# Параллельные FS с объектной архитектурой

**Андрей Слепухин** – руководитель Центра Кластерных Технологий "Т-Платформы".

***Представлен краткий обзор современных параллельных файловых систем (FS) и систем хранения, используемых в HPC-решениях.***

## Системы хранения для суперкомпьютеров: основные проблемы

В процессе вычислений мы всегда имеем дело с данными, и производительность любых вычислительных систем, в конечном итоге, неизбежно зависит от эффективности доступа к данным.
В эпоху, предшествовавшую расцвету параллельных вычислений и высокопроизводительных кластеров, расчеты выполнялись на одном многопроцессорном сервере, который обменивался данными с хранилищем по единственному каналу связи. Однако уже в начале 90-х годов стало очевидно, что наиболее эффективный путь дальнейшего роста производительности компьютеров лежит через объединение многих серверов, работающих с задачей в параллельном режиме. В этот самый момент производители компьютерных систем немедленно столкнулись с необходимостью организации одновременного доступа нескольких компьютеров к общему хранилищу.

Сетевые файловые системы NFS и CIFS, которые во многих случаях и сейчас успешно применяются многими производителями сетевых систем хранения, организуют доступ клиентов (компьютеров, серверов, вычислительных узлов) к хранилищу через выделенный сервер. Однако при увеличении количества вычислительных узлов в таких системах канал связи между клиентами и сервером быстро становится "узким местом", сильно затрудняющим масштабируемость конфигурации. Одним из путей решения этой проблемы стало создание таких файловых систем, как CXFS от SGI, GFX от Red Hat и OCFS от Oracle, поддерживающих обмен данными между клиентами и хранилищем без использования выделенного сервера через сеть хранения данных SAN (Storage Area Network).

Такой подход устранил "бутылочное горлышко" на пути от клиента к системе хранения данных, однако, он не лишен некоторых недостатков, особенно актуальных для систем, занятых высокопроизводительными вычислениями. Во-первых, прямое подключение всех узлов к сети SAN требует наличия адаптера (например, Fiber Channel) в каждом узле кластера, что увеличивает сложность и стоимость вычислительного решения. Но основной недостаток подхода заключается в том, что он не решает проблемы масштабируемости доступа к данным при увеличении количества вычислительных узлов. Дело в том, что при работе нескольких клиентов с одними и теми же файлами, расположенными на одном из серверов хранения, проблемы синхронизации доступа решаются на уровне ОС на самих вычислительных узлах. При этом, если число клиентов превышает несколько десятков, происходит значительное падение скорости обмена данными. Такие файловые системы успешно работают с кластерами баз данных и небольшими вычислительными кластерами, но для более масштабных систем они не подходят.

## Пути развития FS для HPC

Для суперкомпьютерных систем, имеющих сотни и тысячи вычислительных узлов, существует не так много решений. Одно из них – параллельная файловая система GPFS от IBM. Архитектура сетевой системы хранения, организованной с помощью GPFS, не требует подключения всех узлов кластера к сети SAN. Клиентская часть файловой системы обеспечивает доступ узлов кластера через высокоскоростную сеть, которой они объединены, к серверам NSD (Network Shared Disk server), которые делают устройства хранения видимыми для клиентов. Серверная часть GPFS обеспечивает доступ серверов NSD к дисковым массивам через сеть SAN. Такая система легко расширяема посредством добавления новых серверов NSD – на текущий момент существуют примеры подобных инсталляций с количеством вычислительных узлов более 2 тыс. Файловая система GPFS также имеет хорошо развитые возможности администрирования сети хранения.

К недостаткам такой архитектуры нужно отнести ее сложность. В особенности это касается клиентской части файловой системы, т.к. клиенты (узлы кластера) "видят" систему хранения как физическое устройство и работают с ней, как с локальными дисками. Клиентам приходится иметь дело с жесткой блоковой структурой дисковых массивов, а механизмы синхронизации доступа реализованы непосредственно на кластерных узлах. В конечном итоге такая структура получается недостаточно гибкой: распределение данных и политики обеспечения целостности полностью зависят от параметров и настроек дисковых массивов.

Новым витком развития параллельного доступа к данным стало изобретение объектной архитектуры хранения. Такая архитектура позволила абстрагироваться от физических устройств хранения и представить данные в виде логических объектов. Файл с данными разбивается на части – объекты. У каждого объекта есть индекс, который указывает, на каком устройстве хранения он расположен, а также – уникальный идентификатор. В идеальной объектной архитектуре устройство хранения может самостоятельно найти любой объект по заданному идентификатору. Файловая система при этом ничего не знает о физическом уровне хранения, имея дело только с логической структурой объектов.

Однако таких "интеллектуальных" систем хранения на аппаратном уровне пока не существует, хотя стандарт Object Storage Device (OSD), являющийся расширением стандарта SCSI, был принят еще в 2004 г. Выходом из этой ситуации стала эмуляция "интеллектуальных" СХД на программном уровне. В результате появились две параллельные файловые системы, программно реализующие объектную архитектуру хранения – свободно распространяемая Lustre и PanFS с протоколом параллельного доступа к данным DirectFLOW от компании Panasas, Inc.

Эти файловые системы разделяют уровни данных и метаданных – служебной информации о размере и свойствах файла. Если в обычных сетевых файловых системах данные и метаданные хранятся вместе и работа с ними происходит через один канал связи, то параллельные файловые системы с объектной архитектурой работают с этими уровнями по отдельности. Когда вычислительный узел кластера начинает работать с файлом, сначала он обращается к управляющему серверу за метаданными файла, имеющими очень небольшой размер. Вместе с данными о свойствах файла сервер метаданных выдает карту расположения объектов, на которые разбит файл, с информацией об их идентификационных номерах и месте хранения. После этого клиент работает напрямую с теми устройствами хранения, на которых лежат объекты нужного файла.

Поскольку файловые системы с объектной архитектурой реализуют единое глобальное пространство имен, то местонахождение файлов не привязано к конкретным физическим устройствам, и каждый файл представляет собой множество объектов, распределенных между ними, обеспечивая различные пути доступа к ним. Таким образом, параллельные файловые системы с объектной архитектурой обеспечивают одновременный прямой доступ множества клиентов к данным, и если несколько клиентов одновременно обращаются к одному и тому же файлу, "узкое место" устраняется за счет множественных путей доступа. Чем больше в системе хранения серверов OSD, тем больше существует параллельных путей к данным и, соответственно, тем выше суммарная скорость доступа к файлам. Такую систему очень легко масштабировать по производительности – достаточно добавить новые серверы хранения (рис. 1).

Файловая система Lustre обеспечивает основную функциональность параллельного доступа и объектной архитектуры хранения, но настройка этой системы для обеспечения максимальной производительности иногда занимает недели кропотливого труда. С этой точки зрения, гораздо удобнее пользоваться PanFS – все настройки выполняются автоматически. Кроме того, компания Panasas предоставляет готовое программно-аппаратное решение, объединяющее аппаратные модули управления DirectorBlade, модули хранения StorageBlade и параллельную файловую систему с объектной архитектурой. При этом инсталляция и настройка системы занимают не более получаса; еще меньшее время требуется на расширение системы: перераспределение объектов в едином глобальном пространстве имен происходит автоматически. Дополнительная функциональность решения Panasas включает встроенные средства мониторинга и динамической балансировки загрузки, адаптивную поддержку RAID на уровне файлов, дополнительные средства обеспечения отказоустойчивости и развитую, простую в использовании систему управления.

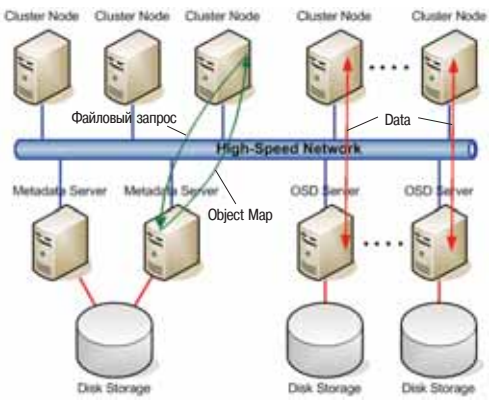


**Рис. 1.** Особенности систем хранения с параллельной файловой системой с объектной архитектурой: 1) данные на устройствах представляются "объектами", не связанными с настройками самих систем хранения; 2) разделенный доступ к метаданным и данным; 3) параллелизм в доступе к данным.

Бурное развитие высокопроизводительных вычислений и рост числа крупных кластерных систем привели к новым разработкам в области масштабируемых файловых систем и у лидеров рынка систем хранения данных, таких, как например, EMC и NetApp.

Компания NetApp предлагает решение Data ONTAP GX, которое обеспечивает доступ к данным через несколько NFS-серверов, объединенных между собой высокопроизводительной сетью. Каждый сервер в Data ONTAP GX обслуживает только часть единой файловой системы, но при этом обеспечивает переадресацию файловых запросов к другим серверам. Однако такое решение тоже имеет свои "минусы" – например, при одновременной работе многих клиентов с одним большим файлом или со многими файлами в одном каталоге, что достаточно типично для высокопроизводительных вычислений, сохраняется "узкое место", присущее традиционным сетевым системам хранения: все данные передаются по каналу связи между NFS-сервером, управляющим дисковым массивом, где хранится файл, и клиентом. При этом максимальная скорость чтения/записи ограничена производительностью сетевых интерфейсов этого сервера. В отличие от этого, в файловых системах с объектной архитектурой большой файл разбивается на множество объектов, к каждому из которых существует независимый путь доступа, и суммарная производительность может достигать десятков гигабайт в секунду. Чем больше объем данных, с которыми работают пользовательские приложения, тем на большее количество логических объектов они могут быть разбиты, и тем выше агрегированная производительность системы хранения. Поэтому такие решения, как Panasas, выгоднее всего использовать с большими кластерными системами с количеством узлов от нескольких сотен, где необходима максимальная скорость обмена данными.

Благодаря параллельной файловой системе с объектной архитектурой и единому глобальному пространству имен это решение не имеет ограничения по объему и может быть легко расширено до петабайт и более путем простого добавления модулей хранения и управления, причем агрегированная пропускная способность решения будет расти пропорционально объему. Недавно введенная Panasas поддержка 10G Ethernet позволяет получить агрегированную пропускную способность более 500 Мбайт/с для одного шасси.

Компания EMC также представила свое решение для высокопроизводительных вычислений – файловую систему MPFSi, обеспечивающую параллельный доступ к данным. Однако реализован этот доступ не за счет объектной архитектуры, а с помощью промежуточного запатентованного ПО, осуществляющего разделение потоков данных и метаданных на клиентской стороне и позволяющего получать данные непосредственно из системы хранения через протокол iSCSI. По производительности и масштабируемости EMC MPFSi может составить серьезную конкуренцию файловым системам с объектной архитектурой, но по простоте использования и управляемости решение от Panasas пока является одним из лучших, а последние разработки Panasas в области обеспечения целостности данных на всех уровнях, начиная от физических дисков и заканчивая передачей по сети, позволяют на сегодняшний день считать это решение наиболее оптимальным для использования в мультипетафлопных суперкомпьютерных системах. Например, в Национальной лаборатории Министерства энергетики США в Лос-Аламосе используется хранилище Panasas объемом более 2 Пбайт, которое демонстрирует пропускную способность более 50 Гбайт/с. Корпорация Boeing использует более 2 Пбайт СХД Panasas в качестве унифицированного хранилища для всего производственного цикла моделирования новых самолетов. И, наконец, один из транспетафлопных суперкомпьютерных проектов – суперкомпьютер Roadrunner, проектируемый IBM для Национальной лаборатории в Лос-Аламосе, который уже в 2008 г. покажет реальную производительность до 1.7PFlops – будет использовать более 2 Пбайт хранилища Panasas с агрегированной пропускной способностью более 100 Гбайт/с.

Многие компании при оценке производительности сетевых систем хранения ссылаются на результаты стандартного теста SPECsfs97. Однако их вряд ли можно считать показательными для таких систем, как Panasas. Во-первых, эти тесты сравнивают производительность NAS-систем, работающих с использованием стандартного протокола NFS, в то время как Panasas гораздо производительнее с "родным" параллельным протоколом DirectFlow: с его использованием система демонстрирует семикратный рост пропускной способности на той же конфигурации. Во-вторых, тесты SPECsfs97 измеряют производительность работы СХД с большим количеством маленьких файлов, что является нетипичной задачей для высокопроизводительных вычислений, где клиенты СХД, как правило, работают с большими файлами в параллельном режиме. Поэтому эти тесты не отражают реальную производительность систем хранения, предназначенных для использования в высокопроизводительных суперкомпьютерных системах с кластерной архитектурой.

***Андрей Слепухин,***
*руководитель Центра Кластерных Технологий "Т-Платформы".*