EMC Celerra MPFSi для HPC-применений

В конце 2000 г. компания EMC представила технологию EMC Celerra HighRoad – метод доступа к данным, позволяющий объединить лучшие черты NAS- и SAN-решений. Метод был реализован на основе мультиплексной файловой системы (MPFS – multiplex file system), включенной в файловый сервер Celerra DART (Data Access in Real Time), анонсированный EMC Corp. в ноябре 2000 г. В конце 2006 г. EMC выпустила новый релиз этого решения, ориентированного специально для HPC-применений и получившего название MPFSi (Multi-Path File System с возможностью использования iSCSI-протокола), развиваемой на базе файловой системы pNFS – параллельной NFS – и анонсированной*) IETF-комитетом (The Internet Engineering Task Force) в 2005 г.

Введение

Решение EMC Celerra HighRoad появилось в конце 2000 г. (SN № 4/5, 2000; http://www.storagenews.ru/05/Celerra_HighRoad.pdf), основной целью которого являлось увеличение пропускной способности файловых серверов. Это достигалось за счет того, что параллельно NAS-серверу подключалось FC система хранения. При этом NAS-сервер обрабатывал только команды управления файлами, а сами данные непосредственно пересылались от/к FC системе хранения.

После появления стандарта параллельной NFS – pNFS) компания EMC выпустила версию Celerra HighRoad специально для HPC-применений (High Performance Computing), основу которой составила технология MPFSi.*

Системы с разделяемым доступом к данным для HPC-применений

Большинство файловых серверов строятся на стандартных компьютерах/серверах. Основной их недостаток при использовании в составе HPC-решений – низкая производительность. Поэтому многие компании, работающие в области высокопроизводительных вычислений ищут пути решения этой проблемы, которых в настоящее время имеется уже около двух десятков. В общем семействе архитектур систем для разделяемого доступа к данным для HPC-решений технология EMC Celerra MPFSi позиционируется в качестве четвертой среди следующих:

- асимметричные out-of-band shared storage (или объектно-ориентированные, такие как OSD и OSS – Panasas, Lustre);
- симметричная in-band shared storage (или симметричные кластеры FS, такие разработки, как IBRIX, GFS, Polyserve);
- forwarding NFS-серверы (или NFS-кластеры – разработки компаний NetApp/Spinnaker);
- pNFS- или MPFSi-решения.

HPC-приложения часто требуют, чтобы большие файлы читались или писались вычислительными узлами одновременно. В то же время каждый индивидуальный узел может управлять только малой частью большого файла. При этом в целях обеспечения целостности данных требуется "жесткое" блокирование файла для выполнения всей последовательности операций от всех узлов. Это приводит к задержкам выполняющегося приложения, поскольку узлы ждут запросов ввода-вывода от других узлов. Для совместного использования больших файлов, было бы желательно делить набор данных на меньшие куски, которые могли бы блокироваться индивидуально, тем самым обеспечивая больший параллелизм узлам при доступе к данным. Но во многих приложениях деление полного набора данных или одного большого файла на части невозможно из-за непрерывности вычислительного процесса – в случаях, когда результаты одной стадии вычисления необходимы в качестве входных данных для следующей. Например, проведение анализа геологической информации в нефтегазовой промышленности обычно требует решения некоторого непрерывного волнового уравнения, в качестве входных данных для которого являются сейсмоданные объемом до 1 Тбайт и более.

Архитектура и особенности решения EMC Celerra MPFSi

Решение EMC Celerra MPFSi – комбинация патентованной технологии и NFS-протокола. MPFSi-решение значительно расширяет возможности NFS-серверов и делает возможным совместное использование файлов от сотен до тысяч клиентских HPC-узлов.

Основными компонентами MPFSi-решения (рис. 1) являются:

- файловый сервер Celerra со встроенным FMP-протоколом (file mapping protocol);
- множество HPC-узлов с установленным MPFSi-агентом и iSCSI-инициатором. Среди поддерживаемых ОС: HP-UX, IBM AIX, Sun Solaris, Red Hat RHEL, SuSE SLES, MS Windows 2000, MS Windows 2003;
- MPFSi-клиент LAN-инфраструктуры;
- одна или несколько FC систем хранения (например, CLARiiON или Symmetrix);

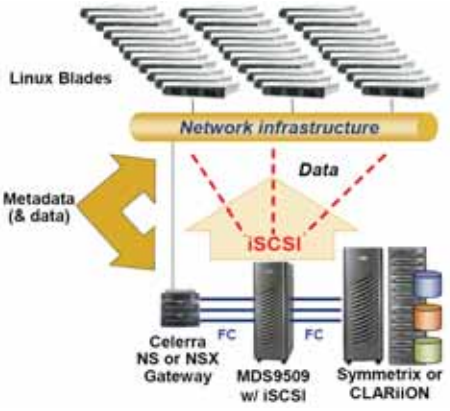

Рис. 1. Основные компоненты MPFSi-решения.

*) G., Gibson, "Parallel NFS (pNFS)", SNIA Developers Solutions Conference, http://www.snia.org/events/past/developer2005/pNFS.pdf, August 3, 2005

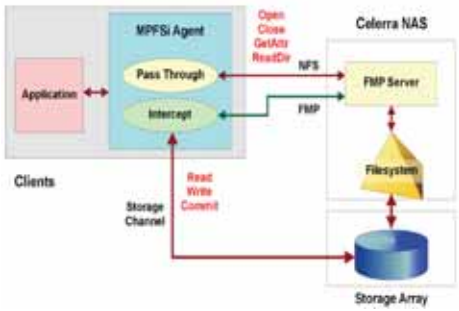

Рис. 2. Разделение потоков управления и данных файлового контента в MPFSi-решении.

– сетевая инфраструктура с iSCSI- и FC-портами.

На каждый HPC-узел инсталлируется MPFSi-агент, при этом каких-либо изменений в приложении не требуется. Этот агент, взаимодействуя с файловым сервером Celerra через FMP-сервер (в составе Celerra), отделяет поток данных в общем файловом контенте от потока NFS-метаданных, делая возможным напрямую перемещать данные между HPC-узлами (MPFSi-клиентами) и высокоскоростной FC системой хранения, используя iSCSI-соединения.

Все операции файловой системы, включая выделение (allocation) блока, блокирование файла, операции с метаданными и логи выполняются Celerra-сервером (рис. 2). Но перемещение данных файлового контента происходит непосредственно между MPFSi-клиентом и FC системой хранения без участия файлового сервера Celerra. Такое разделение до нескольких раз повышает скорость выполнения файловых операций, снижает задержки и до 10 раз увеличивает число HPC-узлов в системе при их подключении только к NFS-серверу.

MPFSi-агент

Программный MPFSi-агент выполняется на инсталлированной файловой системе на клиентской Linux-платформе. MPFSi-агент взаимодействует с файловым сервером Celerra для синхронизации, управления доступом и управления метаданными. MPFSi-агент обеспечивает следующие преимущества для HPC-узлов:

– *высокую пропускную способность:* *во-первых*, в отличие от NFS-протокола файловая система MPFSi имеет гораздо большие блоки данных (рис. 3), что значительно уменьшает накладные издержки при передаче данных; *во-вторых*, подключение клиентов непосредственно к дисковому массиву уменьшает TCP/IP накладные задержки, что также повышает пропускную способность; *в-третьих*, MPFSi обеспечивает высокопроизводительный разделяемый доступ к общим данным. В типовой HPC-среде, большое количество клиентов читают

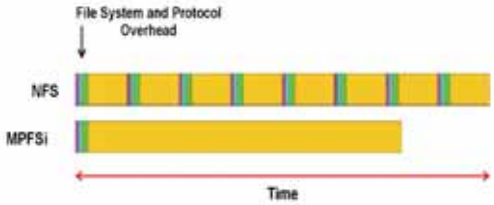

Рис. 3. Уменьшение накладных издержек в MPFSi-протоколе за счет большей длины блока данных.

одни и те же данные и даже один файл, и в этих случаях преимущества многоуровневого кэширования могут играть заметную роль. За счет того, что каждый узел имеет собственную низкую задержку, агрегированная пропускная способность HPC-кластера будет ограничиваться производительностью системы хранения;

– *улучшенное управление метаданными*: MPFSi-клиент также получает преимущества от очень эффективного механизма кэширования для схем файловых блоков (file block maps). Клиент при каждом обращении к FMP-серверу делает предварительную выборку большого числа схем файловых блоков, которые кэшируются и не требуют повторных обращений при передаче данных (рис. 4);

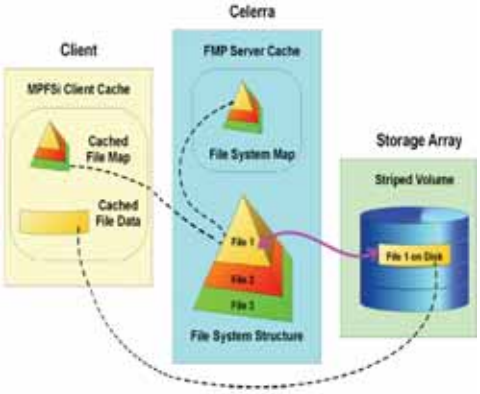

Рис. 4. За счет многоуровневого кэширования в MPFSi-решении не требуется повторных обращений при передаче данных.

– *параллельный доступ к данным:* когда HPC-узлы обращаются к одному файлу одновременно, данные уже располагаются в кэше дискового массива, тем самым обеспечивая очень высокую скорость доступа к единственному файлу. Далее, MPFSi-клиенты расширяют полосу пропускания за счет того, что при обращении к большому файлу они используют намного больше обращений ввода-вывода к дисковому массиву, чем NFS-пользователи;

– *эффективность небольших операций ввода-вывода*: когда приложение обращается с запросом по чтению маленького файла (меньше, чем блок файловой системы – 8 Кбайт), MPFSi-агент будет "просить" сервер послать файловый контент, используя один NFS-запрос. В результате данные будут доставлены вместе со схемой файла;

– *интеллектуальное кэширование данных*: MPFSi-агенты используют память узла для оптимизации доступа к данным и кэширования. Это, например, делается за счет накопления выполнения множества мелких операций в одной большой. Подобный анализ проводится и при чтении.

Файловый сервер Celerra

FMP-сервер, реализованный на сервере Celerra, – центральный компонент MPFSi-архитектуры. FMP-сервер имеет собственную файловую систему, которая ориентирована на формат собственных данных и управляет всем доступом к файловой системе. Файловые сервисы Celerra обеспечивают клиентов однородным представлением данных файла независимо от протокола (NFS) или клиентской операционной системы. Также Celerra обеспечивает параллельный доступ к одним и тем же данным множественных клиентов.

FMP-сервер вместе с другими базовыми файловыми сервисами Celerra обеспечивают следующее:

– резервирование дискового пространства для файловой системы;

– параллельный доступ к одним и тем же файлам через NFS;

– параллельный доступ к одним и тем же файлам для MPFSi-, NFS- и CIFS-клиентов;

– поддержка мелких операций ввода-вывода для MPFSi-клиентов по NFS, используя сетевую фабрику.

FMP-сервер может иметь и обслуживать одну или более файловых систем. Сервер также выполняет резервирование и предраспределение пространства для блоков файловой системы с отсроченными политиками выполнения. Это дает возможность MPFSi-записям распределять большие куски непрерывных блоков файла и затем освобождать их, когда файл закрыт. Такая технология поддерживает более высокую производительность для последовательного доступа, который, в свою очередь, может использовать кэши систем хранения при помощи более агрессивных параметров настройки выбора с упреждением. Наконец, FMP-сервер содержит арбитражную логику, чтобы поддерживать параллельный доступ к файловым данным от больших совокупностей вычислительных узлов. FMP-сервер связывается с клиентом посредством FMP и FMP-уведомлениями.

FMP-сервер управляет атрибутами inode-уровня для его файловых систем. Это дает возможность клиентам использовать их собственные кэши, чтобы считать метаданные и данные ввода-вывода локально, таким образом увеличивая параллелизм и производительность, как на уровне одного клиента, так и для множества узлов/клиентов.

В дополнение к основным функциям, отмеченным выше, MPFSi Metada Server поддерживает следующие сервисы для HPC-кластера:

– *аутентификация:* важная роль FMP-сервера идентифицировать и аутентифицировать MPFSi-клиентов при их доступе к системе хранения. Когда MPFSi-клиент хочет обратиться к файловой системе, FMP-сервер проверяет его идентичность и проводит его авторизацию, в соответствии с чем ему назначаются политики безопасности для его среды. Подобным образом файловый сервер Celerra аутотентифицирует NFS- и CIFS-клиентов;

– *управление доступом к файлам и директориям:* FMP-сервер обеспечивает MPFSi-клиентов структурой файлов в файловой системе. Когда MPFSi-клиент выдает запрос на операцию с каким-то файлом, сервер проверяет

разрешение и права его доступа к файлу, после чего посылает клиенту полную схему файла. Схема состоит из списка идентификаторов томов (volumes ID), смещений, а также количества блоков и посылается клиенту в оптимизированной форме, благодаря чему клиент может обратиться к массиву хранения и дисковым блокам непосредственно;

– *когерентность кэша*: MPFSi-клиент будет кэшировать данные локально в кэше узла. То, что соединение между FMP-сервером и клиентом существует, проверяется постоянными тестовыми импульсами между ними. Если соединение "падает", клиент останавливает любую активность ввода-вывода и лишает законной силы все кэшируемые карты файла и блоки данных. FMP-сервер управляет доступом к файлу на основе управления блокировками к разрешениям доступа к каждому блоку файловой системы. Каждый блок может быть разблокирован, блокирован для записи единственным клиентом или блокирован для чтения одним или множеством клиентов. FMP-сервер разрешает множество операций записи в случае, если блок изменяется одним клиентом. В этом случае сервер уведомляет всех клиентов, которые имеют открытый файл, используя FMP Notify протокол. Клиенты, которые обращаются к тому же самому блоку, уведомляются, что их локальный кэш устарел. Поэтому им следует еще раз обратиться к FMP-серверу для обновления схемы файла и далее обратиться к системе хранения, чтобы прочитать соответствующие блоки данных;

– *масштабируемость:* FMP-сервер выполняет две задачи по управлению метаданными: управление доступом к файлам/директориям и выделение блоков файловой системы, тем самым управляя трафиком данных от/к дисковому массиву для всех клиентов. Тестирование показывает, что, сравнивая подобные NFS-нагрузки, на первую задачу уходит 7% CPU (от полной нагрузки) и на вторую задачу – 4% CPU (от полной нагрузки). Оставшиеся 89% обычно используются на передачу данных. Поэтому число клиентов, которые могут быть обслужены одним MPFSi-решением, подобным Celerra, примерно в 10 раз больше, чем одним Celerra NFS-сервером (рис. 5). Что это означает на практике? Так, если NFS-сервер способен поддерживать работу 110 вычислительных узлов, то тот же сервер в составе MPFSi-решения сможет обслуживать уже 1000 вычислительных узлов при сохранении общего управления и административных затрат.

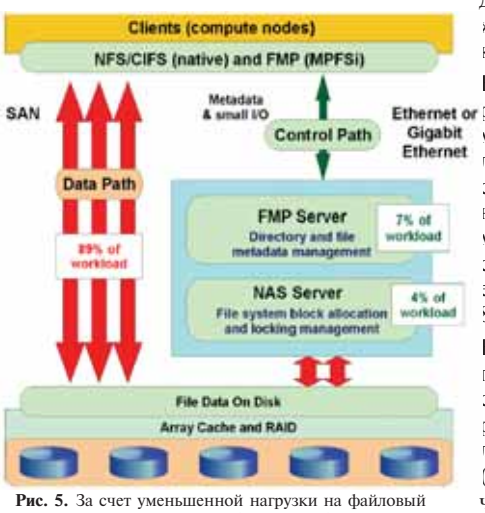

Рис. 5. За счет уменьшенной нагрузки на файловый сервер в MPFSi-конфигурации (в сравнении с NFS-конфигурацией) она может обслуживать массив HPC-узлов почти в 10 раз больший.

Измерение и сравнение MPFSi- и NFS-производительности

Для измерения*) MPFSi-производительности был использован Celerra NS704 файловый сервер с 4 блэйдами, соединенными с 1, 2, 3 и 4 CX700 массивами, каждый с более чем 100 дисками (146 Гбайт, 10 000 rpm). Синтетические генераторы нагрузки ввода-вывода (например, IOzone) выполнялись на 10–50 x86 Linux (RHEL 3.0) серверах.

MPFSi- и NFS-конфигурации были полностью идентичны, только в одном случае при обработке запросов от узлов управляющий поток и данные разделялись (данные напрямую пересылались между узлами и CX700 массивами), а в другом случае (NFS-конфигурация) весь файловый контент полностью проходил через Celerra NS704 файловый сервер.

На рис. 6 показано измеренная пропускная способность на случайных (random) операциях чтения для NFS- и MPFSi-конфигураций в зависимости от размера блока данных. В этом тесте каждый из 10 узлов выполнял 14 потоков команд. NFS-сервер был конфигурирован с 3 файловыми системами и управлял 32 потоками. NFS-конфигурация показала агрегированную пропускную способность – от 30 до 194 Мбайт/с. Самая большая разница (в 4 раза) между производительностями NFS- и MPFSi-конфигураций была зафиксирована для 2 Мбайт блоков.

На рис. 7 покана разница производительностей NFS- и MPFSi-конфигураций на последовательных операциях чтения в зависимости от размера блока данных. Производительность MPFSi-конфигурации выше – от 3 до 4 раз.

На операциях записи (рис. 8) производительность MPFSi-конфигурации также до 3 раз выше показанной на NFS-конфигурации.

В последнем тесте (рис. 9) измерялась разница в производительности в зависимости от числа HPC-узлов. Как показали дополнительные измерения, на последовательных операциях чтения с одним 1GE-SAN соединением (по iSCSI) максимальная пропускная способность составляет 130 Мбайт/с. При трех клиентах – 468 Мбайт/с, а при 10 – свыше 800 Мбайт/с.

В целом, MPFS позволяет повысить производительность обычной Celerra-системы до **30** раз (по заявлениям разработчика) за счет организации параллельного доступа приложений к данным (возможности читать большой файл частями), кэширования метаданных и использования нескольких путей доступа клиентов к массивам одновременно. При этом HPC-система может конфигурироваться до тысяч вычислительных узлов в составе до 4 Symmetrix и/или CLARiiON массивов.

*) Все измерения проводились в лаборатории компании EMC, *прим. ред.*

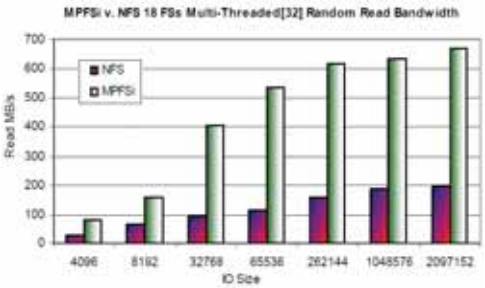

Рис. 6. Сравнение производительности NFS- и MPFSi-конфигураций на случайных операциях чтения в зависимости от размера блока данных.

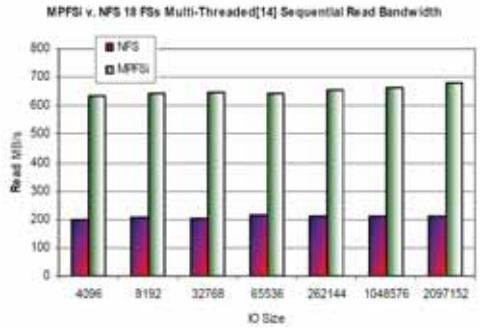

Рис. 7. Сравнение производительности NFS- и MPFSi-конфигураций на последовательных операциях чтения в зависимости от размера блока данных.

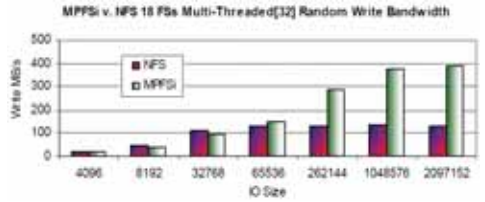

Рис. 8. Сравнение производительности NFS- и MPFSi-конфигураций на случайных операциях записи в зависимости от размера блока данных.

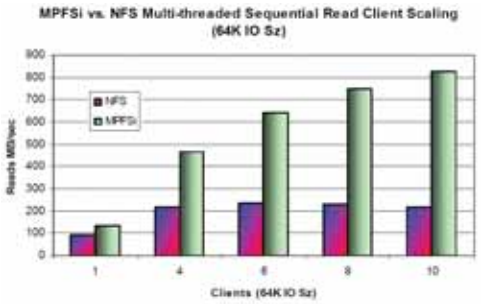

Рис. 9. Сравнение производительности NFS- и MPFSi-конфигураций в зависимости от числа HPC-узлов.

Заключение

HPC-решения во все большей степени начинают зависеть от производительности систем хранения, обеспечивающих раздельный доступ к данным для HPC-узлов. MPFSi-решение позволяет от 2 до 3 раз повысить производительность обычных NFS-серверов и до 10 раз – масштабируемость числа вычислительных узлов, поддерживаемых NFS-сервером. В наибольшей степени MPFSi-решение соответствует таким применениям, как: CAD-CAM-проектирование, газодинамические расчеты, обработка сейсмоданных, BI-анализ и др., которые в операциях ввода-вывода используют блоки 32 Кбайт и более.