大模型-AI集群 (存)

大模型遇到存储

ZOMI

大模型业务全流程

数据 & 模型算法

2. 数据处理

- 开源数据
- 数据预处理
- 向量数据库

3. 模型算法

- LLM模型架构
- 多模态一切皆Tokens

模型训练 & 微调

4. 模型训练

- 混合精度
- 梯度检查
- 梯度累积
- …
- 5. 分布式并行
- 训练集群稳定性

6. 模型微调

- 全参微调
- 低参微调
- 指令微调

模型验证 & 推理部署

7. 模型验证

- 下游任务
- 测评标准
- bench mark

8. 推理与智能体

- 量化压缩
- 推理加速
- 9. Agent智能体

集群算力准备

1. AI 集群建设

- 计算、存储、网络
- AI 集群机房建设
- AI 集群上线与运维

存储的现状

- 数据存储经历了文件存储、数据库存储、大数据存储、数据湖存储、湖仓一体存储、云原生存储等不同发展阶段。

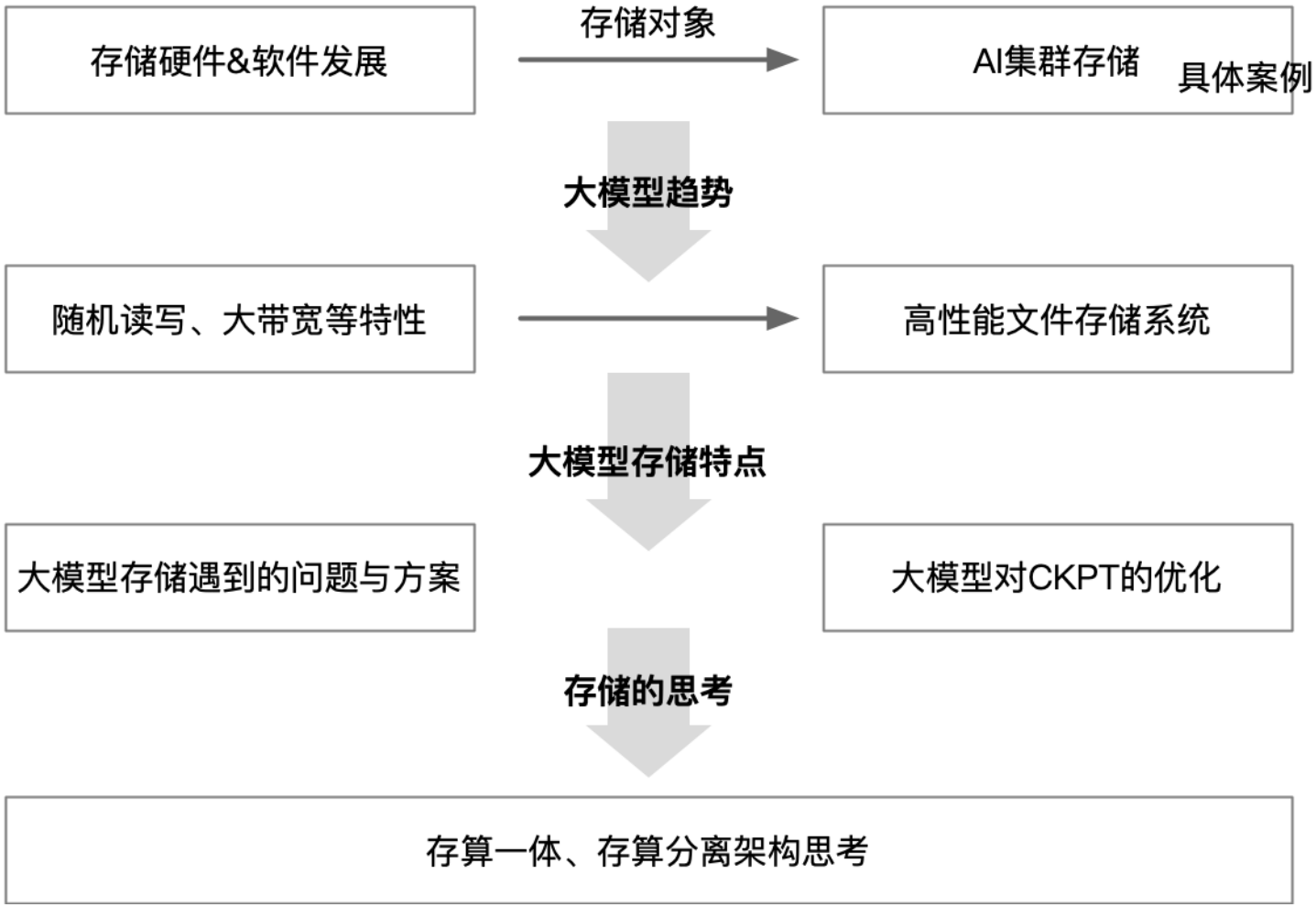
存储的新技术点

- 从而带来新的存储技术点，存算一体、存内计算、异构存储、多模态存储、小文件存储、分布式存储、高性能存储、增量存储、向量存储等。
- 但是对于 AI 大模型而言，有些是 AI 的未来，有些并不是 AI 所需要的，有些只是作为增值服务，有些是落后的技术，因此需要了解 AI 对于存储的需求。

关于本内容

- **内容背景**
  - AI 集群 + 大模型
- **具体内容**
  - **数据存储现状和场景**：存储软件类型、存储硬件类型的发展
  - **大模型对存储的挑战**：存储性能指标、存储遇到大模型挑战与新机会点
  - **大模型训练CKPT优化**：大模型训练过程、CKPT过程分解、CKPT优化
  - **大模型时代对存储的思考**：什么样的存储架构才是AI大模型时代的选择？

关于本内容


存储硬件&软件发展
存储对象
AI集群存储
具体案例
大模型趋势
随机读写、大带宽等特性
高性能文件存储系统
大模型存储特点
大模型存储遇到的问题与方案
大模型对CKPT的优化
存储的思考
存算一体、存算分离架构思考

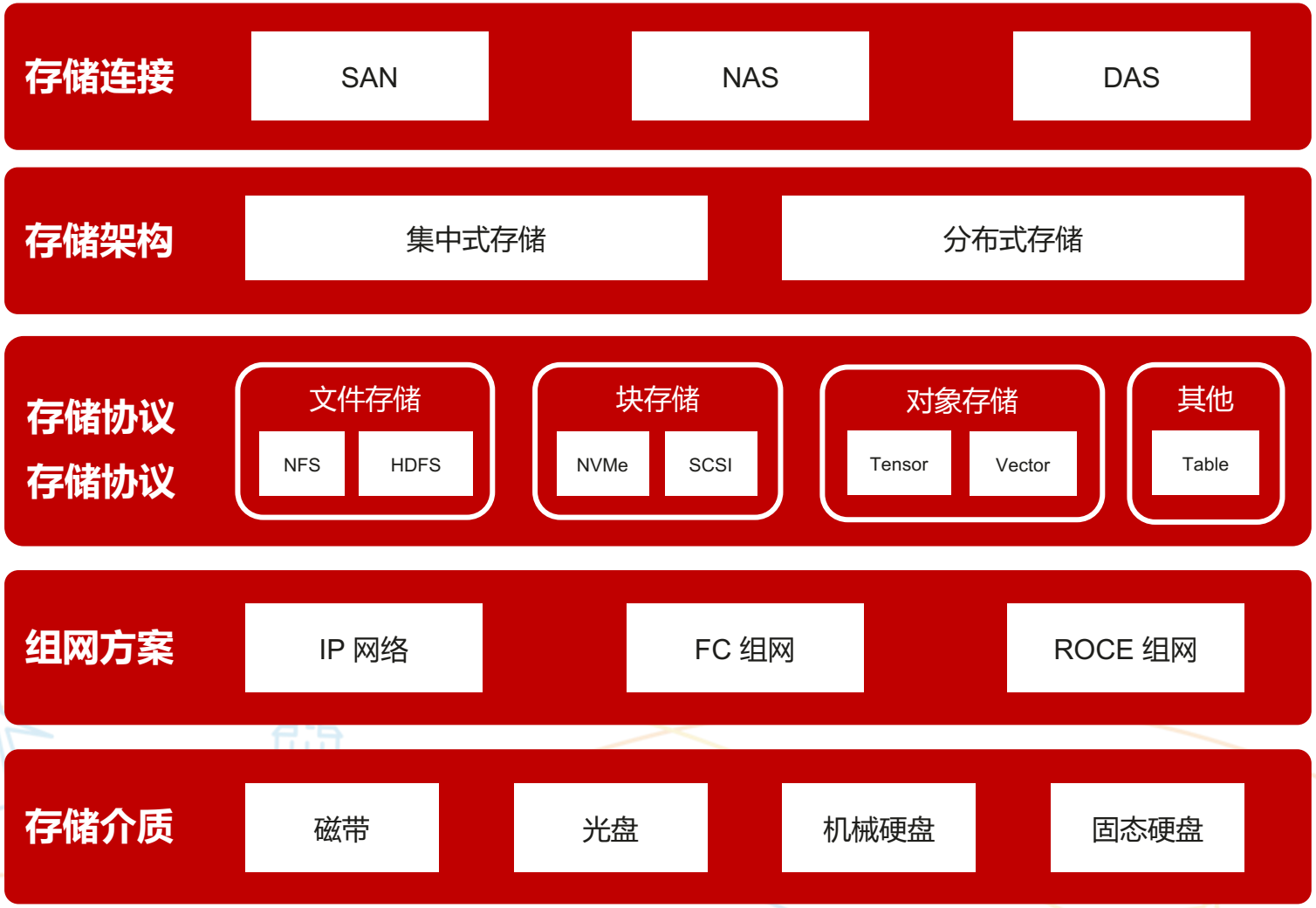
存储使用场景变化

1. 存储系统随着应用的发展在不断进化
   - 从早期的数据库应用催生的集中式存储
   - Web2.0 应用催生分布式存储架构
   - 电商、视频等移动应用产生软件定义存储、融合存储架构

2. 面向 AI 大模型时代产生新应用和海量数据，对数据存储带来怎样的挑战和机遇？存储系统面向AI 的新场景应该如何进化？

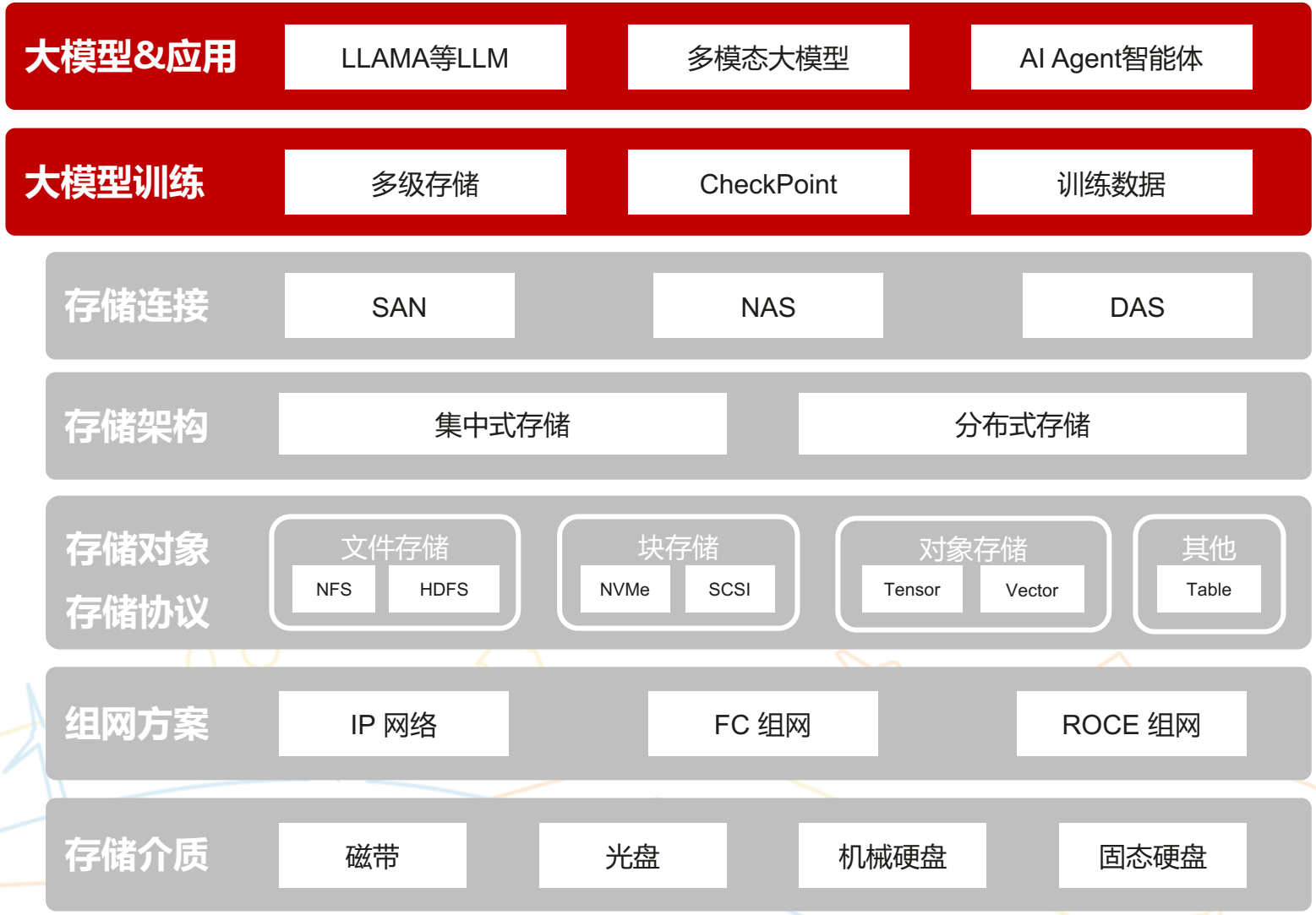
增长的存储主要靠数据中心和存储服务器拉动

存储技术架构



增长的存储主要靠数据中心和存储服务器拉动

存储技术架构



Reference 引用&参考

1. https://www.linkedin.com/pulse/dram-scaling-challenges-grow-aken-cheung
2. https://nn.wikipedia.org/wiki/Random_Access_Memory
3. https://www.youtube.com/watch?v=OwtV0NrTLHA
4. https://www.agrade.com.cn/IndustryNews/683.html
5. https://www.cnblogs.com/FireLife-Cheng/p/17183491.html
6. https://www.wpgdadatong.com/blog/detail/43099
7. https://www.sohu.com/a/476456430_121149658
8. https://www.programmersought.com/article/66421832717/
9. https://blog.router-switch.com/2021/03/storage-architecture-nas-vs-san-vs-das/
10. https://www.linkedin.com/pulse/dram-scaling-challenges-grow-aken-cheung
11. https://36kr.com/p/1933583910734464
12. https://lvxixiao.gitbook.io/blog/os/store/hardware
13. https://www.redhat.com/zh/topics/data-storage/network-attached-storage
14. https://community.fs.com/cn/case-study/the-common-types-of-storage.html
15. https://help.aliyun.com/zh/ecs/user-guide/nvme-protocol
16. https://blog.csdn.net/fgf00/article/details/52592651

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

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem