
ZOMI
AI 计算集群
挑战
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

Question? ？？？！！！！

1. 建设一个 AI 计算集群，软硬件栈分为哪几层？AIInfra 到底是怎么分层，硅基流动说自己做的是 AIInfra，芯片公司说自己做的也是 AIInfra。到底什么是 AIInfra？
2. 你觉得在 AIInfra 的这么多公司里面，有哪些公司能够脱颖而出？是寒武纪，是华为，还是天数，最近的 PPU 阿里平头哥等都出来了，还有谁？！

Content

AI 计算集群：

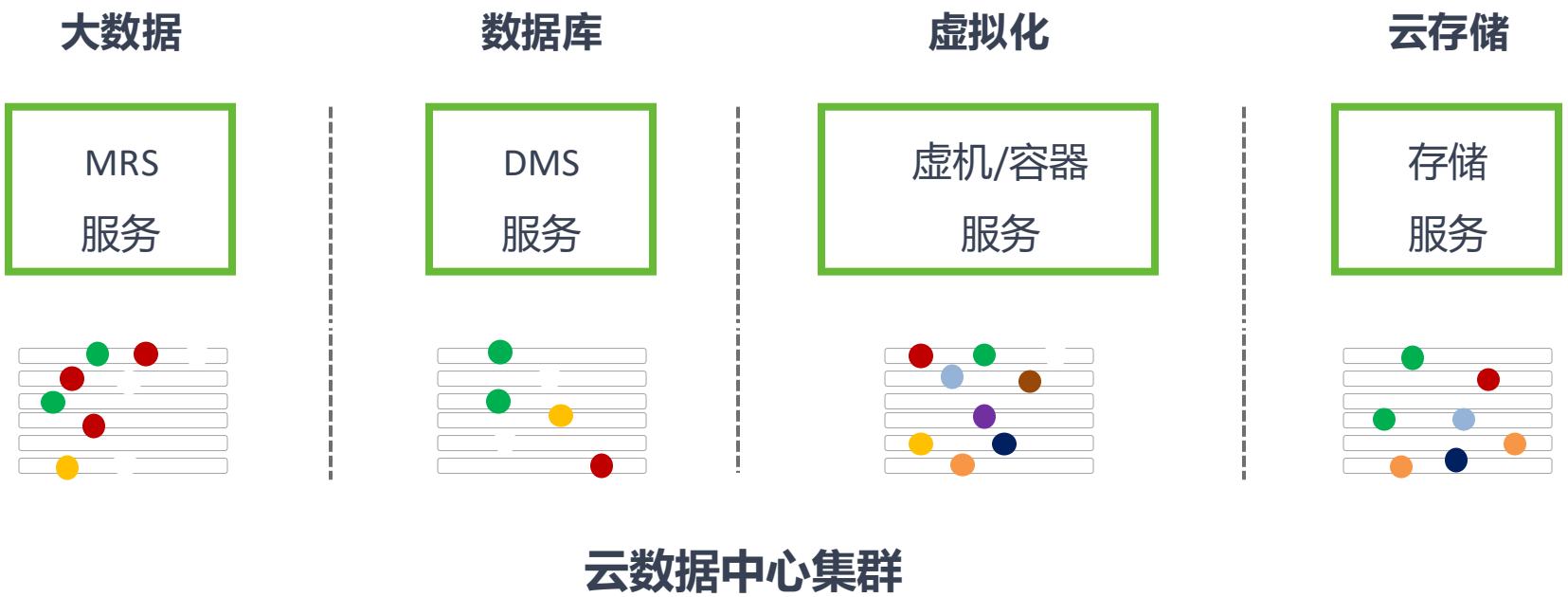
1. 算力挑战（大规模训练、并发负载、集群组网）
2. AI 集群架构（AIInfra 整体解决方案、软件平台与应用）
3. 建设目标（有效算力、可用度）

01

AI计算集群挑战

通算业务负载

- 负载多样、分散、任务间关联少；多数负载在单个服务器内闭环
- N个负载运行在M个服务器上，服务器之间松耦合
- 负载类型多样化，多数负载在单服务器内闭环，难以做针对性集群设计



通算业务负载

- 负载多样、分散、任务间关联少；多数负载在单个服务器内闭环

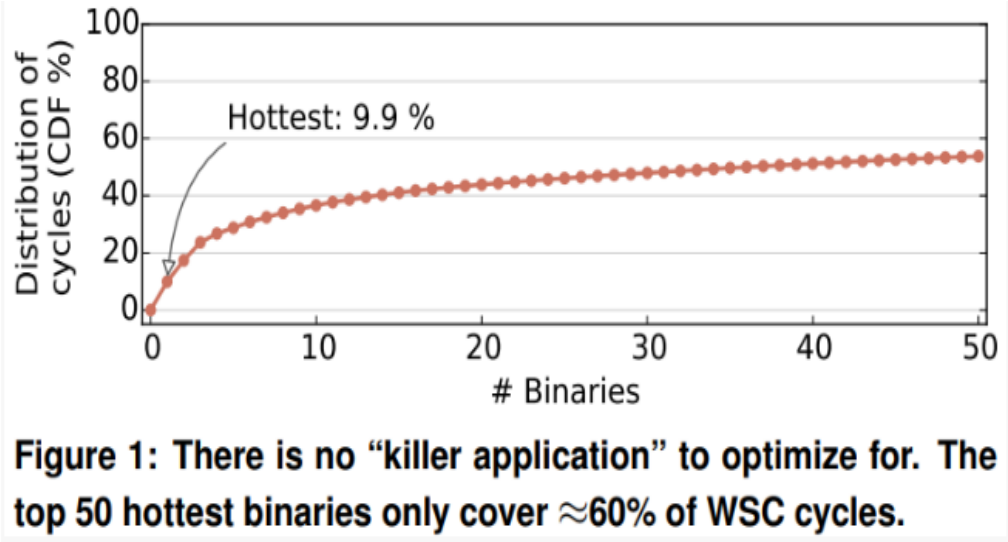


Figure 1: There is no "killer application" to optimize for. The top 50 hottest binaries only cover ≈60% of WSC cycles.

*Source: Profiling a warehouse-scale computer

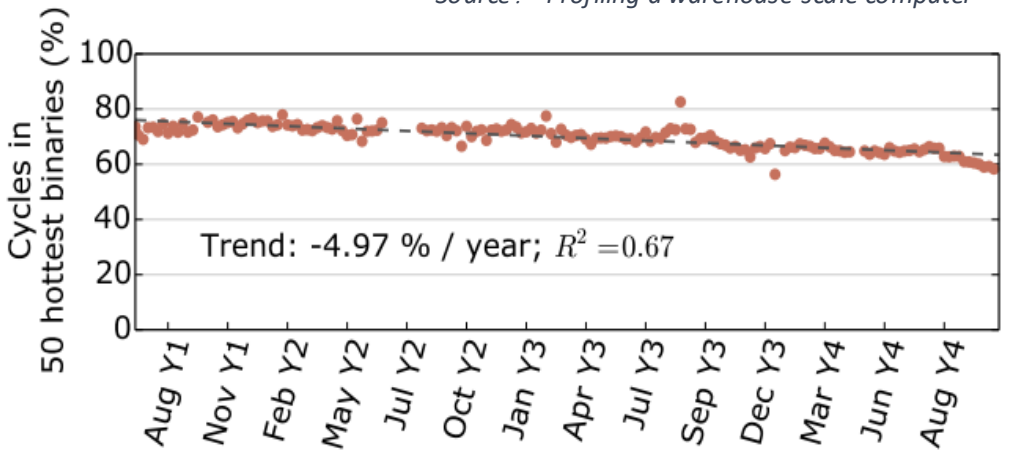


Figure 2: Workloads are getting more diverse. Fraction of cycles spent in top 50 hottest binaries is decreasing.

- 传统 DC 应用分散，TOP50 热点应用只占用了 60% 的 cycle，且 4 年还下降了 4.97%
- 只能针对特定应用进行软件优化，不足以针对某一类应用进行专门集群设计

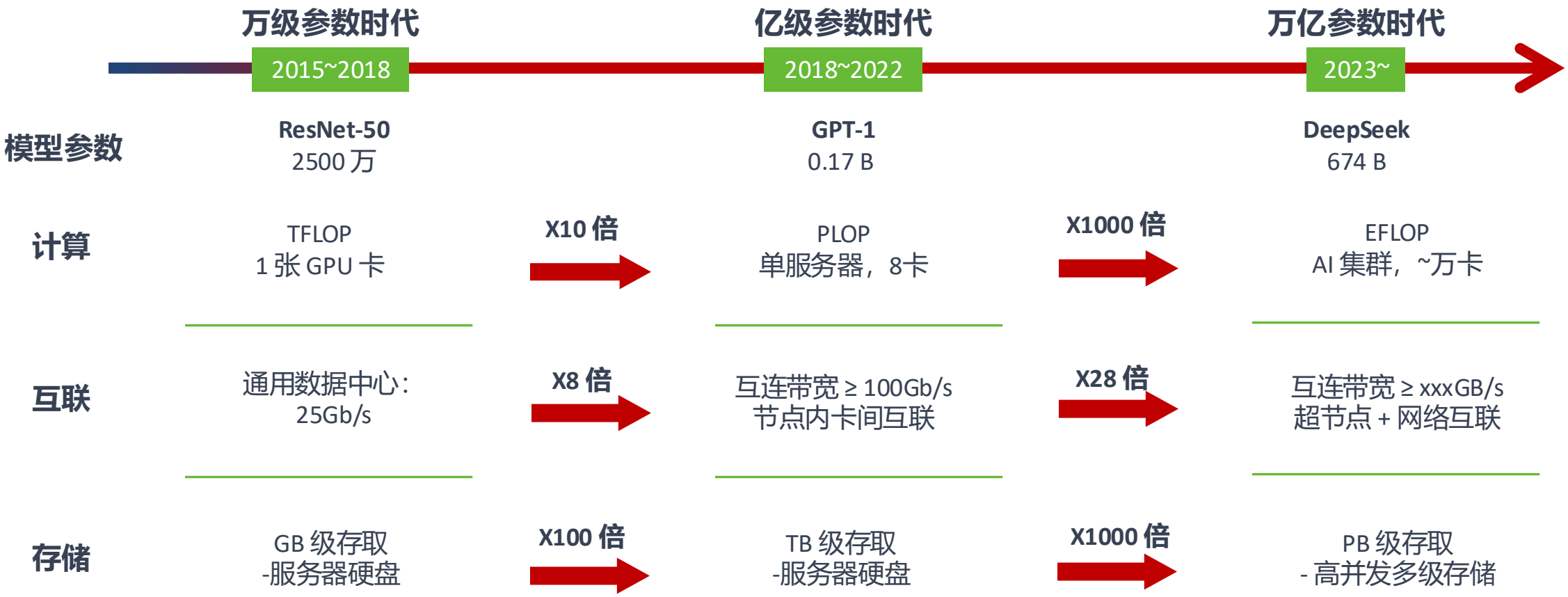
AI 集群负载特征

- AI 业务呈现出 高并行&网络化 的特征，集群成为大模型训练最佳算力平台

	万级参数时代		亿级参数时代		万亿参数时代
	2015~2018		2018~2022		2023~
模型参数	**ResNet-50** 2500 万		**GPT-1** 0.17 B		**DeepSeek** 674 B
计算	TFLOP 1 张 GPU 卡	**X10 倍** →	PLOP 单服务器，8卡	**X1000 倍** →	EFLOP AI 集群，~万卡

AI 集群负载特征

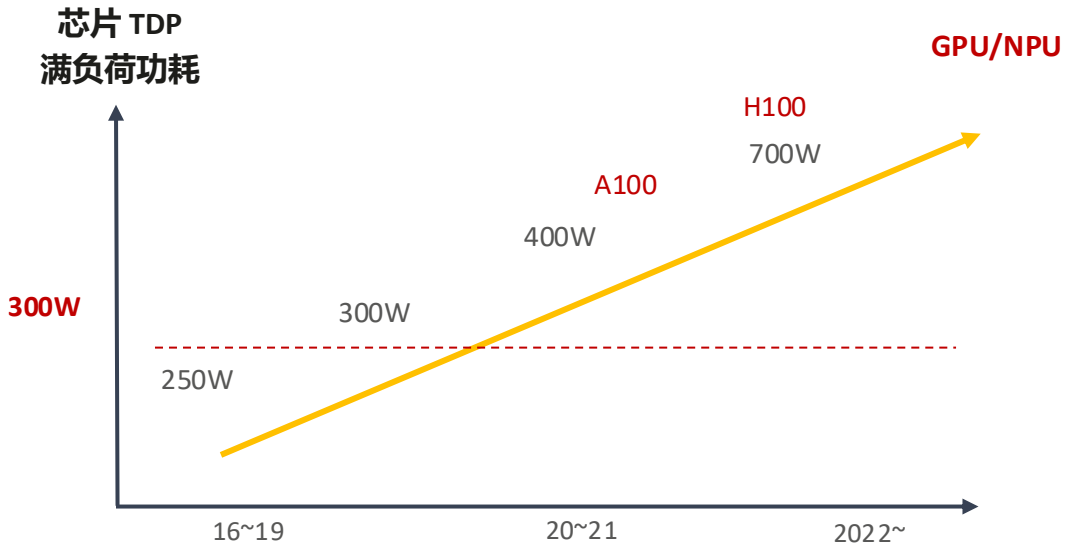
- AI 业务呈现出 高并行&网络化 的特征，集群成为大模型训练最佳算力平台



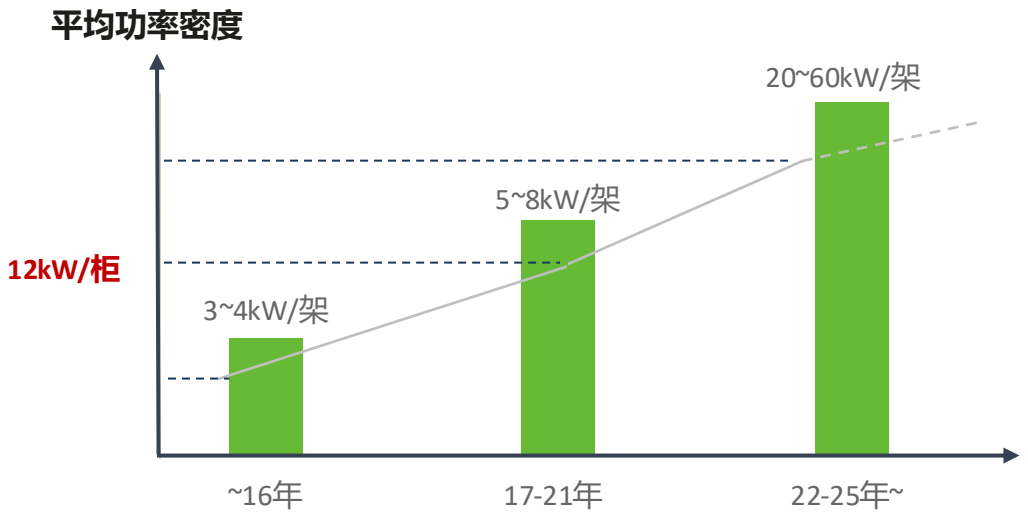
建设领先 AI 集群需要考虑的 4 大关键要素

1. 基础设施的先进性：

- AI 计算中心基础设施不同于传统机房，走向以密集 NPU/GPU 算力为中心进行规划设计
- 需考虑单位功率密度提升、规模液冷技术的应用和网络工程部署



芯片TDP功耗的上升，使液冷技术应用成为必然



智算进一步拉高功耗密度需求，单机柜超过50kW

建设领先 AI 集群需要考虑的 4 大关键要素

2. 超大规模集群互联技术：

- 大模型从千亿稠密走向万亿稀疏，从单模态走向多模态
- 模型结构从 Transformer Attention 到 MOE、RWKV、JEPA、MAMBA、Linear 等
- 带来对于大规模组网、负载均衡、多级存储，及未来超节点的新技术需求

PCIE

NVLINK to NV Fusion

RDMA base IB/RoCE

建设领先 AI 集群需要考虑的 4 大关键要素

3. 极致算力使用效率：

- AI 集群需要通过软硬协同、算存网云全栈协同来综合提升算力使用效率

- **单机执行效率（有效算力）：** 伴随模型规模增大，单机执行内存和 I/O 瓶颈、计算交互损耗、资源调度不均等问题，影响 MFU
- **集群并行效率（线性度）：** 集群规模增大，各类并行通信额外开销，影响算力利用率的保持度，即线性度，进而影响训练效率和成本
- **单 Step 训练时长（Token/s）：** = 计算时长 + 通信时长 + 存储 IO 时长 - 可隐藏通信时长 - 可隐藏 IO 时长。1）提升带宽降低通信时长；2）提升存储性能降低 IO 时长；3）计算+网络协同，隐藏通信；4）计算 + 存储协同，隐藏 IO 时长。

建设领先 AI 集群需要考虑的 4 大关键要素

4. 集群高可用 & 易运维：

- 考虑跨产品故障定界定位、训练任务及时恢复等多个难题；
- 单任务数万卡并行，任一部件/节点故障将导致训练中断；
- 需要系统级可靠性设计，实现故障实时感知、智能定界定位、快速恢复；

千万器件全机运行
故障频次高，管理难

软硬件故障
模式复杂、种类多

万卡级超复杂应用
问题定界定位复杂

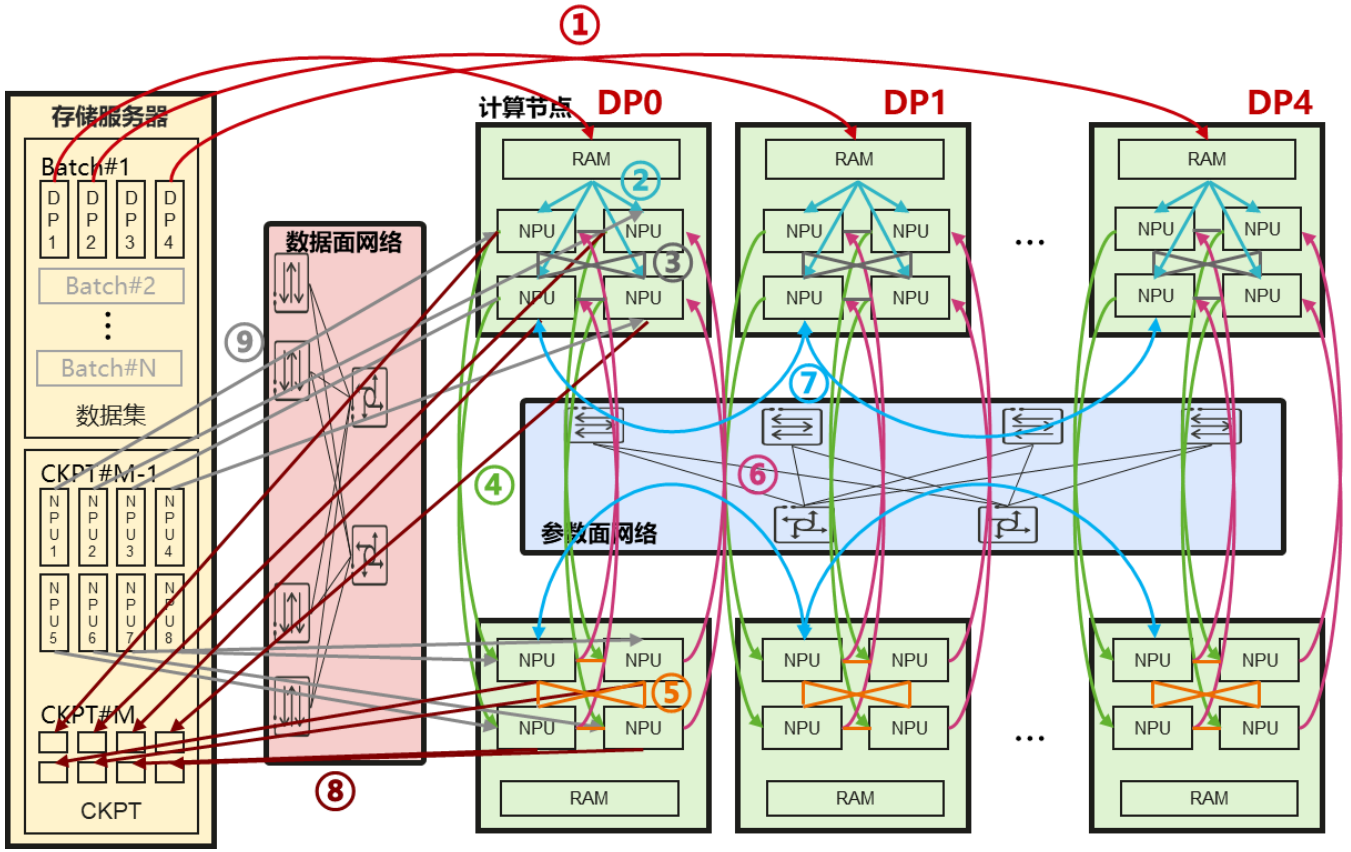
典型硬件故障定位x天
应用类复杂故障定位时间达xx天

万亿参数大作业
断点续训恢复缓慢

万卡集群长稳仅x级
平均恢复时长xx级

智算业务负载：多维并行

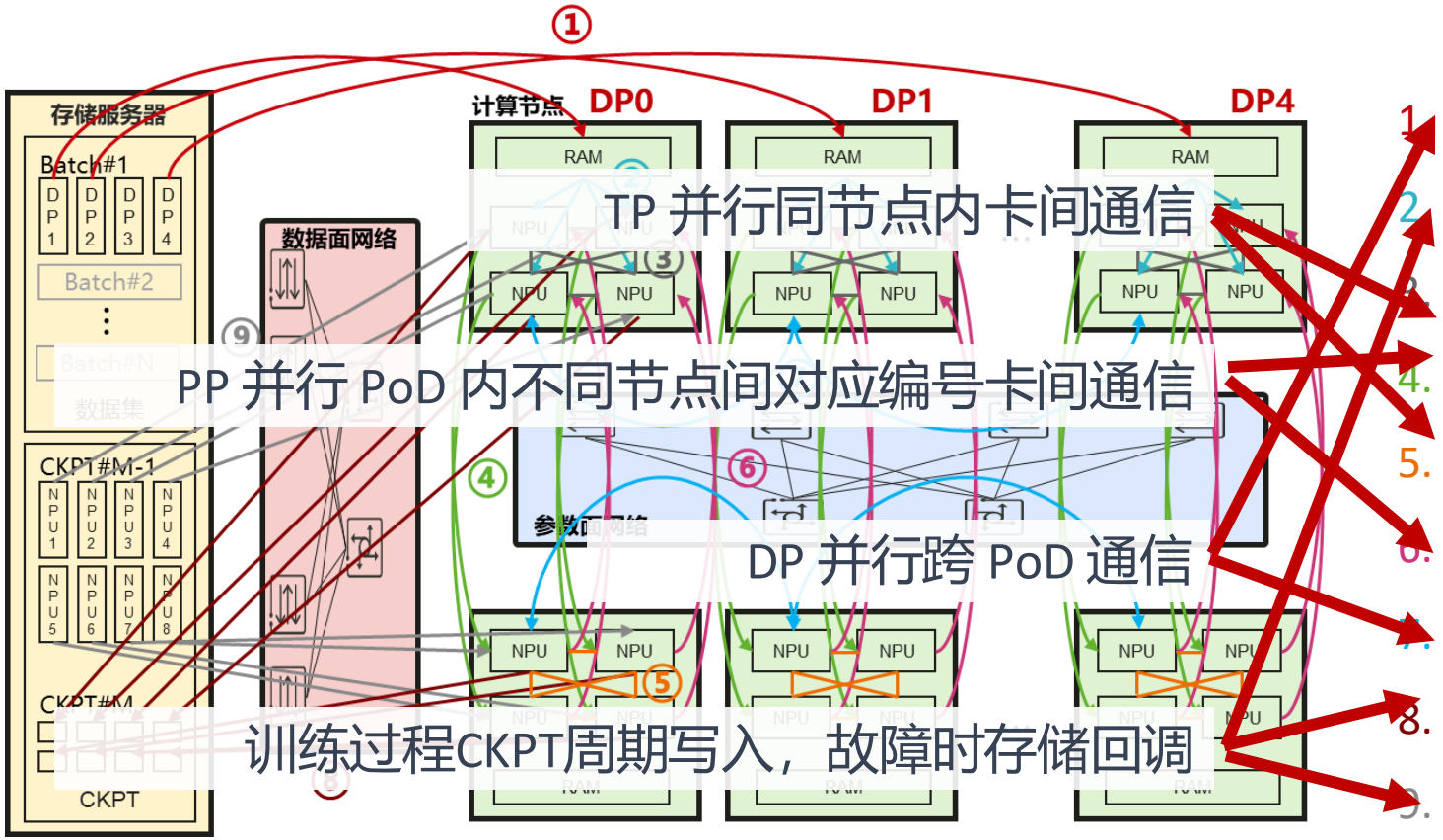
- 1 个 AI 业务负载运行在全部服务器上，计算 + 通信 + 存储 IO 紧耦合



1. 当前 Batch 根据 DP 切分并传输到对应节点
2. 数据加载到 NPU 上运算
3. 前向传播 TP 通信
4. 前向传播 PP 通信
5. 反向传播 TP 通信
6. 反向传播 PP 通信
7. DP 通信，同步梯度数据
8. [周期性] 保存 CKPT
9. [故障时] 恢复加载 CKPT

智算业务负载：多维并行

- 1 个 AI 业务负载运行在全部服务器上，计算 + 通信 + 存储 IO 紧耦合

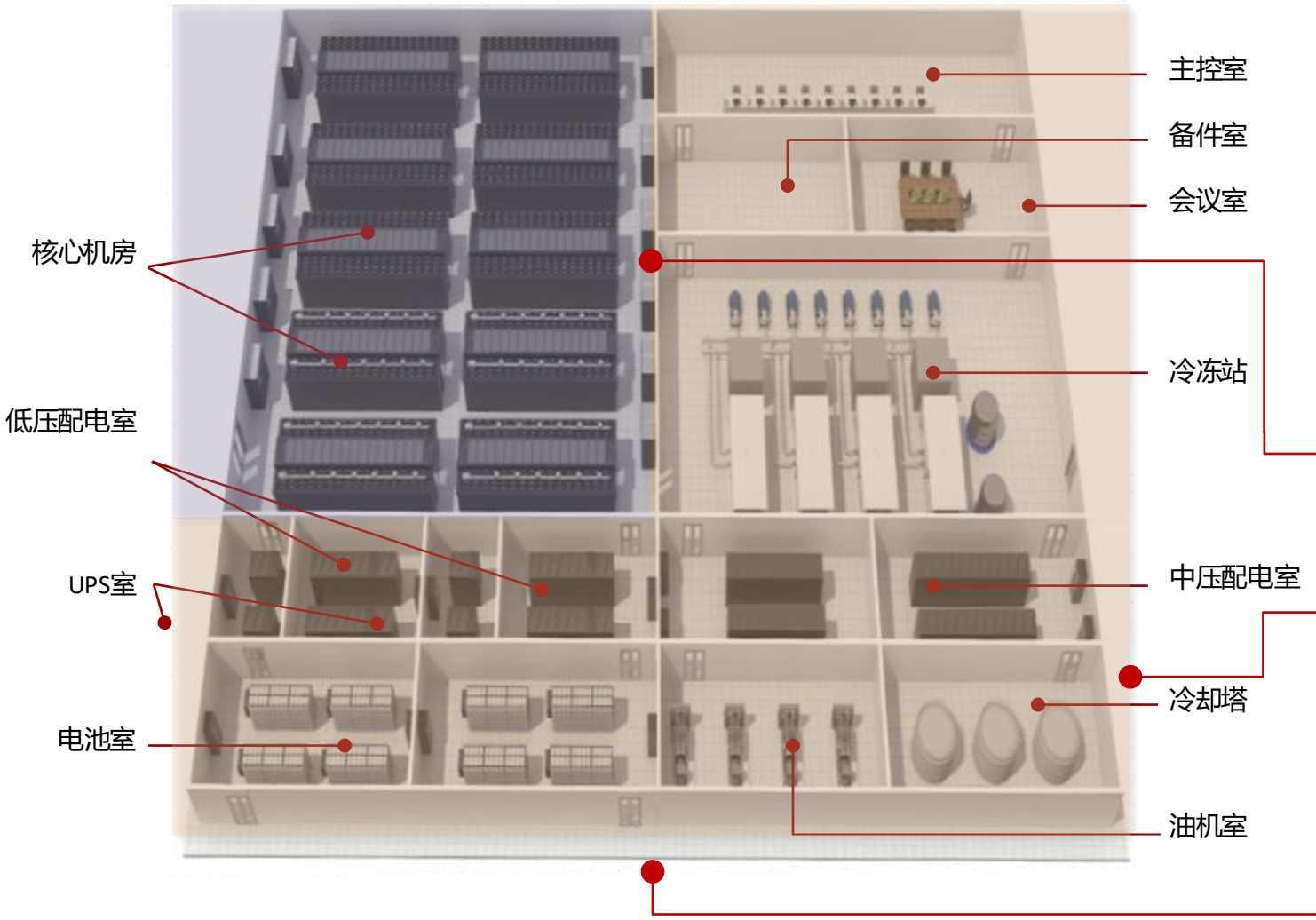


1. 当前 Batch 根据 DP 切分并传输到对应节点
2. 数据加载到 NPU 上运算
3. 前向传播 TP 通信
4. 前向传播 PP 通信
5. 反向传播 TP 通信
6. 反向传播 PP 通信
7. DP 通信，同步梯度数据
8. [周期性] 保存 CKPT
9. [故障时] 恢复加载 CKPT

02

AIInfra/计算集群整体架构

AI 集群硬件解决方案



集群系统架构分层

L4 应用与服务
数据业务及应用（大数据、互联网、HPC、AI等）

L3 算力使能平台
多租户、虚拟化、分布式并行计算、运维&运营等

L2 算力底座
服务器、存储系统、网络系统及组网等

L1 物理基础设施
供电/备电、制冷、布线、机柜、安防等

L0 基建楼宇系统
机房、配套楼宇土建、风火水电等

AI 集群系统软件平台与应用

集群系统架构分层

L4 应用与服务

数据业务及应用（大数据、互联网、HPC、AI等）

L3 算力使能平台

多租户、虚拟化、分布式并行计算、运维&运营等

L2 算力底座

服务器、存储系统、网络系统及组网等

L1 物理基础设施

供电/备电、制冷、布线、机柜、安防等

L0 基建楼宇系统

机房、配套楼宇土建、风火水电等

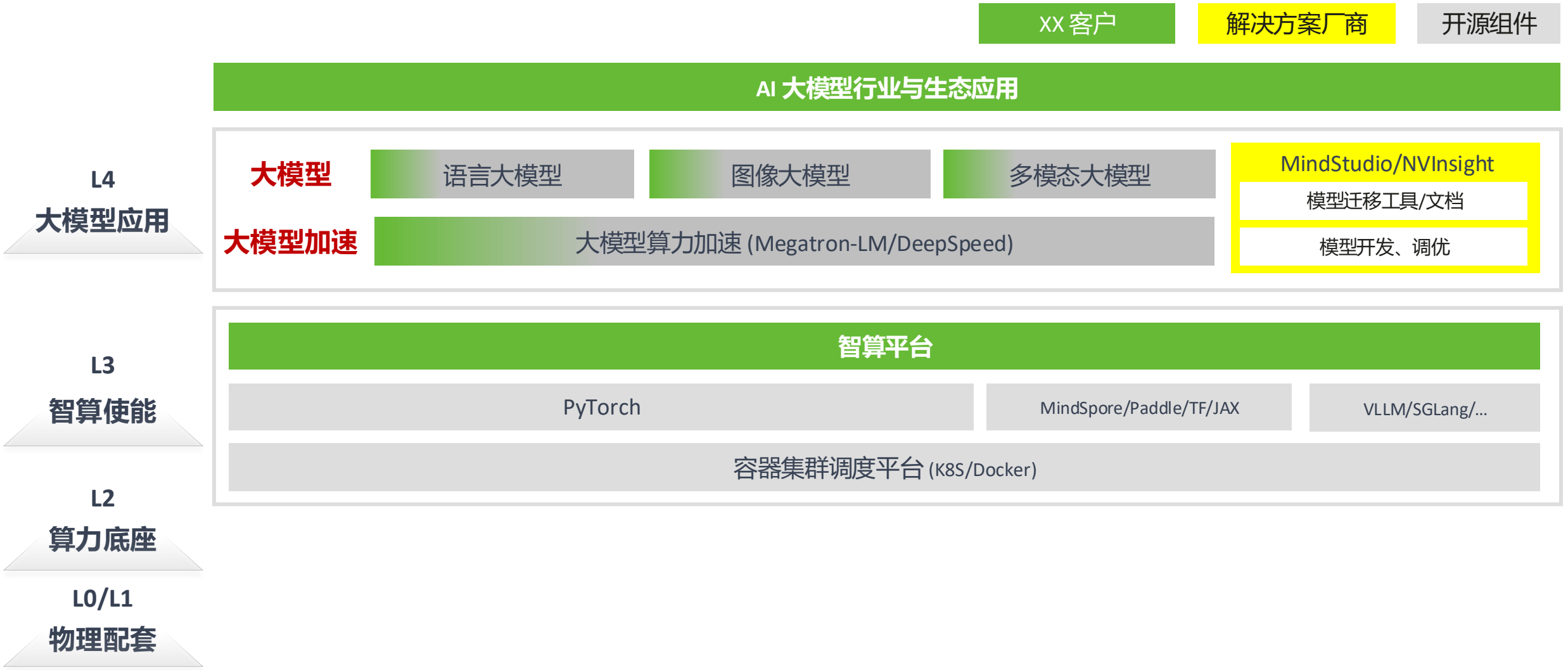
AI 集群架构重点

- 超大规模高速互联，提升有效算力，AI负载更亲和

L3 算力使能平台 调度平台 | PaaS | 智能运维	公有云多租算力存力运力可分可合 算力池化支撑 **AI 高效训练**
	大模型训练对计算 & 通信架构强亲和 故障管理支撑**集群高可用**
L2 算力底座 计算 | 网络 | 存储	算网/算存/存网协同、软硬协同、全栈优化 **长稳算力底座，达成极致算力效率 MFU**
L0/L1 物理配套 能效 | 供电 | 承重	高密、高功、液冷配套 = **低PUE，高能效**

AI 集群硬件解决方案参考架构

软件平台与应用参考架构



参考架构

XX 客户 | 解决方案厂商 | 开源组件

AI 大模型行业与生态应用

L4 大模型应用

大模型：语言大模型 | 图像大模型 | 多模态大模型

大模型加速：大模型算力加速 (Megatron-LM/DeepSpeed)

MindStudio/NVInsight
模型迁移工具/文档
模型开发、调优

L3 智算使能

智算平台

PyTorch | MindSpore/Paddle/TF/JAX | VLLM/SGLang/...

容器集群调度平台 (K8S/Docker)

L2 算力底座

云管服务

计算管控服务 | 网络管控服务 | 存储管控服务

异构计算架构 (NVIDIA CUDA、 Ascend CANN)

操作系统 OS (Ubuntu、 Linux)

计算 AI计算

网络 参数面 | 业务面 | 数据面

存储

运维服务

运维管理服务

集群运维系统

NCE 网络系统

L0/L1 物理配套

液冷方舱 (液冷/风冷柜、CDU、列间空调等) | **液冷机柜** (Super PoD) | **智能配电小母线** | **末端空调**

基础机房配套设施 (冷源、高低压配电、发电)

03

AI计算集群目标

AI 集群设计目标

- **设计目标：围绕集群规模 × 计算效率，实现持续提升有效算力**

$$\text{AI 集群有效算力} = (\text{单卡算力} \times NPU\text{卡数}) \times \text{计算效率} \approx \frac{N \times \text{大模型参数量} \times \text{有效样本量}}{\text{训练时长}}$$

$$\text{计算效率} = \text{单机执行最优 \& 集群并行最优 \& 中断时间最短}$$

AI 集群设计目标

- **计算效率** = 单机执行最优 & 集群并行最优 & 中断时间最短

模型算力利用率（MFU）	◦ **算的更快**：提升算子、算法效率 ◦ **算的更省**：降低调度和计算开销 ◦ **运的高效**：提升参数交换带宽和网络利用率
万卡集群线性度	◦ **高效并行**：架构亲和的模型切分策略 ◦ **高效通信**：1）增大参数面通信带宽和覆盖范围；2）降低冲突，持续提升网络利用率
故障恢复时间（MTTR）	◦ **少出故障**：硬件高可靠、训前压测检查 ◦ **快速恢复**：断点续训，快速定界+快换快修

总结与思考

面向十年百倍增长，未来计算集群 6 大技术特征

数据中心 算力和能效 实现十年百倍增长

全球AI算力规模

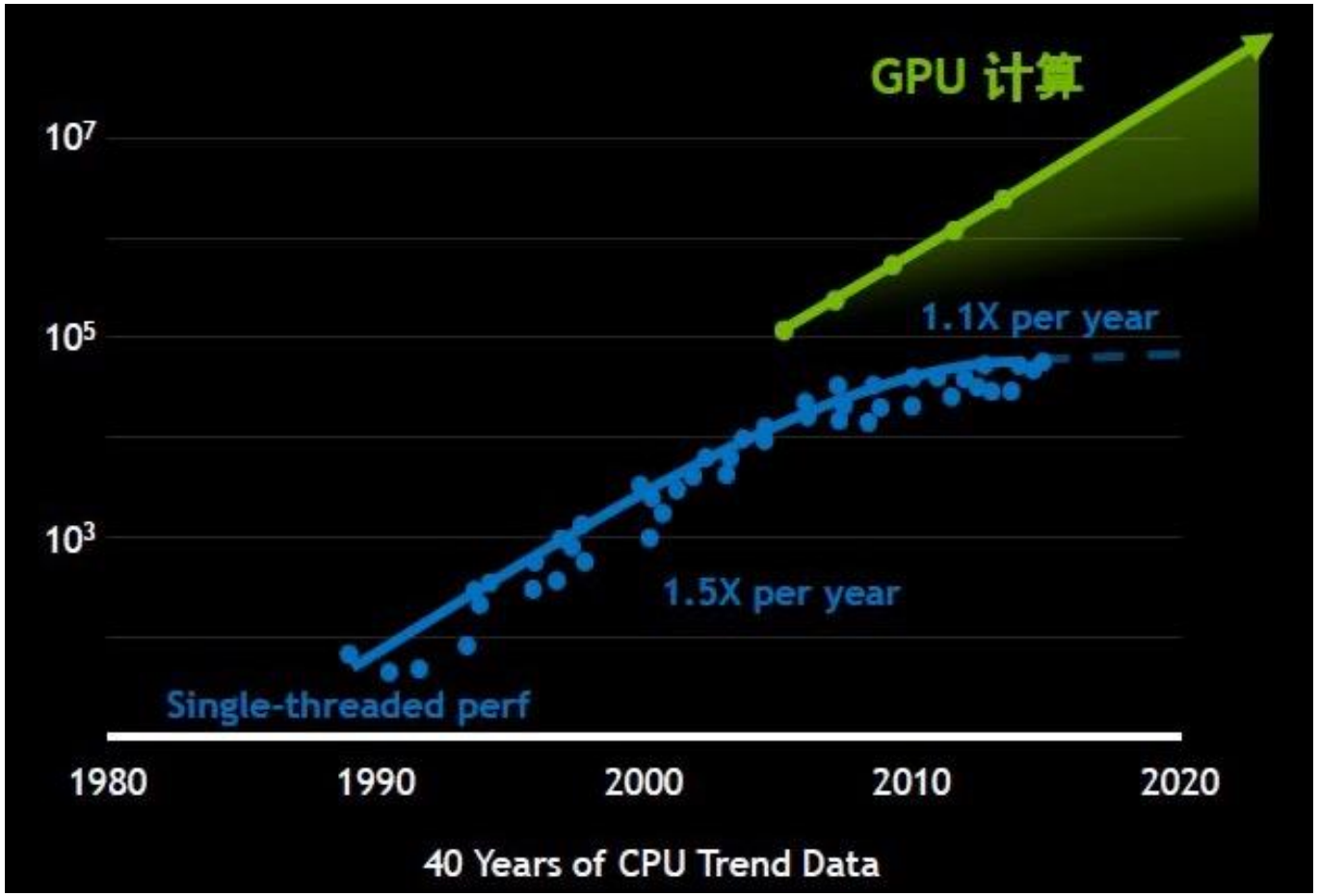
105 ZFlops（FP16），十年增长500倍

算力能效

350TFlops（FP32）/KW，十年增长百倍

集群网络互连速率

3.2Tbps，增长10倍



面向十年百倍增长，未来计算集群 6 大技术特征

柔性资源

全池化；柔计算；泛协作；

多样泛在

大集群；新型态；融算力；

负载亲和

大小芯；新算力；新存储；

计算集群 2030

安全智慧

高安全；高可靠；高智能

对等互联

超融合；高性能；光内生

零碳节能

绿供电；新储能；液制冷

Summary

- 大模型训练负载呈现出 高并行 &网络化 的特征，集群成为最佳算力平台
- 集群建设关键要素：基础设施先进性、超大规模互连、极致算力效率、集群高可用&易运维
- 围绕集群规模、计算效率、长稳运行发力，打造应用亲和 AI 集群架构，最大化使能有效算力

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