2019数据库系统实现技术

期中作业



学院：计算机科学与工程学院

专业：计算机科学与技术

班级：计算机1606班

姓名：戚子强

学号：20164625

# 一、OceanBase数据库的分析

## 1.1 系统概述

在过去半个世纪左右的时间里，数据库系统从无到有，经历了层次数据库、网状数据库和关系数据库等几个发展阶段。时至今日，主要数据库系统基本都是关系数据库，并广泛应用于银行、信用卡交易、商品销售、航空与铁路运输、电信、电力系统、教育、医疗与健康、电子商务等领域，成为了信息社会最关键的基础设施之一。关系数据库系统也成为了最稳定可靠、最核心的系统之一。传统关系数据库本质上是单机系统，通常采用昂贵的高端服务器和高端存储，难以应对互联网应用的高可扩展、高性能、高可用和低成本的挑战。阿里巴巴OceanBase团队研制了OceanBase开源 分布式无共享关系数据库，以“OceanBase数据库＋主流PC服务器”取代“商业数据库＋高端服务器＋高端存储”模式，很好地满足了互联网对关系数据库的需求。OceanBase已经用于淘宝、天猫和支付宝的多个生产系统，极大地降低了软件和设备成本，良好的伸缩性和自动的故障恢复不仅很好地支撑了业务，而且大大地降低了运行维护的人力成本。

## 1.2 OceanBase的国内主要市场情况

长期以来，数据库系统（DBMS）的市场规模在软件行业里一直是最大的。在2017年的总市场规模达到368亿美元，相对于2016年有超过8％的增长。然而在最近几年，传统的5家商业数据库厂商Oracle、微软、IBM、SAP、Teradata总市场份额却在持续下滑，从2011年的91%下降到2016年的86.9%。与传统商业数据库的市场占有率持续下降相对应的是，AWS透露在过去的两年中，有超过64,000个数据库从传统商业数据库（Oracle等）迁移到AWS的云数据库服务。而这个数字在2017年4月份的AWS峰会上还是23,000。虽然AWS没有透露迁移的数据库规模有多大，属于核心业务还是边缘业务，但是从数量和增速上看还是十分惊人的。

OceanBase是由蚂蚁金服、阿里巴巴完全自主研发的金融级分布式关系数据库，始创于2010年。OceanBase具有数据强一致、高可用、高性能、在线扩展、高度兼容SQL标准和主流关系数据库、低成本等特点。OceanBase至今已成功应用于支付宝全部核心业务：交易、支付、会员、账务等系统以及阿里巴巴淘宝（天猫）收藏夹、P4P广告报表等业务。除在蚂蚁金服和阿里巴巴业务系统中获广泛应用外，从2017年开始，OceanBase开始服务外部客户，客户包括南京银行、浙商银行、印度Paytm、人保健康险。经过一年多的积累，我们决定将OceanBase开向下一站的时候，选择的是先满足金融行业，尤其是专有云的场景需求。

在市场化方面，OceanBase将会以蚂蚁金服为始发站，开向下一站金融行业。首先，金融业更关注整个系统的可靠性、扩展性，OceanBase最重要的几个核心特性与金融行业的需求高度契合。其次，目前OceanBase具备的应用场景优势，是在蚂蚁内部积累的核心金融业务场景的经验，映射到金融行业显然是最容易被客户接受的，目前已经实施的几个银行客户案例也证明了这一逻辑。第三，OceanBase会在合适的时间选择上公有云。相对来说公有云的客户众多，对产品的广度要求远远高于深度，这并非我们当前最擅长的领域，团队需要继续成长，如果现在直接上公有云，会拖慢OceanBase团队在核心技术创新方面的脚步。对于蚂蚁金服来说，不是只有OceanBase在考虑怎么去做商业化的问题，而是整个蚂蚁金服都在加速将内部技术创新赋能给金融行业。在这个大背景下，OceanBase跟着集团军一起作战显然是更好的选择。从客户角度看，通常也不会只关注一个单独的数据库产品，而不去考虑整体架构、研发效能、服务治理，不去考虑金融场景、产品营运、甚至是组织文化的创新。这时蚂蚁金服的整体解决方案价值就格外重要，整个金融业正在逐步形成一个共识——金融科技是未来，分布式架构是未来。一眨眼，OceanBase也到了七年之痒的时间点，需要一些变化。对蚂蚁金服来说，OceanBase的商业化进程非常重要，需要通过更好地理解外部客户的业务场景、具体需求，帮助客户的新业务能稳定运行，让更多的业务能够跑在OceanBase上，逐步建立起品牌和声誉。让蚂蚁金服的金融科技输出战略因为有了OceanBase而不同。当然，一切的前提是把产品和生态做好，才有能力越过OceanBase的下一个山丘。

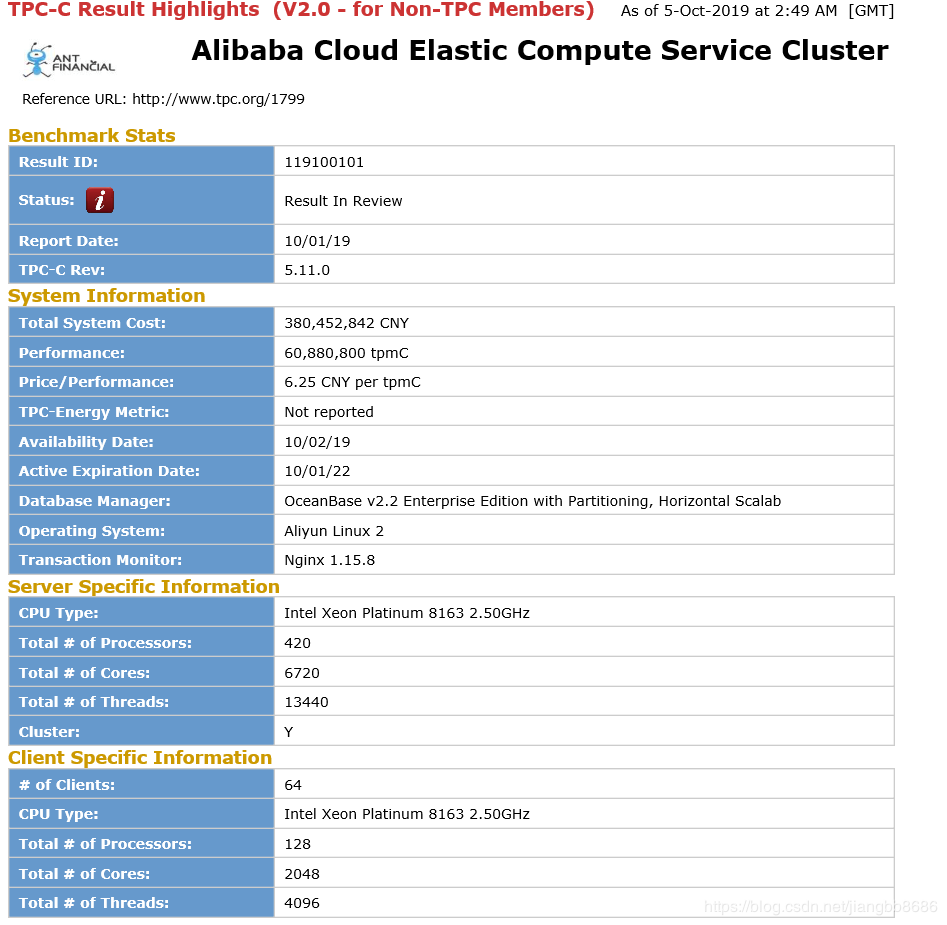
## 1.3 性能对比分析

2019年10月2日，数据库领域最权威的国际机构国际事务处理性能委员会（TPC，Transaction Processing Performance Council）在官网发表了最新的TPC-C 基准测试结果。蚂蚁金服自主研发的金融级分布式关系数据库OceanBase以两倍于Oracle（甲骨文）的成绩，打破数据库基准性能测试的世界记录，成为全球数据库演进史的重要里程碑。

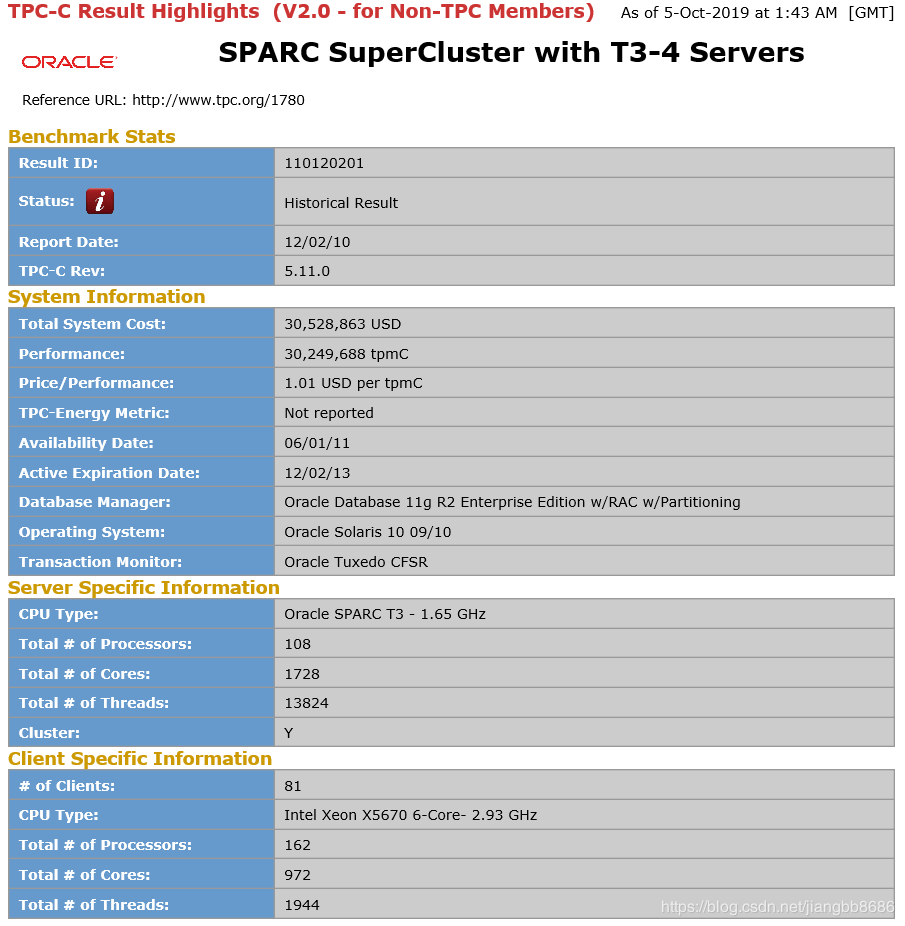
官方提供的测试数据如下：



可以看到，阿里的OceanBase的性能排在第一位，每分钟6088万笔交易，后面依次是Oracle 11g的3000万，IBM DB2 9.7的1000万，看这个数据确实还是非常不错的，但同时也应当看到另一个方面。如下图所示：



阿里的OceanBase服务器使用了6720个2.5GHz的处理器，整体系统价值3.8亿元人民币，总体性能达到6088万笔/分钟，平均每笔交易成本为6.25元人民币，按资源计算，平均8928笔/核。

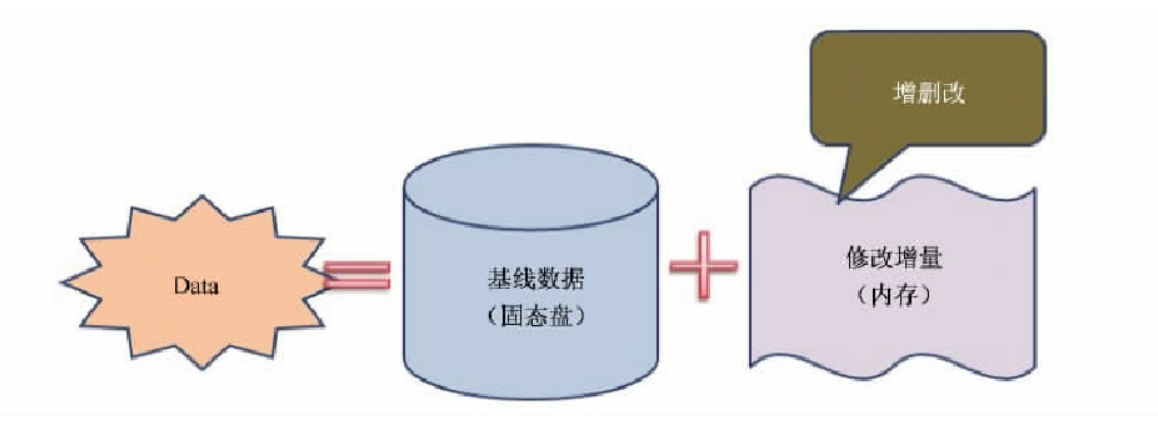


Oracle 11g服务器采用了1728个1.65GHz的处理器，整体系统花费3000万美元，按照2011年的汇率计算约2亿元人民币，总体性能达到了3000万笔/分钟，平均每笔交易成本为1美元，约7元人民币，按资源计算，平均17361笔/核。

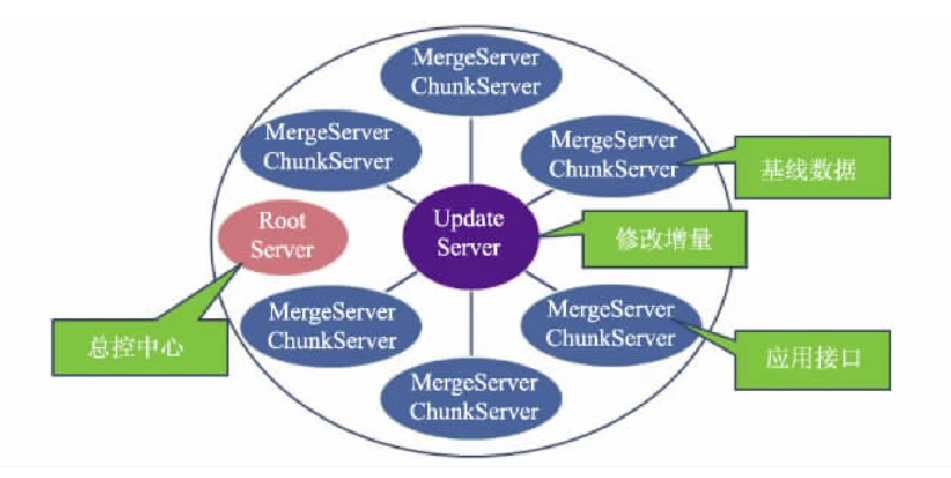
综合来看，一方面，阿里的OceanBase是分布式数据库，分布式数据库相比传统的集中式数据库有先天的优势，性能优于集中式数据库是必然的结果，理论上来讲，分布式数据库是传统集中式数据库性能的数倍甚至数十倍，但是我们看到的仅仅是两倍的性能，可以说，从这一点上来看，阿里的分布式数据库没有体现出分布式数据库的优势，或者说，阿里的分布式数据库技术还有很大的上升空间。另一方面，从成本的角度来考虑，OceanBase的测试结果构建在近4亿元系统的之上，平均每笔交易成本6.25元人民币。Oracle 11g的测试结果构建在3000万美元的基础上，平均每笔交易成本7元人民币。Oracle无论是硬件还是软件，都是基于自家生产，成本本来就很高，而阿里云是基于PC Server，操作系统使用的免费的Linux，成本理应很低，而现在的硬件成本相比2011年要便宜不止数倍的价格。资源使用效率上，OceanBase平均8928笔/核/2.5GHz，Oracle 11g平均17361笔/核/1.65GHz。这样算下来的话，阿里的资源使用效率要远远低于Oracle，因此还有很大空间很多地方需要提升和改进，来不断提升性能的同时降低价格，达到更好的性价比。

## 1.4 技术特点分析

由于大多数数据库一天的增删改数据量相对于数据库的记录数比例较小，因此 OceanBase以增量方式把当天增删改的数据保存在服务器内存中（Redo log保存在磁盘），称为MemTable；对应的基线数据（即数据库在MemTable开始时刻的快照）则分割后保存在磁盘（通常是固态盘）上，称为Sstable。



由于增删改数据MemTable通常较少，所以MemTable通常保存在服务器（称为 Up-dateServer）的内存，以增量方式保存，多次修改之间以指针连接；基线数据 Sstable通常保存在多台服务器（称为ChunkServer）的磁盘，以数据页（page，例如８KB）方式存储和访问。如图所示是单机群OceanBase的架构图。



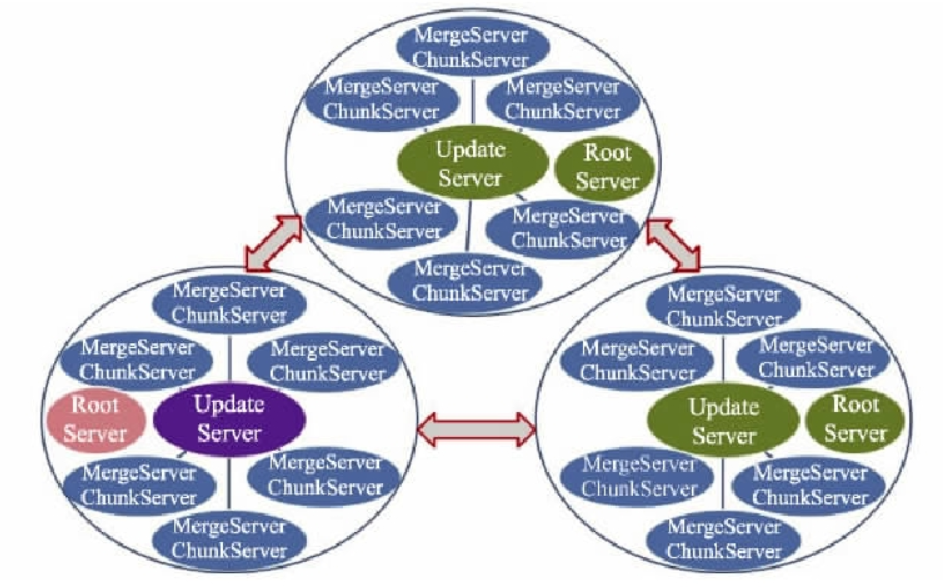
OceanBase系统架构中，包括四类服务器：主控服务器 RootServer、基线数据服务器ChunkServer、合并服务器 MergeServer，以及更新服务器 UpdateServer。

（１）主控服务器 RootServer。该服务器是OceanBase的总控中心，负责ChunkServer/MergeServer的上线、下线管理以及Sstable的负载均衡。

（２）基线数据服务器ChunkServer。ChunkServer保存基线数据Sstable并提供读访问。为了防止由于ChunkServer故障导致服务不可用甚至数据丢失，每个 Sstable保存了多个副本并分布在不同的ChunkServer上。

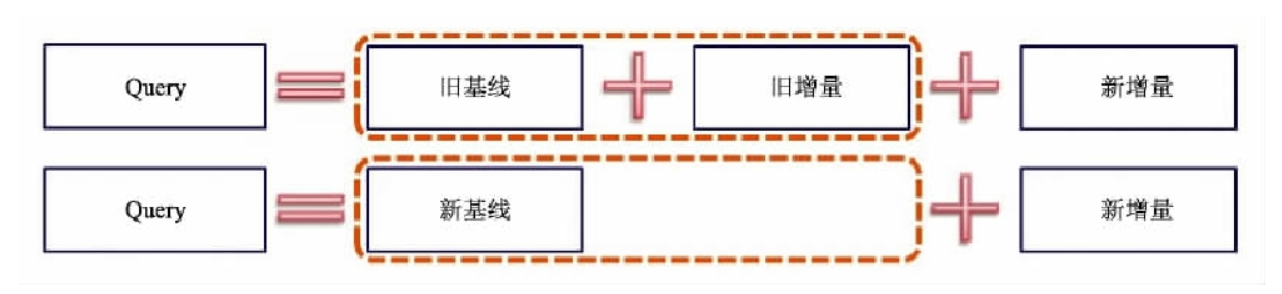
（３）合并服务器 MergeServer。OceanBase的应用接口，支持JDBC/ODBC协议，接收并解析用户的SQL请求，经过词法分析、语法分析、生成执行计划等一系列操作后，发送到相应的ChunkServer。

（４）更新服务器 UpdateServer。执行写事务、保存修改增量（到内存）、写 Redo log（到磁盘）并提供读服务。UpdateServer通常也有多个副本。由于 UpdateServer内存是有限的，修改后的增量无法一直保存在内存中，因此 OceanBase通常每天 在 某 个 时 刻 （例 如 后 半 夜业 务 低 谷 期）冻 结 当 前MemTable并 开 启 新 的MemTable；此后新的增删改写入新的MemTable，然后系统在后台把冻结的MemTable与当前基线数据融合，生成新的基线数据，这个过程称为每日合并。每日合并完成后，冻结的MemTable以及旧的Sstable即可释放，其占用的内存和磁盘空间也被回收。为了防止 UpdateServer故障或者机房故障导致服务不可用甚至数据丢失，OceanBase通常部署多个机群（例如一主两备），主机群执行写事务并且至少同步到一个或多个备用机群（超过半数）。这样任何一个机群异常都不会导致服务不可用，从而保证了数据库的高可用性。其架构如图所示。传统关系数据库通常采用主备库镜像，主库备库之间的网络异常以及备库异常都可能导致主库备库不同步；OceanBase的多机群（≥３）方式使得不仅单个备机群异常不会影响业务，而且使得，即使主机群突然故障，数据库系统也会在最多若干秒后自动恢复服务，不会造成任何数据丢失。



### 1.4.1 读事务

对 于 用 户 的 读 取 任 务，MergeServer 接 受 SQL 请 求 并 解 析 生 成 SQL 执 行 计 划。MergeServer确定事务中数据的基线数据在哪些ChunkServer上，然后通知这些 ChunkServer执行对应的读请求。由于基线数据与修改增量的分离，如图所示，读事务需要融合基线数据和修改的更新数据，返回满足查询条件的数据。一个特殊的情况是在每日合并期间，让所有ChunkServer同步完成每日合并或者同步启用新的Sstable可能非常复杂甚至无法完成（比如某个ChunkServer网络异常），因此某些ChunkServer的部分Sstable完成了与冻结的MemTable的融合，生成了新的Sstable，其他Sstable尚未完成融合，有些查询可能用到旧的 Sstable，有些查询可能用到新的 Sstable，还有的 查 询 可 能 同 时 用 到 旧 的 Sstable和 新 的 Sstable。相 同 内 容 的 SQL 语 句，到 达 的MergeServer不同或到达时间不同，使用新旧 Sstable的情况可能有所不同，那么查询的结果是否一致呢？答案是肯定的。因为查询使用旧基线数据时，会融合旧增量数据（即冻结的MemTable）和新增量数据（即新的MemTable），查询使用新基线数据时，则只融合新增量数据（新的MemTable），因此得到的结果是一致的，如图所示。



尽管使用了与传统关系数据库相似的方式存储，Sstable的随机读并没有成为 OceanBase的瓶颈，这得益于OceanBase通常采用固态盘作为存储并且没有随机磁盘写，一个没有随机磁盘写的固态盘可以提供每秒几万次的随机读（一个机械盘每秒只能提供几百次的随机读）。传统的基于磁盘的关系数据库，其读事务操作的基本步骤是先从磁盘中读出数据页（page），再从中取出需要的内容，然后根据用户的SQL进行操作得到结果。OceanBase的读事务操作与它们类似，不同的是OceanBase从读出的数据页中取出需要的内容时，还得与对应的修改增量融合。由于被修改数据所占的比例比较小、修改增量以及融合操作都在内存中，这个操作对性能的损耗很小。与此同时，由于使用固态盘并且没有随机写，OceanBase能够充分利用固态盘优异的随机读性能，因此能够获得很好的读性能。

### 1.4.2 写事务

MergeServer解析SQL请求生成SQL执行计划后，如果不是只读事务，则向 ChunkServer获取对应的基线数据，然后连同执行计划提交给主机群 UpdateServer执行。主机群UpdateServer执行写事务，生成 Redo log，同步给备机群并持久化到各自磁盘，如果成功者（包括自己）超过了半数，则在MemTable中提交该事务并应答客户。尽管有大量的随机访问，由于增删改的修改增量放置在内存，Redo log是顺序写，所以OceanBase系统中没有随机写磁盘。不仅如此，OceanBase的 UpdateServer服务器通常配置带电池或电容的 RAID 卡，这种 RAID 带有内存（例如１GB）并且在服务器断电时能够保持其中的数据不丢失，当小块的 Redo log写入磁盘时，其实只是写到了 RAID 卡的内存中，该RAID 卡稍后批量把其内存中数据写到磁盘上，这样不仅缩短了日志刷盘的时间（大约0.1ms），而且降低了小块 Redo log的写入对固态盘的性能和寿命的影响（因为固态盘的写入也是以页（page）为单位的，例如４KB，并且写入前需要先擦除，在已有页上即使追加一个字节也需要把原有页读出来，与追加字节合并，然后写入一个新的页）。每日合并期间把修改增量与旧基线数据合并生成新基线数据时，对磁盘的访问是顺序读和顺序写，当施加一定的流量控制后，这种顺序读和顺序写对磁盘的随机读的影响也是可控的。传统的基于 磁 盘 的 关 系 数 据 库，其 写 事 务 操 作 的 基 本 步 骤 是 从 磁 盘 中 读 出 数 据 页（page），再从中取出需要的内容，然后根据用户的SQL进行修改，再把修改后的结果与原数据页融合生成新的数据页，写日志（Redo log、Undo log）并把新的数据页刷到磁盘，由于每次修改通常在几十到几百字节，而数据页的大小通常在几KB（例如８KB），这就出现了明显的写放大（８KB/100≈80X）。OceanBase的写事务操作也是先从磁盘中读出数据页，再从中取出需要的内容，然后根据用户的 SQL 进行修改操作以生成增删改的增量，写日志（Redo log）并且把修改增量加入到MemTable。由于内存中修改增量没有也不需要做成数据页，因此OceanBase不仅省去了写新的数据页到磁盘的操作（随机写磁盘），也避免了传统数据库的写入放大（每日合并通常在系统低谷期间进行，对业务影响很小，更不会影响高峰期的业务）。这使得OceanBase在写事务性能上有明显的优势。

### 1.4.3 性能经济分析

目前，国内外的学术界和工业界在支撑互联网应用的数据库产品方面做了大量努力。Google的 Bigtable是一个基于 Google File System的分布式表格系统，只支持单行事务。Google的 Spanner实现了跨越全球的分布式事务，但复杂 SQL 性能较低，且对时钟有很强的依赖。OceanBase系统是在保证“强一致性”的前提下，追求系统的“最大可用性”，实现面向互联网业务需求的强伸缩性、高性能、高可用和低成本的通用数据库。

根据数据存储方式的不同，可以把目前的数据库分为三类：外存型数据库、全内存数据库以及内外存混合型数据库。

（１）外存型数据库。这种类型的数据库的读写事务都基于外存（磁盘），内存作为缓存（Cache），数据的存储基于数据页（page），读和写都有一定的放大效应，写入放大和磁盘随机写性能限制了数据库的性能。多数传统的关系数据库都属于这个类型。

（２）全内存数据库。这种类型的数据库数据全部在内存中（log除外），没有数据页的概念，数据读写都在内存中进行（写日志除外），性能很高。VolteDB (http://voltdb.com)和MemSQL (http://www.memsql.com)属于这个类型，并且都号称是全世界最快的内存数据库。

（３）内外存混合型的数据库。这种类型的数据库部分数据在外存（磁盘），部分数据在内存，在磁盘上的数据以数据页为单位存储，内存中的数据不使用数据页，内存也不仅仅作为缓存。OceanBase属于这个类型，其基线数据以数据页方式保存在磁盘上，而增删改的增量则在内存中，性能介于磁盘型的数据库与全内存数据库之间。

从存储方式来看，OceanBase属于内外存混合型数据库；与内存数据库相比，OceanBase在读写事务性能上的劣势不大，但在经济性上优势十分明显，具体在以下几方面。

（１）读事务性能。内存数据库完全避免了磁盘I/O（写日志除外）并且也没有数据页导致的读放大。然而，由于固态盘的随机读响应时间（0.1 ms）仅为机械盘（3~5 ms）的几十分之一，并且由于二八法则，OceanBase可以用相对较小的内存代价（比如全量数据的20%）缓存少部分比较热的数据，再加上query Cache，因此读事务并不会成为OceanBase的性能瓶颈。

（２）写事务性能。OceanBase的写事务也是内存操作，与全内存数据库类似，其中的读操作，由于SSD优异的随机读性能以及缓存，因此OceanBase与全内存数据库的写性能的差距不大。

（３）经济性。二八法则表明，一段时间（例如小时或者天）内数据库中的数据只有小部分会被频繁访问，大部分数据被很少访问或者根本没有被访问。把这些很少或没有被访问的数据与频繁访问的数据同等对待并一样占用昂贵的内存资源，显然是不经济的。由于固态盘容量大大高于内存容量，同等尺寸的数据量，内存数据库需要的服务器数量也比内外存混合型的OceanBase数据库需要的服务器数量多得多。

## 1.5 前景展望

最近，即使是数据库这样非常成熟的细分领域也发生了不少动荡：谷歌凭借 Spanner 从一招鲜玩家杀入到远见者，阿里云一举跻身远见者，且拥有最多的 DBMS 服务品种；亚马逊连年快速上升，如今已经跟 Oracle、微软非常接近。

OceanBase 当初没有选择基于开源或已有的技术思路开发，而是选择走分布式自研这条路，虽然走得艰难，但做成之后就会成为不可替代的优势。过去这十来年正好是分布式系统发展的十来年，转型到分布式已经成为所有人都认可的一个选择。如今，以 Google Spanner、蚂蚁金服的 OceanBase 为代表的分布式关系数据库，不仅解决了关系数据库的扩展性问题，也极大地降低了关系数据库的成本（数量级的硬件成本的降低），还提升了可用性。

现在，兼容 Oracle 的工作是 OceanBase 的重中之重。OceanBase 团队的目标是，用两年时间做到 Oracle 业务的平滑迁移，不需要修改一行代码、不需要业务做任何调整就能够将数据库迁移过来。

对于数据库的未来，蚂蚁金服高级研究员阳振坤表示：“尽管今天在业界，数据仓库主要依赖的不是关系数据库，但可以看看 Google。今天 Google 的大数据分析/数据库仓库基本都统一到了 Spanner，这应该是 5-10 年后产业界的写照。”未来，OceanBase 还会走得更快、更远。