Параллельная файловая система GPFS

Сергей Горбас – руководитель руководитель группы продаж IBM System x; IBM East Europe/Asia

Традиционные файловые системы в основном разработаны для среды одного сервера, так, например, популярная файловая система в Linux/Unix операционных системах – NFS – в высокопроизводительных вычислениях страдает тем, что в данный момент на запись с файлом может работать только один из вычислительных узлов (блокируя весь файл), а все остальные должны ждать освобождения очереди, чтобы записать свою порцию информации. При этом все остальные вычислительные узлы простаивают, не выполняя никаких полезных действий. А если этих узлов десятки или сотни? Простой может оказаться существенным, что значительно понизит эффективность использования такого кластера, который стоит достаточно дорого.

Поэтому многие разработчики пришли к выводу о необходимости создания новых специализированных файловых систем для вычислительных многоузловых кластеров, которые бы предоставляли кластерным приложениям параллельный, быстрый и равнозначный доступ к общим данным всем узлам одновременно. Одной из первых такого рода специализированной файловой системой стала IBM General Parallel File System (GPFS), появившаяся в 1998 году благодаря одному исследовательскому проекту в лаборатории IBM Almaden Research Center *(http://www.almaden.ibm.com)* и изначально разработанная как файловая система для обеспечения высокой производительности приложениям мультимедиа. Текущая версия – это уже 9 релиз файловой системы.

GPFS вобрала в себя многие идеи, разработанные изначально в университетском сообществе в последние несколько лет, включая технологии распределенных блокировок и восстановления, благодаря чему она, собственно, и является лидером в практически достигнутой производительности (>100 Гбайт/с на чтение/запись) и масштабируемости (более 2000 узлов в файловой системе GPFS). Не зря GPFS используется во многих суперкомпьютерах из списка Top500 *(www.top500.org)*.

Другие показатели масштабируемости GPFS:

- дизайн файловой системы позволяет расширять ее до 2^{99} байт ~ 633,825 Йотабайт (2^{80});
- максимальная протестированная система – 2 Пбайт (2000 Тбайт);
- максимальное количество файлов в файловой системе – 2,147,483,648 (для версий 2.3 и более поздних).

Функционально узлы в файловой системе GPFS делятся на два типа: те, которым необходим доступ к данным – это обычные GPFS-узлы, и те, которые предоставляют доступ к данным – Узлы Доступа или Network Shared disk/NSD в терминологии GPFS. Наиболее распространенная конфигурация файловой системы GPFS следующая: два или более Узла Доступа соединены по IP со всеми вычислительными узлами с одной стороны и по Fibre channel с несколькими системами хранения – с другой. Причем каждый Узел Доступа может задействовать два или более порта Fibre channel и два или более порта IP для увеличения пропускной способности каналов. При использовании нескольких Узлов Доступа отказ одного из них не ведет к потере доступа к данным файловой системы, доступ будет обеспечиваться через оставшиеся Узлы. IP каналы могут быть как по Ethernet, так и поверх Infiniband и Myrinet сетей. Начиная с версии 3.2, GPFS поддерживает родной протокол Infiniband – RDMA, т.е IP пакеты больше не нужно инкапсулировать в пакеты Infiniband, что даст только прирост производительности и снизит нагрузку на сеть Infiniband. Сеть доступа к данным может быть как совмещенной с сетью интерконнекта всех вычислительных узлов (используемой для счета), так и раздельной (для максимальной производительности).

Большее количество операций ввода/вывода в GPFS обеспечивается за счет разбиения данных файлов на блоки (stripes), которые варьируются в диапазоне от 16 Кбайт до 1 Мбайт, и разнесения их по различным дисковым массивам (RAID) и системам хранения, в результате чего операции чтения и записи блоков происходят в параллель – до нескольких тысяч дисков в самых больших установках GPFS, что позволяет при чтении/записи даже одного файла получить суммарную производительность всех дисковых систем, входящих в файловую систему GPFS.

Любая файловая система состоит минимум из двух составляющих: собственно данных и информации о их размещении на системе хранения (метаданные). GPFS обеспечивает полностью параллельный доступ как к данным, так и к метаданным, т.е каждый Узел Доступа имея доступ к каждому разделу в каждой системе хранения, может работать с разными частями данных и метаданных параллельно. Причем нет специализации серверов на работу с одним типом данных – все серверы работают с обоими типами данных, что обеспечивает максимальные масштабирование, отказоустойчивость и производительность.

Параллельные операции чтения и записи в файловой системе, осуществляемые множеством узлов, должны быть соответствующим образом синхронизированы, чтобы данные или метаданные не были повреждены. GPFS использует специальные протоколы распределенных блокировок для синхронизации доступа к разделяемым дискам, которые обеспечивают целостность данных, с одной стороны, и высокую пропускную способность файловой системы – с другой. Это означает, что при работе с файлами на запись GPFS может блокировать часть файла на уровне блоков для одного узла, а другие части – для других узлов, тем самым обеспечивая возможность многим узлам работать конкурентно с одним файлом параллельно как на запись, так и на чтение.

Файловая система GPFS имеет встроенный механизм репликации данных в дополнение к RAID-массивам дисковых систем для повышения отказоустойчивости. Когда он включен, то GPFS резервирует место для каждого блока данных и метаданных на двух разных дисковых системах, запись каждого блока в этом случае осуществляется в два места – в две разные системы хранения. Репликация может быть включена отдельно для данных и метаданных, причем, если репликацией данных будут заниматься сами системы хранения, например, через механизм PPRC, то это только разгрузит GPFS от необходимости записывать второй раз блок данных, при этом в случае выхода из строя одной из систем хранения, она прочитает данные со второй. После восстановления системы хранения данные автоматически будут синхронизированы.

Масштабирование важно не только для обычных файловых операций, но и для операций по обслуживанию самой файловой системы, особенно, если она большого объема. GPFS позволяет увеличивать/уменьшать объем файловой системы, производить замены дисковых систем благодаря возможности перераспределять полезные данные "на лету". При этом количеством вовлеченных в эти операции Узлов Доступа администратор может самостоятельно управлять, сохраняя при этом необходимый уровень производительности файловой системы для пользователей GPFS.

Одна из полезных особенностей – это объединение нескольких файловых систем GPFS в одну.

Когда это может быть выгодно?

- Кластеры могут монтировать файловые системы, принадлежащие и управляемые другими кластерами.
- Кластеры могут разделять данные и таким образом вычислительные ресурсы могут быть использованы более эффективно.
- Дисковые подсистемы могут быть объединены в группы в зависимости от их производительности.
- Когда требуется разделить вычислительные площадки и площадки данных (Grid).
- Объединение множества кластеров в единый суперкластер для решения более сложной задачи.

Немаловажно также, что при работе с файловой системой GPFS не требуется никаких изменений со стороны программного обеспечения: используется стандартный доступ к файлам. GPFS поддерживает операционные системы RedHat, SuSe и AIX.

Благодаря огромному количеству возможностей, которые предоставляет GPFS, она нашла свое применение не только как файловая система для высокопроизводительных кластерных систем, но и для таких бизнес-задач, как Oracle RAC, SAP Business Intelligent, видео-серверы и др., т.е., где необходима высокая производительность дисковой подсистемы.