File Systems and Benchmark Tools for AI Storage

鲁蔚征 中国人民大学

Weizheng Lu

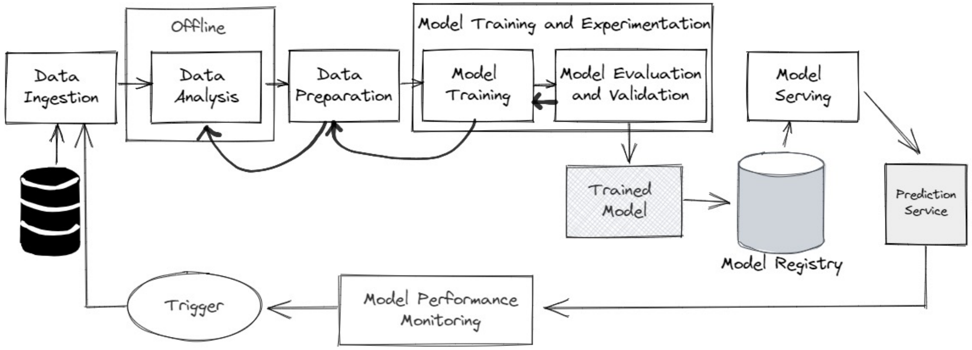
Renmin University of China

Outlines

- ML/AL Workloads
- Distributed File Systems for AI
- Benchmark Tools & Results

ML/AI Workflows

- Training
  - **Preprocessing**
  - **Data Loader**
    - small or big files
    - TFRecord or raw file
  - **Checkpoint**
- Inference



Typical AI Datasets

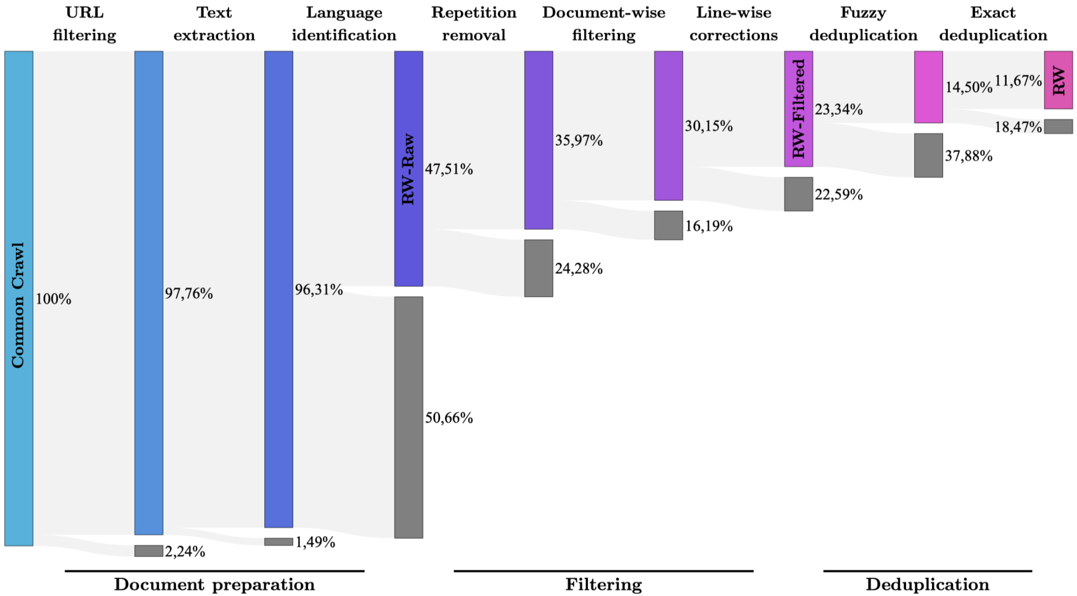
- Images and Videos
  - ImageNet: 14M small files
  - youtube-8M: 1.53TB
- Text
  - C4
  - The Pile
  - Falcon-RefinedWeb
- Recommendation Systems

ML/AI Training Characterizations

- **Big Data**
- **Same Data** Multiple Training Jobs
  - hyper-parameter tuning
    - different model archietcure, different parameters (learning rate, loss function)
  - Fluid, Microsoft Quiver
- Compute Nodes’ SSD or RAM
  - global storage v.s. compute nodes’ **local** storage
  - Lustre PCC, Alluxio, JuiceFS,

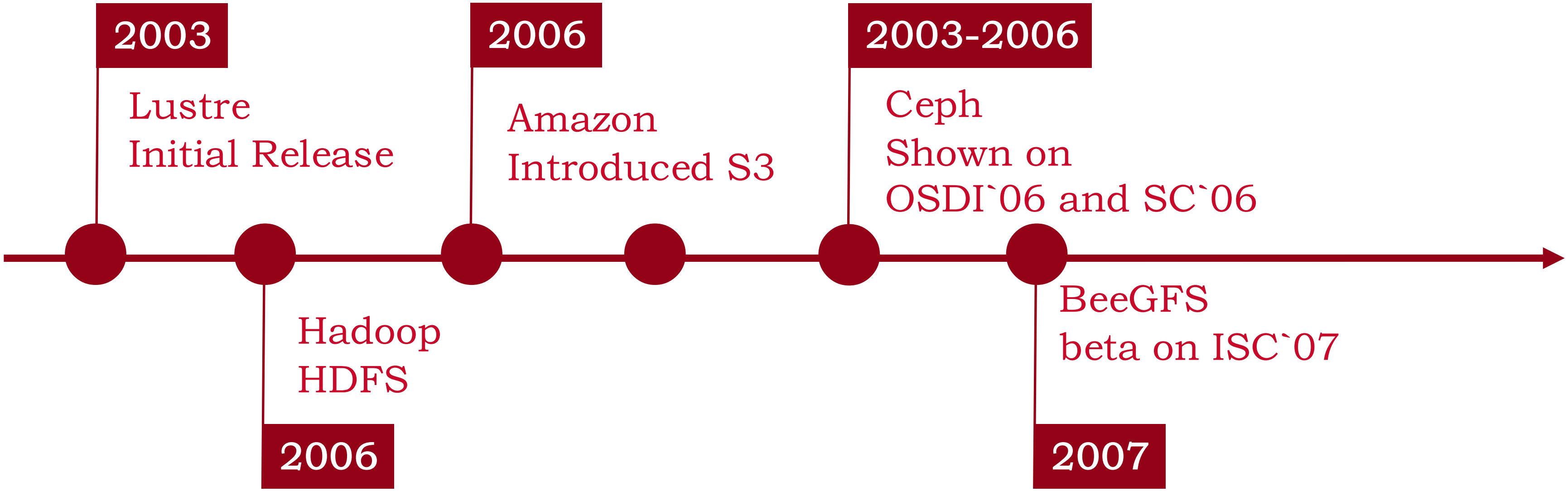
Preprocessing: Falcon-RefinedWeb

- 2.8TB extracted
- Falcon 180B LLM
- Pipelines
  - Preparation
  - Filtering
  - Deduplication



https://arxiv.org/abs/2306.01116

Timeline of Distributed File Systems



2003

Lustre
Initial Release

2006

Hadoop
HDFS

2006

Amazon
Introduced S3

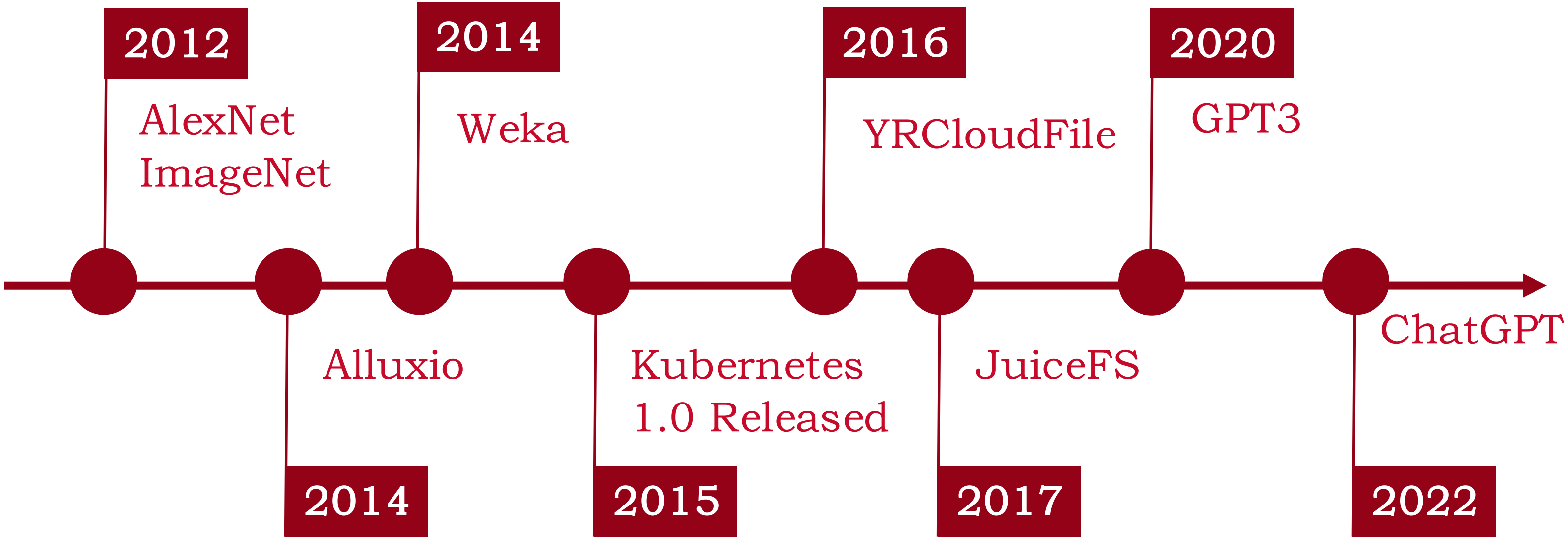
2003-2006

Ceph
Shown on
OSDI`06 and SC`06

2007

BeeGFS
beta on ISC`07

Timeline of Distributed File Systems (cont.)



2012 AlexNet ImageNet

2014 Alluxio

2014 Weka

2015 Kubernetes 1.0 Released

2016 YRCloudFile

2017 JuiceFS

2020 GPT3

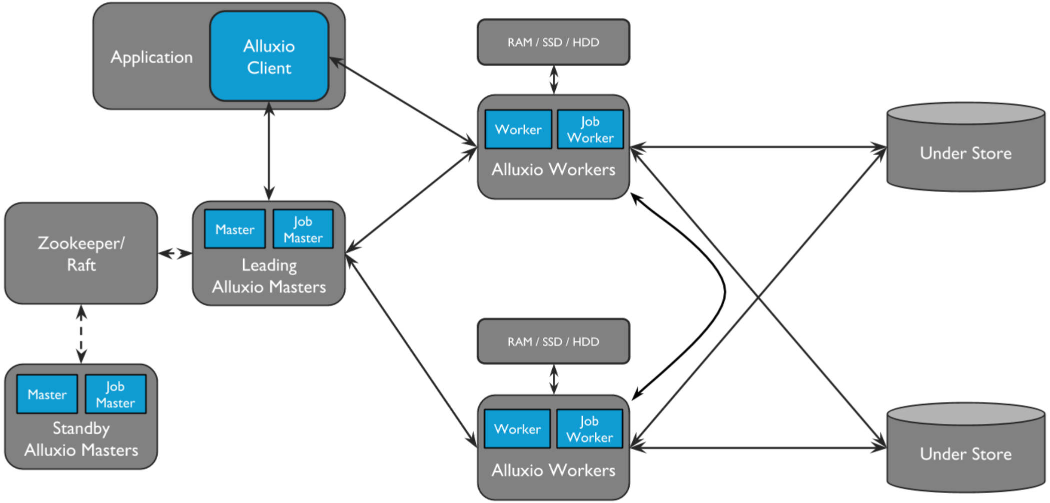
2022 ChatGPT

POSIX or not

	Pros	Cons	File Systems
POSIX	devs, ops, and software rely on POSIX Portable	overhead	Lustre, JuiceFS
non-POSIX	low cost	limited abilities additional code	HDFS, S3

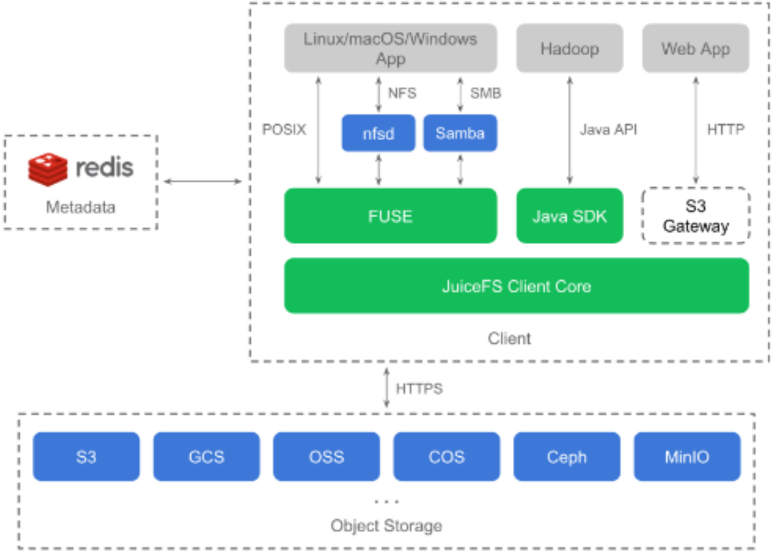
Case Study: Alluxio

- Goal: Data Orchestration
- Under Store
  - S3, HDFS, POSIX FS
- Workers
  - Cache on RAM or SSD
- Client
  - fuse



Case Study: JuiceFS

- Goal: high-performance, cloud native
- data is chunked on S3, HDFS
- metadata is in redis, MySQL, PostgreSQL
- client mount fuse

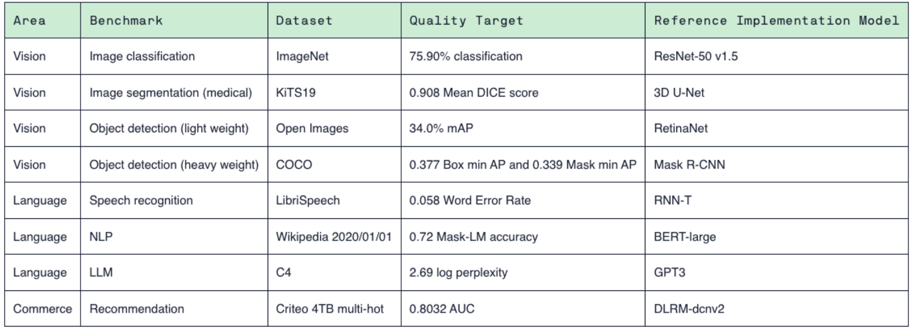


Common Benchmark Tools

- **traditional tools**
  - **IOPS & BandWidth (BW)**
  - fio
  - mdtest
  - iozone
- real-world workloads
- ML benchmark
  - MLPerf

MLPerf

- a suite contains mainstream AI workloads
  - MLPerf Training
  - MLPerf Storage
- MLPerf Storage
  - synthetic random data
  - simulate AI accelerators



Area	Benchmark	Dataset	Quality Target	Reference Implementation Model
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5
Vision	Image segmentation (medical)	KiTS19	0.908 Mean DICE score	3D U-Net
Vision	Object detection (light weight)	Open Images	34.0% mAP	RetinaNet
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN
Language	Speech recognition	LibriSpeech	0.058 Word Error Rate	RNN-T
Language	NLP	Wikipedia 2020/01/01	0.72 Mask-LM accuracy	BERT-large
Language	LLM	C4	2.69 log perplexity	GPT3
Commerce	Recommendation	Criteo 4TB multi-hot	0.8032 AUC	DLRM-dcnv2

MLPerf Training Workloads

Benchmark Results

	Lustre + all flash	Lustre + HDD	JuiceFS + S3	xfs + local SSD
fio IOPS READ	2700k	20k	14k	40k
fio BW READ	30GB/s	12GB/s	2.6GB/s	0.9GB/s
ImageNet PyTorch	1600s	1640s	1570s	1570s
LLM checkpoint (LLaMA 70B)	1 min	10 min		
MLPerf Storage UNet3D				

fio results are based on a script file from DDN
all flush: 24 * NVMe (DDN AI400) + IB
HDD: Metadata - 7 * SSD, Object - 50 * HDD (DDN 7990) + IB
JuiceFS: Metadata – redis, Object - S3 + 10Gb Eth

Discussion

- Workload < -- > Filesystem
- Benchmark Result < -- > Real Performance
- Cost < -- > Performance

Thanks