

# 存储系统在大模型训练与推理中的实践经验

苏锐 Juicedata





# 自我介绍

- 苏锐
- 2017 参与创办 Juicedata, 负责商业化与社区发展工作



[linkedin.com/in/suave](https://linkedin.com/in/suave)



[x.com/suavesu](https://x.com/suavesu)



# Content 目录

- 01 背景：基础模型与训练集的变化
- 02 存储的三种类型，与 AI 业务中的选型
- 03 JuiceFS 的设计思路
- 04 案例：JuiceFS 在 AI 数据管道中的实践



# Part 01

## 背景：基础模型与训练集的变化



# Scaling Law 下的模型与训练数据集变化

## CV 领域

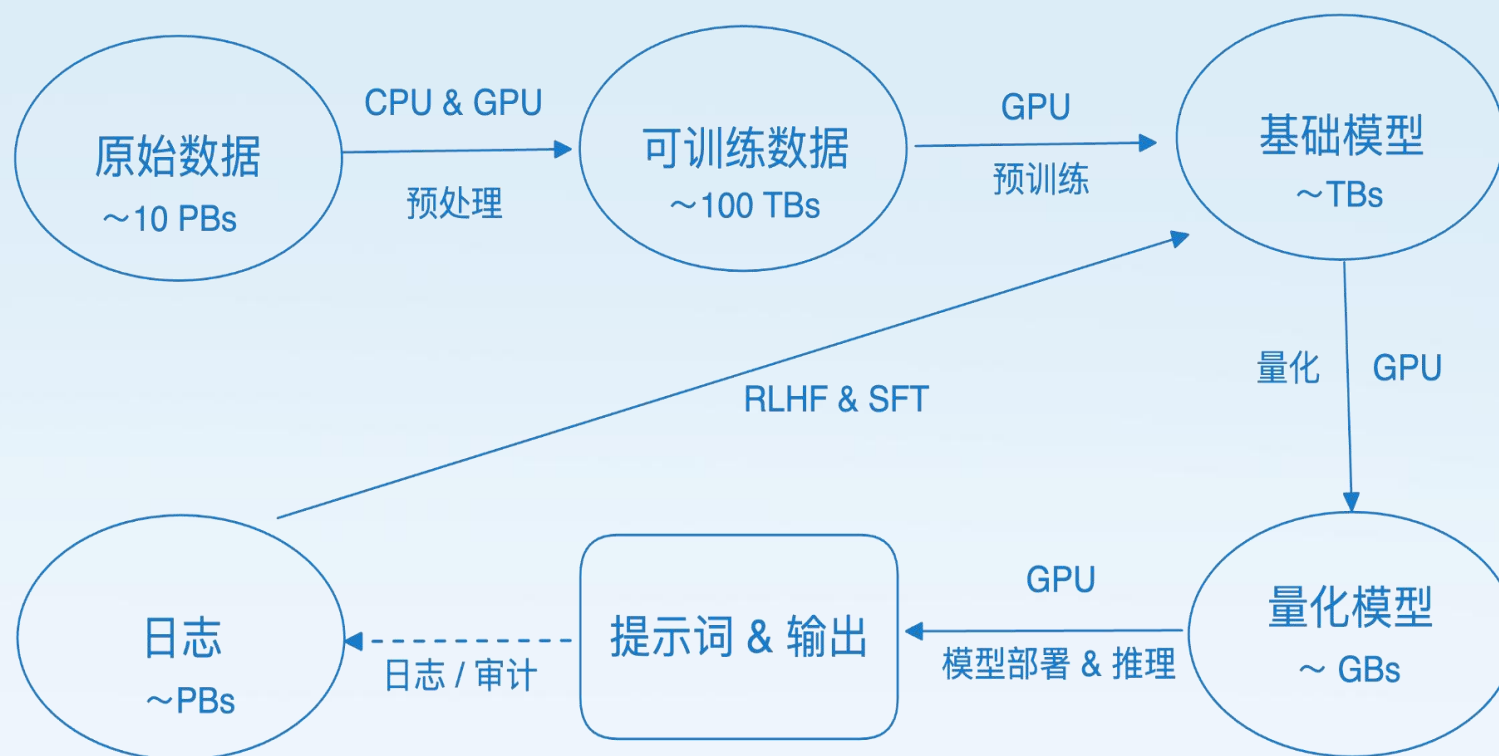
- MNIST, 70K imgs, 50MB
- ImageNet, 1.5M imgs, 150GB
- OpenImages, 9M imgs, 500GB
- SDXL Base+Refiner model, ~26GB

## LLM 领域

- GPT, 参数 110M, 文本 5.7G
- GPT-2, 参数1.5B, 文本 40G
- GPT-3, 参数 175B, 文本 45TB
- GPT-4, 参数 1800B, 文本 1PB
- Yi Model, 3T tokens ~ 6TB

数据集越来越大，模型和 **Checkpoint** 也越来越大。  
单机存储必须转为分布式存储，单机训练也必须转为多机训练。

# AI 数据管道



# Part 02

## 存储的三种类型，与 AI 业务中的选型





# 存储系统的三种类型

	块存储 Block Storage	对象存储 Object Storage	文件存储 File Storage
产品举例	EBS, SSD裸盘	S3, MinIO	EFS, CephFS
多机访问	✗	✓	✓
POSIX 兼容	✓	✗	✓
容量	✗ 有上限	✓ 弹性	部分产品弹性
时延	✓ 低	✗ 高	中
吞吐	✗ 固定	与数据量相关	与数据量或磁盘数相关



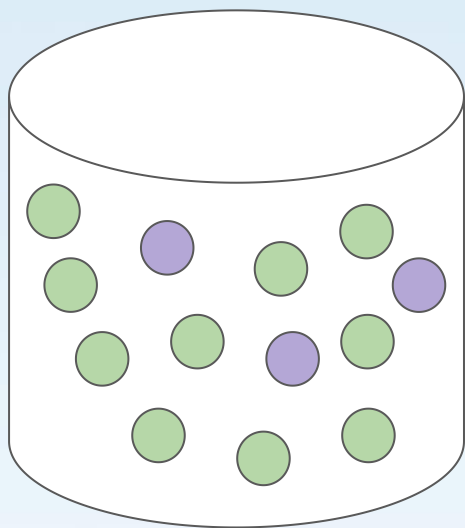


# 存储系统的三种类型

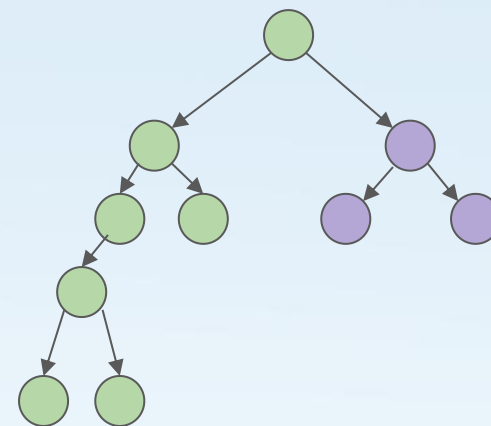
	块存储 Block Storage	对象存储 Object Storage	文件存储 File Storage
产品举例	<del>EBS, SSD裸盘</del>	S3, MinIO	EFS, CephFS
多机访问	分布式训练需要 共享存储。	✓	✓
POSIX 兼容		✗	✓
容量		✓ 弹性	部分产品弹性
时延		✗ 高	中
吞吐		与数据量相关	与数据量或磁盘数相关

# 存储系统的三种类型

对象存储 vs. 文件存储，怎么选？



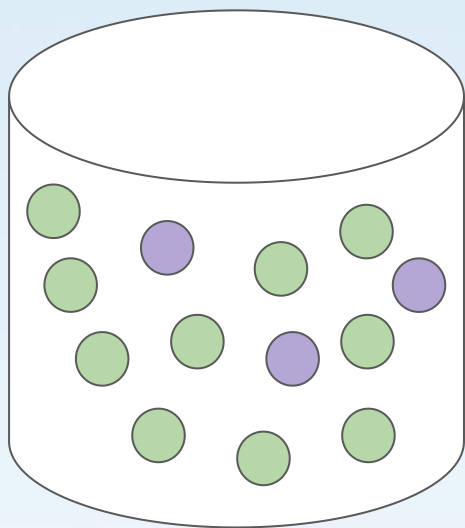
对象存储的命名空间  
Bucket



文件存储的命名空间  
Volume

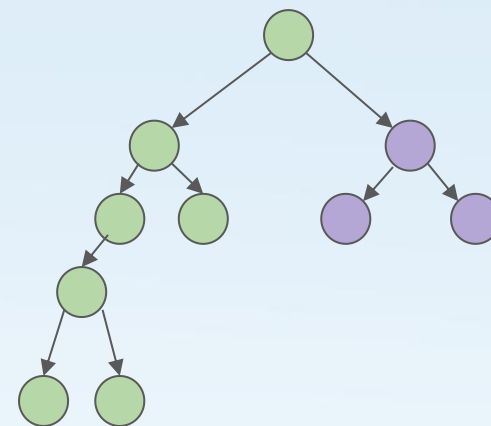
# 存储系统的三种类型

对象存储 vs. 文件存储，怎么选？



对象存储的命名空间  
Bucket

- POSIX 兼容
- 追加写，覆盖写
- 预读
- 目录遍历
- 原子改名

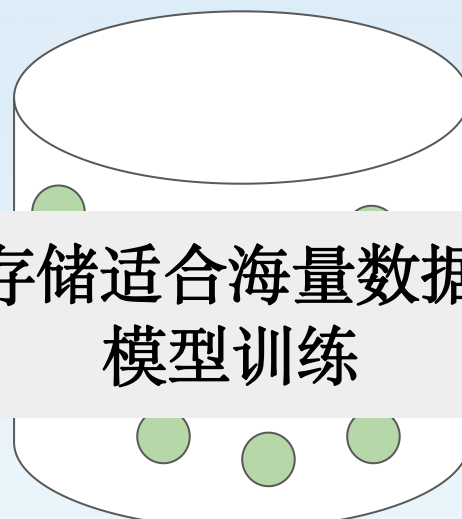


文件存储的命名空间  
Volume

- POSIX 兼容
- 追加写，覆盖写
- 预读
- 目录遍历
- 原子改名

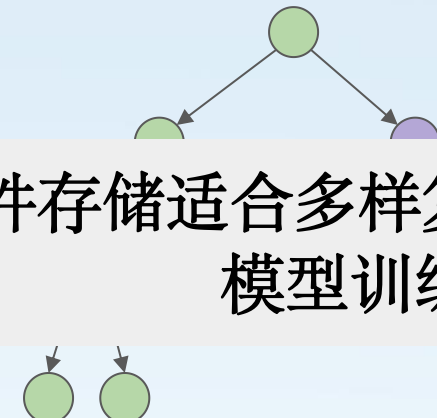
# 存储系统的三种类型

对象存储 vs. 文件存储，怎么选？



对象存储适合海量数据归档，  
模型训练

对象存储的命名空间  
Bucket



文件存储适合多样复杂的计算需求  
模型训练


文件存储的命名空间  
Volume

# 存储系统的三种类型

	块存储 Block Storage	对象存储 Object Storage	文件存储 File Storage
产品举例	<del>EBS, SSD 硬盘</del>	<del>S3, MinIO</del>	EFS, CephFS
多机访问	分布式训练必须使用可以多节点访问的共享存储。	对象存储无法支持AI 应用对数据的复杂访问需求。	✓
POSIX 兼容			✓
容量			部分产品弹性
时延			中
吞吐			与数据量或磁盘数相关



# 文件存储架构变迁

	NAS	第一代分布式文件存储	云原生分布式存储
年代	1990 年代	2005 年	2017 年
产品举例	EMC / NetApp	Ceph、HDFS、Lustre	 JuiceFS
特点	<ul style="list-style-type: none"> <li>• 单点故障</li> <li>• 控制器瓶颈</li> <li>• 共享受限</li> <li>• 横向扩展困难</li> </ul>	<ul style="list-style-type: none"> <li>• 运维成本高</li> <li>• 容量规划和扩容复杂</li> <li>• TCO 高</li> </ul>	<ul style="list-style-type: none"> <li>• 弹性伸缩</li> <li>• 弹性性能扩展</li> <li>• 全托管服务</li> <li>• TCO 低</li> </ul>
是否适用于AI训练?	硬件方案，集群扩展能力有限。	性能与容量绑定，虽然能扩展，但运维复杂度高。	利用云上基础设施，性能与容量解耦，为模型训练提供数据规模与性能的弹性扩展能力。



# Part 03

## JuiceFS 的设计思路

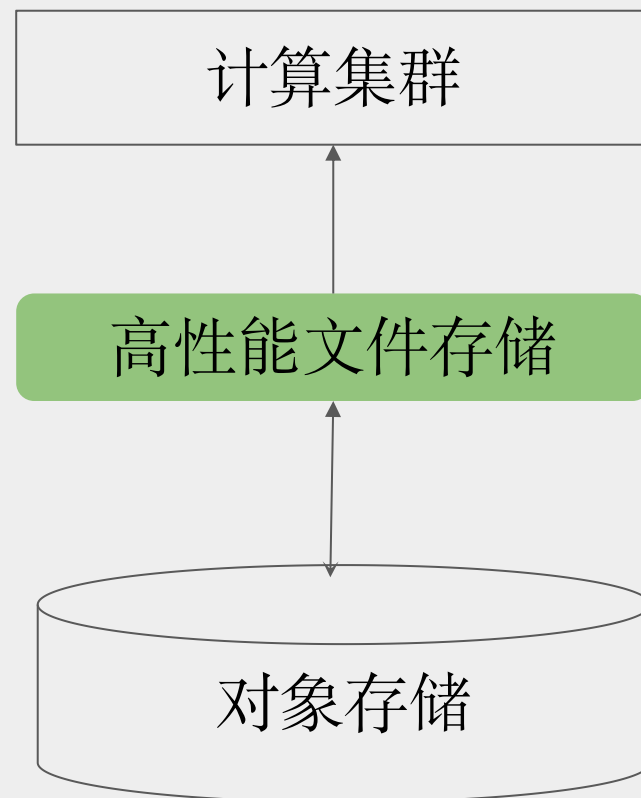




# 一种常见的用法

用户目前解决成本问题的（无奈）方法

- 数据要在两套系统中手工迁移，效率低

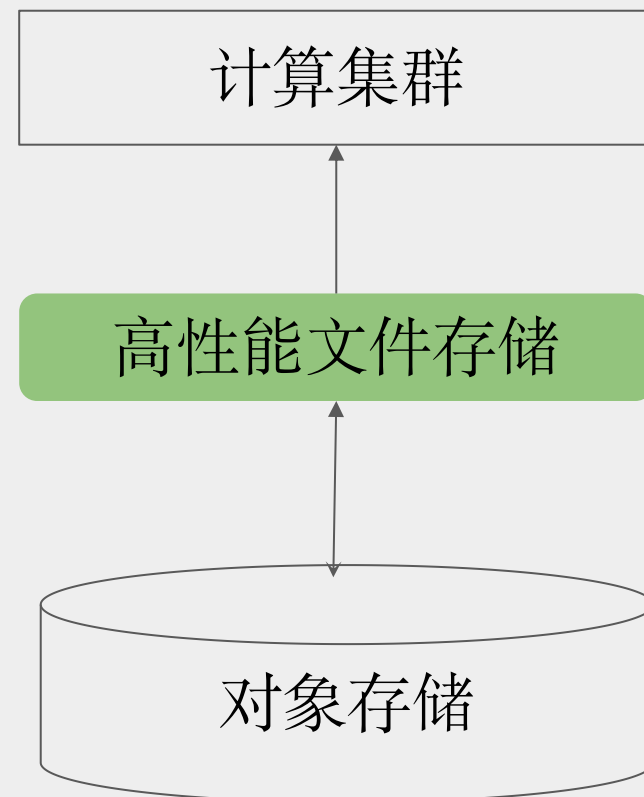


# 常见用法的痛点

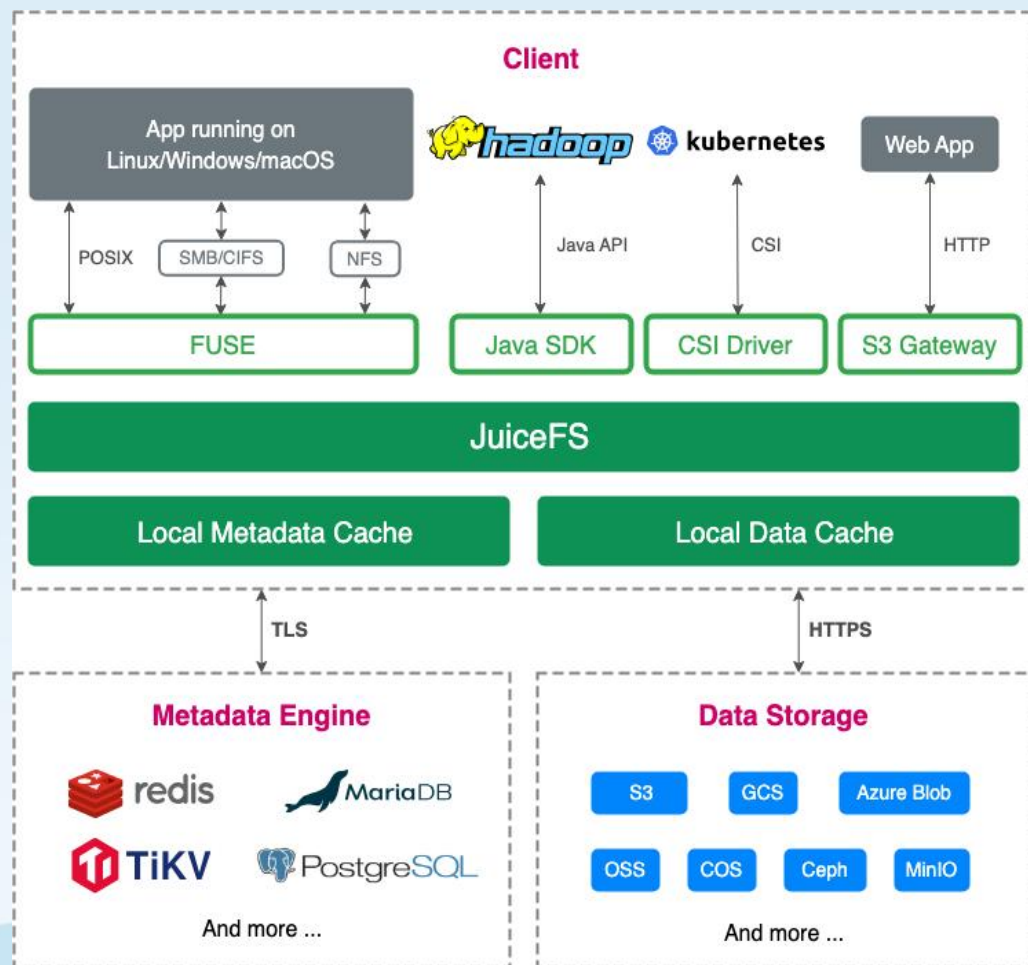
- 不仅在两套存储系统中迁移数据效率低
- 也难以应对弹性负载对存储系统的需求
- 只能按最大值预估



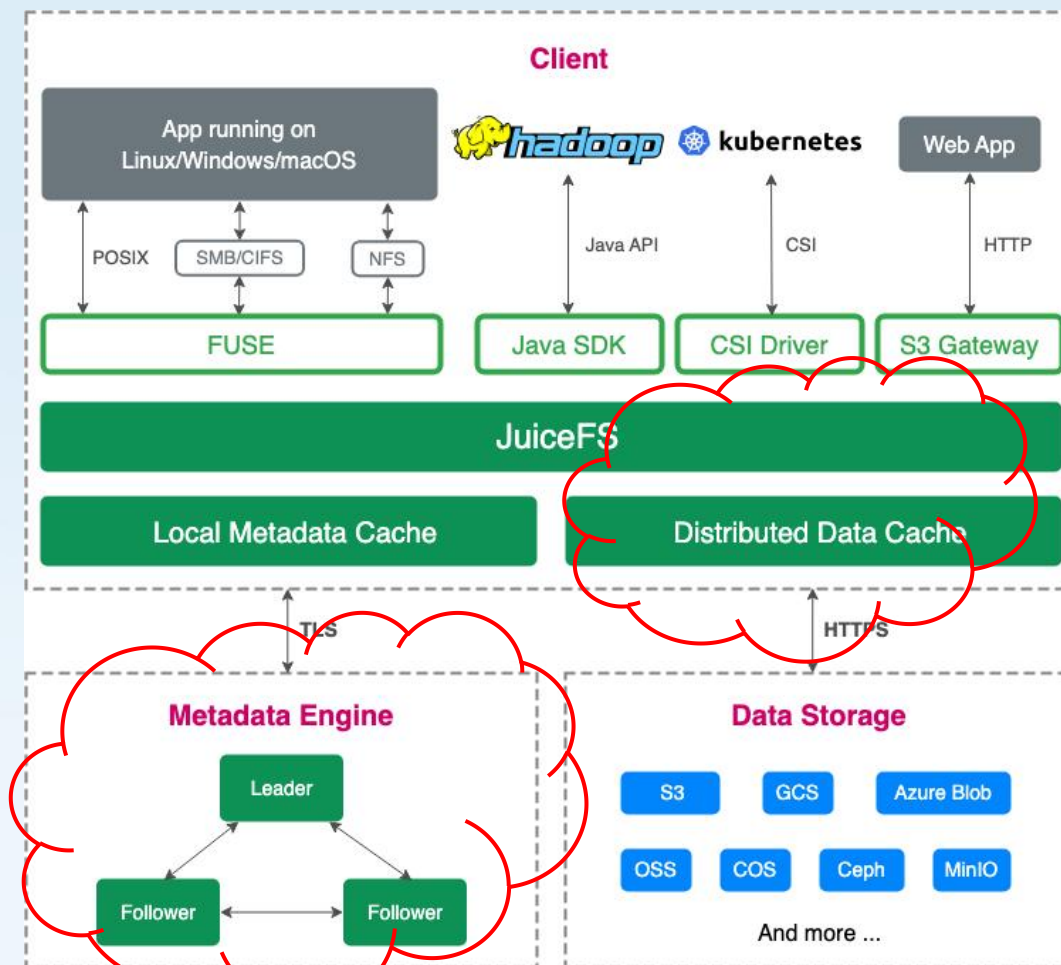
工作负载：突发任务多，弹性要求高



# JuiceFS 架构设计

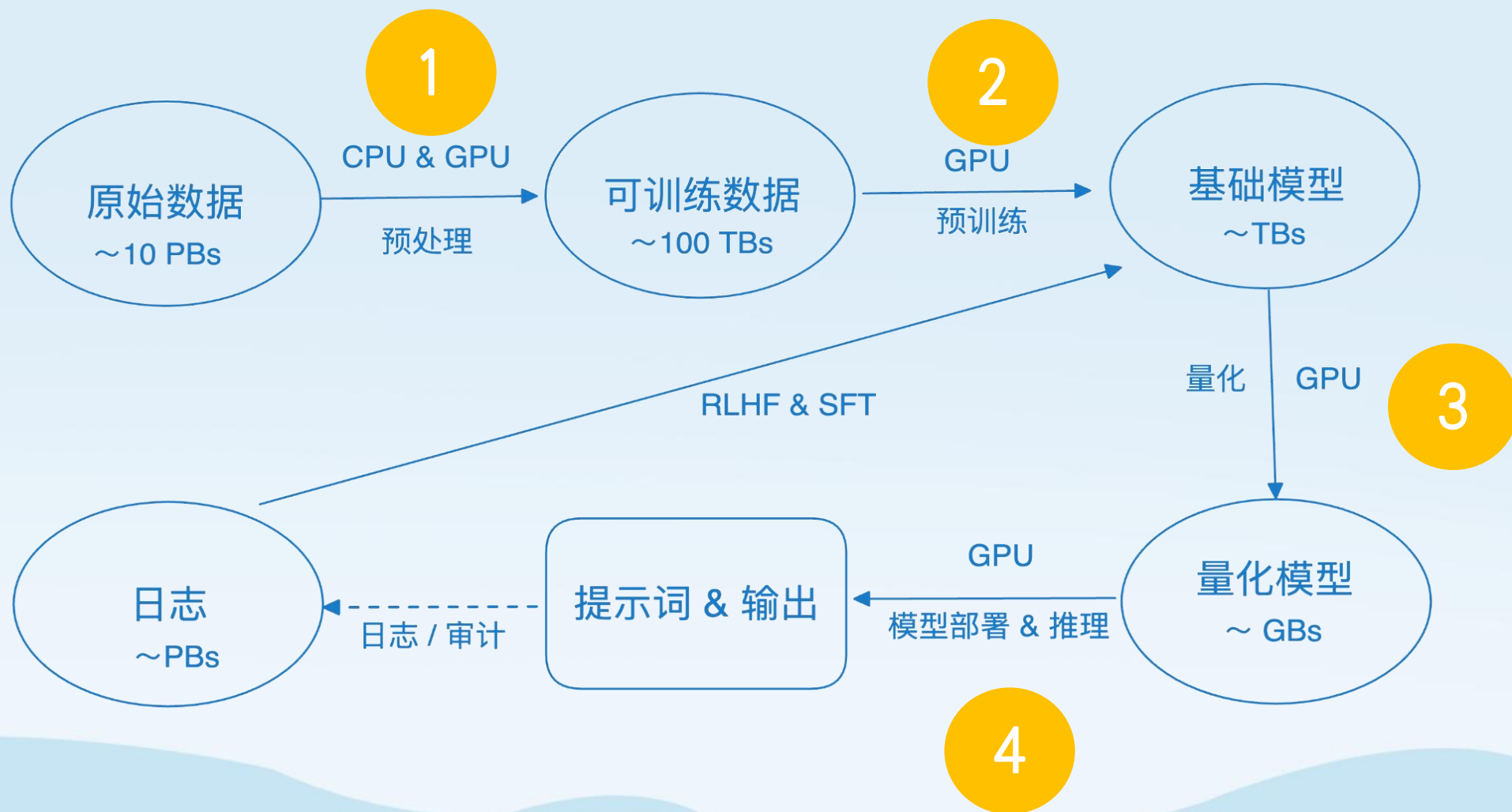


JuiceFS 社区版

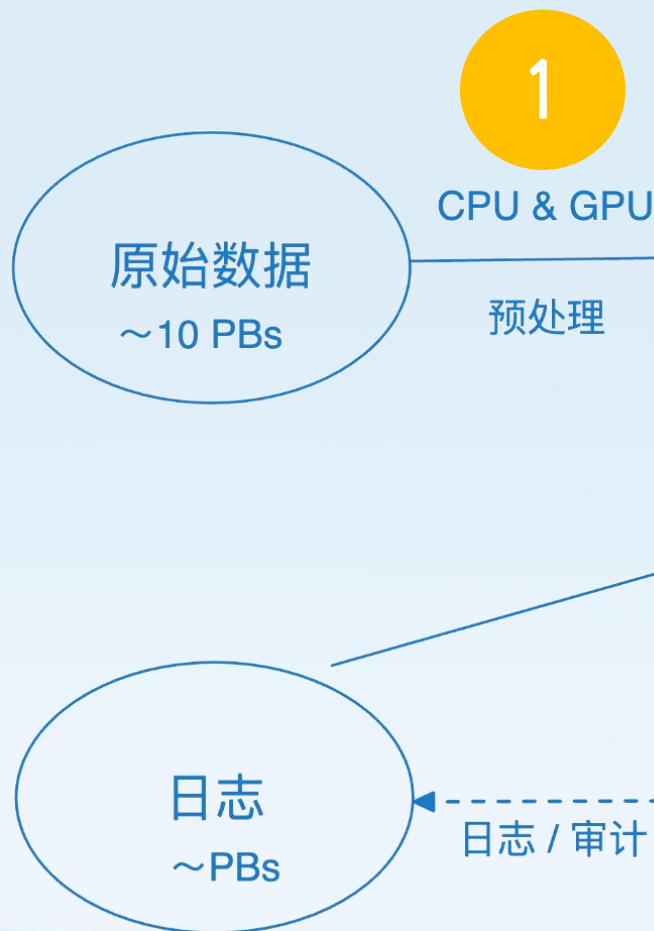


JuiceFS 企业版

# AI 数据管道



# AI 数据管道



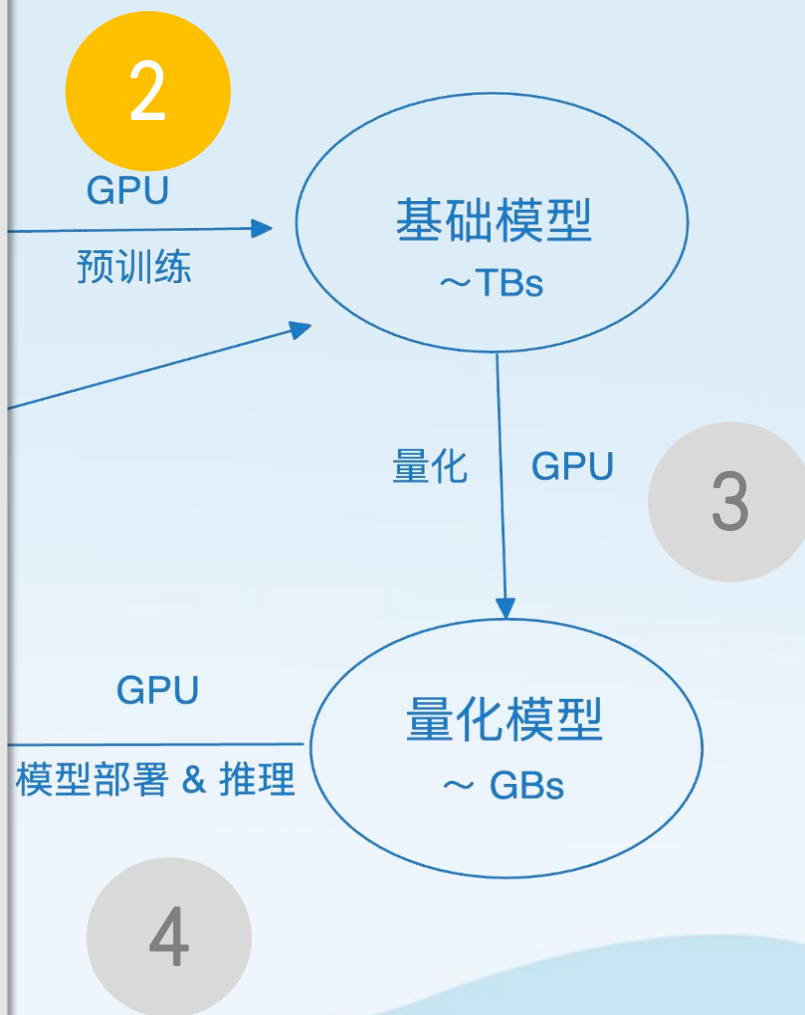
## 存储系统的性价比

- **经济**：用对象存储做数据持久层，便宜且可靠；如阿里云的 CPFS，其按官网报价（1.4 元/GB/月）来算，10PB 数据的月成本将达到千万级别，**JuiceFS TCO 仅为 20%**；
- **弹性**：用分布式缓存和对象存储提供了性能与容量的弹性扩展；
- **高性能**：相较于高性能的专用文件系统，分布式文件的性能是用户最关心的问题之一。JuiceFS 通过多级缓存加速架构为预训练提供充足的读写吞吐能力。**用户生产环境中的 I/O 吞吐量监测数据，峰值超过了 340GB/s**。很多文件存储的性能是每 TB 提供多少吞吐量，比如每 TB 容量提供 250MBps 吞吐，此时小集群无法提供高吞吐。

# AI 数据管道

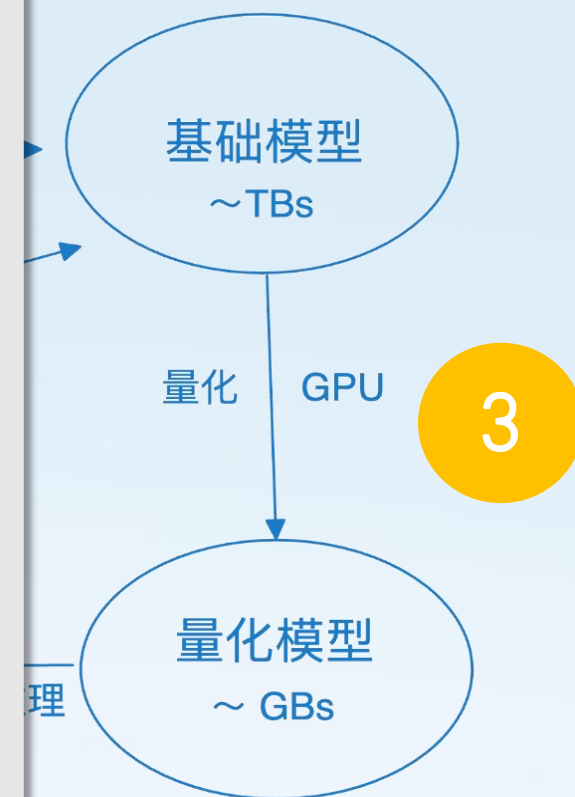
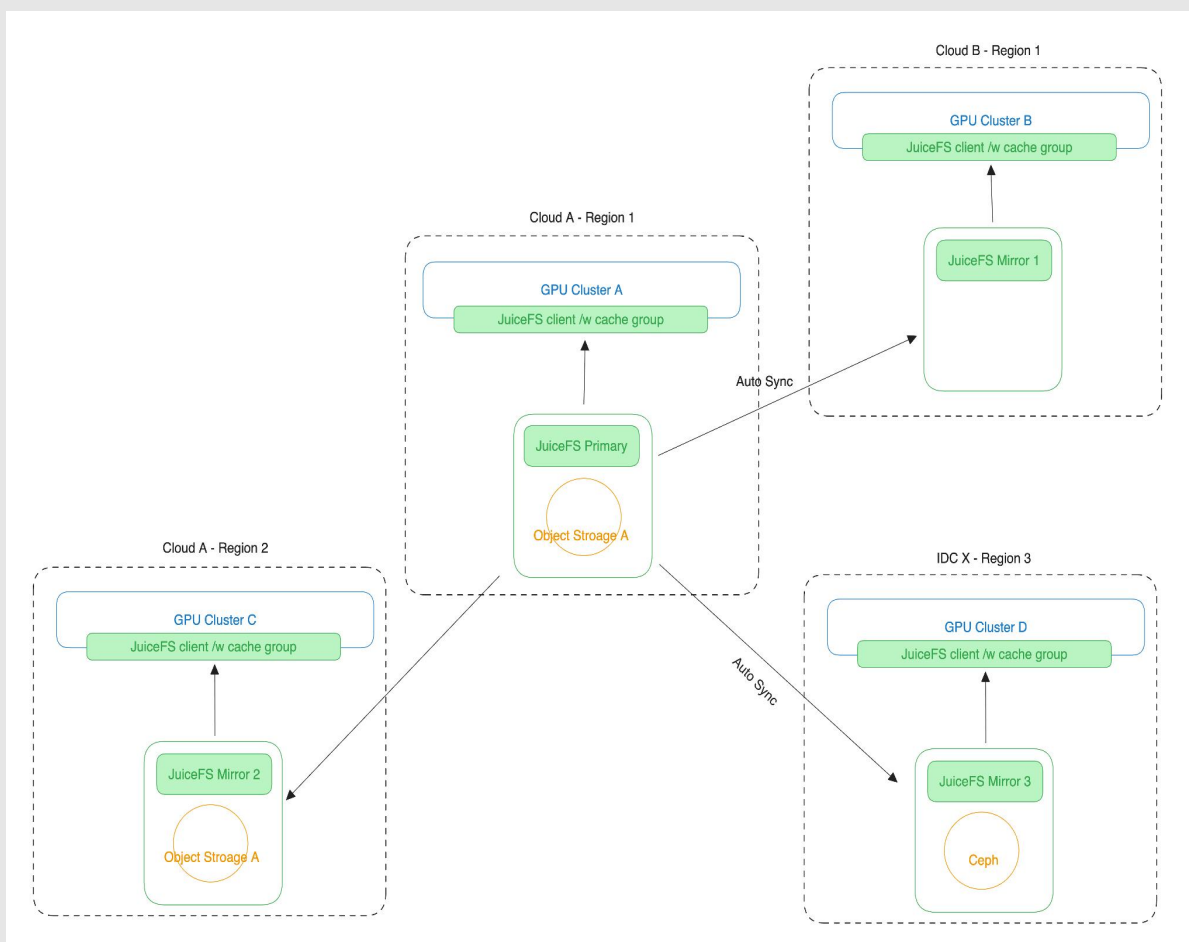
## POSIX 在 AI 业务中不可或缺

- 数据工程师几乎都在用 Python, POSIX 提供了文件操作最好的灵活性, 能满足各种数据处理需求。
- **HDFS 只支持追加写**, 无法支持需要覆盖写的数据处理方法, 比如 Pandas。同时, HDFS 的 Python SDK 也不够成熟。
- **S3 等对象存储不支持高效的追加或者修改** (只能整体覆盖), 不支持重命名操作。目录操作的性能会很慢。另外, 数据处理容易遇到对象存储的带宽限制, 和 API QPS 限制。
- **一定要注意 POSIX 兼容性**, 可以用 pjdftest。S3FS, Goofys, Alluxio 都是部分 POSIX 兼容。



# AI 数据管道

## 多云，多区域的数据访问与分发



理



# AI 数据管道

原始数据  
~10 PBs

日志  
~PBs

## 加速模型部署，提升 GPU 利用率

- 加载过程需要从存储系统中多线程顺序读取，影响速度的关键因素是多线程顺序读取时的吞吐量，JuiceFS 当前版本加载模型吞吐性能为 1500MB/s。经过为模型加载场景的优化后，读吞吐可以提升至 3GB/s。
- 以 PyTorch 加载 pickle 格式模型的过程为例，在顺序读取模型文件的同时会完成 pickle 数据的反序列化，也会消耗时间。在我们的测试中，从内存盘加载 Llama 2 7B 全精度模型，pickle 格式，26GB 大小，吞吐性能是 2.2GB/s。因为内存是最快的存储介质，所以我们将其视为极限值。从 JuiceFS 加载同样的模型，吞吐性能为 2.07GB/s，是极限值的 94%。



3

4

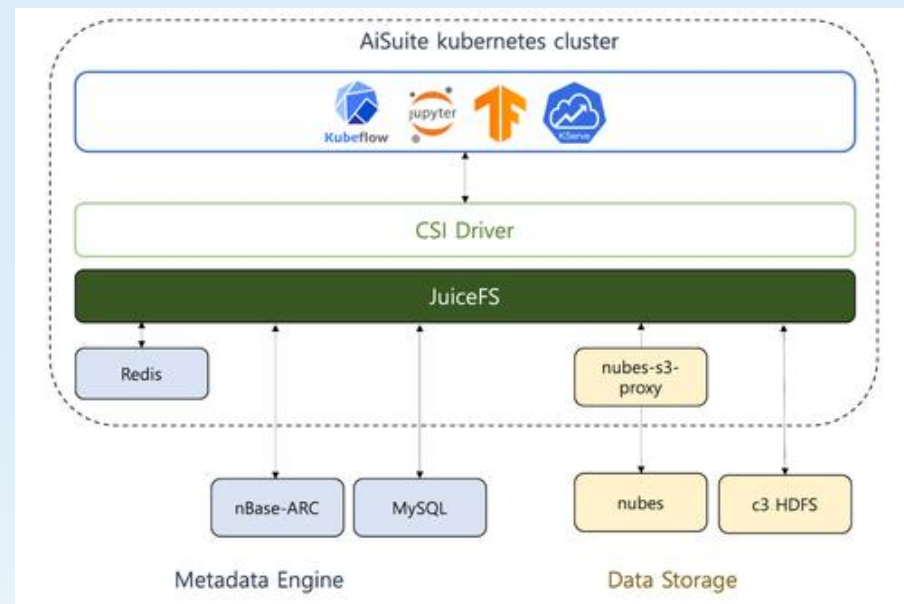
# Part 04

## 案例：JuiceFS 在 AI 数据管道中的实践



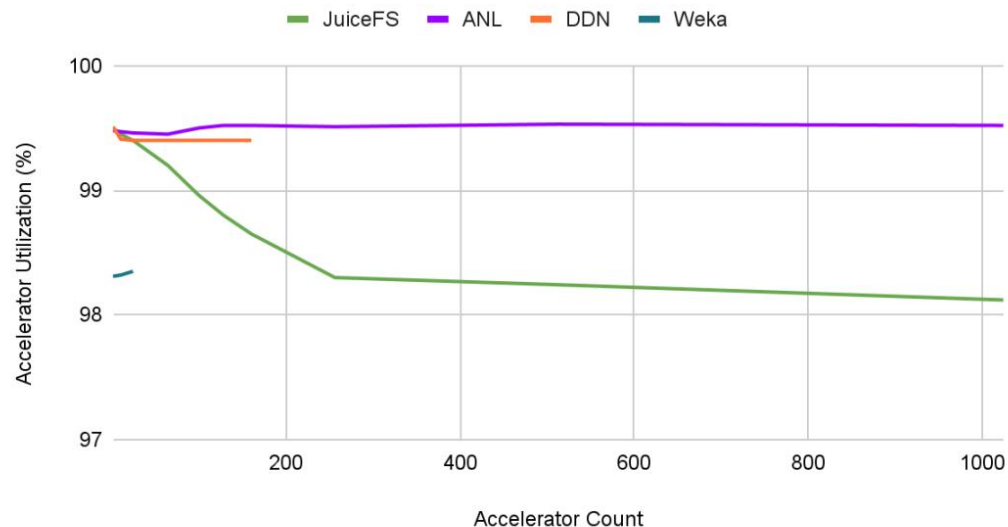
# Case Study **NAVER**

- **选型:**
  - HDFS 不支持 CSI; CephRBD 不支持 ReadWriteMany; 对象存储不支持 POSIX; NFS 缺少 HA;
- **最初引入 Alluxio:**
  - POSIX 兼容不足; 数据不一致; 运维压力;
- **选择 JuiceFS:**
  - 完全兼容 POSIX; 强一致性; 减轻运维负担;
  - 在 AI 分布式学习中, 可以作为共享的工作区、checkpoint、日志存储;
  - 可以使用大容量、可共享 (ReadWriteMany, ReadOnlyMany) 卷;
  - 高性能 (缓存), 可以替代 hostPath、local-path。可以轻松实现有状态应用的云原生转换;
  - 支持多种数据存储和元数据引擎, 适用于大多数 k8s 环境;
  - 可以替代高成本的共享存储, 如 AWS EFS、Google filestore、DDN exascaler。

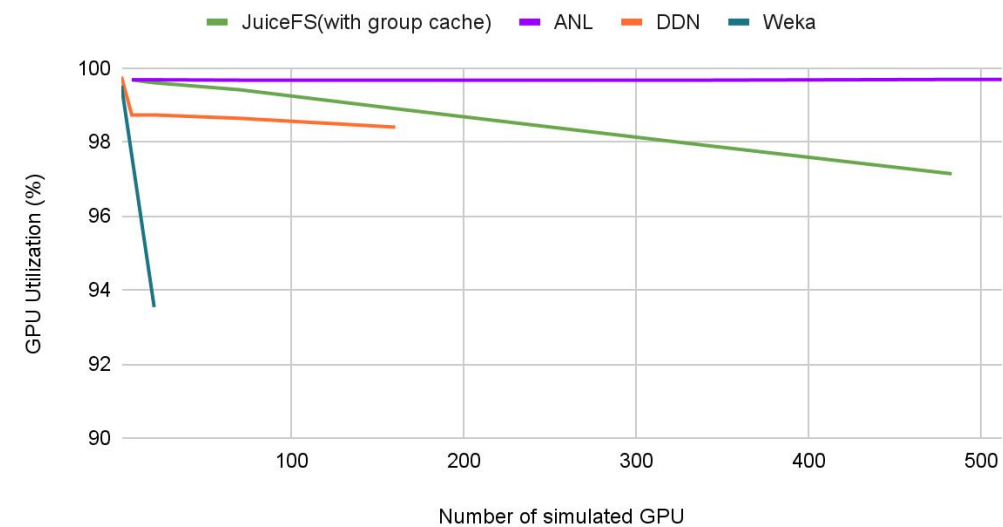


# Case Study MLPerf

BERT



Unet3D



- JuiceFS 在 1000 GPU 规模下保持 98% 以上 GPU 利用率。
- ANL 的结果依然非常优秀，考虑到 ANL 测试的网络条件是高带宽低延迟的 Slingshot 网络，能有这样的成绩也是意料之中的。

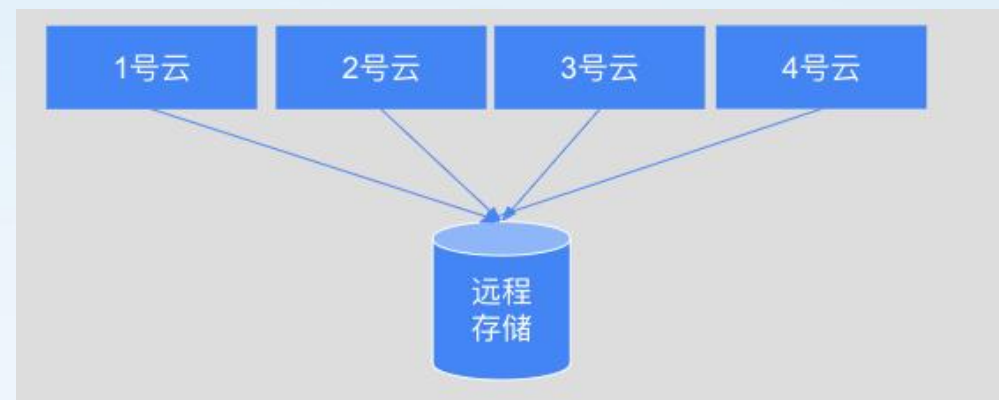
- JuiceFS 的 GPU 利用率随着集群规模变大，缓慢线性下降，在 500 卡规模时保持 97% 以上。
- 缓存节点的机型数量和网络带宽有限，本次测试达到的最大规模为 483 卡。在这种规模下，JuiceFS 集群的聚合带宽为 1.7 Tb，而 ANL 集群的带宽是 5.2 Tb。

# Case Study 知乎

社区版 JuiceFS 与企业版 JuiceFS 迁移方案对比：在执行文件系统的迁移过程中，我们同时对 JuiceFS 社区版和企业版进行了迁移操作。社区版分别采用了以 Redis 和 MySQL 作为元数据管理的两种配置。经过全面的比较后发现，**社区版在迁移期间的业务影响时间较长，且迁移过程极易受到增量数据量的影响。**

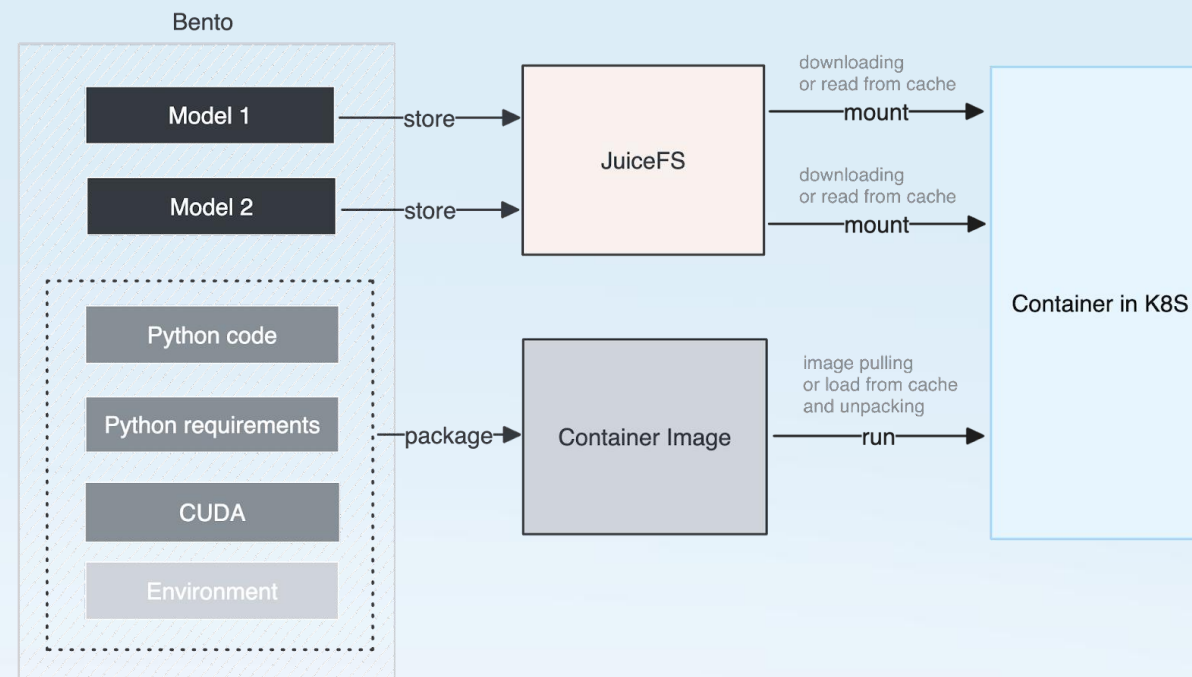
与此相反，**企业版的迁移能够保持 JuiceFS 服务的持续可用性**，尽管这要求业务方进行 3 次重启。正确选择重启时机是至关重要的，如果处理得当，对业务的影响可以降至最低。

搜广推	视觉/NLP	大语言模型	...		
机器学习平台UI与CLI命令行终端工具					
数据集管理	模型管理	笔记本	模型训练	推理服务	镜像构建
BMTrain	DeepSpeed	TensorFlow		Serving	
多个K8s集群					
HDFS		JuiceFS		云盘	



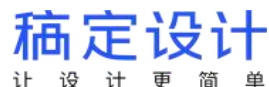
# Case Study BENTOML

JuiceFS 的 POSIX 兼容性和数据分块使我们能够按需读取数据，读取性能接近 S3 能提供的性能的上限，有效解决了大型模型在 Serverless 环境中冷启动缓慢的问题。  
使用 JuiceFS 后，模型加载速度由原来的 20 多分钟缩短至几分钟。





# 他们在使用 JuiceFS







# Thanks.

- [juicefs.com](https://juicefs.com)
- [github.com/juicedata/juicefs](https://github.com/juicedata/juicefs)



公众号



# JuiceFS 苏锐

