第2章中已经讨论了Oracle数据库的体系结构，并且在第7章中介绍了Oracle如何使用硬件资源。硬件体系结构的选择和部署将最终决定你可用的特定扩展性、性能优化、管理以及可靠性等选项。事实上，系统有时配置得很糟糕，没有针对计划的工作量恰当地考虑CPU、内存和I/O之间的平衡。如果后来大仙性能出现问题，这可能限制数据库调整的选择。

本章将介绍各种硬件体系结构，从而为读者理解Oracel如何利用它们打下一个基础。具体包括下面列出的硬件系统类型以及Oracle如何利用每个平台的固有的特性：

* 单处理器（包括多核）。
* 对称多处理（Symmetric Multiprocessing，SMP）系统。
* 集群。
* 不一致内存访问（Non-Uniform Memory Access，NUMA）系统。
* 网络计算。

还将讨论各种磁盘技术的运用，以及如何根据自己的目的选择硬件系统。

## 1、系统基础

所有关于硬件系统的讨论往往都要首先回顾构成硬件平台的组件，以及这些组件对整个系统所产生的影响。你将发现所有计算机系统内部的主要成分都是相同的：

* 一个或多个CPU，用来执行构成计算机程序的基本命令，可能采用多核提供额外的处理能力。
* 内存，用来存储最近访问的指令和数据。
* 输入/输出（Input/Output）系统，通常是由键盘、用于把数据和程序从物理介质导出到设备控制器，以及将系统连接到网络上其他系统的网络控制器组成。

每个组件的数量和单个组件自身的能力决定了系统最终的成本和扩展性。带有四个处理器的主机通常比单处理器的主机贵，但具有更多的功能：新型部件（例如，CPU芯片）比起老一代而言速度更快但并不更贵。

在线事务处理系统（OLTP）的吞吐量通常最需要精心设计。在商业智能或数据仓库系统中，通常认为CPU和内存限制了性能。然而，CPU的处理能力和内存的容量在最近几年已经得到了显著的提升，而且目前I/O的性能也得到了很大改进，这些系统开始得到特别的关注。

每个系统组件都需要时间访问和传送数据，也就是存在一个延迟开销。组件的延迟开销就是将组件引入到系统后所带来的延迟时间，换句话说，就是下一级组件比起上一级组件要慢多少。每个组件的容量都是有限的。

CPU和CPU上的一级（L1）缓存具有最小的延迟时间，如表，但是它们的容量也最小。磁盘的容量最大，但是它的延迟时间最长。

|  |  |  |
| --- | --- | --- |
| 组件 | 典型存储容量 | 典型延迟时间 |
| CPU | 无 | 无 |
| L1高速缓存（集成在CPU中） | 10~100KB | 10ns |
| L2高速缓存（与CPU集成在同一主板上） | 1MB | 40~60ns |
| L3高速缓存（与CPU在同一主板上） | 10MB | 120ns |
| 主内存 | 1MB到几TB | 1000~10000ns |
| 磁盘 | 1GB到几百TB | 1~10μs |

注意：

有几种不同种类的内存：

L1高速缓存，集成在CPU芯片中；

L2（二级）高速缓存，与CPU集成在同一块电路板上；

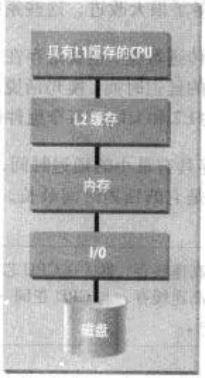
L3高速缓存，与CPU在同一块主板上，主内存，它是系统保留的内存，通过内存总线来访问。

主内存，是系统保留的内存，通过内存总线来访问。

对任何Oracle数据库进行调优的一个重要部分就是要尽可能地减少从延迟事件最大的来源（如磁盘）读取数据。如果必须要访问磁盘，就要保证尽可能地减少I/O子系统的瓶颈。由于Oracle数据库主要从内存而不是从磁盘中访问数据，因而系统的整个延迟事件相应地减少了，并且性能增加了。

## 2、单处理器系统

单处理器系统，如图，从体系结构的角度来说是一种最简单的系统。其中的每个系统（通常是一台标准的个人计算机）包含了单一CPU和单一I/O通道，并且完全是采用标准的工业化组件组成。它们通常用于单用户主机（例如，用于数据库开发或提供浏览器访问网络）。一些单处理器主机也可以被用做数据库的小服务器，特别是装有多核处理器的主机。



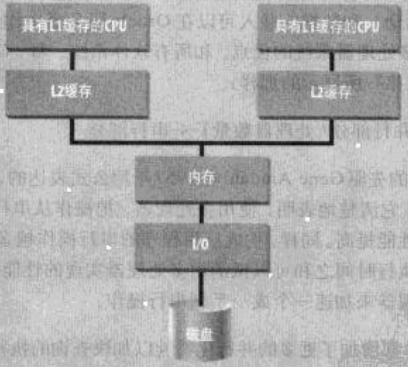
直到20世纪90年代，单处理器系统一直频繁地用做服务器，这是由于它们的价格低廉，并且关系型数据库不能在其他类型的系统中得到充分利用。然而，Oracle通过改进的并行化和更高级的优化使得它能够充分利用多CPU的系统。同时，对称多处理（SMP）系统的价格急剧下降，这使得SMP系统成为数据库服务器硬件的一种选择。

即使是在单处理器系统中，这些系统使用的服务器操作系统都支持多线程。而多核处理器正在普及，它进一步支持多个任务的同时处理。多核处理器是包含两个或多个处理器的集成电路。硬件平台供应商争相提供更多内核的处理器，一球自己的平台领先于其他公司。

服务器操作系统中的每个线程都能够用于支持一个并发的进程，它们可以并行地执行。在默认情况下，对Oracle运行的大多数平台，初始化文件中的PARALLEL\_THREADS\_PER\_CPU参数被设置成2。基于初始化文件中设置的参数，或者利用自适应并行度特性，Oracle能够进一步确定并行度。自适应多用户特性采用的算法将会考虑线程的数量。附加的调优参数同样会影响并行性，尽管最近发布的Oracle版本已不再需要使用这些调优参数。

## 3、对称多处理系统（SMP）

早期单处理器系统的局限性因素之一就是处理器的最大速度——所有的应用程序都必须共享这个资源。为了尽力克服这个局限性，发明了对称多处理系统，它通过在内存总线上添加CPU，来解决这个问题。



每个CPU都有自己的高速缓存。驻留在一个CPU高速缓存中的数据有时需要被另一个CPU处理。由于存在对数据共享的潜在需求，这种机器的CPU必须能够探查内存总线，以确定数据的拷贝驻留在哪里以及数据是否被更新过了。SMP系统的操作系统可透明地管理这种探查。Oracle第一标准版、标准版和企业版可以用在这些平台上。

SMP系统从20世纪80年代开始已作为中等规模的平台可用，主要是用于基于UNIX的机器。今天存在一类具备64位CPU（替代上一代32位CPU）的入门级服务器。运行在这类服务器上最普遍的操作系统是Windows的各种变体和Linux。

在这种基础设计上，能被平台供应商（如HP、IBM和Sun）扩展到更大规模的SMP服务器具有各种变体。例如，SMP系统可能包括多核CPU，一个更大的2级缓存，更快的内存总线和（或）多个高速I/O通道，每秒改进都试图消除在的可能影响性能的瓶颈。在高端群服务器上的Oracle实现中，UNIX和Linux是最常见的操作系统。

SMP系统中可能的CPU数量受系统（内存）总线的扩展性限制。当更多的CPU添加到总线上时，连接到总线的CPU间的流量会使总线本身到达饱和。

其具备64位CPU的系统能比以前的32位CPU更高效地处理大量数据；它们在具有数百GB内存的单个系统上支持许多个CPU。

当然，数据库也必须具有并行化特性，以充分利用SMP架构。Oracle操作（例如，查询执行和其他的DML活动）和数据的载入可以在Oracle服务器中作为并行的进程运行，从而允许Oracle利用多处理器系统的优点。和所有软件系统一样，Oracle受益于并行操作，正如“Amdahl定律”所展示的那样：

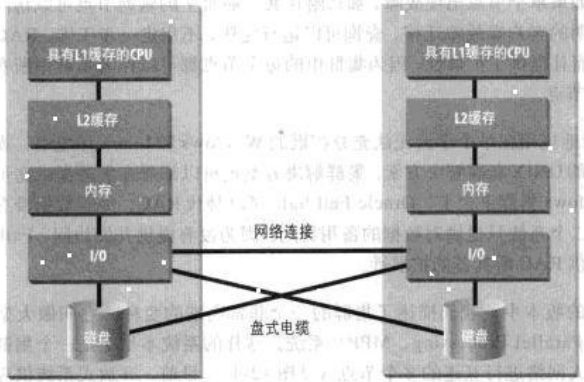
总执行时间 = （并行部分/处理器数量） + 串行部分

Amdahl定律是大型机的先驱Gene Amdahl在1967年用公示表达的，以描述混合了并行和串行的负载的性能，它清楚地表明，使用多处理器，把操作从串行执行部分移到并行部分可以得到预期的性能提高。同样，构成应用程序的串行操作越多，执行的事件越长，因为所有串行操作的执行时间之和可以抵消用多处理器实现的性能提高。换句话说，不能通过增加更多的处理器来加速一个或一系列串行操作。

Oracle的并行操作利用了可用的CPU资源。对一个CPU资源已经耗尽的系统来说，这种并行化无助于提高性能；实际上，增加管理并行进程所需的对CPU能力更多的需求，甚至可能损害性能。Oracle的自适应并行度特性能减少操作的并行度，以避免这种情况的发生。

## 4、集群（RAC）

自从20世纪80年代在DECV AXcluster匹配值中首次出现以来集群系统提供了一种高可用性和高扩展性的方案。集群能够把分离的机器所有组件联合起来，包括CPU、内存和I/O子系统，构成一个单一的硬件实体。但集群通常是通过连接到多个节点（计算机系统）上的共享磁盘建立的。系统间的高速互连提供了一种交换数据和指令的方案，而不需要写磁盘（如图）。每个系统都运行自己的操作系统拷贝和数据库实例。网格通常是由几个非常巨大的集群构成的。



Oracle对集群的支持可以追溯到VAXcluster。Oracle提供一种复杂的锁模型，从而使得多个节点能够访问磁盘上的共享数据。集群需要这样一种锁模型，因为集群中的每台机器都必须知道被物理上分离的其他机器所持有的集群中的数据锁。

目前，Oracle解决方案已经发展到了实时应用集群（它取代了Oracle 9i以前的Oracle并行服务器）阶段。RAC（real application clusters）通常用于Windows、Linux或基于UNIX的集群中。Oracle提供了一个集成锁管理器，用来协调不同的服务器或节点对同一块中数据的更新。

RAC提供了对高速缓存Fusion的完全支持，这里，锁被维护在内存中，不需要频繁地写磁盘。高速缓存Fusion不同于第8章描述的标准锁机制，因为它应用于数据块，而不是行。这种协调是必要的，因为两个不同的节点可能试图访问同一物理块中的不同行，而行是Oracle使用的最小数据单位。

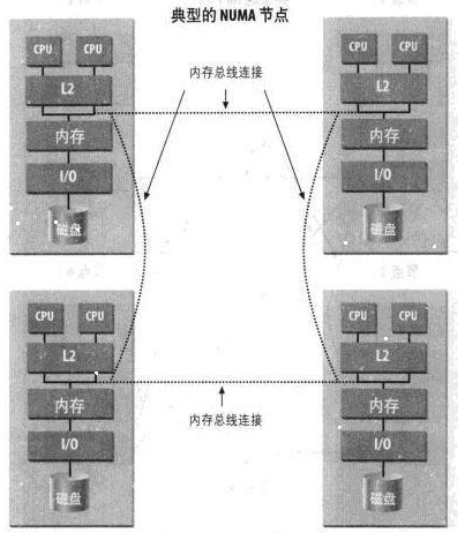
高速缓存Fusion最初大大地增强了读/写操作的性能（和OPS相比），后来对Oracle 9i中的写/读操作做了改进，提高了它的性能。目前，Oracle支持SDP和异步I/O协议，和以前传统的基于TCP/IP的RAC实现相比，可以提供更轻量级的传输。通过利用更快的互连技术，例如，支持RDS的Infiniband网络，Oracle更新的数据库版本进一步改善了性能。例如，Infiniband的节点到节点的延迟事件大约是千兆以太网延迟时间的十分之一（通常大约是70~80微妙）。

在RAC之前，可通过配置集群来为系统提供更高的吞吐量或更高的可用性。在高可用性的情况下，如果单个节点出现故障，那么附在共享磁盘上的辅助节点可以访问相同的数据。通过透明的客户端故障迁移，查询可以运行完毕，不用进一步干涉。RAC不仅提供了可用性，而且提供了扩展性，因为集群中的每个节点都可以作为集群中所有其他节点的故障恢复节点。

RAC正不断地应用在单个平台无法充分扩展的Windows和Linux环境中，或者用来替代花费高昂的UNIX高端解决方案。集群解决方案也可以部署在需要提供高可用性的场合。在Windows集群平台上，Oracle Fail Safe可以替代RAC，尽管数据没有被两个系统共享，第二个系统只提供对数据的备用访问。因为没有提供并发访问，Fail Safe解决方案并不提供对数据的备用访问。因为没有提供并发访问，Fail Safe解决方案并不提供RAC所具备的扩展性。

## 5、非一致内存访问系统（NUMA）

90年代中期引入的非一致内存访问（Non-Uniform Memory Access，NUMA）计算机提供了比SMP更大的吞吐量，它通过分布式内存将多个SMP组件连接起来而实现，如图，与集群类似，这些系统除了提供CPU的扩展性，还提供内存和I/O子系统的扩展性。一个关键的不同之处在于，它需要由单一的操作系统的副本来管理整个平台，而且还需要基于目录的高速缓存一致性机制来保持数据的同步。节点之间的内存访问只需要几百个微妙，比访问集群配置中的磁盘快得多，并且只比单一SMP系统中的本地内存总线的速度慢一点。内存容量可以扩充到几个G。



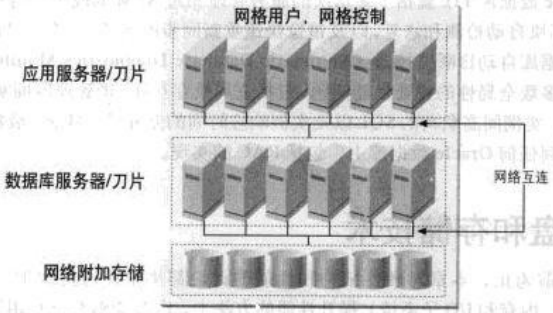
相对于集群方案，NUMA具有几个主要优点：

* 不需要开发和验证运行在这些机器上的应用程序的并行版本（虽然这些应用程序针对NUMA优化时，能够获得额外的性能）。
* NUMA系统上的管理比集群上要简单得多，因为只有一份操作 系统拷贝需要管理，而且通常只部署一个数据库实例。

目前，HP Superdome是NUMA系统的一个例子，它的可扩展性在产品数据库中得到了证明，它可扩展到数十TB的数据。由于这种平台和SMP系统的行为相似，管理也一样，因而对它们的权衡是类似的。

## 6、网格计算

Oracle数据库的命名规则中“g”意味着Oracle公司关注网格计算能力的支持。网格是一个简单的计算机池，基于“按需分配”给应用程序提供所需的资源。其目的是提供透明地扩展到用户社区的计算资源，如同电力公司通过电力网从其他电力供应商那里接入电力以满足高峰需求。计算网格使得能够动态提供CPU和数据资源（如图）。带有RAC的Oracle数据库构成了提供这些资源的基础。



Oracle 10g引入了几种重要的特性，使其能够在需要时通过网格提供资源：

动态服务提供：

Web服务：

滚动升级：

自动存储管理：

企业管理器网格控制：

Oracle 11g囊括了多层次的服务管理和对ASM的改进，这些改进包括对滚动升级、坏块自动检测和修复，以及镜像快速重新同步的支持。对多实例的数据库，用于RAC的数据库自动诊断监控器（Automatic Database Diagnostics Monitor，ADDM）可以确定大多数全局性的重要性能问题，包括全局缓存互连、所管理器拥塞、全局资源竞争（如I/O）、实例间高负荷的SQL以及实例响应时间的差异等。目前，最新的数据库补丁可以应用到任何Oracle数据库上，包括RAC的实现。

## 7、磁盘和存储技术

到目前为止，本章对硬件结构和性能的讨论都集中在通过增加可用的系统资源（如CPU、内存和I/O子系统）提供性能的方法上，以及能够充分利用这些资源的并行特性上。一种重要的提升性能的方法就是调整I/O，包括把数据分布在各个磁盘上和提供足够多的访问通路用于访问数据。根据经验来说，每4个CPU每秒最少有2GB的数据传送，I/O系统应具有每秒1GB的吞吐量。由于磁盘访问具有最大的时间延迟，因而调整I/O系统的另一个重点是在内存中保存从磁盘取到的数据。

Oracle将具有适当的I/O（特别是能够提供数据存储访问通路的磁轴），内存以及CPU的配置称作均衡配置。正如前面所提，针对磁盘，Oracle 10g和之后的版本现在都包括了ASM（自动存储管理），它大大简化了对数据存储的日常管理。但咨询硬件提供商以获得适当的数据存储配置有时候被证明是很困难的，尤其是对数据仓库来说。并且，虽然磁盘的容量增加了，但是访问时间并没有随之改进。因而Oracle为几个主要的硬件平台和存储供应商开发了参考配置，以帮助管理员通过Information Application Initiative提供更准确的初始大小。

## 8、磁盘部署策略

磁盘经常直接附加在系统上，更贵的系统提供更快的磁盘控制器和I/O。随着网络带宽的提高，网络附加存储（Network Attached Storage，NAS）和存储区域网络（Storage Area Networks，SAN）作为更合算的选择而出现。磁盘用各种不同的方法配置，以达到冗余性，并消除单个磁盘故障点导致的数据访问丢失的可能性。

磁盘通常部署成阵列，其工业标准是RAID（Redundant Array of Inexpensive/Independent Disks）。可以把RAID用作任何已经讨论过的用以提供高性能和高可靠性配置的一部分。从Oralce 10g以来，使用普通的磁盘集，自动存储管理（ASM）提供了RAID阵列的许多功能。

Oracle 9i在数据库中引入了表压缩，主要是在数据仓库中用于减少数据的存储需求。因为重复的值被存储在数据块头上的符号表中，而所有另外再出现的都用一个指向符号表的连接替代了，这样数据库中重复的值就被取消掉了。Oracle 11g也具有一个针对OLTP中重要的插入、更新和删除操作的高级压缩选项。现在实现50%的数据压缩是很常见的。除了减少磁盘存储之外，当压缩数据全部在缓存中（而不用请求访问磁盘）时，对性能也是有利的。

因为存储容量增长，而且使用磁盘的成本降低，现在很多机构把数据仓库和商业智能实现的所有相关的数据都联机存储在磁盘存储中。倘若说具有最好性能的磁盘通常更贵并且容量更低，目前许多人针对不常被访问的数据部署的是那种高容量低性能（但便宜）的磁盘。Oracle数据库在2007年首次引入信息生命周期管理，特别是ILM助手，提供了管理这样一个环境的能力。

## 9、选择哪种平台部署方案

如果花在硬件上的钱没有限制，那么可以简单地确定最合适的硬件；只要选择所需的吞吐量和可靠性的级别，然后就去购买！不幸的是，你会发现必须面对现实，硬件方案的选择通常都是折中的。由于一直以来相关价格的持续下跌，这使得做出选择更加容易。

### 9.1平台比较

Oracle服务器最常用的硬件平台是SMP系统，它在性价比上达到了很好的平衡。SMP（对称多处理）系统广受欢迎，存在下面一些原因：

* 与单处理器系统相比，SMP系统为将来提供了更多的和更简单的可扩展性选项。
* 带有大内存支持的64位处理器和操作系统允许SMP系统处理巨大的数据库。
* 不同于集群，SMP系统只需管理和维护单个操作系统和单个Oracle实例。
* 比起集群，SMP系统能被更多的应用使用。
* 在相似的CPU配置中，SMP系统比NUMA、集群或者网格配置便宜，因为内存和I/O子系统不必复制成一样。

这并不是说其他配置不应被考虑。当然，如果扩展性的需求超过了SMP主机的能力，集群或者网格可能就是仅有的解决方案。通过使用RAC配置中的“commodity”节点，集群可以变得便宜些。通过谨慎的规划和企业计算管理，这样的配置可以提供强大且高效的解决方案。

下表提供了不同部署平台在可扩展性、可管理性以及可用性方面各自优点的对比。

|  |  |  |  |
| --- | --- | --- | --- |
| 等级 | 可扩展性 | 可管理性 | 可用性 |
| 最好 | 网格 | 单处理器 | 网格 |
| 集群 | SMP | 集群 |
| SMP | 网格 | SMP |
| 最差 | 单处理器 | 集群 | 单处理器 |

应该基于性能和恢复的需求及预算选择存储技术。通常来说，更贵的方案提供更好的性能和更灵活的可用性选项。选择存储方案时别忘了考虑吞吐量。

### 9.2选择平台的方法

在选择一种部署方案时，大多数机构会选择对将来能满足预期的性能和可扩展性需求的系统，同时考虑可管理性和可用性的需求。但还有两种另外的方法可以考虑。

第一种就是我们都熟悉的真理：等得越久，计算机硬件（和相关的组件）就越便宜。根据摩尔定律，每隔18~24个月，芯片的计算能力就会翻倍，而且每次都会在性能上由巨大的飞跃。这样的性能提高是通过提高时钟速度和处理器中引入更多的内核来实现的。

在计算机硬件工业中，价格持续降低而性能特征的持续提升是一个不争的事实。但是，怎样利用这个事实为机构的系统架构规划部署策略呢？

需要的时候就去购买，并且当硬件不再满足单独应用的需要时，就制定计划把退役的硬件回收再利用到机构的其他地方去。例如，今天的部门服务器可能变成明天的Web服务器。对网格结构来说，可能会不停地把老的硬件用到现存的计算方案中。

其他，在非网格方案中，记住考虑硬件升级的作用尤其是CPU的升级。SMP系统和节点需要所有的CPU都一样，因此如果升级一个CPU就不得不把所有的升级。有时供应商会推荐一个新的系统，因为其他的内部特性（如内存和I/O总线技术）已经得到改善，可以部分地满足新CPU增加的能力。

网格很有诱惑力，因为当新的机器类型出现时，它们可以被添加到网格中去。Oracle 10g中出现的自我优化和高级管理的能力，在Oracle 11g中得到进一步的提高，这使得网格计算更加实用，省去了艰难的手工优化工作，而这些工作以前需要考虑系统中的变化。