InComparison

Bloor

IBM pureScale Application System 与 Oracle Exadata X2-2

Bloor Research 的 InComparison 白皮书
作者：Philip Howard
出版日期：2010 年 11 月

我们开展此项活动是为了能够找出 IBM 胜出的一些领域，以及 Oracle 所专长的其他领域。令我们震惊的是，几乎在我们研究的各个领域，IBM pureScale Application System 都胜过 Oracle Exadata。

Philip Howard

目录

执行摘要

本白皮书由两部分构成：本章节连同结论部分，适用于对技术没有或仅有有限的知识或兴趣的高管；其他章节，适用于对技术有一定了解的人士。特别是本章节的信息，在其他章节中也有重复，但此处较为简略。

本白皮书连同其姊妹篇（《IBM Smart Analytics Systems vs Oracle Exadata X2-2》）的基本主题是提供一个 IBM 和 Oracle 产品分别在在线事务处理 (OLTP) 和数据仓库上的全面对比。

Oracle 认为对于这两种需求只需一个解决方案即可解决，Oracle Exadata 可同时满足上述两种需求；即便如此，我们认为（我们认为 Oracle 也会同意）两种环境的需求有很大不同。IBM 的观点有所不同，他们认为要对不同的领域有不同的侧重，因此针对 OLTP 推出了 IBM pureScale Application System，针对数据仓库推出了 IBM Smart Analytics System。

实际上，IBM 的方法并非如此简单。在 OLTP 方面，有两种可行的方法：授权 DB2 pureScale 或 IBM pureScale Application System，其中后者包含前者，但还（可选的）包含 WebSphere Application Server，并且其整个的软件包围绕 AIX 操作系统和 IBM POWER7 服务器而构建。事实上，如果想要更偏重 DIY 的方法或想要在基于 x86 的硬件或 Linux 操作系统上运行，则使用 DB2 pureScale 授权，但如果想要一个可直接运行的完善系统，则使用 IBM pureScale Application System。类似的概念也适用于 IBM 的数据仓库产品。

IBM 称这些方法为“工作负载优化系统”。也就是说，这些产品和特定的软件包，已根据其具体的环境进行设计并优化。

我们在此白皮书中想要弄清楚的问题是：这两种方法哪一个最好？当然，您可以通过理论的争论支持 Oracle 或 IBM 的方法，这种争论可以一直持续下去，没完没了。一个概念是否更好与理论上的观点无关，哪一个具有最好的性能、可扩展性、更易于管理并节约成本才能说明问题。

我们主要关心的是 Exadata 对于超大组织之外所有的 OLTP 环境而言超规格。尽管我们知道有些用户使用 20Tb OLTP 环境，但很少有企业需要千兆字节以上。然而最小的 Exadata 配置就有 21Tb 的磁盘容量，能够提供 6Tb 的可用容量，这还是未考虑（可选）压缩之前的情形。因此，Exadata 的容量比大多数用户需要的容量都要大，同时，您当然也要为多余的容量买单。此外，Exadata 中的存储经过了特定的配置，用以提高数据仓库环境下的性能，此环境下对大量数据进行的是典型的顺序读取，而不是 OLTP 中典型的随机访问。Exadata 配置中包含有所谓的闪存，这确实提升了 OLTP 的性能。但是，我们认为这种性能提升不足以抵消那些额外费用（上文提到超大环境除外）。如果我们只是讨论 OLTP，那么也许在 DB2 pureScale 和不包含 Exadata 的 Oracle Real Application Clusters 之间进行比较会更恰当。如此，Oracle 环境就不算超规格。

执行摘要

还应该认识到事务处理系统不（即使有也不多）是孤立的存在。实践中，总是有用于支持事务处理的大量报告。这里我们希望 Exadata 的子系统相较 pureScale 能够提高性能。此外，实际情况经常是其他的应用程序也和基于 OLTP 的应用程序安装在同一个系统上（一种统一的方法）,这些应用程序也可从支持数据查询的非事务密集型 Oracle 环境的功能中受益。此外，如果多个应用程序运行在单一 Exadata 系统中，那么超规格问题就定将不复存在。

除了容量问题，IBM 的集群技术与 Oracle 相比又会出现什么结果？Oracle 的技术显然不够成熟，尽管绝大多数技术都源于 IBM 大型机数据库安装经验，并且这些经验得到了广泛认可。例如，IBM 使用集中锁定管理，该技术用于 IBM 在大型机上的 DB2 实施。与之相对应，Oracle 使用分布式锁定管理。对比结果显示，IBM 集群的网络流量更少，并且故障恢复速度更快。也有人认为 IBM 具有后动优势。例如，Exadata 使用 InfiniBand 互连。IBM 也如法炮制。然而，因为总是用于实时应用集群，为支持后向的兼容性，Oracle 在 InfiniBand 上使用相同的传输协议。这与 IBM 的协议相比，需要在更高层的通信栈操作，这意味着 Oracle 运行环境网络需要更大流量。在性能方面如果与 IBM 的方法相比显然有所降低。在该层面上，还有另外一个特点值得注意，那就是用于 DB2 的应用程序无需知道集群的配置及其详细信息。相反，用于实时应用集群的应用程序则需要具有集群感知功能，以进行性能优化。

就 OLTP 的性能而言，IBM 目前在业内拥有 TPC-C（参见第 7 页中的注释）基准测试的最佳性能和最佳性价比两项记录，这两项记录之前为 Oracle 所有。供应商总是在统计中交替获胜，但从 Oracle 的上一次记录到 IBM 的记录（2009 年末至 2010 年仲夏）中间间隔不到 9 个月，并且，两项记录的差距较大。所有这些都说明，基准测试是人为定制的，并不代表您的数据或环境。这些测试确实有利于帮助供应商在客户体验产品之前确定极端性能，但是，至于对产品和供应商的对比而言，也只是象征性的，不能太过当真。

另外一个区别在于 IBM 升级 pureScale Application System 的方法。Oracle 对 Exadata 的规定是：您可以使用 ¼ 机架、半机架或多个全机架。如果您想要添加新的处理节点，就必须添加一个或多个附加存储服务器，除了支付硬件费用之外，您还必须支付增加的运行 Exadata 的授权费用，每个磁盘 1 万美元（每个服务器 12 个磁盘），外加 22% 的维护费，所有这些都是因为您需要更多处理能力，尽管您并不需要额外的磁盘容量！客观上说，Oracle 已经认识到这一点并且宣布 Exadata X-2-8 将解决此问题，但是：a) 目前还无法实现 b) 只适用于全机架，因此将需要巨大的容量来启动。

执行摘要

相反，IBM 未强制要求任何特别的存储。如果愿意，可以继续使用现有的基于 SAN 的存储。如果需要可以选择固态硬盘（闪盘），但不是强制要求的。而且，IBM 提供所谓的“按需工作负载”：可以增加额外的服务器到集群（除了建立物理连接外仅需一步完成），并可升级内核数量（根据系统，从 4 或 8 升级到 16 个内核）和内存容量（从 32Gb 到 48Gb 或 64Gb），并可临时使用任何一个服务器，然后将其关闭或从集群中移除该节点（同样只需一步）。至于授权费，仅当额外的容量启用时向您收取费用，一旦不再使用，授权费用则恢复到之前的水平。

就定价本身，IBM 针对不需要任何软件许可的 Oracle 环境下的 pureScale Application System 收费。也就是对于那些使用通用授权协议 (ULA) 的客户。对于用户，一个小型的 pureScale Application System 同一个 ¼ 机架的 Oracle Exadata 系统拥有几乎相同的处理器性能和存储容量，两者的价格也相当，但中型和大型配置的系统则比对方 ½ 机架和全机架有更大的价格优势。当然，如果没有 ULA，那么价格比较中 IBM 则占足优势。但是请记住，价格清单常常是变幻无常，有时不能反映实际情况。

也许两个系统之间最大的不同表现在可管理性和灵活性方面。我们已经就后者在可用的配置和磁盘容量以及按需工作负载方面进行过讨论。至于可管理性，首先 pureScale Application System 更易于安装、扩展和缩减，并且通过使用集中锁定管理，在节点出现故障时不会导致集群冻结，而这种情况是 Oracle Exadata 无法避免的。最后这一点对于能对业务产生影响的 OLTP 环境相当重要。

最起码同 pureScale 相比在查询和报告方面 Oracle Exadata 应提供性能效益，这样 Oracle 才会适合在整合环境中运行。然而，我们不认为其闪存（用于提供 OLTP 性能效益）带来的优势能补偿使用分布式锁定管理而非集中锁定管理所带来的不足。在其他方面，pureScale Application System 似乎能够提供比 Oracle Exadata 更加显著的优势。

系统说明

在进行任何类型的对比前，我们需要清楚地了解每个产品的架构以及每个产品中包含和不包含的功能。

IBM DB2 pureScale

DB2 pureScale 是 DB2 数据库的许可选项，在 IBM Power System 或 System x 系列上运行，前者以 IBM 的 POWER7 硬件（运行 AIX）为基础，后者则以 x86 处理器（运行 Linux）为基础。

DB2 集成了 pureScale 的新功能后，可在集群硬件上运行。除支持集群环境所需的功能外，DB2 的功能并无任何变化。最值得注意的是，这表示 DB2 pureScale 拥有了一个共享磁盘架构，而非在分布式环境中一直沿用的无共享架构。然而，这并不是说 DB2 发生了全新的转变，因为在大型机 System z 上的 DB2 始终在采用共享磁盘方法，而且，负责 pureScale 的 IBM 开发团队也确实大量借鉴了公司现有的大型机技术，用以引进 pureScale 产品。请注意：Oracle 始终使用共享磁盘架构，因此两家公司目前在这一问题上已无区别。

DB2 pureScale 架构包括：

- 成员，成员是 DB2 引擎的地址空间，可能存在于自身的服务器之上或存在于逻辑分区之中。用户可以在一个服务器或逻辑分区中设置多名成员，这项功能非常适用于测试或开发，但不建议在现场安装时使用。每个成员均拥有自己的缓冲池、内存区域，并可编辑内容到自己的日志文件。只有在一名成员出现故障后，另一名成员才可访问该成员的日志文件。

- PowerHA pureScale 实例（也称作集群缓存工具，或称 CF），这些实例代表设计用于协助全局缓冲一致性管理以及全局锁定管理的软件。尽管该项并非强制要求，但建议选择其中的两个实例，设定为一级和二级实例，形成双套系统，互为备份。

- InfiniBand 互连。可选择其中的两项，但在目前尚未形成双套系统。

- 存储，除与每名成员相关的独立日志文件外，存储进行共享，只有在一名成员出现故障后，其所拥有的日志文件才能被其他成员访问。

- 集群服务（由 IBM 系统和技术团队提供）和 Tivoli，前者提供 GPFS（通用并行文件系统）和 RSCT（可靠、可扩展的集群技术），而后者提供用于多平台的 Tivoli 系统自动化 (TSAMP)。

IBM pureScale Application System

IBM pureScale Application System 与 DB2 pureScale 的区别在于，前者仅运行于 Power 770 硬件之上，而后者则可以在除 Blade 之外的所有 POWER7 服务器，以及 POWER6 550 和 595 服务器、System x 服务器上运行。除此以外，如果运行 DB2 pureScale，则可以在相同配置中设置不同的 Power System 或 System x 服务器，而如果运行的是 Application System，就无法做到这一点，未来，随着更多 Power System 服务器的引进，这个目标也有可能实现。

那么，pureScale Application System 的优势是什么？首先，DB2 和 WebSphere Application Server（作为一个选项，可以与 pureScale Application System 一并获得许可）都具备对 AIX 以及 Power System 架构进行特定使用的相关功能。不过，您可能会说这是使用 Power System 相较于使用 System x 的优势，而且无论使用的平台是什么，都可以应用 DB2 与 WebSphere 间的紧密集成。因此，主要的优势（在未来版本中可能有变）在于便捷性：所有功能都预先捆绑在一起，能够随时为需要 IBM 堆栈的用户所使用，进而加快价值实现的速度。

系统说明

Oracle Exadata X2-2

Oracle Exadata Database Machine 产品实际上有两种，分别为 Exadata X2-2 和 Exadata X2-8。后者是一种仅限于全机架的系统，主要用于最大的 OLTP 和整合环境。该产品尚未推出，因此我们将重点放在 Exadata X2-2 上。这款产品包括 Oracle Database 11g Release 2、Oracle RAC (Real Application Clusters) Database 服务器网格、InfiniBand 互连、Oracle Enterprise Linux 操作系统以及 Exadata Storage Server 网格，该网格使用高性能 (600Gb) 或高容量 (2Tb) 磁盘存储，后者容量更大，但性能较低。

系统的操作方式为：数据存储在 Exadata Storage Server 网格中，存储服务器充当一定形式的预处理器，以便在将结果传递到数据库之前，通过 Oracle 称为智能扫描的优化方式访问磁盘数据。这样大大减少了数据库必须处理的数据量，在数据存储环境中效率尤为突出。为提高在 OLTP 环境中的性能，Oracle Exadata 还采用闪存用以缓存热数据。

可以在一个 Exadata 环境中运行多个数据库，也可以在一个 RAC 节点上部署多个小型数据库，也可以部署跨多个节点的较大数据库。也就是说，可以部署一个 OLTP 系统，使之与数据存储实施共享 Exadata 环境。不能与 Smart Analytics System 以相似方式共享 pureScale。这是因为 Oracle 在整个环境中都使用的是共享磁盘环境，而 IBM 使用共享磁盘的范围仅局限在 OLTP，对于数据存储，IBM 采用的是无共享架构。比较不利的一点在于，无法重新设定 Exadata Storage Server 的用途，当然这个问题并不大。如果将来决定选用其他供应商，可以重新使用 RAC 服务器和 pureScale 服务器，而如果使用的是 Exadata Storage Server，重新使用就没有那么容易，因为其功能设计比较特殊。

扩展系统

Oracle Exadata X2-2 的实施选项见表 1。

	¼ 机架	½ 机架	全机架	2-8 个机架
数据库服务器	2	4	8	16–64
Exadata Storage Server	3	7	14	28–112

请注意，当升级是唯一可用选项时：¼ 机架可以升级到 ½ 机架，½ 机架可以升级到全机架；不能使用 ¾ 机架，在全机架以上，只能使用整数机架。¼ 机架配置储存有 21Tb 的原始数据，全机架（使用高性能驱动）含有大约 100Tb 的原始磁盘容量。如果使用高容量驱动，则全机架的容量为 336Tb。每个 Exadata Storage Server 还包括 4 个闪存卡，每个容量为 96Gb，在全机架上最高可扩展到 5Tb。请注意，如不增加额外磁盘，则无法向上扩展：这意味着，如果您存在 CPU 瓶颈等问题，则无法简单地增加新的处理能力，也就是说，即便不需要，也必须拥有更多的存储容量。

当然，在实践当中，实际磁盘容量与可用磁盘容量完全是两码事。首先是确保弹性所需的磁盘镜像，磁盘镜像会将可用容量减半，其次，考虑到日志、临时空间、索引等需要，也要预留出一部分空间。Oracle 自己估计，在考虑镜像之前，有 55% 的磁盘容量实际可用于存储数据，也就是说，¼ 机架实际提供大约 6Tb 的可用空间，½ 机架提供 14Tb，而全机架则提供 28Tb。当然，这些数据均在未考虑压缩的情况下得出。

IBM 同样也提供高性能和高容量驱动，也可以扩展到大量服务器（最高可支持 128 个服务器的配置）。与 Oracle 不同的是，IBM 并不使用直接附加存储，而是提供各类基于 SAN 的存储方法，方便用户采用，特别是其 XIV 集群架构以及 SONAS 网络附加存储。这还意味着您可以重新使用现有的 SAN 类存储。在稍后就闪存的使用展开讨论时，我们将介绍 IBM 对固态硬盘及其 Easy Tier 功能的使用。

另一点与 Oracle 不同的是，您可以由内含数据少于 1Tb、规模极小的系统开始部署。我们将在适当时间讨论定价问题，但这一点仍然有力地证明，Oracle Exadata 无法缩小规模，不适用于部门环境、中小型企业甚至是一些大型公司，又或者成本过于昂贵，因为许多公司都不需要对 OLTP 配备 6Tb 的（未压缩）数据容量。在需要 Exadata 环境同时托管 OLTP 和存储能力时例外，但即便如此，仍有许多公司都不需要 6Tb 的容量，尤其 6Tb 还仅是压缩前的原始容量。

扩展系统

IBM pureScale 在服务器升级方面也不同于 Oracle。pureScale Application System 可用于小型、中型和大型配置，以两个 Power 770 3.1GHz 处理器为基础，通过 InfiniBand 互连进行关联。不同的选项区别在于，在每个服务器上激活 4 个、8 个或 16 个内核（共有 16 个内核），而激活的内存则分别为 32、48 或 64Gb（共有 64Gb 内存）。16 个内核和 64Gb 安装在每种选项当中。不仅可以增加新的节点，还可以提高服务器内激活内核和内存的数量。

此外，采用 IBM 的许可模式，您可以临时性升级内核以及增加服务器。假设在年末时您需要额外的系统容量；那么就可以在需要额外容量的时间内，在 pureScale 实施中增加另一个服务器，待高峰过后再将之移除。当然，您必须为此准备一个备用服务器，但在 DB2 的范围内，只需针对该有限时间段收取额外的许可费用。内核升级也采用相同的方法：因此，如果您为每个服务器购买了 4 个内核的许可，但一年中有一个星期需要 8 个内核，那么您就可以临时性激活这些内核，只需要为这一个星期支付额外的许可费用。有关成本问题我们将在单独的章节中介绍，但这种方法所具有的潜在优势是显而易见的。

压缩

在此，我们要补充一项关于压缩的注释。Oracle 使用两种不同的压缩类型：一种专为数据存储和存档环境所设计（详细内容将在压缩一节中介绍），另一种则被称作“高级压缩”，Oracle 将其用于 OLTP 用途的压缩。Oracle 提出对事务数据进行 2 到 4 倍的压缩。在实践中，最佳压缩率要在对数据进行预先分类后实现，而对于实时的事务数据，这一点很难实现（或难以保持）。Oracle 也可压缩索引，不过索引的压缩率稍显逊色。

IBM 不仅采用一种不同的技术处理压缩（一种字元化形式），还可对临时数据进行压缩。在此，不对两家公司所采用的不同压缩技术逐一分解，通过一个例子即可有效说明：假设您在压缩客户索引，那么每当“Bloor”出现时，Oracle 都会存储“Bloor”并附加一个行 ID。首先要了解如何操作索引。如果是 Oracle，假设有 250 条“Bloor”的条目，就会有 250 个单独的 Bloor 行 ID 对。而如果是 IBM，则只会将 Bloor 存储一次，后面随附一个由 250 个行 ID 组成的串。因此，IBM 方法是更为高效的起步方法。至于压缩，Oracle 压缩每一个 Bloor 行 ID 对，而 IBM 则对“Bloor”和行 ID 单独进行压缩，因此效率也更高（因为可以根据数据类型采取不同算法），因此我们预计 IBM 的索引压缩率也相应稍高。此外，IBM 的压缩范围还包括临时数据，因此，我们预计 IBM 的压缩率将在整体上超过 Oracle 的各项压缩率。

InfiniBand

然而，扩展并不仅仅关乎要增加多大的容量，以及能够扩展到多大的磁盘空间，还关乎于如何增加容量以及能够以多大的效率完成扩展。为了了解其中的操作方式，并且说明这两个竞争产品间的一项主要区别，有必要了解这两家公司如何使用其 InfiniBand 互连，以及如何实施锁定。

正如我们所指出的那样，两个系统均采用 InfiniBand（因为 InfiniBand 比以太网的带宽容量更高，且延迟率也明显较以太网低）。然而，IBM pureScale 是专为使用 InfiniBand 而设计，而 Oracle RAC（此处的重要元素）的设计却要回溯到上个世纪，当时 InfiniBand 还未面世。因此，两家公司在使用 InfiniBand 的方法上存在很大的不同：Oracle 采用 RDS（可靠数据报套接字），这是一种类似于 TCP/IP 的协议，依靠消息传递；而 IBM 则采用 RDMA（远程数据内存访问），这种协议的级别要低得多，专为 InfiniBand 使用所设计，顾名思义，通过这种协议能够直接访问存储于网络中不同服务器之上的主内存和缓存。后一种方法的优势包括：免除上下文切换；无需中断或消息处理；在需要通知成员页面更新时无需占用 CPU 周期（因为使用的是内存）。因此，来回响应时间一般在 10–15 微秒，而预计 Oracle 的响应时间在几百微秒或毫秒之间。另外，当对一个成员的事务进行了更新并提交了新的数据后，pureScale 能就此新提交数据通知集群中所有其他成员，而不会在其他成员上引起任何托管周期。该功能适用于所有集群规模。pureScale 通过“沉默失效”技术完成该操作，这一技术源自于 pureScale 对 RDMA 的应用。随着集群规模发展为几十或数百个成员，这一技术对应用程序的透明扩展性显得至关重要。即使在存在 100 个成员的情况下，当一个成员提交了新的数据后，其他所有成员都会接到相关通知，且不会引起任何托管周期。对于 Oracle，由于这家公司不具备与“沉默失效”相当的技术，因此在操作过程中所需要的托管周期会随着集群的增多而增长。

扩展系统

互连使用方式的进一步结果，反映在这两家公司所采取的不同锁定方法当中。简单来说，Oracle 使用的是分布式锁定管理器，而 IBM 使用的是全局性锁定管理器，后者位于 PowerHA pureScale 节点上。换言之，在 Oracle RAC 环境中，每个节点都负责系统所持有的一部分锁定，而 IBM pureScale 则集中管理所有锁定。

采取这种方法将导致两个结果。第一个结果（为保证完整性，我们将这部分内容纳入本章加以介绍）是，如使用 Oracle Exadata，在某节点出现故障时，有可能导致集群冻结，而剩下的节点要重新构建锁定列表，解决由故障节点持有的锁定。另一方面，当 DB2 pureScale 节点故障时，唯一的影响是，故障成员中的活动数据在恢复过程（自动启动）中仍保持锁定，但其他处理仍正常进行。

第二个结果是，在 Oracle 环境中，需要处理大量的信息传递工作。不过这并不是唯一涉及的因素，我们没有数据来支持这种说法，根据其各自对 InfiniBand 的使用以及实施日志管理的方式，预计 pureScale 的扩展效果要比 Oracle Exadata 更好一些（以增加额外服务器能够实现的额外性能为标准）：随着节点的增加，互连访问量将呈指数倍上升，而唯一后果就是导致性能降低（可扩展性也将因此而受到影响）。然而，应该注意的是，对于只限读取访问，而不存在锁定的存储环境，这个问题并不适用。

最后，在系统扩展环境内，存在的问题是向集群增加节点后将产生什么效果。IBM 提供自动化的加载平衡功能，能够自动识别向集群添加或从中移除节点的时间，此外，它能够酌情重新安排事务：无需进行调整。而采用 Oracle Exadata 就无法达到这一点。Oracle Exadata 也提供相似的加载平衡功能，不过在向集群添加新节点或从中移除节点时，一般需要进行初始调整。关于添加新节点的便利性，我们将在后面的章节中加以介绍。

性能

Oracle 已在运行 Siebel 软件的 ¼ 机架系统上实施内部基准测试，模拟高流量的呼叫中心环境。该基准测试模拟每小时 30,000 余位用户提出 400,000 多件事务，平均响应时间为 0.12 秒，其中 75% 的事务已从闪存分配出去（详情请见下文）。这确实令人印象深刻，但最令人感兴趣的也许是（每个数据库服务器节点的）CPU 利用率仅为 22%。换言之，如果要求的操作仅限于此，那么所测试的系统在环境方面已经超标。

一般而言（以及当我们质疑未专门针对客户自有数据及工作负载实施基准测试时），有必要记录当前的 TPC-C* 基准测试。在 2009 年即将结束之际，Oracle 以每 tpmC 2.36 美元的成本创下 7,646,486 tpmC（每分钟类型 C 的事务处理量）的记录性能。与（IBM 保持的）以往记录相比，性能提高约 25%，成本降低 16%。最近在 2010 年 8 月，IBM 以每 tpmC 1.38 美元的成本创下 10,366,254 tpmC 的性能率，一举打破了之前由 Oracle 创造的记录。与 Oracle 的记录相比，性能提高约 35%，成本降低约 42%。尤其在性价比方面，这是一次非比寻常的跨越。

然而，这些数据仅仅具有象征意义。在此之外应该注意到，任何性能评估不仅仅是各部分性能的总和，而是软件、操作系统和硬件协同工作所达到的最佳性能。从这个角度来看，重要的是认识到虽然 Oracle 11g Release 2 具备开发 Exadata 的特点，然而从本质而言，却不是针对该目的设计的。尤其是（并以此为例）Real Application Clusters 的原有特性意味着 Oracle 无法充分利用 InfiniBand。乍一看，似乎 IBM 和 DB2 在这一点上是一样的。然而，借助集中锁定和集中缓冲区，主框架上的 DB2 始终与 z 系列操作系统和硬件紧密结合：IBM 与 pureScale 进行协作（在开始于 6 年前的一个项目中），在分布式系统上采用相同的原理。

现在，我们将继续讨论每个系统上对良好性能起积极影响的特定要素。部分功能我们之前已讨论过，尤其是 IBM 在 InfiniBand 和锁定管理方面的优势。我们需要讨论的另一个硬件基础架构（与 Oracle 11g 相比对应于 DB2，稍后介绍）特点是闪存的使用。

***IBM POWER7 基准测试结果：**
IBM Power 780：成本 1.38 美元/tpmC，性能率为 10,366,254 tpmC，2010 年 10 月 13 日上市，运行节点数为 3 个，共 24 个处理器、192 个内核及 768 个线程。

Oracle Sun 基准测试结果：
Sun SPARC Enterprise T5440：成本 2.36 美元/tpmC，性能率为 7,646,486 tpmC，2010 年 3 月 19 日上市，运行节点数为 12 个，共 48 个处理器、384 个内核及 3,072 个线程。
自 2010 年 8 月 17 日起的当前结果。

TPC、TPC Benchmark、TPC-C 和 tpmC 是 Transaction Processing Performance Council 的商标。TPC-C 结果可在 www.tpc.org 上查看。

性能

闪存

IBM pureScale Application System 未配备固态硬盘，然而，您也可以选择使用。借助 DS8700 存储系统（后来在 2010 年用于中型存储服务器），可与 IBM 的 Easy Tier 技术共同使用。所应用的理念是将部分（常用）数据存储在固态硬盘（SSD 阵列）上，将其余数据存储在常规硬盘驱动器上，然后将数据酌情迁移至 SSD 阵列或从其迁移至硬盘，数据的位置部署由软件自动处理。

IBM（配备 Easy Tier）和 Oracle 在闪存使用方面有两大不同之处。首先，相对于固态硬盘，Oracle 使用的是 PCIe 闪存卡。其优点是闪存和处理器之间没有磁盘控制器，如果磁盘控制器未设计为以闪存的速度运行，则可潜在地减慢环境速度。Oracle 和 IBM 之间的另一个不同之处在于两家公司使用闪存的方式。Oracle 将其技术称为 Exadata Smart Flash Cache，在实际应用中用作读缓存。即将常用数据从存储器复制到缓存中，而对于 IBM 的 Easy Tier，其常用数据存储在 SSD 上。尽管用户可在数据库表、索引或段级定义指令，但系统会自动确定在缓存中存储哪些数据，以确保特定应用程序数据存储在缓存中，前提条件是软件具有足够的智慧，能够确定数据何时不适合存储在缓存中。基于缓存的方法的缺点是所有 Oracle 锁定位于数据库中，这意味着更新后的页面在缓存中不再有效。从独立的角度来看，这可能意味着将必须等待直到从磁盘刷新缓存，或需要直接从磁盘读取数据，但是假如我们讨论的是最近更新，则更新内容可能位于内存的数据库缓冲区，因此无需从闪存或磁盘访问数据。尽管如此，定义适当的指令是重要的考虑因素，正是变化不大的数据，才使得您将数据存放在缓存中获益最多。当然，IBM 的方法也有不足之处，即数据会从 SSD 阵列移至硬盘，或者反之，但这可作为不妨碍正常操作的后台任务执行。

如上所述，必须牢记 pureScale Application System 的标准配置中不包括固态硬盘。因此在大多数情况下，将在闪存与便捷的缓冲区之间进行比较。对于适当的应用程序，尤其当环境非常大时，闪存盘的使用可显著提高 Oracle 的优势。对于较小环境的适用度则完全是另外一个问题。

数据库

在本文之前，Bloor Research 在 2003、2005 和 2007 年定期对 DB2 和 Oracle 数据库系统进行过性能比较。我们不打算详述这些报告，否则本文的篇幅可能会翻倍！总的说来，每种产品都有各自的特点，两者不分伯仲。当这两家供应商都试图通过发布相应数据库系统的新版本超越彼此时，仍是如此。对于属于 OLTP 的元素，与数据仓库相比，以往我们或多或少将这两种引擎进行比较：我们欣赏 Oracle 的索引和集群功能，同时又喜欢 IBM 对 XML 的支持，及其调整和管理功能。就 pureScale 而言，除了集群支持之外，我们基本同意以上观点。然而，Oracle 在减少管理要求方面取得了显著进步，已缩短了差距，不过，我们认为其仍落后于 IBM。这两种产品间的显著差异在于对 XML 的支持，而且这种情况似乎不会改变。两家公司都声称原生支持 XML，但各自的声明也有不同之处：Oracle 支持原生 XML 数据类型，而 IBM 除此之外还原生支持存储 XML。这意味着并不需要因为要存储数据或在检索时重新合并而拆分 XML 文档，从而，在读取及写入 XML 数据时，DB2 应比 Oracle 更为出众。当然，这并非与所有用户有关，但如果 XML 对您的企业而言至关重要，则另当别论。

性能

SAP

虽然这并不适用于 Exadata 或 pureScale 的所有潜在用户，仍有必要指出 DB2 内置大量性能优化以支持 SAP 应用环境。尤其是，DB2 了解其工作的 SAP 环境，例如可在系统初次安装时有针对性地重新组织正在使用中的 SAP 配置相关详细信息。作为 SAP 安装过程的一部分，您还可以安装 DB2。DB2 还了解 SAP 工作负载，且数据库的内置调整功能可在提供建议时充分利用这一点；这同样适用于故障排除，因为诊断程序也了解 SAP 环境。

OLTP 环境中的非 OLTP

我们从未在仅处理事务的数据库环境下运行过。始终存在需要定期生成的各种报告（例如以往借项分析）。此外，其他应用程序也很可能利用相同环境。例如，使用 ERP 的制造商不仅会进行销售订单处理之类的事，还会完成生产能力规划等任务。在这种情况中，Oracle Exadata 应在某些条件下比 pureScale 环境更有优势，特别是在查询中需要对（若干）所有表进行扫描的情况。这些情况在数据仓库中更为常见，但有时也在操作环境中发生，在这种情况下，与 IBM 相比，Oracle 具有明显的性能优势。有关 Oracle 的技术如何在这方面发挥作用的详细讨论囊括在本文的姊妹篇中。

管理发展

我们已讨论过添加节点到集群的物理方面，以及两家供应商用于发展的常规方法，Oracle 提供基于机架的发展而 IBM 采用更具模块化的方法，允许您根据需要添加内核及/或附加节点。然而，我们还应考虑添加节点在软件方面的影响。

在 pureScale 环境中，添加节点意味着已安装操作系统，已将节点与网络物理连接且已启用对共享盘的访问。完成后，可输入命令“**db2iupdt -add -m <MemHostName:MemIBHostName> InstName**”添加新成员，然后 DB2 就可以为您处理所有事务（复制图像与响应文件、运行安装、设置对集群文件系统的访问等）。删除成员，或者添加或删除 PowerHA pureScale 服务器也使用类似的流程。然而在本版本中（稍后将进行更改），扩展或缩减实例是离线流程。

此外，与本流程有关的是，应该知道 pureScale 支持应用程序透明性。即由于 DB2 可自动处理提供支持的硬件环境的相关信息，因此在其上运行的应用程序无需再了解任何内容。这意味着添加或删除节点时无需更改编码，同样，也无需进行应用程序测试或基础架构调整。

另一方面，添加新节点到 Oracle RAC 实施意味着制备新节点、安装 CRS、安装 RAC 软件、添加 LISTENER 至新节点、添加数据库软件、手动添加 ASM 实例及手动添加数据库测试实例，这一过程似乎更为繁琐。此外，使用 Oracle RAC，应用程序需要具有集群感知功能，以优化与集群相关的性能优点，这意味着应随着环境的扩展（或缩小）对应用程序进行相应调整。

管理

我们已谈到虽然 Oracle 在近期版本大大减少了管理需求，但其自动和自我调节功能无法与 IBM 匹敌。除此之外，也许最大的差异存在于实施和高可用性方面。

实施

虽然我们没有关于安装 Oracle Exadata 系统的数据，却拥有在 4 节点集群上实施 Oracle Database 11g Release 2 的数据：根据 Winter Corporation 进行的独立研究，安装这类系统需要实施 208 个步骤。通过比较，IBM pureScale Application System 在交付之前预先经过搭建和测试且软件已安装在 IBM Premises 上。在物理装运过程中拆开运输，到货后工程师将为您组装。然后您就可以开始加载数据。换言之，无需安装即可使用 pureScale，而安装 RAC 需要 208 个步骤，可能高于安装 Exadata 所需的步骤数量。升级及修复时，还可能产生类似（虽然并非如此极端）差异，而仅需单一安装流程即可安装 IBM 的所有软件组件。

高可用性

Oracle 使用直接附加存储，而 IBM 使用基于 SAN 的存储方法。因此，Oracle 依赖于基于软件的磁盘镜像：磁盘出现故障时，由软件进行自动检测，并在其他无故障磁盘重新创建（及重新平衡）新的镜像。此外，Oracle 还提供高冗余选项，可通过此功能设置三份数据副本，以避免双重故障造成的问题。通常，两家供应商都声称没有故障源。事实并非如此。IBM 仅有一个互连（您也可以配置两个互连，但二者并不能互为备份）；相反，Oracle 确实装配两个具有独立链接的互连，可进行绑定以互为备份。另一方面，在每个 Exadata Storage Server 中只有一个磁盘控制器，这意味着如果该磁盘控制器发生故障，必须故障转移至另一个 Storage Server：严格而言，这并非单一故障点，而是代价不菲的备份方案。而 IBM 可通过双适配器、双控制器和双电缆等确保冗余。

此外，由于之前讨论过的锁定问题，如果发生节点故障，还可能造成 Oracle 环境的集群冻结。

DB2 的一个主要特征是支持 Oracle 环境。

可将 Oracle 模式直接导入到 DB2 数据库，且 DB2 可对 Oracle 并发控制（DB2 以其他方式进行此操作以避免造成 Oracle 环境性能降低的锁定问题）、SQL、PL/SQL、软件包、内置软件包、OCI（Oracle 调用接口）、JDBC、在线模式更改及 SQL*Plus 脚本提供原生（而非仿真）支持。

这意味着大多数基于 Oracle 数据库的应用程序、存储过程及其他用于在 Oracle 数据库上运行而编写的结构无需更改就可运行，还可能会因为锁定得以改进而在 DB2 数据库上提高性能。

IBM 公司表示，已测试 750,000 余条 PL/SQL 线路，平均兼容性已达 98.43%。

真的令人印象十分深刻。

成本

Oracle 为 Exadata 采取非捆绑式的定价结构，而 IBM 为 pureScale Application System 采取的是捆绑式的定价方法。因此，如选用前一种定价模式，则必须为数据库本身、RAC、分区、高级压缩以及调整和诊断包单独授权。表 2 列出了不同 Exadata 配置的标价，不包括这些附加组件（尽管其中的部分组件包括在 IBM 的标准配置中），此外，表中还列出了第一年的维护和支持，对比 pureScale Application Server 配置的比较数据，该配置拥有相似的服务器性能特性和存储容量，尽管 IBM 数据不包括固态硬盘的使用，而 Oracle 数据确实包括闪存盘。请注意，所列磁盘容量指的是可用的未压缩容量。

Oracle 系统	Oracle 标价	Oracle 列表无标重	IBM 系统	IBM 标价
¼ 机架 (6Tb)	153 万美元	73 万美元	小型	74 万美元
½ 机架 (14Tb)	311 万美元	151 万美元	中型	122 万美元
全机架 (28Tb)	610 万美元	290 万美元	大型	221 万美元

请注意，我们也收录了 Oracle 价格，不包含任何 ULA 类软件的成本（即：不包括数据库许可，但包括存储服务器许可）。这是因为，当前与 Oracle 间签有通用许可协议的客户不必支付部分或全部附加费用（具体取决于协议），因此，对于这些组织，此 Exadata 定价公式的元素可能部分或全部不相关。值得注意的是，尽管所有数据库软件许可费已经排除在外，IBM 仍有意将 pureScale Application System 的价格定位在与同类的 Exadata 配置相当或更低。

当然必须说明的是，这些价格仅为标价，可能会受到大幅折扣的影响。

我们已经提到的另一项主要的成本要素即为 IBM 针对临时附加要求的灵活定价方法（称作“按需容量”），采用这种方法，用户只需为实际使用的软件支付许可费用。此外，对于高可用性环境中的空闲待机 DB2 系统，IBM 不收取任何费用，而在 VMware 虚拟化环境中，只收取在服务器上执行 DB2 工作的相关费用。而在 Oracle，情况却全然不同。此外，正如我们前文所述，对于 Oracle，用户只能通过添加附加机架从 ¼ 机架升级到 ½ 机架，然后再升级到全机架。不能在存储服务器之外单独添加附加的数据库服务器，另外，每个磁盘驱动的存储服务器许可成本为 10,000 美元（另加 22% 的维护费），如果您只需要额外的计算能力，这个选择显然成本不菲。

最后，如果我们关于 DB2 较之 Oracle，pureScale 环境较之 Exadata 更易管理的论断是正确的，那么我们可以预计后者所需要的附加管理要多于 IBM。而这本身就是一笔支出。

结论

我们开展此项活动是为了能够找出 IBM 胜出的一些领域，以及 Oracle 所专长的其他领域。令我们震惊的是，几乎在我们研究的各个领域，IBM pureScale Application System 都胜过 Oracle Exadata。唯一的例外就是性能。闪存在 OLTP 环境中性能出色，这一点是毋庸置疑的，尽管 IBM 也能够达到这个水平，但并非标准情况。另一方面，IBM 的锁定更为高效（在 OLTP 环境中锁定非常重要），在互连环境中通迅量也更少。在这些问题当中，哪一项在 OLTP 环境中最为重要还值得商榷，取决于用户的具体环境。如若不然，我们大可期待 Oracle Exadata 为查询和报告等辅助功能提供性能优势，以及为运行于同一平台上的非 OLTP 整合应用程序提供支持。

除此以外，在其他所有方面，从可扩展性到灵活性，从易用性到高可用性，再到成本（至少从标价看来），IBM 似乎都占据不小的优势。在某些整合功能（尤其是复杂查询）上，Oracle 可能要比 IBM 更为出众，但相较于 IBM 在事务处理上的优势，就又显得有些微不足道了，因为事务处理一般是这类环境中一项关键的购买标准。

更多信息

有关此主题的更多信息，请登录
http://www.BloorResearch.com/update/2063

Bloor Research 概述

Bloor Research 是欧洲领先的 IT 调查、分析和咨询组织之一。我们说明如何通过有效地治理、管理和使用信息，为企业的 IT 系统带来更高的灵活性。通过以独立、智能、清晰的通信内容和出版物，“真实讲述”ICT 行业方方面面的情况，我们已经建立起良好的组织声誉。我们认为，讲述实情的目标在于：

- 在业务价值以及与之交互的其他系统和流程的环境下对技术进行说明。
- 了解最新的创新技术如何与现有的 ICT 投资相互适应。
- 纵观整个市场，说明所有可用的解决方案以及如何对其进行更加有效的评估。
- 过滤“噪音”，使查找支持投资和实施的附加信息或新闻变得更加简单。
- 确保通过最适当的渠道提供我们的所有内容。

我公司成立于 1989 年，已经花费二十多年的时间，通过在线订阅、定制调查服务、活动和咨询项目，为全球的 IT 用户和供应商组织发布调查和分析结果。我们的宗旨，就是将我们的知识转化为您的业务价值。

关于作者

Philip Howard
调研总监 - 数据

Philip 在计算机行业的奋斗史要追溯到 1973 年，他曾先后从事过系统分析师、编程员和销售等各类工作，在市场营销和产品管理领域也曾出现过他的身影，其供职的公司数量众多，其中包括 GEC Marconi、GPT、Philips Data Systems、Raytheon 和 NCR。

在度过了 25 年的打工生涯后，Philip 于 1992 年创立了今天的 P3ST (Wordsmiths) Ltd 公司，他的第一个客户就是 Bloor Research（然后是 ButlerBloor），在此期间，Philip 担任该公司的助理分析师。他与 Bloor Research 的合作关系从那时起延续至今，如今，他的职务是调研总监。他的执业范围涵盖任何与数据和内容有关的领域，另有五名分析师与他共同从事同一领域的工作。在保持对整个领域监管的同时，Philip 本身还专门从事数据库、数据管理、数据集成、数据质量、数据联合、主数据管理、数据管理和数据存储方面的工作。他对事件流/复杂事件处理也颇有兴趣。

除代表 Bloor Research 撰写大量报告外，Philip 还为 www.IT-Director.com 以及 www.IT-Analysis.com 这两家网站定期撰稿，此前，也曾代表剑桥市场情报 (CMI) 担任过《应用程序开发新闻》和《操作系统新闻》的编辑。此外，他还为众多杂志供稿，发表了大量报告，这些报告分别由 CMI 和《金融时报》等公司出版。

工作之余，Philip 主要的休闲活动包括划平底船、滑雪、打桥牌（他是桥牌高手）和溜狗。

版权及免责声明

2nd Floor,
145–157 St John Street
LONDON,
EC1V 4PY, United Kingdom

电话：+44 (0)207 043 9750
传真：+44 (0)207 043 9748
网址：www.BloorResearch.com
电子邮件：info@BloorResearch.com