原子钟在数据中心的作用：原子从对数据造成不利影响到带来各种益处的转变过程

Microchip Technology Inc.
频率与授时系统业务部
产品营销经理
David Chandler

利用原子钟授时现已成为数据中心不可或缺的组成部分。目前，通过全球定位系统（GPS）和其他全球导航卫星系统（GNSS）网络传输的原子钟时间已使全球各地的服务器实现了同步，并且部署在各个数据中心的原子钟可在传输时间不可用时保持同步。

无论是由于系统需求还是合规性，这种出色的同步性能都至关重要，可确保每年在全球范围内收集的数据（以泽字节为单位）能够得到有效存储并用于许多应用。原子的量子性质可保持精确的时间，是确保未来能够以更快的速度处理更多数据的关键所在，而具有讽刺意味的是，就在几年前，原子的量子性质还被视为提升数据处理能力和速度的最大阻碍。

1965 年，Gordon Moore 预测集成电路上的晶体管数量每年翻一番。这一数字最终被修改为每两年翻一番。随着晶体管密度的增加，速度有了显著提升，成本和功耗也不断下降。

在 1965 年，人们可能很难想象，2021 年时在一个半导体上布置 500 亿个晶体管是一种现实需求，但正如半导体技术随着时代不断发展，应用需求也在不断变化。手机、金融交易和 DNA 图绘制等应用都非常依赖单片机每秒可执行的运算次数，而这一数字与芯片上的晶体管数量密切相关。

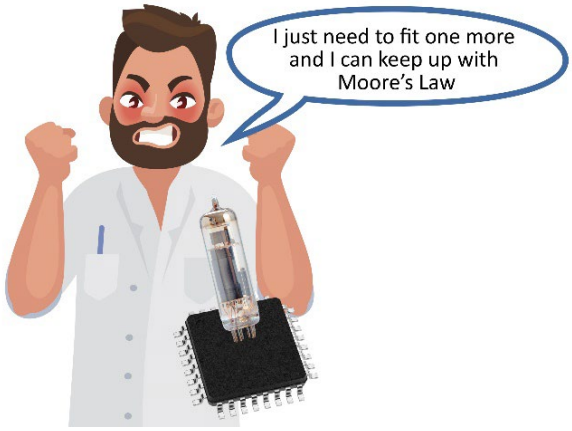

图 1. 极具讽刺意味的图片：工程师试图遵循摩尔定律

摩尔定律的消亡

遗憾的是，由于物理学限制，摩尔定律正在迅速走向终结。随着晶圆生产工艺节点现已达到 10 纳米以下，晶体管的大小仅为硅原子的 10 到 50 倍左右。在这个尺度上，原子和自由电子的大小以及量子特性显著阻碍了晶体管大小的进一步缩减。从本质上讲，可以将原子视作推翻这一定律的最终原因。

尽管摩尔定律终将消亡，但是，对提高处理能力的需求却不断增加。随着物联网（IoT）、信息流服务、社交媒体帖文和自动驾驶汽车的出现，每天产生的数据量会继续呈指数增长。

据估计，2021 年每天产生的数据量为 2.5 艾字节（2,882,303,761,517,120,000 字节）。当前使用的艾字节数据库每秒可处理超过 10 万个事务（一个事务包含许多次运算），而在可预见的将来，数据库的规模和每秒处理的事务数将持续增长。

同步机器

数据量的这种爆炸式增长，再加上数据必须达到的写入、读取、复制、分析、操作和备份速度，这些因素要求数据中心架构师找到一种能够绕过摩尔定律终结的方法。对于采用分布式数据库的数据中心，架构师采用了水平扩展方法，即将数据库分布在一个集群中的多个服务器上，而不是整个数据库驻留在一个服务器上。

在这种配置下，集群本质上用作一台巨型机器，因此系统的大小和速度现在受到数据中心的外形尺寸而非原子大小的限制（接招吧，原子！）。

软件工程师现在的职业是编写能够实现水平扩展的代码。但是，要使各种软件都正常工作，所有机器都必须同步，否则会违反因果关系的概念。

什么是因果关系？举个最简单的示例。假设您用两台摄像机来记录 100 米短跑的图像，每台摄像机都有自己的内部时钟。第一台摄像机位于起跑器上。第二台摄像机位于终点线上。两个传感器都在进行连续拍摄，并用各自时钟的时间给每个图像添加时间戳。

要确定比赛中获胜的短跑选手的正式成绩，将检查第一台摄像机的图像以了解第一位选手离开起跑器时的时间点，然后用终点线上的摄像机图像上该选手冲过终点线时的时间减去该时间戳。

要实现此目的，两台摄像机的同步精度必须都达到可接受的偏差水平。如果时钟的同步精度只有±0.05 秒，那么便无法确定成绩为 9.6 秒的选手是否确实打破了 9.58 秒的世界纪录。如果它们与体育场时钟的同步精度只有±5 秒怎么办？

想象一下这样的场景：从体育场的主时钟观察，一场比赛正好在下午 12:00:00:00 开始。第一位选手在下午 12:00:09:60 时冲过终点线。从体育场主时钟的角度来看，正式比赛成绩是 9.6 秒。

但是，如果第一台摄像机的时钟正好快 5 秒，而第二台摄像机的时钟正好慢 5 秒呢？比赛将在下午 12:00:05:00 正式开始，在下午 12:00:04:60 结束。比赛将在开始前 0.4 秒正式结束，这会打破世界纪录并推翻物理定律，目前的纪录保持者很有可能会不公正地遭到所有赞助商的弃用。

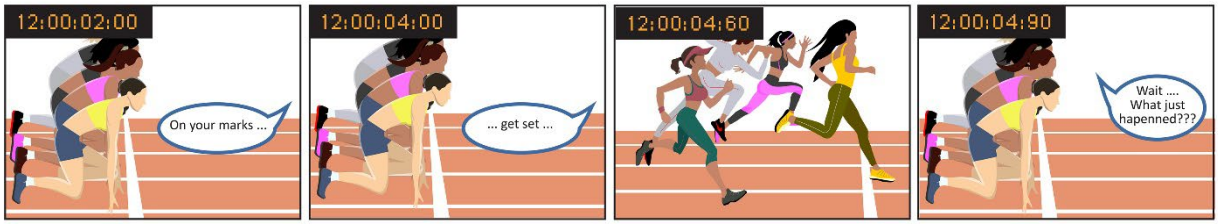

图 2. 时钟偏差会导致因果关系问题。在这种情况下，比赛在开始前就正式结束了。

将因果关系应用于数据库

同样的因果关系原则在数据库中也十分重要。事务记录更新必须按照它们发生的顺序出现在数据库中。如果您期望在通过直接取款支付每月房贷之前直接存入自己的工资，而银行的数据库没有按正确的顺序记录这些事务，那么您可能会被收取透支费。在一台机器上，因果关系错误很容易防止，但在多个服务器上，每个服务器都有自己的内部时钟，服务器必须同步并为每个事务加上时间戳。

要实现此目的，必须有一个服务器充当参考时钟，就像体育场的时钟，它必须采用最大程度减小每个服务器时钟的时间误差的方式，将时间分配给每个服务器。每个时间戳的偏差（比赛中为±5 秒）形成一个时间包络，其长度为时钟偏差的两倍（比赛中为 10 秒）。对于分布式数据库，一秒内可以容纳的非重叠时间包络数量应当至少与系统预期的每秒事务数量大致相同。

概率、因果关系的关键性和实现成本最后都会在最终解决方案中发挥作用，但这种关系是一个很好的起点。时间戳偏差为±1 毫秒的系统将具有 2 毫秒的时间包络，一秒内最多可容纳 500 个非重叠时间包络。此系统可以支持每秒执行约 500 个事务。

NTP 和 PTP 的不足

以太网授时技术也称为网络时间协议（NTP）和精确时间协议（PTP），用于同步数据中心的分布式数据库中的所有服务器。这些协议可以确保局域网能够以亚毫秒（NTP）或亚微秒（PTP）的偏差来分配时间，从而支持每秒执行数千（NTP）或数百万（PTP）个事务。

遗憾的是，即使凭借这些解决方案可以绕过原子带来的摩尔定律消亡，物理学仍以光速的形式在分布式数据库的道路上设置了另一个障碍。

试想一下，一个使用 PTP 进行准确同步的分布式数据库在加州圣何塞运行，每秒可轻松执行 100,000 个事务，且不会产生任何因果关系问题。一位数据库架构师正坐在自己位于纽约的办公室里，他的老板要求他更新大量记录。

这名架构师希望能够充分利用其新数据库并展示系统的能力。他计划每秒执行 100,000 个事务。

为了根据请求更新记录，他创建了一个简单的事务，即仅当第一个记录的值大于第二个记录时，才会将第一个记录的值与第二个记录相加。如要达到这一目的，他必须对这两个记录发出读取请求。然后，他在纽约的本地机器对这些值进行比较，然后在需要时向第二个记录发送写命令。

完成此操作后，他想要接着执行下一个事务，即将第三个值与新的总和进行比较。如果新的总和大于第三个记录，那么将使用第三个记录替换总和。他想对 600 万条记录重复此操作。由于数据库每秒能够处理 100,000 个事务，他认为此任务将在大约一分钟内完成。他告诉老板，他将在五分钟内更新记录，然后离开去喝杯咖啡。

喝咖啡的时候，他读到一个故事，内容是新的百米短跑成绩是负 0.4 秒，这违背了物理定律，并且之前的纪录保持者因为失去了所有的代言费正在起诉体育场负责人。架构师自顾自地笑了起来，认为体育场应该聘请他作为同步专家。

五分钟后他回到办公桌前，沮丧地发现他的数据库更新只完成了不到 1,500 个事务。他难过地意识到自己的错误，并准备将自己的简历发给那个体育场，他希望他的 PTP 部署不会出现同样的问题。

问题出在哪里？光速将纽约和圣何塞之间理论上最快的数据传输速度限制在 13.7 毫秒。

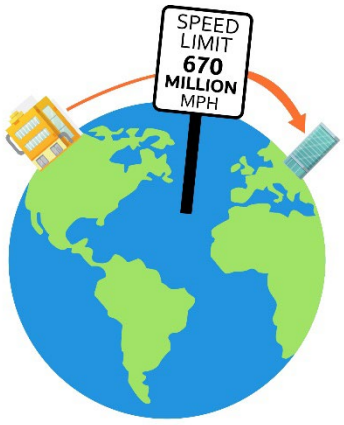

图 3. 光速对两点之间的数据传输速度施加了理论上的限制

距离问题

遗憾的是，现实世界的事务处理速度甚至更慢。即使两个地点之间有专用的光纤链路，光纤的折射率、光纤的实际路径和其他系统问题也会延长传输时间。因此，仅仅从纽约传输一次，就需要 40 到 50 毫秒的时间才能到达圣何塞。

但是，此事务中有四个独特的操作。有两个可以同时发生的读操作，随后必须将它们发送回纽约。往返过程需要 80 到 100 毫秒。然后，在对两个值进行比较后，就会发出写操作，并且必须发回写确认以指示写操作已完成，然后才能开始下一个事务。

突然之间，数据库每秒能否执行 100,000 个事务已无关紧要，因为距离将系统每秒的处理能力限制为不超过 5 个事务。要完成 600 万个事务，此系统需要 13 天的时间，这样便有足够的时间再喝几杯咖啡，甚至更新一份简历。这种延迟称为通信延迟。

规避延迟

但就像摩尔定律一样，数据库架构师想出了规避延迟的方法。在用户附近创建数据库副本，这样他们便可随意使用数据，而不必将信号发送到全国各地。

定期比较和协调复制以确保一致性。在协调过程中，事务时间戳用于确定事务的实际顺序，并且当存在不可协调的差异时（例如事务时间包络重叠时），有时会回滚记录。减少时钟偏差可以减少复制的实例中不可协调的差异数量，因为时间包络增多会减少重叠的概率。这可提高效率并降低数据损坏概率。

但现在，时间戳不仅在每个数据中心内部必须做到精确，在不同的数据中心之间也必须精确，这些数据中心可能相隔数千英里，并通过云相互连接。由于需要一个偏差极低且在两个地点均可随时获得的外部参考，因此这项任务变得愈加困难。

下至原子级别

此时，数据库架构师以前的敌人“原子”登场。当原子忙于废除摩尔定律时，其亚原子粒子却在忙于自旋。原子核内的中子和质子一直在旋转，而与此同时电子则一边忙于围绕原子核公转，一边自旋。这类似于地球在绕太阳公转的同时自转。

电子可以围绕自身的轴顺时针或逆时针自旋。考虑到人体内约有 7*10^27（7 后面有 27 个零）个原子，所有亚原子粒子都在我们体内自旋，令人惊讶的是我们并没有一直头晕目眩。（注：亚原子粒子并不是真的在忙着自旋和公转，它们实际上是在忙着给我们提供概率波函数和磁相互作用，这会让我们获得类似于它们进行自旋和公转时的结果。但是，如果想到所有的自旋会让您头晕目眩，那么试图理解量子物理学的现实肯定会更令人厌恶。）

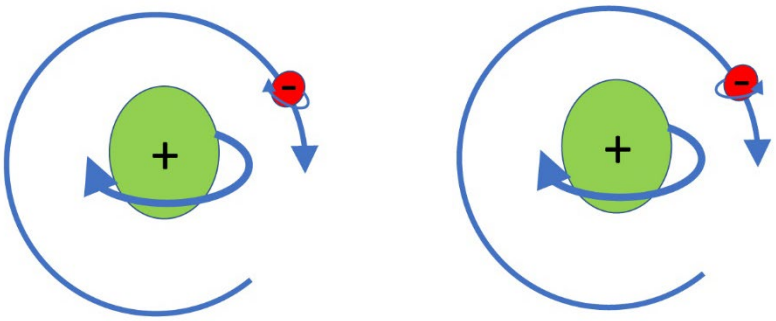

图 4. 具有核和价电子的概念性原子，具有核自旋、电子自旋和轨道自旋

如果电子吸收特定精确频率的微波辐射，绕电子轴的自旋方向会改变。如果地球上发生这种情况，太阳会突然从东方落下，从西方升起！

原子钟这种机器专门用于检测电子自旋状态，然后通过微波辐射改变方向。频率变化取决于元素、同位素和电子的激发态。

在机器确定频率（即所谓的超精细跃迁频率）后，便可将周期确定为频率的倒数，这样便可计算周期数来确定经过的时间。国际上对秒的定义是诱导铯原子轨道外层内电子的超精细跃迁所需的 9,192,631,770 个辐射周期。

原子钟是世界上最稳定的商用时钟。一副纸牌大小的原子钟称为芯片级原子钟（CSAC），其 24 小时内的漂移为百万分之一秒，而冰箱大小的原子钟称为氢微波激射器，其 24 小时内的漂移仅为十万亿分之一秒。巧合的是，十万亿分之一也大约是氢原子半径与百米短跑选手和现已失业的纽约数据中心架构师身高的比值。

凭借这些原子钟提供的精度，可以为在东京、伦敦、纽约、廷巴克图或世界其他任何地方的数据中心运行的分布式数据库提供大约 50 万到 500 亿个非重叠时间包络。

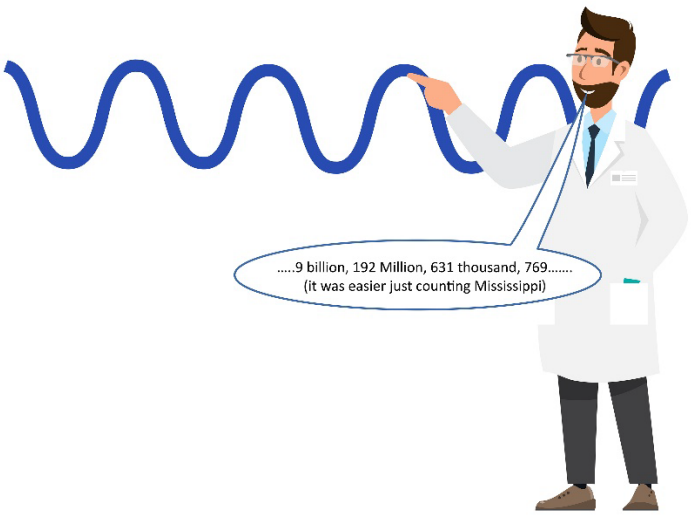

图 5. 单位“秒”是通过计算铯超精细透射辐射频率的 9,192,631,770 个周期来定义的

时间的分配

时间如何从这些原子钟到达所有数据中心？协调世界时（UTC）是通过卫星、光纤网络甚至互联网分配的全球时间。UTC 本身源自位于世界各地的国家实验室和授时站的一系列高精度原子钟。UTC 的提供组织会收到一份报告，其中载明了源自这些时钟的 UTC 时间以及它们各自与计算出的 UTC 的偏移量。然后，这些实验室和其他设施将时间传送到世界各地。

UTC 报告每月公布一次，告诉这些国家实验室他们在上一个月与 UTC 的微小时间偏移量。从技术上讲，直到事发一个月后，我们才知道准确的时间偏差。更糟糕的是，由于地球自转和我们与可观测恒星的相对位置的变化，UTC 会定期增加额外的秒数，即跃迁秒。虽然这可使地球与宇宙保持一致，但它会引起数据中心和 100 米短跑成绩的混乱。

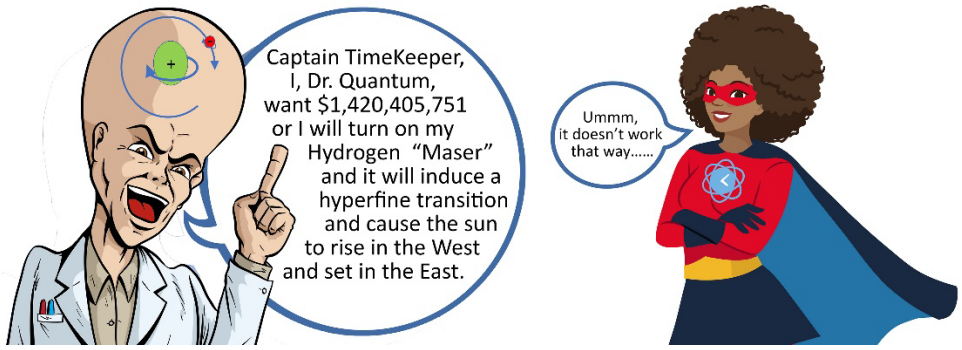

图 6. 氢微波激射器中产生的超精细跃迁频率为 1.420405751 GHz，将导致电子自旋反转

GNSS 登场

数据中心用来获取 UTC 的常用方法有两种：通过互联网使用公开的 NTP 时间服务器，以及通过卫星使用 GPS 或 GNSS 网络。虽然在分布式数据库的早期部署期间，通过互联网上的公共 NTP 时间服务器进行授时很常见，但固有的性能、可追溯性和安全问题已经促使人们放弃了这种解决方案。

尽管 GPS 和其他 GNSS 通常被视为定位和导航系统，但它们实际上是精确授时系统。接收器的位置和时间取决于信号以光速从多个卫星传输到接收器的传输时间。极具讽刺意味的是，这是物理学原理引发问题的又一个案例（此案例中是光速而不是原子），但也有助于解决问题。

这些卫星有自己的机载原子钟，这些原子钟与从地面站传输到卫星的 UTC 同步。利用这种方法获取 UTC 可以提供 5 纳秒范围内的时间偏差，从而实现每秒 1 亿个时间包络。

这种方法比公共 NTP 服务器更可靠、更精确，虽然这些信号可能会被太阳风暴或蓄意的信号干扰等事件中断，但在出现这些信号时，可以在每个单独的数据中心放置与卫星信号同步的备份时钟，以便在中断期间提供所需的偏差水平。

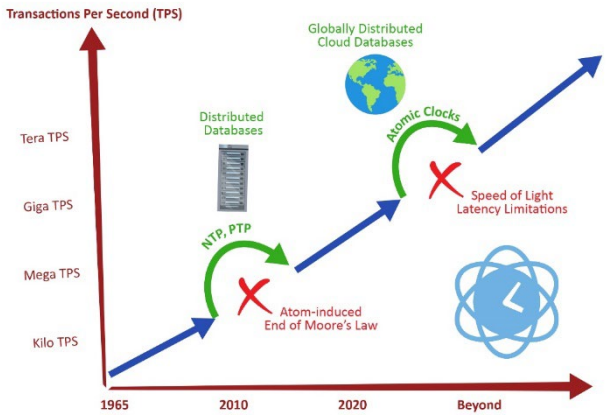

图 7. 数据库事务速率的发展历程以及实现和禁用的技术

下一步：跃迁电子

随着未来对获取、存储和处理数据的需求不断增加，我们需要具有极低偏差的新型原子钟技术和时间传输系统。目前，国家授时实验室正在开发一种新型原子钟，用于研究电子跃过轨道层时发生的光学跃迁。这些原子钟的频率稳定性可达到万亿分之一赫兹，最终将用于重新定义*秒*这个单位。

通过专用光纤链路或机载激光器实现的信号传输已经显著提高了传输精度。凭借这些不断涌现的创新数据，原子和光将继续它们之间复杂的爱恨交织关系，从而能够以更快速度处理越来越多的数据，而不会出现一致性或因果关系问题。