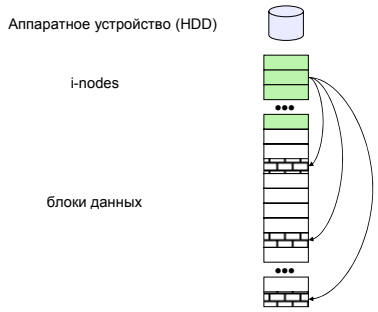
Лекция 16:

Файловые системы: заключение
RAID (Redundant Array of Inexpensive Disks)

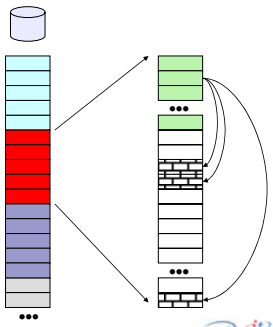
Алексей Линёв
Александр Мощук
Кирилл Погорельский

somes slides are adapted from the OS course at the University of Washington

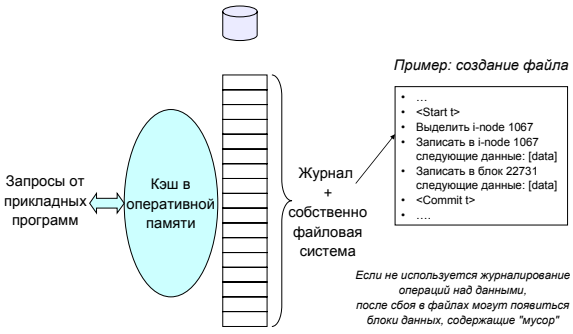
Оригинальная UNIX ФС (UFS)



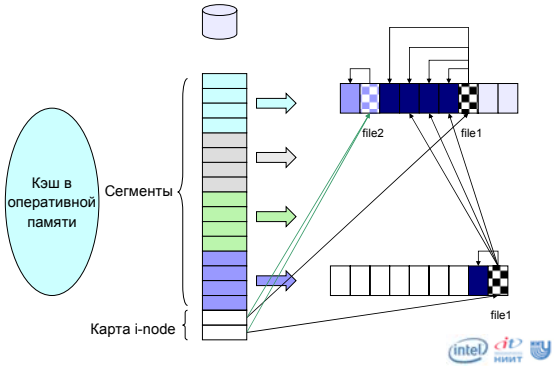
Fast File System (FFS)



Journaling File System (JFS)



Log-Structured File System (LFS)



Управление аппаратными ресурсами

- Каждый диск может быть разбит на несколько логических частей (разделов, partitions)
  - Уменьшаются потери при сбоях
  - Являются единицей резервного копирования
  - Могут содержать файловые системы различных типов
- Как совместно использовать несколько дисков?
  - Ни одна из рассмотренных файловых систем не рассчитана на размещение на нескольких дисках
    - Почему?
  - Вообще, имеет ли смысл использовать несколько дисков одновременно?
    - Увеличение объема
    - Увеличение производительности

Производительность

1. Скорость дискового ввода-вывода улучшается, но все-таки она растет намного медленнее производительности центральных процессоров
2. Для увеличения производительности мы можем использовать несколько дисков
   - разбив файл на части и разместив части на различных жестких дисках (чередуя использование различных дисков), мы сможем распараллелить ввод-вывод и улучшить время доступа к данным файла
3. Чередование (striping) уменьшает надежность
   - 10 дисков имеют значение характеристики MTBF (mean time between failures, среднее время между сбоями) примерно в 10 раз меньшее, чем 1 диск

Надежность

- Как правило, достаточно обеспечить устойчивость к сбою одного диска
  - Теоретически, вероятность того, что в ходе замены одного диска произойдет сбой в работе другого диска – низка
    - Практически, такое все-таки встречается
- Для увеличения надежности, на диски записывают избыточные данные
  - Мы вскоре обсудим этот вопрос
- Итак
  - Чередование может обеспечить производительность
  - Использование избыточных данных может обеспечить надежность, но вызовет потерю производительности

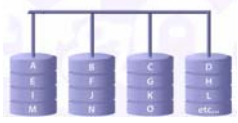
RAID

- RAID – Redundant Array of Inexpensive Disks (избыточный массив недорогих дисков)
- Диски невелики и недороги, достаточно просто разместить несколько дисков (десятки или сотни) в один блок, таким образом обеспечив увеличение объема, производительности и работоспособности
- Данные и некоторая избыточная информация некоторым образом (с чередованием) размещены на дисках
- Способ чередования размещения данных между дисками – ключевой момент в обеспечении производительности и надежности

RAID: выбор параметров

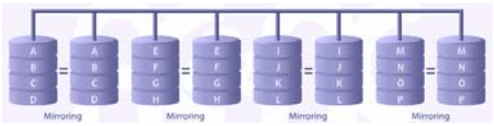
- Гранулярность (granularity)
  - мелкозернистый (fine-grained): каждый файл размещается на всех дисках
    - высокая пропускная способность для каждого файла
    - в каждый момент может передаваться только 1 файл
  - крупнозернистый (course-grained): каждый файл размещен на малом числе дисков
    - ограничивает пропускную способность при работе с одним файлом
    - позволяет одновременно получить доступ к нескольким файлам
- Избыточность (redundancy)
  - равномерное распределение избыточной информации по дискам
    - позволяет избежать проблем, связанных с балансировкой загрузки
  - сконцентрировать избыточную информацию на небольшом количестве дисков
    - разделить диски на 2 категории: диски с данными и диски с избыточной информацией

RAID Level 0: Non-Redundant Striping (чередование без избыточной информации)



- RAID Level 0 – это дисковый массив без избыточной информации
- Файлы распределены по нескольким дискам, избыточная информация отсутствует
- Высокая пропускная способность операций чтения (для одного файла)
- Наилучшая пропускная способность операций записи (не нужно записывать избыточную информацию)
- Сбой в работе любого диска ведет к потере данных
  - Что теряется?

RAID Level 1: Mirrored Disks (зеркалирование дисков)

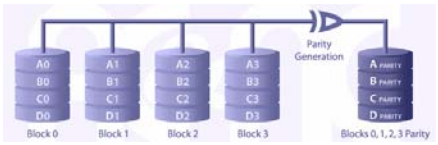


- Файлы распределены по половине дисков, и зеркалированы на другую половину
  - требуется в 2 раза больше дискового пространства
- Чтение:
  - Можно выполнять чтение с любой копии
- Запись:
  - Нужно записывать в обе копии
- В случае сбоя диска, просто используем его выжившую копию

Как это влияет на производительность?

Сколько одновременных дисковых сбоев может "пережить" массив?

RAID Levels 2, 3, 4: Striping + Parity Disk (чередование с диском контрольных сумм)



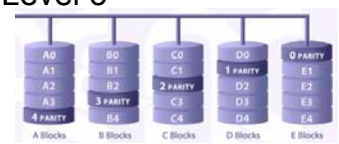
- RAID levels 2, 3, 4 используют диски, содержащие ECC (error correcting code, код корректировки ошибок) или код контроля четности
  - Например, каждый байт диска контроля четности содержит значение функции контроля четности от соответствующих байт со всех остальных дисков
- Объемные операции чтения выполняют доступ к данным со всех дисков
  - При чтении одного блока выполняется доступ только к одному диску
- При выполнении операций записи обновляется содержимое одного или нескольких дисков и диска, содержащего контрольные суммы
- Данные могут быть восстановлены при сбое на одном диске (Как?)
- Чем лучше ECC, тем больше устойчивость к сбоям, но требуется больше дисков для хранения контрольных сумм

Что такое код контроля четности?

1	0	1	1	0	1	1	0	1

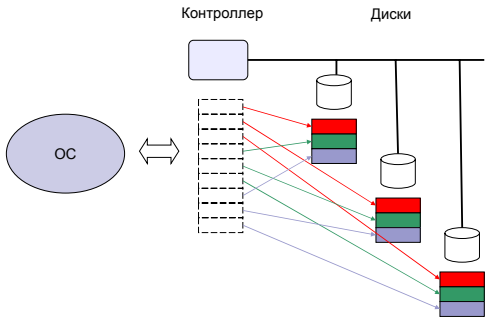
- Каждый байт дополняется битом, установленным таким образом, чтобы общее количество битов, содержащих единицы, стало четным
- Значение любого одного утерянного бита может быть восстановлено
- (Почему контроль четности при обращении к памяти не работает в точности так же?)
- Будем считать, что ECC работает примерно так же, но обладает большими возможностями

RAID Level 5

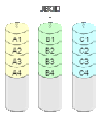


- RAID Level 5 использует контроль четности с чередованием размещения блоков контрольных сумм по дискам
- Похоже на схему с хранением кода контроля четности, но расположение контрольных сумм, как и данных, распределено по всем дискам
  - для каждого блока: один диск содержит код контроля четности, все остальные диски – данные
- Значительно лучшая производительность
  - отсутствует "бутылочное горлышко" в виде диска, хранящего коды контроля четности

Типичная реализация



JBOD – Just a Bunch Of Disks (группа жестких дисков)



- Фактически, это не RAID – нет ни чередования, ни избыточных данных
- Блоки различных дисков последовательно сцепляются и образуют единое логическое устройство
- Допускается использование дисков различного размера
- Поддерживается на аппаратном уровне (контроллером RAID)
  - Еще один популярный подход: LVM - Logical Volume Manager (например в Linux), но там переадресация блоков реализуется в ОС

Выводы

- Если вы используете RAID с достаточной степенью избыточности, нужно ли вам резервное копирование (backup)?
  - В чем разница между RAID'ом и резервным копированием?
- Обеспечивает ли RAID "достаточную" надежность?
  - Представьте, что вы поддерживаете базу данных Amazon.com

Уровень I
Единственный способ обеспечения подачи электроэнергии и охлаждения, отсутствует избыточность компонент, работоспособность – 99.671%
Уровень II
Единственный способ обеспечения подачи электроэнергии и охлаждения, имеются избыточные компоненты, работоспособность – 99.741%
Уровень III
Несколько путей подачи электроэнергии и охлаждения, но только один активный в каждый момент, имеются избыточные компоненты, работоспособность – 99.982%
Уровень IV
Несколько активных путей подачи электроэнергии и охлаждения, имеются избыточные компоненты, работоспособность – 99.995%