МГУ ВМиК
Лаборатория Вычислительных Комплексов

UNIX

Системы хранения данных

30.05.08

Александр Герасёв <gq@cs.msu.su>

Содержание

- Информация правит миром
- Способы использования физических носителей (RAID, логические тома)
- Классы СХД
- Стандартные интерфейсы и протоколы
- Практический пример

Данные

- Рост объемов
- Децентрализация
- Необходимость масштабирования
- Стоимость
- Надежность
- Безопасность
- Сложность управления

Характеристики СХД

- Объем
- Механизм доступа
- Скорость доступа
- Отказоустойчивость
- Доступность
- Безопасность
- Сложность управления

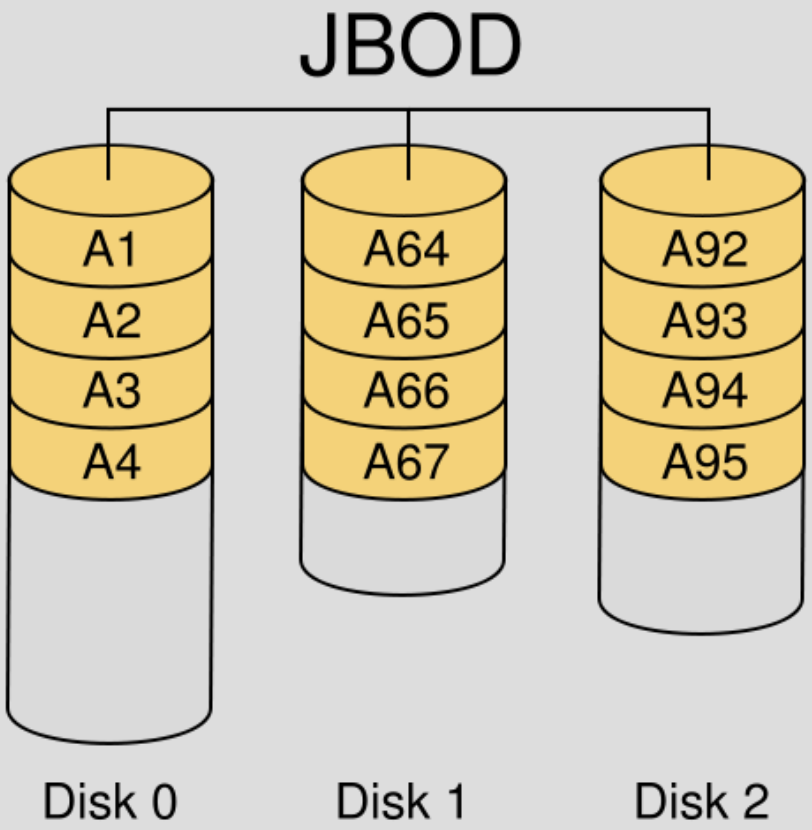
Отказоустойчивость

- Задачи
  - Сохранность данных
  - Обеспечение доступности
- Методы обеспечения отказоустойчивости:
  - Дублирование узлов
  - Избыточность

RAID

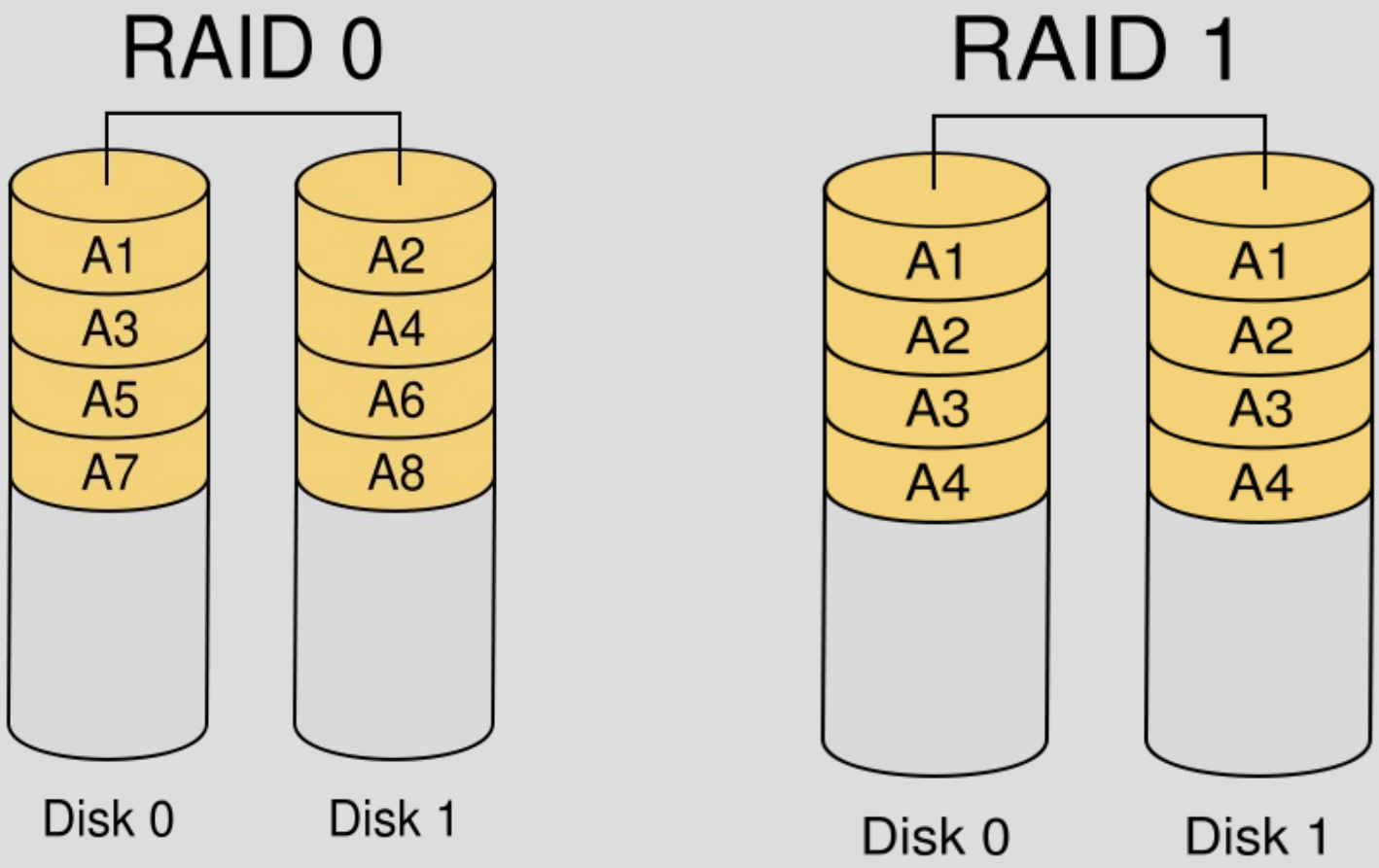
- JBOD
- RAID0
- RAID1-6
- Hot-spare
- Комбинации уровней

RAID

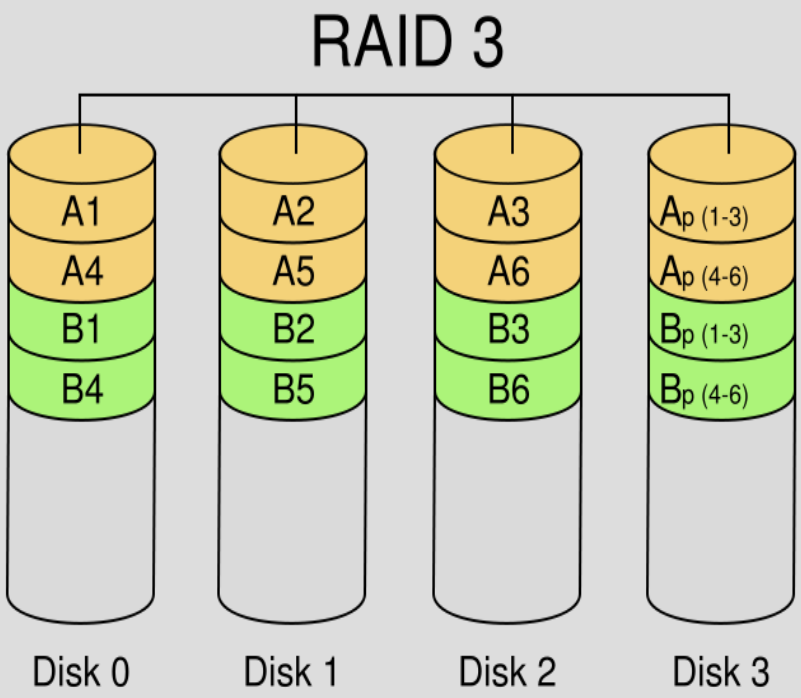


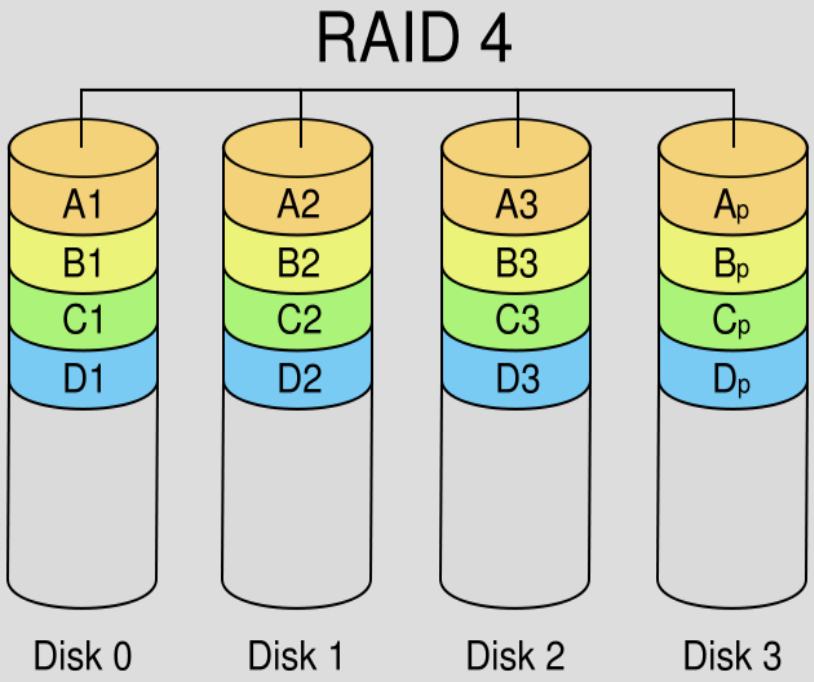
- RAID - Redundant Array of Independent/Inexpensive Drives/Disks

RAID 0 & RAID 1

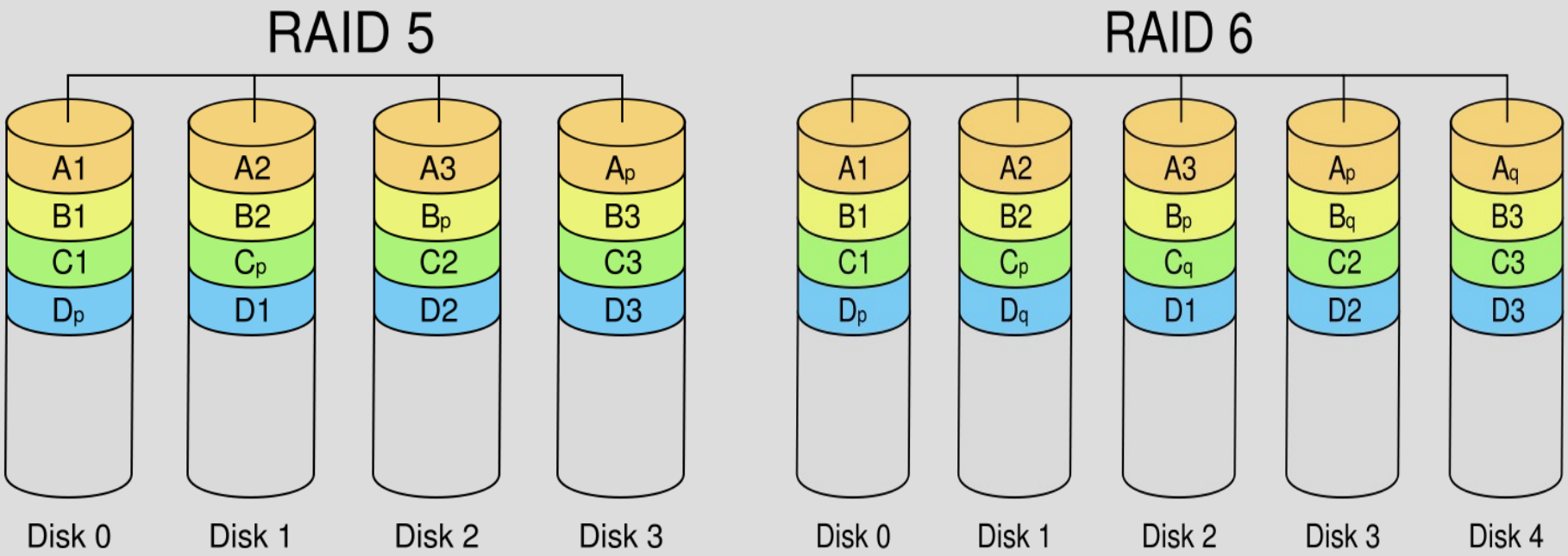

RAID 0
A1
A3
A5
A7
A2
A4
A6
A8
Disk 0
Disk 1
RAID 1
A1
A2
A3
A4
A1
A2
A3
A4
Disk 0
Disk 1

RAID 3 & 4


RAID 3
A1
A4
B1
B4
A2
A5
B2
B5
A3
A6
B3
B6
Ap (1-3)
Ap (4-6)
Bp (1-3)
Bp (4-6)
Disk 0
Disk 1
Disk 2
Disk 3


RAID 4
A1
B1
C1
D1
A2
B2
C2
D2
A3
B3
C3
D3
Ap
Bp
Cp
Dp
Disk 0
Disk 1
Disk 2
Disk 3

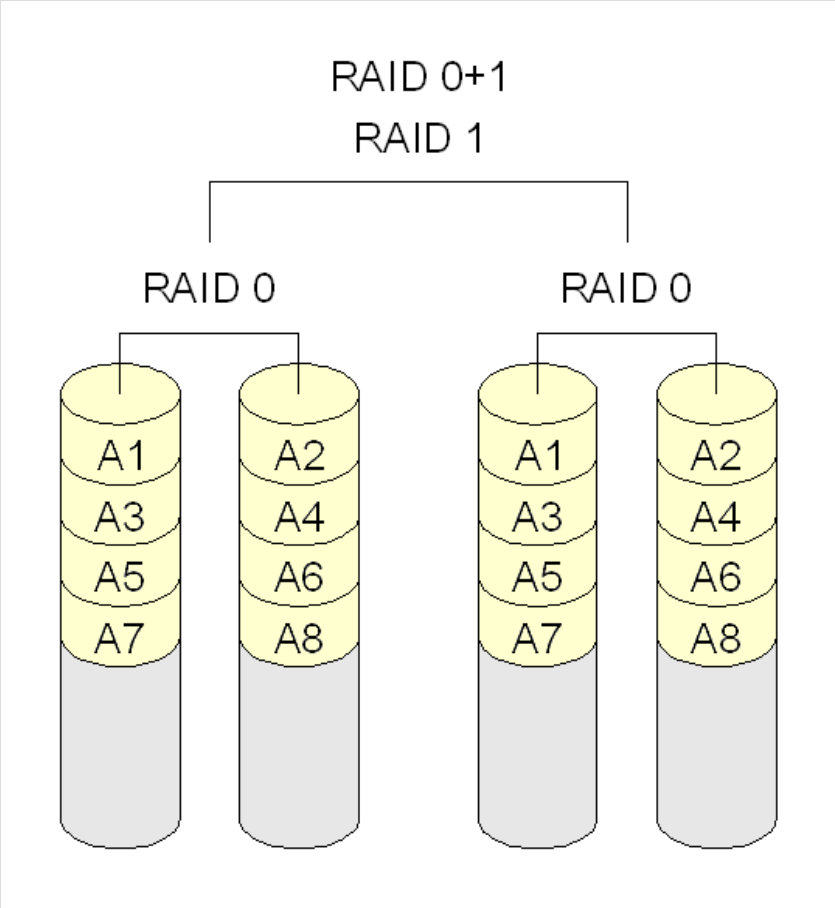
RAID 5 & 6

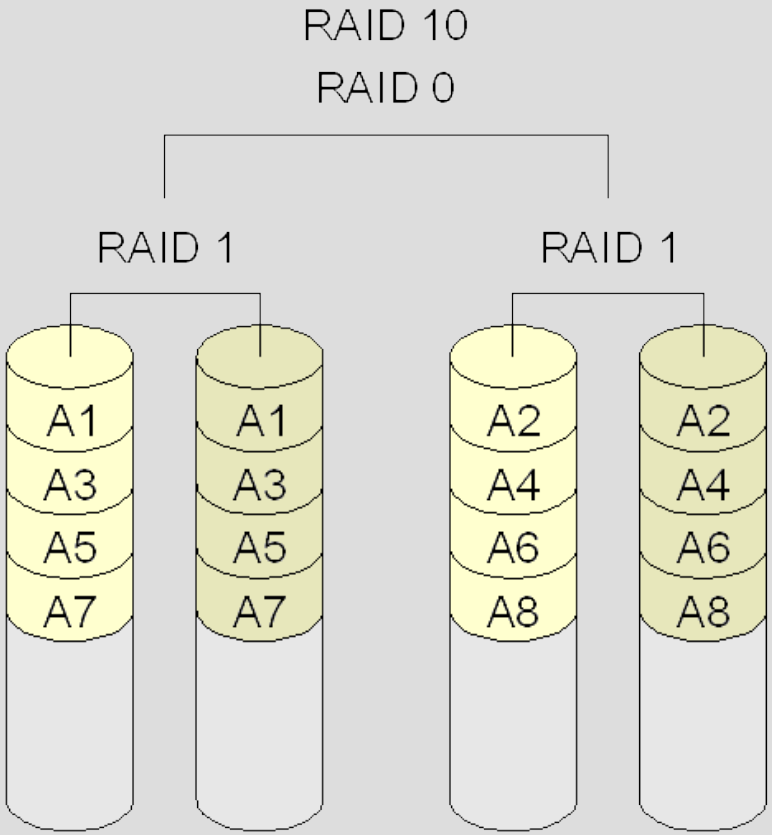


S=(N-1)*s

S=(N-2)*s

Комбинации уровней RAID





Комбинации уровней RAID

- RAID 0+1 vs RAID 10
- RAID 50
- RAID 100

- Hot-spare диски

MultiPath

- Отказоустойчивость
- Балансировка нагрузки

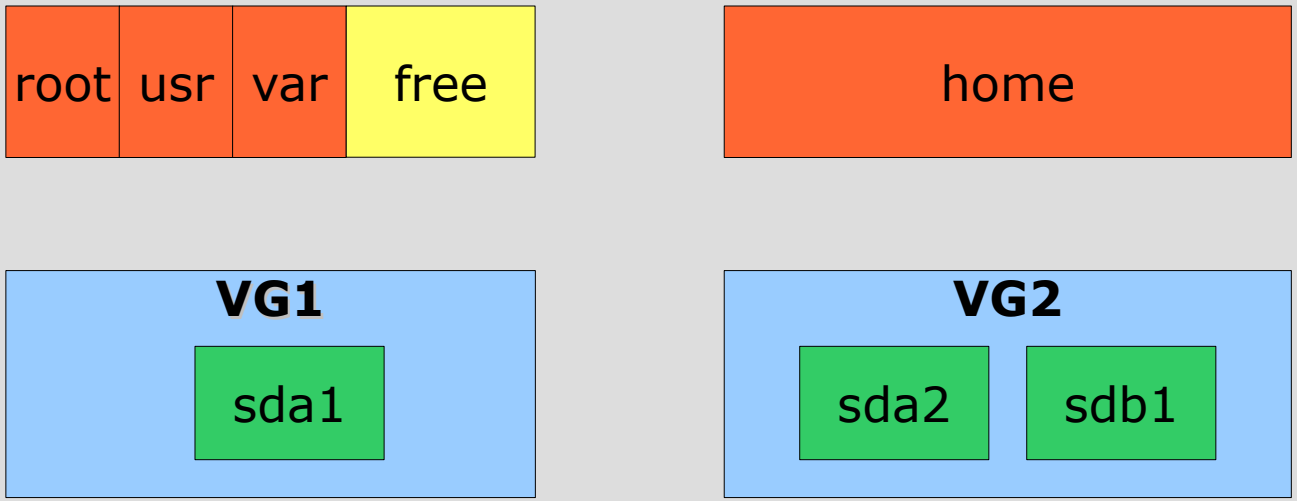
RAID-контроллеры

- Аппаратные
- Программные
- FakeRAID
- Встроенные в СХД

- Кэширование

LVM

- sda 200Gb
- sdb 200Gb
- /home 300Gb



Логические тома

Основные понятия:

- Физический том
- Группа логических томов
- Логический том

Свойства:

- Независимость от носителя и его размеров
- Динамическое изменение томов
- On-line изменение размеров
- Миграция между физическими устройствами
- Снэпшоты
- Зеркалирование

СХД

- Классы СХД
- Стандартные интерфейсы
- Пример СХД и практика использования

Классы СХД

- DAS
- NAS
- SAN

DAS

- Массивы дисков
- RAID
- Интерфейсы
- Логические диски

NAS

Основной тип доступа – доступ к файлам.

Протоколы:

- HTTP/FTP
- NFS
- SMB/CIFS
- AFS etc.

SAN

Storage Area Network

- Доступ к устройствам (RAW)
- Децентрализация
- Поставщики и потребители объединены сетью
- Возможность использования одного устройства несколькими потребителями

Тип сети

Физические интерфейсы:

- Ethernet
- FibreChannel

Протоколы:

- ATA over Ethernet
- NBD
- iSCSI
- FCP

iSCSI

- SCSI over IP
- Среда передачи данных
- initiator
- target
- Топология и гетерогенность

FibreChannel

- Физическая среда: оптоволокно
- Скорости: 1, 2, 4 Gbps
- Протокол: FCP – SCSI over FC
- Коммутация и топология
- FC-свитчи

Топология

- Точка-точка
- Каждый с каждым
- Switched Fabric
- Arbitrated Loop

Подключение СХД

- Аппаратные решения
  - SCSI/ATA
  - Host Bus Adapder (HBA)
- Программные решения
- Поддержка со стороны ОС

- Множественные пути доступа

Типы дисков, используемые в СХД

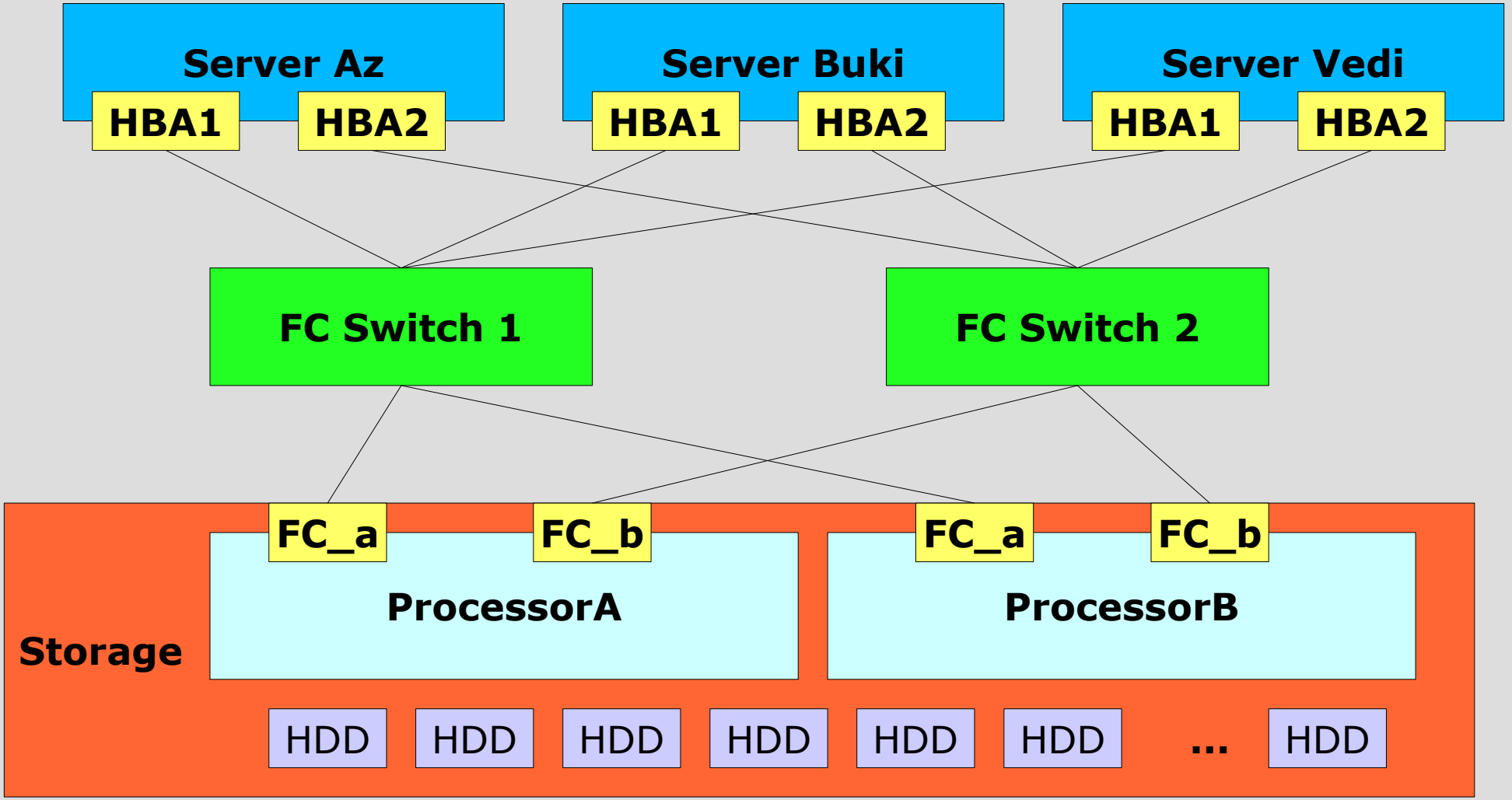
- IDE (PATA)
- SCSI
- SATA
- SAS
- FC

Пример СХД

- EMC CLARiiON AX150 2 processors + ИБП
- FC switches EMC (MCData) x2
- HBA Emulex 2Gbps 1port x6

- Сервер x86 x3

Аппаратная архитектура



Программная архитектура

- Диски объединены в RAID5 :*(
- Всё пространство разбито на 2 логических тома
- Тома экспортируются на сервера
- 8 томов объединены в 2 устройства при помощи MultiPath
- Поверх этих 2х устройств работает LVM (clvm)
- Тома на LVM используются Xen

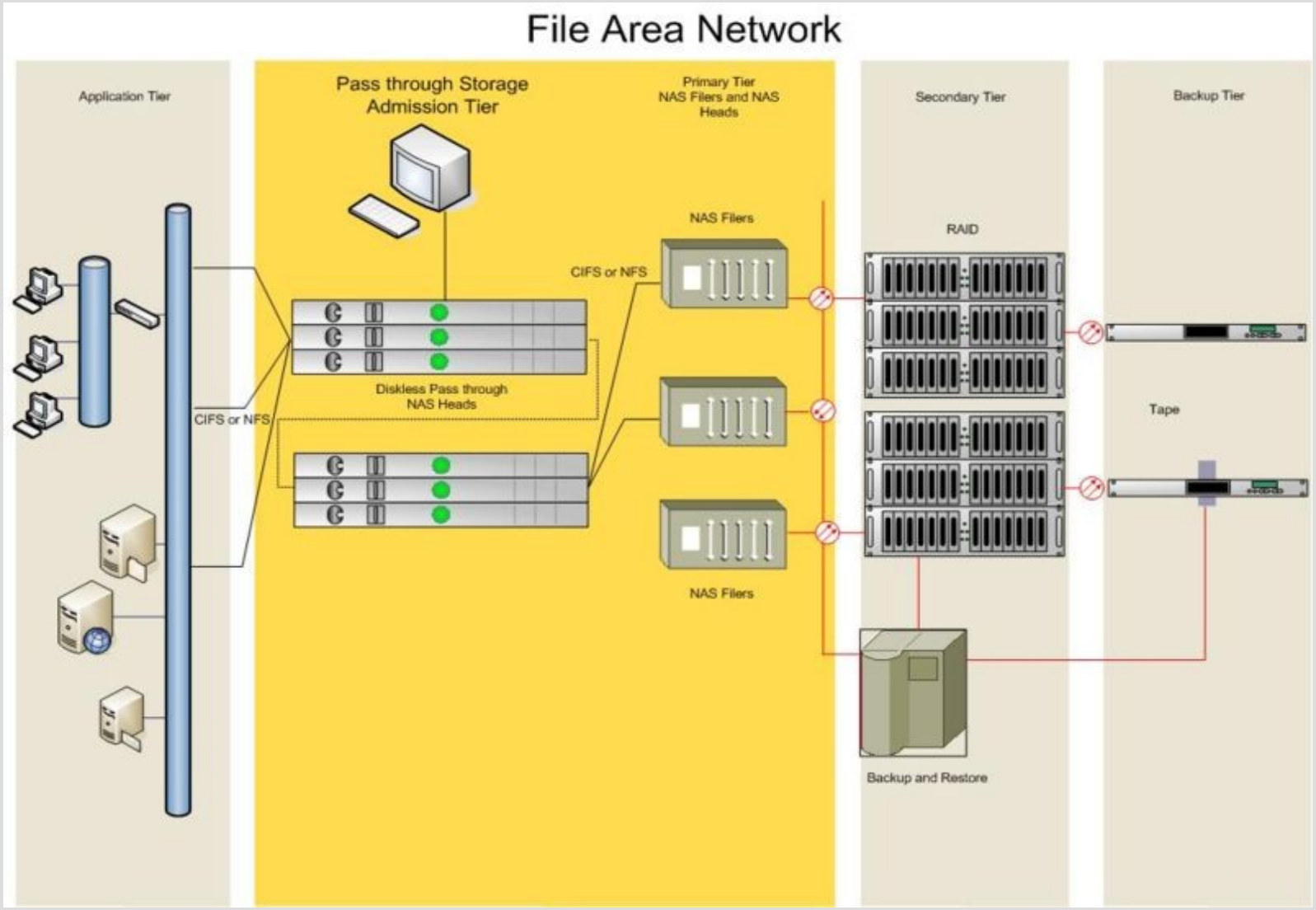
«Профессиональные» (enterprise) решения

- Простота развертывания
- «Лёгкость» управления vs. закрытость проприетарных инструментов и протоколов
- Производительность
- Аппаратные решения в области обеспечения отказоустойчивости

Технологии будущего

- Hierarchical Storage Management (HSM)
- Information Lifecycle Management (ILM)
- Tiered Storage Model
- Storage Admission Tier (SAT)
- File Area Network

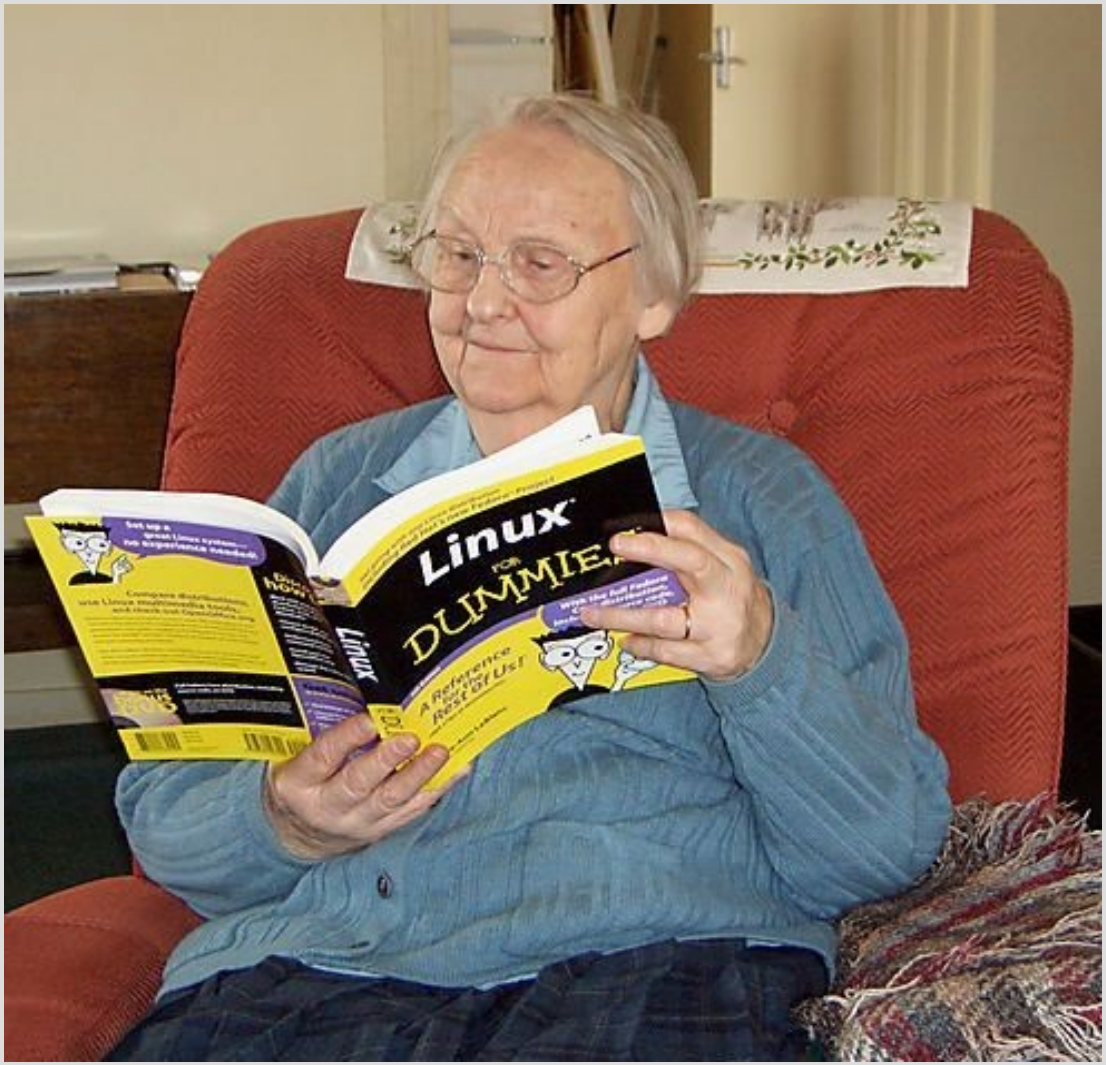
File Area Network



Цены

- DAS
  - Домашние: $100-$1000
  - Профессиональные: $500-$3000
- NAS
  - Домашние: $200-$1000
  - Профессиональные: $1000-$10000 (и выше)
- SAN
  - iSCSI: от $3000 и выше
  - FC: от $10000 до ∞

Спасибо за внимание

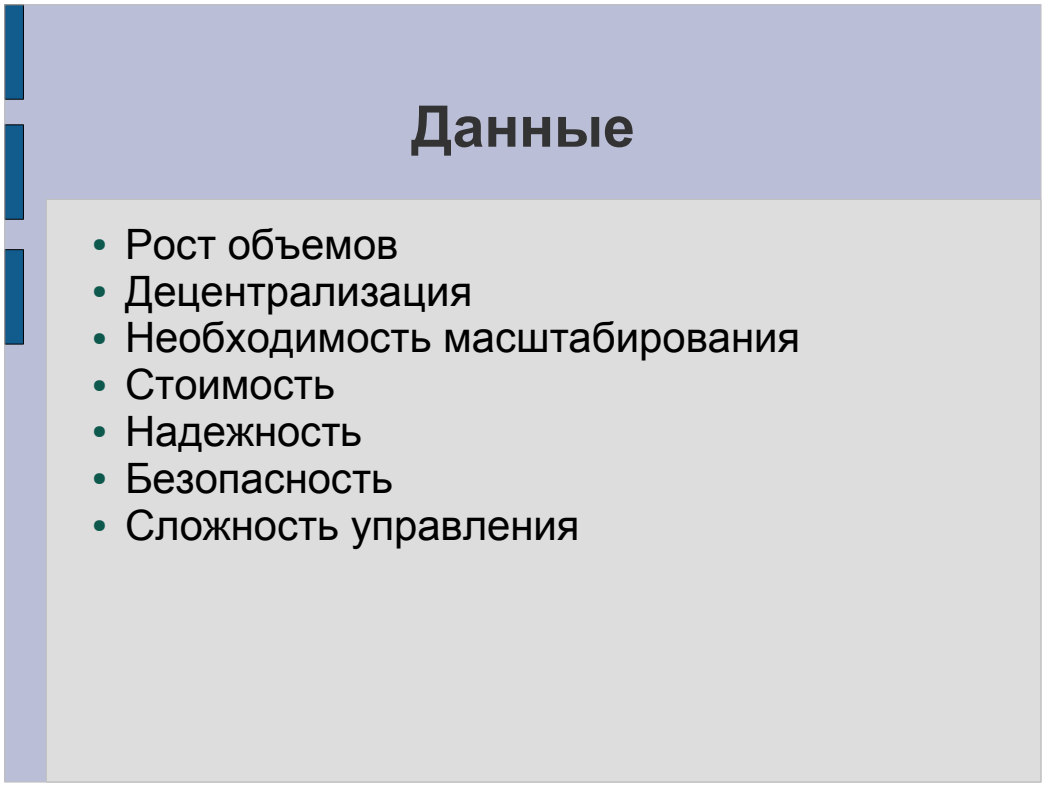


С чего вдруг лектор рассказывает про такие вещи: дело в том, что мы, то есть ЛВК — бюджетная организация, и как это бывает, внезапно под новый год сказали, что есть полтора миллиона рублей, и их надо срочно потратить. Поскольку списка, чего надо, не было, решили, что места мало и пользователи любят хранить всякую мультимедию, в итоге возникла потребность во внешнем хранилище. Поэтому пришлось окунуться в то, что такое СХД, конкретно SAN, и очень много было вопросов, потому что господа, которые при ..., у них немножко другая терминология, а нас интересовали больше всякие разные технические вещи, как оно работает и как его использовать в той инфрастурктуре, которую мы себе видели. Поэтому пришлось покопаться. Неоторое время назад лектор делал доклад на ЛВКшном студенческом семинаре о том, что я узнал, а это такая изменённая версия, без совсем простых вещей, но тем, не менее, версия того доклада.

Содержание

- Информация правит миром
- Способы использования физических носителей (RAID, логические тома)
- Классы СХД
- Стандартные интерфейсы и протоколы
- Практический пример

Попытаемся разобраться, для чего это надо, зачем хранить данные, откуда всё это взялось. Для начала несколько простых вещей, которые, тем не менее необходимы:
•Что такое RAID
•Что такое управление логическими томами (LVM)
•Про СХД, какие они бывают, какие там интерфейсы, протоколы используются, как всё это выглядит. В частности, расскажу, как мы это используем

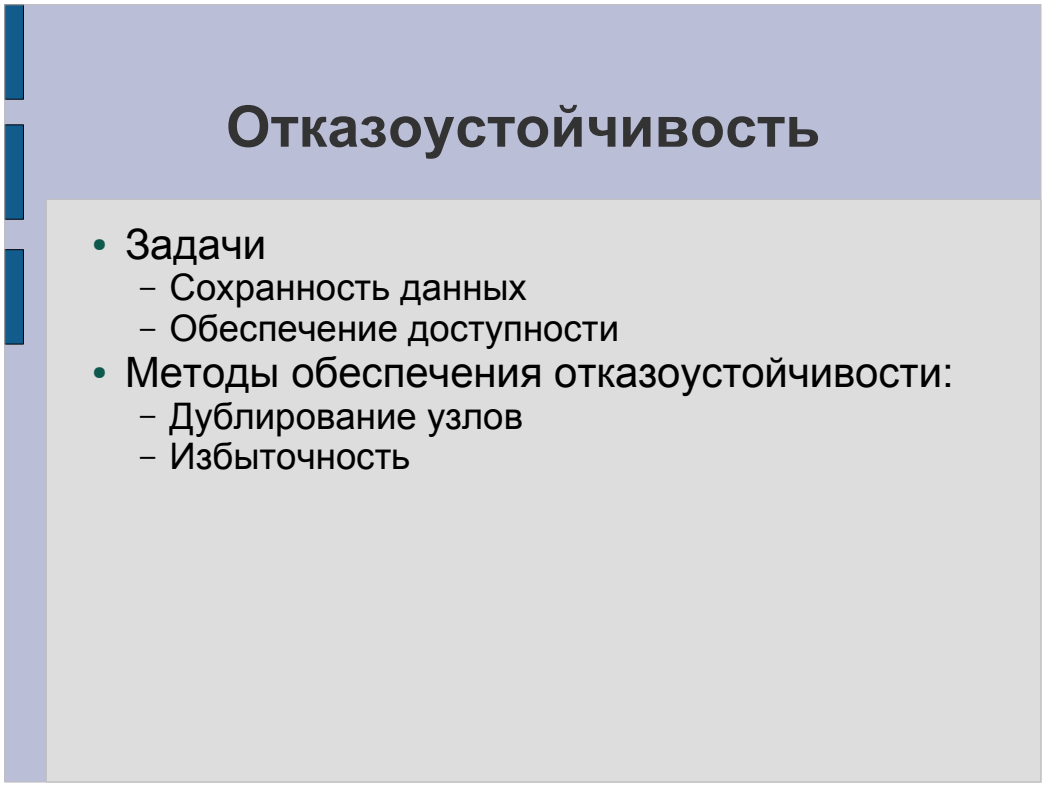


Понятно, что данных много, они всё время растут, их надо где-то хранить, при чём хранить надо так, чтобы они были доступну было из разны мест. При этом, мы не можем сразу решить, какой необходим объём дискового хранилища: сейчас нужен терабайт, через год надо будет два терабайта, сразу покупать 2 терабайта бессмысленно и дорого, поэтому возникает задача масштабирования, в итоге всё упирается в вопрос денег — мы не сразу можем приобрести большое хранилище. Помимо этого, требуется, чтобы данные хранились надёжно, безопасно, и чтобы всем этим можно было как-то управлять так, чтобы это не было чёрной магией. Для решения этих задач используют специализированные системы, которые занимаются как раз тем, что хранят данные, и дальше есть всевозможные потребители этих данных, которые их по-разному используют.

Отдельно отметим пункт с надёжностью, поскольку данные имеют очень высокую стоимость: это могут быть финансовые данные, данные по проектам, которые выполняются для того дяди, и если всё накроется, будет очень грустно.

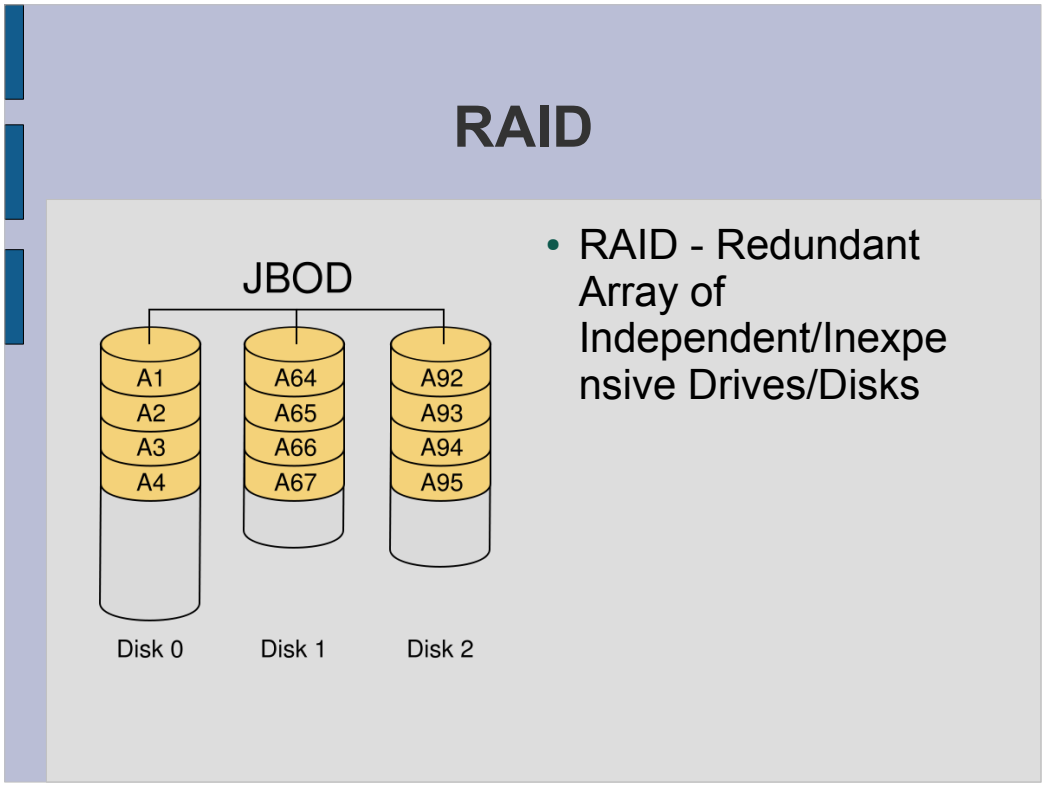
Характеристики СХД

- Объем
- Механизм доступа
- Скорость доступа
- Отказоустойчивость
- Доступность
- Безопасность
- Сложность управления

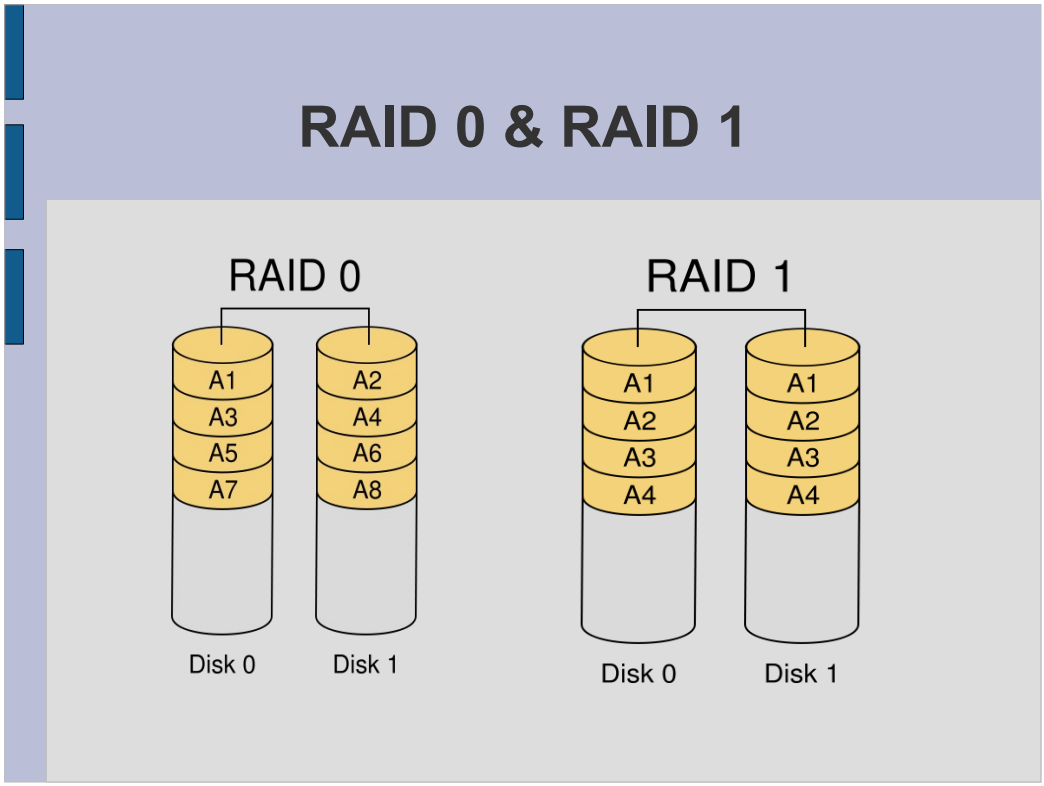


По поводу отказоустойчивости надо отдельно остановиться. Собственно, она делится на две основные задачи:
•Сохранность данных, то есть, чтобы наши данные, которые где-то хранятся, вдруг никуда не исчезли, не получилось так, что мы их потеряли
•Обеспечение доступности. Если есть какая-то система, которая эти данные как-то хранит, то возникает желание, чтобы эта система была доступна всегда. Если у нас, например, 10 серверов, которые используют эти данные, чтобы не было какого-то слабого звена
При этом избыточность может применяться на разных уровнях.

•JBOD (Just Bunch Of Disks). Есть просто много дисков, разных размеров, а надо одно большое пространство, мы все вместе объединяем и получается один виртуальный диск большого размера. Понятно, что объединять много дисков последовательно неинтересно, потому что когда данные читаются, они читаются последовательно, и всё читается с одного диска
•Но дисков много, их можно читать параллельно. Тогда и придумали RAID 0, когда диски делятся на страйпы, и подряд идущие данные лежат часть на одном, часть на другом, на третьем, и так далее. Это к обеспечению отказоустойчивости никак не относится, поскольку избыточности там нет никакой. Избыточность появляется дальше.
•Полное зеркалирование всех данных. Это RAID 1.
•Дальше начали применять коды Хэмминга (RAID 2). Когда есть несколько дисков, можно данные так распределить, чтобы избыточность образовалась.
•Сейчас наиболее используемые RAID 5—6, где данные блоками распределены между несколькими дисками

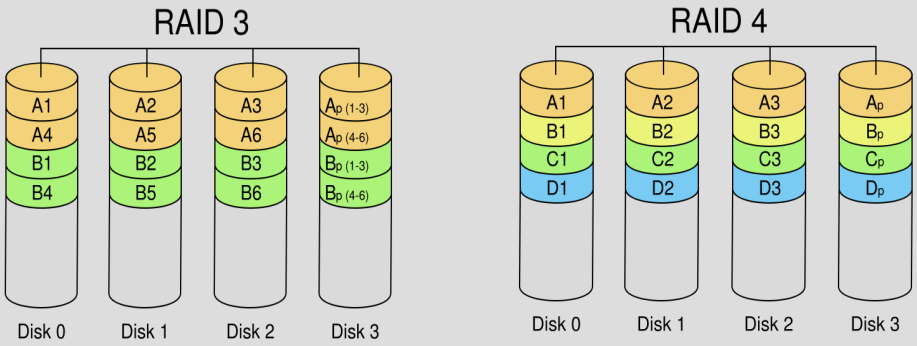


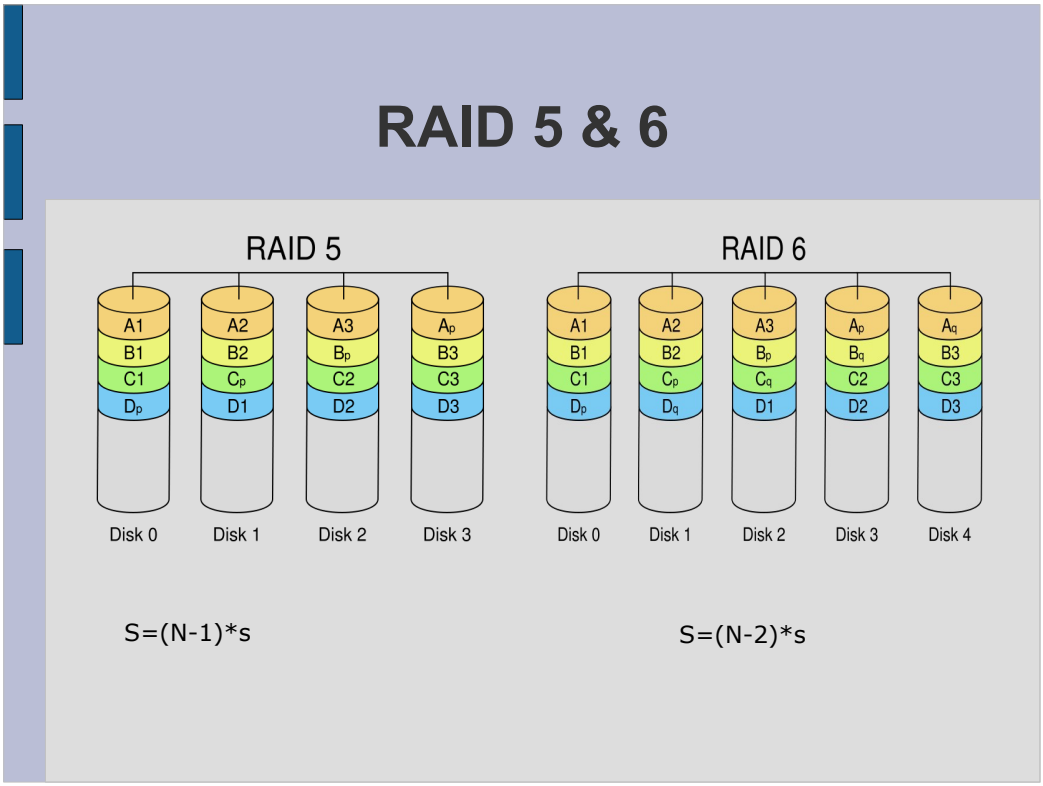
•JBOD (Just Bunch Of Disks). Есть просто много дисков, разных размеров, а надо одно большое пространство, мы все вместе объединяем и получается один виртуальный диск большого размера. Понятно, что объединять много дисков последовательно неинтересно, потому что когда данные читаются, они читаются последовательно, и всё читается с одного диска



•Но дисков много, их можно читать параллельно. Тогда и придумали RAID 0, когда диски делятся на страйпы, и подряд идущие данные лежат часть на одном, часть на другом, на третьем, и так далее. Это к обеспечению отказоустойчивости никак не относится, поскольку избыточности там нет никакой. Избыточность появляется дальше.
•Полное зеркалирование всех данных. Это RAID 1.

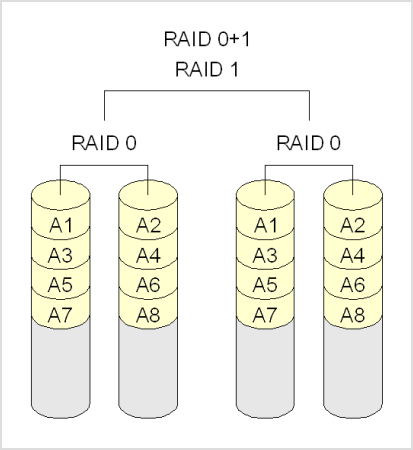
RAID 3 & 4

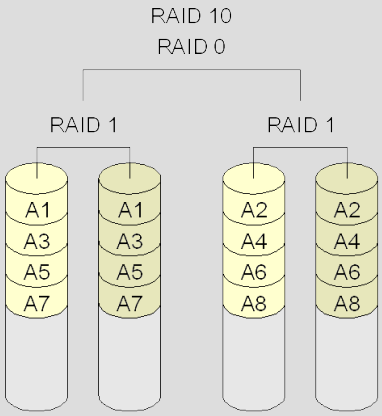

RAID 3
A1
A4
B1
B4
A2
A5
B2
B5
A3
A6
B3
B6
Ap (1-3)
Ap (4-6)
Bp (1-3)
Bp (4-6)
Disk 0
Disk 1
Disk 2
Disk 3
RAID 4
A1
B1
C1
D1
A2
B2
C2
D2
A3
B3
C3
D3
Ap
Bp
Cp
Dp
Disk 0
Disk 1
Disk 2
Disk 3



•Сейчас наиболее используемые RAID 5—6, где данные блоками распределены между несколькими дисками

Комбинации уровней RAID





Комбинации уровней RAID

- RAID 0+1 vs RAID 10
- RAID 50
- RAID 100

- Hot-spare диски

MultiPath

- Отказоустойчивость
- Балансировка нагрузки

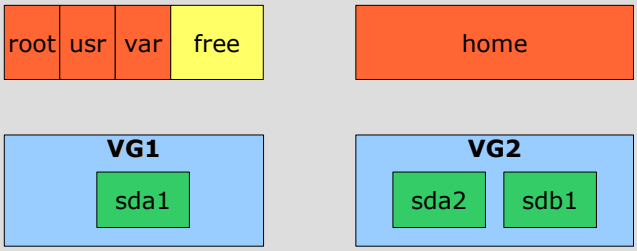
RAID-контроллеры

- Аппаратные
- Программные
- FakeRAID
- Встроенные в СХД

- Кэширование

LVM

- sda 200Gb
- sdb 200Gb
- /home 300Gb



Логические тома

Основные понятия:
- Физический том
- Группа логических томов
- Логический том

Свойства:
- Независимость от носителя и его размеров
- Динамическое изменение томов
- On-line изменение размеров
- Миграция между физическими устройствами
- Снэпшоты
- Зеркалирование

СХД

- Классы СХД
- Стандартные интерфейсы
- Пример СХД и практика использования

Классы СХД

- DAS
- NAS
- SAN

DAS

- Массивы дисков
- RAID
- Интерфейсы
- Логические диски

NAS

Основной тип доступа – доступ к файлам.

Протоколы:
- HTTP/FTP
- NFS
- SMB/CIFS
- AFS etc.

SAN

Storage Area Network

- Доступ к устройствам (RAW)
- Децентрализация
- Поставщики и потребители объединены сетью
- Возможность использования одного устройства несколькими потребителями

Тип сети

Физические интерфейсы:
- Ethernet
- FibreChannel

Протоколы:
- ATA over Ethernet
- NBD
- iSCSI
- FCP

iSCSI

- SCSI over IP
- Среда передачи данных
- initiator
- target
- Топология и гетерогенность

FibreChannel

- Физическая среда: оптоволокно
- Скорости: 1, 2, 4 Gbps
- Протокол: FCP – SCSI over FC
- Коммутация и топология
- FC-свитчи

Топология

- Точка-точка
- Каждый с каждым
- Switched Fabric
- Arbitrated Loop

Подключение СХД

- Аппаратные решения
  - SCSI/ATA
  - Host Bus Adapder (HBA)
- Программные решения
- Поддержка со стороны ОС

- Множественные пути доступа

Типы дисков, используемые в СХД

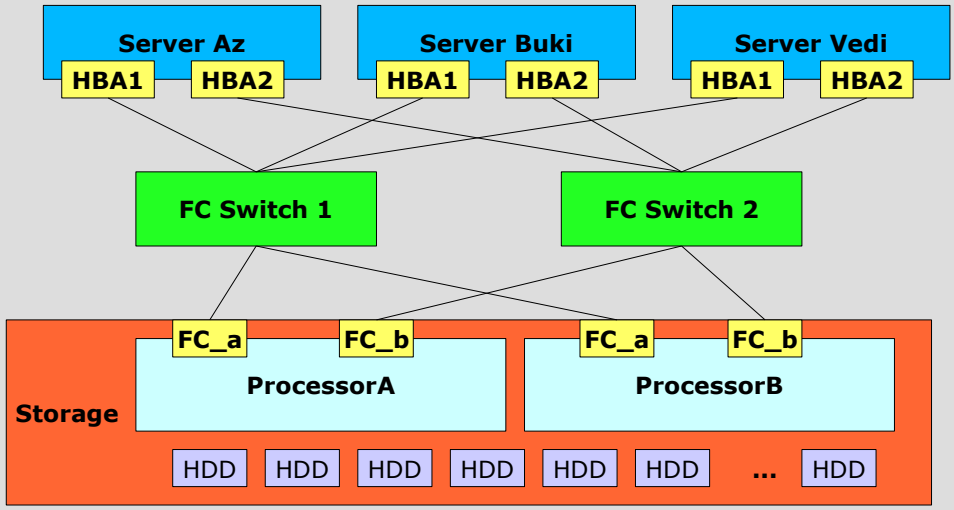
- IDE (PATA)
- SCSI
- SATA
- SAS
- FC

Пример СХД

- EMC CLARiiON AX150 2 processors + ИБП
- FC switches EMC (MCData) x2
- HBA Emulex 2Gbps 1port x6

- Сервер x86 x3

Аппаратная архитектура



Программная архитектура

- Диски объединены в RAID5 :*(
- Всё пространство разбито на 2 логических тома
- Тома экспортируются на сервера
- 8 томов объединены в 2 устройства при помощи MultiPath
- Поверх этих 2х устройств работает LVM (clvm)
- Тома на LVM используются Xen

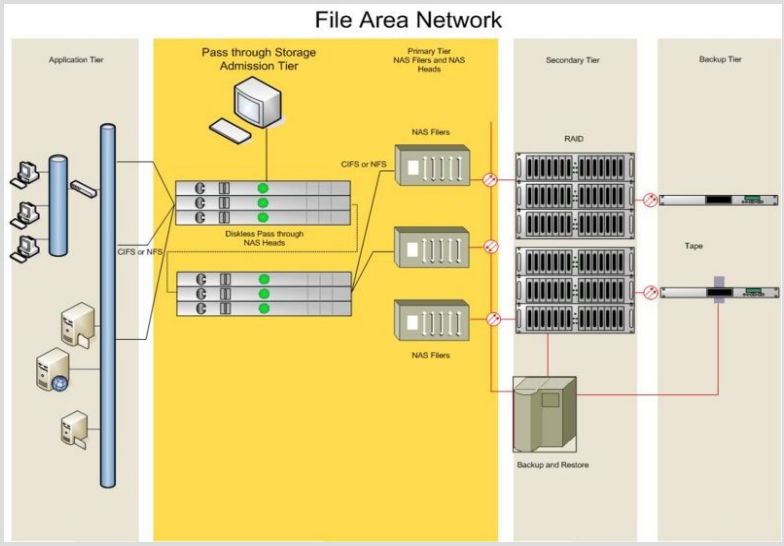
«Профессиональные» (enterprise) решения

- Простота развертывания
- «Лёгкость» управления vs. закрытость проприетарных инструментов и протоколов
- Производительность
- Аппаратные решения в области обеспечения отказоустойчивости

Технологии будущего

- Hierarchical Storage Management (HSM)
- Information Lifecycle Management (ILM)
- Tiered Storage Model
- Storage Admission Tier (SAT)
- File Area Network

File Area Network



Цены

- DAS
  - Домашние: $100-$1000
  - Профессиональные: $500-$3000
- NAS
  - Домашние: $200-$1000
  - Профессиональные: $1000-$10000 (и выше)
- SAN
  - iSCSI: от $3000 и выше
  - FC: от $10000 до ∞

Спасибо за внимание