Часть I. Введение в файловую систему Lustre*

Часть I содержит справочную информацию, которая поможет помочь вам архитектуру файловой системы Lustre и как основные компоненты сочетаются друг с другом. В данном разделе вы найдете информацию о:

- Понимании архитектуры Lustre
- Понимании Lustre Networking (LNET)
- Понимании об отказоустойчивости файловой системы Lustre

Глава 1. Понимание архитектуры Lustre

В данной главе описывается архитектура Lustre и особенности файловой системы Lustre. Она включает в себя следующие разделы:

- Раздел 1.1, “Чем является файловая система (и чем нет)”
- Раздел 1.2, “Компоненты Lustre”
- Раздел 1.3, “Хранилища данных и ввод/вывод файловой системы Lustre ”

1.1. Чем является файловая система (и чем нет)

Архитектура Lustre является архитектурой хранения для кластеров. Центральной компонентой архитектуры Lustre является файловая система Lustre, которая поддерживается операционной системой Linux и поддерживает интерфейс совместимый со стандартом POSIX* файловой системы UNIX.

Архитектура хранения Lustre используется для многих различных видов кластеров. Она является самой известной для поддержания высокопроизводительных вычислительных кластеров (HPC) во всем мире с десятками тысяч клиентских систем, петабайт (PB) данных и сотен гигабайт в секунду (GB/sec) пропускной способности ввода/вывода. Многие HPC сайты используют файловую систему Lustre как свою глобальную файловую систему, обслуживающую десятки кластеров.

Способность файловой системы Lustre масштабировать емкость и производительность для любых потребностей уменьшает необходимость развертывания большого количества отдельных файловых систем, по одной для каждого вычислительного кластера. Упрощается управление хранением, исключая необходимость копирования данных между кластерами. Помимо агрегации емкости хранения многих серверов, также объединяется пропускная способность ввода/вывода масштабируемая дополнительными серверами. Более того, пропускная способность и/или емкость может легко увеличиваться добавлением серверов динамически.

В то время как файловая система Lustre может функционировать во многих рабочих средах, это не обязательно самый лучший выбор для всех приложений. Она лучше всего подходит для применений когда превышаются возможности, которые может предоставить один сервер, хотя в отдельных случаях файловая система Lustre может работать лучше, чем другие файловые системы и на одном сервере благодаря своей сильной синхронизации и когерентности данных.

Файловая система Lustre в настоящее время не особенно хорошо подходит для одноранговых (“peer-to-peer”) моделей использования, когда клиенты и серверы работают на одном и том же узле, причем каждый использует небольшой объем хранения, в связи с отсутствием репликации данных на уровне программного обеспечения Lustre. При таком использовании, если один клиент / сервер выходит из строя, то данные, хранящиеся на этом узле, не будет доступны, пока не будет перезапущен узел.

1.1.1. Характеристики Lustre

Файловая система Lustre работает на различных ядрах ОС различных производителей. Для получения дополнительной информации см. матрицу тестирований Lustre Раздела 8.1, “Подготовка к установке программного обеспечения Lustre”.

Установку Lustre можно масштабировать в сторону увеличения или уменьшения по отношению к количеству клиентских узлов, дисковой памяти и пропускной способности. Масштабируемость и производительность зависят от имеющихся хранилищ и пропускной способности сети, а также вычислительной мощности серверов в системе. Файловая система Lustre может быть развернута в широком разнообразии конфигураций, которые могут масштабироваться далеко за пределы размеров и производительности, наблюдаемых в имеющихся на сегодняшний день системах.

Таблица 1.1, “Масштабируемость и производительность файловой системы Lustre” показывает практические диапазоны показателей масштабируемости и производительности файловой системы Lustre и некоторые результаты произведенных систем.

Таблица 1.1. Масштабируемость и производительность файловой системы Lustre

Показатель	Текущий практический диапазон	Проверено на практике
Масштабируемость клиентов	100-100000	50000+ клиентов, многие с диапазоном от 10000 до 20000
Производительность клиентов	*Отдельный клиент:*	*Отдельный клиент:*
	ввод/вывод 90% производительности сети	ввод/вывод 2 ГБ/с, 1000 оп/с метаданными/с
	Совместная:	*Совместная:*
	ввод/вывод 2.5 ТБ/с	ввод/вывод 240 ГБ/с
Масштабируемость OSS	*Отдельный OSS:*	*Отдельный OSS:*
	1-32 OST на OSS,	8 OST на OSS,
	128ТБ на OST	16ТБ на OST
	количество OSS:	*количество OSS:*
	500 OSS, имеющих до 4000 OST	450 OSS с 1000 4ТБ OST
		192 OSS с 1344 8ТБ OST
Производительность OSS	*Отдельный OSS:*	*Отдельный OSS:*
	5 ГБ/с	2.0+ ГБ/с
	Совместная:	*Совместная:*
	2.5 ТБ/с	240 ГБ/с
MDS Scalability	*Отдельный MDS:*	*Отдельный MDS:*
	4 миллиарда файлов	750 миллионов файлов
	количество MDS:	*количество MDS:*
	1 основной + 1 резервный	1 основной + 1 резервный
Введено	**в Lustre 2.4**	
	Начиная с версии 2.4 ПО Lustre:	
Введено	**в Lustre 2.4**	
	До 4096 MDS и до 4096 MDT	
Производительность MDS	35000/с операций создания, 100000/с стат. операций с метаданными	15000/ с операций создания, 35000/ с стат. операций с метаданными
Масштабируемость файловой системы	*Отдельный файл:* макс. Размер файла 2.5 ПБ *Совместная:* объем 512 ПБ, 4 миллиарда файлов	*Отдельный файл:* макс. Размер файла много-ТВ *Совместная:* объем 10 PB space, 750 миллионов файлов

Другие показатели Lustre:

- **Файловая система с расширенной производительностью ext4:** Файловая система Lustre использует улучшенную версию файловой системы ext4 с журналированием для хранения данных и метаданных. Данная версия, называемая `ldiskfs`, была расширена для улучшения производительности и поддержки дополнительной функциональности, необходимой файловой системе Lustre.

- **Совместимость со стандартом POSIX:** Был проведен полный набор тестов POSIX способом, аналогичным файловой системе ext4, с ограниченными исключениями для клиентов Lustre. В кластере, большинство операций неделимо, следовательно клиенты никогда не видят устаревшие данные или метаданные. Программное обеспечение Lustre использует файловый ввод/вывод mmap().

- **Высокопроизводительные гетерогенные сети:** Программное обеспечение Lustre поддерживает широкий набор высокопроизводительных, низколатентных сетевых средств, а также допускает Remote Direct Memory Access (RDMA) для InfiniBand* (использующей OpenFabrics Enterprise Distribution (OFED*) и других сетевых средств для быстрого и эффективного сетевого обмена. Сложные сети RDMA могут сопрягаться с использованием маршрутизации Lustre для максимальной производительности. Программное обеспечение Lustre также включает интегрированную сетевую диагностику .

- **Высокая доступность:** Файловая система Lustre поддерживает активно/активное восстановление после сбоев применяя совместно используемые разделы целей OSS (OST). Программное обеспечение Lustre версии 2.3 и более ранних версий предлагало активно/пассивное восстановление после сбоев с применением совместно используемого раздела хранения для цели MDS (MDT).

Введено в Lustre 2.4

Начиная с версии программного обеспечения серверов и клиентов Lustre 2.4 и выше, стало возможным конфигурировать активно/активное восстановление после сбоев нескольких MDT. Это позволяет приложениям выполнять прозрачное восстановление. Файловая система Lustre может работать с различными средствами управления высокой доступности (HA, high availability) допуская автоматическое восстановление и исключая единую точку отказа (NSPF, no single point of failure). Защита от множественного монтирования (MMP, Multiple mount protection) обеспечивает комплексную защиту от ошибок в системах с высокой доступностью, которые могли бы привести к повреждениям в файловой системе.

- **Безопасность:** По умолчанию соединения TCP разрешаются только с привилегированных портов. Членство в группах UNIX проверяется в MDS.

- **Список управления доступом (ACL, Access control list), расширенные атрибуты:** модель безопасности Lustre следует аналогичной в файловой системе UNIX, расширенной списками ACL POSIX. Заслуживают внимания дополнительные возможности, включающие дефрагментацию корневой директории.

- **Совместимость:** Файловая система Lustre работает с различными архитектурами CPU и с кластерами со смешанным порядком записи байт, а также совместима с последовательными основными версиями программного обеспечения Lustre.

- **Основанная на объектах архитектура:** Пользователи изолированы от дисковой файловой структуры, что позволяет обновлять архитектуру систем хранения без воздействия на пользователей.

- **Блокировка файлов на уровне байт и метаданных на уровне мелких единиц:** Многие пользователи могут одновременно читать и модифицировать одни и те же файлы или директории. Администратор распределенных блокировок Lustre (LDLM, Lustre distributed lock manager) гарантирует согласованность файлов между всеми пользователями и серверами в файловой системе. MDT LDLM управляет блокировкой прав доступа к индексным дескрипторам (inode) и именам путей. Каждый OST имеет собственный LDLM, который блокирует хранящиеся на нем полосы (stripes) файлов, что масштабирует производительность блокировок по мере роста файловой системы.

- **Квоты:** для файловой системы Lustre доступны квоты пользователей и групп.

- **Рост емкости:** Размер файловой системы Lustre и общая пропускная способность кластера могут увеличиваться без прерываний системы путем добавления новых OSS с OST в кластер.

    - **Управляемое чередование:** Схему размещения файлов между OST можно настроить на основе каждого файл, директории или файловой системы. Это позволяет настраивать файловый ввод/вывод на конкретные требования приложений в рамках отдельной файловой системы. Файловая система Lustre использует RAID-0 чередование и балансировку использования пространства между OST.

    - **Защита целостности сетевых данных:** Контрольные суммы всех передаваемых от пользователя к OSS данных защищают их от разрушения при обмене данными.

    - **MPI ввод/вывод:** Архитектура Lustre имеет специальный выделенный уровень MPI ADIO который оптимизирует параллельный ввод/вывод в соответствии с базовой архитектурой файловой системы.

    - **Экспорт в NFS и CIFS:** Файлы Lustre могут быть реэкспортированы в NFS (с применением Linux knfsd) или в CIFS (с применением Samba), позволяя их совместное использование с не-Linux клиентами, такими как Microsoft* Windows* и Apple* Mac OS X*.

    - **Инструменты аварийного восстановления:** Файловая система Lustre предоставляет онлайновую проверку распределенной файловой системы (LFSCK), которая может восстанавливать согласованность между компонентами системы хранения в случае серьезной ошибки файловой системы. Файловая система Lustre может работать даже при наличии несогласованностей файловой системы, и LFSCK может работать в то время, когда файловая система используется, следовательно, не требуется выполнения LFSCK перед возвратом файловой системы к режиму работы.

    - **Текущий контроль производительности:** Файловая система Lustre предлагает разнообразные механизмы для изучения и настройки производительности.

    - **Открытый исходный код:** Программное обеспечение Lustre регулируется лицензией GPL 2.0 для использования с операционной системой Linux.

1.2. Компоненты Lustre

Установка программного обеспечения Lustre включает в себя сервер управления и одну или несколько файловых систем Lustre, соединенных сетью Lustre (LNET).
Базовая конфигурация компонентов файловой системы Lustre показана на Рисунке 1.1, “Компоненты файловой системы Lustre в базовом кластере”.

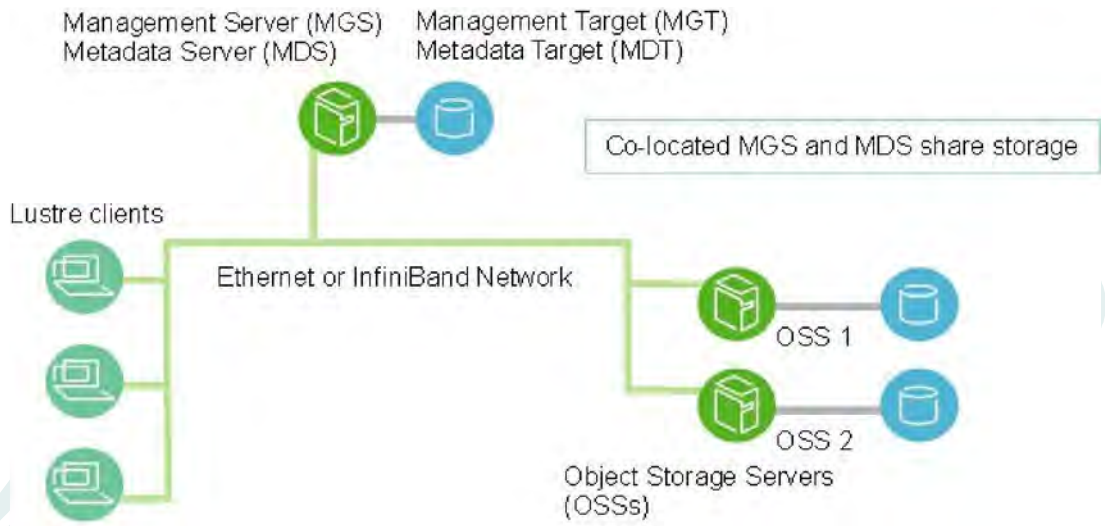

Рисунок 1.1. Компоненты файловой системы Lustre в базовом кластере

1.2.1. Сервер управления (MGS, Management Server)

MGS хранит информацию о конфигурации для всех файловых систем Lustre в кластере и предоставляет эту информацию другим компонентам Lustre. Каждый приемник (цель, target) Lustre связывается с MGS для предоставления информации, а пользователи Lustre связываются с MGS для получения информации.
Желательно чтобы MGS имел свое собственное пространство для хранения данных, чтобы им можно было управлять независимо. Однако, MGS может быть совмещен и разделять пространство хранения данных с MDS как показано на Рисунке 1.1, “ Компоненты файловой системы Lustre в базовом кластере ”.

1.2.2. Компоненты файловой системы Lustre

Каждая файловая система Lustre состоит из следующих компонентов:

- **Сервер метаданных (MDS, Metadata Server)** - MDS делает доступными пользователям метаданные, хранящиеся на одном или более MDT. Каждый MDS управляет именами и каталогами в файловой системе(ах) Lustre и обеспечивает обработку сетевых запросов для одного или нескольких локальных MDT.

- **Приемник (цель) метаданных (MDT, Metadata Target**) – Для программного обеспечения версий Lustre 2.3 и более ранних, каждая файловая система имела один MDT. MDT хранит метаданные (такие как имена файлов, каталогов, разрешений и местоположения файла) в хранилище, подключенном к MDS. Каждая файловая система имеет один MDT. MDT на совместно используемом приемнике (цели) может быть доступен многим MDS, хотя только один может получить к нему доступ в определенное время. Если активный MDS вышел из строя, находившийся в режиме ожидания MDS может обслужить MDT и сделать его доступным для пользователей. Это называется восстановлением MDS после отказа.

Введено в Lustre 2.4

Начиная с версии 2.4 программного обеспечения Lustre, поддерживается несколько MDT. Каждая файловая система имеет по крайней мере один MDT. MDT на совместно используемом хранилище приемника может быть доступен через несколько MDS, хотя только один MDS может экспортировать MDT пользователям в определенное время. Две машины MDS совместно используют для двух или более MDT. После отказа одного из MDS, оставшиеся MDS начинают обслуживать MDT вышедшего из строя MDS.

- **Сервер хранения объектов (OSS, Object Storage Servers)** : OSS предоставляет услуги файлового ввода/вывода и обработки сетевых запросов для одного или более OST. Обычно сервер OSS обслуживает от двух до восьми OST, причем каждый объемом до 16 ТБ. Типичной конфигурацией является MDT на выделенном узле, два или более OST на каждом узле OSS, и по пользователю на каждом из большого числа вычислительных узлов.

- **Приемник (цель) объектов хранения (OST, Object Storage Target)**: Данные пользовательских файлов хранятся в одном или нескольких объектах, причем каждый объект на отдельном OST в файловой системе Lustre. Количество объектов на файл настраивается пользователем и может регулироваться для оптимизации производительности для данной рабочей нагрузки.

- **Пользователи Lustre**: Пользователями Lustre являются узлы для вычислений, визуализации или настольные компьютеры на которых работает клиентское программное обеспечение Lustre, позволяющее им монтировать файловую систему Lustre.

Пользовательское программное обеспечение Lustre обеспечивает интерфейс между виртуальной файловой системой Linux и серверами Lustre. Пользовательское программное обеспечение включает в себя клиент управления (MGC, management client), клиент метаданных (MDC, metadata client) и несколько клиентов хранения объектов (OSC, object storage clients), по одному для каждого соответствующего OST в файловой системе.
Логический том объектов (LOV, logical object volume) объединяет OSC обеспечивая прозрачный доступ ко всем OST. Таким образом, пользователь со смонтированной файловой системой Lustre видит единое, согласованное, синхронизованное пространство имен. Несколько пользователей могут писать в разные части одного итого же файла одновременно, в то же время, когда другие пользователи могут читать из этого файла.
Таблица 1.2, “Требования к хранилищам данных и аппаратным средствам для компонентов файловой системы Lustre” обеспечивает требования для подключаемых хранилищ данных для каждой компоненты файловой системы Lustre и описывает желательные характеристики используемых аппаратных средств.

Таблица 1.2. Требования к хранилищам данных и аппаратным средствам для компонентов файловой системы Lustre

	Требуемое подключаемое хранилище данных	Желательные характеристики аппаратных средств
MDS	1-2% от емкости файловой системы	Адекватная производительность CPU, много памяти, быстрые диски
OSS	1-16 ТБ на OST, 1-8 OST на OSS	Хорошая пропускная способность шин. Рекомендуется равномерная сбалансированность хранения по OSS.
Пользователи	Нет	Низкая латентность, высокая пропускная способность сети.

Дополнительные требования и соображения по оборудованию см. в Главе 5, *Определение требований конфигурации аппаратных средств и параметров форматирования*.

1.2.3. Сеть Lustre (LNET)

Сеть Lustre (LNET) является обычным сетевым API, который обеспечивает инфраструктуру обмена данными, которая обрабатывает метаданные и файловый ввод/вывод данных для серверов и клиентов файловой системы Lustre. Дополнительные сведения о LNET, см. в Главе 2, *Понимание сети Lustre (LNET)*.

1.2.4. Кластер Lustre

В масштабе, кластер файловой системы Lustre может включать сотни OSS и тысячи пользователей (см. Рисунок 1.2, “Кластер Lustre в масштабе”). В кластере Lustre может быть использовано более одного типа сети. Совместное использование хранилищ OSS обеспечивает отказоустойчивость. Более подробная информация об отказоустойчивости OSS приводится в Главе 3, *Понимание отказоустойчивости файловой системы Lustre*.

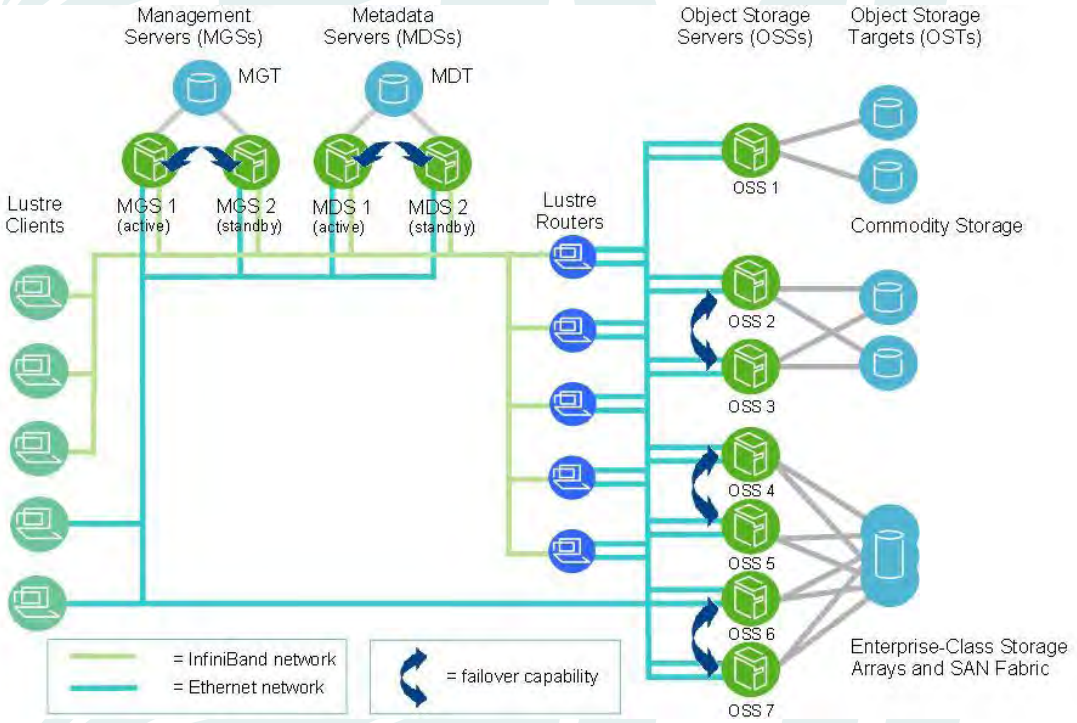

Рисунок 1.2. Кластер Lustre в масштабе

1.3. Хранилища данных и ввод/вывод файловой системы Lustre

В версии 2.0 программного обеспечения Lustre, были введены файловые идентификаторы Lustre (FID, file identifiers), чтобы заменить номера индексных дескрипторов файлов UNIX (inode) при идентификации файлов или объектов. FID является 128-битным идентификатором, который содержит уникальный 64-битный порядковый номер, 32-битный ID объекта (OID) и 32-битный номер версии. Порядковый номер уникален по всем приемникам Lustre в файловой системе(OST и MDT). Это изменение позволило в дальнейшем поддержку нескольких MDT (введенную в версии 2.3 программного обеспечения Lustre) и ZFS (введенную в версии 2.4 программного обеспечения Lustre).
Также в версии 2.0 введена функция, называемая *FID-in-dirent* (также известная как *dirdata*), при которой FID сохраняется как часть имени файла в текущем каталоге. Эта функция существенно увеличивает производительность выполнения команды `ls` путем уменьшения количества дисковых операций ввода/вывода. FID-in-dirent генерируется во время создания файла.

Замечание

Функция FID-in-dirent не совместима с форматом программного обеспечения Lustre версии 1.8. Поэтому, при обновлении программного обеспечения Lustre версии 1.8 на программное обеспечение Lustre версии 2.x, функция FID-in-dirent не включается автоматически. Для программного обеспечения Lustre версии 1.8 обновленного на программное обеспечение Lustre версий с 2.0 до 2.3, FID-in-dirent может быть включена вручную, однако она будет иметь эффект только для новых файлов.
Для получения более подробной информации об обновлении программного обеспечения Lustre версии 1.8 и включении FID-in-dirent для существующих файлов, см. Главу 16, “*Обновление файловой системы Lustre*”.

Введено в Lustre 2.4
в версии 2.4 программного обеспечения Lustre

В версии 2.4 программного обеспечения Lustre был выпущен инструмент администрирования файловой системы LFSCK 1.5 для обеспечения функциональности, которая включает FID-in-dirent для существующих файлов. Он содержит следующую функциональность:

- Генерирует IGIF состояния FID для существующих файлов версии 1.8.
- Проверяет FID-in-dirent для каждого файла, чтобы определить случаи, когда она не существует или является недопустимой и, если необходимо, повторно вырабатывает FID-in-dirent.
- Проверяет запись linkEA для каждого файла, чтобы определить случаи, когда она не существует или недопустима и, если необходимо, повторно вырабатывает linkEA. *linkEA* состоит из имени файла плюс родительский FID и хранится в виде расширенного атрибута в самом файле. Таким образом, linkEA может быть использован для разбора полного пути к файлу от корня.

Информация о том, где находится данный файл на OST хранится в виде расширенного атрибута, называемого EA расположения (layout EA) в объекте MDT определяемом FID для данного файла (см Рисунок 1.3, “Layout EA в MDT указывает на данные файла на OST”). Если файл является фалом данных (не директорией или символьной ссылкой), объект MDT указывает на 1-к-N OST объект(ы) на OST(ах) которые содержат данные файла. Если файл MDT указывает на один объект, все данные файла сохраняются в этом объекте. Если файл MDT указывает более чем на один объект, данные файла *чередуются (striped)* между объектами, используя RAID 0, и каждый объект хранится на своем OST. (Для получения дополнительной информации о том, как осуществляется чередование в файловой системе Lustre, см. Раздел 1.3.1, “Файловая система Lustre и чередование”.

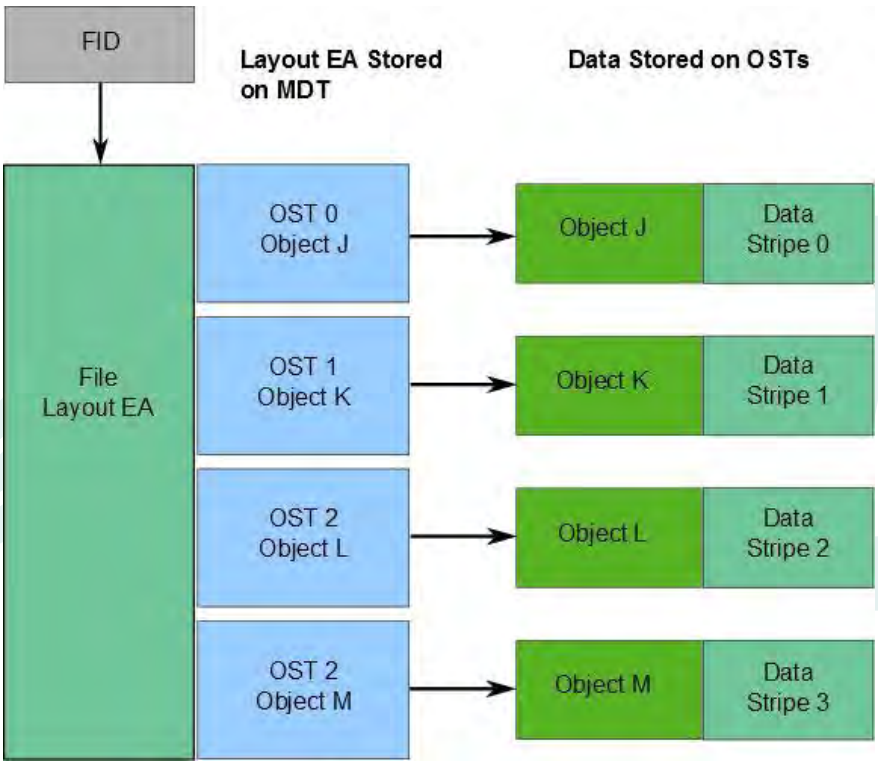

Рисунок 1.3. Layout EA в MDT указывает на данные файла на OST

Когда клиент хочет прочитать или записать в файл, он сначала выбирает layout EA из объекта MDT для файла. Затем клиент использует эту информацию для выполнения ввода / вывода в файле, непосредственно взаимодействуя с узлами OSS, на которых хранятся объекты. Этот процесс проиллюстрирован на рисунке 1.4, "Запрос клиентом Lustre данных файла".

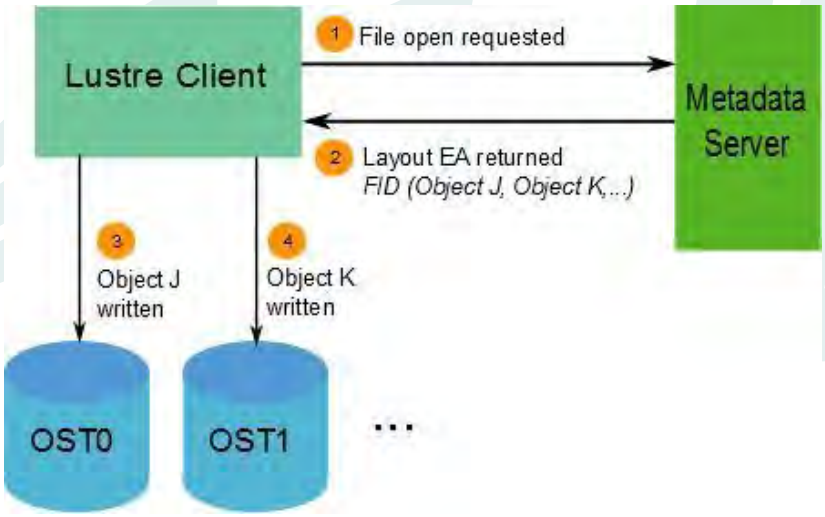

Рисунок 1.4. Запрос **клиентом** Lustre данных файла

Доступная пропускная способность файловой системы Lustre определяется следующим образом:

- *Сетевая пропускная способность* равна объединенной пропускной способности OSS к приемникам (целям).
- *Дисковая пропускная способность* равна сумме пропускных способностей дисков хранилищ приемников (OST), но не выше предела сетевой пропускной способности.
- *Совокупная пропускная способность* равна минимуму из дисковой и сетевой пропускной способностей.
- *Доступное файловой системе пространство* равно сумме свободных пространств всех OST.

1.3.1 Файловая система Lustre и чередование

Одним из основных факторов, приводящих к высокой производительности файловых систем Lustre, является возможность чередования данных между несколькими OST в циклическом режиме. Пользователи могут опционально настроить для каждого файла количество полос, размер полосы и используемые OST.

Чередование может быть использовано для повышения производительности, когда совокупная пропускная способность к одному файлу превысит пропускную способность одного OST. Возможность чередования также полезна, когда один OST не имеет достаточно свободного места для размещения всего файла. Дополнительные сведения о преимуществах и недостатках чередования файла, см. раздел 18.2, "Вопросы размещения файла (чередования) в Lustre"

Чередование позволяет сохранять сегменты или 'куски' данных файла на различных OST, как показано на Рисунке 1.5, “Чередование файла в файловой системе Lustre”. В файловой системе Lustre, используется модель RAID 0, согласно которой "чередуются" между определенным числом объектов. Число объектов в одном файле называется `stripe_count`.

Каждый объект содержит кусок данных из файла. Когда кусок данных, подлежащих записи в определенный объект превышает `stripe_size`, следующий кусок данных в файле сохраняется в следующий объект.

Значения по умолчанию для `stripe_count` и `stripe_size` устанавливаются для файловой системы. Значением по умолчанию для `stripe_count` является 1 полоса для файла значение по умолчанию для `stripe_size` это 1МБ. Пользователь может изменить эти значения на основе значений для каталога, или для файла. Для дополнительной информации, см. Раздел 18.3, “Установка конфигурации расположения файла/чередования (`lfs setstripe`)”.

Рисунок 1.5, “Чередование файла в файловой системе Lustre”, `stripe_size` для Файла C больше чем `stripe_size` для Файла A, позволяет сохранять больше данных в одной полосе для Файла C. `stripe_count` для Файла A 3, что имеет результатом чередование данных между тремя объектами, в то время как `stripe_count` для Файла B и Файла C 1.

Для не записанных данных на OST не резервируется свободное пространство. Файл A на Рисунке 1.5, “Чередование файла в файловой системе Lustre”.

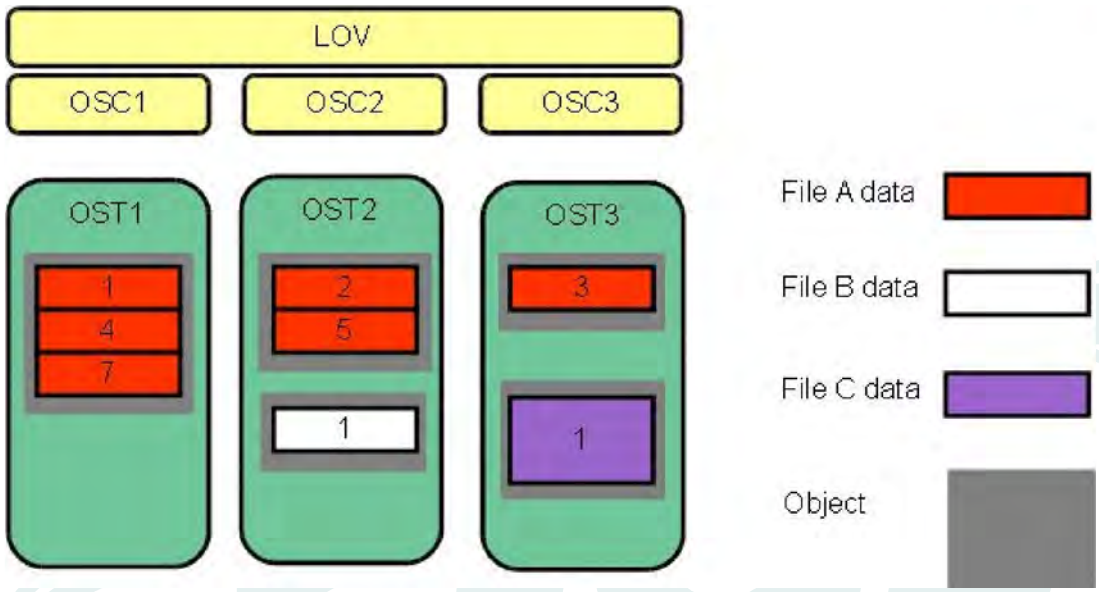

Рисунок 1.5, "Чередование файла в файловой системе Lustre"

Максимальный размер файла не ограничивается размером отдельного приемника (цели). В файловой системе Lustre, файлы могут быть разделены между многими объектами (до 2000), и каждый объект может быть в размере до 16 ТБ при ldiskfs. Это приводит к максимальному размеру файла в 31.25 ПБ. (Заметим, что файловая система Lustre может поддерживать файлы до 2^64 байт в зависимости от поддерживаемых OST дисковых систем.)

Замечание

Версии программного обеспечения Lustre до 2.2 ограничивали максимальное количество чередований для одного файла 160-ю OST.

Хотя один файл может быть разделен на 2000 объектов, файловая система Lustre может иметь тысячи OST. Пропускная способность ввода/вывода для доступа к одному файлу составляется из пропускных способностей ввода/вывода к объектам в файле, которая может составлять пропускную способность до 2000 серверов. В системе с более чем 2000 OST, пользователи могут выполнять ввод/вывод с использованием нескольких файлов для полного использования пропускной способности файловой системы.
Для дополнительной информации о чередовании, см. Главу 18, *Управление расположением файла (чередование) и свободное пространство*.

Глава 2. Понимание сети Lustre (LNET)

Данная глава описывает сеть Lustre (LNET). Она включает следующие разделы:

- Раздел 2.1, “ Знакомство с LNET”
- Section 2.2, “Основные свойства LNET”
- Section 2.3, “Сети Lustre ”
- Section 2.4, “Поддерживаемые типы сетей”

1.1. Знакомство с LNET

В кластере, использующем одну или более файловых систем Lustre, сетевая инфраструктура для обмена данными, необходимого файловой системе Lustre реализуется с использованием функциональных возможностей сети Lustre (LNET).
LNET поддерживает многие типы часто используемых сетей, такие как InfiniBand и сети IP, а также допускает одновременную доступность различных типов сетей с маршрутизацией между ними. При поддержке лежащих в основе сетей с помощью соответствующих сетевых драйверов Lustre (LND, Lustre network driver), допускается удаленный прямой доступ к памяти (RDMA). Высокая доступность и функции восстановления позводяют прозрачное восстановление в сочетании с отказоустойчивостью серверов.
LND является подключаемым драйвером, который обеспечивает поддержку для определенного типа сети, например, `ksocklnd` - драйвер, реализующий LND сокет TCP, который поддерживает сети TCP. LND загружаются в стек драйверов с одним LND для каждого используемого типа сети.

Для информации о конфигурировании LNET, см Главу 9, *Настройка сети Lustre (LNET)* .

Для информации об администрировании LNET, см. Часть III, “Администрирование Lustre”.

1.2. Основные свойства LNET

Основные свойства LNET включают:

- RDMA, если он поддерживается лежащими в основе сетями
- Поддержка многих часто используемых типов сетей
- Высокая доступность и отказоустойчивость
- Одновременная поддержка многих типов сетей
- Маршрутизация между различными сетями

LNET допускает сквозную пропускную способность чтения/записи равную или близкую к пиковым значениям для различных сетевых интерконнектов.

1.3. Сети Lustre

Сеть Lustre состоит из клиентов и серверов, работающих под управлением программного обеспечения Lustre. Нет необходимости ограничиваться одной подсетью LNET и можно охватывать несколько сетей при условии, что возможна маршрутизация между ними. Аналогичным образом, отдельная сеть может иметь несколько подсетей LNET.
Сетевой стек Lustre состоит из двух уровней, модуля кода LNET и LND. Уровень LNET работает над уровнем LND аналогично тому, как сетевой уровень работает над канальным уровнем. Уровень LNET является сободным от установления соединения, асинхронным и не проверяет передачу данных, в то время как уровень LND ориентирован на соединение и, как правило, осуществляет проверку обмена данными.
Сети LNET уникально идентифицируются меткой, состоящей из строки, соответствующей LND и числа, например, o2ib0 или o2ib1 уникально идентифицирует каждую LNET. Каждый узел в LNET имеет по крайней мере один сетевой идентификатор (NID, network identifier). NID представляет собой сочетание адреса сетевого интерфейса и метки LNET в виде: *address@LNET_label*.

Примеры:

```
192.168.1.2@tcp0
10.13.24.90@o2ib1
```

При определенных обстоятельствах для файловой системы Lustre может быть желательным прохождение траффика по нескольким LNET. Это возможно при с помощью маршрутизации LNET. Важно понимать, что маршрутизация LNET не то же самое, что сетевая маршрутизация. Более подробную информацию о маршрутизации LNET, см в Главе 9, *Настройка сети Lustre (LNET)*

1.4. Поддерживаемые типы сетей

Модуль кода LNET содержит драйверы LND для поддержки многих типов сетей, включая:

- InfiniBand: OpenFabrics OFED (o2ib)
- TCP (любая сеть, передающая траффик TCP, включая GigE, 10GigE и IPoIB)
- Cray: Seastar
- Myrinet: MX
- RapidArray: ra

Quadrics: Elan

Глава 3. Понимание об отказоустойчивости файловой системы Lustre

Данная глава описывает отказоустойчивость файловой системы Lustre file system. Она включает:

- Раздел 3.1, “ Что такое отказоустойчивость?”
- Раздел 3.2, “Функциональность отказоустойчивости в файловой системе Lustre”

3.1. Что такое отказоустойчивость?

В высокодоступных (HA, high-availability) системах незапланированные простои сводятся к минимуму с помощью резервных аппаратных и программных компонентов, а также программных средств, которые автоматизируют восстановление в случае возникновения сбоя. Если возникли условия, вызывающие неполадку, такие как потеря связи с сервером или устройством хранения данных, или же сбой в программном обеспечении, системные службы продолжатся после минимального перерыва. Как правило, доступность определяется как процент времени, когда система должна быть доступна.
Доступность достигается путем репликации аппаратных средств и/или программного обеспечения, так что, когда основной сервер выходит из строя или не доступен, на его место может быть включен резервный сервер для запуска приложений и связанных с ними ресурсов. Данный процесс, называемый *восстановлением после отказа*, является автоматическим в HA системах и, в большинстве случаев, полностью прозрачен для приложений.
Настройка отказоустойчивого оборудования требует пары серверов с совместно используемым ресурсом (как правило, физическое устройство хранения данных, которое может быть основано на SAN, NAS, аппаратных RAID, SCSI или Fibre Channel (FC) технологиях). Метод совместного использования хранимых данных должен быть существенно прозрачным на уровне устройств; одни и те же физические номера логических устройств (LUN, logical unit number) должны быть видны с обоих серверов. Для обеспечения высокой доступности на уровне физического хранения, мы поощряем использование RAID массивов для защиты от сбоев на уровне дисков.

Замечание

Программное обеспечение Lustre не обеспечивает избыточности данных; она зависит исключительно от резервирования устройств поддержки хранения данных. Поддерживающие хранение OST должны быть RAID 5 или, еще предпочтительнее, RAID 6 хранилищами данных. Хранилище данных MDT должно быть RAID 1 или RAID 10.

3.1.1. Возможности восстановления после сбоев

Чтобы установить высокодоступную файловую систему Lustre, для восстановления после сбоев используются программные и аппаратные средства управления питанием, а также программное обеспечение высокой доступности (HA) для обеспечения следующих возможностей:

- **Ограждение ресурсов** – Защита физических хранилищ от одновременного доступа с двух узлов.
- **Управление ресурсами** – Запускает и останавливает ресурсы Lustre как часть восстановления после сбоев, поддерживает состояние кластера и выполняет другие задачи управления ресурсами.
- **Мониторинг работоспособности** – Проверяет доступность аппаратных и сетевых ресурсов и отвечает на показания работоспособности, предоставляемые программным обеспечением Lustre.

Эти возможности могут быть обеспечены с помощью различных программных и/или аппаратных решений. Для получения дополнительной информации об использовании программного обеспечения или оборудования управления питанием высокой доступностью (HA) в файловой системе Lustre, см. Главу 11, *Настройка отказоустойчивости в файловой системе Lustre*.
Программное обеспечение высокой доступности (HA) отвечает за обнаружение выхода из строя основного узла сервера Lustre и управление восстановлением после сбоя. Программное обеспечение Lustre работает с любым программным обеспечением высокой доступности (HA) которое содержит ограждение ресурсов (ввода/вывода). Для правильного ограждения ресурсов, программное обеспечение высокой доступности (HA) должно быть в состоянии полностью выключить отказавший сервер или отключить его от общего устройства хранения данных. Если два активных узла имеют доступ к одному и тому же устройству хранения, данные могут быть серьезно повреждены.

3.1.2. Виды отказоустойчивых конфигураций

Узлы в кластере могут быть настроены для восстановления после сбоев несколькими способами. Обычно они конфигурируются парами (например, два OST присоединенных к общему запоминающему устройству), однако и другие отказоустойчивые конфигурации также возможны. Отказоустойчивые конфигурации включают в себя:

- **Активно/пассивная** пара – В этой конфигурации активный узел предоставляет ресурсы и обслуживает данные, в то время как пассивный узел, как правило, стоит на холостом ходу.
- **Активно/активная** пара – В этой конфигурации, оба узла активны, причем каждый обеспечивает подмножество ресурсов. В случае отказа, второй узел берет на себя ресурсы отказавшего узла.

В версиях программного обеспечения Lustre, предшествующих версии 2.4программного обеспечения Lustre, MDS может быть сконфигурирована как активно/пассивная пара, в то время как OSS могут быть развернуты в активно/активной конфигурации, которая обеспечивает избыточность без дополнительных накладных расходов. Часто находящийся в ожидании MDS является активным MDS для другой файловой системы Lustre или MGS, так что никакие узлы не простаивают в кластере.

Введено в Lustre 2.4

Программное обеспечение Lustre версии 2.4 вводит приемники (цели) для индивидуальных подкаталогов. Активно-активные конфигурации восстановления после отказов доступны для MDS, которые обслуживают MDT на совместно используемом хранилище..

3.2. Функциональность отказоустойчивости в файловой системе Lustre

Функциональность восстановления после отказов обеспечиваемая программным обеспечением Lustre можно использовать для следующих сценариев восстановления после отказов. Когда пользователь пытается выполнить ввод/вывод на вышедшем из строя приемнике (цели) Lustre, он продолжает попытки пока он получает ответы от любого из сконфигурированного отказоустойчивого узла для приемника (цели) Lustre. Приложение на пользовательской стороне не обнаружит ничего необычного, кроме того, что ввод /вывод может занять больше времени на выполнение.
Восстановление после сбоев в файловой системе Lustre требует, чтобы два узла были сконфигурированы как отказоустойчивая пара, которая должна использовать одно или более устройств хранения данных. Файловая система Lustre может быть сконфигурирована для обеспечения восстановления после сбоя MDT или OST.

- Для восстановления после сбоев MDT, два MDS могут быть сконфигурированы для обслуживания одного и того же MDT. Только один узел MDS может обслуживать MDT в одно и то же время.

Введено в 2.4

Программное обеспечение Lustre версии 2.4 допускает использование нескольких MDT. Размещая два или более раздела MDT на хранилище данных, совместно используемом двумя MDS, один MDS может выйти из строя и оставшийся MDS может начать обслуживать MDT. Это описывается как активно/активная отказоустойчивая пара.

- Для восстановления после сбоев OST, несколько узлов OSS могут быть сконфигурированы так, чтобы обслуживать один и тот же OST. Однако только один узел может обслуживать OST в одно и то же время. OST могут перемещать между узлами OSS которые имеют доступ к одному и тому же устройству хранения с помощью команд `umount/mount`.

Опция `--servicenode` используется при настройке узлов в файловой системе Lustre для восстановления после отказов во время создания (с использованием `mkfs.lustre`) или позже, когдафайловая система Lustre активна (с использованием `tunefs.lustre`). Для объяснения этих утилит, см. Раздел 37.14, “ mkfs.lustre” и Раздел 37.18, “ tunefs.lustre”.
Возможность восстановления при отказах в файловой системе Lustre можно использовать для обновления программного обеспечения Lustre между последовательными младшими версиями без простоя кластера. Для получения дополнительной информации, см. Главу 16, *Обновление файловой системы Lustre*.
Для дополнительной информации о настройке восстановления после сбоев, см. Главу 11, *Настройка отказоустойчивости в файловой системе Lustre*.

Замечание

Программное обеспечение Lustre обеспечивает функциональность восстановления после отказов только на уровне файловой системы. Для полного отказоустойчивого решения, функциональность восстановления после сбоев для компонентов системного уровня, таких как обнаружение неисправностей или управление питанием должны быть обеспечены с помощью инструментария сторонних производителей.

Предостережение

Функциональность восстановления после сбоев OST не обеспечивает защиты против разрушений, вызванных отказами дисковых устройств. Если выходят из строя носители (т.е. физические диски) используемые для OST, он не может быть восстановлен с помощью функциональных возможностей программного обеспечения Lustre. Мы настоятельно рекомендуем использовать некоторые формы RAID для OST. Функциональность Lustre предполагает, что хранилища данных являются надежными, так что не добавляет дополнительных свойств надежности.

3.2.1. Конфигурирование восстановления после сбоев MDT (Активно/Пассивное)

Два MDS обычно настраиваются как активно/пассивная пара как показано на Рисунке 3.1, “Конфигурирование восстановления после сбоев Lustre для активно/пассивного MDT”. Обратите внимание, что оба узла должны иметь доступ к общему хранилищу для MDT(ов) и MGS. Основной (активный) MDS управляет ресурсами метаданных системы Lustre. Если основной MDS выходит из строя, второй (пассивный) MDS берет на себя эти ресурсы и обслуживает MDT и MGS.

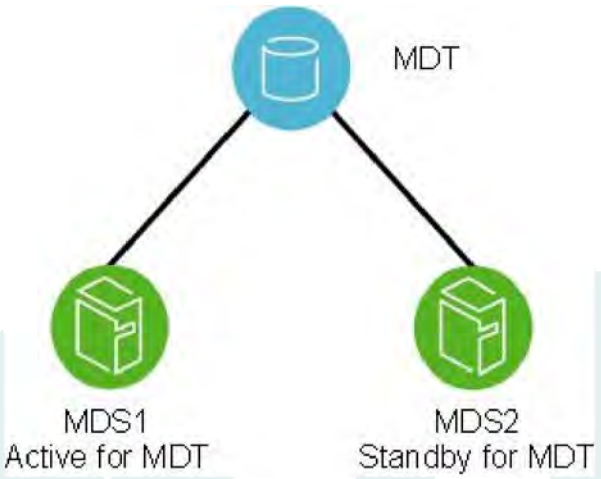

Рисунок 3.1. Конфигурирование восстановления после сбоев Lustre для активно/пассивного MDT

Замечание

В среде с несколькими файловыми системами, MDSs могут быть сконфигурированы в квази активно/активной конфигурации, с каждым MDS управляющим метаданными для подмножества файловой системы Lustre.

Введено в Lustre 2.4

Несколько MDT стало доступно с появлением программного обеспечения Lustre версии 2.4. MDT теперь может быть установлен в активно/активной отказоустойчивой конфигурации. Отказоустойчивый кластер строится из двух MDS как показано на Рисунке 3.2, “Конфигурирование восстановления после сбоев Lustre для активно/активного MDT ”.

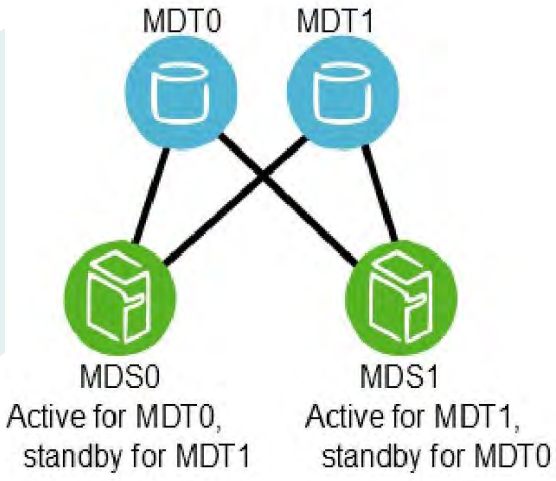

Рисунок 3.2. Конфигурирование восстановления после сбоев Lustre для активно/активного MDT

3.2.3. Конфигурирование восстановления после сбоев OST (Активно/ Активное)

OST обычно настраиваются в активно/активной отказоустойчивой конфигурации с балансировкой нагрузки. Отказоустойчивый кластер строится из двух OSS, как показано на Рисунке 3.3, “Конфигурирование восстановления после сбоев Lustre для активно/ активного OST”.

Рисунок 3.3. Конфигурирование восстановления после сбоев Lustre для активно/ активного OST

Замечание

Сконфигурированные как отказоустойчивая пара OSS должны иметь совместно используемые диски/RAID.

В активной конфигурации, 50% из имеющихся OST назначаются одной OSS, а остальные OST назначены другой OSS. Каждый OSS служит основным узлом для половины OST и резервным узлом для остальных OST.

В этом режиме, если один OSS отказывает, другой OSS берет на себя все вышедшие из строя OST. Клиенты пытаются подключиться к каждой OSS, обслуживающей OST, пока один из них не отвечает. Данные на OST записываются синхронно, и клиенты повторяют транзакции, которые выполнялись и не были зафиксированы на диске до сбоя OST.

Для более детальной информации о настройке восстановления после сбоев, см. Главу 11, *Настройка отказоустойчивости в файловой системе Lustre*.