

# Jet Info

ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ

№ 5 (156)/2006

## Распределенные центры обработки данных

ВЫЧИСЛИТЕЛЬНЫЕ  
КОМПЛЕКСЫ

# Распределенные центры обработки данных

Голубев Денис Леонидович,  
руководитель экспертной группы  
отдела вычислительных комплексов

## СОДЕРЖАНИЕ

---

Введение .....	3
Резервные центры как устаревший метод .....	3
Современные технологии изменяют схемы резервирования .....	7
Ближайшие изменения .....	10
Использование распределенных центров в IT-системах большого масштаба .....	10
<i>Врезка 1: CWDM</i> .....	13
<i>Врезка 2: Infiniband</i> .....	14

---

## Введение

Резервный вычислительный центр (РВЦ) — это одно из решений, направленных на обеспечение доступности данных и информационных служб в целом. РВЦ гарантирует непрерывность работы ИТ-инфраструктуры в случае выхода из строя основного вычислительного центра, когда время его восстановления превышает допустимое время недоступности информационной системы.

Схема резервных центров применяется в ИТ-отрасли уже много лет для защиты критических приложений от разного рода серьезных аварий и катастроф. Предполагается, что резервный центр позволит восстановить работу основных ИТ-сервисов даже при полном отказе основного центра обработки данных в случае наводнения, пожара, отключения электропитания и т.д.

Все резервные центры организованы похожим образом — выбирается площадка, расположенная на безопасном расстоянии от основного центра, и на ней устанавливается оборудование, необходимое для работы основных приложений и ИТ-сервисов. Основной и резервный центры соединяются каналом связи для передачи данных приложений.

Схема миграции данных представляет собой репликацию данных приложений из основного центра в резервный или выполнение резервного копирования из основного центра в резервный. Такие схемы выбирались из-за ограничения пропускной способности каналов связи (в том числе и по причине высокой стоимости последних).

Когда происходит авария в основном центре, в резервном производится создание актуальных данных приложений (восстановление из резервных копий или перевод в рабочий режим имеющихся реплик) и запуск приложений на ресурсах резервного центра.

После запуска приложений в резервном центре пользователи перемещаются на рабочие места, имеющие доступ к резервному центру, и их деятельность возобновляется.

Концепция резервных центров была разработана в конце 80-х — начале 90-х годов прошлого века и с тех пор значительно не изменялась. Но если строить ИТ-инфраструктуру предприятия сейчас, нужно ли следовать концепции, раз-

работанной более пятнадцати лет назад и базирующейся на тогдашнем уровне технологий?

В предлагаемой статье автор представляет свой взгляд на то, какими должны быть современные решения по обеспечению отказоустойчивости ИТ-инфраструктуры, а также дает прогноз достижений в этой области на ближайшие годы.

## Резервные центры как устаревший метод

Современные резервные центры, как метод обеспечения непрерывности производственных и других процессов, не являются продуктом ИТ. Они известны довольно давно, еще со времени холодной войны, и являются в определенной степени копией армейских центров управления операциями на случай войны «горячей». (К счастью, подобные сооружения так и не пригодились военным.) Армейские резервные центры необходимо было поддерживать в постоянной готовности, оборудование должно было полностью соответствовать основным центрам. Узнать наверняка, насколько они были готовы к выполнению своих задач, можно было бы только в случае глобальной войны, вполне возможно, что заявленные функции не выполнялись полностью.

Резервные центры ИТ-систем включаются в работу в случае менее серьезных происшествий, но вопросы готовности и соответствия заявленным функциям также имеют ключевое значение.

При оценке применимости РЦ, кроме прочего, следует учитывать конкретные политические и экономические особенности нашей страны. Резервные центры ИТ-систем — дорогие конструкции, и применялись они изначально транснациональными корпорациями, которые ведут свой бизнес на нескольких континентах. В случае возникновения проблем в Европе такие



Рис. 1. Битва за Москву. Сентябрь – декабрь 1941

компании могут руководить своими операциями из Азии или Америки. В Соединенных Штатах Америки управление компанией может быть перенесено с западного побережья на восточное или наоборот. В нашей же стране государственное управление традиционно централизовано, и в последние годы наблюдается усиление этой централизации.

Трудно представить, каким образом можно было бы перенести управление компанией в Санкт-Петербург в случае какой-либо катастрофы в Москве. Нам пришлось бы решать более серьезные задачи, чем восстановление IT-струк-

тур. Сейчас Москва — не только столица, это еще и крупнейший транспортный узел, и узел связи. В 1812 году М.И. Кутузов, оставляя город Наполеону, мог говорить: «С потерей Москвы не потеряна Россия», — поскольку в ту пору она (Москва) была провинциальным городом. Но спустя более столетия, во время другой Отечественной войны, в 1941 году Москву упорно защищали. И это было необходимо не только потому, что столица государства — символ, и необходимость отстоять этот город от вражеской оккупации имела огромное политическое и даже психологическое значение. Дело в том еще, что при

Табл. 1. Изменение технологий связи

Технология	1990-е годы	2006 год
Сеть	<ul style="list-style-type: none"> <li>• OC-3 (155 Mbit)</li> <li>• OC-12 (622 Mbit)</li> </ul>	<ul style="list-style-type: none"> <li>• 10 Gigabit Ethernet</li> </ul>
Ввод-вывод	<ul style="list-style-type: none"> <li>• SCSI, ESCON (200 Mbit)</li> </ul>	<ul style="list-style-type: none"> <li>• 4 Gigabit Fibre Channel</li> <li>• 10 Gigabit Fibre Channel</li> </ul>
Линии связи	<ul style="list-style-type: none"> <li>• 1 пара оптоволокна – 1 линия связи</li> </ul>	<ul style="list-style-type: none"> <li>• CWDM – 8 линий связи в одной паре оптоволокна</li> <li>• DWDM – до 32 линий связи в одной паре оптоволокна</li> </ul>

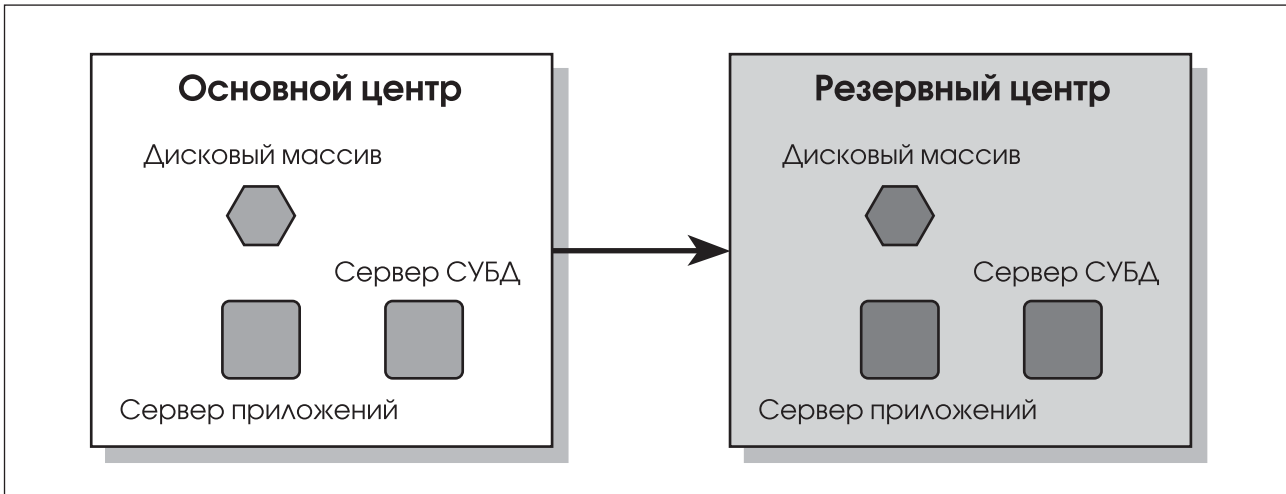


Рис. 2. Схема резервного центра

потере Москвы как транспортного узла, фронт разваливался на две не связанные между собой части, а это грозило катастрофическими последствиями. С тех пор значение Москвы как крупнейшего узла транспорта, связи и управления (сравните: в Москве четыре аэропорта, в Санкт-Петербурге – один) только возросло.

Подобная же централизация управления, транспорта и связи наблюдается и в российских регионах. Ни один город областного, республиканского или краевого подчинения не сравнится по возможностям со своим административным центром. В таких условиях резервный центр, построенный в другом городе, не гарантирует непрерывность процессов, т.к. в чрезвычайной ситуации, прежде чем восстанавливать ИТ-систему, придется решать проблемы восстановления страны или региона. Эффективнее и дешевле разместить площадки в разных районах города и подключить их к разным подстанциям,

городским узлам связи. (Вспомним, что даже во время отказа электросети, произошедшего в Москве в мае 2005 года, в северных районах электроснабжение сохранилось, а на юге всего московского региона и в соседних областях оно было прервано).

Кроме того, при размещении площадок внутри города потребуются оптоволоконные коммуникации протяженностью максимум 35 – 40 км, поэтому гораздо эффективнее было бы использовать сами площадки, чем создавать на одной из них резервный центр.

Со времени появления идеи схемы резервных центров технологии связи претерпели серьезные метаморфозы, которые существенно расширили возможности связи. (Табл. 1). Теперь каналы связи существенно более производительны, а если нет необходимости разносить площадки на значительные расстояния, когда задержка распространения света в оптоволо-

<sup>1</sup> Схема резервного центра предполагает, что:

- существует резервная площадка, не используемая в производственной деятельности;
- на резервной площадке установлено выделенное оборудование, находящееся в «холодном» или «теплом» резерве;
- данные из основного центра переносятся при помощи методов репликации или резервного копирования

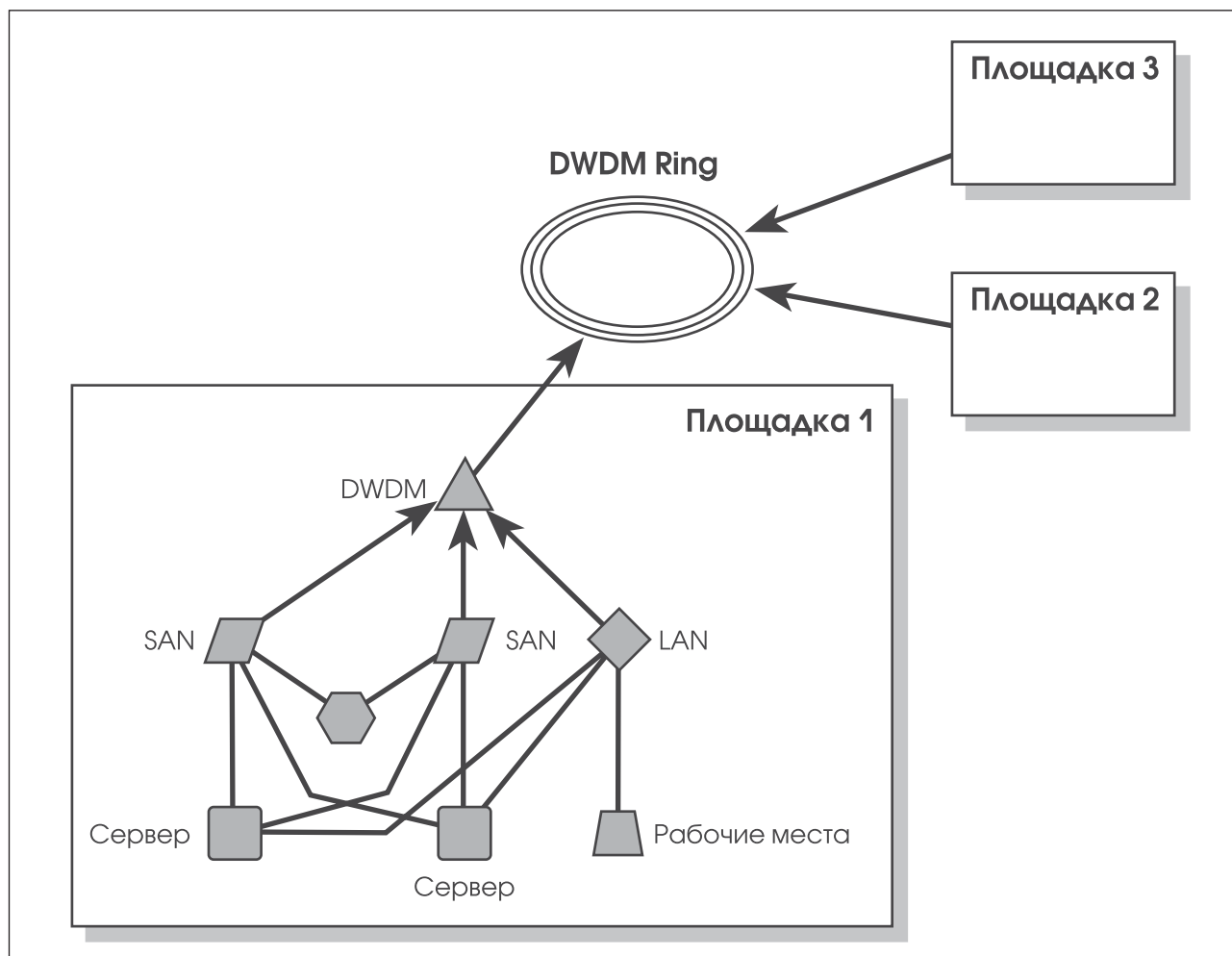


Рис. 3. Распределенный вычислительный центр

конном кабеле (~ 5 с/км) оказывает влияние на время выполнения операции ввода-вывода (~ 5–10 мс), то разумно ли организовывать связь площадок по схеме резервного центра<sup>1</sup>?

Эксплуатация резервного центра создает дополнительные проблемы, которые нужно решать. Так еще на стадии его построения необходимо обеспечивать **подходящее** резервирование инфраструктуры для работы приложений. На каждый установленный в основном центре сервер надо отвечать сервером, который сможет обеспечить работу определенного приложения, в резервном центре. Каждому гигабайту полезной емкости в основном центре должен соответствовать гигабайт в резервном. Можно устанавливать в РЦ менее производительные и более дешевые дисковые массивы, но все необходимые ресурсы для запуска приложений

должны присутствовать. Все это оборудование не используется и стоит в резерве.

Основные центры не могут оставаться неизменными, они должны развиваться — появляются новые приложения, увеличивается объем существующих данных и т.д. Все изменения, происходящие в основном центре обработки данных, нужно отражать в резервном, чтобы он полностью соответствовал решаемым задачам, иначе этот резервный центр перестанет быть таковым.

Сопровождение резервного центра предполагает решение следующих задач:

- Обеспечение резервирования новых приложений — установка нового оборудования и программного обеспечения в резервном центре.
- Когда для работы приложения в основном центре приходится устанавливать более

<sup>1</sup> См. сноску на стр. 5

производительный сервер, дисковый массив или увеличивать пропускную способность сети, то такие же меры проводятся и в резервном центре.

- Поддержание одинаковых версий программного обеспечения в основном и резервном центре — после изменения версии приложения или СУБД либо операционной системы следует провести изменения и в резервном центре также. Обновление версий происходит относительно редко, а установка программных коррекций (patches) и изменение конфигурации производятся часто; если не произвести изменение в резервном центре, то приложение не сможет быть там запущено.
- Изменение схем резервирования, чтобы они работали с меняющимися приложениями. Например, возможна ситуация, когда данные передаются в резервный центр при помощи синхронной репликации средствами дисковых массивов, и в результате возросшей нагрузки выбранная схема перестает справляться с работой и начинает замедлять работу приложения. Очевидно, требуется переработка схемы резервирования и выбор другого решения.
- Обязательное тестирование всех изменений, чтобы быть уверенным в работоспособности резервирования. Тестирование — это перевод приложений в резервный центр и проверка его работоспособности — процедура трудоемкая и, может быть, даже опасная, в случае сбоя во время этой операции восстановление штатного режима может занять много времени.
- Поддержание в актуальном состоянии большого объема документации по резервному центру и процедуре его запуска. Вполне вероятно, что строить резервный центр будут одни люди, вносить изменения — другие, а воспользоваться им придется уже третьим. К моменту, когда потребуются запускать резервирование, создатели этого РЦ могут работать в другой компании или даже в другой стране, и возможности получить их помощь уже не будет. Отсутствие актуальной документации резервного центра может превратить процедуру перехода в увлекательную

игру «русская рулетка» с неизвестным числом патронов в барабане револьвера.

Кроме перечисленных работ, для поддержания резервного центра в актуальном состоянии необходимы и определенные организационные мероприятия, освещение которых требует отдельной публикации.

## Современные технологии изменяют схемы резервирования

Проблемы резервных центров очевидны и даже решаемы, если уметь вовремя выявлять их, иметь ресурсы для их решения и прилагать адекватные усилия. И если проблемы устранены или их острота значительно ослаблена, то организовать центр обработки данных можно иначе, чем принято сейчас. Уровень современных технологий открывает новые возможности для этого.

Площадки соединяются производительными каналами связи с пропускной способностью в десятки гигабит (см Табл. 1). Если площадки находятся в одном городе, и расстояние между ними не превышает 50 км, то из нескольких площадок можно создать единый вычислительный центр, рассматривая их просто как разные комнаты в одном здании.

На Рис. 3 представлена примерная схема распределенного вычислительного центра. Локальные сети (LAN) и сети хранения данных (SAN) всех площадок связаны между собой. Сегменты локальных сетей, к которым подключены серверы, объединены в домены второго уровня модели OSI, и это позволяет прозрачно для приложений перемещать IP-адреса серверов с площадки на площадку. Благодаря объединению SAN всех площадок, серверы могут использовать ресурсы хранения данных на любой из них.

Площадки объединены при помощи технологии DWDM. Для надежности соединения используется схема «кольцо» (DWDM Ring). Оборудование DWDM обеспечивает защиту каналов связи на физическом уровне — импульсы света могут передаваться между двумя точками на кольце по двум разным маршрутам (условно назовем их «короткий», когда расстояние между точками на кольце минимально, и «длинный»). В случае разрыва кольца между двумя площадками «короткий» маршрут становится недоступным, но световые импульсы продолжают передаваться по «длинному» маршруту и связь между площадками не теряется.

При правильной конфигурации оборудования локальной сети и сети передачи данных<sup>2</sup> разрыв кольца и изменение маршрута передачи световых импульсов между площадками не приводят к потере информации и происходят прозрачно для оборудования SAN и LAN.

Прозрачный доступ между площадками позволяет использовать вместо схемы «Резервный центр» более простые способы локального резервирования: серверы, резервирующие друг друга, могут находиться на разных площадках, а данные располагаются на «зеркале» из двух дисковых массивов, также расположенных на разных площадках.

Локальные схемы резервирования предусматривают гораздо больше вариантов, чем схема резервного центра, такие как:

- Параллельные кластеры для самых критичных приложений (например, Oracle Real Application Cluster для экземпляров СУБД с наивысшими требованиями к готовности).
- Кластеры приложений для программного обеспечения, имеющего собственные механизмы резервирования. В качестве примера назовем кластеры таких приложений, как монитор транзакций Tuxedo, сервер приложений Websphere AS, сервис передачи сообщений Websphere MQ.
- Кластеры высокой готовности (HA-cluster) для критичных приложений (сейчас многие

производители серверов и поставщики программных систем предлагают кластерные решения для популярных приложений). Для приложений собственной разработки можно разработать самому или заказать специализированный модуль для выбранного варианта кластера.

- «Теплый» или «холодный» резерв для остальных систем.

Очевидные преимущества предложенного подхода — гибкость и упрощение процедуры резервирования. Снижается время готовности приложений, поскольку перевод приложения между узлами кластера или переключение на другой узел Oracle RAC занимает гораздо меньше времени, чем переход на приложения резервного центра. Не только упрощается процедура перехода на резервный узел, но и уменьшается объем документации, которую нужно поддерживать в актуальном состоянии.

В решении производственных задач участвует оборудование всех площадок. Можно использовать все территории для размещения оборудования. Обычно после замены производственного сервера в основном центре планируют перемещение предыдущей версии оборудования в резервный центр, для обеспечения там необходимой производительности. В случае же локального резервирования достаточно подключить новое оборудование в нужном месте и обеспечить доступ к ресурсам (сеть, ресурсы хранения данных).

При отсутствии разделения оборудования на основной и резервный центры упрощаются процедуры управления ресурсами, поскольку обеспечивается доступ ко всем ресурсам из одной точки, где расположены рабочие места обслуживающего персонала.

Объединение площадок при помощи DWDM — это достаточно дорогое решение, строить распределенные центры обработки данных таким образом могут только большие компании с соответствующими ИТ-бюджетами. Ис-

<sup>2</sup> При использовании режима Trunking (несколько портов коммутатора объединяются в т.н. транк и составляют один логический порт, пропускная способность которого равна сумме пропускных способностей входящих в транк портов) в коммутаторах Fibre Channel производства компании Brocade Communications изменение длины маршрута вызовет потерю синхронизации в транке и его реконфигурацию. Для избежания такой ситуации можно выключить режим Trunking или создать между коммутаторами два транка, один из которых использует «короткий», а другой «длинный» маршрут.



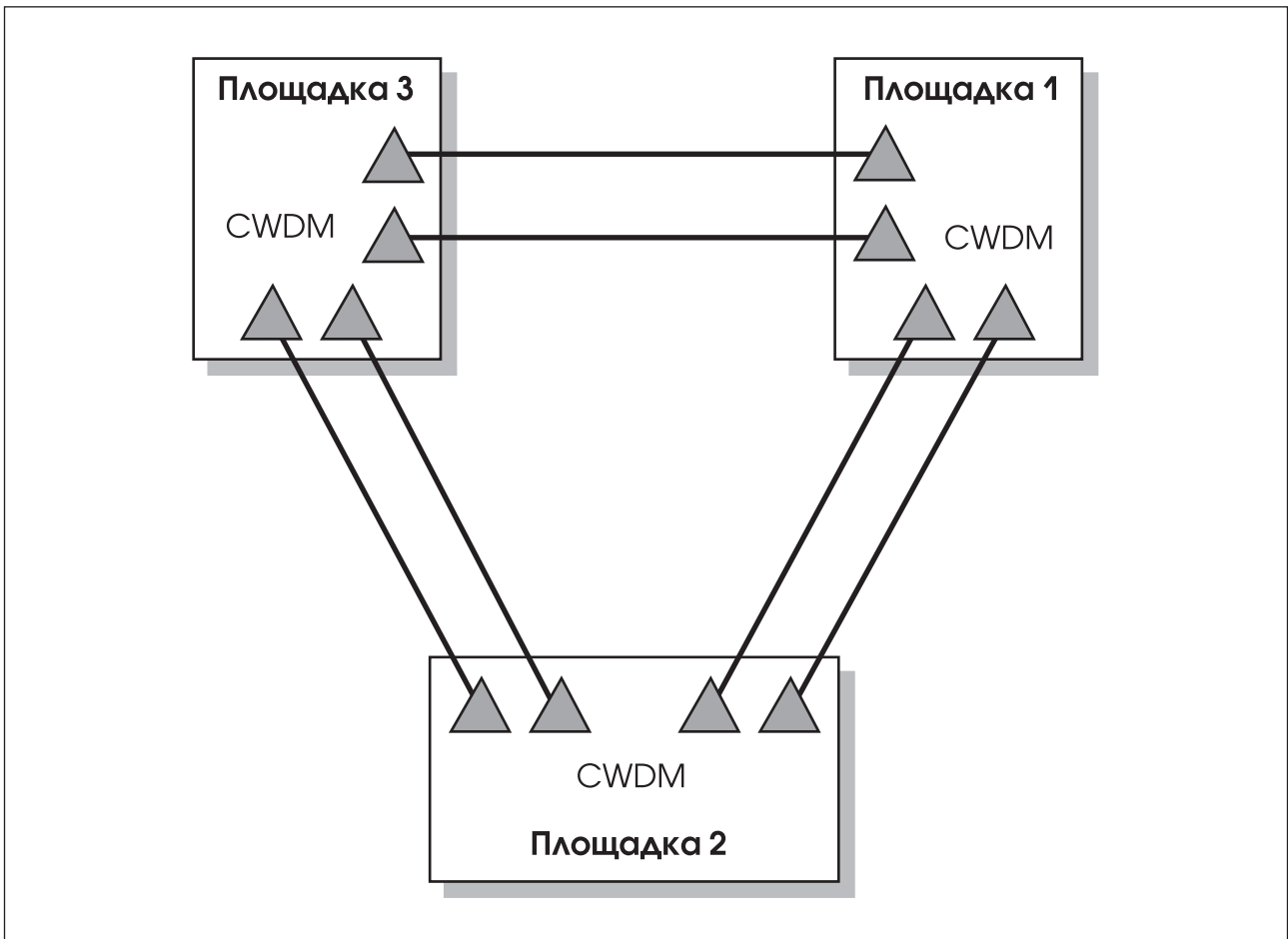


Рис. 4. Объединение трех площадок при помощи оборудования CWDM

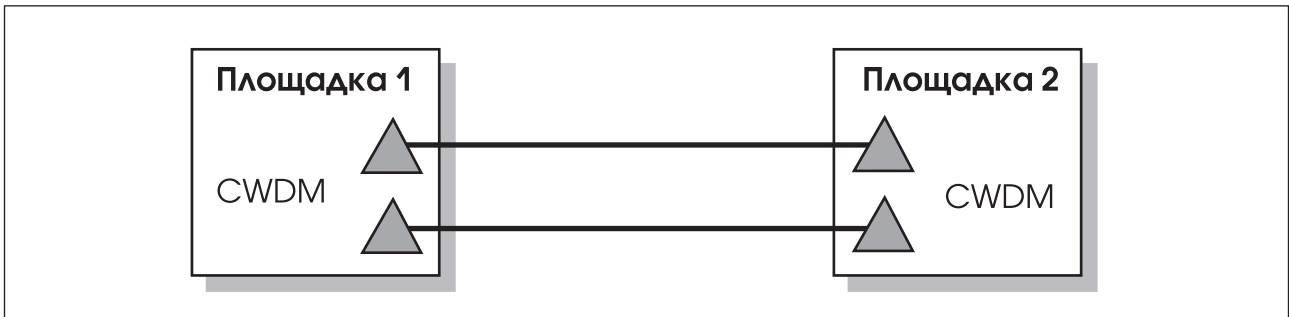


Рис. 5. Соединение двух площадок при помощи оборудования CWDM

пользование более дешевого оборудования CWDM (см. врезку 1. CWDM) позволяет организовать распределенные центры и менее крупным компаниям.

На Рис. 4 представлен вариант модернизации соединения трех площадок с использованием CWDM. Между каждой парой площадок при помощи двух пар оптоволокна и мультиплексов/демультиплексов CWDM проводятся

16 каналов связи. Это могут быть каналы Gigabit Ethernet и 2 Gigabit Fibre Channel. Такого количества каналов должно хватить для создания распределенного вычислительного центра в компании среднего размера.

Для соединения площадок потребуется 12 мультиплексов/демультиплексов CWDM и 96 приемопередатчиков CWDM с 8-ю разными длинами волны. Но все равно цена решения по-

лучается значительно ниже, чем в случае DWDM. Соединение четырех площадок указанным образом будет выглядеть уже сложно, а вот для соединения двух площадок CWDM — почти идеальный вариант (см. Рис. 5).

Следует отметить, что при использовании CWDM отсутствует защита от обрыва соединения на физическом уровне. Резервирование линий связи в этом случае производится на уровне оборудования локальной сети или сети хранения данных. Если разрыв соединения произойдет, то будет потеряно несколько фреймов данных, которые передавались в момент аварии. Однако потери данных не случится, если нарушатся не все связи между площадками. Контроль целостности данных на уровне протоколов Fibre Channel (или FCP) и TCP/IP приведет к тому, что пропавшие команды и данные будут повторены<sup>3</sup>.

- приемопередатчики 4 Gigabit для работы на больших расстояниях — ELWL и CWDM;
- приемопередатчики 10 Gigabit для работы на больших расстояниях.

Благодаря Infiniband можно будет объединять серверы, расположенные на разных площадках, в сильно связанные конфигурации: не только параллельные кластеры, но и комплексы, работающие под управлением одной операционной системы, смогут быть разнесены на значительные расстояния.

Остальные усовершенствования позволят увеличить пропускную способность каналов, соединяющих площадки распределенного вычислительного центра. Площадки распределенного вычислительного центра, удаленные на 45–50 км (длина по оптоволоконному кабелю) можно будет соединить каналами связи с пропускной способностью десятки и сотни гигабит. Использовать полученную мощность для репликации или резервного копирования будет совершенно неразумно.

## Ближайшие изменения

Технологии настоящего времени уже позволяют успешно строить распределенные вычислительные центры, и можно с уверенностью утверждать, что технологии ближайшего будущего должны поставить точку в схеме резервных центров.

По мнению автора статьи, значительное влияние на концепции резервирования окажут следующие новинки:

- появление Infiniband для объединения серверов (см. врезку 2. Infiniband);
- 4 Gigabit Fibre Channel на серверах и устройствах хранения данных (и это уже реальность);
- 8 Gigabit Fibre Channel для соединения коммутаторов — построения ISL;

## Использование распределенных центров в IT-системах большого масштаба

Распределенные вычислительные центры можно использовать как основу больших IT-систем. При построении такой системы в российских компаниях необходимо учитывать особенности, о которых уже упоминалось выше — высокий уровень централизации систем связи и управления в нашей стране. В этих условиях IT-система предприятия или организации тоже должна

<sup>3</sup> Например, в журнале ОС Solaris могут появиться сообщения вида «SCSI transport failed: reason 'tran\_err': retrying command», говорящее о том что пришлось повторить команду SCSI в результате потери соединения.

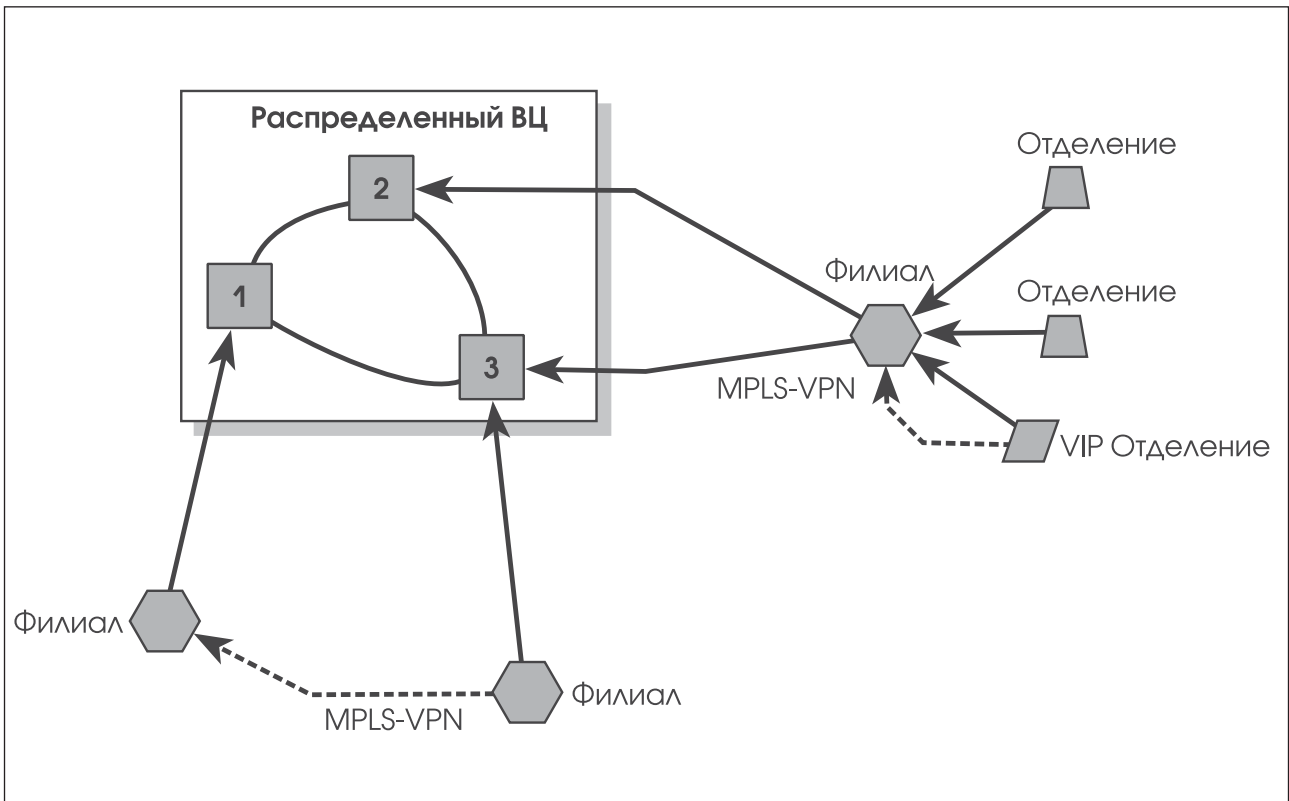


Рис. 6. IT-система, обслуживающая несколько регионов

быть централизованной, а кроме того, конечно, и надежной.

Итак, IT-система, обслуживающая несколько территорий, может быть организована следующим образом. (Рис. 6).

Подходящей архитектурой IT в этом случае может быть такая, которая опирается на сильный центр и базируется на следующих принципах:

- сосредоточить в центре самые значимые функции IT;
- вынести на периферию функции IT, размещение которых в центре приводит к ненужным затратам;
- сосредоточить в центре все сложные решения, периферийные конструкции максимально стандартизовать.

Такой подход позволяет сосредоточить в центре системы решения по повышению производительности и надежности. Для реализации этих задач необходим квалифицированный персонал (во многих городах нашей страны пока еще трудно найти IT-специалистов с квалификацией, выходящей за рамки Windows-Intel-Cisco).

IT-инфраструктура имеет три уровня иерархии – центральный офис, филиал и отделение; отделения подключены к своим филиалам, у центрального офиса могут быть свои отделения, необходимые для работы компании в центре.

Средства поддержки основных приложений – серверы СУБД и приложений – располагаются только в центральном офисе, что облегчает решение следующих задач:

- обеспечение производительности приложений;
- обеспечение доступности приложений;
- обновление версий и внедрение новой функциональности приложений.

Эти задачи придется решать только в одном месте – в центральном офисе.

В филиалах располагаются средства поддержки работы офиса – электронная почта, обработка документов, клиентские места основных приложений и средства связи с центральным офисом и отделениями. Для упрощения администрирования можно размещать средства поддержки работы офиса на терминальных серверах Windows. Такое решение позволит уменьшить количество людей, необходимых для об-

служивания IT- системы в филиалах, упростит задачу обновления версий программного обеспечения в филиалах, а также упростит задачу унификации IT в филиалах и увеличит скорость развертывания филиала.

Отсутствие основных приложений и максимальная унификация IT в филиалах должны компенсировать отсутствие уникальных, незаменимых специалистов и упростить поддержку IT. При такой схеме филиалам не нужна поддержка в режиме 24x7 и не потребуются выезд специалистов внешних служб поддержки (производителей оборудования или специализированных сервисных организаций) в филиалы для решения проблем.

В отделении устанавливаются только удаленные терминалы для доступа к офисным ресурсам филиала и клиентам основных приложений, расположенным в филиале. Квалификация IT-персонала в отделениях самая низкая, поэтому в отделении не должно быть самостоятельных серверов и локально расположенных приложений — только средства связи с филиалом.

Схема, опирающаяся на сильный центр, страдает очевидным недостатком — наличием узла, отказ в работе которого приводит к выходу из строя всей системы. Но этот недостаток не содержит фатальной угрозы для IT-системы, поскольку ее центральная часть организована как распределенный вычислительный центр, а глобальная сеть, соединяющая центральный офис и филиалы должна содержать резервные пути доступа.

Глобальная сеть организована таким образом, что точки выхода в нее расположены на всех площадках распределенного центра обработки данных. Такое решение обеспечит существование сети даже в случае выхода из строя площадок ЦОД. Кроме того, согласно этой схеме, отделения подключаются к филиалам, а филиалы к центральному офису. Филиалы исполняют роль концентраторов для отделений.

Каждый филиал имеет основной и резервный каналы, связывающие его с центральным

офисом. Чтобы глобальная сеть оставалась в рабочем состоянии и в случае отказов площадок центра обработки данных, резервный и основной каналы подключаются к разным площадкам центра.

Резервные каналы могут быть организованы по двум схемам.

1. Основной и резервный каналы соединяют филиал и центральный офис и принадлежат разным провайдерам. Такой вариант подойдет для филиалов, расположенных в центральной части России, где существуют развитая сеть каналов связи и несколько телекоммуникационных провайдеров со своей инфраструктурой.
2. Резервный канал соединяет филиал с другим филиалом, расположенным в соседнем регионе, обслуживаемом другим телекоммуникационным провайдером. Так, например, проще получить канал от Иркутска до Томска, чем второй независимый канал от Иркутска до Москвы. Филиал-сосед будет пропускать через себя транзитный трафик в случае отказа основного канала. Резервный канал может быть организован по более дешевой технологии, чем основной. Основной канал от филиала до центрального офиса реализован по выделенной линии, а резервный, связывающий филиалы, — по технологии MPLS VPN, широко применяемой сейчас телекоммуникационными провайдерами России.

Для работы схемы с резервированием каналов необходимо обеспечить динамическую маршрутизацию (OSPF, EGRP) в глобальной сети компании. Отделения подключаются к филиалам по нерезервируемым каналам связи местных телекоммуникационных провайдеров. Для отделений класса VIP может быть предусмотрено резервирование канала связи с филиалом путем организации VPN-соединения через Интернет. В этом случае не гарантируется пропускная способность канала, но сохраняется возможность работы отделения.

## CWDM

CWDM — более дешевое решение спектрального уплотнения каналов связи, чем DWDM. На Рис. 7 показаны основные компоненты CWDM:

- оптические мультиплексоры/демультиплексоры;
- приемопередатчики с разными длинами волны.

Мультиплексоры/демультиплексоры CWDM — полностью пассивные устройства, обеспечивающие объединение и разделение нескольких сигналов. CWDM позволяет передавать по одному оптоволокну до 8 сигналов из диапазона C-band с длинами волны 1470 нм, 1490 нм, 1510 нм, 1530 нм, 1550 нм, 1570 нм, 1590 нм и 1610 нм. Все сигналы расположены вокруг длины волны 1550 нм — области мини-

мальных потерь при передаче света в одномодовом оптоволоконном кабеле.

Приемопередатчики CWDM позволяют передавать сигналы на расстояние до 75 км по оптоволоконному кабелю. Передатчики выполняются в формате SFP и устанавливаются непосредственно в сетевое оборудование Ethernet или Fibre Channel. В настоящее время существуют CWDM SFP, способные передавать сигнал с пропускной способностью 2.5 Gigabit/s, т.е. могут быть использованы для передачи 2 Gigabit Fibre Channel, Gigabit Ethernet и Infiniband.

Решения CWDM сертифицированы для применения с оборудованием Brocade Communications, Cisco Systems, Nortel Networks и других ведущих производителей сетевого оборудования.

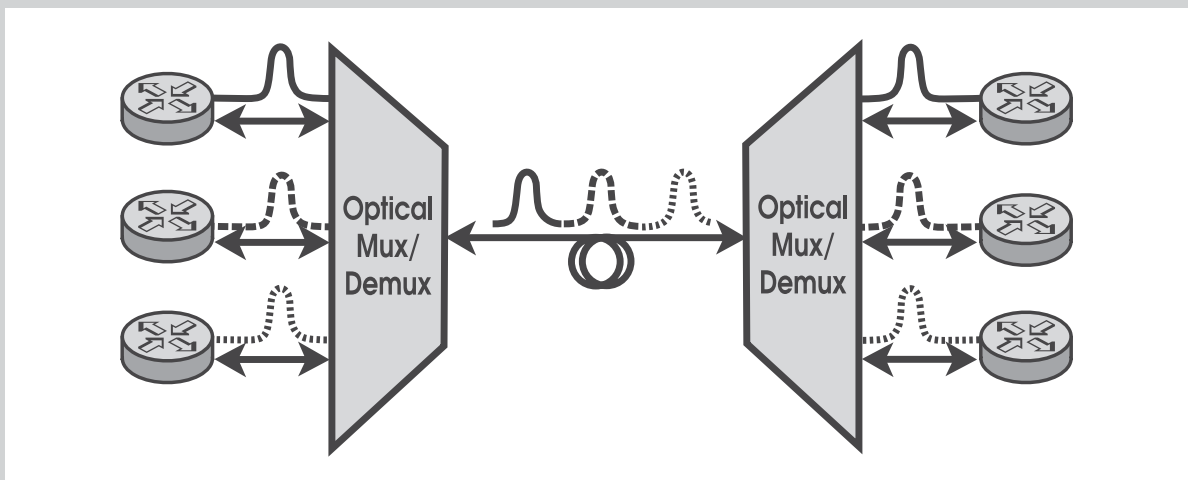


Рис. 7. Компоненты CWDM

## Infiniband

Протокол Infiniband проектировался, чтобы получить сетевую среду, в которой объединяется большое количество узлов, обеспечивается низкая задержка передачи информации и упрощаются протоколы связи между узлами сети.

Существующие протоколы имеют известные недостатки. Так в Fibre Channel не предусмотрена маршрутизация, а передача данных через TCP/IP сопровождается большими накладными расходами и высокой задержкой передачи данных. Авторы Infiniband стремились сделать свой вариант сетевого соединения свободным от указанных недостатков.

Основные компоненты Infiniband показаны на Рис. 8.

- Channel Adaptor — устройство, устанавливаемое в оборудование для связи по Infiniband; различают Host Channel Adaptor (HCA), который установлен в компьютере (т.н. Processor Node), и Target Channel Adaptor (TCA), обеспечивающий подключение к Infiniband систем хранения данных.
- Коммутатор (switch) — объединяет устройства (HCA и TCA) внутри одной подсети (IB Subnet).
- Маршрутизатор (router) — соединяет разные подсети Infiniband. Использование маршрутизации позволяет объединять в сети Infiniband неограниченное количество узлов.

Схема адресации в Infiniband позаимствована из IPv6.

Для уменьшения задержки передачи данных Infiniband спроектирован с небольшим набором основных операций:

- Send — передать данные другому Channel Adaptor;
- RDMA Write — записать данные в память другого Channel Adaptor;
- RDMA Read — прочитать данные из памяти другого Channel Adaptor;
- Atomic — записать в память другого Channel Adaptor 32-разрядное слово и вернуть предыдущее значение — используется в примитивах синхронизации.

Все операции разработаны таким образом, чтобы во время их выполнения не происходило копирования данных внутри одной системы — в каждой команде указывается только адрес блока данных в памяти. Именно поэтому на Рис. 8 HCA показаны присоединенными непосредственно к блокам памяти компьютеров.

Спецификация Infiniband составлена так, что в качестве основного канала связи используется канал 2.5 Gigabit/s — это так называемый однократный (1X) интерфейс. Объединение четырех линий связи 2.5 Gigabit/s составляет четырехкратный (4X) интерфейс с пропускной способностью 10 Gigabit/s, а объединение двенадцати линий 2.5 Gigabit/s — двенадцатикратный (12X) интерфейс с пропускной способностью 30 Gigabit/s.

Однократный и четырехкратный интерфейсы используются для соединения HCA и TCA к сети Infiniband, а двенадцатикратный — для соединения коммутаторов и маршрутизаторов.

Для работы четырехкратного интерфейса используются 4 пары оптоволокна, а для работы двенадцатикратного — 12. Таким образом, для связи площадок сетями Infiniband можно использовать существующие технологии для передачи сигналов 2.5 Gigabit (темная оптика, технологии спектрального уплотнения DWDM и CWDM).

Также в спецификации Infiniband предусмотрены решения по резервированию путей между устройствами Infiniband и решения по управлению пропускной способностью — объединение пропускной способности Channel Adaptors, установленных в одном узле, и ограничение пропускной способности Infiniband для разных приложений (QoS).

В настоящий момент Infiniband находится в том же состоянии, в котором Fibre Channel был в 2001 году — появляются промышленные решения на Infiniband у производителей серверов. Вполне вероятно, что следующее поколение серверов Sun, IBM и HP уже будет обладать встроенными HCA, а устройства хранения данных следующего поколения также будут поддерживать подключение по Infiniband в дополнение к Fibre Channel и iSCSI.

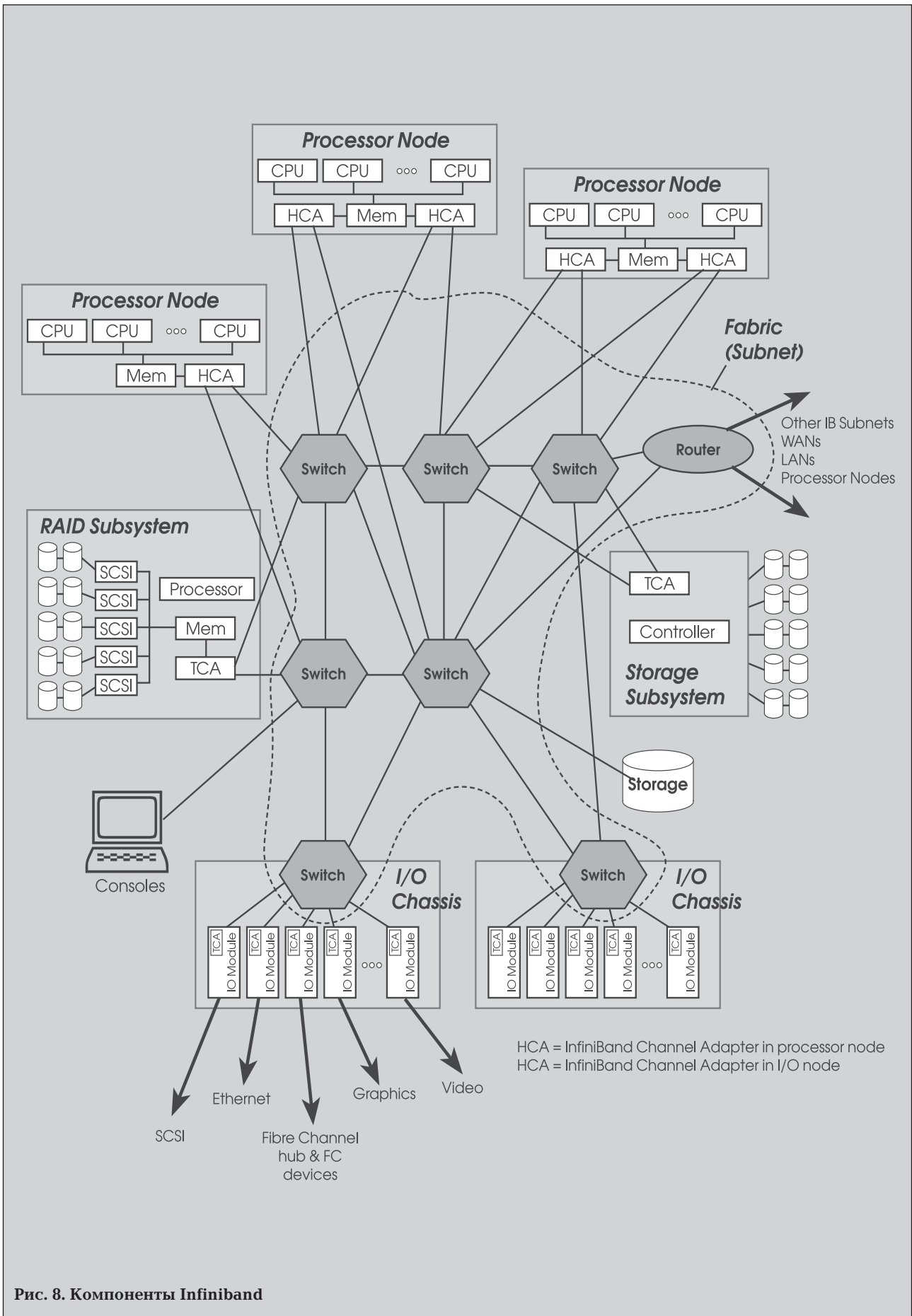


Рис. 8. Компоненты Infiniband

---

---

# Jet Info

ИНФОРМАЦИОННЫЙ БЮЛЛЕТЕНЬ

Издается с 1995 года

Издатель: компания «Инфосистемы Джет»

Главный редактор: Дмитриев В.Ю. ([vlad@jet.msk.su](mailto:vlad@jet.msk.su))  
Технический редактор: Лапина И.К. ([lapina@jet.msk.su](mailto:lapina@jet.msk.su))  
Россия, 127015, Москва, Б. Новодмитровская, 14/1  
тел. (495) 411 76 01  
факс (495) 411 76 02  
email: [JetInfo@jet.msk.su](mailto:JetInfo@jet.msk.su) <http://www.jetinfo.ru>

Подписной индекс по каталогу Роспечати

**32555**

