Хранение данных в HPC-комплексах

Рынок систем для высокопроизводительных вычислений переживает бум. Подобные системы становятся востребованы во многих отраслях промышленности, в науке и образовании. Немаловажным аспектом создания высокопроизводительной системы является построение системы хранения данных, обладающей должной емкостью, производительностью и масштабируемостью. Статья рассматривает технологии построения систем хранения для высокопроизводительных вычислительных комплексов, а также принципы, тенденции и их реализацию.

Системы хранения для суперкомпьютеров: основные проблемы

Главная задача вычислительного комплекса, как, собственно следует из его названия, – проведение ресурсоемких вычислений над некоторыми данными. Данные могут представлять собой модель лопатки авиационного двигателя, результаты сейсмической разведки, модель нефтяного резервуара и др. Объем этих данных может быть очень и очень велик, и количество этих данных в подавляющем большинстве случаев будет заметно расти в процессе эксплуатации системы. Но самое главное – для вычислительного узла для быстрого проведения расчетов, как правило, требуется высопроизводительный доступ к данным, равно как и запись результатов вычислений с узла на систему хранения. Еще один момент: система хранения для вычислительного кластера не просто должна быть быстрой и емкой – она должна быть разделяемой, причем не только на уровне сети хранения данных, но и на уровне единой файловой системы – ведь заранее практически невозможно сказать, какой узел с какими данными будет работать.

Подходы к построению кластерных систем хранения

Традиционное решение – это выделение сервера хранения данных и предоставление доступа к его ресурсам посредством сетевых файловых систем, например, NFS. Для небольших систем с невысокими требованиями к производительности системы хранения простота, универсальность и невысокая стоимость традиционного решения являются, как правило, решающими аргументами при выборе. Тем не менее, производительность и масштабируемость этого решения сильно ограничены из-за природы протоколов NFS, IP, Ethernet, из-за чего применение NFS-шлюзов для систем среднего и старшего уровня как минимум затруднительно.

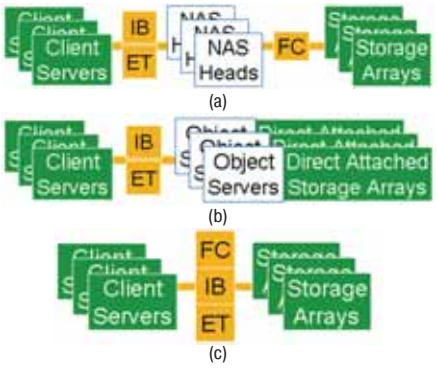

Рис. 1. Типы архитектур систем хранения для HPC-систем: (a) – кластерные системы NAS с поддержкой сетевых файловых систем; (b) – объектные системы хранения на базе параллельной файловой системы; (c) – масштабируемые системы SAN.

Возможные альтернативные решения при построении систем хранения данных для вычислительных систем, позволяющие получить более высокую производительность, хранить большие объемы данных и обслуживать одновременно сотни узлов, это (рис. 1):

- кластерные системы NAS с поддержкой сетевых файловых систем;
- объектные системы хранения на базе параллельной файловой системы;
- масштабируемые системы SAN.

Масштабируемые NAS-решения на базе кластерных файловых систем – HP Clustered Gateway

Одним из вариантов достижения высокой производительности при сохранении прозрачной совместимости со стандартами сетевых файловых систем является построение кластера файловых серверов с возможностью обслуживания единой файловой системы одновременно несколькими серверами, подключенными к единому пулу дисковых систем хранения за счет применения кластерной файловой системы. Примером такой реализации может служить решение HP Clustered Gateway[1]. Это решение отлично подходит для средних и больших вычислительных кластеров, позволяя строить большие хранилища данных, объединяя несколько дисковых массивов корпоративного уровня (таких, как HP EVA или XP) и обеспечивать высокую производительность доступа, объединяя до 16 серверов-шлюзов, параллельно работающих с одной файловой системой с использованием механизмов прозрачной балансировки нагрузки и схем обеспечения отказоустойчивости. При этом решение не требует никакого специального программного обеспечения на вычислительных узлах за счет поддержки повсеместно распространенных стандартов NFS или CIFS. Кроме того, использование стандартных протоколов позволяет обеспечить доступ к хранилищу данных не только с вычислительных узлов, но и с рабочих мест пользователей.

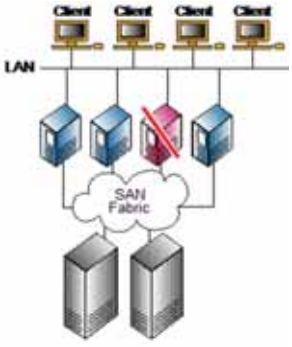

Рис. 2. Топология масштабируемых NAS-решений на базе кластерных файловых систем – HP Clustered Gateway.

1) Развивается на базе файловой системы "PolyServ", *прим. ред.*

Несколько ограничивает повсеместное применение Clustered Gateway в высокопроизводительных вычислительных системах использование в качестве back-end дисковых массивов относительно дорогих дисковых систем хранения корпоративного уровня. Кроме того, поддержка стандартных средств NFS и CIFS имеет и обратную сторону – протоколы Ethernet и IP менее производительны, чем альтернативные решения, построенные на базе технологий FibreChannel или Infiniband и RDMA, что не позволяет получить очень высокую производительность из расчета на один вычислительный узел.

HP Scalable File Share – масштабируемость прежде всего

Для тех вычислительных систем, которым требуется максимальная производительность и масштабируемость, наилучшим образом подходят решения на базе параллельных файловых систем – HP SFS (Scalable File Share)[2] – и использующих высокоскоростные сети в качестве среды для обеспечения доступа к данным. Система SFS строится на базе архитектуры "ячеек". Каждый узел хранения (ячейка) включается в коммуникационную сеть вычислительного кластера, например, Infiniband, что позволяет достичь очень высокой производительности доступа вычислительного узла к данным. В состав каждой ячейки входят два сервера, обслуживающих запросы пользователей, и подключенные к ним дисковые полки, на которых хранятся данные. Благодаря этому емкость и производительность системы легко и практически линейно масштабируются простым добавлением новых ячеек. Так, производительность системы может достигать 35 Гбайт/с, а размер одной файловой системы – 512 Тбайт. При необходимости возможно получение и более внушительных характеристик, архитектурных ограничений масштабируемости практически нет.

Рис. 3. Внешний вид HP StorageWorks Scalable File Share (HP SFS) на базе параллельной файловой системы.

Все ячейки, входящие в состав системы, обслуживают единую общую файловую систему, доступную со всех узлов вычислительного кластера. Распределение данных по ячейкам может осуществляться по различным принципам, в зависимости от природы данных и схем доступа к ним: возможно как "размазывание" каждого файла на все ячейки, входящие в состав системы (страйпинг), так и хранение отдельных файлов на отдельных ячейках. При этом настройки страйпинга могут быть заданы не только для файловой системы, но и для каждого каталога и для каждого конкретного файла.

Применяемые в параллельных файловых системах принципы хранения информации, с одной стороны, позволяют достичь большей производительности и практически линейной масштабируемости, но, с другой, предъявляют серьезные требования к надежности хранения данных и обеспечения непрерывности доступа. Действительно, сбой в работе ячейки, на которой хранится часть данных общей файловой системы, особенно при использовании страйпинга, может привести к серьезным последствиям.

Решение HP SFS предусматривает ряд мер по обеспечению доступности системы. Каждая ячейка представлена двумя параллельно работающими серверами, дисковые полки подключены одновременно к двум серверам. Применение кластерных технологий позволяет сохранить доступность данных даже в случае выхода из строя одного из серверов, адаптера, обрыва кабеля и т.д. Для защиты самих данных применяются как технологии RAID различных уровней защиты (RAID5, RAID6), так и средства зеркалирования полок, востребованные для конфигураций, предназначенных для хранения критически важных данных.

Высокая производительность системы объясняется, прежде всего, использованием низкоуровневого API Infiniband, что позволяет сократить до минимума накладные расходы на передачу данных. С другой стороны, использование низкоуровневых средств несколько ограничивает применимость решения: в качестве клиентов поддерживаются только узлы с ОС Linux с рядом установленных дополнительно модулей ядра. Впрочем, это не является большой проблемой при построении интегрированного решения, а для тех клиентов, которым все-таки требуется доступ через Ethernet, NFS или CIFS, в составе SFS предусмотрены соответствующие средства.

В качестве серверов хранения применяются стандартные, широко распространенные серверы HP ProLiant, в качестве дисковых систем хранения – недорогие системы HP SFS20 или EVA4100. Применение стандартных серверов, систем хранения и дисков позволяет достичь очень привлекательных соотношений цена/емкость, цена/производительность, при этом обеспечивая максимальные производительность и масштабируемость. Совокупность этих качеств и позволяет SFS быть наиболее привлекательным и распространенным вариантом для построения масштабируемых кластерных систем хранения данных.

Масштабируемые решения на базе SAN с поддержкой блочного доступа

Решения на базе технологий SAN (Storage Area Networks) наиболее востребованы в тех вычислительных системах, где используется относительно небольшое количество узлов, с высокими требованиями к надежности хранения и максимальной производительности доступа к данным из расчета на каждый вычислительный узел.

Эти решения предусматривают оснащение каждого вычислительного узла адаптерами FibreChannel и прямое подключение всех узлов к пулу дисковых массивов. В качестве дисковых массивов для решений SAN, как правило, используются дисковые массивы корпоративного уровня (например, семейств HP EVA или XP), что позволяет достигать высокой надежности и производительности. Главная сложность при построении таких систем – организация единой файловой системы, к которой имеют доступ все вычислительные узлы и которая задействует несколько дисковых массивов одновременно.

В качестве примера решения на базе SAN с поддержкой блочного доступа и единой файловой системы можно привести Quantum StorNext, отличительной особенностью которого является поддержка иерархического управления жизненным циклом информации, что позволяет, например, осуществлять на основе политик миграцию редко используемых данных на более дешевые диски или ленточные накопители. Кроме того, файловая система StorNext поддерживает возможность предоставления доступа к данным, не только подключенным непосредственно к узлам, но и через выделенные серверы-шлюзы и сеть Ethernet.

Одним из основных ограничений использования масштабируемых решений SAN является их относительно высокая стоимость за счет использования дисковых массивов корпоративного уровня. Наиболее востребованы такие решения в вычислительных системах с "толстыми" узлами с большим числом процессоров в каждом, с большим объемом хранимых данных, с высокими требованиями к производительности доступа от каждого узла к данным. Широкое применение такие системы нашли также и в вещательных компаниях, где количество серверов, работающих с информацией, не так велико, но при этом объемы данных весьма большие.

Заключение

Выбор архитектуры системы хранения очень важен и во многом определяет эффективность работы вычислительного комплекса, поэтому подходить к вопросу выбора стоит с максимальной серьезностью. Система хранения должна рассматриваться как неотъемлемая, интегрированная часть вычислительной системы, она должна быть сбалансированной и соответствовать по характеристикам используемым приложениям, вычислительным узлам, наборам данных.

Современные вычислительные системы предъявляют очень серьезные требования к системам хранения данных, что привело к возникновению и развитию новых технологий и решений, позволяющих им соответствовать. Системы хранения для высокопроизводительных вычислительных систем строятся на базе разных подходов, "заточенных" под различные применения, что позволяет подобрать на этапе проектирования системы оптимальное решение по хранению данных для каждого конкретного Заказчика, для каждой конкретной задачи.

Евгений Лагунцов
системный архитектор,
Enterprise Servers & Storage, HP Россия

2) Развивается на базе файловой системы Lustre, *прим. ред.*