Лекция. **КОИУ повышенной надежности** (продолжение)

1. RAID-технологии
2. Реализация отказоустойчивых комплексов

1. RAID-технологии

Примером практической реализации концепции отказоустойчивости компьютерных систем является RAID-технология. RAID — это аббревиатура от англ. *Redundant Arrays of Inexpensive Disks*, что переводится как «избыточные массивы недорогих дисков». Термин RAID был впервые использован в статье «*A Case for Redundant Arrays of Inexpensive Disks* (RAID)», написанной сотрудниками Калифорнийского университета и опубликованной в 1987 году. В этой статье были описаны различные типы дисковых массивов, получивших сокращенное название «RAID-массивы». Основная идея RAID-массивов заключается в том, чтобы построить дисковый массив большой емкости, используя много недорогих дисков небольшого объема, с целью получения большую производительность, чем у одного дорогого диска значительной емкости, при меньшей стоимости. Такой массив дисков, будучи подключенным к компьютеру, распознается системой как один диск большой емкости. Отказоустойчивость дискового массива может быть увеличена за счет избыточности хранимой информации.

RAID-массив строится на основе распределения[1] данных между дисками. Пространство каждого диска разбивается на сегменты (в англоязычном варианте используется термин «*stripes*» — в дословном переводе «полосы»), размер которых может составлять от одного сектора (512 байт) до нескольких мегабайт; он зависит от типа операционной системы, если RAID-массив реализуется программными средствами или конструктивных особенностей RAID-контроллера в случае аппаратной реализации RAID-массива. Дисковое пространство RAID-массива есть объединение сегментов данных всех дисков массива.

RAID-массивы позволяют увеличить скорость доступа к данным. Поскольку дисковое пространство равномерно распределено между всеми дисками массива, можно одновременно читать и записывать данные на несколько дисков массива.

Отказоустойчивость RAID-массивов может быть увеличена либо за счет зеркалирования – создания дополнительной копии данных, либо за счет использования контрольных сумм для выявления и исправления ошибок и восстановления данных. В первом случае данные при записи тиражируются и разносятся по дискам: на каждом диске содержится по одной копии данных. Во втором случае при записи данных производится вычисление контрольной суммы. Алгоритм вычисления контрольной суммы, вообще говоря, меняется от одного типа RAID-массива к другому. Данные и контрольная сумма записываются на разные диски. В обоих случаях выход из строя одного (или более, что зависит от типа RAID-массива) диска не приведет к потере данных, а содержимое этого диска может быть восстановлено.

[1] Свободный перевод термина «*striping*». Дословно переводится как «деление на полосы».

Типы RAID-массивов

На сегодняшний день известно девять типов или уровней RAID-массивов, различающихся по скорости, надежности и стоимости изготовления: RAID 0,1,2,3,4,5,6,7,10,53. Наибольшее распространение получили массивы типов RAID 0, RAID 1 и RAID 5.

RAID 0 – *дисковый массив без дополнительной отказоустойчивости.*

RAID 0 – это пример распределения данных между дисками массива в «чистом» виде. Поток данных разбивается на блоки. Блоки последовательно записываются на диски (рис. 1).

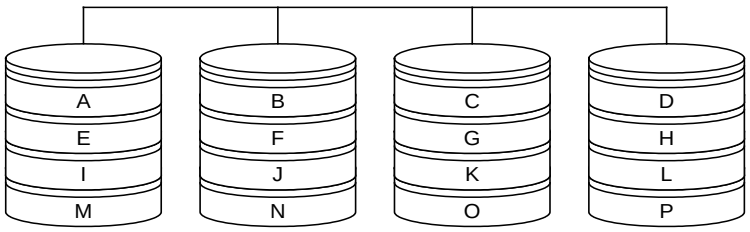


Рис. 1. Организация дискового массива RAID 0

***Достоинства* данного тип массива:**

- высокая производительность за счет распределения операций ввода/вывода между всеми дисками массива;
- отсутствие операций вычисления контрольных сумм, что также увеличивает производительность;
- простое конструктивное исполнение;
- хорошая промышленная технологичность.

Недостатком RAID 0 следует считать отсутствие средств реализации отказоустойчивости: выход из строя одного из дисков приводит к потере всех данных, хранящихся в дисковом массиве.

RAID 1 – *дисковый массив с зеркалированием данных.*

В RAID 1 блок данных записывается в двух экземплярах – каждый на свой диск. Для наилучшей производительности контроллер ввода/вывода должен поддерживать одновременное выполнение двух разных операций чтения и одной дуплексной операции записи для пары зеркалированных дисков (рис. 2).

Достоинства этого варианта организации RAID-массива:

- скорость записи та же, что и для случая использования одного диска;
- скорость чтения в два раза выше, чем для случая использования одного диска;
- высокая скорость восстановления данных из-за их 100% избыточности. Данные восстанавливаются простым копированием с одного диска на другой. При этом не расходуется время на дополнительные вычисления, как в случае с другими типами RAID-массивов;
- наиболее простая конструктивная реализация по сравнению со всеми другими типами RAID-массивов;
- это единственный RAID-массив, позволяющий получить отказоустойчивую дисковую подсистему на двух дисках.

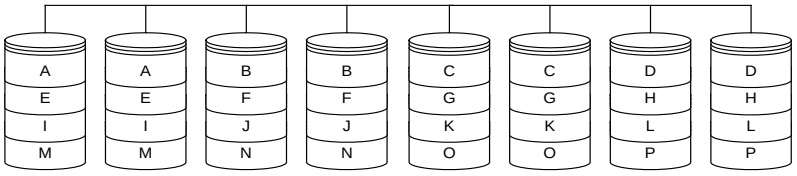


Рис. 2. Организация дискового массива RAID 1

Недостатком следует считать низкий коэффициент использования дискового пространства. Под коэффициентом использования дискового пространства понимается отношение объема полезных данных к суммарному объему дисков массива. В случае RAID 1 он равен 0,5.

RAID 5 – *дисковый массив с независимыми дисками данных и равномерным распределением контрольных сумм между дисками*.

Блоки данных последовательно записываются на диски (рис. 3). Контрольная сумма для блоков одного ряда вычисляется во время операции записи. Контрольные суммы размещаются последовательно по всем дискам. Проверка контрольной суммы производится во время операции чтения.

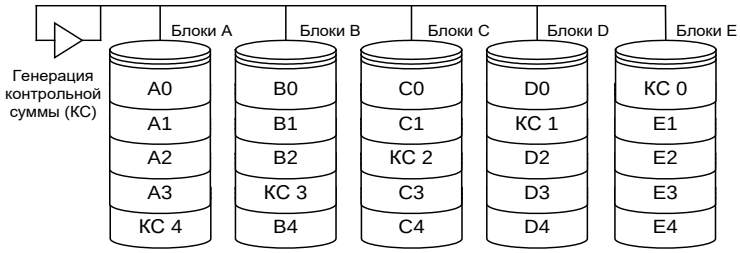


Рис. 3. Организация дискового массива RAID 5

Достоинства RAID 5:

- высокая скорость чтения и записи данных;
- высокий коэффициент использования дискового пространства.

Недостатки RAID 5:

- выход из строя одного из дисков оказывает заметное влияние на производительность;
- сложное конструктивное исполнение контроллера;
- сложный алгоритм восстановления данных в случае выхода из строя одного из дисков.

Массив RAID 0 наиболее быстрый и дешевый, но не обеспечивает отказоустойчивости. Его лучше всего использовать для дисковой подсистемы мощных рабочих станций, особенно при работе с приложениями по обработке графики, аудио- и видеоинформации и с системами CAD/CAM. Массив RAID 1 используется чаще всего в тех случаях, когда необходимо зарезервировать информацию, помещающуюся на одном диске. Он идеально подходит для серверов с небольшим объемом данных. Если же объем данных превышает емкость одного диска, то нужно выбирать между RAID 1 и RAID 5 исходя из следующих соображений:

- RAID 1 и RAID 5 имеют одинаковую отказоустойчивость;
- RAID 1 имеет большую скорость чтения и записи, чем RAID 5;

• RAID 1 быстрее восстанавливает диск в случае отказа одного из них, чем RAID 5;

RAID 1 дороже RAID 5 при одинаковом объеме полезных данных; другими словами, коэффициент использования дискового пространства у RAID 1 ниже, чем у RAID 5.

RAID-технологии могут быть реализованы как программно, так и аппаратно. Аппаратный метод RAID более быстродействующий и надежный, однако, его реализация обходится дороже. Некоторые операционные системы, такие, как Windows NT Server и Windows 2000 Server содержат встроенные средства поддержки программного RAID.

2. Реализация отказоустойчивых комплексов

На ранних стадиях работы над отказоустойчивыми вычислительными системами типичным подходом к обеспечению их высокой отказоустойчивости было использование избыточности в виде *n*-кратного резервирования отдельных подсистем и блоков. Это требовало значительного увеличения количества аппаратуры.

Решения Tandem Computer. Одна из первых отказоустойчивых вычислительных систем Tandem NonStop обладала оригинальной архитектурой, позволявшей решать задачу после отказа при ее работе в реальном масштабе времени.

Это стало возможным благодаря трем особенностям организации работы системы:

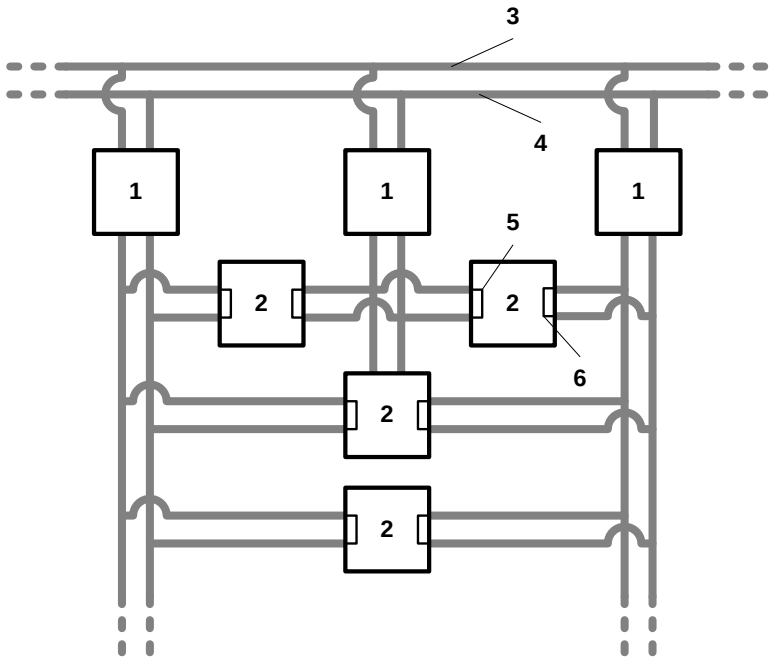
1) дублированию процессоров;

2) дублированию входов устройств управления вводом-выводом;

3) сохранению основной (базовой) операционной системы, многократно копируемой в памяти отдельных процессоров.

Система Tandem NonStop включает в себя от 2 до 16 микропроцессоров (МП), которые могут функционально заменять друг друга. Каждый из них имеет собственную память объемом 8 Мбайт и канал ввода-вывода. Эти процессоры соединены дублированной системой высокоскоростных параллельных шин Dynabus. Каждое устройство управления вводом-выводом имеет два канала связи и доступно для двух процессоров. Накопители на магнитных дисках (НМД) также имеют два канала связи. каждый из которых соединен с двумя устройствами управления. Таким образом, база данных доступна даже в том случае, когда и процессор, и устройство управления НМД отказали. При отказе НМД база данных может быть восстановлена, если на другом диске сохранились все необходимые данные. После отказа система автоматически восстанавливает две одинаковые копии рабочих файлов на двух независимых НМД. Затем восстанавливается остальная информация, необходимая для продолжения работы системы.

Система Tandem NonStop состоит из трех процессоров, четырех устройств управления вводом-выводом, имеющих по два канала связи, и параллельной дублированной системы шин (рис. 4).



1 – процессор;
2 – устройство управления вводом-выводом;
3, 4 – дублированная система шин;
5, 6 – каналы связи

Рис. 4. Структурная схема системы Tandem NonStop

Каждый процессор этой системы имеет свою собственную копию операционной системы Guardian, а также таблицу, где отражаются все ресурсы системы на текущий момент. Процессоры периодически (ежесекундно) оповещают друг друга о своем состоянии, передавая сообщения «я жив» – по шинам Dynabus. Если такие сообщения не поступают от каких-либо процессоров, то остальные изменяют свои таблицы ресурсов, исключая процессоры, не пославшие сообщения.

Основным механизмом восстановления в системе Tandem NonStop является использование контрольных точек. Каждый рабочий (основной) процесс имеет идентичный, но не активный процесс (процесс «поддержки») в другом процессоре. При нормальном функционировании основной процессор в очередной контрольной точке передает своему «дублеру» информацию о своем состоянии и состоянии вычислительного процесса, после чего продолжает выполнение задания до следующей контрольной точки. В случае отказа рабочего процессора его функции берет на себя процессор «поддержки». Он вместе с операционной системой подводит «итог» работы основного процесса (хода выполнения им задания) и продолжает вычисления, начиная с последней контрольной точки, в которой информация не была разрушена.

Несмотря на внешнюю простоту, реализация такого подхода – создание специального программного обеспечения – весьма сложна, поэтому в настоящее время он используется только для внутренних компонентов системы Tandem NonStop, т.е. для операционной системы. Для внешних компонентов (прикладных программ) применяется простая схема, по которой восстановление происходит путем возвращения к более ранней контрольной точке или даже путем повторного выполнения задания.

Процесс пользователя (прикладная программа) благодаря системным сообщениям изолирован от изменяющейся конфигурации системы. Если этому процессу требуются некоторые данные с диска, то он формирует сообщение, которое обрабатывается локальной копией операционной системы (ОС) Guardian, а она, в свою очередь, по таблице ресурсов определяет путь доступа к НМД. Процесс пользователя не решает, от какого из двух процессоров, соединенных с НМД, он получит ответ или какой из них будет играть роль «лидера». Это позволяет производить постепенное наращивание системы путем добавления процессоров.

Система Tandem NonStop занимает лидирующее положение среди ОУВС реального времени.

Решения Stratus Computer Inc. Одним из интересных принципов построения ОУВС является контроль дублированием, воплощенный в системе Stratus/32. В этой системе каждая основная функция выполняется четыре раза объединенными по входам четырьмя независимыми процессорами. Устройства сравнения генерируют сигнал ошибки в случае расхождения двух результатов в одной из двух пар. Пара процессоров с несовпадающими результатами отключается, а другая пара продолжает работу. При подключении отремонтированной пары процессоров необходима синхронизация с дублирующей парой.

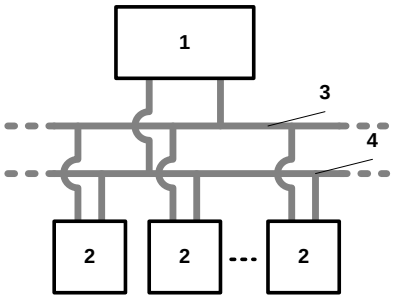
Такой подход – «спаривание пар» – имеет два преимущества:

1. Не требуется времени для автоматического восстановления после отказа, т.е. работа продолжается без задержки. Короткое прерывание необходимо только при подключении отремонтированного процессора для синхронизации его работы.

2. Поскольку не требуется восстановления информации, следовательно, нет необходимости использовать контрольные точки. Отметим, что в этой системе, как и в любой ОУВС, существуют внутренние средства диагностики и восстановления

Система Stratus/32 характеризуется довольно большой избыточностью. Ее основным недостатком является значительное усложнение базового устройства, которое в этой системе называется процессорным модулем (ПМ). Система может насчитывать до 32 МП (объемом памяти 8 Мбайт каждый), соединенных попарно кольцом локальной сети. Один процессорный модуль состоит из 4 МП типа Motorola 68000 и содержит свою копию ОС VOC. Три четверти ресурсов системы являются резервными и поэтому не повышают ее производительность.

Решения Synapse Computer Corp. Этот недостаток преодолен в системе Synapse N+1, основанную на МП типа Motorola 68000. Микропроцессоры взаимодействуют через дублированную параллельную высокоскоростную шину и имеют общую разделенную память. На рис. 5 изображена структурная схема этой системы, состоящая из общей памяти, в которой хранятся программы для всех системных заданий и заданий пользователя, а также список заданий (очередность работ), из универсальных процессоров и параллельной дублированной системы шин.



1 – общая память;
2 – процессор;
3, 4 – дублированная система шин

Рис. 5. Структурная схема системы Synapse N+1

Система содержит да 28 МП. Одни процессоры предназначены для операций ввода-вывода, другие – для обработки процессорных сообщений. Сами процессоры выступают в качестве диспетчеров, обращаясь к общей памяти и самостоятельно выбирая задания. Операционная система Syntesis хранится в общесистемной памяти объемом 16 Мбайт.

В название системы Synapse входит выражение N+1, так как добавление к N работающим процессорам еще одного обеспечивает такую же отказоустойчивость системы, как у некоторых ОУВС с 2N процессорами (т.е. в схемах, где каждый процессор «оглядывается на дублера»). При этом один из работоспособных процессоров продолжает выполнять задание отказавшего процессора, используя информацию, хранящуюся в общей памяти системы. Данный подход позволяет повышать производительность системы в такой же степени, в какой растет число ее процессоров (в отличие, например, от систем типа Stratus/32).

Однако общая разделенная память является «узким местом» в смысле отказоустойчивости. Существуют два способа, позволяющие преодолеть этот недостаток. Первый основан на отделении отказавшего модуля памяти и реализован в системе Synapse. Второй заключается в дублировании памяти системы и реализован в процессоре 3B20D Bell Laboratories Western Electric.

Решения Auragen Systems Corp. В System 4000 процессорные элементы соединены высокоскоростной дублированной параллельной шиной в так называемую «гроздь». Каждая «гроздь» объемом памяти 8 Мбайт состоит из трех МП типа Motorola 68000. Два из них выполняют задание пользователя, а третий реализует функции ОС. Каждая «гроздь» содержит свою копию операционной системы Unix System III.

В System 4000 использован принцип контрольных точек (точек синхронизации). После выполнения части задания происходит синхронизация: дублирующий процессор получает и сохраняет все сообщения, посланные основным процессором, а также запоминает пути следования этих сообщений с момента последней синхронизации. В случае сбоя или отказа «дублер» обрабатывает входные сообщения, блокируя выходные сигналы основного процессора до тех пор, пока он не начнет нормально функционировать. В каждой подсистеме («грозди») производится периодическое самотестирование. При его успешном завершении остальным подсистемам передается сигнал о работоспособности. Если такой сигнал не

вырабатывается или не поступает к другим подсистемам, то данная «гроздь» считается (остальными подсистемами) отказавшей.

Решения Tolerant Systems Inc. Фирма Tolerant Systems Inc. создала ОУВС Plus 32, в которой осуществляется синхронизация основного и дублирующего процессоров. Базовый элемент структуры состоит из двух микропроцессоров фирмы National Semiconductor Corp.'s NS 16000, связанных дублированной локальной сетью Ethernet. Каждый 'из микропроцессоров содержит свою копию операционной системы и имеет оперативную память до 4 Мбайт. Диагностирование отказов производится так же, как в системе Synapse N+1, – по превышению времени выполнения отдельных операций и по отсутствию сигнала о работоспособности.

Решения British Telecom. Фирмой разработана ОУВС Power 5/55, включающая до восьми микропроцессоров Motorola 68000 объемом памяти до 4 Мбайт. Она обеспечивает полную взаимосвязь между любой парой процессоров, имеющих свою копию ОС Perpos, и между всеми устройствами управления НМД. Это значительно уменьшает вероятность изоляции одного из компонентов (МП, НМД) или части системы. Так как информация, не влияющая на содержание базы данных, хранится на всех НМД, значительно сокращается время поиска и обработки ряда сообщений. Отказ процессора или НМД диагностируется по превышению времени выполнения операций. В этом случае любой из работоспособных МП и НМД благодаря их полной связанности может продолжить выполнение задания.

Наиболее общими чертами всех описанных систем являются:

1. Максимальная попарная связанность отдельных компонентов системы высокоскоростными дублированными шинами.

2. Наличие копии ОС в памяти каждого процессора или группы процессоров (подсистемы).

3. Возможность контроля работоспособности и восстановления системы при работе в реальном масштабе времени.

4. Использование универсальных, следовательно, и взаимозаменяемых микропроцессоров.

5. Разбиение основного процесса на ряд подпроцессов (по количеству МП или подсистем), в каждом из которых используется принцип дублирования.

Получили также распространение дублированные и троированные вычислительные системы.

Кластеризация серверов

Основой всех методов защиты данных от аварий является *избыточность.* При резервировании электропитания используются избыточные источники электроэнергии – устройство бесперебойного питания наряду с электросетью. Резервное копирование данных предполагает создание избыточных копий ценных файлов на дополнительных носителях. В системах отказоустойчивых дисков данные записываются на избыточных дисках. Крайней степенью избыточности является кластеризация. Предпосылки для ее реализации создает RAID-технология.

В общем случае *кластеризация* означает группирование серверов в *кластеры.* В сети кластер серверов виден пользователям как один сервер. Если один сервер выходит из строя, его обязанности выполняет другой сервер кластера. Пользователи не замечают этот переход. Кластер серверов показан на рис. 66.

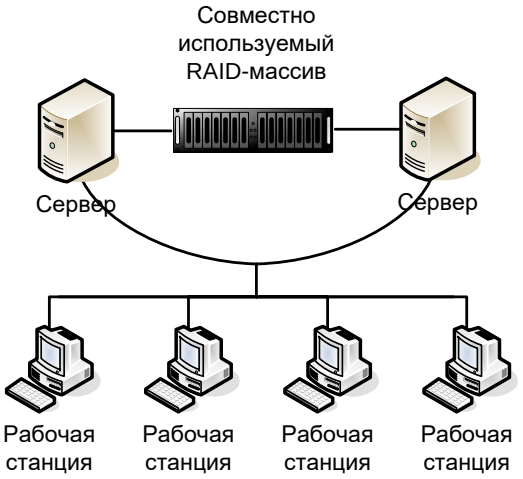


Рис. 6. Объединение серверов в кластер

Средства поддержки кластеризации встроены в такие операционные системы, как, например, Windows 2000 Advanced Server. Существует также программное обеспечение кластеризации, предназначенное для конфигурирования серверов в других операционных системах. Кроме повышенной отказоустойчивости, применение кластеризации имеет и другие преимущества, например, выравнивание загрузки серверов.

Базовая модель VAX/VMS кластеров. Компания DEC первой представила концепцию кластерной системы, определив ее как группу объединенных между собой вычислительных машин, представляющих собой единый узел обработки информации. По существу, VAX-кластер представляет собой слабосвязанную многомашинную систему с общей внешней памятью, обеспечивающую единый механизм управления и администрирования.

VAX-кластер обладает следующими свойствами:

Разделение ресурсов. Компьютеры VAX в кластере могут разделять доступ к общим ленточным и дисковым накопителям. Все компьютеры VAX в кластере могут обращаться к отдельным файлам данных как к локальным.

Высокая готовность. Если происходит отказ одного из VAX-компьютеров, задания его пользователей автоматически могут быть перенесены на другой компьютер кластера. Если в системе имеется несколько контроллеров накопителей и один из них отказывает, другие контроллеры автоматически подхватывают его работу.

Высокая пропускная способность. Имеется возможность параллельного выполнения заданий на нескольких компьютерах кластера.

Удобство обслуживания системы. Общие базы данных могут обслуживаться с единственного рабочего места. Прикладные программы могут инсталлироваться только однажды на общих дисках кластера и разделяться между всеми компьютерами кластера.

Расширяемость. Увеличение вычислительной мощности кластера достигается подключением к нему дополнительных VAX-компьютеров. Дополнительные накопители на магнитных дисках и магнитных лентах становятся доступными для всех компьютеров, входящих в кластер.

Работа VAX-кластера определяется двумя главными компонентами. Первым компонентом является высокоскоростной механизм связи, а вторым – системное программное обеспечение, которое обеспечивает клиентам прозрачный доступ к системному сервису. Физически связи внутри кластера реализуются с помощью трех различных шинных технологий с различными характеристиками производительности.

Основные методы связи в VAX-кластере представлены на рис. 7.

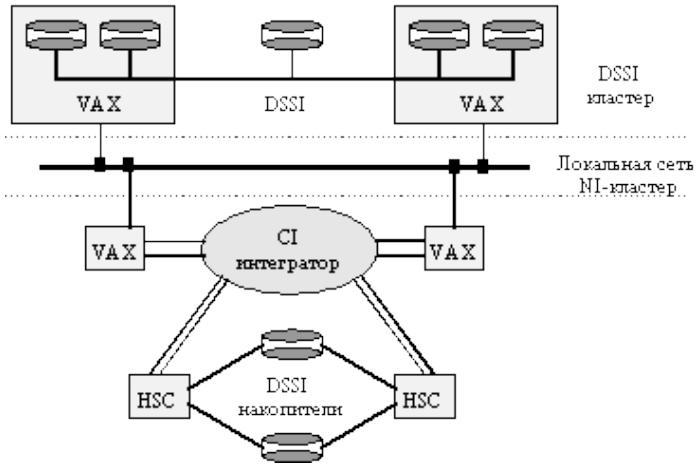


Рис. 7. VAX/VMS-кластер

Шина связи компьютеров CI (Computer Interconnect) работает со скоростью 70 Мбит/с и используется для соединения компьютеров VAX и контроллеров HSC с помощью коммутатора Star Coupler. Каждая связь CI имеет двойные избыточные линии, две для передачи и две для приема, используя базовую технологию CSMA, которая для устранения коллизий использует специфические для данного узла задержки. Максимальная длина связи CI составляет 45 метров. Звездообразный коммутатор Star Coupler может поддерживать подключение до 32 шин CI, каждая из которых предназначена для подсоединения компьютера VAX или контроллера HSC. Контроллер HSC представляет собой интеллектуальное устройство, которое управляет работой дисковых и ленточных накопителей.

Компьютеры VAX могут объединяться в кластер также посредством локальной сети Ethernet, используя NI - Network Interconnect (так называемые локальные VAX-кластеры), однако производительность таких систем сравнительно низкая из-за необходимости делить пропускную способность сети Ethernet между компьютерами кластера и другими клиентами сети.

Кластеры Sequent Computer Systems. Компания Sequent выпускает UNIX-кластеры баз данных. Она предлагает решения, соответствующие среднему и высокому уровню готовности своих систем. Первоначально Sequent Hi-Av Systems обеспечивали дублирование систем, которые разделяли общие диски. Пользователи могли выбирать ручной или автоматический режим переключения на резерв в случае отказа. Программный продукт ptx/CLASTERS, который может использоваться совместно с продуктом Hi-Av Systems, включает ядро, отказоустойчивый распределенный менеджер блокировок, обеспечивающий разделение данных между приложениями. Продукт ptx/NQS предназначен для балансировки пакетных заданий между узлами кластера, а ptx/LAT расширяет возможности управления поль-

зовательскими приложениями в режиме on-line. Продукт ptx/ARGUS обеспечивает централизованное управление узлами кластера, а ptx/SVM (распределенный менеджер томов) представляет собой инструментальное средство управления внешней памятью системы. Hi-Av Systems обеспечивает также горячее резервирование IP адресов и позволяет кластеру, в состав которого входят до четырех узлов, иметь единственный сетевой адрес (рис. 8).

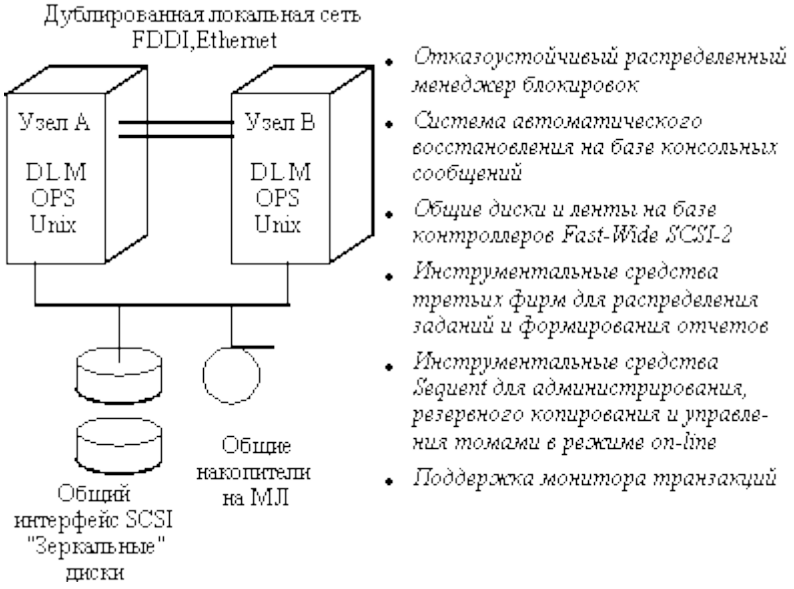


Рис. 8. Архитектура двухмашинного кластера SE90 компании Sequent

Компания Sequent одной из первых освоила технологию Fast-Wide SCSI, что позволило ей добиться значительного увеличения производительности систем при обработке транзакций. Компания поддерживает дисковые подсистемы RAID уровней 1, 3 и 5. Кроме того она предлагает в качестве разделяемого ресурса ленточные накопители SCSI. Модель SE90 поддерживает кластеры, в состав которых могут входить два, три или четыре узла, представляющих собой многопроцессорные системы Symmetry 2000 или Symmetry 5000 в любой комбинации. Это достаточно мощные системы. Например, Sequent Symmetry 5000 Series 790 может иметь от 2 до 30 процессоров Pentium 66 МГц, оперативную память емкостью до 2 Гбайт и дисковую память емкостью до 840 Гбайт.

При работе с Oracle Parallel Server все узлы кластера работают с единственной копией базы данных, расположенной на общих разделяемых дисках.

Кластеры Marathon Technologies. Для решения проблем отказоустойчивости Marathon Technologies использует специальную конфигурацию, которая позволяет сравнивать результаты вычислений с нормальным выполнением. Этот процесс прозрачен как для операционной системы, так и для приложения. Система работает, выполняя одно и то же приложение и транзакции на двух идентичных синхронно работающих серверах одновременно. Если один из серверов остановится из-за неисправности или физической угрозы, другой будет продолжать работать без простоев, не теряя времени на восстановление, и без пропусков транзакций.

Кроме того, каждый из двух серверов физически и логически разделяется на два, выполняющих два основных типа операций – это манипулирование данными

с их преобразованием (обработка данных) и перемещение данных на системы хранения и обратно по сетям и через иные устройства ввода-вывода.

Как показано на рис. 2.13, функция обработки данных выполняется на вычислительных элементах (Compute Element – CE), а функция ввода-вывода - на элементах процессоров ввода-вывода (I/O Processor - IOP). Два сервера соединены при помощи высокоскоростных интерфейсных плат PCI, называемых Marathon Interface Cards (MIC), и волоконно-оптического кабеля. Элементы MIC пересылают данные и получают их от двух систем одновременно. MIC также обеспечивает логику сравнения и тестирования, которая следит за идентичностью результатов от двух систем.

Каждый сервер представляет собой законченное устройство, с серверной версией ОС Windows, выполняющейся как на CE, так и на IOP. Все запросы на операции ввода-вывода от CE перенаправляются для обработки на IOP. На IOP выполняется ПО Marathon в виде приложения, контролирующего процесс управления всеми неисправностями, зеркалированием дисков, системным управлением и задачами ресинхронизации.

Соединив два идентичных сервера, как показано на рис. 9, легко получить конфигурацию системы Marathon Assured Availability. На двух CE-элементах этой конфигурации работает патентованная технология синхронизации компании Marathon, а также операционная система и приложения.

Система обеспечивает функциональность RAID-массива уровня 1 без специального контроллера RAID путем репликации операций записи на диски на каждом элементе IOP (зеркалирование дисков). В случае отказа одного из CE другой сервер продолжает работать, причем пауза будет некритичной и составит всего несколько мс (они требуются системе для удаления неисправного сервера из конфигурации).

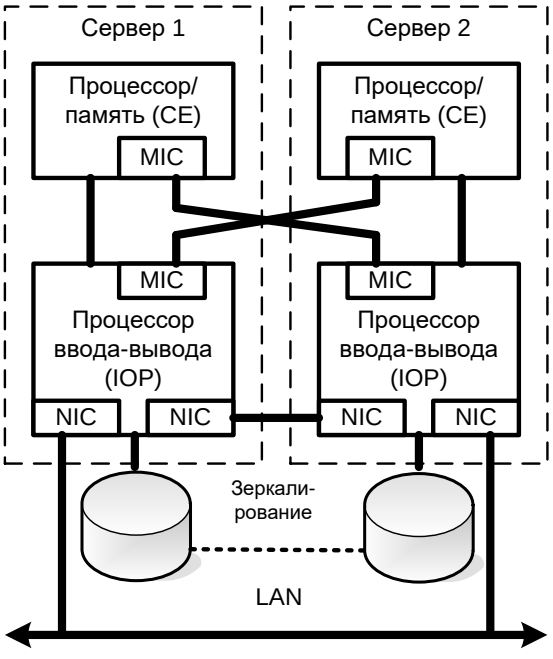


Рис. 9. Архитектура Marathon Assured Availability.

Затем неисправный СЕ можно физически удалить, отремонтировать и вновь включить в систему. Далее через высокоскоростное соединение СЕ автоматически возвращается в конфигурацию путем переноса данных и ресинхронизации статуса работающего СЕ и восстановленного сервера. Статус операционной системы и приложений устанавливается в течение нескольких секунд, требуемых на ресинхронизацию двух СЕ, не оказывая воздействия на работу пользователей.

Этот подход в корне отличается от ресинхронизации кластера. То же относится и к IOP – при отказе одного из них другой продолжает обслуживать систему. Неисправный IOP физически удаляется из системы, ремонтируется и возвращается в систему. После начала работы программного обеспечения Marathon восстановленный процессор IOP автоматически включается в конфигурацию. Повторное зеркалирование дисков выполняется в фоновом режиме через сеть Ethernet, соединяющую два процессора IOP. Отказ одного из зеркалированных дисков обрабатывается при помощи того же процесса.

Сетевые соединения системы также полностью избыточны. Сетевые соединения от каждого процессора IOP реагируют на тот же самый MAC-адрес, при этом только одному процессору разрешается передавать сообщения, но принимают их оба. Таким образом, одно сетевое соединение осуществляет мониторинг другого по частной сети Ethernet.

Если какое-либо сетевое соединение перестает работать, это обнаруживается IOP, и оставшееся соединение будет обрабатывать рабочую нагрузку. Чтобы инициировать ремонт, система отправляет уведомление администратору.

На рис. 9 оба соединения принадлежат единому сетевому сегменту, но это необязательно. Сетевое соединение каждого процессора IOP может принадлежать разным сегментам одной сети. Система также способна работать с несколькими сетями, каждая со своими собственными избыточными соединениями. Чтобы система Marathon удовлетворяла дополнительным требованиям устойчивости к катастрофам, необходимо лишь оптоволоконное соединение. Поскольку СЕ остаются синхронизированными даже на таком расстоянии, отказ компонента или всего локального комплекса проходит незамеченным пользователями.

Помимо перечисленных примеров кластеризации имеются и другие решения, такие как кластеры Alpha/OSF компании DEC, UNIX-кластеры компании IBM, кластеры AT&T GIS, Sun Microsystems, Data General и др