Программируемая логика для высокопроизводительных NAS-платформ

В продуктовой линейке HDS до декабря 2006 г. отсутствовал продукт для виртуализации хранения на уровне файлов. Глобальное 5-летнее OEM-соглашение между компаниями Hitachi Data Systems и BlueArc Corporation устранило этот пробел и дает возможность HDS уже сейчас предлагать высокопроизводительные NAS-решения на базе платформы BlueArc Titan, позволяющие осуществлять доступ к консолидированному объему данных на файловом уровне. HDS будет поставлять OEM-версию решения BlueArc под названием Hitachi High-performance NAS Platform, Powered by BlueArc, в котором системы хранения HDS будут выполнять роль физического хранилища данных.

Введение

Требования к файловым хранилищам (network attached storage – NAS) компаний постоянно возрастают. Это связано, прежде всего, со значительным увеличением числа конечных пользователей IT-инфраструктуры, а также количественным и качественным ростом запросов с их стороны; наращиванием производительности рабочих станций и серверов; ростом объемов обрабатываемых данных и распространением HPC-систем (High Performance Computing) для проведения расчетов; развитием инфраструктуры передачи данных и т.п. Это приводит к тому, что классические файловые серверы уже просто не способны справляться с возросшей нагрузкой.

Большая часть NAS-серверов строится на основе стандартных компонентов или, проще говоря, на базе стандартных серверных платформ. И, несмотря на то, что производительность NAS-серверов постоянно возрастает за счет использования современных CPU, шин/интерфейсов, высокопроизводительных дисков/средств хранения данных, в ряде случаев требования к файловым хранилищам могут значительно превосходить их возможности.

Один из наиболее простых способов решения этой проблемы – развертывание множества NAS-устройств в сетевой инфраструктуре компании. Однако это приводит к децентрализации данных и усложнению управления хранением. Частично эти недостатки можно устранить, объединяя файловые хранилища в один общий пул на основе средств виртуализации, однако проблемы производительности при этом остаются.

Многие пользователи ищут решение проблемы производительности в переходе к SAN-инфраструктурам. Но здесь появляются препятствия, которых не было в NAS-инфраструктурах. Одно из наиболее существенных – это более дорогая реализация SAN-инфраструктуры в сравнении с NAS, но основная – это то, что SAN не обеспечивает стандартно реализуемого разделяемого доступа к файлам, необходимого для простого управления данными.

Другой подход к решению выше отмеченных проблем – это переход к концепции кластеров хранения. В дальнейшем для этого будем использовать термин грид-хранение или storage-grids. Однако длительное время из-за ряда проблем, связанных с работой с файловыми системами, эта концепция не имела промышленных реализаций. К ним можно отнести: блокировку файлов при их совместном использовании, когерентность кэш-памяти, кэширование с клиентской стороны и много других аспектов совместного использования файловой системы требующих решения для реализации этой концепции.

Часть этих проблем для большинства применений разрешимы, но для приложений, связанных с вычислительными кластерами, требуются значительные усилия. Поэтому несколько лет назад IETF-комитет (The Internet Engineering Task Force) предложил новый стандарт, названный pNFS и призванный упростить развитие NAS-серверов для параллельных вычислений. pNFS представляет собой расширение NFS v4 и призван стандартизовать усилия промышленности в этой области. В архитектуре BlueArc полностью поддерживается этот новый стандарт и ряд других протоколов для ускорения потока данных.

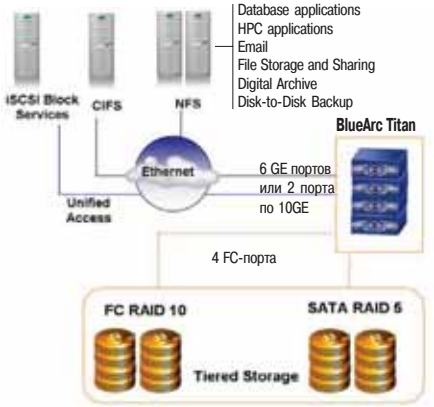

Рис. 1. Физически NAS-решение BlueArc представляет устройство управления потоком файлового контента, которое с одной стороны подключается к LAN по шести 1GE- или двум 10GE-портам, а с другой стороны, имеет возможность подключения систем хранения по 4 FC интерфейсам. Поддерживаются 3 типа входящих протоколов: NFS, CIFS и iSCSI.

Оценив ограничения по масштабированию типовых NAS-серверов, BlueArc в 1998 г. выбрал новый подход к проблеме развития файловых сервисов, суть которого – в "погружении" части файловых функций в аппаратуру и использовании массового параллелизма при обработке файловых операций.

Архитектура NAS-решений BlueArc

Физически NAS-решение компании BlueArc представляет собой NAS-gateway или специализированную систему, управляющую потоком данных и файловых операций, которая для подключения к прикладным серверам и пользователям имеет или 6 1GE портов, или 2 порта по 10GE с одной стороны, а с другой – 4 FC-порта (1, 2 или 4 Gbps) для подключения системы хранения (рис. 1). Для управления имеются еще четыре 100 Mbps порта.

Концепция, заложенная в основу архитектуры решения от компании BlueArc, – это аппаратная реализация файловых систем как локальной SiliconFS, так и сетевых NFS, CIFS*).

За счет использования специализированных чипов FPGA и высокой параллелизации в обработке потоков в NAS-системе BlueArc удалось достичь производительность одну из самых высоких на тесте SPECsfs97_R1.v3 – 98131 NFS SpecOps/Sec *(http://www.spec.org/sfs97r1)* на 1 узел. Число узлов может масштабироваться от 1 до 4 (в ближайшее время – 8). При этом общий объем поддерживаемых ресурсов хранения в одной файловой системе может достигать 512 Тбайт (в ближайшей перспективе – до 2 Пбайт), а число файлов в одной директории – 4 млн и до 60 тыс. одновременно поддерживаемых соединений.

Одно из основных преимуществ, помимо названных, – возможность поддержки самых разных по требованиям приложений на одной консолидированной системе хранения, что обеспечивается за счет возможности выделения для них разных уровней RAID, типов дисков (FC, SATA), а также возможность назначения разных политик управления данными.

Программируемые вентильные матрицы – Field Programmable Gate Arrays (FPGA) – основа BlueArc Titan архитектуры

NAS-решения BlueArc строятся на т.н. Titan SiliconServer архитектуре. В ее основе лежит имплементация функций файлового сервера в аппаратуру (чипы и специализированные устройства). Технология SiliconServer представляет множество машин с массовым параллелизмом для реализации функциональности, которая в стандартном файловом сервере выполняется на уровне операционной системы, но с гораздо более высокой производительностью и надежностью. Однако при этом поддерживаются все стандартные протоколы для коммуникации с существующими клиентскими компьютерами.

*) CIFS – протокол, контролирующий состояние соединения на нескольких уровнях, поэтому только частично реализован аппаратно. Блокировки, история соединения и т.п. реализованы программно.

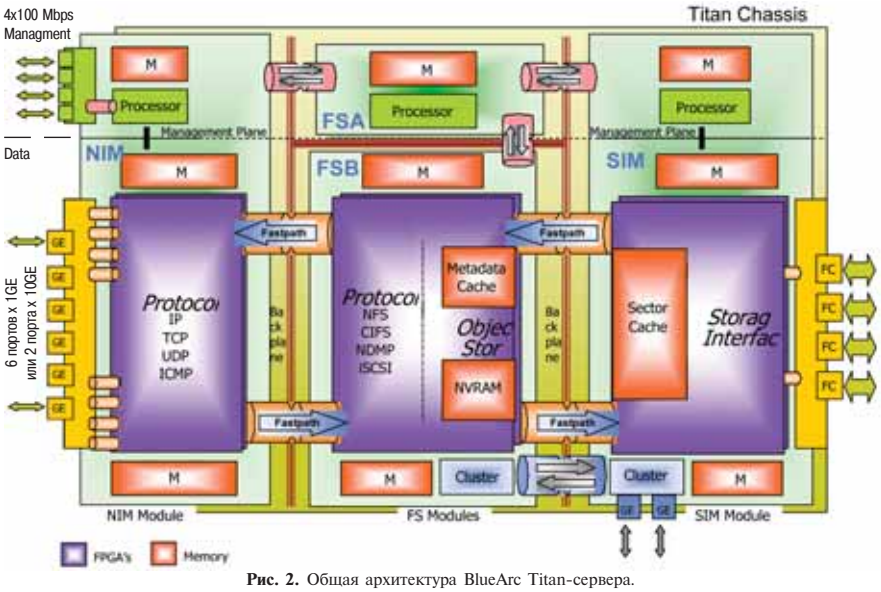

Рис. 2. Общая архитектура BlueArc Titan-сервера.

Программируемые вентильные матрицы – FPGA – или параллельные конечные автоматы составляют основу архитектуры BlueArc Titan. Современные FPGAs – перепрограммируемые интегральные схемы, представляющие собой высокоэффективные аппаратные компоненты/чипы с собственной памятью, буферами ввода-вывода и внутренней синхронизацией. FPGAs подобны специализированным интегральным схемам (Application Specific Integrated Circuits – ASIC), используемым, например, в высокоскоростных коммутаторах и маршрутизаторах, но, в отличие от ASIC, являются перепрограммируемыми. Возможность перепрограммирования FPGA во время эксплуатации дает повышенную гибкость решения. Например, возможность его изменения для выполнения новых или модифицированных задач, а также для поддержания новых протоколов, появляющихся на рынке, или решения проблем апгрейдов.

Разработчики аппаратуры иногда используют FPGA для предварительного дизайна, поскольку они позволяют быстро ("на лету") проводить все изменения в течение стадии проектирования и короткой эксплуатации. Как только логика отработана, вся схемотехника переносится на ASIC. В BlueArc Titan архитектуре на FPGA реализована финальная имплементация Titan-сервера.

За счет параллельной работы нескольких FPGA и их специализации производительность одноузлового Titan-сервера (по заявлениям разработчика) превосходит производительность одноядерного 3,8 ГГц микропроцессора Intel на однотипных операциях более чем в 10 000 раз.

Концептуально NAS-платформа компании BlueArc была разработана в трех отдельных секциях, названных подсистемами (рис. 2).

Первая подсистема – Network Interface Module (NIM) – ответственна за управление всеми Ethernet функциями ввода-вывода, соответствующими OSI уровню 1-4. Функции, имплементированные в NIM-модуль, включают управление Ethernet и Jumbo Ethernet фреймами до 9000 байт, ARP, IP-протоколом и роутингом, а также TCP- и UDP-протоколом.

Вторая подсистема, состоящая из двух модулей – File System Modules (FSA & FSB), поддерживает различные сетевые файловые системы, включая NFS, CIFS, FTP и NDMP, а также объектную файловую систему.

Третья секция – Storage Interface Module (SIM) – обеспечивает кэширование и управление для подключенных систем хранения.

Каждая подсистема имеет собственные процессорные мощности и память для управления всеми задачами параллельно. Такая архитектура позволяет всем процессам работать параллельно без влияния на другие процессы, что существенно отличается от архитектуры традиционных NAS-серверов.

Каждая из подсистем связана с соседними через 2 высокоскоростных канала, каждый из которых может писать и читать данные со скоростью до 20 Гбит/с без влияния друг на друга, что позволяет реализовать преимущества полной дуплексной связи.

Конструктивно NAS-платформа компании BlueArc представляет собой 4U модульную систему (рис. 3), в которую устанавливаются 4 специализированных модуля, функциональность которых описана выше. Также в состав решения входит выделенный сервер управления, т.н. SMU.

Для физической структуры хранения BlueArc имеет два требования. *Во-первых*, использование технологий, поддерживающих защиту данных и, *во-вторых*,

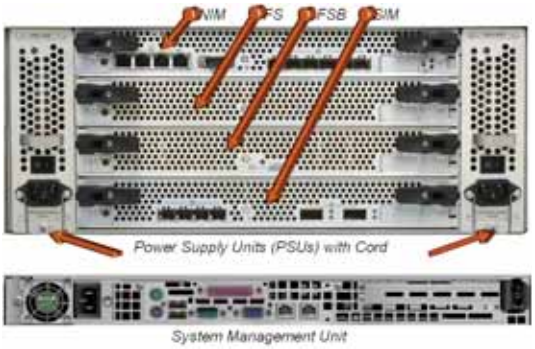

Рис. 3. Конструктивно NAS-платформа компании BlueArc представляет собой 4U модульную систему из 4 модулей и сервера управления (SMU).

обеспечение высокой потоковой производительности на операциях ввода-вывода. С целью поддержания высокой доступности подключение систем хранения к Titan-серверу обычно производится через два FC-коммутатора.

RAID-тома обычно конфигурируются как RAID-5. SIM-модуль затем стрипует до 32 RAID-томов в большой логический модуль, называемый stripe (страйп). Страйпы организованы в высокоуровневый объект, называемый span. Новые страйпы можно добавить к span в любой момент без какого-либо останова системы, позволяя динамически масштабировать тома. Такой дизайн Titan-сервера позволяет проводить его масштабирование как по емкости, так и производительности. Также можно масштабировать производительность добавлением RAID-массивов, а емкость – простым добавлением дисков к существующим RAID-контроллерам.

Ввод-вывод, идущий к span, который, в свою очередь, распараллеливается на страйпы, и все дисководы в пределах страйпа позволяет достигать наивысшей производительности. Такой параллельный RAID-стриппинг назван SiliconStack.

Архитектура SiliconServer также поддерживает уровневое хранение с использованием FC- и SATA-дисков в одной системе и возможностью миграции данных (на основе ПО) между уровнями. Также поддерживается LAN-free backup на основе использования стандартного NDMP-протокола.

Виртуализация ресурсов в рамках архитектуры BlueArc

Виртуализация ресурсов на всех уровнях – вторая отличительная составляющая архитектуры BlueArc (развивающая функциональность FPGA и специализированных аппаратных решений), значительно упрощающая управление файловыми сервисами в рамках BlueArc Titan архитектуры.

Интегрированная виртуализированная структура BlueArc состоит из нескольких уровней или компонент (рис. 4):

- виртуальные серверы;
- виртуальная файловая система с глобальным пространством имен (Global Name Space – GNS или Cluster Name Space – CNS);
- виртуальные тома;
- виртуальные пулы хранения с параллельным RAID-стриппингом.

Виртуальные серверы дают возможность разделить каждый физический Titan-сервер до 32 логических серверов. Каждый виртуальный сервер может иметь отличный IP-адрес и различные политики управления (организации) данных. Виртуальные серверы дают возможность разделения различных групп пользователей/приложений/проектов и назначения каждой группе своих политик управления производительностью, емкостью и доступностью данных.

Виртуальная файловая система от BlueArc представляет собой патентованную файловую систему, основные функции которой, а также файловые функции ОС реализованы на базе специализированных чипов. На основе программируемых логических чипов (микропроцессоров) и патентованной методологии была создана объектно-ориентированная ОС и файловая система, реализованная на множестве программируемых микропроцессоров, каждый из которых представляет специализированную вычислительную систему (с собственной памятью) для выполнения специфической задачи. Это позволяет организовать массивный параллелизм в обработке всех типов запросов: сетевых, от файловой системы и storage-запросов, которые могут обрабатываться одновременно и в одном цикле.

Примерное сравнение показывает, что в традиционных NAS-серверах по мере заполнения директорий время ответа возрастает. И хотя, теоретически, порог снижения времени реакции определяется десятками тысяч, на практике он намного ниже. Специализированная файловая система BlueArc может обрабатывать без замедления директории с числом файлов свыше 4 млн. Это может быть критично в таких областях как поддержание интернет-сервисов или в научных исследованиях, связанных с изучением жизни (life sciences – генетика, биология, иммунология и др.), где одновременно могут храниться миллиарды маленьких файлов.

Сама BlueArc файловая система представляет собой объектно-ориентированную файловую систему, где объект один или более блоков в древовидной структуре. Каждый элемент или объект называется Onode. Первичный элемент называется корневой Onode (Root Onode) и имеет уникальный 64-битный объектный идентификатор (Object Identifier – OID) и метаданные, связанные с ним. Этот корневой Onode может непосредственно указать на данные Onodes или на указатель Onodes в зависимости от количества содержания, которое будет сохранено. Через эту расширяемость файловая система может управлять/обращаться с единственным объектом такого размера, как полная файловая система или миллиарды меньших объектов, одинаково эффективно. Вся эта функциональность реализована на уровни микропрограммирования. К этому следует добавить, что такая функциональность как виртуальные тома и мгновенные снимки (snapshots) также встроены аппаратно и выполняются с высокой производительностью даже во время операций по защите данных.

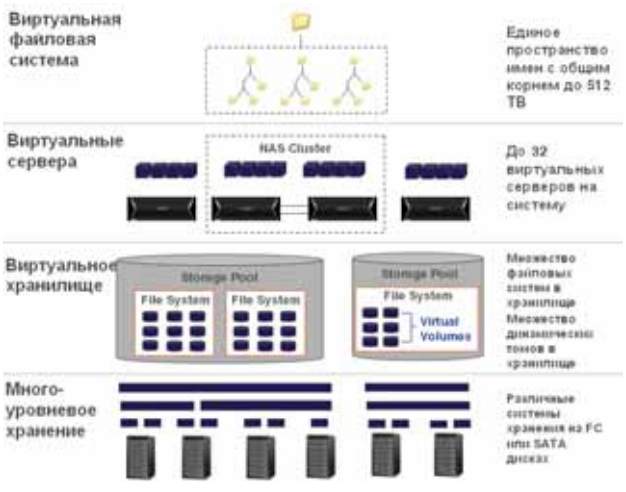

Рис. 4. Уровни виртуализации в рамках архитектуры BlueArc Titan-сервера.

Для пользователей и прикладных серверов не существует понятия объектов – это только внутренняя функция Titan-серверов. Несмотря на это, весь доступ к данным осуществляется на основе использования стандартных промышленных протоколов, таких как CIFS, NFS, iSCSI и NDMP, что устраняет потребность в специализированных драйверах или аппаратных средствах.

Также одной из основных особенностей файловой системы BlueArc является поддержка глобального пространства имен для инфраструктуры Titan-сервера (включая кластеризованные инфраструктуры), что позволяет объединить все файловые системы и делает возможным доступ к данным из любой точки. CNS позволяет администраторам управлять физическими пулами хранения и файловыми системами без влияния на то, как пользователи получают доступ к их файлам. Администраторы просто создают директорию виртуальной файловой системы, связывая ее с реальной файловой системой в пределах Titan-сервера или кластера. Общий объем такой виртуальной файловой системы может составлять до 512 Тбайт. К любому файлу или папке можно тогда обратиться через CIFS или NFS.

CNS работает совместно с т.н. BlueArc Tiered Storage функциональностью, обеспечивая консолидированное пространство имен для файловой системы, использующей различные дисковые технологии. Например, администраторы могут создать псевдодиректорию со связью к первому уровню (Fibre Channel) для “рабочей” директории и другую, ко второму уровню – для архива.

Виртуальные тома – инструмент администрирования системы, разрешающий системному администратору логически группировать директории внутри тома, с которым можно связывать политики, свойства, сервисы. Виртуальные тома – логические объекты/контейнеры, которые находятся на вершине основного тома. Виртуальные тома могут быть динамически расширены, зарезервированы, экспортированы или разделяться. Внутри одного тома может быть создано более 2 000 виртуальных.

Клиентские машины видят изменения размеров виртуальных томов немедленно. Суммарное пространство управляемого виртуального тома может быть больше, чем размер файловой системы. Этот подход, называемый тонкий provisioning, обеспечивает дополнительную

гибкость, когда прогнозируемый объем тома не определен. Это позволяет администраторам сразу представлять максимальный объем хранения и физически расширять его по мере необходимости.

Пулы хранения виртуализируют физическое SAN-хранение и позволяют автоматически изменять размер файловой системы в соответствии с заданными правилами, избавляясь от необходимости предварительного выделения ресурсов хранения. В один пул хранения могут быть объединены системы хранения разного класса, например, с разными дисковыми интерфейсами – FC и SATA с целью более оптимального использования ресурсов для разных файловых сервисов. Архитектура BlueArc может поддерживать пулы хранения до 512 Тбайт и более.

Интеграция решений BlueArc с системами HDS

Использование систем хранения Hitachi Data Systems для построения комплекса Hitachi High performance NAS (HNAS) имеет ряд уникальных особенностей.

Прежде всего, это централизованное управление как системой хранения, так и HNAS с помощью пакета HiCommand. Построение LUN(ов), на основе использования RAID6, даже при выходе из строя двух дисков в одной дисковой группе, позволяет гарантировать сохранность данных. В случае неисправности в системе хранения программа мониторинга Hi-Track автоматически создаст заявку на сервисное обслуживание, что позволит моментально проинформировать организацию, оказывающую сервисные услуги и зарезервировать необходимые компоненты на ближайшем складе запасных частей.

При использовании дисковых массивов High End класса можно воспользоваться технологией Thin Provisioning, что позволяет приобретать жесткие диски по мере необходимости.

Если в составе комплекса используются модульные системы хранения, можно воспользоваться аппаратным решением Hitachi Cache Partition Manager и предоставить каждому из 32-х виртуальных серверов EVS (Enterprise Virtual Server) свою область в кэш-памяти дискового массива, определив при этом размер страницы для операций ввода-вывода.

Любые системы хранения в составе Hitachi High performance NAS можно использовать совместно с другим оборудованием, при этом часть емкости выделить для HNAS и воспользоваться файловым доступом, а часть емкости использовать как хранилище с блочным доступом.

Hitachi High performance NAS поставляется с оптическими SAN-коммутаторами, соединенными по схеме Dual fabric, эта инфраструктура может быть использована для построения отказоустойчивой сети SAN.

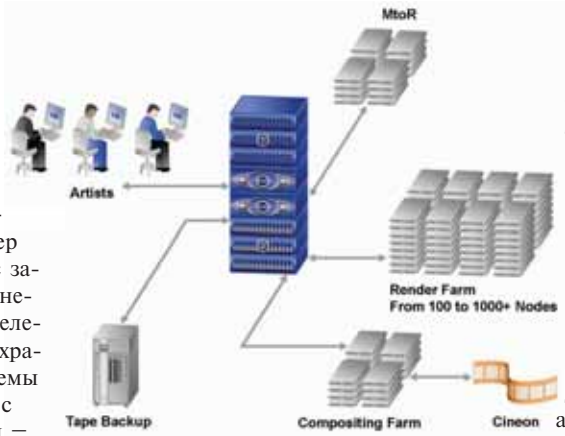

Рис. 5. Использование BlueArc Titan-сервера для CGI-анимации позволяет поддерживать более 1000 вычислительных узлов и более 50 рабочих станций художников-аниматоров.

Примеры использования

Имея очень хорошую масштабируемость по NFS-производительности и поддерживаемой емкости, использование BlueArc Titan сервера особенно эффективно по двум направлениям: 1) в качестве системы для организации высокопроизводительного доступа к данным в составе HPC-систем и им подобным; 2) в качестве корпоративного хранилища для консолидации данных очень широкого круга приложений – от HPC-систем до поддержки критичных БД и совместного использования файловых хранилищ со всеми типами хранения – NAS, SAN и частично – CAS. При этом полностью обеспечивается централизованное консолидированное управление всеми данными с одной консоли.

Большое число инсталляций BlueArc Titan сервера связано с применениями для 3D-моделирования (или для computer generated imagery CGI – производство цифровых анимационных фильмов и спецэффектов). Так, внедрение BlueArc Titan на студии Giant Killer Robots (GKR, Сан-Франциско, США) позволило поддерживать работу более 1000 вычислительных узлов для рендеринга и более 50 художников-аниматоров (рис. 5).

Опыт инсталляции BlueArc Titan в другой студии анимации – Rhythm & Hues – позволил уменьшить время рендеринга ряда сцен в 24 раза – с 6 часов до 15 минут, при этом стоимость обслуживания ресурсов хранения (в расчете за мегабайт) была снижена в 2 раза.

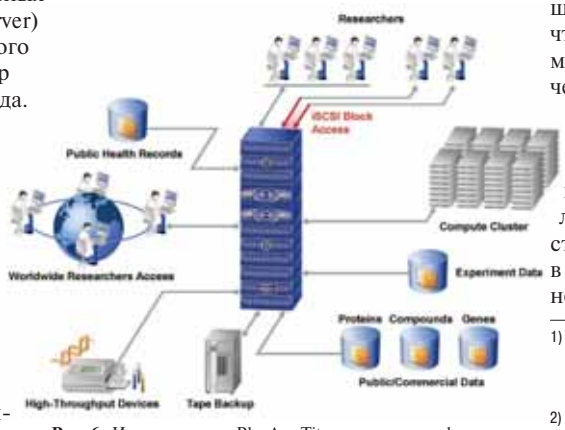

Рис. 6. Использование BlueArc Titan-сервера для drug design в Гарвардском университете в составе высокопроизводительных HPC-систем при проведении 20 000 сравнений позволило сократить время с 4 дней до 4 часов.

Еще пример: "Имплементация BlueArc Titan позволила сократить время считывания файлов с 15 минут до нескольких секунд" – Jami Levesque, Director of Technology, Meteor Studios.

Успешными были внедрения BlueArc Titan в Гарвардском университете в составе высокопроизводительных HPC-систем (рис. 6) для исследований в области молекулярной биологии, медицинской и общей генетики, биохимии для анализа геномов, сравнительного анализа белков и ДНК, изучения пространственной структуры и моделирования белков, разработки новых лекарственных средств, в основе которых лежал т.н. компьютерный drug design[1]. В результате, в среднем, при проведении 20 000 сравнений время сократилось с 4 дней до 4 часов. Базы данных, которые при этом обрабатывались, в объеме превышали более 100 Тбайт.

Заключение

Интеграция решений от BlueArc и HDS позволит консолидировать на ограниченном числе систем хранения очень широкий круг приложений, значительно отличающихся производительностью, емкостью, требованиями к хранению и управлению данными. Наличие единого центра управления и высокой масштабируемости подобных платформ дает возможность эффективного управления данными для самых разных секторов рынка.

Самая маленькая головка чтения для HDD

Октябрь 2007 г. – Компании Hitachi, Ltd. и Hitachi Global Storage Technologies (Hitachi GST) объявили о завершении разработки новой технологии создания самой маленькой в мире головки чтения для жесткого диска, что в 4 раза увеличит его емкость – до 4 Тбайт на жестком диске для настольных систем и до 1 Тбайт на жестком диске для ноутбуков.

Исследователям Hitachi удалось уменьшить размер головки более чем в 2 раза, что соответствует размеру 30-50 нанометров, что примерно в 2000 раз тоньше человеческого волоса (его толщина составляет около 70-100 микрон). Ожидается, что новая головка Hitachi типа CPP-GMR[2], использующая гигантский магниторезистивный эффект и технологию потока перпендикулярно к пластине, впервые появится в устройствах в 2009 г., а пика популярности достигнет в 2011 г.

1) drug design – компьютерное моделирование пространственной структуры белка. Требует использования больших БД с сотнями миллионов записей и перебором/сравнением огромного числа вариантов, *прим. ред.*

2) CPP-GMR: альтернатива традиционным головкам на основе туннельного магниторезистивного эффекта (TMR). Головки, разработанные на основе технологии CPP-GMR, обладают меньшим уровнем электрического сопротивления, так как их действие основано не на туннельной, а на металлической проводимости. Они обеспечивают более высокую скорость работы и дальнейшее уменьшение размеров.