澳門城市大學
Universidade da Cidade de Macau
City University of Macau
明德・博學・尚行

云计算原理与实践-第五章 分佈式存儲
Introduction to Cloud Computing(21/22)

Assistant Prof. Andy Lau 劉文堅 | IDS City University of Macau

Outline

- 5.1 分布式存储的基础
- 5.2 文件存储
- 5.3 从单机存储系统到分布式存储系统
- 5.4 实践：分布式存储系统Ceph

Domain expertise

Mathematics

Machine Learning

Data engineering

5.1 分布式存储的基础

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

5.1.1 基本概念

5.1.2 分布式存储分类

5.1.3 分布式存储的发展历史

5.1.1 基本概念

- 分布式存储系统的定义：分布式存储系统是将为数众多的普通计算机或服务器通过网络进行连接，同时对外提供一个整体的存储服务。
- 分布式存储系统包括以下几个特性：
  - 高性能：无论是针对整个集群还是单台服务器，都要求分布式存储系统具备高性能。
  - 可扩展:分布式存储系统可以扩展到几百台甚至几千台的集群规模，而且，随着集群规模的增长，系统整体性能表现为线性增长。
  - 低成本：分布式存储系统的自动容错、自动负载均衡机制使其可以构建在普通PC机之上。另外，线性扩展能力也使得增加、减少机器非常方便，可以实现自动运维。
  - 易用性：分布式存储系统需要能够提供易用的对外接口，另外，也要求具备完善的监控、运维工具，并能够方便地与其他系统集成，例如，从Hadoop云计算系统导入数据。
- 分布式存储系统的技术挑战包括：数据和状态信息的持久化、数据的自动迁移、系统的自动容错、并发读写的数据的一致性等方面。

5.1.2 分布式存储分类

- 分布式存储面临的应用场景和数据需求都比较复杂，根据数据类型，可以将其分为非结构化数据、结构化数据、半结构化数据三类。
  - 非结构化数据：包括所有格式的办公文档、文本、图片、图像、音频和视频信息等。
  - 结构化数据：一般存储在关系数据库中，可以用二维关系表结构来表示。结构化数据的模式和内容是分开的，数据的模式需要预先定义。
  - 半结构化数据：介于非结构化数据和结构化数据之间，HTML文档就属于半结构化数据。
- 正因为数据类型的多样性，不同的分布式存储系统适合处理不同类型的数据，因此可以将分布式存储系统分为四类：
  1. 分布式文件系统
  2. 分布式键值（Key-Value）系统
  3. 分布式表系统
  4. 分布式数据库

1. 分布式文件系统

- 分布式文件系统存储三种类型的数据：Blob对象、定长块以及大文件。

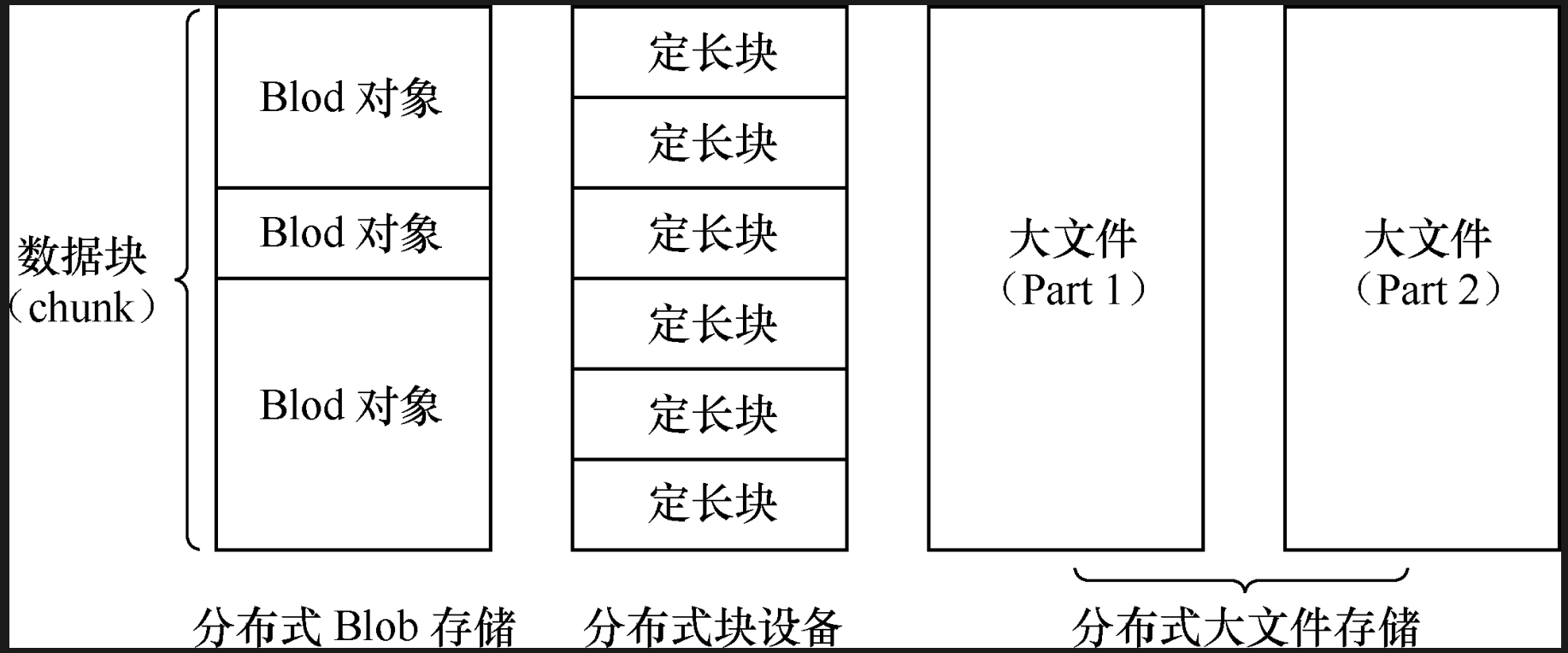


图5.1 数据块与Blob对象、定长块、大文件之间的关系

2 分布式键值（Key-Value）系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

- 分布式键值系统用于存储关系简单的半结构化数据，它提供基于主键的CRUD（Create/Read/ Update/Delete）功能，即根据主键创建、读取、更新或者删除一条键值记录。典型的系统有Amazon Dynamo。
- 分布式键值系统是分布式表系统的一种简化，一般用作缓存，比如Memcache。
- 从数据结构的角度看，分布式键值系统支持将数据分布到集群中的多个存储节点。
- 一致性散列是分布式键值系统中常用的数据分布技术，由于在众多系统中被采用而变得非常有名。

3 分布式表系统

- 分布式表系统主要用于存储半结构化数据。
- 与分布式键值系统相比，分布式表系统不仅仅支持简单的CRUD操作，而且支持扫描某个主键范围。
- 分布式表系统以表格为单位组织数据，每个表格包括很多行，通过主键标识一行，支持根据主键的CRUD功能以及范围查找功能。
- 典型的分布式表系统包括Google Bigtable、Microsoft Azure Table Storage、Amazon DynamoDB等。

4 分布式数据库

- 分布式数据库是从传统的基于单机的关系型数据库扩展而来，用于存储大规模的结构化数据。
- 分布式数据库采用二维表格组织数据，提供经典的SQL关系查询语言，支持嵌套子查询、多表关联等复杂操作，并提供数据库事务以及并发控制。
- 关系数据库是目前为止最为成熟的存储技术，功能丰富，有完善的商业关系数据库软件的支持。
- 随着大数据时代的到来，为了解决关系数据库面临的可扩展性、高并发以及性能方面的问题，各种各样的非关系数据库不断涌现，这类被称为NoSQL的系统，可以理解为“Not Only SQL”的含义。

5.1.3 分布式存储的发展历史

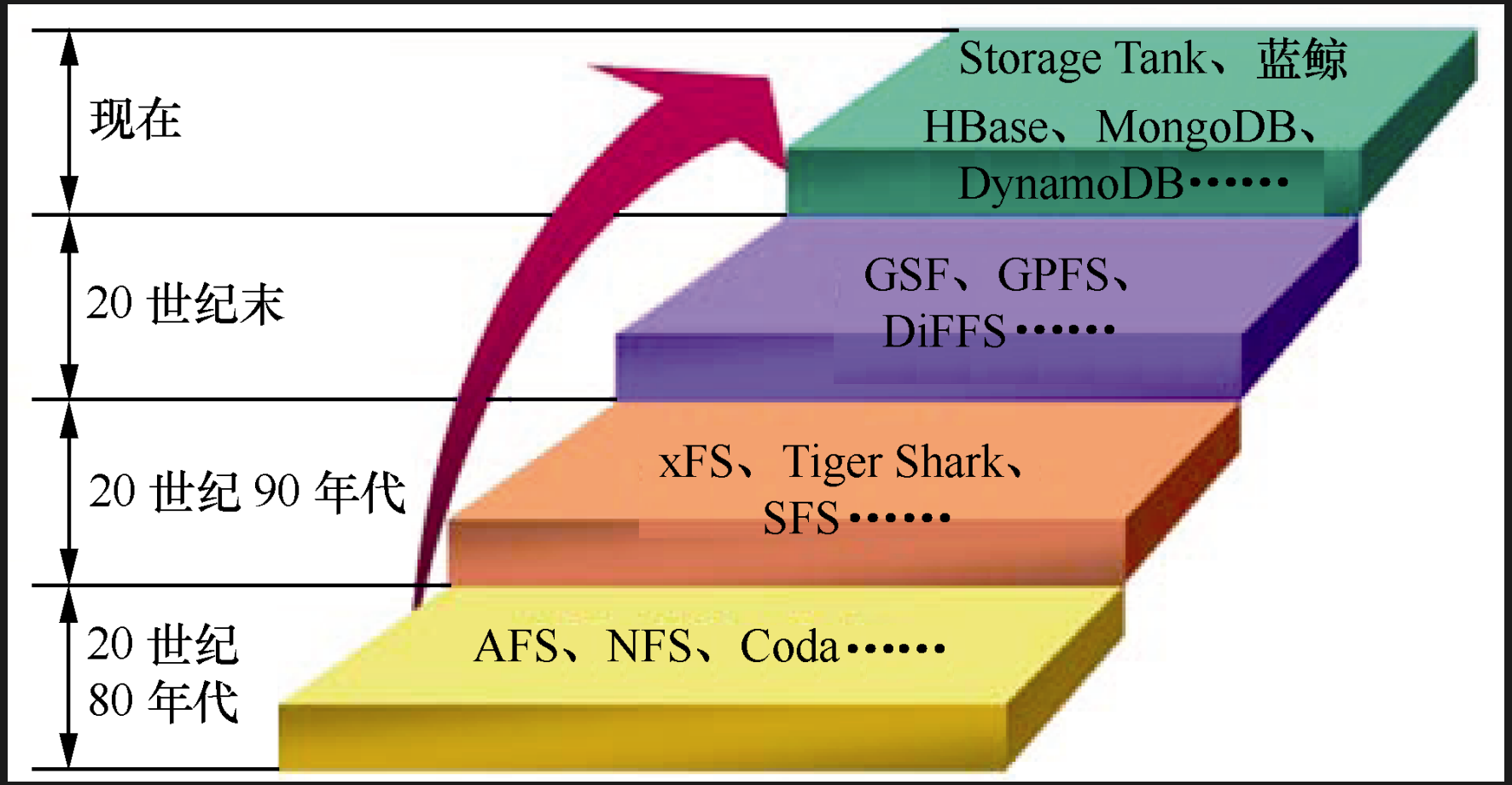


图5.2 分布式文件系统的发展

5.1.3 分布式存储的发展历史

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1．20世纪80年代的代表：AFS、NFS、Coda

（1）AFS：1983年CMU和IBM共同合作开发了Andrew文件系统（Andrew File System，AFS）

（2）NFS：1985年，Sun公司基于UDP开发了网络共享文件系统（Network File System, NFS）

（3）Coda：1987年，CMU在基于AFS的基础上开发了Coda文件系统

5.1.3 分布式存储的发展历史

2．20世纪90年代的代表：XFS、Tiger Shark、SFS

XFS：加州大学伯克利分校（UC Berkeley）开发了XFS文件系统，克服了以往分布式文件系统只适用于局域网而不适用于广域网和大数据存储的问题，提出了广域网进行缓存较少网络流量设计思想，采用层次命名结构，减少Cache一致性状态和无效写回Cache一致性协议，从而减少了网络负载，在当时获得了一定的成功。

5.1.3 分布式存储的发展历史

3．20世纪末的代表：

（1）SAN（Storage Area Network）

（2）NAS（Network Attached Storage）

（3）GPFS（General Parallel File System）

（4）GFS（Google File System）

（5）HDFS（Hadoop Distributed File System）

（1）SAN（Storage Area Networ

- 通过将磁盘存储系统和服务器直接相连的方式提供一个易扩展、高可靠的存储环境，高可靠的光纤通道交换机和光纤通道网络协议保证各个设备间链接的可靠性和高效性。设备间的连接接口主要是采用FC或者SCSI。

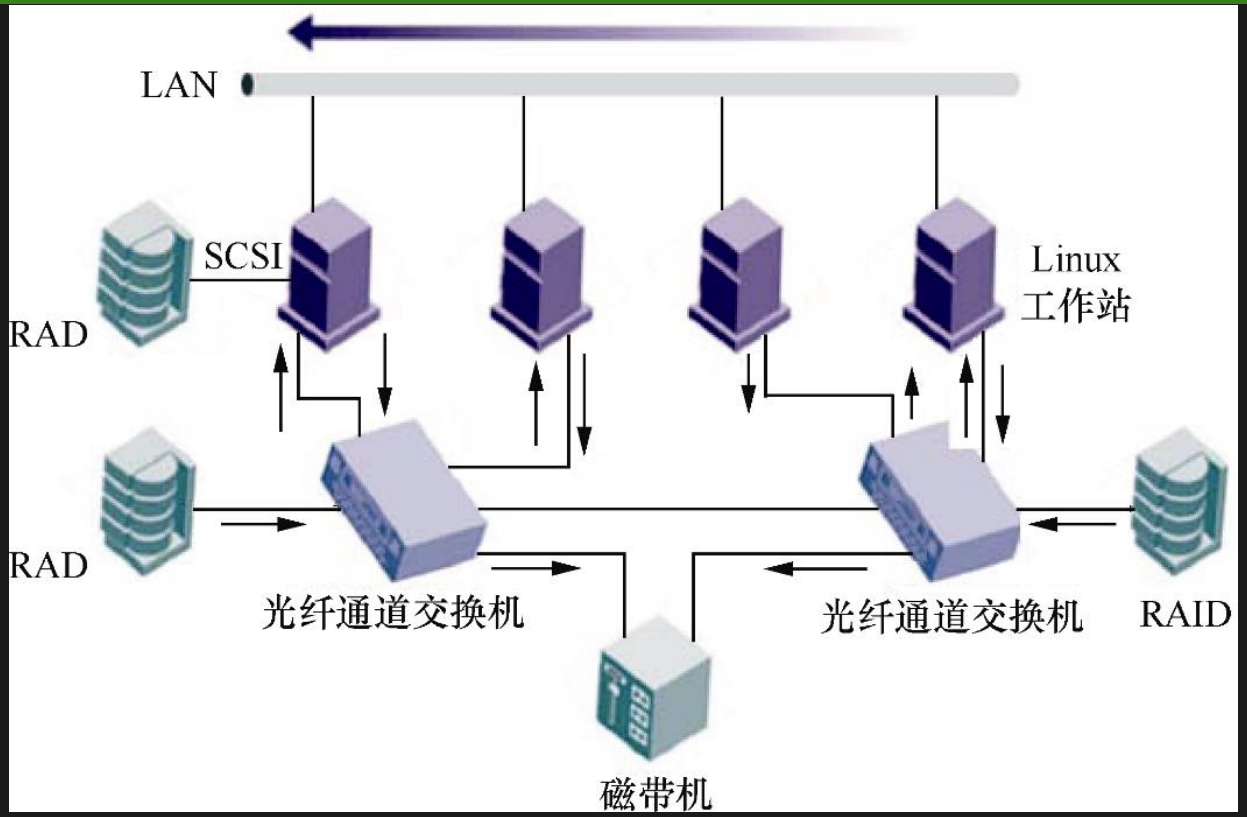


图5.3 SAN网络结构

（2）NAS（Network Attached Storage）

- 通过基于TCP/IP的各种上层应用在各工作站和服务器之间进行文件访问，直接在工作站客户端和NAS文件共享设备之间建立连接，NAS隐藏了文件系统的底层实现，注重上层的文件服务实现，具有良好的扩展性

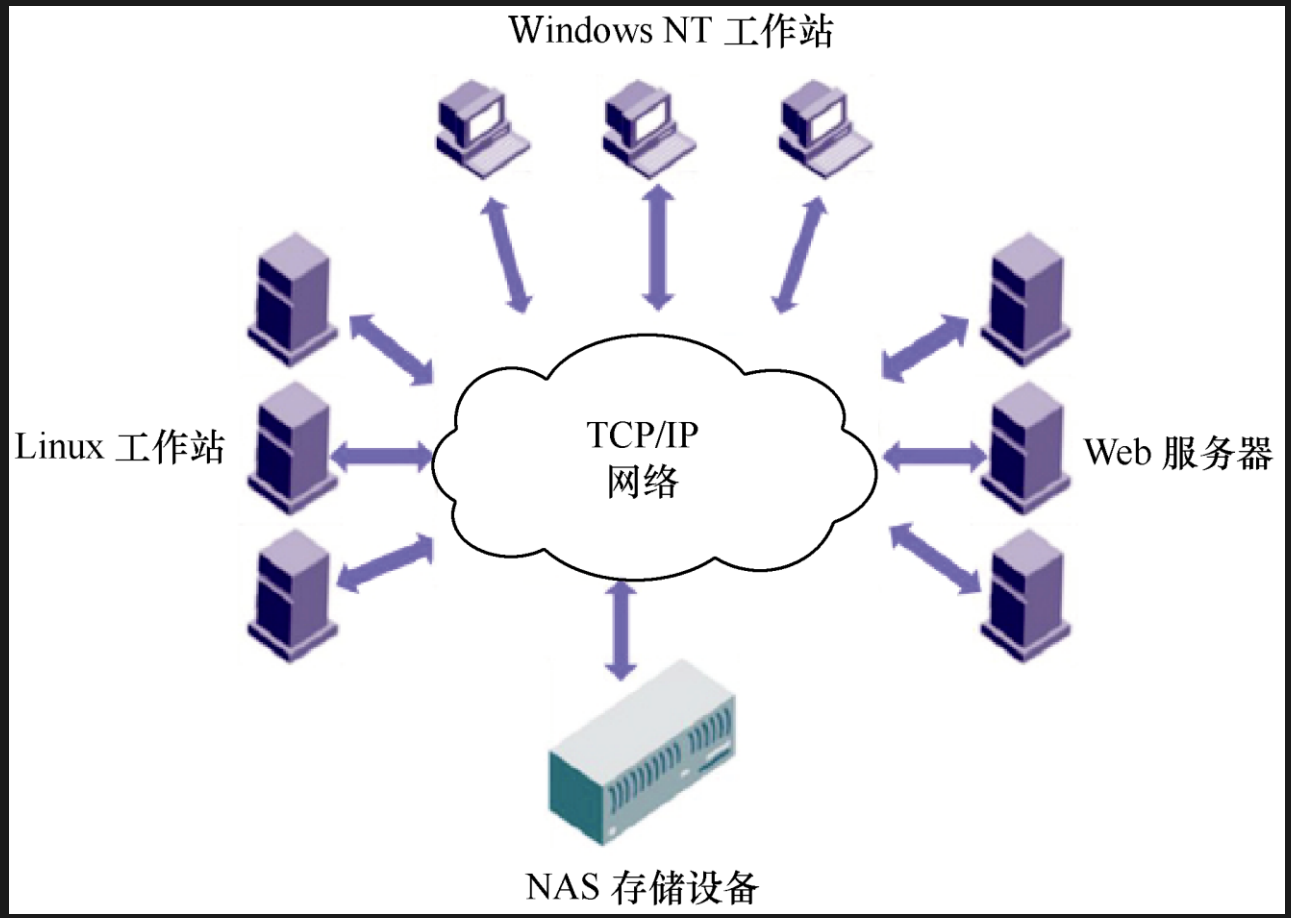


图5.4 NAS存储网络结构

（3）GPFS（General Parallel File System）

- GPFS是IBM公司开发的共享文件系统，起源于IBM SP系统上使用的虚拟共享磁盘技术。
- GPFS是一个并行的磁盘文件系统，它保证在资源组内的所有节点可以并行访问整个文件系统。
- GPFS允许客户共享文件，而这些文件可能分布在不同节点的不同硬盘上。它同时还提供了许多标准的UNIX文件系统接口，允许应用不需修改或者重新编辑就可以在其上运行。

（4）GFS（Google File System）

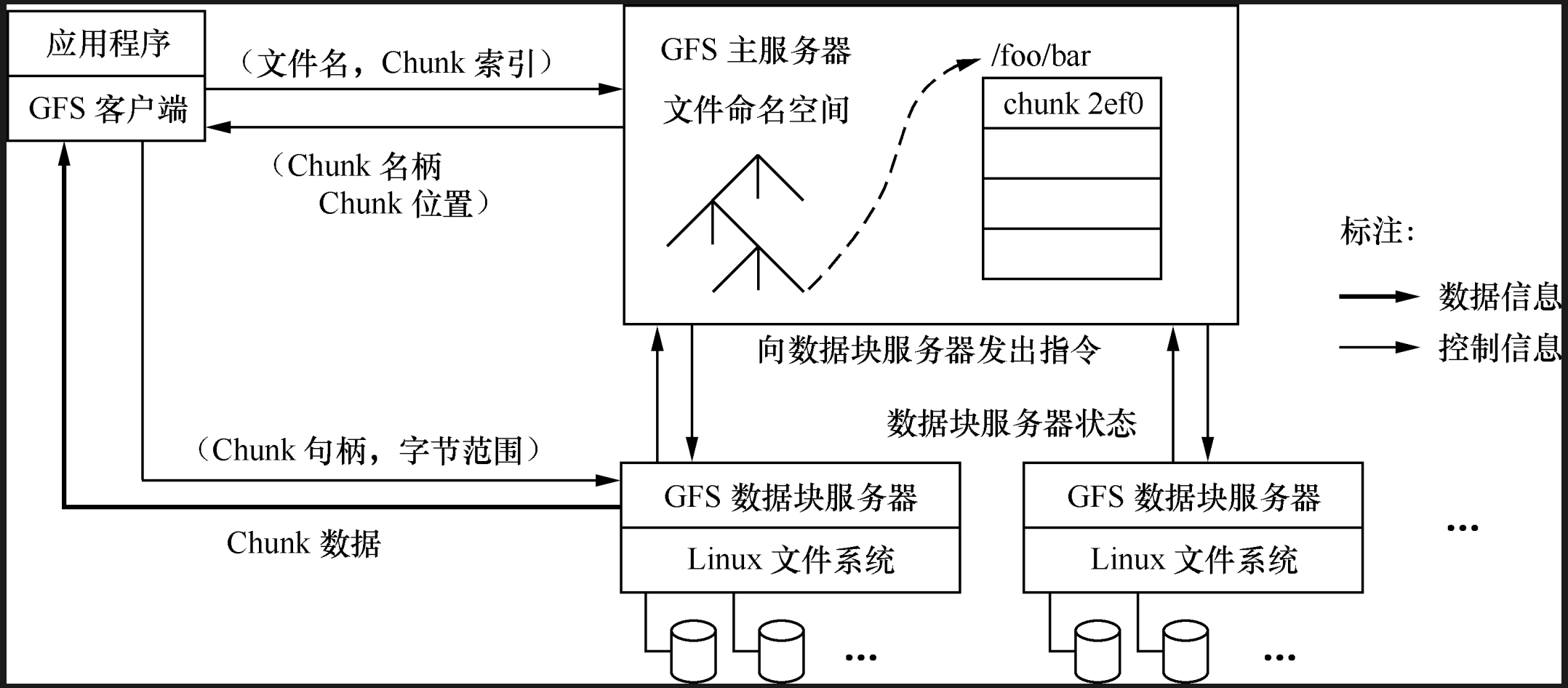


图5.5 GFS架构图

尋找/讀/寫
Video:深入浅出Google File System

（5）HDFS（Hadoop Distributed File System）

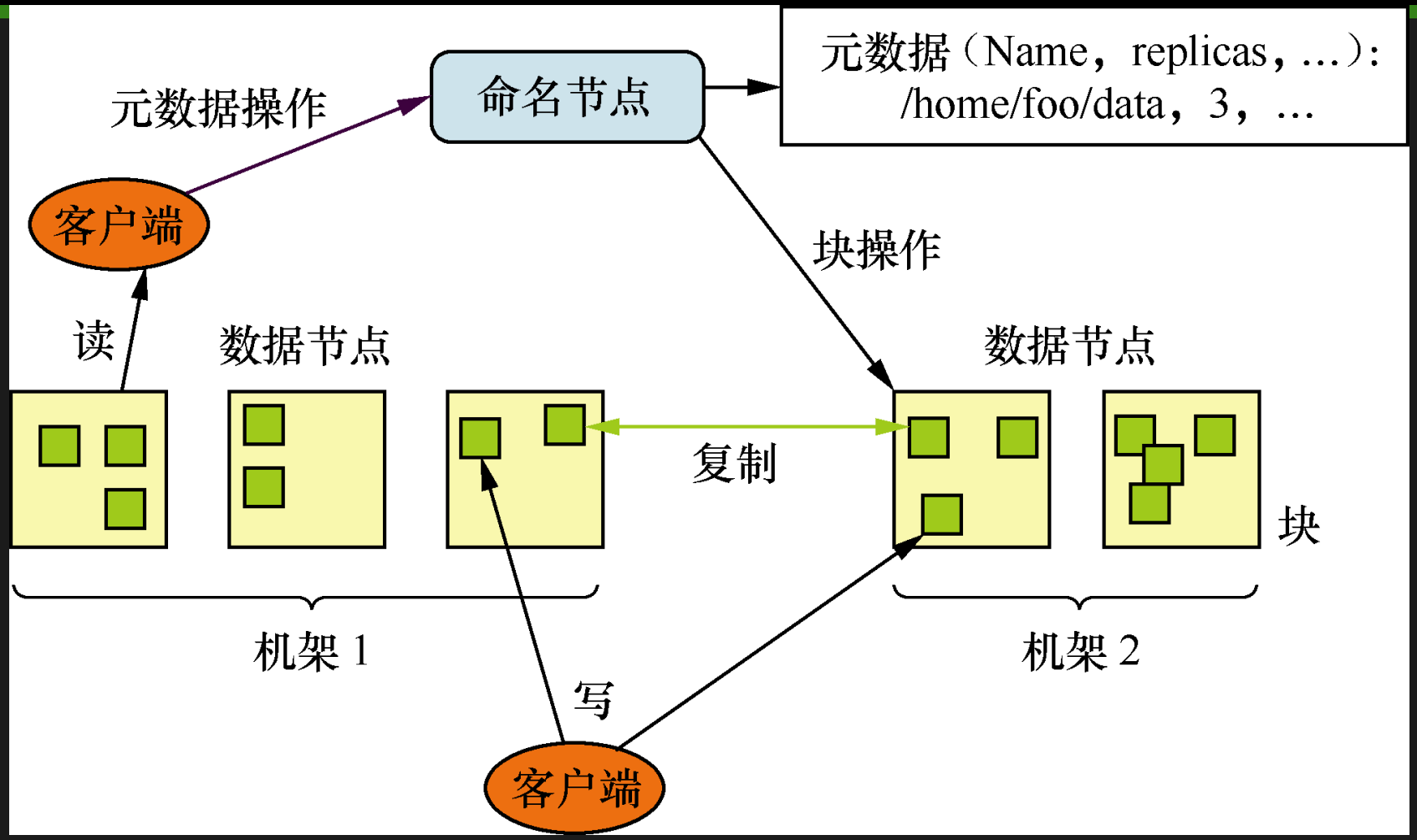


图5.6 HDFS总体结构示意图

5.1.3 分布式存储的发展历史

4．21世纪的代表：Cassandra、HBase、MongoDB、DynamoDB

（1）Cassandra：是一套开源分布式NoSQL数据库系统，最初由Facebook开发，用于储存收件箱等简单格式数据，集GoogleBigTable的数据模型与Amazon Dynamo的完全分布式的架构于一身。

（2）HBase：列存储数据库，擅长以列为单位读取数据，面向列存储的数据库具有高扩展性，即使数据大量增加也不会降低相应的处理速度，特别是写入速度。

5.1.3 分布式存储的发展历史

4．21世纪的代表：Cassandra、HBase、MongoDB、DynamoDB

（3）MongoDB：文档型数据库同键值（Key-Value）型的数据库类似，是键值型数据库的升级版，允许嵌套键值，Value值是结构化数据，数据库可以理解Value的内容，提供复杂的查询，类似于RDBMS的查询条件。

（4）DynamoDB：Amazon公司的一个分布式存储引擎，是一个经典的分布式Key-Value存储系统，具备去中心化、高可用性、高扩展性的特点。

5.2 文件存储

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

5.2.1 单机文件系统

5.2.2 网络文件系统

5.2.3 并行文件系统

5.2.4 分布式文件系统

5.2.5 高通量文件系统

5.2.1 单机文件系统

- 现代文件系统的起源要追溯到分时操作系统时期。1965年，在Multics操作系统中首次提出使用树型结构来组织文件、目录以及访问控制的思想。这些思想被后来的UNIX文件系统（1973年）所借鉴。从结构上看，它包括四个模块：引导块、超级块、索引节点和数据块。
- 为解决UNIX文件系统I/O性能低的问题，先后出现了1984年的快速文件系统（Fast File System，FFS）和1992年的日志结构文件系统（Log-Structured File，LFS）。
- 20世纪90年代至今，出现了很多单机文件系统。包括SGI公司于1994年发布的XFS，以及Sun公司于2004年发布的ZFS。

5.2.2 网络文件系统

- NFS（Network File System，网络文件系统）由Sun公司在1984年开发，被认为是第一个广泛应用的现代网络文件系统。NFS的设计目标是提供跨平台的文件共享系统。由于NFS的实现和设计思想都相对简单，该协议很快被纳入到RFC标准，并开始大量应用。然而，NFS单一服务器的结构也决定了它的扩展性有限。
- AFS（Andrew File System）是美国卡耐基·梅隆大学1982年开发的分布式文件系统。其设计目标是支持5000～10000个节点的集群，扩展性是首要考虑的因素。与NFS等系统不同的是，AFS中有多个服务器，整个命名空间被静态地划分到各个服务器上，因此，AFS具有更好的扩展性。

5.2.3 并行文件系统

- 早期的并行文件系统有BFS（Bridge File System）和CFS（Concurrent File System）等。它们运行在MPP（Massively Parallel Processing，MPP）结构的超级计算机上。。
- 20世纪90年代中期，开源的Linux操作系统逐渐成熟并得到广泛使用，为了能在越来越多的Linux集群上运行，出现了以PVFS和Lustr为代表的Linux集群上的并行文件系统。它们吸收了MPP并行文件系统的很多思想，包括采用一个专门的元数据服务器来维护和管理文件系统的命名空间，以及将文件数据条带化并分散存储在所有的存储服务器上等。

5.2.4 分布式文件系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

- 20世纪90年代后期，随着互联网的发展，出现了搜索引擎这样的海量文本数据检索工具。搜索引擎需要高吞吐率、低成本、高可靠的系统，而非高峰值处理性能的系统。于是产生了以谷歌的Google File System（GFS）、MapReduce 为代表的新型数据处理架构。
- GFS的底层平台是大规模（数千台到数万台）的、廉价的、可靠性较低的PC集群，存储设备是集群中每个节点上的多块IDE磁盘
- 谷歌架构被互联网企业广泛采用，现在流行的Hadoop就是GFS和MapReduce的一种开源实现，被很多企业采用。

5.2.5 高通量文件系统

- 高通量文件系统是为大型数据中心设计的文件系统，它将数据中心中大量低成本的存储资源有效地组织起来，服务于上层多种应用的数据存储需求和数据访问需求。
- 随着云计算技术的发展，数据中心的数据存储需求逐渐成为数据存储技术和文件系统发展的主要驱动力，高通量文件系统将成为一种重要的文件系统。
- 大型数据中心在数据存储和数据访问方面有着与先前的应用非常不同的需求特征，主要包括：数据量庞大、访问的并发度高、文件数量巨大、数据访问语义和访问接口不同于传统的文件系统、数据共享与数据安全的保障越来越重要等。

表5.1 文件系统的发展脉络

阶段	产生的技术背景	负载特征	典型代表	主要的创新技术	性能评价标准
单机文件系统	分时操作系统 多用户共享 磁盘	多用户并发访问 多进程并发访问	Unix FS FFS LFS JFS WAFL XFS ZFS	树型目录结构 索引节点（i-node） 流式访问接口 柱面组 元数据修改日志 B+树组织 写时复制 存储池	I/O 请求响应时间 聚合 I/O 带宽
网络文件系统	局域网 TCPP/IP 协议 RAID FC 网络	多客户端共享访问 多用户共享访问	NFS AFS NAS SAN 文件系统	XDR RPC VFS 无状态服务器 多服务器结构	聚合 I/O 带宽
并行文件系统	MPP 超级计算机高性能互连网络并行编程	一个作业的多任务对同一文件不现位置的并行访问 一个 I/O 请求的并行处理	Concurrent File System Vesta PVFS Lustre	文件的条带化存储 并行 I/O 接口 元数据管理与数据存储分离	并行 I/O 带宽
分布式文件系统	搜索引擎 互联网服务 Google 架构 大规模 PC 集群	数千万在线并发访问 数万并发大粒度访问	GoogleFS HDFS Haystack TFS	非 POSIX 接口和语义集中管理、分散存储全内存元数据处理多个复本	I/O 请求响应时间 并发访问吞吐率 聚合 I/O 带宽

5.3 从单机存储系统到分布式存储系统

5.3.1 单机存储系统

5.3.2 分布式存储系统

5.3.1 单机存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1．硬件基础

- 简单来说，单机存储就是散列表、B 树等数据结构在机械硬盘、SSD 等持久化介质上的实现。单机存储系统的理论来源于关系数据库，是单机存储引擎的封装，对外提供文件、键值、表或者关系模型。

- 由摩尔定律可知，相同性能的计算机等 IT产品，每18个月价钱会下降一半。而计算机的硬件体系架构却保持相对稳定，一个重要原因就是希望最大限度地发挥底层硬件的价值。计算机架构中常见硬件的大致性能参数如表5.2所示。

参考
https://blog.csdn.net/weixin_38499215/article/details/103219257?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-1.control&dist_request_id=&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFrom

https://blog.csdn.net/ChenVast/article/details/72866755

表5.2 常用硬件性能参数

类别	消耗的时间
访问 L1 Cache	0.5ns
分支预测失败	5ns
访问 L2 Cache	7ns
Mutex 加锁/解锁	100ns
内存访问	100ns
千兆网络发送 1MB 数据	10ms
从内存顺序读取 1MB 数据	0.25ms
机房内网络来回	0.5ms
异地机房之间网络来回	30～100ms
SATA磁盘寻道	10ms
从 SATA磁盘顺序读取 1MB 数据	20ms
固态盘 SSD 访问延迟	0.1～0.2ms

5.3.1 单机存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

2．存储引擎

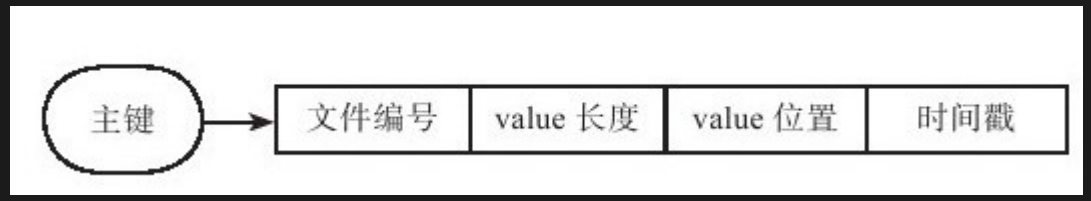
- 存储引擎直接决定了存储系统能够提供的性能和功能，其基本功能包括：增、删、改、查，而读取操作又分为随机读取和顺序扫描两种。
- 散列存储引擎是散列表的持久化实现，支持增、删、改，以及随机读取操作，但不支持顺序扫描，对应的存储系统为键值（Key-Value）存储系统。
- B树（B-Tree）存储引擎是树的持久化实现，不仅支持单条记录的增、删、读、改操作，还支持顺序扫描，对应的存储系统是关系数据库。
- LSM树（Log-Structured Merge Tree）存储引擎和B树存储引擎一样，支持增、删、改、随机读取以及顺序扫描，它通过批量转储技术规避了磁盘随机写入问题，广泛应用于互联网的后台存储系统，例如 Google Bigtable、Google LevelDB 以及Cassandra系统等。

存储引擎

存储引擎是存储系统的发动机，直接决定了存储系统能够提供的性能和功能。常见的存储引擎：哈希、B树、LSM树。

1、哈希（Key-Value存储系统）

Bitcask是一个基于哈希表结构的键值存储结构。它的特点是写时追加，也就是说它每次在文件中只会追加数据而不会修改，所以文件大小超过限制时，会新建一个活跃数据文件，而达到大小限制的文件就叫作老数据文件。由于Bitcask系统中的记录删除后者更新后都会让原来的数据变为越来越大的垃圾文件，所以要进行定期合并，对同一个key只保留最新的一个。哈希索引表存在内存中，一旦断电，重建这个哈希索引表需要扫一遍磁盘中的数据文件，为了加快重建速度，Bitcask通过索引文件来提高重建hash表的速度。索引文件其实就是内存中的哈希索引表转储到磁盘生成的结果文件。



存储引擎

2、B树存储引擎

相比哈希存储引擎，B 树存储引擎不仅支持随机读取，还支持范围扫描。关系数据库中通过索引访问数据，在 Mysql InnoDB 中，有一个称为聚集索引的特殊索引，行的数据存于其中，组织成 B+树（B 树的一种）数据结构。

MySQL InnoDB 按照页面（Page）来组织数据，每个页面对应 B+树的一个节点。其中，叶子节点保存每行的完整数据，非叶子节点保存索引信息。数据在每个节点中有序存储，数据库查询时需要从根节点开始二分查找直到叶子节点，每次读取一个节点，如果对应的页面不在内存中，需要从磁盘中读取并缓存起来。B+树的根节点是常驻内存的。

存储引擎

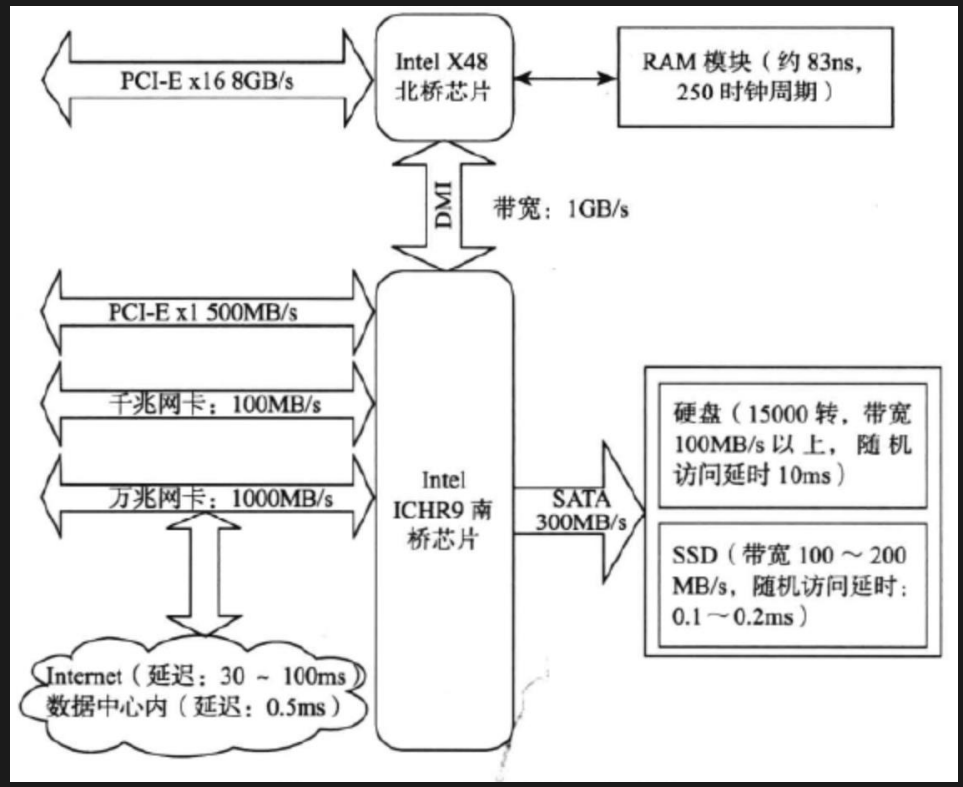
3、LSM树存储引擎

LSM 就是将对数据的修改增量保持在内存中，达到指定的大小限制后将这些修改操作批量写入磁盘，读取时需要合并磁盘中的历史数据和内存中最近的修改操作。LSM 树的优势在于有效地规避了磁盘随机写入问题，但读取时可能需要访问较多的磁盘文件。

LevelDB 存储引擎主要包括：内存中的 MemTable 和不可变 MemTable 以及磁盘上的几种主要文件：当前（Current）文件、清单（Manifest）文件、操作日志（Commit Log，也称为提交日志）文件以及 SSTable 文件。当应用写入一条记录时，LevelDB 会首先将修改操作写入到操作日志文件，成功后再将修改操作应用到 MemTable，这样就完成了写入操作。

单机存储系统瓶颈

- 存储系统的性能瓶颈一般在于IO性能。
  - 同一个数据中心内部的传输延时是比较小的，网络一次来回的时间在1毫秒之内。
  - 数据中心之间的传输延迟是很大的，取决于光在光纤中的传输时间。
- 存储系统的性能瓶颈还在于磁盘随机读写。
  - 设计存储引擎的时候会针对磁盘的特性做很多的处理，比如将随机写操作转化为顺序写，通过 缓存减少磁盘随机读操作。
  - 固态磁盘用来做缓存或者性能要求较高的关键业务。



5.3.1 单机存储系统

3．数据模型

- 如果说存储引擎相当于存储系统的发动机，那么，数据模型就是存储系统的外壳。
- 存储系统的数据模型主要包括三类：文件、关系以及键值模型。
- 传统的文件系统和关系数据库系统分别采用文件和关系模型。关系模型描述能力强，生态好，是目前存储系统的业界标准。
- 而新产生的键值模型、关系弱化的表格模型等，因为其可扩展性、高并发以及性能上的优势，开始在越来越多的大数据应用场景中发挥重要作用。

5.3.2 分布式存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1. 基本概念

（1）异常

（2）超时

（3）一致性

（4）衡量指标

- 性能
- 可用性
- 一致性
- 可扩展性

分布式存储系统，是将数据分散存储在多台独立的设备上。传统的网络存储系统采用集中的存储服务器存放所有数据，存储服务器成为系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要。分布式网络存储系统采用可扩展的系统结构，利用多台存储服务器分担存储负荷，利用位置服务器定位存储信息，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

5.3.2 分布式存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

2．性能分析

- 性能分析是用来判断设计方案是否存在瓶颈点，权衡多种设计方案的一种手段，也可作为后续性能优化的依据。
- 性能分析与性能优化是相对的，系统设计之初通过性能分析确定设计目标，防止出现重大的设计失误，等到系统试运行后，需要通过性能优化方法找出系统中的瓶颈点并逐步消除，使系统达到设计之初确定的设计目标。
- 设计之初首先分析整体架构，接着重点分析可能成为瓶颈的单机模块。系统中的资源（CPU、内存、磁盘、网络）是有限的，性能分析就是需要找出可能出现的资源瓶颈。

5.3.2 分布式存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

3．数据分布

- 分布式系统能够将数据分布到多个节点，并在多个节点之间实现负载均衡。其方式主要有两种：
  - 散列分布，如一致性散列，代表系统为Amazon的Dynamo系统；
  - 顺序分布，即每张表格上的数据按照主键整体有序，代表系统为Google的Bigtable。
- 将数据分散到多台机器后，需要尽量保证多台机器之间的负载是比较均衡的。分布式存储系统需要能够自动识别负载高的节点，当某台机器的负载较高时，将它服务的部分数据迁移到其他机器，实现自动负载均衡。

5.3.2 分布式存储系统

4．复制

- 为了保证分布式存储系统的高可靠和高可用，数据在系统中一般存储多个副本。当某个副本所在的存储节点出现故障时，分布式存储系统能够自动将服务切换到其他的副本，从而实现自动容错。
- 分布式存储系统通过复制协议将数据同步到多个存储节点，并确保多个副本之间的数据一致性。
- 同一份数据的多个副本中往往有一个副本为主副本（Primary），其他副本为备用副本（Backup），由主副本将数据复制到备用副本。当主副本出现故障时，分布式存储系统能够将服务自动切换到某个备用副本，实现自动容错。

5.3.2 分布式存储系统

5．容错

- 分布式存储系统首先需要能够检测到机器故障，然后需要将服务复制或者迁移到集群中的其他正常节点。

表5.3　　Google某数据中心第一年运行故障

发生频率	故障类型	影响范围
0.5	数据中心过热	5 分钟之内大部分机器断电，1～2 天恢复
1	配电装置（PDU）故障	500～1000 台机器瞬间下线，6 小时恢复
1	机架调整	大量告警，500～1000 台机器断电，6 小时恢复
1	网络重新布线	大约 5%机器下线超过两天
20	机架故障	40～80 台机器瞬间下线，1～6 小时恢复
5	机架不稳定	40～80 台机器发生 50%丢包
12	路由器重启	DNS 和对外虚 IP 服务失效约几分钟

5.3.2 分布式存储系统

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

6．可扩展性

- 可扩展性的实现手段很多，如通过增加副本个数或者缓存来提高读取能力，将数据分片使每个分片可以被分配到不同的工作节点以实现分布式处理，把数据复制到多个数据中心等。
- 同时，衡量分布式存储系统的可扩展性应该综合考虑节点故障后的恢复时间、扩容的自动化程度、扩容的灵活性等。

5.3.2 分布式存储的发展历史

7．分布式协议

（1）两阶段提交协议（Two-Phase Commit，2PC）：由阶段1请求阶段（Prepare Phase）和阶段2提交阶段（Commit Phase）组成，经常用来实现分布式事务，以保证跨多个节点操作的原子性。

（2）Paxos协议：用于解决多个节点之间的一致性问题。Paxos协议考虑到主节点可能出现故障，系统需要选举出新的主节点的问题，该协议可以保证多个节点之间操作日志的一致性，并在这些节点上构建高可用的全局服务，例如分布式锁服务、全局命名和配置服务等。

5.4 实践：分布式存储系统Ceph

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

5.4.1 概述

5.4.2 设计思想

5.4.3 整体架构

5.4.4 集群部署

5.4.1 概述

- Ceph最初是一项关于存储系统的研究项目，由塞奇·维尔（Sage Weil）在加州大学圣克鲁兹分校（UCSC）开发。Ceph是一个统一的、分布式的存储系统，具有出众的性能、可靠性和可扩展性。其中,“统一”和“分布式”是理解Ceph的设计思想的出发点。

  ① 统一：意味着Ceph可以以一套存储系统同时提供“对象存储”“块存储”和“文件系统”三种功能，以满足不同应用的需求。

  ② 分布式：意味着无中心结构和系统规模的无限（至少理论上没有限制）扩展。在实践当中，Ceph可以被部署于成千上万台服务器上。

5.4.2 设计思想

- Ceph最初设计的目标应用场景就是大规模的、分布式的存储系统，是指至少能够承载PB量级的数据，并且由成千上万的存储节点组成。
- 在Ceph的设计思想中，对于一个大规模的存储系统，主要考虑了三个场景变化特征：存储系统的规模变化、存储系统中的设备变化以及存储系统中的数据变化。
- Ceph的设计思路基本上可以概括为以下两点。
  - 充分发挥存储设备自身的计算能力
  - 去除所有的中心点

5.4.3 整体架构

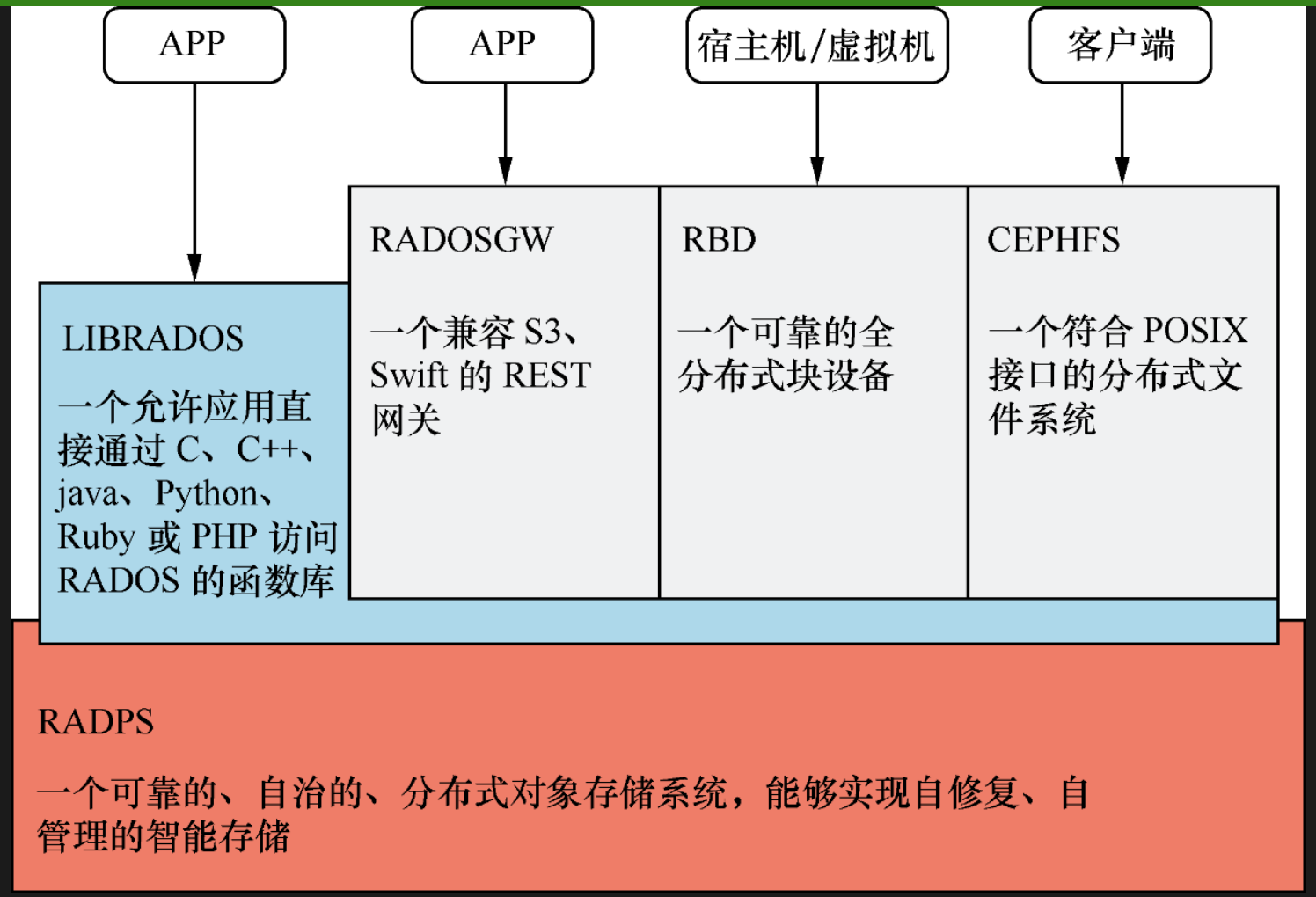


图5.7 Ceph存储系统整体架构

5.4.4 集群部署

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1．环境准备

2．安装Ceph部署工具（仅主控节点）

3．Ceph节点配置

4. Ceph安装

5. 添加OSD节点

图5.8 Ceph存储集群准备

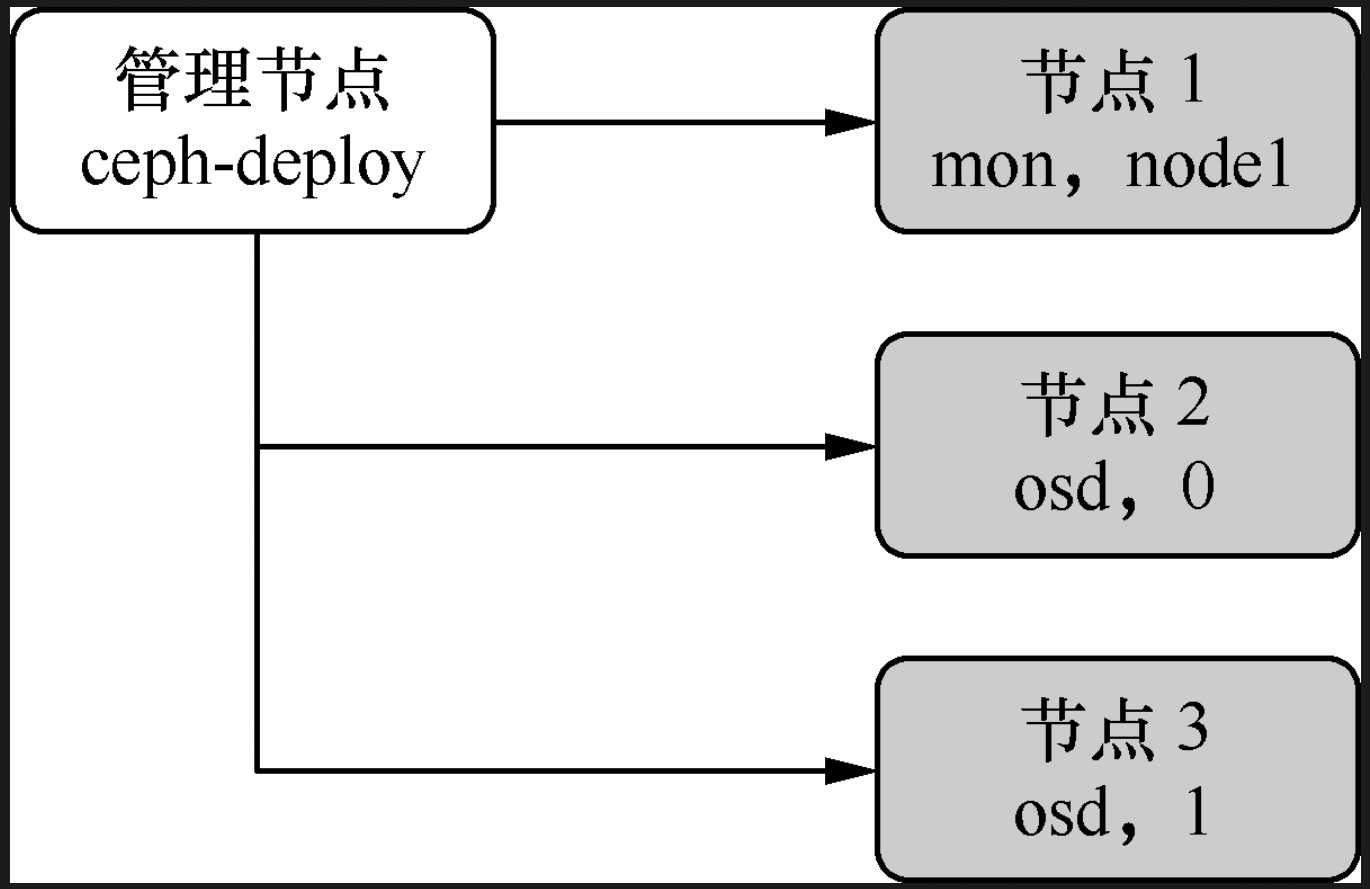


表5.4 主机信息

hostname	IP	配置
admin-node	172.20.0.195	4 核，4GB 内存，CentOS 7
node1	172.20.0.196	4 核，4GB 内存，CentOS 7
node2	172.20.0.197	4 核，4GB 内存，CentOS 7
node3	172.20.0.198	4 核，4GB 内存，CentOS 7

小结

- 分布式存储的基础
- 文件存储
- 从单机存储系统到分布式存储系统
- 实践：分布式存储系统Ceph

课内复习

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1．分布式存储的定义是什么？

2．分布式存储有哪几种类型？

3．SAN和NAS的区别是什么？

4．比较不同文件系统的特点。

课外思考

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

1．是否存在一种文件系统能够应对所有类型的文件存储？为什么？

2．Paxos的原理和机制是什么？

动手实践1

澳門城市大學
Universidade da Cidade de Macau
City University of Macau

- Ceph从2004年提交了第一行代码，至今为止已经十多年了。这个起源于Sage博士论文，最早致力于开发下一代高性能分布式文件系统的项目，现在也成为了开源社区众人皆知的明星项目。随着云计算的发展，Ceph乘上了OpenStack的春风，受到各大厂商的欢迎，成为IaaS三大组件计算、网络、存储之一。
  - 任务：通过Ceph的官方网站下载并安装使用最新的软件，进一步了解Ceph的原理。
  - 任务：理解并实践CRUSH（Controlled Replication Under Scalable Hashing）算法。

动手实践2

- Hadoop分布式文件系统（HDFS）是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。HDFS是Apache Hadoop Core项目的一部分。
  - 任务：通过Hadoop的官方网站下载并安装使用最新的Hadoop软件，进一步了解HDFS的工作原理。

讀城大.成大器

Thank you for listening.

主講人

澳門城市大學
City University of Macau
劉文堅 助理教授
andylau@cityu.mo