NVIDIA GPU 微架构与 NVL72 AI 工厂：从基础概念到 GB300 网络设计

> 面向工程/研究人员的中文技术教程。读者不需要任何计算机体系结构背景，我们将从"流量方向""扩展方式"这类最基础的概念讲起，一步步深入到 GB300 NVL72 整机柜的网络设计，最后逐词拆解一段 NVIDIA 官方反馈中的每一个技术点。

---

0. 引言：为什么要写这篇文档

0.1 缘起：一段看不懂的官方反馈

假设你是一位工程或研究人员，正在参与一个基于 NVIDIA GB300 NVL72 的 AI 集群项目。某天你收到 NVIDIA 官方团队关于 NVL72 网络问题的一段反馈：

> "网卡标记 ECN 解决网络到 Host Memory 因为 PCIe Gen5 导致 PFC 的问题。这个功能目前在 CX8 网卡上还没有实现，CX7 上有但没啥用处。正在和网卡架构团队讨论实现需要的资源。"

短短四句话，里面塞满了缩写和机制：**ECN、Host Memory、PCIe Gen5、PFC、CX8、CX7**。不懂体系结构的人读完只有一个感觉——每个字都认识，连起来完全不知道在说什么。更麻烦的是，这句话不是教科书上的知识点罗列，而是一个真实的工程权衡：为什么 CX7 上有这个功能却"没啥用"？为什么更新的 CX8 反而还没有？为什么 PCIe Gen5 会和 PFC 这种以太网二层流控纠缠在一起？

本文档就是为回答这些问题而写的。我们不会假设你懂任何体系结构知识，而是从零开始，把这段话里的每个概念拆开、讲透，最后回到这段话本身，做一次逐词拆解。

0.2 本文的主要参考来源

本文档的技术事实主要来自 NVIDIA 官方企业参考架构文档《NVL72 AI Factory — Enterprise Reference Architecture with NVIDIA GB300 NVL72 and NVIDIA Spectrum-X Networking Platform》（2026 年 3 月版，共 38 页，以下简称"官方 RA 文档"）。文中嵌入的整机柜实拍图、计算 tray 架构图、网络拓扑图均出自该文档，并在图注中注明页码与图号。官方 RA 文档的在线地址为：https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory-with-gb300-nvl72-dual-plane-networking-architecture.pdf

0.3 阅读路线图

全文按"由浅入深"组织为六部分加附录：

- **第一部分 基础概念**：南北向 vs 东西向流量、Scale-up vs Scale-out、RDMA 与 RoCEv2 入门、GPU Direct RDMA。没有体系结构背景的读者从这里开始。
- **第二部分 GPU 体系结构基础**：CPU 与 GPU 的本质差异、SM、HBM、Tensor Core、NVLink 五代演进、NVSwitch、Grace-Blackwell 超级芯片、NVLink-C2C、MGX 机柜形态。
- **第三部分 深入 GB300 NVL72**：整机柜解剖（配官方实拍图与逻辑架构图）、计算 tray 各组件、NVLink Switch tray、DPU、三张物理分离的 fabric、双平面设计、2/4/8 机柜扩展拓扑。
- **第四部分 无损网络深挖**：为什么 RoCE 需要无损网络、PFC 的机制与缺陷、ECN/DCQCN 闭环、PFC 与 ECN 的门限配合、headroom buffer、Spectrum-X 的自适应路由与 ECN 的关系。

- **第五部分 逐词拆解 NVIDIA 反馈**：把引言里那段话逐词拆开，画出"网卡→PCIe→Host Memory"的拥塞传导链，解释 CX7 有功能却没用、CX8 反而没实现的来龙去脉。
- **第六部分 扩展概念**：与 InfiniBand（Quantum）方案对比、SHARP 网内计算、NVL72 vs 单机 8 卡 HGX 的取舍、CPO/硅光、下一代 Rubin 展望。
- **附录**：术语表（40+ 条）、FAQ 汇总（15+ 问）、参考链接。

每一部分末尾都有**小结**和**思考题/FAQ**模块，帮助读者自检。如果你已经熟悉 RDMA 和无损网络，可以直接跳到第四、五部分。

0.4 小结

本文档围绕一个真实工程问题展开：读懂 NVIDIA 关于"ECN 标记解决 PCIe Gen5 引发的 PFC 问题"的反馈。我们将用六个部分，从流量方向这种最基础的概念，一路走到 GB300 NVL72 的三网分离架构与 ECN/PFC 门限配合的细节，最后再回到这段话。

0.5 思考题

1. 为什么一段关于"网络"的反馈里会出现"PCIe"这个主机总线术语？这暗示了什么问题？
2. 在继续阅读之前，先写下你对"无损网络"这个词的直觉理解，读完第四部分后回来对比。

第一部分 基础概念

1.1 南北向 vs 东西向流量

数据中心网络工程师用地理方位比喻流量方向。

南北向流量（North-South Traffic） 指进出整个数据中心的流量：用户请求从外部互联网进来（由北向南），响应再出去（由南向北）。在 AI 集群语境下，南北向还包括集群与存储系统之间的流量——训练数据从存储读进来、checkpoint 写回存储，以及集群对外的管理、客户网络连接。

东西向流量（East-West Traffic） 指数据中心内部服务器与服务器之间的横向流量。在传统 Web 业务里，东西向通常是微服务之间的调用；在 AI 训练集群里，东西向流量有一个非常具体且压倒性的含义：**GPU 与 GPU 之间的集合通信流量**——比如 data parallel 训练中的 AllReduce、专家并行（EP）中的 All-to-All。AI 集群的主体流量是东西向的，一次大模型训练的 step 里，几千颗 GPU 要反复交换梯度或激活值，流量规模远超南北向。

为什么这个区分在工程上如此重要？因为两类流量的**服务质量目标完全不同**：

- 东西向训练流量要求极高带宽、极低延迟、零丢包（原因见第四部分），任何拥塞都会让全集群 GPU 空转等待；
- 南北向存储/管理流量要求稳定可靠，但对微秒级延迟不敏感。

如果把它们混在同一张物理网络上，存储的大块读写可能挤占训练流量的带宽，训练流量的突发又可能饿死管理流量。所以现代 AI 集群的做法是**物理分离**：为东西向计算流量建一张专门的 fabric，为南北向存储/管理流量建另一张 fabric。GB300 NVL72 官方 RA 正是这么做的，而且还有第三张专门的带外管理网（见第三部分）。

1.2 Scale-up vs Scale-out

扩展算力有两种基本路线。

Scale-up（纵向扩展）：把单一紧耦合计算域做大——在同一台服务器或同一个机柜内放更多 GPU，用超高带宽、极低延迟的互连（NVLink/NVSwitch、PCIe、CXL、UALink）把它们连起来，让它们协同工作时像"一台更大的机器"。

Scale-out（横向扩展）：把多个节点、多个机柜连成更大的集群——通过 InfiniBand（Quantum 系列）或以太网（Spectrum-X / RoCEv2）连接成百上千个节点。

两者的工程特征对比如下：

维度	Scale-up	Scale-out
范围	节点内 / 机柜级域	跨节点、跨机柜、集群级
核心指标	极致带宽、ns/μs 级延迟	高吞吐、可扩展性、无阻塞
拓扑	Mesh / 全互联（NVSwitch）	Spine-Leaf / Fat-Tree / Clos
协议	NVLink、PCIe/CXL、UALink	InfiniBand、RoCEv2、TCP/IP
典型带宽	每 GPU 1.8 TB/s（NVLink5）	每 GPU 400–800 Gb/s（CX-7/CX-8）
主要挑战	信号完整性、供电散热、铜连距离	拥塞控制、丢包、负载均衡、光模块成本

注意一个关键的数量级差异：Scale-up 域内每 GPU 的互连带宽是 1.8 TB/s，而跨域 Scale-out 每 GPU 只有 800 Gb/s（= 100 GB/s），相差约 18 倍。这个带宽落差决定了一条重要的工程规律：**张量并行（TP）、专家并行（EP）这类细粒度、高频次的集合通信要尽可能留在 Scale-up 域内**；更大规模才用 Scale-out 叠加数据并行（DP）。GB300 NVL72 把 Scale-up 域做到了 72 颗 GPU 一个机柜，正是为了让尽可能多的通信留在高速域内。

1.3 RDMA 与 RoCEv2 入门

RDMA（Remote Direct Memory Access，远程直接内存访问） 是一种让一台机器直接读写另一台机器内存、而不需要对方 CPU 参与的网络技术。传统 TCP/IP 通信要走完整的内核协议栈：数据先拷贝到内核 buffer，再逐层封装、发送，接收方再逐层解封装、拷贝到用户态——每一跳都消耗 CPU 和延迟。RDMA 则允许网卡（HCA/NIC）绕过操作系统内核，直接对远端应用的内存做 DMA 读写，把网络传输的延迟压到微秒级，同时把 CPU 完全解放出来。

RDMA 有三种主流实现：

- **InfiniBand（IB）**：原生为 RDMA 设计的专用网络，天然无损（基于 credit 的流控），性能最好，但需要全套专用交换机和网卡，生态封闭、成本高，由 NVIDIA（Mellanox）主导。
- **RoCE（RDMA over Converged Ethernet）**：把 RDMA 跑在标准以太网上。RoCEv1 工作在二层、不可路由；**RoCEv2** 把 RDMA 封装在 UDP/IPv4（或 IPv6）之上，可以跨三层路由，是当前 AI 数据中心以太网方案的事实标准。本文讨论的 Spectrum-X 网络跑的就是 RoCEv2。
- **iWARP**：跑在 TCP 之上，性能与生态均不如前两者，AI 领域基本不用。

RoCEv2 有一个根本性问题：以太网默认是**有损**的——交换机 buffer 满了就直接丢包，靠上层重传兜底。而 RDMA 对丢包近乎零容忍：RoCEv2 的重传机制是 go-back-N，丢一个包就要从该包开始重传整个窗口，一次丢包足以拖垮一次集合通信（几千颗 GPU 一起等）。所以 RoCEv2 必须搭配"无损以太网"技术（PFC、ECN 等）使用——这正是第四部分的主题，也是那段 NVIDIA 反馈的核心背景。

RDMA 的编程模型：Verbs 与 Queue Pair。 RDMA 不是套接字编程，而是一套称为 Verbs 的语义：应用在网卡上创建 Queue Pair（QP，一对发送队列 SQ 和接收队列 RQ）与完成队列（CQ），然后向队列里投递工作请求（WQE）——Send/Receive 用于消息语义，Write/Read 用于直接读写远端内存（RDMA 的精髓），Atomic 用于远端原子操作。网卡硬件异步执行这些请求并把结果放进 CQ。每个 QP

有服务类型：RC（Reliable Connected，可靠连接，AI 训练最常用）、UC、UD、XRC。理解 QP 有助于理解后文：DCQCN 的调速就是以 QP 为粒度进行的——收到 CNP 后，被降速的是具体的某条 QP，而不是整张网卡。

RoCEv2 的报文长什么样。 一条 RoCEv2 报文从外到内依次是：以太头（VLAN/PCP 里携带 CoS）→ IP 头（DSCP 与 ECN 位在这里）→ UDP 头（目的端口 4791）→ IB BTH 基本传输头（携带 QP 号、序号 PSN、操作码）→ 载荷。所以二层用 PCP/CoS、三层用 DSCP，两套 QoS 标记必须映射一致，这正是 4.4 节"惯例 DSCP26→CoS3、DSCP48→CoS7"的来源。

1.4 GPU Direct RDMA（GDR）是什么

RDMA 解决的是"机器到机器"的零拷贝，但在 AI 服务器里还有一个更细的问题：数据到了网卡之后，怎么到 GPU？

没有 GDR 的传统路径是：网络数据 → 网卡 → CPU 内存（拷贝一次）→ 再从 CPU 内存经 PCIe 拷到 GPU 显存（拷贝第二次）。两次拷贝都由 CPU 驱动，既慢又占用 CPU。

GPU Direct RDMA（GDR） 让网卡的 DMA 引擎可以**直接读写 GPU 的 HBM 显存**，完全绕过 CPU 内存和 CPU 本身：

- 接收方向：网线 → 网卡（如 ConnectX-7/8）→ 网卡内 DMA 引擎 → PCIe（Gen5/Gen6 x16）→ GPU HBM，全程零拷贝、不经过 CPU 和内核协议栈。
- 发送方向反之：GPU HBM → PCIe P2P → 网卡 → 网络。

GDR 是 NCCL 跨节点通信的底层支柱。请特别注意这条路径里出现的两个域：**以太网链路段**（网线到网卡，由 PFC 等二层流控管辖）和 **PCIe 主机总线段**（网卡到 GPU/内存，由 PCIe credit 流控管辖）。两者物理上不相交，却在 GDR 数据路径上首尾相接，通过网卡内部的 buffer 耦合在一起——任何一端成为瓶颈，都会以"反压"的形式传导到另一端。请记住这句话，第五部分拆解 NVIDIA 反馈时它就是钥匙。

1.5 小结

- 东西向流量是集群内部 GPU 间通信，是 AI 集群的主体流量；南北向是集群与存储、外部的流量。两者 QoS 目标不同，工程上物理分离。
- Scale-up 追求单一域内极致带宽（NVLink，1.8 TB/s/GPU），Scale-out 追求规模（以太网/IB，800 Gb/s/GPU），带宽落差约 18 倍，决定并行策略的域内/域外划分。
- RDMA 让网卡绕过 CPU 直接读写远端内存；RoCEv2 把它跑在可路由的标准以太网上，但以太网有损，必须配套无损技术。
- GPU Direct RDMA 让网卡直接读写 GPU 显存；其数据路径跨越以太网链路段与 PCIe 主机段，两段通过网卡 buffer 耦合，反压可以跨域传导。

1.6 思考题 / FAQ

Q1：为什么 AI 训练流量主要是东西向而不是南北向？
因为训练的主体计算是分布式的：每个 step 里几千颗 GPU 要互相交换梯度（AllReduce）或激活值（All-to-All），这些 GPU 间通信的流量规模比"从存储读数据"大几个数量级，且每个 step 都发生。

Q2：RoCEv2 为什么选 UDP 而不是 TCP 封装？
RDMA 语义要求低延迟、无连接式的高效传输；UDP 开销小、无 TCP 的复杂状态机，配合无损以太网后可靠性由链路层和 RDMA 自身的 ACK/重传机制保证，不需要 TCP。

Q3：GDR 是不是只用于 GPU 之间的通信？
不是。GPU Direct Storage（GDS）让存储设备也能直接读写 GPU 显存；host memory 路径（如存储流

量经 DPU 到 CPU 内存）同理也走零拷贝 DMA。

第二部分 GPU 体系结构基础

2.1 CPU 与 GPU 的本质差异

CPU 为"通用、低延迟"设计：少量强大的核心，每核有大缓存、复杂分支预测、乱序执行，擅长串行逻辑和控制流。GPU 为"吞吐、并行"设计：成千上万个相对简单的核心，放弃复杂控制逻辑，把晶体管预算全部堆给算术单元，擅长对海量数据做同一种运算（SIMT 模型）。大模型训练恰好是"对海量矩阵做同一种乘加运算"，因此 GPU 成为 AI 算力的载体。

2.2 SM、HBM 与 Tensor Core

现代 NVIDIA GPU 内部的三个关键概念：

- **SM（Streaming Multiprocessor，流式多处理器）**：GPU 的基本计算单元，相当于一个"小核心集群"，内含 CUDA Core、Tensor Core、共享内存、调度器。一颗 Blackwell Ultra GPU 有上百个 SM。
- **Tensor Core**：SM 内部专为矩阵乘加（D = A×B + C）设计的硬件单元，支持 FP16/BF16/FP8/FP4 等低精度格式，是大模型算力（FLOPS）的主要来源。Blackwell Ultra 相比 Hopper 大幅增强了 FP4 推理能力。
- **HBM（High Bandwidth Memory，高带宽显存）**：与 GPU 芯片通过硅中介层 2.5D 封装的堆叠式显存。B300 配备 HBM3e，单颗 GPU 显存容量达 288 GB 级别、带宽数 TB/s。HBM 容量直接决定单机能放多大的模型，HBM 带宽直接决定访存密集型算子（如 decode 阶段）的速度。

2.3 NVLink 的五代演进

PCIe 是通用总线，带宽有限（Gen5 x16 单向约 63 GB/s）。当多颗 GPU 需要像一个整体一样交换数据时，PCIe 远远不够，NVIDIA 为此设计了专用互连 **NVLink**：

代际	首发平台	每 GPU 总带宽（双向）
NVLink 1	Pascal P100（2016）	160 GB/s
NVLink 2	Volta V100（2017）	300 GB/s
NVLink 3	Ampere A100（2020）	600 GB/s
NVLink 4	Hopper H100（2022）	900 GB/s
NVLink 5	Blackwell / GB300（2024-）	1800 GB/s

第五代 NVLink 每 GPU 提供 1.8 TB/s 双向带宽——是 PCIe Gen5 x16 的约 28 倍。这是"Scale-up 域"成立的物理基础。

在 GB300 NVL72 中，这一带宽的物理实现是：每颗 B300 GPU 对外引出 **18 条 NVLink5 链路**，每条 100 GB/s 双向，合计 1800 GB/s。这 18 条链路不是随便接到邻居 GPU 上，而是全部拉进 NVSwitch 交换层（见下一节）——这是 NVL72 能把 72 颗 GPU 编成一个域、而 HGX 只能编 8 颗的关键差异之一。NVLink 承载的是**内存语义**：GPU 可以像执行普通 load/store 指令一样直接访问域内任何一颗 GPU 的 HBM，而不需要显式的消息发送——这使得宽 TP/EP 的通信开销远低于基于网络的方案。

2.4 NVSwitch：从点对点连线到全互联交换

仅有 NVLink 直连，GPU 之间只能 mesh/环形互连，跳数随 GPU 数增长。**NVSwitch** 是一颗专门的交换芯片，把所有 GPU 的 NVLink 链路汇聚到交换机上，实现域内任意两颗 GPU 之间**全带宽、无阻塞、单跳**

的通信。在 GB300 NVL72 中，72 颗 GPU 各自拉出 18 条 NVLink5 链路，接入 9 个 NVSwitch tray（每个 tray 内 2 颗 NVSwitch ASIC），整机柜聚合计 130 TB/s 交换容量——72 颗 GPU 因此"像一颗 GPU 一样"工作。

2.5 Grace-Blackwell 超级芯片与 NVLink-C2C

Grace 是 NVIDIA 自研的 ARM 架构服务器 CPU（Neoverse V2 核，72 核，配最高 1TB LPDDR5 内存），其存在的意义不是跑通用计算，而是给 GPU 当"超大容量、超高带宽的内存侧伙伴"。

NVLink-C2C（Chip-to-Chip） 是 NVLink 技术的片间版本：在 Grace CPU 与 Blackwell GPU 之间提供 900 GB/s 级别的一致性互连——CPU 和 GPU 可以像访问自己的内存一样访问对方的内存（硬件 cache 一致性），彻底消除了传统 CPU-GPU 之间 PCIe 拷贝的开销。一颗 Grace + 两颗 B300 组成一个 Grace-Blackwell 超级芯片模组，这正是 GB300 计算 tray 的核心。

2.6 MGX 机柜形态

MGX 是 NVIDIA 提出的模块化服务器/机柜参考设计规范：定义了计算 tray、交换 tray、供电、液冷、母线的标准机械与电气接口，让 OEM/ODM 可以快速组合出整机柜产品。GB300 NVL72 采用液冷 MGX 机柜形态：整柜功耗约 142 kW，由 8 个 33 kW 电源 shelf 供电，机柜内交换设备直接挂 DC 母线（busbar）供电。液冷不是可选项而是必需品——这个功率密度下风冷完全无法胜任。

2.7 小结

- GPU 以吞吐换延迟，SM 是计算单元，Tensor Core 提供矩阵算力，HBM 提供容量与带宽。
- NVLink 五代演进把每 GPU 互连带宽从 160 GB/s 推到 1.8 TB/s；NVSwitch 让域内 72 颗 GPU 全互联、无阻塞。
- Grace-Blackwell 通过 NVLink-C2C 实现 CPU-GPU 内存一致性；MGX 定义了液冷整机柜的模块化形态。

2.8 思考题 / FAQ

Q1：为什么 Scale-up 域内的通信尽量用 NVLink 而不是以太网?
因为带宽差约 18 倍、延迟差近一个数量级，且 NVLink 支持内存语义（load/store 直接访问对端显存），以太网 RoCE 只能走消息语义。

Q2：Grace CPU 为什么用 LPDDR5 而不是常规 DDR5 RDIMM?
LPDDR5 封装更近、能效更高、带宽密度更大，配合 NVLink-C2C 给 GPU 提供大容量一致性内存池；代价是不可扩展、不可现场更换。

Q3：NVSwitch 和普通以太网交换机本质区别是什么?
NVSwitch 交换的是内存语义的 NVLink 报文，域内任意 GPU 间单跳全带宽；以太网交换机交换 IP 报文，受拥塞控制、路由、队列等一整套网络问题支配——这正是第四、五部分的主题。

第三部分 深入 GB300 NVL72

3.1 整机柜解剖

GB300 NVL72 是一个把 72 颗 Blackwell Ultra（B300）GPU 和 36 颗 Grace CPU 装进一个液冷 MGX 机柜的 rack-scale 系统。官方 RA 文档称其相对 Hopper 平台可带来最高 50 倍的 AI 工厂整体产出提升。整柜的关键数字：

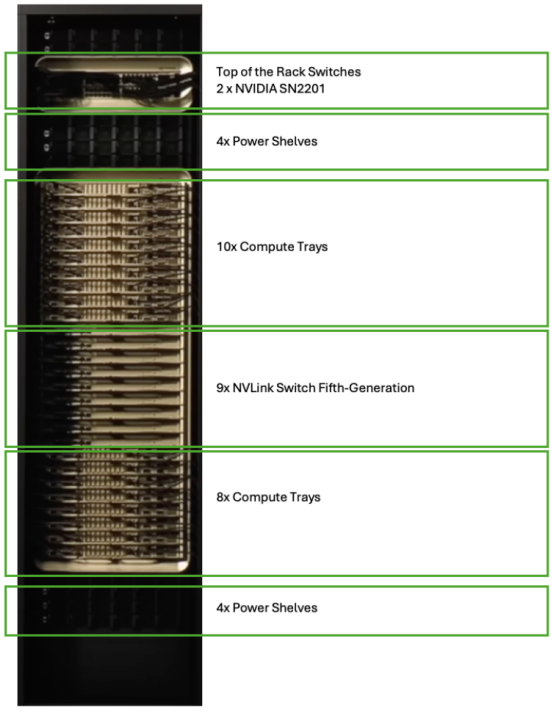
- 72 × B300 GPU + 36 × Grace CPU（每颗 Grace 为 72 核 ARM Neoverse V2，配 1 TB 级 LPDDR5，经 NVLink-C2C 与 GPU 互连）
- 18 个计算 tray（每个 tray 2 颗 Grace + 4 颗 B300）
- 9 个 NVLink Switch tray，第五代 NVLink 全互联，机柜聚合带宽 130 TB/s
- 液冷，整柜约 142 kW，8 × 33 kW 电源 shelf；柜内交换机挂 DC 母线供电
- 2 台 SN2201 带外管理交换机集成在机柜内

System Hardware & Components

This section describes the hardware elements of the Enterprise RA.

NVIDIA GB300 NVL72The NVIDIA GB300 NVL72(Figure 1) delivers breakthrough AI performance and can deliver up to a 50x overall increase in AI factory output performance compared to NVIDIA Hopper-based platforms. The NVIDIA Blackwell Ultra GPU architecture provides the latest technologies that bring months of computational effort down to days and hours on some of the largest AI/ML workloads.

Figure 1 NVIDIA GB300 NVL72



To accommodate the massive compute power within a limited space, the GB300 NVL72 rack is liquid cooled, based on the MGX architecture. Some key highlights of the GB300 NVL72 compute rack include:

图：GB300 NVL72 整机柜实拍与部件标注（出自 NVIDIA 官方 RA 文档第 6 页 Figure 1）。可见自上而下排布的计算 tray、NVLink Switch tray、电源 shelf 与机柜级管理交换机。

3.2 计算 tray：逐一组件

计算 tray（compute tray）是 GB300 NVL72 的基本服务器单元，官方文档把每个 tray 定义为一个"Node"——一个可运行裸金属操作系统的独立硬件单元。整柜 18 个计算 tray，每 tray 的组成：

组件	数量	说明
NVIDIA Grace CPU	2	合计 72 核 ARM Neoverse V2，NVLink-C2C 互连
NVIDIA Blackwell Ultra GPU（B300）	4	每颗经 NVLink5 出 18 条链路
ConnectX-8 SuperNIC	4	每颗 800 Gb/s，GPU:NIC = 1:1
BlueField-3 B3240 DPU	1	双 400 Gb/s 口，跑南北向存储/带内网络
系统内存	1 TB 级 LPDDR5	随 Grace 封装
存储	NVMe	数据缓存盘
BMC	1	带外管理，Redfish

- 9 NVSwitch trays for full non-blocking P2P connectivity of all 72 Blackwell Ultra GPUs, with a total aggregated bandwidth of 130 TB/s
- 2 SN2201 OOB switches for integrated management access of rack components. The GB300 NVL72 in-rack switches use DC power since they are connected to the DC busbar; all other SN2201 switches use AC power
- Integrated tray-level and rack-level liquid leakage detection
- 8 power shelves of 33 kW, with each shelf having six 5.5 kW PSUs
- Full rack requiring up to 142 kW

NVIDIA GB300 NVL Compute Tray

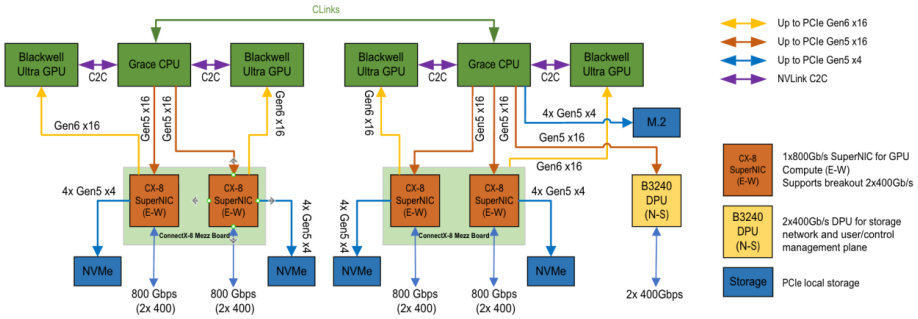
The compute building block for the NVIDIA GB300 NVL72 is the GB300 NVL compute tray/node (Figure 2), with 4 Blackwell Ultra GPUs and 2 Grace CPUs per tray. A logical diagram is shown in Figure 3.

Figure 2 NVIDIA GB300 NVL Compute Tray

The compute tray is a fully integrated NVLink scale-out architecture, so that each GPU in the entire rack is directly connected to every other GPU via NVLink. For compute network connectivity, each compute tray includes 2 Mezzanine Network Boards with 2 ConnectX-8 silicon chips in each, for a total of 4 ConnectX-8 Host Channel Adapters (HCA). Each tray also includes 1 M.2 NVMe device for Operating System (OS) storage and 4 E1.S NVMe storage devices typically used as a fast, local cache.

图：计算 tray 实物（出自 NVIDIA 官方 RA 文档第 7 页 Figure 2）。液冷冷板覆盖 GPU/CPU 模组，前方为网络与存储接口区。

Figure 3 Sample logical design for the compute tray



To ensure balanced compute performance with the GPUs, each CPU is paired with a NVIDIA ConnectX-8 Mezzanine Board. Additionally, to support advanced storage acceleration, each tray includes a NVIDIA Bluefield-3 DPU.

Table 1 Compute Tray Components

Component	Quantity
NVIDIA Grace Processor, with a total of 72 ARM Neoverse V2 cores, connected via NVLink C2C, and 1TB aggregated LPDDR5 CPU main memory	2
NVIDIA B300 GPU, with 1,152 GB aggregated HBM3 memory	4
NVIDIA ConnectX-8 Mezzanine Boards, with 2 ConnectX-8 network adapters each	2
Dual-port QSFP112 NVIDIA BlueField-3 B3240 DPU	1

Control Plane/Management Nodes

The customer should have guidance from the OEM regarding an IaaS-managed control plane nodes. The High Availability (HA) features of the control plane nodes should be enabled to ensure reliability for the enterprise customer. In this Enterprise RA, we assume a control plane composed of 12 management nodes, providing sufficient capacity and HA for large enterprise deployments. Table 2 outlines the specifications of the x86 Management Nodes while Table 3 outlines the specifications of the ARM Management Nodes.

Table 2 Management Node Components (x86 control node)

Component	Configuration
CPUs	Intel based x86 2x Socket, 32c
System Memory	512 GB (8x DDR5 64GB)

图：计算 tray 逻辑设计（出自 NVIDIA 官方 RA 文档第 8 页 Figure 3）。可见 4×B300 GPU、2×Grace CPU、4×ConnectX-8（经 Mezzanine 板）、1×BlueField-3 DPU 以及它们之间的 PCIe 链路标注。

Figure 3 值得细看，它是理解第五部分那段反馈的硬件底图：

- **GPU:NIC = 1:1**。每颗 B300 对应一颗专属 CX-8。这个"1:1 rail"设计的目的，是让每颗 GPU 的跨节点通信都有自己独占的 800G 网络通道，NCCL 可以把每颗 GPU 的流量钉在固定的 rail 上，避免多 GPU 抢一张网卡造成的不可预测拥塞。

- **CPU 与 NIC 配对**。官方文档原文："To ensure balanced compute performance with the GPUs, each CPU is paired with a NVIDIA ConnectX-8 Mezzanine Board."——每颗 Grace 配一块 CX-8 Mezzanine 板（每板 2 颗 CX-8 ASIC），保证 host 侧 PCIe 带宽与网络带宽对称。
- **BlueField-3 DPU 独立承担南北向**。存储与带内管理流量不碰 CX-8，物理上走另一条 PCIe 路径和另一组光口。
- **PCIe 链路标注**。图中 CX-8 与 Grace/GPU 之间、BF-3 与 CPU 之间的连线就是 PCIe 总线——GB300 平台上 CX-8 走 PCIe Gen6 x16。这条链路正是"网卡到 Host Memory"路径的物理载体，请记住它。

3.3 NVLink Switch tray

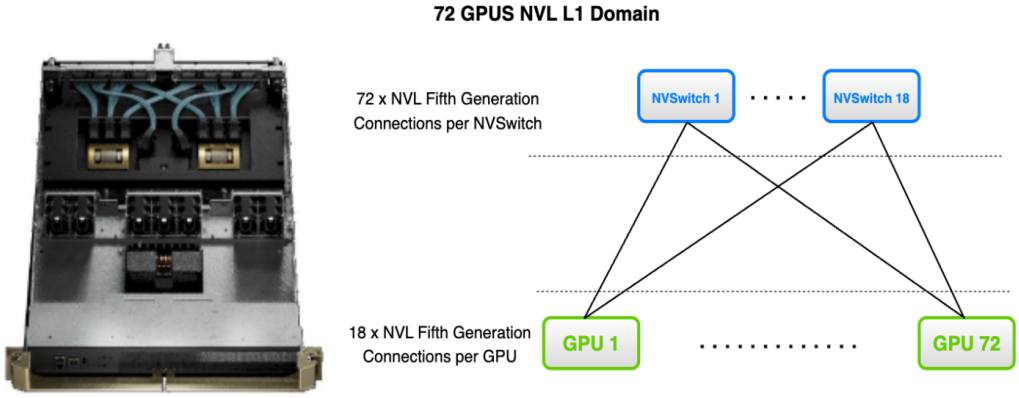
每柜 9 个 NVLink Switch tray，每 tray 内 2 颗 NVSwitch ASIC。72 颗 GPU 每颗出 18 条 NVLink5 链路，全部接入这 18 颗交换芯片，构成一个单一 NVLink 域：任意两颗 GPU 之间单跳、全带宽（1.8 TB/s 双向）、无阻塞通信，机柜聚合 130 TB/s。官方 RA 文档指出，这使得 72 颗 GPU 能作为"单一多 GPU 计算单元"工作，支撑万亿参数模型的实时推理。

Since the GB300 NVL72 integrates 36 NVIDIA Grace CPUs with 72 NVIDIA Blackwell Ultra GPUs in a rack-scale design, the fifth-generation NVLink enables all 72 GPUs to function as a single multi-GPU unit of compute enabling 30x faster real-time trillion-parameter inference compared to prior generations.

NVIDIA NVLink Switch Tray

Each GB300 NVL72 rack includes 9 NVLink Fifth-Generation switch trays, with 2 NVSwitch ASICs per tray. Each GPU has 18 NVLink Fifth-Generation links, one per in-rack NVSwitch via the copper backplane. This way a single rack may form a fully connected L1 Domain with 72 GPUs. The NVLink Switch is a managed switch running the NVOS networking operating system.

Figure 4 NVLink Switch



NVIDIA GB300 NVL72 Networking

The NVIDIA GB300 NVL72 platform networking configuration enables tremendous AI performance and scale. It leverages the NVIDIA expertise in AI cloud data centers and optimizes network traffic flow:

- **East/West East/West(Compute Network) traffic**: This refers to traffic between the compute nodes/trays within the GB300 NVL72 cluster, typically for multi-node AI training, HPC collective operations, and other workloads.
- **North/South (Customer and Storage Network) traffic**: This involves traffic between the GB300 NVL72 and any external resources including cloud management and orchestration systems, remote data storage nodes, and other parts of the data center or the Internet.

图：NVLink Switch tray（出自 NVIDIA 官方 RA 文档第 10 页 Figure 4）。

3.4 BlueField-3 B3240 DPU

每个计算 tray 配 1 颗 BlueField-3 B3240 DPU，双 400 Gb/s 端口。注意一个官方文档明确说明的细节：虽然每个端口支持 400 Gb/s，但这颗卡的聚合带宽低于 800 Gb/s（节点存储带宽约 40 GB/s 量级）。它在 NVL72 平台中承担：

- **南北向存储网络**：RoCE、NVMe-oF、SNAP 存储加速——把远端存储虚拟成本地 NVMe 盘；
- **带内管理（in-band management）**：控制面、编排、遥测；
- **零信任安全**：IPSec/MACSec 线速加密、微分段，把安全策略从 host CPU 卸载到 DPU；
- **带外供给（OOB provisioning）**：作为数据中心控制面的优化计算平台，支持自动化部署。

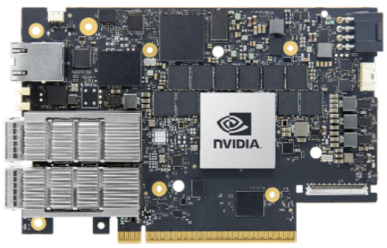
Networking Hardware

This section describes the networking hardware used to connect components of the Enterprise RA

NVIDIA BlueField-3 B3240 DPU

The NVIDIA BlueField-3 B3240 DPU with two 400 Gb/s ports (Figure 5) is used in the in--band management and storage fabric. Even though the B3240 DPU supports 400Gb/s per port, the card only supports an aggregate bandwidth across both its ports of approximately 480Gb/s. The Enterprise RA recommends the use of dual ports for the best performance and flexibility. Each compute tray has one B3240 DPUs.

Figure 5 NVIDIA Bluefield-3 B3240 DPU



NVIDIA Mezzanine Network Board with ConnectX-8

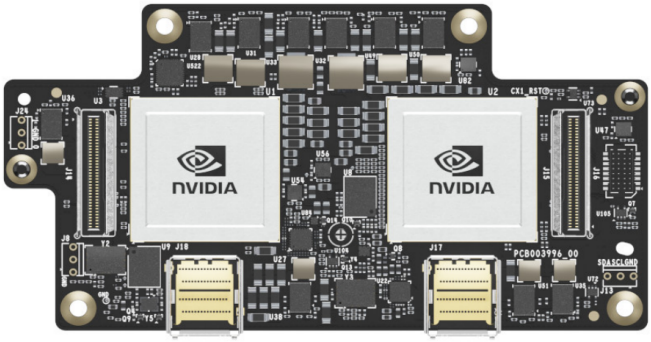
Each GB300 compute tray has a Mezzanine Network Board (**Figure 6**) that is used for Compute (East/West) network. The tray has two such boards, with two ConnectX-8 devices on each board.

图：BlueField-3 B3240 DPU（出自 NVIDIA 官方 RA 文档第 13 页 Figure 5）。

3.5 ConnectX-8 Mezzanine 板与 Spectrum-X 交换机

每计算 tray 装 2 块 Mezzanine Network Board，每块板 2 颗 CX-8 ASIC，合计 4 颗 CX-8。每颗 CX-8 出 800 Gb/s，经 OSFP 光口拆成 2 × 400G，分别接入双平面 leaf 交换机（详见 3.7）。

Figure 6 NVIDIA Mezzanine Network Board with two ConnectX-8 ASICs



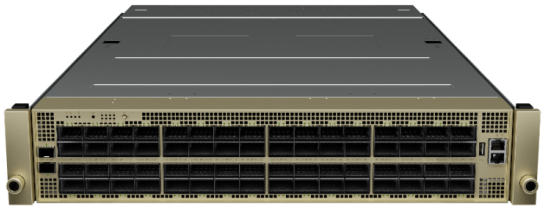
NVIDIA Ethernet Interconnects

The Enterprise RA configurations must be installed with NVIDIA Ethernet interconnects. NVIDIA Ethernet interconnects make extensive use of pluggable optical transceivers and detachable optical fibers for easier installation, inspection, and debugging.

NVIDIA SN5610 Switch

The NVIDIA SN5610 switch (Figure 7) both offer 64 total ports of 800 Gbps to provide connectivity for Compute (East/West), Customer and Storage (North/South), in-band management, and storage in the Enterprise RA.

Figure 7. NVIDIA SN5610 Switch



The NVIDIA SN5610 adds two SFP28 ports and makes switch testing easier since the ports are in pairs.

NVIDIA Enterprise RA 2026 March | 14

图：上为装载两颗 ConnectX-8 ASIC 的 Mezzanine Network Board（Figure 6），下为 SN5610 Spectrum-X 交换机（Figure 7），出自 NVIDIA 官方 RA 文档第 14 页 Figure 6/7。

柜间网络交换的核心是 **SN5600 / SN5610 Spectrum-X 交换机**：单机 64 × 800G 端口，51.2 Tb/s 交换容量，基于 Spectrum-4 ASIC，专为 AI 以太网 fabric 设计（自适应路由、遥测、与 CX-8 联动的拥塞控制，见第四部分）。

3.6 三张物理分离的 fabric

官方 RA 文档明确采用三张物理上完全分离的网络（three physical network fabrics）：

1. **GPU Compute 网（东西向）**：承载训练集合通信。每 tray 4 × CX-8，每颗 800G 拆 2 × 400G 进双平面，rail-optimized fat-tree，RoCEv2 无损，由 SN5600/SN5610 Spectrum-X 交换。这是性能最敏感的一张网。
2. **CPU Converged 网（南北向）**：承载存储 + 带内管理 + 客户网络连接。每 tray 1 × BF-3 B3240 双 400G 口；节点存储带宽约 40 GB/s。管理/控制节点（x86 或 Grace 服务器）则用 4 × ConnectX-7 200G 单口卡接入这张网——CX-7 与 CX-8 在同一集群共存。
3. **OOB 管理网**：1 GbE（SN2201，48 口），连接所有 BMC 口、交换机管理口、DPU 管理口，与业务面物理隔离，供基础设施管理使用。

分离的理由：三类流量的服务质量目标互相冲突——训练流量要零丢包低延迟，存储流量要大块吞吐，管理流量要永远可达。混在一张网上，任何一类流量的异常都可能波及其他两类；物理分离把故障域和拥塞域各自关进笼子。

管理与支持节点的配置也值得一提（官方 RA 文档附录 B）：集群中的管理/控制节点（x86 或 Grace 服务器）配备 2×960 GB M.2/SATA SSD（RAID1）作启动盘、2×7.68 TB NVMe 作数据缓存盘，网络侧用 **4×ConnectX-7 200G 单口卡**接入存储/带内网络，配 TPM 2.0 与支持安全启动的 BMC（Redfish 1.4+）。这解释了为什么一个 CX-8 主导的集群里仍然大量存在 CX-7——CX-7 在南北向 200G 场景下依然是主力，两代网卡在同一集群长期共存。这一点对理解第五部分"CX7 上有这个功能"也很关键：CX-7 没有退役，它就在 NVL72 集群的南北向网络里跑着。

3.7 双平面（Dual Plane）设计的动机

东西向计算 fabric 采用**双平面拓扑**：建两张一模一样的平面（Plane 0 / Plane 1），每颗 CX-8 的 800G 拆成 2 × 400G，各接一个平面。动机有三：

- **消除单点故障**：任一平面的 leaf/spine 故障，只损失一半带宽，作业不中断（吞吐降级而不是任务失败）；
- **负载均衡**：NCCL 知道双平面拓扑，会把集合通信的流量在两个平面间均衡打散；
- **扩大 NIC 与交换机的 fan-out**：更高 radix 的接口让 Clos 拓扑用更少的层级覆盖更多端点，降低网络成本和跳数。

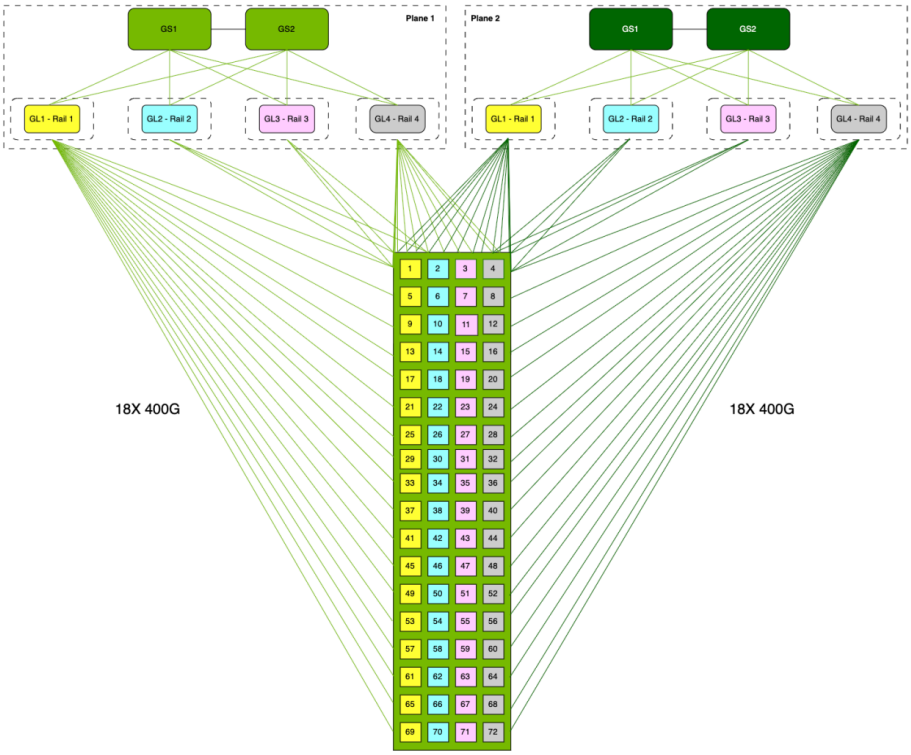
代价是光模块和线缆数量翻倍，以及调度复杂度上升——但对于一个单柜 142 kW、按 MW 级投资的 AI 工厂，这是值得的保险。

3.8 扩展拓扑：从 1 柜到 8 柜

官方 RA 文档以 1 个机柜为 1 个 **SU（Scalable Unit，可扩展单元）** = 18 tray = 72 GPU，给出 2/4/8 柜的标准扩展方案。

单柜（1 SU = 72 GPU）：柜内 72 GPU 已在一个 NVLink 域内；对外每 GPU 一颗 CX-8，800G 拆 2 × 400G 分别进双平面的 4 个 leaf。leaf 按"GPU 位置"组织：每个 leaf 连接所有 tray 上同一位置的 GPU，形成 4 条 rail × 2 平面。

Figure 10 Example diagram of 18-Tray SU connected to 2 planes (4 rails each) of the compute fabric



The single rack scalable unit provides the following connectivity building blocks:
Within the rack: As shown in

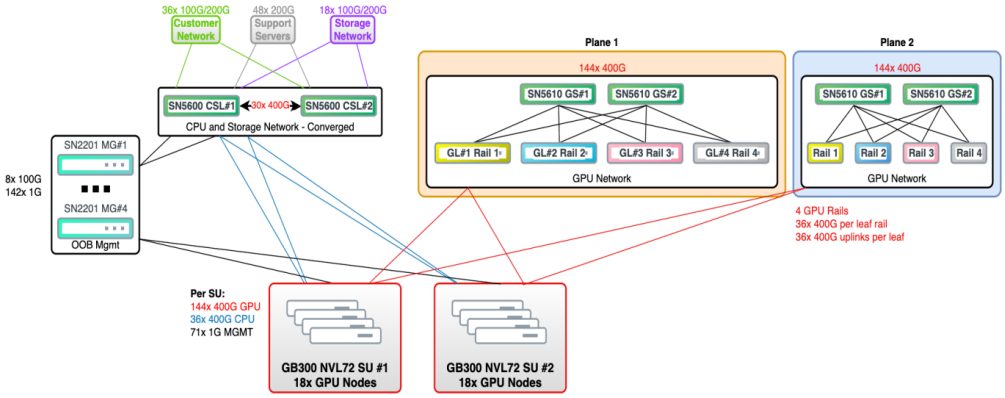
- Figure 10 all 72 GPUs are interconnected in a single NVLink domain, allowing them to function as a single multi-GPU unit of compute with a bandwidth of 900GB/s (1800 GB/s bi-directional)
- For the Compute (East/West) fabric: 18 trays, each with 4 x single-port NVIDIA ConnectX-8 NICs and a total aggregate bandwidth of 3200 Gb/s
- For the Converged (North/South) fabric: 18 trays, each with 1x B3240 DPU providing 2x 400Gb/s connections and a total aggregate bandwidth of 800 Gb/s
- For the Out-of-band Management fabric, 18 trays, each with 3x 1Gb/s connections providing 54 x 1Gb/s for management

图：18-Tray SU 连接双平面（每平面 4 条 rail）计算 fabric 的示例拓扑（出自 NVIDIA 官方 RA 文档第 21 页 Figure 10）。

2 柜（144 GPU）：2 个 SU、36 个节点。每平面 4 个 leaf，leaf 接线仍按 GPU 位置 rail 对齐；每柜另需 2 台 SN5600 承担 CPU/存储流量。spine-leaf 无阻塞 fat-tree。

2 Racks, 36 Trays with144 Blackwell Ultra GPUs

Figure 11 Two Rack GB300 NVL72 with SpectrumX Enterprise RA (144 GPUs)



Architecture Overview

- Two GB300 NVL72 scalable units (SUs) composed of 36 server nodes total (18 nodes per SU)
- 144 NVIDIA Blackwell GPUs connected through the Spectrum-X compute fabric
- Each rack requires 2x SN5600 switches for CPU and storage connectivity and up to 12x SN5600 switches for the dual-plane GPU network
- Infrastructure supports scaling from one to two SUs; leaf switches are intentionally underpopulated to align with GB300 NVL72 software design criteria

Network Design

- Cost-optimized collapsed spine-leaf architecture for all required fabrics
- Spectrum fabric implemented as a rail-optimized, non-blocking 64-port switch design
- Port breakout used to consolidate connections without reducing resiliency
- Each leaf switch is capable of serving three racks but is configured for two to maintain balance and performance

Connectivity (Under Optimal Conditions)

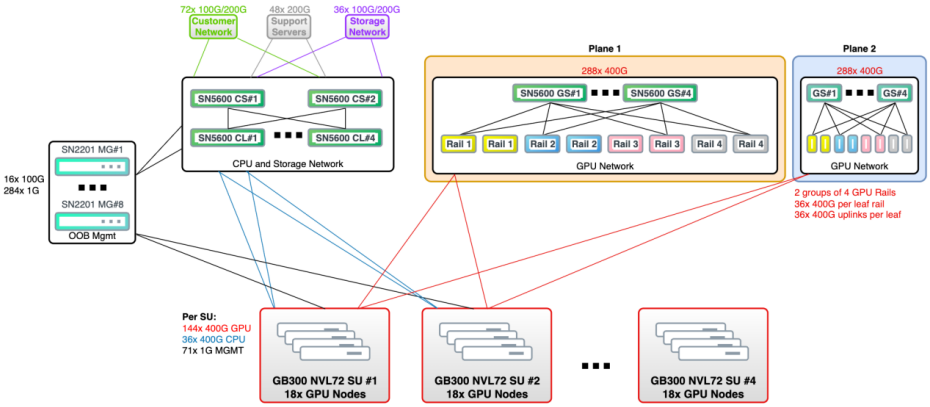
- 144x 400G uplinks per SU for GPU Compute traffic
- 36x 400G uplinks per SU for CPU fabric
- Up to 12 support server nodes, quad-connected at 200G

NVIDIA Enterprise RA 2026 March | 25

图：双机柜 GB300 NVL72 与 Spectrum-X 企业 RA 拓扑，144 颗 GPU（出自 NVIDIA 官方 RA 文档第 25 页 Figure 11）。

4 柜（288 GPU）：4 个 SU、72 个节点、288 颗 GPU。网络按 3~4 个 SU 规模预置，保持每 SU 18 tray 的一致粒度，leaf 仍 rail-optimized，双平面提供高可用与负载均衡。

Figure 12 Four Rack GB300 NVL72 with Spectrum-X Enterprise RA (288 GPUs)



Architecture Overview

- Four GB300 NVL72 scalable units (SUs) with 72 trays/nodes and 288 Blackwell Ultra GPUs interconnected via the Spectrum-X compute fabric
- Networking sized to scale between three and four SUs while keeping a consistent per-SU node count of 18 trays
- Design references switch, transceiver, and cable requirements as summarized in the associated bill-of-materials tables (Table 6 and Table 7)

Network Design

- Cost-efficient collapsed spine-leaf architecture for all fabrics except the GPU Compute (E-W) network
- GPU Compute (E-W) fabric implemented as a separate, isolated leaf-spine network to deliver high-bandwidth, low-latency node-to-node communication
- Port breakout used wherever possible to consolidate ports without compromising resiliency
- Leaf switches can support up to three racks but are intentionally underpopulated to serve only two racks to satisfy GB300 NVL72 software requirements

Connectivity (Under Optimal Conditions)

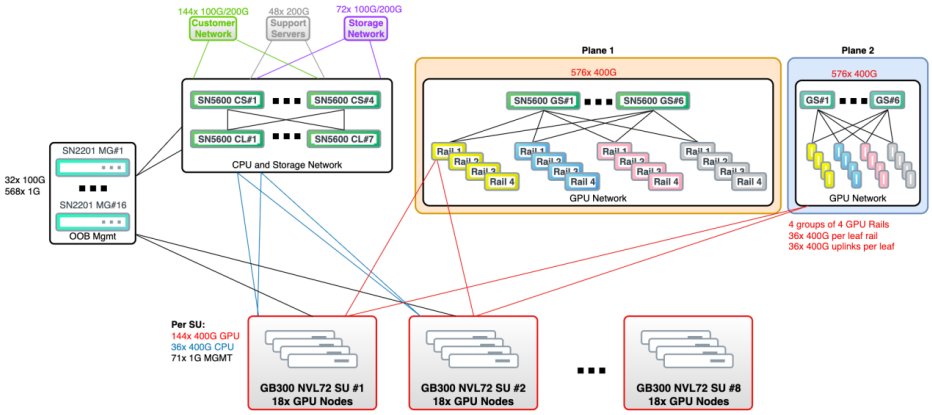
- 144x 400G uplinks per SU for GPU Compute traffic
- 36x 400G uplinks per SU for CPU traffic
- Up to 12 support server nodes, all quad-connected at 200Gb
- 72x 100G/200G connections toward the customer network (minimum 25Gb bandwidth per GPU)
- 36x 100G/200G storage connections (minimum 12.5Gb bandwidth per GPU)

图：四机柜 GB300 NVL72 与 Spectrum-X 拓扑，288 颗 GPU（出自 NVIDIA 官方 RA 文档第 27 页 Figure 12）。

8 柜（576 GPU）：8 个 SU、144 个节点、576 颗 GPU。规模超过单层 spine 的端口容量后，引入 **super-spine** 层，构成三级 Clos。客户网络方向要求每 GPU 至少 25 Gb、存储方向至少 12.5 Gb 的最低带宽。

8 Racks, 144 Trays with 576 Blackwell Ultra GPUs

Figure 13 Eight Rack GB300 NVL72 with Spectrum-X Enterprise RA (576 GPUs)



Architecture Overview

- Eight GB300 NVL72 scalable units (SUs) with 144 trays/nodes and 576 Blackwell GPUs interconnected via Spectrum-X compute fabric.
- Networking sized to scale between seven and eight SUs while keeping 18 trays per SU.
- Switch, transceiver, and cable requirements are defined in the associated bill-of-materials tables (Table 6 and Table 7).

Network Design

- Cost-efficient spine-leaf architecture for all fabrics except the GPU Compute (E-W) fabric.
- GPU Compute (E-W) network implemented as a separate, isolated leaf-spine fabric to deliver high-bandwidth, low-latency node-to-node communication.
- For very large deployments (around 1024 nodes and above), a super-spine layer is introduced to maintain non-blocking point-to-point connectivity.
- Port breakout is used wherever possible to consolidate ports while preserving resiliency.
- Each leaf switch can support up to three racks but is intentionally underpopulated to serve only two racks to meet GB300 NVL72 software requirements.

Connectivity (Under Optimal Conditions)

- 144x 400G uplinks per SU for GPU Compute traffic.

NVIDIA Enterprise RA 2026 March | 29

图：八机柜 GB300 NVL72 与 Spectrum-X 拓扑，576 颗 GPU，引入 super-spine 层（出自 NVIDIA 官方 RA 文档第 29 页 Figure 13）。

3.9 小结

- GB300 NVL72 = 72×B300 + 36×Grace，18 计算 tray + 9 NVSwitch tray，液冷约 142 kW，柜内 NVLink 域聚合 130 TB/s。

- 计算 tray 的关键设计是 GPU:NIC 1:1（4×CX-8）与 DPU 独立承担南北向（1×BF-3）。
- 三张物理分离 fabric：东西向计算网、南北向 converged 网、OOB 管理网，各自隔离故障域与拥塞域。
- 双平面设计消除单点故障并做负载均衡；CX-8 的 800G 拆 2×400G 各进一个平面。
- 官方扩展方案：1 SU = 1 柜 = 72 GPU，标准 2/4/8 柜（144/288/576 GPU），8 柜引入 super-spine。

3.10 思考题 / FAQ

Q1：为什么 CX-8 要做 1:1 而不是 2 颗 GPU 共享 1 颗 NIC?
共享会让两颗 GPU 的通信流量在同一 NIC 队列里互相干扰，尾延迟不可预测；1:1 rail 让 NCCL 可以把流量确定性钉扎，拥塞定位也更简单。

Q2：双平面是不是等于"带宽翻倍"?
不是。单颗 CX-8 的物理上限是 800G，拆成 2×400G 后总带宽不变；翻倍的是路径冗余和故障域隔离能力。

Q3：为什么 OOB 管理网要物理隔离，不能在业务网上划 VLAN?
当业务网本身出故障（比如 PFC 风暴打瘫了 fabric）时，你还需要一条一定能通的路径去登录设备排查——这就是 OOB 存在的意义。

第四部分 无损网络深挖

这一部分是全文的枢纽：NVIDIA 那段反馈里的 ECN 和 PFC，都是无损以太网的核心机制。我们先把它们彻底讲清楚。

4.1 为什么 RoCE 需要无损网络

回顾第一部分：RoCEv2 把 RDMA 封装在 UDP/IP 上跑在标准以太网里，而以太网交换机 buffer 满了就丢包。对 TCP 业务这没关系（重传即可），但对 RDMA 是灾难：

- RoCEv2 网卡的重传机制是 **go-back-N**：丢一个序号 N 的包，就要把 N 之后已发的所有包全部重发，带宽利用率瞬间崩塌；
- AI 集合通信是**同步**的：一次 AllReduce 里任何一个流被打断，几千颗 GPU 一起等——一次丢包放大成一次全集群抖动；
- 重传引起的延迟尖峰（tail latency）比平均带宽下降更致命，因为 step 时间由最慢的那次通信决定。

结论：跑 RoCEv2 的以太网必须被改造成"无损"——宁可不发，也不丢包。实现无损有两层机制：**PFC 兜底**和 **ECN+DCQCN 主控**。

4.2 PFC：逐跳反压的最后防线

PFC（Priority Flow Control，IEEE 802.1Qbb） 是二层逐跳（hop-by-hop）反压机制。工作过程：

1. 交换机的每个端口按优先级（PCP/CoS）划分独立队列，各自有 buffer；
2. 当某优先级队列的缓存占用超过阈值，交换机立刻向**上游直连设备**发送 PAUSE 帧；
3. 上游收到 PAUSE 后暂停发送该优先级的流量（其他优先级不受影响）；
4. 队列水位回落后发送 RESUME（或 PAUSE 计时器到期），恢复发送。

PFC 的定位是**最后防线（backstop）**：它防住 buffer 溢出丢包，把以太网链路变成"无损链路"，但它完全不动拥塞的根因。它的缺陷在生产环境中臭名昭著：

- **PFC 风暴（pause storm）**：反压一跳跳向上游扩散，形成"拥塞树"，一片区域的端口互相 PAUSE，把本来不拥塞的邻居也拖下水；
- **队头阻塞（HoL blocking）**：被 PAUSE 的队列排在端口最前，后面去其他不拥塞方向的流量也被堵住（同优先级内）；
- **公平性差**：长流挤压短流，incast 场景下各流速率失衡；
- **死锁风险**：PFC 是无条件反压，配置不当（尤其存在环路或多层无损优先级映射错误）可能全网互相等待、永久停转。

一句话：PFC 治标（不丢包）不治本（不消拥塞），健康网络应该"绝大部分时间运行在 ECN 反馈上，很少触发 PFC"。

一个 PFC 风暴的具体场景，帮助建立直觉：某台存储服务器向 32 个训练节点同时推数据（incast），其中一个节点 A 的接收链路瞬时拥塞 → 连接 A 的 leaf 交换机队列超限 → leaf 向 spine 发 PAUSE → spine 队列被这些"去 A 的流量"堵住 → spine 再向其他 leaf 发 PAUSE → 去往节点 B、C、D 的正常训练流量在 spine 里被队头阻塞 → 训练集合通信超时抖动。拥塞源头只有一个端口，受害面却蔓延整张 fabric——这就是"反压沿拥塞树扩散"的含义。PFC 的 PAUSE 不区分"谁是罪魁"，它只认队列。

4.3 ECN 与 DCQCN：端到端的主控机制

ECN（Explicit Congestion Notification，显式拥塞通知） 复用 IP 头 DS 字段的最后 2 bit（ECT/CE）。交换机在启用 ECN 的队列上按类 WRED 曲线打标：

- 队列深度 < Kmin：不标记；
- Kmin ~ Kmax 之间：按概率把报文的 CE 位置位（随深度增加概率上升）；
- 队列深度 > Kmax：全部标记。

注意：标记不是丢弃，报文正常转发，只是"贴了张便签"告诉接收端"路上开始堵了"。

DCQCN 是把 ECN 信号转化为发送端调速的闭环协议，分三段：

1. **拥塞点 CP（交换机）**：队列超 ECN 阈值时给报文打 CE 标记，正常转发不丢包。
2. **通知点 NP（接收端网卡）**：收到 CE 报文后，向发送端回一个 **CNP（Congestion Notification Packet）**——这是 RoCEv2 新增的控制报文，约每 50 μs 每流最多发一个。
3. **响应点 RP（发送端网卡）**：收到 CNP 后按 α 参数**乘性降速**：RC = RC × (1 − α/2)；随后按时间/字节阈值**加性增速**恢复（类似 TCP 的 AIMD，但在网卡硬件内完成）。

这套闭环在队列还**没**涨到 PFC 门限之前就开始压低发送速率，从源头消除持续性拥塞——它是主控，PFC 只处理瞬时的微突发。

DCQCN 的关键参数与调优复杂度。 α 是降速因子（越大降得越狠），g 是恢复增益，还有快速恢复的时间阈值、加性增速的字节计数器、CNP 最小间隔（约 50 μs/流）等十余个参数。参数之间强耦合：α 太小压不住拥塞、太大则吞吐锯齿严重；Kmin/Kmax 设低了有效吞吐上不去，设高了 PFC 抢先生效。这些参数没有普适最优值，随流量模式（AllReduce 的周期性突发 vs All-to-All 的均匀流）、RTT、buffer 大小而变——这就是为什么"无损以太网调优"在 AI 集群运维里是一门显学，也是 AI-ECN/FastECN 这类自动调参方案出现的原因。

4.4 PFC 与 ECN 的门限配合

两种机制要能协同，阈值对齐是关键中的关键：

- **ECN 的 Kmax 必须低于 PFC 触发门限**，且中间要留足 headroom（见 4.5）。否则队列先撞 PFC 门限、PAUSE 先于 ECN 生效，DCQCN 还没来得及降速网络就进入反压状态，退化成"PFC 主导"的不

健康模式。

- **CNP 报文所在的队列绝不可开 PFC**：CNP 是拥塞反馈的性命线，如果它被 PAUSE 住，发送端收不到降速信号、继续猛发，进一步触发更多 PFC，形成正反馈乃至死锁。
- 生产惯例：RoCE 业务流 DSCP 26 → CoS 3（该队列开 PFC + ECN），CNP 报文 DSCP 48 → CoS 7（高优先级、不开 PFC）；交换机端口 trust dscp；全网 DSCP→CoS 映射必须一致，任何一台设备配错都可能把整张 fabric 拖入异常。
- 前沿演进：AI-ECN / FastECN 等方案用交换机遥测数据动态调整 ECN 阈值，降低 DCQCN 十余个参数的手工调优复杂度。

4.4.1 无损网络部署检查清单（工程实务）

把本节内容落成一张可操作的核对表，供真实部署参考：

1. **队列规划**：RoCE 业务流走 CoS 3（DSCP 26），CNP 走 CoS 7（DSCP 48），管理流量走低优先级默认队列；
2. **PFC 只在业务队列开**：CoS 3 开 PFC，CoS 7 及其他队列绝不开；
3. **ECN 门限**：Kmin/Kmax 设在 CoS 3 队列上，Kmax 明显低于 PFC 触发门限；
4. **buffer 预算**：PFC 门限 + headroom（含链路传播与线缆长度）< 端口总 buffer；800G 长距链路要重新核算 headroom；
5. **全网一致**：所有交换机端口 trust dscp，DSCP→CoS 映射表逐台核对，一台配错全网约等于裸奔；
6. **监控基线**：PFC PAUSE 发送/接收计数、ECN 标记率、队列水位、NIC 丢弃计数纳入日常看板；PFC 持续触发即告警；
7. **死锁防范**：避免无损优先级跨多跳成环， Spine/Leaf 升级维护时先排空无损队列流量。

4.5 Headroom buffer

PAUSE 帧从发出到生效有传播时延，这段时间里上游还在按线速往链路里灌数据。交换机的 **headroom buffer** 就是预留来吸收这段"飞行中数据"的空间。链路越长、速率越高（800G），所需 headroom 越大。headroom 与 ECN Kmax、PFC 门限三者的关系是 buffer 规划的核心：PFC 门限 + headroom 必须小于端口总 buffer，而 ECN Kmax 又要明显低于 PFC 门限——一张 800G AI fabric 的 buffer 预算是被这三条线挤着画的。

4.6 Spectrum-X：自适应路由、spraying 与 ECN 的关系

传统以太网用 ECMP 哈希分流：一条流钉死在一条路径上，哈希不均就产生"热点路径"。Spectrum-X 的做法不同：

- **自适应路由 / packet spraying**：以报文（而非流）为粒度在多路径间打散，并根据各路径实时拥塞遥测动态选择出口，天然消除热点、接近理论无阻塞吞吐；
- 副作用是报文乱序到达，由 CX-8 SuperNIC 在端侧做硬件乱序重组（RoCE 的 DD/OOO 能力），对上层不可见；
- **与 ECN 的关系**：spraying 解决的是"负载不均衡引起的局部拥塞"，ECN/DCQCN 解决的是"输入总量超过输出能力引起的真拥塞"。两者互补：spraying 让队列水位分布均匀，使 ECN 标记更准确地反映真实拥塞；ECN 标记又是 Spectrum-X 遥测与路由调整的输入之一。可以说 Spectrum-X 把"避堵"（路由层）和"降速"（端侧拥塞控制层）做成了一个联动系统。

4.7 小结

- RoCEv2 对丢包零容忍，无损以太网是 AI 以太网 fabric 的前提。
- PFC 是二层逐跳反压兜底，防丢包但有风暴、HoL 阻塞、死锁风险；ECN+CNP+DCQCN 是端到端主控，在 PFC 触发前就让发送端降速。
- 门限铁律：ECN Kmax < PFC 门限，中间留 headroom；CNP 队列不可开 PFC；惯例 DSCP26→CoS3、DSCP48→CoS7。
- Spectrum-X 的 spraying 解决负载不均，ECN/DCQCN 解决真实拥塞，二者联动。

4.8 思考题 / FAQ

Q1：既然有 ECN/DCQCN，为什么还要 PFC?
DCQCN 的反馈回路需要至少一个 RTT 才能生效，对微突发（incast 瞬时洪峰）来不及反应；PFC 是零延迟的本地反压，负责兜住这类瞬时溢出。

Q2：为什么 CNP 队列开 PFC 会导致死锁?
CNP 被 PAUSE → 发送端收不到降速信号 → 继续高速发送 → 更多队列触发 PFC → CNP 被更久地堵住——正反馈锁死。所以 CNP 走高优先级且绝不开 PFC 的队列（惯例 DSCP48→CoS7）。

Q3：headroom 不够会发生什么?
PAUSE 生效前的在途数据无处安放，buffer 溢出丢包——无损域破功，go-back-N 重传开始，性能雪崩。

第五部分 逐词拆解那段 NVIDIA 反馈

现在回到引言里那段话，逐词拆解。原文：

> "网卡标记 ECN 解决网络到 Host Memory 因为 PCIe Gen5 导致 PFC 的问题。这个功能目前在 CX8 网卡上还没有实现，CX7 上有但没啥用处。正在和网卡架构团队讨论实现需要的资源。"

5.1 "网卡标记 ECN"指什么

第四部分讲过，标准的 ECN 标记发生在**交换机**上（CP 点）：队列水位超阈值就给报文打 CE 标。但这里说的是**网卡标记 ECN**——把标记点从交换机移到了网卡。

为什么需要在网卡上标记？因为有一种拥塞根本不发生在交换机的队列里，而是发生在**网卡内部到主机的路径上**。交换机的队列干干净净，ECN 自然一个标都不打；但数据在网卡内部的 buffer 里已经堆积如山。标准 ECN 对这种"网络之外的拥塞"是盲的。让网卡自己对经过的报文（或对内部状态）做 ECN 标记，等于把拥塞信号体系延伸到了网卡—主机段：网卡发现自己往 host 方向送不动了，就主动打标，触发 DCQCN 让对端发送端降速，从源头减压。

5.2 "网络到 Host Memory 因为 PCIe Gen5 导致 PFC 的问题"：拥塞跨域传导链

这是整句话的核心，也是最反直觉的部分：PFC 是以太网二层机制，PCIe 是主机内部总线，两者怎么纠缠到一起的?

先明确分层（第一部分 1.4 已埋下伏笔）：

- **PFC 管"网线到 NIC"**：以太网链路段，NIC—交换机之间；
- **PCIe 流控管"NIC 到内存/HBM"**：主机总线段，PCIe 靠 credit-based flow control；
- 两段物理上不相交，但在 GDR 数据路径上首尾相接，**通过 NIC 内部 buffer 耦合**。

先看带宽账，这是"PCIe Gen5 导致问题"的定量根源：

- CX-7/CX-8 世代网卡的网络侧速率是 400G/800G，即约 50/100 GB/s 单向；

- PCIe Gen5 x16 单向约 63 GB/s，双向合计约 126 GB/s；
- 400G 网卡插在 Gen5 x16 上基本刚好喂饱；**800G 网卡（CX-8）插在 Gen5 x16 上，主机侧带宽（约 63 GB/s 单向）喂不饱网络侧（约 100 GB/s 单向）**——网卡能收进来，却送不进主机。这正是 CX-8 要用 PCIe Gen6 x16（单向约 128 GB/s）的原因。

一个更具体的数字推演：假设某 CX-8 以 800G 线速收包（100 GB/s），而它所在的 PCIe Gen5 x16 链路实际可用 DMA 写带宽只有约 55–60 GB/s（还要扣掉 PCIe TLP 开销与共享流量）。那么 NIC 内部 buffer 以每秒约 40 GB 的速度净堆积——NIC 片上+外挂的可弹性缓冲通常只有几十 MB 量级，**不到 2 毫秒 buffer 就会打满**。2 毫秒在网络世界是很长的时间（足以发生几百次 ECN 反馈），但前提是有人打标；标准 ECN 标记点在交换机，交换机此时可能毫无压力，于是 DCQCN 静默，NIC buffer 打满后直接反压链路，PFC 登场。这个推演说明：在这种瓶颈结构里，**唯一能在正确时间发出正确信号的设备就是 NIC 自己**——这就是"网卡标记 ECN"的工程必然性。

现在画出完整的**拥塞传导链**（以接收方向为例，网络 → host memory）：

```
远端发送端 → [以太网链路] → 交换机 egress 队列 → NIC  ingress
                                  ↓
                            NIC 内部 buffer
                                  ↓
                        PCIe Gen5 x16（带宽不足成为瓶颈）
                                  ↓
                          Host Memory / GPU HBM
```

当 PCIe 段喂不饱（或 host 侧其他原因拖慢 DMA）时，链条倒着传导反压：

1. PCIe 段成为瓶颈，NIC 的 DMA 写请求拿不到足够的 PCIe credit，写不动；
2. NIC 内部 buffer 开始堆积（收得快、送得慢）；
3. NIC buffer 占满后，NIC 无法再从线上收包，其 ingress 侧反压到对端交换机；
4. 交换机 egress 队列水位上涨，触发 **PFC PAUSE** 帧回压上一跳；
5. PFC 反压沿拥塞树向上游扩散——一场**根源在主机 PCIe 段**的拥塞，最终以以太网 PFC 风暴的形式表现在网络上。

这就是"因为 PCIe Gen5 导致 PFC"的完整因果：瓶颈在 host 总线段，症状（PFC 触发）却在以太网链路段。由于标准 ECN 只在交换机队列打标，而交换机队列在初期甚至中期都可能不超限（反压是从 NIC 侧"顶回来"的），DCQCN 对这种拥塞反应迟钝甚至无感，网络只能一路滑向 PFC——最不健康的那种运行模式。

"网卡标记 ECN"就是针对这条链的解药：NIC 第一时间知道自己往 host 送不动（它的内部 buffer 水位自己最清楚），由 NIC 直接对相关报文打 CE 标，接收端 NP 回 CNP、发送端 RP 乘性降速——在 NIC buffer 打满、PFC 触发**之前**就把注入速率压下来。本质上是把 DCQCN 的感知范围从"交换机队列"扩展到"交换机队列 + NIC 内部队列"，让端到端拥塞控制覆盖到 PCIe 瓶颈这一段。

同理，GDR 路径（NIC → PCIe → GPU HBM）与 host memory 路径（NIC → PCIe → CPU DDR，比如存储流量经 BlueField-3）共享同一条 PCIe 资源；NUMA 亲和性（NIC 挂在哪颗 CPU 的 root complex 上）也会影响实际可达带宽。所以"网络到 Host Memory 路径"的拥塞，既是带宽问题也是拓扑问题。

5.3 为什么"CX7 上有但没啥用处"

CX-7 是 400G 网卡，主机接口 PCIe Gen5 x16（单向约 63 GB/s），刚好喂饱 400G（约 50 GB/s）。也就是说在 CX-7 的规格组合里，**PCIe 段不是瓶颈**：网络侧送进来多少，host 侧基本都能吃下去。网卡标记 ECN 这个功能在 CX-7 上虽然存在，但它要解决的"PCIe 瓶颈引发的 NIC 内部拥塞"在 CX-7 的正常部署中几乎不会出现——所以"有但没啥用处"。这不是说功能坏了，而是说它在一个不构成瓶颈的平台上没

有用武之地。

5.4 为什么 CX8（Gen6）反而还没实现

表面上看很矛盾：CX-8 是更新的旗舰（800G、PCIe Gen6 x16 AI SuperNIC），怎么反而缺功能？拆解一下：

- **CX-8 是全新芯片、全新 datapath**。800G 数据面、可编程拥塞控制、Spectrum-X 遥测联动都是新设计的硬件/固件；"网卡标记 ECN"这种需要在 NIC 内部 buffer 管理逻辑里新增标记点的功能，要在新架构上重新实现、验证，需要独立的开发资源——这正是反馈最后一句"正在和网卡架构团队讨论实现需要的资源"的含义：功能有价值，但要排期、要面积、要验证预算。
- **Gen6 缓解了瓶颈但没有消灭它**。Gen6 x16 单向约 128 GB/s，理论上喂得饱 800G（100 GB/s）；但实际部署中 NIC 可能插在 Gen5 的 root port 上（老平台、或 Mezzanine 走线约束），或者 host memory 路径还要和 GDR、DPU 流量共享 PCIe 资源，瓶颈只是被推高而没有消失。所以这个功能在 CX-8 上依然必要，甚至因为 800G 的流量绝对值更大而更必要。

5.5 对 NVL72 部署的实际影响与缓解手段

把这个功能缺口映射回 GB300 NVL72 的具体架构：

- NVL72 计算 tray 上 CX-8 与 GPU 1:1，走 PCIe Gen6 x16，纯 GPU 计算流量（GDR 到 HBM）路径的 host 侧瓶颈风险相对较低；
- 但 **host memory 路径**（南北向存储/管理流量、Grace 内存池上的流量）共享 host 侧 I/O 资源，且跨 rack 扩展后存储流量规模上升；一旦 host 侧消化不及，反压照样经 NIC buffer 传导到计算 fabric，以 PFC 形式爆发，拖垮训练流量；
- 在"网卡标记 ECN"落地之前，可操作的缓解手段：
- **门限调优**：压低计算 fabric 上 ECN 的 Kmin/Kmax，让交换机更早打标，为 NIC 侧可能积聚的压力预留反应余量；确保 Kmax < PFC 门限且 headroom 充足；
- **隔离**：严守三网物理分离，存储/管理流量绝不进入计算 fabric（NVL72 的三 fabric 设计天然提供了这道隔离）；
- **限速与调度**：对 host memory 方向的流量（如 checkpoint 写入）做应用层限速，避开训练通信高峰期；
- **监控**：盯 NIC 内部 buffer/PCIe 利用率与 PFC PAUSE 计数，PFC 触发率持续非零即为"PCIe 瓶颈在传导"的预警信号；
- **拓扑核对**：确认 NIC 的 NUMA 亲和性，避免跨 root complex 访问 host memory 进一步压缩有效 PCIe 带宽。

5.6 小结

- "网卡标记 ECN"= 把 ECN 标记点从交换机扩展到网卡，感知网卡—主机段的拥塞；
- PCIe Gen5 x16（约 63 GB/s 单向）喂不饱 800G 网卡（约 100 GB/s），host 侧瓶颈使反压经 NIC buffer 跨域传导，最终以 PFC 形式爆发——这就是"因为 PCIe Gen5 导致 PFC"；
- CX-7（400G + Gen5）host 侧不构成瓶颈，所以该功能"有但没啥用"；CX-8 是新架构，该功能需重新实现，正在向网卡架构团队要资源；
- 落地前的缓解：压低 ECN 门限、三网隔离、应用层限速、监控 PFC 计数、核对 NUMA。

5.7 思考题 / FAQ

Q1：为什么不能让交换机直接感知 NIC 的 buffer 水位?
交换机只能看到自己的队列。NIC 内部状态需要通过带内遥测或协议扩展传递，"网卡自己打 ECN 标"正是利用现有 ECN 语义、无需新协议的最小改动方案。

Q2：把 PCIe 全换成 Gen6 是不是就不需要这个功能了?
缓解但不消灭：老平台混插、Mezzanine 走线约束、GDR 与 host memory 流量共享 PCIe 资源，都可能让有效带宽低于标称。拥塞控制的价值恰恰在"最后一公里"的意外处。

Q3：PFC 触发了是不是就说明网络配置错了?
不一定。偶发 PFC 是兜底机制正常工作；持续、大面积的 PFC 才说明主控（ECN/DCQCN）没拦住——按本文分析，这时要怀疑的方向不止交换机队列，还包括 host 侧 PCIe/内存路径。

第六部分 扩展概念

6.1 与 InfiniBand（Quantum）方案对比

同一颗 CX-8 ASIC 既能跑以太（800GbE）也能跑 IB（XDR/NDR）；NVIDIA 的 IB 方案对应 Quantum-2（NDR 400G）/ Quantum-X800（XDR 800G）交换机。对比：

维度	Spectrum-X 以太网（RoCEv2）	InfiniBand（Quantum）
无损机制	PFC + ECN/DCQCN（需精细调参）	原生 credit 流控，天然无损
拥塞控制	端侧 DCQCN + Spectrum-X 遥测联动	交换机硬件内建，历史更成熟
生态	开放以太网生态、供应链多元	NVIDIA 封闭全家桶，即插即用
运维	需要以太网无损调优经验	需要 IB 专网运维经验
典型规模	云厂商超大集群偏爱以太	传统超算/精专集群偏爱 IB

GB300 NVL72 官方 RA 选择 Spectrum-X 以太方案，面向"企业 AI 工厂"的开放生态定位；NVIDIA 内部超大训练集群则两条路线都有使用。

6.2 SHARP / 网内计算

SHARP（Scalable Hierarchical Aggregation and Reduction Protocol） 把 AllReduce 的归约计算卸载到交换机 ASIC 内：梯度在流经交换机时就地完成求和，既减少网络传输量又缩短归约延迟。CX-7 在 IB 模式支持 SHARPv3，CX-8 支持 SHARPv4；以太模式则依赖 Spectrum-X 与 NCCL 的协同优化。网内计算代表了一个趋势：交换机不再只是"搬运工"，而是计算的参与者。

6.3 NVL72 vs 单机 8 卡 HGX 的取舍

HGX（如 HGX H100/B200，单机 8 卡）是上一代标准形态：8 颗 GPU 经 NVSwitch 全互联，节点间靠 IB/以太。NVL72 把 Scale-up 域从 8 卡扩到 72 卡，取舍如下：

- NVL72 优势：宽 TP/EP 可全部留在 1.8 TB/s 的 NVLink 域内；大模型单实例显存池 72×288 GB；机柜级液冷与供电密度更优；
- NVL72 代价：单点故障域变大（一个机柜的事故事关 72 颗 GPU）；调度要求整柜粒度；采购与部署门槛（约 142 kW/柜、液冷机房）远高于 HGX；
- 经验法则：超大模型训练/推理（万亿参数级）选 NVL72；中等规模、异构负载、机房条件受限选 HGX 集群。

6.4 CPO / 硅光

800G 以上速率下，可插拔光模块的功耗与面板密度成为瓶颈（NVL72 双平面设计使光模块数量翻倍，问题更尖锐）。**CPO（Co-Packaged Optics，共封装光学）** 把光引擎与交换 ASIC 封在同一基板上，省掉 DSP 与长电走线，显著降低每比特功耗。NVIDIA 已在 Spectrum-X Photonics / Quantum-X Photonics 系列交换机上推进硅光 CPO 方案，目标是让万卡级 fabric 的光互连功耗下降一个量级。（本小节为公开行业信息，非官方 RA 文档内容。）

6.5 下一代 Rubin 展望

按 NVIDIA 公开路线图，Blackwell 之后是 **Rubin** 平台（预计 2026 年起）：Rubin GPU + Vera CPU，NVLink 代际继续演进（第六代，带宽预计翻倍至 3.6 TB/s 级），配 Kyber 机柜形态（NVL576 级），网络侧向 1.6T 以太网与更深度 CPO 演进。对本文读者而言，值得关注的延续线是：Scale-up 域继续扩大、host 侧 I/O（PCIe Gen7/CXL）与网络侧速率的"带宽赛跑"仍将继续——第五部分那条拥塞传导链的故事，在下一代平台上只会更重要。（本小节为公开信息，非本文档主要来源。）

6.6 小结

以太（Spectrum-X/RoCEv2）与 IB（Quantum）是两条并行的 AI fabric 路线，取舍在生态开放性与调优复杂度之间；SHARP 把归约计算搬进交换机；NVL72 与 HGX 的取舍核心是 Scale-up 域大小与故障/部署成本；CPO 与 Rubin 指向光互连功耗与更大 NVLink 域的下一代战场。

6.7 思考题 / FAQ

Q1：为什么云厂商偏爱以太而传统超算偏爱 IB?
云厂商有强大的网络工程团队、要多元供应链、规模足以摊薄调优成本；传统超算更看重开箱即用的确定性能与成熟的拥塞控制。

Q2：SHARP 为什么在以太模式不如 IB 模式成熟?
IB 的交换机与 HCA 是同一套封闭协议栈，归约卸载可以端到端硬编码；以太生态开放、设备异构，网内计算需要更复杂的协同（如与 NCCL/遥测联动），落地更渐进。

附录

附录 A 术语表（中英对照）

术语	英文全称	一句话解释
AI Factory	AI Factory	以 token 产出为目标、按工厂流水线理念建设的 AI 数据中心
AllReduce	—	集合通信原语：所有节点梯度求和后广播回所有节点
All-to-All	—	集合通信原语：每对节点互换数据，EP 并行的核心
B300	Blackwell Ultra GPU	GB300 NVL72 采用的 Blackwell Ultra 代 GPU
Backpressure	反压	下游来不及处理时向上游发出的"暂停"压力
BMC	Baseboard Management Controller	服务器带外管理控制器，独立于主 CPU
C2C / NVLink-C2C	NVLink Chip-to-Chip	芯片间一致性互连，Grace↔GPU 间 900 GB/s 级

术语	英文全称	一句话解释
CNP	Congestion Notification Packet	RoCEv2 拥塞通知报文，接收端回给发送端的降速信号
CoS	Class of Service	以太网优先级类别（PCP 字段），PFC/ECN 按 CoS 粒度作用
CP	Congestion Point	DCQCN 中的拥塞点，即打 ECN 标记的交换机
CPO	Co-Packaged Optics	共封装光学，光引擎与交换芯片同基板封装以降低功耗
CX-7 / CX-8	ConnectX-7 / ConnectX-8	NVIDIA 400G 通用 HCA / 800G AI SuperNIC
DCQCN	Data Center Quantized Congestion Notification	RoCE 端到端拥塞控制协议：ECN 标记+CNP 反馈+乘性降速
DP	Data Parallel	数据并行：每节点一份完整模型，切分数据
DPU	Data Processing Unit	数据处理器，卸载存储/网络/安全/管理（如 BlueField-3）
DSCP	Differentiated Services Code Point	IP 头 QoS 字段，用于映射到 CoS（如 DSCP26→CoS3）
ECN	Explicit Congestion Notification	显式拥塞通知：交换机在 IP 头打 CE 标记而不丢包
EP	Expert Parallel	专家并行：MoE 模型把专家切到不同 GPU
Fat-tree	胖树	Clos 拓扑的一种，上行带宽逐级不衰减，可无阻塞
GDR	GPU Direct RDMA	网卡 DMA 直接读写 GPU 显存，绕过 CPU 内存
GDS	GPU Direct Storage	存储直接读写 GPU 显存
go-back-N	—	重传策略：丢包后从该包起全部重发，RoCE 采用
Grace	NVIDIA Grace CPU	NVIDIA ARM 服务器 CPU（Neoverse V2，72 核，LPDDR5）
HBM	High Bandwidth Memory	2.5D 封装堆叠显存，B300 配 HBM3e
HCA	Host Channel Adapter	IB 语境下的网卡，功能同 NIC
Headroom buffer	净空缓冲	为吸收 PAUSE 生效前在途数据预留的交换机 buffer
HoL blocking	Head-of-Line Blocking	队头阻塞：队首报文堵住后续去往空闲方向的报文
Host Memory	主机内存	CPU 侧内存（DDR/LPDDR），与 GPU HBM 相对
IB	InfiniBand	原生无损的专用 RDMA 网络（NVIDIA Quantum 系列）

术语	英文全称	一句话解释
Leaf / Spine / Super-spine	—	Clos 拓扑的接入层 / 汇聚层 / 三级扩展层
MGX	NVIDIA MGX	NVIDIA 模块化液冷机柜/服务器参考设计规范
NCCL	NVIDIA Collective Communications Library	GPU 集合通信库，训练框架的通信底座
NIC	Network Interface Card	网卡
NP	Notification Point	DCQCN 通知点：收到 CE 报文后回 CNP 的接收端网卡
NUMA	Non-Uniform Memory Access	非一致性内存访问架构；NIC 须亲和挂载以保 PCIe 带宽
NVLink	NVIDIA NVLink	GPU 间专用高带宽互连，第五代 1.8 TB/s/GPU
NVSwitch	NVIDIA NVSwitch	NVLink 交换芯片，实现域内全互联无阻塞
OOB	Out-of-Band	带外管理：与业务面物理隔离的管理网络（1GbE）
OSFP	Octal Small Form-factor Pluggable	800G 光模块封装形态（8×100G）
PFC	Priority Flow Control	IEEE 802.1Qbb 逐跳反压：按优先级 PAUSE，构建无损链路
Rail / Rail-optimized	—	按 GPU 位置对齐的 leaf 布线方式，每 GPU 一条专属路径
RDMA	Remote Direct Memory Access	远程直接内存访问，绕过 CPU/内核的零拷贝网络
RoCEv2	RDMA over Converged Ethernet v2	封装在 UDP/IP 上、可路由的以太网 RDMA
RP	Reaction Point	DCQCN 响应点：收到 CNP 后乘性降速的发送端网卡
RTT	Round-Trip Time	往返时延，DCQCN 反馈环的最小反应时间
SHARP	Scalable Hierarchical Aggregation and Reduction Protocol	IB 网内归约协议，把 AllReduce 卸载进交换机
SM	Streaming Multiprocessor	GPU 基本计算单元，内含 CUDA/Tensor Core
Spectrum-X	NVIDIA Spectrum-X	NVIDIA AI 以太网平台（SN5600/SN5610 等）
Spraying	Packet Spraying	报文级多路径打散，消除 ECMP 热点
SU	Scalable Unit	可扩展单元；NVL72 中 1 SU = 1 柜 = 72 GPU
SuperNIC	—	面向 AI 的可编程网卡定位（CX-8），强于传统 NIC

术语	英文全称	一句话解释
Tensor Core	—	GPU 内矩阵乘加专用硬件单元，AI 算力来源
TP	Tensor Parallel	张量并行：单层算子切到多 GPU，通信极频繁
无损网络	Lossless Network	不丢包的网络（PFC/ECN 保障），RoCE 的前提
东西向流量	East-West Traffic	集群内部节点间流量，AI 集群主体
南北向流量	North-South Traffic	集群与存储/外部之间的流量
Scale-up	纵向扩展	单一紧耦合域内加 GPU（NVLink/NVSwitch）
Scale-out	横向扩展	跨节点/跨柜加机器（IB/以太网）

附录 B FAQ 汇总

F1：GB300 NVL72 一柜有多少颗 GPU、多少功耗?
72 颗 B300 + 36 颗 Grace，液冷 MGX 机柜，整柜约 142 kW，8×33 kW 电源 shelf。

F2：NVLink 和 NVSwitch 是什么关系?
NVLink 是 GPU 上的链路技术（第五代 1.8 TB/s/GPU），NVSwitch 是把这些链路汇聚成全互联交换网络的芯片（NVL72 每柜 9 个 switch tray、18 颗 ASIC，聚合 130 TB/s）。

F3：为什么每颗 GPU 要配一颗 CX-8?
1:1 rail 设计让每颗 GPU 独占 800G 通道，NCCL 可确定性钉扎流量，避免多 GPU 抢卡的不可预测拥塞。

F4：三张 fabric 分别是什么、谁来承载?
东西向计算网（CX-8 800G×4/tray，Spectrum-X）、南北向 converged 网（BF-3 DPU 双 400G）、OOB 管理网（SN2201 1GbE）。

F5：双平面是不是带宽翻倍?
不是，单 NIC 800G 拆 2×400G 总带宽不变；翻倍的是冗余路径和故障隔离。

F6：RoCEv2 为什么必须无损?
丢一个包触发 go-back-N 重传，在同步集合通信中放大为全集群抖动。

F7：PFC 和 ECN 谁主谁辅?
ECN+DCQCN 是端到端主控，PFC 是逐跳兜底；健康网络绝大多数时间应只运行在 ECN 反馈上。

F8：PFC 三大缺陷是什么?
PFC 风暴（反压扩散）、队头阻塞、死锁风险。

F9：DCQCN 三步闭环是什么?
交换机打 CE 标（CP）→ 接收端回 CNP（NP）→ 发送端乘性降速后加性恢复（RP）。

F10：ECN 和 PFC 门限怎么配?
ECN Kmax 必须低于 PFC 触发门限，并留足 headroom；CNP 队列（DSCP48→CoS7）不可开 PFC；RoCE 流惯例 DSCP26→CoS3。

F11：headroom buffer 是干什么的?
吸收 PAUSE 帧传播期间的在途数据，防止无损域破功；速率越高、链路越长，需求越大。

F12：Spectrum-X 的 spraying 和 ECN 矛盾吗?
不矛盾：spraying 解决负载不均，ECN/DCQCN 解决真实拥塞；spraying 让队列水位均匀，反而让

ECN 标记更准确。

F13：NVIDIA 反馈里"PCIe Gen5 导致 PFC"的因果链是什么?
Gen5 x16（约 63 GB/s 单向）喂不饱 800G 网卡 → NIC 写不动 host → NIC buffer 堆积 → 反压传回交换机 → 触发 PFC——瓶颈在 PCIe 段，症状在以太网段。

F14：为什么 CX7 有网卡标记 ECN 却"没啥用"?
CX-7 是 400G + Gen5 x16，host 侧刚好喂饱网络侧，PCIe 不构成瓶颈，该功能无用武之地。

F15：为什么 CX8 反而还没实现这个功能?
CX-8 是全新 800G datapath，功能需在新架构上重新实现与验证，NVIDIA 正在与网卡架构团队评估所需资源；Gen6 缓解了瓶颈但未消灭（混插 Gen5 平台、PCIe 资源共享等）。

F16：在该功能落地前，NVL72 部署能做什么缓解?
压低 ECN 门限留余量、严守三网隔离、host 方向流量应用层限速、监控 PFC 计数与 PCIe 利用率、核对 NUMA 亲和性。

F17：NVL72 和 HGX 8 卡怎么选?
超大模型、机房具备液冷与高功率条件选 NVL72（Scale-up 域 72 卡）；中等规模或机房受限选 HGX 集群。

F18：官方 RA 支持多大的扩展?
1 SU = 1 柜 72 GPU；标准方案 2/4/8 柜（144/288/576 GPU），8 柜引入 super-spine 三级 Clos。

附录 C 参考链接

- NVIDIA Enterprise Reference Architecture: NVL72 AI Factory with GB300 NVL72 and Spectrum-X（本文主要事实来源）：https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory-with-gb300-nvl72-dual-plane-networking-architecture.pdf
- NVIDIA Enterprise Reference Architectures 总入口：https://docs.nvidia.com/enterprise-reference-architectures/
- NVIDIA Spectrum-X Networking Platform：https://www.nvidia.com/en-us/networking/ethernet/spectrum-x/
- NVIDIA ConnectX-8 SuperNIC：https://www.nvidia.com/en-us/networking/products/connectx-8/
- NVIDIA BlueField-3 DPU：https://www.nvidia.com/en-us/networking/products/data-processing-unit/
- RoCE 拥塞控制（DCQCN）原始论文：Zhu et al., "Congestion Control for Large-Scale RDMA Deployments", SIGCOMM 2015
- IEEE 802.1Qbb（Priority-based Flow Control）标准文本：https://1.ieee802.org/dcb/802-1qbb/

本文档基于 NVIDIA 官方 GB300 NVL72 企业参考架构文档（2026 年 3 月版）及公开技术资料撰写，供工程与研究人员学习参考。第六部分 6.4、6.5 小节为公开行业信息，非官方 RA 文档内容。