Кластерные NAS-хранилища для неструктурированного контента

В статье представлена концепция высокопроизводительных масштабируемых кластерных файловых хранилищ (развиваемых компанией Isilon Systems с 2001 г.), позволяющих, на базе последней из анонсированных моделей – Isilon IQ 12000, в частности, поддерживать 96-узловую систему с агрегированной пропускной способностью до 10 Гбайт/с, единым файловым пространством и емкостью до 1152 Тбайт.

Введение

Компания Isilon Systems основана сравнительно недавно – в 2001 г. Основная миссия компании – продвижение на рынке собственных решений: высокопроизводительных масштабируемых кластерных файловых хранилищ. Решения Isilon строятся на базе стандартных 2U-серверов с собственным дисковым пулом – до 12 HDD. При этом все они объединяются высокоскоростной связью – 2 порта x 1 GE и опционно еще дополнительным Infiniband-соединением (по 2 порта на узел). Серверам приложений весь многоузловой массив (до 96 узлов) представляется единым файловым пространством с высоким уровнем надежности, масштабируемости, пропускной способности и управляемости.

Сама идея использования стандартных серверов для построения систем хранения продвигается еще с конца 90-х. Однако активное использование ее для разработки высокопроизводительных NAS-хранилищ получило лишь в последние несколько лет, что в целом поддерживалось и тенденциями рынка. Несмотря на уже имеющийся ряд подобных систем на рынке, все они позиционируются по-разному, имеют значительно отличающиеся функциональность и методы управления данными, а также ценовую и лицензионную политику. Это позволяет им иметь в значительной степени собственные сектора рынка.

Тенденции рынка

Тенденции рынка последних лет свидетельствует о резком увеличении доли неструктурированных данных в общем объеме, к которым можно отнести: цифровую графику/видео/аудио, компьютерные модели, результаты компьютерного моделирования, справочную информацию и др.

Так, один из лидеров мирового авиадвигателестроения – компания Pratt & Whitney – проводит большое количество тестов газотурбинных двигателей, используя компьютерное моделирование. В результате при каждом моделировании записывается более чем 100 000 тестовых результатов в секунду, занимающие многотерабайтные емкости хранения. Другой пример. Исследовательский центр рака – Cedars-Sinai – в Лос-Анджелесе (США) только при анализе одной капли крови получает более 60 Гбайт неструктурированных данных для дальнейших исследований. При обслуживании сотен и тысяч пациентов объем данных при комплексном исследовании возрастает до петабайт. При проведении в 2004 г. *(г. Афины, Греция)* летних Олимпийских игр ежедневно генерировалось более чем 250 000 цифровых изображений (со средним размером 18–24 Мбайт) в течение 17 дней.

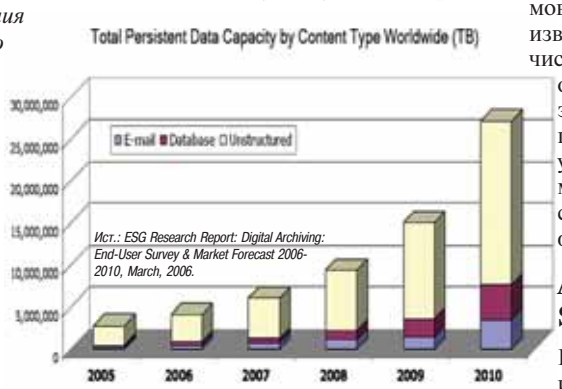

Рис. 1. По исследованиям ESG, общий объем постоянных данных к 2010 г. может вырасти до 26 000 Пбайт.

Исследования, проведенные ESG, показали, что общий объем постоянных данных к 2010 г. составит более 26 000 Пбайт (рис. 1) при соотношении постоянных данных к "динамическим" как 10:1. Однако уже в ближайшей перспективе это соотношение может вырасти до 100:1.

Общей особенностью этих примеров является то, что, помимо необходимости хранения больших генерируемых объемов данных, требуются еще и высокопроизводительные каналы доступа (в том числе и коллективного пользования на основе Web 2.0) по записи/чтению к этим данным. Также для отмеченных применений необходим определенный уровень простоты управления всем массивом данных и соответствующая стоимость за единицу хранения информации.

Архитектура решений Isilon Systems

Решения Isilon ориентированы для использования в составе вышеназванных применений и в настоящее время поставляется уже четвертое по-

коление NAS-систем хранения Isilon Systems, имеющих полностью распределенную кластерную архитектуру.

Все семейство решений строится на интеллектуальной распределенной файловой системе и модульных стандартных аппаратных компонентах, обеспечивающих простоту и высокую масштабируемость решений. Серверы Isilon IQ предназначены для хранения и использования неструктурированных данных/контента в секторах рынка, требующих интенсивного обмена данными, как: медиарешения; 3D-приложения; биология; микробиология; приложения для задач нефтегазовой отрасли, приложений для госструктур и др.

Основой кластерного решения Isilon является патентованная распределенная файловая система – OneFS. Она объединяет три уровня традиционной архитектуры хранения – файловую систему, менеджера томов и принципы RAID – в один программный уровень, создавая единую интеллектуальную полностью симметричную файловую систему, которая охватывает все узлы в пределах кластера. OneFS обеспечивает единую точку управления для большого контентного хранилища, с быстрым доступом к большим файлам, а также высокую доступность, возможность простого масштабирования емкости кластера до 1 Пбайт, поддержку до 10 Гбайт/с полной пропускной способности кластера.

OneFS уникально распределяет файлы и метаданные по множеству узлов в пределах кластера, улучшая традиционные методы стрипования (striping) контента на множестве дисков внутри системы хранения или тома.

OneFS обеспечивает каждый узел знанием о полном размещении файловой системы, а также о том, где постоянно располагается каждый файл и его части. Доступ к любому независимому узлу дает возможность клиенту/приложению обратиться ко всему контенту в объединенном файловом пространстве имен, что означает: отсутствие каких-либо томов/разделов и их ограничений по размеру; необходимость управления множеством сетевых устройств/томов.

Симметричная архитектура

Каждый Isilon IQ кластер имеет от 3 до 96 узлов Isilon IQ узлов, а каждый Isilon IQ узел – дисковую память, CPU, ОП и сетевые устройства – все в отдельной, компактной 2U-системе. При добавлении Isilon IQ узлов к кластеру все показатели симметрично масштабируются, включая емкость, производительность, ОП, CPU и сетевую связность. Isilon IQ узлы автоматически работают вместе, используя общие ресурсы для поддержания доступа к данным в единой объединенной системе хранения, которая является отказоустойчивой к любой части аппаратных средств, включая диски, коммутаторы и даже полные узлы.

В полностью распределенной архитектуре каждый узел должен быть синхронизирован со всеми другими узлами кластера. Isilon IQ узлы используют или Gigabit Ethernet, или высокоскоростную

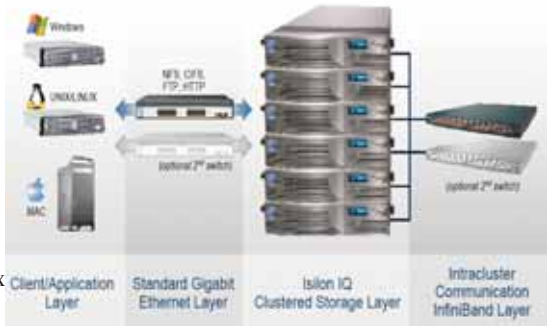

Рис. 2. Основные компоненты Isilon IQ кластера.

с низкими задержками Infiniband-фабрику для межузловой связи, синхронизации и всех внутрикластерных операций (рис. 2). Это дает возможность каждому узлу совместно использовать информацию с каждым другим узлом в системе так, чтобы каждый узел действовал как полностью равный по положению, с законченным пониманием того, что делают другие узлы.

OneFS поддерживает синхронизацию узлов, используя распределенного менеджера блокировки, когерентное кэширование и удаленного блок-менеджера, который поддерживает глобальную когерентность по всему кластеру. Именно эта глобальная когерентность, охватывающая каждый узел, устраняет какую-либо единую точку отказа для доступа к файловой системе. Любой узел в кластере может взять запрос на запись или чтение и каждый узел имеет один и тот же вид полной файловой системы. Все узлы в кластере равны по положению, так что система является полностью симметричной, устраняя иерархию и узкие места.

Внутренняя высокая доступность

Традиционные файловые системы используют отношения "главный/подчиненный", чтобы управлять множественными ресурсами хранения. Такие отношения имеют встроенные зависимости и создают точки отказов внутри системы хранения. Единственный правильный способ гарантировать целостность данных и устранить единую точку отказа состоит в том, чтобы сделать все узлы в кластере равными. Поскольку любой Isilon IQ узел равен по положению другому, то может обработать запрос от любого прикладного сервера. Если какой-либо узел "падает", любой другой узел выполняет его запрос, устраняя любую единую точку отказа.

Защита от множественных отказов

Функциональность кластера Isilon IQ дает возможность сохранять полную его доступность к данным при множественных отказах дисков или узлов. Это поддерживается опцией FlexProtect, использующей коды исправления ошибок Reed Solomon "n+1 и n+2" ECC (error correction code), а также четность стрипинга и четность стрипинга зеркалированного файла (от 2-х до 4-х раз). Политики с этой опцией могут быть установлены на любом уровне, включая кластер, директорию, поддиректорию и даже на индивидуальном файловом уровне. При необходимости политики могут быть изменены/заменены в любое время через WebUI – даже, во время работы системы. Все файлы в кластере стрипуются, что гарантирует отсутствие размещения какого-либо файла на 100% на одном узле. А это, в свою очередь, – его восстанавливаемость в случае сбоев/отказов.

Исследования показали, что среднее время между отказами (mean time between failures – MTBF) в алгоритме n+2 RAID в 100 раз больше, чем при использовании алгоритма с одиночной четностью.

Высокая скорость перестраивания кластера

В случае отказа, OneFS автоматически перестраивает файлы через все доступное свободное пространство в кластере параллельно, устраняя потребность иметь выделенные "диски четности", обычно требующиеся при традиционной файловой архитектуре хранения. За счет возможности использования всего множества процессоров в кластере данные могут быть восстановлены в 5–10 раз быстрее по сравнению с традиционной архитектурой.

Время, которое требуется системе хранения для восстановления данных в случае отказа дисковода, является критическим для надежности этой системы хранения. В традиционных системах хранения, процесс восстановления может требовать многих часов. С увеличением емкости дисков время восстановления, как правило, только возрастает.

В Isilon системах восстановление от отказов дисков происходит намного быст-

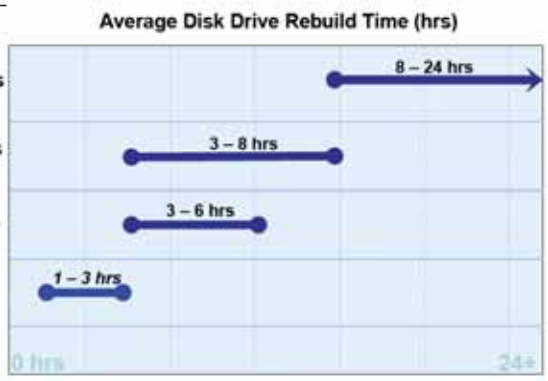

Рис. 3. Сравнение времени кластера Isilon IQ *(по результатам тестирования Isilon, прим. ред.)* с традиционными системами, использующие разные типы дисков. (В тестировании Isilon кластера использовался HDD Maxtor 250GB Serial ATA; диск был на 87% полным).

рее. В зависимости от плотности диска, дисковые отказы в пределах Isilon-кластера могут быть восстановлены в пределах от 1 до 3 часов. В сравнении с традиционными системами (рис. 3) это время значительно меньше.

Самовосстанавливаемость

OneFS постоянно контролирует состояние всех файлов и дисков и анализирует записи smart-статистики (например, восстанавливаемые ошибки чтения), доступные на каждом диске, чтобы преду-

преждать возможные их отказы. Когда OneFS идентифицирует опасные компоненты, она перемещает данные с "опасного" диска на свободное место в кластере автоматически и прозрачно для клиентов. Как только данные перестроены, администратор получает уведомление о возможном отказе подозреваемого диска.

Один уровень управления

Isilon IQ создает единый разделяемый пул всего контента внутри кластера, обеспечивая одну точку доступа для пользователей и одну точку управления для администраторов. Протестированный объем пула составляет более 1,6 Пбайт.

Одно из ключевых преимуществ OneFS – простота увеличения производительности и емкости кластера. Системный администратор просто вставляет новый Isilon IQ узел, соединяет сетевые кабели и включает его. Кластер автоматически обнаруживает добавленный узел и проводит его конфигурирование (на это уходит менее 60 секунд), а далее осуществляет полную перебалансировку данных с учетом добавленного узла.

Isilon IQ кластер позволяет добавлять к существующему узлы с другими дисководами (по емкости или типу) также, например, с большим количеством портов Gigabit Ethernet для более высокой производительности.

Линейная масштабируемость производительности

Каждый добавляемый Isilon IQ узел (по результатам тестовых испытаний Isilon, *прим. ред.*) увеличивает агрегированную производительность кластера примерно на 700 Мбайт/с (рис. 4).

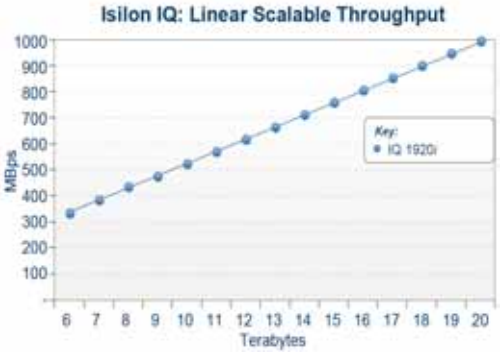

Рис. 4. Каждый добавляемый Isilon IQ узел *(по результатам тестовых испытаний Isilon, прим. ред.)* увеличивает агрегированную производительность кластера на 700 Мбайт/с.

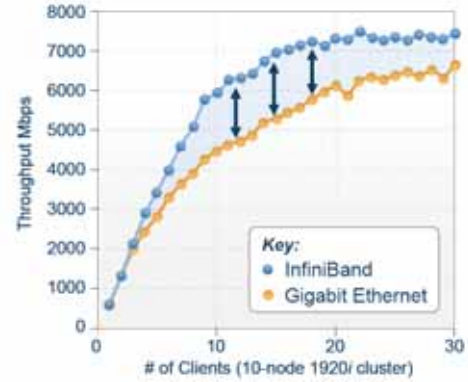

Рис. 5. Внутренний интерконнект на основе Infiniband-технологии позволяет получить более высокую производительность, чем при использовании GigE-коммутаторов.

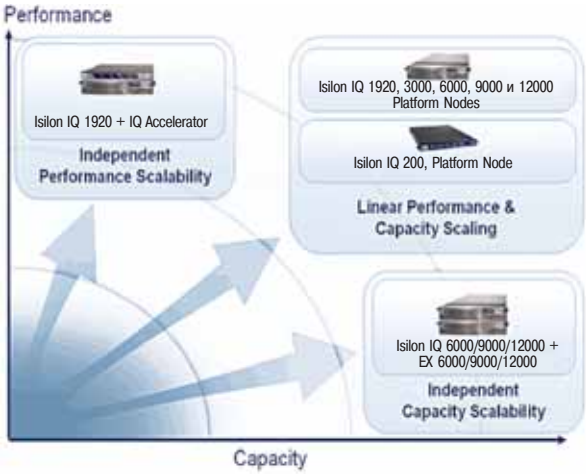

Рис. 6. Наличие ряда моделей в семействе решений Isilon позволяет варьировать производительностью и емкостью.

Линейная масштабируемость Isilon IQ также является следствием использования Infiniband в качестве внутреннего интерконнекта, который позволяет кластеру иметь почти нулевые задержки при синхронизации. Фактически, тестирование Isilon показало, что Infiniband-технология позволяет получить намного более высокую производительность, чем при использовании GigE-коммутатора для внутреннего интерконнекта (рис. 5).

В составе семейства продуктов Isilon 5 моделей узлов (Isilon IQ 200, 1920, 3000, 6000, 9000 и 12000) – один ускоритель производительности (Isilon IQ Accelerator) и три расширителя емкости узлов: Isilon EX 6000, EX 9000, WX 12000, что позволяет достаточно гибко варьировать производительностью и емкостью кластера (рис. 6).

Дополнительное ПО для поддержания высокой доступности

SnapshotIQ ПО расширяет возможности базовой функциональности по поддержанию высокой доступности данных на основе мгновенных снимков. Поскольку томов в Isilon кластере не существует, мгновенные копии производятся для конкретных данных в задаваемые моменты времени.

SmartConnect ПО обеспечивает интеллектуальную балансировку нагрузки соединений клиентов и предотвращает NFS-отказы. Через задаваемые политики SmartConnect упрощает управление и максимизирует управление для большого числа клиентов, использующих кластер. Дополнительно SmartConnect поддерживает динамическую NFS-отказоустойчивость для Linux- и UNIX-клиентов, гарантируя, что даже в случае отказа узла, все незаконченные операции по чтению/записи будут переданы другому узлу в кластере и завершатся без прерывания пользователя/приложения.

SyncIQ ПО – асинхронная репликация для поддержания катастрофоустойчивости (Disk-to-Disk Backup & Restore). На основе набора политик оно обеспечивает LAN- или WAN-репликацию для разных типов файлов в зависимости от их размеров, доступной полосы пропускания и ресурсов хранения.

MigrationIQ – обеспечивает автоматическую и простую миграцию данных между уровнями хранения.

Aspera Enterprise Server For Isilon IQ – осуществляет высокопроизводительную доставку больших файлов и контента по WAN-сетям.

Высокопроизводительные географически распределенные архитектуры Isilon-кластеров строятся на базе ПО Aspera Enterprise Server, интегрированного с технологией Riverbed RiOS Transport Streamlining (технология т.н. WAN-акселераторов, начинающая получать распространение и в России, особенно в последний год).

Основным протоколом передачи, а значит и протоколом синхронизации межкластерных соединений является транспортный протокол TCP. Сам по себе TCP достаточно избыточен и неоптимален. TCP необходимо оптимизировать, чтобы уменьшить количество TCP-пакетов, необходимых для передачи данных. При этом технологии должны оптимизировать соединения с высокой пропускной способностью.

Внедрение системы оптимизации Riverbed SteelhedT в межкластерные связи Isilon исключает ограничения TCP-протокола, оптимальным образом и динамически меняя такие его характеристики как: размер окна передачи, процедуру обработки ошибок, процесс уведомления о перегрузке и многие другие. Технология Riverbed RiOS Transport Streamlining также значительно повышает коэффициент использования соединений с высокой пропускной способностью High Speed TCP и MX-TCP для каналов с высоким коэффициентом потерь.

RiOS разработана таким образом, что параметры соединения меняются на лету в зависимости от возникающих событий, таких как потеря пакета или перегрузка, при этом все характеристики, делающие протокол TCP надежным транспортным протоколом, остаются в силе. По заявлениям Isilon, все распределенные файловые сервисы максимально эффективны даже на каналах ниже 45 Мбит/с.

Заключение

Решения Isilon представляют собой новый класс высокопроизводительных масштабируемых NAS-решений для хранения неструктурированного контента с бюджетными показателями стоимости за единицу хранения, который уже в ближайшей перспективе может занять свою нишу и в России для многих развивающихся отраслей промышленности, науки, а также в медиаиндустрии.