

普华 i-HA 高可用集群软件 V4.0

技术白皮书

普华基础软件股份有限公司

地址：北京市海淀区容达路 7 号院 E 座 3 层。

目录

第 1 章	何谓集群系统.....	5
1.1	集群系统的概要.....	5
1.2	HA(High Availability)集群.....	6
1.2.1	共享磁盘型.....	6
1.2.2	数据镜像型.....	9
1.3	检测故障的原理.....	10
1.3.1	共享磁盘型的各种问题.....	11
1.3.2	网络分区症状(Split-brain-syndrome).....	12
1.4	集群资源的交接.....	12
1.4.1	数据的交接.....	13
1.4.2	应用程序的交接.....	13
1.4.3	失效切换总结.....	14
1.5	Single Point of Failure 的排除.....	15
1.5.1	共享磁盘.....	15
1.5.2	共享磁盘的访问路径.....	16
1.5.3	LAN.....	17
1.6	提高可用性的操作.....	18
1.6.1	操作前测试.....	18
1.6.2	故障监视.....	19
第 2 章	普华 i-HA 高可用集群软件 V4.0 的使用方法.....	20
2.1	普华 i-HA 高可用集群软件 V4.0.....	20
2.2	普华 i-HA 高可用集群软件 V4.0 的产品结构.....	20
2.3	普华 i-HA 高可用集群软件 V4.0 的软件配置.....	21
2.3.1	普华 i-HA 高可用集群软件 V4.0 的故障监视原理.....	21
2.3.2	服务器监视.....	22
2.3.3	业务监视.....	23
2.3.4	内部监视.....	24
2.3.5	可监视的故障和无法监视的故障.....	24
2.3.6	通过服务器监视可以查出的故障和无法查出的故障.....	24
2.3.7	通过业务监视可以查出的故障和无法查出的故障.....	24

2.3.8	网络分区解析.....	25
2.4	失效切换的原理.....	25
2.4.1	失效切换资源.....	26
2.4.2	失效切换型集群的系统配置.....	27
2.4.3	共享磁盘型的硬件配置.....	30
2.4.4	镜像磁盘型的硬件配置.....	31
2.4.5	共享型镜像磁盘型的硬件配置.....	32
2.4.6	集群对象.....	33
2.5	资源.....	34
2.5.1	心跳资源.....	34
2.5.2	网络分区解析资源.....	34
2.5.3	组资源.....	34
2.5.4	监视资源.....	36
2.6	开始使用普华 i-HA 高可用集群软件 V4.0 吧!.....	40
2.6.1	集群系统的构建.....	41

1. 本手册的标记规则

在本手册中，需要注意的事项、重要的事项以及相关信息等用如下方法标记。

注：表示虽然比较重要，但是并不会引起数据损失或系统以及机器的损伤的信息。

重要：表示为避免数据损失和系统、机器损坏所必需的信息。

相关信息：表示参考信息的位置。

另外在本手册中使用以下标记法。

标记	使用方法	例
[]方括号	在命令名的前后， 显示在画面中的字句（对话框、菜单等）的前后。	点击[开始]。 [属性]对话框
命令行中的[] 方括号	表示括号内的值可以不予指定 (可省)。	clpstat -s[-h host_name]
#	表示 Linux 用户正以 root 身份登录的提示符。	clpcl -s -a
等宽字体 (courier)	路径名、命令行、系统输出(消息、提示等)、目录、文件名、 函数、参数。	/Linux/3.2/en/serv er/
等宽字体 粗体 (courier)	表示用户在命令提示符后实际输入的值。	输入以下值。 clpcl -s -a
等宽字体 斜体(courier)	用户将其替换为有效值后输入的项目。	rpm -i expressclsbuilder -<版本编号>-<发行 编号>.i686.rpm

第 1 章 何谓集群系统

1.1 集群系统的概要

在如今的信息时代，持续不中断地提供服务是成功的关键。例如，仅仅由于 1 台机器故障宕机而造成对客户提供的服务全面中断。这样不仅会造成不可估量的损失，还可能失去客户的信赖。

而集群系统就是避免这种情况发生的系统。导入集群系统后，万一发生系统运行停止的情况，集群系统能够将停止时间(Downtime)缩到最短，或者能够通过分散负载的方式回避系统宕机。

所谓集群，顾名思义，就是将多台计算机集成一个群(或多个群)，从而提高可靠性和处理性能的系统。集群系统有很多种，分为以下 3 类。其中，普华 i-HA 高可用集群软件 V4.0 属于高可用性集群。

a) HA (High Availability) 集群

指在正常时将其中一方用作运行服务器，提供业务，运行服务器发生故障时将业务交接给待机服务器的集群系统。该集群主要目的是实现高可用性，可以交接数据。该集群包含共享磁盘型、镜像磁盘型、远程集群型。

b) 负载均衡集群

指能够将客户端发送的请求按照合适的负载均衡原则，分配给各负载均衡主机的集群系统。该集群的主要目的是实现高扩展性，一般情况不能进行数据交接。该集群包含负载均衡集群、并列数据库集群。

c) HPC(High Performance Computing)集群

使用所有节点的 CPU，实现单一业务的集群。该集群的主要目的是实现高性能，应用范围不大。

另外，该集群是 HPC 的一种，将更广泛的节点和计算机集群捆绑在一起的网格运算技术也成为近年来关注的焦点。

1.2 HA(High Availability)集群

为了提高系统的可用性，普遍认为将系统部件冗余化，排除 Single Point of Failure 是至关重要的。所谓 Single Point of Failure 是指因为服务器的物理硬件只有一个，如果硬件发生故障会造成业务中断这一弱点。HA 集群则是通过使用多台服务器使系统冗余化，从而将系统的中断时间控制在最小，提高业务可用性(availability)的集群系统。

HA 集群可以分为共享磁盘型和数据镜像型。下面分别介绍两种类型。

1.2.1 共享磁盘型

在集群系统中，服务器之间必须进行数据的交接。将这些数据放在共享磁盘上，多个服务器共同使用该磁盘的集群形态称为共享磁盘型。

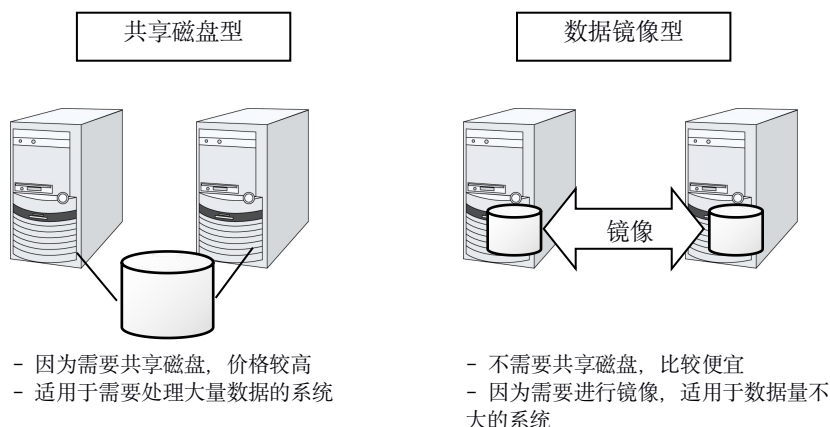


图 1-1 HA 集群配置图

运行业务应用程序的服务器(运行服务器)发生故障时，集群能够查出故障，自动启动待机服务器上的业务应用程序，使业务继续运行。该功能称为失效切换。集群进行交接的业务由磁盘、IP 地址、应用程序等资源配置。

在未进行集群化的系统中，如果在其他服务器上重新启动应用程序，客户端必须重新连接不同的 IP 地址。但是在多个集群系统中，会以业务为单位分配虚拟 IP 地址。因此，客户端不必辨别现在运行的服务器是运行服务器还是待机服务器，可以完全像连接在同一台服务器上一样继续执行业务。

交接数据时必须检查文件系统的一致性。一般使用检查命令(如在 Linux 上会使用 `fsck` 或 `chkdsk`)检查文件系统的一致性，但是文件系统越大，检查所花的时间越长，检查期间会造成业务中断。为了解决该问题，我们通过日志文件系统缩短失效切换时间。

业务应用程序需要对交接后的数据进行逻辑检查。如果是数据库，则需要进行回滚或前滚处理。通过该方法，客户端只需要重新执行未提交的 SQL 文件，就可以使业务不中断运行。

恢复故障服务器时，只要将查出故障的服务器进行物理分离，处理之后，重新连接到集群系统上，就可以作为待机服务器恢复了。在重视业务持续性的实际运用过程中，使用这样的恢复方式就足够了。

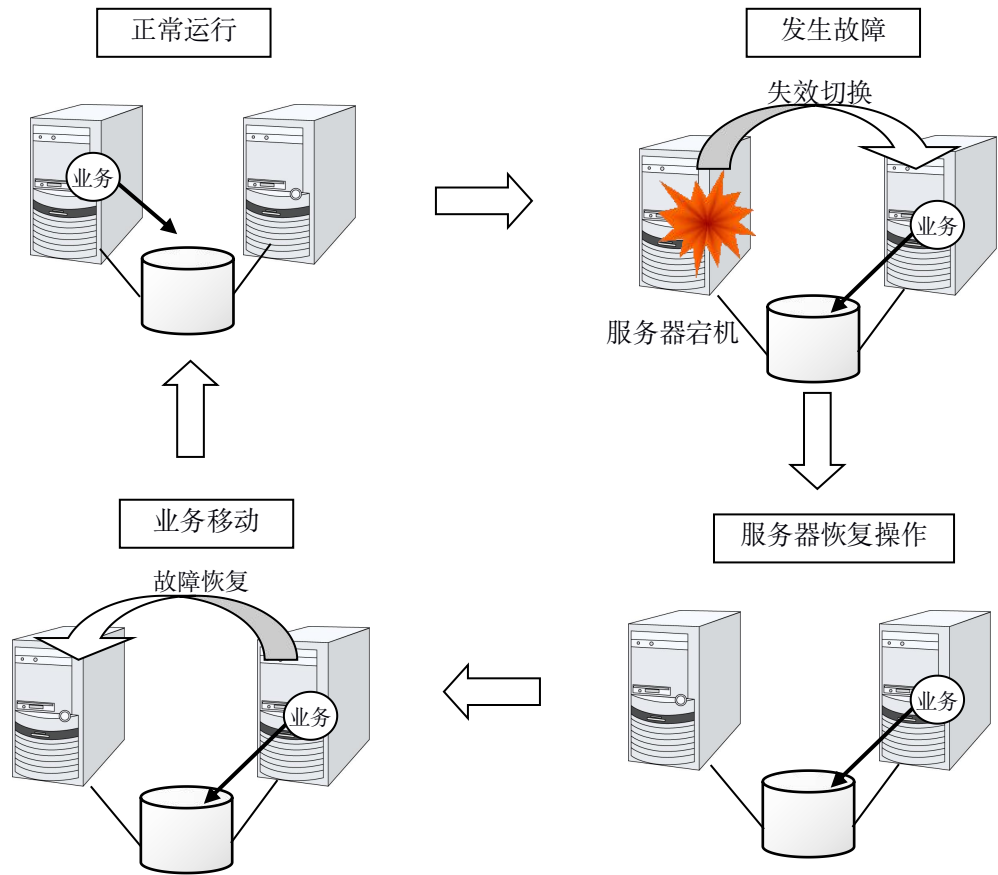


图 1-1 发生故障到系统恢复的流程

如果失效切换到的服务器配置不够，担心双向待机负荷过大，希望在原来的服务器上运行业务时，可以进行故障恢复，重新在原来的服务器上运行业务。

如图 1-2 所示，只有一个业务，在待机服务器上没有运行业务的待机形态称为单向待机。如果有两个以上业务，每个服务器既是运行服务器，又是待机服务器，这种形态称为双向待机。

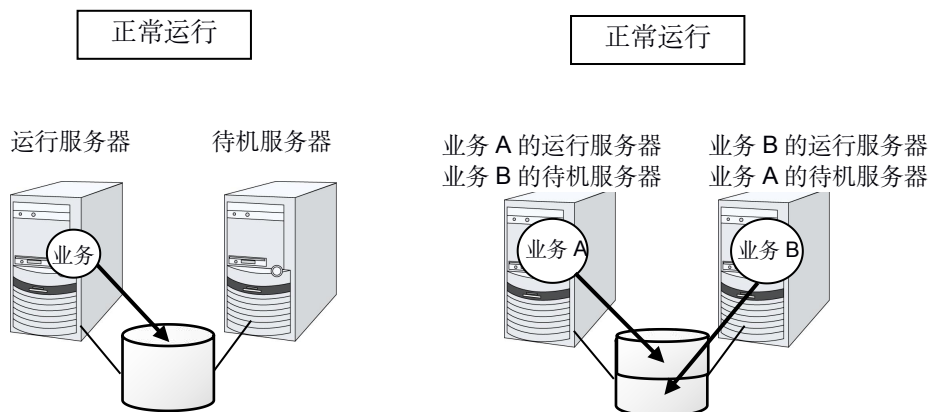


图 1-2 HA 集群的运行形态

1.2.2 数据镜像型

之前介绍的共享磁盘型适用于大规模系统，但是由于共享磁盘大多很昂贵，所以搭建系统的成本也就随之增加。不使用共享磁盘，在各服务器之间对各服务器上的磁盘进行镜像，通过该方法用低廉的价格实现相同功能的集群称为数据镜像型。

但是，由于需要在服务器之间进行数据的镜像，因此，不适用于数据量巨大的大规模系统。

应用程序发出 **write** 请求时，数据镜像引擎会将数据写入本地磁盘的同时，通过心跳线将 **write** 请求同时分发给待机服务器。所谓心跳线是指连接各服务器的网络，在集群系统中需要使用心跳线进行服务器的生存状态监视。在数据镜像类型中，心跳线除了用于生存状态监视，还用于数据传输。待机服务器数据镜像引擎将收到的数据写入待机服务器的本地磁盘，从而实现运行服务器和待机服务器间的数据同步。

应用程序发出 **Read** 请求时，则只需单纯从运行服务器的磁盘中读取数据即可。

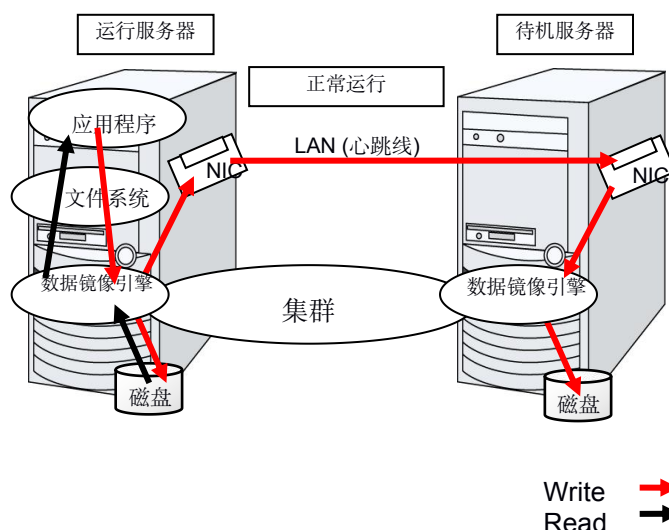


图 1-3 数据镜像的结构

作为数据镜像的一个应用示例，可以使用快照备份。数据镜像类型的集群系统由于同时在 2 处保存共享数据，因此只需将待机服务器从集群中分离开，无需备份，就可以将磁盘作为快照备份保存。

至此，我们介绍了失效切换集群、负载均衡集群、HPC(High Performance Computing)集群等多种集群系统。我们了解了失效切换集群称为 HA(High Availability)集群，其目的是通过将服务器多重化，发生故障时将运行的业务转移到其他服务器上，从而提高业务的可用性(Availability)。

1.3 检测故障的原理

集群软件一旦查出影响业务继续运行的故障，就会进行失效切换。在介绍失效切换处理的详细内容之前，先简单了解一下集群软件是如何检测故障的。

a) 心跳和服务器的故障检测

在集群系统中，最基本的故障是集群内各服务器均停止运行。服务器故障中，包括电源故障、内存错误等硬件故障和 OS 的 panic。要查出该故障，需要在服务器生存状态监视中使用心跳。

心跳也可以是确认 ping 应答这样简单的生存状态监视，但是使用集群软件还能够互相传送自身服务器的状态信息。集群软件进行心跳信息的收发，没有心跳应答时则认为该服务器发生故障，开始进行失效切换处理。但是考虑到服务器也可能是因为负载较高，收发心跳信息有延迟，在做出服务器故障判断之前留有一定的缓冲时间。因此，实际发生故障的时间和集群软件查出故障的时间之间有一定间隔。

b) 资源的故障检测

造成业务中断的原因不仅仅是配置集群的服务器全部中断，还有可能是因为业务应用程序使用的磁盘设备或 NIC 发生故障，亦或是业务应用程序本身发生故障造成业务中断。为了提高可用性，这些资源故障同样需要查出来并进行失效切换。

作为检测资源故障的方法，如果监视的对象资源是物理设备，则采取实际访问的方法。在应用程序的监视中，除了应用程序进程自身的生存状态监视，还考虑在不影响业务的范围内使用服务端口等手段。

1.3.1 共享磁盘型的各种问题

在共享磁盘型的失效切换集群中，多个服务器上共享一个磁盘设备。一般情况下，文件系统会通过保留数据的缓存，发挥出超过磁盘设备物理 I/O 性能界限的性能。

试想一下，如果某个文件系统同时被多个服务器 mount 访问结果会怎样？

一般的文件系统是不考虑其他服务器磁盘上数据的更新的，所以会造成缓存和磁盘上数据的矛盾，最终导致数据被破坏。为了防止下面介绍的网络分区解析资源带来的多个服务器同时 mount 文件系统的问题，在失效切换集群系统中使用了磁盘设备的互斥控制。

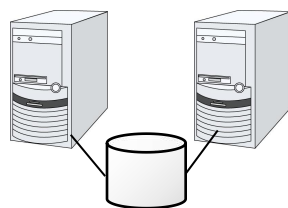


图 1-4 共享磁盘类型的集群配置

1.3.2 网络分区症状(Split-brain-syndrome)

如果断开所有服务器之间连接的心跳线，通过心跳进行的生存状态监视就会互相查出服务器宕机，开始执行失效切换处理。结果就会造成多个服务器上同时 mount 文件系统，导致数据损坏。因此，在失效切换集群系统中，发生故障时必须进行恰当的处理。

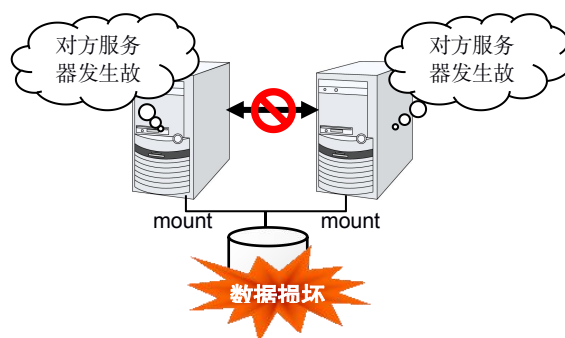


图 1-5 网络分区症状

这样的问题称为“网络分区症状”或 Split-brain-syndrome。现在就要考虑在失效切换集群中，如果所有的心跳线都被切断，该如何实现共享磁盘设备之间的互斥控制。

1.4 集群资源的交接

集群管理的资源中有磁盘、IP 地址、应用程序等。下面介绍失效切换集群系统中提供交接集群资源的功能。

1.4.1 数据的交接

在集群系统中，服务器之间交接的数据存放在共享磁盘的分区中。数据的交接无非就是将保存应用程序所使用文件的文件系统在正常的服务器上重新 mount。共享磁盘设备与交接对象服务器物理连接，集群软件需要做的就是文件系统 mount。

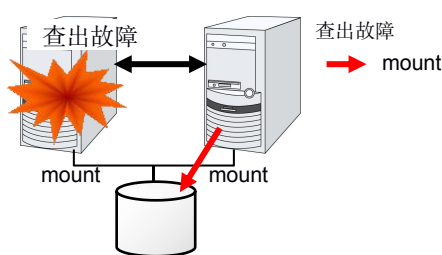


图 1-6 数据的交接

虽然这个过程看似简单，但是在设计构建集群系统时还是有很多需要注意的地方。

首先是文件系统的恢复时间问题。要交接的文件系统在发生故障之前可能正在其他服务器上使用，或者正在更新。所以交接的文件系统一般是会有些垃圾，需要检查文件系统的一致性。文件系统越大，则一致性检查所需的时间就越长，有时甚至需要花几个小时，而这个时间将直接转嫁到失效切换时间(业务的交接时间)中，导致系统可用性下降。

还有一个就是写入保证的问题。应用程序将重要的数据写入文件时，要利用同步写入等方法保证在磁盘中也写入该数据。因此，应用程序认为已经写入的数据在失效切换后该数据也能够被交接。例如，邮件服务器将收到的邮件写入缓冲处理区时，会向客户端或者其他邮件服务器发出收信结束的应答。这样，即使服务器发生故障，在服务器重启后，能够重新传输被缓冲处理的邮件。在集群系统中也一样，必须保证一方服务器写入缓冲处理区的邮件在失效切换后另一个服务器能够读取。

1.4.2 应用程序的交接

集群软件在业务交接的最后一步是交接应用程序。与容错计算机(FTC)不同，在

一般的失效切换集群中，不交接包含正在执行应用程序的内存内容的进程状态。即，在发生故障的服务器上运行的应用程序只能通过正常的服务器上重新执行来完成应用程序的交接。

例如，交接数据库管理系统(DBMS)的实例时，将在启动实例时自动进行数据库的恢复(回滚/前滚等)。数据库恢复所需时间可以通过设置 DBMS 的 Check Point Interval 等进行一定的控制，此过程一般需要几分钟。

很多应用程序只要重新执行就可以重新开始运行业务，但是也有些应用程序在发生故障后需要业务恢复的步骤。对于这样的应用程序，集群软件将业务恢复步骤写在脚本中，在重启时将不启动应用程序，而是启动该脚本。在脚本中，记载了脚本的执行原因、执行服务器等信息。

1.4.3 失效切换总结

至此，大家都已经了解了以下集群软件的运行过程。

- ◆ 查出故障(心跳/资源监视)
- ◆ 网络分区症状解析(NP 解析)
- ◆ 集群资源切换
 - 数据的交接
 - IP 地址的交接
 - 应用程序的交接

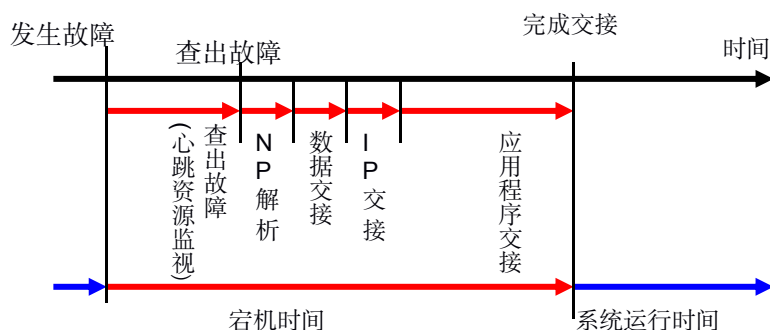


图 1-7 失效切换时间图

集群软件在实现失效切换时，会在短时间内逐一执行这些处理，从而实现了高可用性(High Availability)。

1.5 Single Point of Failure 的排除

在构建高可用性系统时，把握目标的可用性级别是很重要的。在设计系统时，必须考虑到对于可能阻碍系统运行的各种故障，包括我们应该采取的措施，如通过冗余结构保证系统持续运行，或者能够在短时间恢复到运行状态等，以及这些措施的性价比等。

在集群系统中，实现了服务器的多重化，能够排除系统的 SPOF，但是共享磁盘等在服务器之间共享的组建可能造成 SPOF。设计系统时将这些共享部分多重化或者排除是构建高可用性系统的关键点。

集群系统虽然提高了可用性，但是失效切换时，还是需要几分钟的系统切换时间。因此，失效切换时间也是造成可用性下降的原因之一。但是因为在高可用性系统中，ECC 内存或冗余电源技术对于提高单台服务器的可用性原本就是很重要的，在本文中我们暂且不谈这些提高单台服务器可用性的技术，在集群系统中，我们挖掘可能造成 SPOF 的以下 3 点原因，看看对此能够采取什么对策。

- ◆ 共享磁盘
- ◆ 共享磁盘的访问路径
- ◆ LAN

1.5.1 共享磁盘

一般共享磁盘通过磁盘阵列组建 RAID，因此磁盘的成对驱动器是不会造成 SPOF 的。但是由于 RAID 控制器内置，控制器可能会发生问题。很多集群系统中所使用的共享磁盘可以实现控制器二重化。

为了发挥二重化 RAID 控制器优势，一般需要进行共享磁盘访问路径的二重化。如果是二重化的多个控制器能够同时访问同一逻辑磁盘组(LUN)的共享磁盘，可以在每个

控制器上分别连接一台服务器，发生控制器故障时可以通过节点间的失效切换实现高可用性。

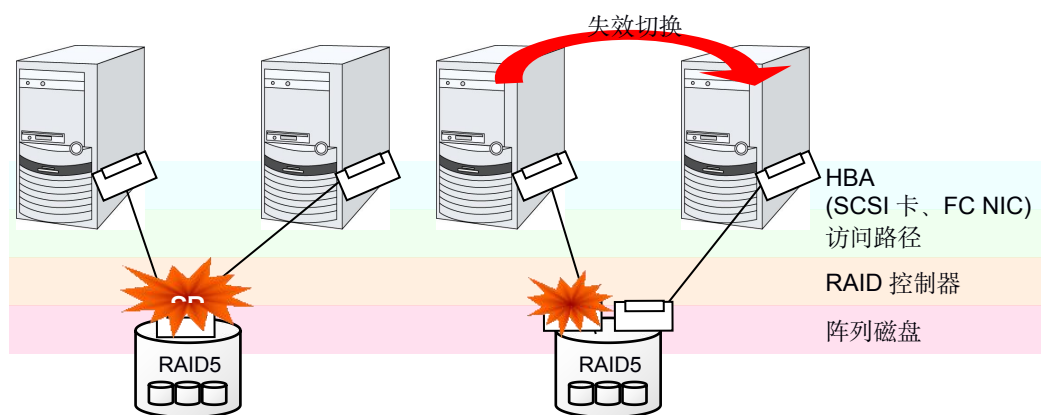


图 1-8 共享磁盘的 RAID 控制器和访问路径发生

SPOF 时的示例(左)和 RAID 控制器和访问路径分离的示例

另外，在不使用共享磁盘的镜像磁盘型失效切换集群中，所有的数据都与其他服务器的磁盘镜像化，能够实现没有 SPOF 的理想系统配置。但是，虽然以下几点不算作缺点，但也需要考虑。

- ◆ 通过网络进行数据镜像化对磁盘 I/O 性能(特别是 write 性能)的影响
- ◆ 服务器故障后恢复时，镜像重新同步过程中对系统性能的影响(镜像复制在后台执行)
- ◆ 镜像重新同步的时间(在镜像重新同步完成之前不会嵌入到集群中)

在数据引用多、数据容量不大的系统中，使用镜像磁盘型的失效切换集群也可以提高可用性。

1.5.2 共享磁盘的访问路径

在普通的共享磁盘型集群结构中，共享磁盘的访问路径通过配置集群的各服务器共享。以 SCSI 为例，就是在一条 SCSI 路径上有 2 台服务器与共享磁盘连接。因此，共享磁盘访问路径的故障可能是造成系统整体中断的原因。

作为其对策，可以考虑准备多条共享磁盘的访问路径，配置冗余结构，而从应用程序来看，共享磁盘的访问路径仍然只有 1 条。实现这一技术的设备驱动程序称为路径失效切换驱动程序(路径失效切换驱动程序多由共享磁盘供应商开发发布，Linux 版的路径失效切换驱动程序好像还未开发完毕，尚未发布。现阶段，如前所述，可以通过给每个共享磁盘的阵列控制器连接一个服务器，分割共享磁盘的访问路径，通过此方法可以确保 Linux 集群的可用性)。

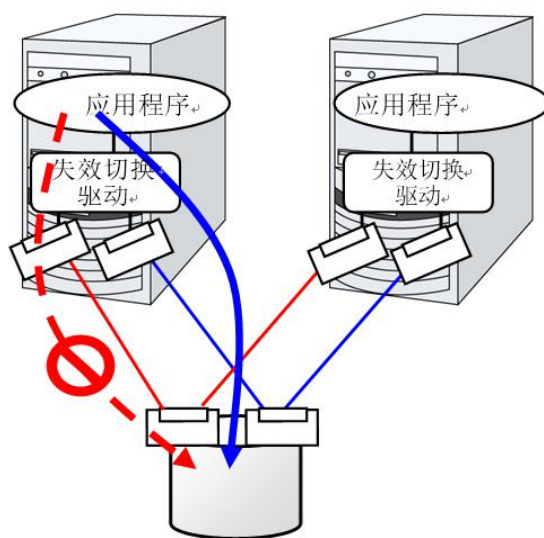


图 1-9 路径失效切换驱动程序

1.5.3 LAN

不仅是集群系统，所有需要在网络上执行某种服务的系统，LAN 故障都是阻碍系统运行的重要原因。在集群系统中，如果配置恰当，可以在 NIC 发生故障时在节点之间进行失效切换，从而提高可用性，但是集群系统外的网络设备如果发生故障仍然会阻碍系统的运行。

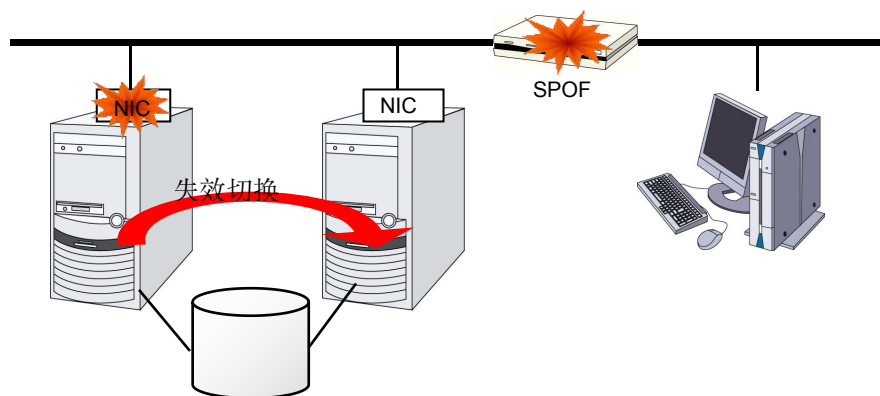


图 1-10 路由器发生 SPOF 的示例

在这种情况下，可以通过 LAN 的冗余化提高系统的可用性。在集群系统中，也可以直接使用单体服务器上的技术提高 LAN 的可用性。如，一种原始方法是准备好备用的网络设备，先不打开电源，发生故障时手动切换，或者冗余配置高性能的网络设备，通过将网络路径多重化实现路径的自动切换等方法。另外，也可以考虑使用类似于英特尔公司 ANS 驱动程序的支持 NIC 冗余结构驱动程序。

负载均衡设备(Load Balance Appliance)和防火墙服务器(Firewall Appliance)都是容易发生 SPOF 的网络设备。这些可以通过标配或可选的软件，使其能够搭建失效切换结构。同时因为这些设备大多在系统整体中的位置非常重要，所以一般必须配置冗余结构。

1.6 提高可用性的操作

1.6.1 操作前测试

系统问题大多起因于配置错误或操作维护。从这一点来考虑，在实现高可用性系统时，操作前的测试和完善故障恢复手册对系统的稳定运行非常重要。作为测试观点，结合实际操作，进行以下操作可以提高可用性。

- ◆ 筛选故障发生位置，讨论对策，进行模拟故障测试验证
- ◆ 进行假定集群生存周期的测试，验证降级运行时的性能
- ◆ 以这些测试为基础，完善系统操作、故障恢复手册

简化集群系统的设计，能够简化上述的验证操作和手册，提高系统的可用性。

1.6.2故障监视

虽然我们已经做了上述的努力，可能还是会发生故障。系统长期持续运行后，必然会发生故障，其原因可能是硬件老化、软件的内存泄漏、或者操作时超过系统当初设计的承受能力等。因此，在提高硬件、软件可用性的同时，需要进一步监视故障，在发生故障时采取恰当的处理，这一点非常重要。例如，万一服务器发生故障，可以通过搭建集群系统，只需要几分钟的切换时间就可以使系统继续运行，但是如果置之不管，系统失去冗余性，发生下一个故障时集群系统就没有任何意义了。

因此，发生故障时，系统管理员必须要采取措施防范下一故障的发生，排除新发生的 SPOF。在支持系统管理业务上，远程维护、故障通知等功能非常重要。毋庸置疑，Linux 在远程维护方面非常优秀，故障通告的机制也在逐渐完备。

以上介绍了使用集群系统实现高可用性时所需的周边技术以及其他的一些要点。简单总结一下，就是要注意以下几点：

- ◆ 排除或掌握 Single Point of Failure
- ◆ 设计简洁的抵抗故障能力强的系统，在操作前测试的基础上完善操作故障恢复手册
- ◆ 尽早查出发生的故障并进行恰当的处理

第 2 章 普华i-HA 高可用集群软件V4.0 的使用方法

2.1 普华 i-HA 高可用集群软件 V4.0

现在大家已经了解了集群，下面开始介绍普华 i-HA 高可用集群软件 V4.0。所谓普华 i-HA 高可用集群软件 V4.0 就是通过冗余化(集群化)的系统结构，运行服务器发生故障时，自动用待机服务器交接业务的软件，该软件实现了系统可用性和扩展性的飞越性提高。

2.2 普华 i-HA 高可用集群软件 V4.0 的产品结构

普华 i-HA 高可用集群软件 V4.0 大致由 3 个模块组成。

- ◆ 普华 i-HA 高可用集群软件 V4.0 Server
- ◆ 普华 i-HA 高可用集群软件 V4.0 主体，包含所有服务器的高可用性功能。还包含 WebManager 的服务器一端的功能。
- ◆ 普华 i-HA 高可用集群软件 V4.0 WebManager

普华 i-HA 高可用集群软件 V4.0 操作的管理工具。使用 Web 浏览器作为用户接口。实体嵌入在普华 i-HA 高可用集群软件 V4.0 Server 中，通过管理终端上的 Web 浏览器进行操作，据此与普华 i-HA 高可用集群软件 V4.0 Server 主体区分。

- ◆ 普华 i-HA 高可用集群软件 V4.0 Builder(Builder)

普华 i-HA 高可用集群软件 V4.0 配置信息的创建工具。与 WebManager 相同，使用 Web 浏览器作为用户接口。在使用 Builder 的终端上，Builder 需要与普华 i-HA 高可用集群软件 V4.0 Server 分开安装。

- ◆ 分为在使用 Builder 的终端上与普华 i-HA 高可用集群软件 V4.0 Server 分开安装使用的脱机版，以及从 WebManager 界面的工具栏点击设置模式图标或点击 [显

示]菜单的[设置模式]来进行转换的在线版。一般情况下无需安装，只有在脱机状态下使用时才需要另行安装。

2.3 普华 i-HA 高可用集群软件 V4.0 的软件配置

普华 i-HA 高可用集群软件 V4.0 的软件配置如下图所示。在 Linux 服务器上安装“普华 i-HA 高可用集群软件 V4.0 Server(普华 i-HA 高可用集群软件 V4.0 主体)”。WebManager 及 Builder 的主体功能包括在普华 i-HA 高可用集群软件 V4.0 Server 内，因此无需另行安装。但是，在无法访问普华 i-HA 高可用集群软件 V4.0 Server 的环境下使用 Builder 时，需要在 PC 上安装脱机版 Builder。除了通过管理 PC 上的 Web 浏览器以外，WebManager 及 Builder 还能通过构成集群的各服务器上的 Web 浏览器进行操作。

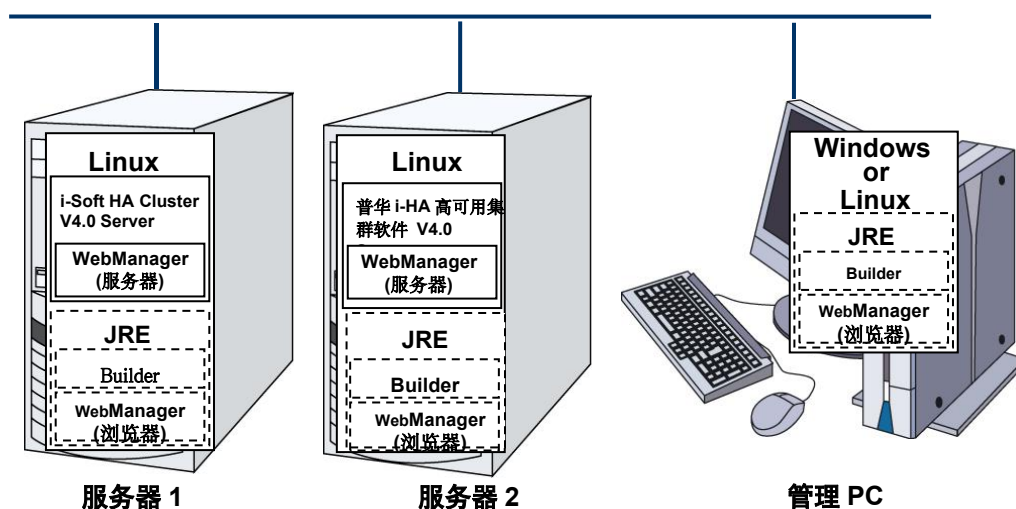


图 2-1 普华 i-HA 高可用集群软件 V4.0 的软件配置

2.3.1 普华 i-HA 高可用集群软件 V4.0 的故障监视原理

通过普华 i-HA 高可用集群软件 V4.0 可以进行服务器监视、业务监视、内部监视等 3 种监视，从而能够准确查出故障。下面详细介绍这三种监视。

2.3.2 服务器监视

服务器监视是失效切换型集群系统中最基本的监视功能，用于监视集群内服务器是否中断。

普华 i-HA 高可用集群软件 V4.0 是为了服务器监视，定期在各服务器之间进行生存确认。该生存确认称为心跳确认。心跳确认通过以下通信线路进行。

◆ 私网

在失效切换型集群专用的通信线路上，使用普通的 Ethernet NIC。除了确认心跳还可以用于服务器之间的信息交换。

◆ 公网

作为备用心跳线，使用与客户端进行通信的通信线路。只要是能够使用 TCP/IP 的 NIC 即可。除了确认心跳，

还可以用于服务器之间的信息交换。

◆ 共享磁盘

在连接到失效切换型集群的所有服务器的磁盘中，创建普华 i-HA 高可用集群软件 V4.0 专用分区(Cluster 分区)，在 Cluster 分区上进行心跳确认。

◆ COM 端口

通过 COM 端口在失效切换型集群的各个服务器之间进行心跳通信，确认其他服务器是否生存。

◆ BMC

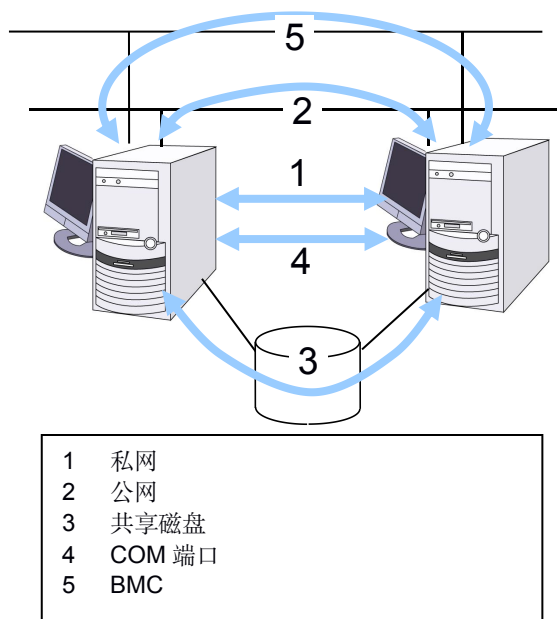


图 2-2 服务器

通过 BMC 在失效切换型集群的各个服务器之间进行心跳通信，确认其他服务器是否生存。

使用这些通信路径能够大幅度提高服务器间通信的可靠性，防止网络分区症状的发生。

注：网络分区解析资源(Split-brain-syndrome)：集群服务器之间所有的通信线路均发生故障，造成网络中断的状态。在不能对应网络分区解析资源的集群系统中，不能区分通信线路故障和服务器故障，多个服务器同时访问同一资源，就可能造成数据损坏。

2.3.3业务监视

业务监视用于监视业务应用程序自身或者监视造成无法执行业务的故障原因。

◆ 应用程序的生存状态监视

能够使用启动资源(称为 EXEC 资源)启动应用程序，通过监视资源(称为 PID 监视资源)定期监视进程的生存。在业务应用程序异常退出造成业务中断时有效。

注：

- 如果普华 i-HA 高可用集群软件 V4.0 直接启动的应用程序为启动、结束监视对象的常驻进程的应用程序，则无法查出常驻进程的异常。
 - 无法查出应用程序内部状态的异常(应用程序的停止、结果异常)。
-

◆ 资源的监视

通过普华 i-HA 高可用集群软件 V4.0 的监视资源能够监视集群资源(磁盘分区、IP 地址等)和公网的状态。在资源异常造成业务中断时有效。

2.3.4内部监视

内部监视是普华 i-HA 高可用集群软件 V4.0 内部模块之间的相互监视，用于监视普华 i-HA 高可用集群软件 V4.0 的各监视功能是否正常运行。

2.3.5可监视的故障和无法监视的故障

普华 i-HA 高可用集群软件 V4.0 中有可监视的故障和无法监视的故障。在构建集群系统时，需要先了解哪些故障能够监视，哪些不能监视。

2.3.6通过服务器监视可以查出的故障和无法查出的故障

监视条件：故障服务器的心跳中断

- ◆ 可监视的故障示例
 - 硬件故障(OS 不能继续运行)
 - panic
- ◆ 无法监视的故障示例

OS 局部功能故障(仅鼠标、键盘等出现故障等)

2.3.7通过业务监视可以查出的故障和无法查出的故障

监视条件：故障应用程序的消失、持续的资源异常、与某网络设备通信的路径中断

- ◆ 可监视故障示例
 - 应用程序的异常退出
 - 共享磁盘访问故障(HBA1 的故障)
 - 公网 NIC 故障
- ◆ 无法监视故障示例

- 应用程序的停止/结果异常

普华 i-HA 高可用集群软件 V4.0 虽无法直接监视应用程序的停止/结果异常，但是可以监视应用程序，在查出异常时能够创建退出自身的程序，通过 EXEC 资源启动该程序，利用 PID 监视资源进行监视，从而使其发生失效切换。

2.3.8 网络分区解析

所谓普华 i-HA 高可用集群软件 V4.0，是指从某一服务器查出心跳中断，并对其原因进行判断：是服务器本身的故障，还是由于网络分区解析资源引起的故障。如果判断为服务器故障，就会执行失效切换(在正常的服务器上激活各种资源，启动业务应用程序)，如果判断为网络分区解析资源引起的故障，保护数据比继续业务更重要，就会执行紧急关机处理。

网络分区解析方式有以下方法。

- ◆ ping 方式

2.4 失效切换的原理

普华 i-HA 高可用集群软件 V4.0 查出故障时，在开始失效切换之前会判断查出的故障是服务器的故障还是网络分区解析资源。之后会在正常的服务器上激活各种资源，启动业务应用程序，执行失效切换。

此时，同时移动的资源集合称为失效切换组。从使用者角度来看，可以把失效切换组看作虚拟的计算机。

注：在集群系统中，通过在正常的节点上重启应用程序来执行失效切换。因此，在应用程序的内存上保存的执行状态不能进行失效切换。

从发生故障到失效切换结束需要几分钟时间。下面是时间图。

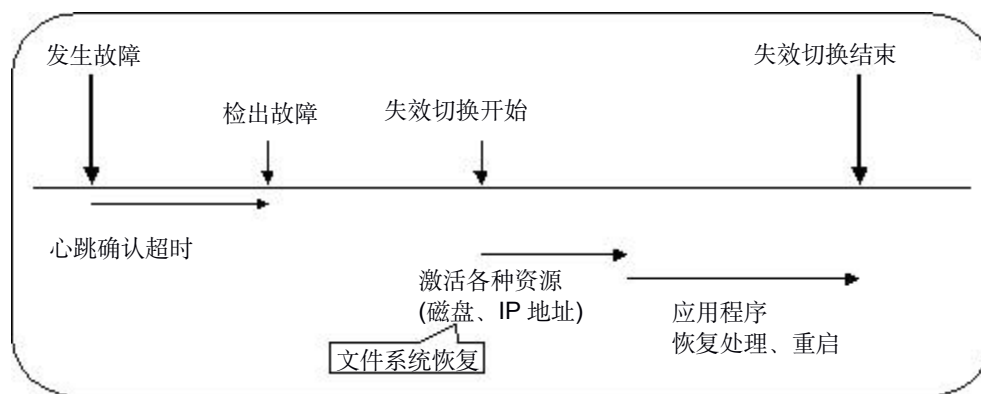


图 2-3 失效切换的时间图

◆ 心跳确认超时

- 正在执行业务的服务器发生故障后，到待机服务器查出该故障之间的时间。
- 可以根据业务的负载调整集群属性的设置值。
(出厂设置是 90 秒。)

◆ 激活各种资源

- 激活业务所需资源的时间。
- 一般配置情况下，激活需要几秒，失效切换组中登录的资源种类数量不同，所需时间会相应变化。

◆ 开始脚本执行时间

- 数据库的回滚/前滚等数据恢复时间和业务中使用的应用程序的启动时间。
- 通过调整 Check Point Interval 时间，可以预测回滚/前滚的时间。详细内容请参考各软件产品的文档。

2.4.1 失效切换资源

普华 i-HA 高可用集群软件 V4.0 能够作为失效切换对象的主要资源如下。

◆ 切换分区(磁盘资源等)

- 保存业务应用程序应该交接的数据的磁盘分区。
- ◆ 浮动 IP 地址(浮动 IP 资源)
- 使用浮动 IP 地址连接业务，这样客户端可以不必考虑失效切换带来的业务执行位置(服务器)的变化。
 - 浮动 IP 地址通过向公网适配器动态分配 IP 地址和发送 ARP 包来实现。大多数网络设备都能通过浮动 IP 地址连接。
- ◆ 脚本(EXEC 资源)
- 在普华 i-HA 高可用集群软件 V4.0 中，从脚本启动业务应用程序。
 - 通过共享磁盘交接的文件虽然作为文件系统是正常的，但是作为数据可能是不完整的。在脚本中，除了启动应用程序，还记载了失效切换时业务特有的恢复处理。

注：在集群系统中，通过在正常的节点上重启应用程序来执行失效切换。因此，在应用程序的内存上保存的执行状态不能进行失效切换。

2.4.2 失效切换型集群的系统配置

失效切换型集群在集群服务器间共享磁盘阵列设备。服务器发生故障时，待机服务器使用共享磁盘上的数据交接业务。

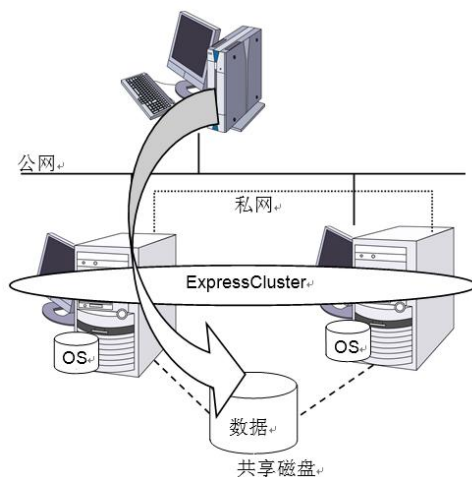


图 2-4 系统配置

失效切换型集群根据运行形态分为以下几类。

2.4.2.1 单向待机集群

其中一个服务器作为运行服务器运行业务，另外一个服务器作为待机服务器不运行业务的运行形态。这是最简单的一种运行形态，使用该形态构建的系统失效切换后性能不会降低，可用性高。

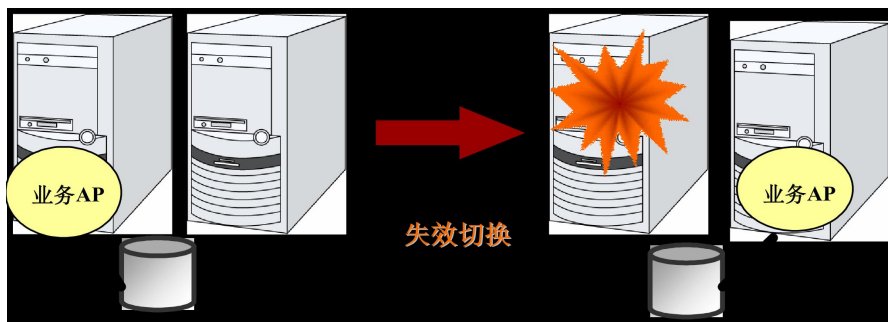


图 2-5 单向待机集群

2.4.2.2 同一应用程序双向待机集群

在多个服务器上同时运行某业务应用程序，互为待机的运行形态。应用程序必须支持双向待机运行。将某业务数据分割成数份时，根据要访问的数据来更改客户端所要连接的目标服务器，可以构建以数据分割为单位的负载均衡系统。

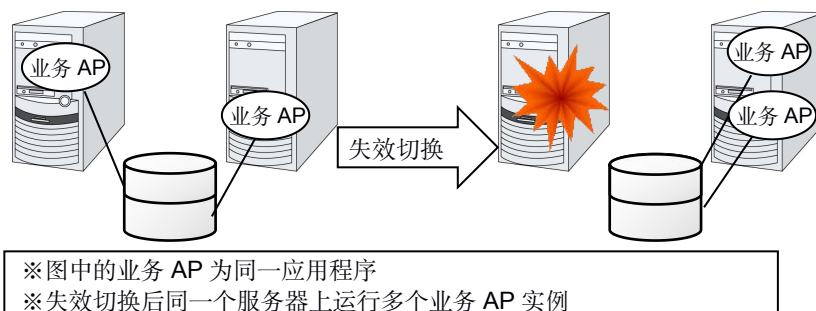


图 2-6 同一应用程序双向待机集群

2.4.2.3 不同应用程序双向待机集群

多种业务应用程序分别在不同服务器上运行，互为待机的运行形态。应用程序没有必要一定支持双向待机运行。能够构建以业务为单位的负载均衡系统。

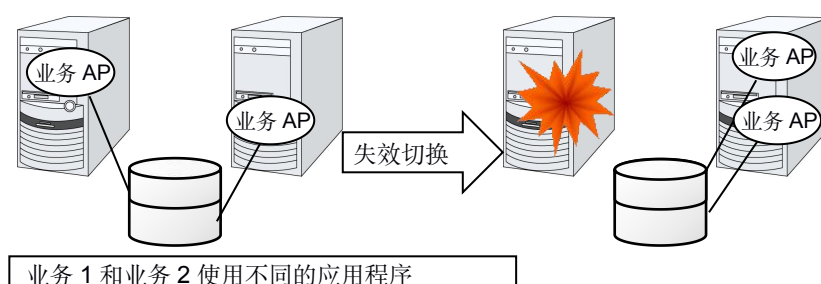


图 2-7 不同应用程序双向待机集群

2.4.2.4 N + N 结构

使用前面介绍的结构，可以将系统扩展为使用更多节点的结构。在下图所示的结构中，在 3 台服务器上执行 3 种业务，一旦发生问题，将把业务交接给 1 台待机服务器。在单向待机中，正常时的资源浪费是 1/2，在此结构中正常时的资源浪费降低到 1/4，而且如果 1 台发生故障，不会造成系统性能下降。

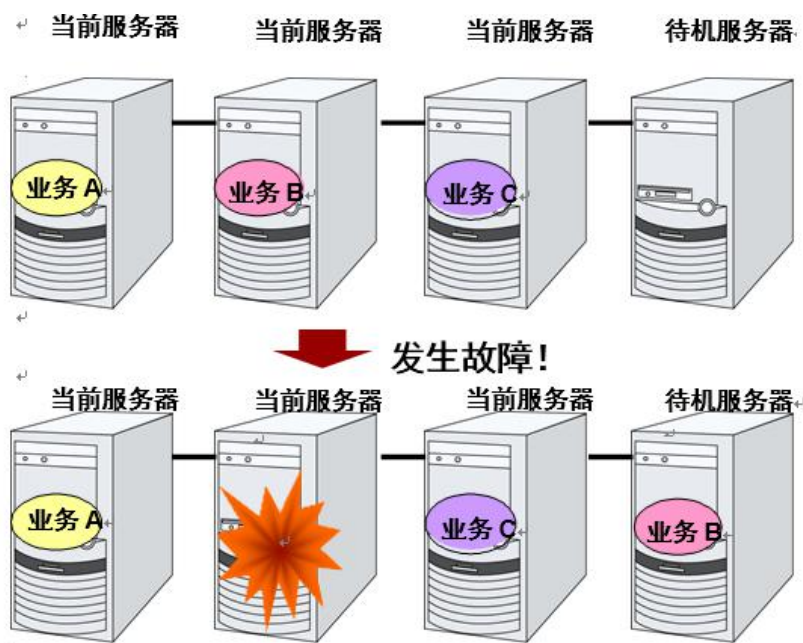


图 2-8 N + N 结构

2.4.3 共享磁盘型的硬件配置

共享磁盘型的普华 i-HA 高可用集群软件 V4.0 的 HW 配置如下图所示。

服务器之间的通信一般使用以下配置：

- ◆ 2 块 NIC(1 块用于与外部通信，1 块为普华 i-HA 高可用集群软件 V4.0 专用)
- ◆ RS232C cross cable 连接的 COM 端口
- ◆ 共享磁盘的指定区域

与共享磁盘连接的接口可以是 SCSI 或 FibreChannel，但是使用 FibreChannel 进行连接比较普遍。

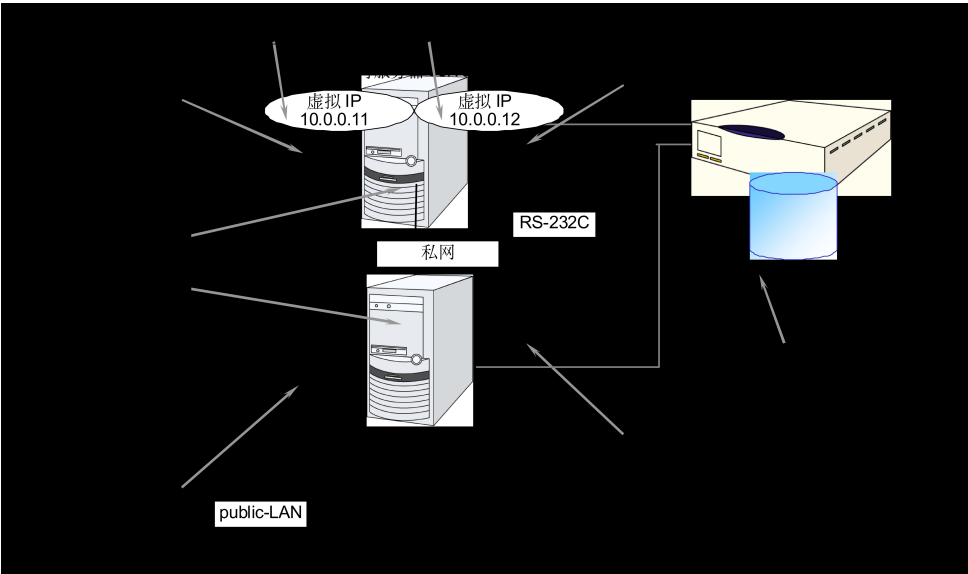


图 2-9 使用共享磁盘时的集群环境示例

2.4.4 镜像磁盘型的硬件配置

镜像磁盘型的普华 i-HA 高可用集群软件 V4.0 配置如下图所示。

与共享磁盘配置相比，需要配备镜像磁盘数据复制所使用的网络，但是一般是使用普华 i-HA 高可用集群软件 V4.0 的内部通信 NIC 来代替。

另外，镜像磁盘不依赖于连接接口(IDE or SCSI)。

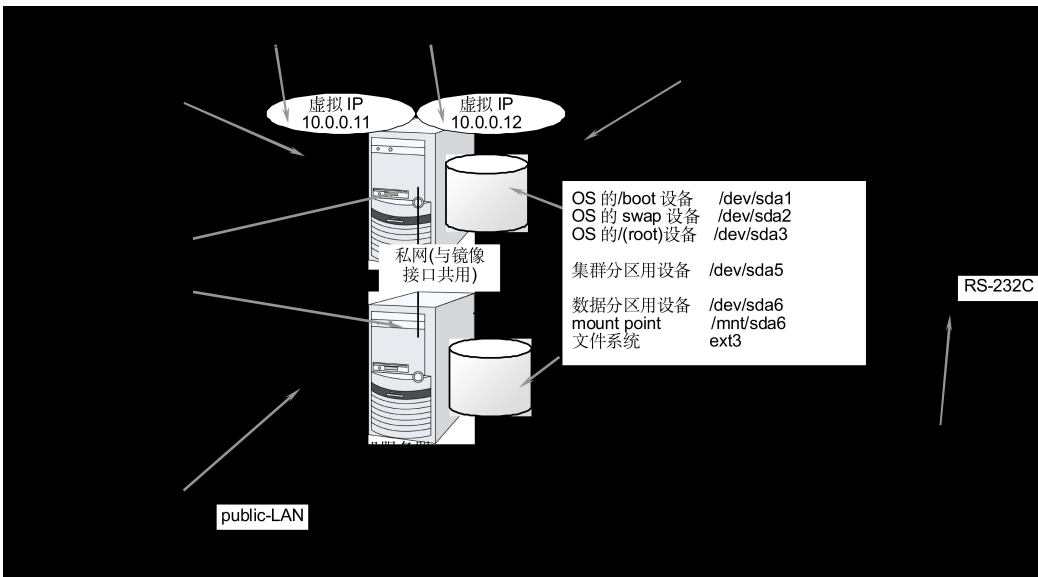


图 2-10 使用镜像磁盘时的集群环境的示例

(安装 OS 的磁盘中能够确保集群分区、数据分区时)

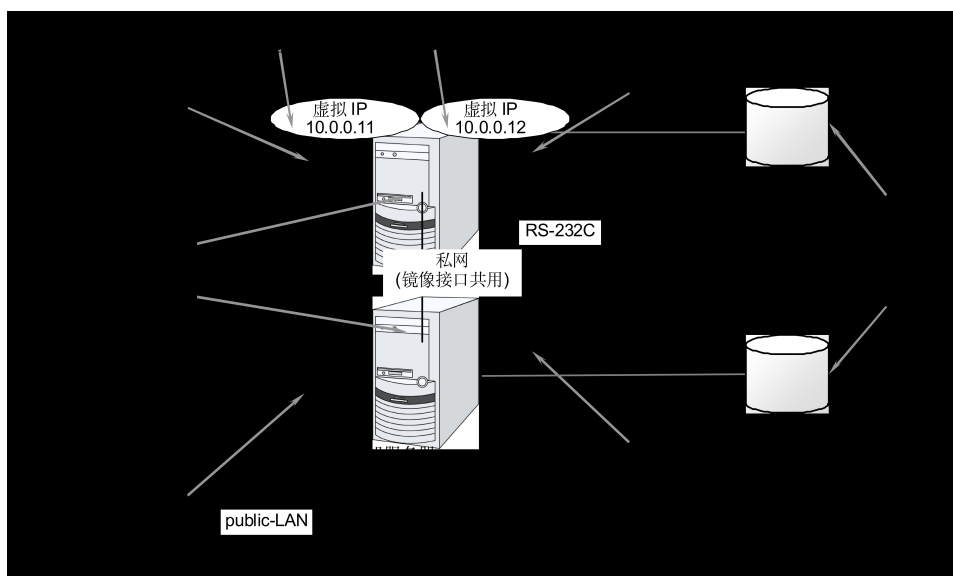


图 2-11 使用镜像磁盘时的集群环境的示例(准备集群分区、数据分区用磁盘时)

2.4.5 共享型镜像磁盘型的硬件配置

与共享磁盘配置相比，共享型镜像磁盘型需要配备复制数据所使用的网络，但是一般是使用普华 i-HA 高可用集群软件 V4.0 的内部通信 NIC 代替。

另外，磁盘不依赖于连接接口(IDE or SCSI)。

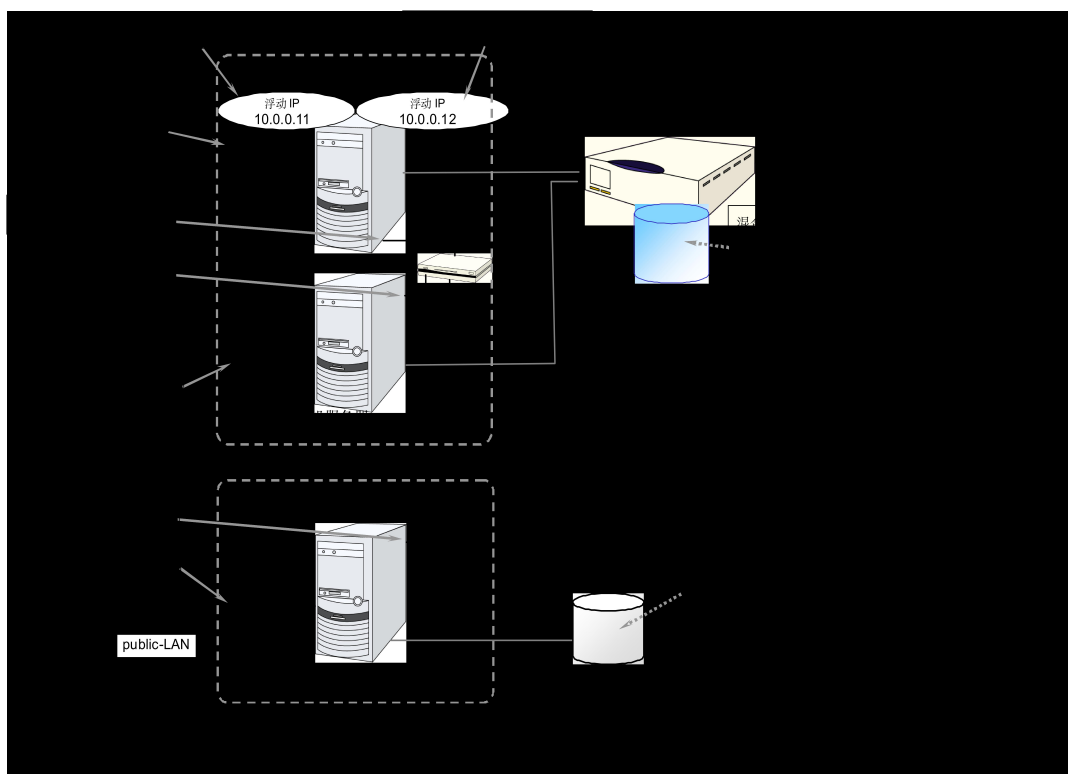


图 2-12 使用共享型镜像磁盘时的集群环境的示例

(在 2 台服务器上使用共享磁盘，向第 3 台服务器的普通磁盘作镜像链接时)

2.4.6 集群对象

在普华 i-HA 高可用集群软件 V4.0 中，使用以下结构管理各种资源。

- ◆ 集群对象配置集群的单位。
- ◆ 服务器对象表示实体服务器的对象，属于集群对象。
- ◆ 服务器组对象捆绑服务器的对象，属于集群对象。
- ◆ 心跳资源对象表示实体服务器的 NW 部分的对象，属于服务器对象。
- ◆ 网络分区解析资源对象表示网络分区解析机构的对象，属于服务器对象。
- ◆ 组对象表示虚拟服务器的对象，属于集群对象。
- ◆ 组资源对象表示拥有虚拟服务器的资源 (NW、磁盘) 的对象，属于组对象。
- ◆ 监视资源对象表示监视机构的对象，属于集群对象。

2.5 资源

在普华 i-HA 高可用集群软件 V4.0 中，监视端和被监视端的对象都称为资源，分类进行管理。这样不仅能够明确区分监视/被监视的对象，还能够使构建集群或故障处理更简便。资源分为心跳资源、网络分区解析资源、组资源和监视资源 4 类。以下简要介绍各类资源。

2.5.1 心跳资源

在服务器之间互相确认生存状态所使用的资源。

现在支持的心跳资源如下所示：

- ◆ LAN 心跳资源，使用 Ethernet 的通信。
- ◆ 内核模式 LAN 心跳资源，使用 Ethernet 的通信。
- ◆ COM 心跳资源，使用 RS232C(COM)的通信。
- ◆ 磁盘心跳资源，使用共享磁盘上的特定分区(磁盘心跳分区)的通信。仅限共享磁盘配置时使用。
- ◆ BMC 心跳资源，经由 BMC 使用 Ethernet 的通信。仅限在对应 BMC 的硬件和固件时使用。

2.5.2 网络分区解析资源

为了处理网络分区解析问题的资源。

PING 网络分区解析资源，使用 PING 方式的网络分区解析资源。

2.5.3 组资源

组成失效切换的单位——失效切换组的资源。

现在支持的组资源如下所示：

◆ 浮动 IP 资源 (fip)

提供虚拟 IP 地址，客户端可以像普通 IP 地址一样访问。

◆ EXEC 资源 (exec)

提供启动/停止业务(DB、httpd、etc..)的机制。

◆ 磁盘资源 (disk)

提供共享磁盘上的指定分区，仅限(共享磁盘)配置时使用。

◆ 镜像磁盘资源 (md)

提供镜像磁盘上的指定分区，仅限(镜像磁盘)配置时使用。

◆ 共享型镜像磁盘资源 (hd)

提供共享磁盘或磁盘上的指定分区，仅限(共享型镜像磁盘)配置时使用。

◆ 卷管理资源 (volmgr)

将多个存储器及磁盘用作一个逻辑磁盘。

◆ NAS 资源 (nas)

连接 NAS 服务器上的共享资源。(集群服务器并不是作为 NAS 的服务器端运行的资源。)

◆ 虚拟 IP 资源 (vip)

提供虚拟 IP 地址。可以像访问客户端的普通 IP 地址一样访问虚拟 IP 地址。用于配置网络地址在不同区间的远程集群。

◆ 虚拟机资源 (vm)

进行虚拟机的启动、停止及迁移。

- ◆ 动态域名解析资源(ddns)

将虚拟主机名及激活服务器的 IP 地址登录在 Dynamic DNS 服务器上。

2.5.4 监视资源

监视资源是集群系统内进行监视操作的主体资源。

现在支持的监视资源如下所示：

- ◆ 浮动 IP 监视资源 (fipw)

提供浮动 IP 资源中启动的 IP 地址的监视机构。

- ◆ IP 监视资源 (ipw)

提供外部 IP 地址的监视机构。

- ◆ 磁盘监视资源 (diskw)

提供磁盘的监视机构，也可以用于共享磁盘的监视。

- ◆ 镜像磁盘监视资源 (mdw)

提供镜像磁盘的监视机构。

- ◆ 镜像磁盘接口监视资源 (mdnw)

提供镜像磁盘接口的监视机构。

- ◆ 共享型镜像磁盘监视资源 (hdw)

提供共享型镜像磁盘的监视机构。

◆ 共享型镜像磁盘接口监视资源 (hdnw)

提供共享型镜像磁盘接口的监视机构。

◆ PID 监视资源 (pidw)

提供 EXEC 资源启动的进程的生存状态监视功能。

◆ 用户空间监视资源 (userw)

提供用户空间的停止监视机构。

◆ NIC Link Up/Down 监视资源 (miiw)

提供 LAN 线缆的链接状态的监视机构。

◆ 卷管理监视资源 (volmgrw)

提供多个存储器及磁盘的监视机构。

◆ Multi-Target 监视资源 (mtw)

提供捆绑多个监视资源的状态。

◆ 虚拟 IP 监视资源 (vipw)

提供送出虚拟 IP 资源 RIP 包的机构。

◆ ARP 监视资源 (arpw)

提供送出浮动 IP 或虚拟 IP 资源 ARP 包的机构。

◆ 自定义监视资源 (genw)

提供有进行监视处理的命令或脚本时，根据其动作结果监视系统的机构。

◆ 虚拟机监视资源 (vmw)

进行虚拟机的生死确认。

◆ 消息接收监视资源 (mrw)

用于实现“设置接收异常发生通知时执行的异常时动作”及“异常发生通知的 WebManager 显示”的监视资源。

◆ 动态域名解析监视资源 (ddnsw)

定期将虚拟主机名及激活服务器的 IP 地址登录在 Dynamic DNS 服务器上。

◆ 进程名监视资源 (psw)

通过指定进程名，提供任意的进程状态监视功能。

◆ BMC 监视资源 (bmcw)

提供配备的 BMC 的状态监视功能。

◆ DB2 监视资源 (db2w)

提供 IBM DB2 数据库的监视机构。

◆ ftp 监视资源 (ftpw)

提供 FTP 服务器的监视机构。

◆ http 监视资源 (httpw)

提供 HTTP 服务器的监视机构。

◆ imap4 监视资源 (imap4w)

提供 IMAP4 服务器的监视机构。

◆ MySQL 监视资源 (mysqlw)

提供 MySQL 数据库的监视机构。

◆ nfs 监视资源 (nfsw)

提供 nfs 文件服务器的监视机构。

◆ Oracle 监视资源 (oraclew)

提供 Oracle 数据库的监视机构。

◆ OracleAS 监视资源 (oracleasw)

提供 Oracle 应用程序服务器的监视机构。

◆ Oracle Clusterware 同步管理监视资源 (osmw)

提供 Oracle Clusterware 联动进程的监视和成员信息同步功能。

◆ pop3 监视资源 (pop3w)

提供 POP3 服务器的监视机构。

◆ PostgreSQL 监视资源 (psqlw)

提供 PostgreSQL 数据库的监视机构。

◆ samba 监视资源 (sambaw)

提供 samba 文件服务器的监视机构。

◆ smtp 监视资源 (smtpw)

提供 SMTP 服务器的监视机构。

◆ Sybase 监视资源 (sybasew)

提供 Sybase 数据库的监视机构。

◆ Tuxedo 监视资源 (tuxw)

提供 Tuxedo 应用程序服务器的监视机构。

◆ Websphere 监视资源 (wasw)

提供 Websphere 应用程序服务器的监视机构。

◆ Weblogic 监视资源 (wls w)

提供 Weblogic 应用程序服务器的监视机构。

◆ WebOTX 监视资源 (otxw)

提供 WebOTX 应用程序服务器的监视机构。

◆ JVM 监视资源 (jraw)

提供 Java VM 的监视机构。

◆ 系统监视资源 (sraw)

提供个别进程资源、系统整体资源的监视机构。

2.6 开始使用普华 i-HA 高可用集群软件 V4.0 吧!

至此普华 i-HA 高可用集群软件 V4.0 的简单介绍就结束了。

下面我们将根据以下流程，一边阅读对应的指南，一边构建使用普华 i-HA 高可用集群软件 V4.0 的集群系统。

2.6.1 集群系统的构建

请参考《普华高可用集群软件 V4.0-用户手册》。