数据库的下一场革命：S3 延迟已降至原先的 10%，云数据库架构该进化了

Original 曹伟（鸣嵩） InfoQ 2023-12-24 10:15 Posted on 辽宁

作者｜曹伟（鸣嵩）

> 众所周知，在数据库的历史上，每次存储介质的变化都会引发软件的变革。从 SAN 存储到 SSD 到大内存到 NVM，都触发了数据库内核从理论到工程的演进。

数据库一直是推动企业数字化和创新的最重要基础设施之一。从关系型数据库到 NoSQL 数据库、分析型数据库、多模数据库，这个领域正在持续的进化与变革，涌现了大量的新型数据库产品，满足不同企业的应用场景和细分市场需求。

然而，关系型数据库（Relational Database Service，简称 RDS）依然占据了数据库整体市场的大半壁江山，根据 IDC 近期发布的《2023 年上半年中国关系型数据库软件市场跟踪报告》，2023 年上半年中国关系型数据库软件市场规模为 17.5 亿美元，其中公有云关系型数据库的市场份额约为 59%。而根据 Garnter《Forecast Analysis: Database ManagementSystems, Worldwide》报告中的预测，2023 年全球关系型数据库总市场达到 838 亿美元，在数据库总盘子里公有云部分占比也约为 59%。

RDS 通常以云盘（即块存储）作为其核心存储基础设施。AWS 的 RDS 服务便是一个例子，其所有实例规格均采用了 Elastic Block Store（EBS）云盘。对于广泛使用 RDS 的用户，以及在公共云上购买虚拟机来自建数据库服务的用户，云盘是否就代表了存储的最终选项呢？答案是“No”。在技术革新缺席的前提下，云盘在性价比和计费策略方面失去了其竞争优势，我们判断其会从云厂商的主导产品降级为边缘选项。

公共云的关系型数据库将会从依赖云盘向利用好对象存储，向采用更加云原生的架构的新时代迈进。为了适应对象存储，充分发挥其优势，数据库的架构也势必需要进行大刀阔斧的改造，水平扩缩容、容灾技术以及存储引擎的数据格式都将随之变化。

云盘存在的问题

云盘的**第一个痛点**是定价比较高。例如，一块 1TB 的标准型云盘（含 5 万 IOPS 及 350MB 带宽），云服务商会收取约 1000 元每月的基础费用。为了服务海量的客户基数，云盘的 IOPS（输入 / 输出操作每秒）和带宽通常会在软件上通过流控进行限制，免费的 IOPS 额度通常在 1000 至 3000 之间，而带宽限额大约在 150MB/s。超出这些限额，用户则需要向云厂商购买额外的 IOPS 和带宽，若要将 IOPS 提高至 10 万（这一

性能水平对于企业级 NVMe SSD 来说并不算特别高），用户需要额外支付 1500 元每月的预配置性能费用。然而，1TB 的企业级 NVMe SSD 的一次性购买成本还不到千元。

第二个限制是云盘的弹性尚未完全向用户开放。主流云厂商的云盘仅支持扩容，不支持缩容。这一限制导致业务在进行缩容时必须采取曲线救国的策略，即通过逻辑复制的方式，先将数据迁移到一块容量更小的新云盘上，然后才能释放原有的较大云盘。另外，部分云厂商，例如 AWS 的 EBS 对控制面还存在限流保护，限制每两次扩容操作之间必须间隔六小时。这样的措施虽然保证了服务的稳定性，但也在一定程度上限制了用户对存储资源的即时调整能力。这些都限制了上层软件基于云盘实现按存储使用量付费的能力。

云盘的**第三个局限性**在于其灾难恢复能力局限于单个可用区（AZ）。云厂商建议的最佳实践是，为了实现更高级别的业务连续性，客户应该采取跨可用区的灾难恢复策略。这意味着，如果客户想要为他们的数据库实现跨 AZ 的灾难恢复，他们不得不购买多个云盘。然而，这种额外投资并非最经济的选择，因为云盘定价已经包含了单 AZ 多副本数据的成本。当用户为了实现跨 AZ 的冗余而购买更多云盘时，存储层面的多副本与数据库层面的多副本机制叠加在一起，便产生了资源上的重复配置。

最后，在面对高性能数据库需求时，云盘的性能也会成为限制整体系统性能的**薄弱环节**。云盘使用分布式架构，通过 Erasure coding 机制将数据分割成多个小片段，并将其冗余存储在多个服务器上。进一步的，还可以对数据进行压缩。这些技术以牺牲一定的性能为代价，换来了显著的可扩展性和成本效益。由于所有 I/O 操作都需要跨越网络，因此云盘的延迟通常比本地盘高一个数量级。

旗舰数据库产品如 Aurora 和 PolarDB 采纳了更新的设计理念，构建了定制化的分布式存储集群，来解决云盘的性能问题。Aurora 采用了日志即数据库的理念来减少数据库节点与存储节点之间的数据传输量，PolarDB 则使用 RDMA 和 NVM 来优化 I/O 延迟，两者都支持多个数据库节点并发访问存储节点的共享数据架构。然而，这些存储系统与数据库之间的通信是通过私有接口实现的，并不对外部用户开放。另外这些专用存储的定价比云盘更高。

其他的云存储选项

和云盘相比，云上的本地盘实例存储性价比要高很多。实例存储采用了类似于 SR-IOV 和 SPDK 的高效虚拟化技术。因此实例存储的延迟和带宽都接近于物理 SSD 的性能。实例存储的价格也经济很多，折算下来 1TB 实例存储每月的单价在 300 元以下。然而，实例存储在数据持久性方面存在局限。由于其设计为单副本存储，如果宿主机发生故障，存储在其中的数据可能会遭受永久性丢失。此外，当虚拟机迁移至另一台宿主机时，实例存储中的数据也将被清除。因此，实例存储不适合在需要高可靠性和数据持久性的应用场景中作为主要存储介质。

对象存储的价格是最低的，1TB 一个月的存储成本约为 120 元。低定价得益于其软硬件的协同优化。在软件层面，采用更激进的 EC 和压缩算法来提高存储密度；而在硬件方面，通过使用高密度盘和专用服务器，进而降低服务器、机架及电力的均摊成本。对象存储系统还利用定制的 FPGA 硬件来减轻 CPU 处理网络和数据的压力。在资源调度上，对象存储一般会采用大集群的部署方案，这有利于降低碎片率，提高系统的整体水位

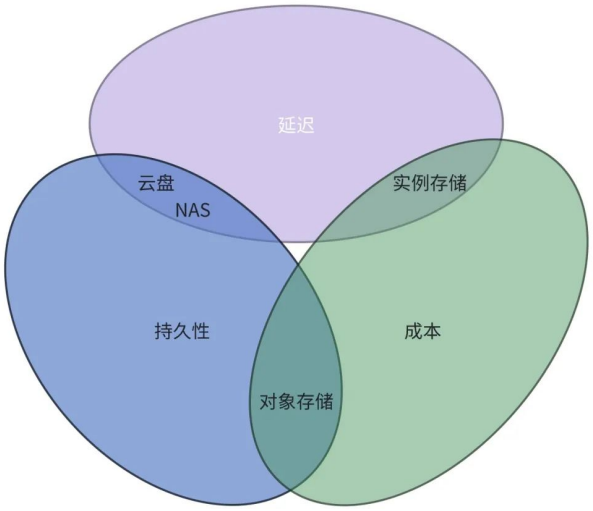
线。得益于其分布式大资源池的设计，对象存储能够支持 10Gb 乃至 100Gb 的访问带宽。此外，对象存储通常还具备跨可用区域（AZ）的灾难恢复能力。

对象存储的缺陷是其延迟比较高，首字节访问延迟可能高达数十毫秒。这一问题部分源于早期对象存储解决方案通常使用 HDD 作为存储媒介，并且在软件层面，I/O 请求的排队处理也会造成一定延迟。然而，高延迟并非对象存储的本质缺陷，而是由于成本考虑和产品定位所做的权衡。

实际上，在技术层面，构建低延迟的对象存储系统是完全可行的。例如最近在 AWS Reinvent 大会上亮相的"S3 Express One Zone"服务，将延迟减少至原先的十分之一，实现了毫秒级的响应速度，接近传统文件存储 NAS 系统的水平。刚才说了，S3 的高延迟是产品定位和技术的权衡，有得必有失，"S3 Express One Zone"不再支持跨 AZ 容灾，所有数据都在单 AZ 内，数据的持久性下降。此外"S3 Express One Zone"是更强了，也更贵了，其价格不仅是 S3 标准版的 7 倍，也超过了自家云盘，达到 EBS gp3 的 2 倍。

解法：持久性与延迟的解耦

用一张图可以总结下几个云存储产品各自的特点和不足：



可以看到，当前市场上尚未出现一种云存储产品，可以同时在低成本、低延迟、高持久性几个维度上都达到令人满意的程度。

随着云计算的渐进普及，降低成本逐渐成为用户的首要诉求，一些公司，比如 37signals 和 X，已经基于成本效益的考量决定下云，从公共云平台转回到传统的 IT 基础设施。在这种趋势下，云数据库服务供应商面临的紧迫挑战是如何在现有的存储 IaaS 产品基础上，构建更有成本竞争力的数据库服务。

一个方案是基于实例存储搭建多副本的数据库系统。前文说过，实例存储是单副本存储，它存在一个风险：一旦托管它的宿主机发生故障或者相应的虚拟机迁移，就可能导致数据丢失。这一方案的理论基础建立在一个假设之上：多个实例存储丢失数据的概率是相互独立的。基于这个假设，增加副本数量，会降低多个实例存储同时丢失数据的概率。

然而，这个假设并不总是成立，在现实里，因为工程的限制，多个实例存储可能会因为共同的原因导致同时数据丢失。例如，可能所有存储都使用了同一批次存在固件缺陷的 SSD，导致多台主机同时下线；或者，云服务提供商的控制系统发生故障，引起大规模的虚拟机迁移。因此，对于那些对数据持久性有极高要求的生产环境来说，这种方案并不适用。

另一个方案是将存储的持久性和延迟两个特性进行分离，通过对象存储实现高持久性，通过实例存储 / 云盘来实现低延迟。尽管对象存储可以提供低成本、高带宽和跨可用区的数据持久性，但在作为关系型数据库主存储时，它的读写延迟成为了一个显著的挑战。为了解决这一问题，我们采用一种分层存储策略，将存储解耦为三个组件：持久化、写缓存和读缓存，分而治之。

在这种设计中，对象存储仅负责数据的持久化，为系统提供灾难恢复保证。写操作会通过缓存层降低写延迟，技术上可以利用如 Raft 这类同步协议，缓存到低延迟的存储介质，例如本地磁盘或云盘，甚至可以考虑追加写入到消息队列服务。读操作也是，通过缓存层降低读延迟，可能是基于本地磁盘、NAS，或者是专为降低延迟而优化的对象存储加速器，将负责实现快速的数据加载。

我在 2021 年 SIGMOD 发表的"LogStore: A Cloud-Native and Multi-Tenant Log Database"论文中就使用了这一方案。LogStore 是一个内部使用的云原生日志数据库，底层采用对象存储，为了降低写入延迟，写入的日志先通过 raft 协议刷到 3 副本的本地 SSD 中即提交，再由 Leader 节点将数据写入到对象存储中。

硅谷初创公司 Neon 也采用了这一方案。Neon 宣称是开源版本的 Aurora Postgresql，采用计算与存储分离架构对开源的 PostgreSQL 数据库进行改造。其存储层由 Safekeeper 和 PageServer 两个组件构成。Safekeeper 负责持久化和复制 WAL 记录，使用 Paxos 协议实现分布式一致性，将 WAL 记录保存在 EBS 上，并将提交的 WAL 记录传递给 PageServer。PageServer 负责将 WAL 记录应用到 Page 上，生成 Page 的快照，并将 Page 存储在对象存储中，缓存 Page 在实例存储上，以提高读取性能。在 Neon 的实现中，数据库的主存储是对象存储，写缓存是 EBS，读缓存是实例存储。

新的商业模式契合对象存储

近年来，公共云数据库服务正日益倾向于提供 Serverless 运行模式，以迎合现代开发者对弹性的需求。早在 18 年，AWS 就推出了 Aurora Serverless，22 年，AWS 又推出了 Aurora Serverless V2。在今年的 AWS Reinvent 大会上，AWS 一口气发布了三款 Serverless 数据库产品：Aurora Limitless Database、ElastiCache Serverless 和 Redshift Serverless，已经隐约摆出了全系产品 Serverless 化的架势。

Serverless 数据库之所以深受青睐，一方面是因为其低门槛的使用成本，另一方面则是因为其能够在面临突如其来的机会时提供快速且灵活的扩展能力。因此，Serverless 数据库成为了初创公司和那些流量增长高度不可预测的创新产品的理想选择。

本质上 Serverless 数据库采用了一种”按实际使用付费 (pay as you go)“的新商业模式。在 Serverless 数据库模式下，云的 dbPaaS 软件会实时追踪用户的活动，从而精确计算其资源使用情况。无论是执行 SQL 命令所消耗的 CPU 时间，还是读写操作产生的 IO 次数，亦或是数据存储的总量，所有这些都被量化为资源单

位（例如计算单元 CU、请求单元 RU），并据此计费。技术如何支撑这种新的商业模式，比如数据库如何根据工作负载的变化智能的扩缩容取决于 dbPaaS 和内核的实现。

在业界实现 Serverless 架构的常见方法之一是通过给基于计算与存储分离的架构增加自动伸缩（Auto-Scale）功能。因此云原生数据库如 Aurora 和 PolarDB 都能比较快的推出其 Serverless 版本。在这样的架构下，存储层通常设计为多租户模式，以实现资源共享和成本效率；而计算层则通常是单租户的，确保了性能和隔离性。随着数据库负载的增加，这种模式允许计算节点通过弹性扩容迅速提升处理能力。相对地，当负载减少时，计算节点可以弹性缩容，甚至完全停止，以优化资源使用并减少成本。

CockroachDB 也采纳了这种 Serverless 架构。为适应这种模式，他们对底层的 KV 存储层进行了重构，转变为支持多租户的架构，并且 SQL 节点和 KV 节点不再运行在同一台服务上，走向计存分离架构。

由此可见，存储池化是 Serverless 数据库架构中的核心设计原则。只有通过存储池化，才能做到按存储的使用量计费。云厂商的产品 Aurora 和 PolarDB 实现这一机制的策略是构建他们自己的分布式存储集群。然而，这种做法要求云厂商在产品推出前进行大规模的一次性投资，并且在产品初期推广阶段承受由于存储使用未达预期而产生的潜在亏损。

然而，对于那些希望利用开源软件自行搭建 Serverless 数据库服务的大型企业用户、以及提供 Serverless 数据库服务的小型的第三方数据库厂商来说，却存在一个潜在的陷阱。如果构建 Serverless 服务之前需要首先投资搭建一个存储池，他们就陷入了传统模式——即预购硬件并在其基础上构建数据库服务，这种做法与 Serverless 的理念相悖。最理想的技术方案是做到不囤货，实际使用多少存储，就向云厂商付多少存储费用。

此外，Serverless 数据库的设计理念自然包含了 Region 级别的容灾功能，这是其另一项核心能力。用户选择 Serverless 服务，意味着他们已经不再关注运行数据库的具体服务器或所在的可用区（AZ）。在这种服务模式下，用户不太愿意去单独从三个不同的 AZ 购买一套 Serverless 数据库，再手动设置数据同步——这种方式与 Serverless 的易用性相悖。因此，Serverless 数据库必须确保其计算节点池和存储池都在跨 AZ 环境中提供无缝的容灾能力。

存储池化、按使用量计费，以及 AZ 级的灾难恢复，Serverless 数据库对存储的需求正是对象存储产品的关键特性。而传统云数据库使用的云盘在成本、弹性能力、容灾能力上均弱于对象存储产品。基于这些考量，我们预见在按使用量计费这种新的公共云商业模式被广泛接受后，未来的 Serverless 数据库都会把对象存储做为首选。

展望：以对象存储为中心的新架构

随着数据库存储基础设施向对象存储的逐步迁移，我们还可以预期新架构里会出现以下几个方面的变化：

行列混合存储格式

数据库存储引擎的数据格式将从纯行存向行列混合格式变化。基于 B+ 树的存储引擎可以采用调高数据页大小，并采用页内列式存储的技术来实现。而基于 LSM 的存储引擎在这块具有更大优势。

将行存数据转换为 Pax 格式（行列混合存储格式）存入对象存储有多重好处：

首先，它显著降低了存储成本并减少了数据加载所需的时间。例如，以行列混合存储格式进行数据转换，可以将 1TB 的数据压缩到仅为原始体积的 20% 至 40% 存储在对象存储中。借助 25Gb 的内网带宽，加载并预热这些压缩数据到缓存的过程大约只需要 100 秒。

其次，采用 Pax 格式还能减少数据库内存缓存的消耗，比如说原来计算节点需要 32GB 的内存，现在只需要 12GB 内存就能达到同样性能了，进一步的降低计算节点成本。

最后，数据库引擎可以借助 Pax 格式实现混合事务 / 分析处理（HTAP），Google 的 Spanner 数据库就是一个例证，它采用了基于 Pax 的 Ressi 存储格式，支持 HTAP 操作。

OLTP 与数据湖的深度融合

传统上，将在线事务处理（OLTP）数据迁移到在线分析处理（OLAP）系统的常规方法依赖于 ETL（提取、转换、加载）流程。然而，ETL 的过程不仅管理负担重，而且增加了一个容易发生错误的环节。鉴于这些挑战，近年来业界出现了向"NoETL"解决方案的转变，将 ETL 过程内置在数据库中，以提高用户进行数据管理和分析的易用性。例如去年 AWS 推出了从其 Aurora 服务直接到 Redshift 数据仓库的 NoETL 集成。

而在 OLTP 数据库内核中，原生支持将全量数据以行列混合存储格式持久化并写入对象存储，会更近一步，促进 OLTP 数据库与现代数据湖技术的协同工作。利用技术如 Iceberg、Hudi 和 Delta Lake，OLTP 数据库可以无缝地将在线数据直接写入数据湖环境，帮助企业能够更简单、实时地管理和分析其数据资产。

使用 K8s 管理数据库变得普及

自从 K8s 问世以来，管理有状态服务，尤其是数据库，一直是个充满挑战的领域。一方面，随着 KubeBlocks 以及各种数据库 Operator 等开源的数据库控制面管理软件的发展，我们拥有越来越多的工具支持在 K8s 上对数据库进行高效管理。此外，当数据库的持久性是通过 K8s 外部的对象存储来保证时，对 K8s 中的数据库 Pod 进行高可用切换、节点迁移、数据迁移、备份等各种管理任务的复杂性会得到进一步减轻，执行效率也会更高。两个方向的发展相辅相成，会推动在 K8s 上管理数据库的普及。

公共云中的关系型数据库正处在一个转型的临界点，即从依赖云盘向利用对象存储的新时代迈进。过去的 RDS 模式是云托管，客户需要预先选择硬件配置（云服务器、云盘、VPC 和负载均衡器），然后在这些硬件上部署数据库内核。而在对象存储时代，没有预置 IaaS 的限制，数据库内核进一步云原生化更有弹性更加强大，这会是数据库领域近期最大的技术变革。

作者简介：

曹伟（鸣嵩），云猿生数据创始人 & CEO，发起了开源云原生数据库管控平台 KubeBlocks 项目。前阿里云数据库总经理 / 研究员，云原生数据库 PolarDB 创始人。中国计算机学会数据库专委会执行专委，中国计算机学会开源专委会执行专委，获得 2020 年中国电子学会科技进步一等奖，在 SIGMOD、VLDB、ICDE、FAST 等数据库与存储国际顶级学术会议发表论文 20 余篇。

今日好文推荐

网游新规致腾讯网易市值半天蒸发5200亿；吴泳铭“爆改”淘天：管理层全换成有功绩的年轻人；字节年收入超腾讯、逼近Meta | Q资讯

创始人 3 天狂砍 5 万行代码后，应用程序更快、更易使用了

选择哪种编程语言已经不重要了，只提倡程序员下班后“多看看书”提升竞争力是误人子弟 | 独家专访亚马逊 CTO

一代更比一代强，AI 时代的至强如何为云服务保驾护航?

活动预告

12 月 28-29 日， 2023 年最后一场 QCon 全球软件开发大会 & QCon 中国 15 周年 Party 即将落地上海。除了精彩演讲之外，还有 7 大亮点活动，等你一起来玩～

① 承载着最前沿生成式 AI 技术之旅「 下一站 GenAI 」；

②「云原生时代的数据架构与性能提升」专场免费报名；

③五场高端闭门交流会议；

④大模型精彩公开路演，免费参与；

⑤大模型展区新升级，10+ 大模型及应用厂商现场 Battle；

⑥「2023 数字化践行者年度力量榜」榜单评选结果正式发布；

⑦ 两大抽奖活动，100% 中奖率!

InfoQ
将在12月28日 09:00 直播

预约

QCon 15周年｜年终盛会落地上海站，主题演讲免费对外直播中～

视频号

现日程已全部上线，点击「阅读原文」即可自行定制您的参会议程，更多大会相关资讯可扫描下方二维码进行了解。咨询购票可联系票务经理 18514549229，锁定最新优惠。

Read more

People who liked this content also liked

选择哪种编程语言已经不重要了，提倡程序员下班后“多看看书”提升竞争力是误人子弟｜独家专访亚马逊 CTO
InfoQ

MindsDB，一个超神奇的python库
聪聪编程

一个数据实时同步方案
小羊架构