

ExpressCluster[®] X 3.0 *for Windows*

开始指南

2010.10.01
第1版



改版履历

版号	改版日期	内 容
1	2010.10.01	第一版

© Copyright NEC Corporation 2010. All rights reserved.

免责声明

本手册内容如有更改，恕不提前通知。

日本电气株式会社对本手册技术上或是编辑上的错误、遗漏等概不负责。

此外，用户若为获得某些预期效果而根据本手册购买产品并使用，属于用户的责任，并由用户对其效果负责。

本手册所述内容的著作权属于日本电气株式会社。未经日本电气株式会社允许，禁止对本手册内容进行复制、更改以及翻译。

商标信息

ExpressCluster® X 为日本电气株式会社的注册商标。

Intel、Pentium、Xeon为Intel Corporation的注册商标或商标。

Microsoft、Windows为美国Microsoft Corporation在美国及其他国家的注册商标。

Java为Sun Microsystems, Inc.在美国及其他国家的商标或注册商标。

本手册中涉及到的其它产品名及标语为各公司的商标或注册商标。

目录

前言	vii
读者对象和用途	vii
本手册的构成	vii
ExpressCluster X 手册体系	viii
本手册的标记规则	ix
最新信息来源	x
第 I 篇 ExpressCluster 的概要	11
第 1 章 何谓集群系统	13
集群系统的概要	14
HA (High Availability) 集群	14
共享磁盘型	15
镜像磁盘型	17
系统构成	18
查出故障的原理	21
共享磁盘的互斥控制	21
网络分区症状 (Split-brain-syndrome)	22
集群资源的交接	22
数据的交接	22
IP 地址的交接	23
应用程序的交接	23
失效切换总结	24
Single Point of Failure 的排除	24
共享磁盘	25
共享磁盘的访问路径	26
LAN	26
支持可用性的操作	27
操作前测试	27
故障的监视	27
第 2 章 关于 ExpressCluster	29
何谓 ExpressCluster	30
ExpressCluster 的产品结构	30
ExpressCluster 的软件配置	31
ExpressCluster 的故障监视原理	31
何谓服务器监视	32
何谓业务监视	32
何谓内部监视	33
可监视的故障和无法监视的故障	33
通过服务器监视可以查出的故障和无法查出的故障	33
通过业务监视可以查出的故障和无法查出的故障	33
网络分区解析	34
失效切换的原理	34
由 ExpressCluster 构建的共享磁盘型集群的硬件配置	36
用 ExpressCluster 构建的镜像磁盘型集群的硬件配置	37
用 ExpressCluster 构建的共享磁盘型集群的硬件配置	39
何谓集群对象	41
何谓资源	41
心跳资源	41

网络分区解析资源	42
组资源	42
监视资源	43
开始使用ExpressCluster吧!	46
最新信息的确认	46
集群系统的设计	46
集群系统的构建	46
开始操作集群系统后的故障对策	46
第 II 篇 发布备忘录（ExpressCluster最新信息）	47
第 3 章 ExpressCluster的运行环境.....	49
硬件运行环境	50
所需规格	50
ExpressCluster Server的运行环境.....	50
支持OS	50
所需内存容量和磁盘容量	51
监视可选软件已经确认完毕的应用程序信息	52
虚拟机资源的运行环境.....	54
Builder的运行环境.....	55
运行确认完毕的OS、浏览器.....	55
Java执行环境	55
所需内存容量/磁盘容量.....	55
支持的ExpressCluster版本	55
WebManager的运行环境	56
运行确认完毕的OS、浏览器.....	56
Java执行环境	56
所需内存容量/磁盘容量.....	56
统合WebManager的运行环境.....	57
运行确认完毕的OS、浏览器.....	57
Java执行环境	57
所需内存容量/磁盘容量.....	57
第 4 章 最新版本信息.....	59
ExpressCluster和手册的对应一览表	60
最新版本.....	60
功能强化信息	61
第 5 章 注意限制事项.....	62
研究系统配置时	63
关于Builder、WebManager的运行OS	63
关于镜像磁盘/共享型镜像磁盘的需求	63
关于共享磁盘的需求	64
NIC Link Up/Down监视资源.....	64
关于镜像磁盘/共享型镜像磁盘的write性能.....	64
关于非同步镜像的历史文件	64
关于多个非同步镜像间的数据一致性	65
关于Multiboot.....	65
安装ExpressCluster前	65
关于文件系统	65
通信端口号	66
时钟同步的设置	67
关于共享磁盘	67
关于镜像磁盘用的分区.....	67
关于共享型镜像磁盘用的分区.....	68

OS启动时间的调整	68
网络的确认	68
关于与ESMPRO/AutomaticRunningController结合使用	69
关于ipmiutil	70
关于在Server Core的安装	70
创建ExpressCluster的配置信息时	71
组资源非激活异常时的最终运行	71
延迟警告比例	71
关于磁盘监视资源和共享型镜像磁盘TUR监视资源的监视方法TUR	71
关于WebManager的画面更新间隔	71
关于心跳资源的设置	72
关于能用于脚本注释等的2字节系字符编码	72
操作ExpressCluster后	72
恢复运行中的操作限制	72
关于命令篇中没有记载的可执行文件或脚本文件	73
集群关机/集群关机重启	73
特定服务器的关机、重启	73
从网络分区状态的恢复	73
关于WebManager	74
关于Builder	74
关于ExpressCluster Disk Agent服务	75
关于镜像构建中的集群配置信息的变更	75
关于镜像磁盘的待机服务器的集群恢复	75
关于镜像磁盘、共享型镜像磁盘间的配置变更	75
关于chkdsk命令和磁盘碎片整理	75
关于索引服务	76
关于Windows Server 2008环境中的应用程序资源/脚本资源的画面显示	76
关于Windows Server 2008环境中启动程序的运行	76
与旧版的兼容性	77
与ExpressCluster X 1.0/2.0/2.1的兼容性	77
关于与ExpressCluster Ver 8.0以下版本的兼容功能	77
关于兼容API	78
关于脚本文件	78
附录	79
附录 A 词汇表	81
附录 B 索引	83

前言

读者对象和用途

《ExpressCluster® X 开始指南》分为第I篇和第II篇两部分。第I篇以初次使用ExpressCluster的用户为对象，就ExpressCluster的产品概要和基本使用方法进行说明。

第II篇以导入ExpressCluster前的用户以及导入系统后进行升级的用户为对象，就最新的运行环境及限制事项等进行说明。

本手册的构成

第 I 篇 ExpressCluster 的概要

第 1 章 “何谓集群系统”：就集群系统的概要进行说明。

第 2 章 “关于ExpressCluster”：就ExpressCluster的使用方法以及相关信息进行说明。

第 II 篇 发布备忘录

第 3 章 “ExpressCluster的运行环境”：就导入前需要确认的最新信息进行说明。

第 4 章 “最新版本信息”：显示有关ExpressCluster 最新版本的信息。

第 5 章 “注意限制事项”：就已知的问题和限制事项进行说明。

附录

附录 A “词汇表”

附录 B “索引”

ExpressCluster X 手册体系

ExpressCluster X 的手册分为以下 4 类。各指南的标题和用途如下所示。

《ExpressCluster X 开始指南》 (Getting Started Guide)

本手册的读者对象为使用ExpressCluster的用户，对产品概要、运行环境、升级信息以及现有的问题等进行了说明。

《ExpressCluster X 安装&设置指南》 (Installation and Configuration Guide)

本手册的读者对象为导入使用ExpressCluster构筑集群系统的系统工程师以及导入集群系统后进行维护和操作的系统管理员，对使用ExpressCluster导入集群系统后到开始操作前的必备事项进行说明。手册按照实际导入集群系统时的顺序，对使用ExpressCluster的集群系统的设计方法、ExpressCluster的安装设置步骤、设置后的确认以及开始操作前的测试方法进行了说明。

《ExpressCluster X 参考指南》 (Reference Guide)

本手册的读者对象为管理员以及导入使用ExpressCluster的集群系统的系统工程师。对ExpressCluster的操作步骤，各模块的功能、维护的相关信息以及疑难解答信息等进行了说明，是对《安装&设置指南》的补充。

《ExpressCluster X 统合WebManager 管理员指南》 (Integrated WebManager Administrator's Guide)

本手册的读者对象为通过ExpressCluster统合WebManager管理使用ExpressCluster的集群系统的系统管理者，以及安装统合WebManager的系统工程师。本手册就导入安装统合WebManager的集群系统的操作步骤，进行了具体说明。

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[启动]。 [详细信息]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]
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	c:\Program files\EXPRESSCLUSTER
等宽字体 (courier) 粗体	表示用户在命令提示符后实际输入的值。	输入以下内容。 clpcl -s -a
等宽字体 (courier) <i>斜体</i>	用户将其替换为有效值后输入的项目。	clpstat -s [-h host_name]

最新信息来源

关于最新的产品信息，请参考以下Web站点。

<http://www.nec.com/global/prod/expresscluster/>

第 I 篇 ExpressCluster 的概要

本篇就 ExpressCluster 的产品概要和运行环境进行说明。

- 第 1 章 何谓集群系统
- 第 2 章 关于 ExpressCluster

第 1 章 何谓集群系统

本章就集群系统的概要进行说明。

本章说明的项目如下所示。

- 集群系统的概要 14
- HA (High Availability)集群 14
- 系统构成 18
- 查出故障的原理 21
- 集群资源的交接 22
- Single Point of Failure的排除 24
- 支持可用性的操作 27

集群系统的概要

现在的计算机社会中，持续的提供不停止的服务已经成为通往成功的关键。例如仅由于 1 台机器故障或超负荷而宕机就导致对客户的服务全面停止。这样的话，不但会带来莫大的损失，还会失去客户的信任。

随着集群系统的导入，发生意外事故时会将系统停止时间(宕机时间)降低到最小限度、使负载均衡，提高其可用性。

所谓集群，有“集团”、“团”的意思，顾名思义是“将多个计算机汇集成一群(或者多群)，谋求提升可靠性及处理性能的系统”。集群系统有多个种类，可分为下列 3 种。其中，ExpressCluster 属于 High Availability 集群。

◆ HA (High Availability) 集群

是平时作为运行服务器作业，在运行服务器发生故障时将业务交接到待机服务器的集群。是以高可用性为目的的集群。包括共享磁盘型、镜像磁盘型。

◆ 负载均衡集群

是将客户端的请求遵从恰当的负荷均衡原则分配给各节点的集群。是以高扩展性为目的的集群、一般无法进行数据交接。包括 load balance 集群、并列数据库集群。

◆ HPC (High Performance Computing) 集群

是指计算量非常大的集群。是为使用超级计算机执行单一业务的集群。使用所有节点的 CPU 来执行单一业务的网格计算技术近年来已成为热点。

HA (High Availability) 集群

一般提升系统可用性时，会认为对构成系统的组件进行冗余化，消除 Single Point of Failure 是重要的。所谓 Single Point of Failure 指的是计算机的构成要素(硬件的组件)因为只有一个，所以该处发生故障时会有使业务停止的弱点。所谓 HA 集群是使用多台节点进行冗余化操作、将系统停止时间控制在最小限度，提升业务可用性(availability)的集群系统。

某些骨干业务系统等的宕机会给业务带来很大影响，因此此类系统不允许停止，需要导入 HA 集群。

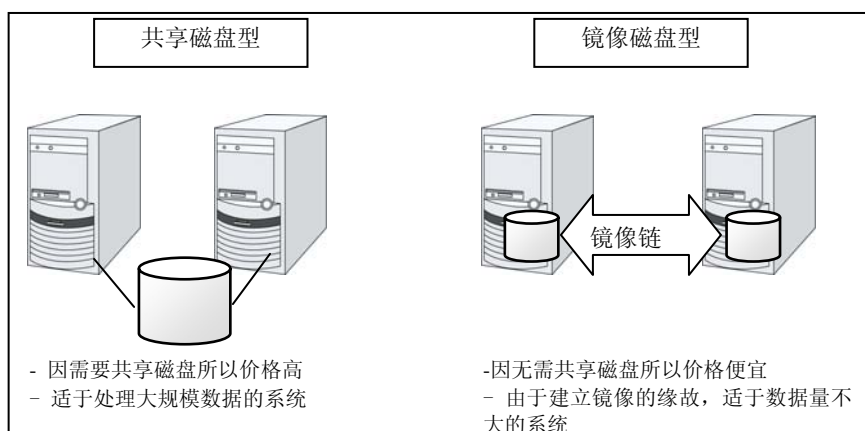


图 1-1 HA 集群构成图

HA 集群可分为共享磁盘型和镜像磁盘型。以下开始逐一进行说明。

共享磁盘型

集群系统中服务器间必须要交接数据。将这些数据从多个服务器放到像以 SAN 连接的 FibreChannel 磁盘阵列装置这样可以访问的外置磁盘（共享磁盘）上，通过该磁盘在服务器间交接数据即称为共享磁盘型集群系统。

运行业务应用程序的服务器(运行服务器)发生故障时，集群系统查出故障并在交接业务的服务器(待机服务器)中自动启动业务应用程序，交接业务。这称为失效切换。集群系统交接的业务由磁盘、IP 地址、应用程序等资源构成。

在没有集群化的系统中使用其它服务器重启应用程序时，客户端必须再次连接不同的 IP 地址。但是多数的集群系统里不是以业务为单位给服务器分配 IP，而是分配其他网络的 IP 地址(虚拟 IP 地址)。因此客户端没必要去区分正在执行业务的是运行服务器还是待机服务器，如同连接同一个服务器一样，可以持续地运行业务。

由于运行服务器宕机发生失效切换时，共享磁盘上的数据没有进行妥当的结束处理就交接给待机服务器。因此待机服务器中有必要对交接的数据进行逻辑检查。这与一般未集群化的系统宕机后重启时进行的处理是一样的。例如，如果是数据库就需要回滚及前滚的处理。由此客户端仅运行未确认的 SQL 就可以继续业务。

故障发生后，被查出故障的服务器经过物理隔离修复后，只要连接集群系统就可以作为待机服务器恢复运行。重视业务连续性的实际操作中，也可以不进行组的故障恢复。如果必须要在原来的服务器上运行业务时，请移动组。

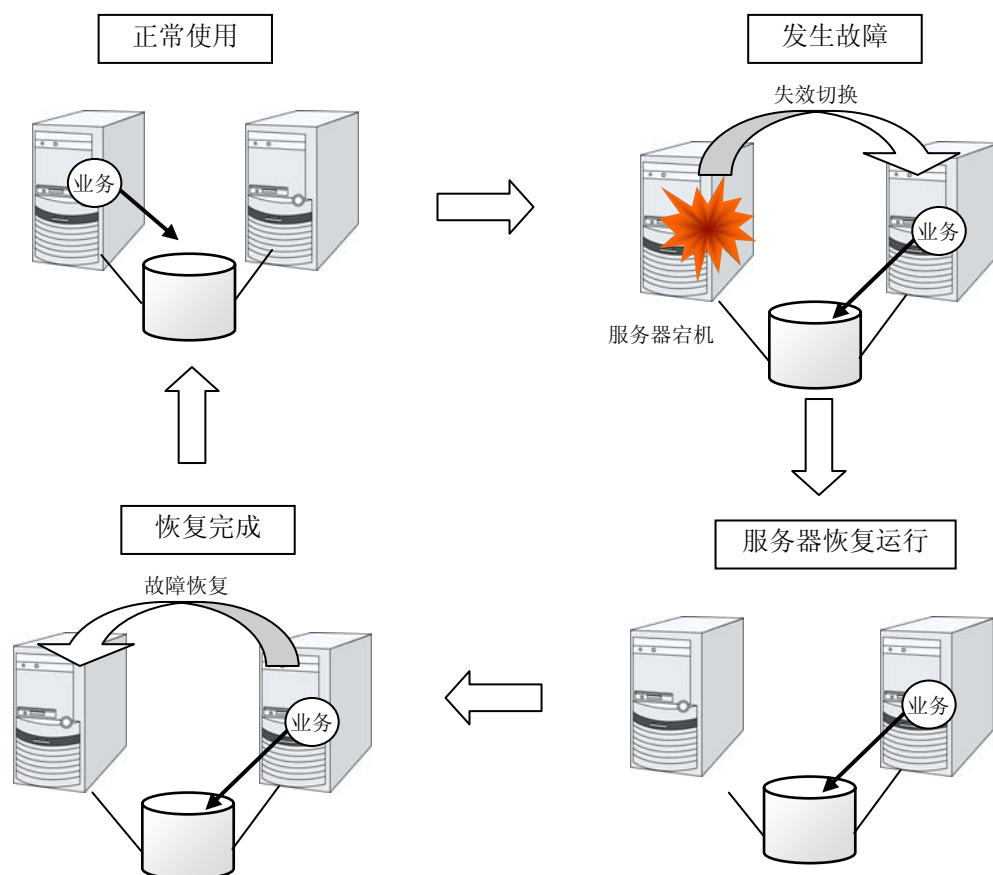


图 1-2 从发生故障到恢复的流程

由于失效切换处的服务器配置不够、双向待机引起超负荷等原因而希望在原来的服务器上运行业务时，原来节点的恢复作业完成后将暂时停止业务，在原来的节点上重启业务。将失效切换的组返回到原来的服务器称为故障恢复。

如图 1-3 所示，只有 1 个业务，待机服务器为不运行业务的待机形态称为单向待机。有 2 个以上业务，各自的节点既为运行服务器又为待机服务器的形态称为双向待机。

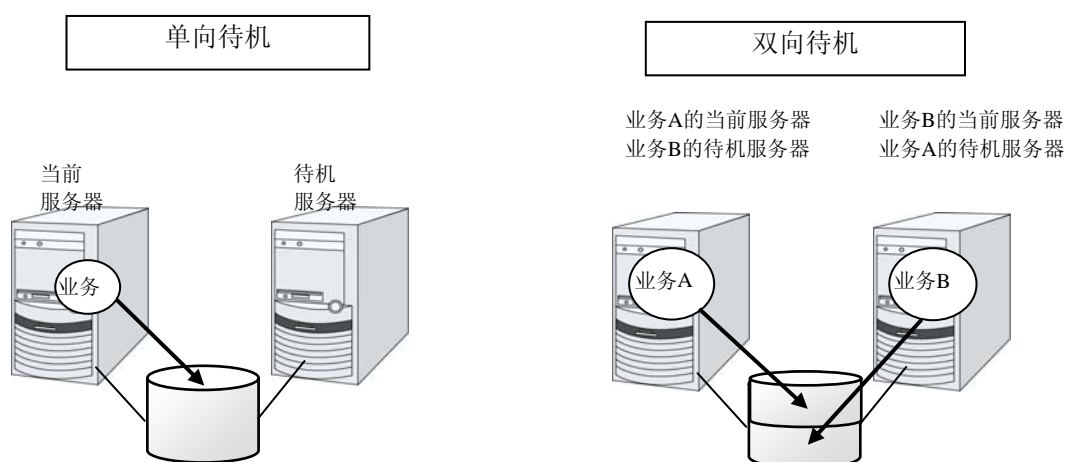


图 1-3 HA 集群的运行形态

镜像磁盘型

上述共享磁盘型适用于大规模系统，但共享磁盘大体价格较高因此构筑系统的成本也会增大。所以，不使用共享磁盘，而通过将各服务器磁盘在服务器间建立镜像，可以以更低的价格实现相同功能，这类集群系统称为镜像磁盘型集群系统。

但是，由于需要在服务器间为数据建立镜像，因此不适用于需要大量数据的大规模系统。

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** 请求分发给待机服务器。所谓心跳线是连接在服务器之间的电缆，在集群系统中用于服务器的死活监视。数据镜像型中除死活监视外还用于数据的传送。待机服务器的数据镜像引擎通过将接收的数据写入待机服务器的本地磁盘中，使运行服务器和待机服务器间的数据实现同步。

对于应用程序发出的 **Read** 请求，仅从运行服务器的磁盘中读取。

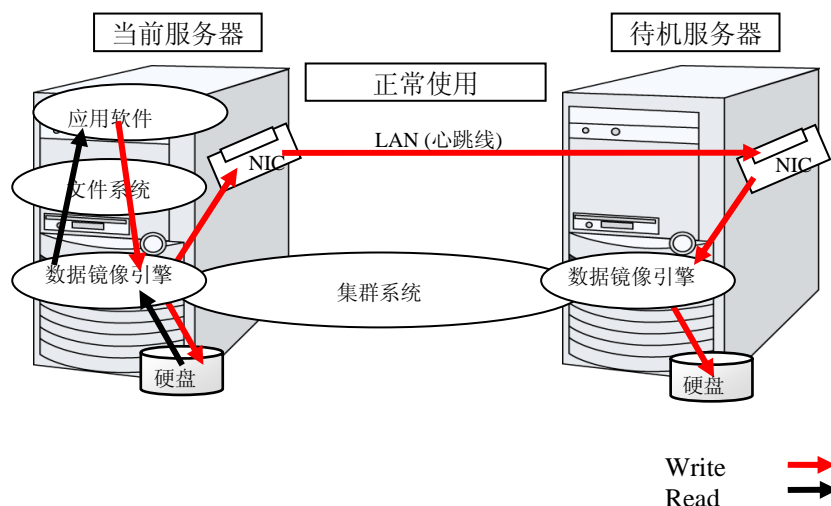


图 1-4 数据镜像的结构

快照备份就是使用数据镜像的例子。因数据镜像型的集群系统在 2 个地方持有共享数据，只需将待机服务器从集群分离，即可作为快照备份来保存数据。

HA 集群的机制和问题点

下面就集群的实现和问题点进行说明。

系统构成

共享磁盘型集群将磁盘阵列装置在集群服务器间实现共享。服务器发生故障时待机服务器使用共享磁盘上的数据实现数据的交接。

镜像磁盘型集群是经过网络将集群服务器上的数据磁盘镜像化的构成。服务器故障时使用待机服务器上的镜像数据交接业务。数据的镜像化以 I/O 为单位进行，因此从上层应用程序看是与共享磁盘相同的。

下图是共享磁盘型集群的构成图例。

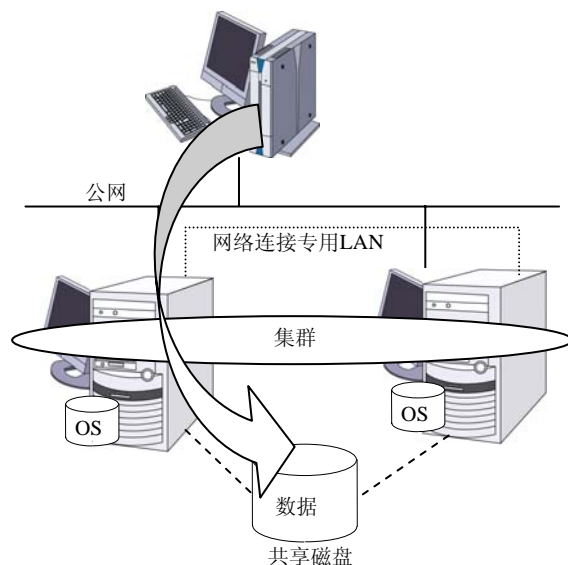


图 1-5 系统构成

根据运行形态可以将失效切换型集群分为下列几类。

单向待机集群

将其中一个服务器作为运行服务器作业、另一方的服务器作为待机服务器不进行作业的运行形态。可以以最简单的运行形态构建失效切换后没有性能下降的高可用性系统。

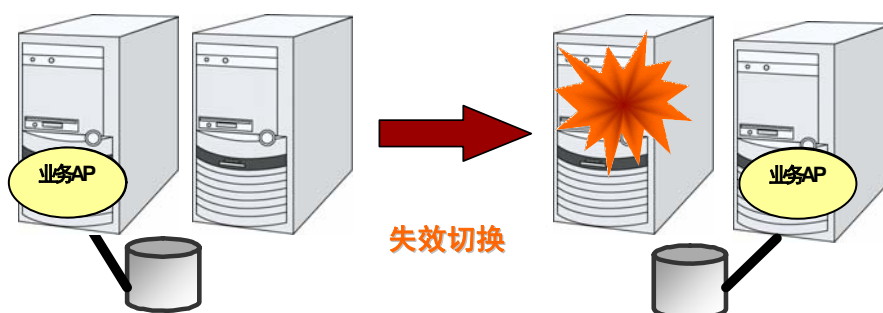


图 1-6 单向待机集群

同一应用程序双向待机集群

在多个服务器上运行相同的业务应用程序并互相待机的运行形态。各业务应用程序独立运行。失效切换时在一台服务器上运行多个同一业务应用程序，所以必须是能够实现这种运行的应用程序。可以将某个业务数据分割成多个时，按照想要访问的数据变更客户端连接的服务器，可以构筑以数据分割为单位的负载均衡系统。

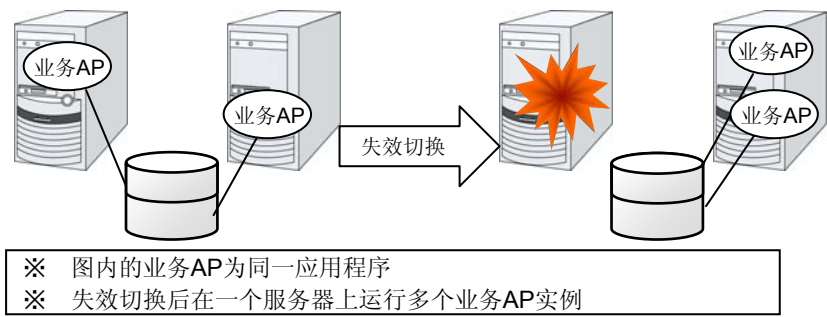


图 1-7 同一应用程序双向待机集群

不同应用程序双向待机集群

是使多种业务应用程序在不同的服务器上运行并互相待机的运行形态。失效切换时在一台服务器上运行多个业务应用程序，因此必须使这些应用程序可以共存。可以构筑以业务为单位的负荷均衡系统。

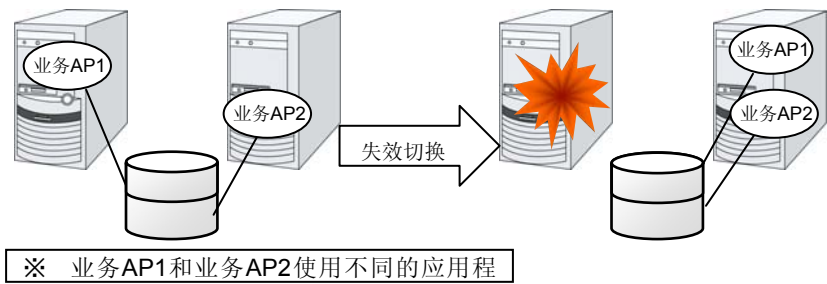


图 1-8 不同应用程序双向待机集群

N 服务器 + M 业务结构

应用以上的结构，可以扩展到使用更多的节点。下图是在 3 台服务器上运行 3 种不同业务，一旦发生问题时向 1 台待机服务器交接该业务的结构。在单向待机中正常时待机服务器不进行任何业务，因此闲置的资源比率为 1/2。如果采用此结构，闲置的资源比率为 1/4，可以降低成本。另外，另外如果有 1 台服务器发生异常，也不会降低性能。

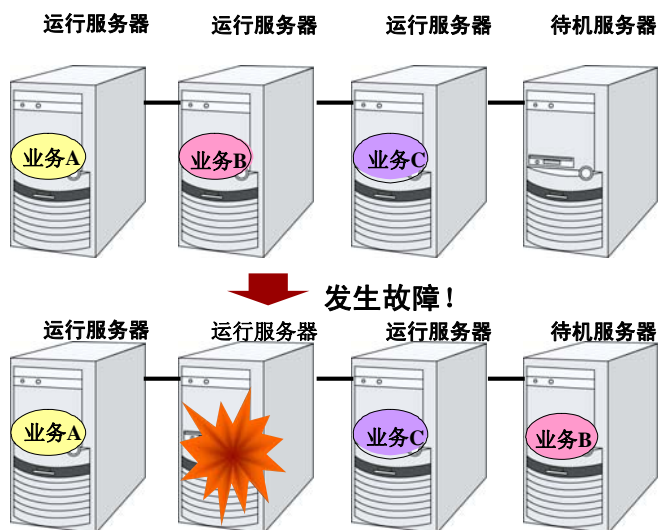


图 1-9 N + M 结构

查出故障的原理

集群软件一查出给持续业务带来问题的故障就会执行业务交接(失效切换)。在进入失效切换处理的具体内容之前，先简单地介绍一下集群软件是怎样查出故障的。

心跳和服务器故障的查出

集群系统中应该查出的最基本的故障是构成集群的服务器的宕机。服务器的故障中包含电源异常以及内存错误等硬件故障或者 OS 的崩溃等。为了查出此类故障，使用心跳来监视服务器的死活。

心跳可以仅通过确认 ping 的应答这种死活监视方式，根据集群软件不同，也可以传送本地服务器的状态信息等内容。集群软件收发心跳，在心跳无应答时视作该服务器故障并开始失效切换处理。但考虑到由于服务器的高负荷等原因会导致心跳的收发延迟，所以到判定为服务器故障需要一定程度的缓冲时间。因此实际发生故障的时间和集群软件查出故障的时间会存在时间差。

资源故障的查出

业务停止的主要原因不仅仅是由于构成集群的服务器的宕机。例如，由于业务应用程序使用的硬盘装置以及 NIC 的故障、或者业务应用程序本身的故障等也会导致业务停止。为了提升可用性，也必须查出这样的资源故障，实施失效切换。

作为查出资源异常的手段，所要监视的资源为物理设备时可以采用实际尝试访问的方法。应用程序的监视中，除了应用程序进程本身的死活监视以外，在对业务没有影响的范围内，也可以考虑尝试服务端口。

共享磁盘的互斥控制

共享磁盘型的失效切换集群中，多个服务器上共享磁盘装置。一般而言，文件系统通过保持服务器内数据的缓存，可以激发超出磁盘装置物理 I/O 性能界限的文件 I/O 性能。

如果从多个服务器上 mount 并访问某个文件系统的话会怎么样呢？

通常的文件系统不考虑在自身以外的服务器上更新磁盘上的数据。因此缓存和磁盘上的数据会产生矛盾，最终将导致数据遭到损坏。失效切换集群系统中，为了防止后面提到的网络分区症状所引起的多个服务器同时 mount 文件系统，会对磁盘装置采取互斥控制。

网络分区症状 (Split-brain-syndrome)

切断所有连接在服务器间的心跳线时，仅根据心跳所进行的死活监视将不能与服务器宕机进行区分。这种状态下会被当作是服务器宕机、实行失效切换处理，在多个服务器上同时 mount 文件系统，将损坏共享磁盘上的数据。

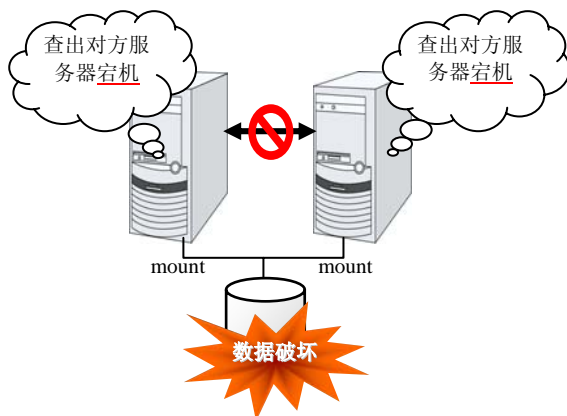


图 1-10 网络分区症状

将这样的问题称为“网络分区症状”或者 Split-brain-syndrome。为了解决这个问题，在失效切换集群中所有的心跳线都被切断时，可以考虑运用能确切实现共享磁盘装置互斥控制的各种对策。

集群资源的交接

集群管理的资源中有磁盘、IP 地址、应用程序等内容。下面说明用于交接这些集群资源的失效切换集群系统的功能。

数据的交接

在共享磁盘型集群中，共享磁盘装置上的分区保存服务器之间的交接数据。也就是说，所谓的交接数据，就是在正常的服务器上对应用程序使用的文件所在分区的文件系统重新 mount。由于共享磁盘装置与交接目标服务器是物理连接，因此，集群软件应该做就只有文件系统的 mount。

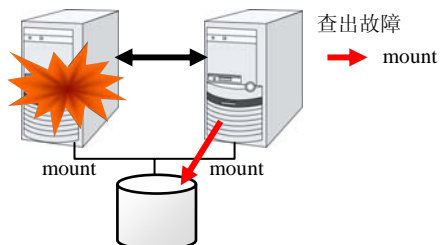


图 1-11 数据的交接

看似简单，但是在设计/构筑集群系统时有几点必须要注意。

一个是文件系统以及数据库的恢复时间问题。要交接的文件在故障发生的前一刻被其他的服务器所使用，或者也许正在被更新。因此，有些文件系统，有时需要交接时进行一致性的检查，如果是数据库就需要进行回滚等处理。这种情况与电源故障造成宕机重启单个服务器时是一样的。此类恢复处理需要较长时间时，若就此追加在失效切换时间(业务的交接时间)上，将成为系统可用性低下的主要原因。

还有一个就是写入保证的问题。应用程序向共享磁盘里写入数据时，通常是通过文件系统写入。应用程序即使写入完成，文件系统仍会保留在磁盘缓存上；所以未写入共享磁盘，运行服务器就宕机的情况下，磁盘缓存上的数据将不能交接给待机服务器。因此，发生故障时，需要切实交接给待机服务器的重要数据，必须要通过同步写入等方法，切实地写入到磁盘中。这与单个服务器宕机时数据不因断电而丢失一样。也就是说，在设计集群系统时要考虑到：交接给待机服务器的只有共享磁盘中记录的数据，象磁盘缓存这样的内存上的数据不予交接。

IP地址的交接

集群软件进行的下一个任务是 IP 地址的交接。失效切换时、通过 IP 地址的交接，所以不用在意业务在哪个服务器上运行。集群软件交接 IP 地址的目的就在于此。

应用程序的交接

集群软件业务交接的最后任务是交接应用程序。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应用程序运行中内存内容的进程状态等。也就是说，通过将故障服务器上运行的应用程序，在正常的服务器中重新执行，来完成应用程序的交接。

例如对 DB 实例进行失效切换时，并非以发生故障前一刻的状态重启，而是和先宕机再启动一样，进行事务的回滚，并且需要从客户端再次连接。该数据库恢复所需要的时间可以通过 DBMS 的 Check Point Interval 的设置某种程度上进行控制，但是一般都需要几分钟。

多数应用程序仅通过再次执行就可以再续业务，但是也有的应用程序需要在故障发生后进行业务恢复操作。为此，集群软件通过启动脚本替代应用程序以便能够描述业务恢复步骤。脚本中以描述脚本执行的主要原因以及执行服务器等信息为主，根据需要，也描述清除正在更新的文件等恢复步骤。

失效切换总结

从上述内容可以理解以下集群软件的运行。

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

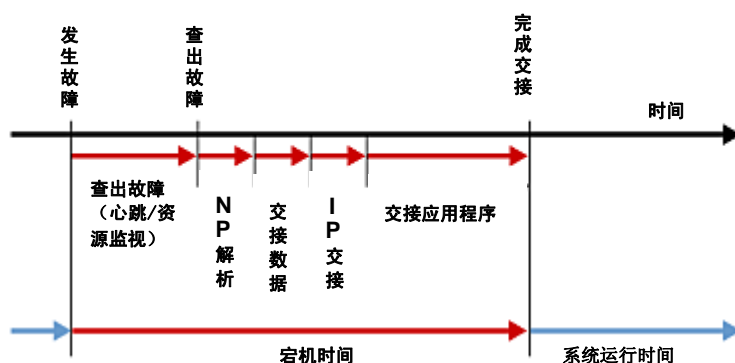


图 1-12 失效切换时间一览表

集群软件为了实现失效切换，就是这样将所有处理一一在短时间内予以实施，来实现高可用性（High Availability）的。

Single Point of Failure 的排除

在构筑高可用性系统时把握所要求的或者说目标的可用性水平是很重要的。也就是说对于可能阻碍系统运行的各种故障，应采用冗余结构，以使运行得以持续并在短时间内能够恢复，讨论或设计系统时就需要从上述策略的性价比方面进行考虑。

所谓 Single Point of Failure(SPOF)，前面讲过是指和系统停止相关联的部位。集群系统可以实现服务器的多重化，排除系统的 SPOF。但是共享磁盘等服务器间共享部分会成为 SPOF。设计多重化或者排除该共享部分的系统成为构筑高可用性系统的要点。

集群系统要提升可用性，但是失效切换中也需要几分钟的系统切换时间。因此也可以说失效切换时间是可用性低下的一个主要原因。为此在高可用性系统中，提高单个服务器可用性的 ECC 存储以及冗余电源等的技术是重要的。但是在此并不涉及单个服务器的可用性提升技术，而是深入研究一下集群系统中容易造成 SPOF 的下述 3 项内容，看看能有怎样的对策。

- ◆ 共享磁盘
- ◆ 通往共享磁盘的访问路径
- ◆ LAN

共享磁盘

通常共享磁盘按照磁盘阵列组合 RAID，所以磁盘的成对驱动器不会成为 SPOF。但是由于内置 RAID 控制器，所以控制器会有问题。多个集群系统中采用的共享磁盘里可以实现控制器的二重化。

为了发挥二重化 RAID 控制器的长处，通常有必要实现共享磁盘访问路径的二重化。但是如果是二重化的多个控制程序能够同时访问同一逻辑磁盘单元(LUN)的共享磁盘，每个控制器上连接 1 台服务器则在控制器发生异常时在节点间实施失效切换，就可以实现高可用性。

※HBA: 是 Host Bus Adapter 的略称，不是指共享磁盘端的而是指服务器本体端的适配器。

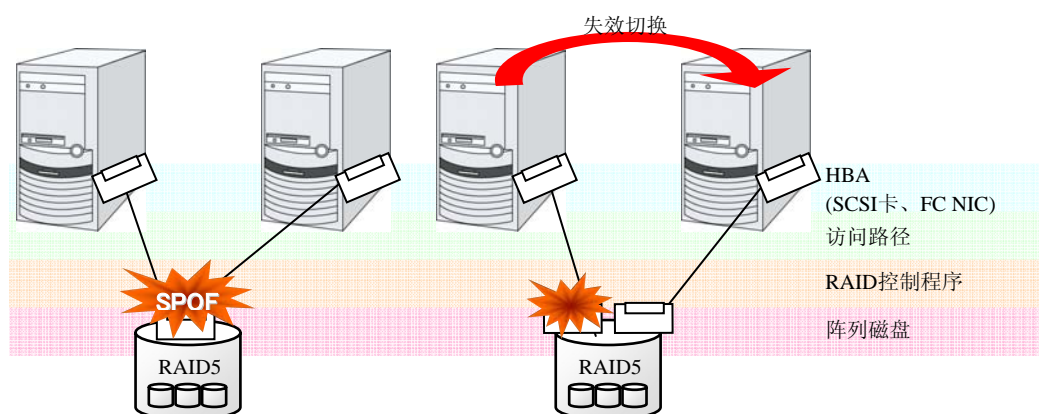


图 1-13 共享磁盘的 RAID 控制器和访问路径变成 SPOF 的例子(左)以及 RAID 控制器和访问路径分开的例子

一方面，在不使用共享磁盘的数据镜像型失效切换集群中，由于将所有数据都在其他服务器的磁盘上建立镜像，所以可以实现不存在 SPOF 的系统构成。但是需要考虑下列几点。

- ◆ 通过网络实现数据镜像化导致磁盘 I/O 性能(特别是 write 性能)低下
- ◆ 服务器故障恢复时、镜像在再次同步中的系统性能(镜像复制通过后台实施)低下
- ◆ 镜像的再次同步时间(镜像再次同步完成为止无法失效切换)

也就是说数据的链接多而数据容量不大的系统中，采用数据镜像型失效切换集群对于提升可用性是有效的。

共享磁盘的访问路径

共享磁盘型集群的一般结构中，共享磁盘的访问路径在构成集群的各服务器中是共享的。以 SCSI 为例，1 条 SCSI 路径上可以连接 2 台服务器和共享磁盘。因此，对共享磁盘的访问路径异常是整个系统瘫痪的重要原因。

解决办法：是准备多条访问共享磁盘的路径的冗余结构，使应用程序对共享磁盘的访问路径看起来像 1 条。实现该结构的设备驱动程序被称为路径失效切换驱动程序等。

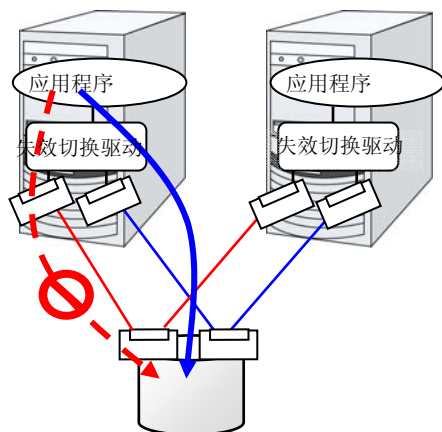


图 1-14 路径失效切换驱动

LAN

不仅限于集群系统，在网络上执行某些服务的系统中，LAN 的故障也是阻碍系统运行的重要因素。集群系统中，进行适当的设置，可在 NIC 故障时进行节点间失效切换，提高可用性；但是集群系统外的网络机器发生故障时，仍然会阻碍系统工作。

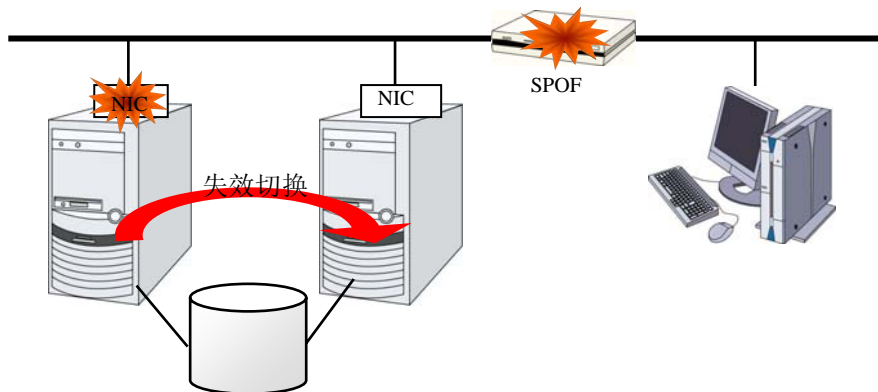


图 1-15 路由器为 SPOF 的例子

在这种情况下，通过 LAN 的冗余化，可以提高系统的可用性。在集群系统中，提高 LAN 的可用性时，可以同样利用单个服务器下的技术。比如，可以考虑通过不接通预备的网络机器的电源，在发生故障的情况下手动进行切换的原始方法，以及冗余配置高性能的网络机器，多重化网络路径来自动切换路径的方法。另外，可以考虑利用如英特尔公司的 ANS 之类，支持 NIC 的冗余配置的驱动程序。

LoadBalance 装置 (Load Balance Appliance) 和防火墙服务器 (Firewall Appliance) 也是容易导致 SPOF 的网络机器。这些通常使用标准或可选软件，来构建失效切换结构。同时，这些机器一般在系统整体中处于非常重要的位置，所以必须要考虑搭建冗余结构。

支持可用性的操作

操作前测试

我们常说产生系统故障的主要原因大多为设置的错误以及操作的维护。从这一点考虑，实现高可用性系统时，操作前的测试和恢复故障手册的完备对于系统的稳定运行是很重要的。作为测试观点，结合实际操作，执行下列工作成为提升可用性的要点。

- ◆ 找出故障发生位置，商讨对策，进行模拟故障测试，并实际验证。
- ◆ 进行假设的集群“一系列状态变化”的测试，对降级运行时的性能进行验证。
- ◆ 以这些测试为基础，完善系统操作/恢复故障手册。

设计简单的集群系统，可以简化上述的验证和手册，是提升系统可用性的要点。

故障的监视

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。

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
- ◆ 进行不容易出故障的简洁设计，基于操作前的测试，完善系统操作/恢复故障手册
- ◆ 及早查出发生的故障并进行恰当的处理

第 2 章 关于 ExpressCluster

本章介绍 ExpressCluster 各个组件的说明、从集群系统的设计到运行步骤之间的流程。

本章将介绍以下内容。

• 何谓ExpressCluster.....	30
• ExpressCluster的产品结构.....	30
• ExpressCluster的软件配置.....	31
• 网络分区解析.....	34
• 失效切换的原理.....	34
• 何谓资源.....	41
• 开始使用ExpressCluster吧!.....	46

何谓 ExpressCluster

对集群有了一定的理解后，让我们介绍一下 ExpressCluster。所谓 ExpressCluster 是用于实现 HA 集群系统的软件。

ExpressCluster 的产品结构

ExpressCluster 可以大致分为 3 个模块。

- ◆ ExpressCluster Server

是 ExpressCluster 的主体。安装在构成集群系统的各服务器上。在 Server 中包含了所有 ExpressCluster 的高可用性功能。另外，也包含 WebManager 以及 Builder 的服务器端的功能。

- ◆ ExpressCluster X WebManager (WebManager)

是进行 ExpressCluster 操作管理的管理工具。以 Web 浏览器作为用户界面。实体嵌入在 ExpressCluster Server 中，但操作是在管理终端的 Web 浏览器上进行，这点与 ExpressCluster Server 不同。

- ◆ ExpressCluster X Builder (Builder)

- ◆ 是创建 ExpressCluster 配置信息的工具。与 WebManager 相同，使用 Web 浏览器作为用户界面。分为在使用 Builder 的终端上与 ExpressCluster Server 分开安装使用的脱机版，以及从 WebManager 界面的工具栏左侧的下拉菜单选择设定模式或点击 [表示] 菜单的 [设定模式] 来进行转换的在线版。通常不需要安装，只有脱机使用时才需要另外进行安装。

ExpressCluster 的软件配置

ExpressCluster 的软件构成如下图所示。在构成集群的服务器上安装“ExpressCluster Server(ExpressCluster 主体)”。WebManager 以及 Builder 的本体功能包含在 ExpressCluster Server 内，因此无需另外安装。但是在不能访问 ExpressCluster Server 的环境中使用 Builder 时，需要在 PC 上安装脱机版 Builder。WebManager 以及 Builder 除了使用管理 PC 上的 Web 浏览器，也可以使用构成集群的各服务器上的 Web 浏览器。

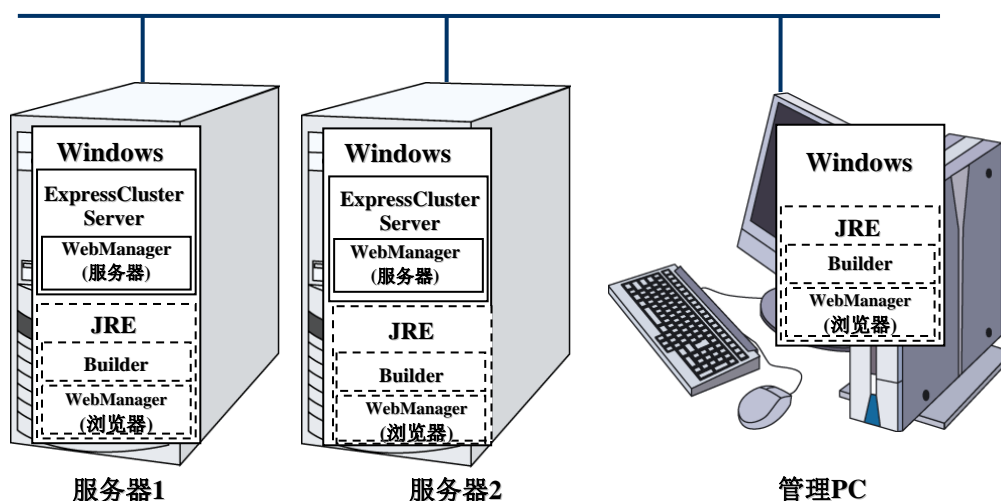


图 2-1 ExpressCluster 的软件构成

注：JRE 是指 Java Runtime Environment。

ExpressCluster 的故障监视原理

在 ExpressCluster 中通过执行服务器监视、业务监视、内部监视 3 个监视任务，可以迅速确切地查出故障。以下介绍这些监视的详细内容。

何谓服务器监视

所谓服务器监视是失效切换型集群系统最基本的监视功能，是监视构成集群的服务器是否停止运行的功能。

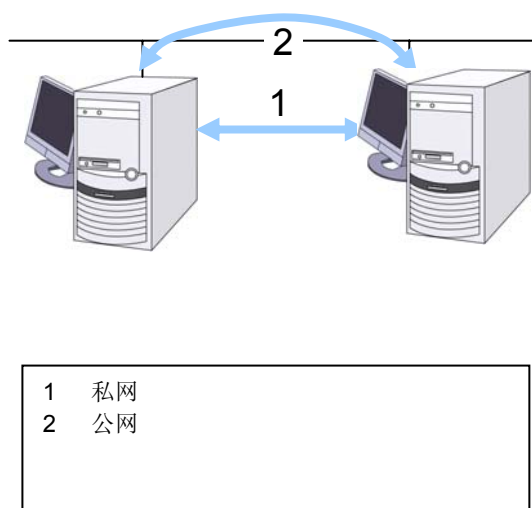
ExpressCluster 为监视服务器，定期地与伙伴服务器进行生存确认。将此生存确认称为心跳。心跳使用下列通信路径。

- ◆ 私网

是集群服务器间通信专用的 LAN。进行心跳的同时在服务器间交换信息。

- ◆ 公网

用于与客户端通信用路径。也用于服务器间的信息交换以及心跳线的备份。



何谓业务监视

所谓业务监视是对业务应用程序本身，以及使业务陷入无法执行状态的故障主要原因进行监视的功能。

- ◆ 按照监视可选软件对应用程序/协议的停止/结果异常进行监视

虽然需要另外购买 License，但是可以监视数据库应用程序(Oracle,DB2 等)、协议(FTP,HTTP 等)、应用程序服务器(WebSphere, WebLogic 等)的停止/结果异常。详细信息请参考《参考指南 第 II 篇》的“第 6 章 监视资源的详细信息”。

- ◆ 应用程序的生存状态监视

能够使用启动资源(称为应用程序资源、服务资源)启动应用程序，通过监视用资源(称为应用程序监视资源、服务监视资源)定期监视进程的生存。在由于业务应用程序异常退出造成业务中断时有效。

注 1: 如果 ExpressCluster 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。

注 2: 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 ExpressCluster 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。在由于必须资源异常造成业务中断时有效。

何谓内部监视

内部监视是 ExpressCluster 内部模块之间的相互监视,用于监视 ExpressCluster 的各监视功能是否正常运行。

ExpressCluster 内部进行以下监视。

- ◆ ExpressCluster进程的生存状态监视

可监视的故障和无法监视的故障

ExpressCluster 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,而哪些不能监视。

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: 故障服务器的心跳停止

- ◆ 可以监视的故障示例
 - 硬件故障(OS 无法继续运行)
 - STOP 错误
- ◆ 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)

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的故障等)
 - 公网 NIC 的故障
- ◆ 无法监视的故障示例
 - 应用程序的停止/结果异常

应用程序的停止/结果异常在ExpressCluster里无法直接监视²,但是监视应用程序查出异常时,会创建自动结束程序,在应用程序资源下启动该程序、在应用程序监视资源内予以监视、则可能发生失效切换。

¹ Host Bus Adapter的缩写,不是指共有磁盘一端的适配器,而是指服务器主机一端的适配器。

² 监视选项中有关操作、数据库应用程序(Oracle,DB2等)、协议(FTP, HTTP等)、应用程序服务器(WebSphere, Weblogic等),可以进行停止/结果异常的监视。

网络分区解析

ExpressCluster 查出某个服务器心跳停止时，要判断其真正原因是服务器故障、还是网络分区症状造成的。判定为服务器故障时执行失效切换(健全的服务器上活用各种资源，启动业务应用程序)，判定为网络分区症状时比起持续业务更优先的是保存数据，因此要执行紧急关机 etc 处理。

网络分区解析方式有下列方法。

- ◆ COM方式
- ◆ ping方式
- ◆ 共享磁盘方式
- ◆ COM+共享磁盘方式
- ◆ ping+共享磁盘方式
- ◆ 多数决定方式
- ◆ 不解析网络分区

相关信息：关于网络分区解析方法设置的详细内容，请参考《参考指南 第 II 篇》的“第 8 章 网络分区解析资源的详细信息”。

失效切换的原理

一旦查出其他服务器的心跳中断，ExpressCluster 将在开始失效切换前判断是服务器的故障还是网络分区症状。然后，在正常的服务器上激活各种资源，启动业务应用程序来执行失效切换。

这时，同时移动的资源集合称为失效切换组。失效切换组从使用者的角度看，可认为是虚拟计算机。

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，应用程序在内存上存放的执行状态无法失效切换。

从发生故障到失效切换结束要花费数分钟。以下是时间表。

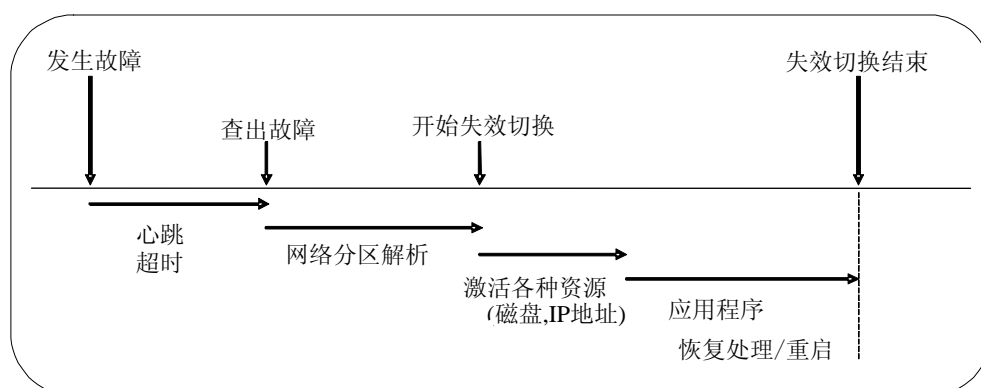


图 2-2 失效切换的时间表

◆ 心跳超时

- 执行业务的服务器发生故障后，直至待机服务器查出该故障为止的时间。
- 考虑到业务负荷等导致的延迟，调整集群属性的设置值。
(默认值为 30 秒。)

◆ 网络分区解析

- 是为了确认对方服务器的心跳的中断(心跳超时)是由于网络分区症状引起还是实际上对方服务器发生故障引起的所需时间。
- 将网络分区方式指定为共享磁盘方式时，考虑到磁盘 I/O 的延迟所需的等待时间，默认值要设置为 30 秒~60 秒左右。这个时间随着访问集群分区的时间和心跳超时值等的变化而变化。而其它分区方式时，通常是瞬间完成确认。

◆ 各种资源的激活

- 激活业务所需资源所用的时间。
- 按照一般的设置只需几秒即可激活，但激活时间也会根据登录到失效切换组上的资源的种类和数量发生变化。
(具体内容请参考《安装 & 设置指南》。)

◆ 应用程序的恢复处理/重启

- 启动业务所用的应用程序所需时间。也包括数据库的回滚/前滚等数据恢复处理的时间。
- 回滚/前滚时间等通过 Check Point Interval 的调整，可在某种程度上预测。具体内容请参考各软件产品的文档。

由ExpressCluster构建的共享磁盘型集群的硬件配置

共享磁盘型集群的 ExpressCluster 的 HW 构成如下图所示。

用于服务器间的通信

- ◆ NIC 两块 (1块用于与外部通信、1块为ExpressCluster专用)
- ◆ 用RS232C Cross Cable连接的COM端口
- ◆ 共享磁盘的特定空间

通常使用以上的配置。

与共享磁盘的连接接口使用 SCSI 或者 FibreChannel，最近多使用 FibreChannel。

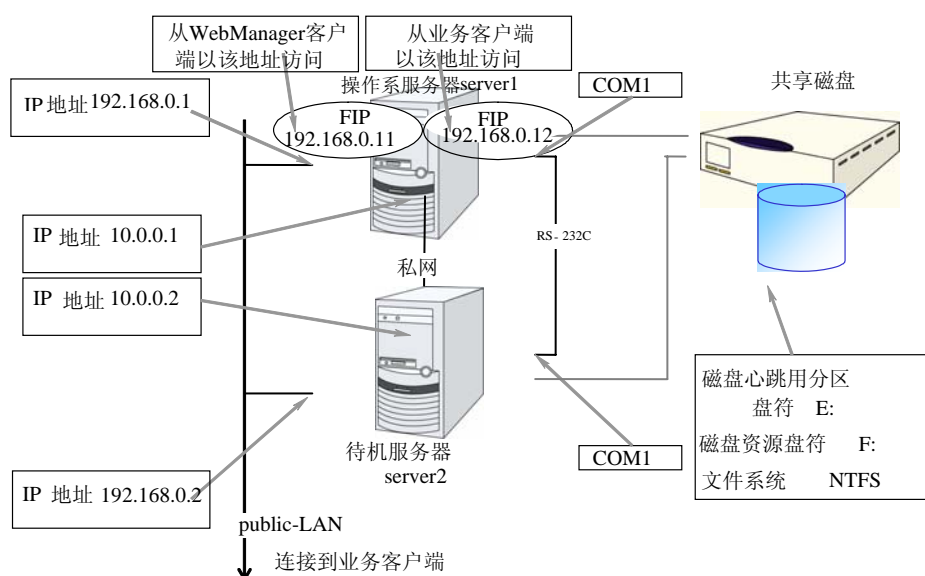


图 2-3 使用共享磁盘时的集群环境示例

以上是使用共享磁盘时的集群环境示例。

用ExpressCluster构建的镜像磁盘型集群的硬件配置

通过为各服务器的磁盘上的分区建立镜像，可替代共享磁盘装置。与共享磁盘型相比适合规模小、预算少的系统。

注： 使用镜像磁盘，您需要购买 Replicator 可选软件或 Replicator DR 可选软件。

复制镜像磁盘数据所用的网络，通常用心跳线(ExpressCluster 的内部通信用 NIC)来兼用。

用 ExpressCluster 构建的镜像磁盘型集群的硬件配置如下图所示。

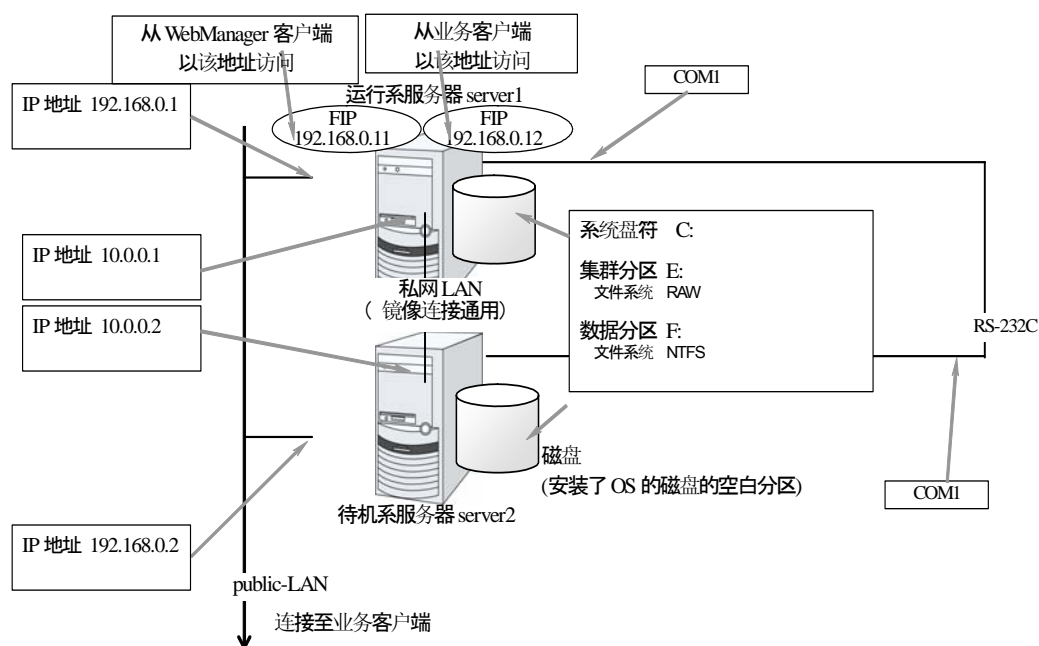


图 2-4 使用镜像磁盘时的集群环境示例

以上是使用镜像磁盘时的集群环境示例(在安装了 OS 的磁盘上确保集群分区、数据分区)。

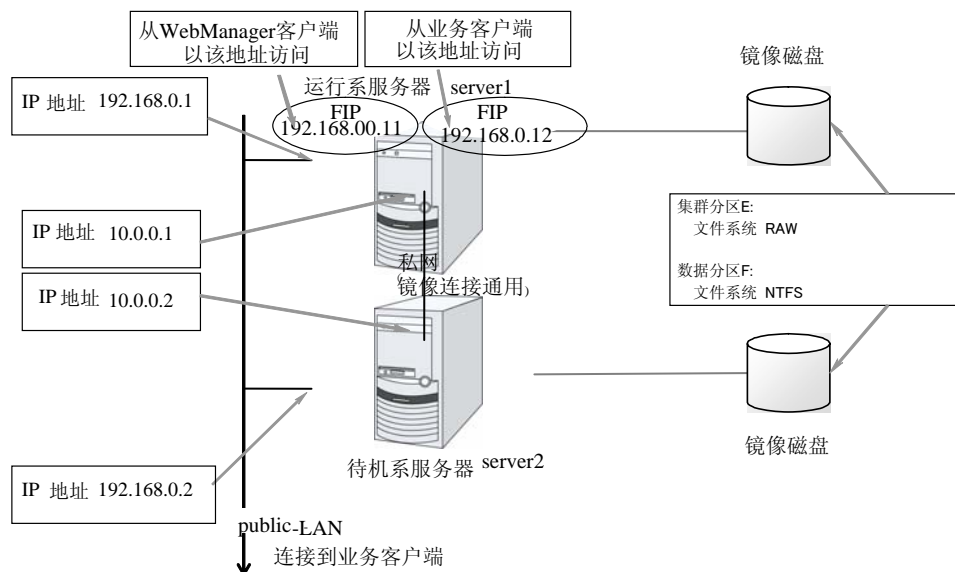


图 2-5 使用镜像磁盘时的集群环境示例

使用镜像磁盘时的集群环境的示例(备有集群分区、数据分区用的磁盘时)。

用ExpressCluster构建的共享磁盘型集群的硬件配置

该配置通过组合共享磁盘型和镜像磁盘型，镜像化共享磁盘上的分区，从而实现在共享磁盘设备出现故障时也能够不中断业务。该配置也可以利用在远程站点之间，通过与远程站点的镜像化对应灾害的发生。

注：使用共享型镜像磁盘，您需要购买 Replicator DR 可选软件。

与镜像磁盘一样，需要复制数据用的网络，一般可以由私网（ExpressCluster 内部通信使用的 NIC）兼用。

用 ExpressCluster 构建的共享磁盘型集群的硬件配置如下图所示。

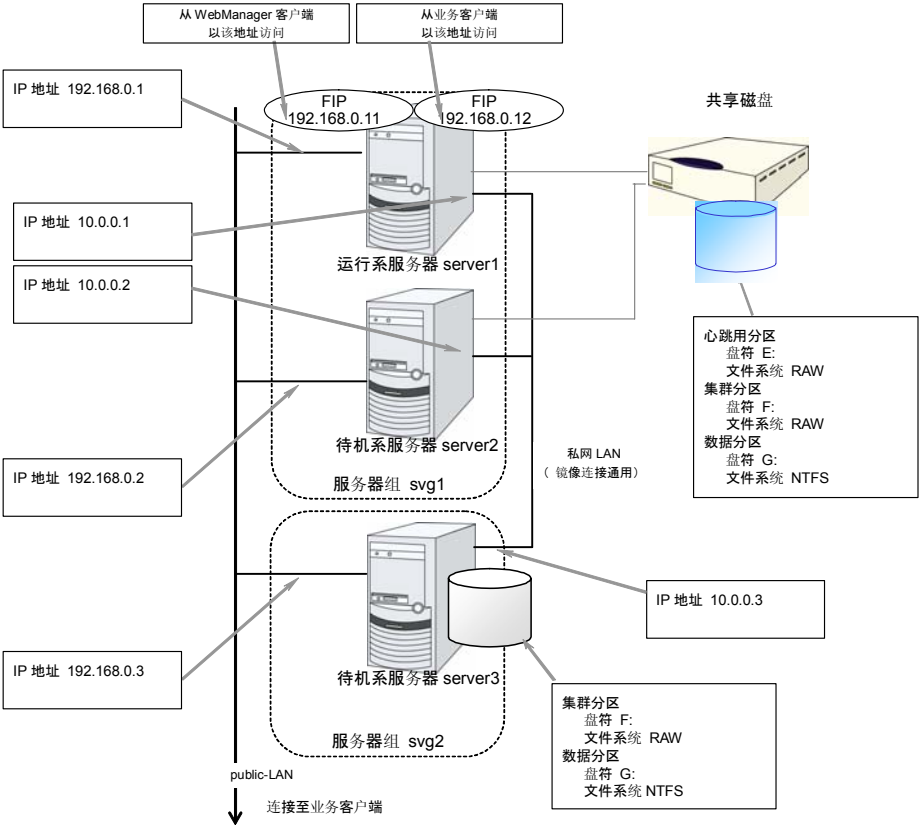


图 2-6 使用共享磁盘时的集群环境示例

以上是在同一网络内镜像化共享磁盘时的集群环境示例。如果使用共享型磁盘，则在连接在同一共享磁盘设备的服务器组之间进行镜像化，但在上述示例中，共享磁盘将镜像化到 server3 的本地磁盘上，因此待机系服务器组 svg2 的成员服务器只有 server3 一台。

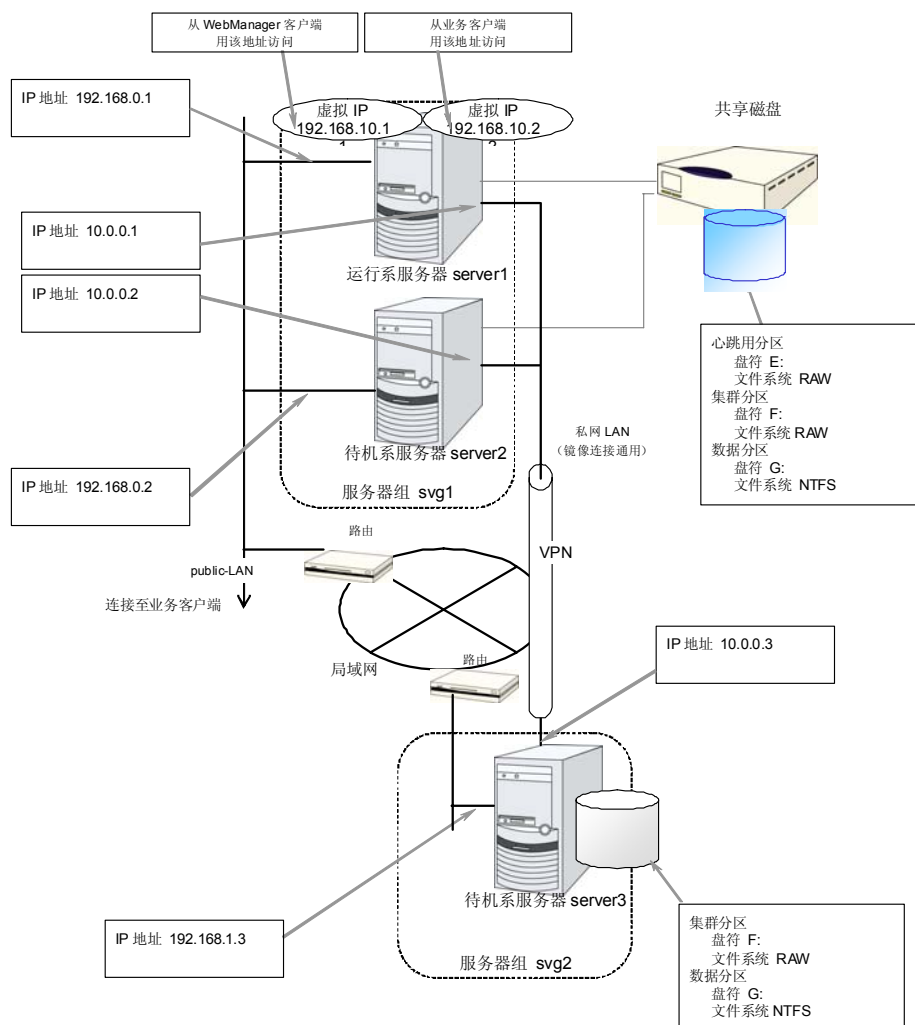


图 2-7 使用共享磁盘时的远程集群环境的示例

上图是与远程站点之间进行镜像化时的集群环境示例。在该例中，因为服务器组之间的 Public-LAN 网段不同，因此没有使用浮动 IP 地址，而是使用了虚拟 IP 地址。使用虚拟 IP 地址时，需要将中途的路由全部设置为传播主路径。并且推荐将镜像模式设置为非同步，开启数据压缩功能。

何谓集群对象

在 ExpressCluster 中按照以下结构管理各种资源。

- ◆ 集群对象
集合了一群服务器的集群系统。
- ◆ 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
捆绑了服务器的对象，属于集群对象。使用共享磁盘资源时需要使用该对象。
- ◆ 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
表示网络分区解析构造的对象，属于服务器对象。
- ◆ 组对象
表示虚拟的服务器的对象，属于集群对象。
- ◆ 组资源对象
表示虚拟服务器的资源(NW、磁盘)的对象，属于组对象。
- ◆ 监视资源对象
表示监视构造的对象，属于集群对象。

何谓资源

在 ExpressCluster 中，监视方和被监视方均被称作资源，并将监视方和被监视方的资源分类管理。这样，除能更明确的区分监视/被监视对象，在构建集群和查出故障时还可更容易应对。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。以下说明其概要。

相关信息：关于各资源的具体内容，请参考《参考指南 第 II 篇》。

心跳资源

是在服务器间用来确认彼此存活的资源。

以下是当前支持的心跳资源。

- ◆ LAN心跳资源
表示使用Ethernet的通信。

网络分区解析资源

以下是解析网络分区症状的资源。

- ◆ **COM网络分区解析资源**
COM方式的网络分区解析资源。
- ◆ **DISK网络分区解析资源**
DISK方式的网络分区解析资源。仅共享磁盘结构时可用。
- ◆ **PING网络分区解析资源**
PING方式的网络分区解析资源。
- ◆ **多数决定网络分区解析资源**
多数决定方式的网络分区解析资源。

组资源

进行失效切换时的单位，即构成失效切换组的资源。

以下是当前支持的组资源。

- ◆ **应用程序资源 (appli)**
启动 / 停止应用程序(包括用户创建的应用程序)。
- ◆ **浮动IP资源 (fip)**
提供虚拟的IP地址。从客户端访问时与普通IP地址相同。
- ◆ **镜像磁盘资源 (md)**
提供本地磁盘上的特定分区的镜像建立和访问控制功能。仅镜像磁盘结构时可用。
- ◆ **注册表同步资源 (regsync)**
在多个服务器上对特定的注册表进行同步，以实现在构成集群的服务器间令应用程序和服务按照同样的设置来运行。
- ◆ **脚本资源 (script)**
启动 / 停止用户创建脚本等脚本(BAT)。
- ◆ **磁盘资源 (sd)**
提供对共享磁盘上的特定分区的访问控制功能。仅当连接了共享磁盘装置时可用。
- ◆ **服务资源 (service)**
启动 / 停止数据库和Web等服务。
- ◆ **Print Spooler资源 (spool)**
提供对Print Spooler进行失效切换的功能。
- ◆ **虚拟计算机名资源 (vcom)**
提供虚拟的计算机名。从客户端访问时与普通计算机名相同。
- ◆ **虚拟IP资源 (vip)**
提供虚拟的IP地址。从客户端访问时与普通IP地址相同。在网络地址不同的网段间配置远程集群时使用。
- ◆ **CIFS资源 (cifs)**
提供共享磁盘/镜像磁盘上的文件夹共享功能。
- ◆ **NAS资源 (nas)**
提供文件服务器上的共享文件夹以网络硬盘的方式加载的功能。

- ◆ 共享磁盘资源 (hd)
是组合了磁盘资源和镜像磁盘资源的一种资源，提供对共享磁盘或本地磁盘上特定分区进行镜像或访问控制的功能。
- ◆ 虚拟机资源 (vm)
进行虚拟机的启动、停止及迁移。

注：使用镜像磁盘资源，需要《ExpressCluster X Replicator》或《ExpressCluster X Replicator DR》的 License。

使用共享磁盘资源时，需要《ExpressCluster X Replicator DR》的 License。

监视资源

集群系统内进行监视的主体资源。

以下是当前支持的监视资源。

- ◆ 应用程序监视资源 (appliw)
提供应用程序资源启动的进程的死活监视功能。
- ◆ 磁盘RW监视资源 (diskw)
提供对文件系统的监视。此外，当文件系统I/O停止 (Stall) 时，提供通过有意的STOP错误或者HW重置实施失效切换的功能。也可用于对共享磁盘的文件系统的监视。
- ◆ 浮动IP监视资源 (fipw)
提供对浮动IP资源启动的IP地址的监视。
- ◆ IP监视资源 (ipw)
提供对网络通堵的监视。
- ◆ 镜像磁盘监视资源 (mdw)
提供对镜像磁盘的监视。
- ◆ 镜像连接监视资源 (mdnw)
提供对镜像连接的监视。
- ◆ NIC Link Up/Down监视资源 (miiw)
提供对LAN线缆的链接状态的监视。
- ◆ 多目标监视资源 (mtw)
提供将多个监视资源集中在一起的状态。
- ◆ 注册表同步监视资源 (regsyncw)
提供对注册表同步资源进行的同步处理的监视。
- ◆ 磁盘TUR监视资源 (sdw)
对通过SCSI的TestUnitReady命令向共享磁盘访问路径发出的动作进行监视。也可用于FibreChannel的共享磁盘。
- ◆ 服务监视资源 (servicew)
提供对服务资源所启动的进程进行死活监视。
- ◆ Print Spooler监视资源 (spoolw)
提供对Print Spooler资源所启动的Print Spooler的监视。

- ◆ 虚拟计算机名监视资源 (vcomw)
提供对虚拟计算机名资源启动的虚拟计算机的监视。
- ◆ 虚拟IP监视资源 (vipw)
提供对虚拟IP资源启动的IP地址的监视。
- ◆ CIFS监视资源 (cifsw)
提供对CIFS资源公布的共享文件夹的监视。
- ◆ NAS监视资源 (nasw)
提供对NAS资源加载的网络硬盘的监视。
- ◆ 共享型镜像磁盘监视资源 (hdw)
提供对共享型镜像磁盘的监视。
- ◆ 共享型镜像磁盘 TUR 监视资源 (hdtw)
通过 SCSI 的 TestUnitReady 命令，提供对用作共享型镜像磁盘的共享磁盘设备访问路径运行的监视。也可以用于 FibreChannel 的共享磁盘。
- ◆ 自定义监视资源 (genw)
有进行监视处理的命令或脚本时，提供通过其动作结果进行监视系统的功能。
- ◆ DB2监视资源 (db2w)
提供对IBM DB2数据库的监视。
- ◆ ODBC监视资源 (odbcw)
提供对ODBC可访问的数据库的监视。
- ◆ Oracle监视资源 (oraclew)
提供对Oracle数据库的监视。
- ◆ OracleAS 监视资源 (oracleasw)
提供对 Oracle 应用程序服务器的监视。
- ◆ OfficeScan CL监视资源 (oscw)
提供对OfficeScan.客户的监视。
- ◆ OfficeScan SV监视资源 (oss w)
提供对OfficeScan.客户的监视。
- ◆ PostgreSQL监视资源 (psqlw)
提供对PostgreSQL数据库的监视。
- ◆ SQL Server监视资源 (sqlserverw)
提供对SQL Server数据库的监视。
- ◆ FTP监视资源 (ftpw)
提供对FTP服务器的监视。
- ◆ HTTP监视资源 (httpw)
提供对HTTP服务器的监视。
- ◆ IMAP4监视资源 (imap4w)
提供对IMAP服务器的监视。
- ◆ POP3监视资源 (pop3w)
提供对POP服务器的监视。

- ◆ **SMTP监视资源 (smtpw)**
提供对SMTP服务器的监视。
- ◆ **Tuxedo监视资源 (tuxw)**
提供对Tuxedo应用程序服务器的监视。
- ◆ **Websphere监视资源 (wasw)**
提供对WebSphere应用程序服务器的监视。
- ◆ **Weblogic监视资源 (wlsw)**
提供对WebLogic应用程序服务器的监视。
- ◆ **WebOTX监视资源 (otxw)**
提供对WebOTX应用程序服务器的监视。
- ◆ **虚拟机监视资源 (vmw)**
提供对通过虚拟机资源启动的虚拟机的监视。
- ◆ **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。

注：使用 DB2 监视资源、ODBC 监视资源、Oracle 监视资源、PostgreSQL 监视资源和 SQL Server 监视资源，需要《ExpressCluster X Database Agent》的 License。

使用 FTP 监视资源、HTTP 监视资源、IMAP4 监视资源、POP3 监视资源和 SMTP 监视资源，需要《ExpressCluster X Internet Server Agent》的 License。

使用 OracleAS 监视资源、Tuxedo 监视资源、Weblogic 监视资源、Websphere 监视资源、WebOTX 监视资源，需要《ExpressCluster X Application Server Agent》的 License。

使用 OfficeScan CL 监视资源、OfficeScan SV 监视资源，需要《ExpressCluster X Anti-Virus Agent》的 License。

开始使用 ExpressCluster 吧!

以上是对 ExpressCluster 的简要说明。

下面，按照以下顺序一边阅读相应的指南一边构建使用了 ExpressCluster 的集群系统。

最新信息的确认

请参考本指南的第 II 篇《发布备忘录 (ExpressCluster 最新信息)》。

集群系统的设计

请参考《安装&设置指南》的“第 I 篇 设计集群系统”以及
《参考指南》的“第 II 篇 资源的详细信息”。

集群系统的构建

请参考《安装&设置指南》的全篇。

开始操作集群系统后的故障对策

请参考《参考指南》的“第 III 篇 维护信息”。

第 II 篇 发布备忘录

（ExpressCluster 最新信息）

本篇中收录了 **ExpressCluster** 的最新信息，包括它所支持的硬件和软件的最新详细信息。此外，对限制事项和现存的问题以及防止方法也进行了说明。

- 第 3 章 **ExpressCluster**的运行环境
- 第 4 章 最新版本信息
- 第 5 章 注意限制事项

第 3 章 ExpressCluster 的运行环境

在本章中对 ExpressCluster 的运行环境进行说明。
本章中说明的项目如下。

- 硬件运行环境 50
- ExpressCluster Server的运行环境 50
- Builder的运行环境 55
- WebManager的运行环境 56
- 统合WebManager的运行环境 57

硬件运行环境

ExpressCluster 在以下结构的服务器上运行。

- ◆ IA32
- ◆ x86_64

所需规格

ExpressCluster Server 所需要的规格如下。

- ◆ RS-232C端口1个 (构建3节点以上的集群时不需要)
- ◆ Ethernet端口2个以上
- ◆ 共享磁盘、镜像用磁盘或者镜像用空分区 (使用镜像磁盘时)
- ◆ CD-ROM驱动器

脱机使用 **Builder** 时，为处理集群配置信息，需要下述方法之一(联机时不需要)。

- ◆ FD驱动器，USB内存等可移动媒介
- ◆ 安装了ExpressCluster Server的服务器以及文件共享方法

ExpressCluster Server 的运行环境

支持OS

ExpressCluster 支持以下 OS。

IA32 版

OS	支持 Replicator / Replicator DR	备注
Microsoft Windows Server 2003, Standard Edition SP2以上	○	
Microsoft Windows Server 2003, Enterprise Edition SP2以上	○	
Microsoft Windows Server 2003, Standard Edition R2 SP2以上	○	
Microsoft Windows Server 2003, Enterprise Edition R2 SP2以上	○	
Microsoft Windows Server 2008 Standard (x86版)	○	
Microsoft Windows Server 2008 Enterprise (x86版)	○	

x86_64 版

OS	支持 Replicator / Replicator DR	备注
Microsoft Windows Server 2003, Standard x64 Edition SP2以上	○	
Microsoft Windows Server 2003, Enterprise x64 Edition SP2以上	○	
Microsoft Windows Server 2003, Standard x64 Edition R2 SP2以上	○	
Microsoft Windows Server 2003, Enterprise x64 Edition R2 SP2以上	○	
Microsoft Windows Server 2008 Standard (x64版)	○	
Microsoft Windows Server 2008 Enterprise (x64版)	○	
Microsoft Windows Server 2008 R2 Standard	○	
Microsoft Windows Server 2008 R2 Enterprise	○	

所需内存容量和磁盘容量

	所需内存容量		所需磁盘容量		备注
	用户模式	Kernel模式	安装后	操作时最大	
IA32 版	110MB	32MB + 4MB(*)×(镜像磁盘资源数+共享型镜像磁盘资源)	27MB	650MB	
EM64T 版	190MB	32MB + 4MB(*)×(镜像磁盘资源数+共享型镜像磁盘资源)	41MB	650MB	

(*) 镜像磁盘资源及共享型镜像磁盘资源所需的内存容量。

非同步方式地更改或者队列大小变化时，都需要添加配置时指定大小的内存。此外，由于使用内存与镜像磁盘的 I/O 相对应，随着磁盘负荷的增加，使用的内存大小也会增加。

监视可选软件已经确认完毕的应用程序信息

监视可选软件已经将下列应用程序作为监视对象进行了运行确认。

IA32 版

监视资源	监视对象 应用程序	ExpressCluster Version	备注
Oracle监视	Oracle 10g Release 2 (10.2)	11.00~	
	Oracle 11g Release 1 (11.1)	11.00~	
	Oracle 11g Release 2 (11.2)	11.00~	
DB2监视	DB2 V9.5	11.00~	
	DB2 V9.7	11.00~	
PostgreSQL监视	PostgreSQL 8.1	11.00~	
	PostgreSQL 8.2	11.00~	
	PostgreSQL 8.3	11.00~	
	PostgreSQL 8.4	11.00~	
	PowerGres on Windows V6.0	11.00~	
	PowerGres on Windows V6.1	11.00~	
	PowerGres on Windows V6.2	11.00~	
	PowerGres on Windows V6.3	11.00~	
	PowerGres on Windows V7.0	11.00~	
	PowerGres on Windows V7.1	11.00~	
SQL Server监视	SQL Server 2008	11.00~	
	SQL Server 2008 R2	11.00~	
Tuxedo监视	Tuxedo 10g R3	11.00~	
OracleAS监视	Oracle Application Server 10g Release 3 (10.1.3)	11.00~	
	Oracle Application Server 10g Release 3 (10.1.4)	11.00~	

Weblogic监视	WebLogic Server 10g R3	11.00~	
	WebLogic Server 11g R1	11.00~	
	WebLogic Server 11g R2	11.00~	
Websphere监视	WebSphere 6.1	11.00~	
	WebSphere 7.0	11.00~	
WebOTX监视	WebOTX V8.0	11.00~	
	WebOTX V8.1	11.00~	
	WebOTX V8.2	11.00~	
OfficeScan CL监视	OfficeScan Corporate Edition 8.0 SP1 Patch 1	11.00~	
OfficeScan SV监视			

x86_64 版

监视资源	监视对象 应用程序	ExpressCluster Version	备注
Oracle监视	Oracle 10g Release 2 (10.2)	11.00~	
	Oracle 11g Release 1 (11.1)	11.00~	
	Oracle 11g Release 2 (11.2)	11.00~	
DB2监视	DB2 V9.5	11.00~	
	DB2 V9.7	11.00~	
PostgreSQL监视	PowerGres on Windows V7.0	11.00~	
	PowerGres on Windows V7.1	11.00~	
SQL Server监视	SQL Server 2008	11.00~	
	SQL Server 2008 R2	11.00~	

Weblogic监视	WebLogic Server 10g R3	11.00~	
	WebLogic Server 11g R1	11.00~	
	WebLogic Server 11g R2	11.00~	
Websphere监视	WebSphere 6.1	11.00~	
	WebSphere 7.0	11.00~	
WebOTX监视	WebOTX V8.0	11.00~	
	WebOTX V8.1	11.00~	
	WebOTX V8.2	11.00~	
OfficeScan CL监视	OfficeScan Corporate Edition 8.0 SP1 Patch 1	11.00~	
OfficeScan SV监视			

虚拟机资源的运行环境

进行虚拟机资源运行确认的虚拟基础架构的版本信息如下所示。

虚拟基础架构	版本	备注
Hyper-V	Hyper-V 1.0	
	Hyper-V 2.0	

Builder 的运行环境

描述运行 ExpressCluster Builder 所需的环境。

运行确认完毕的OS、浏览器

最新信息请参考 ExpressCluster 的主页上公开的最新文档。当前的支持情况如下。

OS	浏览器	语言
Microsoft Windows XP SP2	IE6 SP2	日语/英语/中文
Microsoft Windows Vista™	IE7	日语/英语/中文
Micorsoft Windows 7	IE8	日语/英语/中文
Microsoft Windows Server 2003 SP1以上	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2003 R2	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2008	IE7	日语/英语/中文
Microsoft Windows Server 2008 R2	IE8	日语/英语/中文

Java执行环境

使用 Builder 时需要 Java 执行环境。

Sun Microsystems

Java(TM) Runtime Environment

Version6.0 Update 20 (1.6.0_20) 或者 Version 6.0 Update 21 (1.6.0_21)

所需内存容量/磁盘容量

所需内存容量 32MB 以上

所需磁盘容量 5MB(Java 执行环境除外)

支持的ExpressCluster版本

请按照下表的组合使用 Builder 版本和 ExpressCluster 版本。如果使用下述以外的组合方式可能出现运行异常。

Builder版本	ExpressCluster X Server内部版本
3.0.0-1	11.00

WebManager 的运行环境

描述运行 ExpressCluster WebManager 所需的环境。

运行确认完毕的OS、浏览器

最新信息请参考在 ExpressCluster 的主页上公布的最新文档。当前的支持情况如下。

OS	浏览器	语言
Microsoft Windows XP SP2	IE6 SP2	日语/英语/中文
Microsoft Windows Vista™	IE7	日语/英语/中文
Micorsoft Windows 7	IE8	日语/英语/中文
Microsoft Windows Server 2003 SP1	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2003 R2	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2008	IE7	日语/英语/中文
Microsoft Windows Server 2008 R2	IE8	日语/英语/中文

Java执行环境

使用 WebManager 时需要 Java 运行环境。

Sun Microsystems

Java(TM) Runtime Environment

Version 6.0 Update 20 (1.6.0_20) 或者 Version 6.0 Update 21 (1.6.0_21)

所需内存容量/磁盘容量

所需内存容量 40MB 以上

所需磁盘容量 300KB 以上(Java 执行环境所需容量除外)

综合 WebManager 的运行环境

描述运行 Java 的 Applet 小程序版综合 WebManager 所需的环境。关于 Java 程序版的综合 WebManager，请参考《ExpressCluster X 综合 WebManager 管理员指南》。

运行确认完毕的OS、浏览器

最新信息请参考在 ExpressCluster 的主页上公布的最新文档。当前的支持情况如下。

OS	浏览器	语言
Microsoft Windows XP SP2	IE6 SP2	日语/英语/中文
Microsoft Windows Vista™	IE7	日语/英语/中文
Micorsoft Windows 7	IE8	日语/英语/中文
Microsoft Windows Server 2003 SP1	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2003 R2	IE6 SP1	日语/英语/中文
Microsoft Windows Server 2008	IE7	日语/英语/中文
Microsoft Windows Server 2008 R2	IE8	日语/英语/中文

Java执行环境

使用综合 WebManager 时需要 Java 运行环境。

Sun Microsystems

Java(TM) Runtime Environment

Version6.0 Update 20 (1.6.0_20) 或者 Version 6.0 Update 21 (1.6.0_21)

所需内存容量/磁盘容量

所需内存容量 40MB 以上

所需磁盘容量 300KB 以上(Java 执行环境所需容量除外)

第 4 章 最新版本信息

在本章中对 **ExpressCluster** 的最新信息进行说明。为您介绍新发布版本中强化和改善的内容。

- **ExpressCluster和手册的对应一览表** 60
- **最新版本** 60
- **功能强化信息** 61

ExpressCluster 和手册的对应一览表

本手册以以下版本的 ExpressCluster 为前提进行说明。请注意 ExpressCluster 的版本和手册版数的对应关系。

ExpressCluster 内部版本	手册	版数	备注
11.00	安装 & 设置指南	第1版	
	开始指南	第1版	
	参考指南	第1版	
	统合WebManager管理员指南	第3版	

最新版本

2010 年 10 月 ExpressCluster X 3.0 for Windows 的最新内部版本为 11.00。

关于最新信息，请参考 ExpressCluster 主页上公布的最新文档。

ExpressCluster 的内部版本，请通过 WebManager 进行确认。

从 WebManager 的树形视图中选择服务器的图标，该服务器的内部版本就会在列表视图中表示出来。

功能强化信息

对于各个版本，分别进行了以下的功能强化。

1	11.00	能通过同一浏览器界面操作WebManager及builder。
2	11.00	更新了集群构建向导。
3	11.00	能自动获取集群构建向导中的部分设置项目。
4	11.00	添加了上传设置信息时检查设置IP地址的功能。
5	11.00	发生故障时对于各服务器的监视资源状态或者集群的启动状态能自动选择失效切换目标。
6	11.00	添加了抑制跨服务器组进行失效切换的功能。
7	11.00	能选择[所有组]作为发生故障时的失效切换对象。
8	11.00	添加了集群启动时同时等待其他服务器启动的取消命令clpbwctrl。
9	11.00	添加了处理集群外部以及ExpressCluster以外的监视工具所发生的故障通信的命令clprexec，以及对已被通知的故障的恢复用的信息接收监视资源。
10	11.00	Oracle监视对象应用程序发生超时时，能获取监视进程的dump信息。
11	11.00	Oracle监视资源查出异常时，能获取Oracle的详细信息。
12	11.00	增加了支持的网络警告灯。
13	11.00	镜像恢复以及非同步镜像时，能压缩并转发镜像数据。
14	11.00	提高了镜像磁盘/共享型镜像磁盘完全拷贝的速度。
15	11.00	Hyper-V的主机OS集群化后，客机OS能作为资源使用。
16	11.00	Hyper-V的主机OS集群化后，能在查出故障或进行操作时运行快速迁移。
17	11.00	能进行镜像连接的冗余化设置。
18	11.00	能判别通过应用程序资源启动的应用程序以及脚本资源启动的批处理文件的结束代码并以此判断有无异常为进行失效切换的时机。
19	11.00	添加了对未进行资源登录的共享磁盘上的卷也能进行访问控制的clpvolctrl命令。

第 5 章 注意限制事项

在本章中对注意事项、现有问题及其避免方法进行说明。
本章中说明的项目如下。

- 研究系统配置时 63
- 安装ExpressCluster前 65
- 创建ExpressCluster的配置信息时 71
- 操作ExpressCluster后 72
- 与旧版的兼容性 77

研究系统配置时

对 HW 的准备、系统配置以及共享磁盘配置时需要留意的事项进行说明。

关于Builder、WebManager的运行OS

- ◆ 要在 x86_64 的机器上运行 Builder 以及 WebManager，需要使用 32bit 的 Java Runtime。

关于镜像磁盘/共享型镜像磁盘的需求

- ◆ 不能使用动态磁盘。请使用基本磁盘。
- ◆ 使用镜像磁盘资源/共享型镜像磁盘资源需要镜像用的分区(数据分区和集群分区)。
- ◆ 镜像用的分区的磁盘上的配置没有特别限制、数据分区的大小按字节单位必须完全一致。此外，集群分区需要17MB以上的容量。
- ◆ 数据分区作为扩展分区上的逻辑分区建立时，请确认两个服务器上都建立逻辑分区。即使主分区与逻辑分区指定了相同的容量大小，实际大小也有可能不同。
- ◆ 为负载均衡建议将集群分区和数据分区分别设在不同的磁盘上（虽然创建在同一块磁盘上也能够运行，但非同步镜像和镜像中断状态下的写入性能会有些许下降）。
- ◆ 在磁盘上要确保用镜像资源建立镜像的数据分区，则两台服务器的磁盘类型必须相同。

例)

组合	服务器1	服务器2
OK	SCSI	SCSI
OK	IDE	IDE
NG	IDE	SCSI

- ◆ 用“Disk Management”等确保的分区大小，按照磁盘柱面周围的块(单元)数来分配。因此，如果服务器间用作镜像磁盘的磁盘的几何数据不同，可能无法使数据分区的大小完全一致。为了避免此类问题发生，对于确保数据分区的磁盘，建议两台服务器上的HW构成，包括RAID结构等在内，应该一致。
- ◆ 两台服务器的磁盘类型和几何数据不匹配时，在设定镜像磁盘资源/共享型镜像磁盘资源前，请通过clpvolsz命令确认两台服务器数据分区的正确大小，如果大小不一致，请再次使用clpvolsz命令缩小较大的分区。
- ◆ 为RAID结构的磁盘建立镜像时，对磁盘阵列控制器的缓存以Write-Thru方式使用则写入性能大大下降，建议您使用Write Back方式。但使用Write Back方式时，要使用搭载了电池的磁盘阵列控制器或者同时使用UPS。
- ◆ 有OS的页文件的分区不能建立镜像。

关于共享磁盘的需求

- ◆ 不能使用动态磁盘。请使用基本磁盘。
- ◆ 不能使用软件RAID(Stripe Set、带校验的Stripe Set)和Volume Set。

NIC Link Up/Down监视资源

由于 NIC 的板卡、驱动器而不支持必要的 DeviceIoControl()的情况是很少见的。如不支持时, 则该监视资源无法使用。使用该监视资源时, 请先用试用版 License 等, 事先进行运行确认。

关于镜像磁盘/共享型镜像磁盘的write性能

镜像磁盘资源/共享型镜像磁盘资源的磁盘镜像建立有同步镜像和非同步镜像 2 种方式。

同步镜像时, 每当向镜像化对象的数据分区发出写入请求, 都会向两台服务器的磁盘执行写入, 并等待其结束。虽然向各服务器的写入是同时执行的, 但向其它服务器的磁盘写入是通过网络进行的, 所以与不进行镜像化的普通本地磁盘相比写入性能会降低。特别是网络通信速度慢, 延迟时间长的远程集群结构, 性能会大幅下降。

非同步镜像时, 向自身服务器写出是立即执行的, 而向其它服务器的写出会先保存到本地队列中, 在后台写出。因为不需要等待向其他服务器的写出结束, 即使网络性能较差时写出性能也不会有大幅下降。但是即使是非同步镜像时, 也会在每次发出写入请求时将更新数据保存在队列中, 与不建立镜像的普通本地磁盘和共享磁盘相比, 写入性能会有降低。因此, 向磁盘的写入处理需要较高吞吐量的系统(更新服务器较多的数据库系统等), 推荐使用共享磁盘。

此外, 非同步镜像时, 虽然保证写入顺序, 但运行服务器如宕机, 可能会丢失最新的更新内容。因此, 如想在故障发生前一刻保证此刻信息切实被交接, 需要使用同步镜像或者共享磁盘。

关于非同步镜像的历史文件

在非同步模式的镜像磁盘/共享型镜像磁盘中, 内存上的队列里记录不下的写入数据, 会临时作为历史文件记录到被指定为历史文件夹的目录下。没有设置该历史文件的大小限制时, 可以不受限制地写出到所指定的文件夹内。因此在这种设置的情况下, 如果线路速度比业务应用程序的磁盘更新量低很多, 则向远程服务器的写入处理赶不上磁盘更新的速度, 磁盘里就会充斥历史文件。因此, 远程集群结构也需要根据业务 AP 的磁盘更新量保证通信线路的速度。

此外, 为了避免长时间通信延迟、连续发生磁盘更新导致历史文件保存文件夹溢出, 需要为写出历史文件的磁盘确保充足的空余容量, 或者设置历史文件的大小限制, 又或者指定系统驱动器之外的其它驱动器。

关于多个非同步镜像间的数据一致性

在非同步模式的镜像磁盘/共享型镜像磁盘中，将写入运行服务器的数据分区的内容也按照同样顺序写入待机服务器的数据分区中。

除了镜像磁盘的初始构建过程中和镜像化中断后的恢复（复制）过程中以外，都可以保证该写入顺序，因此可以保证待机服务器的数据分区上的文件间的数据一致性。

但多个镜像磁盘资源/共享型镜像磁盘资源间由于无法保证写入顺序，比如数据库的数据库文件和 **Journal**（日志）文件，如果一方的文件比另一方陈旧，则数据一致性无法确保的文件被分散到多个非同步镜像磁盘上后，可能会由于服务器宕机等引起失效切换时业务应用程序的非正常运行。

因此，这类文件必须放置在同一非同步镜像磁盘/共享型镜像磁盘上。

关于Multiboot

如果从其它启动磁盘启动，镜像和共享磁盘的访问限制将被撤销，镜像磁盘的一致性和共享磁盘的数据保护将无法保证，使用这类资源时请不要使用 **Multiboot**。

安装 ExpressCluster 前

OS 安装完成后，在对 OS 和磁盘进行设置时所须注意的事项。

关于文件系统

安装 OS 的分区、作为共享磁盘的磁盘资源使用的分区以及镜像磁盘/共享型镜像磁盘资源的数据分区的文件系统请使用 **NTFS**。

通信端口号

ExpressCluster 中缺省使用以下端口号。该端口号可在 **Builder** 上更改。请不要从 ExpressCluster 以外的程序访问这些端口号。

为服务器设置防火墙时，请使下列端口号可以访问。

[服务器・服务器之间]

From		To		备注
服务器	自动分配 ³	服务器	29001/TCP	内部通信
服务器	自动分配	服务器	29002/TCP	数据传送
服务器	自动分配	服务器	29003/UDP	警告同步
服务器	自动分配	服务器	29004/TCP	磁盘 Agent 间通信
服务器	自动分配	服务器	29005/TCP	镜像驱动程序间通信
服务器	29106/UDP	服务器	29106/UDP	心跳

[服务器・客户端之间]

From		To		备注
客户端	自动分配	服务器	29007/TCP 29007/UDP	客户端服务通信

[服务器・WebManager 之间]

From		To		备注
WebManager	自动分配	服务器	29003/TCP	http 通信

[以集成 WebManager 连接浏览器的服务器以及管理目标服务器之间]

From		To		备注
连接了浏览器的服务器	自动分配	服务器	29003/TCP	http 通信

[其他]

					备注
服务器	自动分配	—	网络警告灯	514/TCP	控制网络警告灯
服务器	自动分配	—	服务器的 BMC 的管理 LAN	623/UDP	控制 BMC（强行停止/机箱灯的联动）
服务器	自动分配	—	监视目标	icmp	IP 监视
服务器	自动分配	—	NFS 服务器	icmp	确认 NAS 资源的 NFS 服务器是否活动
服务器	自动分配	—	监视目标	icmp	Ping 方式网络分区解析资源的监视目标

³ 自动分配在某一时刻未被使用的端口号。

使用镜像连接监视资源时，为了使 ExpressCluster 能在服务器间使用 ping 命令进行通信确认，必须对 icmp 包进行通行设置。使用镜像连接监视资源时，请更改防火墙的设置，以保证服务器之间能使用 ping 命令进行通信确认操作。

时钟同步的设置

集群系统中，推荐将多个服务器的时钟定期同步操作。请使用时间服务器等使服务器的时钟同步。

关于共享磁盘

- ◆ 在没有通过ExpressCluster限制访问的状态下，如果启动多台与共享磁盘相连的服务器，可能导致共享磁盘上的数据被损坏。在实施访问限制前，请务必仅启动一台机器。
- ◆ 将磁盘方式作为网络分区解析方式使用时，请在共享磁盘上创建DISK网络分区解析资源使用的17MB以上RAW分区（磁盘心跳用分区）。
- ◆ 作为磁盘资源，用于服务器之间的数据交接的分区（切换分区）请以NTFS来格式化。
- ◆ 共享磁盘上的各分区在所有服务器上都请设置为同样的盘符。
- ◆ 要在共享磁盘上创建分区或者格式化，请只在某一台服务器上进行。无须在各服务器上重新创建/重新格式化。只是盘符需要在各服务器上进行设置。
- ◆ 由于重新安装服务器要继续使用共享磁盘上的数据时，请不要进行分区的确保和格式化。如果进行分区的确保和格式化将删除共享磁盘上的数据。

关于镜像磁盘用的分区

- ◆ 请在各服务器的本地磁盘上创建17MB以上的RAW分区作为管理镜像磁盘资源的分区(集群分区)。
- ◆ 在各服务器的本地磁盘上创建镜像化对象的分区(数据分区)，并以NTFS格式化（为已有的分区建立镜像时，无须重新创建分区）。
- ◆ 在两台服务器上设置的数据分区的大小应相等。请使用clpvolsz命令确认调整正确的分区大小。
- ◆ 两台服务器上设置的集群分区和数据分区的盘符应该一致。

关于共享型镜像磁盘用的分区

- ◆ 作为共享型镜像磁盘资源的管理用分区（集群分区），请为每个服务器组的共享磁盘（如果服务器组的成员服务器只有1台，则为本地磁盘）创建17MB以上的RAW分区。
- ◆ 请在各个服务器组的共享磁盘（如果服务器组的成员服务器只有1台，则为本地磁盘）创建镜像化对象分区（数据分区）并用NTFS方式进行格式化（如果镜像化已有的分区，则无需重新创建分区）。
- ◆ 在两服务器组上设置的数据分区的大小应相等。请使用clpvolsz命令确认调整正确的分区大小。
- ◆ 请在各个服务器上为集群分区和数据分区设置相同的盘符。

OS启动时间的调整

请将从接通电源到OS启动为止的时间设为比以下的两个时间都长⁴。

- ◆ 使用共享磁盘时，从接通磁盘的电源到可以使用时的时间
- ◆ 心跳超时时间

网络的确认

- ◆ 确认通过心跳线或镜像连接使用的网络。在集群内所有的服务器上确认。
- ◆ 请使用ipconfig命令或ping命令确认网络的状态。
 - 公网 (与其它机器通信的网络)
 - 私网专用 LAN(ExpressCluster 的服务器之间连接的网络)
 - 镜像连接 LAN(与私网共用)
 - 主机名
- ◆ 集群上使用的浮动IP资源的IP地址不用设置到OS端。
- ◆ 只有IPv6环境的场合、如果发生了网卡LinkDown、则该网卡绑定的IP地址会处于不可见状态，直接影响ExpressCluster的运行。请执行下面的命令，解除对媒体连接的限制

```
netsh interface ipv6 global dhcpmediasense=disabled
```

⁴ 如果在BOOT时只选择了一个OS，即使设置了启动等待时间可能也会被忽略。这时，编辑boot.ini文件，在[Operating System]段中添加第2个入口。第2个入口可以通过拷贝第1个得到。

关于与ESMPRO/AutomaticRunningController结合使用

要求与 ESMPRO/AutomaticRunningController(以下简称 ESMPRO/AC)结合使用的时候, ExpressCluster 的构筑/设置有以下的事项。如果不满足这些事项, 与 ESMPRO/AC 结合使用的功能将有可能不能正常工作。

- ◆ 作为网络分区解析资源, 只有DISK方式的资源不能单独指定。指定DISK方式的时候, 必须与PING方式、COM方式等其它的网络分区解析方式资源一起组合指定。
- ◆ 建立磁盘TUR监视资源时, 最终动作的设定值为默认(不进行任何操作), 请不要对该项进行更改。
- ◆ 建立磁盘RW监视资源时, [文件名]的设定值指定为共享磁盘上的路径时, [监视时间]的设定值为默认(激活时), 请不要对该项进行更改。
- ◆ 断电后再启动时, 在ExpressCluster管理器上可能会提示以下的警告信息。根据以上的设定, 在实际工作中并不会发生故障, 所以可以忽略。
 - ID:18
模块名:nm
信息:资源<DiskNP资源名>启动失败。(服务器名:xx)
 - ID:1509
模块名:rm
信息:监视<磁盘TUR资源名>查出异常。(4: 开启设备失败。请确认监视卷的磁盘状态。)
- ◆ 关于ESMPRO/AC的设定方法、注意事项等, 请参考《ExpressCluster X for Windows PP指南》的ESMPRO/AC章节的内容。

关于ipmiutil

- ◆ 在以下功能中，使用了已经作为BSD许可证的开源代码被公开的IPMI Management Utilities (ipmiutil)控制各服务器的BMC固件。因此使用这些功能时，请在各集群服务器上安装ipmiutil。
 - 强行停止功能
 - 机箱 ID 指示灯联动
- ◆ 使用上述功能时，为了保证管理底板管理控制器（BMC）的LAN板卡所使用的IP地址和OS所使用的IP地址之间的通信，请设置各服务器的BMC。服务器上没有安装BMC或者管理BMC所使用的网络处于关闭状态时无法使用该功能。BMC的设置方法请参考各服务器的手册。
- ◆ ExpressCluster没有附带ipmiutil。Ipmiutil的获取和安装方法请参考“ExpressCluster X 安装&设置指南”的“第1章 确定系统配置”中的“9. 设置BMC和ipmiutil”。
- ◆ ipmiutil相关的以下事项本公司恕不对应。请根据用户自己的判断和责任使用。
 - ipmiutil 自身相关的咨询
 - ipmiutil 的运行保证
 - ipmiutil 的问题对应、因其问题造成的故障
 - 各服务器的 ipmiutil 的支持情况咨询
- ◆ 请用户提前确认您计划使用的服务器（硬件）是否支持ipmiutil。请注意，即使硬件本身是基于IPMI标准，实际上也可能无法运行ipmiutil。

关于在Server Core的安装

- ◆ 在Windows Server 2008的Server Core环境下安装ExpressCluster时，通过命令提示符，请执行CD盘路径下的menu.exe。显示菜单界面。
之后的步骤和普通安装相同，但是在license登录时不能选择[通过license文件登录]。
请务必选择[输入license项目后登录]。

创建 ExpressCluster 的配置信息时

在设计和创建 ExpressCluster 的配置信息前，需要根据系统的配置确认并留意以下事项。

组资源非激活异常时的最终运行

查出非激活异常时的最终运行如选择了“无操作”，则组不会在非激活失败的状态下停止。请注意在实际的业务中使用时不要设置为“无操作”。

延迟警告比例

将延迟警告比例设置为 0 或者 100，可进行以下操作。

- ◆ 延迟警告比例设置为0时

对每一监视以 Alert 通告延迟警告。

使用该功能可计算出服务器在高负荷状态下对监视资源的轮询时间，确定监视资源的监视超时时间。

- ◆ 延迟警告比例设置为100时
不通告延迟警告。

除了测试外，请不要设置 0%等低数值。

关于磁盘监视资源和共享型镜像磁盘TUR监视资源的监视方法TUR

- ◆ 不支持SCSI的Test Unit Ready命令的磁盘和磁盘接口(HBA)上无法使用。
有时硬件支持而驱动程序不支持，请同时确认驱动程序的规格。
- ◆ 与Read方式相比对OS或磁盘的负荷小。
- ◆ Test Unit Ready中有时不能查出实际的对媒体的I/O错误。

关于WebManager的画面更新间隔

- ◆ 在WebManager标签页的“画面数据更新间隔”中，一般请不要设置比30秒小的值。
如果设置比30秒小的值，可能会影响ExpressCluster的性能。

关于心跳资源的设置

- ◆ 最少要设置1个内核模式LAN心跳资源。
- ◆ 将私网专用的LAN作为心跳资源登录，然后将公网LAN也作为心跳资源登录(建议设置2个以上的心跳资源)。
- ◆ 心跳超时时间需要比OS重启所需时间短。如果未满足该条件，集群内的部分服务器进行重启时，其他服务器将无法正确检测出该服务器的重启，重启后将发生运行异常。

关于能用于脚本注释等的2字节系字符编码

- ◆ 在ExpressCluster中，Windows环境下编辑的脚本作为Shift-JIS使用，而Linux环境下编辑的脚本则作为EUC使用。如使用其他字符编码，可能因环境不同而出现乱码的情况。

操作 ExpressCluster 后

对作为集群开始操作后发生的事件需要注意以下几点。

恢复运行中的操作限制

在监视资源的查出异常的设置中将恢复对象指定为组资源(磁盘资源、应用程序资源、...), 在监视资源查出异常后的恢复运行变化过程中(再活性化 → 失效切换 → 最终运行), 请不要通过 WebManager 或命令进行以下操作。

- ◆ 集群的停止 / 挂起
- ◆ 组的开始 / 停止 / 移动

由于监视资源异常导致恢复运行时，如在恢复过程中实施上述控制，则该组的其他的组资源可能不会停止。

此外，即使监视资源处于异常状态，只要执行了最终运行即可实施上述控制。

关于命令篇中没有记载的可执行文件或脚本文件

安装目录下存在命令篇中没有记载的可执行文件或脚本文件，请不要从 ExpressCluster 以外的地方执行。

对于执行时产生的影响不属于支持范围。

集群关机/集群关机重启

正在使用镜像磁盘，激活组的过程中请不要执行 `clpstdn` 命令或者从 WebManager 执行集群关机和集群关机重启。

正在激活组时无法令其变为非激活状态。因此，镜像磁盘资源未被正常地非激活时，OS 可能关机或者发生 Mirror Break。

特定服务器的关机、重启

使用镜像磁盘时，执行命令或从 WebManager 执行服务器的关机、关机重启命令会发生 Mirror Break。

从网络分区状态的恢复

在发生了网络分区的状态下，构成集群的服务器之间无法确认彼此的状态，因此在该状态下对组进行操作(启动/停止/移动)或者重启服务器等，服务器之间对集群状态的识别会产生出入。在这种启动了多个状态识别不同的服务器的状态下进行网络恢复，则其后面的组操作会非正常运行，因此在网络分区状态时，关闭与切断网络（无法与客户端通信）的服务器，或者先停止 ExpressCluster Server 服务，待网络恢复后重启，再复原到集群中。万一在启动了多个服务器的状态下恢复了网络，则可通过重启集群状态识别不同的服务器，复原到正常状态。

另外，使用网络分区分析资源时，即使发生了网络分区，也通常会紧急关闭某个（或者所有）服务器、启动多个彼此无法通信的服务器来避免。紧急关闭的服务器以手动重启或者紧急关机时的动作设置为重启时，重启的服务器都会再次紧急关机（Ping 方式或者多数取胜方式下将停止 ExpressCluster Server 服务）。但以 DISK 方式使用多个磁盘心跳用分区时，由于磁盘路径故障导致无法通过磁盘进行通信的状态下而发生网络分区，则两台服务器将会以保留状态继续运行。

关于WebManager

- ◆ 当WebManager 的“客户端数据更新方法”设置为“Polling”时，WebManager上显示的内容会定期更新，即使状态发生变化也不会立即显示。如要获取最新信息，请点击[刷新]按钮。
- ◆ 如在WebManager获取信息过程中发生服务器宕机等，则获取信息失败，部分对象可能无法正确显示。
当WebManager 的“客户端数据更新方法”设置为“Polling”时，请等待至下次自动更新时或者点击[刷新]按钮获取最新内容。设置为“Realtime”时，将自动更新为最新的内容。
- ◆ ExpressCluster的日志采集不能由多个WebManager同时执行。
- ◆ 如果在与连接目标无法通信的状态下进行操作，则等待控制返回可能会花费些许时间。
- ◆ 在鼠标处于表示正在处理的手表或者沙漏状态下，将其移至浏览器外，即使是正在处理光标也会返回到箭头状态。
- ◆ 经由Proxy服务器时请对该服务器进行设置，以便可以中转WebManager的端口号。
- ◆ 对ExpressCluster进行升级后，先退出Web浏览器，清空Java的缓存后再重启浏览器。

关于Builder

- ◆ 与以下产品的集群配置信息不兼容。
 - ExpressCluster X 3.0 for Windows 以外的 Builder
 - ExpressCluster for Linux 的 Builder
 - ExpressCluster for Windows Value Edition 的 Builder
- ◆ 退出Web浏览器(点菜单的[退出]或窗框的[X]按钮等)将丢失当前的编辑内容。即使更改了配置也不会显示提示保存的确认对话框。
需要保存编辑内容时，请在退出前，点击Builder 的菜单栏中的[文件]-[保存]。
- ◆ 重载Web浏览器(点击菜单或者工具栏的[刷新]按钮等)将丢失当前的编辑内容。即使更改了配置也不会显示提示保存的确认对话框。
需要保存编辑内容时，请在退出前，请执行Builder的菜单栏的[文件]-[保存]。

关于ExpressCluster Disk Agent服务

请勿停止 ExpressCluster Disk Agent 服务。停止服务后无法手动启动。这时必须重启 OS，再启动 ExpressCluster Disk Agent 服务。

关于镜像构建中的集群配置信息的变更

镜像构建过程中(包括初始构建)请不要更改集群配置信息。更改集群配置信息后，驱动程序的运行可能出现不正常。

关于镜像磁盘的待机服务器的集群恢复

镜像磁盘激活时待机服务器在停止集群服务（ExpressCluster Server 服务）的状态下运行时，在开始服务之后恢复到集群前请再次启动待机服务器。如果直接就这样恢复的话会出现镜像的差异信息不正确，镜像磁盘会发生不整合。

关于镜像磁盘、共享型镜像磁盘间的配置变更

进行配置变更，将已通过镜像磁盘资源进行镜像化的磁盘更改为通过共享型镜像磁盘资源进行镜像化时，请先上传删除了现有镜像磁盘资源的配置信息，在变更为现有资源已被删除的状态后，再上传添加了共享型镜像磁盘资源的配置信息。将共享型镜像磁盘变更为镜像磁盘时也是如此。

如在不按以上步骤删除现有资源的情况下上传已替换为新资源的配置信息，可能导致磁盘镜像化的设置变更无法正常进行，运行可能出现不正常。

关于chkdsk命令和磁盘碎片整理

对磁盘资源上控制的共享磁盘上的切换分区或镜像磁盘资源上镜像化的数据分区执行 chkdsk 命令或磁盘碎片整理时，需要在已经启动了资源的服务器上执行。如果没有启动资源，由于访问权限的限制，将无法执行。

另外，在修复模式（/f 选项）下执行 chkdsk 命令时，如果对象分区上有打开的文件或文件夹，将会执行失败，请停止失效切换组，在单机启动对象磁盘资源/镜像磁盘资源的状态下执行该命令。如果有对对象分区进行监视的磁盘 RW 监视资源，需要提前暂停监视资源。

关于索引服务

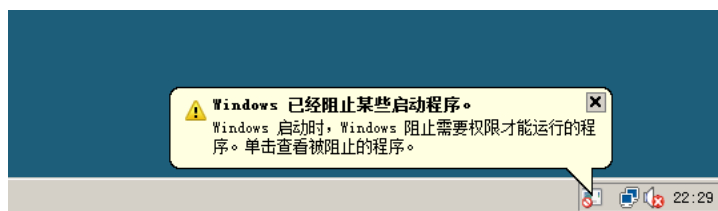
在索引服务的目录中创建共享磁盘/镜像磁盘上的目录、对共享磁盘/镜像磁盘上的文件夹创建索引时，需要从 ExpressCluster 控制，将索引服务设为手动启动，在激活共享磁盘/镜像磁盘后启动。如果将索引服务设为自动启动，则索引服务会 OPEN 对象卷，在之后的激活处理中将造成 mount 处理失败，最后造成从应用程序或资源浏览器访问磁盘因“参数错误”（错误代码 87）而失败。

关于Windows Server 2008环境中的应用程序资源/脚本资源的画面显示

从 ExpressCluster 的应用程序资源、脚本资源启动的进程需要用对话 0 执行，因此如果启动了有 GUI 的进程，在 Windows Server 2008 中将弹出“发现对话型服务对话框”窗口，在该窗口中，如果不选择“显示消息”，则不会显示 GUI。

关于Windows Server 2008环境中启动程序的运行

在 Windows Server 2008 环境中,如果用 built-in Administrator 以外的账号登录,将显示以下界面，登录在启动程序中的 ExpressCluster 的模块（clpalt.exe,clphookstdn.exe）将被阻止运行。



被阻止的模块的功能虽然无法使用，但不会影响 ExpressCluster 的运行。各模块的功能概要如下。

clpalt.exe

显示警告服务的报警图标。详细内容请参考《参考指南》的“第 9 章 其他监视设置信息 警告服务”。

clphookstdn.exe

该功能用于在登录用户从[开始]菜单中选择关闭时，输出警告对话框。

与旧版的兼容性

与ExpressCluster X 1.0/2.0/2.1的兼容性

X 1.0, X 2.0, X 2.1 的 Builder 所创建的集群配置信息在 X 3.0 以上中也可以使用。在 X 2.0 以上中只是组资源、监视资源检测出异常时的失效切换目标服务器的选择方式将变成默认值“稳定运行的服务器”，因此在 3 节点以上配置的情况下，失效切换目标服务器的选择结果可能会与 X 1.0 不同。

失效切换目标服务器存在多个时，如果设置了切换到“稳定运行的服务器”时，则如果失效切换目标服务器也发生了异常，可以优先切换到尚未发生异常的服务器。与此不同，在 X 1.0 中，将会切换到可移动服务器中优先级最高的服务器上，可能造成切换到了最初发生了异常的服务器上，最终造成无法切换到第 3 台服务器上。

因此一般情况下推荐设置为“稳定运行的服务器”。如果需要和 X 1.0 相同的运行，可以在各资源的属性设置标签页中将失效切换目标服务器的设置更改为“优先级最高的服务器”。

关于与ExpressCluster Ver 8.0以下版本的兼容功能

使用下列功能时，集群名、服务器名、组名需要按照 Ver8.0 的命名规则进行设置。

- ◆ 兼容命令
- ◆ ExpressCluster Alert Service (通报图标)
- ◆ ExpressCluster客户端
- ◆ ESMPRO/AC联动功能
- ◆ ESMPRO/SM联动功能
- ◆ 兼容API

原有版本的命名规则如下。

- ◆ 集群名
 - 15个字符以内
 - 可以使用的字符有半角英文数字、中横线 (-)、下划线 (_)。
 - 请不要指定PRN等的DOS输出输入设备名。
 - 不区分大小写。
- ◆ 服务器名
 - 15个字符以内
 - 可以使用的字符有半角英文数字、中横线 (-)、下划线 (_)。
 - 不区分大小写。
- ◆ 组名
 - 15个字符以内
 - 可以使用的字符有半角英文数字、中横线 (-)、下划线 (_)。
 - 请不要指定PRN等的DOS输出输入设备名。
 - 不区分大小写。

关于兼容API

兼容 API 请指定为 ExpressCluster Ver8.0 以下版本可以使用的 API。兼容 API 在 ExpressCluster X 上也可以使用，但是有以下限制事项。

只支持下列资源。其他资源即使设置了也无法从兼容 API 引用。

- 磁盘资源
- 镜像磁盘资源
- 虚拟计算机名资源
- 浮动 IP 资源（仅限 IPv4 地址）
- 虚拟 IP 资源（仅限 IPv4 地址）
- 打印机池资源

集群名、服务器名、组名需要遵循原来版本的规则设置。如果指定了原来版本规则以外的名称，则兼容 API 无法引用。

使用 Builder 所指定的资源名，不能使用兼容 API。

集群事件发生的时机不完全兼容。事件种类虽然相同，通知事件的数量、顺序可能会与原来版本不同。

从常驻进程使用兼容 API 时，[ExpressCluster Server]服务的停止→重启时、执行 ArmTerminateApi → ArminitializeApi，需要对兼容 API 进行重新初始化。原则上，请设置为通过脚本资源的开始和结束脚本启动或停止。

无法使用 Ver3.0 兼容 I/F。

关于脚本文件

需要移动在 ExpressCluster Ver8.0 以下版本所使用的脚本文件时，请将环境变量名从最初的 "ARMS_" 替换为 "CLP_"。

例) IF "%ARMS_EVENT%" == "START" GOTO NORMAL

↓

IF "%CLP_EVENT%" == "START" GOTO NORMAL

附录

- 附录 A 词汇表
- 附录 B 索引

附录 A 词汇表

英文字母和数字

CLUSTER分区

镜像磁盘中设定的分区。用于管理镜像磁盘。
(相关)磁盘心跳用分区

汉字

心跳线

集群服务器之间的通信路径。
(相关) 私网、公网

虚拟IP地址⁵

构筑远程集群时使用的资源(IP地址)。

管理客户端

已启动WebManager的机器。

启动属性

集群启动时，决定是自动还是手动启动失效切换组的失效切换组的属性。
可在管理客户端进行设置。

共享磁盘

可从多台服务器访问的磁盘。

共享磁盘型集群

使用共享磁盘的集群系统。

切换分区

连接到多台计算机的，可切换使用的磁盘分区。
(相关)磁盘心跳用分区

集群系统

通过LAN等连接多台计算机，并作为1个系统进行操作的系统形态。

集群关机

关闭整个集群系统(构成集群的所有服务器)。

集群分区

镜像磁盘中设定的分区。用于管理镜像磁盘。相关
(磁盘心跳用分区)

运行服务器

对某一业务集合来说，正在运行业务的服务器。
(相关) 待机服务器

从服务器 (服务器)

一般使用时，失效切换组进行失效切换的目标服务器。
(相关) 主服务器

服务器组

连接在同一网络或共享磁盘设备上的服务器的集合

待机服务器

非运行服务器。
(相关) 运行服务器

磁盘心跳用分区

共享磁盘型集群中用于心跳通信的分区。

⁵ 虚拟IP地址为仅在Windows版中使用的概念。

数据分区	能像共享磁盘的切换分区一样使用的本地磁盘镜像磁盘中设定的数据用的分区。 (相关) 集群分区
网络分区症状	指所有的心跳中断。 (相关) 心跳线、心跳
节点	在集群系统中，指构成集群的服务器。在网络用语中，指可以传输、接收和处理信号的、包括计算机和路由器在内的设备。
心跳	指为了监视服务器而在服务器之间定期进行相互间的通信。 (相关) 心跳线、网络分区症状
公网	服务器 / 客户端之间的通信路径。 (相关) 心跳线、私网
失效切换	指由于查出故障，待机服务器继承运行服务器上的业务应用程序。
故障恢复	将某台服务器上已启动的业务应用程序通过失效切换交接给其他服务器后，再把业务返回到已启动业务应用程序的服务器。
失效切换组	执行业务所需的集群资源、属性的集合。
失效切换移动组	指用户故意将业务应用程序从运行服务器移动到待机服务器。
失效切换策略	可进行失效切换的服务器列表及其列表中具有失效切换优先顺序的属性。
私网	指仅连接构成集群的服务器的LAN。 (相关) 心跳线、公网
主服务器 (服务器)	失效切换组中的作为基准的主服务器。 (相关) 从服务器 (服务器)
浮动IP地址	发生了失效切换时，可忽视客户端的应用程序所连接服务器发生切换而使用的IP地址。 在与集群服务器所属的LAN相同的网络地址中，分配其他未使用的主机地址。
主服务器(Master Server)	Builder 的 [ExpressClusterproperties]-[Master Server] 中显示在最前面的服务器。
镜像磁盘连接	镜像磁盘型集群中用于进行数据镜像的LAN。可通过和内部主网的通用进行设定。
镜像磁盘系统	不使用共享磁盘的集群系统。 在服务器之间镜像服务器的本地磁盘。

附录 B 索引

2	
2字节系字符编码	72
B	
Builder	49, 55, 63, 74
C	
查出故障	13, 21
产品结构	29, 30
chkdsk命令	75
磁盘容量	51, 55, 56, 57
磁盘碎片整理	75
从网络分区状态的恢复	73
E	
ESMPRO/AC	69
ExpressCluster	29, 30
ExpressCluster Disk Agent服务	75
F	
服务器监视	32
G	
共享磁盘	67
共享磁盘需求	64
共享型镜像磁盘用分区	68
共享型磁盘需求	63
规格	50
故障监视	27, 31
H	
HA集群	14
画面更新间隔	71
恢复运行中的操作限制	72
I	
IP地址的交接	23
ipmiutil	70
J	
Java执行环境	55, 56, 57
兼容API	78
监视资源	43
脚本文件	73
镜像磁盘需求	63
镜像磁盘用分区	67
集群对象	41
集群关机	73
集群关机重启	73
集群系统	13, 14
集群资源的交接	13, 22
旧版本兼容功能	77
K	
可以查出的故障和无法查出的故障	33
可监视的故障和无法监视的故障	33
可执行文件	73
L	
历史文件	64
浏览器	55, 56, 57
N	
内部监视	33
内存容量	51, 55, 56, 57
NIC Link Up/Down监视资源	64
O	
OS	50, 55, 56, 57
OS启动时间	68
R	
软件	49, 50
软件配置	29, 31
S	
Server Core	70
失效切换	24, 29, 34
时钟同步	67
数据的交接	22
数据一致性	65
Single Point of Failure (SPOF)	13, 24
索引服务	76
T	
特定服务器的关机	73
特定服务器的关机重启	73
通信端口号	66
TUR	71

W

网络	68
网络分区解析资源	42
网络分区症状	22
WebManager	49, 56, 57, 63, 74
文件系统	65
write性能	64

X

心跳资源	41, 72
系统构成	13, 18

Y

延迟警告比例	71
业务监视	32
硬件	49, 50
硬件配置	36, 37, 39
应用程序的交接	23
运行OS	63
运行确认完毕应用程序信息	52

Z

资源	29, 41
最终运行	71
组资源	42, 71