

RISC-V走向数据中心之路

武汉元石智算科技有限公司

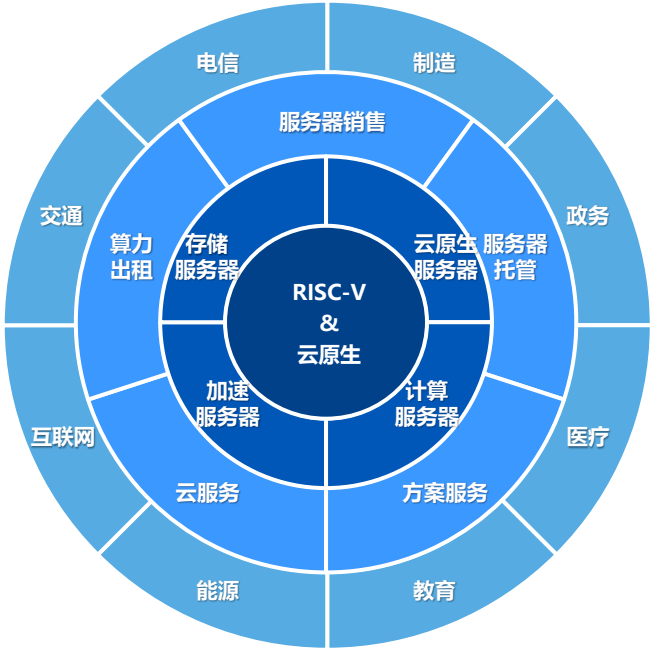
2024.08

- 关于元石智算
- RISC-V的发展及数据中心的演化
- 元石智算RISC-V跨架构云原生平台
- 观察与展望
- 元石智算核心研发平台

元石智算: RISC-V云原生服务器与解决方案提供商 *MetaStone*

武汉元石智算科技有限公司 *MetaStone*

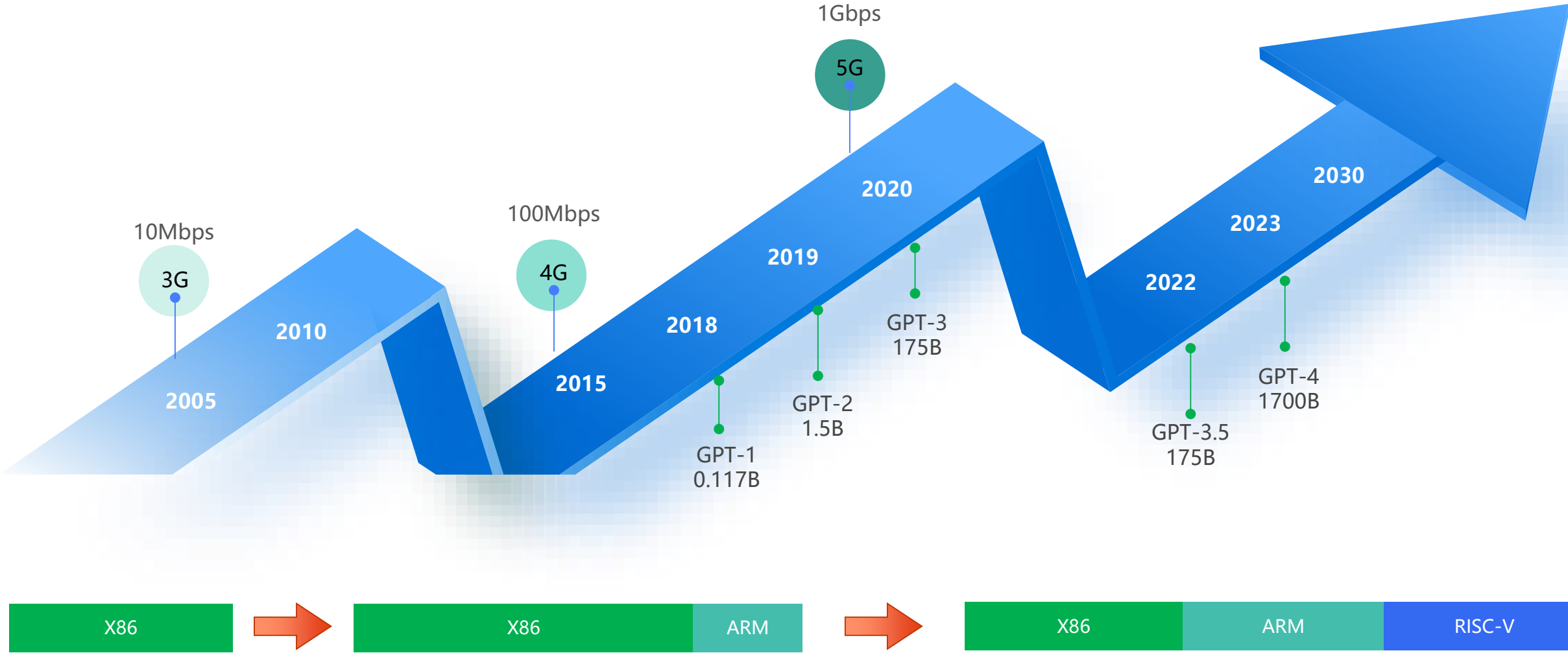
- 打造 “MetaStone” 自主品牌**计算/AI/存储系统级服务器、云原生POD**等产品
- 提供安全可控的**RISMA™跨架构计算集群**和**多元算力解决方案**，以及智算中心、工业、政务等行业解决方案
- 总部位于武汉，在北京、上海设有基地，分别为华中区、华北区、华东区区域总部



公司愿景

值得信赖的新一代RISC-V云原生服务器与解决方案提供商

RISC-V将在数据中心演化中的角色



第一步: 元石智算云原生跨架构集群





网络

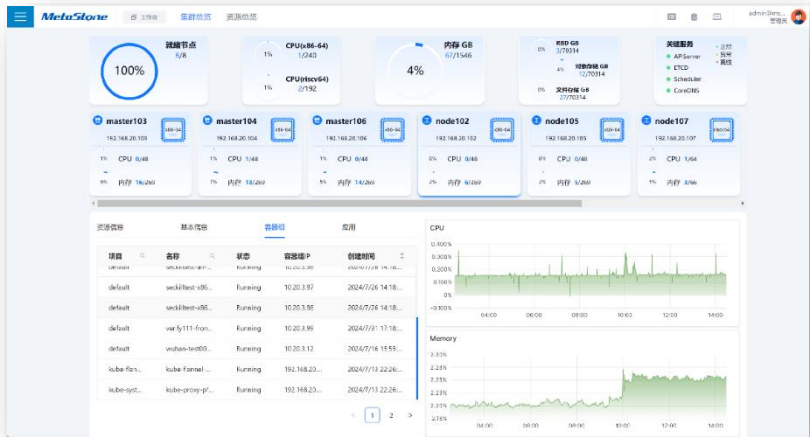
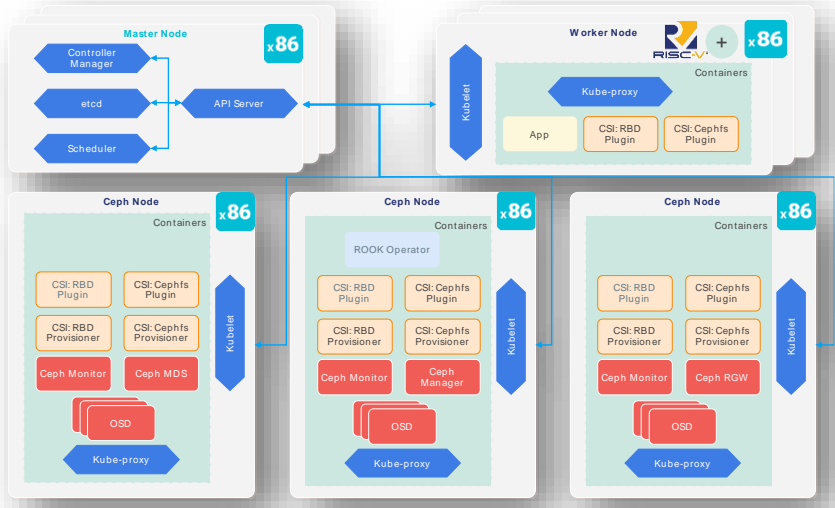
RISC-V

X86

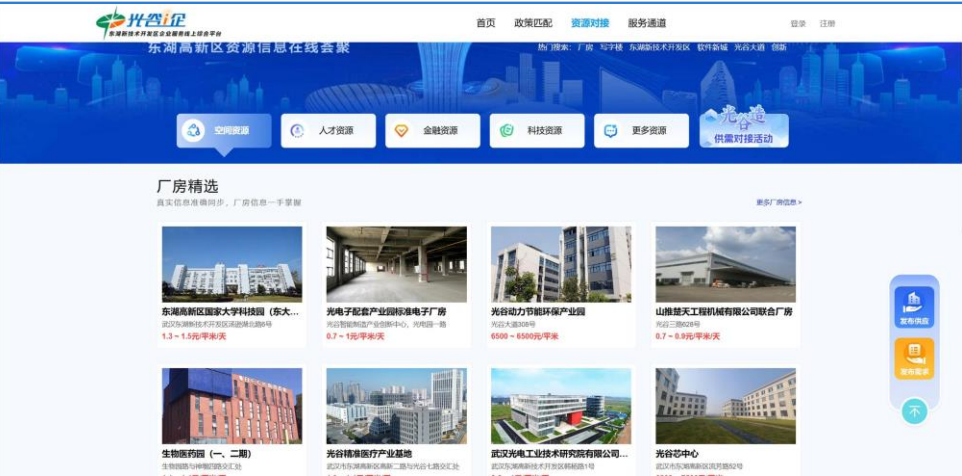
存算

云原生跨架构 POD Computer

- 硬件驱动优化
 - 完善构建工具链
 - 参与OpenEuler RISC-V SIG社区生态, 加固RISC-V服务器操作系统
- ↓
- Kubernetes组件RISC-V移植
 - 主流开源存储客户端移植
 - 多元算力支持
 - 自研开发跨架构云原生OS
 - 数据中心集群交付



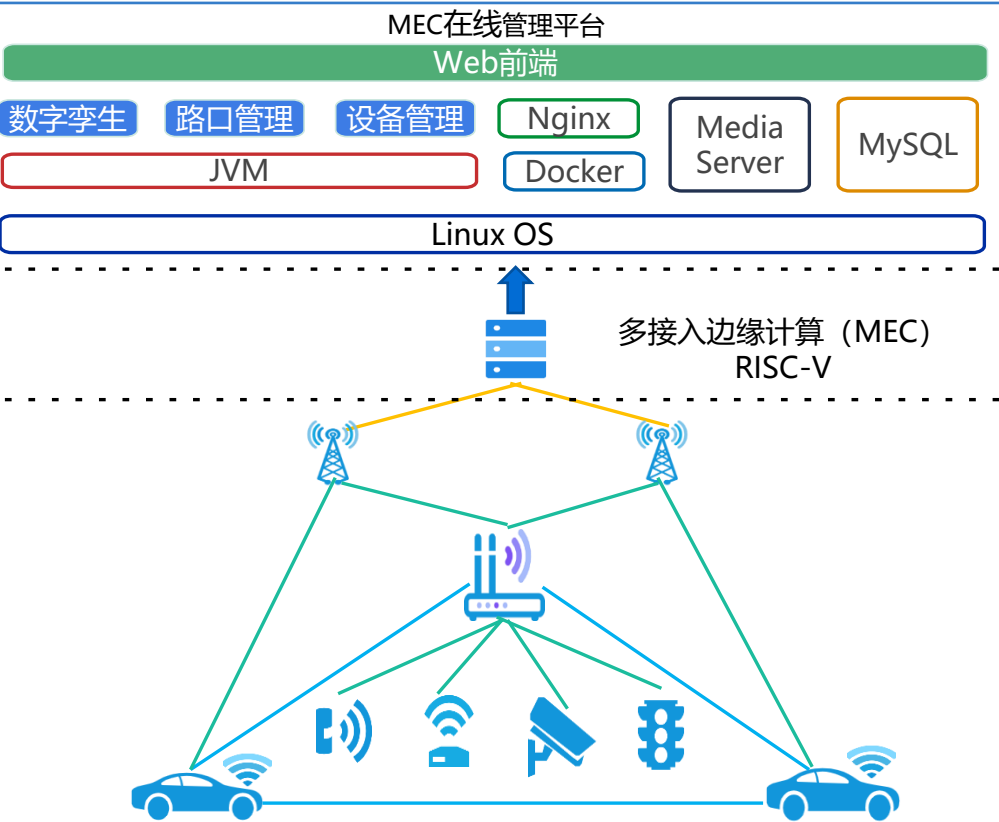
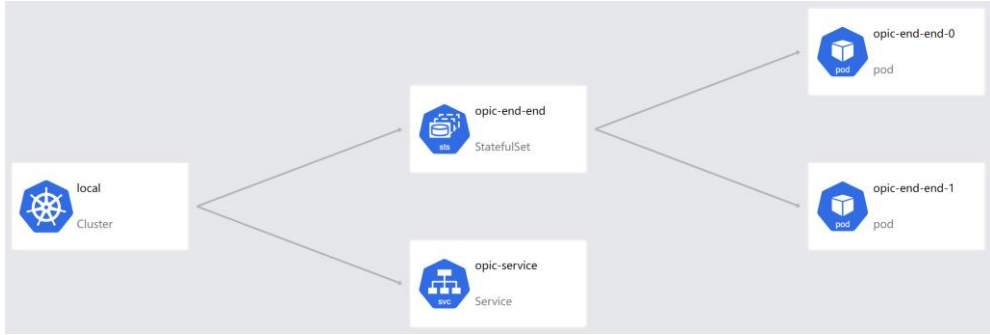
元石智算RISC-V跨架构集群应用案例



Nginx + HTML + JS
前端应用

JDK 11 + SpringBoot
后端应用

MySQL数据库



硬件

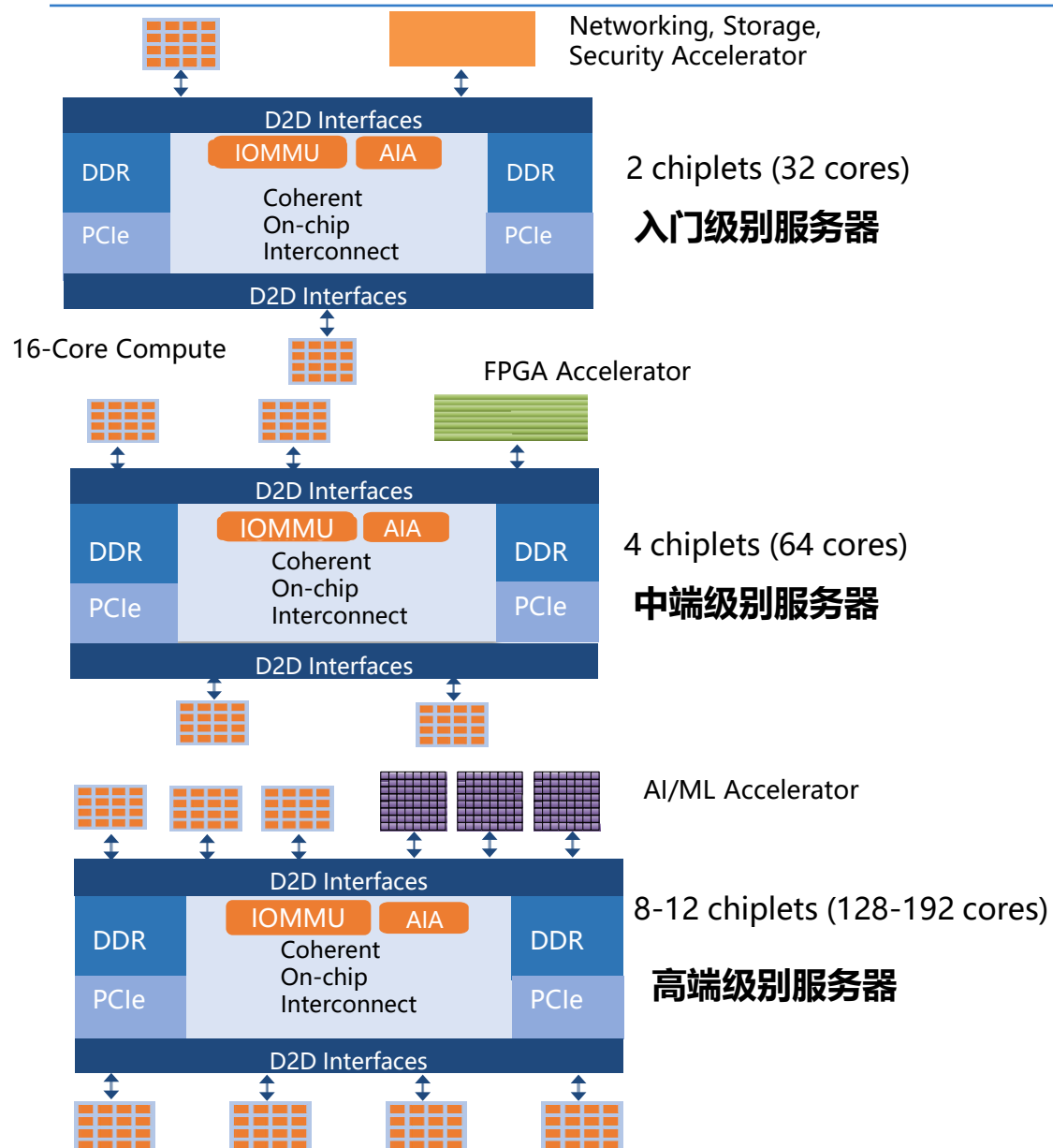
更强的CPU Chiplets
异构 DSA加速器
Vector
Hypervisor
RAS
Security

软件

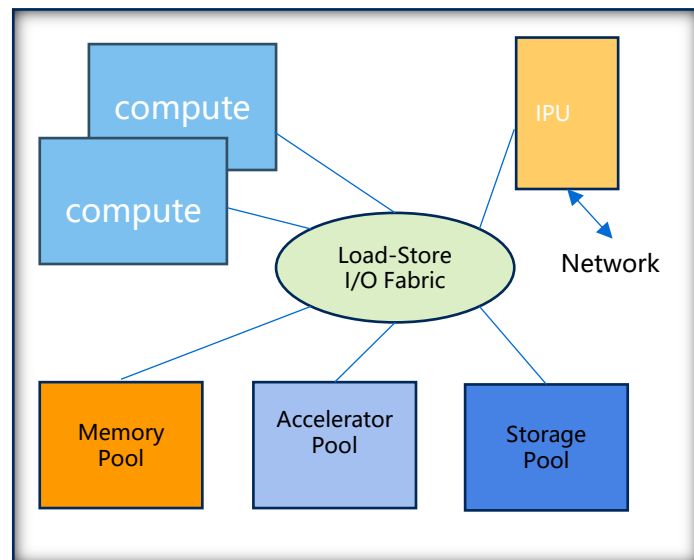
基于Vector的加速库
规模应用的检验
容器镜像

生态

行业的共同努力协作
标准制定



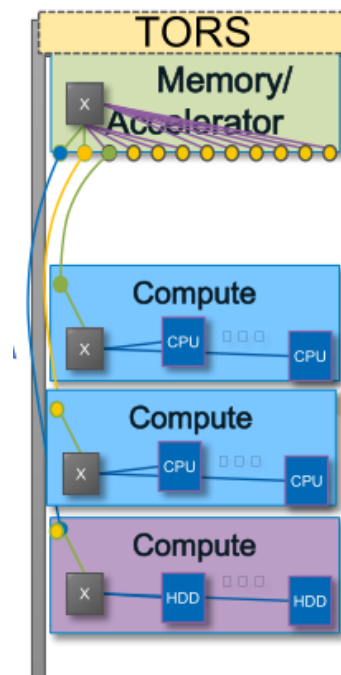
- Chiplet 构建可扩展的架构数据中心的服务器级别的计算
- 标准芯粒可以组合成所需的解决方案，且交付周期短
- 可以在所需的 CPU、内存和加速器配置之间切换
- 可选择标准 IO Hub 或带有集成功能的定制 IO Hub
- IO Hub 和定制加速器可以在 N-1/N-2 工艺节点上设计，以提高成本效益。
- 芯片发布可以缩短为一年以内
- 研发费用可以相对与monolithic die 大幅降低



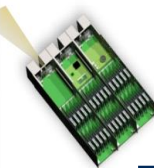
愿景: 将Load-Store I/O(CXL)
作为集群中的结构,提供低延迟和高带宽的资源池化/共享以及消息传递功能



机架集群

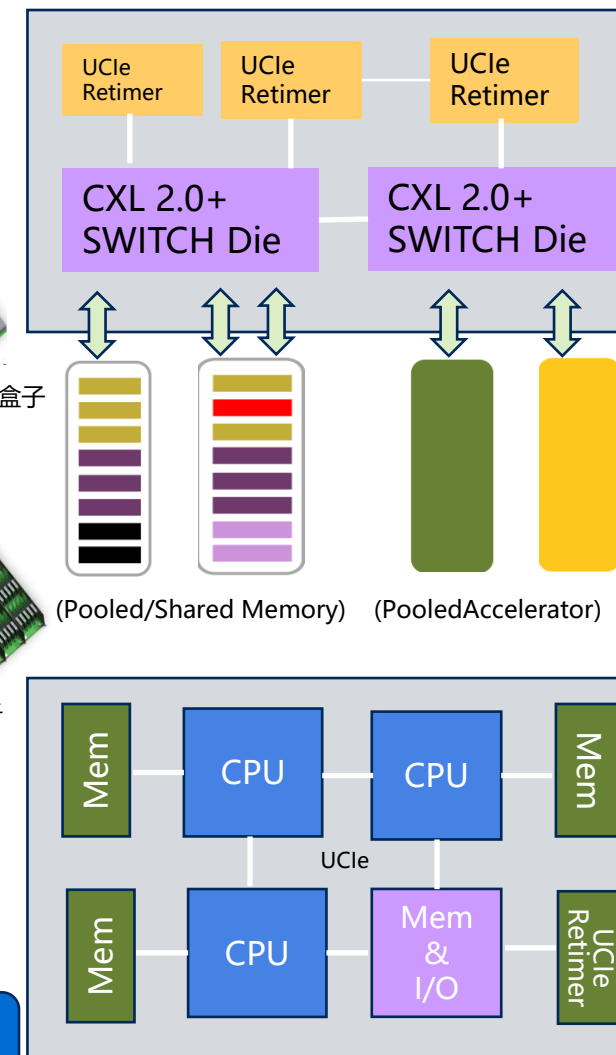


内存/加速盒子



计算盒子

光学连接: 机架和集群内部

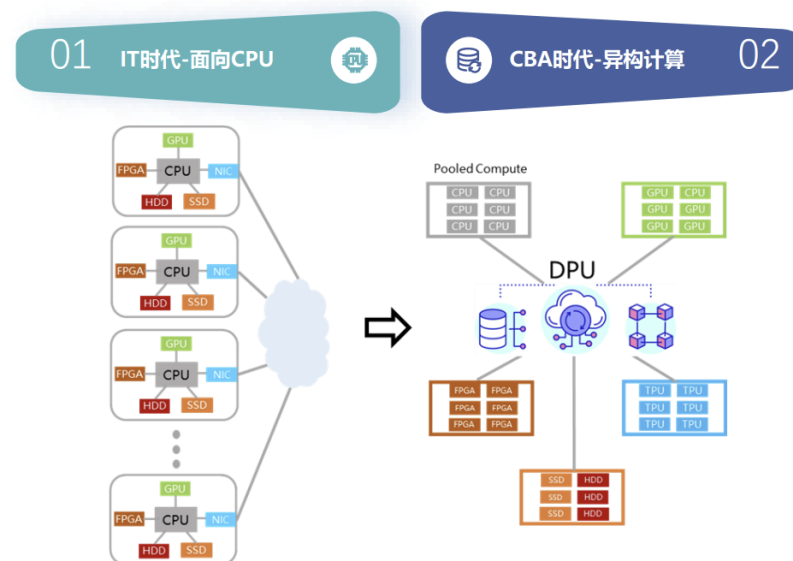


通过UCle Retimer (例如, 共封装光学器件) 使用长距离介质 (电气/光学等) 连接CXL机架/集群级别。物理连接通过基于 UCle Retimer的共封装光学器件实现 (例如光学、电缆、毫米波)的封装外扩展选项

使用UCle Retimer实现封装外连接, 实现机架/集群级别的资源池化共享

- 不同于之前的数据中心, AI数据中心包含4种不同网络, 其中加速器网络尚处于发展早期, 不同厂家有不同的选择, 从而蕴藏着标准化的机会
 - 带外网络 (BMC管理网络)
 - 前端网络 (普通以太网, 用于SLURM/Kubernetes、网络存储、数据加载、模型检查点等)
 - 后端网络 (InfiniBand/RoCE以太网, 用于扩展数百到数千个机架之间的GPU-GPU通信)
 - 加速器网络 (如Nvidia的NVLink、AMD的Infinity Fabric/UALink、Google的ICI等)
- 基于DPU (Data Processing Unit) 的智能网卡是CBA (Cloud Computing, Big Data, AI) 时代数据中心基础架构变革的必要条件

网络	带宽 (Blackwell NVL72)
以太网	18 * 400Gpbs
InfiniBand	72 * 800Gbps
NVLink	72*1.8TBps = 130TBps



展望未来: AI对数据中心的影响—存储

- 当前AI大模型训练耗时长，中断频繁，从而对于数据加载和模型检查点存储提出了更高要求

在规定时间内完成检查点 Checkpoint 的聚合写带宽(GiB/sec)

Model Size (B Parameters)	Checkpoint Size(GiB)	% of Training Time			
		1%	2%	5%	10%
1	13.0	0.18	0.09	0.04	0.02
5	65.2	0.91	0.45	0.18	0.09
10	130.4	1.81	0.91	0.36	0.18
20	260.8	3.62	1.81	0.72	0.36
40	521.5	7.24	3.62	1.45	0.72
175	2281.7	31.6 9	15.85	6.34	3.71
530	6910.4	95.9 8	47.99	19.2	9.60

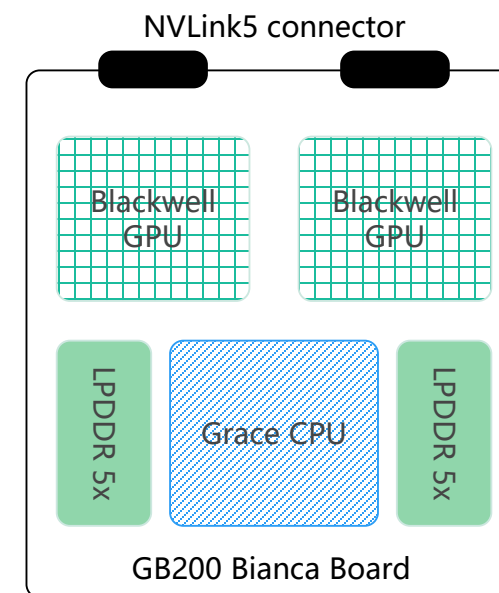
假定:

- 每两小时做一次Checkpoint
- 2 bytes每个模型参数 (BF16)
- 12 bytes每个模型参数用于优化或其他状态

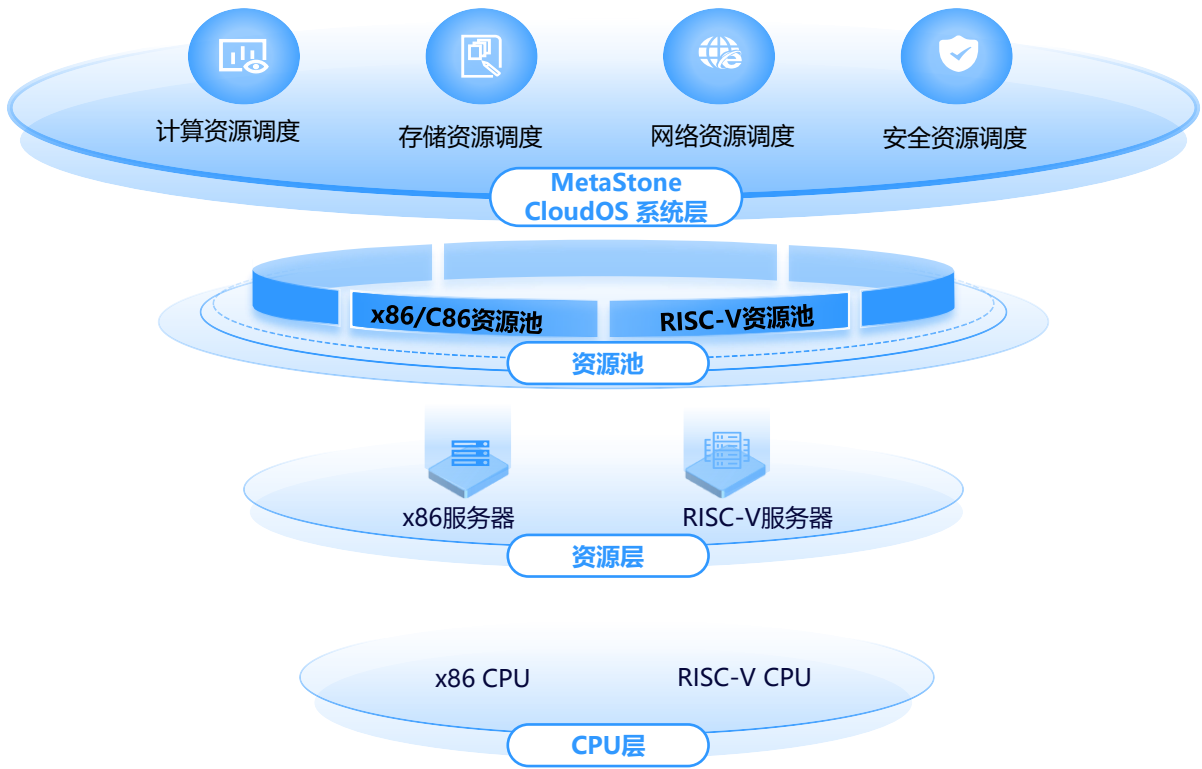


- 针对智算中心对存储的需求，元石智算的存储系统提供热层和温层的分层存储，用户通过元石智算MOD并行文件系统，访问高速存储池时，可以看到所有热层和温层的数据，根据用户的使用情况，平滑地流动。
- 同时，温层的数据可以对外提供块、对象、文件的统一存储并支持多租户访问，数据保护快照，WORM等企业级feature，以满足智算中心对存储的各类需求。

- AI时代各种加速器不断涌现，CPU作为计算机系统​​中的核心地位（尤其是大模型训练场景中）被打破，需要思考匹配加速器的处理器而非匹配处理器的加速器
 - GB200 NVL72的CPU与GPU的比率是1:2
 - HGX H100的CPU与GPU的比率是2:8
 - GH200的CPU与GPU的比率是1:1，但GH200的出货量显著低于HGX H100
- 多元芯片纳管，GPU资源池化，提供对异构算力控制平面和数据平面能力，用户单一界面管理整个数据中心异构算力资源



元石智算 RISMA™ 跨架构计算集群方案



元石智算多元算力集群方案



Thanks

武汉元石智算科技有限公司
WUHAN METASTONE TECHNOLOGIES INC.

www.metastonecorp.com

MetaStone



元石智算

微信扫描二维码，关注我的公众号

