GreatTurbo HA 11

技术白皮书

企业级Linux高可用解决方案

北京拓林思软件有限公司

1. 概述

GreatTurbo HA 11是本公司推出的、为满足 Linux 平台电信级和企业级应用的高可用产品。它提供的“N+1”模式的热备方案能够更好的满足用户业务的连续性和可靠性，可以昼夜不停地提供24×7的服务;并且能够满足不同应用对高可用的要求。 所谓 N+1,即 1 个节点(备节点)备份多个节点(主节点),当某个主节点发生故障时,服务自动迁移到备节点,主节点恢复后,服务再迁移回去。

GreatTurbo HA 11 支持主流 Linux 平台的 N+1 高可用,包括 Greatturbo Enterprise Server 11/12,SUSE Linux Enterprise Server 10/11,Red Hat Enterprise Linux 5/6,CentOS 5/6,以及 Newstart CGSL V3/V4等，适用于i386、x86_64、IA64、 openpower等主流的硬件平台。

从 2000年开始,本公司推出了高可用系列产品-GreatTurbo HA。GreatTurbo HA 11是在以往产品的基础上,根据市场的实际需求和企业级用户多年实践经验的总结,依据已有成熟架构的基础开发的。它能够为LAMP(Linux、Apache、Mysql、Perl/PHP/Python)架构的应用和企业级用户提供更加可靠和可扩展的服务。GreatTurbo HA 11提供了更好的可靠性和可扩展性,更高的性价比,更好的易用性和可管理性,完全满足企业级应用所要求的RASM(Reliability,Availability,Scalability,Manageability)特性。

2. GreatTurbo HA功能简介

GreatTurbo HA 11是专注于Linux上的高可用性产品,提供“N+1”模式的多节点高可用性的双机集群系统。当集群中的某个节点由于软件或硬件原因发生故障时,集群会利用资源切换的方法保证整个系统继续对外提供服务,从而为企业24x7的关键业务应用提供了强大的保障。GreatTurbo HA 11提供对各种应用程序的支持,包括各种数据库应用、WEB应用、Mail应用等等,而其简便的安装和设置、详细的日志信息,减轻了用户日常的维护工作,其中跨平台的远程管理和监控使得系统具有更灵活性。GreatTurbo HA 11同时提供图形化界面和命令行界面两种配置管理工具,使得系统管理员的操作和管理更加方便。

2.1. 应用支持

当我们通过硬件(服务器、交换机、电子开关等)和软件(操作系统平台、HA系统软件、应用软件等)搭建一个高可用群集环境的时候,首先我们需要明确的是,高可用系统软件能否支持和管理我们的应用程序。GreatTurbo HA 11能够支持绝大多数的 Linux 环境下的应用程序,支持的典型应用程序类型如下:

- 通用的,无需修改的应用程序:GreatTurbo HA 11支持大多数Linux平台的应用程序,这些应用大多数是能够接受几秒种的停机时间的业务。
- 数据库应用:GreatTurbo HA 11能够很好的支持各种数据库产品,包括Oracle,MySQL,Sybase和IBM DB2数据库。
- 各种文件服务:GreatTurbo HA 11能够为各种类型的文件服务提供高可用集群功能,如NFS和SMB/CIFS(使用Samba)。
- 主流的商业应用软件:GreatTurbo HA 11能够很好的支持主流的商业应用软件,如SAP,Oracle Application Server和Tuxedo。
- 互联网和开放源代码的应用:GreatTurbo HA 11可以很好的支持各种流行的互联网应用软件和各种开放源代码产品,如Apache,Wu-ftp,VSFTP等。
- 邮件服务软件:如Sendmail和Domino。

2.2. GreatTurbo HA 11 的技术特点

2.2.1. 支持“N+1”模式的多节点热备方式

GreatTurbo HA 11 设计为“N + 1”模式多节点集群系统,集群软件同时运行在多台主机上。对于主机上服务的配置,根据用户的需要,可以设置为N台主机提供服务,一台备机待命的“N + 1”模式。相对于双节

点模式的集群系统，多节点集群系统能够同时为多个服务提供热备，将企业多台提供网络资源的计算机，有效的组合成一个保证多个核心应用服务连续运营的高可用集群系统。

2.2.2. 动态增删节点

在很多情况下，我们需要在高可用集群中现有节点的基础上动态地增加或删除节点，GreatTurbo HA 11 的“N + 1”模式支持动态增加或减少N的数量，即保证现有服务不中断的情况下，在高可用集群系统中增加或删除节点，以满足企业对提供网络服务的计算机动态伸缩的需要。

2.2.3. 备机资源管理

GreatTurbo HA 11 的“N + 1”模式提供多主一备的资源管理模式，备机可对最大资源接管数进行设置，当备机所接管的资源超过最大资源接管数后，备机将不再接管资源，而是将资源设置为等待状态，直到备机所接管的其它资源释放后，再继续接管其它资源。

2.2.4. 支持服务冻结

所谓“冻结(freeze)”,即 HA 冻结服务,不对服务进行任何检测、控制,用户可以对用户程序进行任何操作。Freeze 和 disable 的重要区别在于,freeze 保持服务的当前状态、停止对服务的检测等任何操作,而 disable 是要停止服务,包括停止用户程序、卸载共享存储、停止浮动 IP 等。冻结功能尤其适用于人工测试场景,比如在线升级、调试等;在类似场景下,用户一般要人工测试不同的功能,测试通过后,再正式运行,此时,HA 再进行监控。

2.2.5. 支持浮动IP组

在某些情况下,浮动 IP 需要分组，GreatTurbo HA 11 分别对每组浮动 IP 进行检测,当某个组内的所有浮动 IP 都异常,才会进行切换处理。

2.2.6. 支持磁盘镜像功能

磁盘镜像功能,是一种不需要磁盘阵列的数据共享方案。它的基本原理是通过对主备节点各自的本地磁盘分区进行实时镜像操作,使得这两个本地磁盘对主备节点而言,可以当作一个虚拟的共享磁盘设备来使用。这个虚拟的RAID-1 级别的共享磁盘设备能够作为应用的共享设备,既可以当作共享的裸设备来使用,也可以在其上创建各种 Linux 文件系统。GreatTurbo HA 11本身提供磁盘镜像功能,使得共享数据的应用不需要磁盘阵列也能够搭建高可用方案。

2.2.7. 提供多种类型的磁盘阵列支持

对于需要磁盘阵列的一些应用,如数据库应用等,需要专业的硬件磁盘阵列来保证性能。而 GreatTurbo HA 11 能够支持绝大多数的磁盘阵列设备。

目前业界使用的磁盘阵列一般分两种情况:第一种是带独立 RAID 处理器的磁盘阵列,主流厂商的 SCSI 或光纤磁盘阵列都可以适用于 GreatTurbo HA 11 的要求。第二种是使用主机 RAID 卡和磁盘柜(磁盘柜是指不具备硬件 RAID 处理器的磁盘盒) 的方式, 这种磁盘设备通过硬件 RAID 卡 clustering 技术和 RAID 的用户接口工具等,也可以满足 GreatTurbo HA 11 的共享数据存取的需求,常见的这种类型的“磁盘阵列”的典型产品有 IBM EXP400,DELL 220S,HP MSA500-G2 等。

2.2.8. 多种硬件心跳保证系统一致性

GreatTurbo HA 11 同时支持直连网线、串口和 raw 磁盘分区的方式来同步 HA 节点之间的心跳信息。可同时支持多条直连网线和串口线以及磁盘的 raw 分区作为通道,提供更高可靠性的硬件冗余方式,以保证节点之间不会发生 Split-brain 现象。其中 raw 磁盘分区的心跳通道,保证了只要主备节点能够访问共享数据,就不会发生裂脑,从而有效的确保了共享数据的一致性。即使节点之间的心跳通道都发生故障,GreatTurbo HA 11 还可以通过配置第三方参考 IP 的方式,保证节点之间系统的一致性。GreatTurbo HA 11支持配置多个第三方参考 IP,避免了第三方参考 IP 成为单一故障点。

2.2.9. 可靠的故障时切换策略

无论是否配置第三方 IP,主节点所有的网络都发生故障时,仍能够保证服务切换到正常的备节点上,不影响对外正常提供服务。

2.2.10. 支持 STONITH 技术

Stonith(shut the other node in the head),就是把故障节点重启,以保证资源被完全释放。Stonith 的方式有两种,一种是通过电子开关(power switch)来重启对方;另外一种是通过网络发送命令来重启对方。但是后者通常不起作用。有一些厂家的服务器有类似的管理界面(一种硬件设备),也可以用来作为 stonith 的工 具。 比如 IBM xServer 的 RSA(Remote Supervisor Adapter) 和 Intel 的IPMI 。GreatTurbo HA 11 对以上两种设备都可以支持。

2.2.11. 支持 Watchdog Timers

GreatTurbo HA 11 支持的三种类型的看门狗定时器为系统提供了一个稳健的 I/O barrier。最简单的就是 Linux 内核自带的通过中断处理来实现的软件 softdog 定时器,它被 GreatTurbo HA 11 用来控制后台程序的执行。

Linux 内核也支持一种硬件 NMI (non-maskable interrupt) 看门狗,这种硬件看门狗通常需要专门服务器硬件支持 (常用的是主板上的 Intel 810 TCO 芯片组) NMI 看门狗在没有检测到一个稳定正常的中断发生时就会确发节点服务器的重启动。

最后一种,就是传统的硬件看门狗定时器,它是一种 PCI 设备,在市面上很常见。当 PCI 设备的驱动没有正常的复位时,它就会强迫系统关闭或重启。

2.2.12. 智能的服务回迁以及多服务的负载分担

GreatTurbo HA 11 支持优先节点的设置,可以把一些服务设定到指定的优先节点。当优先节点故障时,服务切换到另一个节点;而当优先节点又恢复时,服务会自动迁移到优先节点。这样当主节点故障时,会切换到备节点,当主节点恢复后,会再切换回来;防止多个主节点的服务都在备节点上运行。

2.2.13. 可以检测更多的故障

GreatTurbo HA 11 能够检测更多的系统故障,从而增强了高可用性集群所提供的可靠性。

- 系统故障 —— 硬件错误
- 系统紊乱 ——系统软件错误
- 存储不可访问 —— 存储错误
- 网络断开 —— 网络错误
- 集群进程故障 —— 集群软件错误
- 服务故障 —— 服务应用程序错误

2.2.14. 应用程序代理检查

GreatTurbo HA 11 通过使用应用程序代理检查某一服务是否运行。应用程序代理用于定期检查某一服务是否正常工作。如果服务没有正常运行,则相应地触发一次切换,使服务在另一节点被恢复。GreatTurbo HA 11 提供用于常用服务的应用程序代理,对于自身没有应用程序代理的服务则可以使用GreatTurbo HA 11 提供的接口进行灵活的按需定制。请同时参见本文中的“应用程序代理 API”一节。

2.2.15. 应用程序代理 API

应用程序代理 API 是一种在应用程序代理或服务检查程序和 GreatTurbo HA 11 服务进程之间的接口,在 GreatTurbo HA 11 用户手册中有详细介绍。按照此接口的规范,您可以为您的特定服务编写定制的应用程序代理。编写定制的应用程序代理的好处在于它可以根据您现场实际的负载情况为应用程序提供更精确的服务检查以及更快的切换。

2.2.16. 图形管理工具

GreatTurbo HA 11 提供了基于 Python 技术的图形管理工具,从而改善了集群的可管理性。既支持本地和远程的监控和管理,又支持 Linux/Windows 客户端的管理。

利用所提供的图形管理工具,可以方便地进行配置更改和状态监测。GreatTurbo HA 11 的图形管理工具提供了初始化集群的功能，即可以使用图形管理工具进行DRBD的配置和member config。除了提供图形管理工具外,GreatTurbo HA 11 还提供有功能同样强大的命令行配置、监控管理工具。

2.2.17. 更好的日志文件系统支持

GreatTurbo HA 11 支持与日志文件系统(诸如 Reiser 和 Ext3 等)的协同工作。这些日志文件系统特别适用于 GreatTurbo HA 11, 因为它们消除了 Ext2 等文件系统中所花费的耗时的文件系统检查,从而减少了切换时间。当系统装载时,日志文件系统仅仅要求恢复其日志。当一种日志文件系统被用于共享存储时,GreatTurbo HA 11 能够自动地进行确认,跳过不需要的FSCK 文件系统检查,并立即装载文件系统用于文件系统日志的恢复。

2.2.18. 更详细的系统故障日志信息

GreatTurbo HA 11 采用的日志函数和 Linux 的 syslogd 是一样的方式,在每个节点均有记录,每个守护进程都有自己的日志级别,可以在配置文件中指定。每一条记录的信息,包括有时间、日志级别、进程名称、进程id、消息等内容,这样可以方便用户进行应用故障现场的保护以及故障后的分析定位。同时日志的级别可以动态进行设置调整,以根据实际需要调整输出日志的信息内容。默认情况下,系统已经将日志级别设置成较为详细的信息输出,包括 HA 启动、停止过程,HA 事件触发原因,服务故障原因,服务切换过程,服务手动操作记录等。同时 GreatTurbo HA 11 还提供日志收集工具,自动收集系统以及 HA 相关信息,以便于进行故障定位。

2.2.19. 报警功能

GreatTurbo HA 11 预留有报警接口,当发生故障切换时,HA 会发送报警信息。报警方式包括短信，邮件或者SNMP TRAP等。

2.2.20. 配置的简化

member config的简化，GreatTurbo HA 11 中用配置工具配置完/etc/hosts文件之后就会有一个默认的member config的配置，如果没有其他修改，只需要直接使用这个默认配置就可以完成member config。

添加服务的简化，在添加新的服务时，可以选择一个服务的类别，如http服务或者sybase服务，HA通过这个服务类别给出一个对应的默认的服务配置，用户可以直接使用这个配置，也可以进行修改。

2.3. GreatTurbo HA 11 高可用功能设计原理

不间断的提供有效、准确的服务是高可用集群软件的设计目标。在保证用户数据完整性的前提下,当系统或服务失效时,及时的将服务切换到正常节点,同时采取必要措施,帮助失败节点能够恢复正常,这就是 GreatTurbo HA 11 作为优秀高可用集群软件所提供的功能。

2.3.1. 主机及服务

GreatTurbo HA 11 设计为“N + 1”模式多节点集群系统,集群软件同时运行在多台主机上。对于主机上服务的配置,根据用户的需要,可以设置为N台主机提供服务,一台备机待命的“N + 1”模式,也可以是两台主机同时提供不同服务,并且互为备份的“主动—主动”模式。如果用户有两台同样高配置的服务器,并希望提供两种或两种以上的服务,则可以采用“主动—主动”模式以提高系统利用率;如果用户有一台高配置的服务器和一台较低配置的服务器,希望建立高可用性服务,则可以采用“主动—被动”模式,并把服务配置成 “回切”型。

2.3.2. 监测

GreatTurbo HA 11 对于主机系统级的失败,主机间通讯的失败和所提供服务的失败都能进行准确的实时监测。

- 系统及通讯监测

  任何操作系统,都有出现死机或系统挂起的可能。系统挂起和死机不同,系统挂起时对用户的输入不再有响应,好像被锁住一样,在有些情况下,系统挂起一段时间后,有可能重又继续工作。GreatTurbo HA 11 可以准确的检测到一台主机系统挂起或死机的发生,并把服务切换到正常工作的主机上。

  为了监测对等主机的状态, GreatTurbo HA 11 集群在主机之间可以建立三种方式的任意多条连接通路,这也被称为“心跳”(Heartbeat)。“心跳”方式有 UDP/IP 连接和串行线连接以及 raw 磁盘设备三种。可以使用多块网卡,在主机间建立多条点对点的 UDP/IP 连接。如有多个串行口,则可以建立多条串行连接,若有共享的磁盘设备可以建立起通过裸设备磁盘的心跳方式。

  使用三种方式的多条连接,也就是利用冗余的硬件,提高主机间通讯的可靠性。只有当所有心跳通路全部失败时,才认为主机在通讯上失效,此时GreatTurbo HA 11 会采取及时有效的应对措施。

- 服务监测

  GreatTurbo HA 11 对于服务的状态也会定时进行监测,监测的时间间隔可由用户指定。GreatTurbo HA 11 提供一个“通用应用程序代理”,可以对各种服务进行一般性的监测。对于常用类型的服务,还有相应的“应用程序代理”可以实现具有针对性的服务监测功能。用户也可以自行编写应用程序代理,以满足特殊的需要。

2.3.3. 切换

GreatTurbo HA 11 检测到一台主机上系统或服务的失败时,正常主机首先会建立 I/O 屏障,保护共用存储设备上的数据不被失败节点修改。然后,会把故障主机上的服务切换到正常主机上,继续对外提供服务。可以把一个或多个 IP 地址绑定在服务上,在服务切换时,IP 地址也随之切换到正常主机上,所以用户仅仅在服务切换的瞬间能感觉到极短时间的服务暂停。

在服务切换的同时,如果故障主机上集群软件仍在运行,则在检测到 I/O屏障后,会把本机重启动,如果重启动后系统恢复正常,则重新加入集群,可以接管服务。这样,即使两台主机都发生故障,只要不在同一时刻发生,集群仍可保证提供服务。