CHIP@net

ТЕХНИЧЕСКАЯ РЕАЛИЗАЦИЯ КЛАСТЕРА НА ПЛАТФОРМЕ INTEL

Серверы объединяются в стаи

В последнее время огромный интерес вызывают кластерные технологии. Особенно много споров и сравнений вызывают решения от Microsoft с кодовым названием WolfPack и от Novell — Wolf mountain. Каким же образом на базе стандартных компьютеров Intel-архитектуры можно построить информационную систему, обладающую высоким уровнем масштабируемости и готовности?

Кластеры как совокупность серверов, решающих совместно определенные задачи, известны примерно с 1985 г. Наиболее известны система VAX под ОС VMS и кластеры Tandem.

В настоящее время большинство систем с ОС UNIX позволяют объединять серверы в кластер. Обычно понятие кластер ассоциируется с очень высокой ценой, что соответствовало действительности.

В последнее время (примерно со средины 1990-х годов), после того как ведущие разработчики операционных систем *Microsoft*, *Novell*, *SCO*, *Sun* анонсировали свои кластерные решения для серверов на платформе *Intel*, ситуация в корне изменилась. Цена кластера стала доступной даже для не очень крупных компаний и организаций. Это и обусловило всплеск интереса к кластерам, наблюдаемый в последнее время.

Что же такое «кластер»

Под кластером понимают совокупность серверов, накопителей и (по крайней мере, в перспективе) рабочих станций, которые *действуют*, *представляются пользователям* и *управляются* как единая система.

При этом обязательно соблюдение всех трех указанных в определении требований. Так, например, *Novell High Availability Server* — решение, созданное в рамках кластерной стратегии компании Novell. Два сервера в нем действуют и управляются как одна система, но пользователям представляются как два отдельных компьютера. Поэтому это решение не является кластером.

Чем же так привлекательны кластерные системы? Они позволяют значительно увеличить общую производительность сети (имеют хорошую *масштабируемость*), уменьшают затраты на администрирование локальной сети (хорошая *управляемость*). Но основное назначение кластера — это обеспечение высокой *доступности сетевых служб*. Даже при отказе одного из серверов кластера все обеспечиваемые кластером службы остаются в распоряжении пользователей.

Поясним функционирование кластерной системы на примере. При традиционном решении, когда в локальной сети находится, например, почтовый и Web-серверы, отказ одного из них приводит к тому, что соответствующая служба становится недоступной для пользователей. Если же в сети организован кластер, то каждый из узлов, до сбоя выполняющий только свою задачу, берет на себя дополнительно нагрузку и отказавшего сервера.

Как устроен кластер?

В современном понятии кластер на платформе Intel содержит несколько стандартных серверов *SHV* (Standard High Volume) и общую дисковую систему (Shared External Storage). Все серверы объединены внутренней (для кластера) локальной сетью *Cluster SAN* (System Area Network) на основе SCSI-, Ethernet-, FDDI-, Fibre Channel или других высокоскоростных интерфейсов. Кроме того, каждый сервер подключен к общей локальной сети, в которой находятся и все клиентские рабочие станции.

CHIP@net

Существует две модели функционирования кластера: с *разделяемыми дисками* и *без совместно используемых ресурсов*.

В первой модели каждый сервер имеет доступ к любому диску общей системы хранения данных. При одновременном запросе на чтение данных от двух серверов данные или читаются дважды, или запрашиваются у одного из серверов. При одновременном запросе на запись данных возникает конфликт. Для его разрешения операционная система кластера с разделяемыми дисками содержит обязательный элемент — *распределенный менеджер блокировок*.

Механизмы блокировок широко применяются во всех СУБД, однако в данном случае конфликты возникают между несколькими экземплярами приложения (например, баз данных), выполняемыми на различных серверах. Поэтому и механизм блокировок должен быть единым для всей системы.

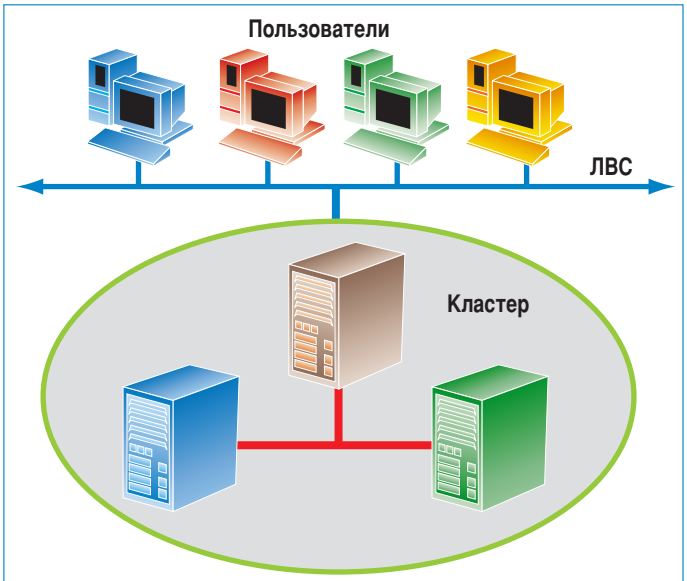
В модели *без совместно используемых ресурсов* в любой момент времени каждый сервер имеет доступ к данным только своей группы дисков. Конфликты как при чтении, так и при записи отсутствуют.

Модель с разделяемыми дисками применяется в тех случаях, когда главное предназначение кластера заключается в обеспечении высокой доступности и масштабируемости *одного приложения* (как правило, СУБД). Именно такая модель, в частности, применена в СУБД *Oracle Parallel Server*.

На различных серверах кластера располагаются не копии одной базы данных, а отдельные части общей базы данных. При необходимости добиться высокой производительности базы данных очень большого объема в систему добавляется еще один сервер со своей дисковой системой.

На дисковой системе вновь введенного сервера располагается часть общей базы данных. Однако пользователи базы данных по-прежнему видят только один сервер базы данных и одну базу данных. Добиться такого результата в модели без разделяемых ресурсов было бы очень трудно. Модель без разделяемых ресурсов применяется в кластерах, основное предназначение которых — обеспечение высокой доступности *нескольких* сетевых служб. Такая модель применяется в *Microsoft Cluster Server* (кодовое название *Wolfpack*). В нормальном режиме на каждом сервере, как правило, выполняется свой набор приложений. Любое приложение, выполняемое только на одном сервере, не требует доступа к данным другого сервера. Это позволяет применить модель без разделяемых ресурсов. Модель без разделяемых ресурсов в данном случае обеспечивает большую производительность, так как каждый сервер работает только со своими дисками.

Представление серверов в кластере для клиентов



Единая система: в кластере совокупность серверов и накопителей действуют, представляются пользователям и управляются как одно целое CHIP

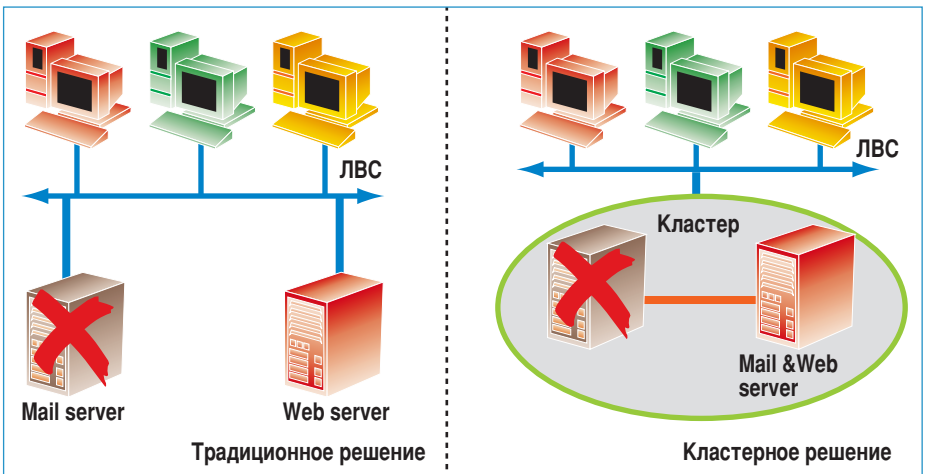
Особенности организации работы кластера

Рассмотрим одну из распространенных схем организации двухузлового кластера с разделяемой дисковой SCSI-системой, которая является своего рода классической для платформ Intel.

Основными элементами являются: 2 сервера (узлы кластера), каждый из которых имеет свой загрузочный диск и по два RAID-контроллера (или один 2-канальный); межузловое соединение (обычно Ethernet); разделяемая дисковая система, в одном корпусе которой размещается 2 независимых дисковых массива, на каждом из которых сформирован RAID (обычно 5 уровня); две SCSI-шины, к которым подключен один дисковый массив и по одному RAID-контроллеру каждого узла кластера.

Данная конфигурация поддерживается *Microsoft Cluster Server* (MSCS), входящим в состав MS Windows NT

Функционирование кластера



Один за всех: в кластерной системе каждый из узлов, до сбоя выполняющий только свою задачу, берет на себя дополнительно нагрузку и отказавшего сервера CHIP

Server Enterprise Edition. На кластере, выполненном по приведенной схеме, хорошо работают и другие решения. В частности, СУБД *Informix Dynamic Server 7.30.TC1* (под ОС Windows NT Server 4.0 Enterprise Edition и MSCS) и *Novell High Availability Server*. Как же организуется работа кластеров в такой схеме?

В нормальном режиме работы каждый сервер кластера конфигурируется для выполнения своих задач. Применительно к MSCS v.1.0, на узлах устанавливаются ресурсы (сетевые приложения, файлы данных, утилиты, обеспечивающие услуги сетевым клиентам). Связанные ресурсы объединяются в группы ресурсов. Однако каждый ресурс или группа ресурсов доступны для клиентов только на одном узле.

Каждый сервер использует только выделенную ему группу дисков. В процессе работы узлы обмениваются информацией о своем состоянии через межузловое соединение (*heartbeat messages*). В случае обнаружения неисправности на одном из узлов кластера, на исправный узел посылается соответствующее сообщение по межузловому соединению. Это является сигналом для исправного узла о запуске процедуры аварийного восстановления. На исправном узле запускаются все приложения, обеспечивающие ресурсы отказавшего сервера. Исправный сервер начинает использовать все диски: как выделенные для него, так и выделенные отказавшему серверу.

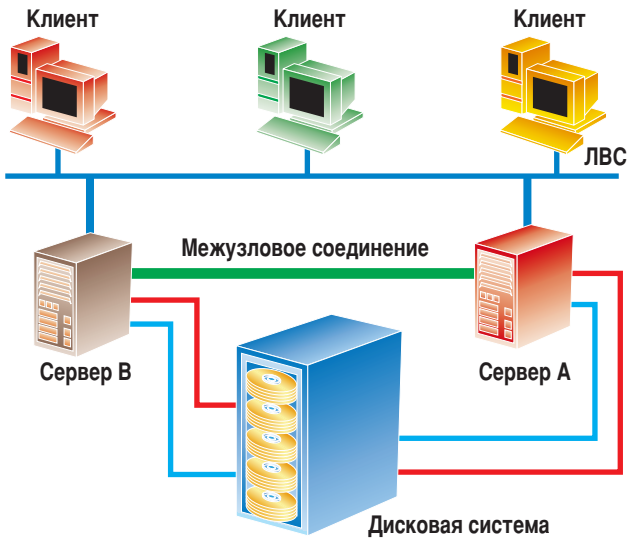
Процедура аварийного восстановления запускается также в случае полного выхода из строя одного из узлов после прекращения обмена сообщениями.

Более сложная ситуация возникает, если произошел отказ соединения между узлами. Оба узла исправны, но поскольку отсутствует межузловой обмен, оба считают, что второй узел отказал. У кластера наступает «раздвоение личности» — «split brain syndrome». Без принятия специальных мер на обоих узлах была бы запущена процедура аварийного восстановления.

Чтобы избежать данной ситуации, один из логических дисков разделяемого массива (так называемый «кворум-диск» — «quorum disk») выполняет особую роль. На этот диск, в частности, записывается журнал работы кластера. Кроме того, еще на этапе объединения серверов в кластер присоединяемый сервер проверяет, имеет ли данный диск владельца, и если нет, то пытается присоединить этот диск себе. После присоединения кворум-диска сервер начинает формировать кластер. Если же этот диск уже имеет владельца, то сервер присоединяется к кластеру.

В случае отказа межузлового соединения кворум-диск используется для предотвращения «раздвоения личности» кластера. Когда обмен по межузловому соединению прекращается, каждый узел после короткого ожидания пытается присоединить кворум-диск себе. Тот узел, который определит, что кворум-диск уже имеет владельца, будет считаться в дальнейшем отказавшим.

Двухузловой кластер с разделяемой дисковой системой



Техническая реализация: основными элементами являются: два сервера, каждый из которых имеет свой загрузочный диск и по два RAID-контроллера; межузловое соединение (обычно Ethernet); разделяемая дисковая система, в одном корпусе которой размещается 2 независимых дисковых массива, на каждом из которых сформирован RAID (обычно 5 уровня); две SCSI-шины, к которым подключен один дисковый массив и по одному RAID-контроллеру каждого узла кластера

Пристальный взгляд на вещи

Приведенная выше схема на первый взгляд кажется очень простой. Более того, ее еще можно упростить. Если установить по одному SCSI-контроллеру в каждый сервер, соединить их шлейфом, к которому подключить 2 диска, то, скорее всего, удастся установить на такую конструкцию Windows NT Server Enterprise Edition, включая MS Cluster Server. Однако в реальной жизни такой кластер «нежизнеспособен».

Ведь основной задачей кластера является обеспечение высокой доступности сетевых ресурсов. Из рассмотренного выше механизма обеспечения высокой доступности ясно, что кластер работает до тех пор, пока работоспособна его дисковая система. Статистика же неумолимо свидетельствует, что в подавляющем большинстве случаев (более 90 %) отказ сервера происходит из-за отказа жестких дисков, источников питания, системы вентиляции или контроллеров. Именно из этих элементов и состоит разделяемая дисковая система. Поэтому в конструкции кластера дисковой системе уделяется повышенное внимание.

Реализация дисковой подсистемы кластера

Диски в кластере, как правило, устанавливаются в отдельную стойку, имеющую встроенную избыточную систему электропитания (с 2 или даже 3 источниками питания с возможностью горячей замены) и вентиляции. Диски объединяются в отказоустойчивый массив, зачастую с дополнительным диском горячего резерва.

Эти меры позволяют сохранить определенное время работоспособность дисковой системы даже при выходе отдельных элементов из строя. Однако отказавший элемент требует как можно более быстрой замены для сохранения отказоустойчивости системы. Для этого дисковые стойки для кластеров оборудуются системой контроля состояния дисков, источников питания, вен-

CHIP@net

Подключение контроллеров с выделенным соединением

Кластер на уровне контроллеров: оба контроллера являются активными, т. е. имеют свой ID-адрес и обслуживают каждый свою группу дисков. Между контроллерами существует выделенное соединение. В случае отказа одного из контроллеров исправный начинает обслуживать оба ID-адреса и обе группы дисков CHIP

тиляторов и температуры, совместимой со спецификацией *SAF-TE*.

SAF-TE (SCSI Accessed Fault-Tolerant Enclosures) — это открытая спецификация, разработанная как стандартизованный метод мониторинга и предоставления всесторонней информации о состоянии элементов высокодоступных серверов и устройств хранения информации. Эта спецификация не зависит от контроллеров, кабелей и операционной системы, так как система представляется отдельным устройством на SCSI-шине. Данные о состоянии элементов собираются встроенной микропроцессорной системой и периодически (обычно каждые 10—20 сек) передаются по SCSI-шине. Благодаря этому значительно облегчается предупреждение администратора сети об отказе элементов и предоставляется возможность удаленного контроля состояния дисковой системы.

Не меньшую, чем при отказе дисковой системы, опасность для кластера представляет также отказ одного из RAID- или SCSI-контроллеров, установленных в узлах кластера. Как правило, при выходе из строя контроллера нарушится согласование SCSI-шины. В отдельных случаях некоторые линии шины могут оказаться замкнутыми накоротко. Это приведет к потере работоспособности общей шины, и исправный сервер не сможет нормально работать с дисками. Произойдет остановка обоих узлов кластера.

Чтобы этого не произошло, разделяемый дисковый массив подключается к контроллерам узлов через специальное устройство, осуществляющее изоляцию отказавшей части шины и динамическое согласование (терминирование) оставшейся ее части. Как правило, это устройство осуществляет также функции расширения шины (дает возможность увеличить общую длину шины).

Реализация RAID-контроллера для кластера

Не меньшие проблемы вызывает также вопрос, какой вариант лучше: RAID-контроллер в каждом узле кластера или SCSI-контроллеры в узлах и внешние SCSI-to-SCSI RAID-контроллеры в дисковой системе?

Конечно, вариант с внутренним RAID-контроллером, как правило, несколько дешевле и, к тому же, обеспечивает превосходную производительность (по крайней мере, при чтении). Скорость передачи данных ограничивается шиной PCI. В варианте же с внешним (SCSI-to-SCSI) RAID-контроллером скорость передачи данных ограничивается интерфейсом SCSI.

Тем не менее, в первом решении с внутренним RAID-контроллером кэширование при записи (*Write-back cashing*) обычно отключается, так как всегда существует опасность потери в момент отказа не сохраненных на диске данных. Отключение же кэширования записи значительно ухудшает производительность в приложениях, интенсивно использующих операции записи на диски (скорость записи снижается от 5 до 30 раз). Наиболее заметно ухудшение производительности при использовании RAID уровня 5. Резервная батарея питания не решает проблемы в кластерной конфигурации, так как данные, сохраненные в кэш, все равно не могут быть использованы другим узлом кластера.

Кроме проблем с кэшированием записи при использовании внутреннего контроллера значительно ограничивается также количество узлов кластера.

При использовании внешнего контроллера возникают другие проблемы. Фактически разделяемыми в кластере оказываются не только дисковая система, но и RAID-контроллеры. Выход из строя контроллера повлечет за собой потерю работоспособности кластера. Поэтому в кластере внешние RAID-контроллеры включаются в дуплексном режиме (один контроллер в горячем резерве). Это, безусловно, значительно повышает стоимость оборудования.

В современных SCSI-to-SCSI RAID-контроллерах, специально разработанных для использования в кластерах (например, фирмы Mylex), внедряется кластерная технология на уровне контроллеров. Оба контроллера являются активными, т. е. имеют свой ID-адрес и обслуживают каждый свою группу дисков. Между контроллерами существует выделенное соединение (*heartbeats* — «пульс»). В случае отказа одного из контроллеров исправный начинает обслуживать оба ID-адреса и обе группы дисков.

Процесс восстановления после отказа является прозрачным для узлов кластера и приводит только к некоторому снижению производительности.

Дальнейшим развитием этой технологии является *зеркалирование кэш-записи — Mirrored Write Caching*. По этой технологии данные, записываемые в кэш записи одного контроллера, немедленно копируются в буферную па-

Отсекаем лишнее: специальное устройство осуществляет изоляцию отказавшей части шины в кластере при выходе из строя RAID- или SCSI-контроллера

Максимальная надежность: диски в кластере, как правило, устанавливаются в отдельную стойку, имеющую встроенную избыточную систему электропитания и вентиляции, систему контроля состояния элементов. HDD объединяются в отказоустойчивый массив, зачастую с дополнительным диском горячего резерва

мять другого контроллера. Если в момент записи на диск один контроллер выйдет из строя, данные на диск будут записаны из буферной памяти другого контроллера. Описанные выше меры позволяют достичь высокой производительности и отказоустойчивости системы в целом.

Что нас ждет в ближайшем будущем?

Перспективы широкого внедрения кластеров следует искать в появлении приложений, специально разработанных для работы в таких системах. Основной предпосылкой этого является разработка кластерных технологий для платформы «Wintel», ведь на сегодня это один из самых дешевых способов построения отказоустойчивых и масштабируемых систем.

Некоторым препятствием могла быть трудность обеспечения обмена сообщениями между программными модулями приложения, исполняемыми на разных узлах. Такой обмен сообщениями требует передачи больших объемов данных между серверами (необходим скоростной канал — *high-bandwidth*) и, что не менее важно, быстрой передачи сообщений между узлами (необходима малая задержка — *low-latency*). Для программной реализации обмена сообщениями в кластере по инициативе Intel, Microsoft и Compaq разработана открытая спецификация, определяющая интерфейс высокоскоростного обмена между серверами и накопителями в пределах кластера *Virtual Interface Architecture* (VIA).

Объединение нескольких узлов в кластер с использованием технологии SCSI встретит серьезные трудности, что повлечет за собой более широкое внедрение технологии Fibre Channel. Однако это существенно только для очень мощных вычислительных систем с жесткими требованиями к надежности. Оборудование же для объединения в кластер 2—3 узлов с использованием SCSI-интерфейса будет применяться еще очень долгое время. Более того, стоимость этого оборудования будет непрерывно снижаться, благодаря чему может уже в следующем году кластеры будут применяться так же широко, как RAID-технология в серверах сегодня.

Вячеслав Овсянников,
ведущий специалист компании «ЕПОС»
slv@eposmail.kiev.ua