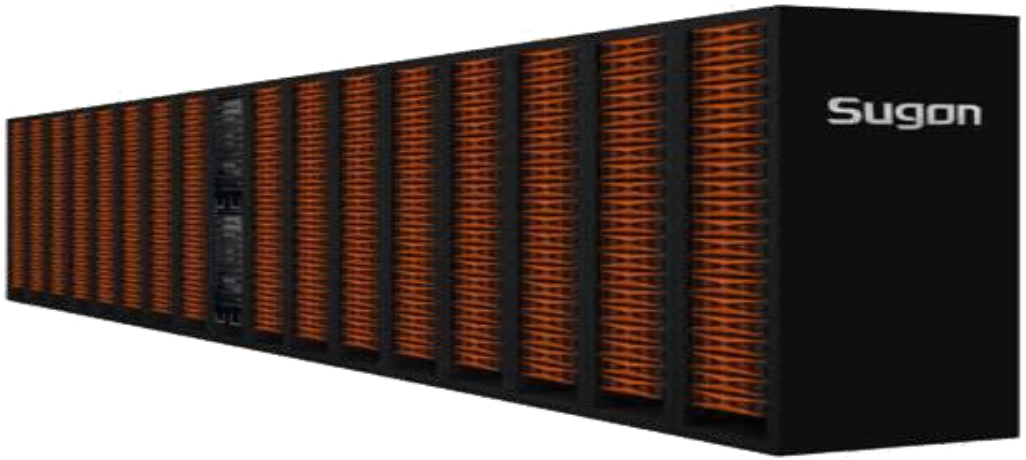
中科曙光
Sugon

ParaStor: 一种海量数据存储系统

王勇
曙光公司
2011年12月

ParaStor：星云存储

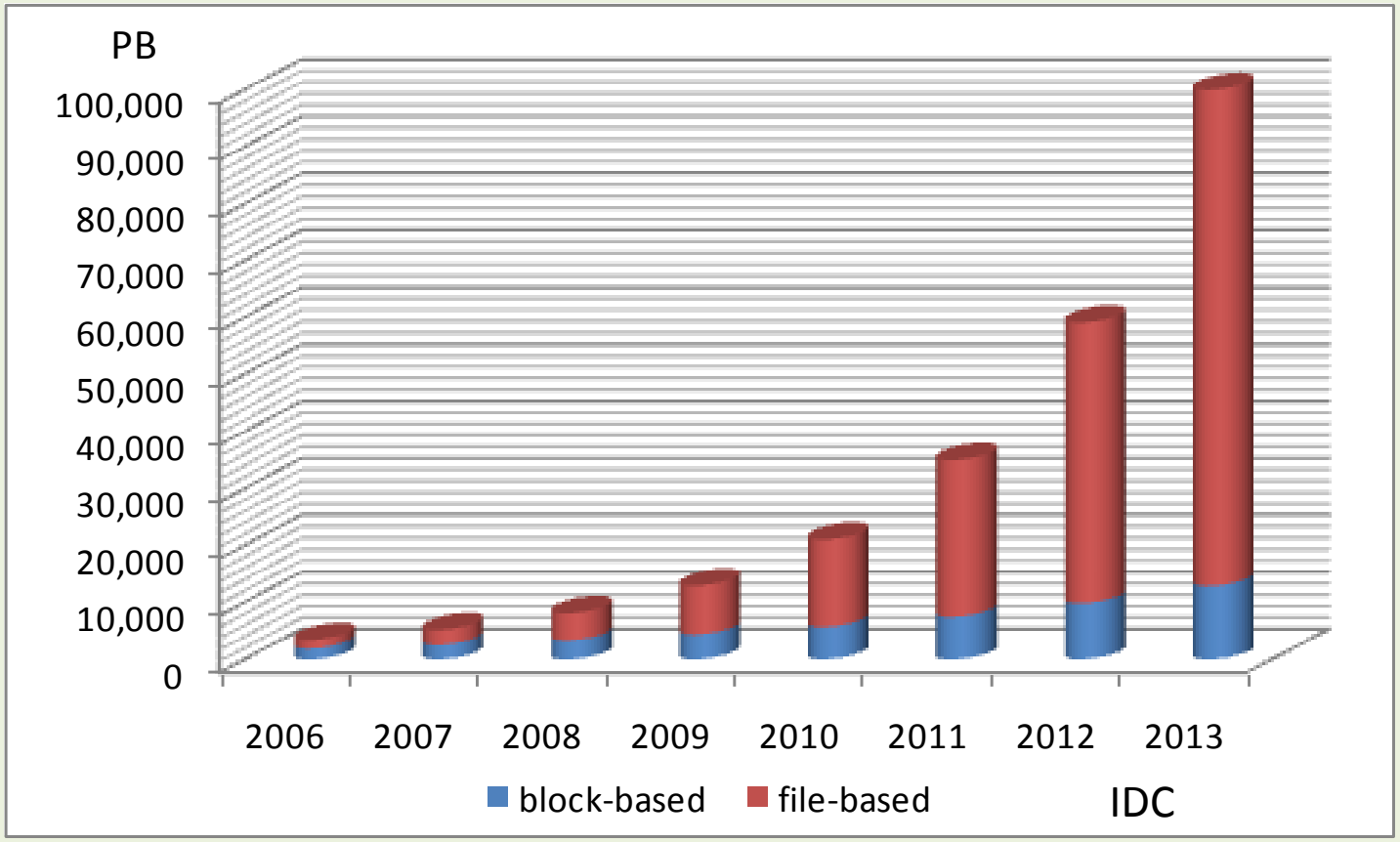
- 存储系统
  - 基于文件的
    - 共享的
      - 海量
- 目标
  - 高可靠
  - 高性能
  - 高容量



目录

海量数据

Web数据　　办公文档　　富媒体　　科学数据



存储问题

- 海量（Volume）
- 多样性（Variety）
- 快速（Velocity）
- 高可靠（Reliability）
- 高并发（Concurrency）

目录

设计理念

商品化部件

网络

备

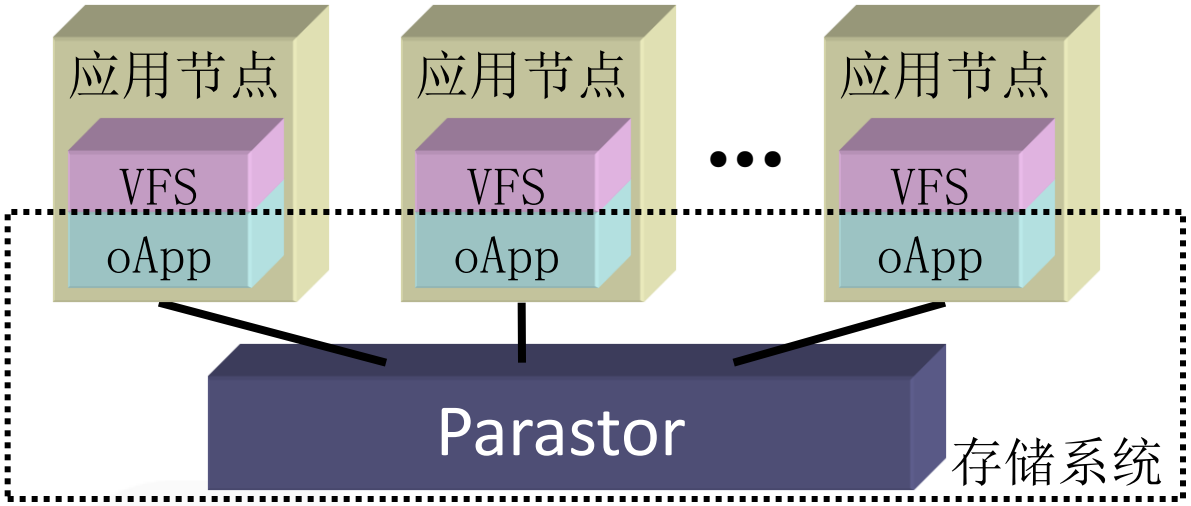
将主要部件失效作为常态考虑

以相对廉价、可靠性不高的工业标准单元，构造高性能、高可用性、低TCO的大规模存储系统

应用程序不需要修改，支持POSIX语义，CIFS

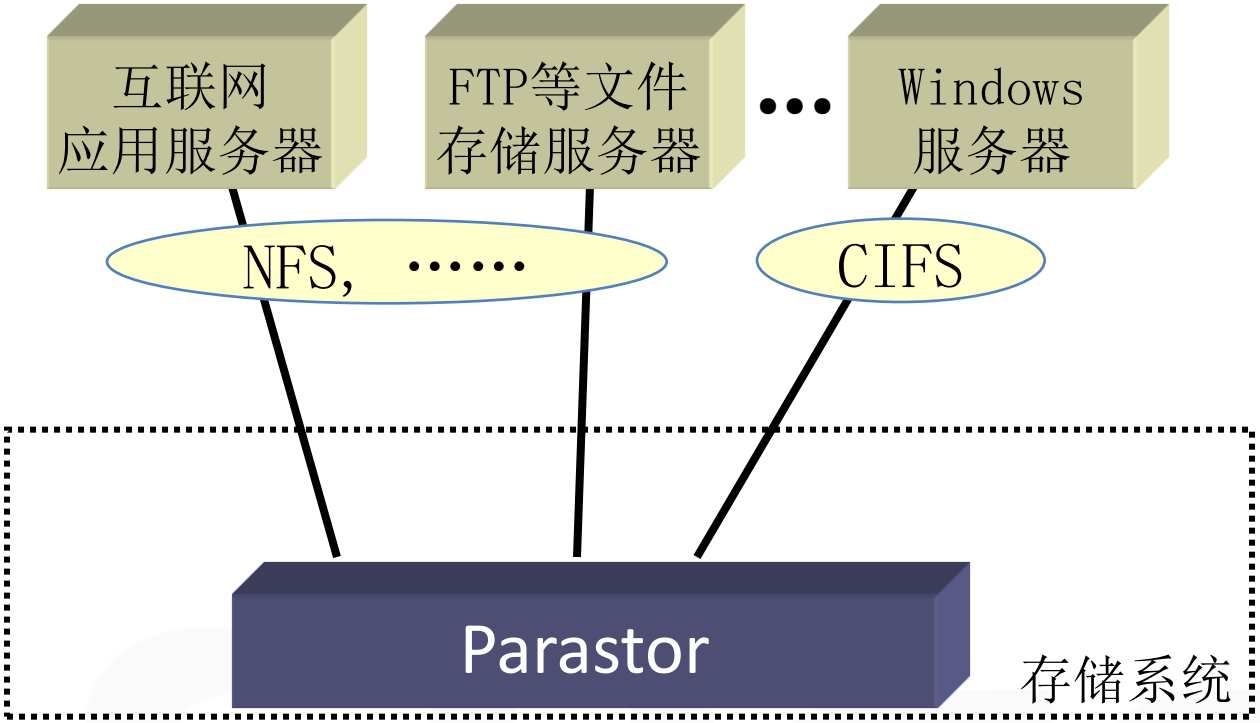
使用方式

私有协议模式

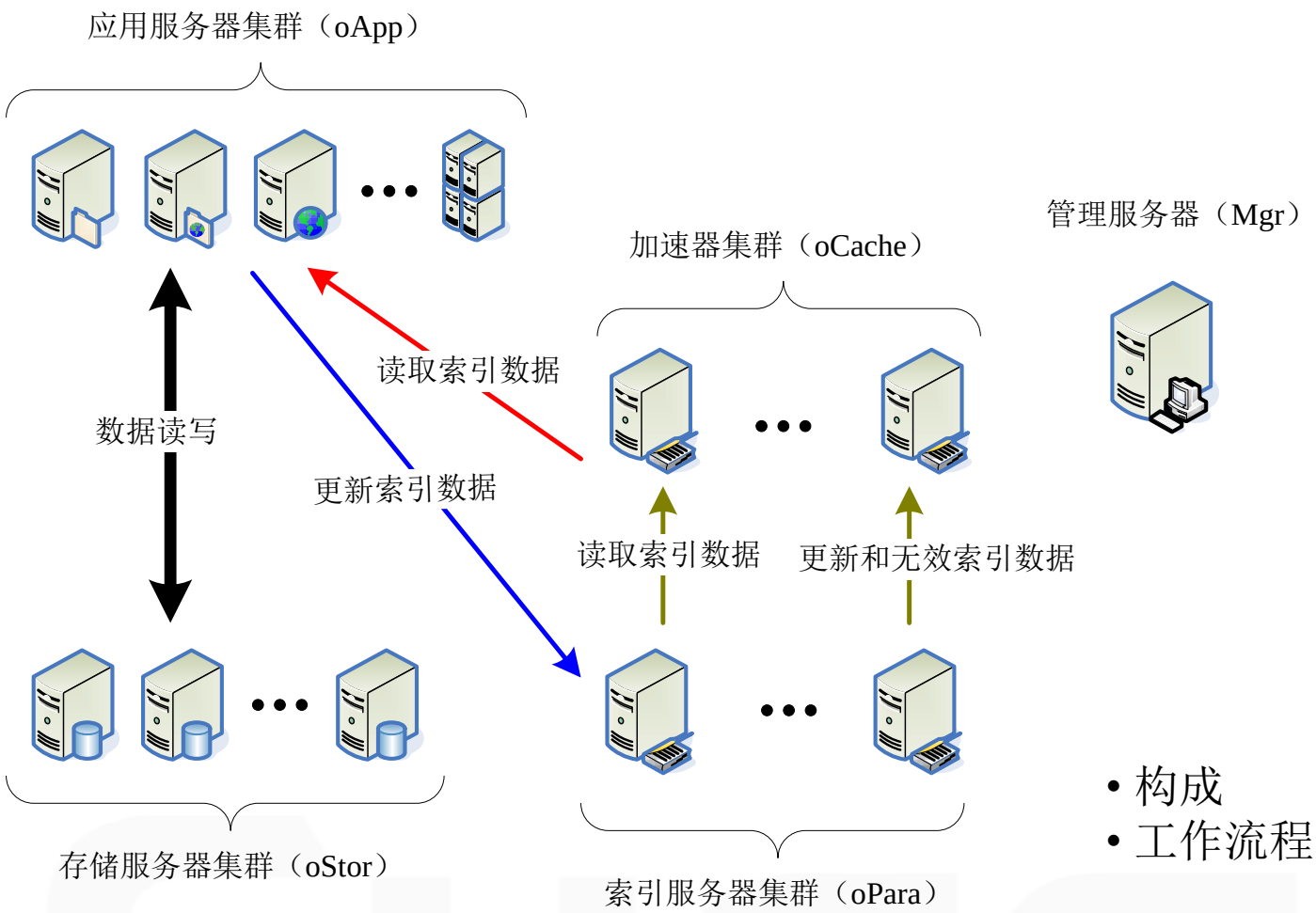


使用方式

NAS方式

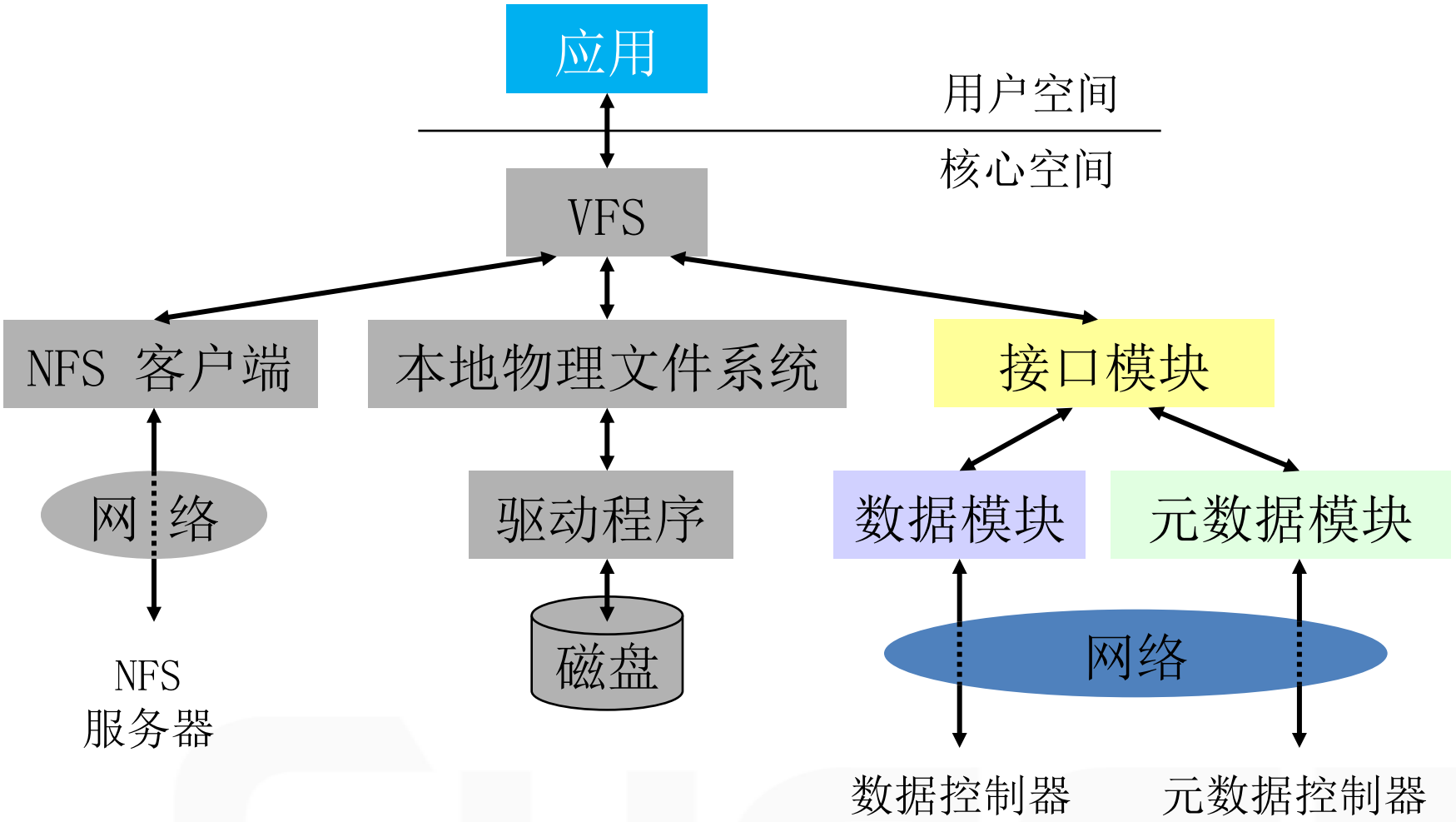


系统结构

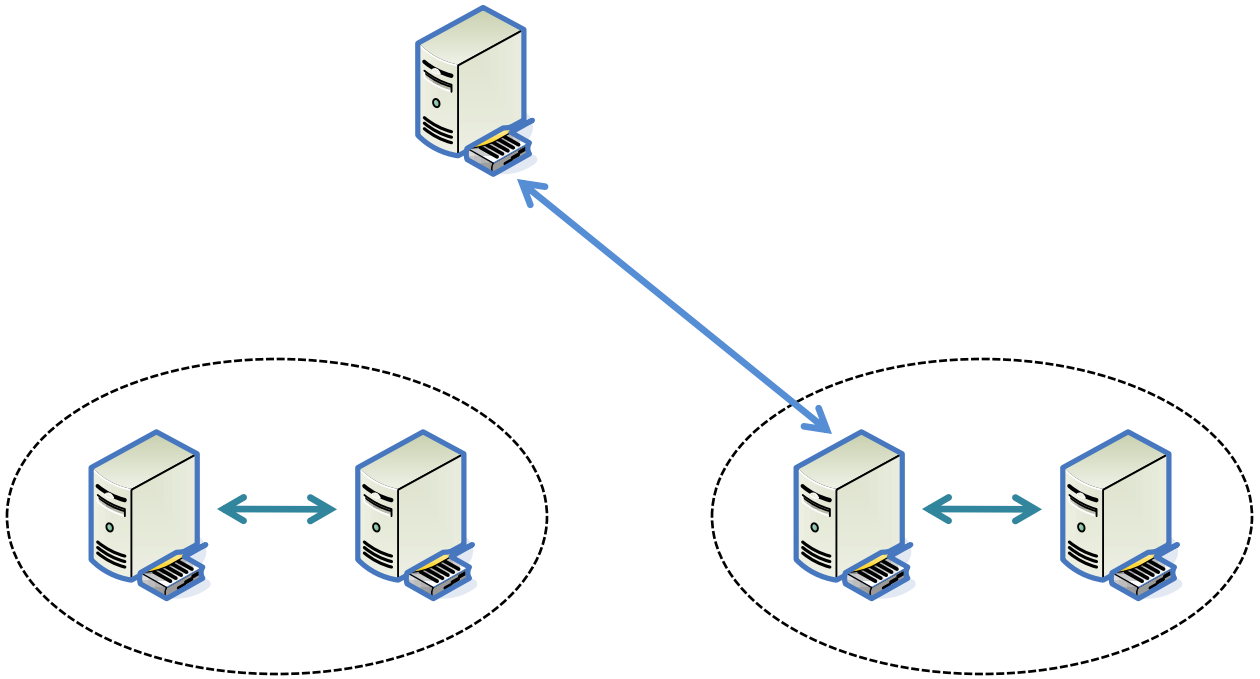

应用服务器集群（oApp）
加速器集群（oCache）
管理服务器（Mgr）
读取索引数据
数据读写
更新索引数据
读取索引数据
更新和无效索引数据
存储服务器集群（oStor）
索引服务器集群（oPara）

- 构成
- 工作流程

应用服务器


应用
用户空间
核心空间
VFS
NFS 客户端
本地物理文件系统
接口模块
网 络
驱动程序
数据模块
元数据模块
磁盘
网络
NFS
服务器
数据控制器
元数据控制器

元数据控制器结构

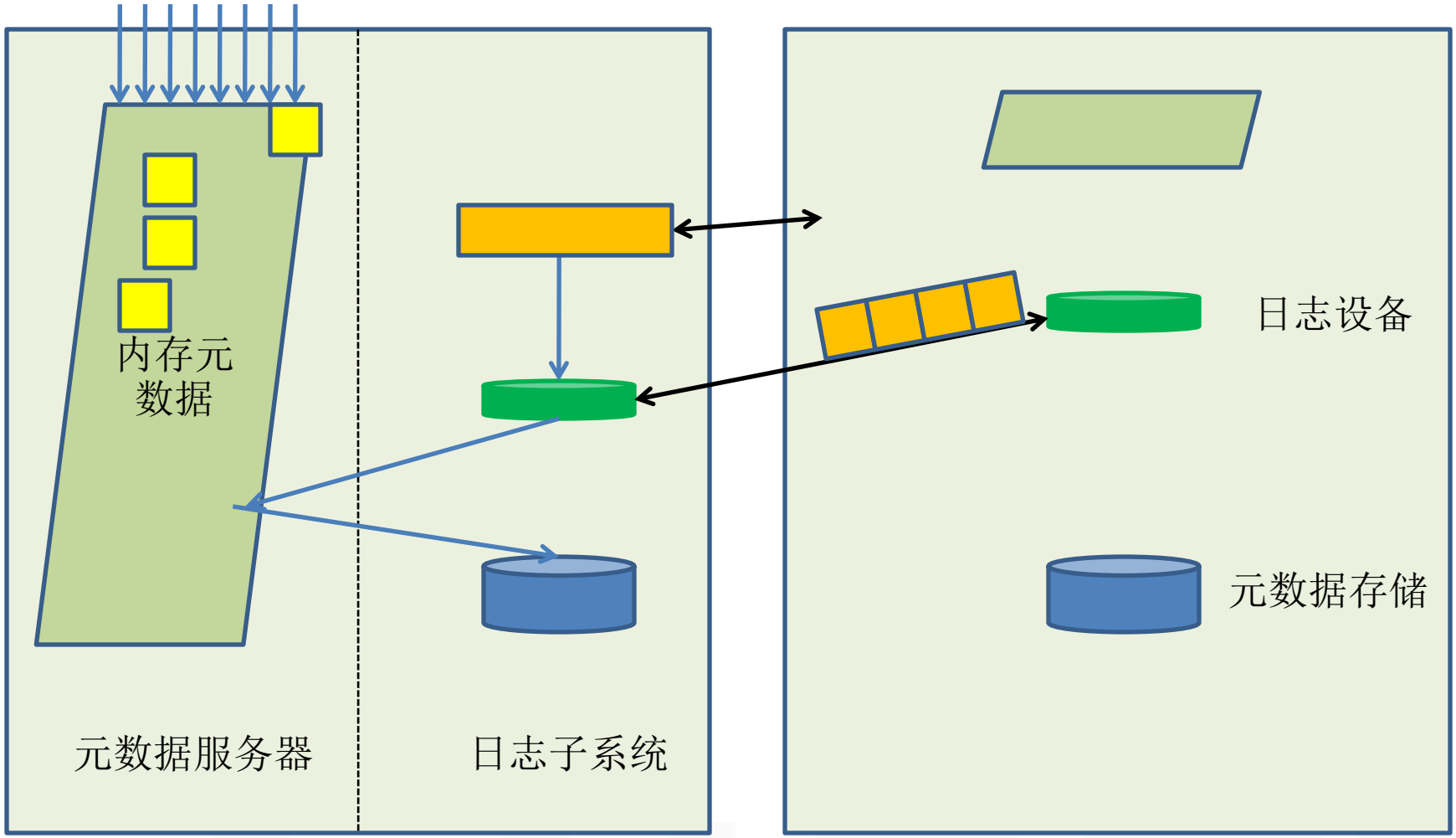


元数据副本

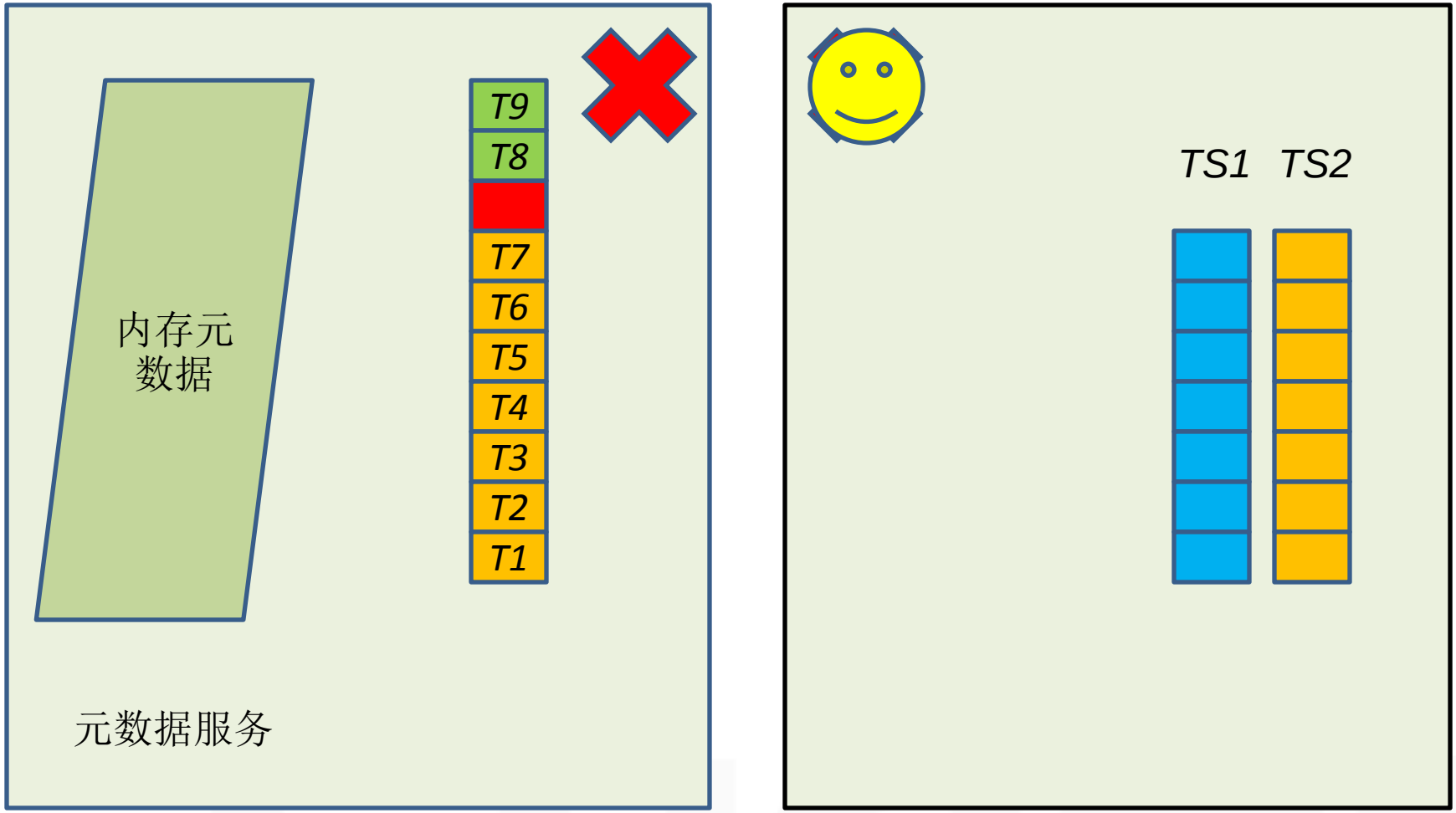
简称	磁盘数据结构		访问	名称
hPara	保存	一个	更新	主元数据控制器
gPara	保存	多个	备用	从元数据控制器
oCache	无		读取	加速服务器

- ✓ 某个元数据控制器只有针对具体的某一个元数据结构而言有具体的意义
- ✓ 同一个物理元数据控制器针对不同的元数据记录可以是不同的身份

元数据日志



元数据日志—恢复



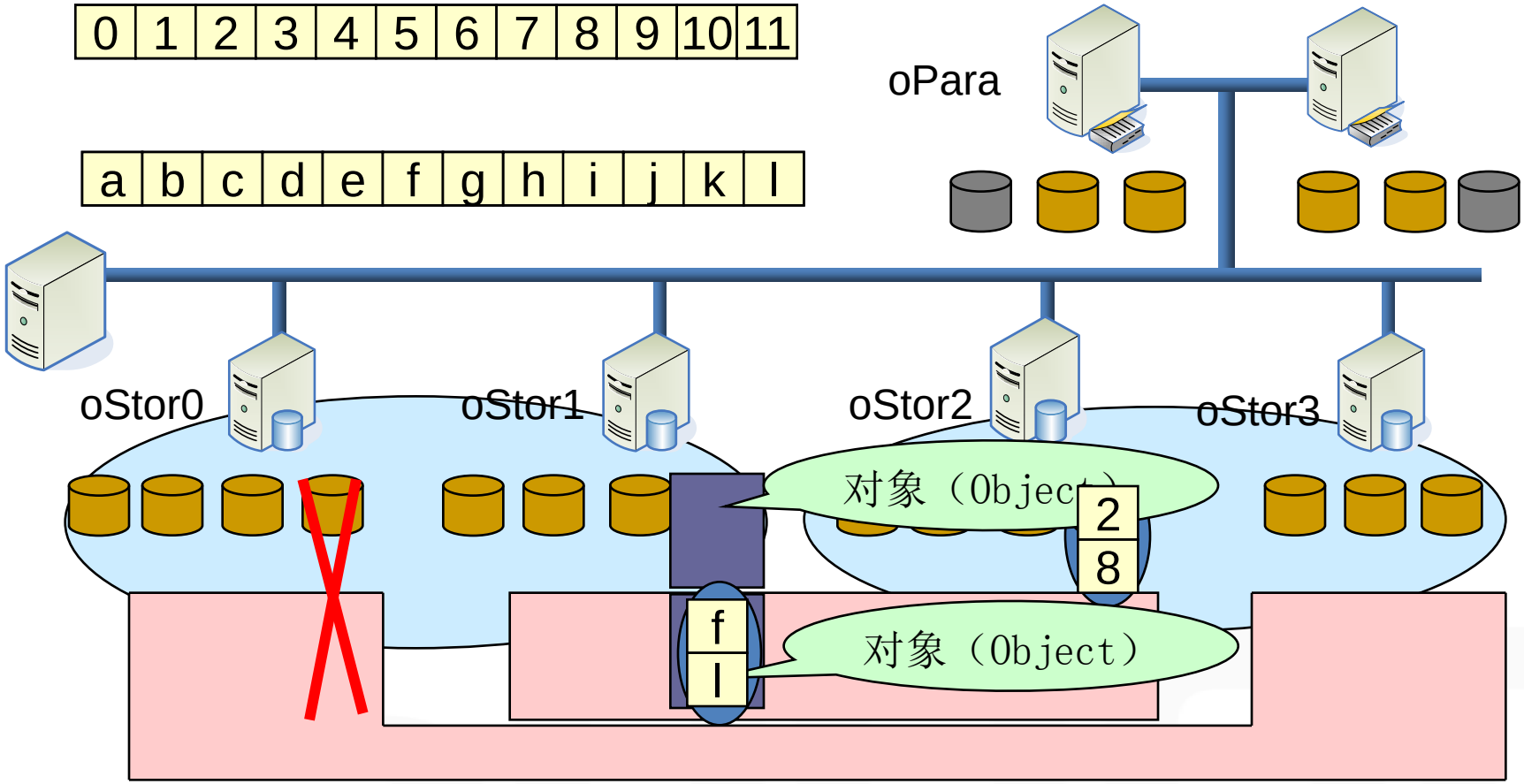
数据控制器

- 用户态
  - 缓存
- 数据分片
- 数据复制
  - 同步

I640r-G

并行数据恢复

重建服务器并行读数据并写入到本地对应磁盘
数据复制完成后，修改对应文件的layout信息



目录

设计经验

- 尽量放在用户态
- 简洁设计
- 灵活、统一的程序结构
- 多线程、无锁设计
- 非阻塞流水
- 关键路径不能有磁盘操作、mutex操作
- 尽量多用Linux自有机制

目录

产品族


ParaStor200I
面向高IOPS应用,满足海量小文件并发随机读写的性能需求
面向高带宽应用，满足视频、测绘等大文件读写的性能需求
ParaStor200W
面向均衡型应用，满足多种应用模式的数据存取需求
ParaStor200B

高性能计算（1）

- 工作流处理
  - 任务由若干个阶段的子任务构成
  - 子任务以中间文件衔接。并发但共享程度低
  - 多种访问模式
    - 随机，顺序
  - 高聚合带宽

高性能计算特点（2）

- 并行IO
  - 多客户端访问同一大文件的不同部分
  - Stride模式
  - 要求高单流带宽

解决方案

- 选择合适的ParaStor系统
  - IB网络
- 预读策略
  - 根据应用设计的预读算法
  - 支持Stride模式
- 大文件采用分片模式
  - 按顺序访问粒度设置分片大小
  - 针对应用特点设定
    - 负载
    - 大文件

视频应用

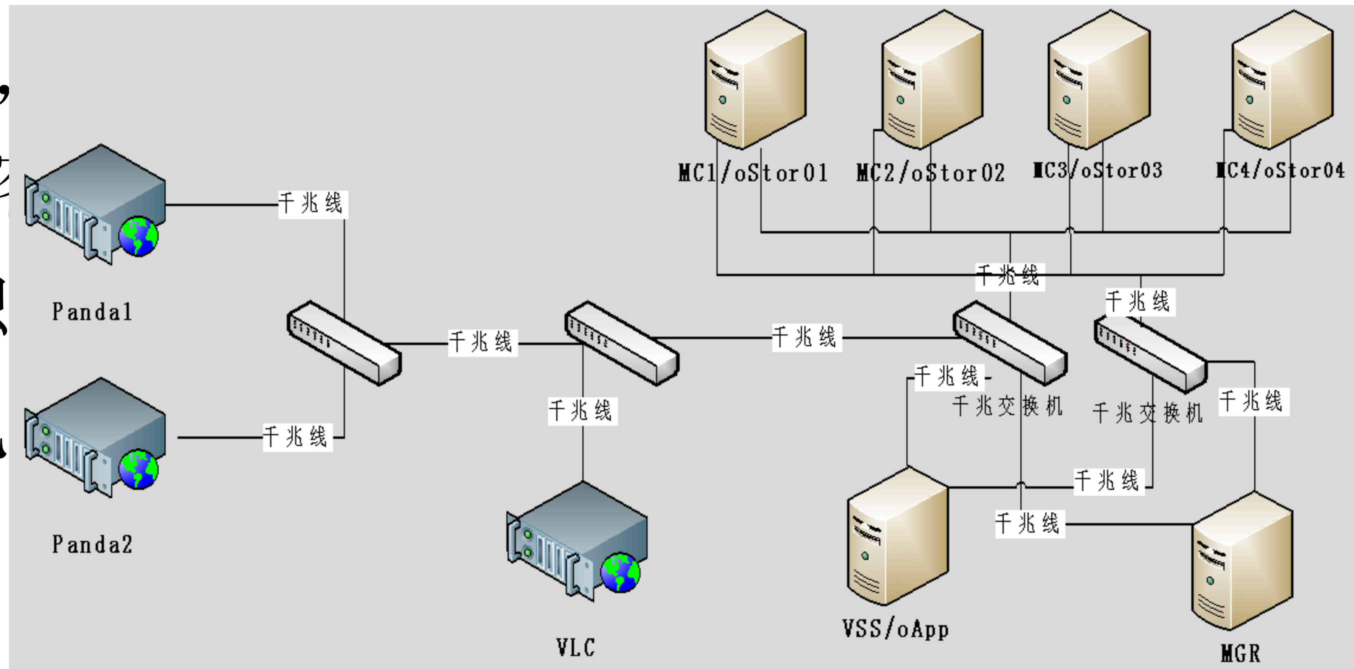
- 应用背景
  - 互动的富媒体
  - 视频分享，VOD，IPTV，节目制作
- 应用需求
  - 并发访问量极高
  - 视频文件数量多、容量大

解决方案

- 互动媒体
  - 同时提供极高带宽和具备灵活可扩展性的大规模存储系统
  - 简化管理
  - 高效文件检索
- 节目制作
  - GB级的聚合带宽
  - 单数据控制器提供超300MB/s的读速度

视频应用－案例

- IPTV
  - 推流，
  - 单个
- 三网融
- 数字电



数字图书馆

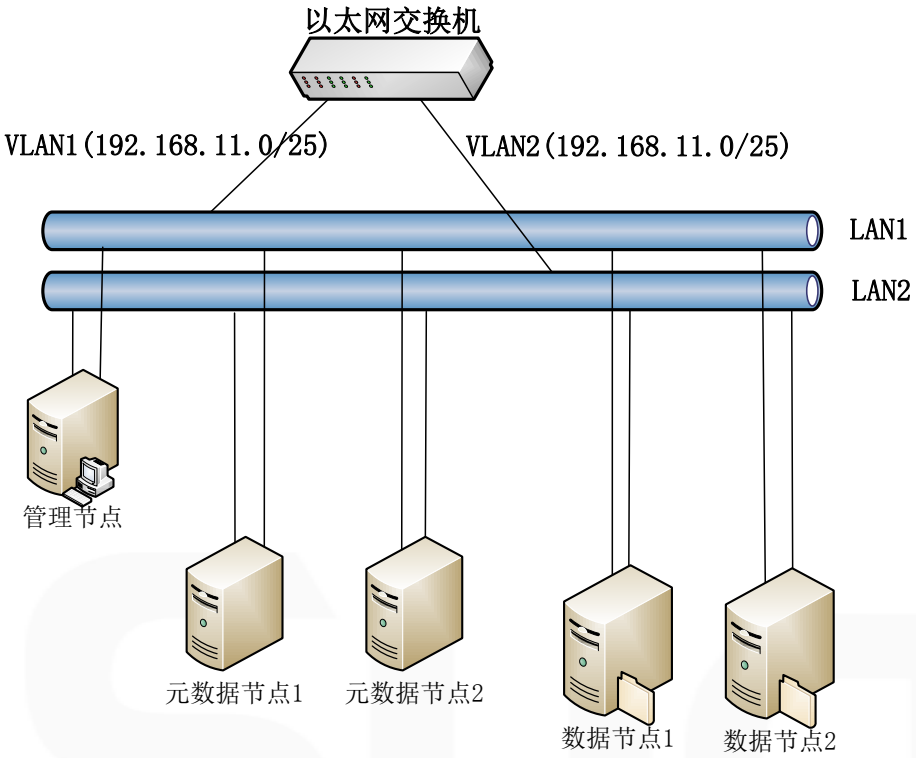
- 应用背景
  - 数字化时代
- 应用需求
  - 百亿级以上的文件数
  - 绝大多数小文件（32KB）
  - 集中高速的文件写入
  - 大文件高带宽

解决方案

- ParaStor-I
- 单一系统轻松管理百亿以上的文件
- 单目录千万级文件，简化目录管理
- 高速小文件写入
- 高速文件检索
- 设置目录分片模式
  - 大文件高带宽

数字图书馆—案例

- 北京某研究院图书馆
  - 创建1TB，32KB的小文件需要10小时



ParaStor-I

Web 2.0应用

- 应用背景
  - 微博、SNS、图片/视频分享、电商
- 挑战
  - 百亿级海量小文件管理
  - 多应用共享存储平台
  - 可用性要求极高
  - 快速扩容

云存储

- 应用背景
  - 高校园区云，空间租赁，远程备份
- 需求
  - 应用种类丰富，数据密集
  - 同时要求高IOPS和高带宽
  - 近乎无限的扩展能力
  - 异地容灾能力

中科曙光
Sugon

谢谢！