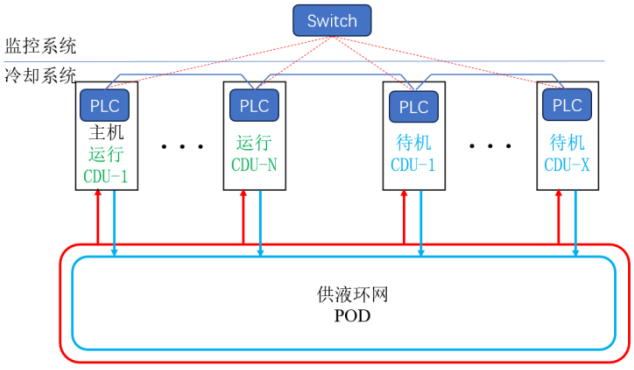


图 1.17 N+X 控制架构

N+X 系统中必有一台 CDU 作为群控主机，当系统处于群控模式时，执行群控主机的命令，CDU 可脱离群控处于单机模式，由自身 PLC 执行控制逻辑。

N+X 系统每台 CDU 内具有独立的 PLC 控制器，各控制器处于平级关系，无单独一对多的主控制器，各 CDU 的控制器均可作为系统群控主机，CDU 控制器之间通过控制网络进行通讯，实现关键数据的同步，进而实现故障投切和主备机人工设定。自控网内任何两个 PLC 都能互相通讯。各 CDU 的控制器与监控系统通讯上传运行数据。

N+X 系统主机设定与切换：N+X 系统主机可通过人工指定，同一时刻系统内有且仅有一

台主机。特殊情况下，如主机 PLC 与其余 PLC 均断开通讯连接或主机 PLC 断电、故障，其余在线 CDU 可自动推选出主机进行群控控制。

3）群控系统控制逻辑

系统内的群控主机统一调配系统所有的CDU机组的启停、故障投切、机组轮巡以及CDU机组内部的二次侧水泵，二次侧二通阀的调整等。

冷备/热备：群控冷备模式下，备用 CDU 停止运行，并按照固定的轮巡间隔轮流运行，以平衡机组的运行时间，延长使用寿命；群控热备模式下，系统内的 CDU 共同运行，均分负荷换热量。

数据同步：同一系统的内的 CDU 的控制模式、参数保持一致，在任一 CDU 修改参数，数据会同步至整个系统。

故障投切：对 CDU 的故障等级进行分级管理，当系统内出现影响机组运行的故障时，需要启动备用机组，停止故障机组，并保证机组在切换过程中保持稳定的流量。

可靠性设计：系统内部设计重要传感器备份，防止因传感器故障造成系统故障。另外，为保证CDU不因控制器故障而失去换热量，当控制器故障后，CDU水泵保持故障前状态继续运行。

1.3.1.3 换热设备

1.集中式 CDU 要点及趋势

集中式 CDU 根据其在智算中心机房内的物理布局分为入列式和非入列式两种类型。入列式CDU指与服务器机柜并列布置的CDU设备，其高度和深度一般与服务器机柜保持一致，为本列服务器机柜提供散热。非入列式 CDU 指布置在服务器机柜列之外、机房内特定区域的 CDU 设备，其尺寸一般不做严格要求，一般设置多台 CDU 共同为整个服务器机房内的服务器机柜提供散热。

图 1.18 集中式 CDU

1）设计要点

a）换热器：当冷板材质为铜时宜选用铜钎焊换热器，当冷板材质为铝时宜选用全不锈钢钎焊式换热器或者可拆式换热器；

b）水泵：水泵机封应选用碳化硅对碳化硅或者碳化硅对碳化钨材质，不宜选用石墨机封，接液材质为 304 不锈钢及以上，为了保证系统可靠性，泵组需要配置 2N 冗余备份，且需具备在线更换的能力；

c）管路流速：CDU 一、二次侧内部管路流速应≤3m/s；

d）过滤要求：一次侧过滤精度宜≥50 目，二次侧过滤精度宜≥270 目，一、二次侧过滤器应支持在线清洗；

e）自动排气功能：CDU 最高点及局部高点应设置自动排气阀；

f）泄压保护功能：CDU 内须配置自动泄压阀，当系统压力大于 5bar 时，应能自动泄压；

g）防凝露功能：CDU 须具备防凝露控制功能，避免二次侧系统发生凝露风险；

h）湿材质：所有湿材质均应满足兼容性要求，管道推荐采用 304 及以上不锈钢，焊料应采用对应的不锈钢焊料，密封垫及软管宜采用 EPDM、PTFE、PVDF 材质。禁止使用青铜、碳钢、锌含量＞15%的黄铜、高铅黄铜、铝及铝合金、生料带、螺纹胶等；

2）发展趋势

a）高热流密度：单台 CDU 功率不断增大，目前已经出现单台 CDU 制冷量 2MW 机型，未来会有 3MW 甚至更高制冷量的 CDU，满足智算中心、超算集群高密度需求；多模块并联，支持线性扩容，适配机房分期建设；

b）全链路智能化：机房级 AI 调度可以根据室外温湿度、IT 负载动态优化供液温度和流量，节能 20%～30%；CDU+管路+机柜全链路仿真，提前预警堵塞、泄漏和水利失衡；泵寿命、换热器结垢、管路腐蚀趋势预测；

c）高温化+余热回收：供水温度提升至 40～55℃，实现全年自然冷却，PUE 降至 1.05～1.1；将冷却液余热用于园区供暖、生活热水、除湿，能源综合利用率≥80%；

d）低能耗、长寿命：高效磁悬浮泵+变频控制，输送能耗降低 40%，耐腐蚀新材料（钛合金、特种塑料），寿命提升至 15 年以上。

2.二次侧管网要点及趋势

二次侧管网通常采用工厂预制化，将管道分组，在工厂内完成大部分管道模组的焊接加工、清洗和密闭性测试等工作，并氮气保压运输至工程现场进行组装。现场只进行少量管道安装，常见为卡扣、法兰等“无动火”的非焊接连接方式。

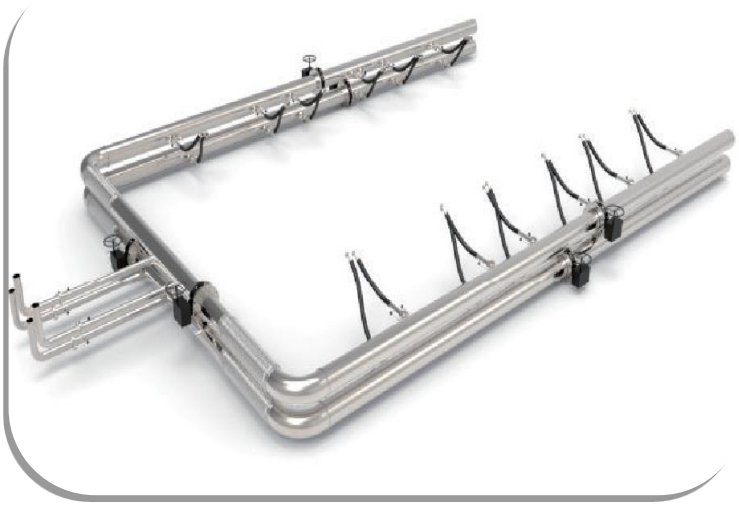


图 1.19 二次侧工程管网

1）设计要点

a）材料选择：液冷 CDU 通常运行在 15℃ 至 45℃之间，工质多为纯水或抑制型乙二醇，对于管路的材料选择须具备耐腐蚀性、机械强度、加工性能及与冷却工质的化学相容性。实际应用中，304L 和 316L 是最为广泛采用的材料，两者的主要区别在于镍（Ni）含量的差异以及钼（Mo）元素的添加，综合性能以及成本，优选 304L 作为二次侧管路材料。

b）柔性连接：管道的设计安装时，也要将热胀冷缩的影响考虑在内，允许其热胀冷缩，不应使与其所连接的管道或设备受力，且无论高温或低温工况，都满足压力密封测试。密封性测试的压力大小要高于设计压力值（例如 1.5 倍设计压力）；

c）可维护：管网应设计为可拆卸、可冲洗结构，关键节点预留检修口；

d）清洁度要求：系统投运前需进行全系统循环冲洗，并取样检测冷却液介电常数、体积电阻率、含水量等指标，且液体无杂质、表面无油污；

e）集水盘：二次侧管网底部需设置集水盘，集水盘四周通常需采用满焊工艺。配套不同分段管路，集水盘也需要进行分段预制，且需设计一定坡度，用来实现定向排水；

2）发展趋势

a）随着智算中心液冷系统对高可靠性、高稳定性的要求，液冷管网正从“被动散热通道”向“智能热管理基础设施”演进，例如：

b）分布式传感网络：在每条支路部署压力、温度、流量、漏液检测传感器，实现实时采样监控，及时发现系统运行存在的故障及风险；

c）数字孪生集成：将液冷管网中的传感数据接入智算中心数字孪生平台，实现故障预判与热流仿真优化。

1.3.2 机房级冷却系统工程验证

1.密封性测试

对于机房级的液冷系统投入运行前，进行整体二次侧密封性测试，保证整体二次侧系统不存在渗液、漏液的情况，通常采用气检或液检的方式。

2.供液特性测试

用于考核二次侧系统水流阻与流量的关系是否满足技术指标要求，根据液冷系统的流量和水压差进行模拟试验，同时也对整体二次侧系统的流量上限及下限进行摸底测试，为系统后期运行提供数据支撑。

3.CDU 故障投切及切换时长测试

模拟 CDU 在发生故障报警时，机组切换稳定的时间，并记录切换过程中流量的波动值，为系统后期运行提供数据支撑。

4.换热能力测试

模拟测试 CDU 的换热能力以及冷却塔的散热能力能否达到设计值。

1.3.3 未来展望

随着 AI 大模型以惊人的速度迭代，算力需求呈指数级攀升，服务器的功耗和散热需求也随之水涨船高。在算力密度与能源政策双重挤压下，液冷由之前的“可选项”逐渐变为“必选项”，液冷逐渐成为刚需标配。未来，吉瓦级智算中心冷却系统的发展核心围绕高密度散热、冷板式+浸没式双轨并行、极致 PUE、全栈智能化等几大方向演进。

1.高密度散热：目前单芯片功耗从 200W 跃升至未来的 3000W（英伟达 GB300/Rubin），

单机柜功耗也随之跃升至 150kW 以上，甚至到兆瓦级别。对于液冷系统的散热功率密度需求越来越高，随之对直接参与到服务器的换热设备例如 CDU、冷板部件的热流密度要求也越来越高。机架式 CDU 热流密度在 4U-150kW 已成为主流，集中式 CDU 换热功率也突破 2MW 的级别，液冷冷板也通过歧管微通道等结构实现热流密度超过 150W/cm²。吉瓦级智算中心，对于各种新技术、新材料和新架构的需求也呼之欲出。

a) 新技术：利用工质相变（沸腾 / 冷凝）汽化潜热的物理特性为基础的两相式液冷系统，进一步提升散热效率，是目前单相液冷散热效率的 1.3~1.5 倍；
b) 新材料：冷板采用金刚石等超高导热率材料与歧管微通道等技术相结合，实现热流密度突破 1000 W/cm²，但是由于加工工艺等问题，目前仍停留在预研阶段；

2.冷板式+浸没式双轨并行：由于冷板式液冷系统具有技术成熟、稳定可靠、成本可控等优势，目前冷板式液冷系统占据主流地位。但是随着热流密度的增高，浸没式液冷系统正在逐步发展。在未来一段时间内会存在冷板式液冷+浸没式液冷两种主流架构并存的状态。未来随着浸没式液冷技术持续迭代优化、标准体系逐步完善、设备互操作性全面兼容、硬件架构深度解耦、商业化规模持续拓展，浸没式液冷将成为高密度智算中心不可或缺的主流冷却方案，支撑 吉瓦级智算中心规模化部署，降低单位算力 Token 成本，助力算力普惠目标落地。

3.全栈智能化：为提高液冷系统的稳定性及可靠性，提升系统的 PUE，未来液冷系统的智能化、精细化管理势在必行。从 CDU 的全参数感知，建立 CDU 系统的健康管理系统，预测 CDU 故障及维护部件，到引入 AI 群控系统，自适应调节液冷系统温度，进行负荷预测以及故障自愈等功能，可使液冷系统具备更加智能的“大脑”，实现真正的“预见性散热”。

4.极致 PUE：液冷系统与余热回收相结合，将冷却液余热用于园区供暖、生活热水、除湿等场景，实现能源的综合利用，综合降低整套智算中心园区的 PUE。

第二章 IT 设施层

2.1 AI 超节点

本章聚焦 AIDC 中承担主要算力负荷的 AI 超节点；通用算力（管理、登录、数据预处理等）作为必要支撑，其架构可复用成熟 IDC 实践，当前版本的这份报告不单独展开。

AI 超节点本质上是一体化超级服务器，依托 UALink、Ethernet、CLink 等高速总线纵向扩容 GPU 等 AI 加速模组；借助架构解耦、液冷散热等前沿技术，实现 AI 加速模组极限高密集成。通过缩短模组间高速信号传输路径，有效提升互连带宽、降低通信时延，可充分适配 MoE、PD 分离、智能体、长序列输入输出、多轮交互等应用场景。

AI 大模型的发展推动了 AI 加速模组性能与功耗持续大幅攀升，而各地存量数据中心供电、制冷基础设施条件参差不齐，为适配现有机房基础设施承载能力，衍生出分布式超节点形态，与整机柜超节点形成互补布局，共同保障上层 AI 业务高效稳定运行。

整机柜超节点以机柜为单元，适用于高功率密度、液冷散热环境数据中心。分布式超节点以机箱为单元，可直接上架已部署的机柜，适用于供电和散热条件受限、或逐步扩展建设的数据中心环境。

2.1.1 整机柜超节点

整机柜超节点新型架构实现了高密度集成，高速互连，液冷散热，集中供电，将 GPU 芯片通过高速总线和交换芯片互连，实现 32 卡、64 卡，128 卡等不同规模 GPU 芯片在单柜内或多柜间的互连，突破了机柜内计算、内存、存储、供电、散热等技术挑战，解决了资源池化和一体化设计难题。

2.1.1.1 计算节点

计算节点是 CPU、GPU、内存及柜内高速 Scale-up 互连接口的核心承载单元，是整机柜超节点的核心算力载体。对性能与算力密度的极致追求，带来了大功率供电与散热的双重挑战。在当前业界超节点计算节点主流方案中，普遍采用集中供电与全液冷冷板散热方案。

以国内互联网头部厂商百度的超节点为例，其计算节点集成三大功能模块：CPU 计算板、PCIe Switch 板及 GPU 板，资源配比上采用 1：2：4，并配置 4 块智能网卡进行扩展。计算节点为 1U 高度设计，宽 21 英寸、深 1.2 米；网卡、NVMe-SSD、M.2 等部件采用前置布局，Scale-up 高密连接器则后置布置。

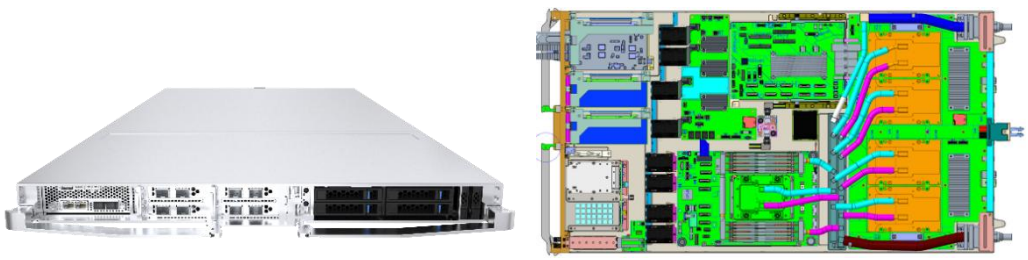


图 2.1 百度整机柜超节点计算节点设计实例

不同的 AI 应用场景对于 CPU 与 GPU 的资源配比需求有着显著差异，也带来截然不同的超节点架构设计。

AI 训练场景中，CPU 主要承担数据加载、任务分发、GPU 算力协同调度等功能，更侧重低时延、高带宽互连能力，而非单纯堆叠核心数来追求更强大的计算能力。NVIDIA GB200 NVL72 采用 36 颗 72 核心的 Grace CPU + 72 卡 Blackwell GPU，CPU 与 GPU 数量配比为 1:2，Vera Rubin 超节点整体架构延续了 1:2 的 CPU 和 GPU 数量配比（CPU 核心数提升至 88），代表了现阶段训练场景超节点的主流架构。

AI 推理场景中，尤其 AI Agent 等复杂应用下，随着任务复杂度提升、多模型调度增多，CPU 在任务编排、内存管理、调度，以及向 GPU 加速器传输数据等环节的负载显著增加。固定 CPU 与 GPU 资源配比的紧耦合架构，易出现资源错配与算力浪费，而 CPU 与 GPU 解耦池化架构能够适配更广泛的应用，显著提升资源利用率，已成为推理集群的重要方向，阿里云磐久 AL128 超节点服务器就是此类架构的典型代表。在 2026 年 GTC 上，NVIDIA 推出 Groq 3 LPX 的 LPU 整机柜方案，进一步丰富了推理场景超节点计算资源类型，组合配比也更加灵活。

2.1.1.2 机柜

机柜是整机柜超节点系统的物理承载与结构核心，用于部署计算节点、Power shelf、分布式 CDU、网络节点、Scale-up 高速网络、Scale-out 互联网络等，并为系统供电铜排、液冷管路提供可靠固定与结构支撑，实现柜内各功能模块间的精密互连与协同，是超节点实现算力聚合的核心物理框架。整机柜超节点机柜在结构强度、加工精度、与内部设备的耦合适配度方面的显著提升是其与通用 ICT 设备机柜的核心差异。

整机柜超节点机柜的规格需匹配超节点系统整体架构设计，目前业界存在多种差异化方案。其中，经定向优化、基于 OCP Open Rack V3.0（ORV3）的机柜在英伟达的推动下产业生态较为成熟；具备更高扩展性的宽体机柜也被推出，OCP Open Rack Wide（ORW）是应用较为广泛的产业实例，由 Meta、AMD 等厂商在 2025 OCP Global Summit 进行了展示。但 Meta 也在其技术分享中提到，更大的机柜意味着更高的设计和制造难度、更高的运输花费和更高的运维挑战，因而未来的方向可能会是多个解耦的低密度 ORV3 机柜通过光互连来实现更高的扩展性。

国内互联网头部厂商百度的超节点机柜传承了天蝎整机柜服务器架构设计理念，机柜高度为 46U，外形尺寸为 2300mm（±2） × 600mm（0,－3） × 1200mm（±2），单元 RU 高为 46.5mm；机柜额定承重不低于 1600kg，支持水、电、网三路盲插拔接口，实现节点盲插拔运维。液冷汇流排（Manifold）采用上下进回液设计，支持整机柜一体化交付，显著提升交付与运维效率。

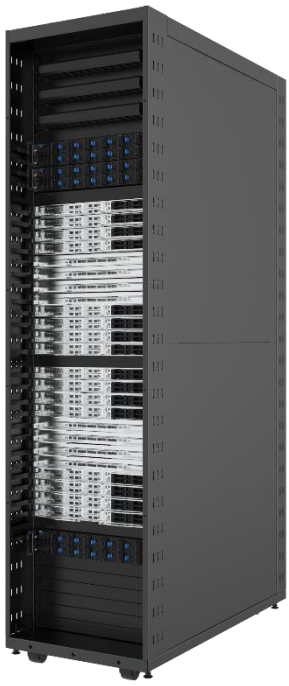


图 2.2 百度整机柜超节点机柜设计实例

2.1.1.3 整机柜供电

在供电接入环节，北美数据中心传统以 480Y/277VAC、208VAC 为主要输入制式，国内数据中心则普遍采用 380Y/220VAC、240VDC 和 336VDC。随着 AI 超节点功率密度快速提升，±400VDC 和 800VDC 等更高电压的直流输入方案将日渐成熟与普及。

1.Power shelf 供电

Power Shelf 是整机柜超节点系统的供电转换枢纽，采用池化架构对柜内电源模块（PSU）进行集中部署，并配置电源监控管理单元（Power Management Controller, PMC），实时监测 PSU 等供电组件的运行状态与健康度。Power Shelf 通过 PDU 接入输入电源，经 PSU 降压转换为 54VDC 或 48VDC 后输出至机柜母线，为柜内设备统一供电。定向优化的 Power Shelf 还可以提供若干标准供电接口，为其他额定电压输入设备供电。

随着单机柜功率快速提升，提升功率密度与供电转换效率已成为 PSU 的核心设计目标。第三代半导体（如氮化镓器件）、高频低损耗磁性元件、先进功率变换拓扑（如飞跨电容四电平 FC4L 拓扑）等新材料与新技术的应用，能够有效缩小器件体积、降低磁滞与涡流损耗、降低器件电压应力、显著提高器件品质因数，最终实现更高的功率密度与转换效率。

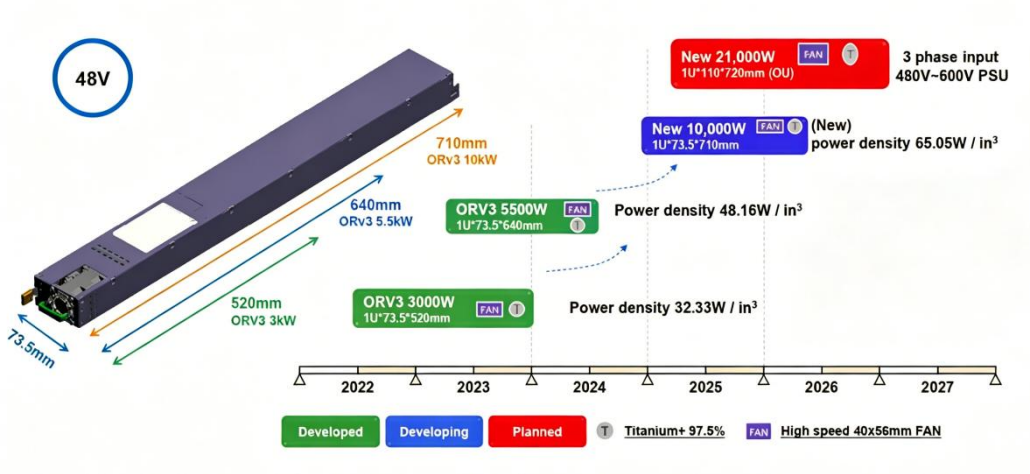


图 2.3 PSU 功率及尺寸发展路线图

柜内集成 Power Shelf 会挤占计算节点与交换节点的部署空间，当单机柜功率突破 200kW 时，若继续沿用 Power Shelf 内置架构，会显著降低算力部署密度，甚至会影响整个计算集群的运行效率。而随着单机柜功率进一步提升，供电单元与计算、网络设备共柜部署的模式将难以为继，电源将不得不从业务机柜中剥离，Sidecar 独立供电柜方案将逐步成为智算中心的主流方案。

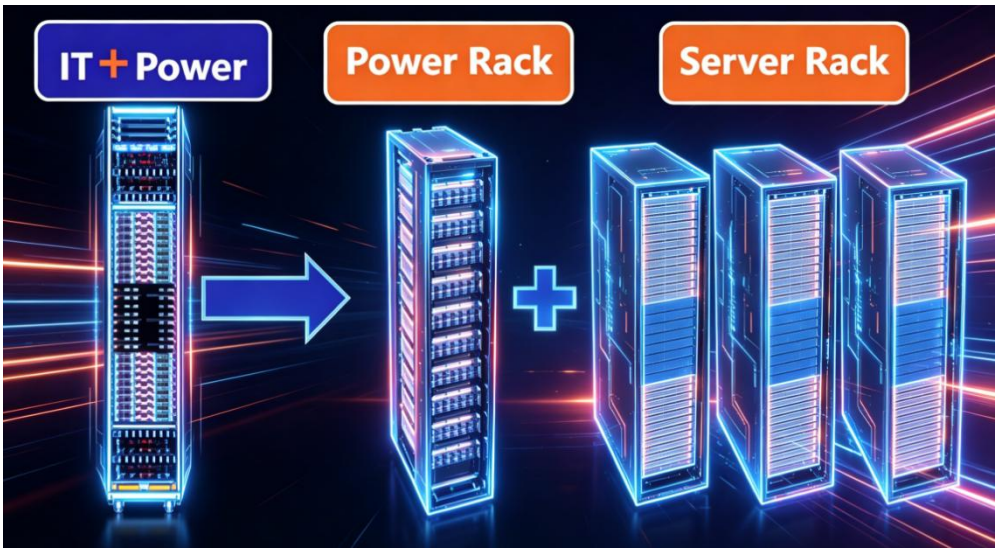


图 2.4 整机柜超节点供电架构由共柜转向 Sidecar 独立供电柜方案

2. 计算节点供电

当前产业实践中，计算节点、网络节点通常通过 Power clip 从机柜 Busbar 直接获取 54VDC 或 48VDC 输入为 GPU 模组供电，为节点内 CPU 及其它组件供电则需进一步通过降压转换为 12VDC。随着±400VDC、800VDC 直流输入方案的应用，Busbar 电压也将升级为±400VDC、800VDC，需要在计算节点、网络节点内部增加 DC 转换模块进行降压转换，输出 54VDC、48VDC、12VDC 等电压，为 GPU 模组和节点内的其它组件供电。随着 GPU、CPU 等芯片功耗持续攀升（目前单颗 GPU 功耗已达 2kW），供电路径设计挑战日益提升，垂直供电等更高效的新型供电架构也随之出现。

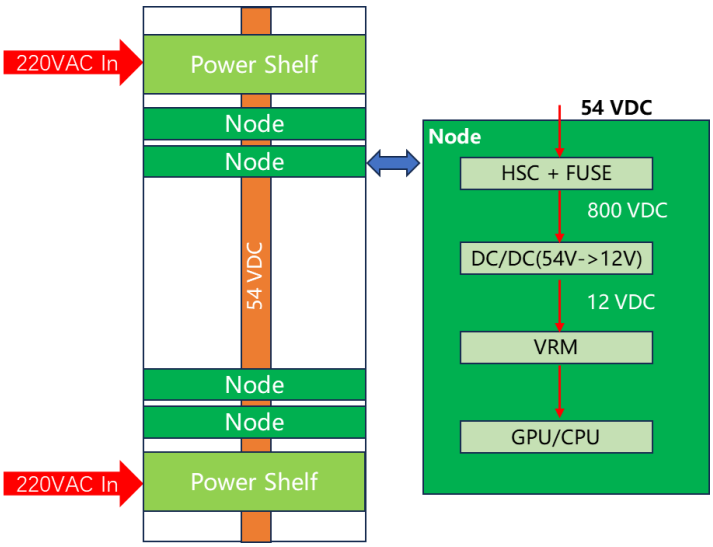


图 2.5 当前 220VAC 整机柜供电拓扑

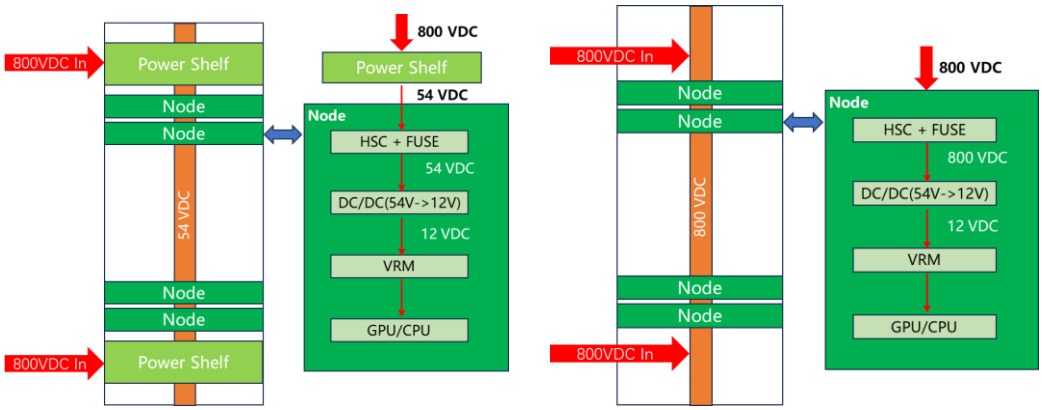


图 2.6 未来 800VDC 供电拓扑

2.1.1.4 整机柜散热

在整机柜系统层面，机柜级散热从风冷散热逐渐迈向风液混合散热，最终走向全液冷散热。计算节点、网络节点以及存储节点随着算力、运力、存力提升和功耗的增加，保持卓越的性能和系统的稳定可靠是超节点必须攻克的技术挑战，而液冷提供了高功率密度场景下更佳的散热方案。

为实现极致算力密度部署，大型云服务商（CSP）及新一代云数据中心用户普遍采用液冷散热架构的整机柜超节点方案。而对于算力部署密度要求不高或数据中心暂不具备液冷条件的中小规模用户，可选用风冷或风液混合散热方案。

当前液冷整机柜超节点普遍采用冷板方案，柜内液冷组件主要包含分水器、冷板、快接头及柜内 CDU（In-rack CDU）。

柜内分水器是用于连接 CDU 与节点冷板的管路系统，根据分水器与冷板间快接头的连接形式，又可分为手插分水器和盲插分水器两种，盲插分水器通常配备防喷溅组件及导流槽。为实现可靠装配，分水器连接器安装面的平面度不可大于 0.1mm。为保障柜内节点散热均衡，分水器支路流量偏差不可大于 10%。

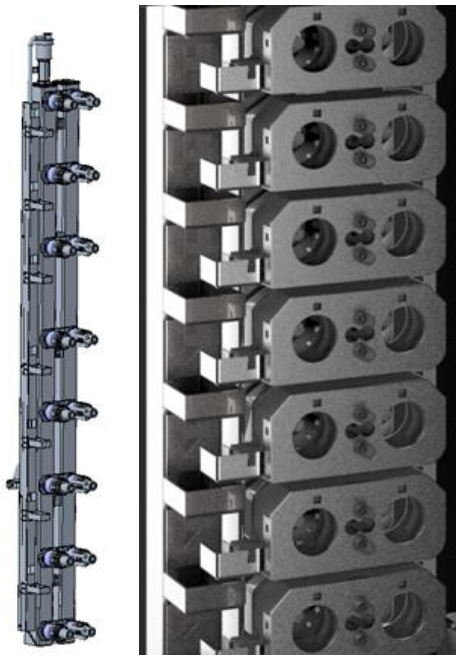


图 2.7 分水器及防喷溅组件示意

冷板位于计算节点、交换节点内，冷板的高效设计、制程工艺，对于 GPU、CPU、ASIC 等高功率组件的长期可靠运行和系统节能至关重要。在设计层面，优化铲齿结构，以增大散热面积并增强流体湍流效应；采用射流腔室设计，实现对热点区域的直接冲击冷却，提升局部散热强度；引入两相冷板技术与高效蒸发换热结构，利用液体相变吸收大量潜热，从而进一步强化散热能力。在制程工艺和检验工序层面，焊接后进行补充加工，保证平面度；对焊缝进行 100%的超声无损检测，剔除缺陷；采用镀镍或钝化处理，来达到长期防护效果。

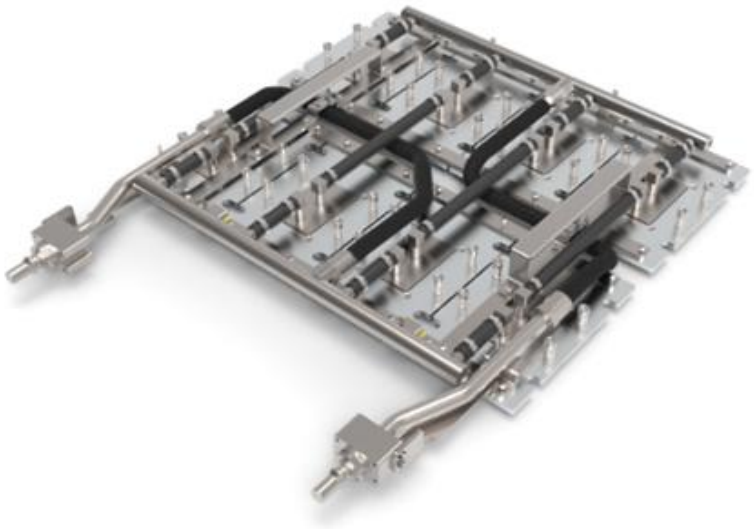


图 2.8 GPU 冷板组件

按照安装部位，快接头可分为柜内安装（装配于分水器侧）和服务器内安装（装配于冷板侧）。按照对接的操作方式，快接头可分为手插快接头和盲插快接头。盲插快接头在超节点系统中应用较为广泛。盲插快接头要实现可靠连接，需要满足严苛的径向、轴向和角

度浮动量要求：径向浮动量应不低于±1.0mm，轴向浮动量应不低于 1.0mm，配合浮动模组，角度浮动量应不低于±2.5° 。

图 2.9 液冷快接头

依据 CDU 的系统整体布局，整机柜超节点液冷方案可分为分布式散热与集中式散热两种架构：分布式散热方案将 CDU 与 IT 设备集成在同一机柜内，更适用于中小部署规模、不追求极致算力密度的场景；而大型云服务商（CSP）与新型云数据中心（Neo cloud）用户，则普遍采用独立 CDU 柜的集中式散热方案，有助于实现更高算力密度，简化液冷管路组网，并大幅提升运维效率。

图 2.10 分布式散热方案整机柜超节点

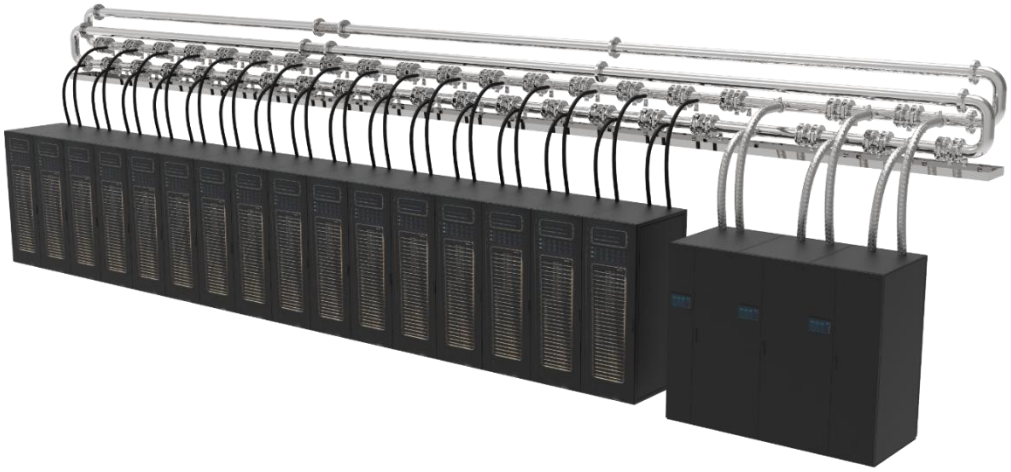


图 2.11 集中式散热方案整机柜超节点集群

2.1.1.5 Scale up 互连

依托 Scale-up 专用互连协议，超节点可构建全局统一编址空间，将系统内所有 AI 加速芯片的内存与显存抽象为统一存储资源池；计算单元能够以内存语义直接访问其他芯片的存储资源，省去复杂的数据拷贝操作，端到端时延可降至亚微秒级。

超节点 Scale-up 互连需重点满足高带宽、低时延、高可靠和高密度等核心需求，可划分为铜互连与光互连两类技术路线。

1. 铜互连技术路线以铜质导体为传输介质，具备可靠性高、时延低、产业链成熟及成本可控等优势，在当前占据主流地位。线缆背板互连和板间正交互连是两类核心互连架构。

1）线缆背板（Cable Tray）互连

通过连接器与铜缆工艺的持续优化，线缆背板大幅提升了链路损耗裕量，互连长度可扩展至 2m。其搭载的盲插自导向浮动机构，彻底消除了传统硬连接因不完全互配导致的高速信号劣化问题，同时支持线对板、线对线等灵活的互连形态转换，具备极高的设计灵活性与可靠性。该方案尤其适配计算节点与网络节点水平布局、垂直堆叠的架构，是当前整机柜超节点柜内应用最广泛的互连方案。

为保障线缆背板的长期可靠连接，需量化对浮动机构与对接力的要求。以百度超节点为例，其线缆背板量化指标如下：

a）面板侧 X/Y 轴浮动：在指定承压力区间内，实现±3.0mm 的自校正导引，机构全程无卡滞、锁死或回弹失效等问题；

b）面板侧 Z 轴行程浮动（弹簧压缩方向）：在指定承压力区间内，实现节点过压状态下最大 3.5mm 的弹簧压缩，机构全程无卡滞、锁死或回弹失效等问题；

c）弹簧弹力性能：弹簧是实现 X/Y/Z 三轴向浮动的核心部件，需在不同设计的承压力区间内稳定满足弹性性能要求。

2）板间正交互连（Orthogonal Interconnect）：

板间正交互连突破了柜内节点水平布局、垂直堆叠的结构限制，可实现计算节点与网络节点前后竖直对插（计算节点与网络节点均采用竖直布局对插），进一步提升了超节点系统算力集成密度。

a）**有中板正交互连（Orthogonal Midplane）：**适用于小规模短距离互连场景，典型应用为 GPU Tray 内部、相邻 GPU Tray 和 Switch Tray 之间的互连（互连距离通常在

0.5m~0.7m），具备架构简单、集成度高、低成本、易维护等优势，在分布式超节点系统中应用较为广泛。

b）**无中板正交互连（Orthogonal Direct）：**依托低损 PCB 板材、高速连接器材料升级与工艺迭代，加之架构优化，彻底消除中置背板信号损耗，大幅延展了高速信号的传输距离。其高速通道分为 PCB 与铜缆两种：铜缆互连距离在 112G PAM4 速率下可以达到 1m~1.3m，224G PAM4 速率下可以达到 0.4m~0.8 m；PCB 走线互连距离则略低于铜缆。无中板正交互连尤其适配超高集成密度的整机柜超节点架构，是下一代 Scale-up 高速互连的重要技术路线。

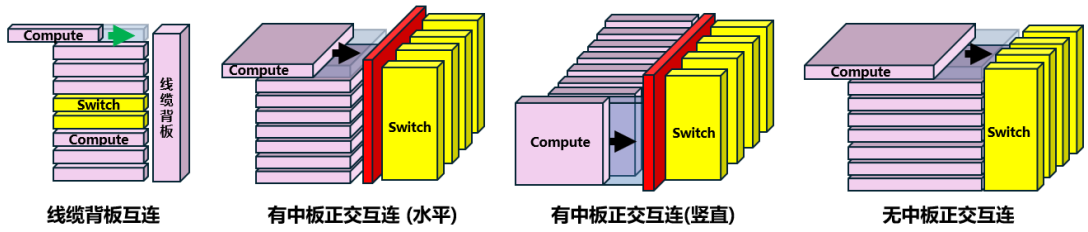


图 2.12 多种板间正交互连方案

随着超节点信号传输速率的不断攀升，要解决 112Gbps 以上速率信号在机箱内的互连和传输瓶颈，NPC 和 CPC 成为最优方案：

NPC（Near Package Copper）：面向机架级长距互连，以损耗优化与距离延展为核心，支持 112G/224G 速率，传输距离 1.5m~2m；

CPC（Co-Packaged Copper）：面向封装级超高密度互连，以速率提升与能耗降低为核心，将 224G/448G 信号缩进封装内部，传输距离 1m。

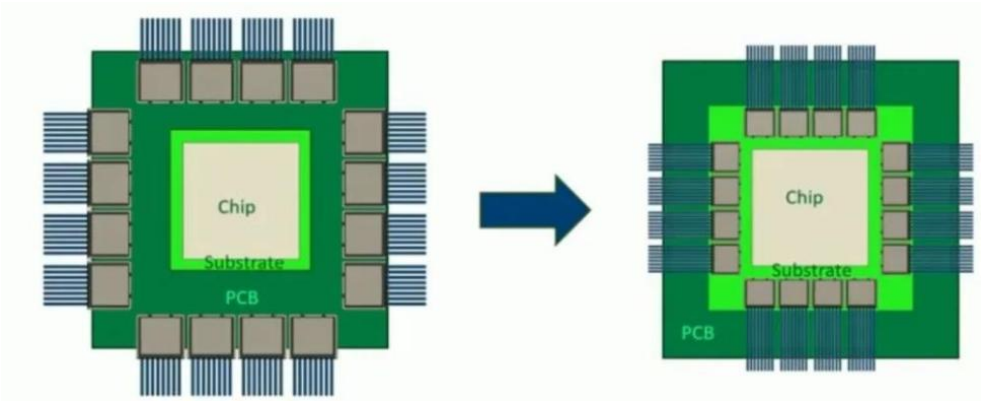


图 2.13 NPC 与 CPC 互连方案

2.光互连技术路线以光纤为传输介质，利用光信号实现远距离、超高带宽的数据传输，是破解超节点机柜内及机柜间互连瓶颈的重要技术路线。

LPO（Linear Pluggable Optics，线性可插拔光模块）沿用传统可插拔模块形态，通过移除内置DSP、将均衡和时钟恢复等复杂功能转移至ASIC侧的SerDes（串行器/解串器）芯片中，成为短距离低功耗可插拔方案的重要过渡形态。

当前超节点光互连领域的热点技术是 NPO（Near Packaged Optics，近封装光互连）与 CPO（Co-packaged Optics，共封装光互连）。两类方案均通过移除光模块中内置 DSP，并与 ASIC 更近距集成，显著降低光模块功耗（单 400G 端口功耗约从 8W 降至 3W），将单端时延降低约 100ns，并显著提升端口密度与系统可靠性。NPO 方案因其产业更成熟、开放，在现阶段获得更多系统厂商与用户的支持。

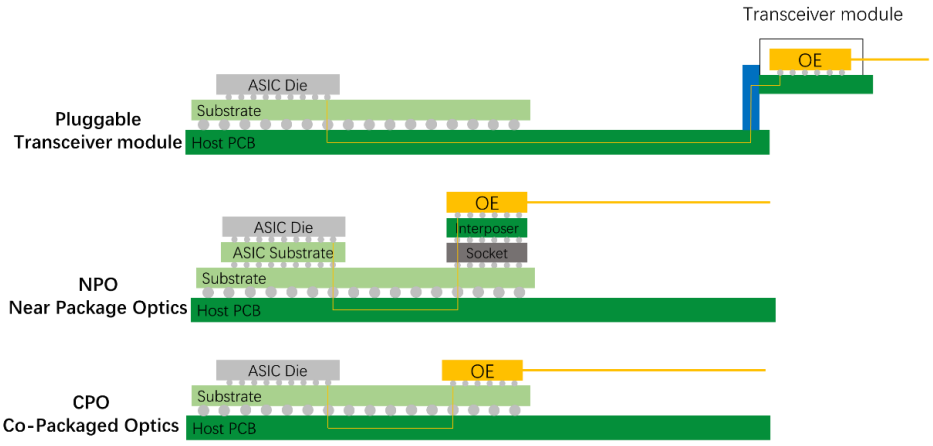


图 2.14 可插拔光模块、近封装光互连、共封装光互连链路图示

国内厂商自 2025 年下半年起，密集开展 NPO 方案技术探索与产品布局：腾讯主导推出 ETX Ultra 512 卡超节点方案，采用 NPO 模组实现超节点内部全光互连；阿里云于 2025 年 9 月云栖大会首次发布 NPO 模组与 102.4T NPO Switch 概念样机，并在新一代国产四芯片交换机中率先应用 3.2T NPO 技术。

2.1.2 分布式超节点

分布式超节点是通过紧耦合架构与高速协议协同，将小规模计算芯片在单一机箱内实现高效聚合的系统形态，其单体规模更小、算力颗粒更精细，并可根据业务场景需求实现灵活扩展，适用于中小部署规模或供电密度受限的已建成存量数据中心场景。

2.1.2.1 模组

1.AI 加速模组

分布式超节点 AI 加速模组在物理形态上可分为 PCIe 插卡型与板载集成型两类形态。

PCIe 插卡型 AI 加速模组：通常以 PCIe 作为卡间互连总线，标准化程度高，但散热与功耗上限相对受限，适用于单模组算力要求适中的计算场景。

板载集成型 AI 加速模组：采用开放标准或厂商自定义接口及协议，突破了 PCIe 插卡型 AI 加速模组的接口和尺寸限制，支持更高功率输入与系统散热，具备更强大的 AI 算力，是当前占据主流地位的 AI 加速模组形态。

图 2.15 板载集成型 AI 加速模组设计实例

以 Rubin/Rubin Ultra 为技术参照，面向训练与高密度推理 GPU 集成模组，功耗在 2027 年将进一步上探至 3600W，体积、热流密度与供电需求将显著提升，对冷板/浸没液冷、供

电稳压、机柜散热与结构设计带来系统性挑战。

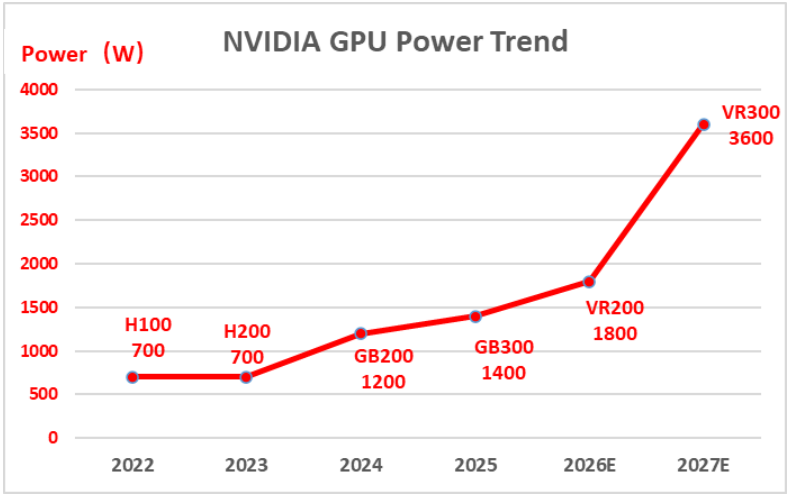


图 2.16 NVIDIA GPU 功耗趋势

2. AI 载板模组

AI 载板模组是用于承载 GPU 等 AI 加速模组的核心基板单元。单块载板可搭载多颗 AI 加速模组，相互之间通过总线协议实现互连，构成基础 AI 加速资源池单元。载板对外提供高速总线及扩展网络接口，支持多载板 Scale-up 或 Scale-out 扩展。

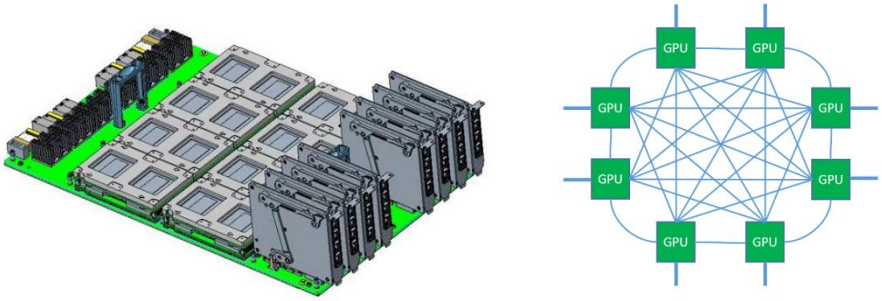


图 2.17 搭载扩展卡的 AI 加速器载板模组及其总线拓扑

随着 GPU 等 AI 加速模组的功耗、尺寸、总线速率高速增长，AI 载板模组的设计也面临着更加复杂的架构、拓扑和工程实现挑战。

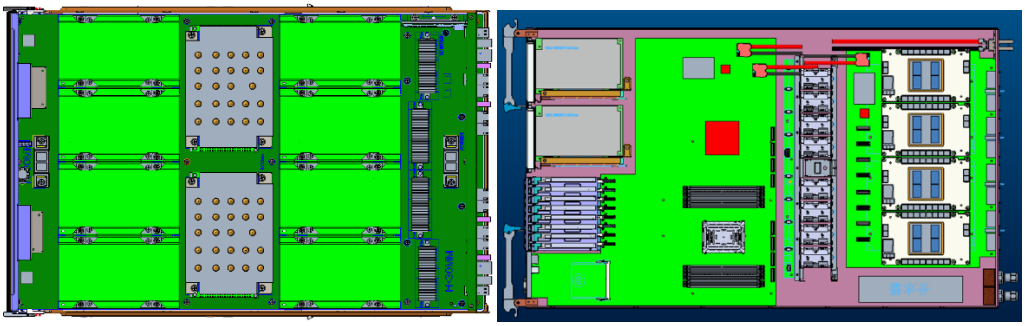


图 2.18 8*GPU 载板与 4*GPU 载板设计对比

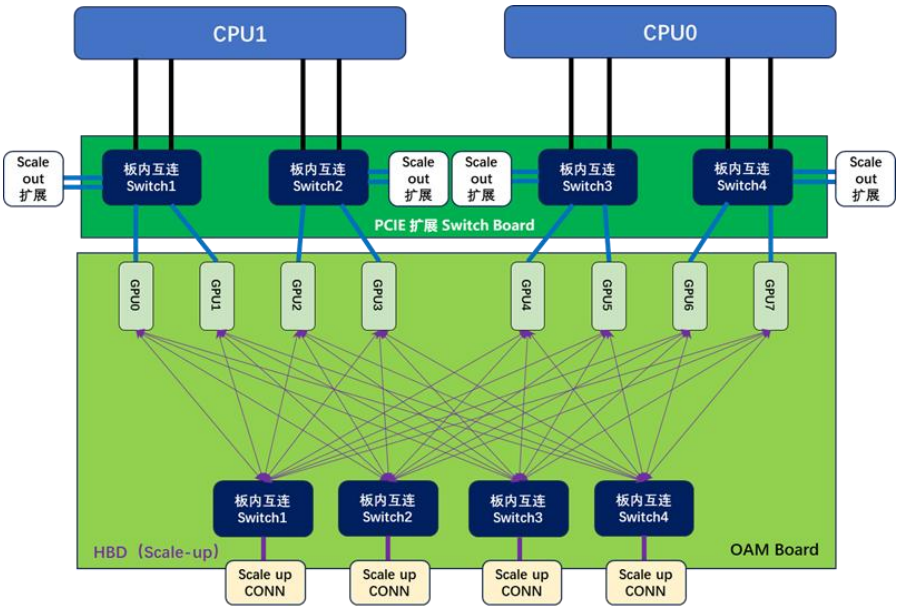


图 2.19 8*GPU 载板系统拓扑

2.1.2.2 拓扑

1.集群层拓扑

在集群层，分布式超节点通过 IB 或 Ethernet 协议进行胖树等网络架构组网，可以灵活配置网卡及数据中心网络拓扑。

典型集群拓扑有三层 CLOS/胖树（Fat-Tree）。

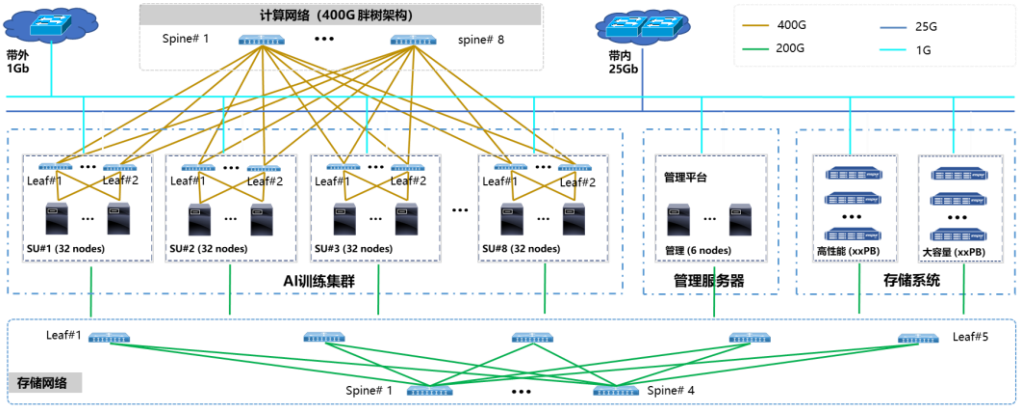


图 2.20 典型 AI 集群组网拓扑示例

2.机柜层拓扑

机柜层有两种主流拓展方案：

a) 机柜内同时集成计算、存储与加速节点，部署更灵活，适配中小规模集群场景；

b) 机柜内仅集成单一类型节点，再以机柜为单元进行扩展集群，适配大规模集群场景。

针对单机柜内同时集成多种节点的方案，宜通过总线架构实现多种 CPU 计算节点间、GPU 节点间（PD 分离等典型场景）、存储节点间的互联互通访问，宜采用包括 Ethernet、OISA、UALink、CLink 等在内的一种或多种开放标准互连协议。

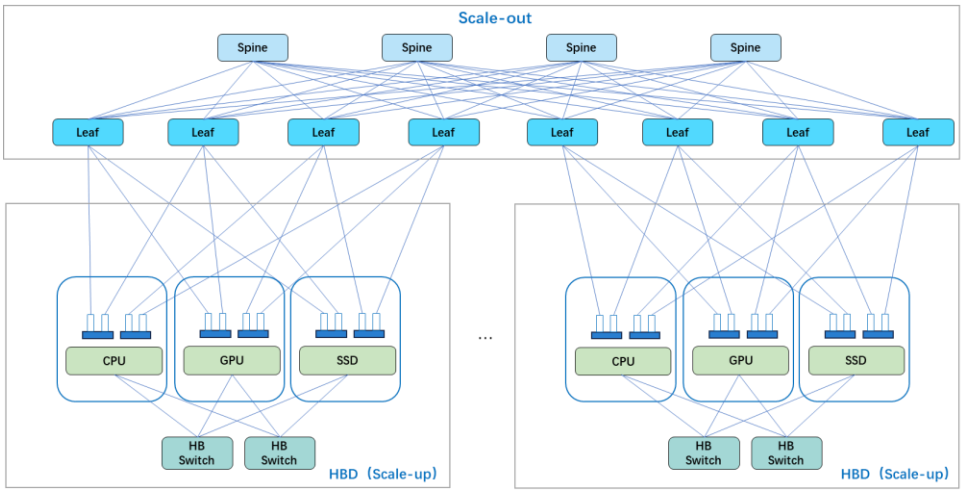


图 2.21 单机柜内同时集成计算、存储、AI 加速节点集群拓扑

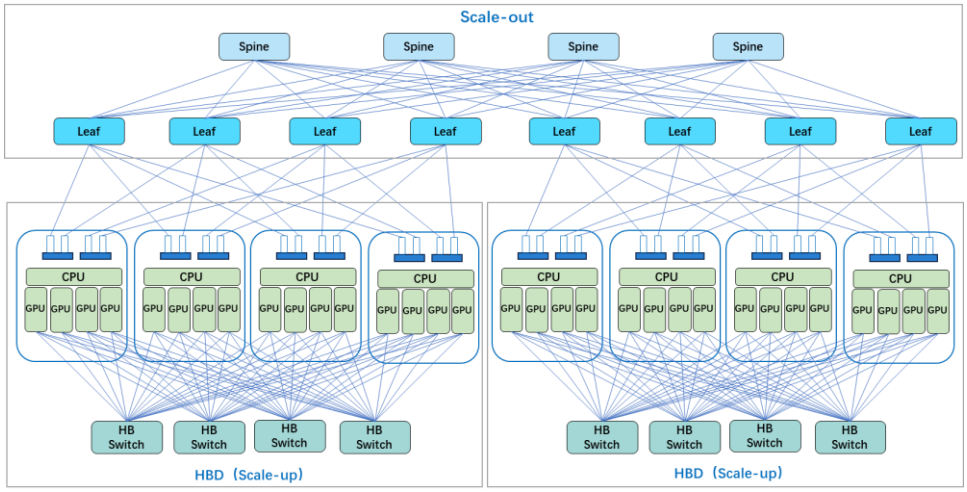


图 2.22 单机柜内仅集成 AI 加速节点集群拓扑

3. 节点层拓扑

包含多种物理拓扑，支持更灵活的 CPU，内存，GPU 配比关系。

a) 传统 CEM 标卡：CPU 通过 PCIe 总线与 GPU 直连，GPU 与 GPU 之间通过 PCIe switch 实现节点内多卡互连。此方案拓扑简单，可靠性高，灵活性高，但节点内及节点间互连带宽性能有限；

b) OAM 卡：以 OCP UBB 规范下的直连拓扑（Full-mesh 全互连、Hybrid Cubic 混合立方）为主。未来将增加载板内 switch 芯片，来实现编程友好及性能进一步提升；

如在以下 NVIDIA BFX NVL16 拓扑实例中，GPU 上行与 CPU 直连，下行通过 Switch 进行卡间互连，有效聚合了 CEM 标卡的灵活性与 NVSwitch 的高互连带宽性能优势

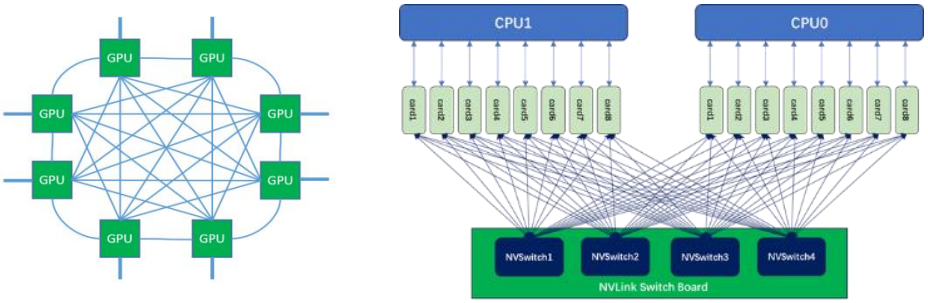


图 2.23 Full-mesh 全互连拓扑与 NVL16 互连拓扑

2.1.2.3 供电

分布式超节点集成电源模块。随着 GPU 模组功耗与部署密度提升，单系统功耗已经达到几十 kW，单颗 PSU 功率输出已经突破 5kW。匹配内部多载板模组垂直堆叠架构，系统普遍采用母线统一供电方案。

2.1.2.4 散热

得益于分布式超节点具有颗粒小、组网部署灵活的特点，可结合基础设施条件，将一定数量的分布式超节点与分布式 CDU 集成于同一机柜内，构成标准机柜单元。再以标准机柜单元为基础进行集群拓展。

当用户部署规模较大时，可采用独立 CDU 柜方案，进行集约化部署。

2.1.2.5 IO 扩展卡

为进一步提升分布式超节点与 GPU 载板模组的扩展灵活性，可开发标准的 Scale-up 扩展卡。Scale-up 扩展卡往往需集成信号恢复、信号放大等功能模块以实现分布式超节点间或载板模组间的 GPU 模组可靠、高效互连。

2.2 高速互连物理层与连接器

2.2.1 吉瓦级智算中心高速互连网络

超节点是由多卡 GPU 通过点对点互连构成的统一计算单元，其核心是通过 Scale-up 技术实现节点内高带宽、低时延通信，解决 AI 计算的通信瓶颈。藉由 Scale-up 优化单节点内多卡协同，Scale-out 支持超节点间扩展，实现万卡级集群的灵活组网。Scale-up 高速互连硬件应用包括：计算节点高速互连、交换节点高速互连、柜内节点间高速互连、机柜到机柜间高速互连。

2.2.1.1 超节点高速互连硬件架构

超节点集成的更大规模的 XPU 卡算力，以及算力需求爆炸式增长所驱动的高带宽和低损耗目标，共同推动着高速互连硬件架构的转变。超节点需综合考虑算力容量、效率、散热及可拓展性，架构正逐步向线缆背板（Cable Tray）架构、有中板正交架构、无中板正交架构、NPO 与 CPO 架构等方向演进。

线缆背板（Cable Tray）架构：节点从前面板水平移入，末端与背板互连。是目前行业应用最广泛的方案，如：英伟达 NVL 72~144.字节大禹、百度天池（线缆架构）等。

有中板正交架构：计算与交换节点分别从前后面板垂直或水平装入，通过中置背板互连。如英伟达 Rubin Ultra NVL 576 等。

无中板正交架构：计算与交换节点直接在前后面板垂直对接。如华为 Cloud Matrix 384.曙光 scaleX640、阿里云磐久 AL128 超节点服务器、百度天池（正交架构）等。

2.2.1.2 计算节点高速互连

计算节点是整机柜超节点的核心算力载体，其中单节点内多张 XPU 卡间互连（行业常见 4 到 16 卡/单节点）。随着 AI 大模型对性能与算力密度的极致追求，基于铜互连的高速互连技术面临着信号衰减、功耗增加等挑战。行业提出了高密度大通道线缆组件、近封装

铜互连（NPC）、近封装光互连（NPO）等技术方向，同时推动标准（PCIe 协议等）高速线缆的速率升级，共同致力于实现信号损耗优化及传输距离延展等核心目标。

2.2.1.3 交换节点高速互连

交换节点是整机柜超节点的核心互连枢纽，随着多组节点间 XPU 卡间互连（16-640 卡/超节点）通信需求的爆发，大量计算单元与交换芯片需在有限空间内实现高密度互连。

在吉瓦级智算中心，AI 大模型对通信时延与网络带宽的极致追求，推动了以太网交换芯片的飞速发展：速率由 56Gbps/lane 向 224Gbps/lane & 448Gbps/lane 提升。在这一高速网络演进趋势下，高速互连面临着信号衰减、功耗增加及节点后窗面板空间布局受限等多重挑战。为此，行业已提出一系列技术路径，如高密度大通道线缆组件、近封装铜互连（NPC）、近封装光互连（NPO）、共封装铜互连（CPC）及共封装光互连（CPO），并通过快速实践与迭代，推动 RNIC 与新型互连方案的协同适配演进，共同保障大规模 AI 网络在高吞吐场景下的稳定与高效运行。

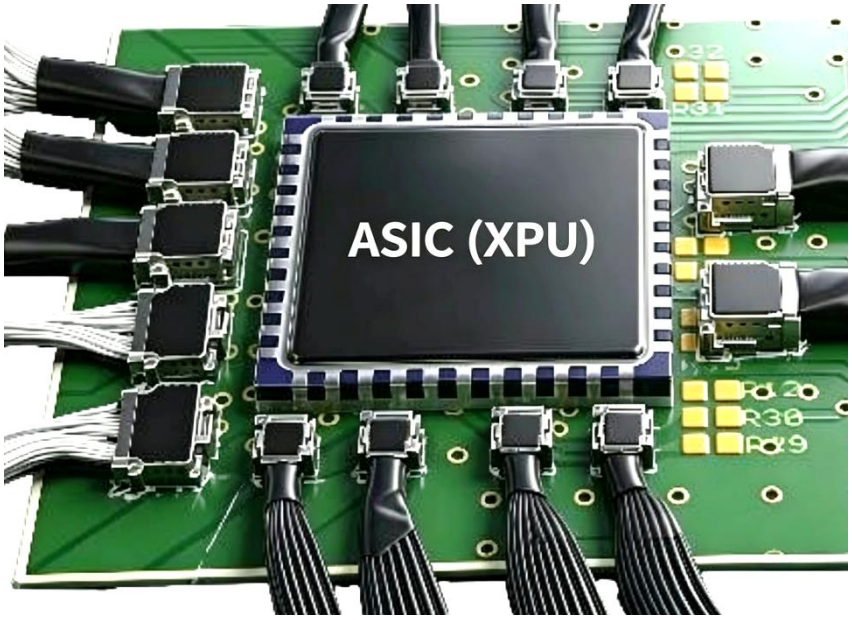


图 2.17 节点高速互连某 NPC 产品应用示意图

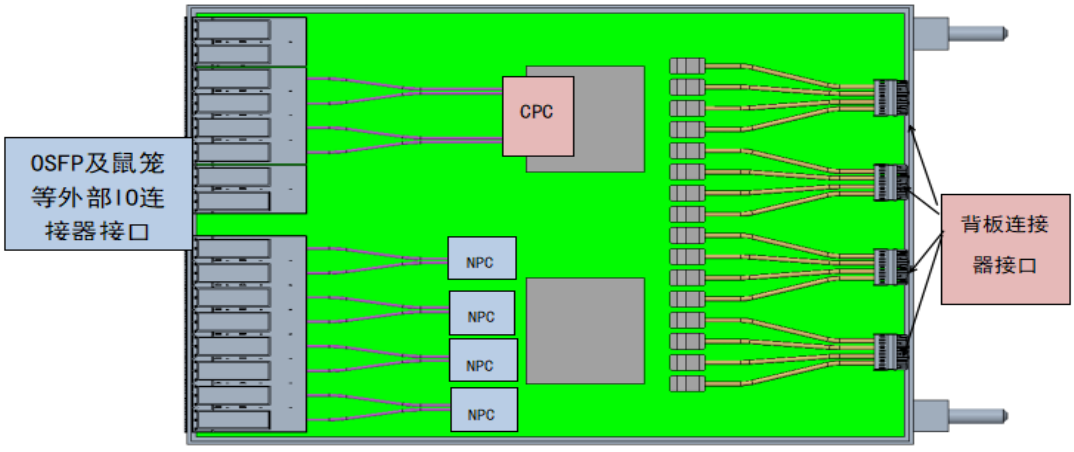


图 2.18 交换节点高速互连应用示意图

2.2.1.4 柜内节点间高速互连

在吉瓦级智算中心机柜的应用场景中，整机柜正从传统 PCB 背板架构向线缆背板（Cable Tray）架构、有中板正交架构、无中板正交架构等方向演进。

1.线缆背板（Cable Tray）架构

线缆背板（Cable Tray）架构采用高速铜缆模组替代传统 PCB 背板，构建计算节点与交换节点的直连，形成灵活的网状拓扑。该方案突破了背板物理尺寸限制，支持高密度、多方向扩展，具备低成本、浮动机构保障的高可靠性及多 XPU 兼容等优势，并可快速更换线缆实现带宽升级。同时，其开放的布局，便利适配风冷（优化风道）与液冷（配合冷板与歧管）等多种散热需求，实现了机柜内部跨节点间的高效互连部署。

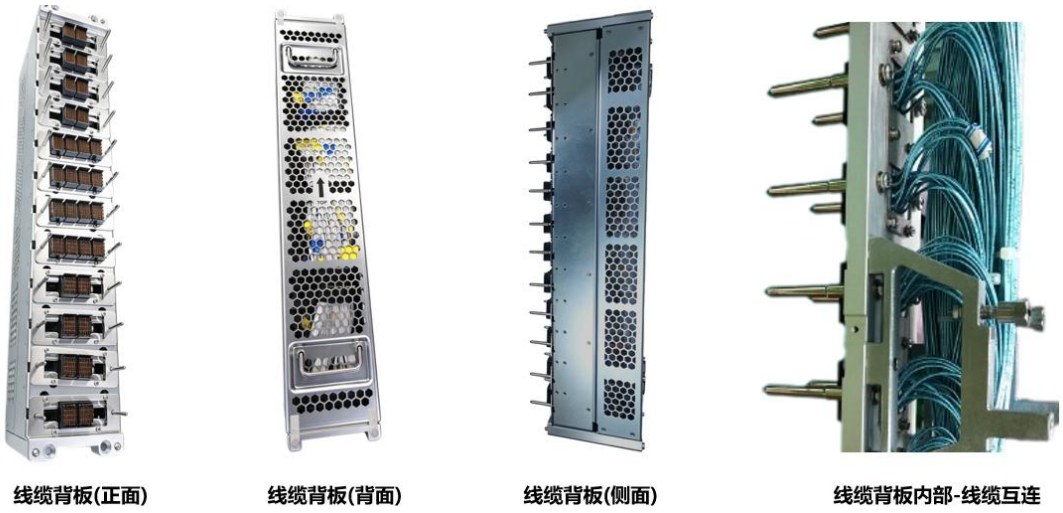


图 2.19 某线缆背板（Cable Tray）产品示意图

2.有中板正交架构

有中板正交架构采用计算节点与交换节点 90° 垂直正交的方式插接于公共中背板之上，构建无阻塞矩形拓扑。该架构凭借极短的信号链路实现超低时延与卓越的信号完整性，配合稳固的连接器设计，连接器可以兼容 PCB 板级方案与 Cable 线缆化方案，支持前后盲插维护与高效的直通风道散热，同时具备高密度、高可靠及良好的跨代兼容性，但也相应在架构与硬件设计方面引入了新的挑战。

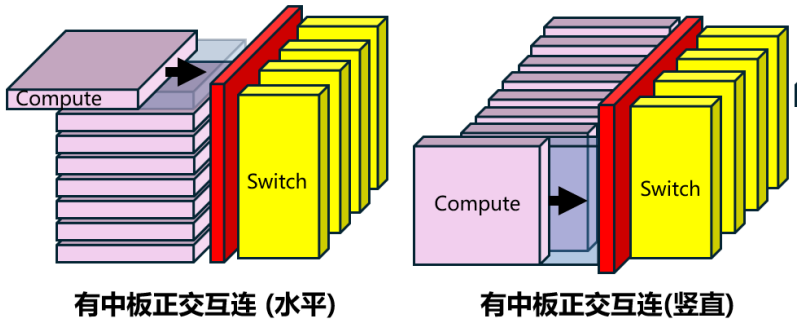


图 2.20 有中板正交架构示意图

3.无中板正交架构

无中板正交架构取消了中间公共中背板，采用计算节点与交换节点 90° 垂直正交直连的硬件结构。该设计突破了传统背板的线路速率瓶颈，在实现超大交换容量与高扩展性的

同时，保持了极低信号衰减。不过，这种高度集成的直连模式也对机构的精度、互连稳定性、散热液冷规划、供电及硬件等设计提出了更大的工程挑战。

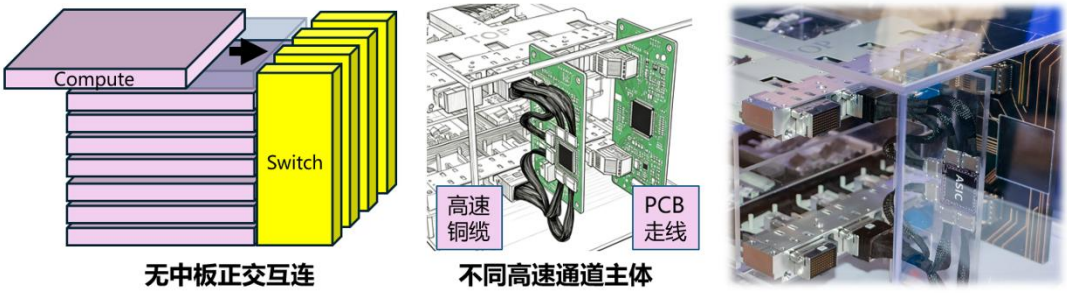


图 2.21 直接正交+PCB 架构互连方案

2.2.1.5 机柜到机柜间互连

设备间互连方案主要分为铜互连与光互连两大技术路线。其中铜互连传输方案主要为 DAC（Direct Attach Cable）、ACC（Active Copper Cable）、AEC（Active Electrical Cable），以及背板连接器直连铜缆等组件。随着数据中心带宽从 100G/400G 向 800G/1.6T 演进，铜互连正由传统 DAC 向具备更强信号补偿能力的 ACC、AEC 升级。ACC 适用于机架内等中等距离场景，AEC 则更适配跨机架等长距互连需求。而在光互连领域，AOC 与可插拔光模块等作为核心方案，其演进重心聚焦于实现更高带宽、更低功耗、更小尺寸及更优成本的系统集成。

2.2.2 互连技术实现路径

面向吉瓦级智算中心，高速互连面临巨大挑战。SerDes 技术已从 28Gbps 演进至 224Gbps，目前 56G/112Gbps 广泛应用，正强力驱动网络架构革新。因此，系统梳理全速率段互连技术路径，是保障系统平滑升级的关键。

2.2.2.1 端接形式：Press-fit/SMT/BGA/LGA

为了满足传输速率的提升，连接器的端接形式发生了演变：Press-fit（压接针脚）的成品孔越来越小，逐渐向 0.25 mm 演变；SMT、BGA、LGA 的间距向 ＜0.6 mm 持续高密化缩小，以解决大带宽场景下的阻抗一致性问题。

表 2.1 端接形式应用信息比对表

端接形式	典型规格范围	优势	劣势
Press-fit	56Gbps 推荐成品孔径：0.31~0.45 mm	机械可靠性高，取卸方便，维修及返工便捷；	孔径受到 PCB 叠层板厚、冲压等工艺限制，空间密度上有局限；
	112Gbps 推荐成品孔径：0.29~0.36 mm		
	224Gbps 推荐成品孔径：0.25~0.36 mm		
SMT	端子间距：0.5～0.6 mm	更高密度及小型化，优于 Press-fit	维修与返工困难；
BGA	端子间距：0.4～0.6 mm	更高密度及小型化	维修与返工困难；
LGA	端子间距：0.4～0.6 mm	更高密度及小型化	机械可靠性稍低于 Press-fit；

2.2.2.2 高速连接器

高速背板连接器实现单节点末端面板与对接节点的互连形式主要是高速背板连接器。随着互连应用差异化诉求，可以挑选板级应用类型或是线缆化应用类型，并延伸不同系列 56Gbps ~ 224Gbps 高速连接器。【2.2.2.2~2.2.2.4 章节-详细连接器选型推荐全景图—参考《智算中心高速互连技术技术报告（2026）》产品表】

图 2.22 某高速背板连接器及线缆组件产品示意图

2.2.2.3 内部高速 I/O

内部高速 I/O 实现单节点内的高速互连，随着互连应用协议的差异化诉求，可以挑选标准类型连接器（MCIO、Gen-Z 等）或是定制化类型连接器（NPC/CPC），并适配不同协议的高速特性，如 PCIe 5.0 ~ PCIe 7.0（单通道 32GT/s ~ 128 GT/s），SerDes 速率则覆盖 56Gbps ~ 224Gbps。

图 2.23 某高速 I/O 连接器与定制化连接器产品示意图

2.2.2.4 外部高速 I/O

外部高速 I/O 实现跨节点与跨柜间的高速互连，随着互连应用协议的差异化诉求，可以挑选标准类型线缆（QSFP、OSFP-XD 等）或是定制化类型背板连接器线缆，并适配不同协议的高速特性。在产品线缆长度方面，无源 DAC 线缆大多 0.5～2.0 米，有源 ACC 线缆大多 2.0～4.0 米，有源 AEC 线缆大多 4.0～7.0 米。

图 2.24 某外部高速 I/O 线缆产品示意图

2.2.2.5 高速 PCB 板材

对于高速 PCB 的板材损耗要求，按照不同速率的信号链路需求如下：

表 2.2 高速 PCB 板材应用信息表

信号速率	介电损耗（Df）	非国产板材	国产板材
56Gbps	0.003 ~ 0.005	MEGTRON6（G），EM891.TU-883.IT-968 等	NY6300S、H360Z、S6 等
112Gbps	0.002 ~ 0.003	MEGTRON7（N），EM890K、TU-933.IT988GSE 等	NY-P2.S6NX 等
112Gbps	0.0015 ~ 0.002	MEGTRON8（N），EM892K、TU943SN、IT998G 等	NY-P4N、S9N 等
112Gbps 224Gbps	0.0015 ~ 0.002	MEGTRON8（N），EM892K、TU943SN、IT998G 等	NY-P4N、S9N 等
224Gbps	0.0001 ~ 0.0015	MEGTRONS（U），EM892K2.IT998GSE 等	NY-P4.S9N2 等
224Gbps	0.00005 ~ 0.0001	MEGTRON9（U/Q），EM896K2/K3.IT999GSE2/3 等	NY-P5N2/Q、S10GN2/Q 等

2.2.2.6 高速 Raw Cable

更高速率的高速裸线要求尺寸更小，损耗 SDD21 更优。为此，设计中常选用介电常数（Dk）与损耗因子（Df）更低的绝缘材料，或通过增强导体间耦合来降低损耗。

实心绝缘由于材料介电常数 Dk 的限制，无法达到 2.0 以下，常规 PE/PP/FEP 材料的 Dk 值在 2.03～2.3 之间。为了降低损耗，高速裸线的绝缘设计从一层绝缘优化为双层绝缘、发泡绝缘或双芯共挤结构。

SI 典型指标也愈发严苛，例如：中心阻抗与连接器匹配，公差从 ±5Ω 收严至 ±2Ω；线对内延时差（Intra pair skew）从原来的 10 ps/m 降低至 2 ps/m；SDD21 曲线平滑，带宽上升到 40 GHz，67 GHz，甚至到 110 GHz 无突波。

下表仅提供了无地线典型高速铜缆的 SI 性能，实际还可根据需求选用双地线、单地线等不同结构。

表 2.3 高速 Raw Cable 应用信息表

AWG	导体 OD	绝缘结构	地线	包带尺寸（mm）	阻抗/欧姆	112Gbps	224Gbps	448Gbps	
						SDD21 @26.56GHz	SDD21 @53.12GHz	SDD21 @84GHz	SDD21 @110GHz

32	0.203SC	FEP 双芯共挤	无	1.55×0.89	92	7.9~8.4 dB	11.5~12.0 dB	16.0 dB	19.5 dB
30	0.285SC	FEP+FEP	无	1.70×1.20	92	6.1 dB	8.8 dB	12.3 dB	14.5 dB
28	0.35SC	Foam FEP+PE	无	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	无	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	无	2.55×1.32	92	4.1~4.4 dB	7.3~7.7 dB	8.4 dB	9.8 dB

2.2.3 下一代互连技术路径展望

随着 AI 大模型与万卡集群爆发，吉瓦级智算中心互连面临高带宽、低时延与低成本的刚性需求，推动单通道速率向 448Gbps 演进。当前互连技术正沿着铜互连与光互连双路径并行突破。

铜互连正经历系统级革新：从编码调制（PAM4/6/8）到架构（向各式正交演进），再到产品结构升级。核心焦点 CPC（共封装铜互连）通过“芯片－连接器－铜缆”一体化设计，与传统方案相比，具有链路损耗降低、功耗降低、系统成本降低、兼具高密度与可维护性等多种优势。同时，PCB 与 Raw Cable 的演进，随着高频板材、发泡绝缘等新材料的引入，一同助力铜互连不断突破互连距离的长度边界，持续攻克 110GHz 及以上的高速互连难题。

光互连则从可插拔光模块向 NPO/CPO 演进。CPO 技术将光引擎与电芯片共封，电信号仅传输数毫米，极大提升了带宽密度并降低了误码率，是突破面板 IO 密度限制的关键，同时受限于行业生态仍处于探索期，整体产业链尚未完成，需要行业上下游伙伴一同催熟产品化落地。

未来，“光铜协同”将是长期格局，铜互连固守机柜内短距场景，光互连承担跨机柜长距通信，二者优势互补，共同构建高效能智算底座。

2.3 AI 网络与交换系统

2.3.1 规模扩展

AI 智算 Scale-Out 网络正朝着更大规模的集群架构演进，其驱动力主要来自两方面。

1.技术需求驱动：以 GPT、Claude、Deepseek 为代表的先进大模型已验证，模型的智能水平与其参数规模、训练数据量及总计算量直接正相关。为追求更强大的推理、泛化及多模态能力，必须构建由更多 GPU 组成的超大规模集群，这对底层网络的 Scale-Out 横向扩展能力提出了极高要求。

2.经济效率驱动：在激烈的全球 AI 竞争中，缩短模型训练周期已成为核心竞争优势。更大规模的 GPU 集群能够显著加速训练任务，其带来的效率提升也直接降低了单位计算任务的运营成本，实现了效率与成本的双重优化。

2.3.1.1 交换芯片带宽快速演进，驱动智算网络代际跨越

智算交换机芯片与组网架构的协同发展，是支撑万卡级集群演进的基石。其演进路径清晰地反映了从”以架构复杂性弥补带宽不足”到”以芯片性能驱动架构简化”的转变。

初期架构补偿阶段：受限于单芯片交换带宽，业界不得不采用框盒组网、三层盒盒组网等复杂架构，以满足早期万卡集群组网需求。

当前芯片驱动简化阶段：随着交换芯片性能的提升，单芯片带宽已从 12.8T 迈向 51.2T 乃至 102.4T 时代。这使得简洁、扁平的二层盒盒组网成为主流，显著降低了管理复杂度和成本。

这一转变的核心驱动力，来源于以博通为代表的芯片供应商在工艺与架构上的持续突破。其通过采用 5nm、3nm 等先进制程，并迭代推出 51.2T 至 102.4T 的系列芯片，提供更高的集成度与带宽性能，为构建简洁、高效的智算网络奠定了坚实的硬件基础。

图 2.25 博通智算芯片 Roadmap

为匹配交换芯片容量从 51.2T 向 102.4T 乃至更高演进的刚性需求，单通道接口速率与信号调制技术正经历关键代际跃迁。

端口速率与 SerDes 技术发展趋势：当前业界端口仍以 400G 为主，800G 即将爆发。受芯片工艺与生态成熟度影响，基于 112G SerDes 的技术平台将在国内持续较长时间。随着 224G SerDes 技术逐步落地，NPO（近封装光学）光模块预计 2027 年起规模部署，CPO（共封装光学）交换机则有望在 2028 年后实现大规模商用。

SerDes 与调制技术协同演进：当前单通道 SerDes 速率正从 56Gbps 向 112Gbps 快速迁移，Tomahawk5 芯片已全面支持 112Gbps。下一代 Tomahawk6 将进一步引入 224Gbps SerDes。调制技术上，PAM4 已成为高速互连的主流标准，其传输效率为传统 NRZ 的两倍。未来在 102.4T 及以上芯片中，PAM4 将继续与 112G/224G SerDes 结合，构建高带宽、高能效的物理层基础。

2.3.1.2 弹性智算网络架构，支撑超大规模扩展

AI 算力集群的 Scale-Out 网络旨在实现 GPU 服务器的高效、无损互联。其组网架构需根据集群规模弹性伸缩，主要分为以下几种模式。

1.单机直连：适用于小规模推理

场景：适用于数十张 GPU 卡的小型推理或开发测试场景。

方案：将所有 GPU 卡直连至单台高性能交换机，实现最简单的无阻塞网络。如 H3C 基于 Tomahawk5 芯片的 S9827-128DH 系列交换机，单台设备最大可支持 128 个 400G 端口，满足百卡级集群需求。

价值：轻松支撑百卡级集群，部署快捷、运维简单。

2.二层 CLOS 架构：支撑中大规模训练集群

场景：支撑数百至上万 GPU 卡的中大规模训练集群，是当前主流方案。

方案：采用经典的 Spine-Leaf 二级扁平架构。通过选择不同带宽的 Spine/Leaf 交换机（如 12.8T 至 102.4T），并线性增加设备数量，可实现从 512 卡到 16K 卡的灵活、线性扩展，同时保持恒定的低延迟与高带宽。

价值：三跳无阻塞网络，低时延、设计、部署、运维更简单，故障定位容易。

表 2.4 二层 CLOS Scale-Out 网络架构规模

	400G 卡接入规模	400G 卡接入规模
Spine：12.8T，32*400G Leaf：12.8T，32*400G	Spine：16 台 Leaf：32 台	512

Spine：25.6T，64*400G Leaf：25.6T，64*400G	Spine：32 台 Leaf：64 台	2K
Spine：51.2T，128*400G Leaf：51.2T，128*400G	Spine：64 台 Leaf：128 台	8K
Spine：102.4T，128*800G Leaf：102.4T，128*800G，接入一分二	Spine：64 台 Leaf：128 台	16K

注：102.4T 方案中，Leaf 交换机使用 800G 端口通过分支线缆（如一分二）连接 400G GPU 卡，实现端口密度翻倍。

3. 三层 CLOS 架构：构建十万卡级超大规模集群

场景：用于构建数万至十万卡级别的单一超大规模型训练集群。

方案：在二层架构之上引入 Core 层，形成 Core-Spine-Leaf 三级架构。该架构基于 POD（性能优化模块） 单元进行横向扩展，每个 POD 内部为二层 CLOS。通过叠加多个 POD 并经由 Core 层互联，可实现近乎无限的线性扩展能力，是承载下一代万亿参数模型训练的基础设施。

价值： 超大规模扩展方案，二层 CLOS 无法替代。但也存在跳数多，增加时延，设计、部署、运维复杂，成本高昂等问题。

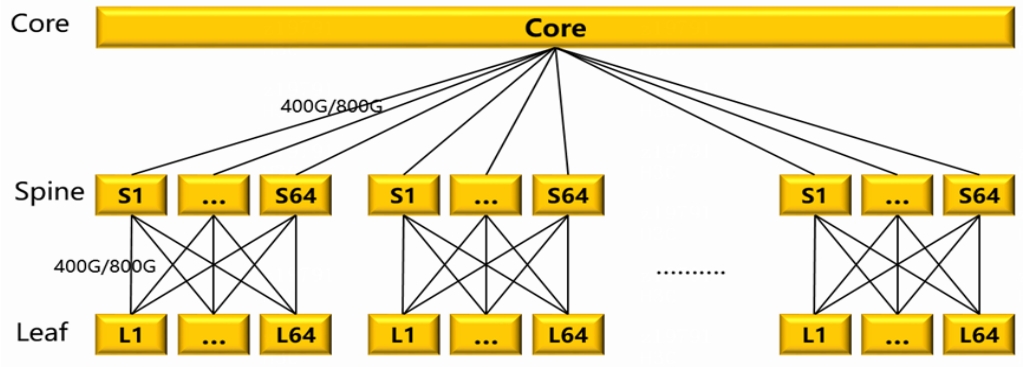


图 2.26 三层 CLOS Scale-Out 网络架构

4. 多平面组网：面向高性能网卡的异构扩展方案

场景：配合支持端口聚合拆分的高性能智能网卡（如 NVIDIA CX8 800G，目前支持 2*400G），使用 2 平面，2 个独立的二层 CLOS 提供数万卡超大集群。

方案：将单台服务器的多个网卡端口分别接入多个独立的二层 Spine-Leaf 网络平面。每个平面为服务器提供一条完整的网络通路，实现跨平面的带宽聚合与故障隔离。基于 Tomahawk6 芯片，单平面可提供 32K 个 400G 接口。

价值：使用简洁低成本的二层 CLOS，依然可以构建数万卡的超大集群。

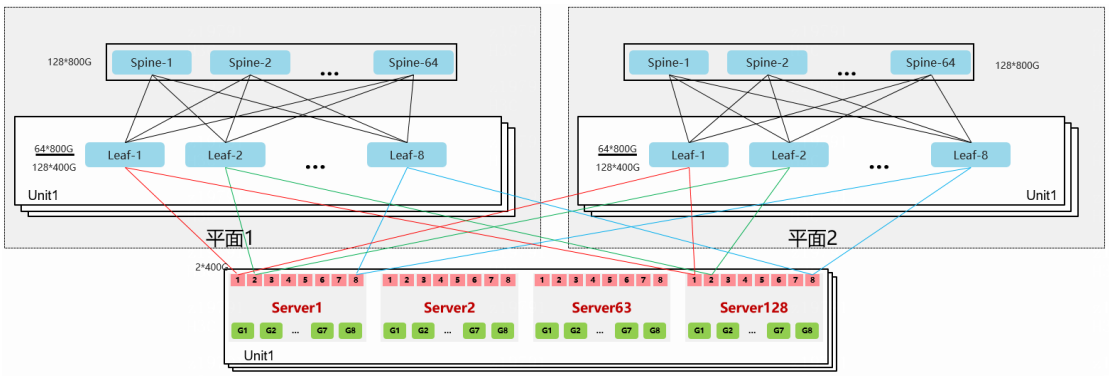


图 2.27 双平面 Scale-Out 网络架构

2.3.2 通信效率

智算网络的流量模型由 All-Reduce、All-to-All 等几种固定的集合通信操作主导。其流量具有低熵（即流数量少、特征单一）、突发、长流的典型特征，极易在交换网络中引发瞬时拥塞。

由于 GPU 训练遵循严格的同步范式，所有节点必须在完成集合通信后，才能进入下一轮计算，容易出现”木桶效应”。因此，由拥塞引发的端到端时延抖动，会直接拖慢整体训练任务，成为智算网络的核心性能瓶颈。

为攻克此难题，业界聚焦于提升网络通信效率，提出了两大技术路径：端网协同调度与纯网络侧调度，旨在实现更精细的负载均衡与拥塞控制。

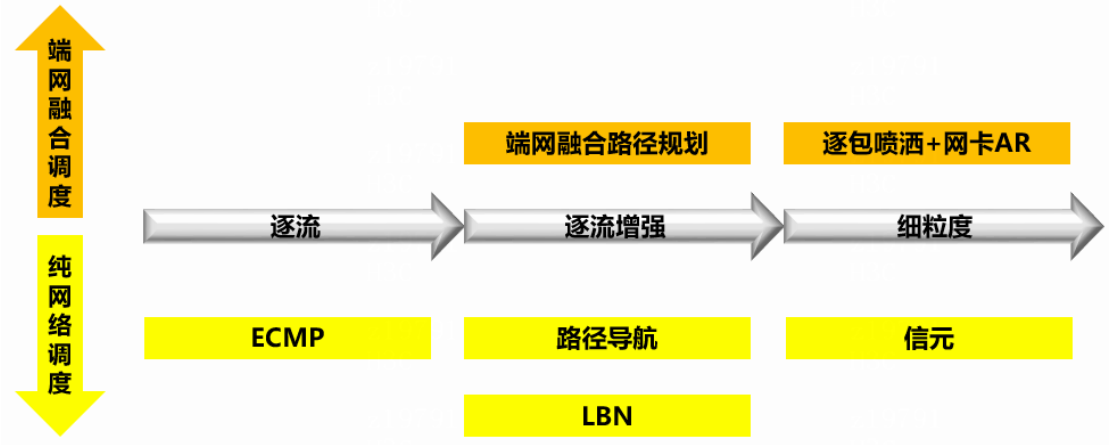


图 2.28 Scale-Out 网络负载均衡技术体系

2.3.2.1 纯网络调度类负载均衡技术

指完全由网络设备（主要是交换机）自主实现流量优化调度的技术方案。不依赖服务器或网卡的任何配合与感知。

表 2.5 纯网络调度类负载均衡技术

项目	描述
ECMP	交换机 N 元组 Hash 算法，目前业界应用最广泛的方式；但由于低熵特性，负载均衡效果可能不佳，可通过多 QP，调整 Hash 算法等方式调优。
路径导航	控制器采集交换机信息，发现拥塞点和贡献流，控制器重规划，将拥塞流调整到轻载路径，只适用于通信关系固定场景，对 MOE 这类流量不可预期场景效果不佳。
LBN	基于入端口 Hash 可确保 Leaf 上行流量均匀，大规模组网下多台 Leaf 经同一 Spine 下行时引发的拥塞概率较高，适用于端侧不支持多 QP 的小型组网场景。
信元	信元交换技术将流量在 Leaf 切分为信元，均匀喷洒至所有 Spine 后再重组，实现负载均衡，天然异构解耦，无需网卡支持且适配任意流量。

2.3.2.2 端网融合调度类负载均衡技术

是一种通过网络侧（控制器、交换机）与计算侧（智能网卡、集合通信库）之间的信息交互与联合决策，实现全局流量优化的技术架构。

表 2.6 端网融合调度类负载均衡技术

项目	描述
端网融合路径规划	控制器通过 CCL 获取 GPU 通信关系，预规划非重叠网络路径以避免拥塞，适用于通信模式固定的训练场景，对 MOE 这类流量不可预期场景效果不佳。
逐包喷洒+网卡 AR（Adaptive Routing，自适应路由）	基于报文级均匀喷洒，搭配接收端网卡 AR 乱序重排能力，负载均衡性能优越，当前主要依赖 NVIDIA 封闭生态，未来有望随国产及博通等网卡支持 AR 而普及。

2.3.3 可靠性与维护

2.3.3.1 交换机节点级可靠性与维护

智算网络通常由数十至数百台盒式架构交换机组成，其硬件设计需确保关键部件（如电源、风扇）支持 M+N 冗余，并具备对 CPU、内存等核心器件的故障监控与运维能力。

2.3.3.2 网络架构级可靠性与维护

网络内部可靠性：核心依赖 ECMP 实现快速自愈。以 H3C S9827-128DH 为例，Spine-Leaf 间可形成 64 路径 ECMP 组，当发生设备或链路故障时，流量自动切换至其他路径，实现无损自愈。结合数据平面快速收敛技术，可将故障切换丢包控制在毫秒级，保障业务连续性。

网络边缘可靠性：多平面冗余架构提升接入可靠性。服务器网卡双归属到 2 个平面的 Leaf，正常流量在双平面隔离转发，当接入链路或设备故障时，依赖端侧 CX8&NCCL 逃生实现故障冗余，避免网络边缘单点故障。

2.3.3.3 链路级可靠性与维护

智算网络依赖数万个光模块互联，其传输稳定性直接关系到 AI 训练与推理任务的连续性。为确保可靠性，业界采取训前与训中两级保障措施。

训前 PRBS 压测：在组网完成、业务上线前，普遍采用 PRBS（伪随机二进制序列）码流对所有光模块进行大规模测试。可提前筛选出性能不达标的光模块，从源头避免因硬件质量问题引发的业务中断。

训中光链路监控：业务运行后，光模块会随时间和环境因素发生概率性性能劣化，可能导致端口故障或链路闪断。为此，业界普遍对光模块的信噪比（SNR）、前向纠错（FEC）计数、符号错误率（SER）与误码率（BER） 等关键指标进行实时监控。通过分析这些指标的劣化趋势，可提前预警并更换潜在故障模块，实现预测性维护，从而保障传输质量，显著降低 AI 任务中断风险。

通过上述”上线前严格筛查”与”运行中持续监控”的组合策略，智算网络能够构建高可靠的光互连层，为大规模 AI 业务提供稳定基础。

2.3.3.4 业务级可靠性与维护

智算网络对拥塞监控提出了更高要求，传统交换机基于秒级的端口统计难以满足需求。

传统监控的局限：交换机通过 ECN 标记、PFC 报文速率等统计可发现网络拥塞，提示性能瓶颈。但这类监控粒度较粗，既无法捕捉毫秒至百毫秒级的瞬时微突发流量，也无法定位引发拥塞的具体业务流，导致突发拥塞难以被有效观测和展示。

微突发监控能力：为解决此问题，智算交换机需以毫秒级粒度统计端口缓存（Buffer）

使用率与转发速率。这使得在微突发发生时，系统能准确捕捉并上报瞬时流量峰值，避免其被秒级平均值”抹平”，从而实现有效的突发监控与问题定位。

流级别监控与调优：为进一步定位拥塞根源，需启用流级别监控。通过为业务流建立流表，并统计每条流的转发时延，可精准识别出时延显著增大的”拥塞贡献流”。结合网络控制器，可将这些流动态重定向至负载较轻的路径，实现基于实时状态的负载均衡调优，从而优化网络性能。

综上，通过毫秒级端口监控与流级时延分析的结合，智算网络能够实现从”发现拥塞”到”定位根源”再到”动态调优”的闭环管理，有效应对高突发、低时延的 AI 流量挑战。

2.3.4 未来展望

2.3.4.1 网络管控软件为智算业务提供支撑

智算网络建设与运维是跨越多层业务的系统工程，需实现网络与交换机、网卡、GPU 等设备的深度融合与协同。网络管控软件以自动化、体系化、智能化为核心，提供三大关键能力。

端网协同，极速开局：通过全局可视化与自动化部署，实现设备、GPU、网卡的统一纳管，支持策略一键下发与健康压测，大幅提升大规模组网交付效率。

全局掌控，网络调优：针对智算流量低熵、突发、大象流等特征，业务级感知拥塞，动态调优实现动态负载均衡与全局路径优化，提升网络吞吐与训练效率。

智能运维，训练护航：构建训前巡检、训中实时监控（路径、拥塞、哈希）、训后故障定界与回溯的全周期保障体系，实现快速感知、定位与优化。

由此形成”快速交付-动态调优-智能运维”闭环，为智算业务提供稳定、高效、可视的网络支撑。

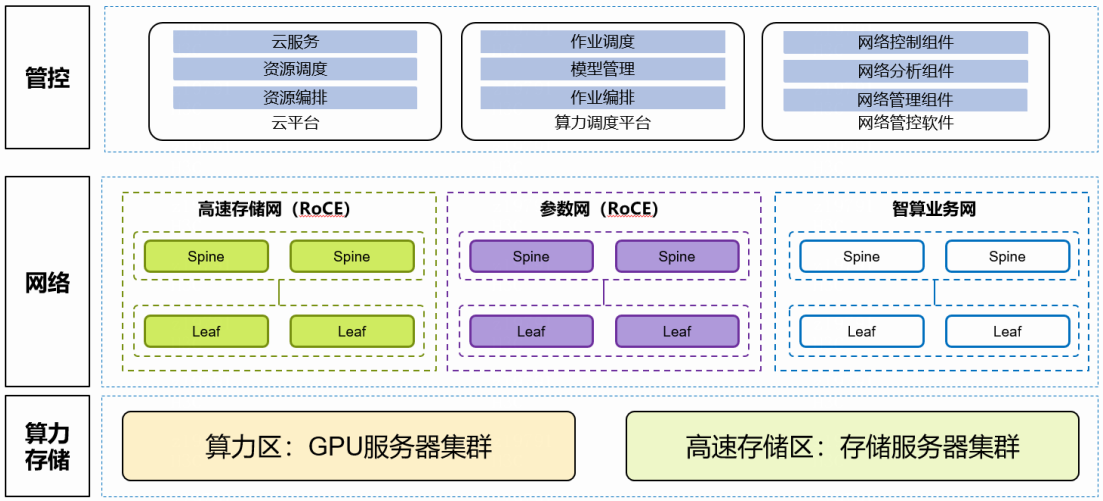


图 2.29 智算网络管控软件架构

2.3.4.2 开放解耦的良好生态环境

智算网络主要由交换机，网卡，GPU 构成。我们以一次典型的 GPU 间 RDMA Write 操作为例，DMA 操作过程如下。

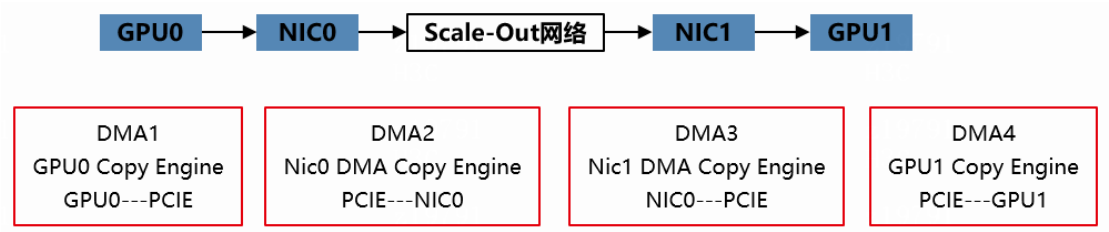


图 2.30 硬件视角 RDMA 流程参与者

智算网络建设常面临生态封闭与集成挑战。当前市场主要由不同厂商提供交换机、网卡、GPU 等独立部件，需通过系统集成构建完整方案。然而，部分大型厂商凭借全栈产品能力，倾向于推广封闭的”全家桶”方案，客观上限制了客户选择业界最优技术与产品组合的机会，也带来了供应链单一、成本缺乏竞争等问题。

为应对这一挑战，DDC（多元动态联接）技术应运而生。其核心价值在于实现网络与上层 AI 框架及异构算力的解耦。DDC 通过将信元喷洒、报文重组与重排序等功能完全集成到网络侧，无论客户使用何种品牌或型号的 GPU 与网卡，即使其本身不具备高级乱序处理能力，也可通过 DDC 网络获得稳定、高性能的传输体验，真正实现”异构兼容、性能无忧”。该技术不仅打破了生态绑定，也为客户构建开放、高效、自主可控的智算网络提供了关键技术支撑。

2.4 AI 存储技术与系统

2.4.1 概述

面向大型机房的 AI 存储基础设施建设，对于承担训练热数据、推理缓存、向量索引文件持久化与高速读取、Checkpoint、并行文件服务、元数据服务的存储节点而言，其**能力边界主要由以下因素共同决定**：一是 CPU 提供的 PCIe 通道规模、内存通道规模及 NUMA 拓扑特征；二是前端 NVMe 介质密度与本地闪存池能力；三是 400G 乃至新兴 800G 级 RDMA 网络的接入方式。

基于当前大型机房的部署实践，高密度全闪节点已成为高性能 AI 存储服务器的重要代表形态，此类节点通常提供热数据层的高带宽数据服务，其设计目标是实现 CPU、SSD、内存与高速 RDMA 网卡之间的系统性匹配。

2.4.2 AI 存储技术

2.4.2.1 高速推理存储架构

高速推理存储架构可以概括为 GPU 显存、CPU 内存、本地 SSD、网络 SSD 四层。GPU 显存承载最热的 KV Cache 和当前请求上下文；CPU 内存作为活跃缓存与换入换出缓冲；本地 NVMe SSD 提供低时延容量扩展；网络 SSD 或远端闪存池用于跨节点共享、弹性扩容和温冷数据承载。

这一架构的意义在于把推理上下文按访问热度分层放置，使 GPU 不必因长上下文、多轮对话或 Agentic AI 任务而频繁等待数据。

图 2.31 高速推理存储四层架构示意图

NVIDIA Inference Context Memory Storage Platform（简称 CMX），是 NVIDIA 在 2026 年初 CES 发布、面向长上下文 LLM / 多轮对话 / 智能体（Agentic）AI 推理的专用 “推理上下文记忆存储平台”，核心是用 BlueField-4 DPU 做硬件加速，在 GPU HBM 和传统存储之间插入一个第 3.5 层（如下表）的 Pod 共享上下文层，专门解决 KV 缓存爆显存、延迟高、功耗高的问题。

表 2.7 NVIDIA 把整个 AI 内存 / 存储层级划成 G1～G4，CMX 插在中间：

层级	位置	介质	延迟	用途
1	GPU 本地	HBM3e	~100ns	模型权重、最热 KV
2	CPU 近内存	Vera CPU+LPDDR5X	~1 µs	溢出 KV、中等热度
3.5（CMX,第 3.5 层）	Pod 共享	BlueField-4 + Flash	~5 µs	长上下文 KV、多轮共享、Agent 记忆
4	集群 / 云	传统 NAS/S3	~ms	冷数据、持久化

2.4.3 AI 存储节点

2.4.3.1 CPU

在存储服务器架构中，x86 CPU 的核心价值主要体现在成熟的 I/O 生态、稳定的企业级 RAS 特性以及对主流存储软件栈的长期适配能力。对于 AI 存储节点，CPU 选择需考虑以下问题：单个服务器如何支持大量 NVMe SSD 的直通接入；以及如何挂接一个或多个高速 RDMA 网卡。

平台规划首先要看 PCIe 通道数量。企业级 NVMe SSD 普遍以 PCIe x4 接口接入；400G/800G 级高速网卡通常以 PCIe x16 插槽接入，此外还需叠加启动盘、BMC、板载控制器等设备的通道占用。就 CPU 可直接引出的 PCIe 5.0 通道数（CPU 原生直连口径，非主板最终可用数）而言：Intel Xeon 6（Granite Rapids，第六代 Xeon）单路最高约 96 条；AMD EPYC 9005（Turin）单路最高 128 条，双路配置下为 160 条。需要注意，双路系统并非简单翻倍——两颗 CPU 间需占用部分通道做 UPI/Infinity Fabric 互联，且实际可用通道还会被启动盘、BMC 等占用。因此不同平台在单路直连能力、双路扩展方式、以及

是否需要外加 PCIe Switch 扩展上，会呈现明显差异。

当节点部署多个高速 RDMA 端口时，CPU 与网卡、SSD 之间的亲和性尤为关键。网卡应尽量与承担主 I/O 的 NVMe 池位于同一 NUMA 域，减少跨 Socket 数据搬运；采用双路架构，则要充分评估跨 Socket 路径对带宽效率和尾时延稳定性的影响。

2.4.3.2 内存

如下两个场景会显著增加内存的容量配置。

一. KV Cache 使用内存缓存时，其容量与 token 数、模型层数、KV 头数、每头维度、数据类型及并发会话数近似呈线性关系。在长上下文或高并发推理场景下，缓存需求极易突破单卡显存上限。为此，业界通常采用 **GPU 显存 → CPU 内存 → 本地 SSD→ 网络 SSD** 的四级分层卸载策略。CPU 内存不仅作为中间缓冲层存放换出的 KV block，还承担预取与异步换入换出的调度职责。在容量规划上，内存一般配置为 GPU HBM 的 **3－6 倍**；以 4 张 H100（共 320 GB HBM）为例，推荐搭配 **1－2 TB DRAM**，专用于存放从 GPU 显存换出的非活跃 KV block。

二. 并行文件系统的元数据服务层需要大量内存，内存容量与 inode、dentry、条带映射和客户端锁表构成的元数据热集密切相关。其中客户端锁表的内存占用通常与并发客户端数近似呈**线性增长**；在高锁争用场景下，由于锁回调和锁升级频繁，增长速度可能超过线性，规划时需预留额外容量。

2.4.3.3 NVMe SSD

在本地持久化介质层面，NVMe SSD 是 AI 存储服务器的核心组成部分。主流配置方向为企业级 NVMe SSD，常见形态包括 U.2/U.3 以及 E1.S、E3.S 等 EDSFF 规格。AI 训练、推理、Checkpoint、向量索引以及高并发元数据访问等场景，对低时延、高并发随机读写和稳定顺序吞吐具有明确要求，因此 SSD 选型应综合评估稳态性能、写入一致性、掉电保护、热管理表现和长期可运维能力。

从接口代际看，PCIe 5.0 x4 已成为高性能企业级 NVMe SSD 的主流基础，其单盘顺序读写、随机 IOPS 和队列并发能力相比 PCIe 4.0 有明显提升，能够更好地承接训练热数据、Checkpoint 写入、向量索引加载和推理缓存扩展等负载。对于 AI 存储节点，Gen5 SSD 的意义不仅是单盘速度更快，还在于更少盘数即可形成更高节点吞吐，从而降低盘位、功耗、PCIe Switch 和网络出口之间的结构性压力。

PCIe 6.0 将链路速率提升到 64 GT/s，并引入 PAM4.FLIT 编码与更强的纠错机制，面向下一代 Gen6 NVMe SSD 提供更高的主机侧带宽上限。随着 Gen6 SSD、PCIe 6.0 Switch、CXL 相关内存扩展和 800G/1.6T 网络逐步成熟，AI 存储节点可以在更少 PCIe 通道、更少盘位或更低访问层级跳数下提供更高聚合带宽，对长上下文推理、KV Cache 分层、实时特征读取和多租户向量检索具有直接价值。

更长期看，PCIe 7.0 规范已面向 128 GT/s 链路速率演进，未来更高代际 SSD 将推动本地闪存层继续靠近内存层的访问能力。对 AI 存储而言，先进 SSD 技术的价值不是简单追求峰值跑分，而是缩短 GPU 等待数据的时间，提升缓存命中后的恢复速度，降低热数据层的扩容成本，并为存储解耦的 GPU 节点（本地少盘/无盘、存储池化到机架级）、网络上下文缓存和机架级闪存池提供更可靠的介质基础。

介质路线方面，TLC 与 QLC 的适用位置应结合工作负载类型进行区分。TLC 更适合训练热数据层、Checkpoint 层、元数据层和混合读写层，其稳态写入能力和耐久性更适合长期重载场景；QLC 更适合读多写少、容量优先的 AI 场景，例如推理缓存扩展层、向量库主数据层、RAG 原始索引层和只读数据集层。

NVMe SSD 的错误诊断和故障提前预测可以通过标准化管理能力系统实现设备监控。NVMe-MI 规范要求 NVMe 设备的发现、监控、配置和更新在带内与带外路径上都具备统一接口，便于大规模节点的预测性维护。另一方面，FDP（Flexible Data Placement）技术通过让主机向 SSD 提供更细粒度的写入放置提示，减少垃圾回收和写放大，进一步改善介质寿命、性能一致性与空间效率。

形态与功耗方面，新型接口例如 E1.S、E3.S 会替换传统盘位描述方式。SNIA 对 EDSFF 的公开说明强调，其相较传统形态在闪存密度、供电能力、热设计和可维护性方面具有更好的适配性，有利于在单节点中容纳更多高性能闪存。单盘功耗方面主要注重更低的写放大、更合理的风道和更稳定的热控制曲线。

表 2.8 TLC 与 QLC 在 AI 存储中的适用位置

介质类型	更适合的负载	主要优势	需要重点控制的风险
TLC	训练热数据、Checkpoint、元数据层、混合读写缓存层	稳态写入更强、耐久性更高、尾时延更易控制	成本更高，应优先用于性能一致性要求最高的层
QLC	推理语料、向量库主体、读多写少数据集、容量扩展层	容量效率更高、单位成本更低	持续重写和高频随机写会更快暴露稳态性能与寿命约束

2.4.3.4 RDMA 网卡

在 AI 存储服务器中，高速 RDMA 网卡是与 CPU 和 NVMe SSD 同等重要的核心部件。随着本地 NVMe 池聚合能力持续提升，400G/800G 级 RDMA 网络接口已经成为高性能 AI 存储节点的基础配置方向，新一代高速网卡会对 PCIe 5.0 或者 PCIe 6.0 的主板与 CPU 平台提出更高要求，网络规划必须与主机平台一体设计。

GPU Direct Storage（GDS）是 NVIDIA 提出的一种存储 I/O 加速技术，其原理是通过在 GPU 显存与存储设备之间建立直接 DMA 数据路径，绕过 CPU 内存中的 bounce buffer，从而降低 CPU 负载并缓解系统带宽瓶颈。在本地存储场景下，GDS 直接打通 GPU 显存与 NVMe SSD 之间的 DMA 通路；AI 存储节点结合 RDMA 网卡进一步缩短 GPU 数据路径、降低存储节点的 CPU 介入。

在介绍 EBOF、JBOF 和 存储解耦架构前，需要先明确这些术语的边界。JBOF（Just a Bunch of Flash）通常指由多块 NVMe SSD 组成的外置闪存框，通过 NVMe-oF 对外提供共享块访问；EBOF（Ethernet Bunch of Flash）可以理解为以以太网/RoCE 承载的 JBOF 形态；存储解耦则是将 GPU 计算节点设计为无本地持久化数据盘，计算侧通过高速网络访问远端闪存池或并行文件系统。三者共同指向“计算与闪存解耦”的资源池化方向。EBOF/JBOF/Diskless,并行文件系统和分布式存储都高度依赖高速 RDMA 网络，以保证多客户端并发访问的持续带宽和低时延。在推理体系中，无论是 KV Cache 的远端网络存储分层，还是上下文共享与迁移，都依赖高速网络的支撑。

2.4.3.5 DPU 网卡

基于 DPU 的高速推理存储架构是 NVIDIA 的 Inference Context Memory Storage Platform 平台。该平台由 NVIDIA BlueField-4 DPU 驱动，面向长上下文、多轮和 Agentic AI 推理，目标是在**机架级乃至跨节点集群层面**扩展 GPU 上下文容量，并支持高带宽 KV Cache 共享。网络底座采用 **NVIDIA Spectrum-X 以太网**，提供基于 RDMA 的 KV Cache 高速访问；软件层面与 **NVIDIA Dynamo** 协同，通过异步传输库 **NIXL** 实现 KV Cache

在 HBM、CPU 内存与片外存储间的无缝流动。该平台的出现说明 DPU 在 AI 存储中的角色，正从传统协议卸载进一步演进到上下文存储、缓存路由和高效数据共享层。

2.4.3.6 存储盘扩展

存储盘扩展可以从“盘柜扩容”进一步理解为“闪存资源池扩展”。JBOF 是把大量 NVMe SSD 放入独立闪存框，通过 NVMe-oF 对外暴露共享访问能力；EBOF 则强调该闪存框主要通过以太网和 RoCE 接入计算网络。两者的核心价值在于将闪存资源从计算节点解耦，作为独立资源池统一管理，通过 NVMe-oF over RoCE 或 NVMe-oF over TCP 向上层节点提供可组合的闪存访问服务。这种架构有利于在扩容、维护、固件升级和介质更换时实现更细粒度的故障域隔离，也更适合大型机房按资源池方式管理闪存。

Diskless 架构是在上述资源池化思路上的进一步推进，即 GPU 计算节点不再配置用于业务数据持久化的本地盘，而是把模型、数据集、向量索引、Checkpoint 或上下文缓存放在远端闪存池、并行文件系统或对象/文件融合存储中。部分实现仍保留少量本地 NVMe 用于系统启动、日志或临时缓存，但核心数据路径依赖高速网络访问远端存储。该方式特别适合需要弹性扩容、快速替换无状态计算节点、统一管理闪存容量，以及按训练、推理、检索等角色拆分故障域的 AI 集群。JBOF/EBOF 与 Diskless 架构结合后，可以形成从本地缓存到远端闪存池的多层 AI 存储体系。

第三章 网络与高速互连层

本报告将高速互连的物理层与连接器（介质、端接、线缆、PCB）归入第二章 IT 设施层，将互连协议与网络范式（Scale-Up/Out/Across 的协议、拓扑与软件栈）归入本章；第二章 2.3 侧重交换系统的硬件形态与部署，本章 3.2 侧重 Scale-Out 的协议与调优，二者互为补充。

智算集群的互连体系可划分为 Scale-up 纵向扩展、Scale-out 横向扩展与 Scale-across 长距离扩展三个层级。其中，Scale-up 架构聚焦单一逻辑计算节点内部的资源整合，实现多 GPU 间统一寻址，突破传统总线的带宽与时延瓶颈。Scale-out 横向扩展依托无损网络完成节点间通信，侧重集群规模扩容。Scale-across 架构则将分布在不同区域的 Scale-out 集群进一步扩展，形成跨区域的算力池，常用于万卡以上超大规模智算中心。三种互连网络形成互补的混合算力架构，大幅提升节点集群算力密度与并行计算效率。

3.1 Scale-Up 技术分析

3.1.1 Scale-up 研究背景与研究意义

当前大模型对底层算力基础设施的计算性能、存储容量与数据交互能力提出了严苛要求。算力需求不仅是单纯的芯片数量叠加，而是对低延迟、强一致性、高算力密度的刚性要求。GPU 计算能力与存储空间的扩展性能，直接决定模型训练、推理及多节点协同计算的效率。传统互联架构普遍存在带宽瓶颈突出、通信时延偏高、扩展能力有限等问题，难以适配模型训推场景下海量数据高频交互的需求。在此产业背景下，Scale-up 纵向扩展成为提升算力的关键技术路径。

在智算集群基础设施发展初期，纵向扩展架构普遍采用 PCIe 总线实现互联。随着算力与互联技术的持续迭代，传统架构逐步向高性能超节点架构演进，当前超节点场景中主流的 Scale-up 互联协议与技术体系主要包括 NVLink、OISA、UALink、SUE、ESUN、CLINK、Co-Fabric 等。

3.1.2 卡间高速互连协议与技术标准

3.1.2.1 NVLink 协议

NVLink 是英伟达于 2014 年提出的高速互连技术，与传统架构中 GPU 必须通过 CPU 转发数据不同，NVLink 采用 GPU 直连设计，配合专用 NVSwitch 交换芯片，可使多颗 GPU 直接进行数据交互，大幅提升通信效率。NVLink 支持内存一致性与统一内存寻址，能够将多颗 GPU 的本地显存整合为一个大容量统一内存池。同时，该技术集成硬件加速协议，对 All-Reduce 等集合通信操作进行优化，降低多 GPU 协同的通信时延，提升数据交互效率与稳定性。

作为 Scale-up 架构的标杆方案，NVLink 历经多次迭代，性能持续跨越式提升。从初代 NVLink 1.0 双向 160GB/s，发展到当前商用的 NVLink 5.0 实现双向 1.8TB/s，性能提升超过 10 倍；下一代 NVLink 6.0 预计 2026 年商用落地，双向带宽将进一步提升至 3.6TB/s。在系统级部署上，英伟达 Blackwell Ultra NVL72 机架可实现 72 颗 GPU 的全互联（all-to-all）拓扑，整机柜通信总带宽可达 130TB/s。

NVLink 虽在性能上表现突出，但也存在明显短板，最核心的问题是生态高度封闭、厂

商绑定强、兼容性差。虽在 2025 年推出 NVLink Fusion 开放互连技术方案，但厂商面临严格的授权准入门槛，目前仍未大规模落地应用。用户一旦采用便深度锁定在英伟达硬件体系内。其次，成本高昂，NVLink 需搭配专用 NVSwitch 芯片、专属背板与供电散热设计，对整机架构要求严苛，部署与运维成本高于其他通用方案。这些局限使其在建设中面临较高的落地与适配门槛。

3.1.2.2 全向智感互联 OISA 技术

中国移动于 2024 年提出面向智算中心 Scale-up 场景的全向智感互联 OISA（Omni-directional Intelligent Sensing Express Architecture）技术体系，并于 2025 年发布更新的 OISA 2.0 协议。该技术可支持超节点内 1024 张 GPU 的高效互联，卡间互联带宽突破 TB/s 级别，时延缩短至数百纳秒，为大模型训练、推理等数据密集型 AI 应用提供有力支撑。

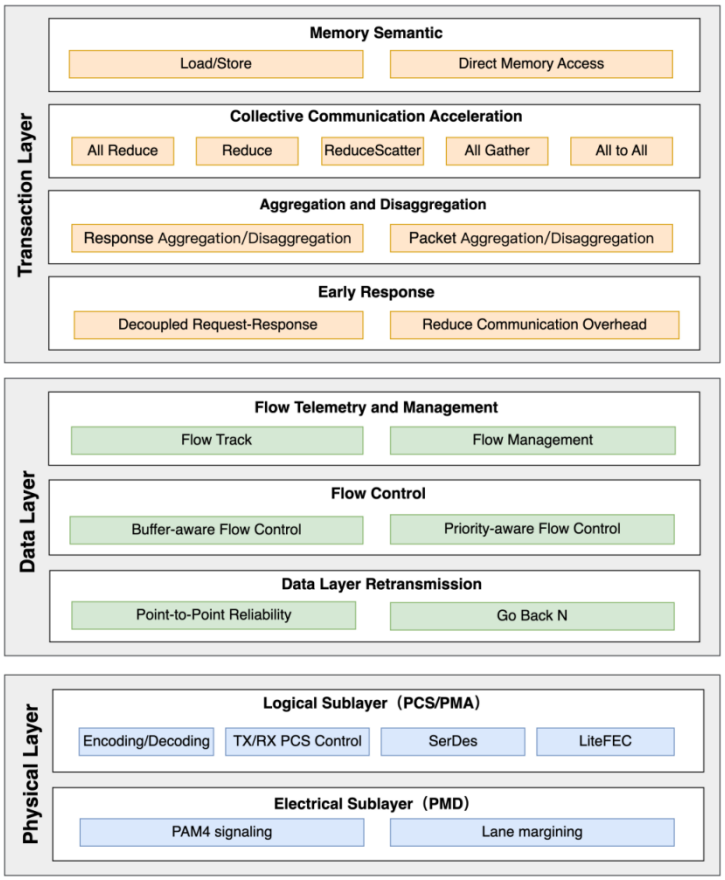


图 3.1 OISA 协议架构

如图 3.1 所示，OISA 的协议栈由事务层（TL）、数据层（DL）和物理层（PL）组成。事务层位于协议栈的最顶层，它与 GPU 的片上网络进行交互，负责接收 GPU 上与事务相关的信号，封装成事务层报文（TLP）后转发到数据层。数据层进行流量控制与智能流量感知，保证传输稳定性。物理层采用基于以太网的实现方式，利用成熟的以太网物理层技术（如 SerDes）来支持 GPU 设备之间的高速传输。

OISA 2.0 协议具备多个特征：一是支持原生内存语义，使 AI 芯片得以跨节点、无障碍地进行数据访问；二是创新 TLP 报文重构技术，可聚合微小计算事务，提高带宽利用效率；三是支持智能感知，实时监测全域 GPU 与交换芯片的互联状态，为动态路由、拥塞控制和

智能运维提供数据支撑；四是集合通信硬件加速，将 AI 训练中的集合通信等复杂操作，通过专用硬件引擎卸载至交换芯片，为大模型训练等场景带来显著的性能提升。

通过构建统一开放的技术平台，OISA 为超节点卡间互联提供全新的国产解决方案，是国内 Scale-up 技术的代表性方案。这将有效激发产业链创新活力，推动相关产品的技术升级与迭代创新，构建良性互动、持续进化的智算产业新生态。

3.1.2.3 SUE 技术

为实现大规模场景下 XPU 之间的高效传输，博通于 2025 年 4 月发布 SUE（Scale-up Ethernet）协议，如图 3.2 所示为 SUE 的协议栈架构。除了标准模式外，SUE 还提供了名为 SUE Lite 的精简模式，精简包格式和包头长度，仅保留用于数据转发的包头以减少硬件开销和传输延迟的目的。

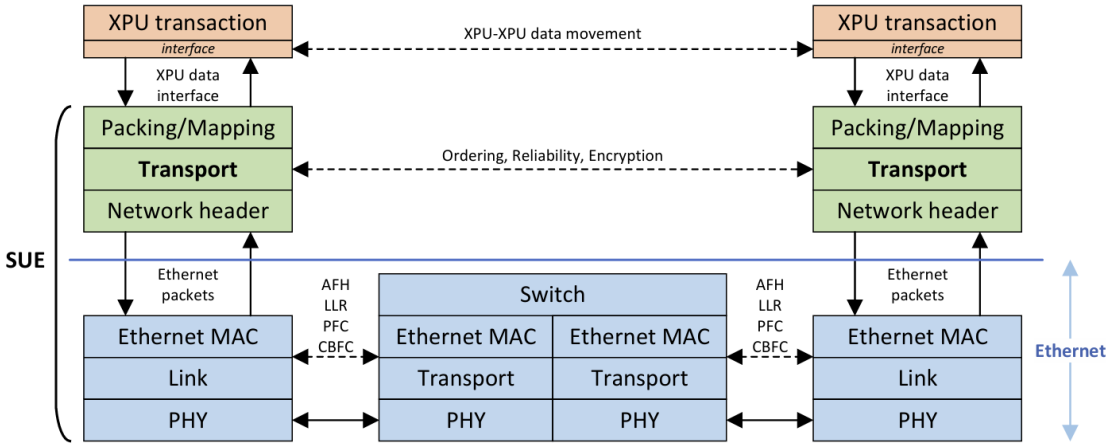


图 3.2 SUE 协议架构

SUE 在传输可靠性设计上支持基于信用的流量控制（CBFC）、基于优先级的流量控制（PFC）与链路层重传（LLR）技术，构建端到端无损传输体系。三者协同工作，显著降低时延波动与丢包率，满足大模型训练与推理场景对传输稳定性的严苛要求。

2025 年 6 月，开放计算项目（OCP）宣布成立 SUE-T（Scale-up Ethernet Transport）工作组，旨在完善 Scale-up 以太网互联的协议体系。SUE-T 将承接并继续推进部分原有的 SUE 工作中关于传输层的核心工作，聚焦端侧传输层协议，实现传输层与交换层的解耦。

在硬件生态层面，SUE 适配博通 Tomahawk Ultra 与 Tomahawk 6 系列交换芯片，依托交换芯片的转发性能，为大规模 Scale-up 集群提供充足带宽和低时延。SUE 与交换芯片的深度协同，让基于以太网的 Scale-up 架构成为当前 AI 超节点从技术方案走向商用部署的核心支撑。

3.1.2.4 UALink 协议

2024 年 10 月，为打破私有协议的垄断，让不同厂商的加速器和交换芯片能够高效协同，AMD、AWS、Astera Labs、思科、谷歌、惠普、英特尔、Meta 和微软等九大公司宣布成立 UALink（Ultra Accelerator Link）联盟。UALink 协议是 UALink 联盟主导研发的核心技术，于 2025 年 4 月正式发布 1.0 版本，于 2026 年 4 月更新 2.0 版本。协议向联盟成员开放。该协议架构如图 3.3 所示，从上到下分别为协议层、事务层、数据链路层和物理层。

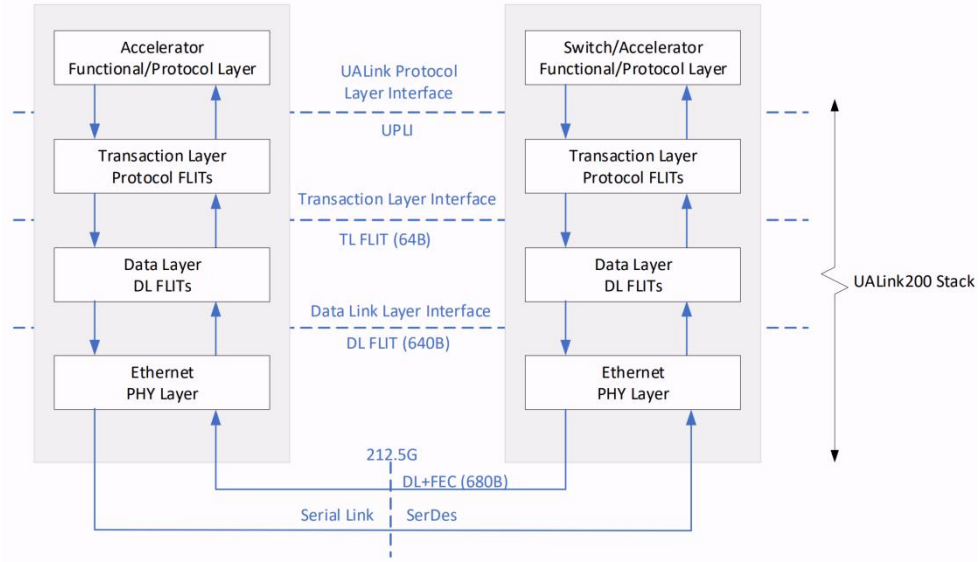


图 3.3 UALink 协议架构

UALink 协议层也称作 UPLI（UALink Protocol Level Interface），是协议和加速器的交互接口。其事务层负责将协议层的指令拆解为标准的事务层传输单元 TL Flit（64Byte），并传递给数据链路层。其数据链路层负责为每个 Flit 添加 CRC 和帧头，封装成数据链路层传输单元 DL Flit（640Byte）并传递至物理层，此外还有链路重传和信用流量控制等机制保障传输可靠性。其物理层基于 IEEE 802.3dj，进行了部分修改，让协议可直接复用以太网成熟的 SerDes、线缆等供应链资源，降低硬件量产与部署成本。

UALink 集成 UALinkSec 安全机制，支持端到端数据加密与可信执行环境适配，可实现租户流量隔离，适配云原生多租户场景的安全合规要求。但是，UALink 不支持一致性协议，数据一致性需要厂商自己来设计实现，例如通过软件或者其他的方式实现数据的一致性。在产业推广上，预计首批支持 UALink 规范的产品将于 2026 至 2027 年商用落地，成为 AI 智算基础设施开放化的核心支撑。

3.1.2.5 其他卡间互联协议

PCIe 技术凭借产业链成熟、成本可控、生态兼容性强的优势，成为智算基础设施发展初期快速落地的技术路线。其典型架构采用单节点内全互连拓扑，节点间通过 PCIe Switch 实现多机通信，解决万亿参数大模型对大显存空间和低延迟通信的核心需求。但受限于协议栈架构与传输特性，导致节点间无法实现原生点对点直连、影响带宽利用率。

浪潮信息对 PCIe 协议进行深度改造，开发 Co-Fabric 互连协议，打造 32 卡、64 卡、128 卡及更大规模 Scale-up 智算系统。Co-Fabric 在通用总线协议基础上增加了代理层（Agent Layer）与适配层（Adapter Layer），突破原生总线协议扩展限制，支持百卡以上全交换无阻塞互连。通过本地虚拟设备实现对远端 GPU 显存的透明化映射，构建全局统一地址空间。通过交换域内的高效数据转发，将通信时延压缩至百纳秒级别。以 64 卡超节点为例，运行 DeepSeek R1 大模型 Token 生成速度达 8.9ms，运行 DeepSeek V3 大模型单卡吞吐达 631.4 tokens/s。

2025 年 9 月，由 ODCC 主导、中国信通院与腾讯牵头，发布 ETH-X 互联协议技术报告。该规范事务层定义了 PAXI（Peer to Peer AXI）的实现规范，将 AXI 总线事务通过以太网扩展至远端节点；数据链路层定义了帧结构和重传机制；物理层兼容标准以太 IEEE 802.3 的 PMD 和 PCS 定义。ETH-X 旨在基于以太网生态提供 Scale-up 互联能力。

2025 年 10 月的 OCP 峰会上，AMD、Arista、博通等 12 家科技巨头共同发起 ESUN

（Ethernet for Scale-Up Networking）联盟，专注于在加速器和交换机中实现较低网络层的标准化，与聚焦传输层的 SUE-T 协同，共同推动开放以太网在超节点中的规模化应用，打破专有协议垄断。

2025 年 11 月，在国家工信部指导下，中国电子标准化研究院发起计算互联总线协议（CLink）联合倡议。通过标准化引领计算技术创新，共同推动计算互联总线协议标准簇持续迭代更新。CLink 涵盖 Scale-up 总体架构设计、通信语义、流量控制等核心内容，为大规模智算集群提供统一技术规范，提升规模化部署效率。

此外，还有阿里云发起的 Scale-up 开放生态 ALS（ALink System，加速器互连系统）；字节跳动提出基于以太网的 EthLink 方案，支持 Load/Store 语义和 RDMA 语义；华为公开了面向超节点的互联协议——UB（Unified Bus，灵衢统一总线）2.0 技术规范，可实现 CPU、GPU、NPU、DPU、Switch 等异构处理单元的对等直接通信；阿里云和计算所发起的高通量以太网 ETH+，协议标准支持开放解耦思想；探微芯联自研的 ACClink，预计相关产品 2026 年底完成技术验证；中兴自主研发的 OLink 技术，兼容 RDMA 标准；海光信息全面开放互联总线协议（HSL），降低 GPU 与 CPU 之间的协同延迟和适配成本。

3.1.3 未来发展趋势与应用潜力

Scale-up 技术未来将沿着物理层、协议栈、边界扩展三大方向持续优化。作为底层核心支撑，基于 OIF CEI-448G 框架的 448Gbps 单通道物理层技术已于 2025 年 11 月完成标准框架发布，该技术可在不改变端口密度的前提下实现链路带宽翻倍，大幅提升单机柜算力密度，但也对 PCB 材料、连接器的信号完整性提出极高要求，须配合低损耗材料与光电融合方案才能稳定部署。承接物理层的演进需求，光互连技术将成为突破电互连极限的核心方案，近封装光学（NPO）、共封装光学（CPO）等技术后续将逐步商用落地，具体内容详见本书 2.2 节。

协议栈中的在网计算技术已在头部厂商大模型训练集群实现初步落地，实现集合通信算子的标准化硬件卸载，减少端侧资源的消耗。但也面临算子可编程性受限、跨厂商兼容性差、芯片设计复杂度与功耗攀升的问题，流控与计算调度的协同失配还可能引发新的拥塞风险。该技术将不断迭代，共建生态。

同步推进的还有架构边界的延伸，Scale-up 将从单机柜向跨机柜的二层网络扩展，可突破单机柜算力上限，简化大规模集群的组网复杂度，但由于跳数增加导致时延增大、多路径负载均衡、自适应路由、报文乱序、广播域扩大引发的广播风暴风险、单点故障的影响范围显著扩大等一系列问题，需平衡集群规模和互联性能。整体来看，Scale-up 技术将持续向更高带宽、更低时延、更优适配的方向演进。

3.2 Scale-Out 网络技术

3.2.1 Scale-Out 网络需求和目标

3.2.1.1 Scale-Out 需求

智算中心的横向扩展（Scale-out）网络需同时支撑大规模分布式训练与推理任务的跨节点通信，核心需求包括超高带宽、极低时延和无损传输。其中，训练任务主要由数据并行的梯度 All-Reduce、流水线并行的跨节点传递产生 Scale-out 流量（张量并行等紧耦合通信则主要由 Scale-up 域承载）；推理任务则随 PD 分离、MoE 专家并行（All-to-All）

与跨节点 KV 缓存共享的兴起，对 Scale-out 提出越来越强的带宽与时延要求。

1.高带宽需求

当网络带宽不足时，数据并行通信导致的等待时间可占模型训练迭代时间的 55-85%，成为主要性能瓶颈。为了在毫秒或微秒级时间内完成训练同步，每个计算节点都需要极高的网络出口带宽，例如 400Gbps、800Gbps 乃至 1.6Tbps。

2.极低时延需求

集合通信常常要求所有工作节点必须等待最慢的通信完成才能进入下一迭代，因此网络尾部延迟直接决定了整个训练作业的进度。另外，推理的 PD 分离与 KV 缓存传输同样对端到端时延和抖动高度敏感。为此，Scale-out 网络需要提供微秒级的端到端延迟并尽量减少抖动。

3.无损传输需求

数据包丢失会触发过多的重传，导致网络性能急剧下降。因而，Scale-out 网络基于高性能通信协议 RDMA，实现工作流量无损传输。

3.2.1.2 Scale-Out 网络的目标

Scale-out 网络的目标是最大化分布式训练的整体通信性能，并最小化因网络导致的 XPU（GPU、TPU、LPU、NPU 等）空闲等待时间，其设计指标包括 1）支持高达 100 GB/s 的带宽，和 2）在数百米范围内实现低于 10 微秒的端到端往返延迟。

3.2.2 网络操作系统 SONiC

SONiC（Software for Open Networking in the Cloud）是由微软于 2015 年发起并开源的基于 Linux 的开源网络操作系统，目前属于 OCP 网络领域的子项目。SONiC 旨在通过软硬件解耦为数据中心提供高度可扩展的网络操作方案，其支持多种 ASIC 芯片和 CPU 架构，打破了传统操作系统的硬件绑定限制。SONiC 具备强大的生态支持，涵盖 Broadcom、Mellanox 等芯片商以及 H3C、Dell、Arista 等设备厂商，并已被百度、腾讯、LinkedIn 等数百家大型组织在生产环境中部署。

SONiC 的核心架构建立在交换机抽象接口（SAI）之上，利用 Docker 容器技术将网络功能模块（如 BGP 等）封装为独立进程，实现了软硬件及协议模块间的完全解耦，允许运营商在多个厂商的硬件上共享相同的软件堆栈，支持快速增加新组件或第三方软件。

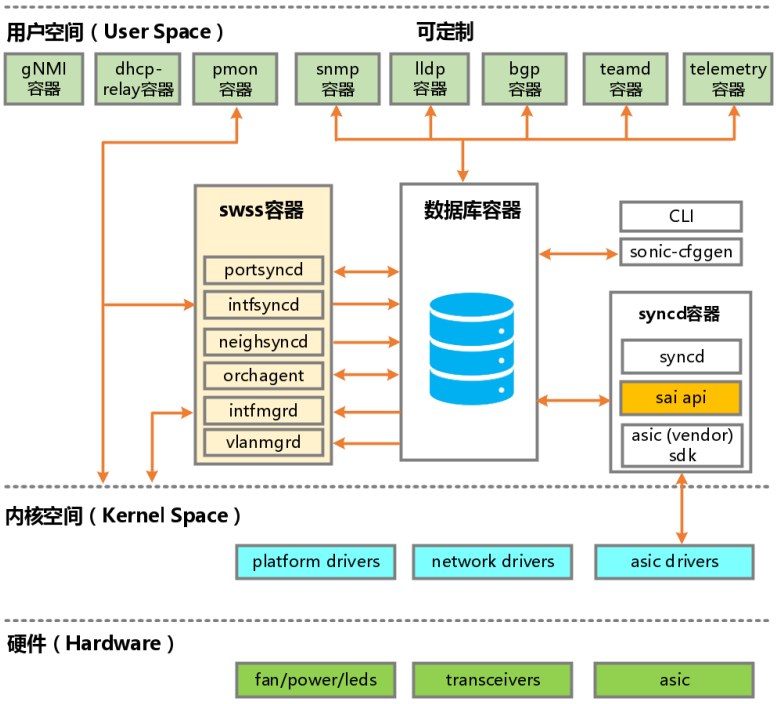


图 3.4 SONiC 系统架构

在 Scale-out 网络中，SONiC 的标准化接口和容器化架构契合 AIDC 大规模集群构建需求，能够应对多节点间的高速互连与动态配置，支持 RDMA 等高吞吐场景，且无需更换软件即可实现硬件节点的线性扩展。随着商业生态日益成熟，SONiC 有望成为 AIDC 的主流网络操作系统，地位类似于服务器领域的 Linux。

3.2.3 Scale-Out 协议与软件栈

3.2.3.1 通信协议

AIDC Scale-out 网络主要采用 InfiniBand、RoCEv2（RDMA over Converged Ethernet v2）、超以太网（Ultra Ethernet，UE）等通信协议，此外还包含多样的应用协议。

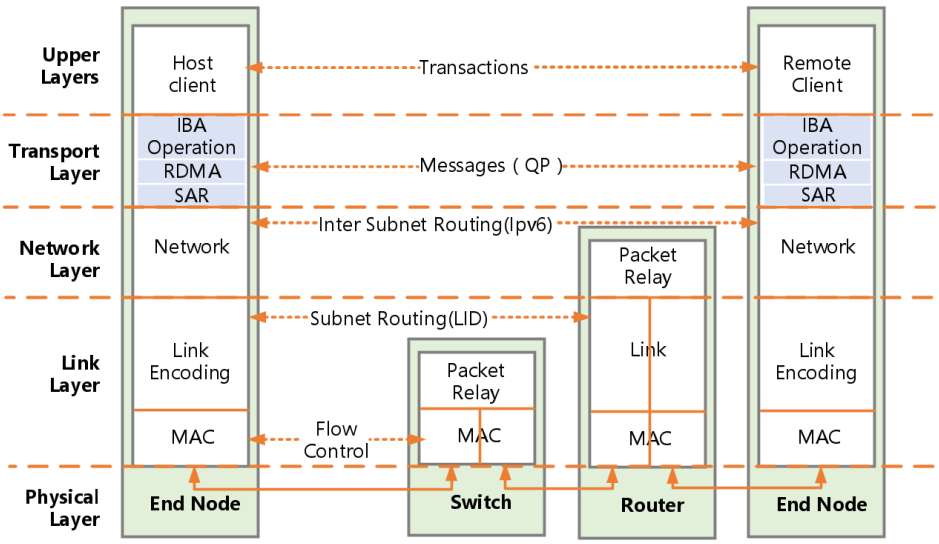


图 3.5 InfiniBand 协议栈

1. InfiniBand RDMA

InfiniBand RDMA 是一种专为高性能计算和 AI 集群设计的网络互连协议，允许一台计算机的网卡直接读写另一台计算机的内存，绕过操作系统内核和 CPU 协议处理，从而减少延迟。InfiniBand 链路层采用基于信用的流量控制协议来实现无损传输，并在单个虚拟通道路径内保证包序。RDMA 操作由 InfiniBand Verbs 接口控制 RDMA 网卡硬件执行，提供低延迟、高带宽的数据传输。此外，协议栈的大部分关键数据传输操作卸载到硬件上进行加速。目前，尽管面临 RoCEv2 和 UEC 等以太网方案的竞争，但在对性能有极致要求的 AI 场景中，InfiniBand RDMA 仍保持着优势。

2. RoCEv2

在智算中心 Scale-out 网络中，RoCEv2 目前是以太网方案的主要标准。RoCEv2 是 RDMA 技术的一种实现，运行在以太网之上，并支持跨 Layer 3 的 IP 路由，传输层采用 UDP 协议，应用层支持集合通讯库和 RDMA API。RoCEv2 相对于 RoCEv1 的最大优势在于可跨子网路由，使其在 Scale-out 环境中能够实现更大范围的互联扩展。

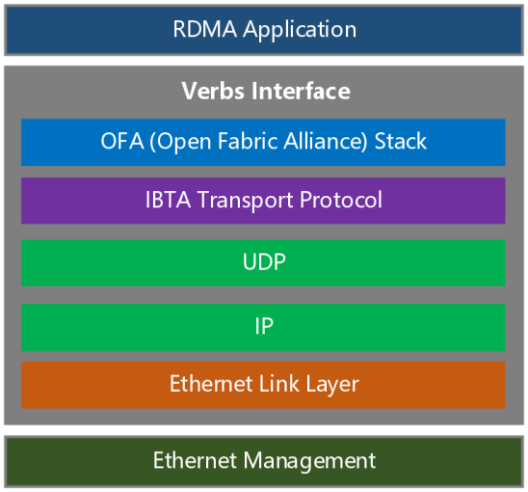


图 3.6 RoCEv2 协议栈

RoCEv2 基于 UDP 无连接传输协议，缺乏类似 TCP 的可靠性与拥塞控制能力，因此通常需要结合 PFC、ECN 以及 DCQCN 等协议，以保障大规模 XPU 集群中的可靠通信与高性能传输。RoCEv2 契合了数据中心基于 IP/Ethernet 的现有协议栈，避免了部署专用 InfiniBand 网络带来的额外成本和复杂性。

3. UE

超以太网（UE）定义了一套基于以太网的标准和规范，使其满足 HPC 和 AI 应用高带宽和低延迟通信需求。特别地，UE 是 OCP Open Systems for AI 推荐的 Scale-out 网络。

UE 软件层定义了 API 和规范，其核心是支持 libfabric(v2.0）接口。

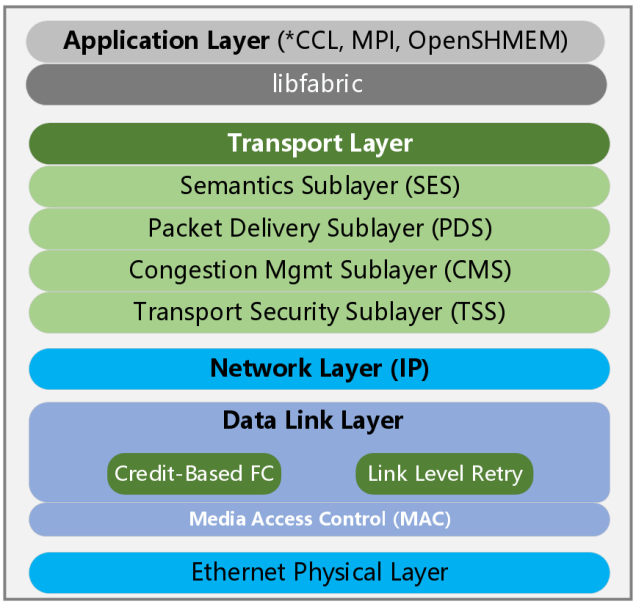


图 3.7 UE 协议栈

如上图所示，传输层是 UE 的创新核心，支持 RDMA，旨在实现百万级端点的超大规模扩展，由语义子层、数据包传递子层、拥塞管理子层等组成。

- 语义子层(SES）将 libfabric API 映射为网络操作（如 RDMA 读/写），支持延迟发送和消息汇合协议，以优化 AI 集合通信库的卸载效率；
- 数据包传递子层(PDS）负责可靠性与排序，通过短暂数据包传递上下文(PDC）实现连接管理，支持零 RTT 上下文构建，即第一个到达的数据包即可建立可靠性上下文；
- 拥塞管理子层(CMS）提供端到端拥塞控制，支持发端和收端拥塞控制。CMS 的关键特性是数据包喷洒，通过改变熵值(EV）将单条流的数据包分布到所有可用路径，消除网络热点。

UE 网络层利用标准的 IPv4 和 IPv6 协议，以确保与现有数据中心路由架构和管理工具的兼容性。数据包修剪(Packet Trimming，PT）是网络层一个关键的可选特性。当交换机缓冲区发生拥塞时，PT 会剪除数据包的负载并仅转发包头，作为一种极快的显式丢包信号，使接收端能立即感知并请求重传，从而大幅降低尾部延迟。

UE 链路层协议保持与 IEEE 802.3 标准的兼容性，同时引入了增强性能的可选功能，包括链路层重试(LLR）协议和基于信用的流控(CBFC）协议。链路层发现协议(LLDP）用于与通信对象协商以启用 LLR 和 CBFC。InfiniBand、RoCEv2 和 UE 的对比如下表。

表 3.1 InfiniBand、RoCEv2 和 UE 的对比

	InfiniBand	RoCEv2	UE
当前阶段	成熟期/高端市场垄断	成熟期/大规模部署	导入期/标准制定
核心优势	极致性能、全栈优化	开放性、成本、生态，迭代速度快	开放架构、面向未来的设计
核心劣势	封闭、昂贵，迭代速度慢	生态开放带来的运维复杂度	尚未大规模落地，存在不确定性
未来发展	市场可能逐渐被 RoCE 占据	长时间事实上的市场主流	重要的未来投资与技术押注

4. 应用层通信协议

应用层协议的核心作用是抽象底层 RDMA 或网络硬件，实现集合通信操作（All Reduce 等），提供高性能、可扩展的通信接口。例如，NCCL 用于自动拓扑感知，高效实现集合通

信功能。此外，还可以通过 RDMA API 直接操作 RDMA 网络，自定义库、存储、网络优化功能。

3.2.3.2 网络拥塞控制

1.自适应路由

自适应路由是一种由交换机和网卡紧密协作实现的端到端细粒度负载均衡协议，旨在解决 AI 训练中常见的大象流冲突和网络拥塞等问题。Scale-out 自适应路由协议支持逐包级的多路径传输，根据出口队列可用情况，同时向可用端口转发数据包，实现数据包喷洒和收端排序，减少链路拥塞，大幅提升网络的有效带宽利用率。

2.闭环控制

对于由多个发送端同时向单个接收端发送数据引起的多对一（Incast）拥塞，Scale-out 网络会配合使用拥塞控制机制如 PFC、ECN、DCQCN、包裁剪等，通过反馈信号控制数据注入速率，从而实现 Incast 场景的拥塞闭环控制。

3.网络遥测

AIDC 的 Scale-out 路由和控制协议依赖对网络状态的测量。交换机通过遥测实时检测拥塞热点，并将信息反馈给网卡。网卡以微秒级延迟执行拥塞控制，对引发拥塞的流量源进行精细化调节，从而避免网络拥塞。此外，实时遥测还确保了多租户环境下的性能隔离。

3.2.4 性能评估与调优

3.2.4.1 Scale-Out 网络性能指标

Scale-out 网络的性能指标包括：1）RDMA bisection 场景下的带宽和时延；2）大规模 All-Reduce 和 All-to-All 带宽；3）有效带宽利用率；4）分布式训练作业的扩展效率，即当 XPU 卡数从 N 扩展到 2N 时训练作业完成时间的提升比例；5）网络韧性，即部分链路断开时的网络性能；6）多租户网络虚拟化前后的性能表现。

3.2.4.2 性能调优技术

Scale-out 网络性能调优技术包括流量调节、拓扑感知路由、端到端协同、RDMA 故障恢复等方向。流量调优针对不同流量模式动态调整 ECN 阈值、PFC 门限等；拓扑感知路由采用基于作业通信矩阵和拓扑信息的路由机制（如多路流量工程），改进负载均衡协议；端到端协同机制通过软件通信库、网卡驱动、交换机 OS 等进行全栈联合调优；RDMA 故障恢复协议提供故障检测和恢复能力，通过备份多路径减少故障对计算任务的影响时间。

3.2.5 业界前沿案例与趋势

3.2.5.1 AIDC Scale-Out 网络实践

1.基于 InfiniBand 的 Scale-out 网络

NVIDIA DGX SuperPOD 利用 InfiniBand 构建大规模的分布式 AI 集群，并通过模块化设计实现 Scale-out 网络，采用自适应路由、多路径流量分担机制，以及基于硬件的网络自愈和错误检测能力，以保障高性能 RDMA 通信。

2.基于 RoCEv2 的 Scale-out 网络

Azure 和 Meta 采用 RoCEv2+定制以太网，实现接近 InfiniBand 性能，但更开放、更低

成本的大规模 AI Scale-out 网络。

3.基于 UEC 的 Scale-out 网络

博通 Scale-out 网络方案基于 UEC，用增强 RDMA、多路径、可编程拥塞控制，把以太网升级为面向百万节点 AI 集群的高性能互联系统。

4.异构互联与定制化方案

Google OCS Scale-out 网络本质是用"可重构光路+SDN 控制"替代传统多级电交换网络，实现 AI 集群的低延迟、高带宽和低成本互联。OCS 属于纯物理层交换，不对数据包进行协议解析或电信号处理，因而 OCS 对封装协议（如 RoCE）透明。

5.世纪互联 Scale-out 案例

世纪互联 AIDC 采用两层无阻塞 Spine-Leaf 架构，打造 1:1 无收敛比的 Scale-out 网络，整网基于 RoCEv2 无损以太网技术，在乌兰察布、张家口、长三角、粤港澳等国家算力枢纽和基地落地千卡商用集群，可平滑扩容至万卡级规模，端到端通信时延稳定控制在 5μs 以内，网络峰值利用率达 90%以上。

6. 紫金山实验室 Scale-out 案例

紫金山实验室建设超宽无损智算中心网络，基于自研 51.2T 以太网白盒交换机，搭载 UniNOS 网络设备操作系统，采用两层 Spine-Leaf 架构，构建端到端 400G Scale-out 超宽 RoCEv2 网络，部署 PFC、ECN 实现无损传输，同时采用自适应路由与逐包负载分担技术，链路带宽利用率不小于 95%。

3.2.5.2 未来趋势展望

InfiniBand 和 RoCEv2 目前仍然是典型的高性能 Scale-out 网络协议，而 UE、MRC（Multipath Reliable Connection）等协议快速迭代，可能成为未来的主流。同时，部署在 OCS 上的 RDMA 相关协议是 AIDC Scale-out 网络重要的演进方向。此外，Scale-up 与 Scale-out 协议存在协同演进的趋势，例如 ESUN 与 UEC 开展协作，以对齐开放标准。

3.3 Scale-Across 网络技术

3.3.1 智算中心互联需求

3.3.1.1 算力扩展

Scale-across 网络将 AI fabrics 扩展到单个 AIDC 之外，连接距离可以跨越数十千米。通过 Scale-across 网络，运营商可以互联多个 AI 数据中心，形成跨越多个地点的超大规模 AI 基础设施。

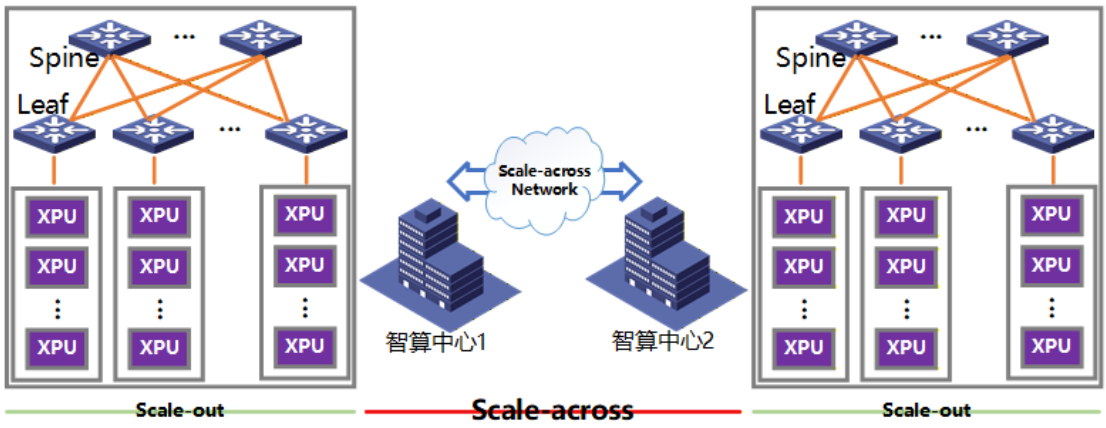


图 3.8 Scale-across 网络示意图

Scale-across 算力扩展是指通过 Scale-across 网络将分布在不同地理位置的多个数据中心的算力连接起来，使其在逻辑上表现为一个统一的、可调度的超大规模计算 Fabric。其需求主要源于单体数据中心在承载万亿参数级大模型训练时面临的物理和工程极限。此外，随着 AI 业务重点转向推理，算力需要分布在离用户更近的都市圈；通过 Scale-across 技术，可以将多个都市圈内的推理节点组织成统一资源池，实现按距离优先调度和动态负载分配。

3.3.1.2 物理限制

除算力扩展需求外，通过 Scale-across 网络连接多个数据中心的需求还来自单个数据中心的物理限制，包括电力、散热和占地限制。

1. 电力供应限制

单机柜功率正从当前 ~150kW 向 200–500kW（乃至更高）演进（详见 1.3.2.1），单个 AIDC 难以支撑数百万颗 XPU 所需的巨大电力和基础设施消耗。

2. 散热挑战

传统风冷已无法满足超高功率机柜的散热需求，需要采用液冷系统，这在物理空间和工程实施上对单体机房提出了更高要求。

3. 占地面积限制

单栋建筑或园区的物理空间有限，难以持续扩展以容纳百万级 XPU 集群。

3.3.2 无损 RDMA

3.3.2.1 距离适应负载均衡

与单数据中心网络相比，Scale-across 网络面临的核心挑战在于链路距离差异显著，导致 RDMA 不同路径存在明显延迟差异，同时负载容易集中在部分链路上，传统静态路由难以充分利用整个网络资源，因此需要高效的动态负载均衡机制来提升性能和稳定性。

距离适应负载均衡（Distance-Adaptive Load Balancing，DALB）协议通过路径权重动态计算和基于距离及负载的自适应流量分配，可显著提升超大规模 Scale-across 网络的性能和可靠性。DALB 协议根据路径距离和当前网络负载自适应地分配流量，并通过动态监控网络状态、定期调整权重，使热点链路得到均衡分配，多路径资源被充分利用，从而增强了 InfiniBand RDMA、RoCEv2 和 UE RDMA 协议在 Scale-across 网络中的稳定性。当某条

链路异常时，DALB 协议能够自动降低其权重或将流量切换到其他路径，保证通信连续性和整体网络的可靠性。

3.3.2.2 端到端拥塞控制

Scale-across 网络涉及跨机架、跨 Pod 或跨数据中心的通信，网络拓扑复杂且链路延迟和带宽差异显著。端到端拥塞控制协议（end-to-end congestion control, E2E CC）通过在通信双方之间监控网络拥塞状况，自适应地调整发送速率，从而避免链路过载。协议通常依赖显式拥塞通知（ECN）机制：当网络设备检测到队列长度过高或潜在拥塞时，通过 ECN 标记数据包，发送端据此降低发送速率，防止丢包和延迟累积。

在大规模、异构和多路径的网络环境中，端到端拥塞控制协议可以结合多路径流量工程（MPTE）协议使用，将流量分散在多个链路上，并根据各路径的拥塞状态动态调整流量分配比例，使低拥塞路径承载更多流量。通过端到端拥塞控制与 ECN 通知结合多路径流量调度，Scale-across 网络能够提前感知和响应拥塞，实现 RDMA 无损传输。

3.3.3 广域确定性互联

3.3.3.1 精确的延迟管理

广域 Scale-across 网络的延迟管理通过网络资源预留、多路径流量分配、负载均衡权重优化、快速拥塞通知、弹性带宽分配等机制实现，依赖于计算与网络调度的紧密协作。网络调度器会提前为每条通信路径预留带宽资源，同时结合多路径流量工程（MPTE）和备份路径策略，实现对流量的精细控制。当某条链路即将出现拥塞时，网络节点能够识别受影响的链路并通知发送节点，发送节点据此动态调节发送速率，抑制延迟波动，同时维持高吞吐率和端到端无损性能。此外，弹性带宽（Elastic Bandwidth）和多租户调度策略可使网络资源随计算任务需求动态增长或缩减。

3.3.3.2 网络状态感知

在 Scale-across 网络中，网络状态感知（如带内遥测）是超大规模分布式算力系统的核心性能保障机制。协议将 Scale-across 链路的物理距离及由此产生的传播延迟纳入流控模型，使每条链路的数据注入速率能够根据其传播特性进行优化，从而最小化尾部延迟；并能自动识别网络拓扑结构，在多条可行路径之间智能分配流量。带内遥测为网络提供亚微秒级的实时监控能力，能将拥塞信号直接反馈至发送端，使网卡快速完成速率调整，并提供端到端的健康监控和故障排除能力。

3.3.4 业界前沿方案与趋势

3.3.4.1 长距跨智算中心的 AI 训练方案

如下图所示，业界将 XPU 部署在多个机房，并通过 DCI 设备将其互联，形成 Scale-across 网络，用于跨 DC 协同训练。

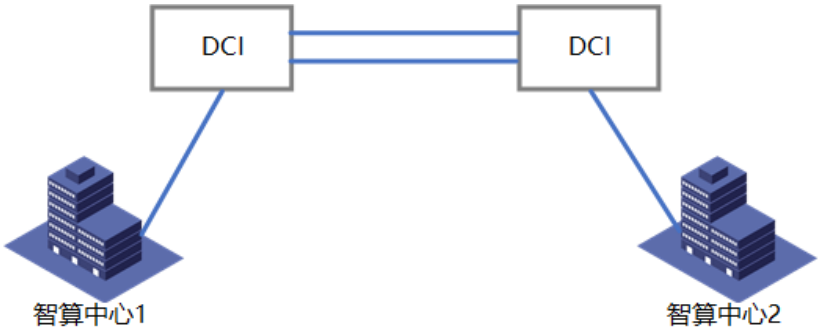


图 3.9 智算中心远距互联示意图

北美已经完成长距跨 DC 协同训练部署，受技术限制，目前距离均在 100km 以内。具体地，100km 长距跨机房训练场景面临的挑战与解决方案如下表所示。

表 3.2 长距跨机房训练面临的挑战与解决方案

100km 场景技术挑战	描述	解决方案
长距拥塞通告效率问题	近端智算中心内设备拥塞 ECN 标记，ECN 标记-CNP 拥塞通告-发送端降速总计需要经过 1ms	近端 DCI 设备支持 FastCNP，代理回复 CNP，us 级实现发送端降速
长距拥塞代理的缓存积压问题	远端智算中心内设备拥塞，CNP 拥塞通告-发送端降速总计需要 1ms，此时远端触发 PFC，远端 DCI 需要等待 1ms 才能释放缓存，单端口需要 50MB 缓存开销，易溢出丢包	DCI 设备使用带 HBM 大缓存款型，比如博通的 StrataDNX（Dune）系列芯片，或者路由器 NP 芯片

3.3.4.2 Scale-Across 网络实践

1.Nvidia Scale-across 案例

Nvidia 推出了专门针对 Scale-across 场景优化的 Spectrum-XGS 以太网技术，使地理分布的设施能够像一个单一的“Giga-Scale”AI 工厂一样运作。在相距 10 公里的数据中心之间进行 NCCL All Reduce 集合通信测试中，Spectrum-XGS 的性能比传统以太网高出 1.9 倍。

Nvidia Scale-across 方案支持距离自适应拥塞控制协议，引入了能感知站点间距离的算法，动态优化流控以最小化长距离传输带来的尾部延迟。方案采用的拓扑感知负载均衡，自动调节 Scale-across fabrics 间的流量分布，消除网络热点并最大化吞吐量。方案通过基于全局遥测的控制逻辑，确保了确定性的时延和抖动服务质量。

2. 博通 Scale-across 案例

博通在 Scale-across 网络领域提出了基于开放以太网标准的方案，旨在通过城市或区域级的长距离互联，将分散的 AI 集群整合为支持百万级 XPU 节点的超大规模算力池。

博通方案结合了交换机端的丢包拥塞通知(DCN）、优先级流控(PFC）以及显式拥塞通知(ECN），配合端到端的流量可见性，解决 Scale-across 链路带宽通常小于集群内带宽带来的超配（Oversubscription）挑战。方案在 Scale-across 层级使用深缓存技术来吸收长距离传输带来的延迟和微突发拥塞，利用 GB 级的缓存能力来确保 RDMA 流量在长距传输中维持无损性能。

3. 世纪互联 Scale-across 案例

世纪互联在北京、苏州等城市范围内建设类似 Core-Spine-Leaf 的三级 CLOS 网络架构的 AIDC 集群：Hub-Spoke-Edge 拓扑，Hub 是核心枢纽型 AIDC，部署在城市近郊电力充足、绿电可及的区域，承载万卡级 GPU 训推集群；Spoke 是中型辐射型 AIDC，是城域算力的中继枢纽，均匀部署在城市各行政区核心位置；Edge 是边缘 AIDC，下沉至城市核心商圈、产业园区，贴近用户与数据源部署，提供亚毫秒级低时延的边缘智能服务。底层通过全光交换网络重构城域算力总线，实现 Hub-Spoke-Edge 全节点一跳直达的全光互连，光层端到端传输时延稳定控制在 0.5 毫秒以内，构建城市范围内的无阻塞 AI Fabric。

世纪互联提出了超互联网络协议 UCC（Universal Cross Connection），基于标准以太网和 RDMA 技术，对网络协议和传输协议进行深度优化与轻量化，并内置了基于公私钥体系的内生网络安全机制，实现城域内连接的可信、可管和可控，保障高速与安全的跨节点算力调度。

4.紫金山实验室 Scale-across 案例

紫金山实验室基于 CENI 网络构建多算力中心跨广域互联 RDMA 网络，在算力中心出口处部署 RDMA 网关设备，网关采用紫金山实验室自研 12.8T 可编程白盒交换机，搭载 UniNOS 网络设备操作系统，采用 400G 网络端口进行网络互联，在网关设备上部署快速拥塞控制技术 FastCNP，实现拥塞控制快速反馈，解决广域长距离传输时延带来的端侧 RDMA 网卡流量控制问题；同时在网关上部署大象流负载均衡技术，基于五元组+QPID 实现 RDMA 业务流量无损拆解，通过 SRv6 多路径策略动态编排，达到业务流精准分割与最优路径匹配，提升广域链路利用率；CENI 网络开通 400G 互联，提供确定性网络传输。最终实现在跨 100 公里的长距传输中，网络吞吐率保持在 90%以上。

3.3.4.3 未来趋势展望

AIDC 的 Scale-across 网络目前正处于方兴未艾的发展阶段，它不仅是 Scale-out 的逻辑延伸，也是突破单体数据中心物理极限、实现万亿参数大模型和多智能体持续演进的关键路径。

随着 Ultra Ethernet Consortium（UEC）等组织的推动，Scale-across 网络协议将更可能基于以太网标准构建。同时，开放生态也是 Scale-across 协议演进的一个重要方向。

第四章 统一运维管理

受限于当前阶段的资源与时间，当前版本的这份报告在运维管理层主要覆盖以 BMC/Redfish 为底座的超节点级管理（机柜级、节点级），对园区级 DCIM、动环监控与集群级智能自治运维（AIOps）等数据中心层面的内容尚未充分展开，有待后续版本补充。我们深知这一层面对 吉瓦级智算中心的重要性，**诚挚欢迎数据中心运维、DCIM、动环、智能运维等领域的专家与厂商加入工作组，与我们共建、共同完善本章**。

4.1 AIDC 运维现状与分析

当前 AI 超节点单机柜最大功率密度约 150kW、并向 200-500kW 演进，液冷、RoCE 无损网络、大规模服务器集群成为标配，业务以大模型训练、推理、算力租赁为主，对设备在线率、故障快速处置、远程管控要求远高于传统 IDC，AIDC 运维正处于从传统 IDC 托管向高密度、高可靠、智能化运维转型的关键期。

1) 运维模式：逐步从人工现场巡检，转向**远程集中运维 + 自动化运维 + AI 预测运维**结合，现场值守仅做硬件更换、应急抢修。

2) 核心诉求：算力集群不能中断，单台服务器 / 卡故障需分钟级定位、隔离、恢复；海量服务器（万级 / 十万级）需统一纳管。
3) 运维分层：行业普遍分为**带内运维**（操作系统、业务、集群软件）与**带外运维（BMC）**（硬件底层、上电、BIOS、健康状态）两层，二者互为补充，BMC 已成为 AIDC 运维底座。

4.2 整机柜超节点管理架构

超节点系统管理软件具备清晰的分层管控与一体化集成特征，技术架构自上而下涵盖机柜级管控、节点级管理、网络管理、集中运维控制四个层级。

1) 机柜级管理将管控范围扩展至机柜级微环境，实现功耗管理、分布式 CDU 散热调控、温湿度等环境传感器统一监控。
2) 节点级带外管理依托主板管理控制器，基于 IPMI、Redfish 等标准协议，实现对服务器硬件状态的深度监控与管控。
3) 网络管理器面向高性能互联网络，提供拓扑自动发现、健康巡检、可视化运维能力，保障数据中心网络稳定高效运行。
4) 最终通过集中运维控制台实现全域管理能力汇聚，为运维人员提供全局视图，支持跨域策略统一下发与自动化运维。

超节点系统管理软件采用统一建模与遥测机制，依托 Redfish 等协议标准构建统一硬件模型，将超节点内 CPU、加速卡、内存、网卡及 PCIe/CXL 设备等所有部件抽象为标准化数据模型，并基于该模型开展细粒度遥测数据采集。

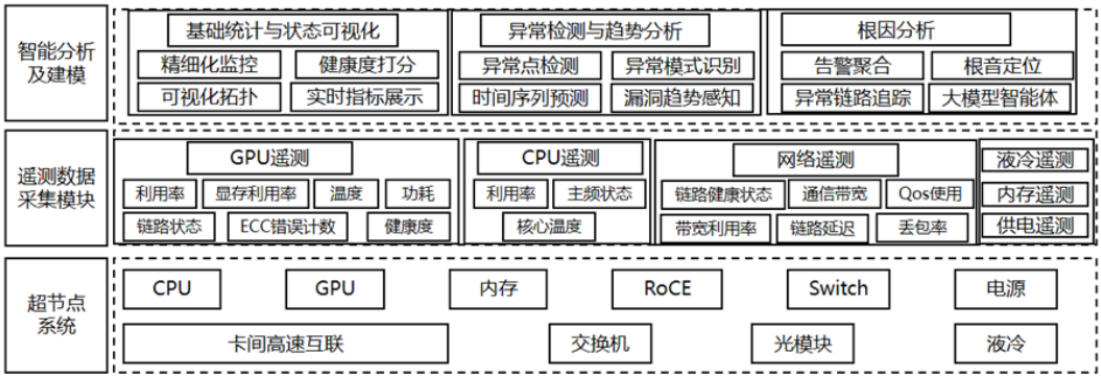


图 4.1 整机柜超节点管理功能框架

4.3 超节点机柜级管理

整机柜通过 PMC（Power Management Controller）管理模块对 Power Shelf 管理、PSU（Power Supply Unit）管理和 CDU（Coolant Distribution Unit）系统管理。

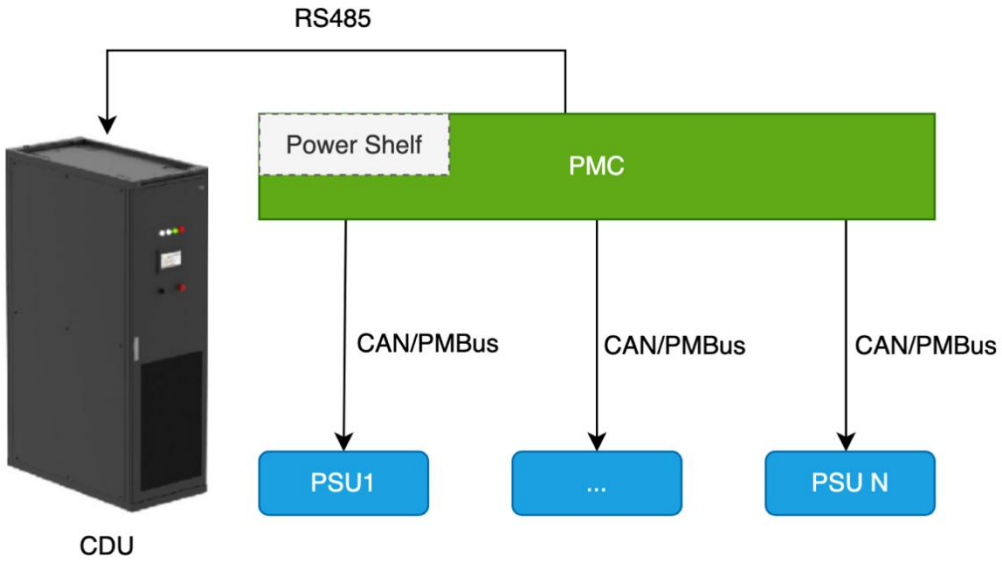


图 4.2 机柜级管理框图

4.3.1 电源系统管理

Power Shelf 需满足核心电压、电流、输入功耗、输出功耗、环境温度、Shelf output busbar clip 温度等传感器监控，以及 Power Shelf 内总体 PSU 数量监控，支持 Power Shelf 健康状态日志告警功能。

针对 Power Shelf 内每个 PSU，PSU 资产支持查询功能，包括 Firmware、生产厂商、序列号、FRU 等，PSU 健康状态监控需支持电压、电流、温度、功耗监控和日志告警，支持 PSU 黑盒日志下载，支持业务无感升级 PSU 固件。

4.3.2 液冷系统管理

整机柜超节点漏液检测源需覆盖整机柜 Manifold 进液和回液快接头、法兰和节点核心部件（CPU/GPU 等），支持漏液监控和日志告警功能。

针对分布式 CDU 独立给整机柜供液场景，需支持 CDU 资产获取，包括厂商、产品名、序列号、固件版本、液位、环温、液体温度、风机、电源管理，支持漏液、温度、风机、电源等核心组件故障监控和日志告警，日志掉电不丢失。CDU 管理需支持远程和本地控制 CDU 运行参数，支持远程升级 CDU 固件，对外接口协议满足 Modbus RTU/TCP。

4.3.3 机柜管理协议

机柜 PMC（Power Management Controller）通过 Redfish、IPMI 和 SMASH CLP 协议对接上层运维管理平台，对机柜健康状态和属性进行监控。

4.4 超节点节点级管理

节点级 BMC 通过对计算 Tray、Switch Tray、电源与散热子系统进行统一编排，实现超节点基础设施的可观测、可控制。

节点级管理主要包括：

核心部件状态监控与生命周期管理

PCIe Switch 管理与拓扑配置

电源与散热控制

图 4.3 节点级管理框图

4.4.1 核心部件管理

1） 主板管理

BMC 负责主板基础资源的统一监控与管理，系统通过 PMBus、I2C、GPIO、ADC 等接口持续采集板级遥测数据，并结合事件驱动机制实现异常状态的快速检测与上报。主要包括：

a） 电压、电流、VR 及功耗监控

b） 上电时序检测与告警

c） FRU 管理，支持主要字段修改

d） 固件生命周期管理，支持 BMC、BIOS、CPLD、PSU、GPU FW 等固件升级和版本管理，其中，BMC 升级对业务运行无影响。

e） BIOS 配置管理，支持带外 BMC 修改 BIOS Setup 选项

2） CPU 管理

CPU 管理主要涵盖 CPU 生命周期管理、运行状态监测以及故障检测等能力，包括：

a） CPU 温度与功耗监控

b） CPU 资产信息管理

c） CPU 在位、故障检测与告警

3） GPU 管理

GPU 是 AIDC 超节点中的核心计算资源，节点级 BMC 需要具备 GPU 识别、状态监控与故障管理能力。

GPU 管理能力包括：

a) GPU 资产管理

b) GPU 温度与功耗监控

c) GPU 在位检测、利用率及显存利用率监控、显存可恢复性错误、不可恢复性错误和 ECC Mode 监控

4) 智能网卡管理

随着 AI 数据中心逐步向智能网络演进，智能网卡已成为超节点中的关键组件。BMC 可通过 MCTP、NC-SI、PCIe VDM 等方式实现对 DPU 的统一管理。

节点级管理系统需要支持：

a) 智能网卡资产管理

b) 智能网卡温度与功耗监控

c) 智能网卡在位检测、端口连接状态、协商速率等信息监控

5) 内存管理

AI 训练场景对内存稳定性要求极高，BMC 需要持续监测：

a) 内存温度监控

b) 内存资产信息监控，包括厂商，序列号，内存容量，位宽，最大频率，当前频率等

c) 内存在位检测、可恢复性与不可恢复性错误监控

6) PCIe Switch 管理

在 AI 超节点场景中，PCIe Fabric 的稳定性直接影响 GPU 间通信性能及整体计算效率。BMC 需持续监测 PCIe 链路健康状态，并对链路降速、链路抖动、错误恢复及异常中断等事件进行实时检测与告警，保障 PCIe Fabric 的高可用性与稳定运行。

PCIe Switch 管理主要包括：

a) Link Width/Speed 状态监控

b) 端口流量检测

c) AER 错误采集

d) Switch Firmware 管理

此外，系统应支持 PCIe 拓扑的动态重建与状态同步能力。在 GPU Reset、PCIe Recovery、链路异常或拓扑变化等场景下，能够自动重新发现并更新拓扑信息，确保 BMC 侧拓扑视图与实际硬件状态保持一致。

PCIe Switch 拓扑管理主要包括以下设备类型：

a) PCIe Switch

b) GPU Endpoint

c) RNIC Endpoint

d) PHY Retimer

e) NVME

图 4.4 PCIe SW 拓扑管理

4.4.2 电源管理

节点 BMC 可对主机服务器执行电源控制，其包括开机、强制关机、软关机、复位重启、强制重启，针对智能卡配置需满足智能卡上下电控制，其具体描述如下。

1） 开机：等同于服务器关机状态下短按 Power Button 使服务器开机，系统各组件按设计时序依次上电开机。

2） 软关机：等同于开机情况下短按 Power Button，需要在 OS 下做正确的配置，将 Power Button 短按动作设置为直接关机此功能才能生效，某些 OS 下在收到软关机信号后会弹出确认提示，用户在 OS 中确认后才能真正关机。

3） 强制关机：等同于长按（5s 以上）Power Button 强制服务器关机，CPU/GPU 处理器同步下电。

4） 复位重启：系统立即重新启动，模拟短按 Power Reset Button。

5） 强制重启：先关机再开机，等同于长按 Power Button 关机，然后等待 10s 以后，短按 Power Button 开机。

6） 智能卡上下电：智能卡承载裸金属和虚拟机业务，当 BMC 侦测到智能卡在位，常规电源操作要求智能卡不下电，当需要对智能卡进行上下电时，需要提供额外指令对智能卡进行上下电，以保障智能卡新增特性生效。

4.4.3 散热管理

节点部件散热除液冷散热部分，其余依赖风扇散热的部件，BMC 通过调整风扇转速满足其正常工作温度，需支持手动和自动 2 种调速模式：

a）手动控制风扇：设置风扇控制模式为手动，可自定义调节风扇转速

b）自动控制风扇：设置风扇控制模式为自动，进行智能自动调速

针对散热异常处理，需覆盖以下场景：

a）BMC 挂死，要求从异常场景拉升至安全转速；

b) BMC 重启或 IPMI 服务异常过程中，要求保持安全转速，直到 BMC 启动完成，散热模块正常工作为止；

c) 固件更新过程中需要确保无散热风险（升级过程中，若散热模块不能正常进行调速，则需要在文件上传开始，拉升至安全转速）；

d) 散热模块异常，在恢复正常调速前，保持安全转速；

核心组件（CPU、GPU、智能网卡等）温度连续 3 次以上温度读取异常(0 值或无法读值)，需要立刻进行异常调速，调整到安全转速。

4.4.4 日志诊断

BMC 需具备一键日志收集能力，自动完成多源日志采集、打包、上传或导出，便于问题快速定位和运维分析。

日志采集内容应包括：

a) 系统事件日志（SEL，System Event Log）
记录主板、CPU、GPU、内存等硬件传感器故障事件。

a) 审计/操作日志
记录对设备进行操作的、非查询类日志。

a) HOST 日志
记录 BIOS、OS 启动和运行 SOL 日志。

a) BMC 日志
记录 BMC 运行日志，包括 dmesg、trace log、内部模块异常 log 等。

a) 设备综合诊断数据
记录系统内部件资产信息、固件版本号、历史传感器数据、BIOS 选项数据、散热调速日志、CPU 关键状态寄存器信息、宕机最后一屏等。

4.4.5 告警设置

BMC 需具备对告警日志进行配置，包括 syslog 告警设置、SNMP 设置、邮箱告警设置，具体要求如下：

1）SNMP 设置

通过 SNMP Trap 将告警日志转发到外部，需支持配置参数主机标识、Trap 版本、告警级别、目标设备 ip 等。

SNMP Get/Set 用于读取设备日志，需支持配置 SNMP V1/V2C/V3.团体名等。

2）Syslog 日志告警设置

通过 syslog 远程转发将告警日志转发到外部，需支持配置目标设备 IP、告警级别、传输协议等。

3）邮箱告警设置

通过邮件将告警日志转发到外部，需支持 SMTP 服务器地址、端口、告警级别等。

4.4.6 远程访问

BMC 具备远程访问和操作服务器功能，包括图形访问、命令行访问、远程介质，无需本地显示器。

1）图形访问

可通过 HTML5 KVM 和 VNC 远程查看、操作主机屏幕，VNC 支持密码修改。

2）命令行访问

通过将服务器的串行端口重定向到 SOL 控制台，进行输入/输出操作。

3）虚拟媒体

支持远程挂载镜像 ISO 文件、CD/DVD、HDD、文件夹挂载镜像，用于主机服务器远程媒体使用。默认关闭虚拟媒体，要求进入到 OS 下自动关闭虚拟 USB 网卡设备，支持 IPMI/Redfish 指令或 web 打开虚拟媒体。

4.4.7 用户与安全

BMC 是 AI 基础设施的重要管理入口。节点级管理系统需要构建完善的身份认证、权限控制、网络访问防护以及证书安全体系，以保障超节点平台的安全性与可靠性。

1）用户与权限管理

节点 BMC 支持多级用户管理与基于角色的访问控制（RBAC，Role-Based Access Control），可针对不同用户角色分配差异化管理权限，可有效降低误操作风险，实现系统资源的安全隔离与精细化访问控制。

节点级管理系统支持典型的三级权限模型，具体权限划分如下：

功能模块	管理员	操作员	普通用户
用户配置	✓	×	×
配置自身账户	✓	✓	✓
常规配置	✓	✓	×
电源控制	✓	✓	×
远程媒体	✓	✓	×
远程 KVM	✓	✓	×
安全配置	✓	×	×
调试诊断	✓	×	×
查询功能	✓	✓	✓

2）密码安全策略

为提升系统安全性，BMC 支持密码复杂度策略配置，并默认启用复杂密码要求。

密码策略支持以下能力：

最小密码长度限制

密码复杂度校验

密码有效期管理

历史密码重复检测

登录失败锁定机制

3）系统防火墙

节点 BMC 支持内置系统防火墙能力，用于限制非法访问与网络攻击风险，管理员可根据数据中心网络规划限制特定来源访问管理接口，提升管理网络安全性。

系统支持以下类型的防火墙规则：

IP 地址防火墙规则，支持基于 IPv4/IPv6 黑/白名单地址进行访问控制。

MAC 地址防火墙规则，支持 MAC 地址黑/白名单进行访问控制。

端口防火墙规则，支持针对管理服务端口进行访问限制。

4）SSL 证书管理

为保障管理接口通信安全，节点级 BMC 支持 SSL/TLS 证书管理能力，用于实现 HTTPS 等安全通信。

系统支持通过 Web 界面生成 CSR（Certificate Signing Request）文件，并支持证书导入与更新。

有效保护管理数据传输安全，防止敏感信息泄露与中间人攻击。

4.4.8 BMC 网络管理

节点 BMC 支持灵活的网络配置与管理能力，满足数据中心不同网络部署场景下的带外管理需求，支持 IPv4 与 IPv6 双栈网络协议，并支持静态 IP 与 DHCP 模式的动态切换。

为提升节点级管理网络的可靠性与可用性，BMC 网络服务具备自动恢复能力（Self-Recovery Mechanism），通过网络自恢复机制，节点级 BMC 能够有效提升管理网络的稳定性与连续运行能力，降低网络异常对远程运维与集群管理的影响。网络恢复方案如下：

1）当系统检测到 IPv4 或 IPv6 DHCP 地址异常丢失时，网络服务将自动周期性发送 DHCP/DHCPv6 请求，以重新获取有效网络地址，确保带外管理网络能够及时恢复。

2）当 IPv4/IPv6 网络服务进程异常退出、崩溃或丢失时，系统可自动执行服务重启与网络恢复操作。

4.5 发展趋势展望

未来，BMC 将从传统的带外管理模块，逐步演进为集标准化接口、硬件可观测性、自动化运维执行、安全可信控制与场景协同于一体的基础设施控制节点。在 AIDC、液冷、GPU 集群等新型算力场景驱动下，BMC 的能力边界将持续外扩，成为服务器、机柜乃至集群级运维体系中的关键组成部分，BMC 正从辅助管理组件，走向智能算力基础设施的核心控制节点。

主要发展趋势可从如下五个方向看：

标准化：从 IPMI 走向 Redfish 主导，全机型统一接口，降低多设备纳管难度。

智能化：从“监控”走向“诊断 + 决策 + 联动”，BMC 不再只是“采数据”，而是逐步具备硬件自治运维能力。

安全化：从管理入口变成安全边界，未来 BMC 的安全要求会越来越高，未来 BMC 不只是“运维入口”，更是服务器硬件信任链的重要组成部分。

场景化：面向 AIDC/液冷/GPU 集群深度演进，BMC 的职责将明显扩展，从“单机视角”走向“节点 + 机柜 + 集群”协同视角。

平台化：从单点管理走向大规模编排，未来 BMC 的价值，不在单台机器点进去看页面，而在于它是否能被批量化、自动化、平台化使用。

结语

GW-scale Open AIDC 框架技术报告，是面向人工智能时代算力基础设施快速演进需求的一次系统性回应。全文以开放、高效、绿色、智能为主线，将吉瓦级智算中心的建设统筹于基础设施层、IT 设施层、网络与高速互连层、系统软件与资源管理层四个相互依存的层次之中，强调在规划阶段即对能源、计算、存储、网络、冷却与运维进行整体协同设计。

贯穿全篇的一个核心判断是：吉瓦级智算中心的竞争力，越来越取决于系统级的可靠性、可维护性、能效与可持续性，而非单点的峰值性能。为应对算力需求的指数级增长，本技术报告提炼出若干具有共性的演进方向：

a) 园区级、可分期、可复制的开放参考架构，以降低跨主体协作成本；
b) 面向 140kW 以上单机柜、吉瓦级园区总功率的高功率密度建筑与结构设计；
c) 以高压直流（HVDC，±400V/800VDC）与“源网荷储”协同为代表的绿色供备电体系；
d) 以冷板、浸没为代表的液冷系统，及其与园区热回收、可再生能源的耦合；
e) Scale-Up/Scale-Out/Scale-Across 多层级、开放解耦的高速互连与网络体系；
f) 以 Redfish 为统一接口、面向超节点的智能自治与统一运维管理。

这些方向既体现了系统级协同设计的内在要求，也反映了中国在新能源供给、高压直流、制造业供应链与大规模工程组织等方面的独特优势。工作组希望以本技术报告凝聚产业共识，**邀请组件设计、芯片、整机、土建、供配电、冷却、网络、存储与运维等各环节的伙伴共同参与**，沿“需求—架构—接口—规范”的路径持续完善，并在对齐 OCP 全球方法论的基础上，推动可被国际理解与采纳的系统性方案，实现中国实践与全球开放体系的协同共建。

术语表（Glossary）

本术语表汇集技术报告中涉及的关键术语与缩略语，供读者统一参照。

AIDC（AI Data Center）：人工智能数据中心。以系统级协同设计为特征、面向大规模 AI 训练与推理负载的新一代数据中心形态。

BBU/BESS：机柜级电池备份单元（Battery Backup Unit）与园区级电池储能系统（Battery Energy Storage System）。

BMC（Baseboard Management Controller）：基板管理控制器，承担服务器硬件的带外监控与管控，是 AIDC 运维底座。

CDU（Coolant Distribution Unit）：冷量分配单元，为液冷二次侧提供受控的流量、压力与温度。

DPU/智能网卡：数据处理单元/智能网络接口卡，承担网络、存储与安全卸载。

GW 级智算中心：园区总功率达到吉瓦（Gigawatt，10^9瓦）量级的超大规模智算中心，需在园区级乃至区域级尺度上进行系统工程统筹。

HBD（High Bandwidth Domain）：高带宽域。为加速器之间提供高带宽、低时延通信的专用网络。

HVDC：高压直流供电（如±400VDC、800VDC），用于降低传输电流与线路损耗、提升系统效率。

NVLink/UALink：加速器卡间高速互连协议；UALink 为开放标准的 GPU-GPU 互连规范。

OAM/UBB：开放加速器模块（Open Accelerator Module）与通用基板（Universal Base Board），OAI 体系的核心规范。

OCP（Open Compute Project）：开放计算项目。面向数据中心的开源硬件与系统设计协作社区。

OCTC：开放计算标准工作委员会，隶属中国电子工业标准化技术协会（中电标协）。

OISA/SUE：全向智感互联（OISA）与可扩展统一以太（SUE）等卡间互连技术路径。

ORv3（Open Rack v3）：OCP 21 英寸开放机架规范，面向高密度、液冷与高功率系统。

Pod：在并行计算语境下被纵向扩展（Scale-Up）以充当单一逻辑单元的一组计算节点/资源域。

PUE/WUE：电能使用效率（Power Usage Effectiveness）与水使用效率（Water Usage Effectiveness），数据中心能效与水效核心指标。

Redfish：开放的 RESTful 硬件管理标准，用于带外管理、遥测与自动化。

RoCEv2：RDMA over Converged Ethernet v2，基于以太网承载 RDMA 的协议。

Scale-Up：在单一 Pod 内通过高带宽互连纵向扩展加速器规模（如 NVLink、UALink、OISA、SUE）。

Scale-Out：通过高性能网络横向扩展，将多个 Pod/机柜互联为更大规模集群（如 RoCEv2、UEC）。

Scale-Across：面向跨智算中心、长距离互联的算力扩展，需解决距离自适应负载均衡与广域确定性互联。

SONiC：开源网络操作系统（Software for Open Networking in the Cloud）。

SST（固态变压器）：Solid State Transformer，可将中压交流直接变换为受控直流，缩短供电链路。

TCS/FWS：技术冷却系统（Technology Cooling System，二次侧）与设施水系统（Facility Water System，一次侧）。

UEC（Ultra Ethernet）：超以太网联盟提出的、面向 AI/HPC 优化的以太网体系。

超节点（Super Node）：由多颗加速器经高带宽域紧耦合、对外呈现为单一逻辑计算单元的整机柜或跨机柜系统。

东数西算：国家算力枢纽节点布局工程。依托东部需求密集、西部资源（尤其绿电与土地）富集的禀赋差异，将东部非实时、时延不敏感的算力需求有序引导至西部，优化全国算力资源的跨域配置。

算电协同：算力负荷与电力系统跨域协同调度的机制与工程实践。将算力中心作为可调节负荷，使其时空分布与电力（尤其新能源）供给相匹配，并参与电网调节，从而提升算力中心能效与电网稳定性。

冷板/浸没（Cold Plate/Immersion）：两类主流液冷方式：冷板直触芯片导热；浸没将 IT 设备浸于介电流体中散热。

参考与延伸阅读（References）

本技术报告在方法论与体例上参考了 OCP Open Systems for AI 战略倡议的相关成果，并与 OCP 中国社区、OCTC 的开放计算工作相衔接。建议读者结合以下来源进一步阅读：

a) Open Compute Project,“Open Systems for AI: Blueprint for Scalable Infrastructure”, v1.0.0, 2025.

b) Open Compute Project 各项目与规范（Server/DC-MHS、OAI、Open Rack v3.Cooling Environments、Hardware Management/Redfish Profiles、Networking 等），https://www.opencompute.org。

c) 开放计算标准工作委员会（OCTC）/中国电子工业标准化技术协会相关开放计算标准与技术报告。

d) GW-scale Open AIDC 工作组公开工作清单与配套技术文档（随正式发布版本一并公布）。

关于 OCTC 与 OCP 中国社区

开放计算标准工作委员会（OCTC）隶属于中国电子工业标准化技术协会（中电标协），致力于推动开放计算相关技术标准的研究、制定与产业落地，促进数据中心从云到边的开放创新。

OCP 中国社区（OCP China Community）是 Open Compute Project 在中国的开放协作平台，汇聚云计算厂商、超大规模数据中心运营方、电信与托管服务商、企业用户及产品与解决方案生态伙伴，围绕影响力、效率、规模与可持续四大准则，推动以 AI/ML、光互连、先进冷却、可组合内存与硅为代表的重大技术演进，将超大规模引领的创新惠及更广泛的产业。

GW-scale Open AIDC 工作组在上述两个组织的共同指导下成立，面向中国国情与工程实践，聚焦吉瓦级智算中心的开放参考架构与规范共建，欢迎产业各界加入与共建。

了解更多：https://www.octc.net 、https://www.opencompute.org