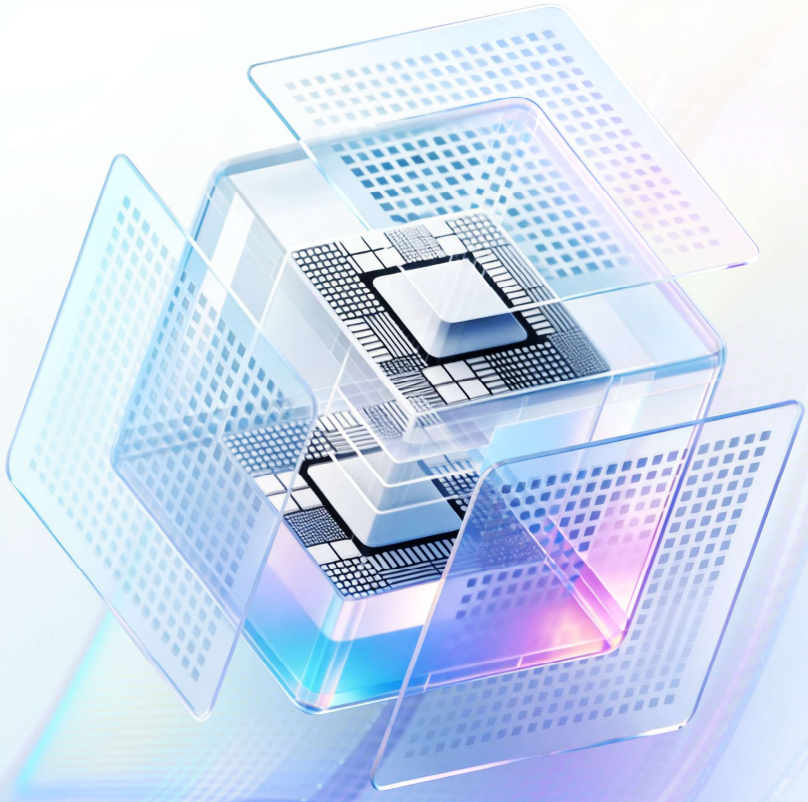
昇腾 950 NPU 架构白皮书



华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process

如企业客户须获取漏洞信息，请参见如下网址：

https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 关键术语

表1-1 关键术语

术语	描述
AIC	AI Cube Core。在 AI Core 分离架构下，一组 Cube Core 和 Vector Core 组合中的 Cube Core。
AIGC	Artificial Intelligence Generated Content，人工智能生成内容，指利用深度学习模型（如 GPT、Diffusion Models）自动生成文本、图像、音频、视频等内容的技术。
AIV	AI Vector Core。在 AI Core 分离架构下，一组 Cube Core 和 Vector Core 组合中的 Vector Core。
AI CPU	芯片内的自研 ARM 架构 CPU 内核，在昇腾 950 芯片中指自研 Linx816 CPU Core。
AI Die	昇腾 950PR 芯片和昇腾 950DT 芯片中的计算 Die。
CANN	Compute Architecture for Neural Networks，昇腾异构计算架构软件栈。
Clos	Clos 组网是一种基于多级交换的无阻塞网络架构，主要用于构建高性能、高扩展性的数据中心网络。其核心特点是通过多级互连和全连接拓扑实现任意节点间的无阻塞通信，同时支持水平扩展和成本优化。
CMO	Cache Maintenance Operations，通过 SDMA 实现的 L2 Cache 管理机制。
CTP	Compact Transport，Unified Bus 的轻量级传输层模式，借助下层协议共同提供可靠和拥塞控制的传输服务。
Device	Host-Device 架构的设备侧，本文指昇腾 950 系列 NPU 芯片。
Die	芯片中具体的晶粒（Die）描述，一般一个芯片中集成一个或者多个 Die。
DVPP	DaVinci Vision Pre-Processing 模块， 图像处理模块，包括图片编解码（JPEGD、JPEGE）、视频编解码（Video Decoder、Video Encoder）

术语	描述
	和视觉预处理核（Vision Pre-Processing Core，VPC）等。
Host	Host-Device 架构的主机侧，本文指 x86 的 CPU 或者是华为自研鲲鹏 CPU。
HSCB	High Speed Control Bus，自定义的 STARS 高速调度总线。
IO Die	昇腾 950PR 芯片和昇腾 950DT 芯片中的 IO 通信 Die。
LLM	Large Language Model 的缩写，大语言模型。
NCA	Non-Cacheable Allocate，强制将 Cacheable 操作转换为 Non-Cacheable 操作。
NDDMA	N-dimensional Direct Memory Access Engine， AI Core 里内置的多维 DMA 数据搬运引擎，支持多维数据 Layout 搬运和变换。
RTP	Reliable Transport，UB（Unified Bus）的标准可靠传输层模式，提供端到端可靠传输服务。
SDMA	System Direct Memory Access，系统 DMA 引擎，可以实现昇腾 950 芯片内和芯片间数据拷贝、以及芯片内高速片上内存和 Cache 之间数据拷贝和管理。
Sector Cache	一种高速缓存设计策略，将缓存划分为多个扇区（Sector），每个扇区包含多个缓存行（Cache Line），以提升存储效率并降低冲突未命中率。
SIMD	Single Instruction Multiple Data，单指令多数据流，一种并行处理架构，一个指令同时操作多个数据元素。
SIMT	Single Instruction Multiple Threads，单指令多线程，一种并行处理架构，一个指令驱动多个线程同时执行，且每个线程可以有不同的独立地址空间。
STARS	System Task and Resource Scheduler，昇腾 950 芯片里的智能调度模块，负责系统任务和资源调度。
UB（Unified Buffer）	统一缓存区，AI Core 内部存储单元，主要用于矢量计算。
UB（Unified Bus）	灵衢总线的简称，华为定义的统一的互联标准协议，更高效地支持计算原生的内存语义、IO 语义和网络通信语义，更加充分地提升集群计算性能，提升资源利用效率，降低软件编程的复杂度。
UBoE	UB over Ethernet，UB 提供的对接以太网的协议接口，可以直接利用现有 Ethernet 交换机进行组网。
UMA	Unified Memory Access，所有处理器共享同一物理内存池，访问延迟一致，无需区分本地与远程内存。
URMA	UB Remote Memory Access，UB 提供的异步内存拷贝语义。

术语	描述
UB Memory	Unified Bus 提供的同步访存语义，支持 Load/Store 操作。

2 引言

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，大幅超越摩尔定律揭示的硬件迭代速度。数据规模持续激增，过去五年全球年新增数据量从 64ZB 飙升至近 500ZB，以前所未有的体量冲击着传统计算架构。在 LLM 大模型预训练和后训练场景中，大量 All-to-All 数据交换使得单次芯片间通信数据达到数十 MB，一次迭代的总通信数据量相比小模型提升近百倍，达到数百 GB，传统的互联带宽难以支撑如此密集的通信。LLM 大模型推理对算力需求的增长速度远大于硬件迭代速度，有必要引入低精度数据格式以提升有效算力。同时多模态生成和多模态理解计算任务的算存比相差巨大，单一类型硬件难以达到最佳性价比。AI Agent 应用需要超长上下文记忆、多轮复杂交互以及长时间的任务规划，导致 KV Cache 存储需求呈指数级增长，单靠 AI 芯片内存存储已无法支撑业务的快速发展。为了应对上述挑战，我们推出了全新的昇腾 950 系列芯片及产品。

昇腾 950 系列是华为面向下一代人工智能应用打造的旗舰级计算芯片，涵盖昇腾 950PR 与昇腾 950DT 两款核心产品。该系列基于全栈自主可控的制造工艺，搭载华为自研的第三代达芬奇（DaVinci）架构，在算力密度、存储带宽及互联拓扑三大维度实现了跨越式升级，能够全面赋能从大模型预训练、微调到推理部署的全生命周期，以及 AIGC、智能推荐、多模态处理等多元化场景。

在计算架构上，昇腾 950 系列通过引入 N 维直接内存访问引擎（NDDMA）、SIMD/SIMT 混合编程模式以及丰富的低精度格式支持（原生支持 MXFP4/MXFP8 和 HiF8），显著提升了 Transformer 类模型的训练与推理效率；同时集成自研 Linx816 CPU、DVPP 媒体处理子系统及安全引擎，构建了“AI+通用+安全”的多元异构计算体系。

在系统效能上，该系列针对不同场景进行了定制化优化：昇腾 950PR 侧重高性能推荐与大模型 Prefill 阶段，配备 128GB、1.6TB/s 高速片上内存，打造极致吞吐能力；昇腾 950DT 则聚焦大模型全量训练与复杂推理，配备 144GB、4TB/s 高速片上内存，突破内存墙瓶颈。配合创新的灵衢（Unified Bus）互联总线与灵活组网技术，昇腾 950 系列可支持超 128K 卡的大规模集群，以高联算比和低时延特性，为万亿及以上参数大模型的规模化落地提供强劲动力。

在软件方面，华为推出了异构计算架构 CANN（Compute Architecture for Neural Networks），以释放昇腾 AI 处理器的澎湃算力，并提供多层次编程 API、支持开发者快速构建 AI 算法和应用。CANN 软件栈主要分为高性能加速库、算子编程体系、编译器、运行系统等层次，并全面开源开放，对开发者开放底层能力和提供代码参考。

在后续章节，本白皮书将介绍昇腾 950 系列芯片的整体架构、详细规格和创新的特性能力，供读者深入了解昇腾 950 系列芯片。

3 架构概述

昇腾 950PR 芯片和昇腾 950DT 芯片采用共架构设计，通过搭载不同的片上内存，实现在细分应用场景上的竞争力提升：

- 昇腾 950PR：配备 128GB、1.6TB/s 高速片上内存，主要面向高性能推荐系统、大模型 Prefill（预填充）阶段及多模态推理场景，兼顾高吞吐与低延迟。
- 昇腾 950DT：配备 144GB、4TB/s 高速片上内存，专为大模型全生命周期打造，覆盖预训练、后训练及推理（含 Decode 与 Prefill）全流程，尤其适用于生成式大模型的训练与推理任务。

昇腾 950 系列基于华为自研的第三代达芬奇（DaVinci）架构，构建了灵活、多样且强大的 AI 算力底座。该架构全面支持 TF32、FP16、BF16、FP8、MXFP8、HiF8、INT8 及 MXFP4 等多种精度格式，能够精准适配不同场景的模型需求。通过支持 SIMD/SIMT 混合编程模式，并配备大容量 L2 Cache 与超高片上访存带宽，昇腾 950 能够最大化释放算力潜能，显著提升计算效率。

除卓越的 AI 算力外，昇腾 950 还集成了强大的通用计算与多媒体处理能力：

- 通用 CPU：集成华为自研的 Linx816 CPU 核心，支持物理双线程技术，提供强劲的通用逻辑处理能力。
- 媒体处理：内置 DVPP（数字视觉预处理）子系统，提供硬件级的图像预处理、编解码加速能力。
- 安全引擎：搭载专用安全算法引擎，确保数据处理的全链路安全。

昇腾 950 具备业界领先的 IO 扩展能力，整芯片集成 72 Lane HiLink SerDes，划分为 18 个 X4 端口。每个端口支持最高达 4×112Gbps 的 HiLink 互联速率，使整芯片对外 IO 带宽峰值达到 2TB/s。在网络协议栈方面，全面支持 URMA、UB Memory、PCIe 5.0 及 UBoE 等多种先进网络协议，为大规模集群组网提供高吞吐、低延迟的连接保障。

昇腾 950 系列芯片的主要特性

昇腾 950 系列芯片搭载华为自研的第三代达芬奇（DaVinci）架构。该架构在继承前代优势的基础上，以 Transformer 为核心，兼顾 LLM、推荐、多模态等多元化算法趋势，从低精度算力、计算效率、编程易用性及系统规模等维度进行了全方位优化，实现了性能与易用性的双重飞跃。

1. **AI 子系统架构演进：计算效率和易用性的双重突破**

- **Cube Core（张量计算单元）升级**
  - 新增精度格式：原生支持 HiF8、MXFP8、FP8、MXFP4 等前沿低精度格式，完美契合大模型量化需求。
  - 算力跃升：相比上一代 BF16 精度，昇腾 950 MXFP4 张量浮点峰值算力提升高达 4 倍。
- **Vector Core（向量计算单元）优化**
  - 架构革新：从传统 SIMD 升级为双发射 Register-Based 的 SIMD 新架构，并首创支持 SIMD-SIMT 混合编程模式，兼顾 SIMD 的高效和 SIMT 的编程灵活性。
  - 场景赋能：显著增强了对推荐算法、多模态处理等需高灵活度 Vector 算力场景的支持，为算法创新提供广阔空间。
  - 格式丰富：全面支持 FP32、FP16、BF16、INT8/16/32/64 等多种格式，并提供丰富的数值转换指令。
- **数据流与调试增强**
  - 支持 NDDMA 灵活的异步数据拷贝机制。
  - 构建 Cube 与 Vector 之间的高效内部数据通路及同步机制，减少数据搬运开销。
  - 大幅增强 Profiling 与 Debug 能力，降低开发门槛，提升调优效率。

2. **存储体系：大容量缓存与高带宽大容量片上存储**
   - **128MB 全局 L2 Cache**
     - 采用 Chiplet UMA 统一内存架构，支持算子可控的 Cache Hint 管理机制。
     - 支持按 Way 的 Cache Lock 与驻留策略，管理粒度精细至 128Byte Sector。
     - 性能提升：在离散小包访问和随机访存场景下，同带宽条件下性能较上一代提升 2 倍以上。
   - **高速片上内存**
     - 昇腾 950PR 单芯片提供最高 128GB 容量与 1.6TB/s 带宽。
     - 昇腾 950DT 单芯片提供最高 144GB 容量与 4TB/s 带宽。
     - 具备完备的 RAS（可靠性、可用性、可服务性）特性，保障大规模集群稳定运行。
3. **互联组网：灵衢总线与超大规模集群**
   - **超高带宽互联**
     - 基于新一代 HiLink SerDes 技术，单芯片互联带宽高达 2TB/s。
     - 支持 PCIe 5.0 x16（兼容 EP/RC 双模式）。
     - 支持 2*400Gbps UBoE（UB over Ethernet），实现 UB 协议无缝接入以太网络。
   - **灵衢（Unified Bus， UB）互联总线**
     - 端口复用：Scale-up（芯片间）与 Scale-out（节点间）端口可灵活复用。
     - 同步语义：通过 UB Memory 技术支持 Load/Store/Atomic 同步操作，实现最高 128TB 的 Host-Device 以及 Device-Device 内存共享访问。
     - 异步语义：支持灵活的 URMA 异步内存访问和消息语义。

- **灵活组网与超大规模扩展**
  - 支持 nD-Mesh、Clos 等多种灵活组网。
  - 通过软硬协同优化，充分利用物理带宽，实现高吞吐、低时延的集合通信。
  - 规模突破：超节点（Super Node）规模从上一代的 384 卡提升至 8K（8192）卡，整体集群支持规模超过 128K 卡。

4. **通信加速：计算通信深度融合**
   - **CCU（集合通信计算加速单元）**
     - 实现计算与通信的深度并行，显著降低对主存（Memory）的占用及 IO 调度延迟。
     - 通过硬件卸载集合通信任务，释放 AI Core 算力。
5. **Transformer 专项优化：大模型训练推理加速器**

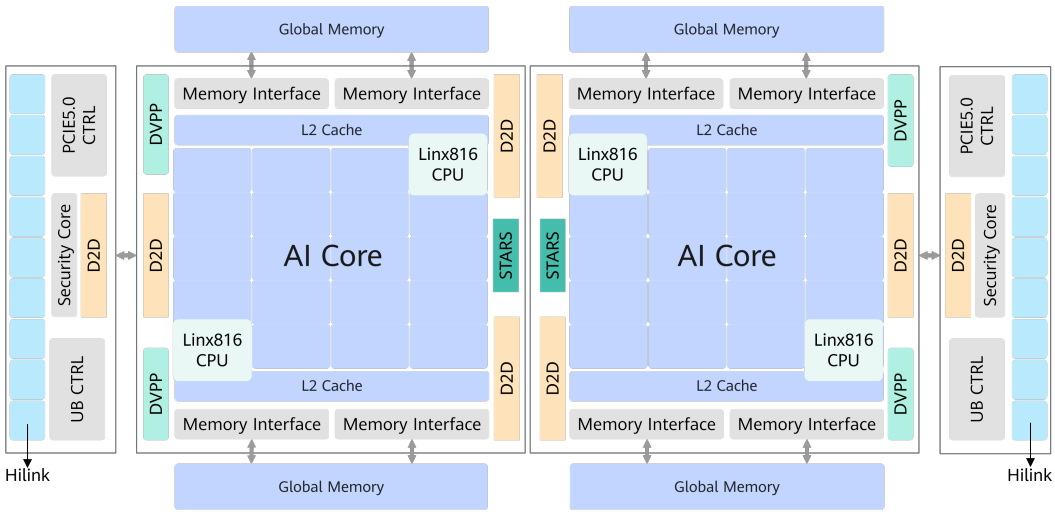
   针对大语言模型（LLM）的关键痛点进行深度硬件优化：
   - 算子加速：针对 FlashAttention 等关键算子，通过融合 Cube-Vector 通路、增强 Softmax 计算效率及支持 MXFP8/MXFP4 格式，单核性能较上一代提升 1.5~2 倍。
   - 全流程提速：结合高效的计算通信并行机制，显著缩短大模型的预训练与推理时间，加速业务落地。

昇腾 950 系列芯片架构和主要规格

昇腾 950 芯片是一个多 Die 合封的芯片，整芯片里合封了 2 个 AI Die、2 个 IO Die 和 8 个（昇腾 950PR）或者 4 个（昇腾 950DT）高速片上内存模块，这些 Die 和片上内存模块通过高速的 D2D Clink 和 Memory Interface 连接在一起，整个 Chiplet 芯片构成一个内存统一访问（UMA）的整体。

昇腾 950 芯片架构示意图如下：

图3-1 昇腾 950 芯片架构示意图



完整的昇腾 950PR 和 950DT 芯片主要包括如下规格：

- 36 个基于第三代 Davinci 架构的 AI 子系统， 每个 AI 子系统包括 1 个 Cube Core 和 2 个 Vector Core。
- 4 个 AI CPU Cluster， 每个 AI CPU Cluster 包括 2 个华为自研的基于 ARMv8-A 架构的高性能、低功耗双线程处理器 Linx816 CPU 和 4MB L3 Cache。
- 4 个 DVPP 子系统，包括 4 个 VPC Core， 4 个图片编码 JPEGE Core 和 8 个图片解码 JPEGD Core。
- 128MB 统一访问的 L2 Cache。
- 新一代高效任务调度系统 STARS2.0。
- 高带宽高容量片上内存：
  - 昇腾 950PR：单芯片提供最高 1.6TB/s 带宽和最高 128GB 容量。
  - 昇腾 950DT：单芯片提供最高 4TB/s 带宽和最高 144GB 容量。
- 华为自研的灵衢互联系统 Unified Bus 2.0，包括：
  - 72x 最高支持 112Gbps 速率的 HiLink ，分为 18 个 Port；
  - UB2.0 协议，支持异步拷贝的 URMA 和 Load/Store 全局内存语义 UB Memory；
  - 通过 Port 兼容复用，芯片对外提供 PCIe 通信协议接口，速率最高支持 GEN5 x16，向下兼容 GEN4/3/2/1，支持 EP 模式与 RC 两种模式；
  - 支持 2*400Gbps 的 UBoE，用于支持 UB 协议接入以太网络，支持 1x4/2x2 模式的 Port Bifurcation，即 1x400/200/100/50/25Gbps 或者 2x200/100/50/25Gbps 端口速率。

在上述完整规格基础上，结合精细的冗余设计，昇腾 950PR 和昇腾 950DT 会进一步衍生出多个不同规格的版本，并应用于不同的产品形态中。昇腾 950PR 和昇腾 950DT 详细的规格信息如下表所示。

表3-1 昇腾 950 系列芯片支持的主要规格

规格项			昇腾 950PR	昇腾 950DT
AI 子系统	Cube Core 数量		32/28	36/32/28
	Vector Core 数量		64/56	72/64/56
	Cube+Vector 总算力	MXFP4（TFLOPS）	1784/1561	2007/1784/1561
		HiF8/MXFP8/FP8（TFLOPS）	919/804	1034/919/804
		INT8（TOPS）	919/804	1034/919/804
		BF16/FP16（TFLOPS）	486/425	547/486/425
		TF32（TFLOPS）	243/212	273/243/212
	Cube 算力	MXFP4	1730/1513	1946/1730/1513

规格项			昇腾 950PR	昇腾 950DT
		（TFLOPS）		
		HiF8/MXFP8/FP8（TFLOPS）	865/756	973/865/756
		INT8（TOPS）	865/756	973/865/756
		BF16/FP16（TFLOPS）	432/378	486/432/378
		TF32（TFLOPS）	216/189	243/216/189
	Vector 算力	FP16/BF16（TFLOPS）	54/47	60/54/47
		FP32（TFLOPS）	27/23	30/27/23
		INT8（TOPS）	54/47	60/54/47
		INT16（TOPS）	27/23	30/27/23
		INT32（TOPS）	13/11	15/13/11
		INT64（TOPS）	6/5	7/6/5
Memory	Memory 容量（GB）		128/112	144/96
	Memory 带宽（TB/s）		1.6/1.4	4
SOC	AI CPU Subsys		Linx816 8C16T/6C12T/4C8T 支持 NEON	Linx816 8C16T/6C12T 支持 NEON
	DVPP Subsys	VPC	4/2 VPC Core 5760/2880FPS@1080P 等效的预处理能力	4/2 VPC Core 5760/2880FPS@1080P 等效的预处理能力
		Image Decoder	8 JPEGD Core 4096FPS@1080P 等效处理能力，支持最大分辨率 32K*32K	8 JPEGD Core 4096FPS@1080P 等效处理能力，支持最大分辨率 32K*32K
		Image EnCoder	4/2 JPEGE Core 1024/512FPS@1080P 等效处理能力，支持最大分辨率 32K*32K	4/2 JPEGE Core 1024/512FPS@1080P 等效处理能力，支持最大分辨率 32K*32K

规格项			昇腾 950PR	昇腾 950DT
	L2 Cache	容量（MB）	128/112	128
		Cache 配置	512B Cache Line 支持 4*128B Sector 支持 L2 Cache Hint 操作 支持 CMO	512B Cache Line 支持 4*128B Sector 支持 L2 Cache Hint 操作 支持 CMO
IO 规格	互联协议		URMA-CTP URMA-TP UB Memory PCIe 5.0 UBoE	URMA-CTP URMA-TP UB Memory PCIe 5.0 UBoE
	Unified Bus 带宽 （URMA-CTP URMA-TP UB Memory）		18Port，112Gbps，2016GB/s 双向 （出框速率受限于光模块）	18Port，112Gbps，2016GB/s 双向 （出框速率受限于光模块）
	UBoE		2* Port 2*400Gbps （与 UB 共用 2 个端口）	2* Port 2*400Gbps （与 UB 共用 2 个端口）
	PCIe		PCIe 5.0 x 16，128GB/s 双向 （与 UB 共用 4 个端口）	PCIe 5.0 x 16，128GB/s 双向 （与 UB 共用 4 个端口）

4 深度剖析

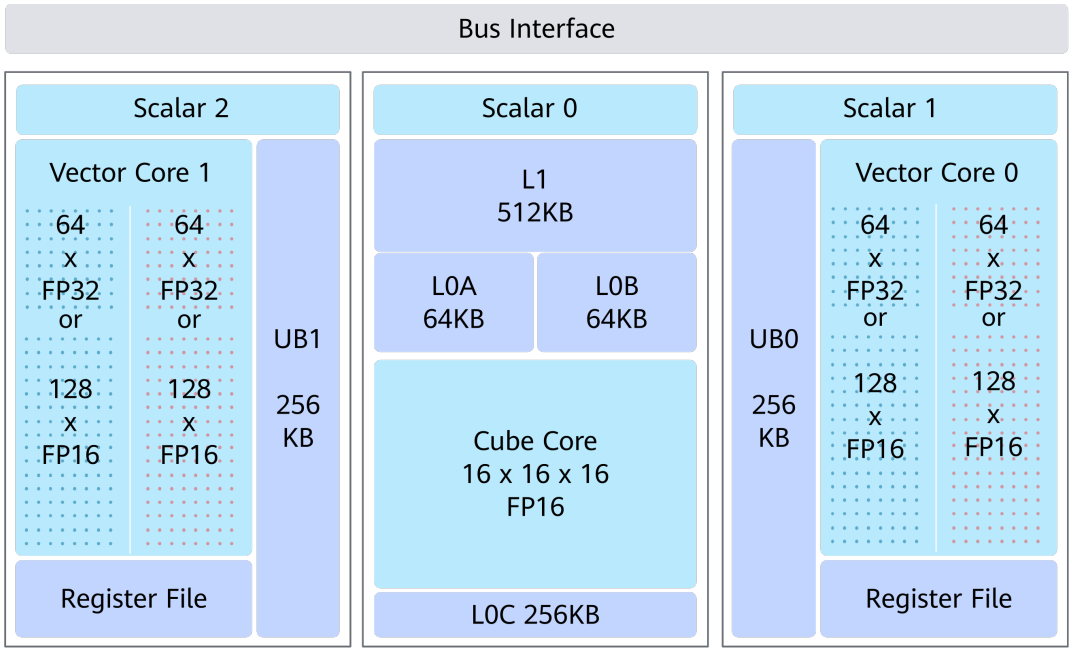
4.1 AI 子系统：第三代 DaVinciCore

第三代 DaVinciCore 通过系统性的协同升级，大幅提升计算性能，并显著改善开发者的编程体验，全面加速人工智能应用的落地。

新增 HiF8、MXFP8、MXFP4 精度格式。张量（Cube 核）浮点峰值算力提升至上一代的 4 倍。为打破计算瓶颈，向量（Vector 核）计算进行了大幅强化，FP32 与 FP16 单核算力提升 100%，提升张量与向量混合计算任务的整体效率。针对 LLM 中的关键算子 FlashAttention，通过深度硬件优化，单核性能相比上一代芯片提升 1.5 到 2 倍，显著加速了模型的训练与推理进程。

昇腾 950 DaVinciCore 采用新同构 SIMD/SIMT 混合编程架构，以 SIMD 为主、SIMT 为辅。在算力承载方面，大部分类型的向量计算由 SIMD 承担，成为硬件算力高效释放的核心路径与主力范式，充分发挥了架构高性能、高算力利用率的优势；SIMT 则作为差异化增强特性，面向离散访问（如 Gather/Scatter 操作）、复杂分支计算（如 Hash Insert 操作）等场景，有效提升特殊场景下的编程易用性和性能表现。此外，通过新增 NDDMA 指令，提供了多维数据排布的直接转换，简化了编程接口；而新增 Cube Core 与 Vector Core 之间的高速数据交换通道，大幅简化了核间数据流转的编程复杂度，显著提升了融合算子性能。

图4-1 AI Core 架构及各层级 SRAM 示意图



4.1.1 CUBE CORE

第三代 Cube 核支持的数值格式上进一步丰富。新增 MXFP8、HiF8 及 MXFP4 等低比特浮点格式的张量计算支持。在使用新的 FP8/FP4 数据类型时，在相同频率下，HiF8/MXFP8/FP8 能够提供 2 倍的 FP16 浮点张量计算 TFLOPS，MXFP4 数据类型能提供 4 倍 TFLOPS。

与上一代芯片 Cube 核相比，第三代进一步优化微架构，提升缓存复用率，面向 GEMM/FlashAttention 等关键算子，提供更高效的矩阵运算。更大的 L0C Buffer，提供更灵活的 Tiling 策略，并提高数据复用率。

提供按需张量结果即时数据格式转换，回写 Unified Buffer 阶段直接完成数据量化（FP32 到 BF16/FP16/FP8）与排布转换（NZ 到 ND/DN），提升开发灵活性；同时支持 L0C Buffer->Unified Buffer 随路量化，能够从 FP32/INT32 量化到 BF16/FP16/FP8/INT8，降低核内缓冲空间及核间带宽占用，提升端到端吞吐与能效。

图4-2 Cube Core 处理架构示意图

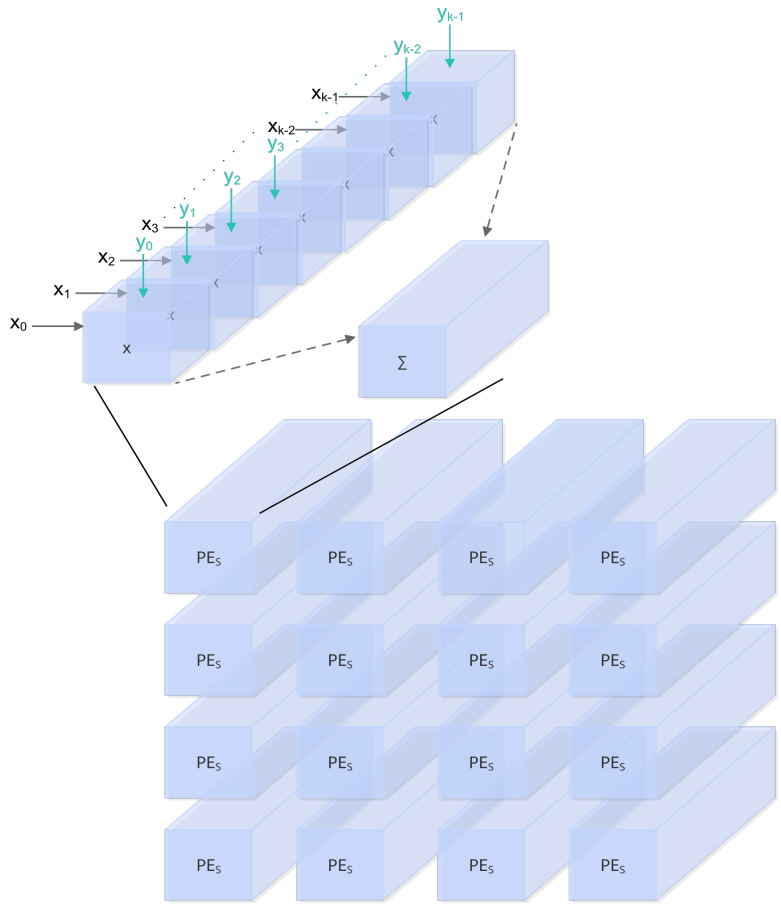
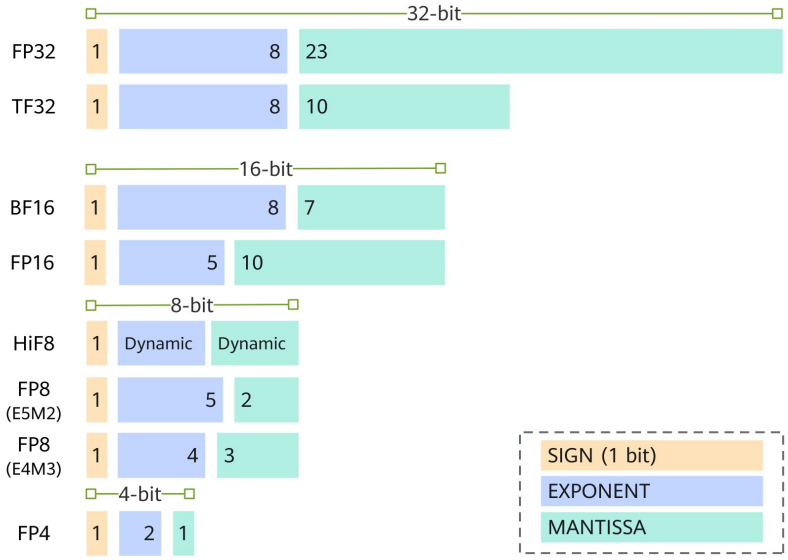


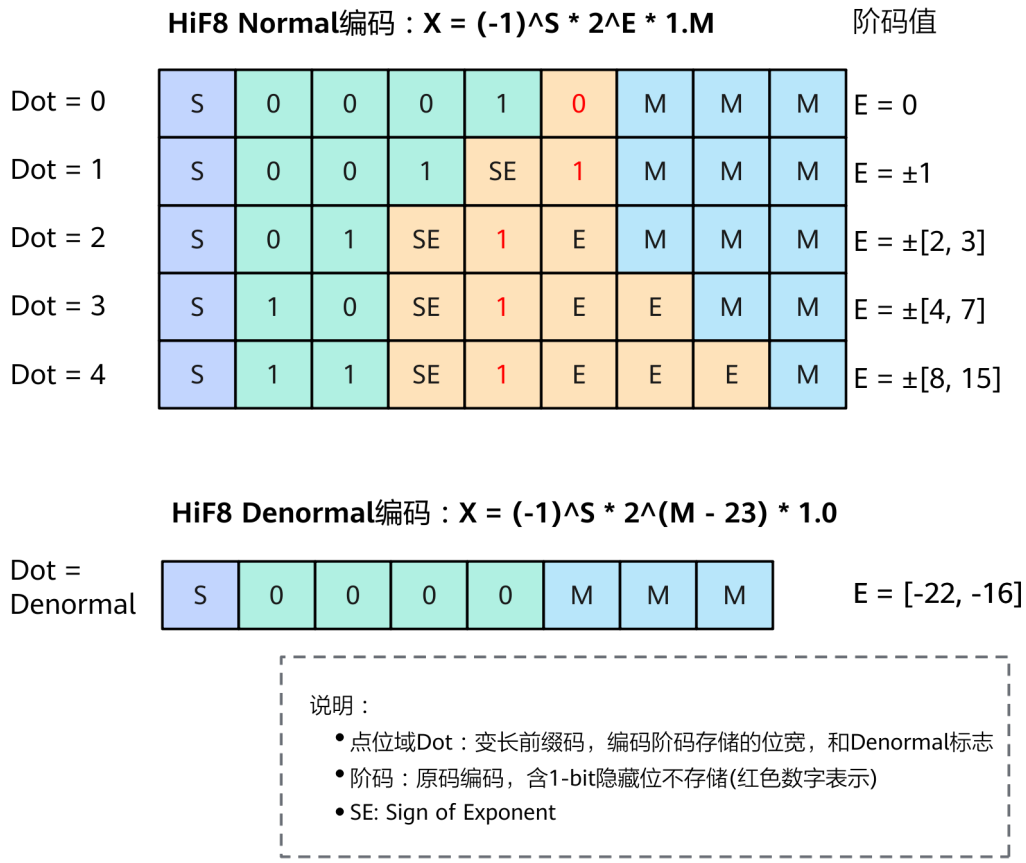
图4-3 Cube Core 支持的数值精度示意



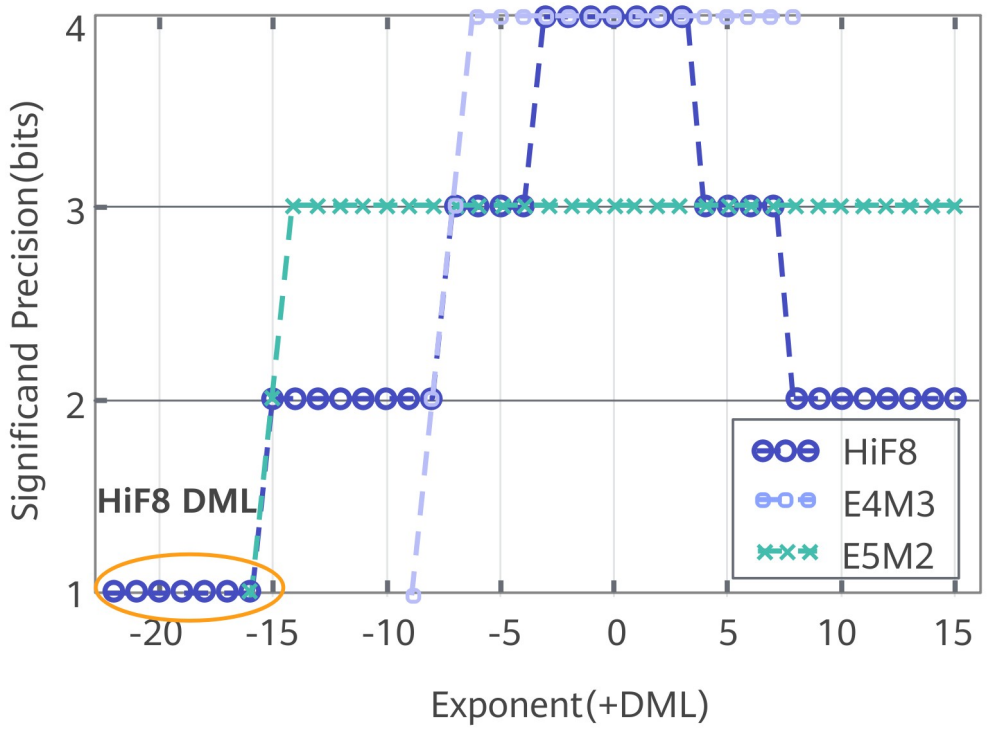
HiF8 兼顾精度和动态范围，为神经网络的训练和推理，提供了能力更全面的 8 比特单数据格式表达。

相比于 FP8 E4M3 的 18 个指数表达，HiF8 支持 38 个指数表达，动态范围成倍提升。与 MXFP8 相比，HiF8 不需要额外的 8 位 MX 缩放因子。在 8 比特的极低开销限制下，HiF8 有效地找到了一条既能满足现代神经网络精度需求，又能提供足够动态范围的路径。

图4-4 HiF8 数值精度



1. HiF8 利用变长前缀码编码的点位域 Dot，显式指示阶码存储的位宽和 Denormal 标志，实现符合 AI 数据分布特征的锥形精度格式。
2. 同时阶码采用原码编码，并隐藏了 1 比特固定值不存储，确保了不同位宽的阶码表达范围不重复，进而实现无冗余编码。
3. 最后通过特殊的浮点 Subnormal number 设计，将综合阶码范围从[-15, 15]提升到了[-22, 15]共 38 个阶码，接近 FP16 的 40 个综合阶码值表达。



HiF8 阶码分布图（锥形精度图）：

- 有效位包含 1.M 的隐藏位 1，因此比 Mantissa 位宽多 1 比特；
- 在数值绝对值靠近 1 的时候，精度高；远离 1，精度逐渐降低。精度不存在跳变，都是比特数为 1 的渐变；
- 综合阶码范围达到[-22, 15]，和 FP16 的[-24, 15]接近，共有 38 个 powers of 2；
- 编码了 4 个特殊值，不区分正负 0。

表4-1 HiF8 特殊值编码

特殊值	编码
ZERO	00000000
NAN	10000000
+INF	01101111
-INF	11101111

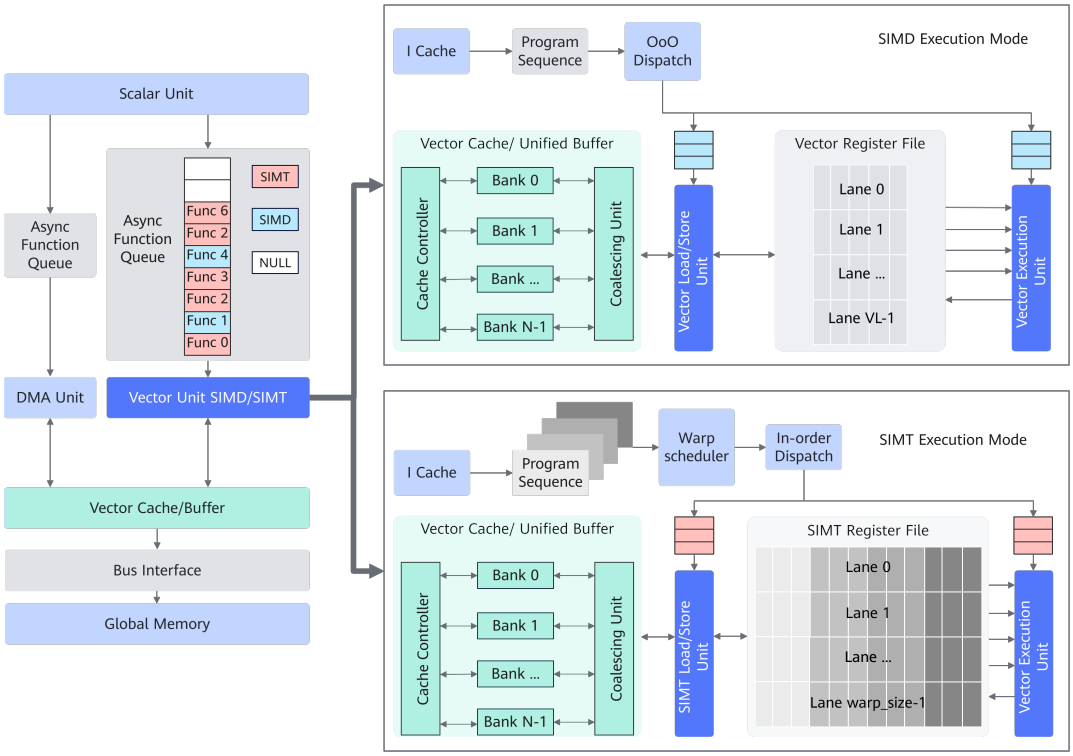
4.1.2 VECTOR CORE

第三代 VECTOR 核从指令集到微架构都进行了全面升级：

- 算力：单核 FP16 和 FP32 的 TFLOPS 较上一版本提升 100%，使 Cube-Vector 融合算子（例如 FlashAttention）的非矩阵类操作不再成为性能短板。

- 指令：新增对 BF16 的原生支持，并扩展多种浮点格式转换指令，强化量化与反量化能力。
- 微架构：针对 Softmax、GELU 等关键函数优化微架构。提升张量 ALU 利用率，减少因数据依赖导致的“气泡”。 在优化频率、提升算力的同时，仍然保持低指令延迟。
- 混合编程：支持 SIMD/SIMT 混合编程，将向量级并行与线程级并行结合，在性能、可编程性与可移植性上形成互补优势。
- 内存架构：在 Unified Buffer 与 Vector ALU 之间引入 RegFile 寄存器作为临时存储，为向量执行进程提供更高的带宽支持与数据复用能力。

图4-5 Vector Core 架构示意图



4.1.3 新同构 SIMD/SIMT 混合编程

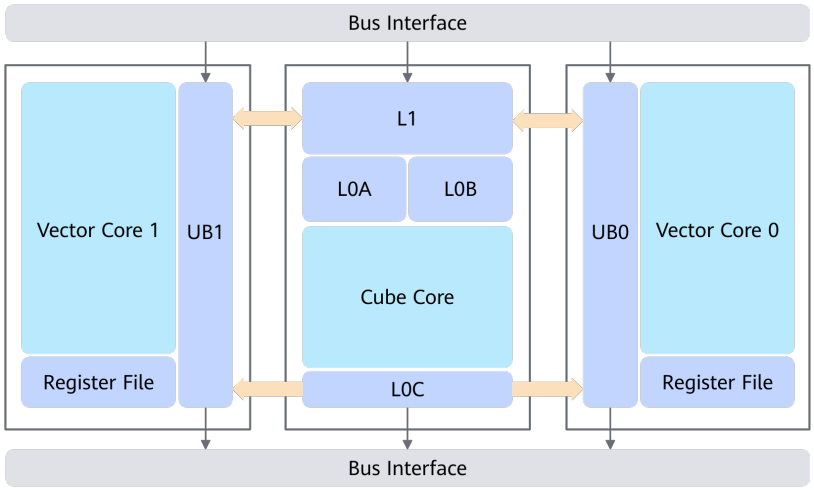
SIMD/SIMT 混合编程将向量级并行与线程级并行相结合，在性能、可编程性与可移植性之间形成互补优势。其中，SIMT 降低了复杂算子与不规则控制流的开发难度；SIMD 则借助双发 ALU 指令与乱序执行（Out-of-Order Execution）能力，实现单指令处理多数据，提升每周期吞吐量。混合编程允许根据算子特征进行精细化映射：例如，对于以规则访存为主的 element-wise 计算，优先采用 SIMD 模式以获得高带宽与高算力利用率；而对于不规则或包含分支的部分，采用 SIMT 模式以缓解 gather/scatter 操作带来的控制复杂度。在系统层面，混合编程有助于提升硬件利用率与能效，同时更便于算子融合、数据复用等优化。对于多架构部署，该模型还能兼顾代码复用与性能可移植性。基本函数块定义为 Vector Function（VF）， 每个 VF 可选择以 SIMD 或 SIMT 方式实现，并支持在两种不同类型 VF 之间快速切换。

总体而言，新同构 SIMD/SIMT 混合编程以 SIMD 为主、SIMT 为辅，在各种工作负载中，能在端到端吞吐、时延与开发效率之间取得更优平衡。

4.1.4 支持 CV 融合

支持 Cube L1 Buffer 和 Vector Unified Buffer 的直接 CV 数据传递通道；提高核内数据复用率，减少 L2 层的数据交换，提升 CV 融合算子效率。

图4-6 AI Core Cube-Vector 融合示意图

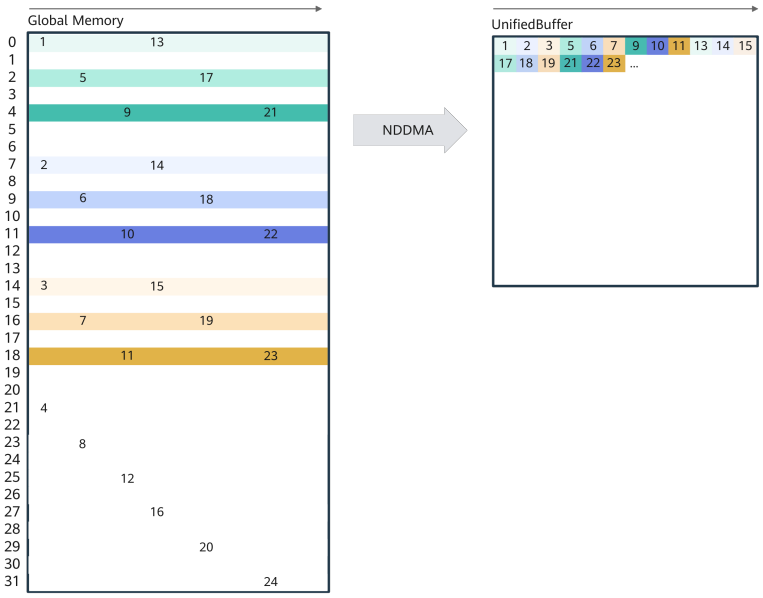


针对 FlashAttention 的带宽瓶颈，支持 Cube Core 和 Vector Core 之间的内部数据通路，实现 Cube 计算和 Vector 计算的高效融合。此外，还支持在 Cube Core 和 Vector Core 之间传输数据时进行随路数值精度和数据排布转换，以进一步提升端到端吞吐与能效。

4.1.5 新增 NDDMA 指令

新增 NDDMA 指令，为数据搬运与数据排布转换提供了更简单的编程接口。开发者可在 Kernel 层以更少的程序语句实现 NCHW/NHWC 等不同数据排布和对齐方式的转换。NDDMA 硬化了地址生成逻辑，从而能更快速地生成地址，且通过内置缓存最大限度地降低冗余访存，加快了数据搬运和数据排布转换的过程。

图4-7 NDDMA 指令

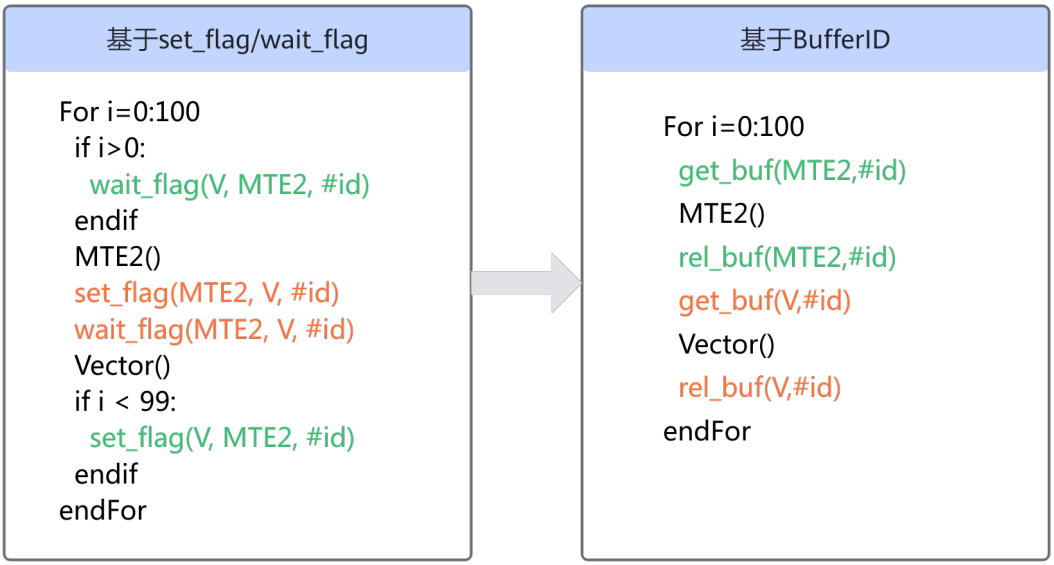


NDDMA 指令可以对全局内存中的数据做最多 5 个维度的重排操作，然后存储至 Vector Core 的 Unified Buffer 中，且可以同时实现数据搬运以及数据重排或转置等操作，降低编程复杂度。假设一个输入矩阵如上图全局内存（Global Memory）所示，需要将几个间隔规律的数据块 1,2,3,5,6,7,... 按顺序读取并存储至 Unified Buffer 中，使用 NDDMA 指令可直接通过配置参数，让硬件自动完成任务。NDDMA 内置缓存能自动发掘数据局部性，并将多个数据元素粒度的读操作转换为 128 字节的读操作，提升内存访问效率。

4.1.6 更易用的同步机制

新增 BufferID 同步机制，其使用方式类似于编程语言中的互斥锁（get_buf()对应加锁，rel_buf()对应解锁），能够直观表达流水线对于 AI Core 内部存储的占用和释放，相对于 set_flag 和 wait_flag 同步机制，该方式内聚性更强，与其他流水线解耦，降低了同步的复杂度。

图4-8 昇腾 950 新同步机制代码示例



4.2 AI CPU 子系统

AI CPU 是昇腾 950 系列芯片的重要组成部分，它提供了强大的通用处理能力，核心承担两大类任务：

- 通用的控制类任务：支持在 NPU 侧运行操作系统，并执行页表管理、性能监控、算子编译、加速器和 IO 调度等一系列通用任务；
- 计算类任务：作为 AI Core 的有效补充，负责执行 CPU 类算子（包括控制类算子、标量和向量等通用计算类算子）。

昇腾 950 系列芯片集成了华为自研的 Linx816 CPU Core。 Linx816 CPU Core 基于 ARMv8-A 架构，支持物理双线程，具有低功耗、高性能的优势。AI CPU 子系统具有如下典型特性：

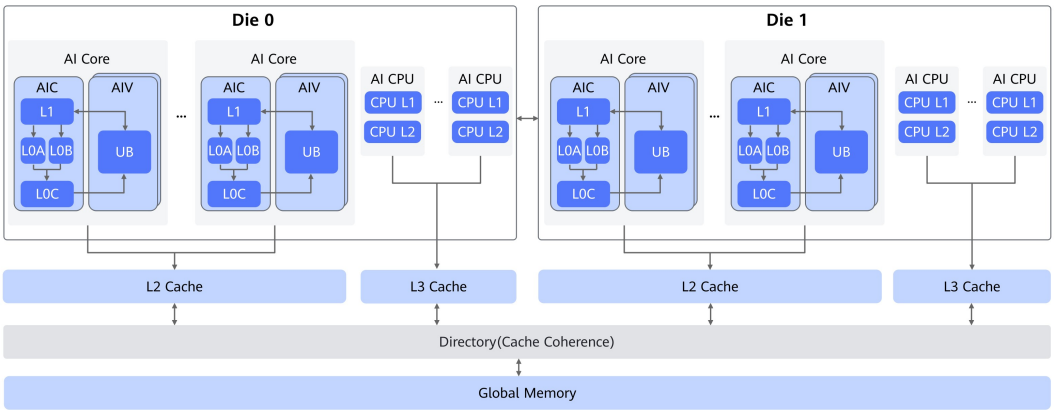
- 每个 Linx816 CPU Core 支持两个独立线程，两个线程共享核内的 CPU L1 Cache/L2 Cache 子系统。每个 CPU Core 可以独立配置为双线程模式或者单线程模式。
- 所有 AI CPU 和 AI Core 共享统一的片上内存；AI CPU 有独立的 CPU L1/L2/L3 Cache 系统，并通过硬件支持的全局缓存一致性设计，可以和 AI Core + L2 Cache 子系统交换数据。具体 Memory 层级结构请参考 4.3 Memory 子系统。

4.3 Memory 子系统

作为 NPU 芯片的重要组成部分，Memory 子系统扮演着缓存计算的输入、输出、中间结果等数据的作用，对芯片的性能、面积、功耗、成本等有着重要影响。在昇腾 950 芯片中，Memory 主要包括高速片上内存、L2 Cache、L3 Cache、AIC 和 AIV 内的各种 Local Memory、AI CPU 的 L1/L2 Cache，各级 Memory 的关系如下图所示。

- 高速片上内存：昇腾 950PR 和昇腾 950DT 采用不同的片上内存，是缓存全局性数据的 DRAM。
- L2 Cache 主要服务于 AIC/AIV 的 AI 计算，可以从高速片上内存读数据到 AIC/AIV，也可以将 AIC/AIV 的数据写到高速片上内存，以其高带宽、低延迟的特点提高了 Memory 子系统的效率；L3 Cache 主要服务于 AI CPU 的通用计算，其作用与 L2 Cache 相同。
- AIC/AIV 内的 Local Memory（主要包含 L1 buffer、L0 buffer、Unified Buffer 等）和 AI CPU 内的 L1/L2 Cache 主要缓存计算过程所需的数据。

图4-9 昇腾 950 内存层次示意图



昇腾 950 系列芯片主要 Memory 的缓存大小如下表所示。

表4-2 昇腾 950 Memory 层次中主要 Memory 及其大小

Memory	Memory size
L1 Buffer	512KB per AI Core
L0A Buffer	64KB per AI Core
L0B Buffer	64KB per AI Core
L0C Buffer	256KB per AI Core
Unified Buffer（UB）	512KB per AI Core
CPU L1 Cache	64KB per CPU Core
CPU L2 Cache	1MB per CPU Core
L3 Cache	4MB per CPU Cluster
L2 Cache	Up to 128MB
昇腾 950PR 片上内存	Up to 128GB
昇腾 950DT 片上内存	Up to 96/144GB

4.3.1 高速片上内存

随着业界网络模型的持续变化，更大的内存容量、更快的内存带宽成为 NPU 的演进目标。为了满足业务需求，昇腾 950PR 最高可以提供 128GB 的容量、1.6TB/s 的内存带宽。昇腾 950DT 最高可以提供 144GB 的容量、4TB/s 的内存带宽。

现代 AI 集群对内存可靠性提出了非常高的要求，为此，高速片上内存实现了如下 RAS（Reliability, Availability and Serviceability）特性：

- 支持 Online ECC。
- 巡检特性，发现薄弱点并回写或隔离。
- 预留部分行，在发现新增行失效时，可以动态隔离问题行，并搬运数据到预留行，从而实现用户无感知的行失效隔离。

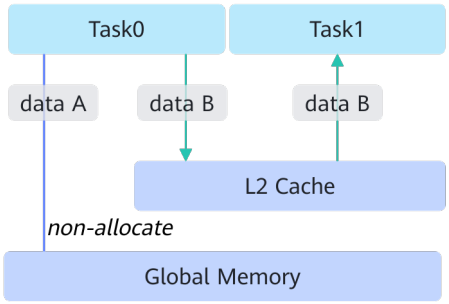
4.3.2 L2 Cache 特性

为了进一步提高 Memory 系统的效率，昇腾 950 芯片配置了 128MB 容量的 L2 Cache。当 AIC/AIV 的输入、输出数据为重复使用时，L2 Cache 能够缓存中间数据并及时供应给 AIC/AIV，避免数据访问片上内存的更大开销。

L2 Cache 具有如下特性：

1. 昇腾 950 芯片是 Chiplet、2 Die UMA 架构，整芯片的地址空间统一管理，可以跨 Die 访问 L2 Cache，L2 Cache 具有局部亲和性。
2. 由硬件维护 2 个 Die 之间的 L2 Cache 一致性，软件不感知。
3. L2 Cache 的微架构
   - 多 Bank 分布式架构，512B 低位交织（采用高位异或交织，交织算法升级）；
   - Cache Line 粒度 512B，新增支持 128B Sector Cache 特性，128B/256B 访问更高效，大大提高 NPU 处理的灵活性和效率；
   - 每个 Bank 支持同时读写。
4. Cache 管理优化策略
   - 针对 AI 业务场景的模型数据流动关系，昇腾 950 芯片为增强 L2 Cache 性能而设计了 L2 hint 策略。程序员可以根据实际需求选择启用不同的 L2 hint 配置，将 L2 hint 配置在用例代码中，NPU 硬件根据上游携带的 hint 信息进行分配（allocate）和替换（victim）。例如，Task0 的输出数据 data B 是下一个 Task1 的输入，data B 缓存在 L2 Cache，提高访存的效率和性能；Task0 的输出数据 data A 不是作为下一个 Task1 的输入，且短期内不会用到，可将 data A 设置为不分配（non-allocate）的 L2 hint，则该数据不会 allocate 进 L2 Cache，而是直接写到 Global Memory，如此可以避免 data A 替换其他短期内会用到的数据，提高 L2 Cache 的局部数据利用率。

图4-10 Non-allocate（L2 hint）典型应用场景示意图

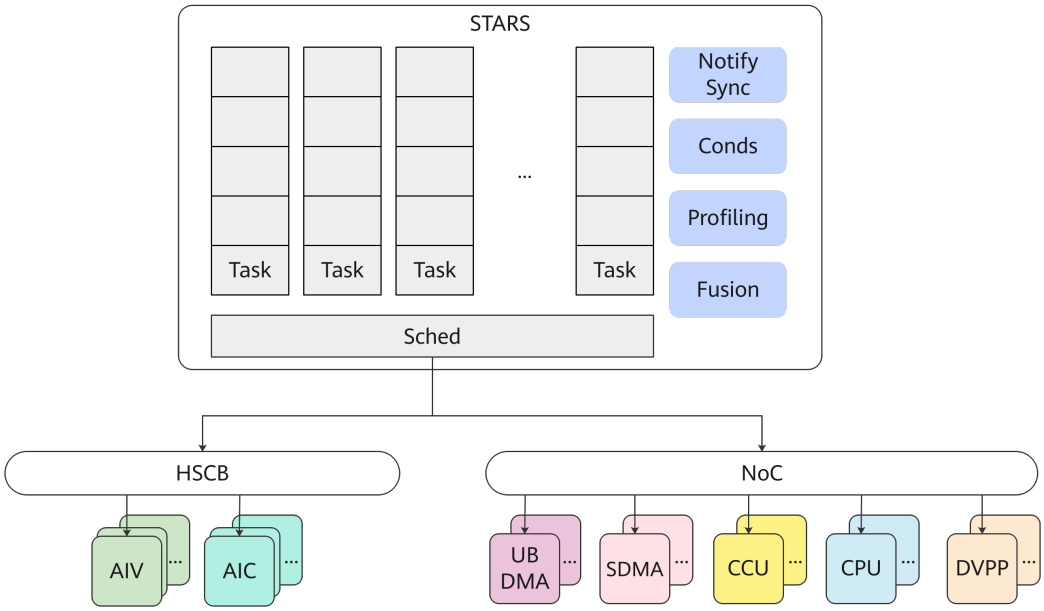


- 昇腾 950 芯片提供针对 SDMA 的 L2 Cache 驻留策略（Cache Maintenance Operation，CMO），包括预取（Prefetch）、预写回（Writeback）、无效化（Invalid）、冲刷（Flush）。程序员可以根据实际需求选择使用不同的 CMO 策略，并且可以修改配置参数调整 CMO 的发生时机、有效范围等。

4.4 软硬协同高效调度：STARS2.0

昇腾 950 系列芯片升级了第二代的硬件调度器 STARS2.0。 System Task and Resource Scheduler（STARS）是全芯片的任务和资源调度处理中心，可以为多种运算核和 IO 资源调度任务、串接数据流以及同步信息，支持高效调度 AIC、AIV、CPU、DVPP、SDMA、UB、CCU 等多个计算和数据搬运引擎。STARS 提供丰富的软硬件接口，支持任务下发、算力切分等特性。STARS 同时支持实时 TOP-DOWN Profiling 分析模型，可以记录 Task 执行时间轨迹、计算开销、带宽、功耗等信息，为整系统调优提供数据支持。

图4-11 STARS2.0 架构示意图



- STARS 支持 Host 下沉 2048 条任务流到 Device 侧，STARS 会根据任务流对应配置预取任务，调度任务，上报执行结果给 Host，整个流程由软硬件协同完成，以降低端到端调度时延，提升芯片计算效率。
- STARS 支持调度多种加速器：
  - 支持调度整芯片的 AIC、AIV、VPC、JPEGD、JPEGE 等加速器；
  - 支持并发调度最多 16 个 AI CPU 任务和 64 个 Host CPU 任务；
  - 支持并发调度最多 64 个 UB jetty；
  - 支持并发调度最多 32 个 CCU 任务；
  - 支持并发调度最多 32 个 SDMA 通道；
  - 支持最多 128K 个单比特的同步标志位或最多 4096 个 32 比特的多比特同步标志位，以方便实现多个任务流之间的同步操作；
  - 支持调度计算融合通信任务，实现计算和通信的任务并发；
  - 支持执行条件算子。
- STARS 使用专用的高速控制总线（High Speed Control Bus，即 HSCB）与 AIC/AIV 交互，调度开销降低至 ns 级，实现 AIC/AIV 计算任务的高效调度。相比 AI Core/AI CPU 等正常数据访问的片上互联网络（Network on Chip，即 NoC），HSCB 速度快、专用不受干扰且支持广播调度能力。
- STARS 支持 Group 调度，支持将 AI Core 等资源分成最多 8 个 Group， 每个 Group 的 AI Core 数量、安全属性等可以软件配置。该特性支持将算子按 Die 分成 Group 进行亲和性调度， 以更好地利用 L2 Cache 的局部性。
- STARS 支持算力切分模式：
  - 支持将 AIC/AIV/SDMA 最多切分成 16 个资源池；
  - 支持将其他加速器资源最多切分成 8 个资源池；
  - 支持配置资源池和虚拟机间的绑定关系，以实现虚拟机之间的资源隔离。

4.5 图片处理子系统

在深度学习与视觉计算平台中，输入图像通常以 JPEG 等标准压缩格式传输。如果解码或预处理性能不足，会成为端到端训练和推理吞吐的主要瓶颈。为此，芯片集成了 DVPP（DaVinci Vision Pre-Processing）子系统，通过专用硬件加速器完成图像的解码、预处理与编码，避免数据搬运对 AI Core 和 AI CPU 的计算占用，实现更高能效和更低延时。

DVPP 子系统由以下模块组成：

- JPEG Decoder（JPEGD）：解码网络端或本地存储中的 JPEG 图像比特流，输出为 YUV 图像帧。
- JPEG Encoder（JPEGE）： 接收原始图像帧或 VPC 预处理结果，将其编码为 JPEG 比特流，用于回传或存储。
- Vision Processing Core（VPC）：提供图像预处理操作，将输入图像转化为 AIC 可直接使用的格式。

芯片集成 4×VPC、4×JPEGE 和 8×JPEGD，实现多流、多格式的并行处理能力，适配 AI 训练、推理、视频分析和多媒体处理等典型场景。

- **VPC （Vision Processing Core）**

  VPC 提供了丰富的视觉预处理功能，算子覆盖范围与 OpenCV、TensorFlow、TorchVision、Pillow、DALI 等主流软件库对标。

  支持的功能包括：

  - 尺度变换：缩放（Resize）、裁剪 （Crop）、填充 （Padding）。
  - 采样操作：上采样（UVDEC）、下采样 （UVUP）。
  - 色彩空间与图像增强：CSC （Color Space Conversion）、HSV 调整、像素增强 （PixAug）。
  - 几何变换：仿射变换（Affine）、透视变换（Perspective）。
  - 在典型多流场景下，可同时处理多路 1080p 输入，保证训练数据加载与 AIC 推理无缝衔接。
  - 支持 STARS 直接硬件调度。

- **JPEGD（JPEG Decoder）**

  JPEGD 负责静态数字图像解码，提供对超高分辨率与多色彩格式的硬件支持。支持的功能包括：

  - 最大分辨率：32768 × 32768。
  - 支持格式：YUV444/422/420/440/400 (8 比特)，输入为 baseline JPEG 比特流，输出为对应的 semi-planar 格式。
  - 区域解码：可对指定图像区域进行解码。
  - 对标库：libjpeg-turbo v2.0.2。
  - 支持 STARS 直接硬件调度。

- **JPEGE （JPEG Encoder）**

  JPEGE 模块提供 JPEG Baseline 编码（Sequential DCT），支持高分辨率和多种 YUV 格式：

  - 最大分辨率：32768 × 32768。
  - 支持格式：YUV420 semi-planar、YUV422 packed、YUV444 planar/packed、YUV422 semi-planar、YUV400。
  - 支持 STARS 直接硬件调度。

4.6 互连子系统

昇腾 950 芯片集成了 Unified Bus 控制器，支持灵衢 2.0 互联技术栈，可跟所有实现 UB2.0 开放标准的部件兼容互通，协同工作。此外还集成了 PCIe Gen5.0 控制器。

昇腾 950 系列芯片整体互连规格和特性如下：

特征	昇腾 950 系列芯片 IO 规格和特性
UB 协议版本	Unified Bus 2.0
UB IO 带	UB：支持 2016GB/s（双向）。

特征	昇腾 950 系列芯片 IO 规格和特性
宽	UBoE：支持 200GB/s（双向，和 UB Link 复用互连 SerDes 端口）。
UB 端口模式	UB：支持 18 个 x4 Port。每 Port 可向下降 Lane 支持到 x2 或 x1。 UBoE：支持 2 个 x4 Port 或 4 个 x2 Port。每 Port 可向下降 Lane 支持到 x2 或 x1（和 UB Link 复用互连 SerDes 端口）。
UB 编程接口	同步访问：UB Memory。 异步访问：URMA。 集合通信加速处理器：CCU。
UB 互联规模	最大支持 8192 卡超节点。
UB 可靠性	支持链路层重传。 支持端到端可靠性重传机制。
UB 灵活拓扑	Clos 组网。 Full Mesh+Clos 混合组网。 nD-Mesh 组网。
UB On Chip 转发	支持。
UB 组网扩展	通过 UB Switch 或通过 UBoE 搭载以太 Switch 进行组网扩展。
PCIe 协议版本	支持 PCIe 5.0。
PCIe IO 带宽	支持 128GB/s（双向）。
PCIe 端口模式	支持 1 个 x16 Port。可向下降 Lane 支持到 x8、x4 端口（Port 数不变）。 支持 RC 和 EP 双模（静态选择）。

昇腾 950 芯片支持 UB 互连，基于 UB 端口提供基于 URMA 的异步访问能力和基于 UB Memory 的同步访问能力。异步访问即基于队列 SQE/CQE 交互机制实现了异步编程接口，同步访问即芯片上的 Core 或调度器可以直接向 UB 端口发起 Load、Store 和 Atomic 访问。URMA 的异步访问能力也可以基于 Ethernet 物理层与外部进行互通。

昇腾 950 芯片提供集合通信加速引擎（CCU）用于加速集合通信交互。

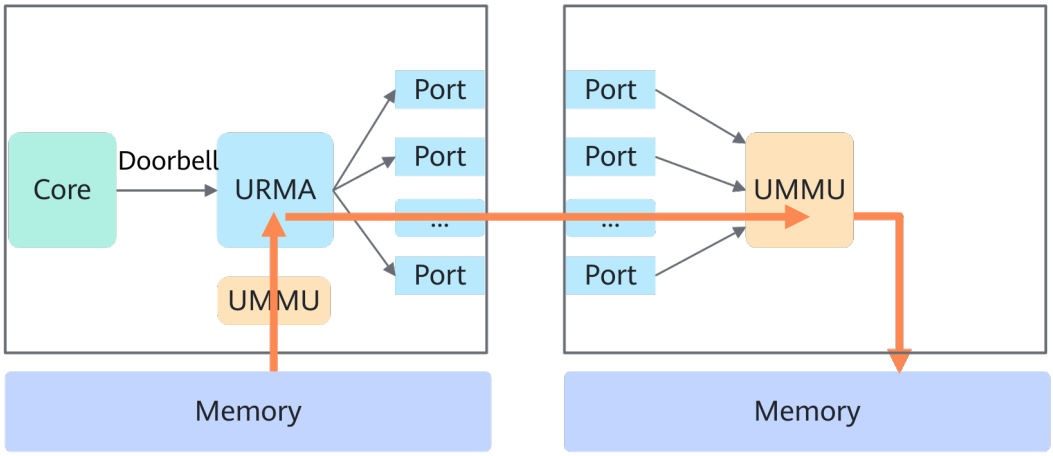
昇腾 950 芯片提供 Die 内多个 Port 的转发能力，可支持更灵活的通信路径规划。

4.6.1 URMA

本芯片提供基于 URMA 的异步通信语义，其交互流程如下图所示，Core（AI Core or AI CPU or STARS）通过 Doorbell 的机制调用 URMA 的 Jetty 队列进行内存搬运，经底

层的 Port 与其他的节点进行通信，UMMU 在通信过程中提供 VA 到 PA 的地址转换和权限控制能力。

图4-12 URMA 异步访存通信的过程示意图



- URMA 异步通信以 Jetty 队列作为入口，支持提供以下通信能力。
  - 支持 Write、Write with ImmediateData 和 Write with Notify 操作。
  - 支持 Read 操作。
  - 支持 Send、Send with ImmediateData 操作。
  - 支持 Atomic FetchAdd 和 CompareAndSwap 操作。
- URMA 异步通信可以基于以下两种传输层服务进行通信。
  - RTP mode：提供可靠传输服务，可支持端到端可靠重传机制。可支持 4 个 Port 的可靠性传输带宽能力。可靠传输服务下基于多个 Transport Channel 可以提供多路径传输服务，有效利用多端口的传输能力。
  - CTP mode：提供简易传输层服务，不支持端到端可靠重传机制。可支持 9 个 Port 的带宽能力。在简易传输服务下也可以支持多路径传输。

4.6.2 UBoE

UB 协议栈用于 Scale-Out 通信的功能可以直接运行在以太网上，即 UB over Ethernet（UBoE），直接利用现有 Ethernet 交换机进行组网。

整芯片提供了 2 个 400G Ethernet Link，用于支持 UB 协议接入以太网络，在底层 SerDes 上 Ethernet Link 和 UB Link 是静态复用的。当该物理层 400G 端口被用作 Ethernet 端口后，UB Link 的能力也就相应减少 1 个 400G 端口。

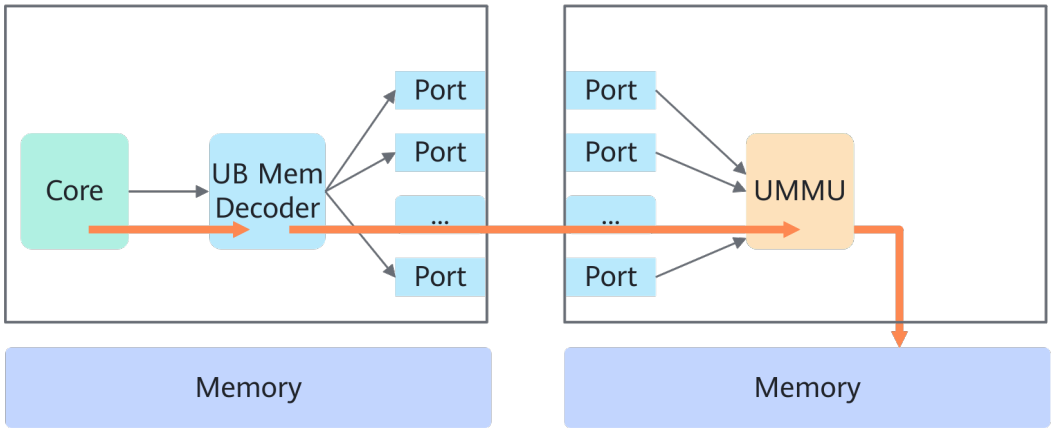
支持 1x4/2x2 模式的 Port Bifurcation，即 1x400/200/100/50/25Gbps 或者 2x200/100/50/25Gbps 端口速率。

4.6.3 UB Memory

本芯片提供 UB Memory 同步访存语义，其交互过程如下图所示，Core（AI Core or AI CPU）发出同步操作送至 UB， 经过 UB 的 Memory Decoder 获取到目的节点信息和地

址信息，然后从底层的 Port 将操作送至对端芯片，经对端芯片 UMMU 地址翻译和权限校验后，可直接访问对端芯片的内存。

图4-13 UB Memory 同步访存语义地址通信过程示意图



- 本芯片 UB Memory 提供的同步操作能力如下：
  - 支持 Write 操作；
  - 支持 Read 操作；
  - 支持 AtomicStore、AtomicLoad、AtomicSwap 和 AtomicCompareAndSwap。

4.6.4 CCU

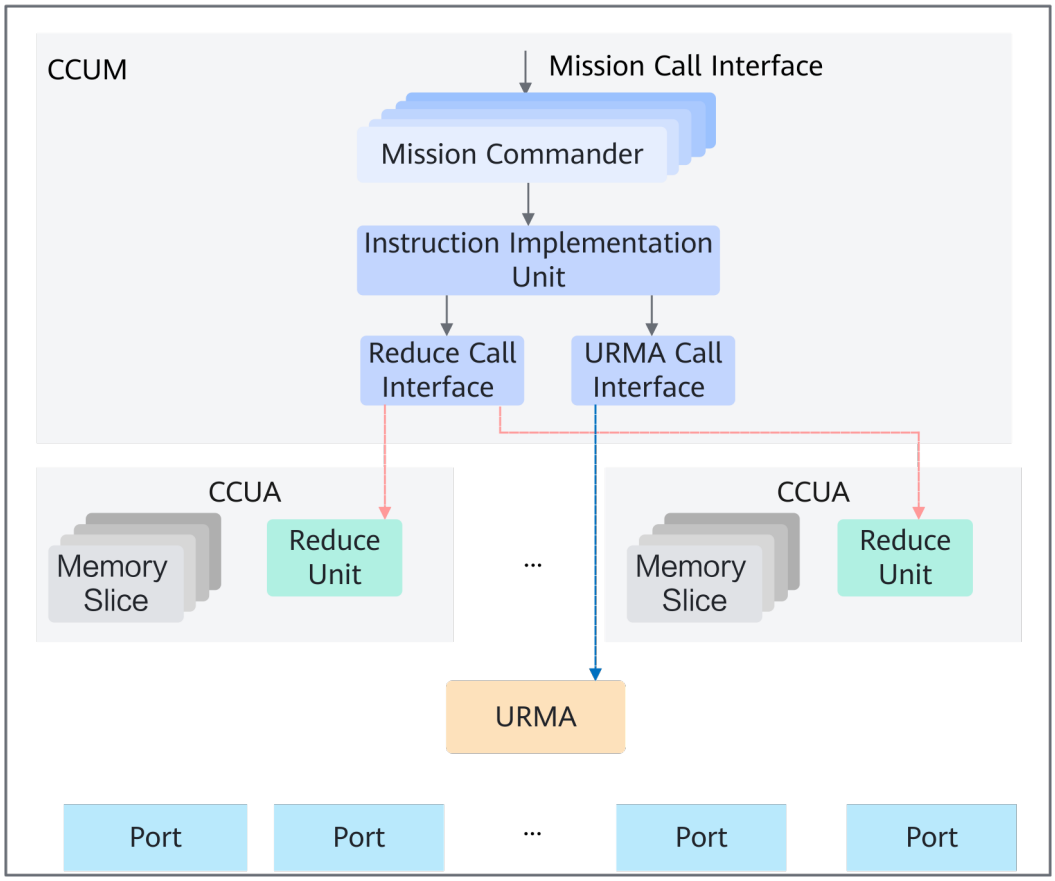
本芯片集合通信处理引擎 CCU（Collective Communication Unit），接收芯片上系统调度器下发的集合通信调用任务，根据软件预置的通信算法程序，自行进行节点之间的数据搬运与同步，自行根据数据状态完成集合通信的计算任务。

通过 CCU 指令实现集合通信，有效减少通信流量对系统总线带宽占用，并提升集合通信性能。

在集合通信处理引擎中，集合通信会被展开为多项并行小颗粒度任务，循环执行，灵活组合支持不同的集合通信算法，支持的典型算法如下：

- Broadcast
- Reduce Scatter
- All Gather
- All Reduce
- All2All
- All2Allv

图4-14 CCU 架构示意图



集合通信软件通过 CCU Management（CCUM）中的 Mission 任务的入口进行软件编程，硬件完成指令的解析和处理，并根据指令判断当前是执行 Reduce 计算还是 URMA 搬运。CCU Agent（CCUA）中集成了 MemorySlice 用作数据存储，集成了 Reduce Unit 用作数据计算。

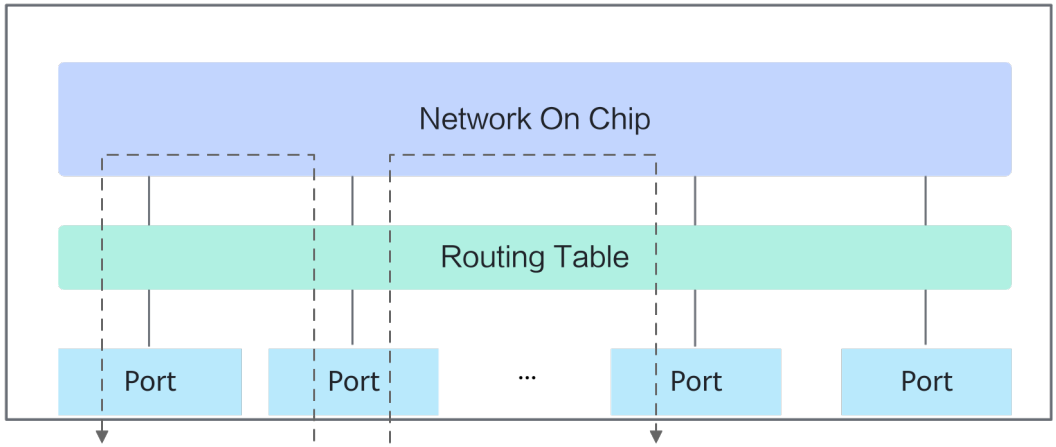
如果是 URMA 搬运则调用 URMA 执行数据搬移，可完成远端节点到本端节点 DRAM 或 MemorySlice 之间的灵活数据搬运。

如果是 Reduce 则调用 CCUA 的计算单元进行计算。

CCU 完成集合通信任务后通过 Mission 任务的编程接口上报任务完成状态。

4.6.5 UB On Chip Switch

图4-15 UB On Chip Switch 转发示意图

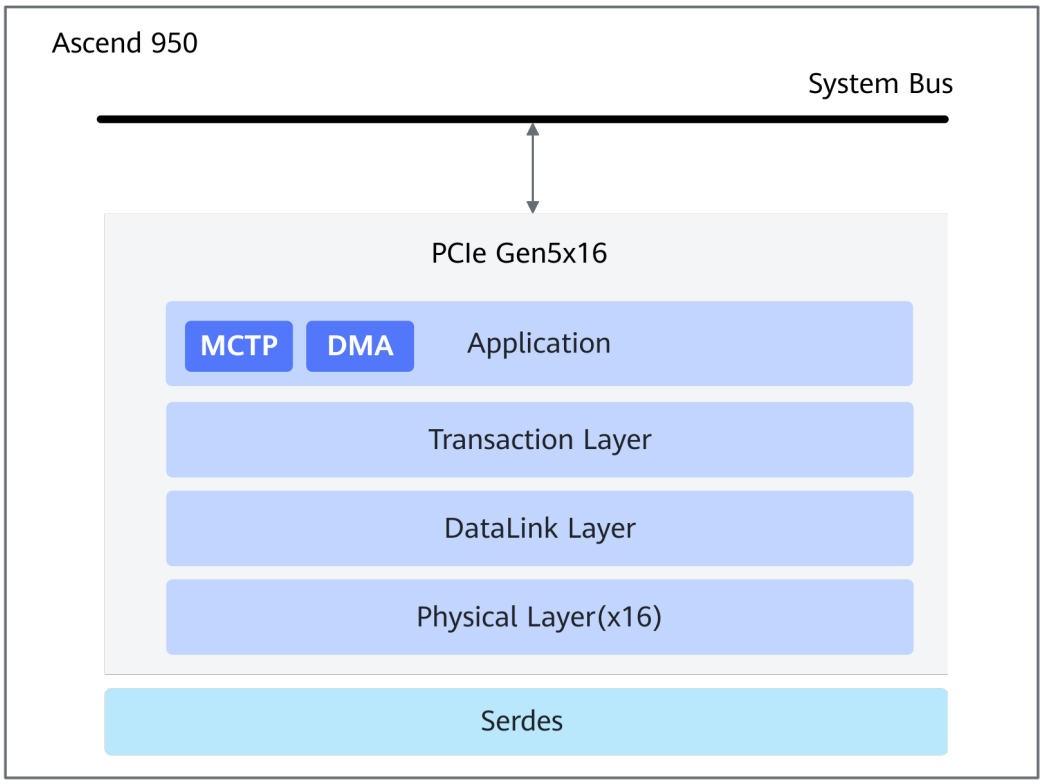


本芯片支持单 IO DIE 内 9 个 x4 Port 之间进行流量转发。从每个端口进入的流量在查询路由表后如判断该流量并非本芯片流量且判断得到转发的出端口，此时该流量会经过片上互联网络（Network on Chip，即 NoC）转发至出口端口送出。此转发流量不会进入计算 DIE，也不会占用 DRAM 带宽，在 IO DIE 上即完成数据转发。

本芯片支持注入流量和转发流量的混合部署，提供更多样的组网和业务规划可能性。

4.6.6 PCIe 5.0

图4-16 PCIe 5.0 架构示意图

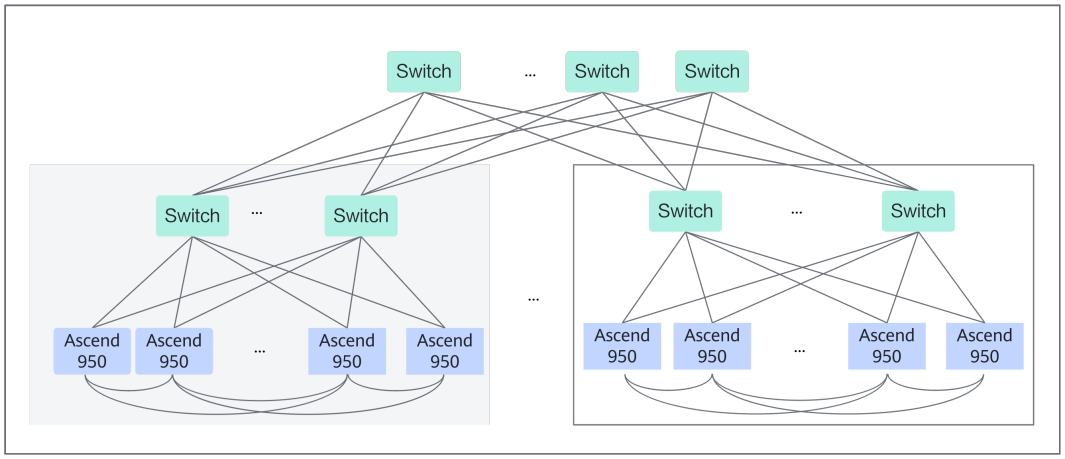


- 支持标准的 PCIe GEN5 协议，向下兼容 GEN4/3/2/1 协议。
- 支持 1x16/1x8/1x4/1x2 Link 模式。
- 支持 EP 与 RC 两种工作模式，模式静态选择。
- 支持 DMA、MCTP 加速器。

4.7 超节点能力

4.7.1 昇腾超节点

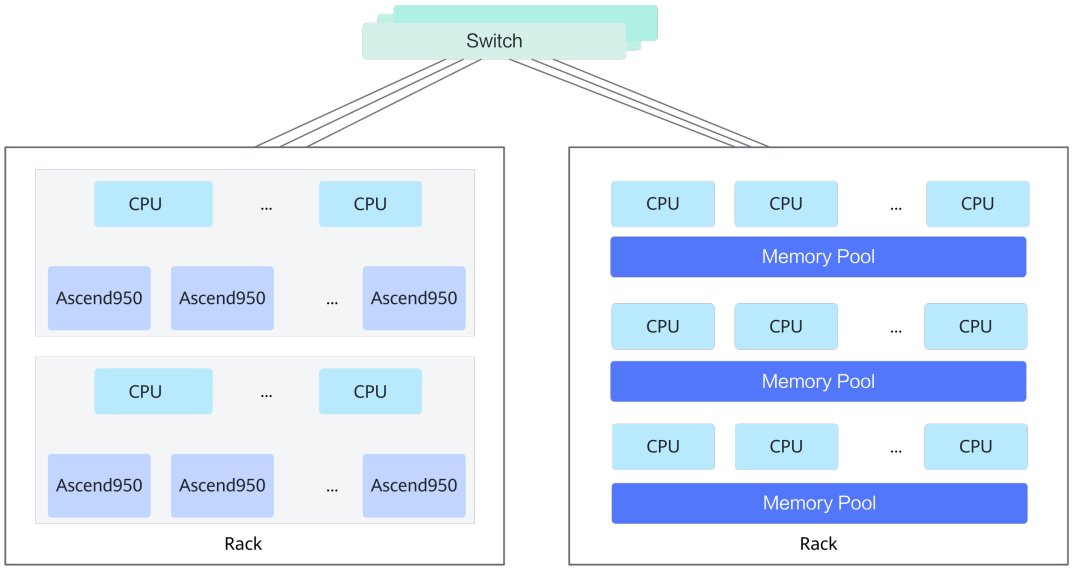
图4-17 昇腾 950 的一种超节点示意图



昇腾 950 芯片可以基于 UB 互连组建超节点，搭配 UB Switch 可以组建多达 K 级别的超节点。在超节点内基于 UB 实现高效的通信。基于 UB 互连协议可以组建多种不同的拓扑结构，比如常见的 Full Mesh 互连和 Clos 互连拓扑，或者是灵活的混合组网拓扑。

4.7.2 昇腾超节点与超大内存池组网

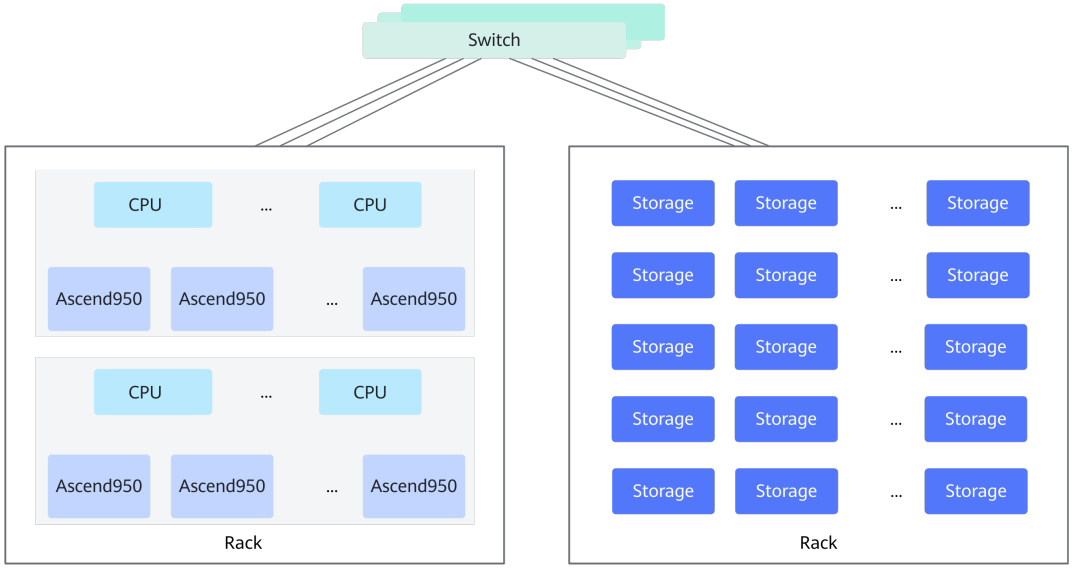
图4-18 昇腾 950 访问 CPU 超大内存池示意图



基于 UB 互连可以构建超大内存池，昇腾 950 Rack/Pod 的计算芯片可以通过 UB 端口直接访问该超大内存池，实现高带宽和低延时的访问，充分有效利用 CPU 的内存资源。

4.7.3 昇腾超节点与超大存储资源池组网

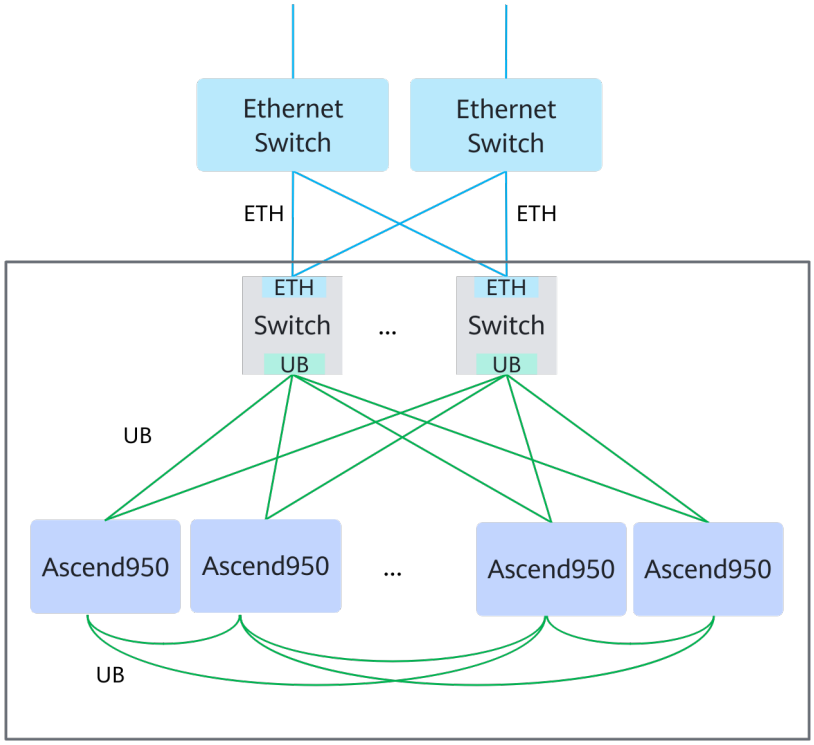
图4-19 昇腾 950 直接访问超大存储资源池示意图



基于 UB 互连可以构建超大存储资源池，昇腾 950 Rack/Pod 的计算芯片可以通过 UB 端口直接访问该超大存储资源池，不需要中间的存储协议转换开销，从而实现高带宽和低成本的存储资源访问。

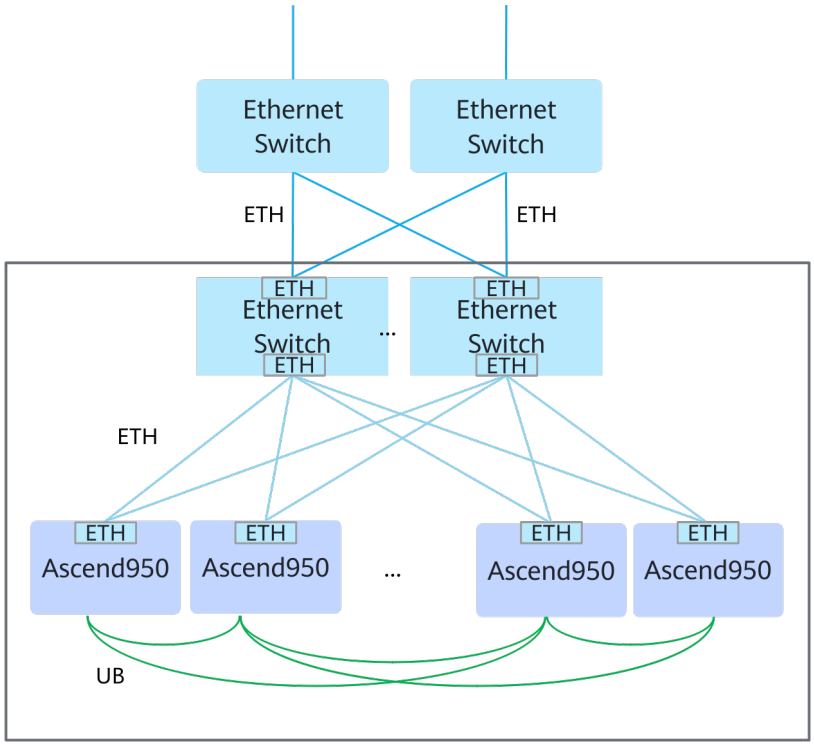
4.7.4 昇腾超节点与以太世界互通

图4-20 昇腾超节点基于 UB Switch 转换为以太网与以太世界互通示意图



基于 UB 组建的昇腾超节点提供了领先的竞争力，为了更便捷地接入到已有数据中心的以太交换网络，当前昇腾超节点基于 UB Switch 提供了从 UB 到 Ethernet 世界的转换功能，可以无缝接入到以太网络，不需要额外的硬件成本。

图4-21 昇腾芯片支持以太网与以太世界互通示意图



当前昇腾 950 芯片可直接基于 UBoE 技术接入到以太网络中，直接利用业界标准的以太网交换机进行组网。

5 更多参考

您可以通过本章节获取更多的参考资源：

- 《Ascend C 编程指南》请参考 LINK。
- 《基于灵衢的超节点参考架构白皮书》请参考 LINK。