ELCOM

Технологии хранения данных

Сергей Головин, ELCOM Ltd.

Sergey.Golovin@Elcom.SPb.Ru

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

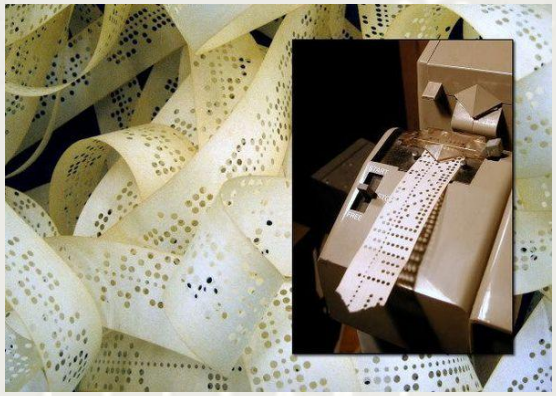
История систем хранения данных

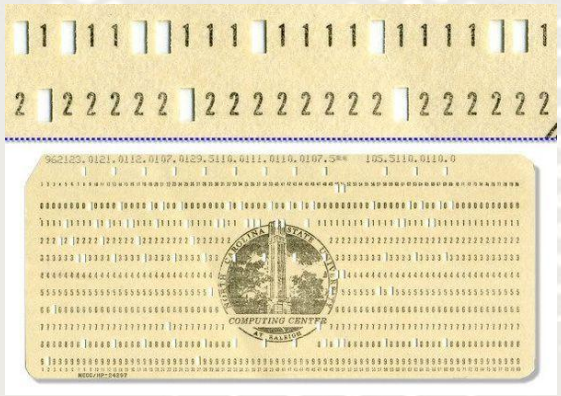
Первый шаг на пути к созданию современных СХД был сделан в конце XVIII века французом Жозеф-Мари Жаккардом, который изобрел перфокарты для управления вышивальным станком.

В 1890 году Герман Холлерит применил перфокарту для обработки данных переписи населения в США. Именно он нашел компанию (будущую IBM), которая использовала такие карты в своих счетных машинах.

В 1950-х годах IBM уже вовсю использовала в своих компьютерах перфокарты для хранения и ввода данных, а вскоре этот носитель стали применять и другие производители. Тогда были распространены 80-столбцовые карты, в которых для одного символа отводился отдельный столбец.

В 2002 году IBM все еще продолжала разработки в области технологии перфокарт. Правда, в XXI веке компанию интересовали карточки размером с почтовую марку, способные хранить до 25 миллионов страниц информации.





ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

История систем хранения данных

Вместе с выходом первого американского коммерческого компьютера UNIVAC I (1951) в IT-индустрии началась эра магнитной пленки. Первопроходцем, как водится, снова стала IBM, потом «подтянулись» другие.

В 1963 году IBM представила первый винчестер со съемным диском – IBM 1311. Он представлял собой набор взаимозаменяемых дисков. Каждый набор состоял из шести дисков диаметром 14 дюймов, вмещавших до 2 Мб информации.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

История систем хранения данных

(другие варианты носителей)

1970 – Дискеты

1976 – ROM картриджи

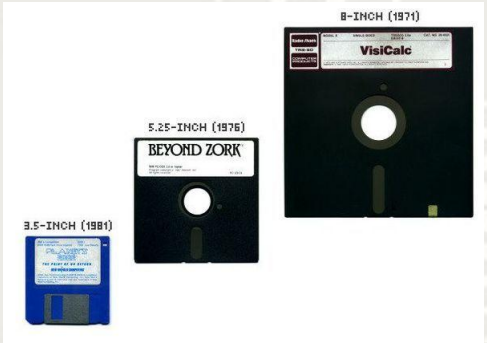
1982 – CD-ROM

1988 – CD-R

1992 – Магнитооптические диски

1995 – Flash карты

2000 – USB Flash



ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Типы интерфейсов жестких дисков

SCSI (Small Computer Systems Interface)

SCSI-1, SCSI-2, SCSI-3, Ultra-2 SCSI, Ultra-3 SCSI, Ultra-320 SCSI, Ultra-640 SCSI

SATA (Serial ATA)

SATA/150, SATA/300 (SATA II), eSATA , NL SATA (Enterprise Near Line)

SAS (Serial Attached SCSI)

FC (Fibre Channel)

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

SCSI

SCSI-1	восьмибитная шина, с пропускной способностью в 3,5 МБайт/сек в асинхронном режиме и 5 МБайт/сек в синхронном режиме
SCSI-2	Fast SCSI и Wide SCSI. Fast SCSI характеризуется удвоенной пропускной способностью (до 10 МБайт/сек). Wide SCSI имеет удвоенную разрядность шины (16 бит), что позволяет достичь скорости передачи до 20 МБ/сек.
SCSI-3	Пропускная способность шины 20 МБайт/сек для восьмибитной шины и 40 МБайт/сек - для шестнадцатибитной
Ultra-2 SCSI	Использует LVDS. Максимальная длина кабеля - 12 метров, пропускная способность - до 80 МБайт/сек.
Ultra-3 SCSI	Имеет удвоенную пропускную способность (по сравнению с Ultra-2 SCSI), которая составила 160 МБайт/сек. В этот стандарт было добавлено использование CRC (Cyclic Redundancy Check), исправление ошибок.
Ultra-320 SCSI	Развитие стандарта Ultra-3 с удвоенной скоростью передачи данных (до 320 МБайт/сек).
Ultra-640 SCSI	Также известен под названием Fast Ultra-320. Удвоенная пропускная способность (640 МБайт/сек). Не получил большого распространения, т.к. поддерживает всего 2 устройства на шлейфе.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Разъемы и кабели SCSI

ULTRA320 SCSI

SCSI SE

SCSI LVD

SCSI LVD

SE

Ultra 160 SCSI

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Протокол команд SCSI

В терминологии SCSI взаимодействие идёт между инициатором и целевым устройством.
Инициатор посылает команду целевому устройству, которое затем отправляет ответ инициатору.
Команды SCSI посылаются в виде блоков описания команды, длина которых может составлять 6, 10, 12 или 16 байт. В последних версиях SCSI блок может иметь переменную длину. Блок состоит из однобайтового кода команды и параметров команды.
После получения команды целевое устройство возвращает значение 00h в случае успешного получения, 02h в случае ошибки или 08h в случае, если устройство занято. В случае, если устройство вернуло ошибку, инициатор обычно посылает команду запроса состояния. Устройство возвращает Key Code Qualifier.

Все команды SCSI делятся на четыре категории: N - non-data, W -запись данных от инициатора целевым устройством, R - чтение данных и B - двусторонний обмен данными. Всего существует порядка 60 различных команд SCSI, из которых наиболее часто используются:

- Test unit ready - проверка готовности устройства.
- Inquiry - запрос основных характеристик устройства.
- Send diagnostic - указание устройству провести самодиагностику и вернуть результат.
- Request sense - возвращает код ошибки предыдущей команды.
- Read capacity - возвращает ёмкость устройства.
- Format Unit
- Read (4 варианта) - чтение.
- Write (4 варианта) - запись.
- Write and verify - запись и проверка.
- Mode select - установка параметров устройства.
- Mode sense - возвращает текущие параметры устройства.

Каждое устройство на SCSI-шине имеет как минимум один номер логического устройства. В некоторых более сложных случаях одно физическое устройство может представляться набором логических номеров. Терминирование Параллельные шины SCSI всегда должны терминироваться с обеих сторон для обеспечения нормального функционирования. Подавляющее большинство контроллеров и многие устройства имеют возможность автотерминирования - использования встроенного терминатора.
Для передачи команд протокола SCSI по IP-сетям используется сетевой протокол iSCSI, утверждённый IETF как стандартный в 2003 году

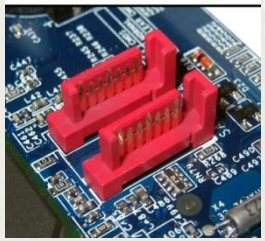
ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

SATA

Последовательный интерфейс обмена данными с накопителями информации. SATA является развитием интерфейса ATA, который после появления SATA был переименован в **PATA**.

SATA/150



Первоначально стандарт SATA предусматривал работу шины на частоте 1,5 ГГц, обеспечивающей пропускную способность приблизительно в 150 МБ/с. Пропускная способность SATA/150 незначительно выше, чем у шины UDMA/133. Главным преимуществом SATA перед PATA является использование последовательной шины вместо параллельной. Несмотря на то, что последовательный способ обмена принципиально медленнее параллельного, в данном случае это компенсируется возможностью работы на более высоких частотах за счет большей устойчивости кабеля к помехам. Достигается это меньшим числом проводников и объединением информационных проводников в 2 витые пары, экранированные заземлёнными проводниками.

SATA/300

Стандарт SATA/300 работает на частоте 3 ГГц, обеспечивает пропускную способность до 300 МБ/с. Весьма часто стандарт SATA/300 называют SATA II. Теоретически SATA/150 и SATA/300 устройства должны быть совместимы за счёт поддержки согласования скоростей, однако для некоторых устройств и контроллеров требуется ручное выставление режима работы. Стандарт SATA предусматривает возможность увеличения скорости работы до 600МБ/с.

eSATA

eSATA или External SATA это интерфейс подключения внешних устройств, поддерживающий режим горячей замены. Он был создан несколько позже SATA, в середине 2004 года. Основные особенности eSATA:

- Разъёмы менее хрупкие и конструктивно рассчитаны на большее число подключений.
- Требует для подключения два провода: шину данных и силовой кабель.
- Ограничен по длине кабеля данных около двух метров.
- Средняя скорость передачи данных выше, чем у USB и IEEE 1394.
- Существенно меньше нагружается центральный процессор.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

6Gb/s Serial Attached SCSI

SAS

Интерфейс SAS или Serial Attached SCSI обеспечивает подключение по физическому интерфейсу, аналогичному SATA, устройств, управляемых набором команд SCSI. Обладая обратной совместимостью с SATA, он даёт возможность подключать по этому интерфейсу любые устройства, управляемые набором команд SCSI - не только жёсткие диски, но и сканеры, принтеры и др. По сравнению с SATA, SAS обеспечивает более развитую топологию, позволяя осуществлять параллельное подключение одного устройства по двум или более каналам. Также поддерживаются расширители шины, позволяющие подключить несколько SAS устройств к одному порту.

Типичная система с интерфейсом SAS состоит из следующих компонентов:

Инициаторы:

Инициатор - устройство, которое порождает запросы на обслуживание для целевых устройств и получает подтверждения по мере исполнения запросов.

Целевые устройства:

Целевое устройство содержит логические блоки и целевые порты, которые осуществляют приём запросов на обслуживание, исполняет их; после того, как закончена обработка запроса, инициатору запроса отсылается подтверждение выполнения запроса. Целевое устройство может быть как отдельным жёстким диском, так и целым дисковым массивом.

Подсистема доставки данных

Является частью системы ввода-вывода, которая осуществляет передачу данных между инициаторами и целевыми устройствами. Обычно подсистема доставки данных состоит из кабелей, которые соединяют инициатор и целевое устройство. Дополнительно, кроме кабелей в состав подсистемы доставки данных могут входить расширители SAS.

Расширители

Расширители SAS - устройства, входящие в состав подсистемы доставки данных и позволяют облегчить передачи данных между устройствами SAS, например, позволяет соединить несколько целевых устройств SAS к одному порту инициатора. Подключение через расширитель является абсолютно прозрачным для целевых устройств.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Fibre Channel

Жесткие диски с протоколом подключения FC - по сути дела повторяют по своей концепции традиционные SCSI диски, за исключением того, что к контроллеру подключается не по электрическому шлейфу SCSI а по оптическому кабелю. Все основные стандарты SCSI сохраняются и в FC, только изменяется среда передачи данных на оптическую. По сути можно говорить об инкапсуляции протокола SCSI по сетям Fibre Channel. Внутренняя механическая структура жесткого диска FC полностью повторяет таковую у SCSI. Так что можно рассматривать эту технологию как развитие SCSI в плане устранения ограничений, накладываемых последовательной шиной и позволяющую организовывать подключения между диском и контроллером в режиме точка-точка, при том, что режим петли (Arbitrated Loop), похожий по принципу на работу шины SCSI, так же оставлен в функционале FC.

В свое время, переход протокола подключения внутренних дисков серверов с SCSI на FC, позволил добиться лучшей надежности работы дисков за счет того, что исчезло ограничение полосы пропускания из за высокочастотных помех на шлефе SCSI и диски стали доступны одновременно сразу для нескольких инициаторов, что дало возможность упростить кластерные конфигурации. Замена SCSI внутренних дисков на FC происходила довольно планомерно и достаточно прозрачно для операционных систем, т.к. не требовала никаких переписываний драйверов и модулей, ибо внутри системы диски выглядели как традиционные SCSI с тем же набором команд и сигналов.

Особенности протокола FC будут рассматриваться далее в разделе про внешнее подключение СХД.

На данный момент диски FC позиционируются как носители данных с наименьшим временем реакции, наименьшим временем ожидания и высокой производительностью. В производстве существенно более дороги, потому используются в основном в системах Enterprise класса. До недавнего времени скорость вращения шпинделя 15000 RPM можно было встретить только у FC дисков.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Infiniband

На данный момент Infiniband контроллеры для жесткого диска существуют, но практически не используются, и считаются избыточными, т.к. механика диска пока неспособна преодолеть границу пропускной способности традиционного FC при стоимости Infiniband контроллера многократно большей.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

RAID и его типы

RAID (англ. redundant array of independent disks — избыточный массив независимых жёстких дисков)

Массив из нескольких дисков, управляемых контроллером, взаимосвязанных скоростными каналами и воспринимаемых внешней системой как единое целое. В зависимости от типа используемого массива может обеспечивать различные степени отказоустойчивости и быстродействия. Служит для повышения надёжности хранения данных и/или для повышения скорости чтения/записи информации

Калифорнийский университет в Беркли представил следующие уровни спецификации RAID, которые были приняты как стандарт де-факто:

RAID 0 представлен как дисковый массив повышенной производительности, без отказоустойчивости.
RAID 1 определён как зеркальный дисковый массив.
RAID 2 зарезервирован для массивов, которые применяют код Хемминга.
RAID 3 и **4** используют массив дисков с чередованием и выделенным диском чётности.
RAID 5 используют массив дисков с чередованием и "невыделенным диском чётности".
RAID 6 используют массив дисков с чередованием и двумя независимыми "чётностями" блоков.
RAID 10 — RAID 0, построенный из **RAID 1** массивов
RAID 50 — RAID 0, построенный из **RAID 5**
RAID 60 — RAID 0, построенный из **RAID 6**

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

RAID и его типы


Standalone


Cluster


Hot swap


RAID 0


RAID 1


RAID 5

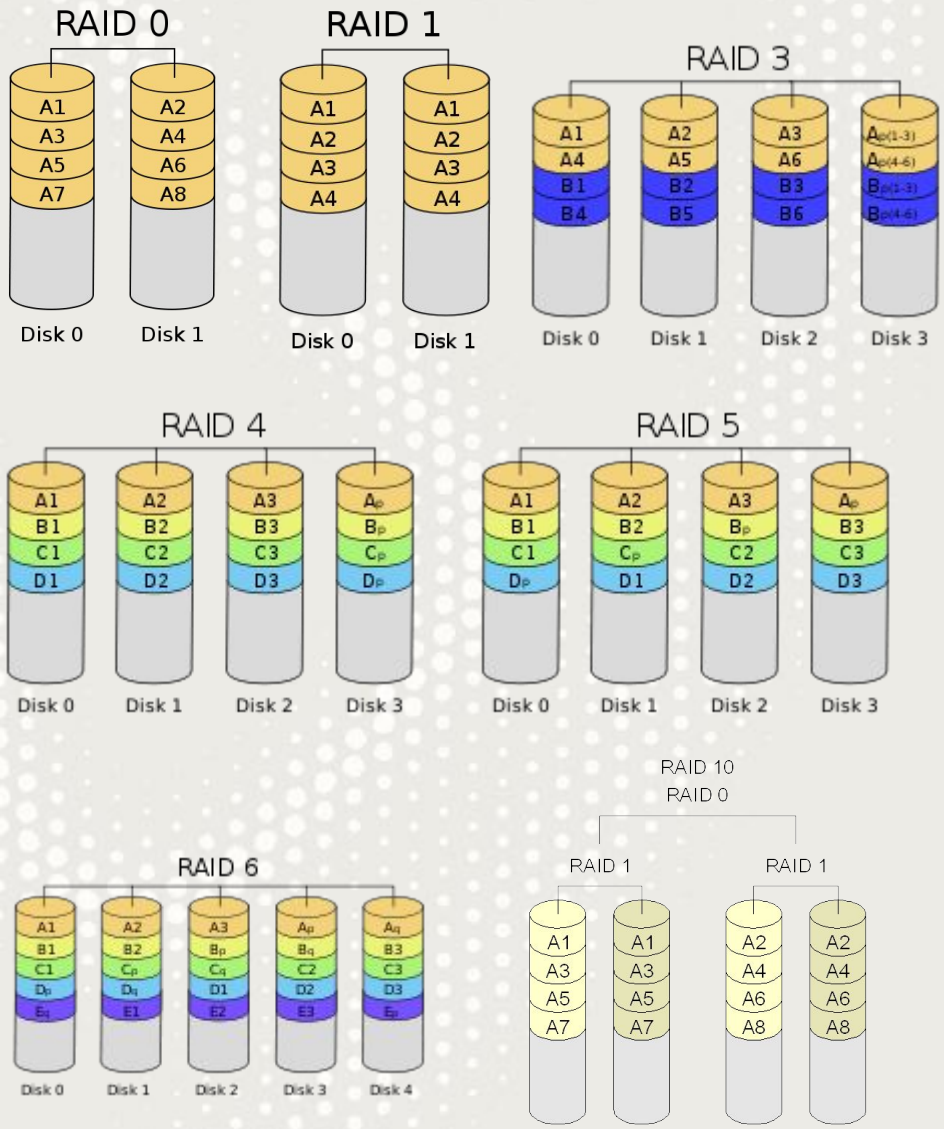

RAID 0+1

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

RAID и его типы

Типы RAID массивов	Требуемое количество дисков	Избыточность	Увеличение скорости при N дисках в массиве*	Накладные расходы**
JBOD (Span)	2+	Нет	Неизвестно (сильно зависит от параметров раздела), но для типичных применения увеличение незначительно.	Нет
RAID 0 (Чередование)	2+	Нет	Ускорение операций чтения/записи в N раз	Нет
RAID 1 (Зеркало)	Ровно 2	Отказ одного диска	Скорость чтения удваивается; Скорость записи без изменений.	50%
RAID 5 (Чередованный раздел с четностью)	3+	Отказ одного диска (с падением скорости)	Скорость чтения увеличивается в (N-1) раз; Скорость записи снижается вплоть до 50% в тяжелых случаях	Емкость, равная емкости одного из дисков составляющих массив (используемого для хранения контрольной суммы).
RAID 0+1 (Mirrored stripe set)	4+	Отказ одного диска и половина случаев выхода из строя двух дисков сразу, в зависимости от расположения/ назначения этих дисков.	Скорость чтения увеличивается в N раз; Скорость записи в N/2 раз	50%

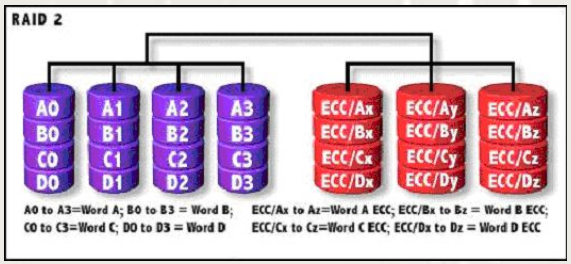


ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

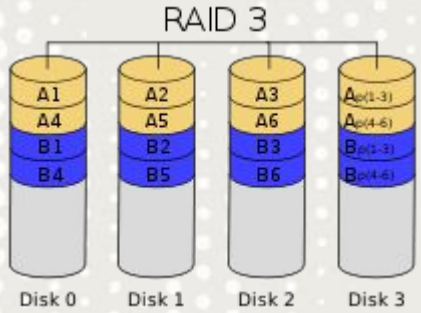
Экзотические и проприетарные типы RAID

RAID 2 (с использованием кода Хемминга)



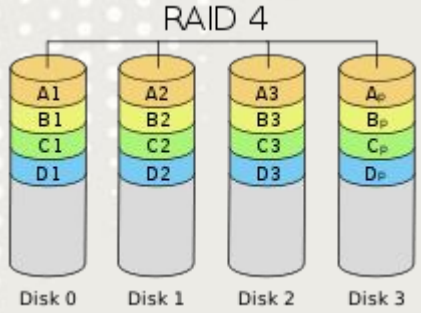
Массив RAID 2 распределяет данные на жестких дисках массива побитно: первый бит записывается на первом жестком диске, второй бит - на втором жестком диске и т. д. Избыточность обеспечивается за счет нескольких дополнительных дисков, куда записывается код коррекции ошибок. Эта реализация дороже, поскольку требует больших накладных расходов: RAID массив с числом основных дисков от 16 до 32 должен иметь три дополнительных жестких диска для хранения кода коррекции. Массив RAID 2 обеспечивает высокую производительность и надежность, но его применение ограничено главным образом рынком компьютеров для научных исследований из-за высоких требований к минимальному объему дискового пространства массива.

RAID 3



Массив RAID 3 распределяет данные на жестких дисках массива побайтно: первый байт записывается на первом жестком диске массива, второй байт - на втором жестком диске и т. д. Избыточность обеспечивает один дополнительный жесткий диск, куда записывается сумма данных по модулю 2 (XOR) для каждого из основных дисков массива. Таким образом, массив RAID 3 разбивает записи файлов данных, храня их одновременно на нескольких жестких дисках и обеспечивая очень быстрые чтение и запись. XOR-сегменты на дополнительном диске позволяют обнаружить любую неисправность дисковой подсистемы, а специальное ПО определит, какой из дисков массива вышел из строя. Использование побайтного распределения данных позволяет выполнять одновременное чтение или запись данных с нескольких дисков для файлов с очень длинными записями. В каждый момент времени может выполняться только одна операция чтения или записи.

RAID 4



RAID 4 похож на RAID 3, но отличается от него тем, что данные разбиваются на блоки, а не на байты. Таким образом, удалось отчасти «победить» проблему низкой скорости передачи данных небольшого объёма. Запись же производится медленно из-за того, что чётность для блока генерируется при записи и записывается на единственный диск. Из систем хранения широкого распространения RAID-4 применяется на устройствах хранения компании NetApp (NetApp FAS), где его недостатки успешно устранены за счет работы дисков в специальном режиме групповой записи, определяемом используемой на устройствах внутренней файловой системой WAFL.

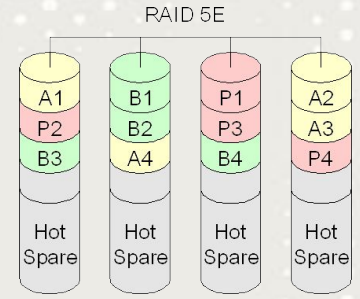
ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Экзотические и проприетарные типы RAID

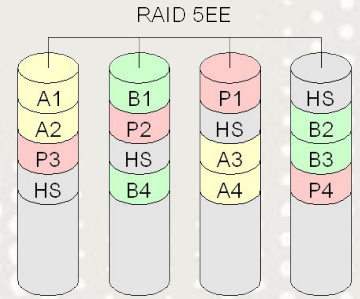
RAID 5E

По сути дела такой же RAID 5, но с дополнительным диском для контроля четности, который по сути находится в состоянии hot-spare. В случае выхода из строя одного из дисков, система по сути деградирует в обычный RAID 5



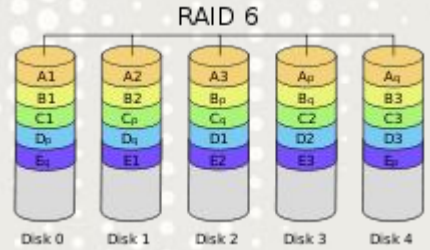
RAID 5EE

Отличается от RAID 5E тем, что на резервном диске располагаются блоки контрольных сумм, которые помогают уменьшить время восстановления данных в случае поломок. В случае выхода из строя одного из дисков, система деградирует в обычный RAID 5



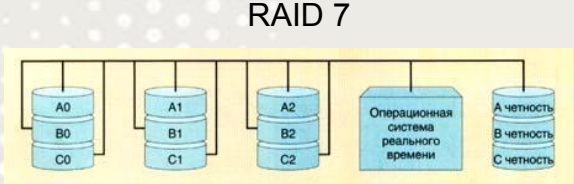
RAID 6

Похож на RAID 5, но имеет более высокую степень надёжности — под контрольные суммы выделяется ёмкость 2-х дисков, рассчитываются 2 суммы по разным алгоритмам. Требует более мощный RAID-контроллер. Обеспечивает работоспособность после одновременного выхода из строя двух дисков — защита от кратного отказа. Для организации массива требуется минимум 4 диска. Обычно использование RAID-6 вызывает примерно 10-15% падение производительности дисковой группы, по сравнению с аналогичными показателями RAID-5, что вызвано большим объёмом обработки для контроллера (необходимость рассчитывать вторую контрольную сумму, а также прочитывать и перезаписывать больше дисковых блоков при записи каждого блока).



RAID 7

Зарегистрированная торговая марка компании Storage Computer Corporation, отдельным уровнем RAID не является. Структура массива такова: на дисках хранятся данные, один диск используется для складирования блоков чётности. Запись на диски кешируется с использованием оперативной памяти, сам массив требует обязательного ИБП; в случае перебоев с питанием происходит повреждение данных.



ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Разновидности СХД и их особенности

JBOD (англ. Just a bunch of disks, просто пачка дисков) — RAID-массив дисков, в которых дисковое пространство распределено по жёстким дискам последовательно. Однако в некоторых контроллерах режимом «JBOD» назван режим, при котором контроллер работает как обычный IDE- или SATA-контроллер, то есть с выключенным RAID, в таком случае каждый диск будет виден как отдельное устройство в операционной системе. Может быть выполнено в разном конструктиве, но суть от этого не меняется, это просто набор дисков, собранных в одном корпусе. В какой-то мере его можно назвать системой прямого доступа **DAS (Direct Access Storage)** но в упрощенном, вырожденном варианте.

DAS (англ. Direct-attached storage) Прямо подключенное хранилище - является запоминающим устройством, непосредственное подключенным к серверу или рабочей станции без помощи сети хранения данных. Способы подключения могут быть различны, SAS, SATA, SCSI, FC

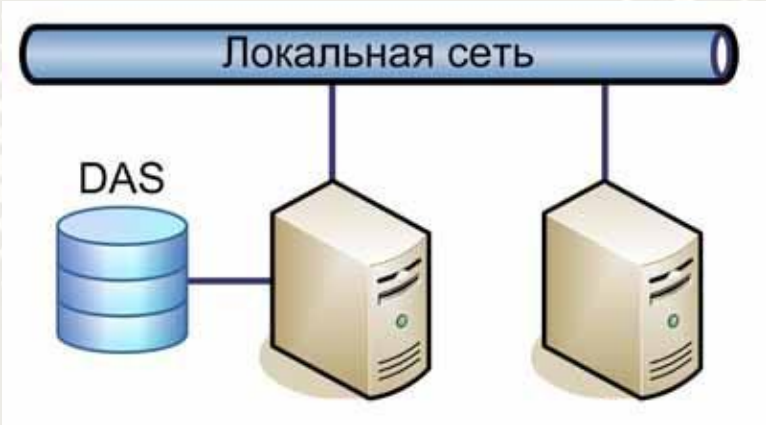
NAS (англ. Network Attached Storage) — сетевая система хранения данных, сетевое хранилище. Зачастую представляет собой просто сервер с подключенным к нему **DAS.** На рынке также присутствуют и специализированные устройства NAS, отличающиеся большим функционалом и автономностью, нежели сервера.

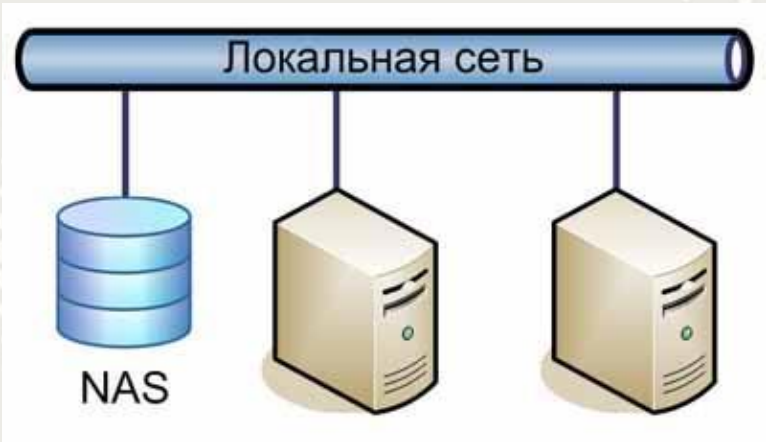
SAN (англ. Storage Area Network, сеть хранения данных) — спредставляет собой архитектурное решение для подключения внешних устройств хранения данных, таких как дисковые массивы, ленточные библиотеки, оптические приводы к серверам таким образом, чтобы операционная система распознала подключённые ресурсы как локальные. SAN характеризуются предоставлением так называемых сетевых блочных устройств (обычно посредством протоколов Fibre Channel, iSCSI или AoE), в то время как сетевые хранилища данных (англ. Network Attached Storage, NAS) нацелены на предоставление доступа к хранящимся на их файловой системе данным при помощи сетевой файловой системы (такой как NFS, SMB/CIFS, или AppleTalk). Следует обратить внимание, что категорическое разделение вида «SAN — это только сетевые диски, NAS — это только сетевая файловая система» является искусственным: с появлением iSCSI началось взаимное проникновение технологий с целью повышения гибкости и удобства их применения.

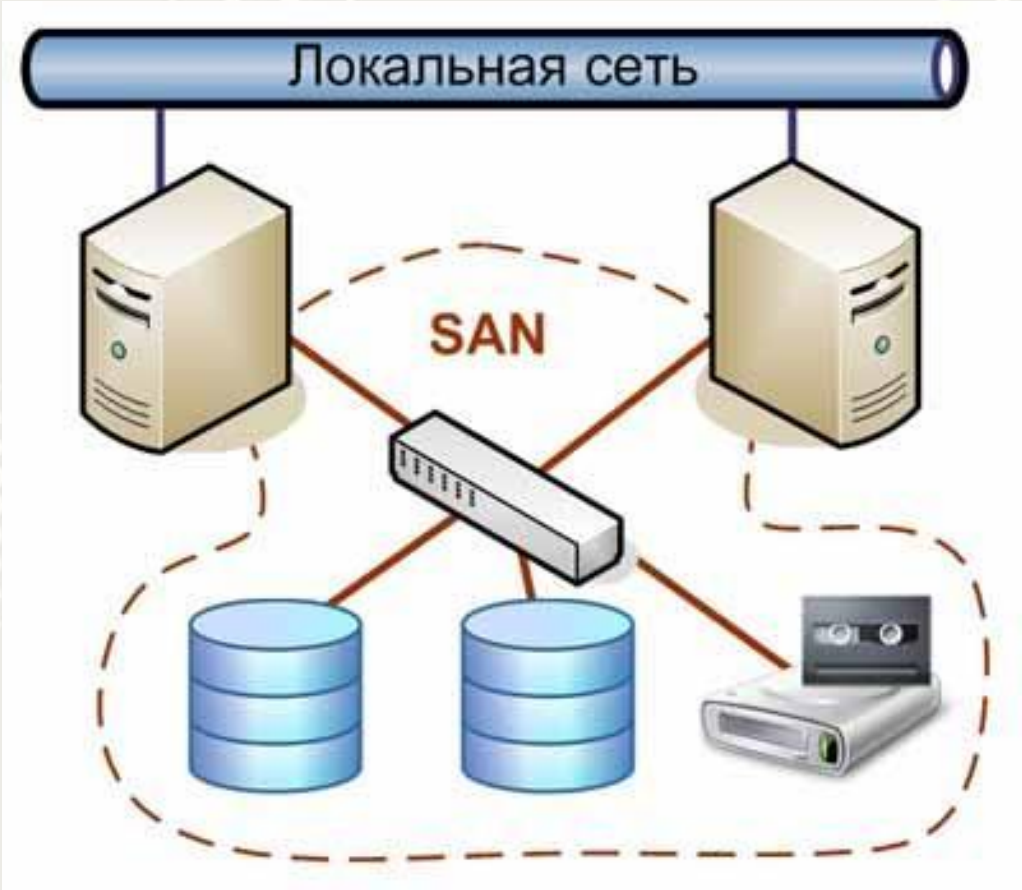
ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Разновидности СХД и их особенности







ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

SAN

Самые распространенные СХД в категории среднего и высшего уровня - это хранилища, которые можно полноправно отнести к категории SAN.
Многие считают, что термин SAN можно применять только для СХД, подключаемым к хостам через FC. Но это на самом деле не так.

Большинство сетей хранения данных использует протокол SCSI для связи между серверами и устройствами хранения данных на уровне шинной топологии. Так как протокол SCSI не предназначен для формирования сетевых пакетов, в сетях хранения данных используются низкоуровневые протоколы:
Fibre Channel Protocol (FCP), транспорт SCSI через Fibre Channel. Наиболее часто используемый на данный момент протокол.
Существует в вариантах 1 Gbit/s, 2 Gbit/s, 4 Gbit/s, 8 Gbit/s, 10 Gbit/s, 16 Gbit/s, 20 Gbit/s.
iSCSI, транспорт SCSI через TCP/IP.
iSER, транспорт iSCSI через InfiniBand / RDMA.
SRP, транспорт SCSI через InfiniBand / RDMA
FCoE, транспортировка FCP/SCSI поверх «чистого» Ethernet.
FCIP и **iFCP**, инкапсуляция и передача FCP/SCSI в пакетах IP.
HyperSCSI, транспорт SCSI через Ethernet.
FICON транспорт через Fibre Channel (используется только мейнфреймами).
ATA over Ethernet (AoE), транспорт ATA через Ethernet.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Основные компоненты SAN

Fibre Channel HUBs (концентраторы) используются для подключения нод к FC кольцу (FC Loop) и имеют структуру, похожую на Token Ring концентраторы. Поскольку разрыв кольца может привести к прекращению функционирования сети, в современных FC концентраторах используются порты обхода кольца (PBC-port bypass circuit), которые разрешают автоматически открывать/закрывать кольцо (подключать/отключать системы, присоединенные к концентратору). Обычно FC HUBs поддерживают до 10 подключений и могут стекироваться до 127 портов на кольцо. Все устройства, подключенные к HUB, получают общую полосу пропускания, которую они могут разделять между собой.

Fibre Channel Switches (коммутаторы) имеют те же функции, что и привычные читателю LAN коммутаторы. Они обеспечивают полноскоростное неблокированное подключение между нодами. Любая нода, подключенная к FC коммутатору, получает полную (с возможностями масштабирования) полосу пропускания. При увеличении количества портов коммутированной сети ее пропускная способность увеличивается. Коммутаторы могут использоваться вместе с концентраторами (которые используют для участков, не требующих выделенной полосы пропуска для каждой ноды) для достижения оптимального соотношения цена/производительность. Благодаря каскадированию свичи потенциально могут использоваться для создания FC сетей с количеством адресов 224 (свыше 16 миллионов).

FC Bridges (мосты или мультиплексоры) используются для подключения устройств с параллельным SCSI к сети на базе FC. Они обеспечивают трансляцию SCSI пакетов между Fibre Channel и Parallel SCSI устройствами, примерами которых могут служить Solid State Disk (SSD) или библиотеки на магнитных лентах. Следует заметить, что в последнее время практически все устройства, которые могут быть утилизированы в рамках SAN, производители начинают выпускать с вмонтированным FC интерфейсом для прямого их подключения к сетям хранения данных.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Основные компоненты SAN

Для соединения компонентов в рамках стандарта Fibre Channel используют медные и оптические кабели. Оба типа кабелей могут использоваться одновременно при построении SAN. Конверсия интерфейсов осуществляется с помощью GBIC (Gigabit Interface Converter) и MIA (Media Interface Adapter). Оба типа кабеля сегодня обеспечивают одинаковую скорость передачи данных. Медный кабель используется для коротких расстояний (до 30 метров), оптический - как для коротких, так и для расстояний до 10 км и больше. Используют многомодовый и одномодовый оптические кабели. Многомодовый (Multimode) кабель используется для коротких расстояний (до 2 км). Внутренний диаметр оптоволокна мультимодового кабеля составляет 62.5 или 50 микрон. Для обеспечения скорости передачи 100 МБ/с (200 МБ/с в дуплексе) при использовании многомодового оптоволокна длина кабеля не должна превышать 200 метров. Одномодовый кабель используется для больших расстояний. Длина такого кабеля ограничена мощностью лазера, который используется в передатчике сигнала. Внутренний диаметр оптоволокна одномодового кабеля составляет 7 или 9 микрон, он обеспечивает прохождение одиночного луча.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Основные компоненты SAN

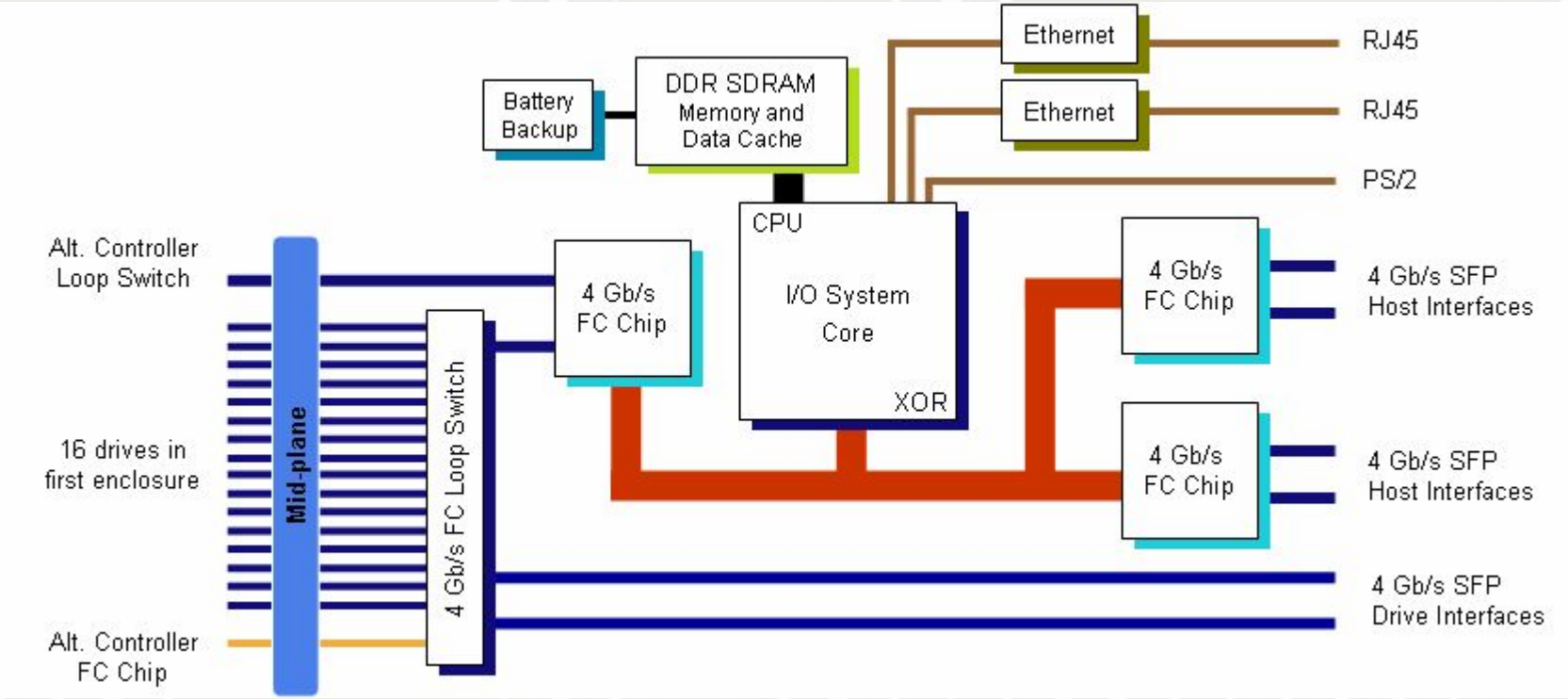
Основные ключевые особенности канальных:
Низкие задержки
Высокие скорости
Высокая надежность
Топология точка-точка
Небольшие расстояния между нодами
Зависимость от платформы
и сетевых интерфейсов:
Многоточечные топологии
Большие расстояния
Высокая масштабируемость
Низкие скорости
Программная загрузка
Большие задержки
объединились в Fibre Channel:
Высокие скорости
Независимость от протокола (0-3 уровни)
Большие расстояния
Низкие задержки
Высокая надежность
Высокая масштабируемость
Многоточечные топологии

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Типовая структура контроллера СХД

Рассмотрим структуру контроллера СХД на примере структурной схемы массива среднего уровня IBM DS4700



Мы видим контроллер среднего уровня, выполненного по классическому однопроцессорному принципу, без применения специализированного сигнального процессора.

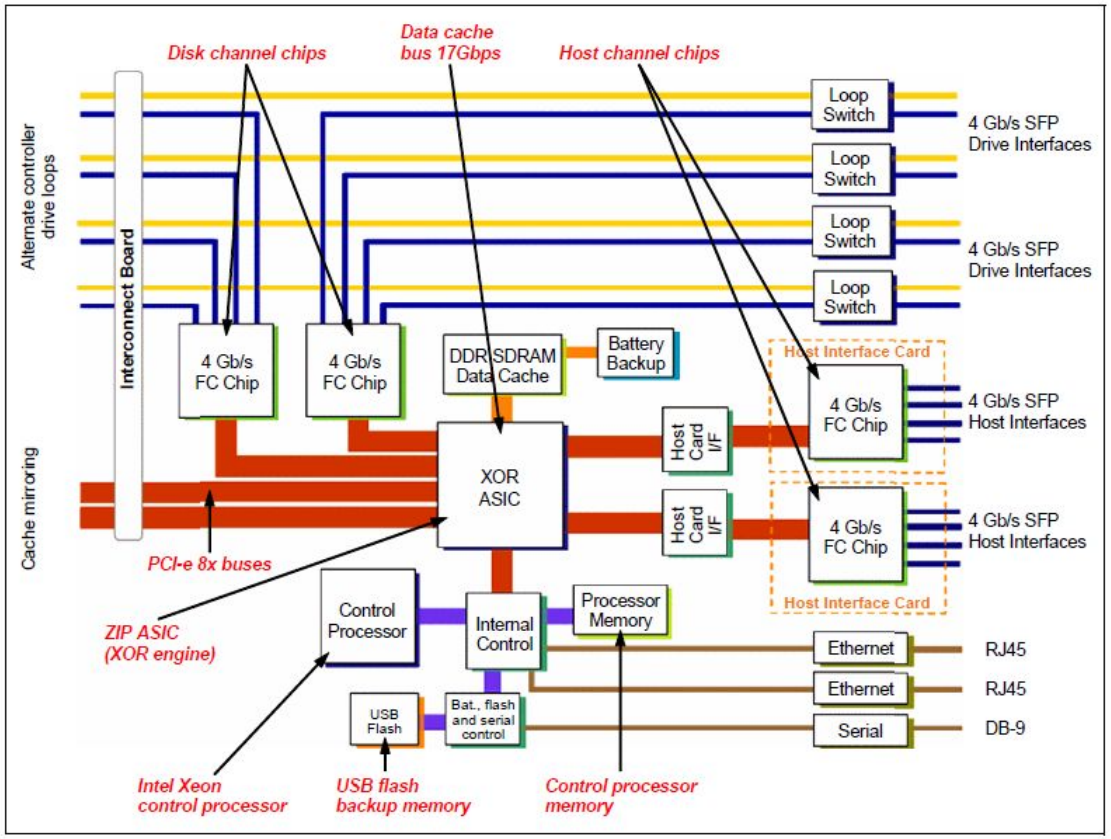
Здесь мы видим основной процессор, выполняющий функции управления и подсчета контрольных сумм, чипы хост-интерфейсов, кэш-память и чип подключения внутренних дисков. У всех основных производителей данная схема практически одинакова. Только могут различаться типы модулей ввода/вывода, тип и мощность процессора и конструктив модуля управления.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Типовая структура контроллера СХД

Рассмотрим структуру контроллера СХД на примере структурной схемы массива IBM DS5000



Здесь мы видим контроллер среднего уровня, выполненного уже в виде раздельных чипсетов управления и обработки сигналов.

Система хост-интерфейсов уже выполнена в виде модулей, что позволяет гибко подстраивать систему под необходимую топологию SAN

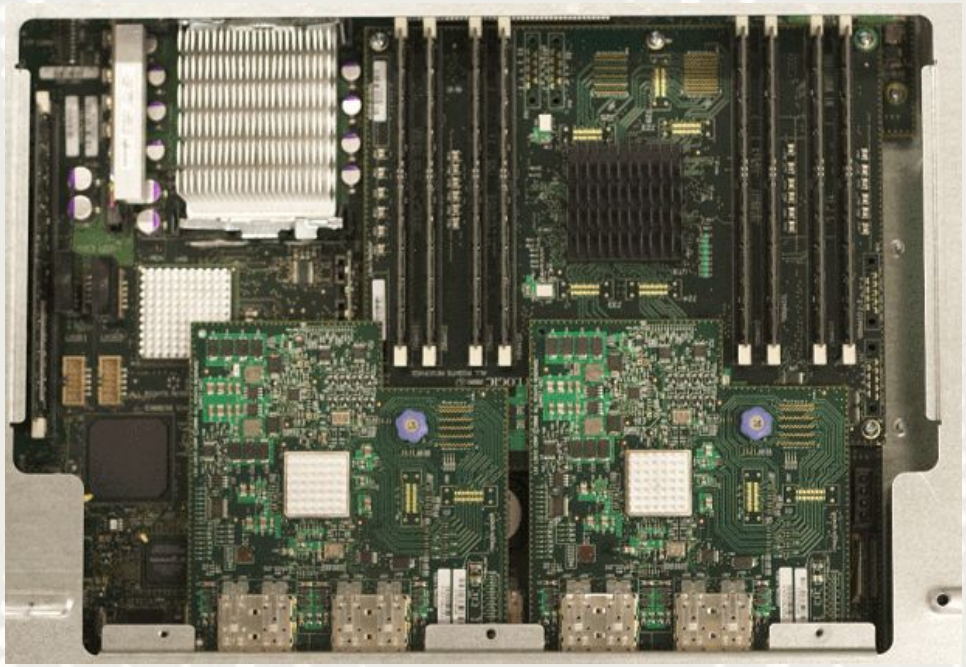
Так же появился мощный интерконнект между двумя контроллерами.

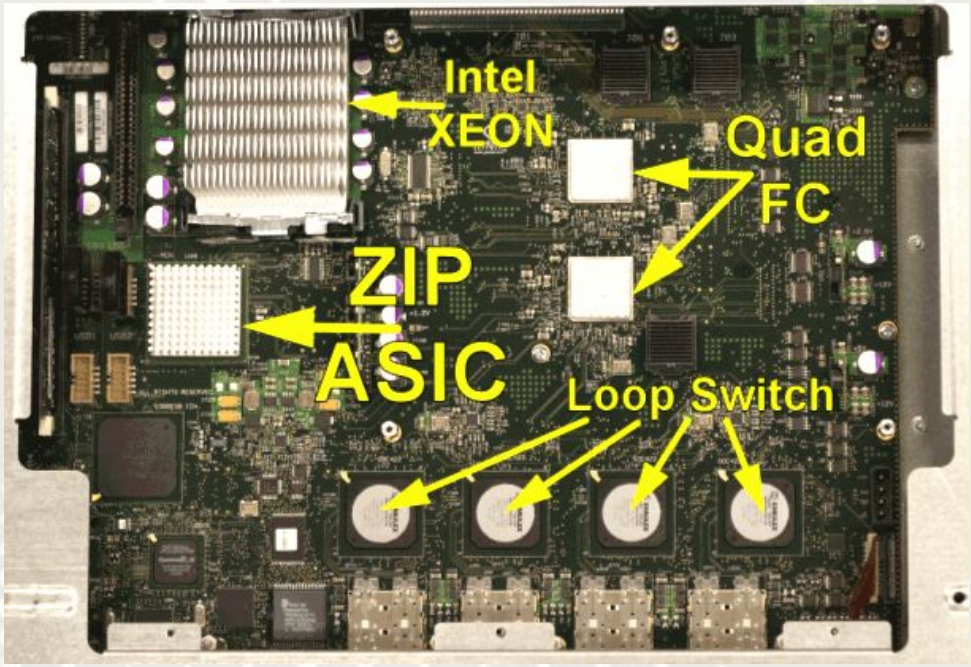
Энергонезависимая Flash-память заменила собой ранее используемые специальные разделы жестких дисков, которые использовались для сброса содержимого кэш-памяти во время пропадания питания.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Типовая структура контроллера СХД





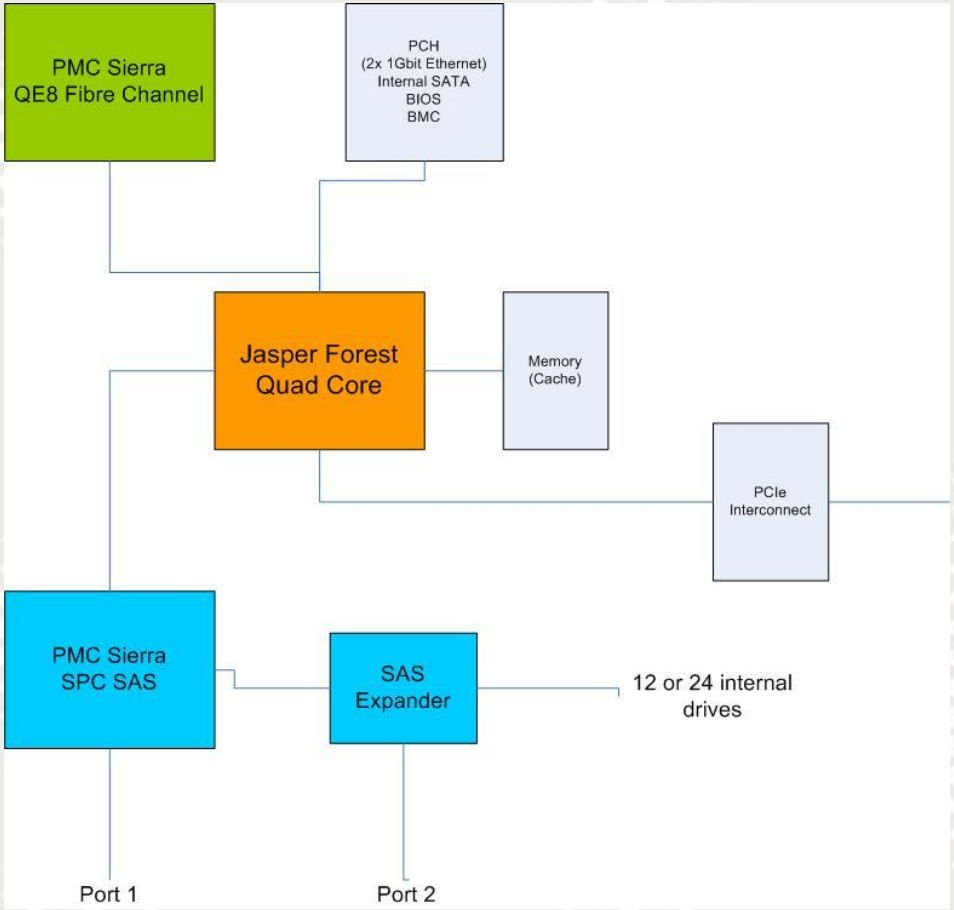
DS5000

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Типовая структура контроллера СХД

Структура виртуализированного модульного хранилища данных на примере IBM Storwize V7000



С повышением мощности процессоров, стало возможным создание модульных систем хранения данных, представляющих собой по сути дела кластер высокопроизводительных серверов с увеличенным кэшем и усиленной системой ввода/вывода

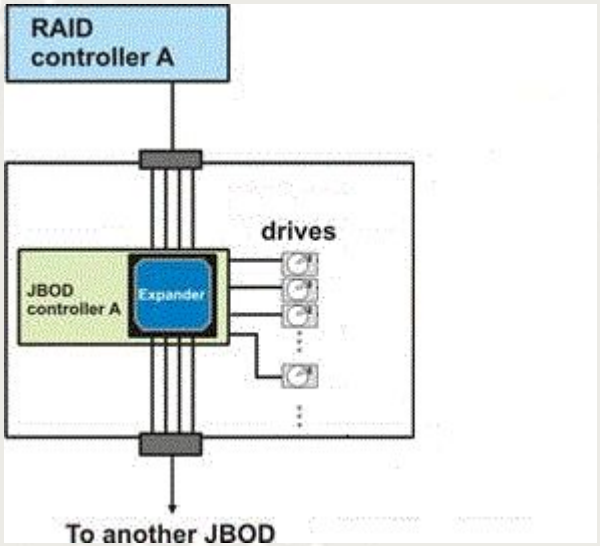
В отличие от традиционных СХД, при добавлении модулей увеличивается не только емкость хранилища, но и процессорная мощность СХД, что позволяет объединять гигантские объемы дисков в единую систему, производительность которой не ограничена полосой пропускания и пределами мощности контроллера.

ELCOM

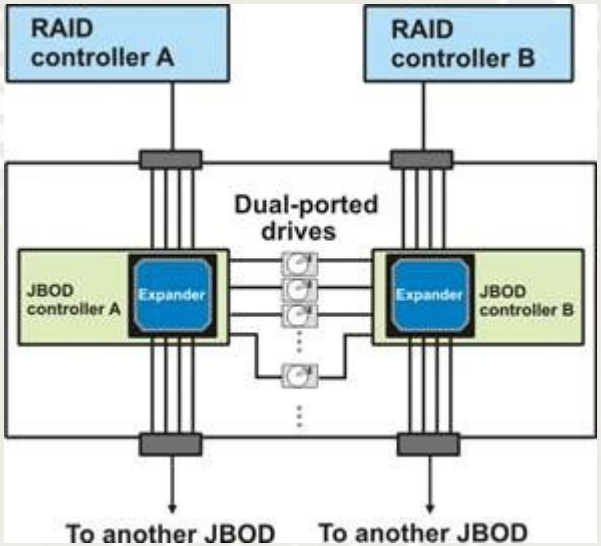
Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Разные типы подключения полок расширения СХД

Традиционно диски подключались на так называемые петли SCSI, и каждая полка JBOD представляла собой одну или две петли, на которые «нанизаны» диски. Сами петли просто выводились наружу на RAID контроллеры хостов и управлением дисками занимался сам хост.



Для того, чтобы можно было добиться отказоустойчивости, появились двухпортовые диски, к которым одновременно могли обращаться два контроллера. Что дало возможность создавать кластерные системы, чего на однопортовых дисках реализовать было невозможно.



При увеличении количества дисков появились проблемы с производительностью, в связи с ограничением шины SCSI. Вследствие чего встала необходимость управления дисками силами самого СХД. А для хостов стал выдаваться LUN, по сути дела представляющий собой виртуальный диск, собранный из RAID групп внутри СХД.

ELCOM

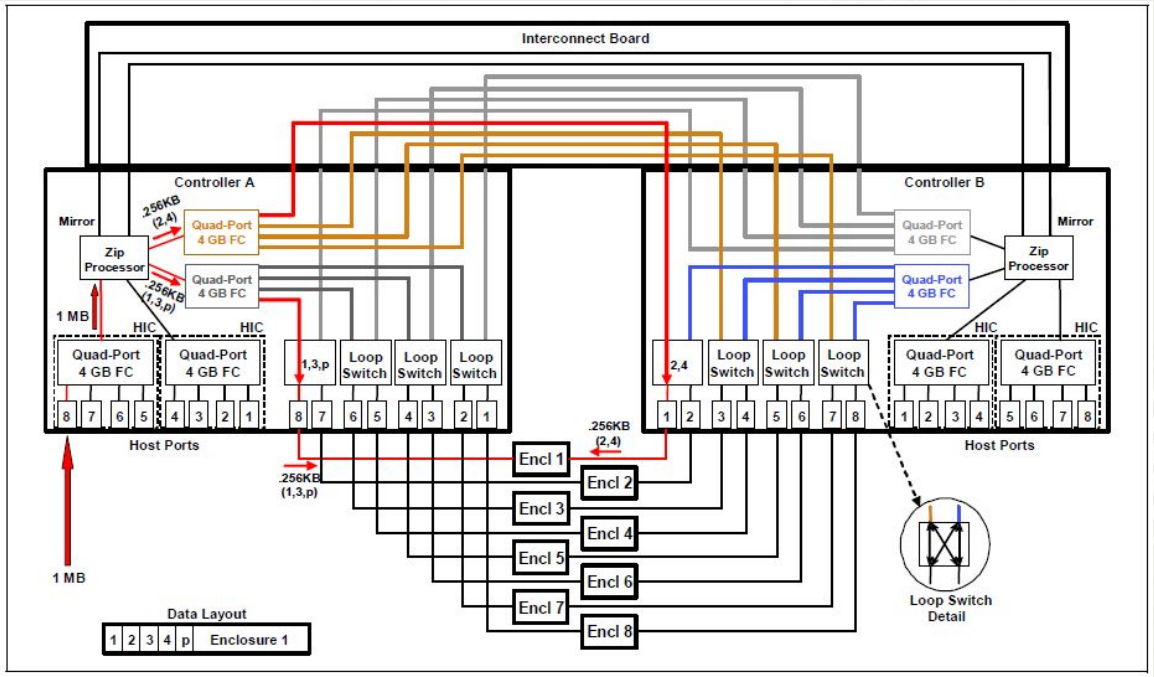
Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

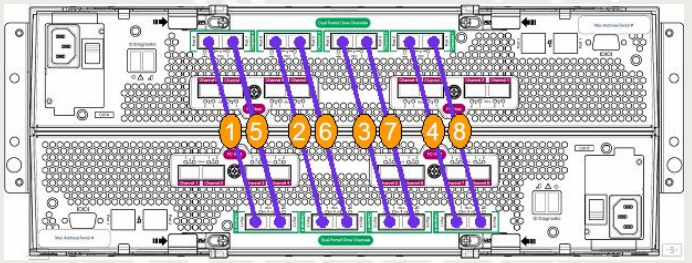
Разные типы подключения полок расширения СХД

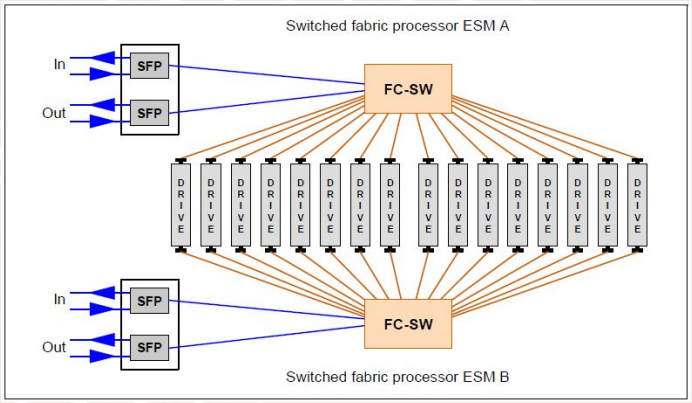
Так как проблема с выносом контроля за самими HDD решилась за счет контроллеров СХД, то стало возможным увеличить количество дисков в разы, что привело к новому ограничению. Внутренние дисковые петли SCSI не стали справляться с большим количеством дисков в силу конструктивных ограничений шины. Увеличение количества петель какое-то время позволяло решить эту проблему. Переход на FC сделал прорыв в технологии, т.к. позволил на одну петлю посадить гораздо больше дисков. Но дальнейшее увеличение количества дисков выявило проблему, которая выразилась в том, что из за «флуда» запросов в петле, ухудшается ширина полосы пропускания. И на рынке появились дисковые полки расширения со встроенными коммутаторами, которые в свою очередь сидят на петле, подключенной к контроллеру.

Коммутатор полки «видит» диски в режиме «точка-точка» либо в виде короткой петли. За счет двухпортовых контроллеров в полке можно разместить два коммутатора, которые будут резервировать друг друга.

В современных системах хранения данных петли «контроллер – коммутатор полки» могут быть выполнены в виде FC, SAS или eSATA. На современных уровнях развития SAS, их скорость (6 Гбит) уже вполне может конкурировать с FC.







ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Резервируемость и мультипассинг в СХД

В СХД любого уровня, одним из главных требований является гарантия работоспособности системы при выходе из строя определенных компонентов.

Практически все современные СХД имеют два контроллера, два блока питания, два контура подключения дисковых полок. В свою очередь в полках стоит по два коммутатора.

Контроллеры как правило имеют систему зеркалирования кэш-данных. Чтобы в случае выхода из строя одного из них, консистентность данных не нарушалась.

В системах с резервированием контроллеров в режиме активный-пассивный, практикуется закрепление дисков за контроллером. Что позволяет раздавать дисковые группы на обработку обоим контроллерам, а не держать один из контроллеров в холодном резерве.

Системы с контроллерами активный-активный позволяют использовать двойной канал для аггрегации потока данных, что позволяет работать с LUNом с удвоенной скоростью и при выходе из строя одного контроллера, не будет требоваться процедура переноса LUNов с неработающего контроллера на резервный, что иногда занимает время.

Системы с двумя активными контроллерами как правило используются на уровне hi-end массивов, но иногда встречаются в системах среднего уровня.

Немаловажную роль играет сама операционная система сервера, к которому подключается дисковый массив. Она должна поддерживать протокол обеспечения мультипассинга, совместимый с хранилищем. Основные вендоры как правило предоставляют к своим СХД пакеты драйверов для поддержки разных операционных систем.

Операционные системы могут обеспечивать мультипассинг различными механизмами, но они в основном все стандартизованы.

Особняком стоит процедура подключения дисков к кластерным системам. В них для работы кластеров используется механизм флагов, унаследованный от SCSI. Который позволяет и драйверу и контроллеру СХД знать, кому стоит отдать диск, а кому отрапортовать, что диск занят другим членом кластера.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

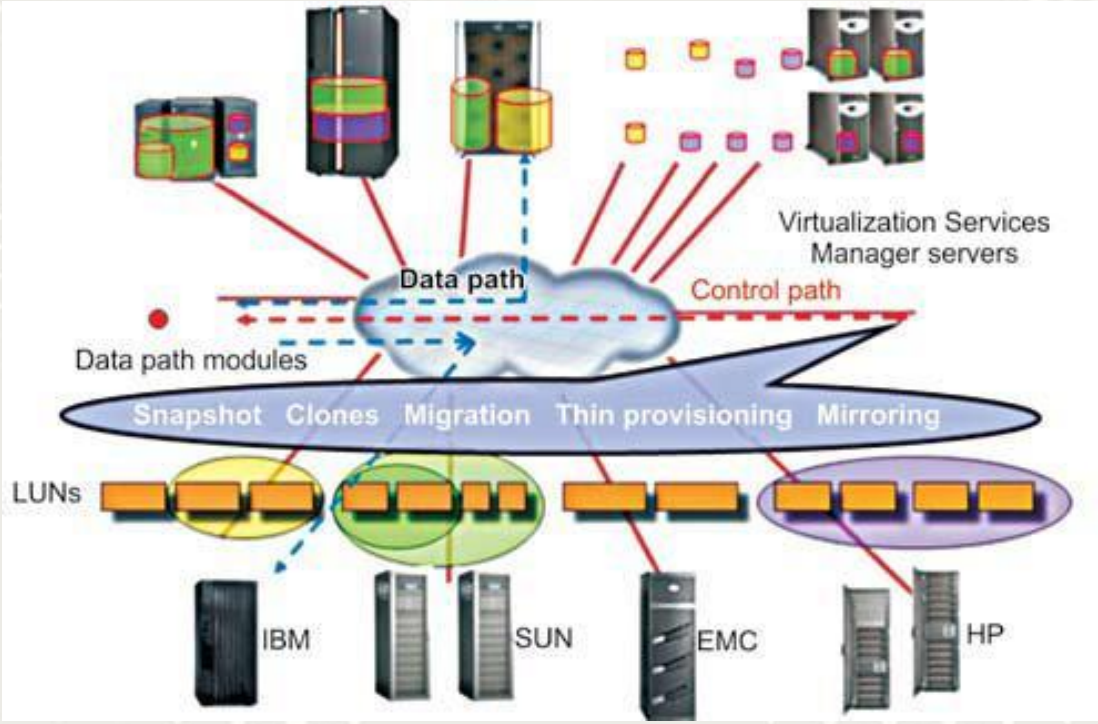
Виртуализация СХД

В настоящее время все большую популярность приобретает виртуализация систем хранения данных.

Главным образом это происходит по причине того, что сами жесткие диски устаревают медленнее, чем контроллеры. И зачастую оказывается, что функционал СХД становится недостаточным для полноценной работы вычислительной системы. Тогда применяются либо специализированные системы виртуализации, представляющие из себя кластерную систему ввода/вывода. Либо используются дисковые массивы, которые уже включают в себя функционал виртуализации.

Виртуализатор за счет своей более широкой полосы пропускания, большего объема кэш-памяти и большей вычислительной мощности, выполняет функции, до этого недоступные для виртуализируемых СХД, точнее их устаревшим контроллерам.

Так же расширяется функционал снапшотов, появляется возможность создания распределенных ЦОДов за счет функционала Metro-mirror и Global-mirror.



ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Дополнительный функционал систем хранения данных

Основные функции, присутствующие в СХД, могут быть дополнены дополнительными «бонусами»

Disk Encryption or FDE - Шифрование данных на лету, для конфиденциальных данных. Однако в России запрещены большинство из предлагаемых стандартов шифрования, предлагаемых ведущими вендорами.

Storage Partitioning - формирование виртуальной петли, в которой идет своя собственная нумерация LUNов. Для работы с кластерами крайне необходимая вещь. К тому же позволяет изолировать диски от не участвующих в кластере хостов, подключенных к СХД.

FlashCopy Logical Drives – функционал создания флэш-копий определенных LUNов, еще называемых снап-шотами. Позволяет делать «на ходу» копии данных на определенный момент. Эти «запаздывающие копии можно подключать как отдельные диски и по сути дела иметь бэкап данных. Очень удобно для девелоперов, т.к. позволяет вносить изменения в рабочую базу, и в случае проблем откатываться на ближайшую копию.

VolumeCopy - создание копий LUNa на другой LUN силами самого массива. Копирование происходит гораздо быстрее, нежели это делалось бы при помощи сервера, т.к. внутренняя шина контроллера гораздо производительней и к тому же нет нагрузки на сервер.

Remote Logical Drive Mirroring - удаленное зеркалирование логического диска.
Различают несколько типов зеркалирования:
Metro-mirror - синхронный метод зеркалирования, данные пишутся одновременно на оба массива.
Global-copy - асинхронный метод зеркалирования, данные пишутся на ведущий массив, хосту рапортуется об окончании записи, но запись на второй массив выполняется позже.
Global-mirror – асинхронный метод зеркалирования, данные пишутся на оба массива, и ведущий массив ждет от ведомого окончания записи, и только потом рапортует хосту об окончании записи.
FC/SATA Intermix - поддержка разных типов дисков в одном массиве или в одной дисковой полке.

Thin Provisioning – обеспечение дискового пространства с расширением объема «по требованию». Можно выделить хосту 3 Терабайта, в системе будет виден LUN на 3 ТБ , но фактически СХД будет хранить не 3 ТБ, а лишь тот объем, который фактически занят данными. В случае заполнения всего объема, масссив сам расширит область хранения, либо предупредит администратора о необходимости добавить дисков.

Дедупликация – автоматическое определение дуплицированных данных и хранение только отличающихся фрагментов файла.

ELCOM

Адрес: 197101, Санкт-Петербург, Кронверкский пр. 49
Телефон: (812) 325-2553
Факс: (812) 325-4115
E-mail: sales@elcom.spb.ru

Вопросы