

Системы хранения данных

Васенина Анна Игоревна

18 мая 2023 г.

Дисклеймер

Английские термины в основном приведены для тех, кто будет готовится к экзамену по презентации — так проще гуглить. Знать термины, не указанные в билете, не обязательно

О докладчике

- В 2018-2021 гг. — R&D разработчик в ООО «Рэйдикс»
- В 2021 г. — и.о руководителя группы R&D
- Бакалавр математического обеспечения, магистр программной инженерии
- В настоящее время аспирант кафедры системного программирования, инженер-программист в ООО «Битблэйз Технологии»

Содержание

1 Определение системы хранения данных

2 Доступность

3 Производительность

4 Надежность

СХД и где они обитают



Горный Китай, монастырь Чжоан-Чжоу, год от Рождества Христова 2004-ый. Некто спросил Лин Цзы, что есть мать. И мастер ответил: «Северный и южный мосты есть мать. И шина есть мать. И ещё два десятка конденсаторов есть мать. И много чего ещё есть мать»

СХД и где они обитают



Горный Китай, монастырь Чжоан-Чжоу, год от Рождества Христова 2004-ый. Некто спросил Лин Цзы, что есть мать. И мастер ответил: «Северный и южный мосты есть мать. И шина есть мать. И ещё два десятка конденсаторов есть мать. И много чего ещё есть мать»

Так же и с СХД. Очень много чего есть СХД

Неформальное определение СХД

Нет!

- Файловая система
- База данных

Да!

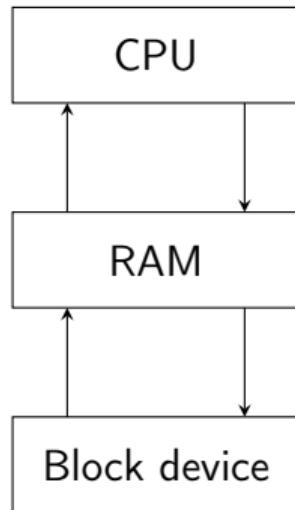
- Сложно организованное блочное устройство

Блоchное ???

Сложно организованное ???

Блочное устройство

- Случайный доступ
- Чтение и запись блоками
- Блок фиксированного размера
 - 512 B
 - 4 KiB
- На современных устройствах большой объем данных
- Работа через оперативную память (DMA — direct memory access)



Сложная организация



Содержание

1 Определение системы хранения данных

2 Доступность

3 Производительность

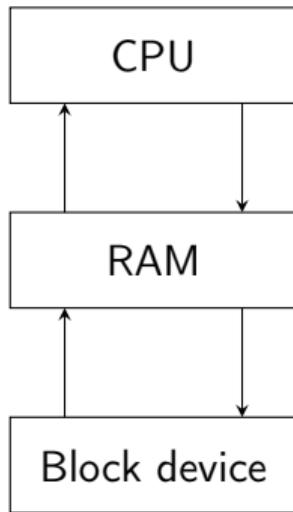
4 Надежность

DAS NAS SAN



Прямое подключение (DAS)

DAS — directly attached storage



Подключение по сети (NAS)

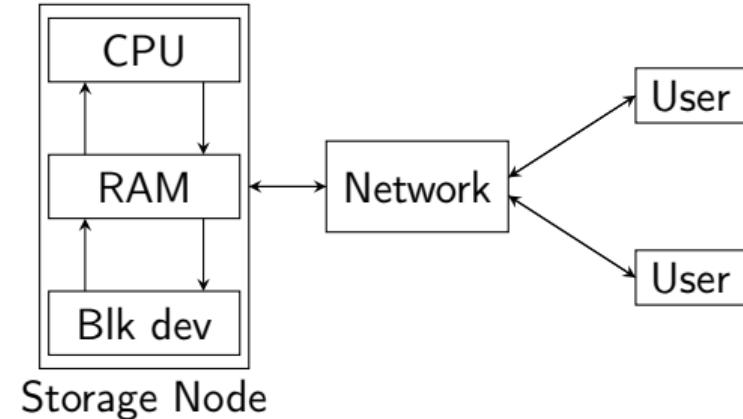
NAS — network attached storage

В домашнем использовании

- Несколько файловых систем разного типа или с полнодисковым шифрованием
- Обход ограничения материнских плат по числу дисков

В промышленном использовании

- Видеонаблюдение
- Сбор информации с датчиков (напр. для автопилота)



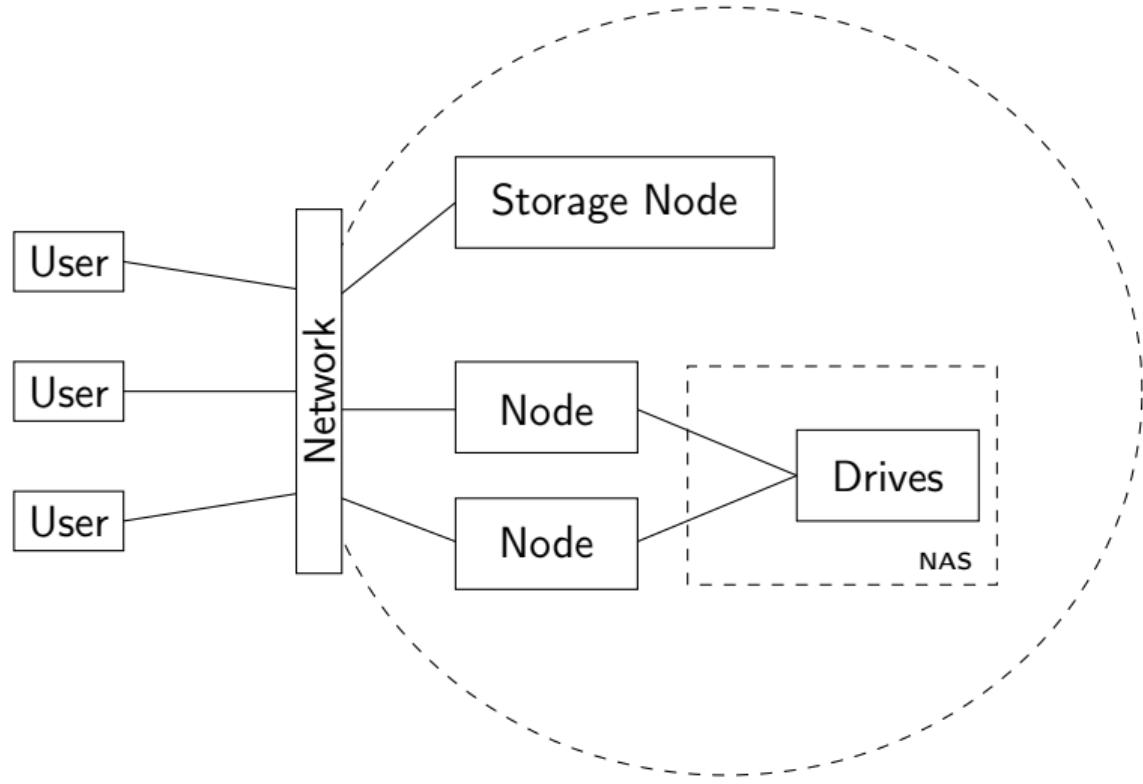
Сеть хранения данных (SAN)

SAN — storage area network

Прежде всего это именно
сетевое ПО

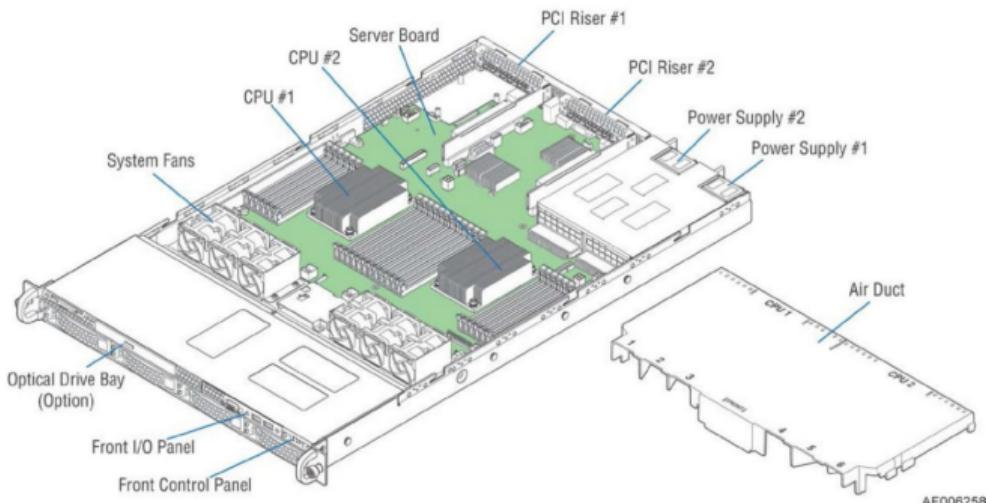
Примеры
функциональности

- Маршрутизация
- Аварийное
переключение
(failover)
- Репликация



Аппаратное обеспечение сервера хранения данных

- Дополнительные возможности подключения
 - Для сетевых карт
 - Для дисков
- Быстрая замена накопителей без разбора сервера и отключения питания
- Несколько сокетов для процессоров
- Несколько блоков питания с возможностью быстрой замены



Источник: Intel® Server System R1000WT Product Family. Intel® Storage System R1000WT Family. Technical Product Specification

Содержание

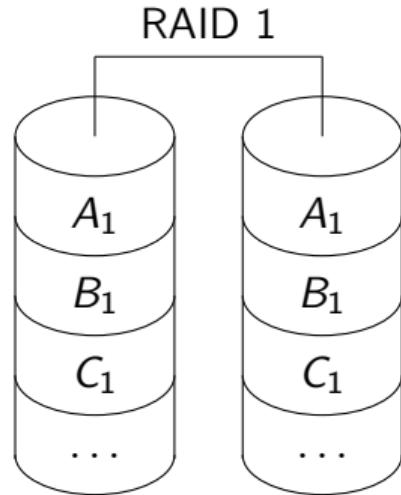
1 Определение системы хранения данных

2 Доступность

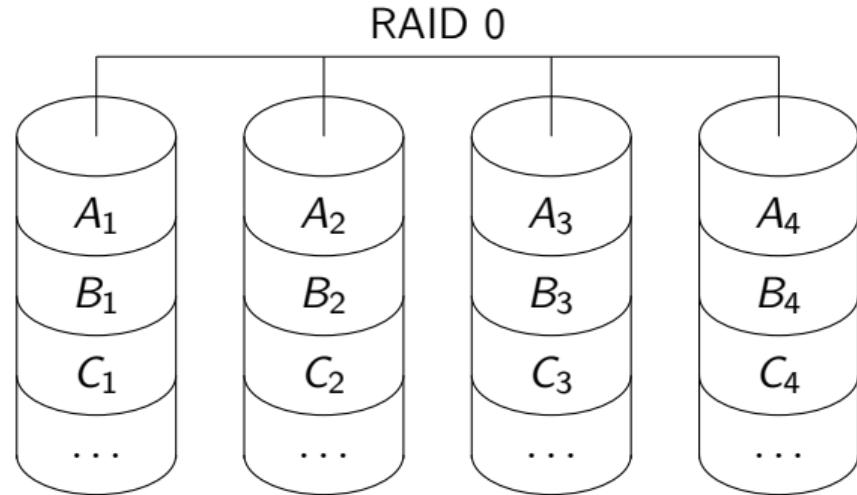
3 Производительность

4 Надежность

RAID для ускорения операций



Зеркалирование, может ускорять чтение



Страйпинг, ускоряет чтение и запись

Бутылочное горлышко — от CPU к CPU

1999 г. SSE

CPU слабые и плохо справляются с логикой СХД
Логика СХД выносится в отдельный аппаратный контроллер
(RAID-контроллер)

На CPU появляются и развиваются векторные регистры
Бутылочным горлышком становится скорость операций с дисками
Логика на CPU позволяет внедрять дополнительную функциональность,
в том числе увеличивающую скорость операций

2011 г. NVMe

Появляются более быстрые диски и задачи в индустрии, для которых
они необходимы (например запись видео высокого разрешения)
Логика на CPU тормозит
Ответ — DPU (data processing unit)

Software vs. Hardware

Software

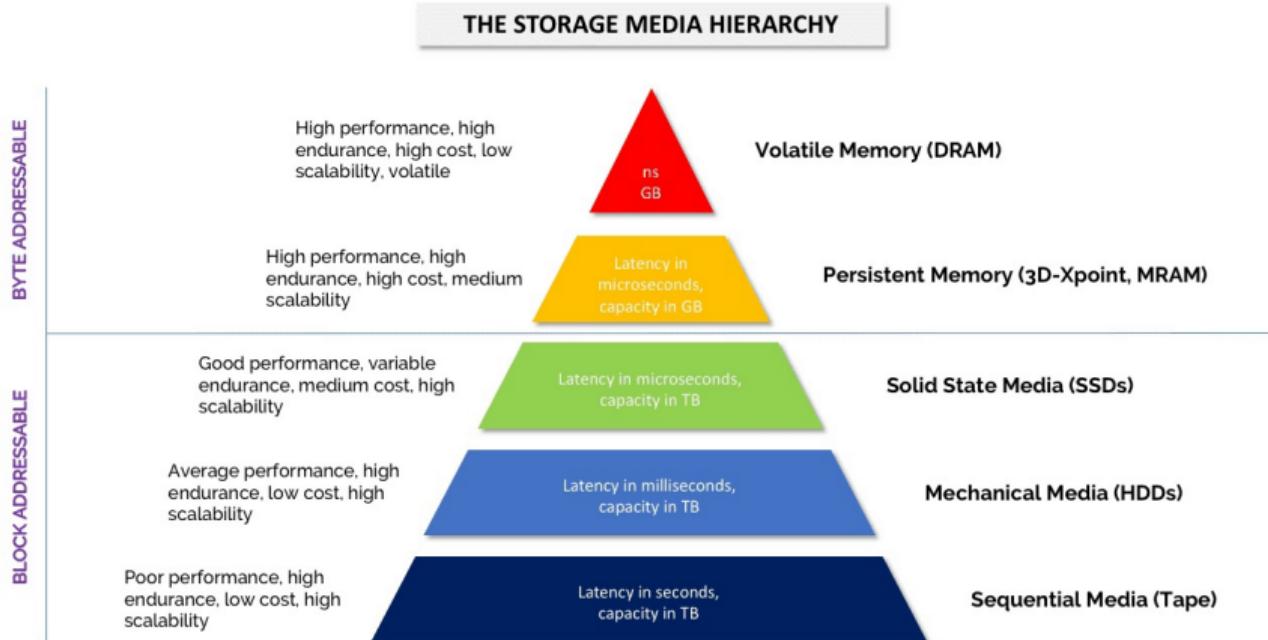
- + Оптимизация блочных операций (кэширование, тиринг, QoS, объединение)
- + Удобное управление
- + Возможность обновления и добавления функциональности
- + Разработка ПО на С
- Занимает ресурсы CPU
- Привязан к ОС

Hardware

- + Позволяет начинать операции до загрузки ОС
- + Встроенный UPS
- Занимает PCIe слот
- Невозможность обновлений
- Для настройки требуется выключать сервер
- Сложно разрабатывать ПО

Важно отметить, что на вопрос о производительности "дискуссионный". Если просто ввести запрос в гугл, то можно найти множество статей, утверждающих, что hardware быстрее, но дороже. Это устаревшая информация, не верьте ей. Hardware рейды до сих пор используются, но скорее маленькими компаниями для внутренних нужд. Например у Dell (<https://www.dell.com/en-us/shop/scc/sc/storage-products#portfolio>) сервера в основном с software рейдом

Об оптимизациях



Источник: <https://www.architecting.it/blog/caching-tiering/>

Содержание

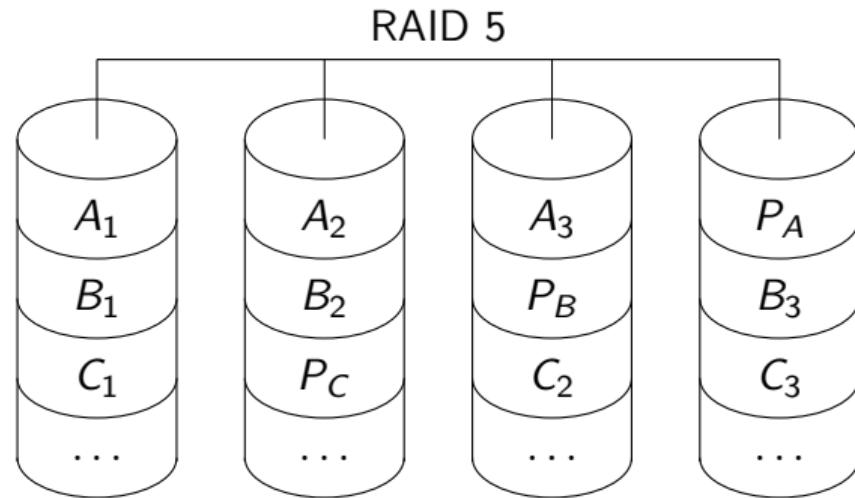
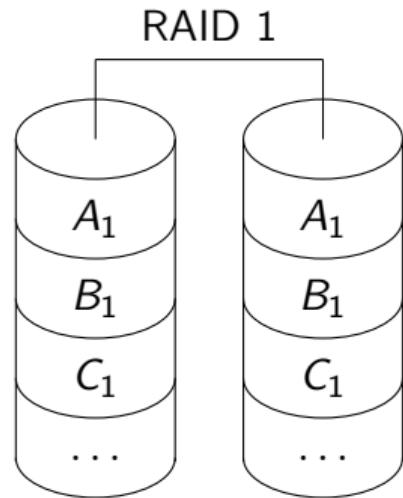
1 Определение системы хранения данных

2 Доступность

3 Производительность

4 Надежность

RAID для надежности



$$RAID \text{ failure rate} = f(ndrives, nparity, capacity, drive \text{ failure rate})^1$$

¹<https://www.ibm.com/support/pages/re-evaluating-raid-5-and-raid-6-slower-larger-drives>

- XOR
 - RAID 5
 - Исправляют одну ошибку
- Коды Хэмминга
 - RAID 2
 - Из $(2^n - 1)$ дисков n дисков под коды
 - Исправляет одну ошибку, обнаруживает двойную ошибку
- Коды Рида-Соломона
 - RAID 6, исправляет 2 ошибки
 - RAID N+M, исправляет M ошибок

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$D_\alpha + D_\beta = P - \sum_{i=0}^{N-1} D_i \quad i \neq \alpha, \beta$$

$$q_\alpha D_\alpha + q_\beta D_\beta = Q - \sum_{i=0}^{N-1} q_i D_i \quad i \neq \alpha, \beta$$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned} D_\alpha + D_\beta &= \bar{P}_{\alpha,\beta} \\ q_\alpha D_\alpha + q_\beta D_\beta &= \bar{Q}_{\alpha,\beta} \end{aligned}$$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned}D_\alpha &= \bar{P}_{\alpha,\beta} - D_\beta \\q_\alpha D_\alpha + q_\beta D_\beta &= \bar{Q}_{\alpha,\beta}\end{aligned}$$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$D_\alpha = \bar{P}_{\alpha,\beta} - D_\beta$$
$$q_\alpha(\bar{P}_{\alpha,\beta} - D_\beta) + q_\beta D_\beta = \bar{Q}_{\alpha,\beta}$$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned}D_\alpha &= \bar{P}_{\alpha,\beta} - D_\beta \\(q_\beta - q_\alpha)D_\beta &= \bar{Q}_{\alpha,\beta} - \bar{P}_{\alpha,\beta}q_\alpha\end{aligned}$$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned}D_\alpha &= \bar{P}_{\alpha,\beta} - D_\beta \\(q_\beta - q_\alpha)D_\beta &= \bar{Q}_{\alpha,\beta} - \bar{P}_{\alpha,\beta}q_\alpha\end{aligned}$$

- $q_\beta \neq q_\alpha$

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned}D_\alpha &= \bar{P}_{\alpha,\beta} - D_\beta \\(q_\beta - q_\alpha)D_\beta &= \bar{Q}_{\alpha,\beta} - \bar{P}_{\alpha,\beta}q_\alpha\end{aligned}$$

- $q_\beta \neq q_\alpha$
- $q_\beta - q_\alpha$ обратим

Коды Рида-Соломона



Страйп RAID 6

Вычисление P, Q (encode)

$$P = \sum_{i=0}^{N-1} D_i$$

$$Q = \sum_{i=0}^{N-1} q_i D_i$$

Восстановление данных (decode)

$$\begin{aligned}D_\alpha &= \bar{P}_{\alpha,\beta} - D_\beta \\(q_\beta - q_\alpha)D_\beta &= \bar{Q}_{\alpha,\beta} - \bar{P}_{\alpha,\beta}q_\alpha\end{aligned}$$

- $q_\beta \neq q_\alpha$
- $q_\beta - q_\alpha$ обратим
- Конечные поля!

RAID 6. Запись

Аналогично RAID 5 при записи данных надо также вычислить и записать контрольно-восстановительные суммы (P и Q)

1 Вычисление P и Q через Encode

- Требуется одинаковый кусок данных с каждого диска
- Чтение с дисков, не участвующих в операции (на которых не будет записи)
- Особенно неэффективно для небольших запросов (напр. 4 KiB)
- Может выполняться в несогласованном состоянии данных и синдромов

2 Вычисление P и Q через Update

- Для вычисления нужно прочитать только старые данные и старые синдромы (те же части, которые будут записаны)
- В некоторых случаях может быть медленнее, чем через encode
- Не может быть выполнено в несогласованном состоянии данных и синдромов, для того, чтобы выполнять такие операции после создания RAID нужно посчитать и записать все контрольно-восстановительные суммы (процесс инициализации)

RAID 6. Запись. Write hole

Записи на каждый диск выполняются независимо. Если происходит отключение питания, то могло записаться разное количество данных на каждом из дисков и может возникнуть несогласованном состоянии данных и синдромов, в таком случае нельзя делать запись через update



Пути решения проблемы

- Единственный бит в метаданных на весь рейд, по которому можно определить, был ли он выгружен корректно
- Битовая карта последних записей
- Журнал