文档编号：**GT-MBS-PMDMDTJDDCJBJMSGZJBDMJZFA-001**

Moebius™ for SQL Server

数据库集群平台技术说明

格瑞趋势 Green Trend Software

北京格瑞趋势科技有限公司

商业信用

文档变更

版本号	修订日期	描述
1.0	2007.4.5	初稿
1.1	2007.5.8	修订
1.2	2008.8.5	修订
1.3	2008.12.5	修订
1.4	2009.7.8	修订
1.5	2009.9.5	修订
1.6	2010.2.8	修订
1.7	2011.5.9	修订
1.8	2012.6.5	修订

目录

第1章 背景

40 年来，在 IT 的影响下，企业在各个方面都发生了翻天覆地的变化， IT 与业务的关系不仅仅是“支持”，而是业务赖以发展的基础 ，Gartner 公司的研究表明，80%的业务流程依赖于 IT。 在这样的背景下，我们对 23 大型家企事业单位的信息化负责人进行了面对面的访谈，从而能直观地获知 IT 在他们实际业务中的作用到底有多大？同时分析大家在 IT 建设中存在的问题，总结共性、避免风险。

新华百货、***CCTV*** 、中国移动、湖北宜化集团、巨人网络集团、路桥建设集团、中国烟草总公司、天津电力建设集团、京东网上商城、湖南省计生委、河南省建设厅、上海交通大学、浙江卫生厅、中国海关、中国中铁、三一重工、九牧王服饰集团、南京市建设交易中心、四川民政、易车网、宏源证券、国泰集团、中石油云南分公司

图 1. 访谈公司名单

经过和客户的交流，这 23 家企业都已经建设了多个信息系统，且这些信息系统都面临着相同的问题，概括如下：

1 随着系统数据量和用户数量的增加，数据库的负载居高不下（如 CPU、内存、IO 等指标高），用户对系统缓慢的响应速度怨声载道；

2 对数据进行集中汇总操作（统计或报表业务）耗费时间很长，管理者无法及时获取决策所需的实时数据；

3 时而发生的故障（如操作系统、数据库、网络、服务器、存储等硬件）致使系统中断，严重影响信息系统的运行，进而影响企业的正常运转；

4 缺乏实时的冗余数据，面临着丢失数据的风险，存在着极大的安全隐患；

5 更换更高配置的硬件来提升性能，扩展能力有限（PC Server4 路为最高配置），无法继承历史投资且回报率逐渐变低；

6 尽管部署了"双机热备"系统，但依然面临着性能瓶颈，对于此结构中资源闲置的节点无能为力，设备利用率低；

7 因性能原因计划数据库迁移（迁移到其他平台），需重构代码、重建系统平台且面临着系统稳定性及安全性的风险。

第2章 IT 的现状及分析

2.1 专业术语

高可用性：High Availability（HA）通常用来描述一个系统经过专门的设计，从而减少停工时间，而保持其服务的高度可用性。

计算机系统的可靠性用平均无故障时间（MTTF）来度量，即计算机系统平均能够正常运行多长时间，才发生一次故障。系统的可靠性越高，平均无故障时间越长。可维护性用平均维修时间（MTTR） 来度量，即系统发生故障后维修和重新恢复正常运行平均花费的时间。系统的可维护性越好，平均维修时间越短。计算机系统的可用性定义为：MTTF/(MTTF+MTTR) * 100%。由此可见，计算机系统的可用性定义为系统保持正常运行时间的百分比。

高伸缩性：指一个系统的持续扩展能力，当一个系统遇到性能瓶颈时，一般有两种扩展方式。

向上扩展：向单一节点添加硬件设备或将其升级为一个大型节点。升级到综合性能更强大的硬件， 带来的问题是硬件的浪费，一次性的投资增加。单节点体系结构最终会达到一个瓶颈并无法实现进一步的有效扩展。具体表现为逐渐缩小的回报率或者价格惊人的昂贵硬件设备。系统得不到可持续的扩展，不能从根本上解决问题。

向外扩展：添加更多节点并将数据及工作负载分布于这些节点当中。

负载均衡：负载均衡有两方面的含义，首先，大量的并发访问或数据流量分担到多台节点设备上分别处理，减少用户等待响应的时间；其次，单个重负载的运算分担到多台节点设备上做并行处理，每个节点设备处理结束后，将结果汇总，返回给用户，系统处理能力得到大幅度提高。

虚拟 **IP** 技术：虚拟 IP 地址(V IP) 是一个不与特定计算机或在一个计算机中的网络接口卡(NIC)相连的 IP 地址。引入的分组被发送到这个 VIP 地址，但是所有的分组旅行通过实际的

网络接口。VIPs 大部分用于连接冗余；一个 VIP 地址可能也在一台计算机或 NIC 发生故障时可用，因为一个可选计算机或 NIC 响应连接。

2.2 企业信息系统的发展阶段

企业的信息化建设按照时间的先后顺序或者用户关注的焦点来分为 3 个阶段:分别是基础建设阶段、应用建设阶段、数据交付阶段。

2.2.1 基础设施阶段

信息化的建设阶段，即从无到有的过程，本阶段，用户主要以机房、网络等基础建设为中心。

2.2.2 应用建设阶段

为了更好地支持企业的业务，大家都纷纷建设了各类应用系统，如：财务管理系统、OA 系统、ERP 系统、物流系统、供应链系统、客户关系管理系统等，到今天为止，大多数企业已经或多或少地建设了一些信息系统。本阶段，用户主要以应用软件（即应用软件的功能或满足业务的逻辑）为中心。

2.2.3 数据交付阶段

随着各应用系统的运行，系统的用户数量和数据量都不断增加，这些数据将成为企业核心资源，不论是事物处理、统计处理还是数据挖掘都将依赖于对数据处理的能力，今天，很多企业的信息系统都处于这一阶段，主要面临的焦点问题如下：

1. 大量系统处于独立分散的单机状态,无法有效地整合

提供数据存取服务的设备依然处于独立、分散的状态，导致一部分设备资源闲置，一部分却能力不足，这样不但增加了系统维护的复杂性，又造成了资源的浪费。

2. 系统面临性能瓶颈,严重影响业务的进行

系统的速度越来越慢，尤其在并发大的时候，几乎无法访问，严重影响业务的进行，用户对缓慢的响应速度怨声载道；尤其在业务高峰，如“五一”、“十一”等高峰时节，会面临很多用户排队等待，严重影响业务的进行。（仔细想象，由于系统慢，您的客户排起长队无法买单，会对您造成多大的经济损失）

数据库的负载居高不下（如 CPU、内存、IO 等指标高），对数据进行集中汇总操作（统计或报表业务）耗费时间很长，管理者无法及时获取决策所需的实时数据；

3. 对数据的重要程度没有足够的认识,安全性严重不足

企业的数据库保存着企业的重要信息，一些核心数据甚至关系着企业的命脉，单一设备根本无法保证数据的安全性，一旦发生丢失，很难找回，造成难以估量的后果。很多管理者对企业数据的重要程度未上升到一定高度，或者根本都没有安全意识，甚至存在侥幸心理，认为出错的可能很小。所以对数据的管理往往由系统管理员进行人工备份。（仔细想象，您核心的购物卡数据丢失一部分，会对您造成多大的经济损失）

4. 对信息系统的健康状况没有足够的认识,系统出现中断的风险

数据库作为信息系统的核心，起着非常重要的作用，单一设备根本无法保证系统的持续运行，时而发生的任何故障（如操作系统、数据库、网络、服务器、存储等硬件）都将致使系统中断，严重影响信息系统的运行，进而影响企业的正常运转。（仔细想象下，您企业的核心 ERP 系统中断一天会对您造成多大的经济损失）

很多企业的现状是对信息化的建设思维停留在以应用为中心，过多地关注业务而忽略了信息系统是一个整体，对信息系统建设的第三阶段没有认识，且未对未来的发展进行合理的规划，聪明的企业主明白，不管经济状况多么不好，提前做好技术方面的准备都非常重要。

2.3 IT的重点从"计算"转向"数据"

当今的企业正面临着不断变化的业务需求和挑战，高并发访问、海量数据处理和严格的实时业务需求对企业内部 IT 系统在性能、可靠性、扩展性和效率上提出了更高的要求。这就要求系统的结构高度灵活，在面对新的应用环境下能快速作出反应。在负载均衡技术的广泛应用下，应用层已经能够胜任这种变化的需要，但是在数据层依然面临着大量问题。

2.4 数据库面临的挑战

数据库作为信息系统的根基，支撑着整个应用系统，发挥着非常重要的作用，因此，它的健壮与否直接决定着整个信息系统能否高效、稳定运行。在许多人看来，当前的数据库技术已经说是非常地成熟了，然而，在满足不断增长、不断变化的应用需求方面，当前的数据库技术其实还存在不少急迫的技术问题。

对于所有的数据库而言，除了记录正确的处理结果之外，它们都面临着四方面的挑战：如何提高处理速度，数据可用性、数据安全性和数据集可扩性。随着 Google、Amazon、Alibaba 等公司在大规模数据库集群上实践的成功，这将对传统的数据库部署模式及数据处理方式带来巨大的变化，数据库集群作为"云计算"的底层支撑平台，它的许多重要特性：可靠性、高性能、易伸缩性和安全性将给企业带来更多新的机遇。理想的数据库集群应该可以做到以下：

1 将多个独立的主机虚拟成一个对应用完全透明的虚拟主机（多虚一），需要更高数据库处理速度，我们只要简单地增加数据库服务器就可以了。这样，不但可以继承历史投资，节约硬件投入成本，而且大幅提升处理速度。

2 在任何时刻需要有多个随时可用的实时同步数据服务，为了快速处理报表业务，最好有多个异步同步的数据。这不仅会增加数据可用性，还会成倍提升查询的速度。

3 数据集的可扩性可能是最简单的要求了，最好能任意增大数据库而没有可用性的负面影响。

有关数据库集群的技术都是非常复杂的，更具挑战性的是，实际的应用要求上述几方面的指标能同时提升，而不是某一指标提升了，另外的指标却下降了。然而，所有的技术都有副作用的，这就是当前数据库集群技术面临的重大问题。

2.5 用户需要"一站式"数据平台

目前 SQL Server 数据库服务器的部署模式主要是"单机模式"；小部分用户选用了传统的"双机模式"，而这仅仅是一种备份的方案，数据库只运行在一个节点上，当出现故障时，另一个节点只是作为这个节点的备份，在性能上是没有提升的。

用户需要的是"一站式"数据服务，一个可以为之稳定提供服务的数据库平台，一个涵盖高可用、数据安全、负载均衡的整体数据库解决方案，而不是一堆零散的"双机备用"、"灾备""数据复 制""负载均衡"软件，或者是它们之间的简单集成。

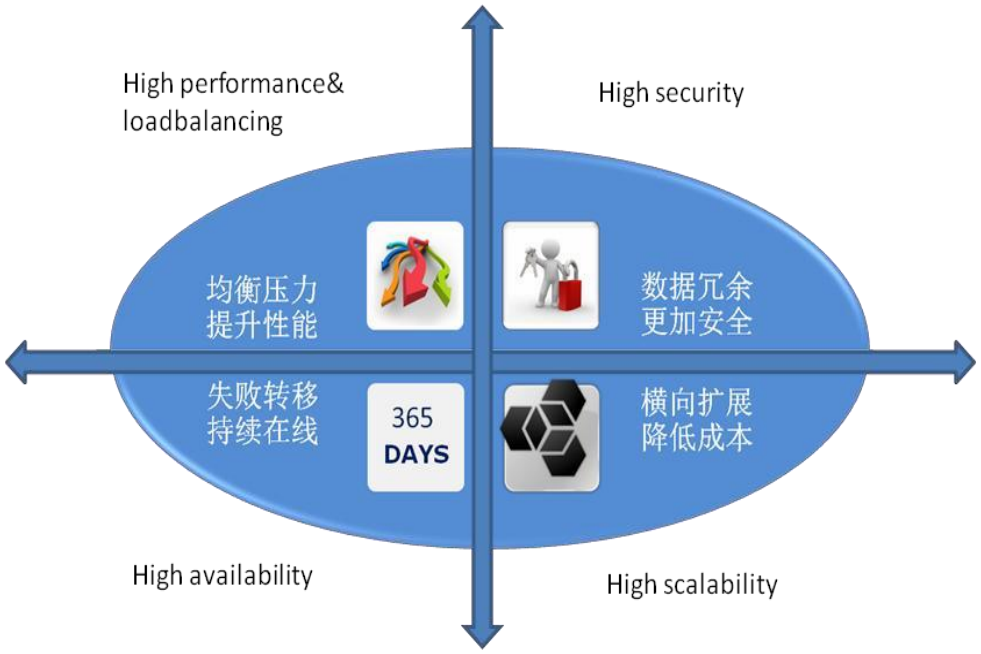

图 2. "一站式"数据服务模式

第3章 Moebius 集群平台软件介绍

3.1 概述

数据库集群技术可以有效地整合和利用现有 IT 资源，提供高效、可靠的数据服务。Moebius® for SQL Server 是格瑞趋势专门针对 Microsoft SQL Server 数据库提供的综合集群平台，利用这一平台，任何企业都能够轻松地构建出适合自身业务的数据库集群，满足用户对负载均衡、可用性、数据安全、扩展性的需要。

Moebius® for SQL Server 基于 SQL Server 的内核实现，核心程序宿主在 SQL Server 的内核之中， Moebius 集群强大的 SQL 解析引擎结合多种负载均衡策略，可以实现 SQL 语句一级的负载均衡；同时将自动故障监测、虚拟 IP 及失败转移技术融入其中，满足企业对高可用系统建设的要求；数据复制时，采用了同步和异步两种复制模式，可实现数据在多台服务器间实时同步，保证事务的一致性和完整性，支持远距离复制；Moebius 集群具有带宽占用少、同步效率高、数据实时性高、数据一致性保障好的特点。

3.2 软件的组成

Moebius 集群平台软件由 3 部分组成：Moebius 集群配置管理器、Moebius for SQL Server Agent、Moebius Core。

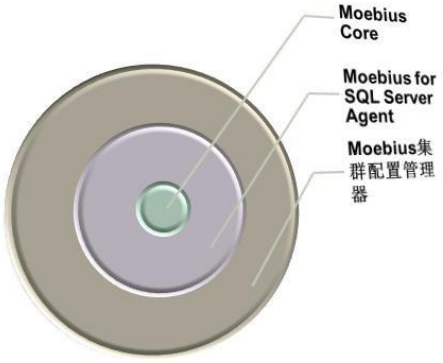

图 3. Moebius 集群平台软件的组成

3.2.1 Moebius集群配置管理器

在整个构建 Moebius 集群的过程中，用户只安装 Moebius 集群配置管理器并通过此配置管理来创建集群的节点、配置功能、管理集群；Moebius for SQL Server Agent 和 Moebius Core 是在创建集群的过程中自动部署到每台 DB Server 上。

Moebius 集群的管理工具集成到 SQL Server 的 Management Studio 管理工具中，这样高度集成的设计更方便用户使用，图形化的管理工具可以轻松地实现集群的创建、节点的扩展、负载设置、虚拟 IP、故障监控、日志记录、性能预警、邮件通知以及更加及时的短信通知等操作，最大限度地降低用户的管理成本。

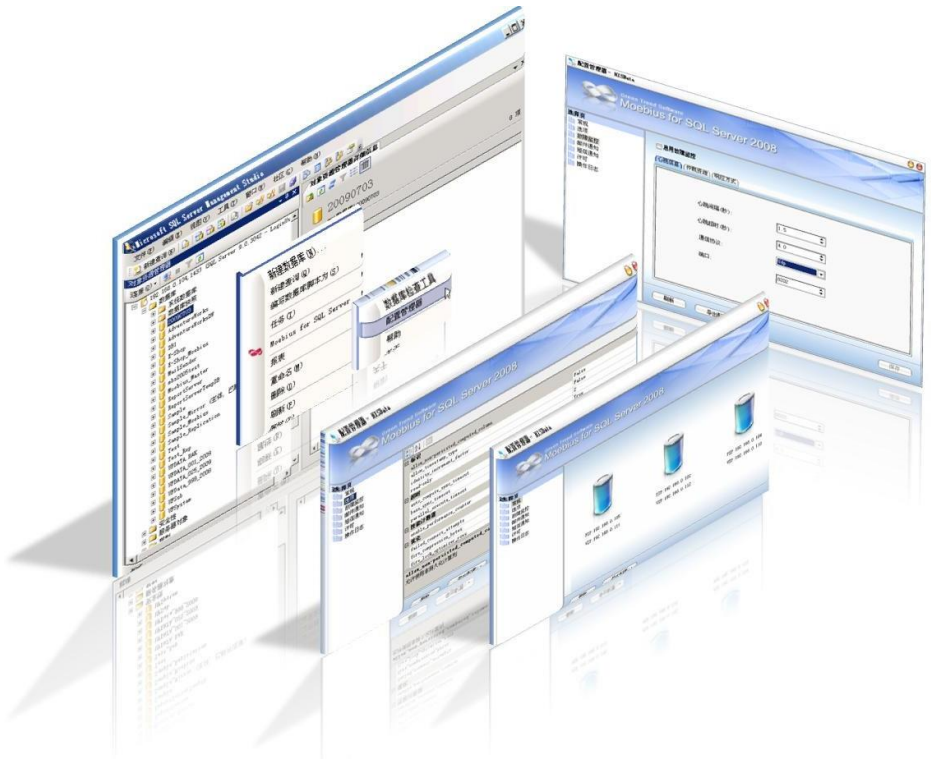

图 4. Moebius 集群配置管理器

Moebius 集群配置管理器可以安装到任何机器上，要求：

1. 此机器安装 SQL Server 的 Mangment Studio；
2. 此机器可以通过 SQL Server 的 Mangment Studio 管理 DB Server；

3.2.2 Moebius for SQL Server Agent

运行在 DB Server 端，在创建集群的过程中自动部署到每台 DB Server 上，Agent 对集群进行了封装，涵盖了 SQL 解析及调度引擎。

3.2.3 Moebius Core

运行在 DB Server 端，在创建集群的过程中自动部署到每台 DB Server 上，Moebius Core 是整个软件的核心，涵盖了同步引擎及故障监控引擎。

3.3 功能介绍

3.3.1 核心功能

数据库的负载均衡

传统的数据库集群都是保证业务持续可用的，有一个主节点，一个备用节点，如 MSCS 或者第三方的 HA 集群，这类集群的共同特点是始终只有一个节点在运行，在性能上得不到提升，系统也就不具备扩展的能力，当现有的机器不能满足应用的负载时只能更换更高配置的机器。这样的系统既不利于扩展，同时硬件资源浪费严重。

在 Moebius for SQL Server 数据库负载均衡集群中，打破了以往主节点和备用节点的概念，集群中的每个节点都具有同等地位，Moebius 可以在多个节点之间实现动态均衡连接请求，实现各节点压力的均衡，进而显著提升数据库系统的性能。

高可用性

在 Moebius for SQL Server 数据库负载均衡集群中，继承了 HA 集群的优点，Moebius 集群为用户提供了多种选择模式，您可以根据对可用性要求程度的不同，采取合适的设置，采用多种故障

监控机制实时监测系统的软硬件健康状况，在 Moebius for SQL Server 负载均衡集群中若某节点发生故障，故障节点的虚拟 IP 会立即飘移到其余健康的节点来响应连接请求，保证业务不中断，同时可以在不影响业务的情况下完成故障节点的修复、重新上线。

数据集的可扩性

传统方案当一台服务器处理能力都用尽时，我们一般会替换成一台新的更强大的服务器，这样的扩展方式我们称之为向上扩展；随着服务器处理能力的增强，它们的价格也会更昂贵。使用 Moebius for SQL Server 负载均衡集群，在需要更高数据库处理速度时，我们只要简单地增加数据库服务器就可以了。这样的扩展方式我们称之为向外扩展，可以大大降低硬件投资的风险，而且大大提高现有服务的质量。

数据集的安全性

Moebius for SQL Server 负载均衡集群采用无共享磁盘架构，这样各个机器可以不连接一个共享的设备，数据可以存储在每个机器自己的存储介质中。集群中各节点在任何时刻具有实时一致的数据，实现了真正的数据冗余，这样冗余的硬件架构不但可以避免单点故障而且提供了杰出的故障恢复能力；不会因为系统故障导致数据的丢失，大大提高了整个系统的可靠性与安全性。

3.3.2 功能明细

表 2. 基本参数

功能名称	**Moebius for SQL Server** 数据库负载均衡集群软件 **V6.0**
最大节点数量	16 节点
每节点最大 CPU 数量	Os 支持最大
吞吐量（或最大连接发数）	Os 支持最大
高速缓存	500M
透明支持读、写分离架构	支持
支持数据库版本	SQL Server2005/2008/2008R2 标准版、企业版（64 位和 32 位）

扩展性能力	高
可靠性	高
易移植特性	完全透明

表 3. 数据同步

功能名称	**Moebius for SQL Server** 数据库负载均衡集群软件 **V6.0**
数据实时同步	支持
并发执行 SQL 语句	支持
同步 SQL 语句	支持
升级数据库锁	支持
智能同步策略	支持
发布订阅等异步同步技术	支持
批量同步数据	支持
检查数据一致性	支持
传输压缩	支持

表 4. SQL 语句调度

功能名称	**Moebius for SQL Server** 数据库负载均衡集群软件 **V6.0**
替换 SQL 语句	支持
制定 SQL 语句的分发规则	支持
自动优化	支持
连接级负载分发	支持
SQL 语句级负载分发	支持
随机负载均衡	支持
权重负载均衡	支持
轮询负载均衡	支持
自适应动态负载均衡	支持
负载均衡外置	支持

表 5. 高可用

功能名称	**Moebius for SQL Server** 数据库负载均衡集群软件 **V6.0**
虚拟 IP 技术	支持
手动故障转移	支持
CPU 监测	支持
内存监测	支持

网络监测	支持
数据库实例监测	支持
操作系统监测	支持
自动故障转移	支持
数据级的高可用	支持
邮件通知服务	支持
短信通知服务	支持

表 6. 集群管理

集群管理	
功能名称	**Moebius for SQL Server** 数据库负载均衡集群软件 **V6.0**
远程管理	支持
中文界面	支持
图形化的管理工具	支持
数据库管理工具	支持
安全认证	支持
计划停机	支持
自动同步差异数据	支持

3.3.3 许可方式

表 7. 许可方式

许可方式	说明
处理器许可证方式	按照组成集群的 SQL Server 数据库服务器的处理器（物理 CPU ）数量授权。（如搭建 2 台数据库服务器的集群，每台服务器 2 个 CPU，则需购买 4 个授权）

3.3.4 运行环境

要求服务器可以正常运行 Windows 及 SQL Server，对构建集群的数据库服务器品牌、型号、配置无特殊要求。

表 8. 运行环境

序号	操作系统	数据库
1	Windows2000（32 位/64 位）标准版/企业版	SQL Server2005（32 位/64 位）标准版/企业版
2	Windows2003（32 位/64 位）标准版/企业版	SQL Server2008（32 位/64 位）标准版/企业版
3	Windows2008（32 位/64 位）标准版/企业版	SQL Server2008R2（32 位/64 位）标准版/企业版
4	Windows2008R2（32 位/64 位）标准版/企业版	

3.4 核心价值

可持续扩展的方案，实现负载均衡 - Moebius 集群提供数据包解析及多种负载分发机制，最终实现 SQL 语句级负载均衡；集群中所有节点处于实时活动状态，可以有效分担系统的压力，进而显著提升数据库系统的访问能力；

保护您的数据安全、可靠 - Moebius 集群中，任何时刻系统拥有多份实时一致的数据，彻底避免系统故障造成关键数据丢失，确保数据安全；

保证应用不间断，支持异地 - Moebius 集群采用非共享磁盘冗余结构设计，快速的故障监测及自动失败转移机制确保系统可靠性，即使某节点发生故障，也不会导致系统中断，保证数据库持续提供服务；

同步效率高 -Moebius 集群采用多种同步策略，更智能；并行复制速度更快；采用数据压缩，带宽消耗更小；

简单易用 -管理工具集成到 SQL Server 中，操作更方便；图形化的界面，使用更轻松；对应用程序透明，无需改动原有程序。

可信赖的解决方案-基于数据库实现的集群技术，专门针对 SQL Server 提供，更专注；提供7*24小时客户支持。

降低系统 **TCO**（总体拥有成本）

1　对硬件的一致性无要求，可以通过增加服务器的数量来提升性能，极大的降低系统投入成本：

2. 集群支持无共享磁盘架构，可以节省存储设备的开销；可以充分利用企业原有设备组建集群，避免资源浪费；
3. 可以用多个廉价 PC 服务器代替昂贵的小型机或大型机，节约硬件成本；
4. 集群支持 SQL Server 各个版本，可以和 SQL Server 标准版搭配节约软件的投资；
5. 将数据库系统统一整合，节约管理成本。

第4章　工作原理

4.1 架构

Moebius 集群采用无共享磁盘架构设计，各个机器可以不连接一个共享的设备，数据可以存储在每个机器自己的存储介质中。这样每个机器就不需要硬件上的偶合，只需要能够互相连通。

注意：*Moebius* 集群并非不允许使用磁盘阵列，使用本地硬盘和存储都可以，若您使用了集中存储，无非多化几个分区而已。

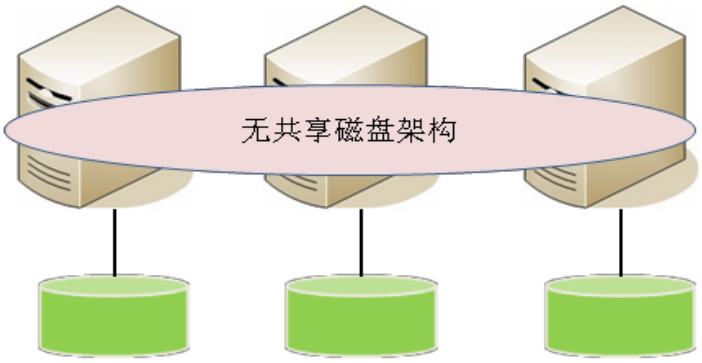

图 5. 无共享磁盘架构

Moebius 集群是一组相互独立的服务器，通过相互协作形成一个统一的整体。集群中多个节点相互连接，这样冗余的硬件架构不但可以避免单点故障而且提供了杰出的故障恢复能力。一旦发生系统失败，Moebius 集群对用户保证最高的可用性，保障关键业务数据不丢失。

一个集群数据库可以看作是一个被多个应用访问的单一数据库。在 Moebius 集群中，每个 SQL Server 实例在各自的服务器上运行。随着应用的增加，当需要添加额外的资源时，可以在不停机的情况下很容易地增加节点。一旦新增节点中的实例启动，可以马上为应用程序提供服务，在此过程中无需对应用程序进行任何修改。

4.2 核心技术

1. Moebius 集群提供强大的 SQL 解析、调度及优化加速引擎，有多达 10 种灵活的算法，将所有的访问均衡地分配到所有数据库服务器上，面对用户只是一台虚拟服务器而已。
2. Moebius 集群通过"网络心跳"及"仲裁机制"可以实现自动故障转移，当侦测到集群中某节点发生故障时，会在最短的时间内发现并通过虚拟 IP 转移技术自动将故障节点的业务转移，同时将此节点剥离出集群。
3. Moebius 集群含"实时"和"准实时"2 套数据同步引擎，可以分别针对交易型业务和报表型业务使用。
4. 在同步数据时会有 6 种同步策略，将变化的数据以最小的消耗、最快的速度同步到伙伴节点。

4.3 构成功能块

如图所示，Moebius 集群的生态系统的组成共分为如下部分：集中管理平台、解析及调度引擎、内核（Moebius Core）、故障监控引擎。

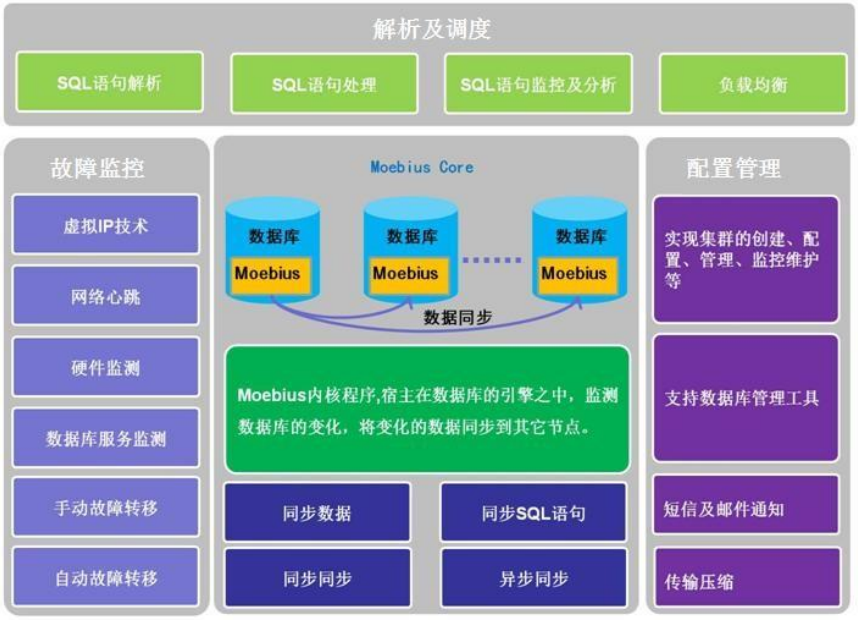

图 6. 核心模块

4.4 基本原理

SQL 解析及调度引擎 **----------**监控 **SQL** 语句，透明地切分应用与数据库

解析：解析应用程序传递的 SQL 语句，并作相应的优化加速及缓存。

调度：按照业务的需要将 SQL 语句调度到相应的服务器上；在对 SQL 语句进行分发时采用多种负载均衡策略，可以实现 SQL 语句一级的负载均衡。

处理：按照业务的需要对 SQL 语句进行相应的处理，包括修改、替换 SQL 语句等等。

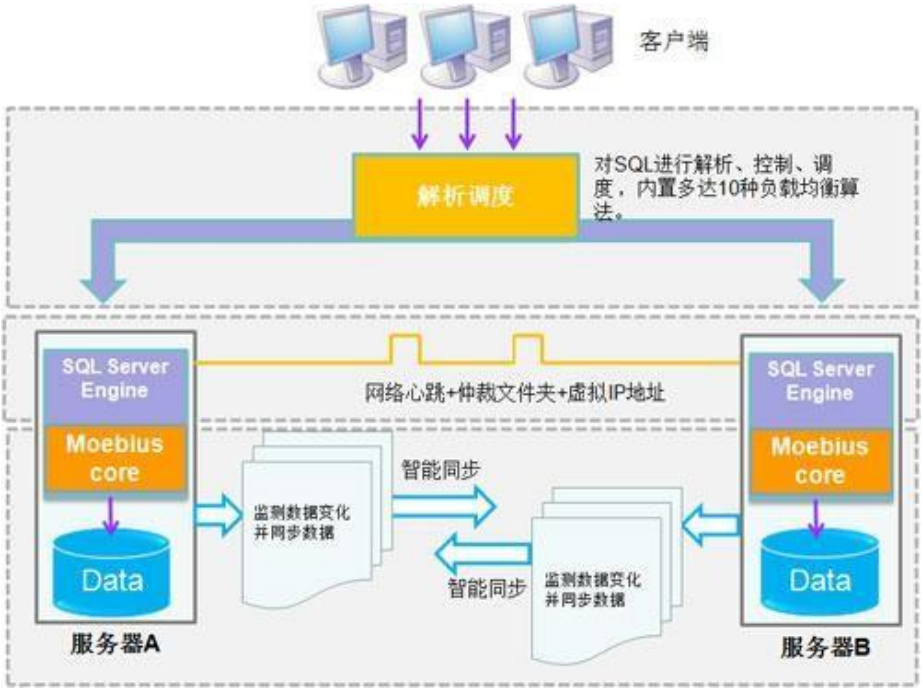

图 7. 工作原理

故障监控引擎**----------**快速发现故障节点并将其剥离

Moebius 集群通过“网络心跳”及“仲裁机制”可以实现自动故障监测，当侦测到集群中某节点发生故障时，会在最短的时间内发现并通过虚拟 IP 转移技术自动将故障节点的业务转移，同时将此节点剥离出集群。

数据同步引擎---------同步数据，保证数据一致性及事务的连续性

数据实时复制是构建多机高可用及负载均衡，系统实时容灾、备份所采用的一种核心技术。Moebius Core 宿主在 SQL Server 数据库引擎中，监测数据库内数据的变化并分析导致数据变化的原因，将变化的数据以最小的消耗同步到其它节点中，保证数据的实时一致性及事务的连续性。

4.5 解析及调度引擎

4.5.1 概述

Moebius 集群的解析及调度引擎对集群进行了统一的封装，通过一个设定的端口结合集群的虚拟 IP 地址来访问集群，对集群的访问和对单个数据库的方法完全一致。如：192.168.1.100，8000（虚拟 IP 地址+端口）。

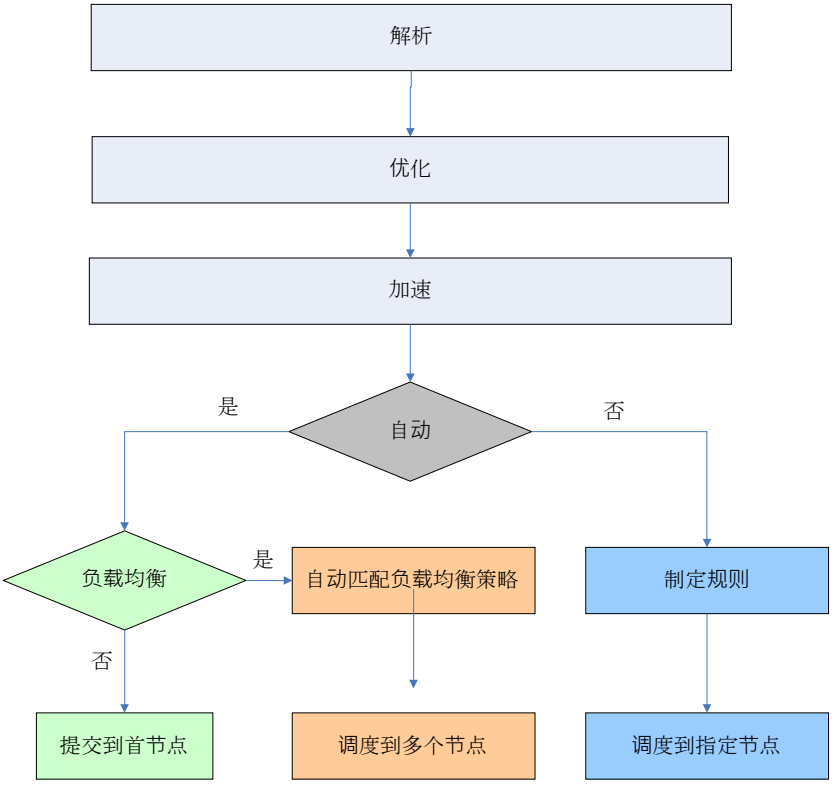

图 8. 调度原理

4.5.2 SQL解析

SQL 解析及调度引擎是在 SQL Server 的驱动层工作的，因此，对应用程序完全透明，能支持所有的 SQL Server 客户端。SQL 解析及调度引擎截断了应用程序和数据库的直接连接，起一个代理的作用，它会监控应用程序传递的数据包，通过网络包反向解析数据库驱动获取 SQL 语句。

对数据包进行解析时包括：解析数据包中 SQL 语句的类型、执行频率等参数，分析多并发是否会产生并发冲突；分析同一数据包的 SQL 语句是否可以分配到多台服务器，分析是否可以从数据缓存中获取数据。

4.5.3 优化

统计分析执行时间超长的 SQL 语句，分析原因，自动对 SQL Server 的执行计划、表的索引等进行优化。

4.5.4 加速

CPU 的快速发展使得数据库技术提供者越来越重视对多 CPU（或多核心）的并行技术的应用，一个数据库的访问工作可以用多个 CPU 相互配合来完成，Moebius 集群中，针对多核服务器提供并行加速技术，自动协调多个资源协同工作，并行加速。

4.5.5 负载均衡

负载均衡，其意思就是将负载（工作任务）进行平衡、分摊到多个操作单元上进行执行，从而共同完成工作任务。Moebius 集群中，在调度 SQL 语句时会采用多达 10 种负载均衡算法，将大量的并发访问或数据流量分担到多台节点设备上分别处理，减少用户等待响应的时间。

4.5.6 静态均衡

1. 轮询：顺序循环将请求一次顺序循环地连接每个服务器。
2. 权重：给每个服务器分配一个加权值为比例，根据这个比例，把用户的请求分配到每个服务器。
3. 随机：把来自网络的请求随机分配给内部中的多个服务器。
4. 优先权：给所有服务器分组,给每个组定义优先权，将请求，分配给优先级最高的服务器组（在同一组内，采用轮询或比率算法，分配用户的请求）；当最高优先级中所有服务器出现故障， 才将请求送给次优先级的服务器组。

4.5.7 动态均衡

1. 最少的连接方式：传递新的连接给那些进行最少连接处理的服务器。
2. 最快模式：传递连接给那些响应最快的服务器。
3. 观察模式：连接数目和响应时间以这两项的最佳平衡为依据为新的请求选择服务器。
4. 预测模式：利用收集到的服务器当前的性能指标，进行预测分析，选择一台服务器在下一个时间片内，其性能将达到最佳的服务器相应用户的请求。
5. 动态性能分配：收集应用程序和应用服务器的各项性能参数，动态调整流量分配。

4.5.8 修改 SQL语句

Moebius 集群提供强大的配置工具，用户可以按照业务的需要修改应用程序提交的 SQL 语句，将调整后的 SQL 语句提交给集群中的数据节点。（更多介绍请参考 Moebius for SQL Server 数据库集群平台用户手册）

4.5.9 替换 SQL语句

Moebius 集群提供强大的配置工具，用户可以按照业务的需要替换应用程序提交的 SQL 语句，将调整后的 SQL 语句提交给集群中的数据节点。（更多介绍请参考 Moebius for SQL Server 数据库集群平台用户手册）

4.5.10 重定向 SQL语句

Moebius 集群提供强大的配置工具，用户可以按照业务的需要将某类 SQL 语句调度到指定的数据节点上。（更多介绍请参考 Moebius for SQL Server 数据库集群平台用户手册）这些功能为用户提供了一个透明地干预应用程序提交的 SQL 语句的机会，避免更改应用程序。

4.6 故障监控引擎

4.6.1 概述

故障监控引擎用于保障 Moeb ius 集群的高可用性，Moebius 集群通过“网络心跳”及“仲裁机制”可以实现自动故障转移，当侦测到集群中某节点发生故障时，会在最短的时间内发现并通过虚拟 IP 转移技术自动将故障节点的业务转移，同时将此节点剥离出集群。

4.6.2 数据库服务与数据双冗余结构

Moebius 集群是由多个独立的服务器组成的，同步引擎使得每个成员都有实时一致的数据，服务器、操作系统、数据库、数据全部是冗余的结构。

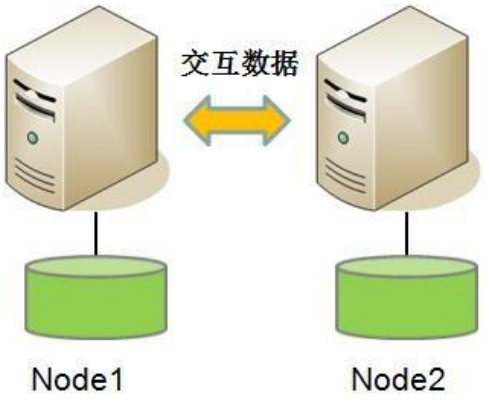

图 9. 完全冗余

4.6.3 心跳监控

在集群节点间保持着间歇的通信信号，也叫做心跳信号，是错误检测的一个机制。即通过每一个通信路径，在两个对等系统之间进行周期性的握手，如果连续没有收到的心跳信号到了一定的数目，Moebius 就把这条路径标示为失效，同时启用仲裁。

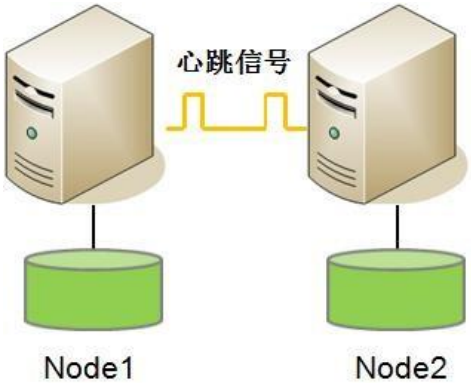

图 10. 心跳监控

4.6.4 仲裁机制

在 Moebius 集群中，当发生心跳超时时， 便立即启用仲裁机制，所有节点（2 节点集群）共同访问一个固定的文件夹（独立于集群之外），共同裁定故障节点。

Moebius 集群采用共享文件夹作为“仲裁”（一些“双机”软件也采用共享磁盘、服务器来仲裁），仲裁文件夹可以创建到任何一台服务器上，只要 Node1 和 Node2 可以访问到即可。

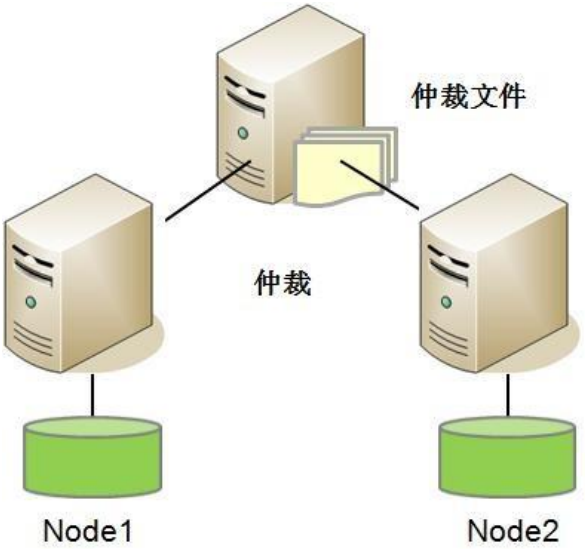

图 11. 仲裁机制

4.6.5 故障类型

Moebius 集群是数据库级别的集群技术，发生故障转移的前提一定是当前数据库不可用，物理故障、操作系统故障、SQL Server 故障等都可能导致数据库的不可用，Moebius 故障检测系统会以最快的时间发现这些故障，并做出及时的转移。常见的错误原因包括（但不限于）下列几种情况：

表 9. 常见故障类型

常见故障类型			
1	未连接或网线断开；	9	操作系统或进程故障；
2	网卡出现故障；	10	Microsoft W indows 防火墙阻止了特定端口；
3	路由器更改；	11	监视端口的应用程序出现故障；
4	防火墙更改；	12	重命名基于 W indows 的服务器；
5	端点重新配置；	13	重新启动基于 W indows 的服务器；
6	事务日志驻留的驱动器丢失；	14	SQL Server 服务故障；
7	CPU 故障；	15	内存故障；
8	主板故障；	16	IO 故障。

4.6.6 虚拟 IP

虚拟IP地址(V IP) 是一个不与特定计算机或在一个计算机中的网络接口卡(NIC)相连的 IP地址，Moebius 集群的虚拟 IP 由集群创建，可以在不同的节点之间转移。192.168.1.8 为 Node1 的实际 IP 地址，192.168.1.9 为 Node2 的实际 IP 地址。

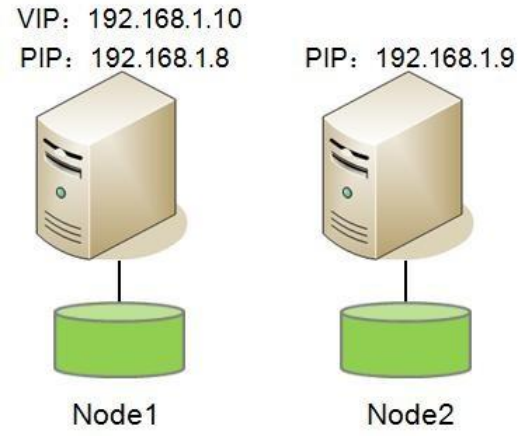

图 12. 虚拟 IP 地址

4.6.7 故障处理

应用程序通过虚拟 IP 地址访问 Moebius 集群，当集群中的某节点发生故障时，虚拟 IP 会自动转移到另一节点。

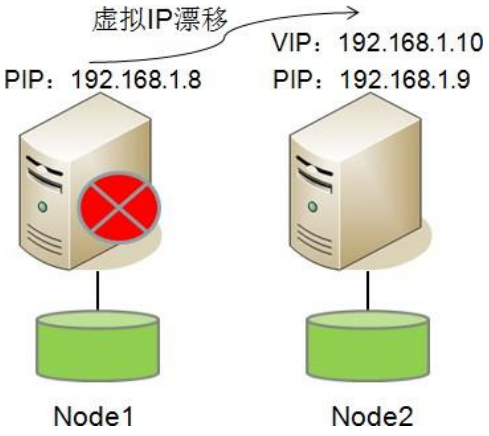

图 13. 虚拟 IP 漂移

4.6.8 故障通知

发生故障或进行切换时，自动向管理员发送故障通知邮件、手机短信。

4.6.9 故障转移时间

由于 Moebius 集群中，所有节点都处于活动状态，故障切换速度比双机软件要快。共分 3 个步骤，故障转移时间 T=T1+T2+T3

1. 查出故障(心跳／资源监视)T1
2. 申请仲裁 T2
3. 转移虚拟 IP 地址 T3

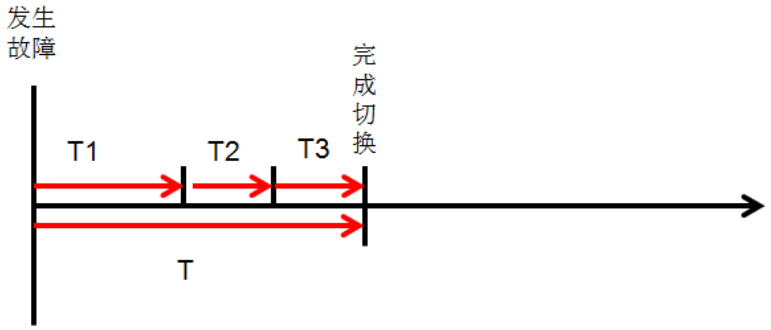

图 14. 故障转移时间

4.7 数据同步引擎

4.7.1 概述

构建 Moebius 集群的每个节点都是独立的 SQL Server 实例，内核程序（Moebius Core）驻留在每个 SQL Server 实例的内核中，监测数据库内数据的变化，同时还要分析引起数据变化的 SQL 语句的类型及其特点，经智能分析后，以最经济的方式完成与其他节点的数据同步，为了满足不同的应用场景，Moebius 集群提供不同的同步技术。

4.7.2 动态并行同步

Moebius Core 驻留在每个机器的数据库引擎中，监测数据库内数据的变化， Moebius 集群采用专有的同步技术，是在 SQL Server 的执行过程中完成数据同步的。同步引起同步数据时，具备 SQL Server 当时的执行环境，因此，会获知导致数据变化的原因，进而采取同步数据与同步 SQL 语句相结合的动态策略。同时，开启多线程，并行同步数据，而非双机软件所采用的串行复制。另外同步的过程是在事务的环境下完成的，保证了多份数据在任何时刻数据的一致性。

4.7.3 实时同步

数据同步采用的是完全同步的方式，数据同步完成后客户端才会得到响应，同步过程如下：

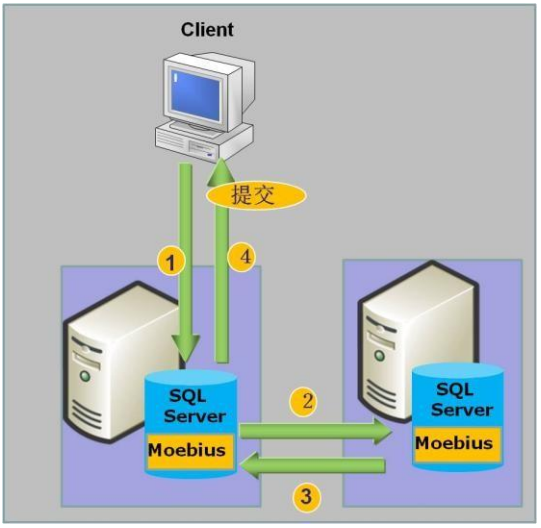

图 15. 实时同步

4.7.4 准实时同步

数据更新和数据同步是两个独立的过程，是一个异步的过程，同步过程如下：

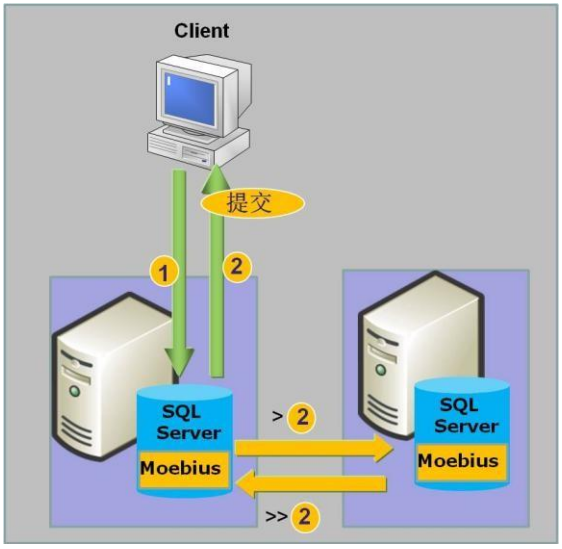

图 16. 准实时同步

4.7.5 同步策略

正因为 Moebius Core 宿主在数据库中的创新，让中间件不但能知道数据的变化，而且知道引起数据变化的 SQL 语句类型，根据 SQL 语句的类型智能地采取不同的数据同步策略以保证数据同步成本最小化。

数据压缩

如果需要同步的数据中包含文本、二进制等大数据类型， 则先对数据进行压缩然后再同步，从而减少对网络带宽的消耗和数据在传输过程中所用的时间。尤其对于网络带宽资源非常稀缺的场景。通过 fire_compression_bytes 参数进行阀值控制。

批量执行重复性数据

如果需要同步的数据中包含重复性的数据，则中间件会把重复性的数据合并到一个同步命令中只执行一次，从而减少执行的次数。

例如：执行语句 UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE LastLoginDate < '2008-08-08' 共修改了 600 条数据，这 600 条数据的 DeleteFlag 列具有相同的值，中间件会把这些相同的值合并到一个同步命令中去，同步的 SQL 语句为：UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID
IN(1, 3, 4, 5, 7, 10, 13, 15, ..., 895, 897, 899, 1000)。而不是逐条的去同步：

```
UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID = 1
UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID = 3
UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID = 4
        .
        .
        .
        .
        .
        .
UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID = 899
UPDATE dbo.UserInfo SET DeleteFlag = 1 WHERE UserID = 1000
```

该策略对数据进行批量更新、批量删除的场景下比较适用。 通过 rows_in_command 参数进行阀值控制。

升级数据库锁（锁优化器）

如果更新的数据量非常大，SQL Server 本身会对锁进行升级，将大量较细粒度的锁（例如行）转换为少量较粗粒度的锁（例如表）从而减少系统开销。中间件在同步之前先检查当前 SQL Server 的锁的粒度，如果锁已经升级，则中间件先对目标数据库直接进行锁升级然后再同步数据。从而避免了目标数据库锁升级的过程。通过 fire_lock_optimizer_rows 参数进行阀值控制。

同步 **SQL** 语句（同步优化器）

如果更新的数据量非常大，超过了设定的阀值，同步大量的数据势必会消耗大量的网络带宽并且延长同步的时间，甚至会造成网络拥堵。这时候中间件首先获取导致数据变化的 SQL 语句，分析该 SQL 语句的类型以及执行成本，并选择是把变化的数据同步过去还是把导致数据变化的 SQL 语句同

步过去。该策略在批量更改数据的时候非常有用，大量的减少网络带宽的消耗，降低同步时间。通过 fire_sync_optimizer_rows 参数进行阀值控制。

并发执行 **SQL** 语句

即使使用了同步 SQL 语句策略，总的执行时间也相当于执行两次 SQL 语句的时间。如果这个时间还是不能接受。可以通过中间件提供的系统存储过程 usp_MBS_CMD 在集群中的各个节点数据库中并发执行 SQL 语句，使执行时间降低到相当于在单机数据库中执行一次的时间。

第5章 部署模式

在系统建设初期，不论是用户还是应用软件厂商，关注最多的是软件功能（业务逻辑的实现）；在系统的运行期，用户关注的是数据及提供数据服务的装备，Moebiusfor SQL Server 作为 SQL Server 数据库上的一个专业集群平台，伸缩度非常广，用户可以从微观（调整 SQL 语句）和宏观（调整系统结构）角度入手，根据自身业务的需要将集群部署成不同的模式。

5.1.1 优化加速模式

信息系统是一个整体工程，在系统的运行中，我们经常发现应用程序中 SQL 写法不合理或者使用低版本的语法规范，此时,需要调整这些语句。可是很多时候应用程序是无法更改的，Moebius 集群提供了透明地干预方式，无需更改应用程序。

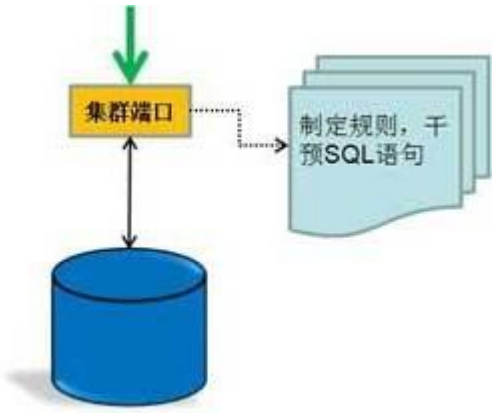

图 17. 调整与优化应用程序

5.1.2 实时灾备模式

企业的数据库保存着企业的重要信息，一些核心数据甚至关系着企业的命脉，单一设备根本无法保证数据的安全性，一旦发生丢失，很难找回。Moebius 集群可以将源系统的数据实时复制到目标系

统，从而建立实时冗余的数据，保证数据的安全。具有带宽占用少、同步效率高、数据实时性高、数据一致性保障好的特点。

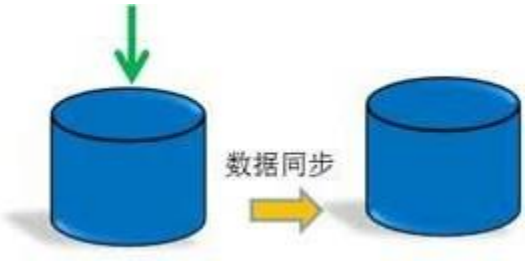

图 18. 灾备模式

5.1.3 高可用模式

数据库作为信息系统的核心，起着非常重要的作用，单一设备根本无法保证系统的持续运行，若发生故障，将严重影响系统的正常运行，甚至带来巨大的经济损失。Moebius 集群通过“网络心跳”+“仲裁机制”可以实现自动故障监测，当侦测到集群中某节点发生故障时，会在最短的时间内发现并通过虚拟 IP 转移技术自动将故障节点的业务转移，同时将此节点剥离出集群，保证系统 7*24 不间断运行。

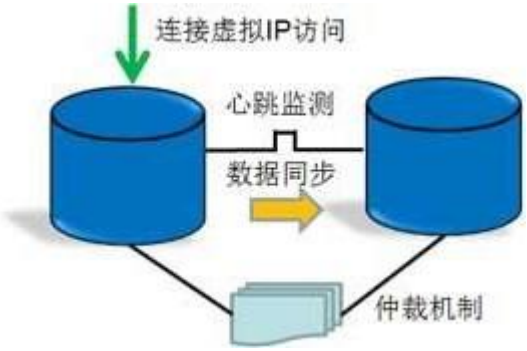

图 19. 高可用模式

5.1.4 负载均衡模式

伴随着企业的成长，在业务量提高的同时，数据库的访问量和数据量快速增长，其处理能力和计算强度也相应增大，使得单一设备根本无法承担。在此情况下，若扔掉现有设备做大量的硬件升级， 势必造成现有资源的浪费，而且下一次业务量提升时，又将面临再一次硬件升级的高额投入。Moebius 集群提供强大的 SQL 解析及调度引擎，有多达 10 种灵活的算法，将所有的访问均衡地分配到所有数据库服务器上，面对用户只是一台虚拟服务器而已，在需要更高数据库处理速度时，只要简单地增加服务器就可以得到扩展。

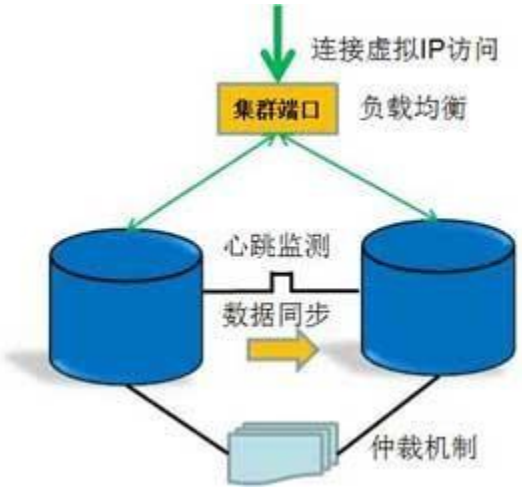

图 20. 负载均衡模式

5.1.5 读写分离架构下的负载均衡模式

读写分离是构建分布式系统的一个重要架构，采用这种架构不但可以有效扩展系统，而且可以避免不同业务对数据库访问时的相互锁定。要想实现这种架构，传统的做法需要应用层针对不同的业务（读写与只读）作分离，分别配置不同的连接串，这就面临着要重新更改应用程序，给用户带来很大的麻烦。

Moebius 集群透明地支持读写分离架构，集群可以解析所有的 SQL 语句并自动分离查询操作，无需改动应用程序。

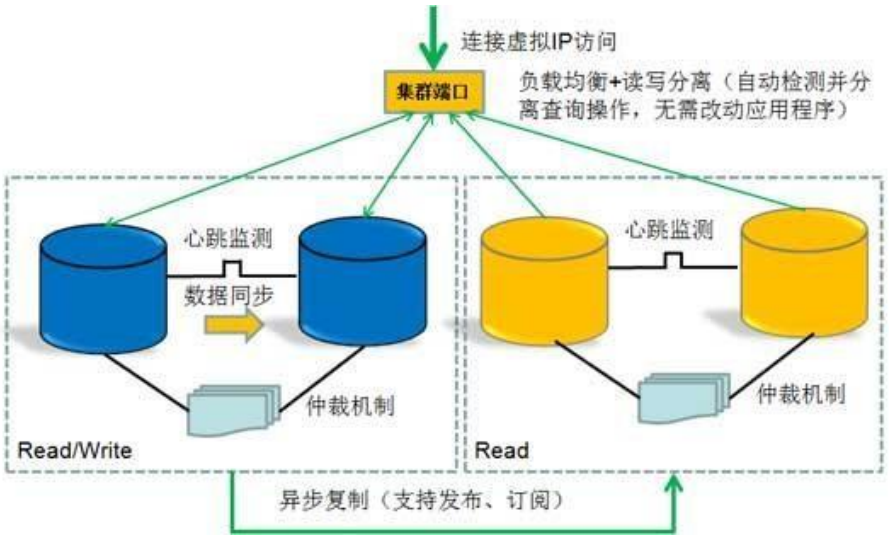

图 21. 读写分离架构下的集群

读写业务组处理和交易相关的实时查询和事务处理，只读业务组处理一些非实时的交易明细、报表类的汇总查询等，这样可以有效地保证事务性操作和统计性操作的速度。读写业务组与只读业务组中具体节点的数量根据用户的实际确定，Moebius 集群在读写分离模式下可以配置成如下几种：

1 主 1 从（读写分离）；

1 主多从（读写分离+只读组的负载均衡+只读组的高可用）；

多主 1 从（读写分离+读写组的负载均衡+读写组的高可用）；

多主多从（读写分离+负载均衡+高可用）。

第6章 Moebius 集群与传统技术的对比

企业在构建数据库平台时，根据侧重的方向和试图解决的问题，往往会选择三大类软件产品：双机热备软件、灾备软件、负载均衡软件。

6.1 双机热备软件

是用双机热备软件的目的是保证数据库的可用性，当数据库服务器发生故障时，备用服务器可以提供服务，避免由于服务器故障而导致业务中断。按照技术实现可以分为两类：基于共享磁盘的双机热备和数据库镜像技术。

6.1.1 基于共享存储的双机热备

结构要求：至少需要 2 台服务器、一台独立存储设备，共 3 台设备。

基本原理：运行时，主节点处于工作状态，另一个处于备份状态，数据存于共享磁盘中，由主节点接管。双方通过“心跳”机制来检测对方的运行状态，当主节点出现故障时，备用节点启动来接管 数据，对外提供服务，这类技术保证了数据库应用的可用性，没有保证数据的可用性。

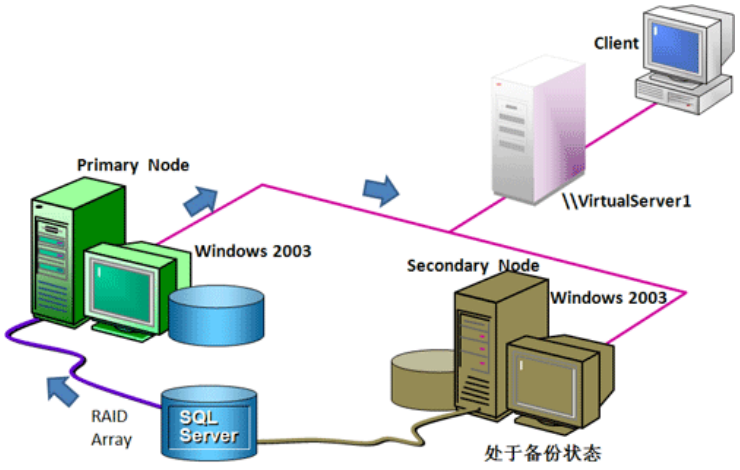

图 22. 基于共享存储的双机热备

这类技术可以说是一个传统的技术了，市场上相当广泛，可以用在应用程序服务器，也可用在数据库服务器或其它服务器，常见的技术如：微软 MSCS、Ver itas、Redhat、RoseHA 等。

6.1.2 数据库镜像

结构要求：至少需要 2 台服务器（自动切换时需要 3 台服务器）。

基本原理：这种技术是把事务先交给主服务器来完成，然后这些事务再被串行地交给备份服务器执行同样的操作以保证数据的一致性，在主服务器出现故障时可以转移到备用服务器。

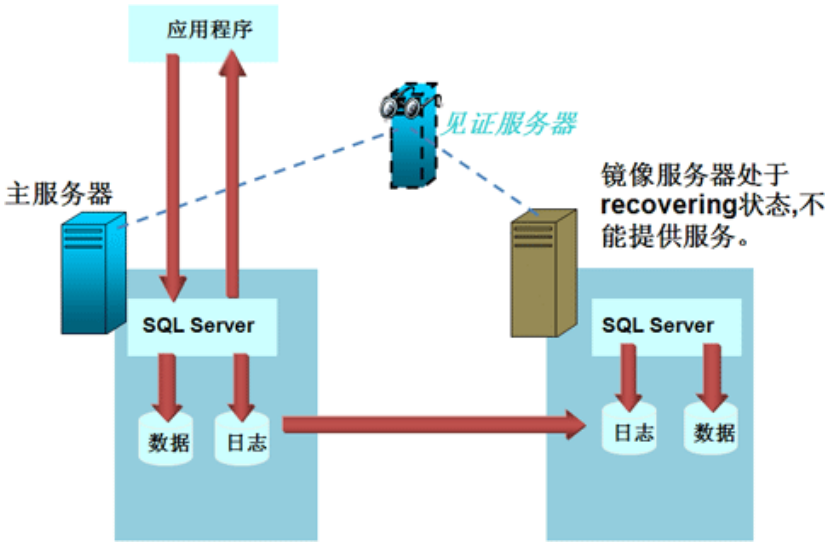

图 23. 数据库镜像

这类技术降低了用户实现可用性的门槛，无需共享磁盘，从构成上讲是一个冗余的结构，即可以同时实现应用和数据的可用性。由于镜像服务器也就处于 recovering 状态，因此不能对外提供服务。常见的产品微软 Mirror、DoubleTake、 Ver itas and Legato、Rose Mirror

6.2 灾备软件

结构要求：至少需要 2 台服务器

基本原理：这类技术侧重于数据的保护，将主数据库的文件通过磁盘镜像技术复制到另一磁盘中， 当主数据库服务器的磁盘发生灾难性事故时，避免数据丢失。常见产品如：EMC 的 TimeFinder 系列

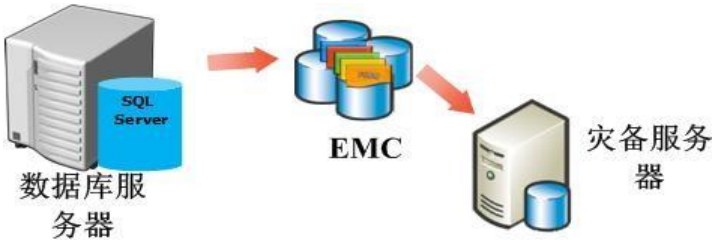

图 24. 磁盘镜像

6.3 对比分析

今天，用户的数据库部署模式分别为：单机、双机热备、灾备、负载均衡四种模式，我们分别针对这几种部署模式分别从性能、可用性、数据安全、扩展性、透明性、易用性 6 个维度来进行对比。

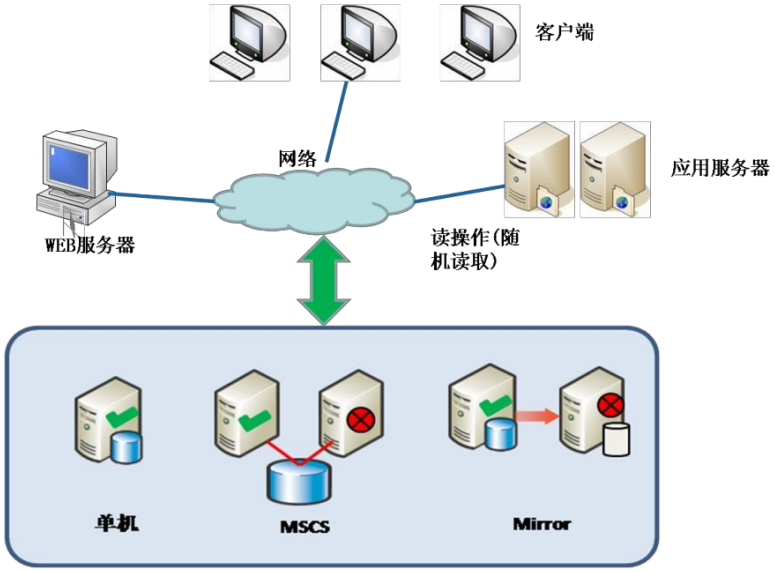

图 25. 当前的数据库部署模式

6.3.1 Moebius集群与双机、灾备的功能对比

表 10. 对比分析

一	性能	单机	共享存储双机	镜像	灾备	**Moebius** 集群
1	SQL 优化及加速		无	无	无	支持
2	负载均衡		无	无	无	静态+动态
3	吞吐量影响（读）		无	无	无	成倍提升
4	吞吐量影响（写）		无	小	小	极小
二	可用性	单机	共享存储双机	镜像	灾备	**Moebius** 集群
1	硬件冗余		部分冗余	完全冗余	部分冗余	完全冗余
2	软件冗余		完全冗余	完全冗余	部分冗余	完全冗余
3	数据冗余		无	冗余	冗余	冗余
4	共享存储		必须	无要求	无要求	无要求
5	故障转移方式		手动+自动	手动+自动	手动	手动+自动
6	数据实时性			准实时	准实时	实时
三	扩展性	单机	共享存储双机	镜像	灾备	**Moebius** 集群
1	扩展方式		向上	向上	向上	向外（支持 16 台）
2	扩展效果		有限	有限	有限	持续扩展
四	透明性	单机	共享存储双机	镜像	灾备	**Moebius** 集群
1	应用透明度					完全透明
五	易用性	单机	共享存储双机	镜像	灾备	**Moebius** 集群
1	操作		复杂	复杂	复杂	简单易用
2	管理工作		学习新工具	学习新工具	学习新工具	不改变用户的习惯

6.3.2 Moebius 集群与双机软件切换速度的对比

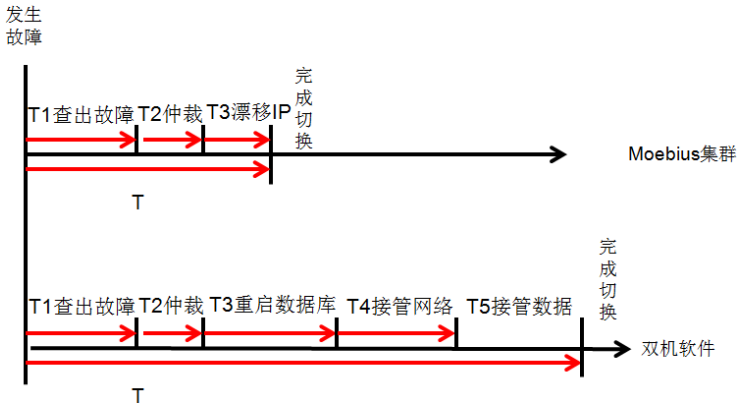

图 26. 与双机软件切换时间对比

6.3.3 Moebius集群与 Oracle's RAC对比

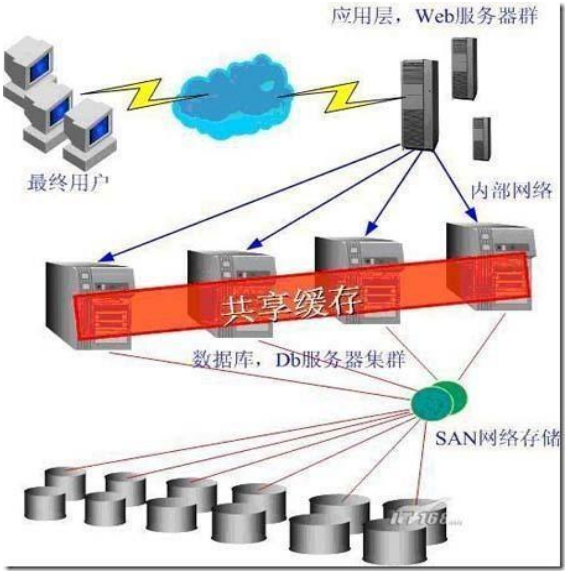

图 27. Oracle's RAC

表 11. 和 RAC 的对比

Moebius（shared-nothing architecture） 1. 节点间是松耦合，无需共享存储，真正的多点并行运算，可以充分使用多个机器的CPU、内存、IO。 2. 针对 PC Server 的方案，软硬件价格相对低廉。 3. 简单易用。
RAC（shared-disk/shared-everything architecture） 1. 结构上要依赖共享存储，多个节点同时访问一份数据，要求 IO 的性能要好，一般选择高性能存储。 2. 一般是用于小机的方案，软硬件价格昂贵。 3. 相对单机，管理更复杂，要求更高，在系统规划设计较差时性能甚至不如单节点。 4. 使用较繁琐。