Кластерные файловые системы: особенности, перспективы

Публикация является продолжением темы "Кластерные HPC-системы: особенности, перспективы" (SN 2/23, 2005). Хотя параллельные файловые системы имеют самостоятельное значение, их использование в настоящий момент, в основном, ориентировано на применение в составе HPC-систем, но в ближайший год-два это одно из самых быстроразвивающихся направлений, по прогнозам, будет широко использоваться в компьютерных системах более общего применения.

Введение

Активное развитие высокопроизводительных вычислительных систем (HPC – High Performance Computing) породило ряд проблем. В частности, рост производительности элементной базы и, в целом, производительности серверов в настоящий момент в области HPC опережает увеличение производительности обмена данными между внешними устройствами хранения (дисками).

Например, если ранее компания Dell при построении своих систем ориентировалась на показатель 1 Гбайт/с производительности ввода/вывода на каждый терафлоп компьютерной мощности, то теперь он уменьшается (Information-Week, June 20, 2005 – Victor Mashayekhi – старший менеджер Scalable Systems Group Dell). Это стало приводить к значительным простоям системы из-за ожидания ввода/вывода при обмене данными.

Далее, большие требования к записываемым данным заставили техасский Advanced Computing Center (университет шт. Техас, Austin, US) перейти на файловую систему Fusion от компании Ibrix вместо ранее использовавшейся NFS. Данный Центр поддерживает вычислительной мощностью исследования более 1700 ученых в области изучения турбулентности в аэродинамике (приложения жидкой динамики). "В среднем каждый процесс в программе требовал записи файла 300-400 Мбайт. Использование NFS при записи всех требуемых данных – приблизительно 20 Гбайт – составляло около 50 мин (по словам Tommy Minyard, менеджера группы Центра), что в целом составляло – 1 час при 10-минутной загрузке процессорной системы. С использованием технологии Irbix время записи для этой программы сократилось до 5 мин".

Или другой пример. "В течение последних 10 лет, – говорит Scott Studham, главный менеджер по технологиям National Center For Computational Sciences, Oak Ridge National Laboratory, – мы договаривались с нашим поставщиком систем хранения о цене за 1 Гбайт. Начиная с этого года и дальше, мы будем договариваться уже о стоимости за единицу пропускной способности" (InformationWeek, June 20, 2005).

А так прокомментировала использование Lustre, Michelle Butler, технический программный менеджер по хранению (technical program manager for storage) из NCSA (National Center for Supercomputing Applications, 1240-узловой, 9,8-teraflops кластер, названный Tungsten, – 20 место в списке top500 – июнь 2005): "Мы не хотим, чтобы компьютер за $8 млн простаивал из-за ввода/вывода. И если 5–10 лет назад мы не знали, что такое ввод/вывод данных, то теперь это – все".

Традиционно существует несколько способов масштабирования систем хранения. Высокостандартизированные сетевые системы хранения (NAS – Network-Attached Storage) используют популярные протоколы для совместного использования файлов в локальной сети, типа Microsoft CIFS или the Network File System (NFS – сетевая файловая система является стандартом на Unix- и Linux-системах). NAS использует недорогие Ethernet-связи между компьютерами, но передает данные со скоростью 1 Gbps (реальная скорость передачи данных много меньше из-за большой доли служебной информации в пакете протокола TCP/IP), что для большинства высоконагруженных приложений становится недопустимым.

В большинстве подобных ситуаций альтернативой NAS является SAN, в которых физической стандартной скоростью является 2 Gbps. При этом размер блока данных в пакете протокола FC на порядки превышает блок данных в TCP/IP, что создает хорошие условия практически для потоковой передачи данных без многочисленных подтверждений между ис-

точником и получателем данных в TCP/IP (что было необходимо для обеспечения надежности передачи данных в слабозащищенных сетях 60-70-х годов прошлого столетия). Развитие технологии iSCSI, являющейся мостом между FC и TCP/IP, также не снимает проблемы.

Обобщая, можно выделить следующие требования, которые предъявляются к HPC файловым системам/системам хранения:

- *множественный параллелизм для ввода/вывода данных;*
- *возможность объединения больших и малых операций ввода/вывода;*
- *возможность обслуживания множества (сотни и тысячи) распределенных вычислительных узлов;*
- *отсутствие единой точки отказа при интенсивной продолжительной работе;*
- *поддержка простых стандартных интерфейсов ввода/вывода, согласованные с POSIX-стандартом API, а также поддержка ввода/вывода с MPI;*
- *использование протестированных файловых систем.*

Обзор рынка

Все предложения на рынке в области глобальных файловых систем для параллельных кластеров можно классифицировать по нескольким направлениям. Строго говоря, все глобальные (кластерные) файловые системы делятся на две группы: параллельные (parallel) и разделяемые (shared SAN).

Параллельные файловые системы используют архитектуру клиент-сервер. Серверы подключаются к дисковым накопителям и обеспечивают доступ к данным для клиентов. Серверы, как правило, не имеют общих дисков, у каждого есть собственное дисковое пространство. Клиенты параллельных файловых систем, в отличие от серверов, имеют доступ ко всем серверам и, соответственно, ко всему дисковому пространству одновременно.

В *разделяемых* файловых системах нет серверов, к которым обращаются клиенты. Каждый сервер запускает ПО файловой системы и может иметь доступ ко всему дисковому пространству. В таком решении у серверов нет приватного дискового пространства, как у серверов в параллельных файловых системах. Необходимо заметить, что серверы могут экспортировать файловую систему через различные интерфейсы, например, как NFS – через Ethernet, таким образом, разделяемая файловая система может функционировать аналогично параллельной.

Исходя из приведенных архитектур и их реализаций, можно увидеть, что параллельные файловые системы лучше масштабируются, чем разделяемые. Это обусловлено более сложной архитектурой параллельных файловых систем, что в свою очередь ведет к несколько более сложному внедрению. Надо отметить, что разделяемые файловые системы, как правило, более гетерогенны, чем параллельные. То есть под каждую конкретную задачу архитектуру глобальной файловой системы надо выбирать индивидуально.

Глобальные файловые системы сами по себе интересны только с академической точки зрения. Практический интерес к ним возникает лишь при использовании в составе конкретного кластера, ориентированного на определенный тип задач, совместно с конкретным решением системы хранения. Необходимо отметить, что имеющиеся на рынке глобальные файловые системы на текущий момент мало стандартизованы и требуют определенных усилий по инсталляции в каждом проекте. Поэтому наибольший интерес представляют законченные решения: глобальная файловая система, интегрированная с системой хранения. На текущий момент такие решения продвигаются вендорами – непосредственными разработчиками и поставщиками кластерных систем, а также независимыми компаниями, являющимися интеграторами подобных решений. При этом следует учитывать специфику каждого региона в плане сервисной поддержки и обслуживания.

С учетом вышесказанного, рассмотрим более подробно 3 возможных варианта решений: PolyServe, решения на основе IBM General Parallel File System (GPFS) и HP StorageWorks Scalable File Share (HP SFS).

PolyServe Matrix Server

PolyServe – легкоустанавливаемая масштабируемая разделяемая кластерная NAS система (рис. 1). Данный продукт активно продвигают как IBM, так и HP.

PolyServe Matrix Server for Linux (PMS) – кластерное ПО для разделения данных, позволяющее клиентам заменять UNIX SMP серверы на более дешевые кластеры Linux-серверов. PMS позволяет множеству Linux-серверов функционировать как одна легкоуправляемая высокодоступная система. PMS имеет интегрированные сервисы для поддержания высокой доступности (Matrix HA) и балансировки нагрузки (Multi-Path I/O) и требует присутствия SAN.

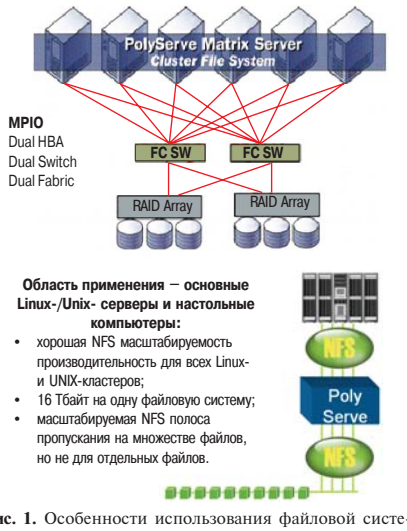

Рис. 1. Особенности использования файловой системы PolyServe на базе NFS.

Используемая PMS версия 3 NFS имеет ограничения по параллельности ввода/вывода, масштабируемости и надежности, но эти ограничения в 4 версии NFS (2006 г.) во многом будут сняты.

Ядром PMS является полностью распределенная и полностью журналируемая кластерная файловая система, которая поддерживает онлайновое добавление/удаление узлов и параллельный многоузловой доступ к общим данным. Все серверы в кластере PMS полностью равноправны.

PMS полностью поддерживается на Windows and Linux, а также на NFS- и Samba-серверах. В частности, PMS работает со всеми стандартными SAN-системами хранения HP: StorageWorks XP/EVA/MSA на протоколах FC или iSCSI с возможностью использования таких HA-сервисов, как мгновенные снимки и клонирование. Это хорошее решение для масштабирования стандартных SAN и NAS технологий. Текущие ограничения PMS: масштабируемость полосы пропускания – до 3 Гбайт/с, размер файловой системы – до 16 Тбайт, максимальное число узлов – 16.

IBM GPFS (General Parallel File System)

Общая параллельная файловая система – GPFS – впервые была анонсирована вместе с проектом ASCI (Accelerated Strategic Computing Initiative – ускоренная стратегическая вычислительная инициатива), инициированным для реализации оборонных программ Министерства энергетики США в сотрудничестве с лабораториями Lawrence Livermore и Los Alamos (США) с целью перехода от ядерных испытаний к методам, основанным на численном моделировании создания ядерного оружия, оценки его производительности и т.п.

В середине 2000 г. в рамках этого проекта была разработана суперкомпьютерная система ASCI White *(включающая IBM RS/6000 SP с 512 симметричными мультипроцессорными машинами – SMP-узлами; в каждом узле 16 POWER3 процессоров, с общим числом – 8192; система имеет общую память 4 Тбайт и дисковую память 150 Тбайт, с пиковой производительностью не менее 9,8 Тфлоп/с.; на июнь 2005 г. – 35 строка списка top500)*, на которой и была установлена GPFS. Дополнительно ASCI White была оснащена внешней дисковой памятью, архивной системой хранения данных HPSS (High Performance Storage System – для защиты вычислительных средств) и средствами визуализации.

Некоторое время спустя GPFS была доработана для Linux-систем. Сегодня она

представляет файловую систему в качестве основы высокопроизводительных кластерных решений IBM для обеспечения масштабируемого высокопроизводительного разделяемого доступа к данным от всех узлов в гомогенном или гетерогенном кластере, построенном на IBM UNIX-серверах под управлением операционных систем AIX 5L или Linux. Существует несколько вариантов реализации GPFS:

- GPFS для AIX 5L на процессорах POWER;
- GPFS для Linux на процессорах Intel и AMD (серверы xSeries);
- GPFS для Linux на процессорах POWER.

GPFS позволяет параллельно работающим приложениям осуществлять одновременный доступ к набору файлов (и отдельному файлу) от любого узла, на котором установлена GPFS, обеспечивая при этом высокий уровень управления по всем операциям файловой системы. Это, например, дает возможность множеству 3D-аниматоров или редакторов одновременно работать над различными частями одного видео-файла, что устраняет дополнительные издержки на хранение, управление и объединение многочисленных копий.

GPFS поддерживает стандарты файловой системы UNIX, и в отличие от большинства UNIX-систем, разработанных для функционирования в среде одного файлового сервера (добавление большего количества файловых серверов обычно не улучшает производительность файлового доступа), поддерживает намного более высокую производительность, масштабируемость и отказоустойчивость при параллельной работе множества файловых серверов.

GPFS обеспечивает высокопроизводительный ввод/вывод за счет разделения отдельных файлов на блоки (striping) и записи/чтения их параллельно на/с множество дисков (дисковых систем). Чтобы гарантировать последовательность данных в течение параллельного доступа, GPFS использует блок-уровневое блокирование, основанное на продуманной эстафетной системе управления. Это предотвращает одновременный доступ множества приложений/пользователей к одной и той же части (блоку) файла в один и тот же момент.

Кроме того, GPFS может читать или записывать большие блоки данных одной операцией ввода/вывода, минимизируя т.о. накладные расходы ввода/вывода. Помимо этого, GPFS имеет следующие преимущества:

- множество клиентов (приложения или пользователи) могут обращаться к отдельному файлу (в том числе и к отдельным его частям) одновременно. При этом продуманное управление блокировками предотвращает столкновения/конфликты и гарантирует целостность данных;
- поддержка блоков данных: 16/64/256/512/1024 Кбайт для оптимизации работы различных приложений;
- распознавание различных типов доступа к данным: последовательный (потоковый); случайный; нечеткий последовательный (strided) с целью оптимизации коэффициента стрипования;
- GPFS кэширует данные на стороне клиента с целью минимизации обращений к системе ввода/вывода;
- пользовательские данные и метаданные (данные транзакций файловой системы) могут реплицироваться, что обеспечивает их надежность хранения.

Отказоустойчивость GPFS поддерживается за счет множественности путей доступа к данным, которая также используется и для автоматической балансировки нагрузки. Кроме того, GPFS обеспечивает поддержку катастрофоустойчивости и высокой доступности решений в целом на его основе за счет:

- синхронизации зеркалирования на базе GPFS репликации;
- синхронизации зеркалирования на базе IBM TotalStorage Enterprise Storage Server/Metro-Mirror replication (прежнее название – Peer-to-Peer Remote Copy, PPRC);

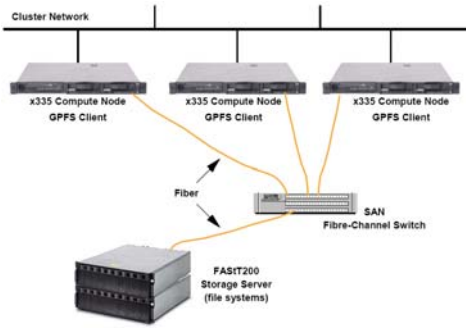

Рис. 2. Модель подключения дисковых систем, поддерживаемых GPFS, через SAN.

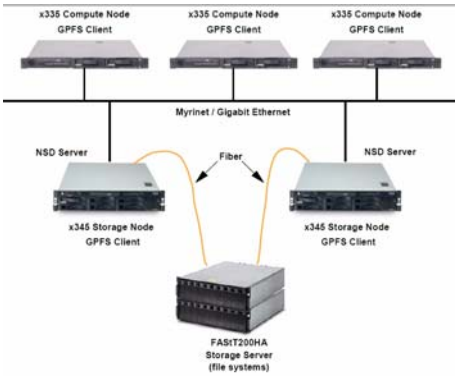

Рис. 3. Модель подключения дисковых систем, поддерживаемых GPFS, через NSD-серверы.

- асинхронного зеркалирования на базе IBM TotalStorage ESS FlashCopy.

Также GPFS дает возможность создания логической копии или "мгновенного снимка" полной файловой системы в указанное время, обеспечивая резервное копирование или зеркалирование в онлайновом режиме.

В целом GPFS можно масштабировать до сотен вычислительных узлов и более чем до 1000 дисков с общим объемом хранения свыше 200 Тбайт. В результате она может превосходить по быстродействию Network File System (NFS), Distributed File System (DFS) и Journaled File System (JFS).

Так называемые сетевые разделяемые дисковые компоненты (NSD – Network Shared Disk) в GPFS поддерживают 3 конфигурации: модель с подключением через SAN, NSD серверную модель и комбинацию обеих. В первом случае (рис. 2) NSD физически подключаются ко всем узлам кластера, которые получают доступ к файловой системе. В этой конфигурации NSD-компонент проверяет, что все узлы видят общедоступный диск и назначает ему общее имя. Далее оно транслируется GPFS к соответствующему локальному имени на каждом узле. Основное преимущество использования данной конфигурации в том, что GPFS не должен использовать кластерную сеть для пересылки данных от одного узла другому. В этом случае она может быть представлена дешевой IP-сетью и использоваться GPFS только для пересылки метаданных, а высокоскоростная используется лишь для соединения с дисками. Эта модель требует Fibre Channel SAN или IP SAN оборудования.

Во втором случае (рис. 3) некое подмножество полной совокупности всех узлов определяется как NSD-узлы хранения, и GPFS-диски подключаются только к этим узлам. NSD реализует программный уровень, который направляет требования ввода/вывода от вычислительных узлов к фабрике межсоединений (interconnect fabric), которая затем посылает требования к NSD узлу хранения, чтобы выполнить операцию ввода/вывода и переслать данные обратно вычислительному узлу. Этот удаленный ввод/вывод прозрачен для приложения, и конфигурация может быть более рентабельна (за счет экономии на FC-оборудовании), чем при полнодоступной SAN, поскольку кластер становится очень большим, но вероятность того, что ввод/вывод даже при ограниченном числе узлов может быстро стать узким местом весьма велика.

Список протестированных продуктов для GPFS представлен на "GPFS FAQ"

по адресу: http://publib.boulder.ibm.com/infocenter/clresctr/topic/com.ibm.cluster.gpfs.doc/gpfs_faqs/gpfs_faqs.html.

С момента поддержки GPFS стандартов файловой системы X/Open 4.0 большинство AIX 5L UNIX и Linux приложений могут использовать данные от GPFS без модификации и наиболее распространенные Linux- и UNIX-утилиты будут работать без каких-либо изменений. Кроме того, GPFS файлы могут экспортироваться в NFS, что делает их доступным клиентам вне кластера.

GPFS реализует Data Management API (DMAPI), который разрешает DMAPI-клиентам, таким агентам, как Tivoli Storage Manager Hierarchical Storage Manager, выполнять комплексные задачи по управлению данным на основе политик.

HP StorageWorks Scalable File Share (HP SFS)

Высокопроизводительные кластерные решения от HP на основе стандартных серверов и технологии Lustre (HP SFS) стали доступны сравнительно недавно – с конца 2004 г. Параллельная файловая система Lustre разрабатывается компанией Cluster File Systems (*www.clusterfs.com*) с 2000 г. и свободно распространяется *(www.lustre.org),* а компания HP является одним из основных ее спонсоров. На конец 2004 г. было свыше 100 коммерческих инсталляций Lustre, включая 8 из 16 самых больших Linux-кластеров в мире.

В основе технологии Lustre так же, как и в GPFS, лежит идея разбиения файла на несколько частей или потоков, так же, как это делается в обычных дисковых RAID. В настоящее время поддерживается уровень RAID-0, планируется поддержка RAID-1 и RAID-5. Lustre-протокол представляет собой Object Based Storage протокол, который обеспечивает значительно более высокую полосу пропускания для данных в системах хранения, ориентированных на HPC Linux- и Unix-кластеры с числом узлов 32-512. Область применения Lustre в составе HP SFS – высоконагруженные кластерные HPC Linux/Unix системы.

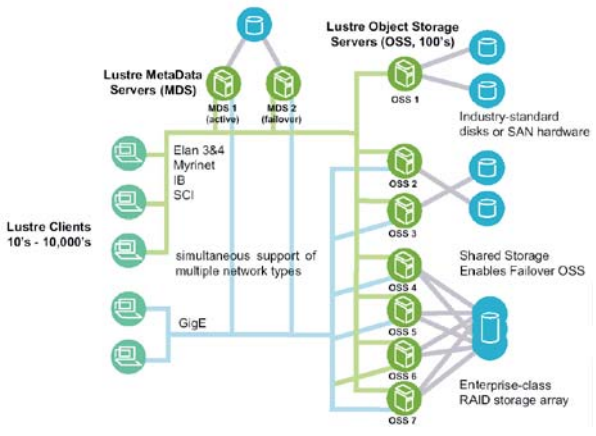

Рис. 4. Топология HP HPC-кластера на базе стандартных серверов и технологии Lustre.

Управление файловой системой осуществляется на уровне Object Storage Target (OST – дисковое пространство, экспортированное Object Storage Server'ом, аналогично NFS export, размер ограничен размером 2 Тбайт на Linux 2.4 или 4 Тбайт на Linux 2.6). Однако сами OST могут объединяться в одну агрегированную файловую систему с максимальной емкостью – 800 Тбайт. Среди других показателей Lustre:

- масштабирование системы по емкости – до 256 Пбайт (возможно и больше);
- пропускная способность – до 12 Гбайт/с (возможно и больше);
- масштабируемое число вычислительных узлов (клиентов) – до 10 000;
- масштабируемая надежность.

Технология Lustre, в отличие от ряда других параллельных файловых систем, обеспечивает построение очень гибких параллельных систем хранения (рис. 4) с использованием самых разных по емкости и производительности единичных Lustre Object Storage Servers (OSSs) – от low-end до high-end систем хранения и, что важно, без использования дорогих для интерконнекта компонент. Другой ее особенностью является то, что Lustre с подсистемой для обмена данными является не некой "примочкой" в дополнение к основному вычислительному кластеру, а основной компонентой, полностью интегрированной в систему.

Например, кластерное хранилище с общим объемом 100 Тбайт и пропускной способностью 10 Гбайт/с можно построить несколькими способами.

Вариант 1. 100 серверов хранения (OSS), соединенные по одному Gigabit Ethernet адаптеру в сервере. К каждому подключены системы хранения по 1 Тбайт со скоростью 100 Мбайт/с

Вариант 2. 4 мощных сервера хранения (OSS), соединенные по три канала Quadrics-Elan 4 в сервере. К каждому с помощью 16 FC 2 Gb каналов подключены системы хранения по 25 Тбайт с общей пропускной способностью 2,5 Гбайт/с.

Общая пропускная способность Lustre здесь ограничивается только пропускной способностью интерконнекта.

В конце 2005 г. и в 2006 г. в HP SFS будут добавлены некоторые особенности, которые уже доступны, например, в решениях PolyServe от HP. В частности – масштабируемая NFS полоса пропускания и поддержка защиты данных (мгновенные копии, зеркалирование и др.).

Также возможно, например, объединение HP SFS и HP PolyServe с целью использования HP SFS для расширения полосы пропускания для данных, а PolyServe – для доступа к данным, которые обычно совместно используются среди Linux-кластеров и других Linux- и Unix-клиентов.

Базовым элементом HP SFS является OSS, который строится на базе управляющего сервера – DL380 (с соответствующими адаптерами для интерконнекта) и подключаемого модуля – EVA3000 или SFS20 – специализированного низкостоимостного модуля для HP SFS (табл. 1). В реальности управляющий сервер и Lustre поддерживают все типы блокового хранения: FC, SCSI, SATA, ATA и, например, NVRAM. Отметим здесь, что для поддержки высокой доступности решения также не требуется дорогих коммутаторов и дополнительных компонент.

Табл. 1.

SFS Storage types	Disk type	Enclosure usable capacity (TB)	Read (MB/s)	Write (MB/s)	Disk capacities (GB)	Disks per enclosure
EVA3000 (2C2D)	FC	1,1	180	100	72	28 (2 полки)
EVA3000 (2C2D)	FC	2,3	180	100	146	28 (2 полки)
EVA3000 (2C2D)	FC	4,7	180	100	300	28 (2 полки)
SFS20	SATA	1,6	140	85	160	12
SFS20	SATA	2	140	85	250	11

Общие характеристики HP SFS на базе SFS20 с различными интерконнекторами приведены в табл. 2.

Табл. 2.

Interconnect type	Data protection	Usable capacity per SFS (TB)	Read per SFS (MB/s)	Write per SFS (MB/s)	Число SFS20
GbE	RAID 5	4	160	160	2
GbE	RAID 5	8	320	320	4
GbE	RAID 5	16	640	640	8
GbE	RAID 5	32	1280	1280	16
GbE	RAID 5	64	2560	2560	32
GbE	RAID 5	256	3200	3200	128
Quadrics ELAN4	RAID 5	25	1600	1300	16
Myrinet E (2XP)	RAID 5	51	2800	2600	32
Quadrics ELAN4	RAID 5	102	6400	5200	64
Quadrics ELAN4	RAID 5	204	12800	10400	128
GbE	RAID 5+1	32	1600	1600	32
Myrinet E (2XP)	RAID 5+1	51	2800	2600	64

Исходя из имеющийся экспертизы компании ЛАНИТ, можно сказать, что для небольших кластеров с ограниченными требованиями к пропускной способности на узел будут использоваться разделяемые файловые системы, так как они проще в построении при приемлемой масштабируемости и пропускной способности. Параллельные файловые системы применяются в больших вычислительных системах и являются своеобразным венцом развития глобальных файловых систем. Относительно основных игроков на этом рынке – Lustre и GPFS – можно сказать, что GPFS является более зрелой, однако, Lustre уже доказала свою работоспособность в реальных решениях, и, являясь независимой разработкой, обладает лучшей совместимостью с различным оборудова-

нием (системы хранения, серверы, интерконнект), тогда как GPFS изначально практически со 100%-ной вероятностью определяет поставщика всего решения, что не всегда удобно. Принимая во внимание планы развития Lustre *(http://www.clusterfs.com/roadmap.pdf)*, можно сказать, что это достойный, быстроразвивающийся продукт, который обязательно следует иметь в виду при разработке HPC-решений.

Дополнительные замечания

Помимо рассмотренных HPC-кластерных файловых систем, имеется достаточно большой спектр предложений других, менее известных и протестированных решений. В частности:

- *GFS (Global File System)* – изначально разработана компанией Sistina, приобретенной Red Hat в 2004 г. Исходные тексты GFS после этой покупки стали доступны для свободного использования. GFS это распределенная кластерная файловая система, работающая в среде SAN, предоставляющая возможность любому из узлов кластера работать с любыми объектами единой файловой системы, распределенной по различным устройствам хранения;
- *ADIC StorNext (имеется в портфеле HP)* – легкоустанавливаемая масштабируемая разделяемая SAN файловая система с расширенными HSM особенностями для HPC (неоткрытый стандарт, аналогична закрытым SAN-технологиям типа IBM TotalStorage SAN FS, Sun SAM-FS, SGI CXFS);
- *Parallel Virtual File System (PVFS – разработка Clemson University и Аргонской национальной лаборатории, http://www.parl.clemson.edu/pvfs/index.html)*. Под параллельностью здесь понимается то, что вывод каждого процесса параллельной задачи пишется поочередно на диски нескольких вычислительных узлов, и хранится каждый файл на нескольких узлах сразу, кусками фиксированного размера. Возможны 3 способа использования PVFS: работать с ней, как с обычной файловой системой (OPEN и пр. в Фортране); использовать стандартизованный MPI-2 интерфейс к параллельному вводу-выводу в реализации ROMIO; через собственные вызовы PVFS *(http://parallel.uran.ru)*;
- *параллельная файловая система (PFS)* является компонентом пакета программного обеспечения *Sun HPC Cluster Tools 4*. Она позволяет приложениям с параллельной архитектурой выполнять высокопроизводительные и масштабируемые операции ввода/вывода при работе с большим количеством параллельных систем хранения данных;
- *параллельная файловая система Fusion* от компании Ibrix;
- *разделяемая файловая система CXFS* от компании Silicon Graphics. Данная файловая система построена на основе 64-битной файловой системе SGI XFS с журналированием и предоставляет единую файловую систему, доступную для большого числа гетерогенных хостов (IRIX, Solaris, AIX, Windows, Linux), с возможностью разделения дисков в рамках сети хранения;
- *параллельная файловая система Terragrid* от компании Terrascale выделяется своей интересной реализацией – используется iSCSI протокол. На серверах хранения применяется специализированный iSCSI сервис, на клиентах – оптимизированный iSCSI инициатор, а для создания параллельных блочных устройств – драйвер md (программный RAID для Linux).

При выборе того или иного решения необходимо руководствоваться всей совокупностью факторов: совместимостью с прикладным ПО и ОС, поддержкой/сервисом, перспективой развития, требуемой надежностью и др., а не только основной заявляемой функциональностью.

Заключение

Появление параллельных кластерных файловых систем для HPC – значительный шаг вперед с точки зрения доступности и продвижения high-end технологий. И хотя сейчас их ориентация в основном на HPC-рынок, в ближайшей перспективе, по мнению аналитиков, их использование может быть в значительной степени смещено в сторону коммерческих систем.

Василий Кострюков,
vkostr@lanit.ru
Руководитель направления систем хранения данных, ЛАНИТ