DTCC 2010-2022

第十三届中国数据库技术大会

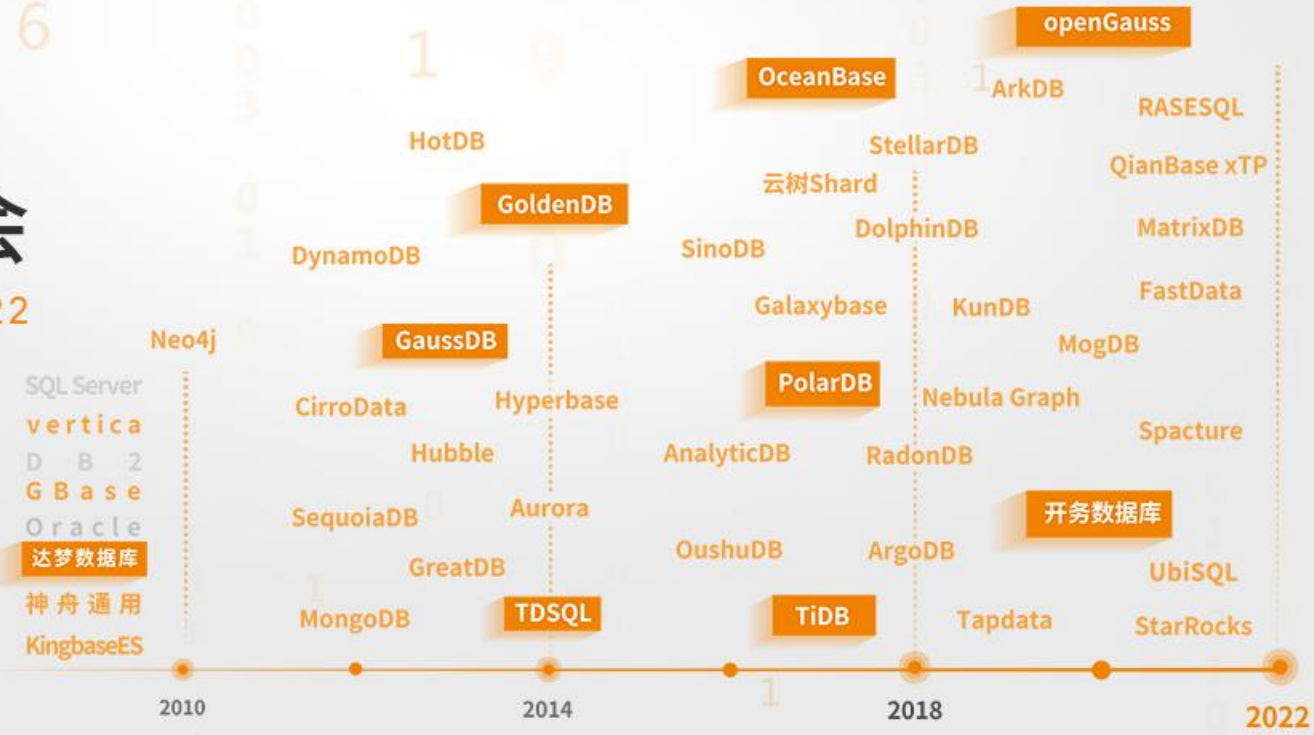
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



百亿级分布式文件系统 FastCFS架构与实现

余 庆　FastDFS & FastCFS创始人

自我介绍

- 分布式文件系统 FastDFS & FastCFS 作者
- 曾任职于新浪、雅虎中国和阿里巴巴
- 对分布式架构和高性能编程有着深入的研究和丰富的实践经验

为什么要研发FastCFS？

- 几款开源分布式文件系统：GlusterFS、MooseFS、Ceph
- 缺乏一款好用的DFS
- 数据库云化是趋势

数据库存储面临的挑战

- 数据一致性
- 系统可用性
- IO性能

数据库对分布式存储要求

- 硬盘好
- 网络好
- 软件好

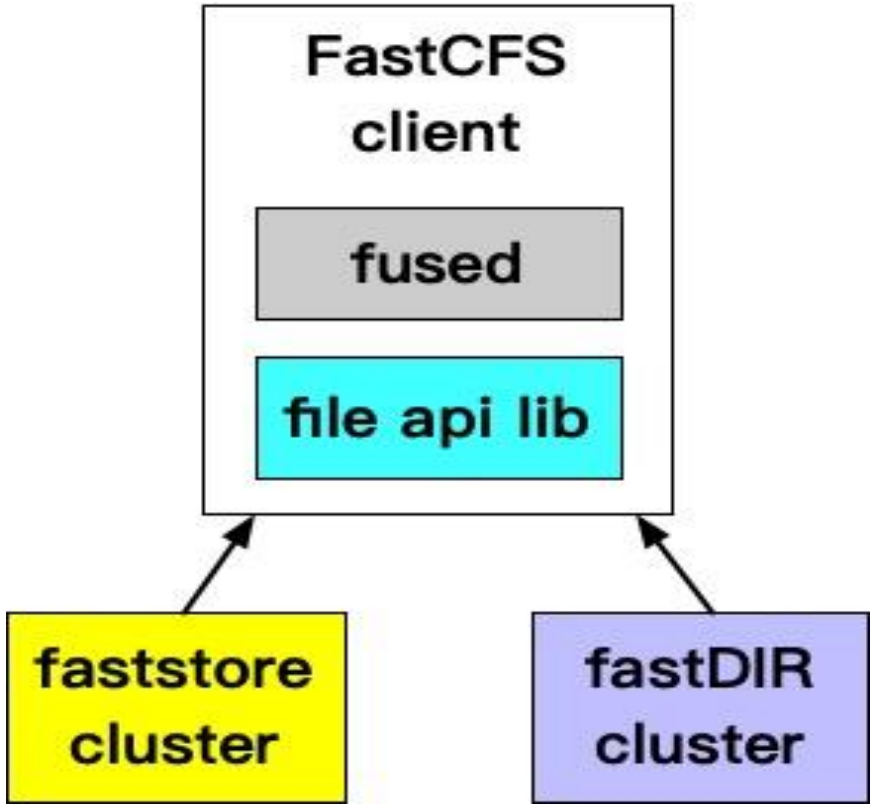
FastCFS的定位

FastCFS 是一款强一致性、高性能、高可用、支持百亿级海量文件的通用分布式文件系统，可以作为MySQL、PostgresSQL、Oracle等数据库，k8s，KVM，FTP，SMB和NFS等系统的后端存储。

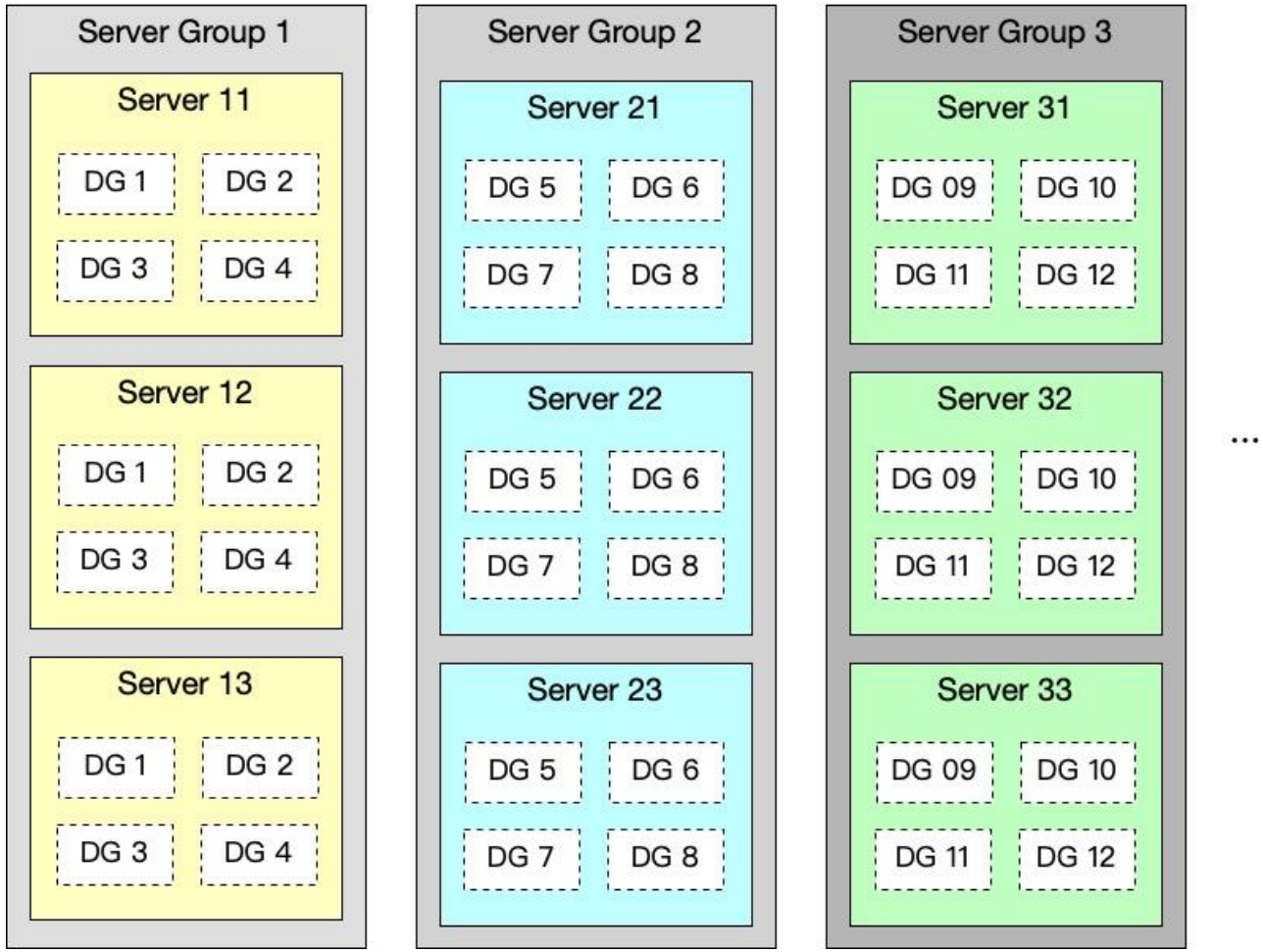
FastCFS版本历史

- V1.0：2020年12月第一个版本
- V2.0：2021年4月支持k8s
- V3.0：2021年12月实现存储插件
- V3.3：2022年4月生产环境可用
- V3.7：2022年11月当前最新版本

FastCFS核心模块



faststore架构



FastCFS架构特点

- 有中心和无中心结合
- 分组方式，简单高效
  - 服务器分组
  - 数据分组
- 对等结构，自动failover

FastCFS软件特点

- 保证数据强一致前提下实现了高性能
- 完全兼容POSIX文件接口，支持文件锁，支持百亿级海量文件
- 高可用：不存在单点，自动failover
- 简洁高效的架构和原生实现，不依赖第三方组件
- 数据写入性能强悍

FastCFS如何做到数据强一致

- 数据版本号
- 集群动态拓扑信息
- 多数派机制，特有的公共选举节点
- 幂等机制

FastDIR如何实现高性能

- 支持命名空间
- 采用跳表（skiplist）
- 数据线程无锁化

FastDIR如何支持百亿级海量文件

- binlog + 存储插件：异步持久化
  - 修改的inode数目达到阈值
  - 超过特定时间间隔
- 按目录结构淘汰
- 按数据线程淘汰

FastCFS性能对比数据（一）

读写方式	并发数	FastCFS 2.2.0	Ceph 13.2.10	比值
顺序写	4	126.0	20.0	630%
	8	216.0	32.7	661%
	16	300.0	45.2	664%
随机写	4	24.9	17.4	143%
	8	44.0	25.0	176%
	16	65.9	27.7	238%
顺序读	4	136.2	58.0	235%
	8	245.1	97.2	252%
	16	337.9	152.0	222%
随机读	4	48.9	47.5	103%
	8	92.2	86.8	106%
	16	163.2	143.0	114%

注：性能指标为bw（吞吐量），单位为MiB。

FastCFS性能对比数据（二）

读写方式	并发数	V3.6	V2.2	比值
顺序写	4	351	126	279%
	8	351	216	163%
	16	347	300	116%
随机写	4	31.9	24.9	128%
	8	57.7	44.0	131%
	16	90.1	65.9	137%
顺序读	4	109	57.7	189%
	8	205	112	183%
	16	374	199	188%
随机读	4	61.6	43.4	142%
	8	114	81.9	139%
	16	211	137	154%

注：性能指标为bw（吞吐量），单位为MiB。

FastCFS性能对比数据（三）

读写方式		FastCFS 吞吐量（MB/s）		MooseFS 吞吐量（MB/s）	FCFS与MFS百分比	
		V3.5	V3.6		V3.5	V3.6
1MB顺序写	队列深度 1	4.61	256.58	152.14	2%	169%
	队列深度 8	4.82	316.43	171.01	2%	185%
1MB顺序读	队列深度 1	8.78	350.05	357.35	3%	98%
	队列深度 8	8.79	448.85	430.85	2%	104%
4KB随机写	队列深度 1	0.15	8.58	5.12	2%	168%
	队列深度 32	0.16	18.98	13.41	1%	142%
4KB随机读	队列深度 1	0.61	11.41	11.68	5%	98%
	队列深度 32	0.64	20.53	20.57	3%	100%

FastCFS如何做到极高性能

- 简洁高效的架构和原生实现
- 内存池、连接池、线程池等
- 客户端读写缓存

FastCFS后续工作计划

- 支持集群在线扩容
- 分级存储 & slice数据合并：支持两级存储（如SSD + HDD）
- S3 、块设备、NBD等接口方式


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS