



# 算力发展趋势分析& 高算网与广域网融合的挑战

——发展网内安全先进算力

2025年12月  
集成业务部

- 算力发展趋势分析
- 高速计算网架构设计
- 高速计算网与广域网融合
- 总结



目录

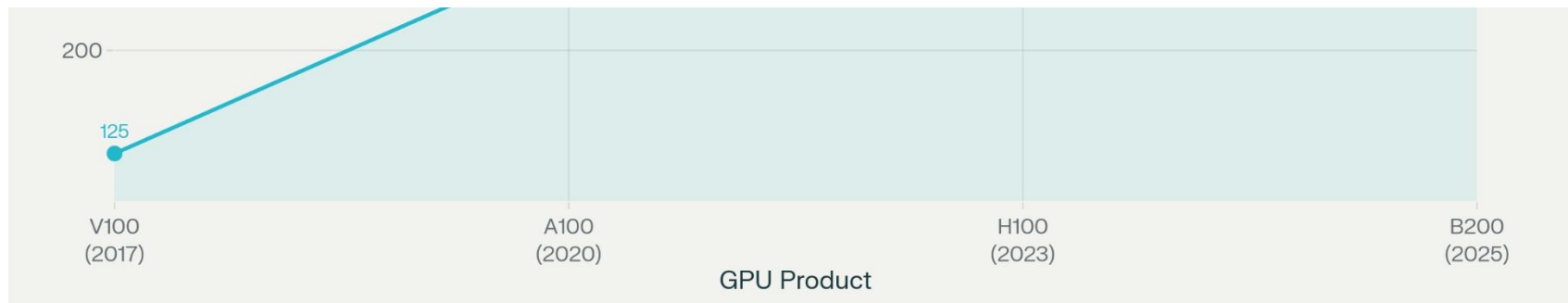
CONTENTS

### 三 英伟达2017-2025旗舰产品算力发展趋势

## 功能齐全的架构

Turing 之前的 GPU 架构（计算能力 7.5）被视为功能齐全，因此，我们在 CUDA 13.0 中取消了对计算能力 7.5 之前的 GPU 的离线编译支持。此外，NVIDIA 驱动程序的 R580 分支将是支持这些架构的最后一个驱动程序分支。

这意味着，希望支持 7.5 之前的架构的应用开发者需要使用 CUDA 12.9 或更早版本来构建应用，而这些应用的用户需要使用 580 驱动分支。所有之前发布的 CUDA 工具套件均可在我们的 [CUDA 下载存档页面](#) 上获取。请注意，580 分支是 [长期支持分支](#)，将维护和支持三年。我们将在 [博客文章](#) 中更详细地讨论这些更改。

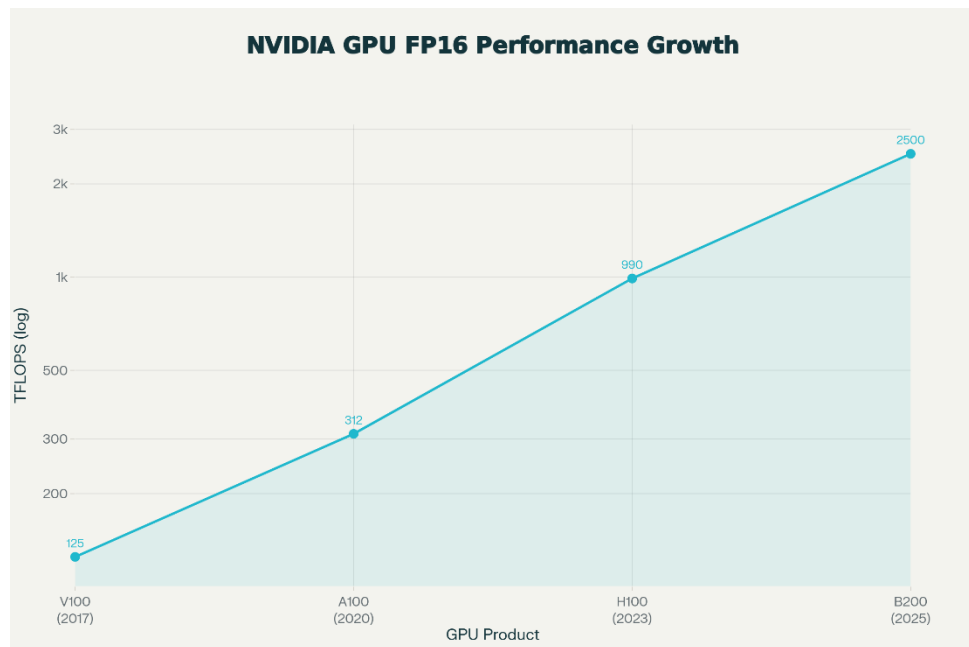


英伟达 GPU 从 V100 到 B200 的算力提升呈现**指数级增长**，**B200 的 FP8 算力比 V100 提升约 80 倍**，比 H100 提升 2.5 倍。

这一飞跃主要归功于：

- 计算精度不断降低 (FP16→FP8→FP4)
- 架构革新 (Volta→Ampere→Hopper→Blackwell)
- 内存和互联技术突破

# 三 英伟达2017-2025旗舰产品算力发展趋势



英伟达 GPU 从 V100 到 B200 的算力提升呈现**指数级增长**，**B200 的 FP8 算力比 V100 提升约 80 倍**，比 H100 提升 2.5 倍。这一飞跃主要归功于：

- 计算精度不断降低 (FP16→FP8→FP4)
- 架构革新 (Volta→Ampere→Hopper→Blackwell)
- 内存和互联技术突破

## 精度革命：

- 从 FP32→FP16→FP8→FP4，每代精度降低使算力呈指数级增长

## 架构创新：

- Volta→Ampere: CUDA 核心从 5120→6912 (+35%)，引入稀疏计算
- Hopper: Transformer 引擎 + FP8, FLOPS 利用率从 16.6% 提升至 42.0%
- Blackwell: 双芯片 + FP4, 推理性能比 H100 提升 15 倍

## 内存与互联：

- 带宽从 V100 的 900GB/s 飙升至 B200 的 8TB/s (+789%)
- NVLink 从 300GB/s→1.8TB/s (+500%)，大幅减少通信瓶颈

### 三 国内厂商旗舰算力产品参数介绍

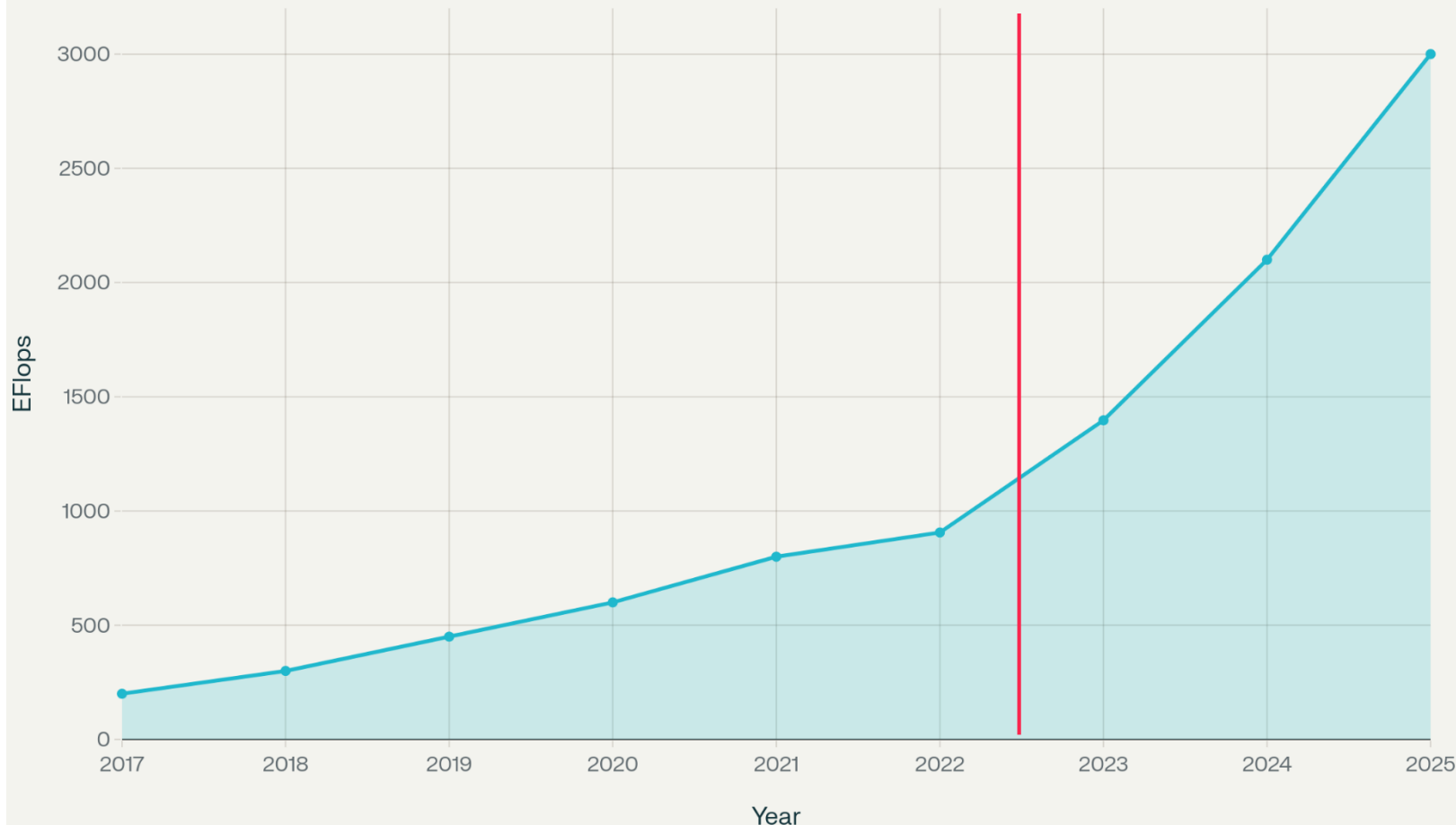
厂商	产品	关键算力指标	对标 N 卡	工艺 / 特点
华为	昇腾 950DT	FP8: 2 PFLOPS	H100	7nm+, SuperPoD 支持 15k 卡
壁仞	BR100	BF16: 1,024 TFLOPS	H100	7nm+ Chiplet, 利用率 + 18%
海光	深算三号	FP16: 192 TFLOPS	A100	7nm, 类 CUDA 生态
天数	智铠 100	INT8: 384 TOPS	RTX 4090	7nm, 推理专用
浪潮	元脑 SD200	支持 64 路国产 GPU	DGX SuperPod	开放架构, 支持多厂商芯片

国内厂商也涌现出一批对标N卡，性能达到国际先进水准的产品，华为的**昇腾 950**系列单芯片性能突破 PFLOPS 级，壁仞、海光也都推出了对标 H 系和 A 系显卡的对应产品；集群计算方面，昇腾 384 超节点支持 **10** 万卡集群，互联带宽可达 **784GB/s**，浪潮也对标 DGX SuperPod 推出元脑 **SD200**，其优点在于采用开放架构，能够集成多厂商芯片。



# 2017-2025全球设备算力总量增长趋势

## Global Computing Power Growth



受 AI 需求驱动，全球算力设备总算力从 2017 年的约 **200 EFLOPS** 增长至 2025 年的约 **3000 EFLOPS**，呈现爆发式增长。

该图基于中国信通院等数据估算，2022 年达 906 EFLOPS，2023 年跃升至 1397 EFLOPS，年增速超 50%，2024 - 2025 年预计继续高速扩张至超 3 ZFLOPS。

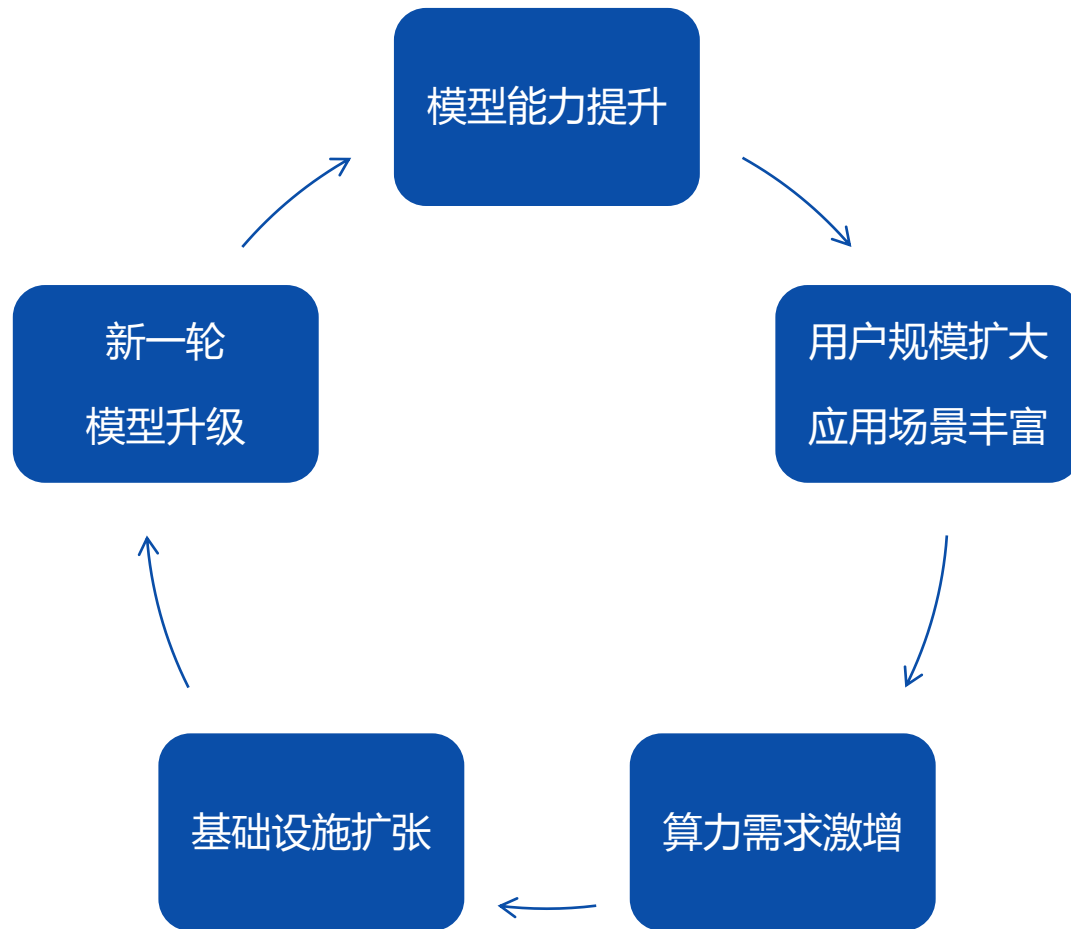
# OpenAI的算力扩张与全球算力需求爆发

- 算力投资巨额增长
- 算力规模惊人扩张
- 从模型驱动转向算力驱动

OpenAI 算力需求的 "黑洞效应": 从**训练**到**推理**的全链路爆发

- 参数规模爆炸
- 多模态融合
- 训练策略升级

与此同时, 推理算力需求以高于训练需求的惊人速度增长, 2025 年推理增速为训练的 4 倍



# 最新 AI 旗舰模型对比

参数	GPT-5.1	Gemini 3 Pro	趋势说明
发布时间	2025.11.12	2025.11.18	
上下文长度	128K tokens	100 万 tokens	为支撑更庞大的信息处理需求，相比前代模型上下文窗口显著增大
- Humanity's Last Exam	26.5% (无工具)	37.5% (无工具)	复杂推理能力提升 41.5%，算力需求激增
- GPQA Diamond (科学)	87.60%	91.90%	博士级科学推理能力提升 4.9%，需更强算力支持
- MathArena Apex (地狱数学)	~1.0%	23.40%	数学推理能力提升，算力需求呈爆发式增长
推理速度	150 秒 / 题 (ScienceQA)	49 秒 / 题	处理速度提升 67%，但需更高效的硬件支持
架构创新	自适应推理：简单任务减少 57% token，复杂任务增加 71%	MoE 架构：万亿参数仅激活 2%，推理延迟降至前代 1/3	模型通过创新架构优化算力使用，但整体需求仍大幅提升

根据最新两代 AI 旗舰模型的参数，可以发现，模型逐渐往**上下文窗口增大、推理复杂度提高和多模态融合**的方向发展，这意味着未来的 AI 模型训练将需要十万卡级超算集群、专用计算架构和革命性的算力优化技术，才能满足不断攀升的算力需求，这也将持续推动全球 AI 基础设施投资进入新一轮高潮。

### 算力市场未来展望：

- 2025 年全球算力规模已达约 **3300 EFLOPS**，其中 AI 算力占比 60%。据预测，2030 年全球总算力将达到 **50-60 ZFLOPS**，主要取决于 AI 算力的增长速度。
- 中国将成为全球算力中心，2030 年市场份额达 **55%**，智能驾驶、工业 AI、医疗影像三大场景贡献 62% 算力消耗。
- 推理算力预计将在 2028 年超过训练算力，成为市场主导，规模达 **2,000+ EFLOPS**。

国内算力厂商正通过差异化路线追赶英伟达：华为依靠集群优势，壁仞专注单芯片性能，海光发力生态兼容，天数智芯主打训推一体，浪潮提供全栈基础设施。虽然单芯片性能与英伟达仍有差距 (约 70-90%)，但在特定场景和集群规模下已展现出竞争实力。



- 算力发展趋势分析
- 高速计算网架构设计
- 高速计算网与广域网融合
- 总结



# 目录

CONTENTS

### 三 高速计算网区别于一般局域网的核心逻辑

- 传输协议差异：从 **TCP** 到 **RDMA**（远程直接内存访问）—— 绕开 CPU 干预，降低协议栈开销
- 网络特性优化：
  - 带宽设计：从 “共享带宽” 到 “专属通道”，支持多路径聚合（MPT）
  - 延迟控制：**硬件卸载**（CRC 校验、流量控制）、短帧传输优化
  - 拥塞处理：基于优先级的流量调度（QoS）、无损传输机制（LRO/TSO 禁用、PFC 流控）
- 应用场景适配：计算密集型、数据密集型场景的传输需求与普通办公场景的本质区别



# 三 高速计算网区别于一般局域网的核心逻辑



对比维度	TCP (传输控制协议)	RDMA (远程直接内存访问)
协议本质与定位	面向连接的传输层协议，适用于通用数据传输，核心目标是“可靠传输”（兼容各类网络环境）	跨节点内存直接访问技术，核心目标是“高速低耗”（为计算密集 / 数据密集场景设计）
核心传输机制	基于“端到端确认 + 重传”：滑动窗口、拥塞控制（TCP Cubic/BBR）、CRC 校验、超时重传	基于“硬件卸载 + 零拷贝”：绕过 CPU/OS 协议栈，直接通过网卡访问远程内存，支持 Read/Write/ 原子操作
CPU 资源占用	高：需 CPU 处理协议栈解析、数据拷贝（用户态 → 内核态 → 网卡）、重传确认等操作，占用大量计算资源	极低：数据传输由网卡（RNIC/HCA）硬件完成，CPU 仅初始化连接，负载降低 80% 以上
端到端延迟	高：协议栈处理（数十微秒）+ 重传等待（毫秒级），10Gbps 链路下单程延迟通常 10-50μs	极低：硬件直连 + 无协议栈开销，200Gbps 链路下单程延迟低至 0.5-2μs（亚微秒级）
协议栈复杂度	复杂：包含传输层（TCP）、网络层（IP）、链路层（以太网）完整栈，冗余逻辑多	简洁：原生 RDMA 无 TCP/IP 层（如 InfiniBand），RoCE 仅轻量封装 UDP/IP，协议开销极低
硬件依赖	低：普通以太网网卡（NIC）即可支持，无需专用硬件	高：需支持 RDMA 的专用网卡（RNIC/HCA）+ 兼容交换机（如支持 PFC 的以太网交换机、IB 交换机）

## 核心差异本质：

**TCP** 是“软件层兼容优先”的通用协议，  
**RDMA** 是“硬件层性能优先”的专用技术；

## 性能差异原因：

**RDMA** 通过“零拷贝 + 硬件卸载”解决了 **TCP** 的两大痛点——协议栈开销和数据拷贝延迟；

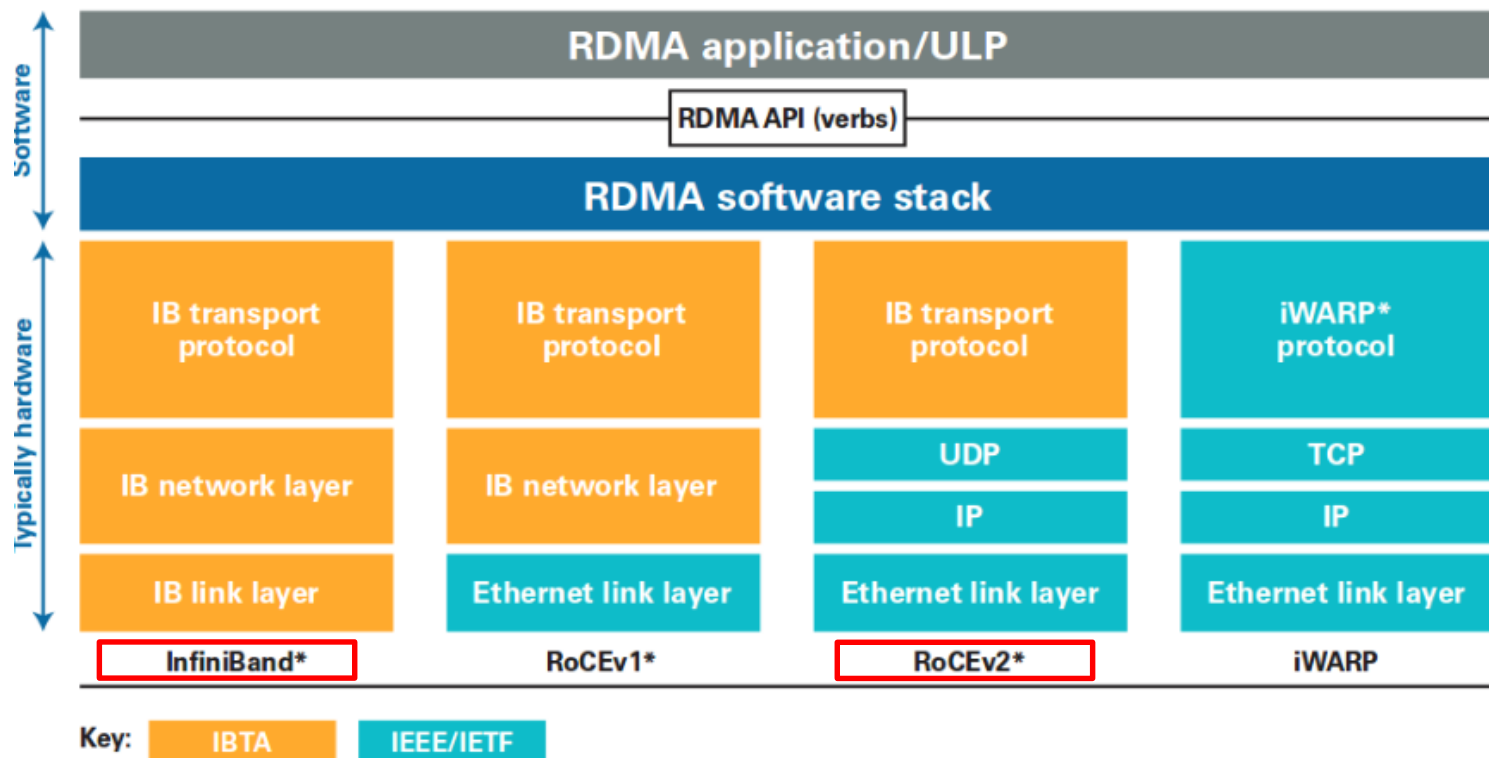
## 融合趋势：

**RoCE** 系列协议正是通过“**RDMA** 核心机制 + 以太网底层”，试图平衡“高性能”与“低部署成本”，成为以太网与高速计算网融合的关键桥梁。

# 三 高速计算网的核心技术：RDMA

## RDMA and RDMA options

RDMA is a host-offload, host-bypass technology that enables a low-latency, high-throughput direct memory-to-memory data communication between applications over a network.

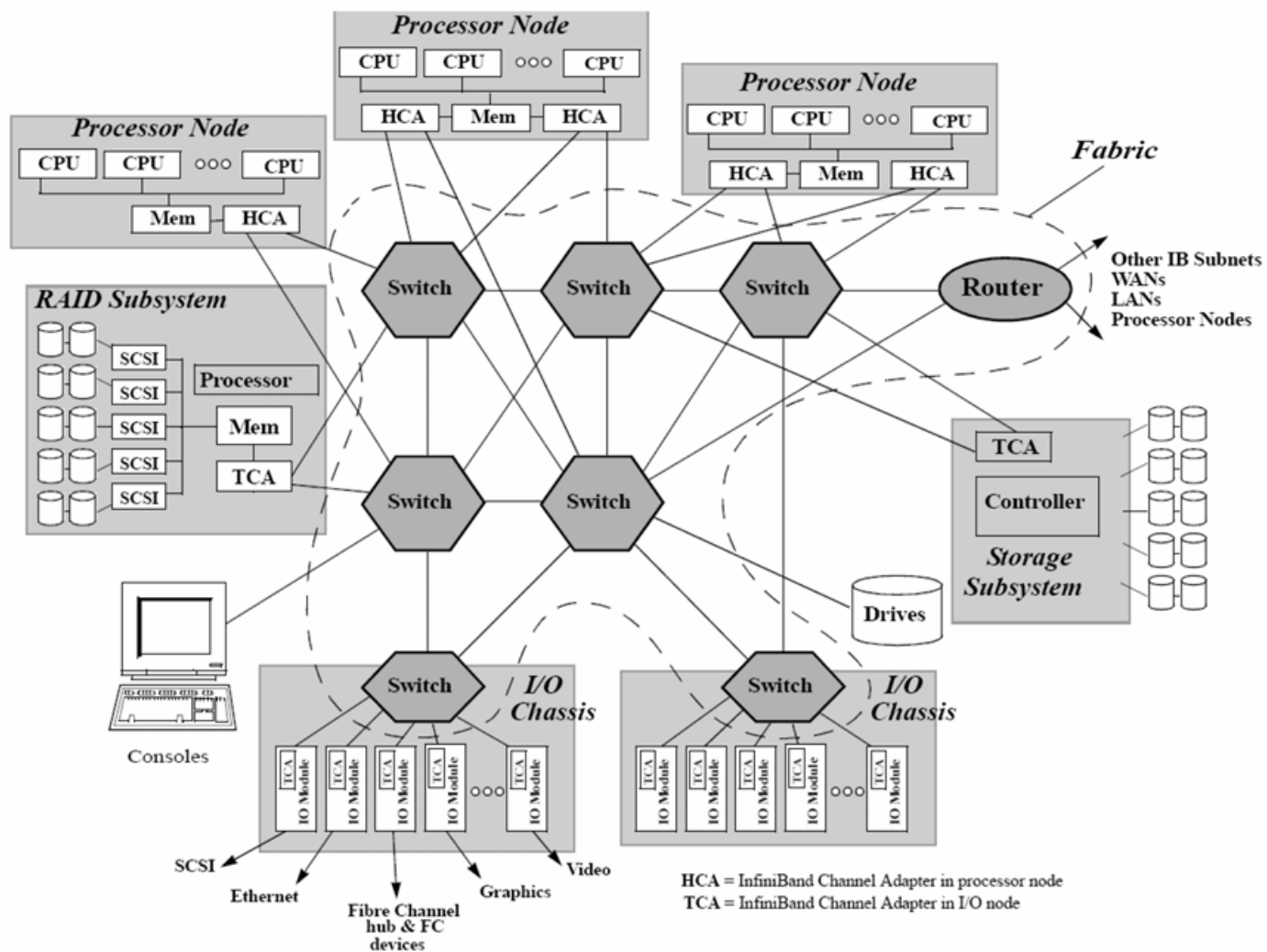


RDMA (Remote Direct Memory Access, 远程直接内存访问) 是一种**高性能网络传输技术**, 核心是允许一台计算机的内存直接访问另一台计算机的内存, 无需双方 CPU 和操作系统内核的介入, 实现“内存到内存”的直接数据传输。

### 核心特点

- 低延迟
- 低 CPU 占用
- 高带宽

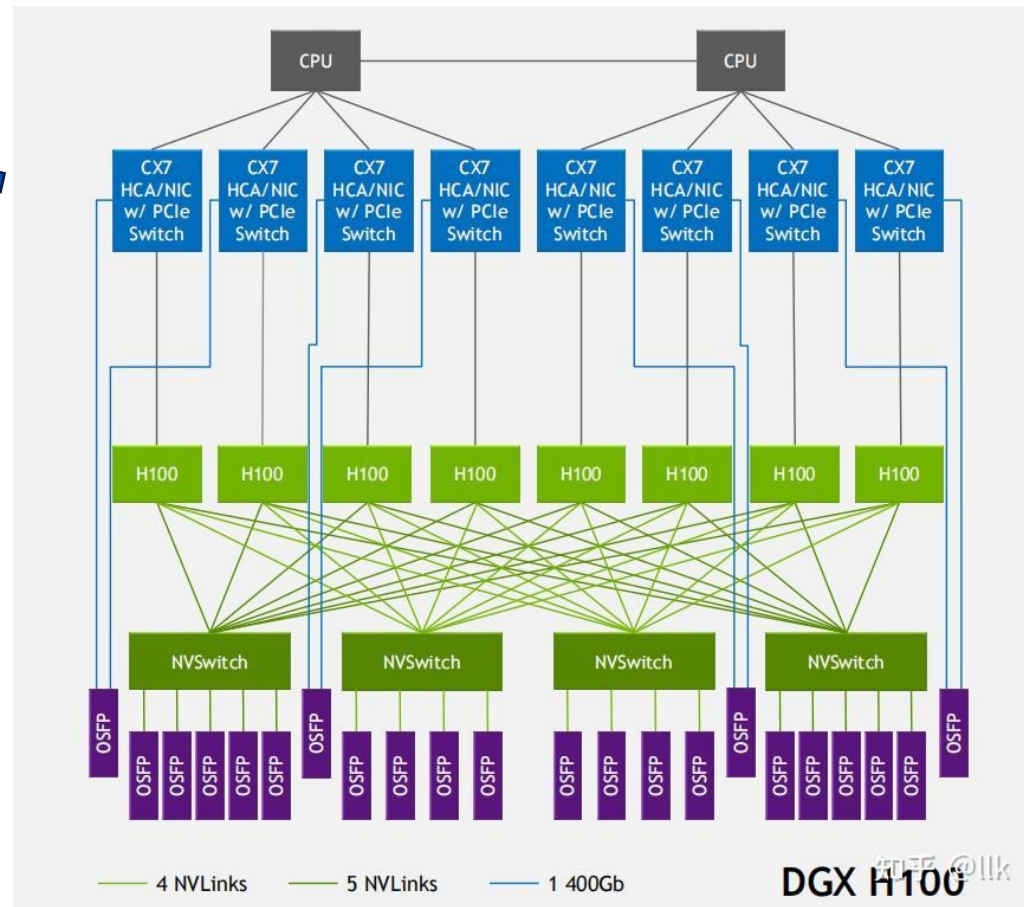
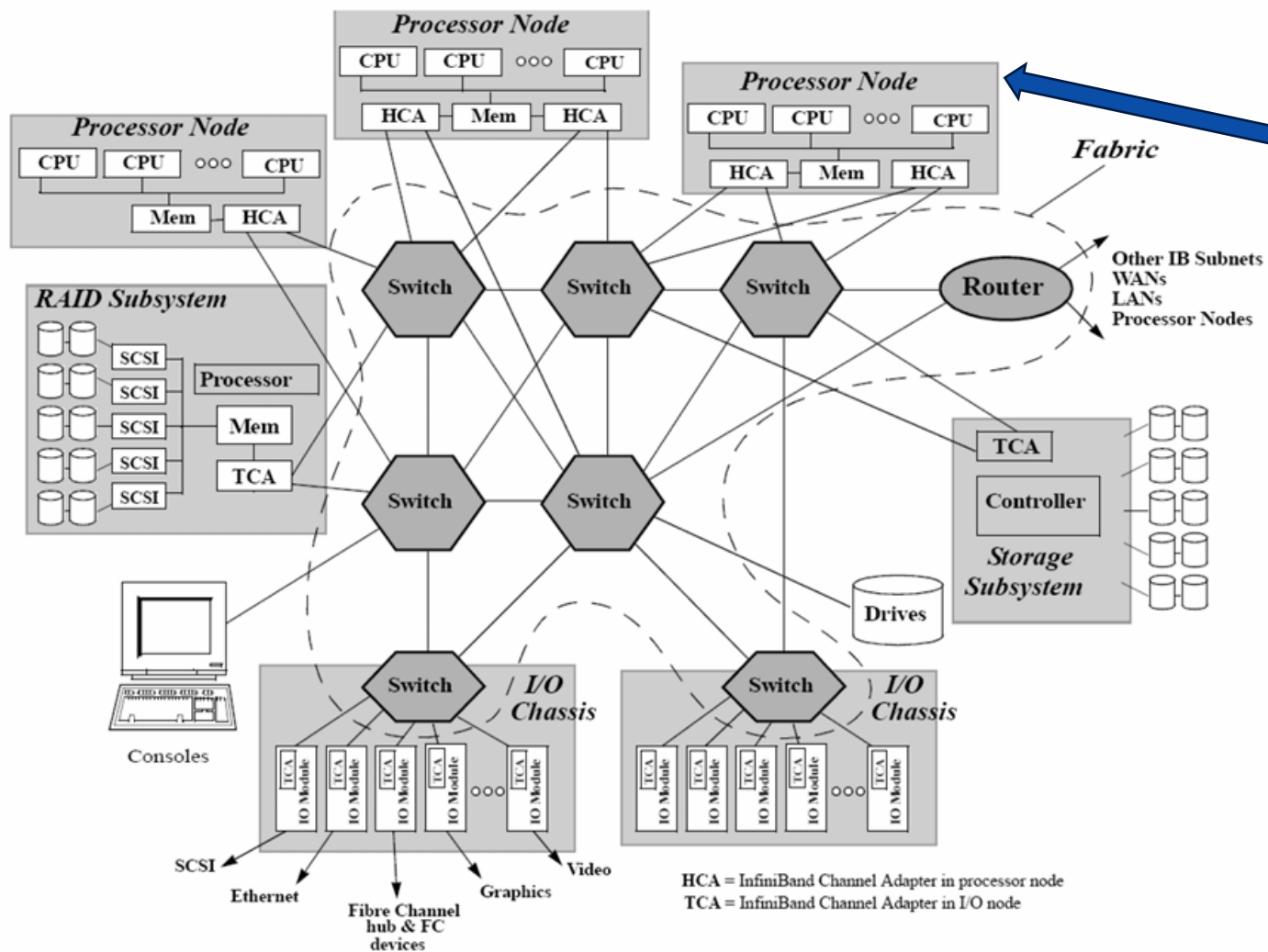
# 三 高速计算网的典型架构：IB



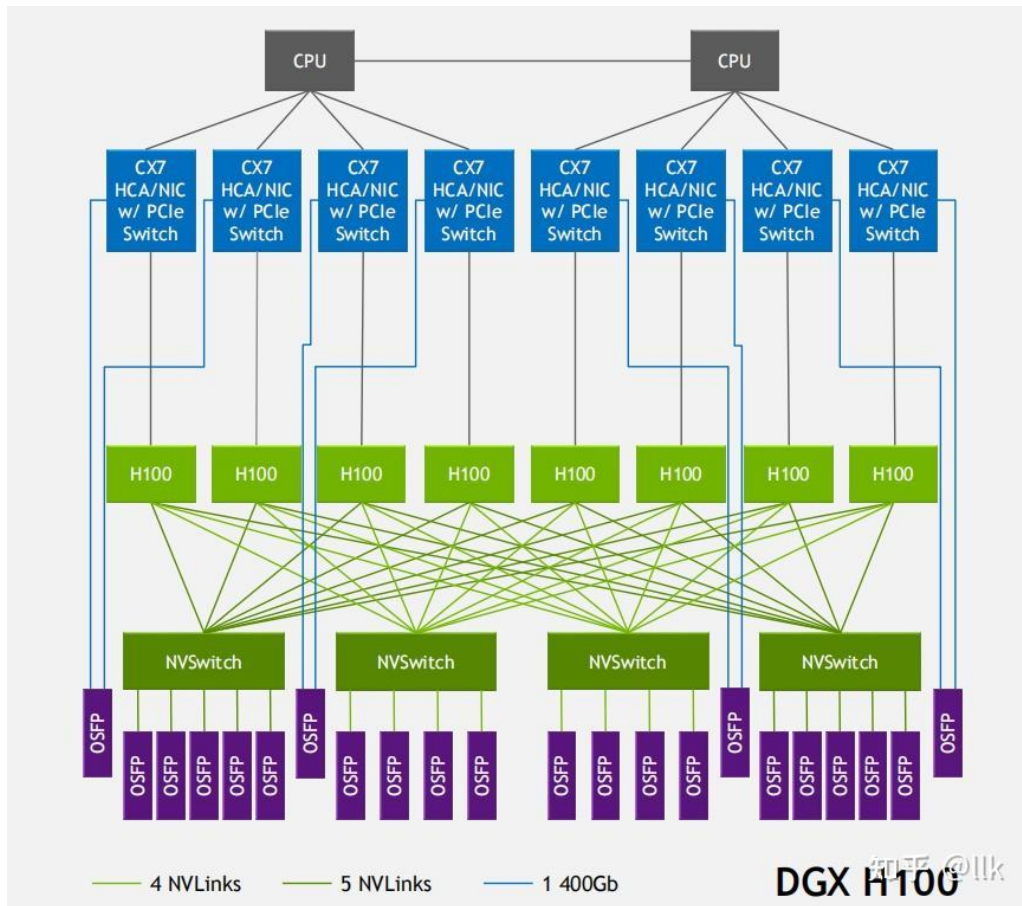
三个主要组件

- Channel Adapters
- Switches / Routers
- Links and connectors

# 三 高速计算网的典型架构：IB



# 三 单算力节点内部结构分析

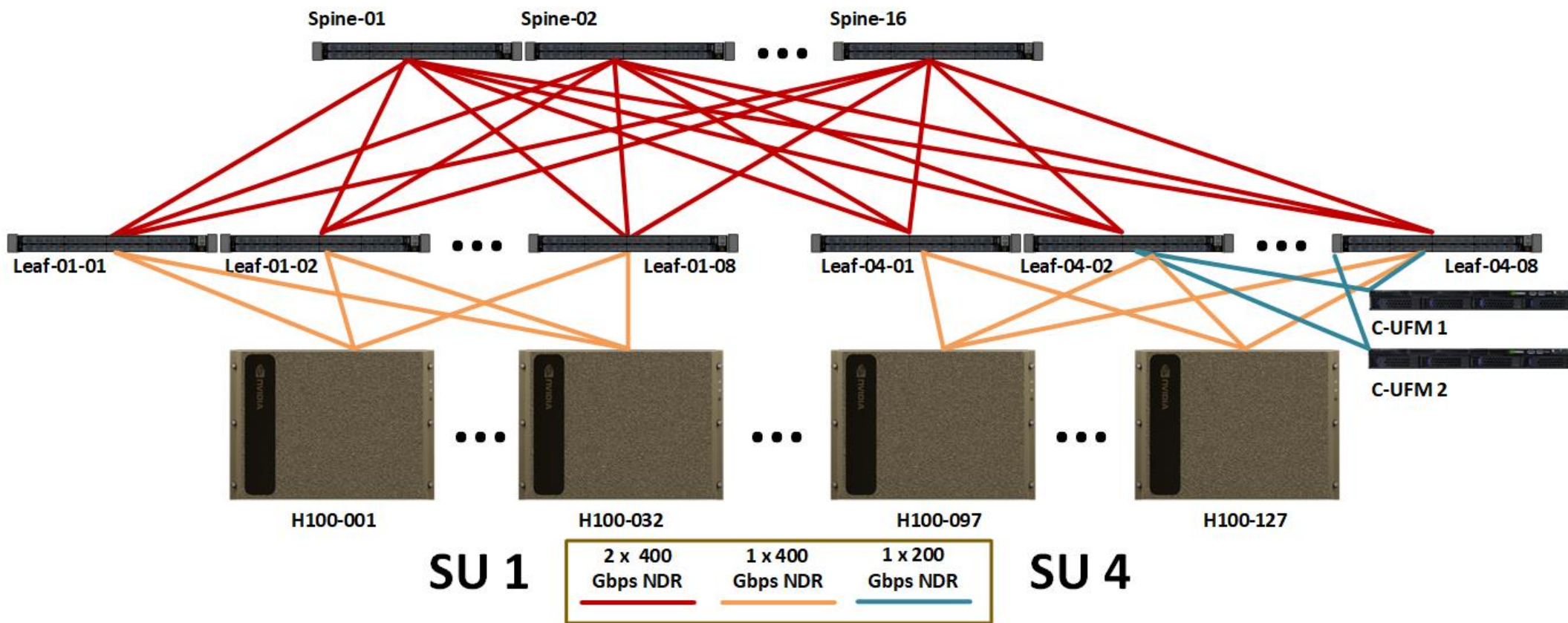


## 以 DGX H100 为例

每台 [DGX H100](#) 服务器里有 8 张 [H100 卡](#)，每个 H100 卡分别连接到一个 [leaf 交换机](#)上，这是为了防止单个 leaf 交换机故障，以及形成无阻塞网络。

并且这样可以保证 SU 里的每台服务器都有一个连接到同一个 leaf 交换机上，编号相同的 GPU 卡（例如，所有服务器里的 3 号卡）都连接到同一个 leaf 交换机，有利于提升模型训练效率（如提升 AllReduce 操作的效率）。

# IB的典型拓扑结构：胖树（Fat-Tree）

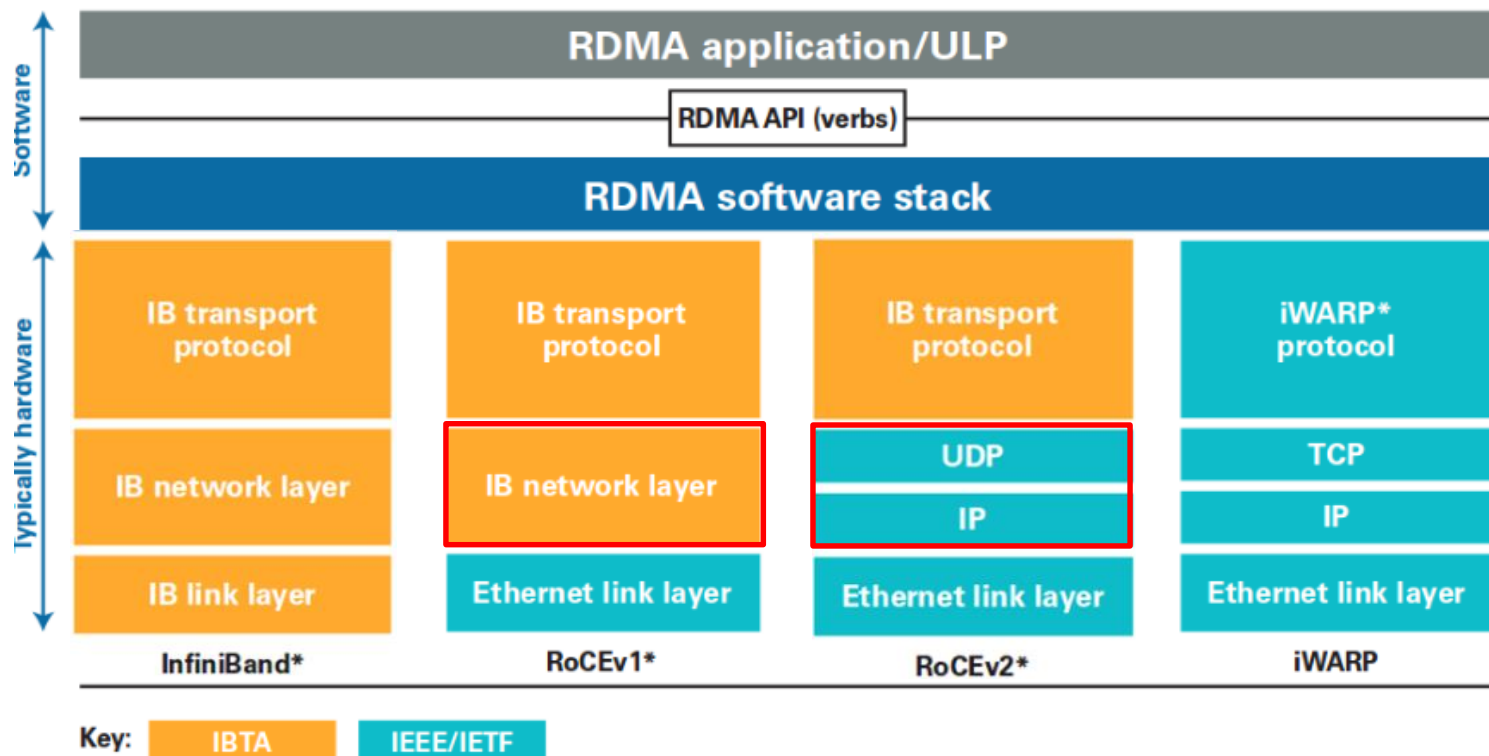


Fat-Tree（胖树）是一种**高带宽、全网络非阻塞**的拓扑结构，核心特点是网络带宽随层级向上扩展（避免上层瓶颈），广泛用于高性能计算（HPC）、数据中心和 InfiniBand 网络中。

# 三 高速计算网的典型架构：RoCEv2

## RDMA and RDMA options

RDMA is a host-offload, host-bypass technology that enables a low-latency, high-throughput direct memory-to-memory data communication between applications over a network.



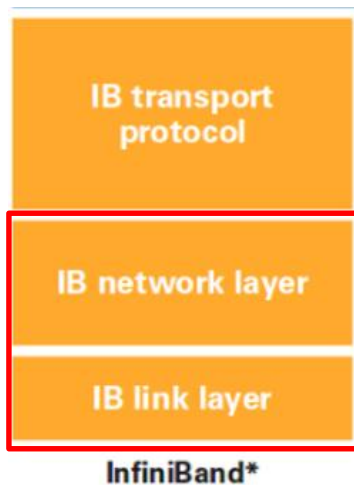
## 核心工作原理:

数据包结构: 在传统以太网帧内封装 IP/UDP 头 (目标端口 4791) 和 RDMA 控制信息, 形成标准兼容的三层数据包

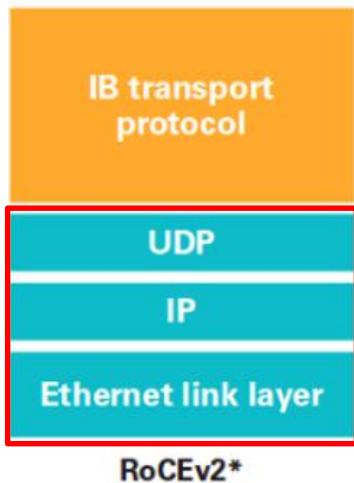
传输机制: 绕过操作系统内核, 网卡 (RNIC) 直接访问远程主机内存, 消除 CPU 参与和数据拷贝

路由能力: 支持跨子网三层路由, 突破 RoCEv1 的二层网络限制, 适应企业复杂网络拓扑

# IB 与 RoCEv2 对比



- **InfiniBand**: 追求最高性能，协议各层级均采用自研架构，在传输过程中能够保证无损稳定、低网络开销，能够满足超算、算力集群部署等场景要求。缺点是部署成本过高，且可扩展性弱，难以与以太网架构为基础的传统局域网融合。



- **RoCEv2**: 与以太网兼容，仅需在原本网络基础上进行少量的硬件升级，部署成本较低，相比于 IB，在网络层使用 UDP/IP 封装，会带来少量额外开销，对于企业级高性能计算、分布式存储等领域较为适配。相比于 IB，RoCEv2 是更适合作为传统广域网向高速计算网过渡时期的协议。

- 算力发展趋势分析
- 高速计算网架构设计
- 高速计算网与广域网融合
- 总结



### 三 高速计算网推广到广域网场景的局限性

搭建高速算力网络的最基本目标是满足高性能计算场景下的高带宽需求，在高算网架构推广到广域网后，或可实现算力的远距离传输，拓展通过云平台远程访问资源的传统算力资源调度模式。但目前仍然面临以下几点限制：

**算力的远程传输受到距离远等空间限制：**即使目前使用光缆传输也仍然存在较长的传输时延，难以避免——例如，而高性能计算对时延的容忍阈值多在毫秒级以下，直接影响实时性算力任务的执行；

**长传输时延使得重传的难度大幅提高：**一旦数据传输出现丢包，需等待原数据反馈后才能发起重传，不仅延长任务耗时，更增加了数据在传输过程中被拦截、篡改的风险，对传输的安全性提出了更高要求；

**传统广域网的基础设施与高算网架构存在显著适配鸿沟：**传统广域网的核心节点设备多基于常规数据传输设计，普遍存在带宽瓶颈——多数传统广域网骨干链路带宽仅能满足 GB 级传输需求，而高算网单算力节点的峰值传输需求常达 TB 级，硬件设备的承载能力不足直接限制高算网的算力传输效率；



# RoCEv2架构推广的可行性分析

## 兼容性优势

- **设备兼容**：使用标准以太网交换机和线缆 (10 Gbps ~ 800 Gbps)，仅需将服务器网卡升级为支持 RoCE 的 RNIC
- **协议透明**：数据包外观与普通 UDP/IP 包无异，不干扰现有网络管理和监控系统
- **路由友好**：可与传统 IP 路由协议 (OSPF/BGP) 共存，支持跨部门、跨区域部署

性能指标	传统 TCP 网络	RoCEv2 网络	提升幅度
端到端延迟	10-100 $\mu$ s	1-5 $\mu$ s	10-20 倍
CPU 负载	高 (数据拷贝 + 协议处理)	极低 (硬件卸载)	降低 80%+
带宽利用率	60-80%	90-95%+	提升 15-30%
数据拷贝次数	2-3 次 (用户态 $\rightarrow$ 内核态 $\rightarrow$ 网卡)	0 次 (直接内存访问)	消除拷贝

## 三 慢数据与快数据分流方案评估

### 分流实现原理:

- 部署智能边缘设备 (如智算 CPE) 识别流量类型
- "快数据"(如 AI 训练样本、大模型参数)→RDMA 高速通道→算力网
- "慢数据"(如普通办公、Web 服务)→传统 IP 路径→广域网
- 网络控制器动态监控、调度, 保障 "快数据" 优先无损传输

### 实施效果与价值:

- 算效保障: 广域 RDMA 使跨百公里训练效率仅下降 1%
- 带宽利用率: 从 60% 提升至 90%, 成本降低 30%
- 传输效率: TB 级数据传输从 "小时级" 降至 "分钟级"

### 存算分离架构:

- 企业原始数据本地存储, 模型训练在云端, 通过智算广域网传输中间结果
- 解决 "数据安全不出域" 与 "算力资源不足" 双重痛点

## 三 案例分析——端到端 400GE 智算试验网



国内已有厂商打造了**业界首张端到端 400GE 的 IP 智算广域试验网络**，为 300 + 企业和 40 + 高校提供智算服务，构建了连接企业园区与运营商智算中心的“算网一体”基础设施。

- **端到端 400GE+RDMA 广域无损传输**：采用RDMA(Remote Direct Memory Access)协议实现“零拷贝”数据传输，将网络吞吐率提升至90%以上（传统网络仅60%）
- **秒级智能调度系统**：对GB级与KB级数据进行有效区分
- **云边协同训推架构**：模型Prefill层部署在本地，Decode层部署在云端，通过智算网传输中间数据，保护隐私
- **跨240公里拉远推理效率100%**：企业本地部署轻量级模型，核心计算在240公里外的智算中心完成
- **TB级模型数据“分钟达”**：1TB模型数据从智算中心到企业园区仅需**数分钟**（传统方式需数小时）
- **算网融合新服务模式**：为企业提供“运力 + 存力 + 算力”的打包服务，支持一条专线同时访问智算、超算等多种算力资源

## 三 案例分析——微软 AI 超级工厂

微软正式推出首座“AI 超级工厂”，将分散在亚特兰大和威斯康星等地的数据中心整合为统一算力系统，构建起行星际规模的分布式网络。该架构成功整合数十万个 Blackwell GPU，并采用液冷高密度设计，通过专用光纤网络实现算力协同，将复杂 AI 训练任务从数月压缩至数周，标志着 AI 基础设施正式进入网络化协同新时代。



### 核心创新

- **专用 AI WAN 高速网络：** 12 万英里专用光纤
- **行星级统一调度架构：** 三层调度体系、单一扁平网络
- **高密度 GPU 集群架构：** 部署 GB200 NVL72 计算单元
- **软件定义的 AI 基础设施：** SONiC 自研网络操作系统



- 算力发展趋势分析
- 高速计算网架构设计
- 高速计算网与广域网融合
- 总结



## 三 总结——广域网与高速计算网融合的未来趋势



**800Gbps RoCE 普及**: 主流厂商推出 QSFP-DD 800G RoCE 网卡和交换机

**无损以太网标准完善**: IEEE 802.1Qbb (PFC)、802.1Qau (ECN) 全面标准化

**NVMe over RoCE 成为存储主流**: 替代传统光纤通道, 成本降低 60%+


*Network will be the new memory.*

**融合网络成为基础设施**: 所有数据中心默认支持 RoCE 等高速计算协议

**RoCE 与传统以太网边界消失**: 标准以太网设备原生支持完整无损特性

**计算 - 存储 - 网络深度融合**: 通过 RDMA 实现 "内存即网络, 网络即内存" 的架构革命





感谢各位领导和嘉宾的支持！