基于灵衢®的

超节点参考架构白皮书

目 录

1 AI 时代的挑战

AI时代的到来为全球计算领域带来跨越式变革。算力需求在大模型爆发式演进下呈指数级增长，平均每6个月实现翻倍，大幅超越摩尔定律揭示的硬件迭代速度；组网规模由过去百卡扩展至当前十万，未来甚至百万卡，需要计算系统实现极致效率与可靠性。通用计算持续追求资源利用率的提升，同时又面临与AI算力紧密结合的新需求，需要加强多芯片池化和通信能力。计算领域的加速演进对各种算力应用场景带来全新挑战：

- **大模型预训练：**Scaling Law曲线持续攀升，SP/TP/EP总通信数据量相比较小模型提升近百倍，单批次达到数百GB，模型跨节点通信带宽增长速率仅为算力增长速率的1/5，成为训练性能关键瓶颈。
- **大模型推理：**Token输出时延（TPOT）是大模型推理核心性能指标，随着多模态与AI Agent发展，TPOT需进一步下降至5ms甚至1ms，序列长度向百万Token增长，KV Cache远超HBM容量，同时不同的推理阶段算力和内存需求差距大。
- **Agentic AI：**AI应用正逐步向Agentic AI演进，这类任务通常需要多组件协同，共同处理包含数据库操作、数据预处理和推理等在内的子任务。
- **虚拟化：**当前通用计算业务主要部署在虚拟机或容器，传统虚机30%-50%内存未被使用，设备无法在节点间池化，整体资源利用率低；同时如果虚拟机热迁移时间达到100~200ms量级，客户将感知到业务中断。从业务连续性视角，期望RPO指标为0，RTO指标小于50ms，以实现虚机热迁移场景业务无损。
- **数据库：**为提升数据库线性度，达到峰值性能最优，数据库需要分层解耦的资源池化架构，而这种架构需要大规模和低时延的内存共享。只有实现百ns量级的跨节点内存访问时延（低于当前RDMA的1/10），才能匹配数据库架构发展趋势，使得数据库峰值性能进一步提升。

为了满足日益增长的业务需求，支持智算、通算及各种融合应用场景，我们针对AI时代数据中心提出超节点参考架构。

2 超节点参考架构

超节点参考架构（SuperPoD Reference Architecture）是面向AI时代数据中心，基于灵衢（UnifiedBus，简称UB）的新型计算系统架构，支持CPU、NPU、GPU、MEM、DPU、SSU（Scalable Storage Unit）和Switch中的一种或者多种组件资源池化和平等协同，构建逻辑上的一台计算机。

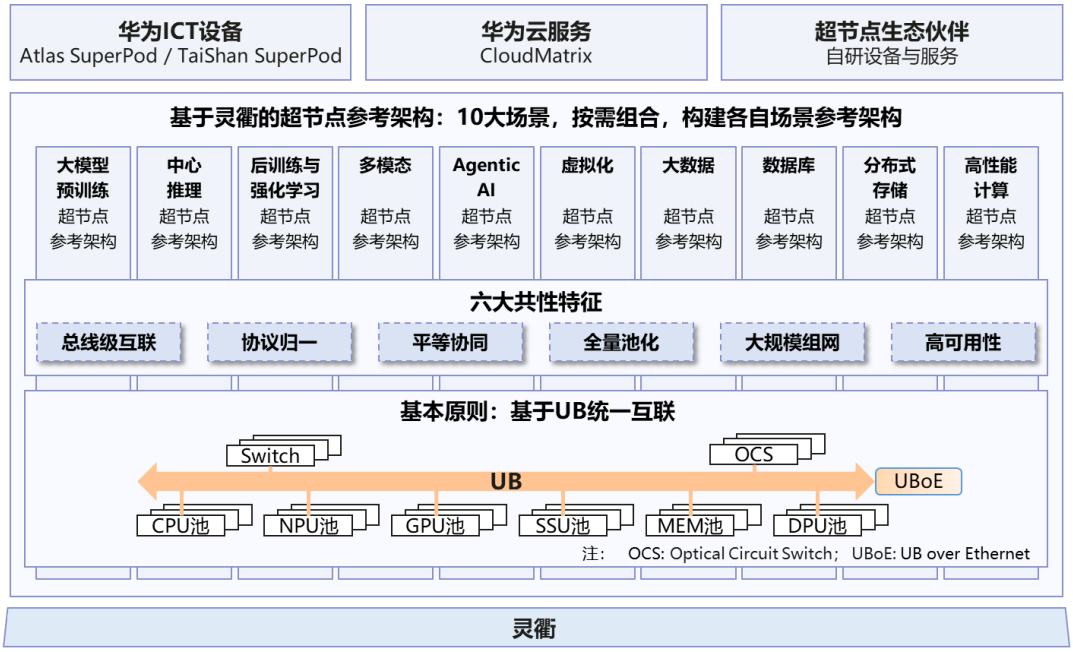


图 2-1 基于灵衢的超节点参考架构

基于灵衢的超节点参考架构具备如下六大特征：

1. 总线级互联

基于灵衢的总线级互联，提供百ns同步内存语义访问时延和2~5us异步内存语义访问时延，满足算力单元高并发的访问需求；提供组件间TB/s级带宽，相比传统数据中心网络带宽至少提升10倍。

2. 协议归一

基于灵衢的协议归一，支持超节点内不同类型、不同距离的组件统一互联，访问无协议转换开销，组件包括CPU、NPU、GPU、MEM、DPU、SSU和Switch等；提供统一的编程模型。

3. 平等协同

基于灵衢的平等协同机制，支持超节点内所有组件去中心化的互相访问、调用和协同工作，提升组件间访存和通信性能。

4. 全量池化

基于灵衢和Linux操作系统的灵衢扩展组件，提供超节点的设备管理、内存管理、通信和虚拟化等功能，支持超节点资源的高效池化管理和调用，提升资源弹性和利用率。

5. 大规模组网

支持超节点以大于90%的线性度从单节点扩展到8192卡，未来还将持续提升至15488卡，甚至更大规模；支持超节点通过UBoE构建百万卡规模的集群，兼容以太组网。

6. 高可用性

基于灵衢的可靠机制，支持超节点内应用无感知的us级检错和容错，在8192卡超节点范围内实现光互连MTBF（Mean Time Between Failures）大于6000小时。

灵衢是一种面向超节点的互联协议，将I/O、内存访问和各类处理单元间的通信统一在同一互联技术体系，实现数据高性能传输、算力高效协同、资源统一管理和灵活组合，是超节点参考架构的基础。

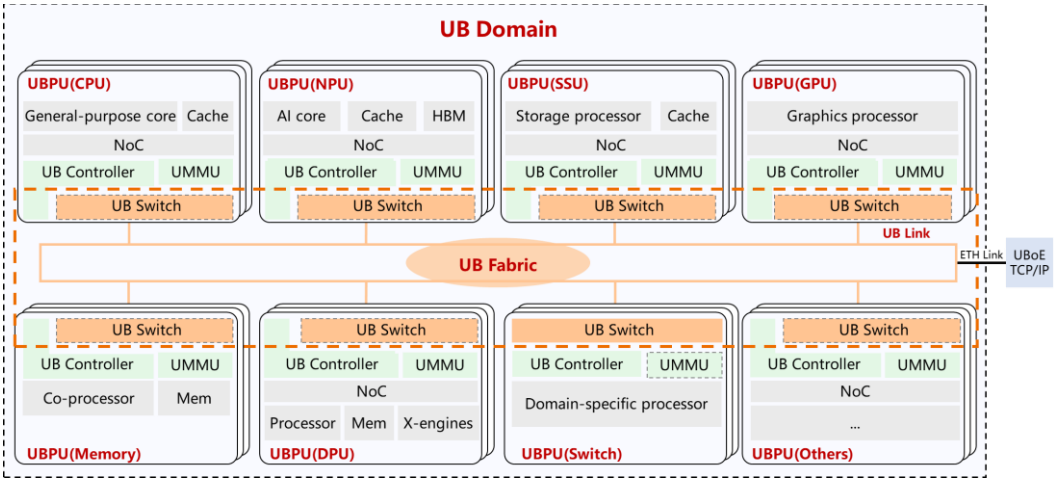


图 2-2 灵衢部署示意

灵衢包含以下要素：

- UB Processing Unit（UBPU）是支持UB协议栈的处理单元，实现特定功能。
- UB Controller是UBPU中执行UB协议栈的组件，并提供软硬件接口。

- UB Memory Management Unit（UMMU）是UBPU中执行内存地址翻译和访问权限控制的组件。
- UB Switch是Switch中的必选组件，在其他UBPU中是可选组件，支持在UB端口间转发报文。
- UB Link是UBPU间的点到点连接。
- UB Domain是一个全部使用UB Link连接起来的UBPU集合。
- UB Fabric是UB Domain内所有UB Switch和UB Link的集合。
- UB over Ethernet（UBoE）通过以太/IP网络承载UB事务，实现跨UB Domain互通。

基于灵衢的超节点参考架构帮助您迎接智能化时代算力基础设施挑战。

3 超节点参考架构的场景化应用

以灵衢为基础构建的超节点参考架构，在面向人工智能计算与通用计算领域的10大核心业务场景，如大模型预训练、中心推理、后训练与强化学习、多模态内容理解与生成、Agentic AI、虚拟化、大数据、数据库、分布式存储和高性能计算等，均可提供领先的系统能力，带来计算业务性能和资源利用率提升。

3.1 大模型预训练

为了提升大模型性能，业界主流基础大模型训练的参数量和数据量快速增长，从早期的数十亿参数发展到如今的万亿级别，如：2024年12月发布的DeepSeek V3参数量达到671B，2025年7月发布的Kimi K2参数量突破万亿大关。大模型预训练的业务负载呈现高并行特征，单一训练任务横跨整个算力中心的所有节点，随着预训练所需要的算力规模持续增长，对计算基础设施高效并行提出挑战：

- **模型参数增长需要多卡并行部署，带来大量集合通信开销**

以GPT4 1.8T为例，模型部署需要超10TB显存占用，远超GPU单卡甚至单服务器容量上限。为了解决这一问题，业界普遍采用张量并行（TP）、流水线并行（PP）、专家并行（EP）、序列并行（SP）和数据并行（DP）等技术进行分布式部署，但会带来巨大通信开销，当采用TP8/PP16/EP8/SP4/DP64@seq8k/bs8k的超参设置时，单次Global Batch Size迭代过程中，每张卡的TP通信量高达336GB，EP通信量为268GB，DP通信量为86GB。如此大量的集合通信对网络基础设施提出了极致低时延和大带宽要求，特别是在并行域内的通信需求更为迫切。

- **传统服务器训练集群节点间通信带宽不满足多卡并行集合通信要求**

近8年间单节点算力增长约40倍，但节点间通信带宽仅增长4~8倍，通信与计算能力的不匹配成为制约大模型训练效率的关键瓶颈，高比例的通信开销无法被计算有效掩盖，严重影响了模型的算力利用率（MFU），使得昂贵的计算资源无法得到充分利用。

大模型预训练超节点参考架构实现NPU-to-NPU的大带宽低时延互联，降低了大模型训练过程中高度并行化部署所带来的通信开销，提升了大模型训练性能。系统在混合专家（MoE）模型、长序列输入等复杂场景下均能取得性能收益。DeepSeek V3为例：在PP8/DP64/EP64典型配置场景下，超节点集群相比传统服务器集群，未被掩盖的通信比例降低80%，在单NPU 2倍裸算力增加的情况下，实现系统最高3倍以上的训练性能提升，充分证明先进架构对大模型训练效率的重要价值。

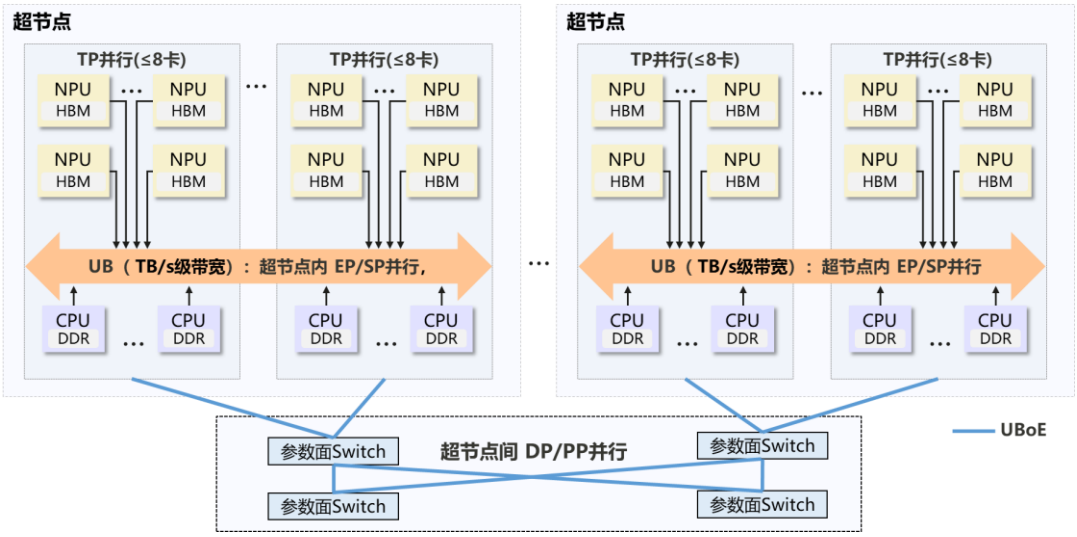


图 3-1 大模型预训练超节点参考架构

3.2 中心推理

主流大语言模型（LLM）向万亿参数稀疏化演进，总参数量与专家数量不断增长，如：DeepSeek V3 671B/256专家和Kimi K2 1.04T/384专家；推理模式也从单卡单机走向多机大专家并行。同时，大语言模型推理序列长度持续增长，豆包1.6模型成为国内首个支持256K序列的商业模型。随着AI应用持续发展，多图+视频等多模态理解类应用预计驱动序列长度进一步增加到M级别。针对大专家并行和长序列输入的发展趋势，中心推理场景的算力基础设施面临如下关键挑战：

- **多专家低时延通信挑战**

多机大专家并行推理，小数据高频通信占比高，以DeepSeek V3为例，每Token推理，58层MoE需要58轮Dispatch和Combine跨节点动态通信，最小粒度仅为7KB，通信量随Batch Size线性增加，最大可达百MB。推理服务要兼顾单用户体验与多用户并发开销，需在TPOT 10ms甚至更低情况下尽量做大Batch Size，那么通信比例应控制在20%以内，这样单次Dispatch和Combine耗时均需小于10us，RoCE技术挑战很大。

- **长序列 KV Cache 缓存挑战**

稀疏化注意力是解决长序列推理的关键技术之一，在推理过程中需要保存用户的全量KV Cache数据，并卸载到节点内Memory或Nand Flash，大幅提升Batch Size及推理吞吐。如：基于DeepSeek V3模型64~128K序列场景测算，通过卸载用户KV Cache，提升4~6倍的Batch Size和30%~40%的推理吞吐。

中心推理超节点参考架构一方面利用NPU-to-NPU的大带宽、低时延和高效统一语义能力实现高性能Dispatch和Combine通信，大幅降低多专家通信时延；另一方面基于全局资源池化架构提供多层次KV Cache卸载和加载能力，纵向多层级介质扩容缓存空间，横向NPU-to-NPU直通加速传输，以查代算提升计算吞吐。

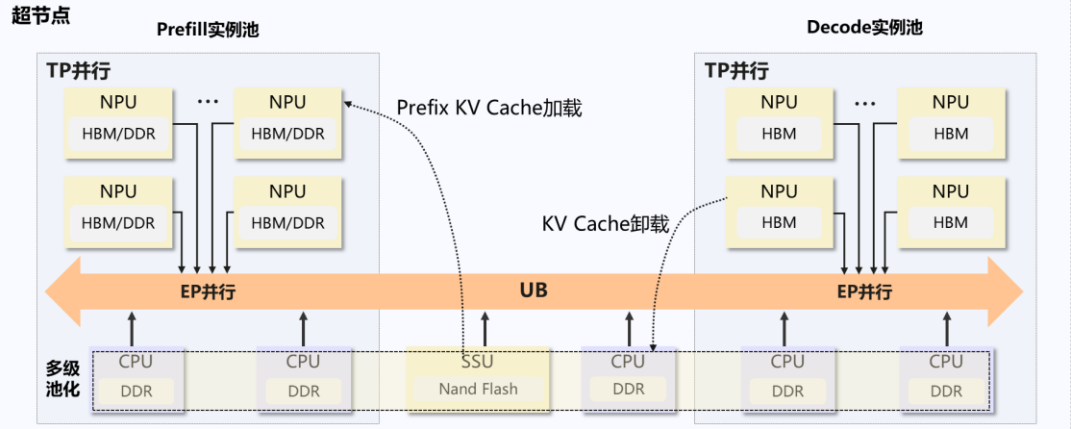


图 3-2 中心推理超节点参考架构

3.3 后训练与强化学习

在大模型强化学习训练场景中，随着语言模型规模的不断扩大和人类反馈优化需求的提升，计算和通信挑战变得更加复杂和严峻。人类反馈强化学习（RLHF）训练流程通常包括监督微调（SFT）、奖励模型训练（RM）和强化学习优化（PPO）三个阶段，每个阶段都需要百亿甚至万亿参数规模的大模型。以ChatGPT和Claude等先进对话系统为例，其RLHF训练过程需要同时维护策略模型、价值模型、奖励模型和参考模型等多个大型神经网络，总参数量可达数万亿级别，对计算资源和通信带宽提出极高要求：

- **大量模型参数与梯度同步通信挑战**

大模型强化学习训练的独特挑战在于其多模型协同和频繁同步的特征，与传统预训练不同，RLHF的PPO阶段需要在策略网络推理、奖励计算、优势估计和策略更新之间进行复杂的数据流转。在分布式RLHF架构中，Actor节点生成响应文本，Critic节点计算价值函数，Reward节点评估响应质量，Learner节点执行策略优化，这些组件之间需要进行大量的模型参数传输和梯度同步。

传统服务器堆叠集群在处理RLHF的多模型通信模式时面临瓶颈，大量参数和梯度信息频繁交换，网络带宽很快成为限制因素。模型间复杂的依赖关系导致通信时延被多次放大。以GPT3的训练配置为例，当使用175B参数的策略模型和6.7B参数的奖励模型时，单次PPO迭代的模型间通信量可达500GB以上，传统服务器堆叠集群的网络带宽瓶颈往往导致整体训练效率下降40%~60%。

- **大批量样本生成与并行推理性能挑战**

大规模RLHF训练时，使用当前策略模型生成大量候选响应，然后通过奖励模型进行评分和排序，通常需要为每个提示生成数十个候选响应，可达数万个样本，这要求系统能够高效处理大规模的序列生成和并行推理任务。同时，由于奖励模型的评分结果需要及时反馈给策略优化过程，任何通信时延都会直接影响训练的收敛速度和稳定性。

大模型强化学习超节点参考架构基于NPU-to-NPU的大带宽显著降低了多模型组件之间的参数同步时延，将原本数秒到数十秒级的模型传输时间降低到毫秒到秒级别，使得PPO和GRPO等算法能够更频繁地进行策略更新，提高学习效率和收敛速度。同时，超节点可支持数百卡以上的大规模并行推理和生成，在相同硬件资源下可以处理更大的Batch Size，从而提高样本生成的吞吐量。在大规模对话模型的RLHF训练中，超节点上部署的四模型系统相比传统服务器堆叠集群，训练的收敛速度有倍级提升，也显著提升了推理吞吐量。以70B参数模型的RLHF训练为例，传统服务器堆叠集群在处理32K Batch Size的候选生成时，推理时延通常在数十秒级别，而超节点系统可以将这一时延降低到秒级，提升样本生成效率。

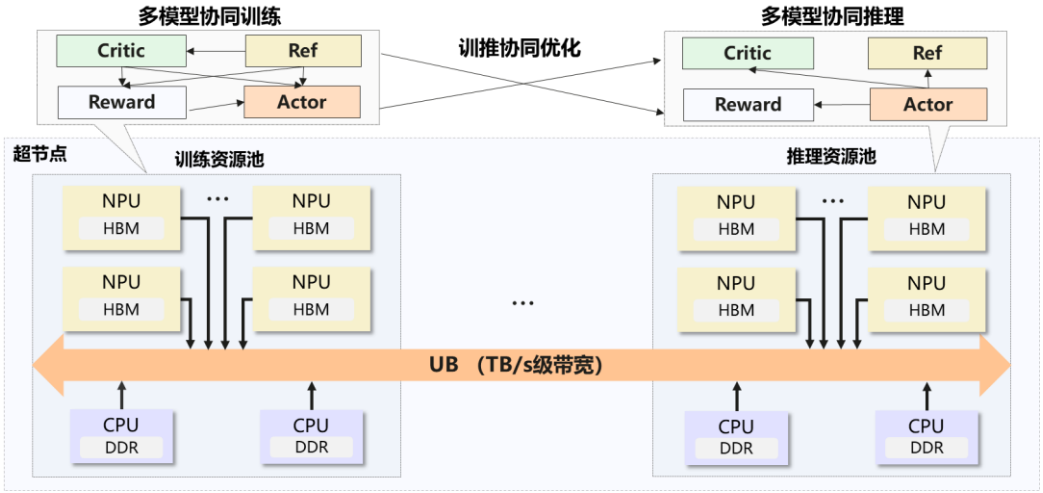


图 3-3 强化学习超节点参考架构

3.4 多模态内容理解与生成

多模态内容理解和生成是当前AI关键应用，演进趋势包括：1) 多模态模型向百B稀疏+实时生成（如5s生成5s 720p视频）演进，部署形态由单卡单机走向多机并行；2) 多模态负载从单一模型/任务走向多模型/多任务，不同任务（如生成/编辑/理解）对算力及访存的需求相差较大，驱动部署方式走向高算力与高带宽NPU组合异构部署；3) 自动驾驶/具身智能多模态仿真生成场景，3D高斯溅射（3DGS）技术成为主流，其中千万级高斯球+分钟级3DGS重建是重要场景。基于演进趋势，算力基础存在以下挑战：

- **多机并行多模态推理挑战**

序列并行和专家并行等策略，需百GB/s多机并行带宽，当前网络难以满足。

- **NPU 异构部署多任务并发挑战**

多任务需要在Token级别对模态进行分组和Re-order，实现多任务计算量动态分配，存在大量小包通信需求，导致通信时延要求高。

- **高性能分布式 3DGS 挑战**

3DGS重建在千万级高斯球场景，单轮迭代需压缩至数十毫秒，同步百万级椭球梯度（位置/协方差/透明度等），GPU间通信需求达到百GB/s。

- **大量分散高斯球的梯度传输挑战**

梯度数据可能分散在离散的显存空间中，需要大量细粒度（如数十个Byte）通信行为，需要支持小包低时延通信。

多模态理解与生成超节点参考架构提供突破性解决方案。GPU资源池通过UB将多个GPU组成逻辑统一资源池，提升分布式3DGS重建的生成效率；NPU异构部署通过UB支持不同类型NPU异构组网，提升资源动态调度能力（如动态配比和动态并行策略寻优等）及灵活性；内存语义通信通过UB实现小包超低时延通信，支持多任务推理Token级别分组和Re-order，提升推理效率，并支持3DGS分散高斯球梯度快速同步，提升3D渲染视图生成效率。

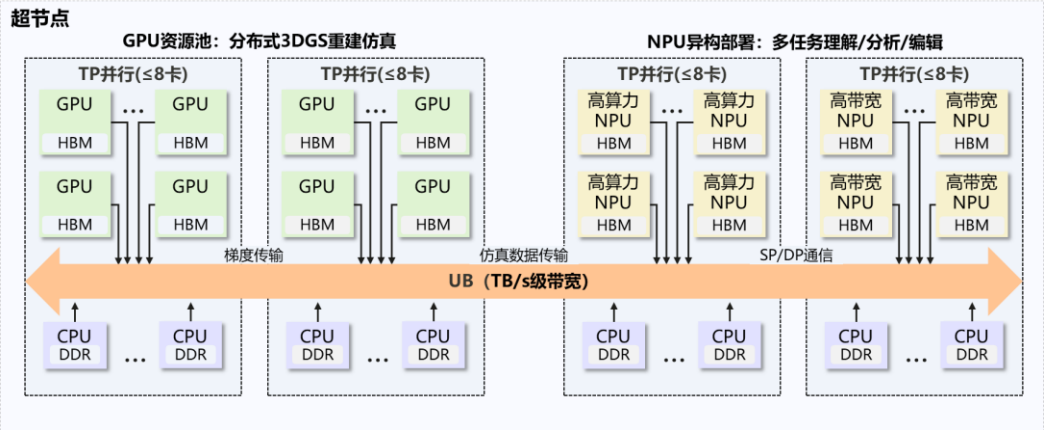


图 3-4 多模态理解与生成超节点参考架构

3.5 Agentic AI

Gartner将Agentic AI列为2025年十大技术趋势之首，并预测到2028年，全球15%的日常工作决策将由具备自主决策能力的AI代理完成，AI逐步演变为能够自主感知、推理、决策并执行任务的智能体Agentic AI。如：进一步融合多模态能力（文本、图像和语音等）和物理世界交互（机器人控制），推动AI从“被动响应”转向“主动行动”。Agentic AI与互联网结合，形成“分布式智能体网络”，支持大规模协作。

- **Agentic AI 带来新的业务负载挑战**

在Agentic AI中，多个LLM共同协作增加了系统的复杂性，结合了训练、推理、通算和数据库管理四个核心环节，除训练、推理、强化学习和多模态等方面的挑战外，还引入长链条推理和长期记忆带来的存储及访存挑战，长尾任务与异构硬件阻塞带来的资源利用率难题，以及Agent多任务多副本带来的快速上下文切换与频繁镜像拉取开销。

- **Agentic AI 负载带来新的计算系统挑战**

Agentic AI的负载Scaling模式的升级换代给系统设计带来了挑战，包括内存需求激增、资源调度复杂化、通信随机性增加，以及存储持久化的管理问题。随着多样化环境的动态变化，系统必须适应不断扩展的负载，确保高效的内存使用、精确的资源调度和持续的数据存储，满足长期运行的需求。

Agentic AI超节点参考架构通过多样性算力平等协同计算实现多模态训练数据高吞吐预处理、检索增强生成（RAG）及慢思考推理等高吞吐优化；基于UB全局资源池化无损热迁移，应对Agent中长期记忆带来KV Cache、RAG等异构数据检索指数级增加的挑战；基于UB实现沙箱镜像的毫秒级热频繁加载，实现Agentic AI负载亲和。

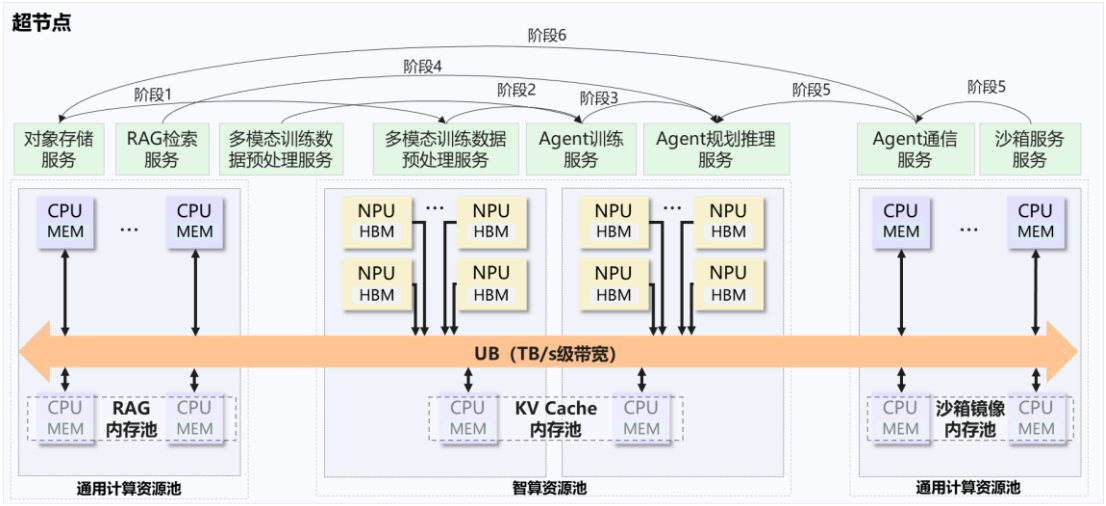


图 3-5 Agentic AI 超节点参考架构

3.6 虚拟化

传统云计算虚拟化技术通过计算、存储和网络资源的抽象，提升数据中心部署灵活性与资源利用率，但实际应用场景仍面临以下关键挑战：

- **资源碎片化降低利用率**

虚拟机的动态创建与销毁导致物理服务器在CPU、内存及存储资源上产生碎片。传统云服务20%~25%内存搁浅，无法有效分配与使用，致使整体资源利用率低于预期水平。

- **虚拟机热迁移效率瓶颈**

高负载虚拟机迁移时，过高的内存脏页率会显著延长迁移时间甚至导致迁移失败。同时，目标主机需满足连续资源需求，资源碎片化问题阻碍迁移的顺利完成。

- **内存资源利用率低下**

受限于内存资源的不可压缩性与实时访问要求，传统云平台普遍采用预留分配机制，导致内存实际利用率低，接近半数虚拟机30%~50%内存未被使用，导致资源浪费。

- **固定资源配置限制灵活性**

通用服务器预设的CPU与内存配比，如1:4，难以适配内存密集型或计算密集型应用的差异化需求，造成资源配置失衡。

虚拟化超节点参考架构，可实现超过25%的内存资源超分，提升资源利用率，为新一代云/虚拟化平台提供优化解决方案：

✓ **资源弹性分配**

基于UB的全局内存资源池，实现超节点内资源灵活调配，内存分配率从80%提升至95%，提升资源调度弹性。

✓ **高效无损热迁移**

基于UB提供的大带宽低时延通信，结合统一语义技术，实现低至50ms的虚拟机极速热迁移，有效规避因内存超分导致的内存不足风险。

✓ **透明内存冷热分级**

基于UB的内存页面访问统计机制，实现高效透明的冷热内存识别与交换，性能损失控制在5%以内，提升内存资源利用率。

✓ **超大内存实例支持**

基于UB，构建超大规格内存（大于3TB）的云主机实例，满足内存密集型服务的多样化资源需求。

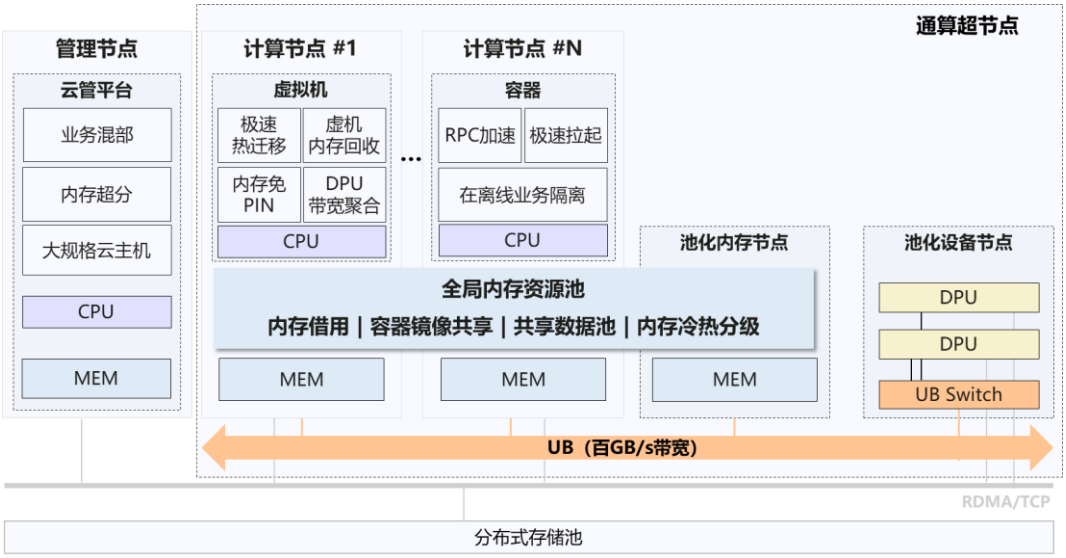


图 3-6 虚拟化超节点参考架构

3.7 大数据

传统大数据平台普遍面临两大关键性能瓶颈，制约了数据处理效率和资源使用效益：

● **资源配置失衡与浪费**

静态的资源分配机制导致资源配置与实际负载需求脱节，资源使用不均，极端场景下，仅30%~40%内存资源被有效使用。同时，为避免任务失败，运维人员倾向于过度申请资源，进一步造成资源闲置。

● **数据交换效率制约整体性能**

跨节点数据混洗（Shuffle）过程中的数据读写与交换效率成为系统性能的主要瓶颈，

网络与磁盘I/O压力，导致CPU的平均使用率仅为峰值的50%。

大数据超节点参考架构通过大带宽低时延通信和全局资源池化等，提供优化方案：

✓ **资源池化与智能超分**

基于UB在超节点域内实现灵活的内存池化借用与共享，结合大数据负载的资源超分策略与池化算子优化，Spark任务执行效率提升30%。

✓ **Shuffle 性能深度优化**

基于UB通信对Spark shuffle进行深度优化，提升10%的Spark处理性能，缩短作业批处理完成时间。

✓ **高效 RPC 通信机制**

基于UB重构远程过程调用（RPC），实现共享内存和大带宽低时延通信，降低序列化和反序列化开销，减少内存拷贝，为实时计算等场景提供极速高效的通信保障。

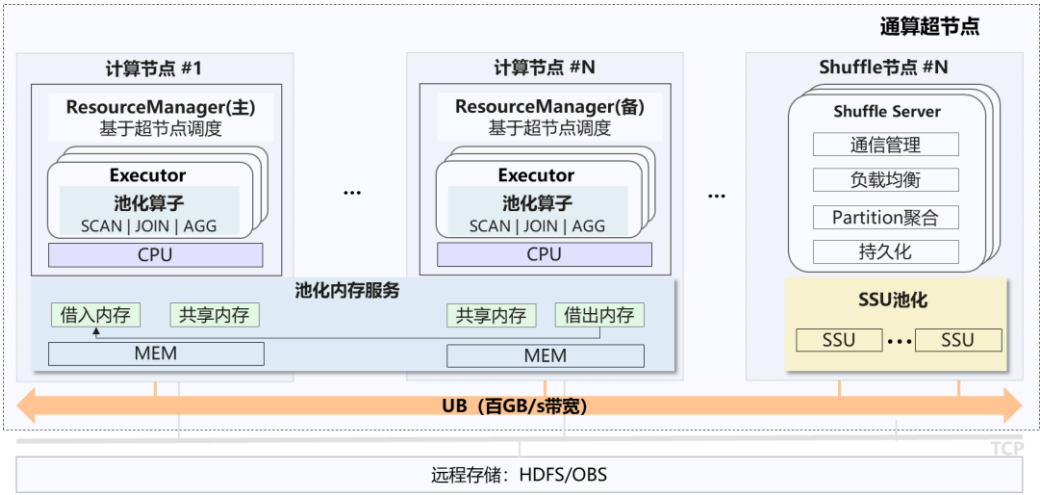


图 3-7 大数据超节点参考架构

3.8 数据库

随着算力需求持续增长，数据库应用在传统计算架构下，面临如下问题：

● **架构局限性与扩展瓶颈**

传统主从架构所有写操作集中于主节点，使其成为性能瓶颈和故障单点，线性度小于70%，制约系统的写扩展性与高可用性。

● **网络与 I/O 性能制约**

跨节点数据同步、分布式事务协调及日志复制带来网络开销和带宽压力，对时延高度敏感，网络协议栈开销最大超过33%，造成CPU资源大量浪费。日志同步的I/O负担进

一步制约了高并发和高吞吐场景下的性能。I/O瓶颈和缓存不命中导致长尾SQL，响应时间大于10ms。

数据库超节点参考架构提供的解决方案包括：

✓ **实现多主架构**

依托UB的全局资源池化和大带宽低时延通讯，实现共享内存服务，支持数据库多主架构。有效消除写性能瓶颈，提升写扩展能力，线性度大于0.75，避免写节点单点故障，支持数据库实例的动态在线扩容，增强系统可靠性与灵活性。

✓ **提升 OLTP 性能**

利用UB全局资源池化与统一语义访问，扩展数据库缓冲池（全局Buffer Pool），构建高效的多级页面缓存。结合本地与远端内存间透明的冷热数据迁移，提升热数据的本地内存访问命中率， TPC-C性能提升20%。

✓ **优化 OLAP 性能**

基于UB全局资源池化与统一语义访问的共享内存服务，加速OLAP SQL算子执行效率，实现行列混合存储引擎间的数据高效共享，提升OLAP查询性能， TPC-H性能提升65%。

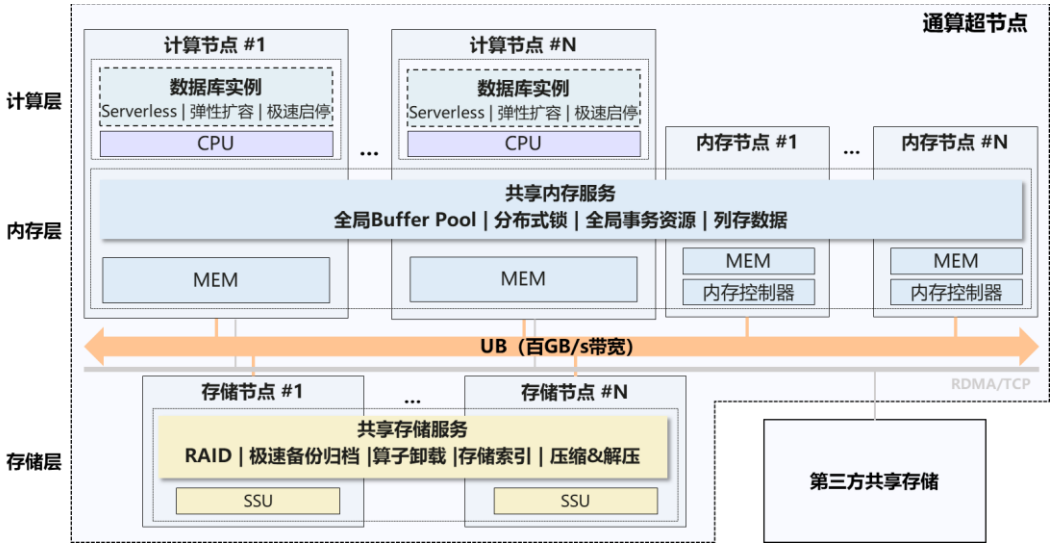


图 3-8 数据库超节点参考架构

3.9 分布式存储

传统分布式存储存在以下问题：

● **硬件资源离散化**

传统存储每台服务器的SSD容量被独立分配，资源利用率受限于单机物理边界，SSD的

带宽利用率约为25%，造成基础设施效率低下，运维复杂。

- **软硬协同能力薄弱**

以CPU为中心的架构设计存在根本性缺陷：EC编解码、CRC校验、数据跨节点搬移等关键任务抢占主机CPU算力，使其成为系统性能瓶颈，制约分布式存储扩展能力。以分布式存储系统垃圾回收任务为例，CPU将数据从源盘读出并重新写入目的盘，期间产生多次网络栈和I/O栈开销，挤压业务侧处理能力，导致有效IOPS下降25%~35%。

分布式存储超节点参考架构基于多样性算力平等协同计算，带来技术创新：

✓ **存储池化架构**

通过UB全局资源池化技术与统一语义访问能力，构建跨节点存储资源池。实现物理介质的逻辑抽象化与统一调度，支持按需动态分配多节点存储容量及性能资源。消除资源孤岛，SSD带宽资源利用率最高提升60%。

✓ **异构算力卸载引擎**

依托多样性算力平等协同计算架构，打通CPU、GPU和SSU等设备的平等互联访问通道。通过硬件级业务卸载引擎，将存储系统侧的垃圾回收等从主机CPU卸载至SSU，实测IOPS性能提升25%~35%，为高密存储场景提供性能保障。将存储系统侧的EC（Erasure Coding）重构任务从主机CPU卸载至SSU中，结合UB超大带宽能力，重构速度提升6倍，为高密存储场景提供可靠性保障。

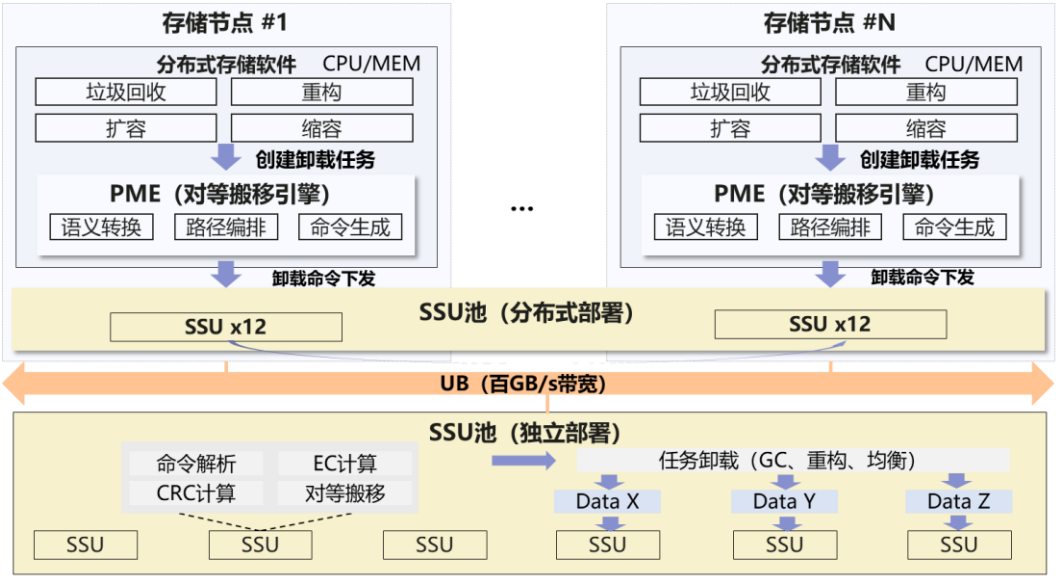


图 3-9 分布式存储超节点参考架构

3.10 高性能计算

随着科研应用计算需求持续攀升，高性能计算集群算力规模持续扩大。同时，近十年人工智能技术的快速发展，催生出AI4S（AI for Science）这一新兴应用范式，带来高性能计算与智算融合，对高性能计算架构和性能提出了更高挑战。

- **通信性能挑战**

随着处理器计算能力不断提升，高性能计算应用通信占比将进一步增加，部分应用可达30%，为降低应用通信占比、提升应用性能和扩展性，面向大包通信场景需提升节点间通信带宽，面向小包通信场景需降低传输时延。AI4S场景，训练和推理存在大量分布式矩阵计算，产生有依赖的计算通信算子。为实现计算和通信隐藏，通常采用矩阵和通信细粒度流水，但会引入KB级中小包通信，导致通信性能劣化，需降低通信时延。

- **超大规模集群面临网络拥塞和组网可靠性挑战**

业界领先的高性能计算集群算力规模已达到E级，正在向10E级演进，需要支持万级到10万级节点超大规模组网。超大规模集群会面临网络拥塞问题，需通过路由和拥塞控制算法优化提升性能。同时，传统胖树网络拓扑面临光模块数量多和网络可靠性挑战，需通过网络拓扑优化提升组网可靠性，减少光模块数量。

- **存储 I/O 过程涉及多次内存拷贝，影响性能**

传统高性能存储采用专有客户端架构，数据落盘存在多次内存拷贝：数据首先从计算节点应用侧内存复制到客户端内存，再通过网络传输至存储节点内存，最后写入硬盘。多次内存拷贝会导致I/O时延增加，需要减少I/O路径的内存拷贝次数，降低I/O时延。

高性能计算超节点参考架构，面向下一代超大规模组网、AI4S场景提供技术创新：

- ✓ **软硬协同通信性能优化**

通过UB提供百GB/s~TB/s级互联带宽，结合拓扑感知的通信优化和作业调度，在分子动力学、气象和流体等通信密集型应用场景实现大包通信加速，典型应用E2E性能可提升10%。同时，通过在超节点内实现基于内存语义的通信优化，使小包通信时延从us级降低至百ns级，典型应用E2E性能可提升8%。面向AI4S应用的分布式矩阵计算场景，通过基于内存语义的融合算子，实现计算和通信细粒度流水线，支持直接访问对端NPU的HBM数据，减少访存次数，降低小包通信时延，提升训练推理性能。

- ✓ **超大规模组网拓扑和网络拥塞控制优化**

高性能计算场景存在大量邻居通信，超节点内可基于UB，通过FullMesh/Torus+Clos拓扑提升进程间通信性能；超节点间可采用较高收敛比的网络架构，支持超大规模组网，

结合UB链路层虚通道避免直连路径绕路引入的死锁，通过时空均衡技术充分利用最短和非最短路径、逐包/逐流动态路由技术，减少网络拥塞，提升网络带宽利用率，实现高性能、大规模和高可靠的组网。

✓ **基于 UB 的数据直通机制和统一存储架构**

通过UB实现计算节点数据直通存储硬盘，不经过存储节点内存，减少内存拷贝带来的额外开销，I/O性能提升20%。面向AI4S场景，智算超节点和高性能计算超节点可通过UB访问共享的高性能分布式存储，减少数据搬移开销。

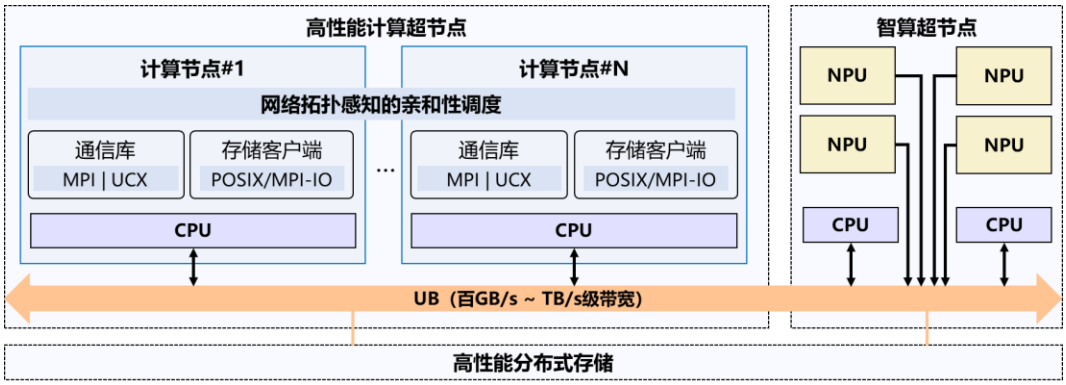


图 3-10 高性能计算超节点参考架构

4 灵衢协议栈和机制

4.1 灵衢协议栈

灵衢提供分层的协议栈，从下到上由物理层、数据链路层、网络层、传输层、事务层、功能层以及UMMU、UBFM（UB Fabric Manager）组成，如下图所示。其中，Entity为功能实体，是全局通信的基本单元；URMA（Unified Remote Memory Access）为统一远程内存访问。

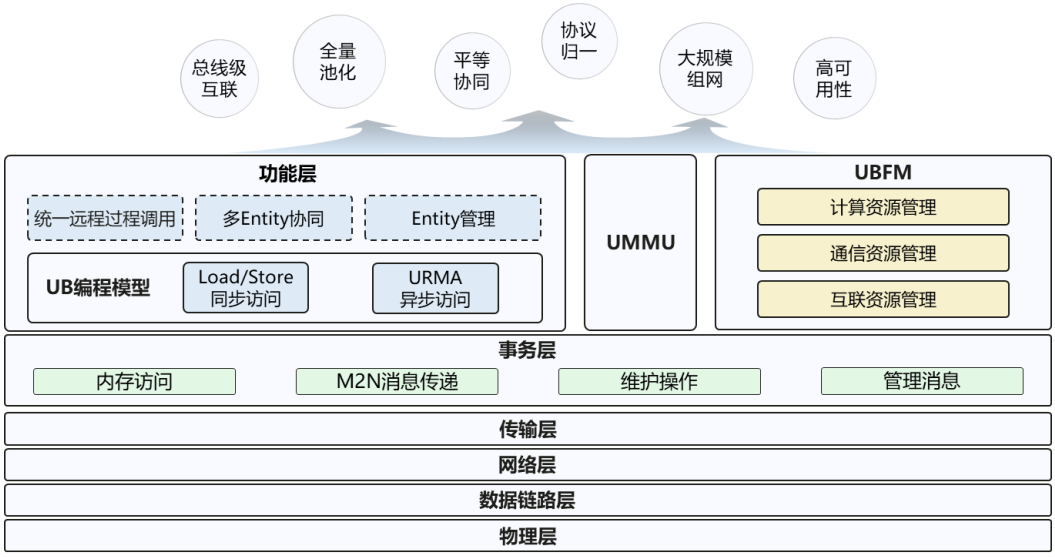


图 4-1 灵衢协议栈

- **物理层：**支持低时延和高速率设计，最大化利用物理信道裕量，支持线性直驱光技术。
- **链路层：**支持点对点流控和重传，提供高效传输格式并支持丰富拓扑结构。
- **网络层：**支持大规模组网、动态路由和多路径均衡，提供最优端到端时延和可用带宽。
- **传输层：**支持带宽聚合、连接共享和端到端可靠传输，实现大规模可靠扩展。

- **事务层：**提供统一的事务操作，支撑稳定的编程接口。
- **功能层：**提供简化的编程接口（如Load/Store和URMA等）和高效的协同机制。
- **UBFM：**支持大规模计算、通信和互联资源管理以及高效资源调用。
- **UMMU：**负责全局地址翻译和内存访问权限检查，实现全局内存共享。

详细灵衢协议栈能力，请参考《灵衢基础规范》。

4.2 灵衢使能超节点机制

4.2.1 总线级互联

- **百ns到us级内存语义时延：**基于统一编址和访问权限控制的基础机制，UB支持UBPU中的计算单元直接发起同步和异步访存指令，减少控制命令交互，实现百ns~us级低时延。UB内存语义支持64B至4KB大小以及变长的单事务操作，降低乱序处理压力，提升传输效率，同时基于链路层的Flit传输机制，实现低时延传输和转发。
- **TB/s级大带宽：**UB面向AI时代大带宽需求进行了单Lane增强速率和多端口多路径聚合带宽设计。通过物理层多种FEC模式和链路层重传技术，降低BER要求，实现单Lane速率增强至118Gbps，优于同代际IEEE Ethernet定义的标准速率。

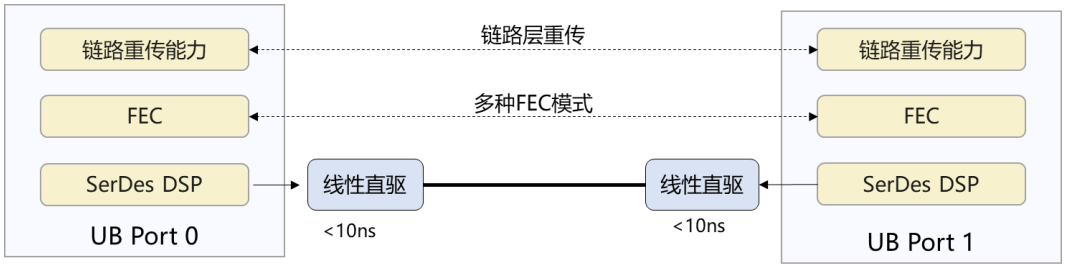


图 4-2 单端口高速率

通过多端口聚合和高密光电互连技术，实现UBPU间TB/s级带宽互连。UB支持Load/Store语义和URMA语义共享多端口带宽，实现多个端口间的多路径传输。

图 4-3 多端口聚合以及高密光电互连

4.2.2 协议归一

UB支持内存和寄存器等资源直接映射到全局地址，计算单元通过UB完成全局数据访问和全局同步；支持计算单元指令调度UB Controller，原生匹配计算单元乱序和并发执行的特征。UB从计算硬件系统角度，用户所有的操作归为内存访问、消息传递、过程调用和资源管理。UB使用单一协议栈支持上述所有操作，避免协议转换开销，为软件开发者提供利用硬件能力的高效编程方式。

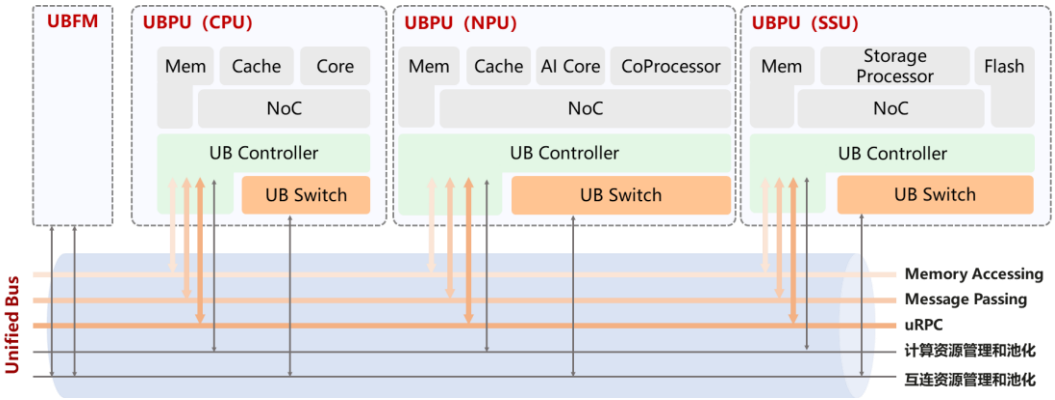


图 4-4 统一编程模型

UB事务层通过内存、消息、维护和管理四大类事务，为统一编程模型提供基础。内存事务用于UBPU间的内存访问，支持同步和异步两种方式；消息事务用于1对1、1对多、多对1和多对多的UBPU间消息传输；维护事务用于更新远端UBPU内部模块的状态，如访问安全状态和缓存状态等；管理事务用于实现UBPU配置以及运行过程中的状态上报，如地址配置和故障上报等。为了在不同场景和通信范围内获得最优性能，UB从时延、带宽和可靠性三个维度提供灵活的配置机制，如下图所示。

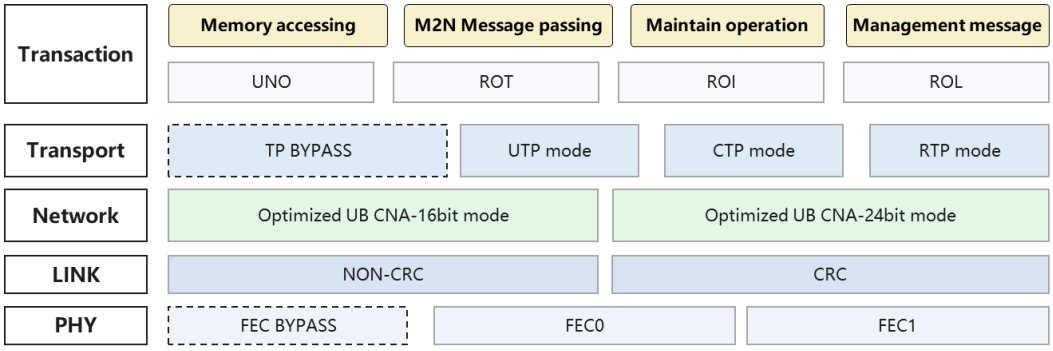
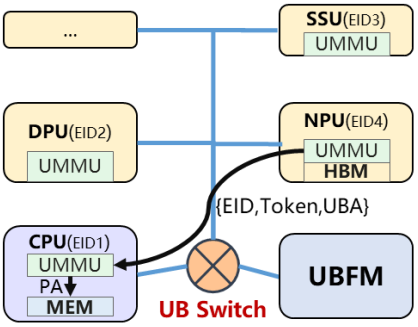


图 4-5 UB 协议配置模式

4.2.3 平等协同

UB支持超节点内所有组件全量平等协同。通过Entity和UMMU实现统一编址和访问权限控制，使能计算资源平等互联访问。每个UB Entity有一个唯一的身份标识EID（Entity ID），是全局通信的唯一标识。构建UMMU，支持内存访问鉴权和地址翻译，

设备间访问无需经过CPU。UBFM管理逻辑实体，负责Fabric域内设备EID、网络地址和路由等互连配置管理及设备资源管理。



通信范围	H2H	H2D	H2N	H2S	D2D	D2N	D2S
节点内	√	√	√	√	√	√	√
柜内跨节点	√	√	√	√	√	√	√
超节点内	√	√	√	√	√	√	√
集群内	√	√	√	√	√	√	√

EID：Entity ID；Token：访问凭证；UBA：UB地址；PA：物理地址　　H：CPU；　D：NPU；　N：DPU；　S：SSU

图 4-6 平等互联访问

4.2.4 全量池化

UB定义新的原生系统管理权限和消息，在超节点中通过UB带内方式完成资源管理；采用UB Partition的方式支持多用户，降低虚拟化开销，通过UBFM将一个UBPU的多个UB Entity注册给不同用户使用，以及将同一个用户的UB Entity加入同一个UB Partition，实现不同用户基于UPI（UB Partition ID）的隔离，如下图所示。

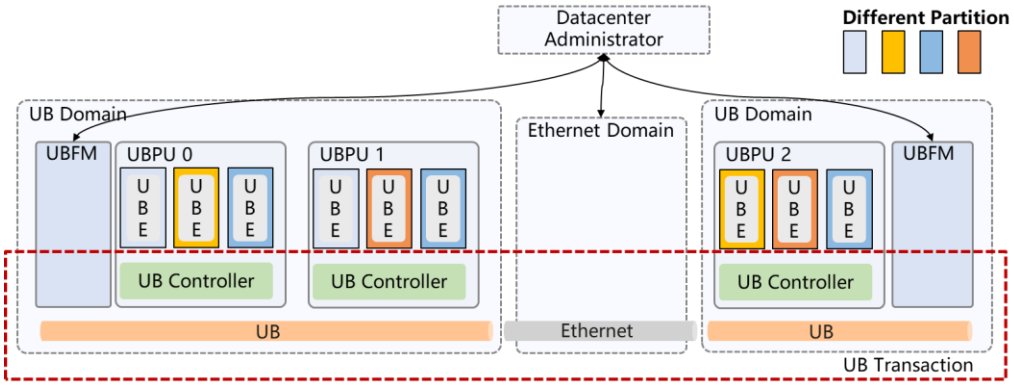


图 4-7 统一资源管理

UB支持将Entity的资源共享给多方，并通过MMIO（Memory-mapped I/O）和消息的方式直接访问和调用，例如UBPU将Entity 0的资源授权给Entity 1，UBPU和Entity 1均可直接通过MMIO或者消息的方式调用Entity 0声明的功能或者服务。

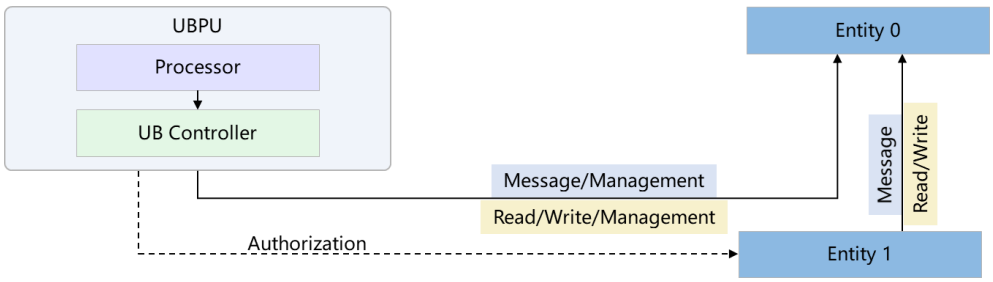


图 4-8 资源高效调用

除了计算和内存等资源的池化之外，UB通过多通道共享机制，实现TB/s级带宽的互联资源池化共享，所有Entity均可使用所有可达路径和端口。

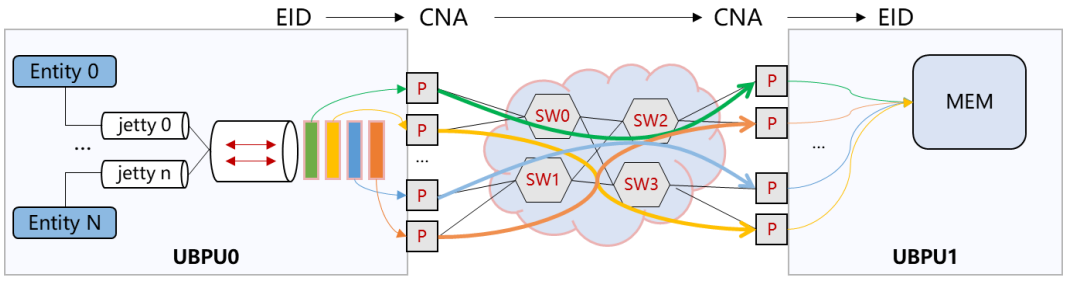


图 4-9 互联资源池化

4.2.5 大规模组网

UBPU内嵌UB Switch，支持UB报文通过UBPU直接转发至直连相邻的UBPU，无需软件中转，同时UB通过链路层虚通道、网络层逐包/逐流多路径路由和传输层通道共享等机制，解决UB报文通过直连路径绕路引入的死锁和带宽利用率下降等难题，为超节点提供 UB-Mesh （UB-Mesh: a Hierarchically Localized nD-FullMesh Datacenter Network Architecture，链接：https://arxiv.org/abs/2503.20377）以及基于光交换的组网技术，实现大规模低成本部署。

UB-Mesh中的nD-FullMesh拓扑充分利用了业务数据局部性，优先考虑短程直接互连路径，以最大限度减少数据移动距离并减少交换机使用为目标，是一种兼具高性能和低成本的拓扑组网，如下图所示。

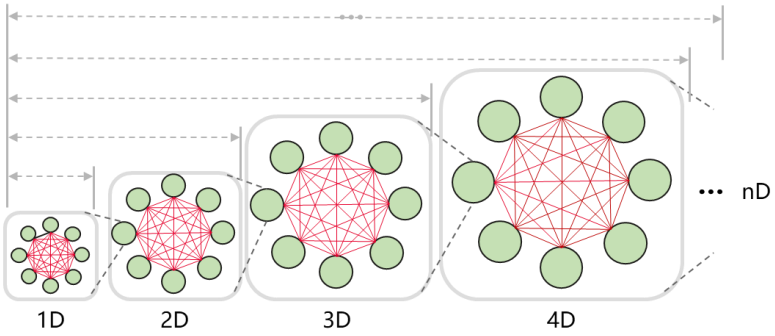


图 4-10 nD-FullMesh 拓扑示意

2D-FullMesh是nD-FullMesh在极致低时延场景的应用，减少交换引入的时延开销，可用于内存共享和内存借用等场景。

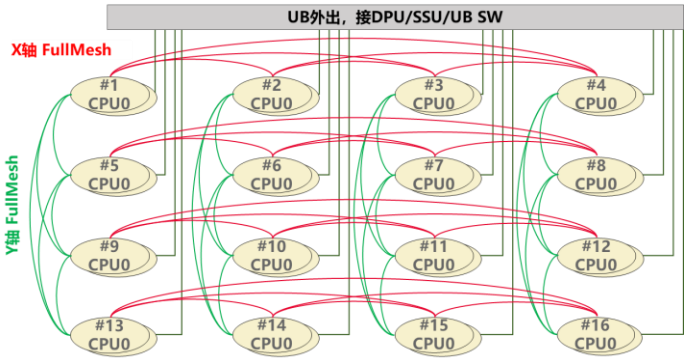


图 4-11 2D-FullMesh 直连拓扑示意

UB-Mesh还支持混合拓扑，例如在Rack内部采用1D/2D-FullMesh拓扑，提供全电缆互连的本地高带宽，在Rack间采用一层交换的Clos拓扑，提供适当收敛或者无收敛的带宽。该拓扑可用于训练和推理等场景。如下图所示，Rack内采用2D-FullMesh组网，Rack间采用一层UB Switch互连，支持从64卡线性扩展到8192卡。

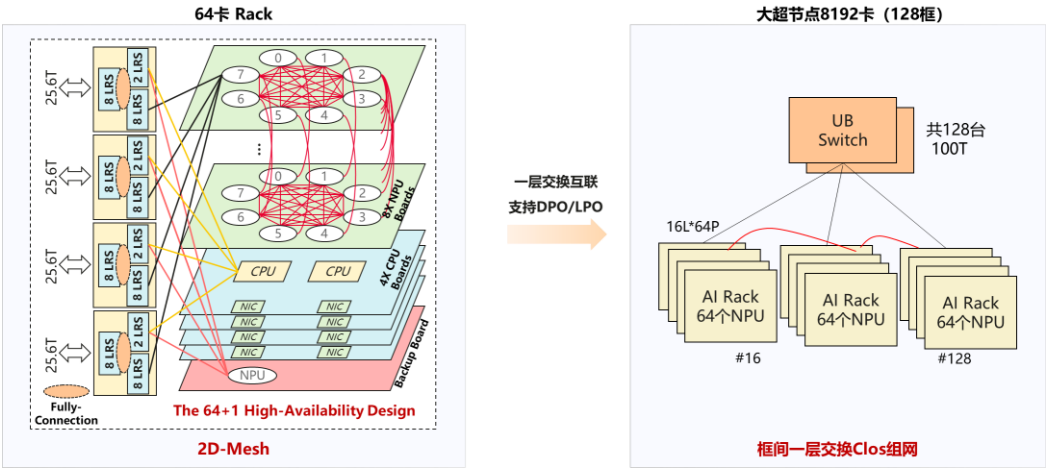


图 4-12 2D-FullMesh + Clos 混合拓扑示意

为了进一步扩大组网规模，UB除了支持采用多级UB Switch扩展组网之外，还支持通过UBoE与以太Switch对接，实现融合组网，以及通过OCS组网，实现可变拓扑，匹配业务动态流量。

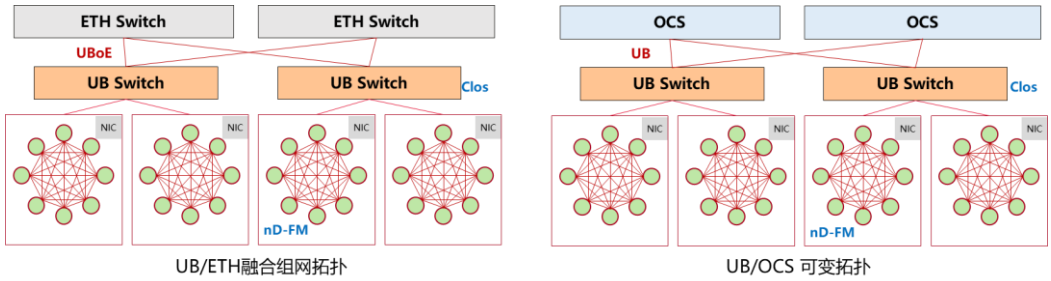


图 4-13 融合组网和光交换组网示意

4.2.6 高可用性

超节点高带宽互连引入数倍的光模块，对系统可靠性带来挑战。UB通过分层可靠协同机制，实现应用无感知的us级故障检错和容错。针对光链路误码闪断场景，通过链路层LLR（Link Layer Retry）技术实现重传；针对单Lane故障场景，通过物理层降Lane和链路层LLR技术协同，实现点对点零丢包；针对单个光模块故障场景，通过2+2光模块备份技术，实现业务无感知的故障恢复；以上技术均可满足内存语义可靠传输。同时UB支持网络多路径切换和可靠传输层，支持端到端的重传。

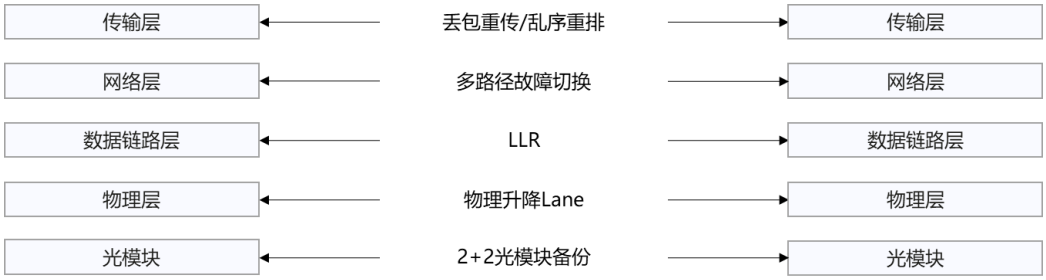


图 4-14 分层可靠机制

- **链路层重传：**UB链路层提供了重传Buffer，远端链路层检测到丢包会发送重传请求到源端，触发源端重新发送，该技术可以解决链路闪断和误码等场景丢包问题。

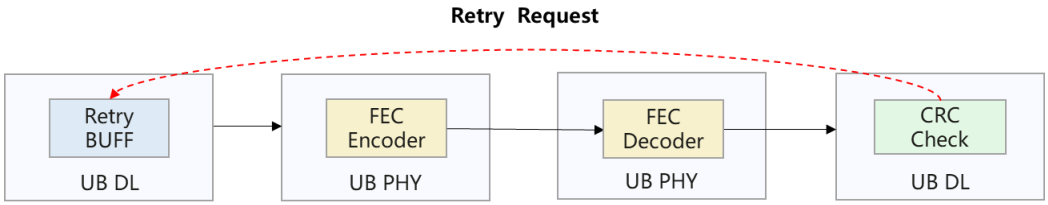


图 4-15 链路层重传

- **物理升降Lane和光模块备份：**UB物理层支持动态升降Lane，协同链路层重传，实现光模块2+2备份技术，如下图所示。

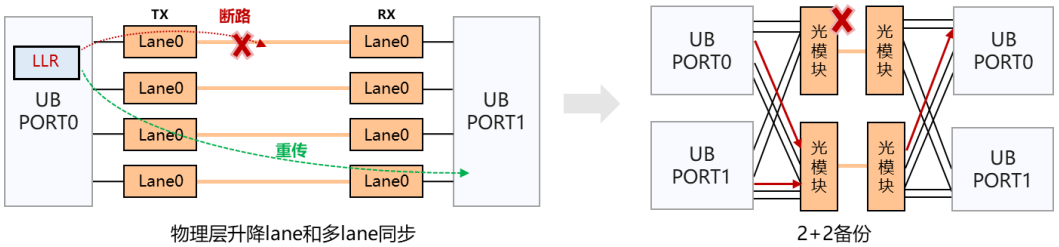


图 4-16 2+2 光模块备份技术

4.3 灵衢软件配套

操作系统只需增加部署灵衢组件软件配套即可扩展支持灵衢。操作系统灵衢组件（UB OS Component）通过异构硬件统一抽象解耦、多级异构资源池化管理、算力/内存/互联

资源策略调度和统一内存地址空间，实现超节点内的池化内存访问、资源全局调度、计算资源动态组合扩展和设备间高性能通信。

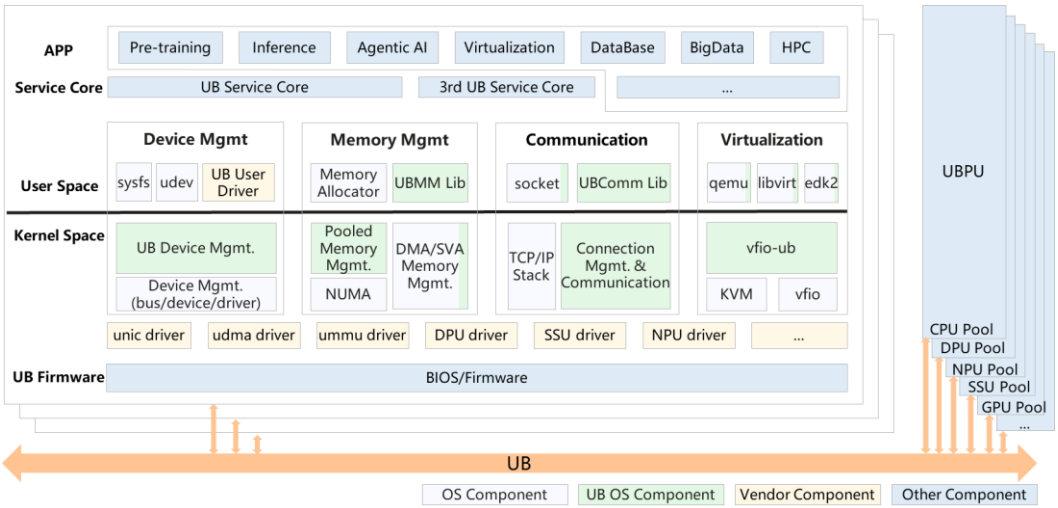


图 4-17 操作系统灵衢组件软件架构

操作系统灵衢组件扩展的4个功能分别是：

Device Mgmt.：提供UB互联和UB设备管理能力，实现计算节点内UB设备热插拔和配置。

Memory Mgmt.：提供超节点内内存语义访问能力，实现跨计算节点的内存借用和共享。

Communication：提供跨计算节点/跨设备的通信和远程调用功能。

Virtualization：提供UB设备直通虚拟机能力。

5 总结

灵衢已在Atlas 900 A3 SuperPoD等产品实践与验证，适合AI技术与产业发展。同时，基于灵衢的下一代际产品也即将上市，为了更广泛地促进互联技术发展和产业进步，华为决定将灵衢系列技术规范对业界开放，欢迎更多客户和伙伴基于灵衢和超节点参考架构实现更先进的算力基础设施，共同推动计算产业发展。

更多详细的信息，请参考灵衢系列规范，包括：《灵衢基础规范》、《灵衢固件规范》、《灵衢使能操作系统参考设计》。文档下载请访问灵衢官方网站：www.unifiedbus.com。