MPP架构简介

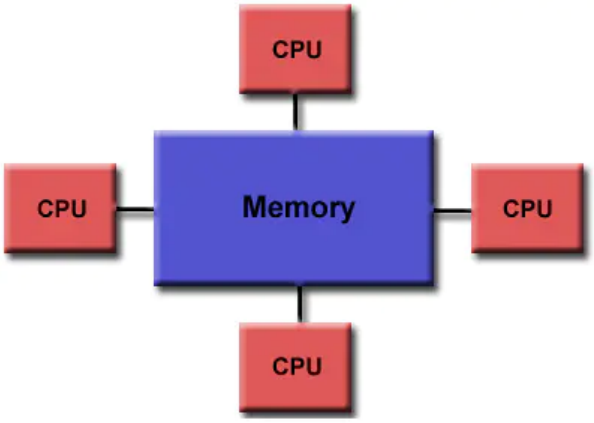
1.什么是MPP架构

MPP是系统架构角度的一种服务器分类方法。

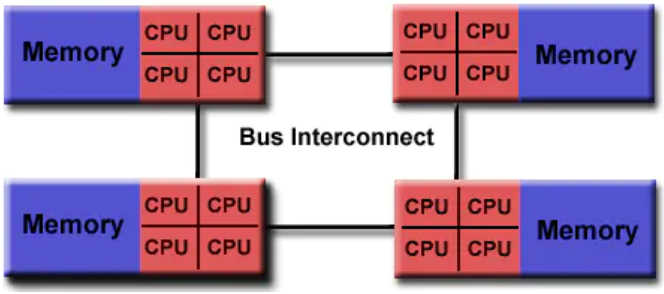
目前商用的服务器分类大体有三种：

- SMP（对称多处理器结构）（Symmetric Multi-Processor）

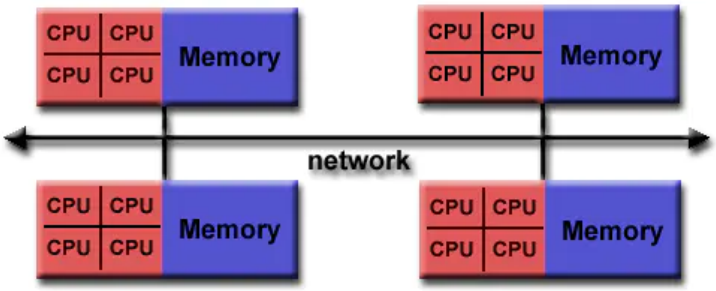
所谓对称多处理器结构，如下图所示，是指服务器中多个 CPU 对称工作，无主次或从属关系。各 CPU 共享相同的物理内存，每个 CPU 访问内存中的任何地址所需时间是相同的，因此 SMP 也被称为一致存储器访问结构（UMA：Uniform Memory Access）。



- NUMA（非一致存储访问结构）（Non-Uniform Memory Access）
  由于 SMP 在扩展能力上的限制，人们开始探究如何进行有效地扩展从而构建大型系统的技术，NUMA 就是这种努力下的结果之一。利用 NUMA 技术，可以把几十个 CPU（甚至上百个 CPU）组合在一个服务器内。其 CPU 模块结构如下图所示，NUMA 服务器的基本特征是具有多个 CPU 模块，每个 CPU 模块由多个 CPU（如4个）组成，并且具有独立的本地内存、I/O 槽口等。



- MPP（大规模并行处理结构）（Massive Parallel Processing）
  和 NUMA 不同，MPP 提供了另外一种进行系统扩展的方式，它由多个 SMP 服务器通过一定的节点互联网络进行连接，协同工作，完成相同的任务，从用户的角度来看是一个服务器系统。其基本特征是由多个 SMP 服务器（每个 SMP 服务器称节点）通过节点互联网络连接而成，每个节点只访问自己的本地资源（内存、存储等），是一种完全无共享（Share Nothing）结构，因而扩展能力最好，理论上其扩展无限制，目前的技术可实现512个节点互联，数千个 CPU。



2.数据库架构简介

数据库构架
数据库构架设计中主要有Shared Everthting、Shared Nothing、和Shared Disk：

Shared Everthting:一般是针对单个主机，完全透明共享CPU/MEMORY/IO，并行处理能力是最差的，典型的代表SQLServer

Shared Disk：各个处理单元使用自己的私有 CPU和Memory，共享磁盘系统。典型的代表Oracle Rac， 它是数据共享，可通过增加节点来提高并行处理的能力，扩展能力较好。其类似于SMP（对称多处理）模式，但是当存储器接口达到饱和的时候，增加节点并不能获得更高的性能 。

Shared Nothing：各个处理单元都有自己私有的CPU/内存/硬盘等，不存在共享资源，类似于MPP（大规模并行处理）模式，各处理单元之间通过协议通信，并行处理和扩展能力更好。典型代表DB2 DPF和hadoop ，各节点相互独立，各自处理自己的数据，处理后的结果可能向上层汇总或在节点间流转。

我们常说的 Sharding 其实就是Share Nothing架构，它是把某个表从物理存储上被水平分割，并分配给多台服务器（或多个实例），每台服务器可以独立工作，具备共同的schema，比如MySQL Proxy和Google的各种架构，只需增加服务器数就可以增加处理能力和容量。

3.MPP和批处理的对比

批处理系统 - 使用场景分钟级、小时级以上的任务，目前很多大型互联网公司都大规模运行这样的系统，稳定可靠，低成本。

MPP系统 - 使用场景秒级、毫秒级以下的任务，主要服务于即席查询场景，对外提供各种数据查询和可视化服务。

批处理架构（如 MapReduce）与MPP架构的异同点，以及它们各自的优缺点是什么呢?

- 相同点:
  批处理架构与MPP架构都是分布式并行处理，将任务并行的分散到多个服务器和节点上，在每个节点上计算完成后，将各自部分的结果汇总在一起得到最终的结果。

- 不同点:
  批处理架构和MPP架构的不同点可以举例来说：我们执行一个任务，首先这个任务会被分成多个task执行，对于MapReduce来说，这些tasks被随机的分配在空闲的Executor上；而对于MPP架构的引擎来说，每个处理数据的task被绑定到持有该数据切片的指定Executor上。

- 批处理的优势:
  对于批处理架构来说，如果某个Executor执行过慢，那么这个Executor会慢慢分配到更少的task执行，批处理架构有个推测执行策略，推测出某个Executor执行过慢或者有故障，则在接下来分配task时就会较少的分配给它或者直接不分配，这样就不会因为某个节点出现问题而导致集群的性能受限。

- 批处理的缺陷:
  任何事情都是有代价的，对于批处理而言，它的优势也造成了它的缺点，会将中间结果写入到磁盘中，这严重限制了处理数据的性能。

- MPP的优势:
  MPP架构不需要将中间数据写入磁盘，因为一个单一的Executor只处理一个单一的task，因此可以简单直接将数据stream到下一个执行阶段。这个过程称为pipelining，它提供了很大的性能提升。

- MPP的缺陷:
  对于MPP架构来说，因为task和Executor是绑定的，如果某个Executor执行过慢或故障，将会导致整个集群的性能就会受限于这个故障节点的执行速度(所谓木桶的短板效应)，所以MPP架构的最大缺陷就是——短板效应。另一点，集群中的节点越多，则某个节点出现问题的概率越大，而一旦有节点出现问题，对于MPP架构来说，将导致整个集群性能受限，所以一般实际生产中MPP架构的集群节点不易过多。

4. 场景MPP架构系统介绍

MPP架构的OLAP引擎采用MPP架构的OLAP引擎有很多，下面只选择常见的几个引擎对比下，可为公司的技术选型提供参考。
采用MPP架构的OLAP引擎分为两类，一类是自身不存储数据，只负责计算的引擎；
一类是自身既存储数据，也负责计算的引擎。
1）只负责计算，不负责存储的引擎

1. Impala
   Apache Impala是采用MPP架构的查询引擎，本身不存储任何数据，直接使用内存进行计算，兼顾数据仓库，具有实时，批处理，多并发等优点。提供了类SQL（类Hsql）语法，在多用户场景下也能拥有较高的响应速度和吞吐量。它是由Java和C++实现的，Java提供的查询交互的接口和实现，C++实现了查询引擎部分。Impala支持共享Hive Metastore，但没有再使用缓慢的 Hive+MapReduce 批处理，而是通过使用与商用并行关系数据库中类似的分布式查询引擎（由 Query Planner、Query Coordinator 和 Query Exec Engine 三部分组成），可以直接从 HDFS 或 HBase 中用 SELECT、JOIN 和统计函数查询数据，从而大大降低了延迟。Impala经常搭配存储引擎Kudu一起提供服务，这么做最大的优势是查询比较快，并且支持数据的Update和Delete。

2. Presto
   Presto是一个分布式的采用MPP架构的查询引擎，本身并不存储数据，但是可以接入多种数据源，并且支持跨数据源的级联查询。Presto是一个OLAP的工具，擅长对海量数据进行复杂的分析；但是对于OLTP场景，并不是Presto所擅长，所以不要把Presto当做数据库来使用。Presto是一个低延迟高并发的内存计算引擎。需要从其他数据源获取数据来进行运算分析，它可以连接多种数据源，包括Hive、RDBMS（Mysql、Oracle、Tidb等）、Kafka、MongoDB、Redis等。
   2）既负责计算，又负责存储的引擎

3. ClickHouse
   ClickHouse是近年来备受关注的开源列式数据库，主要用于数据分析（OLAP）领域。它自包含了存储和计算能力，完全自主实现了高可用，而且支持完整的SQL语法包括JOIN等，技术上有着明显优势。相比于hadoop体系，以数据库的方式来做大数据处理更加简单易用，学习成本低且灵活度高。当前社区仍旧在迅猛发展中，并且在国内社区也非常火热，各个大厂纷纷跟进大规模使用。ClickHouse在计算层做了非常细致的工作，竭尽所能榨干硬件能力，提升查询速度。它实现了单机多核并行、分布式计算、向量化执行与SIMD指令、代码生成等多种重要技术。ClickHouse从OLAP场景需求出发，定制开发了一套全新的高效列式存储引擎，并且实现了数据有序存储、主键索引、稀疏索引、数据Sharding、数据Partitioning、TTL、主备复制等丰富功能。以上功能共同为ClickHouse极速的分析性能奠定了基础。

4. Doris
   Doris是百度主导的，根据Google Mesa论文和Impala项目改写的一个大数据分析引擎，是一个海量分布式 KV 存储系统，其设计目标是支持

中等规模高可用可伸缩的 KV 存储集群。Doris可以实现海量存储，线性伸缩、平滑扩容，自动容错、故障转移，高并发，且运维成本低。部署规模，建议部署4-100+台服务器。Doris3 的主要架构：DT（Data Transfer）负责数据导入、DS（Data Seacher）模块负责数据查询、DM（Data Master）模块负责集群元数据管理，数据则存储在 Armor 分布式 Key-Value 引擎中。Doris3 依赖 ZooKeeper 存储元数据，从而其他模块依赖 ZooKeeper 做到了无状态，进而整个系统能够做到无故障单点。

5. Druid
   Druid是一个开源、分布式、面向列式存储的实时分析数据存储系统。Druid的关键特性如下：亚秒级的OLAP查询分析：采用了列式存储、倒排索引、位图索引等关键技术；在亚秒级别内完成海量数据的过滤、聚合以及多维分析等操作；实时流数据分析：Druid提供了实时流数据分析，以及高效实时写入；实时数据在亚秒级内的可视化；丰富的数据分析功能：Druid提供了友好的可视化界面；SQL查询语言；高可用性与高可拓展性：Druid工作节点功能单一，不相互依赖；Druid集群在管理、容错、灾备、扩容都很容易；

6. TiDB
   TiDB 是 PingCAP 公司自主设计、研发的开源分布式关系型数据库，是一款同时支持OLTP与OLAP的融合型分布式数据库产品。TiDB 兼容 MySQL 5.7 协议和 MySQL 生态等重要特性。目标是为用户提供一站式 OLTP 、OLAP 、HTAP 解决方案。TiDB 适合高可用、强一致要求较高、数据规模较大等各种应用场景。

7. Greenplum
   Greenplum 是在开源的 PostgreSQL 的基础上采用了MPP架构的性能非常强大的关系型分布式数据库。为了兼容Hadoop生态，又推出了HAWQ，分析引擎保留了Greenplum的高性能引擎，下层存储不再采用本地硬盘而改用HDFS，规避本地硬盘可靠性差的问题，同时融入Hadoop生态。

参考文章:
https://blog.csdn.net/wank1259162/article/details/109719031
https://blog.csdn.net/Fei20140908/article/details/115420408
https://zhuanlan.zhihu.com/p/395519072

标签: MPP