EMC Isilon: горизонтально-масштабируемая NAS-система

В конце 2010 года в состав корпорации EMC вошла компания Isilon, являющаяся лидером в производстве горизонтально-масштабируемых NAS-систем. Это приобретение позволило EMC выйти на российский рынок с уникальными технологиями, востребованными для хранения "больших данных".

Евгений Красиков – консультат по технологиям, EMC Россия и СНГ.

Введение

Рост объемов обрабатываемых и хранимых данных наблюдается во всех областях ИТ, вместе с которым увеличивается и сложность традиционных систем (как правило, двухконтроллерных). Такая архитектура разработана более 20 лет назад и пасует перед "большими данными", не справляясь с их эффективным хранением и управлением ими. "Большие данные" требуют нового подхода, и этот подход предлагает EMC Isilon – система, изначально разработанная для работы с "большими данными".

Принципиальное отличие EMC Isilon от традиционных решений – кластерная архитектура. СХД не содержит контроллеров и дисковых полок, а представляет собой набор равнозначных узлов, объединенных с помощью выделенной дублированной сети Infiniband. Каждый узел содержит диски, процессоры, память и сетевые интерфейсы для клиентского доступа. Вся дисковая емкость кластера формирует единый пул хранения и единую файловую систему (ФС), доступ к которой возможен через любой из узлов.

Предложенная архитектура обеспечивает практически безграничное масштабирование (до 144 узлов и 15 Пбайт в единой ФС), простоту использования независимо от объема, линейную масштабируемость по емкости и производительности, а также высокую эффективность управления.

"Большие данные"

По данным исследований IDC, суммарный объем СХД в 2014 г. приблизится к 80 экзабайт, при этом около 80% хранилища – неструктурированные файловые данные. Рост объемов данных заметен везде: видео в интернете; социальные сети с гигантскими объемами загружаемой пользователями информации; новые медиаформаты; видео высокого разрешения и 3D-формата; научные исследования, связанные с расшифровкой генома, компьютерным моделированием сложнейших процессов и др.

При использовании традиционных СХД с ростом объемов данных неизбежно возрастает число контроллеров, дисковых групп, томов, файловых систем, так или иначе появляются "островки" разнородных систем, становится все сложнее поддерживать эффективность их использования. Масштабирование систем требует предварительного анализа, сложных работ по конфигурации и миграции данных.

Для хранения "больших данных" необходимо эффективное и простое решение, которое позволило бы не бороться с их растущими объемами, а извлекать из них выгоду.

Scale-Out

Решить проблемы традиционных систем, ответить на вызовы "больших данных" призвана кластерная архитектура с горизонтальной масштабируемостью, на основе которой строятся вычислительные кластеры, веб-сервисы и др.

Isilon предлагает такой подход для СХД, который обеспечивает ряд преимуществ в масштабируемости, производительности, защите данных и простоте и эффективности системы. В Isilon нет привычных контроллерных пар, подключаемых к ним дисковых полок, RAID-групп, томов и файловых систем, обслуживаемых одним из контроллеров. Вместо этого СХД представляет собой кластер однотипных узлов, объединенных с помощью выделенной сети Infiniband. Каждый узел содержит и диски, и вычислительные ресурсы.

Дисковая емкость всего кластера формирует единый пул и единую файловую систему. Данные автоматически равномерно распределяются между узлами. Память всех узлов объединена в общий когерентный кэш. Доступ к файловой системе возможен через любой из узлов, все узлы кластера равнозначны, отсутствуют какие бы то ни было выделенные серверы управления, контроллеры метаданных или файловые шлюзы.

Масштабируемость

Благодаря горизонтально-масштабируемой кластерной архитектуре, Isilon наращивается простым добавлением новых узлов.

Для увеличения емкости достаточно физически подключить новый узел к кластеру и отдать команду на его добавление. Дальше все происходит автоматически: настройки берутся из существующего кластера, моментально увеличивается доступный объем файловой системы, начинается миграция данных на новый узел. Благодаря тому, что каждый узел содержит и диски, и процессоры, и кэш, добавление нового узла сразу линейно увеличивает и емкость, и производительность кластера.

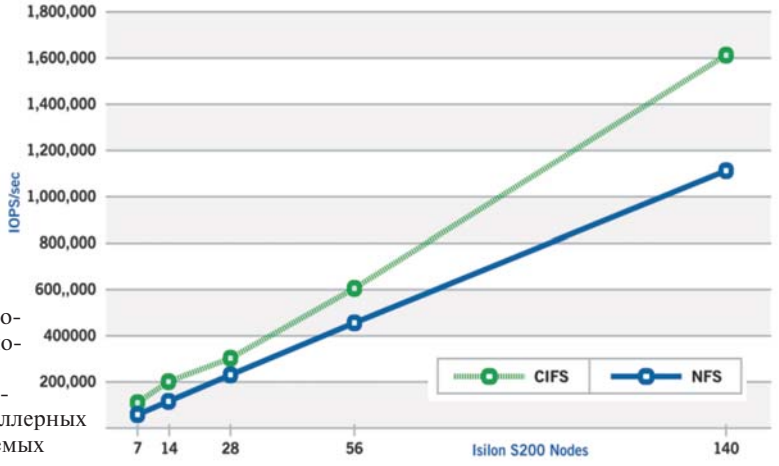

Рис. 1. Линейный рост производительности модели Isilon S200 при добавлении узлов на тестах SPECsfs 2008.

Производительность

Зачастую scale-out решения на самом деле представляют собой объединенные в общее пространство имен отдельные классически дисковые массивы. При этом производительность доступа к отдельному сегменту, файлу или каталогу ограничена его дисковой группой и одним контроллером.

В Isilon все данные равномерно распределены между всеми узлами, и любой узел может на равных обслуживать запросы к любому файлу. Таким образом достигается высокая производительность работы с любым отдельным файлом. Кроме того, все ресурсы объединены в единый пул, архитектура изначально параллельная.

Линейный рост производительности при добавлении узлов подтверждается, например, тестами SPECsfs 2008.

Высокая доступность

В традиционных двухконтроллерных системах отказ одного контроллера означает потерю 50% производительности. В случае отказа обоих контроллеров – недоступность данных.

В Isilon же при потере, например, одного из 10 узлов, теряется 10% производительности. В зависимости от сконфигурированных политик, Isilon может обеспечить защиту данных при выходе из строя одновременно до 4 узлов.

При выходе диска из строя в ребилде участвуют все диски и узлы кластера, а не только диски одной RAID-группы и hot-spare диск, что позволяет значительно сократить время ребилда.

Простота

С ростом объема растет лишь число узлов в кластере. Сохраняются единая точка администрирования и единая файловая система. Таким образом, управление 10 Пбайт и управление 100 Тбайт мало отличаются между собой.

Эффективность

В традиционных системах нужно следить за загрузкой контроллеров, портов, дисковых групп, что становится практически невозможным с ростом объемов и числа систем. Единственный приемлемый подход для “больших данных” – равномерно распределить данные и нагрузку. Это позволяет достичь высокой эффективности как в смысле загрузки всех компонентов, так и в смысле полезной емкости.

Говоря о коэффициенте полезной емкости, часто учитывают только накладные расходы на RAID и ФС, забывая о второй составляющей – общей эффективности. Простой пример – сколько файлов по 11 Тбайт можно хранить на 5 файловых системах по 20 Тбайт? Пять, хотя суммарного объема достаточно и для девяти! Коэффициент полезной емкости, не связанный с защитой данных, составляет в этом случае 55%. Пример примитивный, но иллюстрирует проблему.

Isilon же, благодаря своей схеме защиты данных и единой ФС, может обеспечить реальный коэффициент полезной емкости более 80%.

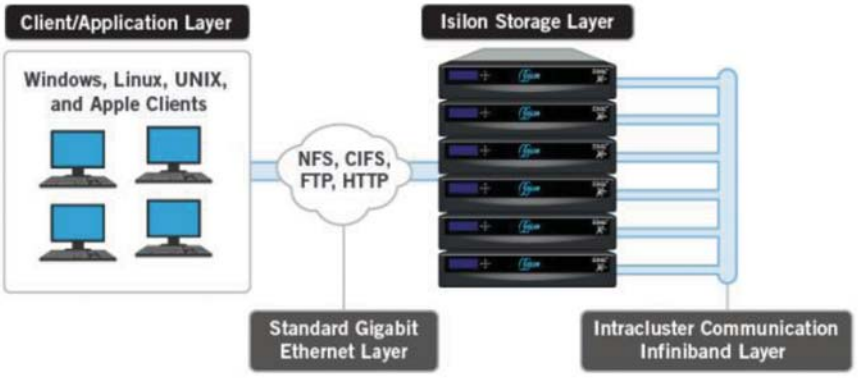

Рис. 2. Архитектура Isilon.

Архитектура Isilon

Кластер

Итак, кластер Isilon строится из узлов стандартной архитектуры. Каждый узел содержит диски, процессоры Intel Xeon, память, NVRAM, Ethernet порты 1 Гбит/с или 10 Гбит/с и два порта Infiniband. Между собой узлы объединены с помощью выделенной сети Infiniband, для отказоустойчивости входящие в состав кластера QDR Infiniband-коммутаторы дублированы (рис. 2).

Кластер может содержать от 3 до 144 узлов.

Isilon – это система хранения с файловым доступом (NAS). Поддерживаются стандартные файловые протоколы SMB (v1, v2), NFS (v3, v4), FTP и HTTP. Таким образом, для работы с Isilon не требуется установка дополнительного ПО на клиентские компьютеры, реализация кластерной ФС скрыта от клиентов, которые просто работают с кластером по привычным файловым протоколам. Также есть поддержка iSCSI и интерфейса кластерной файловой системы Hadoop HDFS.

Все узлы формируют единый пул ресурсов и единую файловую систему. Все узлы равноправны, любой из узлов может обработать любой запрос без дополнительных накладных расходов.

Как выглядит поток данных в Isilon? Клиент устанавливает соединение с одним из узлов кластера и передает файл для записи. Узел, обслуживающий клиентскую сессию, разбивает полученный файл на сегменты, добавляет к ним сегменты четности в соответствии с выбранной политикой защиты, сохраняет часть данных на своих дисках, а также по внутренней сети Infiniband равномерно распределяет данные между остальными узлами кластера. Аналогично, при чтении узел, получивший запрос, имея доступ к единой кластерной ФС OneFS, “знает”, где находятся требуемые сегменты файла, получает их по сети Infiniband, “собирает” их и отдает клиенту.

Типы узлов

Основа кластера – операционная среда OneFS, она одинакова для всех узлов Isilon, а аппаратную часть можно выбрать сообразно задаче. Модели отличаются числом и типом используемых дисков, процессорами, объемом памяти и поддержкой SSD.

В настоящий момент в линейке имеются три типа узлов хранения: S200 – для самых требовательных к производительности и случайных нагрузок, X200 –для потоковых нагрузок с хорошим балансом между емкостью и производительностью, и NL – с самой высокой эффективностью хранения и выгодной стоимостью за терабайт. Характеристики систем представлены в табл. 1.

Узлы разных типов можно объединять в одной ФС. С помощью функционала SmartPools (см. ниже) можно организовать прозрачную миграцию данных между уровнями хранения. Например, перемещать файлы, к которым не было обращений в течение недели, на узлы NL-серии, а небольшие mov-файлы помещать на узлы S-серии.

Файловая система

Одна из главных отличительных особенностей Isilon – единая ФС. Даже операционная система называется OneFS. Весь объем кластера доступен в одной файло-

Табл. 1. Характеристики различных моделей Isilon.

Модель	S200	X200	36NL, 72NL, 108NL
Тип дисков	2.5” SAS	3.5” SATA	3.5” SATA
Объем дисков	300 ГБ или 600 ГБ	500 ГБ, 1 ТБ, 2 ТБ или 3 ТБ	1ТБ (36NL), 2 ТБ (72NL), 3ТБ (108NL)
Число дисков	24	12	36
Поддержка SSD	0, 1, 2 или 6 SSD по 200 ГБ	0, 1, 2 или 6 SSD по 200 ГБ	Нет
Объем	От 7.1 ТБ до 14.4 ТБ	От 4.2 ТБ до 36 ТБ	36, 72 или 108ТБ
Объем памяти ECC	24, 48 или 96 ГБ	6, 12, 24 или 48 ГБ	4 ГБ (36NL, 72NL) или 16 ГБ (108NL)
Объем NVRAM	512 МБ	512 МБ	512 МБ
Процессоры	Два четырехядерных Intel Xeon Westmere	Один четырехядерный Intel Xeon Nehalem	Один (36NL, 72NL) или два (108NL) четырехядерных процессора Intel Xeon
Сетевые порты	4 x 1 GbE	4 x 1 GbE	2 (36 NL, 72NL) либо 4 (108NL) x 1 GbE
	или	или	или
	2 x 1 GbE и 2 x 10 GbE SFP+	2 x 1 GbE и 2 x 10 GbE SFP+	2 x 1 GbE и 2 x 10 GbE SFP+
Занимаемое место	2 RU	2 RU	4 RU
Типичное энергопотребление	456 Вт	408 Вт	720 Вт

вой системе /ifs, в которой можно создавать каталоги и настраивать общие ресурсы. При добавлении новых узлов просто растет объем /ifs.

При этом OneFS обеспечивает не агрегированное пространство имен, которое лишь создает видимость единого пула, но каждый файл или каталог по-прежнему привязан к своему сегменту или ФС. В OneFS каждый файл может располагаться на всех узлах кластера.

Это действительно единая ФС, которая распределена по множеству дисков разных узлов. Блоки данных адресуются с помощью "троек" вида (номер узла, номер диска, номер блока на диске). Метаданные и задачи управления блокировками распределены между всеми узлами. Архитектура файловой системы симметричная, нет выделенных контроллеров метаданных, хранящих информацию о расположении данных в ФС или управляющих блокировками, все узлы в кластере равноправны, каждый узел имеет одинаковый доступ к кластерной файловой системе.

Защита данных

В Isilon не используется привычная технология RAID. В узлах нет аппаратных RAID-контроллеров, все диски видны ОС как отдельные устройства. А данные и метаданные всегда с избыточностью распределяются между отдельными дисками разных узлов.

Для метаданных используется зеркалирование, а сами данные защищаются с использованием кода Рида-Соломона (так называемая защита N+M) или зеркалирования. Каждый файл защищается независимо, политику защиты можно настраивать не только для всего кластера, но и на уровне отдельных файлов и каталогов.

Каждый файл записывается на диски в виде набора страйпов или групп защиты (protection group). Каждая группа защиты располагается на дисках разных узлов в соответствии с выбранной политикой. Уровни защиты с использованием четности (кода Рида-Соломона) обозначаются N+M, где N и M обозначают, соответственно, число блоков данных и четности внутри каждой группы защиты. При этом значение M определяет число дисков или узлов целиком, которые могут выйти из строя без потери данных. Например, политика N+2 защищает данные от отказа одновременно двух любых дисков кластера, или узла целиком и любого диска, или любых двух узлов.

На рис. 3 показана защита данных в кластере из шести узлов при использовании политики N+2. Такая схема обеспечивает коэффициент полезной емкости 66%.

OneFS поддерживает уровни защиты до N+4, защищая данные от выхода из строя одновременно любых четырех дисков или узлов, что значительно превосходит общепринятый RAID6.

В ситуациях, когда требуется защита от выхода из строя двух дисков или одного узла одновременно, OneFS предлагает политику защиты N+2:1. Она особенно актуальна для небольших кластеров, так как обеспечивает лучший коэффициент полезной емкости по сравнению с N+2 и лучшую защиту по сравнению с N+1. Именно такая политика используется по умолчанию.

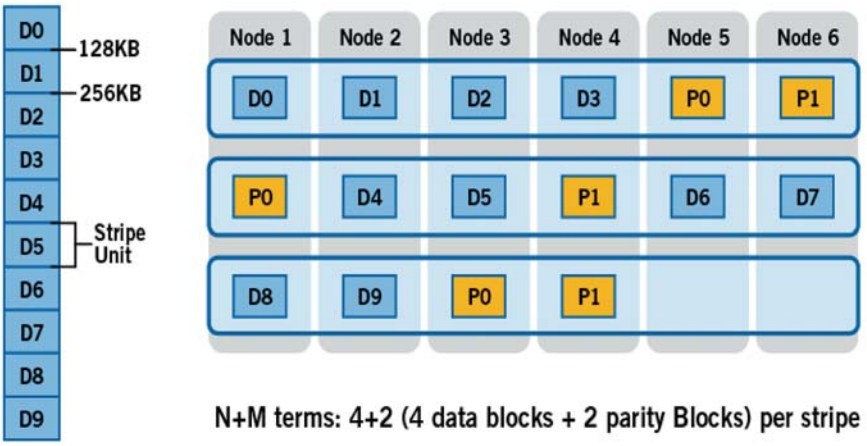

Рис. 3. Защита данных в кластере Isilon из шести узлов при использовании политики N+2.

SSD

Модели X200 и S200 поддерживают SSD. Диски SSD в Isilon могут использоваться двумя способами.

Во-первых, для данных. С помощью политик функционала SmartPools действительно можно указать, какие файлы будут храниться на SSD (например, все *.jpg, все созданные менее суток назад и так далее).

Во-вторых, интереснее другая возможность – использование SSD для метаданных.

При файловом доступе значительную часть операций составляют именно операции с метаданными, их доля часто превышает 50%. Например, прежде чем начать читать блоки файла с сетевого диска, необходимо "пробежаться" по дереву каталогов, получить атрибуты файла, получить список его физических блоков. Все это требует отдельных операций и добавляет задержки.

В Isilon и без SSD метаданные обрабатываются быстро: они равномерно распределены между всеми узлами и эффективно кэшируются на разных уровнях.

Но, кроме того, у нас есть уникальная возможность поместить на SSD только метаданные (это области инодов, карты блоков, сами каталоги и различные системные структуры). В таком случае множество небольших обращений к метаданным моментально обслуживается из SSD, а не теряет время вместе с большими запросами к данным в очереди к медленным дискам.

Таким образом, можно добавить в кластер всего 1–2% емкости на SSD (например, по одному SSD в каждый узел) и получить значительный прирост производительности файловых операций (до 30% на близких к случайным нагрузкам при вдвое меньшем времени отклика).

В первую очередь, SSD для метаданных будут полезны при большом числе небольших файлов, при преобладающем случайном чтении и при любых нагрузках.

Функциональность Isilon

EMC Isilon поддерживает набор дополнительный возможностей, привычных в корпоративных системах хранения данных: уровневое хранение, квоты, репликацию, мгновенные снимки и др.

В последнее время в портфеле решений Isilon произошел ряд существенных изменений: обновилась линейка оборудования, появились новые типы узлов и поддержка SSD, значительные изменения претерпела операционная среда OneFS, расширилась существующая и добавилась новая функциональность. Появились уровневое хранение, позволяющее комбинировать диски и узлы разных типов в одной файловой системе, улучшенные репликация, снэпшоты и квоты, WORM, развитые возможности анализа производительности и использования системы, улучшилась интеграция с VMware vSphere. Все это позволило не только усилить позиции в традиционных для Isilon сферах (медиа, интернет, медицина и т.д.) и показать выдающиеся результаты в тестах производительности SPECsfs 2008, но и начать рассматривать Isilon для "больших данных" в корпоративных ИТ в качестве: больших централизованных файловых хранилищ, больших платформ виртуализации второго и третьего уровня, облачных сред.

SmartPools

SmartPools позволяет объединить в одном кластере узлы разных типов и орга-

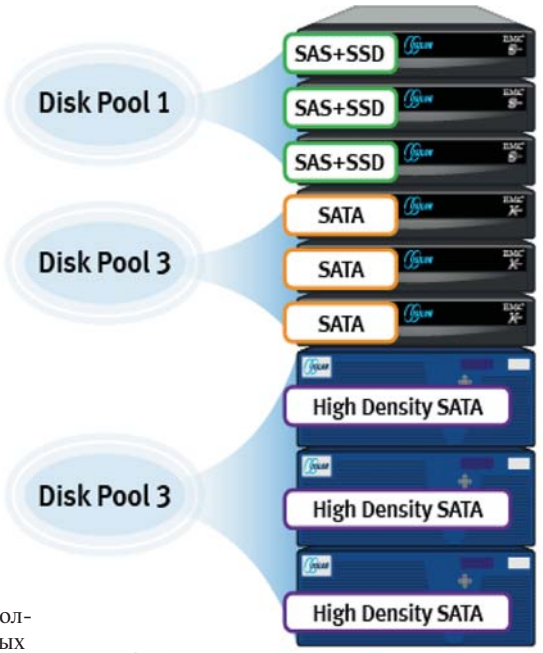

Рис. 4. SmartPools позволяет объединить в одном кластере Isilon узлы разных типов и организовать прозрачное уровневое хранение на основе политик

низовать прозрачное уровневое хранение на основе политик (рис. 4). При использовании SmartPools для узлов разных моделей создаются отдельные пулы, а размещением данных в них можно управлять с помощью политик.

Поддерживаются политики на основе имени файла или каталога (в том числе заданных с помощью шаблонов), типа файлов, размера файлов, дополнительных определяемых пользователем атрибутов, времени создания, доступа или изменения.

Миграция данных происходит по внутренней сети Infiniband прозрачно для клиентов.

Кроме пула для размещения данных, политики SmartPools могут определять уровень защиты файлов, использование SSD, оптимизацию доступа и ряд других параметров.

Даже при использовании нескольких пулов файловая система остается единой, доступ к любым данным по-прежнему возможен через любой узел. Политики SmartPools лишь определяют, диски каких узлов файловая система будет использовать для хранения тех или иных данных.

SmartConnect

Благодаря тому, что все узлы равноправны, для доступа к файловым ресурсам клиенты могут устанавливать соединения с любыми узлами, каждый раз самостоятельно выбирая узел. Но, конечно, это неудобно. Поэтому Isilon предлагает способ обеспечить единую точку входа – функционал SmartConnect. При его использовании клиенты подключаются по единому доменному имени, а встроенный в кластер балансировщик отправляет их к тому или иному узлу, в зависимости от загрузки процессоров, сетевых интерфейсов, числа соединений или просто поочередно. При работе по протоколу, не использующему сессии (NFS v3), этот же функционал позволяет организовать полностью прозрачную для клиентов обработку отказа узла путем переноса его IP-адресов на остальные узлы кластера.

SnapshotIQ

Поддерживаются мгновенные снимки на уровне каталогов. Не требуется выделение отдельной дисковой емкости для хранения данных снэпшотов, пространство выделяется динамически из общего пула. Поддерживается до 1024 снэпшотов для каждого каталога. Есть интеграция с Microsoft VSS.

SmartQuotas

EMC Isilon поддерживает и дисковые квоты. Можно настраивать квоты для пользователей, групп или каталогов. Есть возможность отображать пользователям либо реальное доступное пространство, либо размер их жесткой квоты. В случае задания квоты, превышающей емкость кластера, можно реализовать thin provisioning для NAS.

SmartLock

Для защиты файлов от случайного удаления предусмотрено ПО SmartLock. С его помощью можно определить отдельные каталоги как WORM (write once, read many), после чего располагающиеся в них файлы, имеющие установленный атрибут read only, не смогут удалить даже администраторы.

SyncIQ

Для защиты данных Isilon поддерживает асинхронную репликацию между двумя или более кластерами. Двумя ключевыми отличиями, например, от входящего в различные системы rsync являются работа на блочном уровне и параллелизация.

При использовании SyncIQ нет необходимости в обходе файловой системы при каждой сессии репликации. При большом числе файлов такой обход может оказать существенное влияние на продуктивные нагрузки и занять продолжительное время. В Isilon с помощью функционала снэпшотов на блочном уровне отслеживаются измененные с момента прошлой репликации блоки, и во время сессии репликации передаются только они.

Кроме того, как и все в Isilon, репликация многопоточная. Сразу несколько узлов могут одновременно передавать данные на вторую систему.

Поддерживаются различные топологии: один ко многим, двусторонняя репликация, и так далее.

InsightIQ

Аналитическая платформа InsightIQ позволяет собирать, хранить и анализировать информацию о производительности кластеров Isilon и об использовании файловой системы. С ее помощью можно, например, построить график распределения файлов по размеру, найти 100 самых больших или самых старых файлов, построить распределение подкаталогов каталога по размеру и так далее. Таким образом, можно понять, как используется файловая система. InsightIQ реализован в виде отдельной виртуальной машины, который развертывается в существующей виртуальной среде, что изолирует продуктив от мониторинга.

Позиционирование

Исторически первыми с таким ростом объемов данных, с которым перестали справляться традиционные системы, массово столкнулись медиакомпании. Уже в 2000-х года были ясно видны ограничения традиционных систем по масштабируемости, эффективности и защите данных. И исторически именно медиакомпании (телеканалы, кинокомпании, продакшн-студии, интернет-вещание) были основными потребителями Isilon. В мире медиа Isilon очень хорошо известен, есть множество заказчиков и референсов. Системы EMC Isilon подходят практически для любых задач медиакомпаний (получение контента, обработка, кратковременное хранение, нелинейный монтаж, хранение архивов) и позволяют консолидировать все задачи на единой центральной системе хранения.

С течением времени “большие данные” становились актуальными и для других компаний. Можно выделить следующие направления:

– нефтегазовая отрасль, где Isilon используется, например, для хранения данных сейсмических исследований и их интерпретации;
– промышленность и разработка (электроника, машиностроение, двигателестроение, авиапромышленность), где больших объемов требуют данные различных проектов, симуляций и испытаний;
– хранение и доставка контента во всевозможных интернет-проектах (video on demand, облачные сервисы);
– кластеры высокопроизводительных вычислений (HPC) в любых отраслях;
– геоинформационные системы;
– наука и образование;
– медицина и биологические исследования (медицинские изображения и секвенирование ДНК);
– архивы данных видеонаблюдения.

Однако в последнее время “большие данные” приходят и в корпоративные ИТ. Файловые хранилища крупных предприятий и среды виртуализации уже достигают объемов, при которых традиционные решения теряют эффективность, и преимущества Isilon выходят на первый план. Например, Isilon отлично подходит для консолидации файловых хранилищ предприятия. Благодаря интеграции с VMware vSphere 5, можно рассматривать Isilon для больших 2-tier/3-tier виртуальных сред.

Для аналитики больших данных EMC предлагает комплексное решение в составе Greenplum HD и Isilon, объединяющее в себе продвинутую кластерную файловую систему OneFS и вычислительный кластер Hadoop. В нем Isilon используется как уровень хранения для аналитического кластера Greenplum. Это позволяет достичь высокой эффективности хранения, использовать для данных Hadoop всю богатую функциональность OneFS, независимо масштабировать емкость хранения и вычислительную мощность кластера Greenplum HD, а также избежать продолжительной загрузки данных для аналитики.

На мировом рынке Scale-Out NAS EMC Isilon занимает лидирующее положение, более 2000 крупных заказчиков в различных отраслях используют продукты Isilon. Есть успешные внедрения и в России и СНГ. СХД EMC Isilon используются в российских телеканалах и интернет-компаниях, а также для хранения спутниковых изображений, в образовании и для архивов видеонаблюдения.

Заключение

NAS-системы EMC Isilon изначально созданы для эффективной работы с большими объемами данных. Использование кластерной архитектуры позволило создать, с одной стороны, очень простую, а с другой – очень производительную, линейномасштабируемую и отказоустойчивую систему, идеально подходящую для задач, требующих хранения больших объемов файловых данных.

Евгений Красиков,
EMC Россия и СНГ.