CCF Storage Evaluation

CCF 信息存储技术专委会•存储测试工作委员会

CCF专业存储3系统测评工作进展

张武生 清华大学 2025.10.17



背景介绍

关于CCF专业存储评测

2025年度工作介绍

2025年度测评结果

智能化时代，新联接、新应用对存储能力提出更高的专业要求

超级计算

大规模数据处理和数值计算

- 混合负载模型，同时满足高带宽、高IOPS和极致低时延
- 确保数据的完整性和持久性
- 可扩展性，不影响性能
- 具备高效的数据管理能力

智能计算

模型训练与推理

- 数据的可视性、可管理性和可用性
- 百万级OPS/PB的性能密度和TB级带宽/PB存储容量
- 存储系统稳定可靠和内生安全

大数据

海量、非结构、高吞吐、分布式

- 高效访问，高速读写
- 高并发读写，高IOPS
- 高可扩展，在线弹性伸缩

事务系统(关系数据库)

OLTP 业务运行

- 存储系统具有高度的可用性、事务性和数据完整性
- 要求存储系统具有良好的性能和扩展性



云计算

对象存储，虚拟化

- 实时或近实时访问
- 块对象内的复杂IO
- 高吞吐量和低延迟
- 弹性、可扩展性
- 全生命周期的安全防护

提出新测评需求：单一性能评估→与场景结合的多维度评估

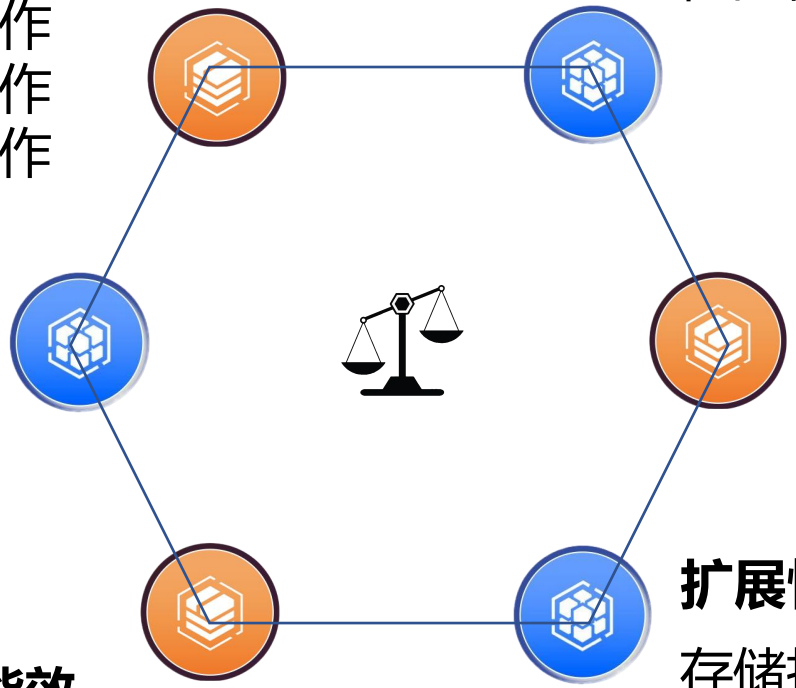
性能

应用场景决定性能需求：

读写比例
文件大小比例
目录深度
元数据操作
块设备操作
对象操作

可靠性和可用性

衡量存储系统持续支撑业务的能力
故障包容能力
性能保持能力



数据保护

防篡改
防劫持
防丢失
防非法访问

确保数据完整、可用和可信

生态协议

网络协议： RoCE、Infiniband...
访问协议： FTP、NFS，
CIFS/SAMBA、定制协议(客户端)

能效

容量/能耗比
性能/能耗比
规模/能耗比

扩展性和可关联性

存储扩展性是指存储系统能够适应数据量增长和业务需求变化的能力。
存储系统的可管理性指标反映了系统在运维、监控、故障排查和性能优化方面的便捷程度。

现状：国外相关测评组织情况

HPC TOP100

- 衡量国内HPC系统及应用发展的重要参考
- 与国际HPC TOP500榜单对标，展现中国在高性能计算领域的进展

AIPerf500

- 主要面向AI智算系统，专注智能超算的设计与优化
- 基于端到端的基准测试工具，采用典型的机器学习技术和实际任务衡量各种AI算力系统的性能

国内：AI、CPU、服务器整机测评陆续发布，存储测评尚缺影响力

IO500（2017）

- 面向超算平台IO系统的性能排行榜单，每年发布两次
- 分为针对存储系统全性能和10节点性能的研究型和生产型系统两种
- 测试负载设计为主要面向HPC的几类计算模式：带宽型、元数据操作型、查找型的性能
- 以所测结果的几何平均作为排名分数

SPC存储性能委员会（1999）

- SPC-1主要针对随机IO应用环境，模拟OLTP、数据库和e-mail等应用
- 主要评测维度有：读写响应时间、工作负载的多样性及动态变化等8种数据访问行为
- SPC-2主要针对顺序I/O应用环境，评估大文件读写、非线性编辑和大表查询，测试带宽负载场景，其指标包括可靠性、IOPS/时延、稳定性和成本

海外：多以性能测评为主，缺乏面向场景的多维度综合评估

SPECstorage Solution 2020 benchmark（2014）

由标准性能评估公司（SPEC）发布最新存储性能基准测试套件，测试文件服务器吞吐量和响应时间，为比较不同供应商平台的性能提供了标准化方法。测试包括AI图像处理、基因领域数据处理、电子设计自动化、视频数据采集等场景的负载。【商业工具， https://www.spec.org/storage2020/ 】

MLPerf Storage（2023）

- 针对AI场景，通过运行一个分布式训练测试程序，模拟GPU计算过程，仿真AI服务器对存储系统的访问模式
- 以模拟AI存储系统能够支撑的最大GPU数量，来评估存储系统测试性能表现
- 最新发布的1.0版本支持3D-Unet、ResNet50和CosmoFlow三种工作模型



背景介绍

关于CCF专业存储评测

2025年度工作介绍

2025年度测评结果

面向场景的存储测评：场景

生命科学 | 气候气象 | 天体物理 | 石油勘探 | 反应堆热功 | 空气动力学 | 人工智能 | ……

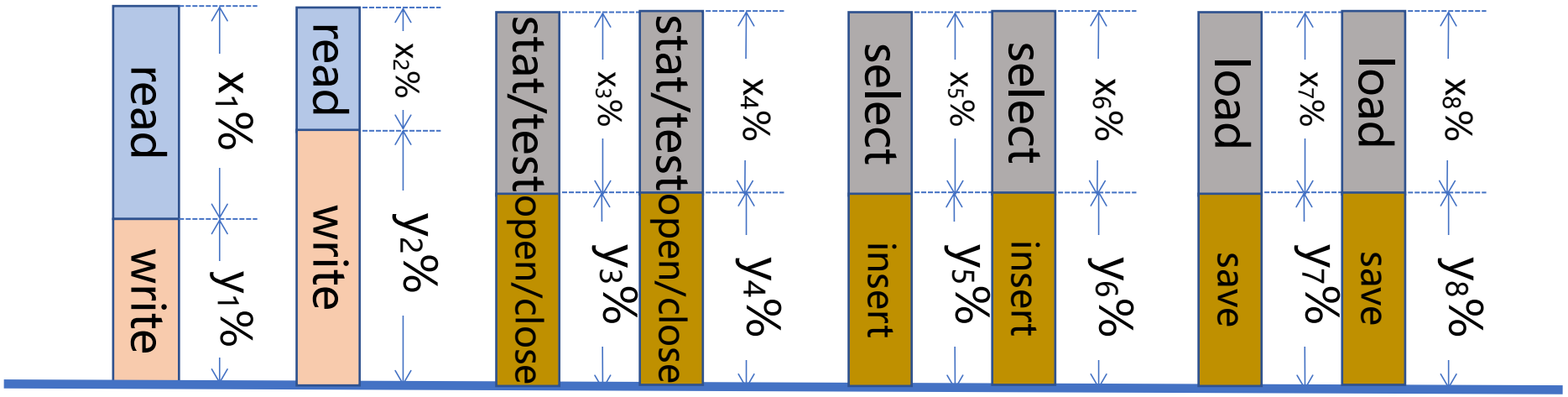
大数据处理 | 事务处理 | noSQL数据库 | 云存储 | 虚拟化/容器 | 虚拟化网络 | 虚拟机

不同类型访问，不同比例读出和写入

单用户单任务多实例

多用户单任务多实例

多用户多任务多实例



单集群

多集群

容器集群

虚拟机集群

不同硬件形式、不同架构

文件接口 | 块接口 | 对象接口 | netcdf/hdf5

不同解决方案、不同技术路线

性能+可靠+可信+可扩展+易用+易管理

面向场景的存储测评：目的、目标

1. 收集整理和建立面向领域的各类负载模型
2. 开发专业存储各类指标的评测方法，数据采集和分析方法
3. 建立可模拟真实场景的测试规范
4. 形成一套参数化、可配置、自动化的负载生成、负载运行、效果评估流程和工具

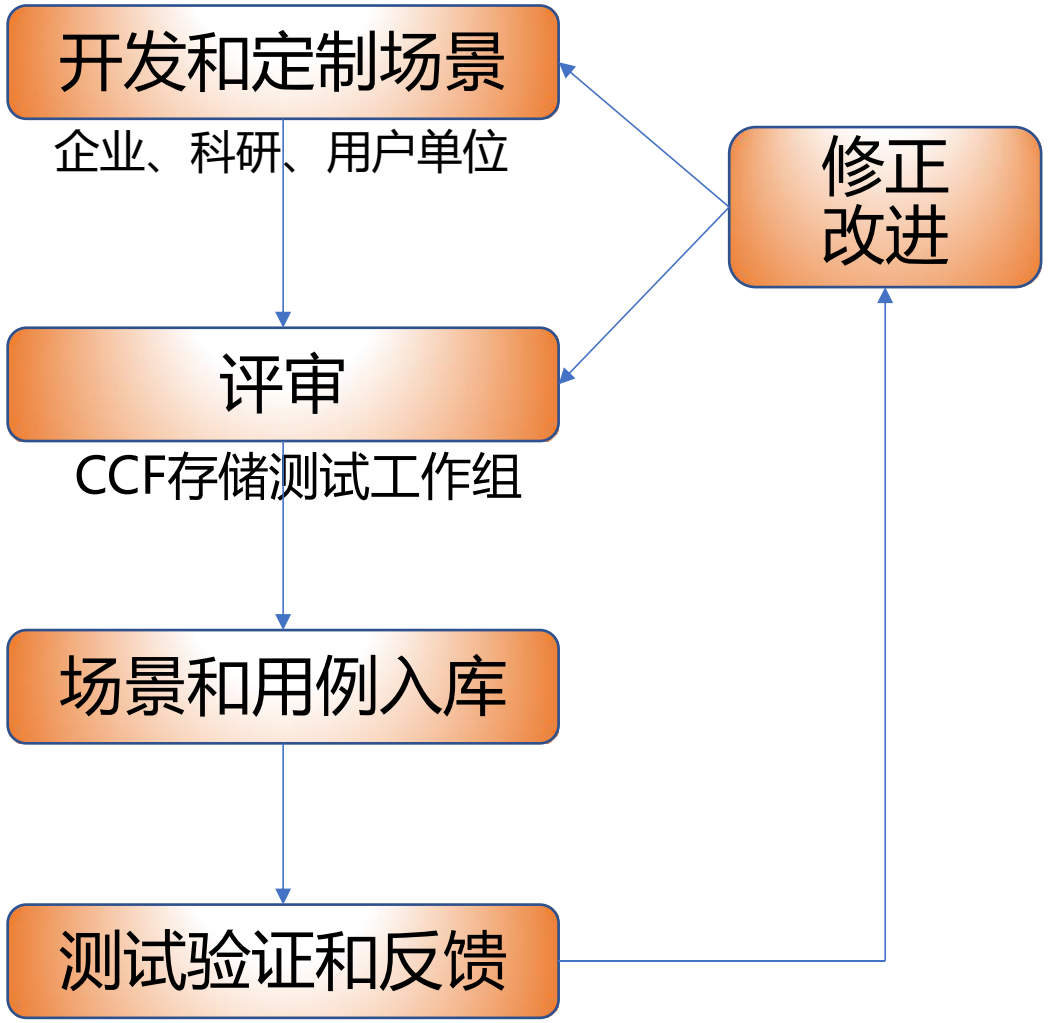
技术层面

1. 对软硬件环境和存储设备给出客观、公平的评价
2. 推动存储、文件系统产品的功能演进、性能提升
3. 为面向各类业务运行场景的设备选型和适配提供依据
4. 为业务系统架构和方案设计提供参考
5. 促进和引导新技术新方法的集成和开发

业务层面

面向场景的存储测评：组织实施

集合企业、科研院所、用户单位力量、定制开发丰富场景
构建开放、可扩展的存储测评社区



形成《xx场景存储测试操作手册》文档

形成《xx场景存储测试指引》文档

特征分析：

实采、公开论文、公开trace等

用例设计：

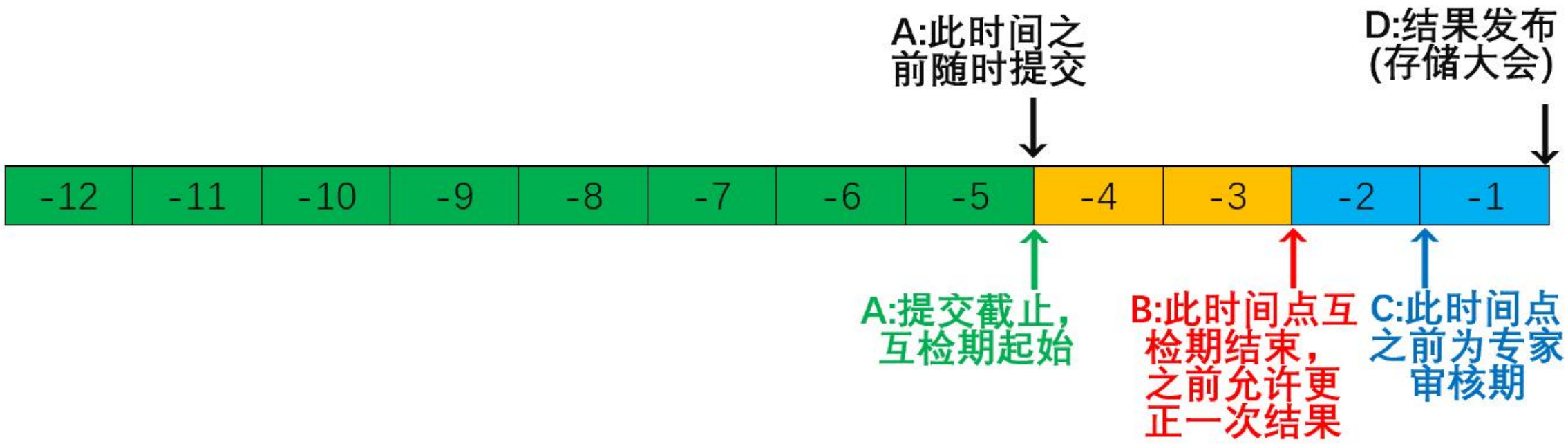
负载模型、参数化、读写比例、
强度、操作频率、测试数据特征

约束条件：

测试场景、测试流程、测试规范、
执行约束、结果生成与收集等

面向场景的存储测评：组织实施

集合企业、科研院所、用户单位力量、定制开发丰富场景
构建开放、可扩展的存储测评社区



- 提交结果的时间为上次发布结束后开始期，至下一次发布前4个月截止
- -4到-3两个月为互相检查期，此期间允许且仅允许更正一次结果
- -2到-1月为互检期和专家审核期，其中-1月内整理结果、分析结果



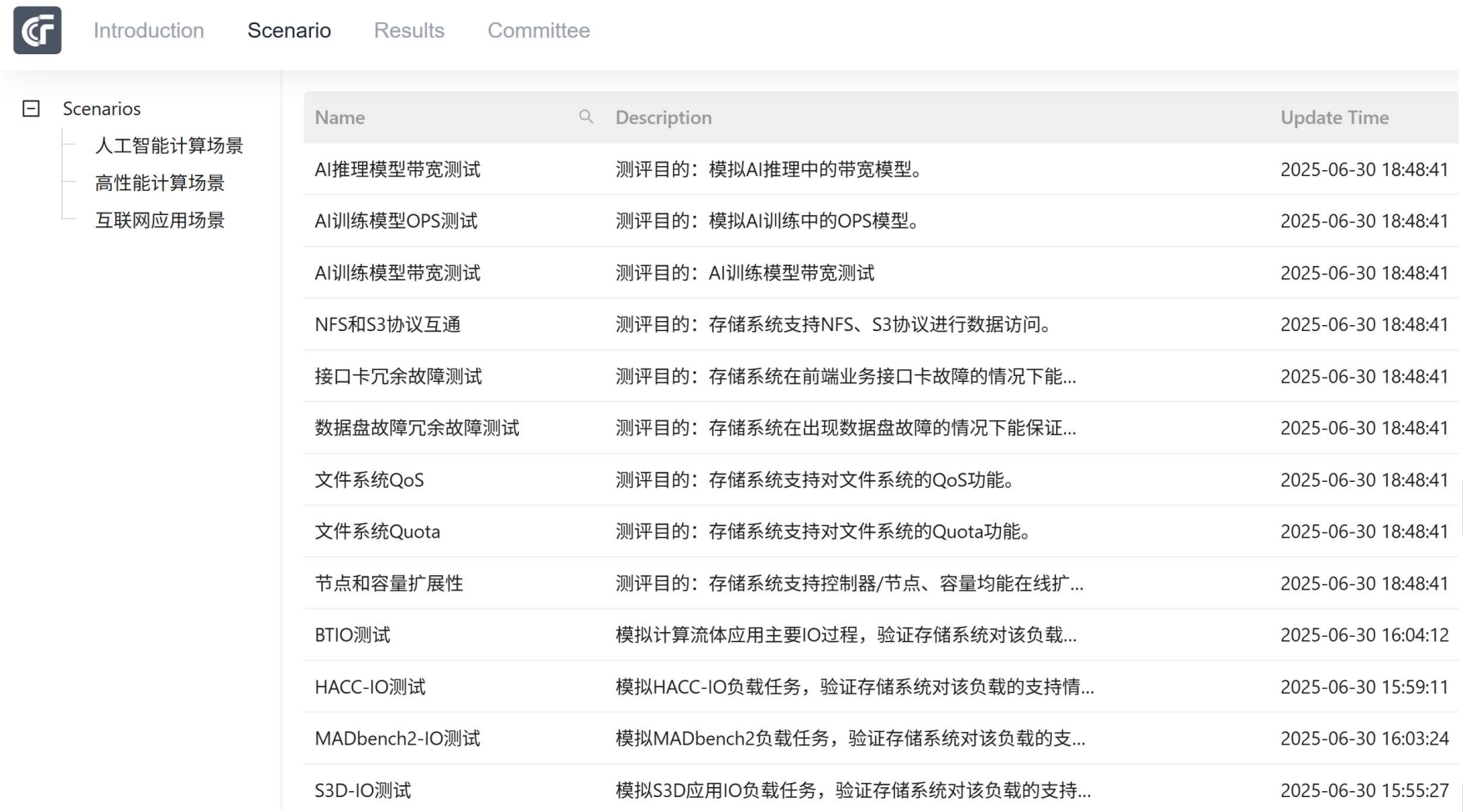
背景介绍

关于CCF专业存储评测

2025年度工作介绍

2025年度测评结果

年度工作: 评测规范、指引、工具



- 人工智能场景
  - 训练
  - 推理
- 高性能计算场景
  - S3D
  - HACC-IO
  - MADbench2
  - BTIO
- 互联网场景
  - 云盘
  - 相册

https://8.145.32.66/wp/apps/eweb/ui/

年度工作: 评测规范、指引、工具

CCF存储测评技术交流研讨





人工智能(推理)

负载模型：

序号	负载	比例
1	128KB 顺序写	10%
2	64KB 顺序写	10%
3	128KB 顺序读	20%
4	64KB 顺序读	20%
5	128KB 随机读	20%
6	64KB 随机读	20%

测量指标：

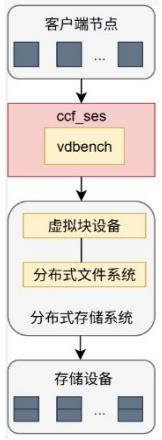
- 单节点带宽
- 系统总带宽

高性能计算场景

- 材料科学S3DIO
- 多体引力模拟 HACC-IO
- 宇宙微波背景辐射MADBench2
- CFD仿真

互联网应用场景

- 云盘
- 相册



年度工作: 评测规范、指引、工具-人工智能场景

人工智能训练

BW

文件模型

- 目录深度：5
- 目录宽度：用户设置最底层目录文件分布如下表
- 调整第一层宽度和下层目录宽度达到数据量不低于预置数据

大小	比例%	
		分配到2目录，360个/目录，目录内占比：
64K	60	67%
128K	30	33%
		分配到2目录，40个/目录，目录内占比：
1M	5	50%
2M	2	20%
8M	1	8%
128M	1	10%
1G	1	10%

负载模型

- 设置IO节点数量和每个IO节点上运行测试线程的并发数产生合适负载
- 测量系统总带宽和单节点带宽
- 运行时间10分钟

负载	比例
32KB顺序读	20%
getattr(32KB)	20%
12KB顺序读	10%
12KB顺序写	10%
64KB顺序读	10%
64KB顺序写	10%
8KB顺序读	10%
8KB顺序写	10%

年度工作: 评测规范、指引、工具-人工智能场景

人工智能训练

OPS

文件模型

- 目录深度：5
- 目录宽度：用户设置最底层目录文件分布如下表
- 调整第一层宽度和下层目录宽度达到数据量不低于预置数据

大小	比例%	
		分配到2目录，360个/目录，目录内占比：
64K	60	67%
128K	30	33%
		分配到2目录，40个/目录，目录内占比：
1M	5	50%
2M	2	20%
8M	1	8%
128M	1	10%
1G	1	10%

负载模型

- 设置IO节点数量和每个IO节点上运行测试线程的并发数产生合适负载
- 通过读写+元数据混合操作测量系统总OPS和单节点OPS

负载	比例
32KB顺序读	20%
getattr(32KB)	20%
12KB顺序读	10%
12KB顺序写	10%
64KB顺序读	10%
64KB顺序写	10%
8KB顺序读	10%
8KB顺序写	10%

年度工作: 评测规范、指引、工具-人工智能场景

人工智能推理

BW

文件模型

- 目录深度：4
- 目录宽度：用户设置各层目录宽度，最底层目录文件分布如下表
- 调整第一层宽度和下层目录宽度达到数据量不低于预置数据

大小	比例%	
		10个/目录
512M	30	
1G	40	
2G	30	

负载模型

- 设置IO节点数量和每个IO节点上运行测试线程的并发数产生合适负载
- 测量系统总带宽和单节点带宽
- 运行时间10分钟

负载	比例
128KB顺序写	10%
64KB顺序写	10%
128KB顺序读	20%
64KB顺序读	20%
128KB随机读	20%
64KB随机读	20%

年度工作: 评测规范、指引、工具-人工智能场景

预埋数据量说明

- **目录总数 * 单目录容量 = 宽度（深度-1） * 第一层目录宽度 * 单目录容量**

- 以训练模型为例
- 根目录下为5层目录
- 一个fsd_group中2个大小文件目录400个文件
- 按照不同大小文件的比例计算得总容量为4706MB
- 第一层目录宽度为fsd_group_number * 2
- 第二到第五层宽度为fsd_width
- 则预埋数据量: fsd_width (5-1)* (fs_group_number*2) * (4706/2) /1024/1024 TB

尽可能复现实际工况:

- 存储中存放了60%-70%的数据
- 有多个作业/任务在同时训练
- 多个研究人员在使用机器
- **训练、推理测试、数据治理、上传、下载、打包、查找**……

年度工作: 评测规范、指引、工具-高性能计算场景

由中山大学开发

S3D-IO

第一原理计算
大量非连续数据写入，借助MPI-IO聚合操作共享文件
四个全局double数组（分别表示质量、速度、压力和温度）
质量、速度是四维double数组
压力和温度是三维double数组

数据规模	网格规模	迭代次数/文件数量	内存需求	文件总数据量（单个文件大小）
tiny	800*800*800	20	64GiB	1.2TiB(61.04GiB)
small	1200*1200*1200	20	256GiB	4.0TiB(206GiB)
large	1600*1600*1600	20	512GiB	9.5TiB(488.28 GiB)
huge	2000*2000*2000	20	1024GiB	18.6TiB(953.67 GiB)
gigantic	2400*2400*2400	20	1800GiB	34TiB(1647.95 GiB)

HACC-IO

N-Body问题
使用MPI-IO的独立模式，非集合模式
顺序读写为主，不涉及复杂的随机访问
数据量为：24MB文件头+2×38×N体×P进程数

数据规模	总粒子规模(G)	迭代次数/文件数量	总内存需求	文件总数据量（单个文件大小）
tiny	1	10	76GiB	380GiB（38GiB）
small	2	20	152GiB	1.5TiB（76GiB）
large	8	40	608GiB	12TiB（304GiB）
huge	16	40	1216GiB	24TiB（608GiB）
gigantic	16	80	1216GiB	48TiB（608GiB）

年度工作: 评测规范、指引、工具-高性能计算场景

由中山大学开发

MADbenche2

宇宙微波背景数据分析
伪数据的大小no_pix，决定矩阵维度，矩阵大小为no_pix * no_pix。角功率谱的分箱数量no_bin。no_bin必须为完全平方数
单进程数据读写总大小为no_pix*no_pix*no_bin*sizeof(double)
文件写入方式可选shared和unique

数据规模	网格规模	迭代次数	总内存需求	文件总数据量
tiny	408*408*408	200	2.5GiB	506GiB
small	1020*1020*1020	100	42GiB	3.9TiB
large	1280*1280*1280	200	83GiB	15TiB
huge	2000*2000*2000	100	320GiB	29TiB
gigantic	2560*2560*2560	100	671GiB	62TiB

尽可能复现实际工况：
- **共享文件、大规模并行访问**

BTIO

流体力学计算
输出四维网格，其中前3维与问题规模相关，第4维是解向量，分量分别为：密度、动量x/y/z和总能量，数据量计算为x*y*z*5*8*i
网格数据输出每次都写入同一个文件，其模式为大文件并发读写，数据量由网格数和迭代次数决定

模	网格规模	迭代次数	总内存需求	文件总数据量
tiny	408*408*408	200	2.5GiB	506GiB
small	1020*1020*1020	100	42GiB	3.9TiB
large	1280*1280*1280	200	83GiB	15TiB
huge	2000*2000*2000	100	320GiB	29TiB
gigantic	2560*2560*2560	100	671GiB	62TiB

年度工作: 评测规范、指引、工具-互联网应用计算场景

云盘

模拟云块存储系统架构
基于分布式文件系统模拟块设备
测试数据量取存储系统内存总和的5倍、主机客户端物理内存总和的2倍、存储池可用用量10%中3个数据的最大值

大小	比例%	读(顺序)%	写(顺序)%
11K	35	36	9
36K	15	15	4
17K	13	9	2
73K	13	8	2
56K	8	5	1
182K	6	3	1
4K	10	4	1

相册

由华中科技大学开发

参考开源trace数据，模拟社交媒体图片服务下的工况
混合执行各规格图片的读写负载，并将访问集中在少数热点文件上
测试该场景下存储系统的通常性能表现

读	
大小	比例%
4K	
8K	
16K	
32K	
128K	

写	
大小	比例%
4K	
8K	
16K	
32K	
64K	
1M	

尽可能复现实际工况：
- **大规模小文件、高并发访问**



背景介绍

关于CCF专业存储评测

2025年度工作介绍

2025年度测评结果

年度测试结果(AI-2025)

厂商	机型	台套	磁盘类型	容量（总）TB*数量	接口类型数量(单节点)	接口带宽(Gbps)	性能-ops(PERF_001)	训练性能-带宽(GB/s)（PERF_002）	推理性能-带宽（GB/s）（PERF_003）	可靠性(DFX_001\2)	扩展性(EXP_001)
天津中存超为	LS-AI4K	2	NVME U.2双端口SSD	7.68*24	IB*4	100	587919	54.21	45.3	PASS	PASS
北京容数	RDXE4000	2	NVMe SSD	3.94*17	IB*6	200	149078	26.74	Skip	PASS	PASS
无锡易存	WS8000-S	2	NVMe SSD	7.68*24	IB*8	100	220356	48.29	Skip	PASS	Skip
北京极客天成	极客追光AI存储全闪一体机AI400X2	2	NVMe SSD	7.68*30	GE*2	400	303923	26.44	23.12	PASS	PASS
济南浪潮	AS13000G7-224	10	NVMe SSD	7.68*120	IB*4	400	7300743	1569.19	1438.5	PASS	PASS
新华三	UniStor Polaris X20828	3	NVMe SSD	3.5*72	GE*8	200	5010386	283.63	215.27	PASS	PASS

年度测试结果(HPC-2025)

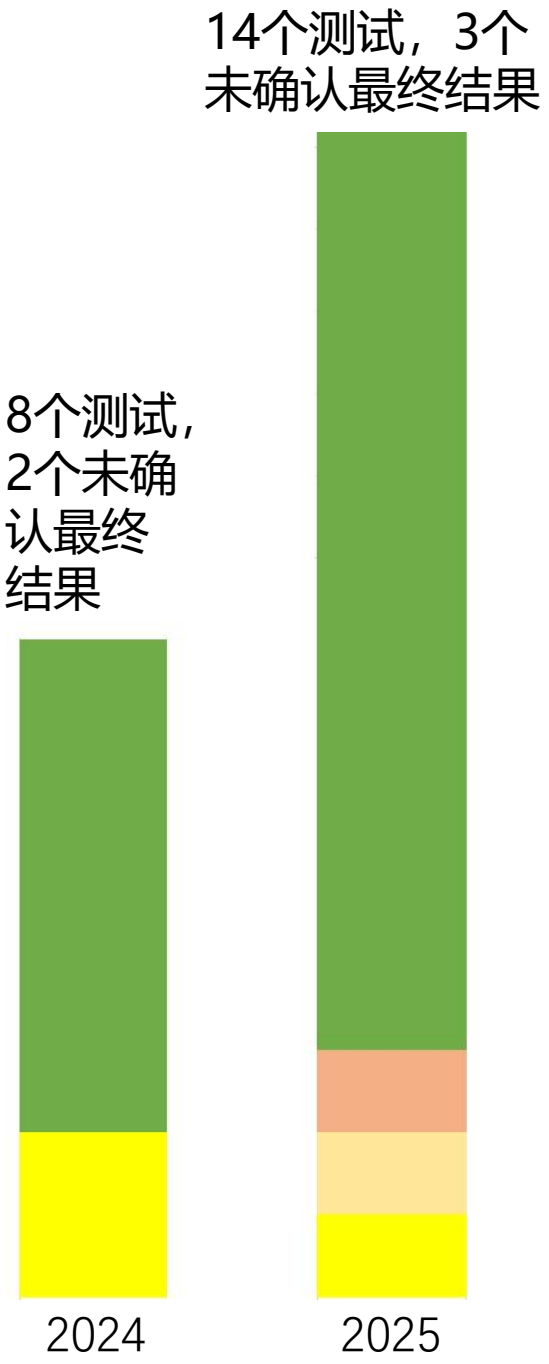
厂商	机型	台套	磁盘类型	容量（总）TB*数量	接口类型数量	接口带宽(Gbps)	Tiny 读/写(GB/s)	Small 读/写(GB/s)	Large 读/写(GB/s)	Huge 读/写(GB/s)	Gigantic 读/写(GB/s)
北京容数	RDXE4000	2	NVMe SSD	3.94*17	IB*6	200	9.76/7.41	3.39/4.40	2.10/6.85	Skip	Skip
无锡易存	WS8000-S	2	NVMe SSD	7.68*24	IB*8	100	2.13/1.66	Skip	Skip	Skip	Skip
浪潮	AS13000G7-224	10	NVMe SSD	7.68*120	IB*4	400	211.68/94.55	248.97/180.23	263.93/213.66	Skip	Skip

年度测试结果(互联网应用-2025)

厂商	机型	台套	磁盘类型	容量（总）TB*数量	接口类型数量(单节点)	接口带宽(Gbps)	用例	平均延迟	性能-ops	性能-带宽(GB/s)	可靠性	扩展性
曙光信息	ParaStor F9000	3	NVMe SSD	14.0*24	GE*4	400	云计算	0.388	129.70	84.87	Skip	Skip
济南浪潮	AS13000G7-224	4	NVMe SSD	3.84*48	IB*4	400	社交媒体图片服务	0.377	1071.86	217.85	PASS	PASS

总结

- 主流设备基本都使用NVMe磁盘了
- 面向场景测试的用例设置较好地反映了业务运行的真实工况：
  - 同一设备在不同场景下的表现出较大差异
  - 理想峰值和理论峰值达成率一定程度上反应存储的实际功效
  - 组合用例测试能较好模拟存储设备的实际工况
- 可靠性测试为模拟磁盘掉线和接口卡掉线，最大程度还原真实运行场景
- 绝对值和相对值基本呈正相关，有些性能指标跟所选设备配置和规模有一定相关性
- 测试仅发布基本指标，派生指标可包含更丰富信息，如：单盘性能达成率，单端口性能达成率，单盘/单设备性能，叠加可靠性、扩展性测试后的综合性能等



当前进展: 下一步工作规划

- 规范化场景开发，持续完善测试规范框架
- 构建测评场景库和用例库(存储测评社区)
- 开发新场景：事务处理(数据库），大数据(hdfs/noSQL)
- 设计适配能力好的测评工具集和框架
- 动员更多、更广泛单位参与；国产存储测试
- 推广应用

THANK YOU

存储测试工作委员会：

肖侬　武永卫　周可　陈志广　张武生　陈俭喜
董勇　张晓　曾令仿　任峰　武林平　杨斌　程耀东

存储测试工作组：

肖侬　武永卫　周可　陈志广　张武生　陈俭喜
董勇　张晓　曾令仿　任峰　武林平　杨斌　程耀东

工程支持：

郭一华　李学龙　尚攸仁　郑谦