

# 大模型系列之AI集群

# AI 集群服务器架构



ZOMI

# 大模型 + AI系统全栈架构



提供大模型算法

提升 AI 集群整体利用率

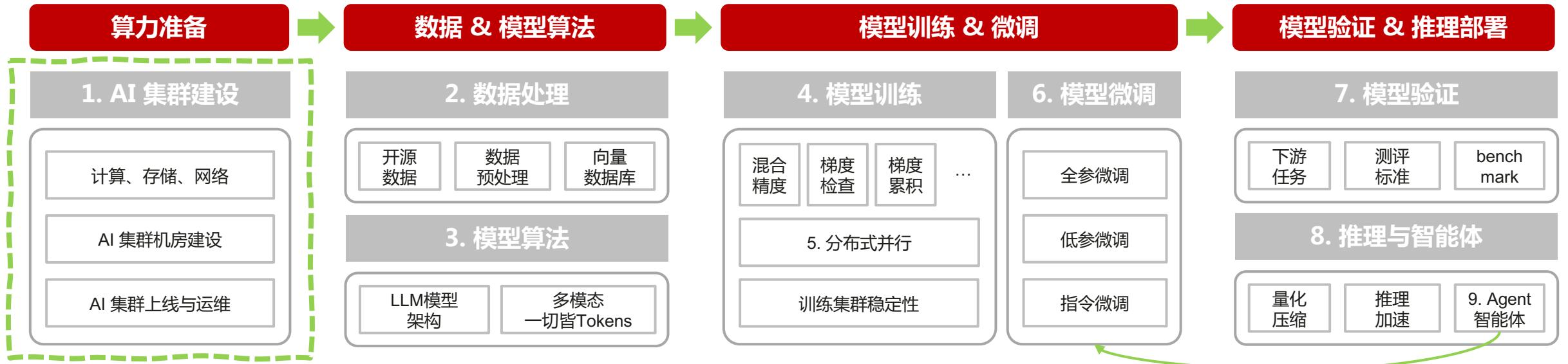
通过软硬件协同优化

基于硬件的编译、框架、使能层

硬件体系结构不仅需要算力

还包括网络、存储组成的超计节点

# 大模型业务全流程



大模型不仅需要 LLM 算法，同时需要提供  
AI 集群、海量数据、分布式并行、推理部署等 AI 系统全栈软硬件协同优化

# 关于本内容

## I. 内容背景

- **AI 集群 + 大模型** : AI集群服务器形态 – 集群训练的指标

## 2. 具体内容

- **AI 集群硬件架构** : AI集群组成 – AI集群硬件 – AI集群软件 – 分布式架构
- **AI 集群通信方式** : 通信硬件实现 - 集群组网 – 集群软件通信 - 通信实现方式
- **分布式通信原语** : 通信源语 - 点对点通信 – 集合通信
- **分布式存储系统** : 大模型权重存储方式 – 多级存储系统
- **AI 集群回顾** : NVIDIA 与 TPU 超级计算节点POD

# Question?

I. AI 集群规模越大越好？大集群拥有大算力？



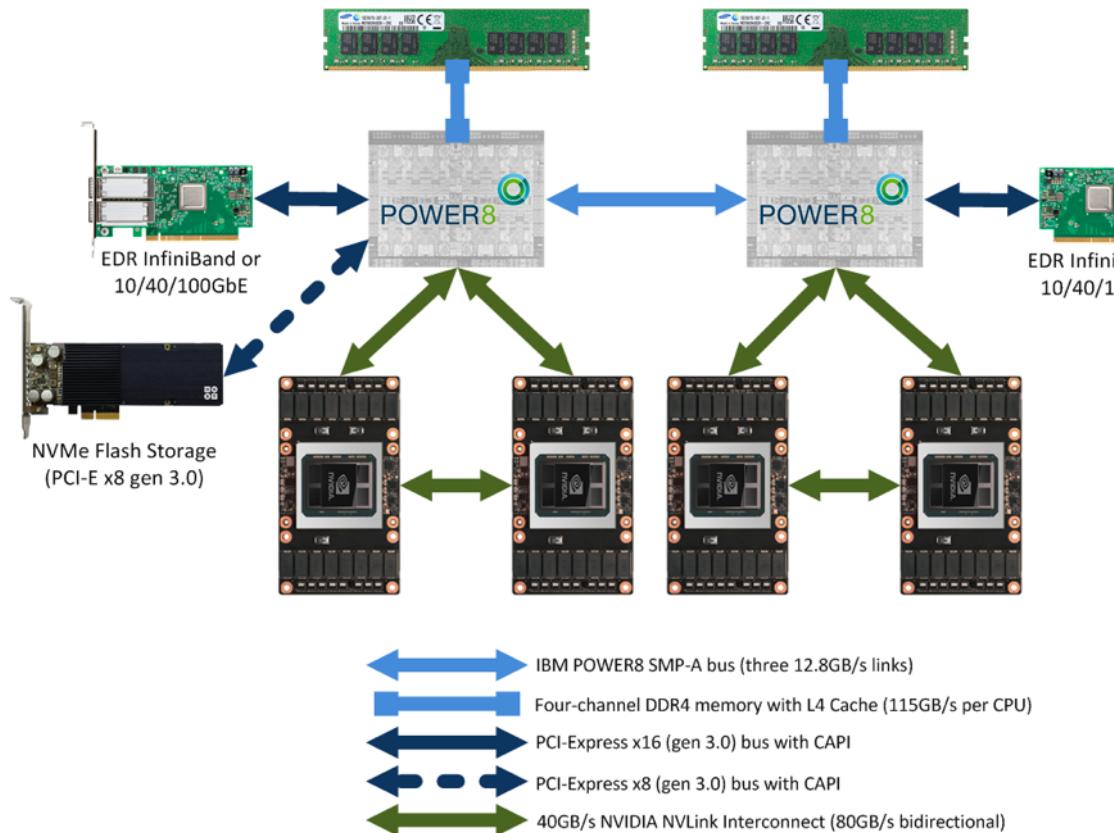
# 3. DGX SuperPOD

看AI集群软件

# 计算节点通讯

## Server Block Diagram

Microway OpenPOWER Server with NVIDIA Tesla P100 NVLink GPUs



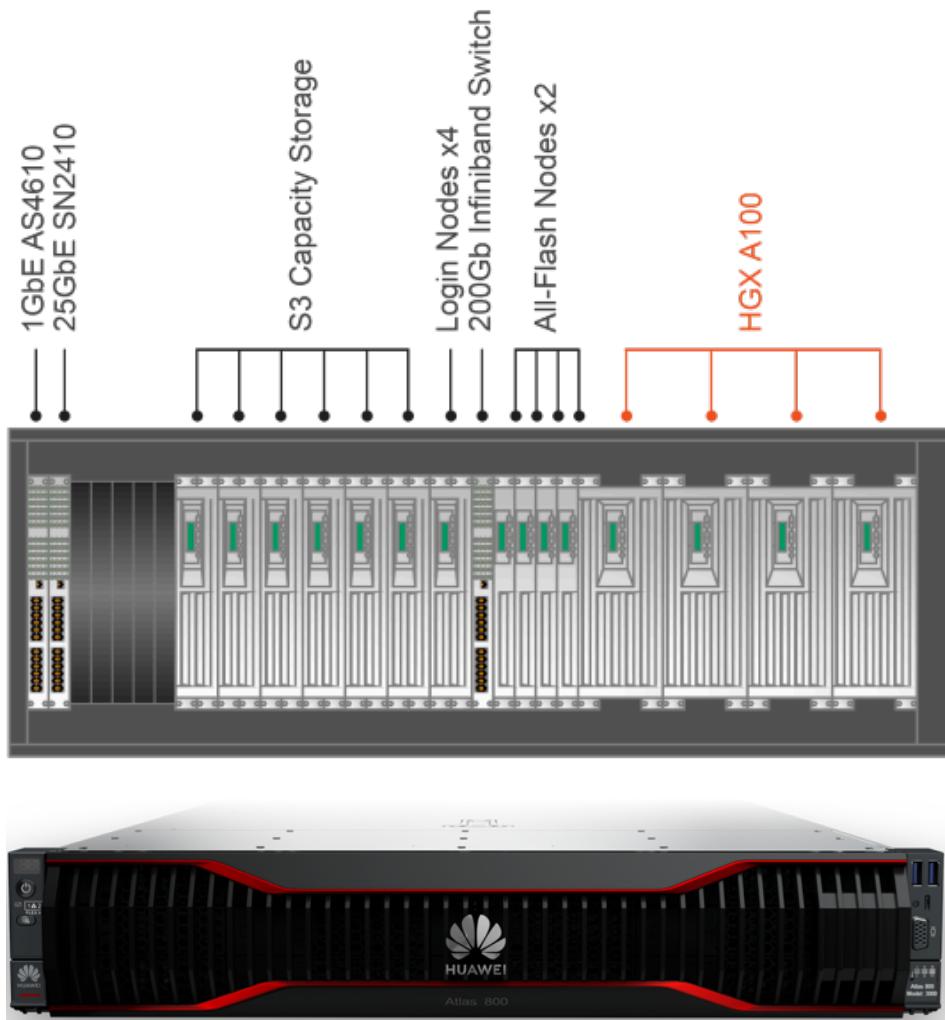
### 机器内通信

- 共享内存
- PCIe
- NVLink ( 直连模式 )

### 机器间通信

- TCP/IP 网络
- RDMA 网络 ( 直连模式 )

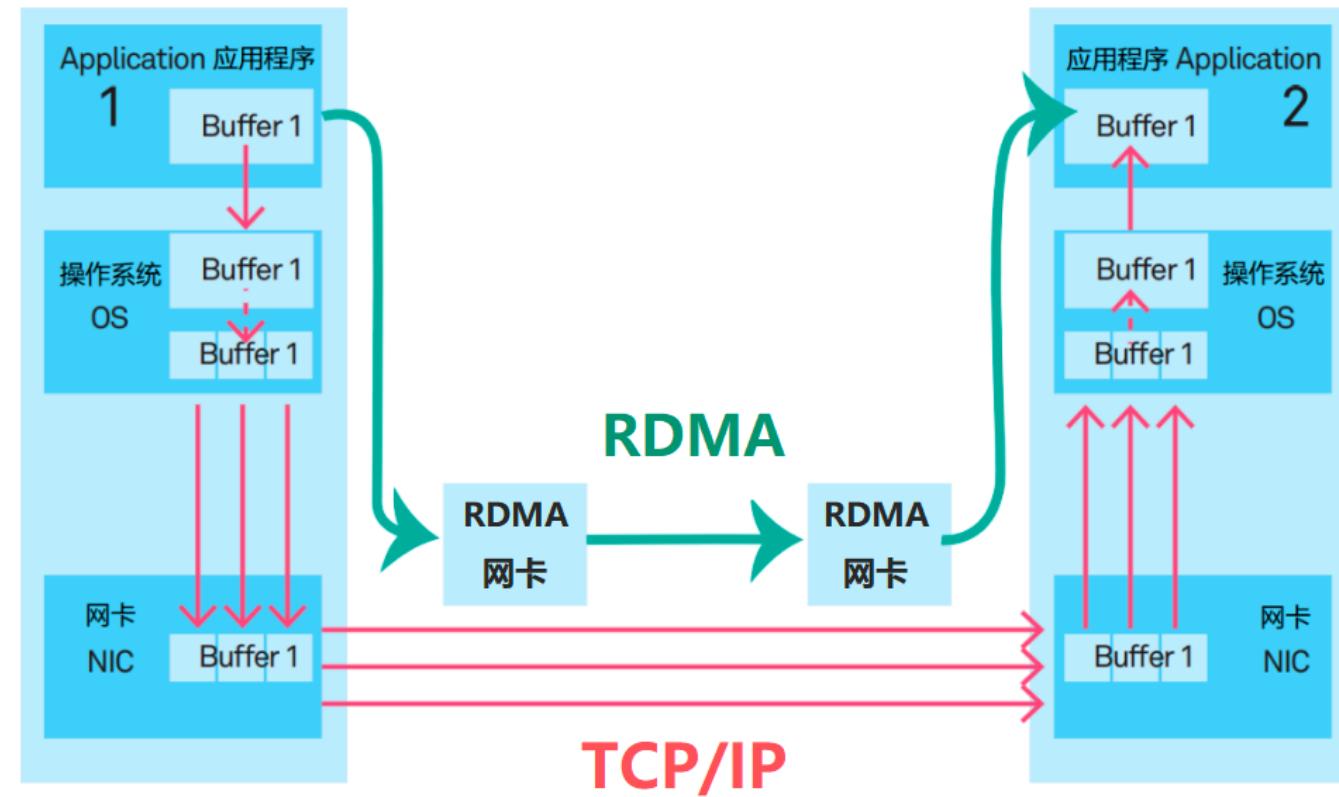
# 计算节点通讯



- **机器内通信**
  - 共享内存
  - PCIe
  - NVLink ( 直连模式 )
  
- **机器间通信**
  - TCP/IP 网络
  - RDMA 网络 ( 直连模式 )

# RoCE

- RoCE ( RDMA over Converged Ethernet , 基于聚合以太网的RDMA ) , 允许计算节点间直接通过内存进行数据传输 , 无需 OS 内核和 CPU 参与 , 大幅减小CPU负荷 , 降低延迟 , 提高吞吐量。

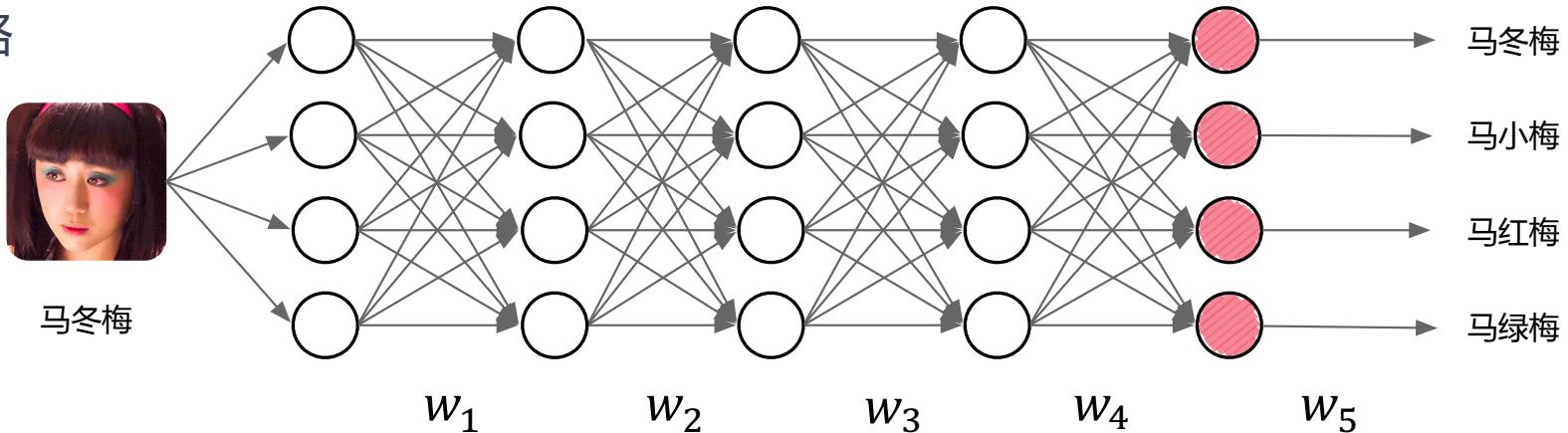


# 4. 分布式架构



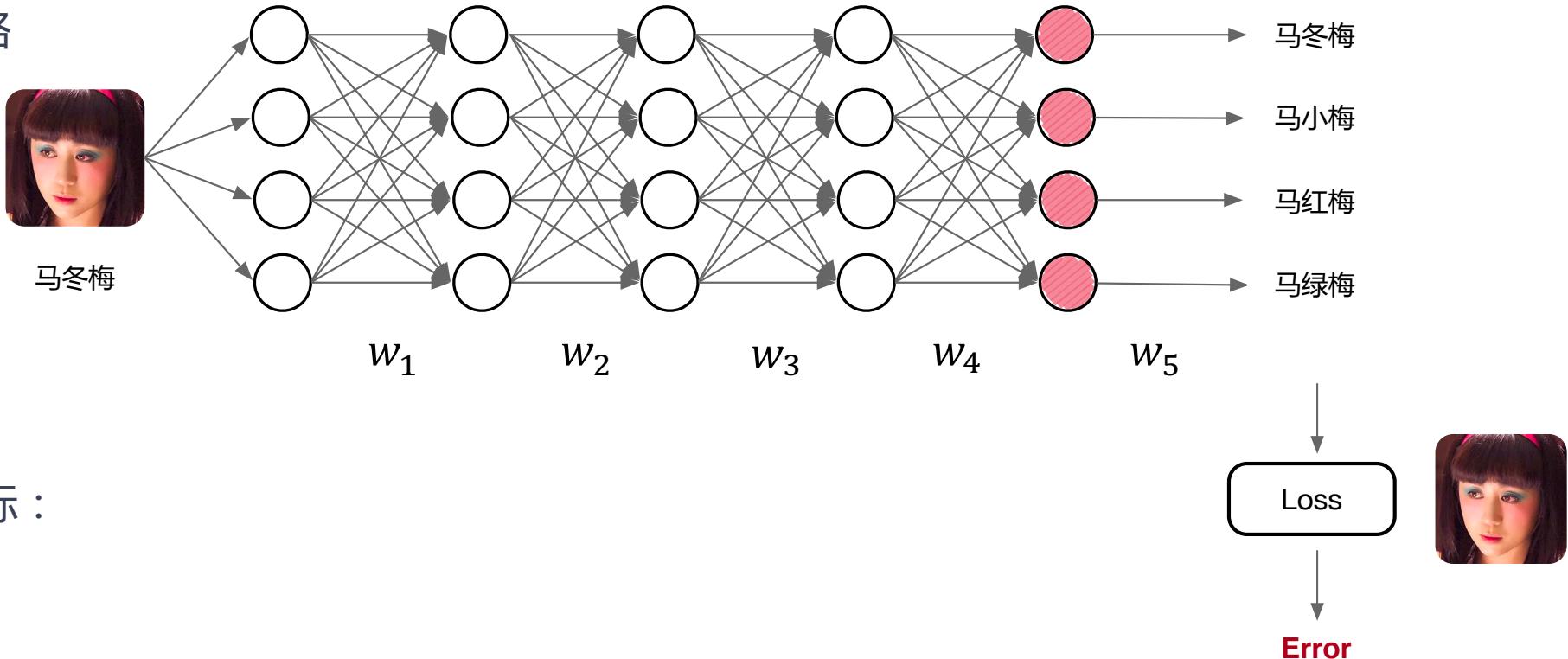
# 回顾：模型训练过程

1. 定义一个神经网络



# 回顾：模型训练过程

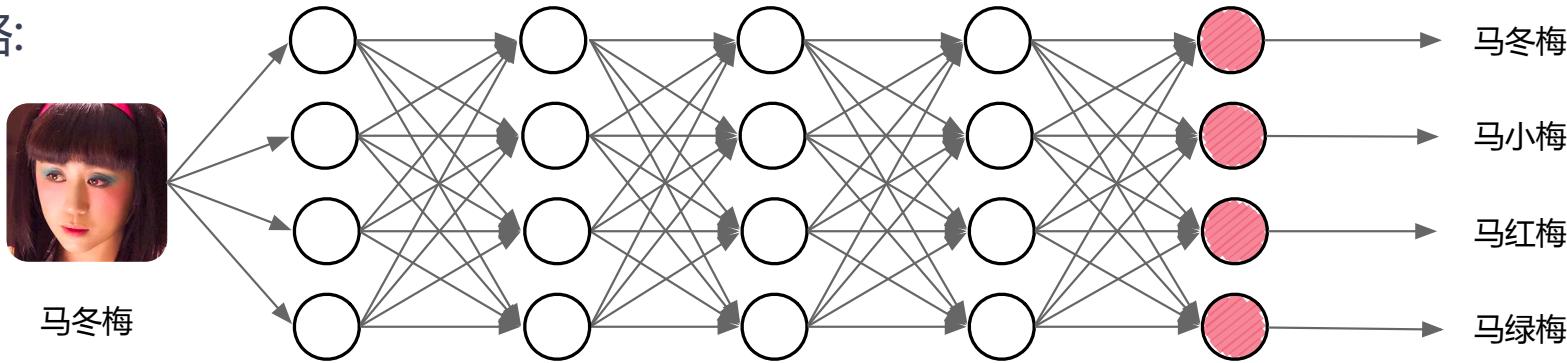
1. 定义一个神经网络



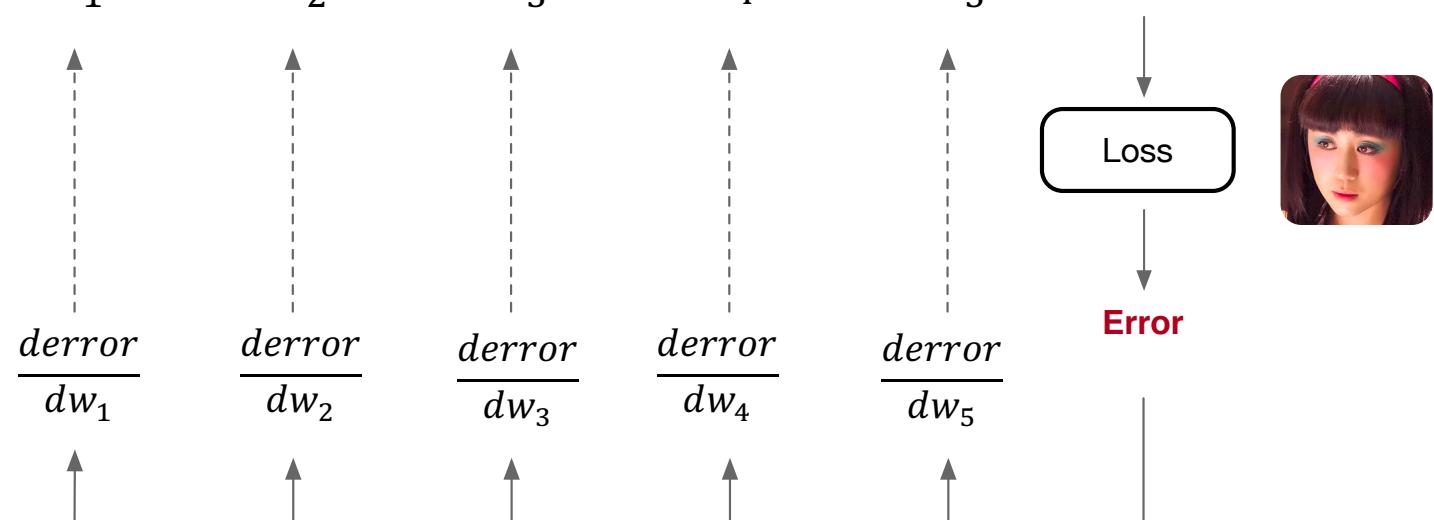
2. 定义训练优化目标：

# 回顾：模型训练过程

1. 定义一个神经网络：



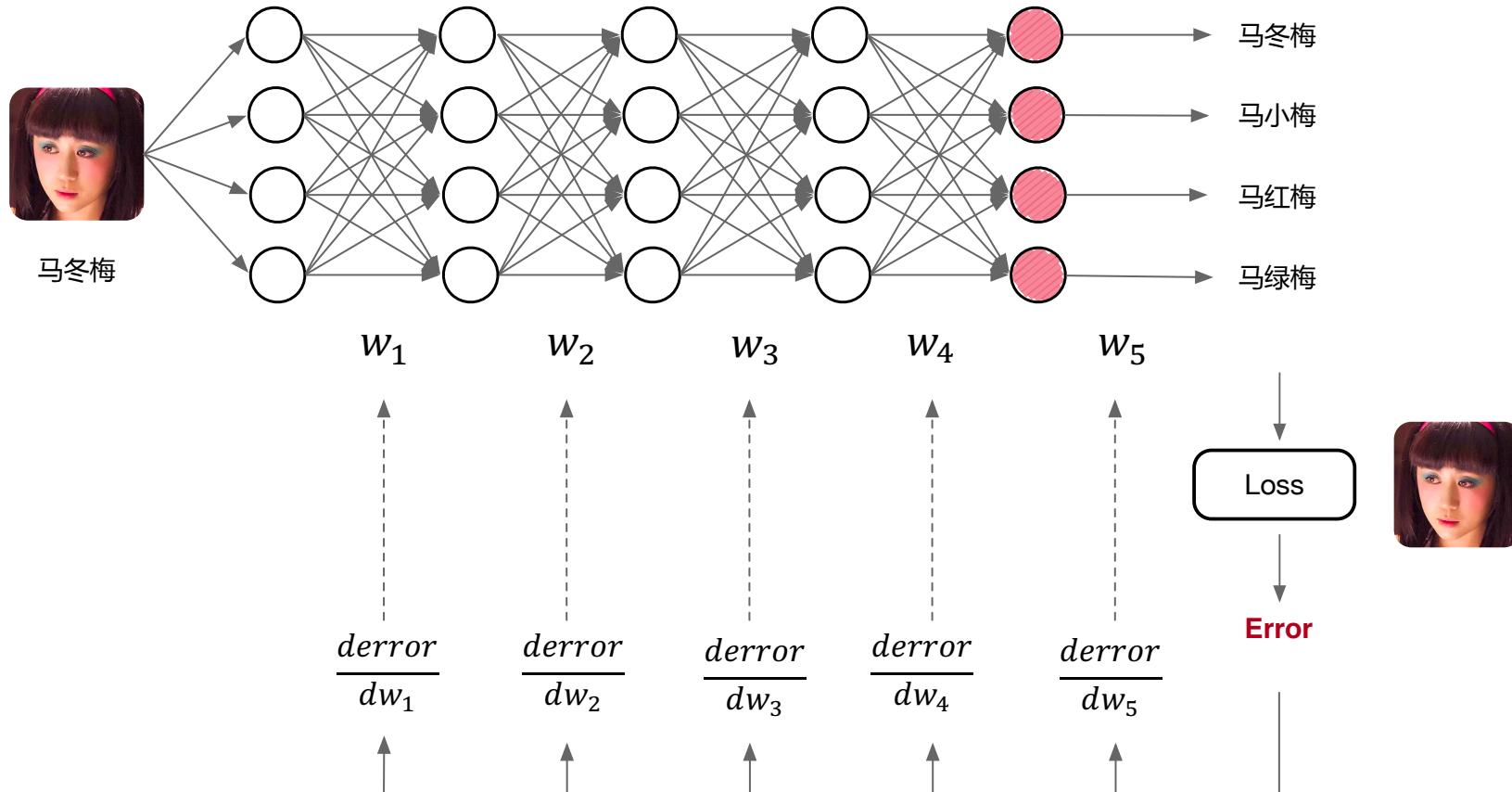
2. 定义训练优化目标：



3. 计算梯度并更新权重参数：

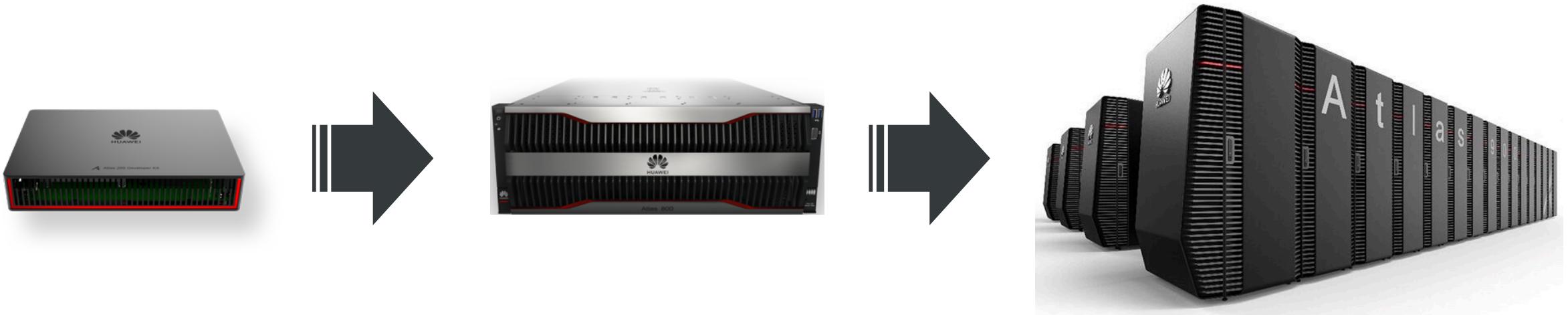
# 回顾：模型训练过程

计算神经网络损失，然后求梯度，最后更新权重参数



# 从单卡 -> 节点 -> 集群

训练从单卡、到单节点（单机8卡）演进到AI集群（百卡），需要集群架构承载训练能力



# 加快大模型训练速率

$$\text{训练时间} = \text{训练数据规模} \times \text{单步计算量} / \text{计算速率}$$

模型相关，相对固定

可变因素

- 加快计算速率 (Computation rate) :

$$\text{计算速率} = \text{单设备计算速率} \times \text{设备数} \times \text{多设备并行效率 (加速比)}$$

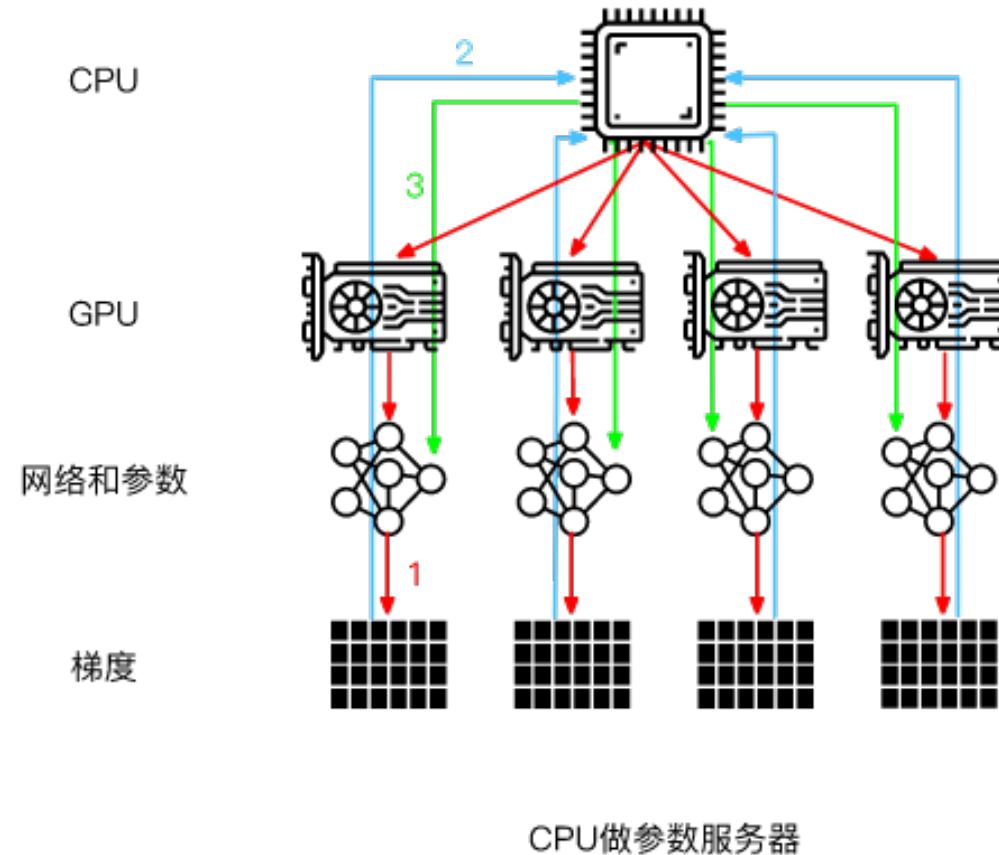
混合精度  
算子融合  
梯度累积

**服务器架构**  
通信拓扑优化  
存储系统优化

数据并行  
张量并行  
流水并行

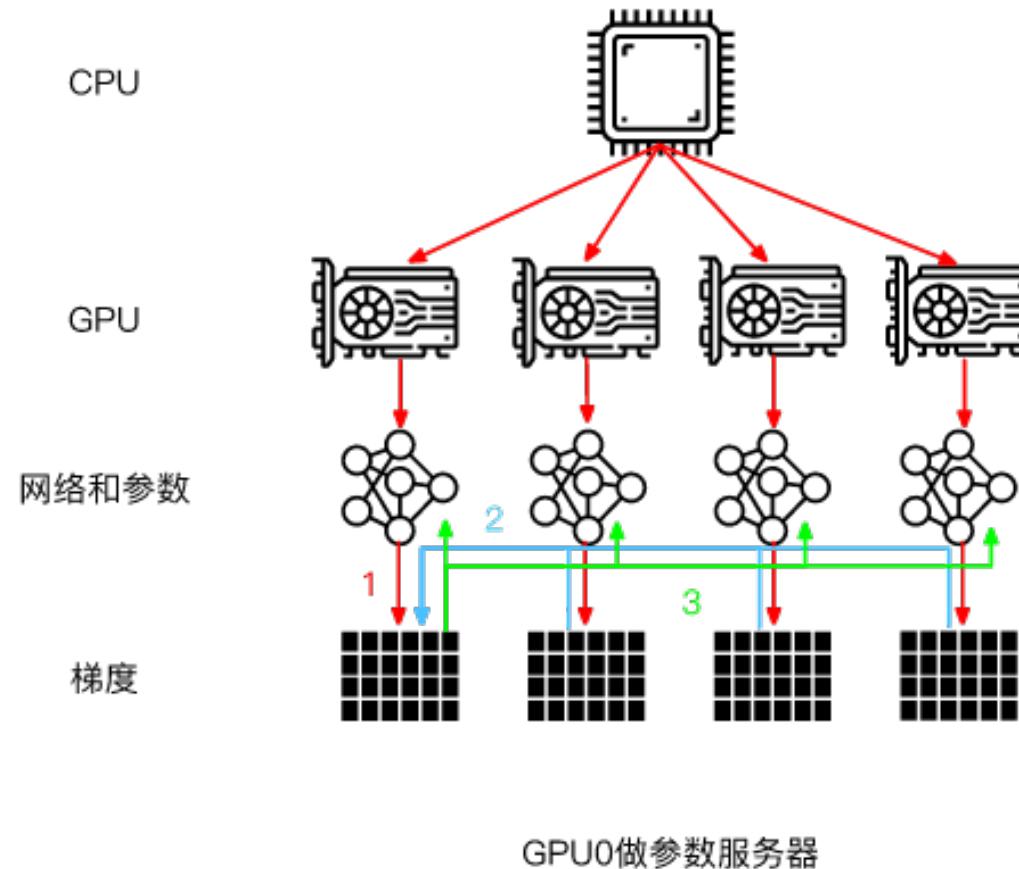
# 分布式架构：参数服务器

( 1 ) 计算损失和梯度 ( 2 ) 梯度聚合 ( 3 ) 参数更新并参数重新广播



# 分布式架构：参数服务器

( 1 ) 计算损失和梯度 ( 2 ) 梯度聚合 ( 3 ) 参数更新并参数重新广播



# Question ?

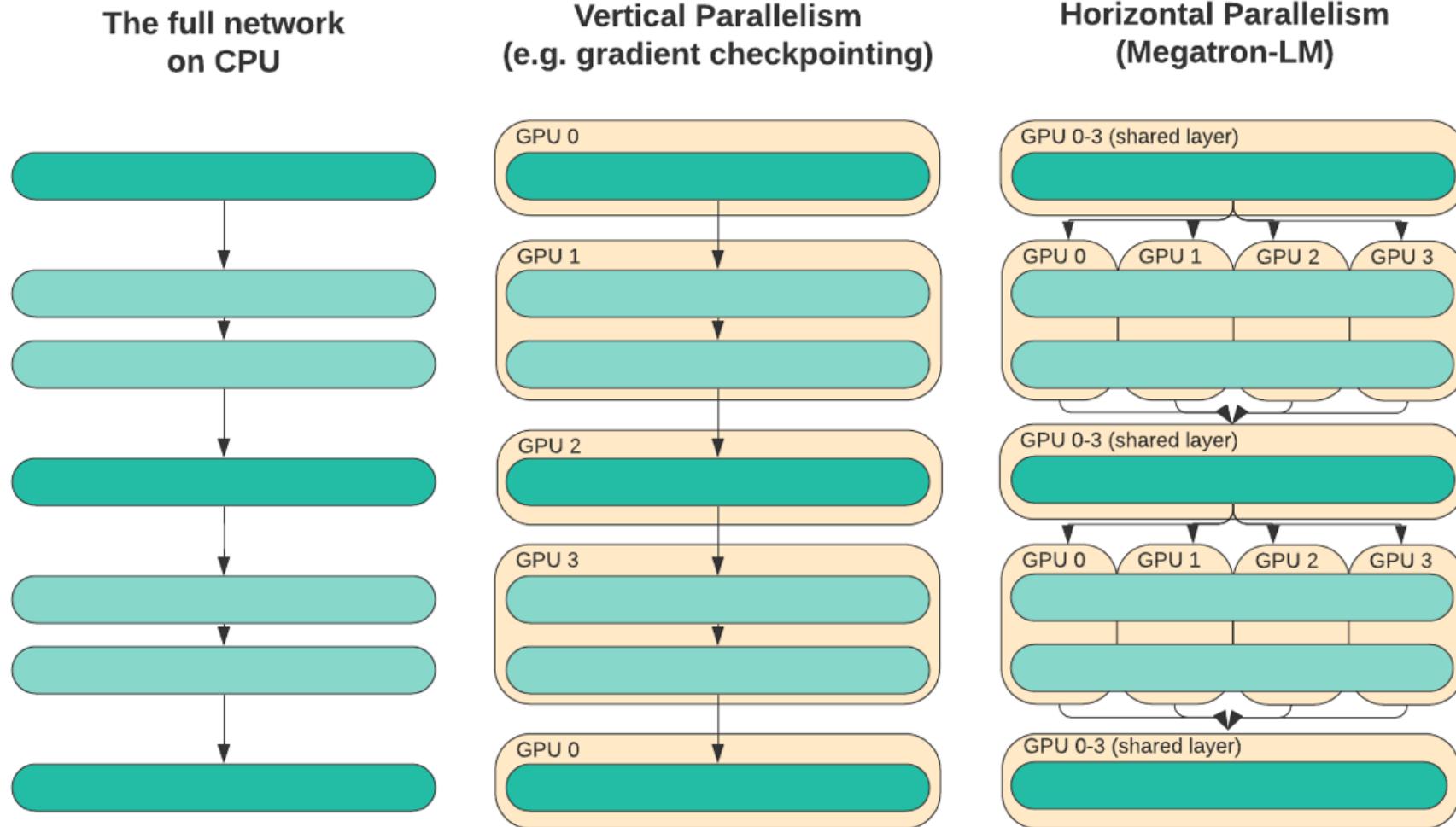
- 大模型训练是不是不适用于参数服务器P-S架构呢？

参数服务器架构

集合通信架构



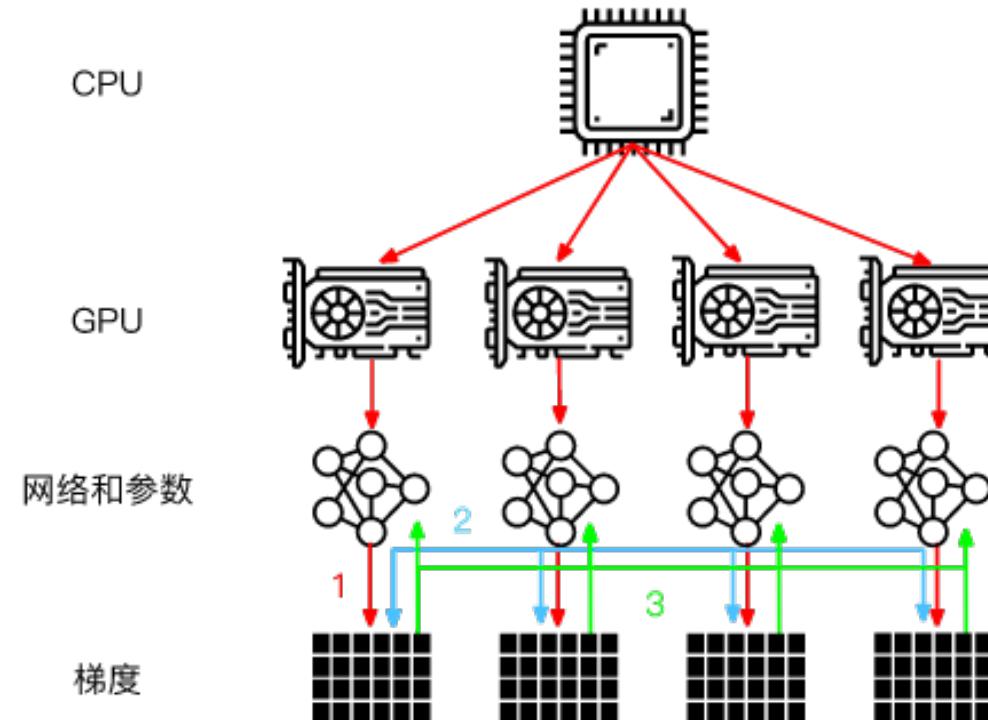
# Megatron-LM 语言大模型



Narayanan, Deepak, et al. "Efficient large-scale language model training on gpu clusters using megatron-lm." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021.

# 分布式架构：参数服务器 + 大模型（集合通信）

( 1 ) 计算损失和梯度 ( 2 ) 梯度聚合 ( 3 ) 参数更新并参数重新广播



参数服务器分布在所有GPU上

# 小结&思考



# Question?

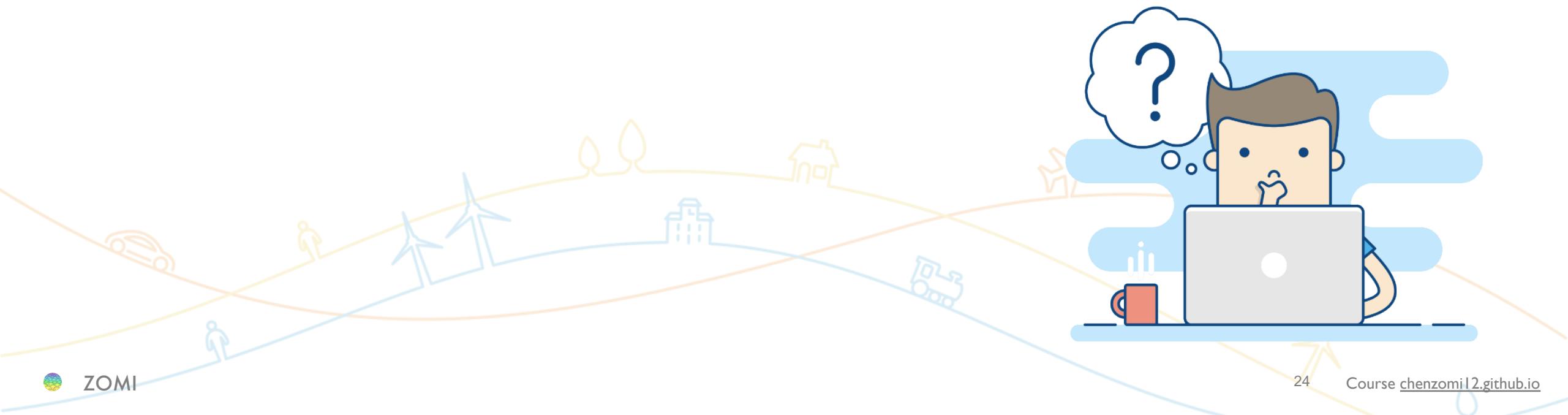
## I. AI 集群规模越大越好？大集群拥有大算力？

- 集群训练引入通信开销，集群的算力并不是线性增长，增加 GPU 计算节点，并不能线性地提升算力收益。要想通过 AI 集群提供更多算力，需要优化服务器间通信、拓扑、模型并行、分布式框架等软硬件协同。
- 对于网络而言，高速、低延迟的网络可以缩短节点间同步梯度的时间，加快训练过程；对于计算而言，降低不必要的计算资源消耗，使计算节点能够专注于训练任务。



# 小结

1. 了解AI集群由从计算、存储、管理节点和集群辅件组成；
2. 深入探讨AI集群服务器的主要硬件和之间的关系；
3. 通过回顾深度学习训练流程，了解从单卡到AI集群服务器架构；





# 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.

Copyright © 2023 XXX Technologies Co., Ltd.  
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course [chenzomi12.github.io](https://chenzomi12.github.io)

GitHub [github.com/chenzomi12/DeepLearningSystem](https://github.com/chenzomi12/DeepLearningSystem)