Распределенное grid-хранение: настоящее и будущее

В середине 2006 г. компания Network Appliance анонсировала выход своей новой файловой системы – Data ONTAP GX, ориентированной на параллельные вычисления. Данное объявление явилось результатом двухлетней работы по интеграции файловой системы SpinOS (технологии компании Spinnaker Networks, перешедшие к NetApp после ее приобретения) с ключевыми особенностями Data ONTAP 7G.

Введение

Компания Spinnaker Networks приобретена Network Appliance в ноябре 2003 г. за $300 млн и была одним из пионеров в области масштабируемой системной архитектуры, распределенных файловых систем, технологий кластеризации и виртуализации, показав еще в 2002 г. на тестах spec.org (SPECsfs97_R1.v3) результаты, которые во многом остаются еще актуальными и сегодня. Данное приобретение дало возможность Network Appliance значительно развить и расширить ее концепцию Storage Grid архитектуры.

В 2004 г. NetApp объявила о начале разработки новой операционной системы следующего поколения – Data ONTAP GX, которая уже в ближайшей перспективе должна стать основой для построения корпоративных распределенных IT-инфраструктур в составе ее ONTAP GX Storage Grid решений (рис. 1). Она позволит консолидированно на основе глобальной виртуализации ресурсов и файловых структур управлять всеми базовыми технологиями хранения (NAS, SAN и др.) на основе распределенной в широких пределах гибкомасштабируемой среды хранения.

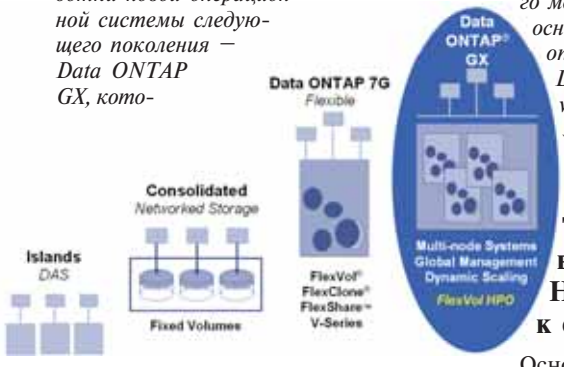

Рис. 1. NetApp "видит" Data ONTAP GX как операционную систему следующего поколения, которая уже в ближайшей перспективе должна стать основой для построения корпоративных распределенных IT-инфраструктур в ее решениях.

Первый этап разработки (до июня 2006 г.) Data ONTAP GX Storage Grid ставил целью представить современное решение для HPC-применений (High Performance Computing) и обработки цифрового медиаконтента на основе интеграции основных функциональных возможностей операционных систем от NetApp – Data ONTAP 7G и от Spinnaker Networks – SpinOS (рис. 2). На втором этапе (2009 г.) Data ONTAP GX уже должна стать базовой ОС для основных разработок NetApp.

Требования высокопроизводительных HPC-систем и медиарешений к системам хранения

Основные параметры ONTAP GX Storage Grid решений определяются требованиями приложений, прежде всего – HPC- и медиазадач, на которые они ориентированы.

Современные HPC-системы отличаются очень широкими требованиями к системам хранения, обслуживающим доступ к данным в составе HPC-кластеров/SMP-систем и медиарешений. В табл. 1 приведены характеристики наборов данных и требования к архитектуре основных отраслевых HPC-приложений и медиарешений.

Одно из *первых* требований HPC-систем к системам хранения – возможность поддержания самой высокой производительности HPC-узлов на уровне файловой системы, а также ее масштабируемость (до десятков гигабайт в секунду и миллионов файловых операций) в зависимости от класса задач, объема вычислений и этапа развития самой HPC-системы. *Во-вторых* – минимизация сложности управления приложениями. Как правило, HPC-центры консолидиру-

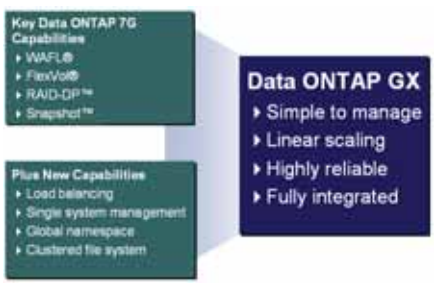

Рис. 2. ОС Data ONTAP GX вобрала в себя лучшее, что было в составе операционных систем от NetApp – Data ONTAP 7G и от Spinnaker Networks – SpinOS .

Табл. 1. Характеристики наборов данных и требования к архитектуре основных отраслевых HPC-приложений

Отрасль	Приложения/задачи	Наборы данных/файлы	Требования
Нефтегазовая	Seismic Processing	Very large image data sets; many intermediate versions	Very high aggregate I/O to storage; multiple jobs generate hot spots
	Reservoir Modeling	Many small files	Massively compute bound
Индустрия развлечений	Renderman; Maya, Softimage; ray tracing	Very large files; 2D and 3D frames; textures; compositing	Hot spots; concurrent access to data sets
Автомобильная и космическая	Computational fluid dynamics; crash simulation; finite element analysis	Large files; many intermediary files	Data availability; performance; storage hot spots
Разработка чипов	Cadence; Synopsys; Mentor	Large files	Scalability; availability; performance; hot spots
Разработка программного обеспечения	Rational ClearCase	Mixed, large, and small files; replicated source trees.	Remote collaboration; hot spots during compilation

ют решения многих вендоров, что создает определенные трудности по их управлению. Другой аспект в том, что развитие самих приложений в минимальной степени должно зависеть от самих уже имеющихся средств управления. *В-третьих* – обеспечение повышенной надежности вычислений. Поскольку отдельные задачи могут считаться днями и даже неделями, то HPC-системы должны иметь необходимый уровень надежности вычислительного процесса, позволяющий проводить реконфигурацию системы и восстановление счета в случае возникновения сбоев/отказов.

Медиарешения менее критичны, но также требовательны к выполнению ряда условий. *Во-первых*, высокой эффективности с точки зрения стоимости ресурсов хранения ($/GB), их использования, управления и масштабирования. *Во-вторых*, возможности использования аналоговых технологий – ленты и JBOD в качестве альтернативы. *В-третьих*, возможность масштабирования по производительности до десятков гигабайт в секунду, а по емкости от десятков терабайт до десятков петабайт. *В-четвертых*, надежное и простое управление контентом на протяжении всего жизненного цикла.

Архитектурные особенности ONTAP GX Storage Grid решений

Не занимаясь непосредственно разработкой аппаратных средств, NetApp всех основных преимуществ своих продуктов добивается на базе специализированного программного обеспечения, инсталлируемого на стандартные аппаратные компоненты. Поэтому вся полнота функциональности NetApp достигается за счет "интеллектуализации" ПО при максимально полном использовании "устоявшихся" аппаратных решений.

Файловая система Data ONTAP GX была развита на основе Data ONTAP 7G с включением всех основных ее преимуществ – адаптивности, управляемости, обработанности, включая такие ее компоненты, как: WAFL, FlexVol, RAID-DP и Snapshot. От SpinFS в GX были добавлены такие возможности, как: глобальное пространство имен, балансировка нагрузки, единая система управления (см. рис. 2). Ряд возможностей был разработан только в процессе интеграции систем.

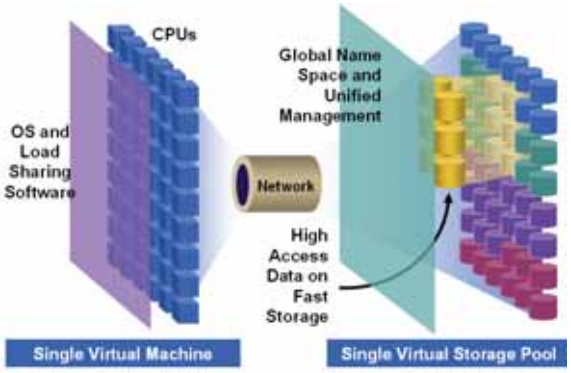

Рис. 4. Логическое представление архитектуры ONTAP GX Storage Grid решений.

Логическое представление архитектуры ONTAP GX Storage Grid решений

Логическое представление архитектуры ONTAP GX Storage Grid решений полностью виртуализировано. Оно включает такие виртуальные понятия, как: FlexVol-том, виртуальные серверы, виртуальные интерфейсы, виртуальный пул хранения и ряд других. Общая цель виртуализации – повышение управляемости системы, улучшение (в ряде случаев – значительное) технических характеристик системы: производительности, надежности, сервиспригодности и др.

Основная единица хранения в ONTAP GX Storage Grid системе – FlexVol-том, или гибкий том, который является логическим контейнером для набора логически связанных директорий и файлов. Гибкие тома создают единую логическую область хранения – *виртуальный пул хранения*, который может бесшовно управляться в составе различных аппаратных платформ и географически распределенных кластеров. Гибкие тома "нарезаются" из т.н. агрегатов – физических "кусков" систем хранения (рис. 4). С точки зрения иерархического представления, гибкий том постоянно находится в пределах агрегата. *Виртуальный сервер* обеспечивает доступ к набору гибких томов через набор *виртуальных интерфейсов*. Виртуальный сервер связан с одним или более агрегатами, одним или более узлами и одним или более физическими интерфейсами через интерконект клиентской сети.

Кластерное распределенное хранение с единственным глобальным пространством имен

ONTAP GX Storage Grid делает общую емкость полного распределенного кластера хранения (до 24 узлов) видимой для клиентов (HPC-узлов) под единственным глобальным именем, что позволяет всем сетевым клиентам обращаться к ресурсам хранения через одну точку (рис. 3). Все сетевые клиенты видят единственную большую файловую систему и могут обращаться к любой части

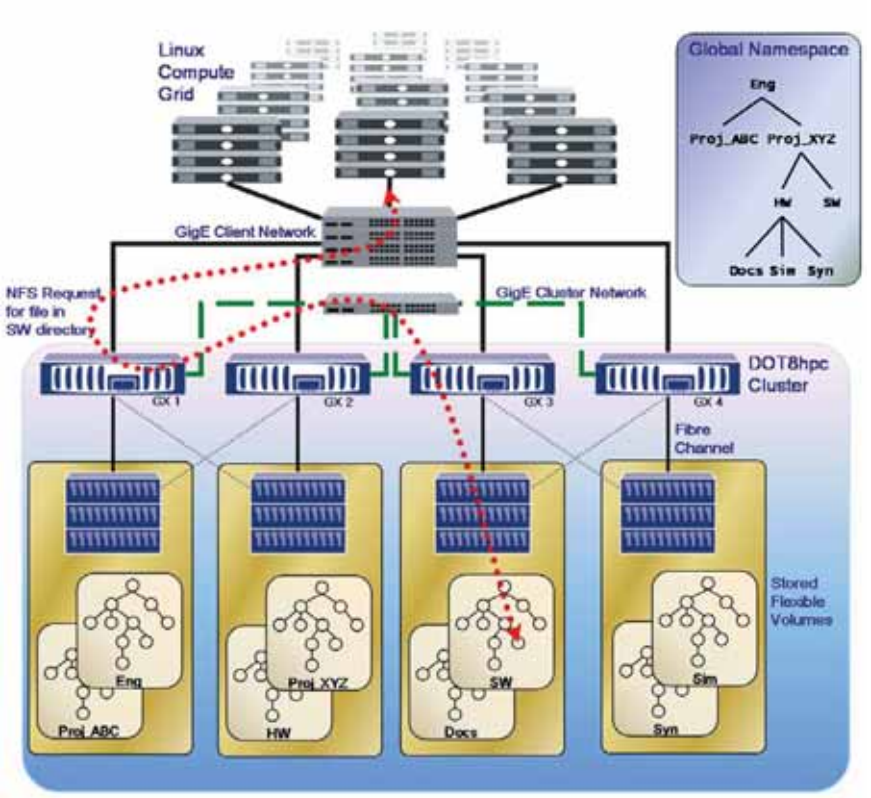

Рис. 3. Система ONTAP GX Storage Grid поддерживает глобальное единое пространство имен для всех вычислительных узлов. При этом полностью решаются проблемы масштабирования узлов хранения, а HPC-клиент полностью абстрагирован от физического места расположения данных.

глобального пространства имен без необходимости установления конкретного узла, на котором физически хранятся требуемые данные. Это решает общую проблему масштабируемости множества файловых серверов и управления консолидированной схемой расположения файлов. Доступ к данным обеспечен через стандартный NFS-протокол, при этом не требуется установка какого-либо клиентского программного обеспечения.

Линейная масштабируемость по производительности и емкости

ОС Data ONTAP GX обеспечивает линейное увеличение производительности

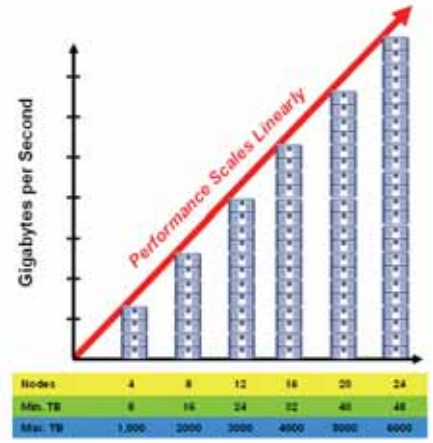

Рис. 5. Одно из основных преимуществ систем на базе ОС Data ONTAP GX – их линейная масштабируемость по производительности *(по результатам тестирования разработчика, прим. ред.).*

ONTAP GX Storage Grid системы (рис. 5) по мере добавления в нее GX-узлов (по результатам тестовых испытаний разработчика, *прим. ред.*). Как только GX-узлы добавляются к кластеру, все физические ресурсы (CPU, кэш-память, сетевая пропускная способность ввода-вывода и дисковая пропускная способность) автоматически балансируются (за счет использования специальных технологий), что делает GX-систему высокоэффективной компонентой для тысяч Linux вычислительных узлов.

Официально зарегистрированная производительность ONTAP GX Storage Grid системы из 24 узлов (FAS6070) на тестах SPECsfs97_R1.v3 (*www.spec.org*) составляет 1032461 sfs-команд/с, что на текущий момент является лучшим показателем в отрасли для данного класса решений. Масштабируемость ONTAP GX Storage Grid по потоковой производительности – от 1 Гбайт/с до 6 Гбайт/с.

Максимальная масштабируемость систем ONTAP GX Storage Grid по емкости хранения составляет 6 Пбайт.

Максимальное число 1GE портов на 1 узел (FAS6070) – 20, максимальное число FC-портов на 1 узел (FAS6070) – 16. В каждый узел (опционно) может включаться до 4 FC-портов для лент.

Балансировка нагрузки

Борьба за производительность в специализированных файловых системах для HPC-решений, по сути – решающая причина их разработки. Главный фактор

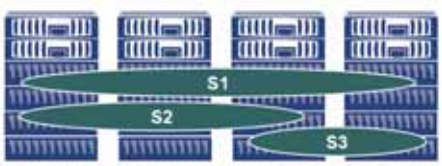

Рис. 6. За счет опции FlexVol HPO Data ONTAP GX происходит "размазывание" томов (S1, S2, S3) по узлам кластера с учетом критичности приложений/директорий, связанных с этими томами.

снижения производительности в системах хранения (в том числе и для HPC) – появление узких/горячих мест в системах хранения. Один из основных способов устранения этого явления – распараллеливание/стрипование запросов к данным. В Data ONTAP GX есть 3 опции FlexVol, позволяющие реализовать этот принцип. *Во-первых*, опция FlexVol HPO (High Performance Option) дает возможность "размазывать" тома критичных приложений/директорий по множеству узлов GX-кластера, обеспечивая тем самым мультигигабайтную пропускную способность для критических задач (рис. 6). По мере увеличения требований к пропускной способности/добавления GX-узлов стрипование тома может автоматически расширяться без прерывания работы приложения.

Во-вторых, FlexVol дает возможность увеличить производительность по чтению за счет асинхронного зеркалирования файла/тома на другие GX-узлы. Это связано с тем, что HPC-приложения часто требуют масштабирования пропу-

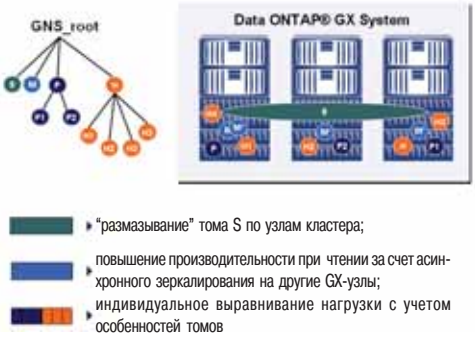

Рис. 7. Три опции FlexVol для выравнивания нагрузки GX-кластера.

скной способности по чтению вне этапа записи файла.

В-третьих, FlexVol дает возможность индивидуально управлять агрегированной производительностью подобных томов и связанных с ними операций. Например, меткой P маркированы тома, связанные с проектами, а меткой H – домашние директории. На рис. 7 все они равномерно распределены по GX-узлам.

ONTAP GX Storage Grid системы могут конфигурироваться всеми тремя опциями одновременно. Одновременно выполняться могут дв или одна опция.

Обеспечение высокой доступности ONTAP GX Storage Grid систем

Поддержание высокой доступности GX-систем строится на базе отсутствия единой точки отказа в GX-системе, а также ряда решений обеспечения доступности.

Во-первых, принципы по поддержанию высокой доступности, заложенные в системе Data ONTAP GX, дают возможность проводить апгрэйд аппаратного или программного обеспечения без останова системы/приложений. В случае какой-либо аппаратной проблемы на одном из GX-узлов данные могут быть автоматически и прозрачно для приложений перенесены на другой узел, пока отказавший узел не будет восстановлен или заменен.

Во-вторых, высокая доступность поддерживаемая ONTAP GX, строится на том, что все GX-узлы соединены в пары, у каждого из которых есть доступ к данным другого. Если отказывает один узел, второй автоматически берет управление на себя. До появления сбоя оба узла выполняют рабочие нагрузки.

В-третьих, надежность от сбоев на дисках стандартно поддерживается на уровне RAID-групп, в частности, на основе RAID-DP (защита от сбоев на двух дисках). Катастрофоустойчивость GX-системы поддерживается на базе FlexVol зеркалирования и организации географически разнесенных GX-кластеров. Также локально поддерживается резервное копирование томов на основе NDMPv4-протокола (сертифицировано с продуктами CommVault, BakBone, Symantec NBU 6.0 и EMC Legato Networker 7.3).

В-четвертых, в ONTAP GX введено понятие виртуального интерфейса. Это означает, что с каждым физическим портом GX-узла связано несколько виртуальных (или виртуальных IP-адресов). Аналогично каждый виртуальный порт может быть связан с несколькими физическими. Виртуальные интерфейсы конфигурируются так, что, если порт узла "падает", виртуальные интерфейсы мигрируют к другому порту. Если бы это не поддерживалось, в ситуациях, когда возникают сетевые аппаратные сбои/отказы или сбои в каналах связи, приложение могло бы стать недоступным.

Поддержка уровневого хранения

ONTAP GX Storage grid может поддерживать несколько уровней хранения, например, первый (один из GX-узлов) реализуется на базе FAS6070 с FC-дисками – для критичных томов, второй

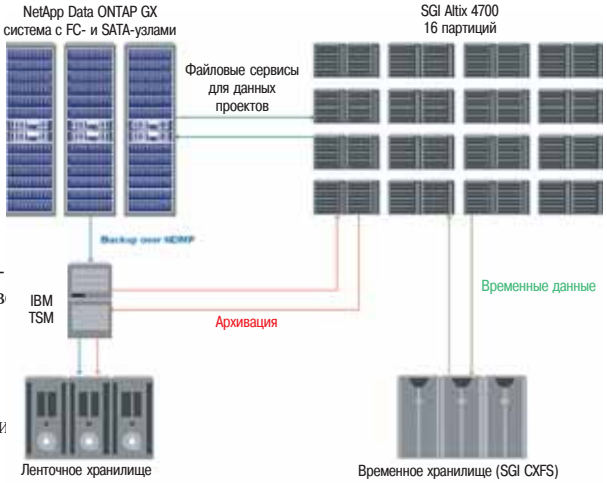

Рис. 8. Конфигурация Linux-кластера – HLRB II – установленного в исследовательском центре Leibniz Rechenzentrum (Германия) в 2007 г. с 6 узлами ONTAP GX Storage Grid системы (62 Тфлоп/с).

Рис. 9. Общий вид кластера HLRB II *(установлен в 2007 г. в исследовательском центре Leibniz-Rechenzentrum (Германия) – 15 строка top500).*

(другой из GX-узлов) – на базе FAS3040 – с лучшим соотношением $/Гбайт. Все манипуляции с томами между уровнями могут совершаться в онлайновом режиме без останова приложений.

Примеры построения ONTAP GX систем

Одна из последних инсталляций ONTAP GX Storage Grid была имплементирована в текущем году в исследовательском центре Leibniz-Rechenzentrum (Германия). Linux-кластер – HLRB II – был реализован на серверах Altix 4700 1.6 GHz с общим числом процессоров (Intel IA-64 Itanium 2 1600 MHz) – 9728 и показал пиковую производительность более 62 Тфлоп/с (рис. 8, 9).

В этой системе кластер, занимающий 15 строку в top500 *(ноябрь 2007 г.)*, поддерживается 6-узловой ONTAP GX системой, часть узлов которой укомплектовывалась SATA-дисками (только для зеркалирования). При этом, помимо SGI-кластера, в состав системы входит хранилище для временных данных – SGI CXFS и ленточное хранилище под управлением IBM TSM.

Заключение

Системы под управлением ОС ONTAP GX – новый класс NetApp-решений, ориентированный для использования в составе высокопроизводительных параллельных систем (кластеры и SMP-системы), а также медиарешений. Они могут поддерживать работу HPC-кластеров из нескольких тысяч узлов в качестве систем хранения обрабатываемых данных первого или второго уровня. Для развиваемых HPC-систем с требованиями высокой масштабируемости системы ONTAP GX Storage Grid могут быть одними из наиболее соответствующих для этого класса применений.