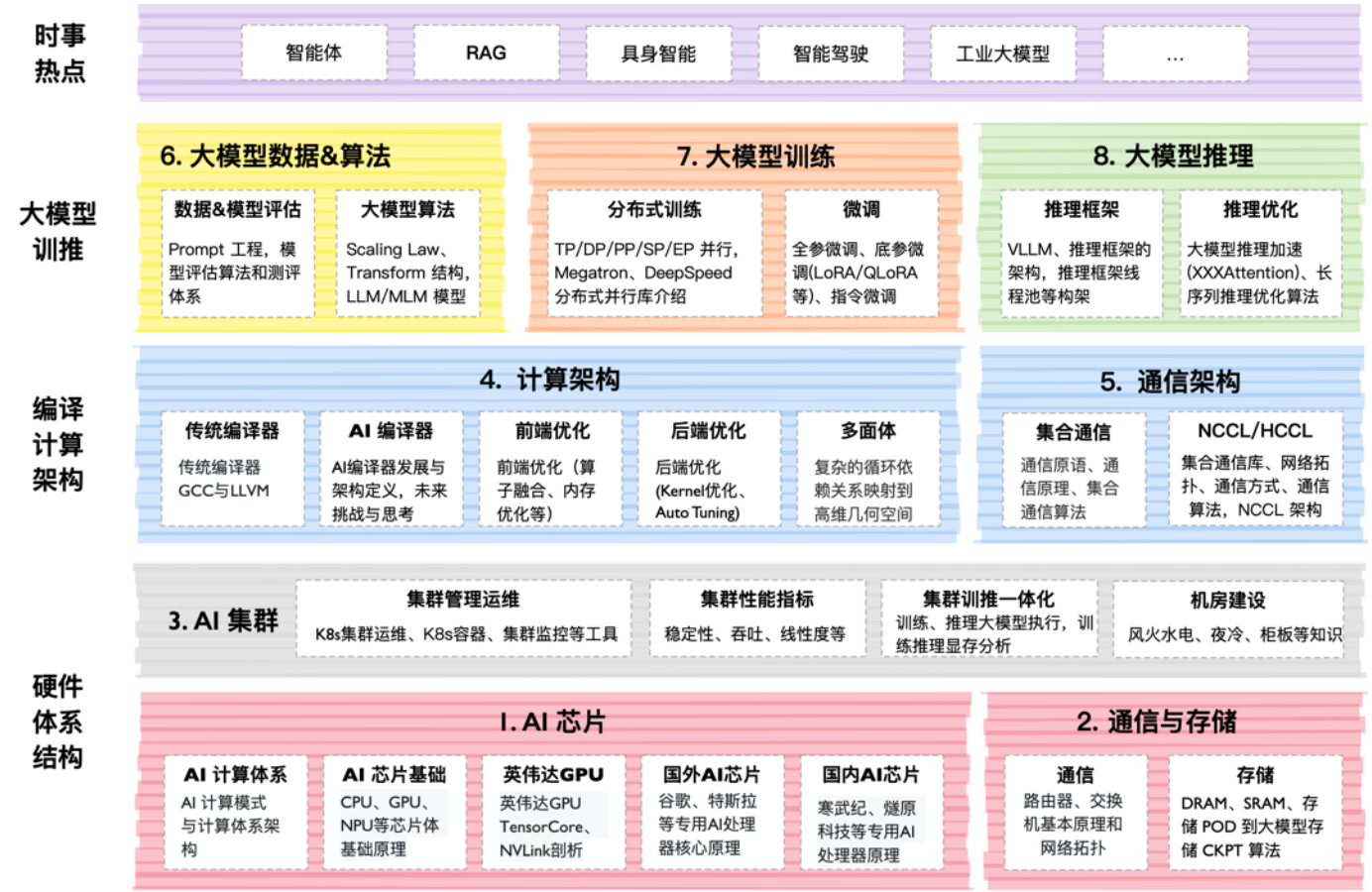

ZOMI
高性能计算
发展趋势
豆包AI生成

Content


计算集群之路
高性能计算定义
算力、分类、集群
高性能计算发展趋势
硬件架构、软件与编程
集群计算：发展历程
整体介绍、初代、二代、三代、未来
集群计算：系统挑战
总览、挑战、架构发展
集群计算：系统架构
L01/L1机房，L2基础，L3 使能

AIInfra Architecture



三类计算集群的主要区别

集群类型	应用目的	数据类型	计算特征	网络特征	存储特征
HPC高性能计算	国防科研	FP64	计算密集，高并行	密集通信（每个应用流量特征不同，部分可隐藏）	密集&复杂 IO
AI人工智能计算	AI训练推理	FP16/BF16/FP8/FP4		密集通信（部分可隐藏）	密集 IO（按节奏迭代）
云数据中心	互联网云计算	INT32 FP32	通用计算，高并发	分散通信	分散/密集 IO

Content

1. 核心硬件（高性能-处理器、存储、网络、服务器）
2. 基础软件（编译器与运行时、计算库、通信中间件、存储系统、调度系统）
3. 应用软件（发展历程、行业应用趋势）

01

硬件

Hardware

硬件 Hardware

- HPC 硬件发展历程和未来趋势从**高性能网络、处理器、服务器、存储器**四个核心维度展开分析。
- 其演进逻辑始终围绕性能突破、能效优化与场景适配展开：

1. **高性能处理器：**从通用多核到异构计算
2. **高性能网络：**从低延迟到高带宽互联
3. **高性能存储器：**从容量扩展到存算协同
4. **高性能服务器：**从单机性能到绿色化集群

01. 处理器

发展历程1：CPU主导时代

- 早期 HPC 依赖于大规模并行 CPU 集群（e.g. Intel Xeon, AMD Opteron 系列）。
- 通过提升主频、增加核心数量（多核/众核）和优化指令集实现性能增长。

Intel Xeon

AMD Opteron

发展历程1：CPU主导时代

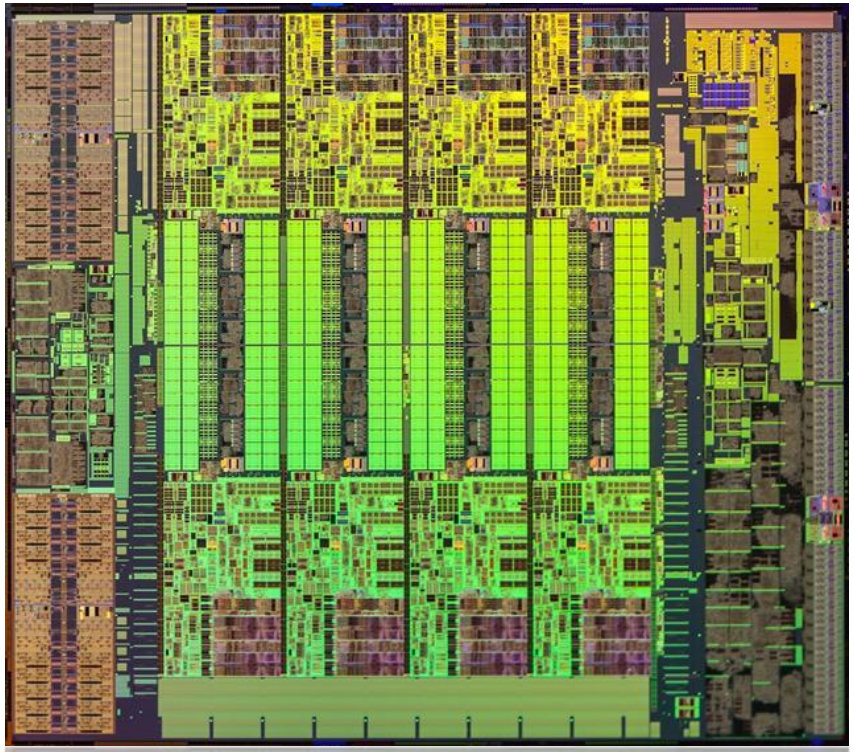
- 早期 HPC 依赖于大规模并行 CPU 集群（e.g. Intel Xeon, AMD Opteron 系列）。
- 通过提升主频、增加核心数量（多核/众核）和优化指令集实现性能增长。

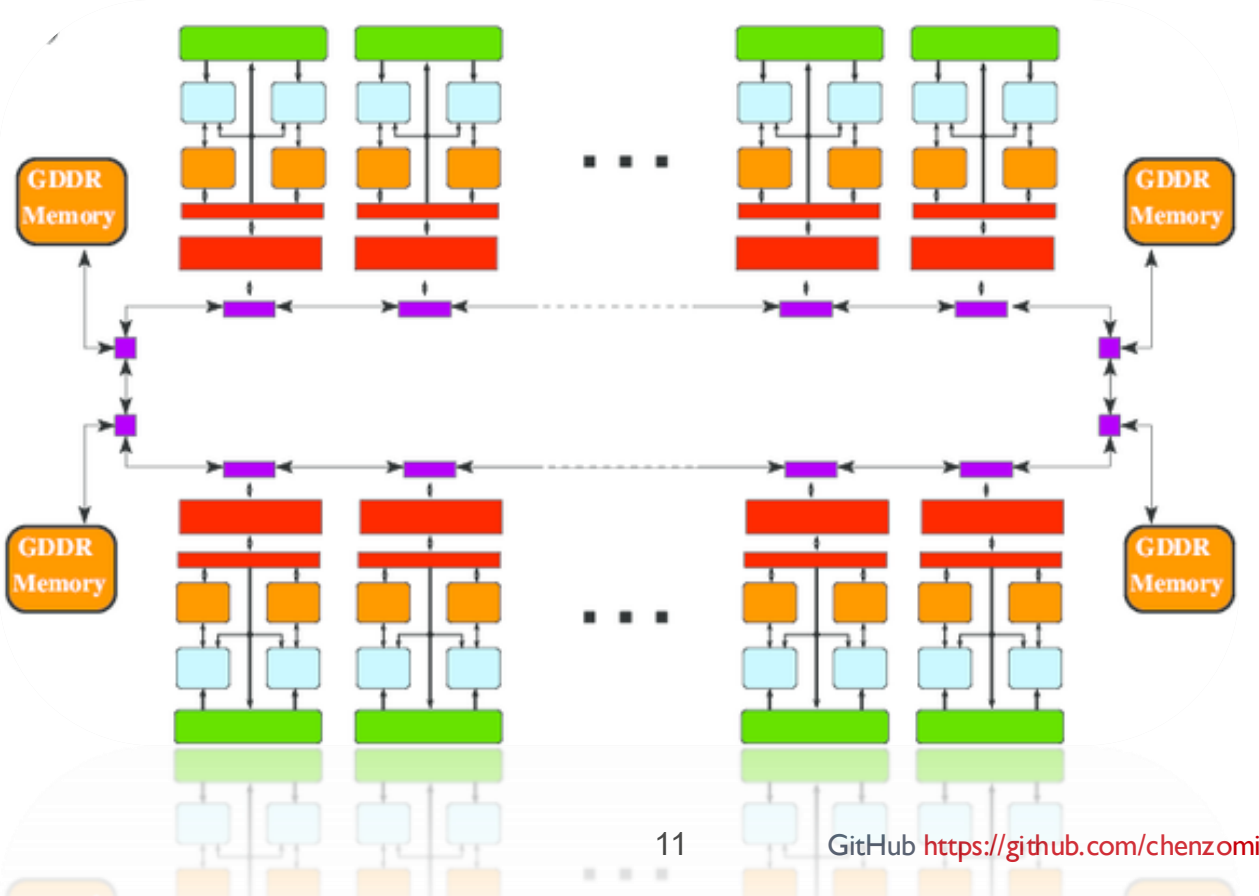
Intel Xeon

AMD Opteron

处理器发展历程：协处理器兴起

- 专用加速：Intel Xeon Phi（MIC架构，2012-2020）尝试众核路线。
- GPU 加速：NV CUDA 革命性地将 GPU 用于通用计算， Tesla 成为通用并行计算标杆。



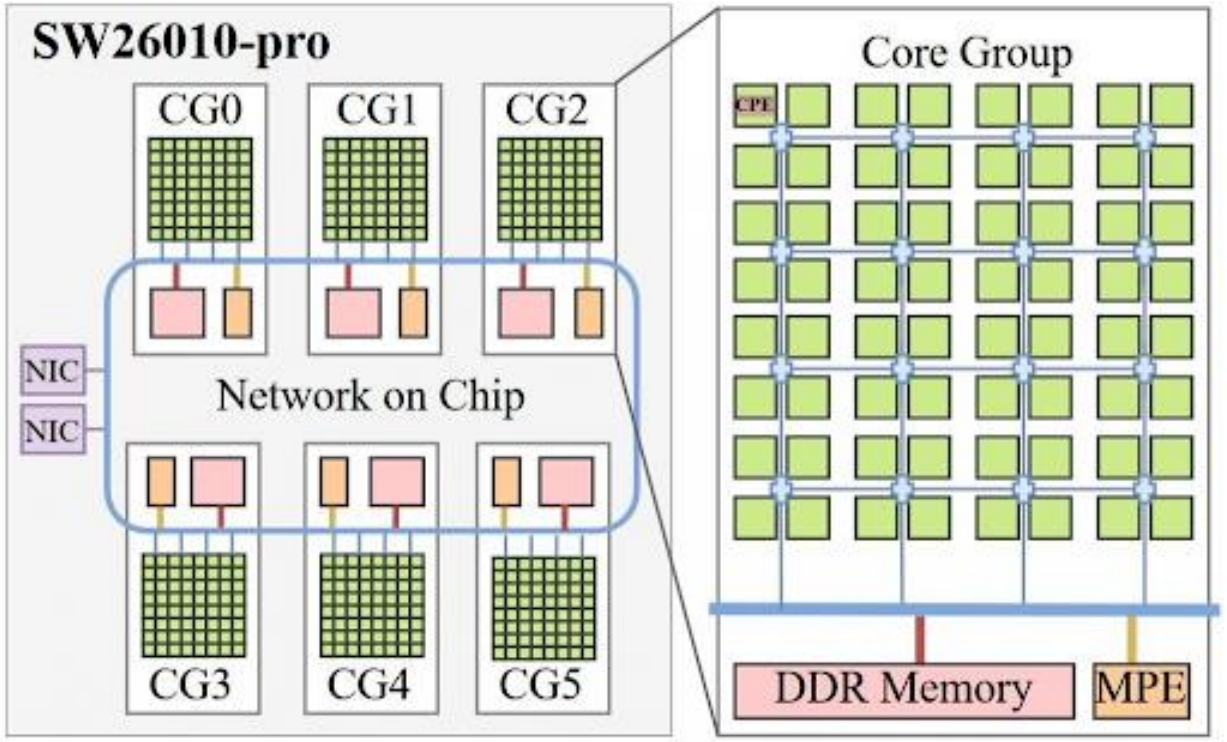


发展历程2：协处理器兴起

- 专用加速：Intel Xeon Phi（MIC架构，2012-2020）尝试众核路线。
- GPU 加速：NV CUDA 革命性地将 GPU 用于通用计算， Tesla 成为通用并行计算标杆。

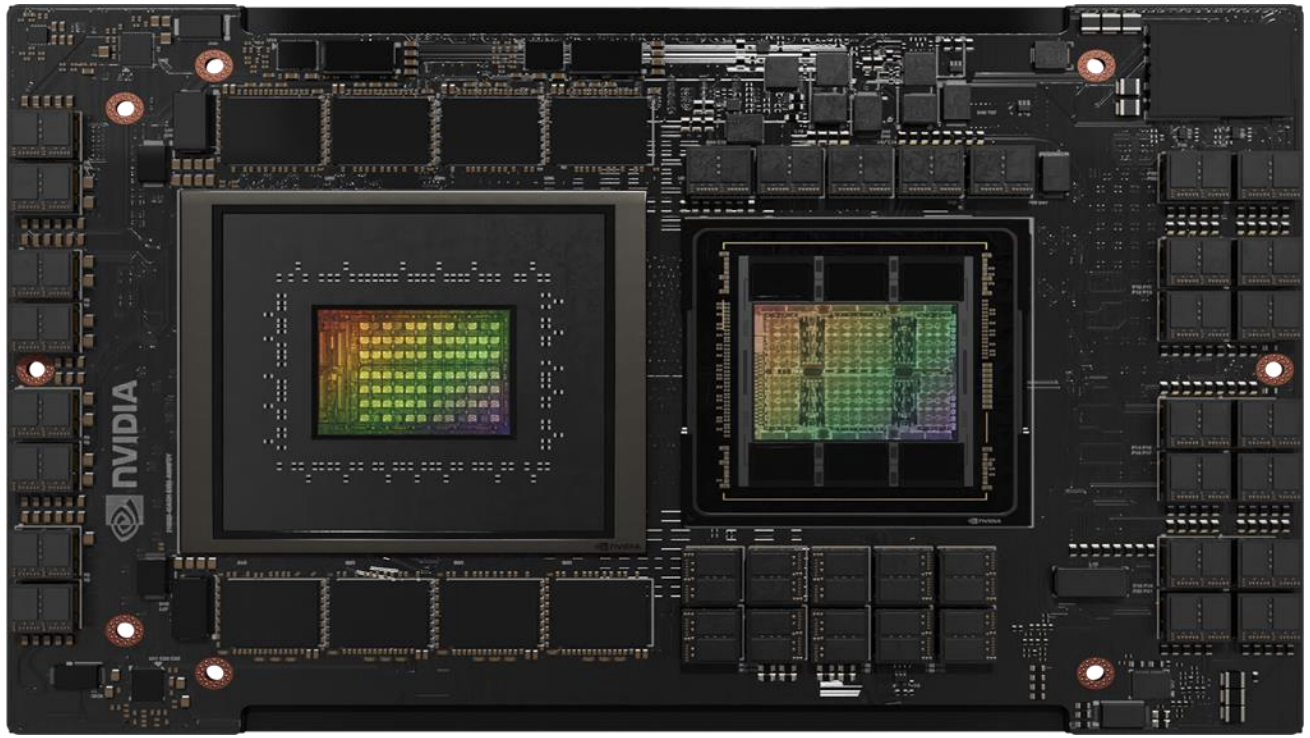
发展历程3：国产突破

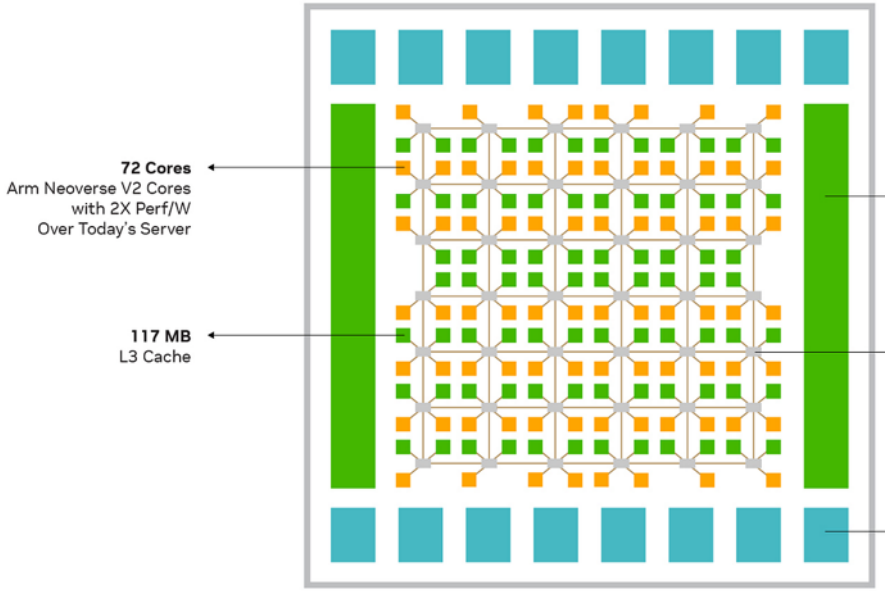
- 申威 SW26010（256核+1主核）实现每秒3万亿次浮点运算，支撑国产E级超算。采用自主指令集，集群“神威·太湖之光”2016 年 TOP500 榜首获全球第一超算。



发展历程3：Arm 架构 HPC 突破

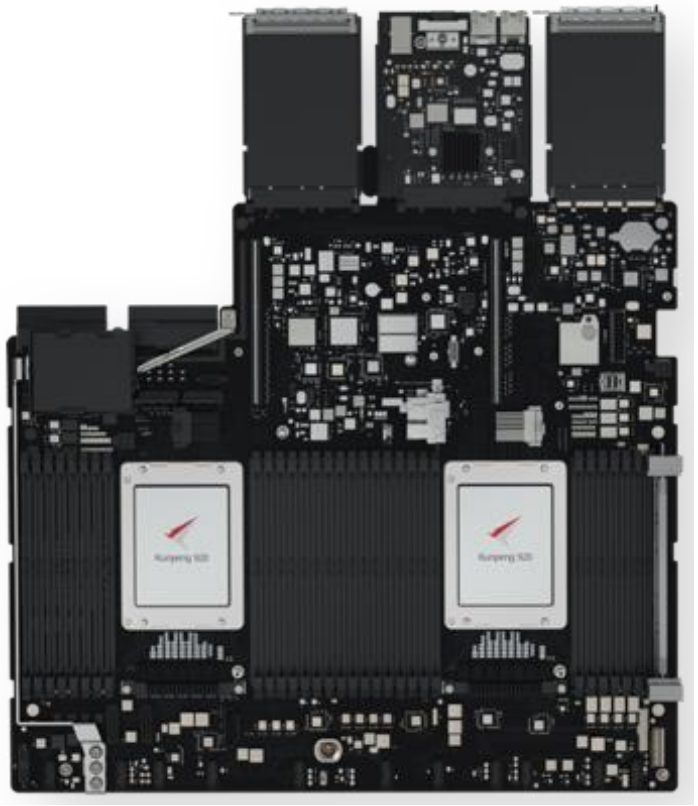
- NVIDIA Grace CPU，提供内存子系统革新：LPDDR5X + 纠错码 → 能效提升2倍；CPU-GPU 一致性缓存：NVLink-C2C 直连延迟降至 1/10。





发展历程3：Arm 架构 HPC 突破

- 鲲鹏 920 ARM-based处理器：7nm 工艺，ARM 架构授权，华为自主设计。通过优化分支预测算法、提升运算单元数量、改进内存子系统架构等一系列微架构设计，提高处性能。

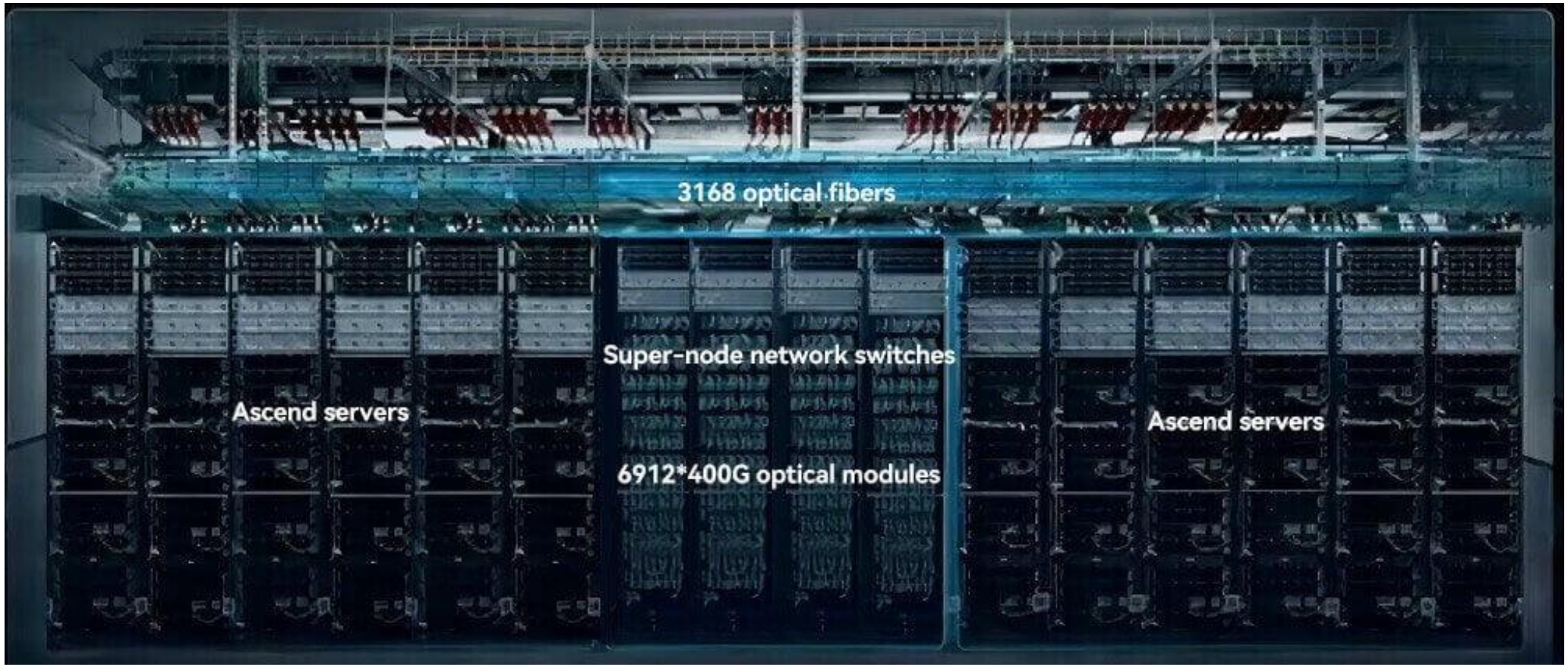


架构优势与改造

特性	传统Arm（移动端）	HPC优化版（服务器级）
指令集	精简指令集（RISC）	扩展SIMD指令（SVE/SVE2）
核心规模	多核低频（能效优先）	512核以上众核架构（Fujitsu A64FX）
内存系统	低带宽LPDDR	HBM2e（>1TB/s带宽）
功耗管理	动态调频（DVFS）	精细功耗门控（Per-core PowerGating）

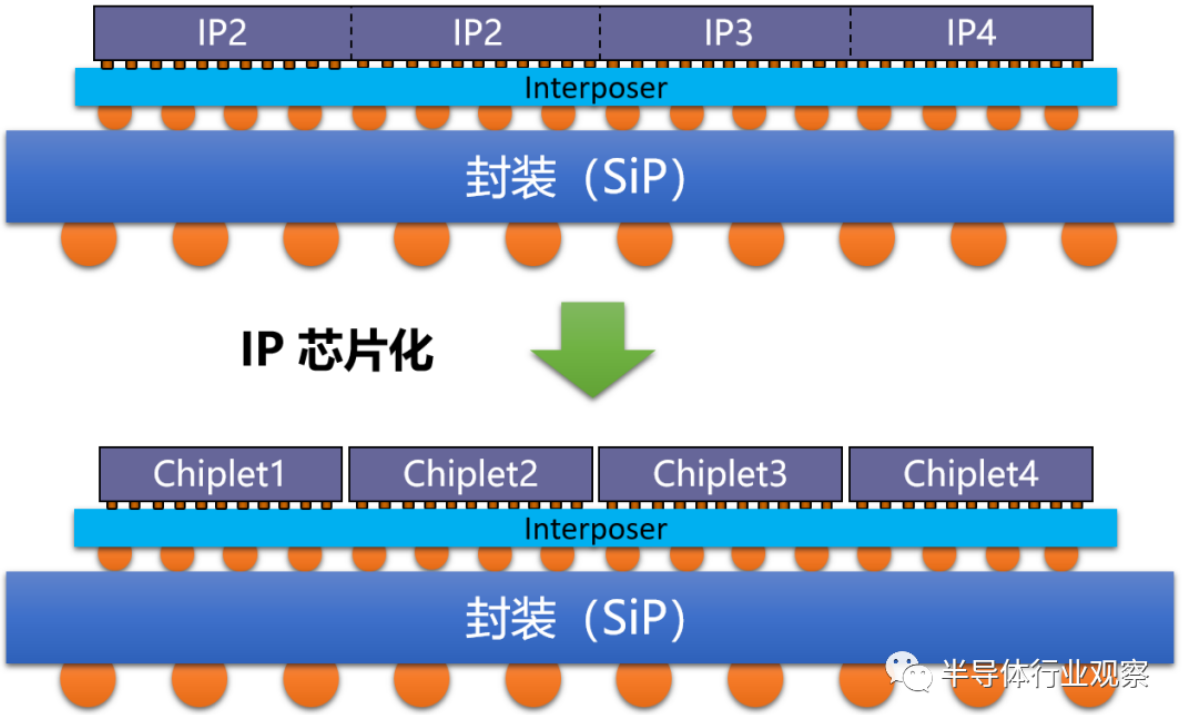
处理器发展趋势：异构多样化与国产化

- **GPU 主导加速市场**：NVIDIA H100/AMD MI300X 成为 AI/HPC核心算力，国产替代（e.g. 寒武纪、华为昇腾）加速发展。



处理器发展趋势：异构多样化与国产化

- **Chiplet 技术：** AMD、 Intel 通过多芯片封装提升集成度和良率（e.g. Intel Ponte Vecchio 含47 颗 Chiplet）。
- **存算一体技术：** 台积电 3D Fabric 技术将计算单元堆叠至存储层（e.g. CXL 协议设备），突破冯·诺依曼瓶颈。



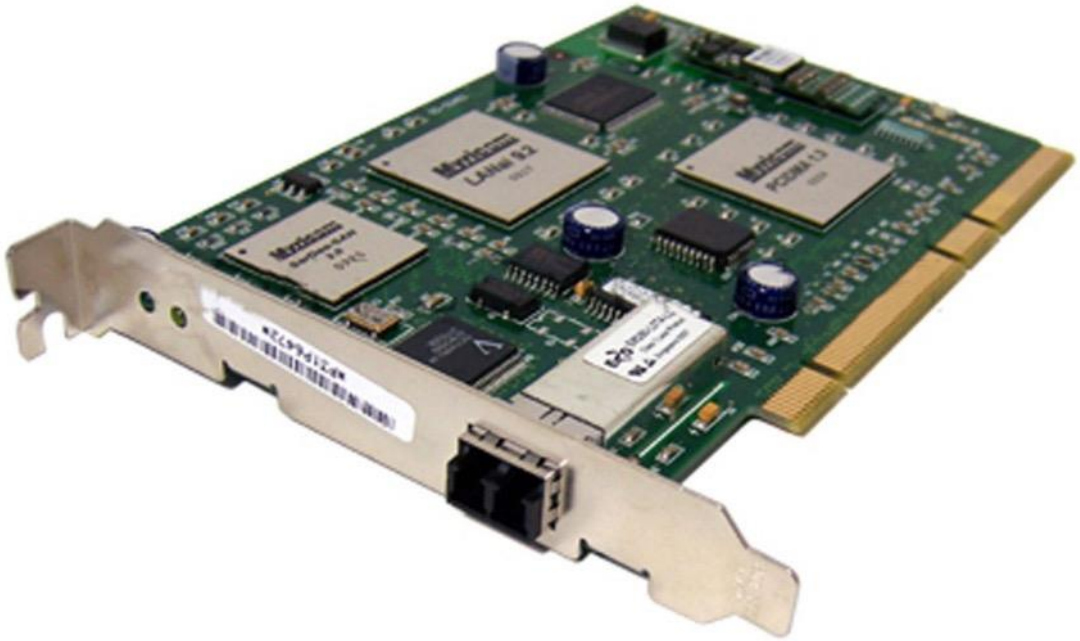
02. 高性能网络

Interconnect

网络发展历程：早期阶段（1990s-2000s）

- 以太网主导： 千兆以太网（GigE）成本低被广泛采用，但延迟高（>100μs）、带宽瓶颈明显。
- 专有网络兴起： Myrinet、Quadrics 等私有协议网络出现，延迟~10μs，但生态封闭制约普及。

Ethernet



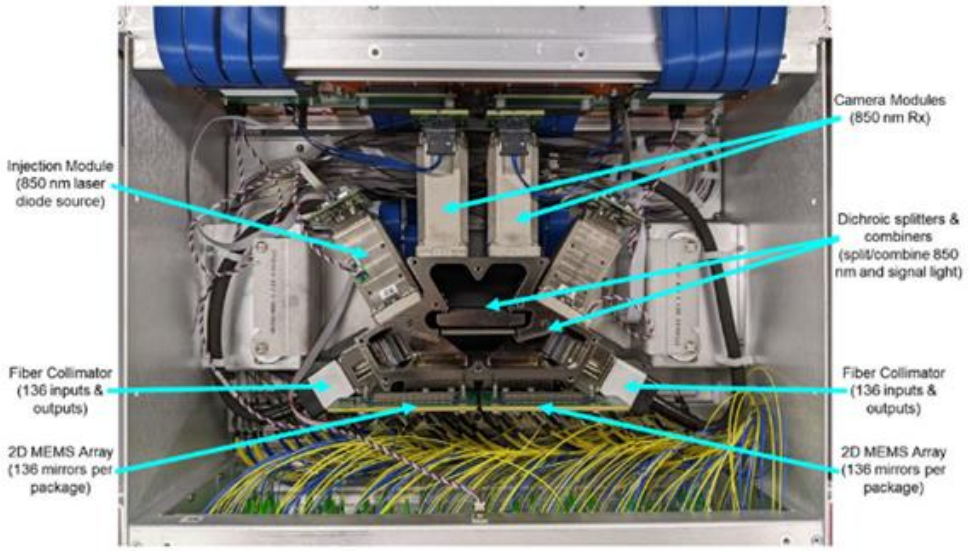
Myrinet

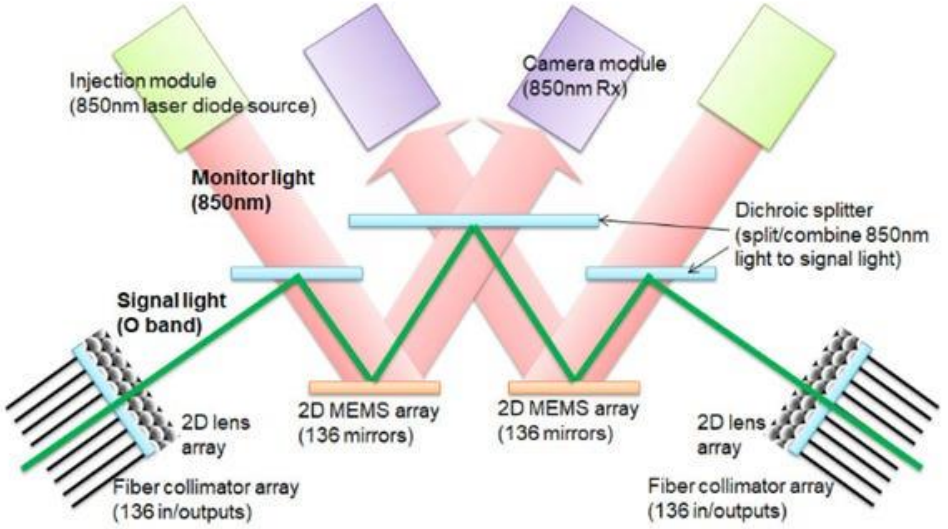
网络发展历程：主流技术成型（2010s至今）

- **InfiniBand：** Mellanox 主导，采用 RDMA 实现微秒级延迟和百 GB/s 带宽，成为 HPC 主流。
- RoCE：基于以太网 RDMA ，兼顾低成本与高性能，华为、阿里云等国产厂商加速布局。
- NVLink ： NV 专为 GPU 互联设计高速总线，用于节点内多卡互联，演进 NV Fusion。

高性能网络：低延迟与融合

- **InfiniBand 持续领先：** NVIDIA Quantum-2 支持 400Gb/s 带宽， NDR 1.6Tb/s。
- **RoCE v2与智能网卡：** 借助 DPU/IPU 降低 CPU 负载，提升以太网竞争力。
- **光互连技术：** 硅光集成、CPO（共封装光学）、 OCS（光学链路开关）成为下一代网络关键。
- **新互联协议：** 华为灵渠总线互联，实现 CPU2NPU、NPU2NPU 新一代的集群互联。





03. 高性能存储

Storage

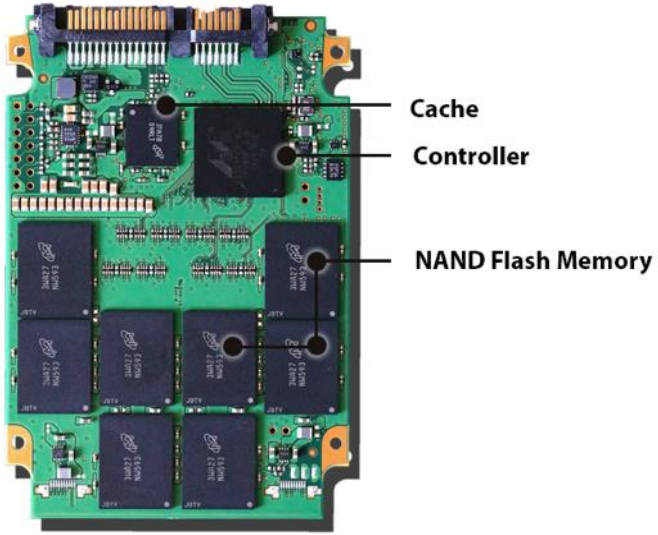
发展历程1：硬盘时代

- **硬盘革命， SSD 取代 HDD 提升 I/O 速度:**
  - 容量优先： 希捷、西数推出 HDD，受限于 IO 延迟（ms 级），难满足 HPC 需求。
  - NVMe SSD 普及： 三星 PM1733（30TB）顺序读写达 7GB/s，延迟降至 50μs。

HDD

Shock resistant up to 55g (operating)
Shock resistant up to 350g (non-operating)

SSD



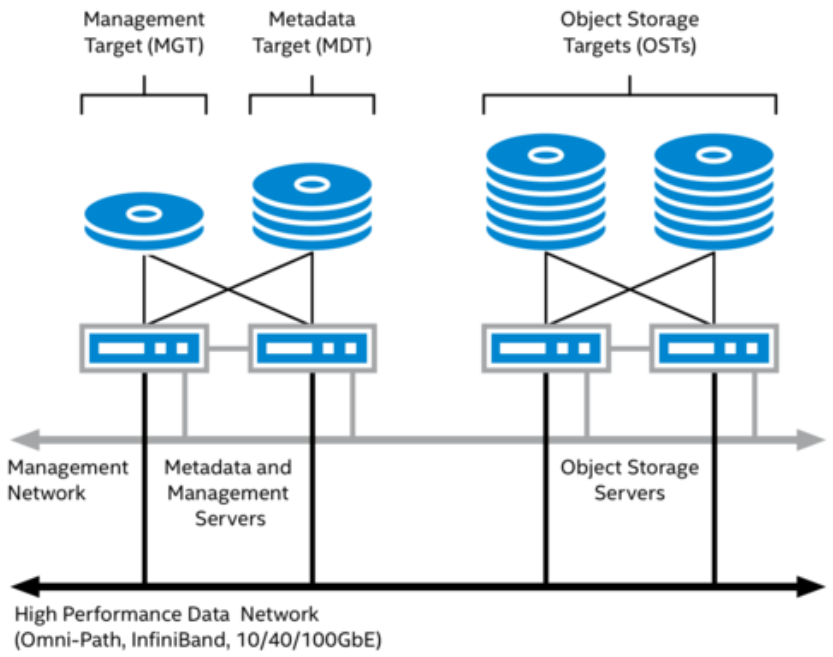
Shock resistant up to 1500g
(operating and non-operating)

Micron SSD

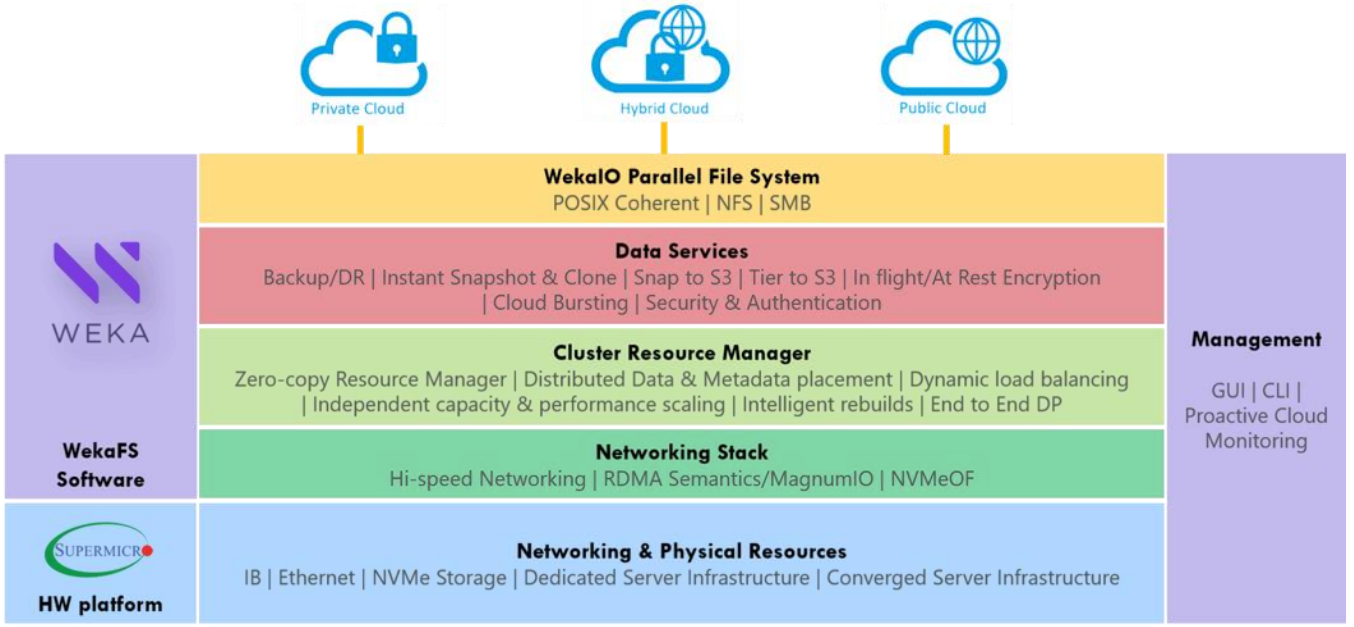
发展历程2：分布式存储 & 新存储FS

- 分布式文件系统：Lustre、Ceph 等 FS 支撑 EB 级数据吞吐，e.g. Lustre 带宽突破 1TB/s。
- AI 驱动存储优化：WekaIO、VAST Data 通过 NVMe-oF + RDMA 实现全闪存存储集群。

Lustre

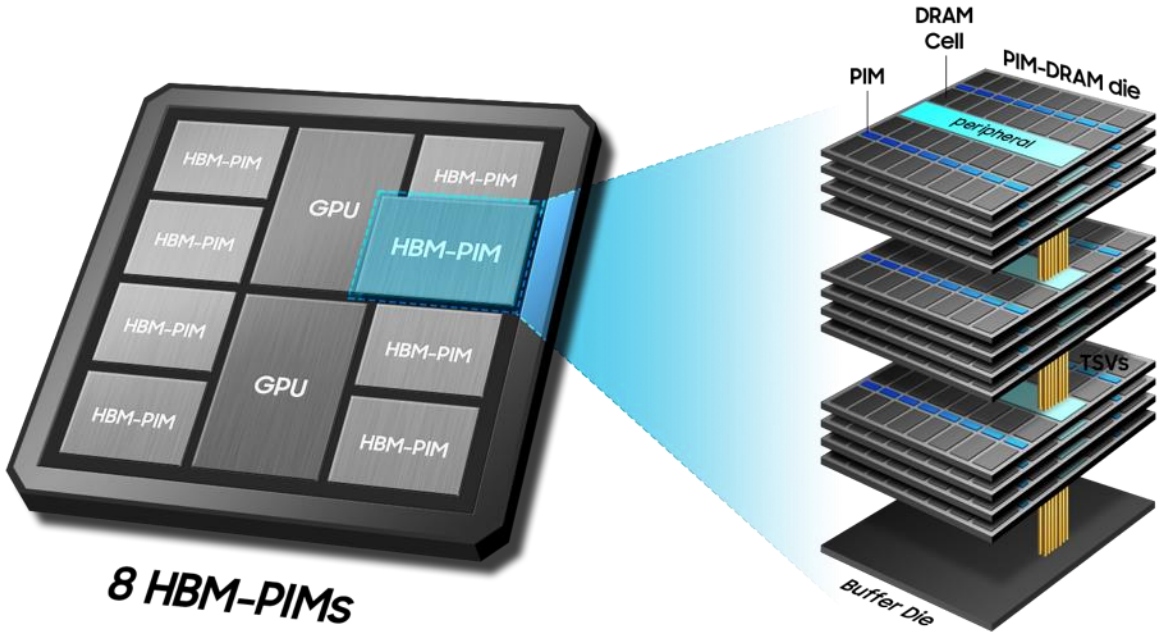


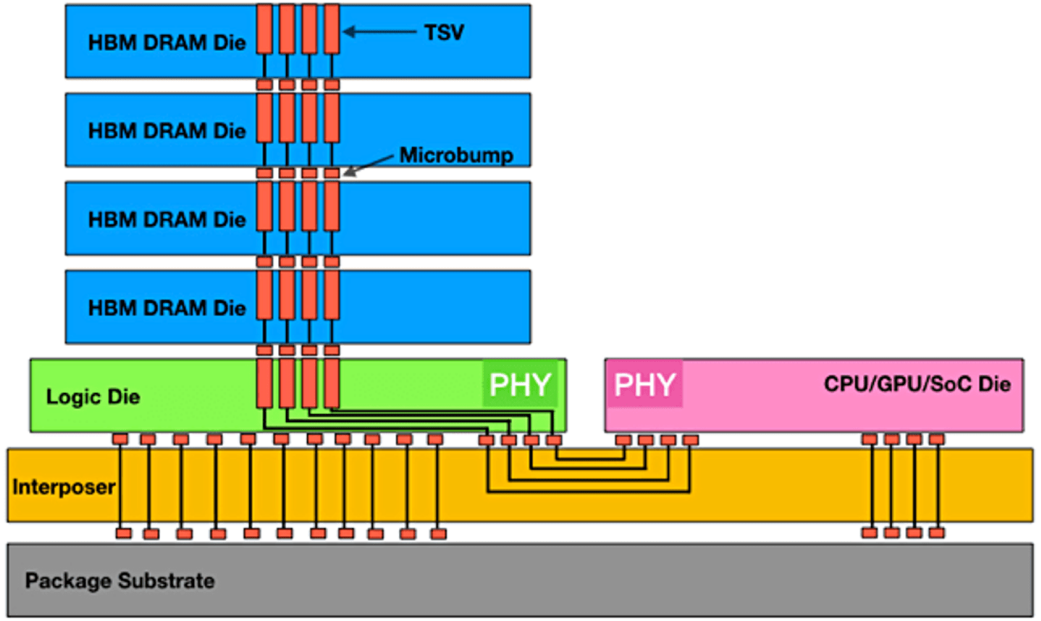
WekaIO



高性能存储发展历程3： 新存储技术革命

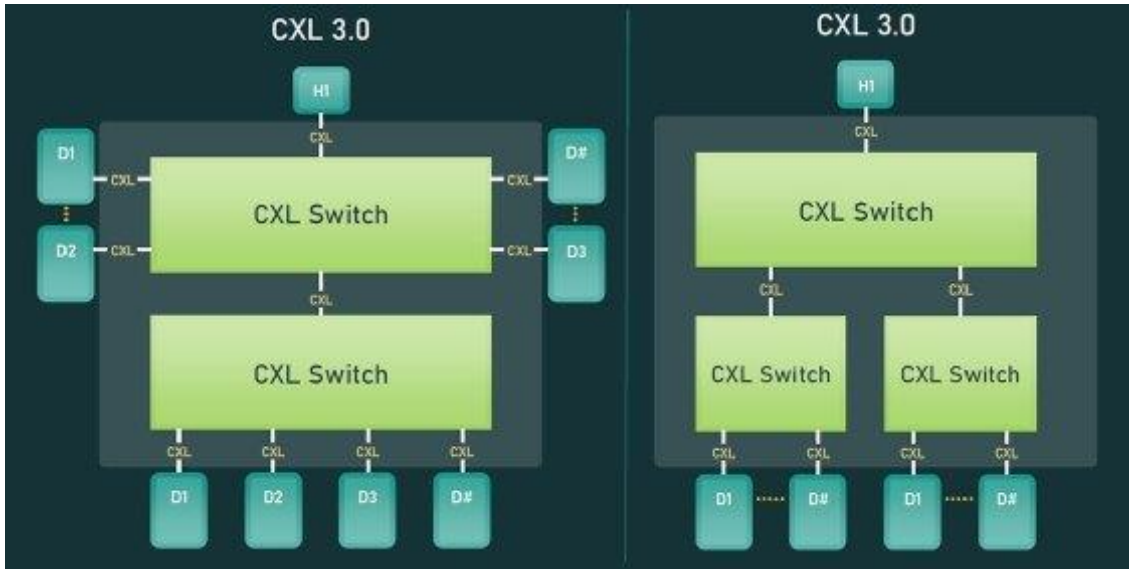
- SCM（存储级内存）： Intel Optane 持久内存试图弥合内存与存储间的性能鸿沟。
- HBM（高带宽内存）： NPU/GPU 高度集成与依赖， HBM3 每引脚数据速率提高到 6.4Gb/s，单设备带宽 819GB/s 。

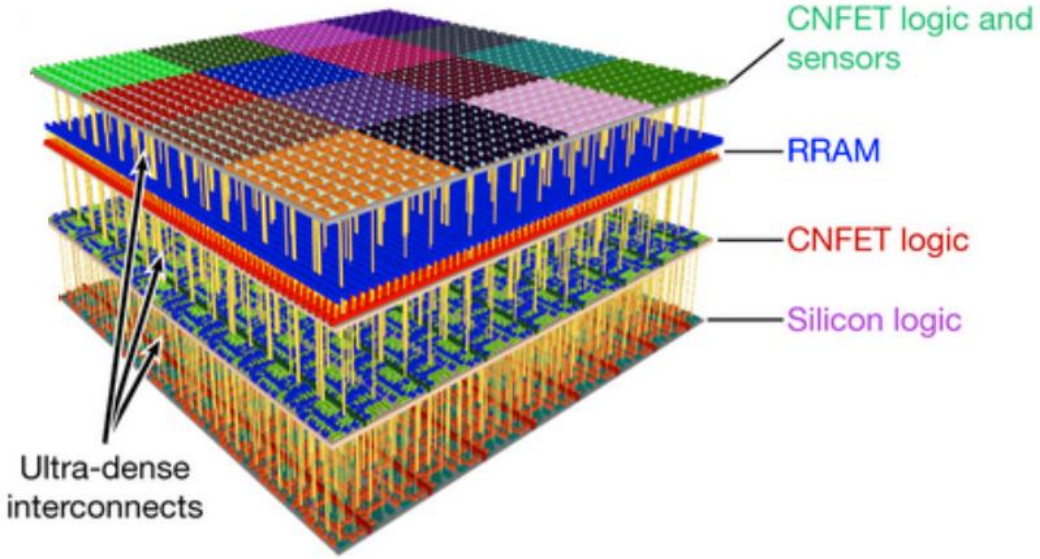




高性能存储：近计算与层级优化

1. **存储级内存 SCM：** CXL 3.0 协议推动存储池化，实现 CPU/GPU 共享内存资源。
2. **HBM3e 与 HBM4：** 2025 年 HBM3 带宽突破 1TB/s，成 AI 芯片标配。
3. **存算一体架构：** 计算单元嵌入存储层，减少数据搬运。
4. **存算一体器件：** 忆阻器 Memristor、相变存储器 PCM 支持内存计算，降低搬运功耗。





04. 高性能服务器

Server & Cooling

发展历程 1：性能导向阶段（2000s）

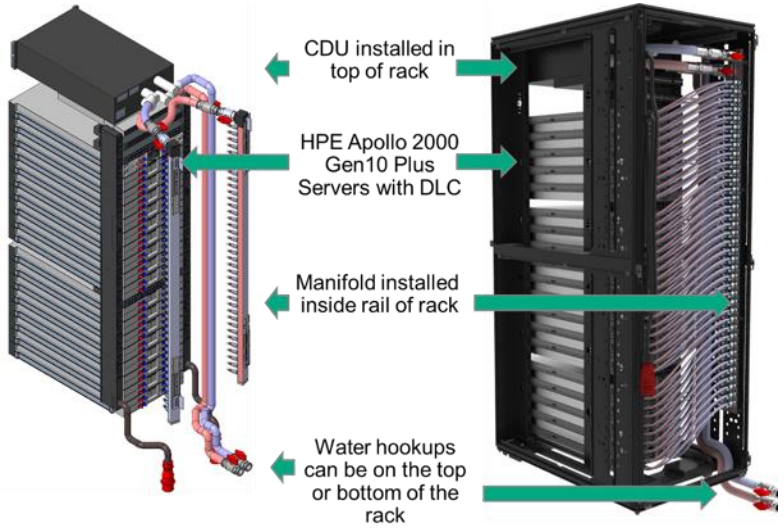
- **刀片服务器普及：** HP BladeSystem、IBM BladeCenter 通过高密度设计提升机架利用率。
- **风冷散热瓶颈：** 单机柜功耗超 20kW 后，传统风冷效率骤降（PUE > 1.5）。

刀片式服务器

机架式服务器

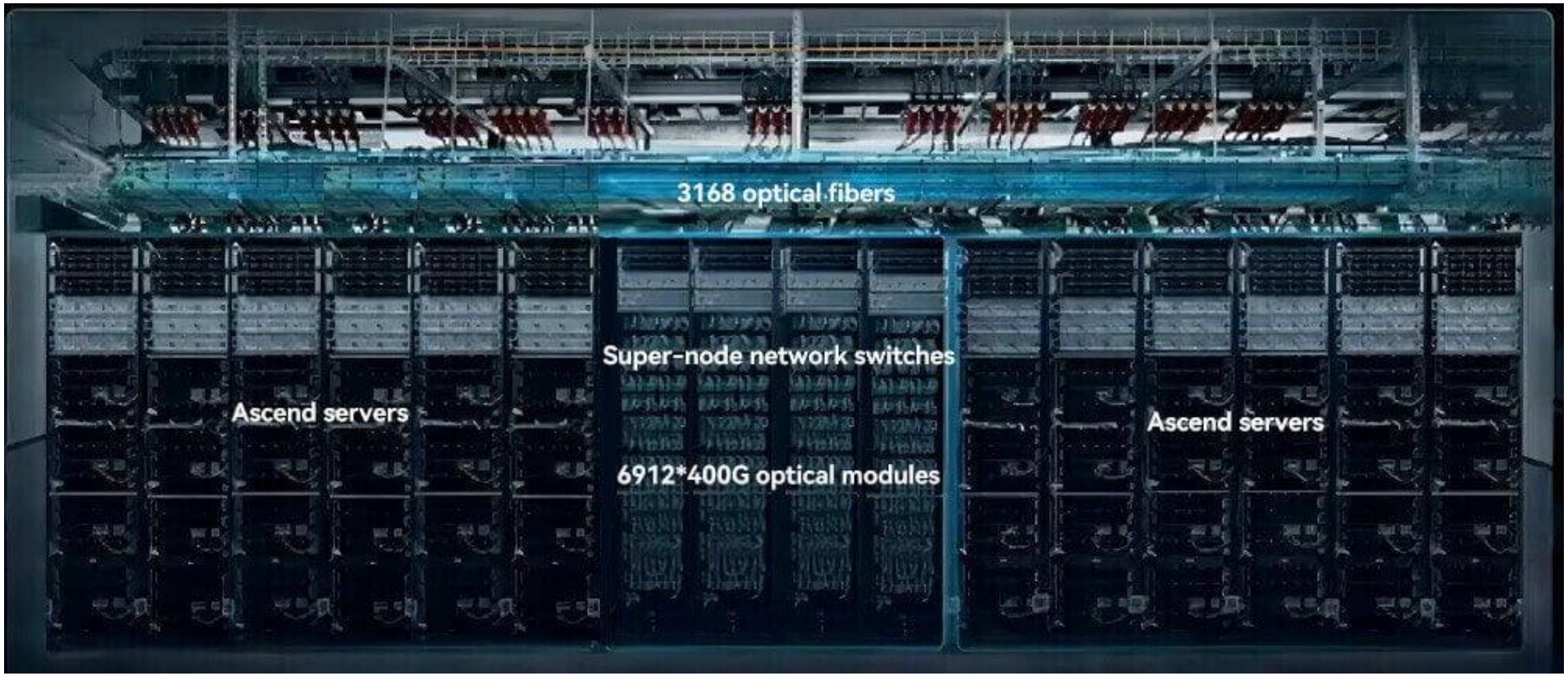
发展历程 2：绿色化阶段（2010s ~ 至今）

- **散热技术突破，传统风冷机架 to 高密度液冷机架演进**：
  - 冷板液冷：直接冷却CPU/GPU（e.g. 曙光浸没式液冷，PUE ≈ 1.04）
  - 浸没式液冷：整机浸入非导电冷却液（e.g. 阿里云“麒麟”数据中心 PUE ≈ 1.09）
- **模块化设计**：超算模块化机柜（e.g. Frontier）实现计算、存储、网络的灵活扩展。



未来趋势

- 液冷普及：欧盟要求 2025 年后数据中心 PUE ≤ 1.3，液冷从实验走向标准方案。
- 整机柜设计：曙光硅立方实现机柜级液冷，昇腾推出 Cloud Matrix 超节点。

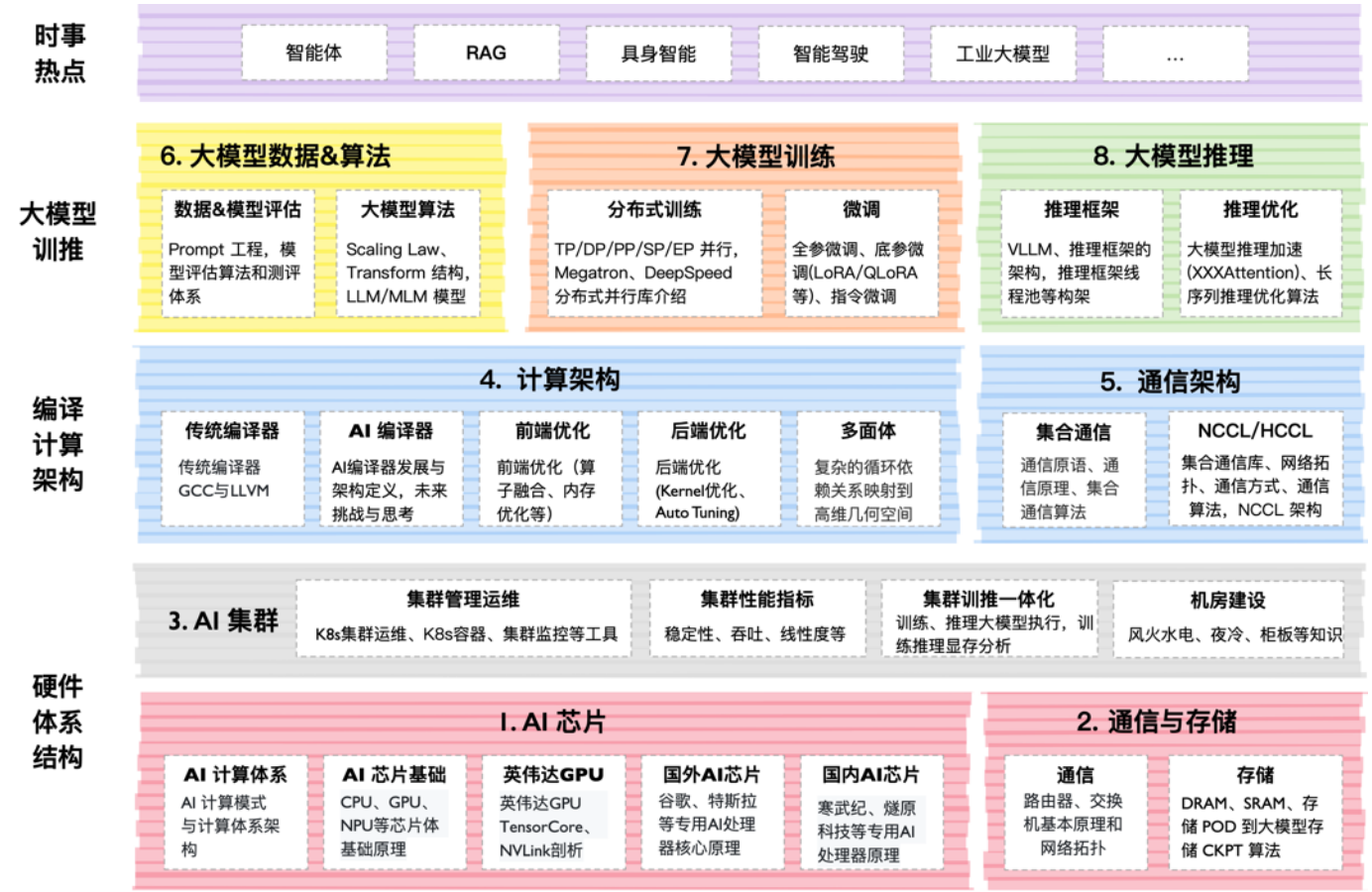


硬件发展驱动力

领域	核心驱动力	代表技术/案例
处理器	提升能效比、支持混合精度计算	ARM/AMD/NVIDIA GPU, Chiplet封装
网络	破解通信瓶颈，降低延迟	InfiniBand NDR, RoCE v2 + DPU
服务器	突破功耗墙，降低PUE	曙光浸没液冷（PUE<1.1）
存储器	打破“内存墙”与“存储墙”	HBM3

总结与思考

Question？ HPC vs AI



Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIInfra

引用与参考

- https://zhuanlan.zhihu.com/p/683671511

- PPT 开源在：https://github.com/chenzomi12/AIInfra