# Обзор существующих распределенных файловых систем хранения.

## NFS.

## ZFS ([Zettabyte](https://ru.wikipedia.org/wiki/%D0%97%D0%B5%D1%82%D1%82%D0%B0%D0%B1%D0%B0%D0%B9%D1%82) File System)

Эта система проектировалась с очень большим запасом по параметрам, на основе совершенно справедливого прогноза огромного роста данных, подлежащих хранению в распределенных системах в будущем.

Функции ZFS:

* Разделение системы хранения на пулы (Pooled storage).
* Копирование при записи (Copy-on-write).
* Снапшоты (Snapshots) системы.
* Верификация целостности данных и автоматическое исправление данных.
* Автоматическая замена на запасной диск (Hot spare).
* Максимальный размер файла 16 эксабайт.
* Объем хранения 256 квадриллионов зеттабайт.

Рассмотрим самые интересные особенности системы.

### Хранение в пулах (Pooled storage)

В отличие от большинства файловых систем, ZFS объединяет функции файловой системы и менеджера томов (volume manager). Это означает, что ZFS может создавать файловую систему, которая будет простираться по многим группам накопителей ZVOL или пулам. Более того, можно добавлять емкость в пул простым добавлением нового накопителя. ZFS сама выполнит партицию и форматирование нового накопителя.

### ****Копирование**** при записи (Copy-on-write)

В большинстве файловых систем при перезаписи данных на то же физическое место носителя ранее записанные там данные теряются навсегда. В ZFS новая информация пишется в новый блок. После окончания записи метаданные в файловой системе обновляются, указывая на местоположение нового блока. При этом, если в процессе записи информации с системой что-то происходит, старые данные будут сохранены. Это означает, что не нужно запускать проверку системы после аварии.

### Снапшоты (Snapshots)

Copy-on-write закладывает основу для другой функции ZFS: моментальных снимков системы (снапшотов). ZFS использует для отслеживания изменений в системе.

Снапшот содержит оригинальную версию файловой системы, и в «живой» файловой системе присутствуют только изменения, которые были сделаны с момента последнего снапшота. Никакого дополнительного пространства не используется. Когда новые данные записываются в «живую» систему, выделяются новые блоки для сохранения этих данных.

Если файл удаляется, то ссылка на него в снапшоте тоже удаляется. Поэтому снапшоты в основном предназначены для отслеживания изменений в файлах, а не для добавления или создания файлов.

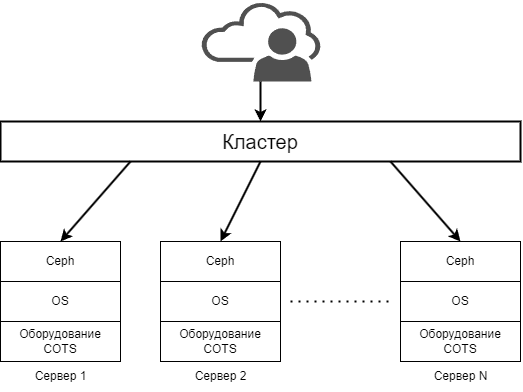
Снапшоты могут устанавливаться в режим read-only (только чтение), чтобы восстановить прежнюю версию файла. Также можно сделать откат «живой» системы на предыдущий снапшот. При этом будут потеряны только те изменения, которые были сделаны после момента этого снапшота.

## GFS

## HDFS

## Ceph

Ceph – распределенная масштабируемая система хранения с дизайном без единой точки отказа. Система разрабатывалась как открытый проект с файловой системой высокой масштабируемости до уровня эксабайт и рассчитанная на работу на стандартном коммерчески доступном серверном оборудовании COTS (Commercial Off The Shelf). Ceph приобретает все большую популярность в облачных системах хранения данных.



Основные принципы построения Ceph:

* Масштабируемость всех компонентов.
* Отсутствие единой точки отказа.
* Использование открытого ПО (Linux).
* Использование стандартного оборудования COTS (Commercial Off The Shelf).
* Автономность в управлении компонентами.

Ceph для клиента выглядит как обычная файловая система с папками и файлами, организованными по принципу иерархии.

Ceph реплицирует данные, за счет чего файловая система устойчива к отказам при использовании обычного стандартного компьютерного оборудования COTS, которое не требует специального обслуживания или администрирования. Поэтому она хорошо подходит для организаций, которые не обладают экспертизой в IT, и в которых недостаточно средств для содержания штата IT-специалистов.

## Lustre

Lustre – это кластерная архитектура для СХД, центральным элементом которой является файловая система Lustre, работающая на ОС Linux и обеспечивающая POSIX-интерфейс для файловой системы UNIX.

Lustre способна масштабировать емкость и производительность практически до любых требуемых величин, устраняя надобность в развертывании многих отдельных файловых систем в каждом компьютерном кластере. Кроме агрегирования емкости хранения многих серверов, пропускная способность ввода-вывода также агрегируется и масштабируется при добавлении серверов – в том числе в динамическом режиме.

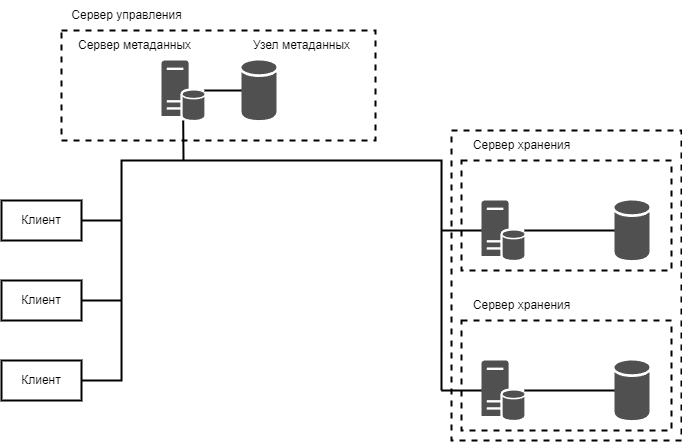
Lustre также не очень подходит для сетевых моделей peer-to-peer, где клиенты и серверы работают на одном узле хранения, и каждый занимает небольшую область емкости, из-за недостаточной репликации на уровне ПО Lustre. При этом, если один клиент или сервер отказывает, то данные этого узла будут недоступны, пока узел не перезапустится.

*Архитектура Lustre*

Сервер метаданных MDS (Metadata Server) – обеспечивает доступ к метаданным файловой системе и управляет запросами клиентов.

Узел метаданных – хранит метаданные файловой системы.

Сервер хранения – хранит данные.



*Архитектура серверов хранения*

Сервер объектного хранения – обрабатывает запросы данных.

Целевой узел объектного хранения – диск с данными.

Любой узел может быть доступен для многих серверов, но сервер может иметь доступ только к одному узлу в какой-то момент времени.

