基于 Lockstep 的容错技术的研究

Study of Fault Tolerance Technology Based on Lockstep

付爱英 周晶晶

Fu Aiying Zhou Jingjing

（南昌大学网络中心，江西 南昌 330031）

（Network Center，Nanchang University，Jiangxi Nanchang 330031）

摘 要：为了更加有效地理解和部署云计算平台的高可用性，本文针对实现云计算平台高可用性的容错技术进行了研究。回顾了容错技术的发展历程，重点研究了 Lockstep 技术的原理；以 Vmware 为例，分析了 Vmware Vlockstep 技术原理及 Vmware 容错实现机制，展示了实现云计算平台高可用性的技术原理。

关键词：容错；锁步；虚拟锁步；云计算

中图分类号：TP391 **文献标识码**：A **文章编号**：1671-4792（2012）07-0070-04

Abstract：The article studies the fault tolerance technology to help us to understand and implement the high availability of cloud computing platform. Firstly, the paper lookes back the development process of the fault tolerance technology and studies the lockstep emphatically, and then analyzes the technology principle of Vmware Vlockstep and the mechanism of Vmware fault tolerance, which showes the technology theory of the high availability of cloud computing platform.

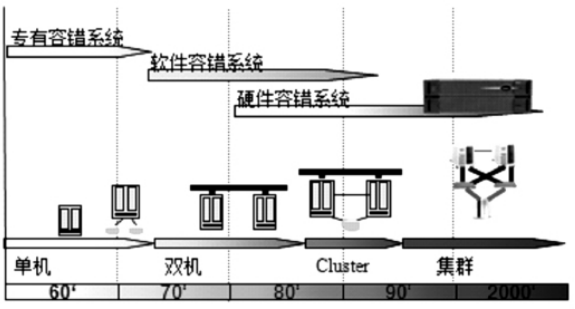
Keywords：Fault Tolerance; Lockstep; Vlockstep; Cloud Computing

0 引言

云计算是一种基于互联网的超级计算理念和模式，它易部署、易扩展，资源共享程度高，数据安全性高，是建设绿色数据中心的最优选择。云计算的虚拟化技术极大程度地提高了物理硬件的整合率和资源的利用率，降低了功耗；分布式资源调度技术则实现了各类资源灵活、快速、动态地按需调度，给实施云计算企业带来了极大的收益。随着人们对云服务的依赖性越来越强，以及越来越多的关键业务部署在云计算平台中，数据安全性、应用程序及业务系统运行的连续性保障是关键。在众多解决方案和实现技术中，容错技术是云计算提供安全、可靠服务的保障。在容错技术发展的历程中，Lockstep 技术是基石也是核心，而 Vmware Vlockstep 是容错技术的进一步发展，也是云计算技术持续发展的一大技术保障。

1 容错技术

容错含义比较广泛，从概念上说，容错是指服务器或运行系统对错误的容纳能力，就是要求系统能容忍任一部件的失效并继续工作，这是应用过程中对应用服务器或者运行系统稳定性追求的一个目标[1]。设计与分析容错计算机系统的各种技术称为容错技术。从发展历程来看，它经历了专有容错系统阶段、软件容错系统阶段和硬件容错系统阶段，如图一所示。



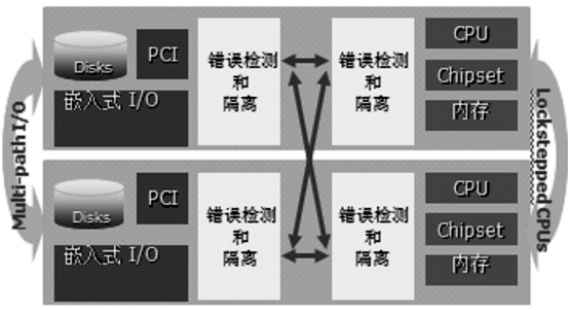
图一 容错技术发展历程图

专有容错系统主要是指采用单机容错技术实现的容错服务器。这类服务器由两套独立的硬件系统构成，但两套独立的硬件是受一套时钟锁步系统控制，即硬件间采用锁步技术实施完全同步。在系统运行的任何时候，要对两组的 CPU、内存、芯片组、磁盘、I/O 等硬件部件的处理结果进行比对，相同则执行下一步，不相同则重新计算来确保系统的可靠性。双机热备技术是基于共同文件系统建立的，也就是两台计算机的内容完全一致，而且依靠两台计算机之间建立心跳线检查对方是否存在、服务是否健全，一旦任何一方心跳消失，另一方立即接替继续提供服务，它需要双机软件来监控和处理心跳、交换资料等[2]。在理论上，这种技术的可靠性最高只能达到 99.9%。集群技术，以 Vmware 为例，它的集群就是把所有的 ESX/ESXi 主机组织起来，形成一个大的资源池，所有的虚拟机可在池中任意主机上自由移动。当集群中的某个主机出现故障时，运行在它上面的虚拟机可自动迁移到其他可用的主机上，保障业务的不间断运行。单机容错技术确保单机环境下的容错能力，而双机热备技术和集群技术则是在多机基础上借助高可用性软件来实现容错功能，从而保障应用的连续性。

在云计算时代，容错技术得到进一步发展。以 Vmware 为代表的云计算应用软件，不但实现了硬件资源的有效整合和资源池化，保障了各虚拟机应用服务的高可用性，而且基于 Vlockstep 技术的容错功能，实现各虚拟机的容错，进一步提升系统的高可用性。

2 Lockstep 技术

Lockstep，锁步，也称时钟同步，是一种容错技术。不管是硬件容错还是软件容错的实现，都依赖于冗余。许多早期的容错系统是基于冗余硬件，如 Stratus Ftserver。它就是一类基于专用硬件的容错服务器，它的 CPU、内存、I/O 等都采用冗余设计。这种容错系统通过硬件锁步技术，它的技术原理就是使相同的、冗余的硬件组件在同一时间内处理相同的指令，在一个组件失效的同时，另一组件作为一个激活的备用组件继续正常运转，并且避免系统的死机[3]。这是部件级别的冗余，即主机内部有冗余的 CPU部件和 I/O 部件，同时 CPU 部件和 I/O 部件交叉通讯，实现方式如图二所示。



图二 部件冗余锁步模式

图二所示的锁步模式可以保持多个 CPU、内存精确的同步，在正确的相同时钟周期内执行相同的指令，从而能够在不间断处理和不损失数据的情况下恢复正常运行。这种方式消除了系统内部包括 CPU、内存、I/O 控制设备以及硬盘甚至底板的单点故障。

锁步技术包括插槽级锁步技术和内核级锁步技术。内核级锁步技术通过消除内核中未检测到的错误来提高数据完整性和应用可靠性，它与现有插槽级锁步技术相结合，还可确保计算结果在内核和插槽间保持一致，从而为平台提供更高的可靠性、可用性和可管理性。

锁步技术是实现容错功能的核心，单机容错是其技术本质的体现。然而，随着应用业务的发展以及更高计算能力要求的提出，单机容错已经向多机方向发展。尤其是云计算时代，虚拟化技术的应用，业务连续性的要求，使得在 Lockstep 容错技术发展起来的 Vmware Vlockstep 容错技术成为云计算的高可用性保障技术。

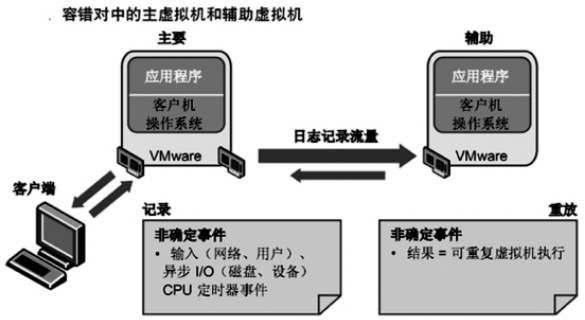
3 Vmware Vlockstep 技术

Vmware 是一个知名的虚拟化软件，也是推进云计算发展的贡献者之一。Vmware 容错是利用虚拟化技术的封装特性来建立硬件式容错虚拟机，它不需要定制硬件或软件，也不需要修改或重新配置

客户操作系统和应用程序，而是依赖两种技术：Vlockstep 技术和故障透明转移技术。

3.1 Vlockstep 技术

Vlockstep 技术，即虚拟锁步技术。在 Vmware 中，容错软件为每个虚拟机（主虚拟机）在其他 ESX/ESXi 主机上建立一个虚拟机副本（辅助虚拟机）。该辅助虚拟机同样处于活动状态，并接收与主虚拟机相同的输入，即它所完成的所有任务都依照主虚拟机的变动，主、辅助虚拟机之间是一种虚拟锁定同步关系。实现主、辅助虚拟机之间这种关系的技术就是虚拟锁步技术。当建立锁步关系后，辅助虚拟机以虚拟锁步方式和主虚拟机一起运行，并在相同的时间周期内严格执行相同顺序的 X86 指令，从而使两机的状态在虚拟机的指令执行的任何时间点均相同。图三所示的 Vlockstep 结构显示主、辅助虚拟机之间的关系。



图三 Vlockstep 结构图

如图三所示，主虚拟机正常启动后，Vsphere 自动在另外一台物理服务器上创建并启动辅助虚拟机（即辅助虚拟机与主虚拟机共享相同的虚拟机档案文件），建立主、辅助虚拟机之间虚拟锁步关系，并透过私有网络保证辅助虚拟机与主虚拟机完全同步。Vlockstep 结构中，只有主虚拟机发送和接收网络包，辅助虚拟机仅仅是个安静的倾听者。Vlockstep 技术只让主虚拟机从处理器以及虚拟 I/O 设备中捕获所有的非确定性事件，如网络及用户的输入、磁盘等设备的异步 I/O、CPU 定时器事件等，并通过私有网络被发送给辅助虚拟机。辅助虚拟机由此获得与主虚拟机同样输入数据，并在一秒内进行重演，从而确保主从数据的一致性。

在 Vlockstep 技术中，因为主、辅助虚拟机都执行相同的指令流，并且同时发起 I/O 操作，所以它们的数据始终是一致的。它们之间的区别在于输出待遇不同：主虚拟机的输出总是有效的、可见的，而辅助虚拟机的输出总是被压制和隐藏着，外界感觉不到它的存在。辅助虚拟机和主虚拟机一起被系统作为一个单元进行管理。Vlockstep 技术提供完整性保障体现在：每一个主、辅助虚拟机的边界是一样的，包括客户操作系统、应用程序状态以及所有硬件的状态；辅助虚拟机必须仅能看见不确定性输入，所以私有网络只需使用常规的 1Gbps 网络接口和交换设备。因为主、辅助虚拟机都是活动的，并且以相似的速度执行相同的指令流，所以整体性能的影响是最小的。

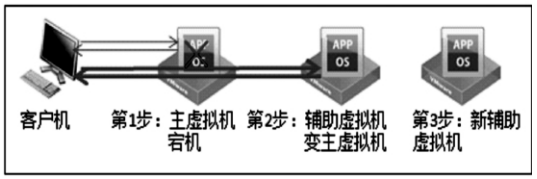
3.2 反关联性规则

关联规则反映一个事物与其他事物之间的相互依存性和关联性，根据事物之间的关联关系，就能从其中一个事物预测到其他事物。而反关联性规则就是要求事物之间遵守无关联性规则，如 Vmware 容错系统就要求主虚拟机和辅助虚拟机遵循反关联性规则，即当主虚拟机位于一个物理主机上时，它的辅助虚拟机必须驻留在另一个物理主机上，要严格遵守辅助虚拟机的不关联性规则，但主、辅助虚拟机是作为一个单元进行管理的[4]。只有遵循这种反关联性规则的主、辅助虚拟机之间，才能够建立虚拟锁步关系。

3.3 故障透明转移

Vlockstep 技术保证了主、辅助虚拟机之间数据的一致性，而且它们是作为一个单元进行管理，但对外只有主虚拟机执行工作负载。通过主虚拟机和辅助虚拟机可持续交换检测信号，使得虚拟机对中的

虚拟机能够监控彼此的状态。容错系统启动后，两台物理机上的管理程序系统之间建立一个心跳线，主、辅助虚拟机通过自身的 Tools 发送心跳到管理中心，如果管理中心检测到主虚拟机所在的主机丢失，则将不再发送心跳。此时，依靠透明故障转移技术，辅助虚拟机立刻接管主虚拟机的一切工作，并成为主虚拟机。所有关于虚拟机状态的信息都会被完整的保留，存储在内存中的数据不需要被 Re-entered 或 Reloaded，服务不会经历任何中断。图四显示了透明故障转移的过程。



图四 透明故障转移过程图

在图四中，假设运行主虚拟机的物理机出现故障宕机了，如第 1 步所示，辅助虚拟机的管理程序接收到失效的通知后，立即松开同主虚拟机建立的 Vlockstep。辅助虚拟机有来自主虚拟机完整的 I/0 操作，并且它提交所有挂起的 I/O 操作，然后执行“去往”操作成为新的主虚拟机，如第 2 步所示。这将结束以前主、辅助虚拟机之间的关系，例如，“去往”之后，新的主虚拟机启动接受来自物理网络接口的输入和启动磁盘的读写，这个过程没有数据丢失和中断，而且是自动完成的。初始故障转移后，系统自动选择一个有足够资源的物理主机建立辅助虚拟机，如第 3 步所示。新的主虚拟机与新的辅助虚拟机再建立 Vlockstep，即建立冗余。这个过程是透明和自动的。反之，当辅助虚拟机发生故障时，主虚拟机收到信息后，松开与辅助虚拟机的 Vlockstep 关系，系统自动选择一个有足够资源的物理主机建立辅助虚拟机，并再建立 Vlockstep，即建立冗余。透明故障转移技术使用文件共享存储的锁定，以协调故障转移，从而保证只有一台继续作为主要运行的虚拟机。

4 结束语

Lockstep 技术是容错技术的基础，Vlockstep 技术是容错技术的发展。基于 Vlockstep 技术的 Vmware 容错系统结合了容错服务器和软件集群的优点，它提供了硬件式的容错系统。在云计算平台中实施和部署容错系统，可以最大限度地提高数据中心的正常运行时间和减少停机管理成本，保障零停机、零数据丢失，并保障了应用程序的连续可用性。容错技术是推进云计算发展的保障之一。

参考文献

[1]陆阳，王强，张本宏，等.计算机系统容错技术研究[J].计算机工程,2010，,36（13）:230-235.

[2]冯海军.双机热备技术的应用[J].科技信息,2008，(14):27-27.

[3]乔正洪，邓新莉，余兵，等.Stratus 容错服务器在电力营销系统中的应用研究[J].重庆工商大学学报(自然科学版),2009，26（03）:277-281.

[4]付飞.浅析云计算节能技术与应用[J].信息通信技术，2012，(01):31-32.

作者简介

付爱英（1972—），女，硕士研究生，工程师，主要从事网络应用安全研究。