Sun N1 DSP – универсальная платформа для систем среднего класса

15 сентября 2004 г. прошло официальное публичное представление продукта (продажи с 13 июля), которого долго ожидали, – Sun StoreEdge 6920 – первая реализация компанией Sun Microsystems своего направления в SAN-виртуализации. Этим объявлением Sun Microsystems заложила основу целой серии своих решений среднего класса на базе т.н. платформы сервисов данных (Data Services Platform – DSP).

Введение

Появление StoreEdge 6920 по сути – заявка Sun предложить свою основу для построения целой серии решений для среднего сектора рынка. Эти решения имеют долгосрочную перспективу развития и строятся на базе относительно недорогого с высокой масштабируемостью по техническим показателям и функциональности аппаратно-программного комплекса, или платформы DSP.

Первая реализация DSP – продукт SE6920, который состоит из основных двух компонент: самого DSP и непосредственно дисковых блоков (в настоящее время работает только с системой SE6020). Спектр поддерживаемых в настоящее время операционных серверных платформ достаточно широк: Solaris 8/9/10, Microsoft Windows 2000 Server/Advanced Server SP4/2003 Server, IBM-AIX 5.1 (64-bit), HP-UX 11.0/11i, Red Hat Linux. ПО-основная заявленная поддержка более широкого спектра дисковых массивов и ряда существенных расширений в функциональности планируется в первой половине 2005 г.

Чего не хватает современным системам хранения среднего класса

Если говорить обобщенно и кратко, то это: гибкости, масштабируемости, управляемости. Современные системы среднего класса на рынке представлены в основном модульными системами. Архитектурно они построены по однотипной схеме: несколько петель FC (от 1 до 4) с пропускной способностью 1/2/4 Гбит/с объединяются общим кэшем и контроллерами в единое целое. Такие системы имеют ограниченную масштабируемость по техническим параметрам: производительности, портам подключения к хостам, не очень высокую управляемость, слабую гибкость в перераспределении ресурсов, а также по функциональности – в плане поддерживаемых сервисов данных, возможности "заточки" под специализированные решения. Эти ограничения диктуются, прежде всего, контроллерами модульных систем хранения. Введение DSP в качестве основы для систем среднего класса, в состав которой входят масштабируемая процессорная система, коммутатор (доступен с 16 и 32 FC 2Gbit портами) и специализированное ПО, которое тоже может масштабироваться и развиваться, кардинально решает проблему. Т.е., в отличие от традиционных модульных систем при использовании DSP (например, в составе SE6920) интеллектуальная часть выносится на отдельную платформу и полностью не зависима от самого массива дисков, за счет чего появляется возможность строить мощные по функциональности специализированные системы на базе недорогих дисков. При этом система имеет гораздо бóльшую масштабируемость, чем традиционные. Безусловно, эта функциональность может быть достигнута на существующих системах, но это потребует их интеграции с другими продуктами и значительных дополнительных вложений средств.

Сама идея DSP не нова и активно развивается в отрасли другими вендорами. Так, год назад IBM объявила о введении SAN Volume Controller – аналогичного по направленности продукта. Вопрос в данном случае – в цене, конкретной физической реализации и возможностях.

Рассмотрим более подробно ограничения, накладываемые современными модульными системами.

Один из основных недостатков таких систем – их ограниченная масштабируемость по производительности. Так, если взять FC-интерфейс с пропускной способностью 2 Гбит/с, то при средней производительности одного HDD, равной 250 операций ввода/вывода в секунду *(SN № 7, апрель-май 2001),* и блоке данных – 8 Кбайт, насыщение интерфейса наступит при 60 дисководах (при этом производительность случайного доступа будет составлять 15 000 IOPS). Дальнейшее увеличение числа накопителей приведет только к повышению общей емкости. Таким образом, при наличии двух петель FC (с пропускной способностью 2 Гбит/с каждая) при равномерной загрузке накопителей случайная производительность системы будет увеличиваться от 1 до 120 дисков (рис. 1) и составит максимум 30 000 IOPS (блок = 8 Кбайт). Но это теоретически максимально возможная производительность, обычно она ограничивается при задержке (при обработке запроса) более чем

Рис. 1. Небольшое число разделяемых FC петель в модульных системах приводит к тому, что производительность не масштабируется с увеличением числа накопителей (пример для одной FC петли).

10-30 мс (величина порога меняется в зависимости от стандартов).

Плохая масштабируемость модульных систем порождает другую проблему – невысокую эффективность их использования, когда рано или поздно потребитель, не имея возможности приобретения одной более мощной системы хранения, вынужден наращивать IT-инфраструктуру новыми модульными системами, что без консолидации ресурсов в единый пул приводит к низкому коэффициенту их использования и более плохой управляемости, несмотря на все усилия IT-персонала.

Низкая масштабируемость модульных систем имеет еще один аспект: с учетом имеющихся предложений на рынке, она фактически создает большой разрыв между модульными и монолитными системами при сравнении на мультипараллельном гетерогенном потоке (комбинация OLTP, DSS и др.) по показателям производительности, которые в отдельных случаях отличаются не в разы, а на порядки. Это также в ряде случаев приводит к необоснованным издержкам со стороны потребителей.

Решение проблем масштабируемости производительности современных модульных систем и их управляемости, в основном, лежит в направлении создания общего виртуального пула из отдельных устройств на основе специализированных программно-аппаратных систем, требующих дополнительных вложений и усилий. Возможность масштабирования (вертикального) за счет частичной замены отдельных компонент (контроллеров) модульных систем более высокопроизводительными, не изменяя все остальное (диски, систему электропитания/охлаждения, конструктив), имеет явную ограниченность и допускает расширение в рамках только одной продуктовой линейки. В обоих случаях потребитель должен нести дополнительные затраты, которые, например, при организации общего виртуального пула сопоставимы (и более) со стоимостью самой модульной системы), не обеспечивая при этом желаемой гибкости и интегрированности решения.

Итогом и некоторой формализацией сказанного является концепция N1™ Grid Data Services Platform, развиваемая Sun Microsystems, которая позволяет интегрировать ранее сложные аппаратно-программные многокомпонентные решения в единое целое с гибкими характеристиками, легко адаптируемыми/развертываемыми в среде использования, с множеством встроенных сервисов данных.

Архитектура и принципы построения SE6920

SE6920 – первый продукт в семействе под флагом N1™ Grid DSP.

В разработке SE6920 Sun Microsystems постаралась объединить в одном продукте многое из лучшего, что предлагается сейчас на рынке. Две ключевые особенности (рис. 2) выделяют SE6920 в общем ряду модульных систем хранения. *Во-первых,*

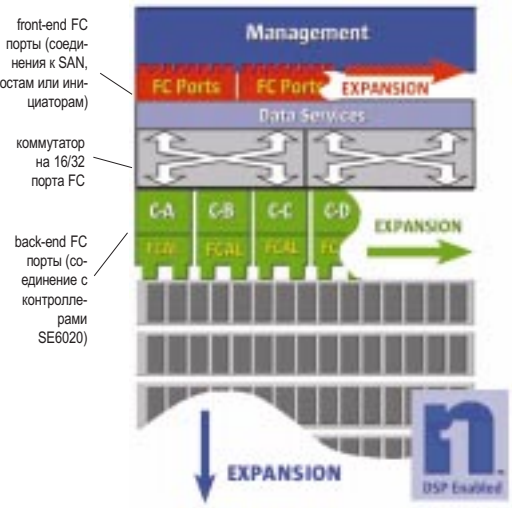

Рис. 2. Архитектура Sun StorEdge 6920.

в SE6920 интегрированы т.н. сервисы данных, или, проще, функциональность, которая раньше предлагалась, в основном, в составе high-end систем или развивалась отдельно. SE6920 имеет два встроенных сервиса данных в своем составе. Это виртуализация томов *(Sun StorEdge Storage Pool Manager software),* что дает возможность SE6920 изначально взаимодействовать с клиентом на уровне виртуальных томов, размеры и профилирование которых для конкретных применений легко изменяются без привязки к физическим ресурсам. Физически это организовано следующим образом. Массивы объединяются в storage pool'ы, точнее, не массивы, а LUN'ы массивов. Поддержкой LUN'ов и их отображением на физические диски через механизм RAID занимается микрокод массивов SE 6020. Сам SE6920, или его виртуализационная часть (DSP), занимается только созданием виртуальных дисков (volume) поверх storage pool'ов, а также раздачей этих volume хостам. Виртуализация значительно уменьшает сложность, затраты на управление, время развертывания ресурсов хранения, а также существенно повышает коэффициент использования ресурсов.

Второй сервис данных – создание мгновенных копий томов *(Sun StorEdge Data Snapshot software).* Эта функциональность поддерживается специализированными процессорами, которые могут масштабироваться от 1 до 16 (в зависимости от требований) и не затрагивают ресурсы (пути доступа, RAID-контроллеры), связанные с хранением/доступом к данных, т.е. практически не влияют на основные характеристики.

Традиционно сервисы данных выполняются или на базе ресурсов дискового массива, или на основе хоста*). В первом случае сервисы данных поддерживаются, используя пути передачи данных хост/массив; во втором – используется мощность сервера. Как в первом, так и втором случаях это отражается на скорости (снижение) обработки приложений.

Во-вторых, SE6920 имеет в качестве ядра (см. рис. 2) FC коммутатор (точка-точка) на 16 или 32 порта (в зависимости от комплектации), что значительно расширило гибкость и масштабируемость системы.

Причем масштабируемость может осуществляться по емкости, производительности, связности с хостами, мощности сервисов данных и размеров кэша на каждую FC петлю (т.н. N-way масштабируемость в терминологии Sun). В плане расширения гибкости использования портов SE6920 представляет попытку симбиоза принципов организации монолитных систем и модульности систем среднего класса. Наличие коммутатора (ранее было, безусловно, признаком high-end системы) позволило в рамках одного устройства значительно увеличить степень варьирования связностью SE6920 с хостами, или front-end портами (изменяя ее от 4 до 28 FC портов) и количеством FC петель, или back-end портами (от 2 до 28 в максимальном варианте).

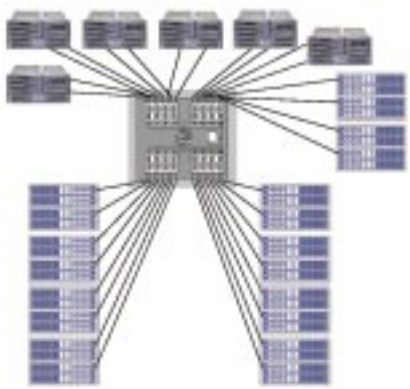

Рис. 3. Конфигурирование Sun StorEdge 6920 для достижения максимальной масштабируемости по производительности на случайных операциях IO (6 серверов, 20 FC петель).

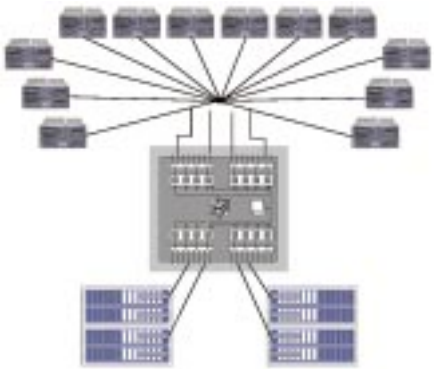

Рис. 4. Конфигурирование Sun StorEdge 6920 для достижения максимальной связности (12 серверов со сдвоенными FC портами).

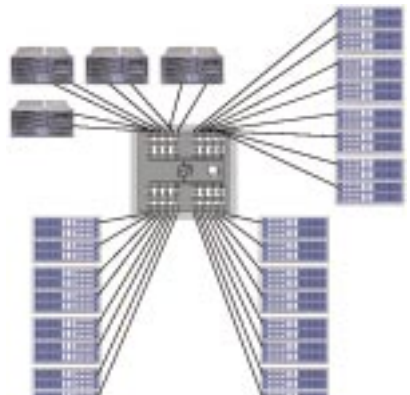

Рис. 5. Конфигурирование Sun StorEdge 6920 для достижения максимальной емкости (4 сервера, 24 trays).

*) Сервисы данных также активно на уровне сети развиваются всеми основными поставщиками SAN-коммутаторов. Но до настоящего времени, полнофункциональные продукты этого класса на рынке отсутствуют – *прим.ред.*

В общем случае соотношение между front-end и back-end портами коммутатора (из общего числа 16/32) можно менять произвольно, однако нужно помнить, что если речь идет о масштабируемой производительности, то она должна быть сбалансирована как со стороны front-end, так и back-end портов одновременно. Как показали тестовые испытания *(http://www.storageperformance.org/results/a00033_sun_SPC1_executive-summary.pdf)*, максимальная производительность на случайных операциях ввода/вывода была достигнута в соотношении front-end /back-end портов, равном 12/20. Т.е. при полной загрузке и двойной избыточности портов к SE6920 можно подключить (рис. 3) до 6 серверов (с двумя 2 Гбит/с HBA на сервер), при этом гарантируется полная масштабируемость системы на случайных операциях IO (увеличивающаяся нагрузка на серверы будет полностью удовлетворяться системой хранения). Естественно, соотношение front-end /back-end портов может меняться в зависимости от типа нагрузки.

Максимальная связность, рекомендуемая разработчиком, – 12 серверов (по 2 HBA на сервер) на 32 FC порта SE6920 (рис. 4).

Максимальная емкость в настоящее время составляет 28 trays = (HDD 146 Гбайт) x (14 дисков в одном tray) x 28 = 57 Тбайт (рис. 5). При этом число серверов – 4 (по 2 HBA в каждом) с большим запасом по производительности уравновешиваются системой хранения.

Производительность и применение SE6920

Вопросы оценки производительности, а также стоимости единицы производительности – одни из самых болезненных для разработчиков. При проведении такого анализа крайне важно сравнивать системы одного класса, в которых производительность поддерживается примерно равными уровнями доступности, надежности и сервиса. В противном – в выигрыше всегда будут самые дешевые системы с невысокими уровнями надежности и поддержки.

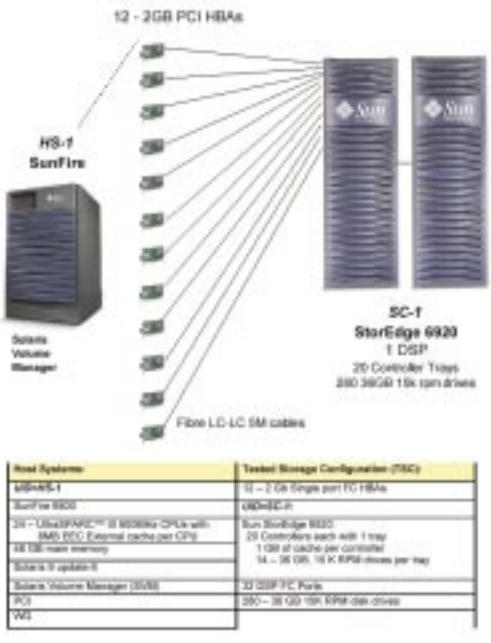

Рис. 6. Конфигурация тестирования Sun StorEdge 6920 на эталонной нагрузке SPC-1.

Табл. 1. Сравнение различных систем хранения на эталонной нагрузке SPC-1

Company	System	SPC-1 IOPS	$/SPC-1 IOPS	ASU Cap. (GB)	TSC Price	Pub. Date
Sun	StorEdge 6920	48,647	$10.73	3022	$522,087	16.08.04
Sun	StorEdge 6320	44,806	$15.41	3276	$690,433	22.03.03
IBM	TotalStorage SAN Volume Contr.	44,508	$15.08	5025	$671,314	06.08.04
StorageTek	StorageTek D280 Disk System	24,507	$12.56	1196	$307,904	17.02.04
IBM	TotalStorage FAStT900 (NCM)	24,507	$12.63	1196	$309,499	26.08.03
HP	EVA Model 2C12D (NCM)	24,006	$19.99	2596	$479,860	10.02.03
IBM	ESS 2105-M800	22,999	$34.88	3207	$802,116	13.05.02
HP	EVA Model 2C12D	20,097	$23.88	2596	$479,860	12.06.02
IBM	TotalSTorage FAStT900	18,448	$16.78	1196	$309,499	26.08.03
StorageTek	StorageTek D280 Disk System	18,448	$16.69	1196	$307,904	17.02.04
Fujitsu	ETERNUS 3000	17,545	$37.92	2076	$665,370	13.02.03
LSI	MetaStor E4600	15,701	$16.01	400	$251,434	13.05.02
3PAR	InServ S800 (2node)	12,906	$19.73	896	$254,638	12.09.02

Комитет по оценке производительности систем хранения в среде SAN (Storage Performance Council – SPC, *www.storageperformance.org*) был создан в 1996 г., но активно стал выдавать результаты только около трех лет назад. SPC поддержан ведущими вендорами отрасли, такими как: IBM, HP, Sun Microsystems, Ngenio (LSI LOGIC), VERITAS и другими и на сегодняшний день является одной из немногих организаций, проводящей подобные измерения и предоставляющей их результаты в открытом доступе. Каждая имитационная нагрузка моделирует определенный класс приложений и фиксирует максимально достижимый уровень производительности для конкретной среды и при равномерной загрузке, нивелируя ряд архитектурных особенностей и алгоритмов управления потоками данных отдельных систем в сравнении с другими. В реальных условиях прежде всего может сказываться: несбалансированность нагрузки; одновременное наличие разнородных потоков, требующих разных решений для их оптимизации; различный уровень накладных издержек, связанных, например, с поддержанием дополнительных сервисов для обеспечения доступности основных процессов и др. Поэтому влияние и важность всех дополнительных факторов при проектировании каждой конкретной системы нужно учитывать отдельно.

SPC-1 (SPC Benchmark-1™) – первый промышленный эталонный тест, разработанный для тестирования в среде SAN корпоративных многопользовательских приложений ввода/вывода типа СУБД/OLTP и почтовых серверов. В настоящее время он наиболее используемый для сравнения систем по производительности на случайных операциях ввода/вывода.

Тестирование SE6920 проводилось в конфигурации, показанной на рис. 6 *(http://www.storageperformance.org/results/a00033_sun_SPC1_executive-summary.pdf)*. Результаты измерений в сравнении с системами, которые были представлены для тестирования, даны в табл. 1. В заданном спектре продуктов SE6920 показал наилучший результат, при этом он обладает некоторым резервом за счет встроенных сервисов данных. При оценке второго показателя – стоимости на единицу производительности ($/SPC-1), SE6920 – тоже в первой строке, но здесь также следует учитывать комплектацию системы, т.е. при прочих равных условиях системы необходимо сравнивать по общей емкости и их потенциалу.

Важно заметить, что ценность подобного тестирования заключается не только в получении абсолютных значений параметров, но и в представлении вендором оптимизированной конфигурации для тестирования и его ценовой политике.

В первой реализации SE6920 ориентирован в бóльшей степени на бизнес-критические приложения типа: управление базами данных – как в системах обработки транзакций (On-Line Transaction Processing – OLTP), так и в системах поддержки принятия решений (Decision Support System – DSS); инженерные расчеты (technical computing); управление почтовыми сообщениями.

Системы управления базами данных для интерактивной диалоговой обработки запросов (OLTP) требуют высокой производительности на случайных операциях ввода/вывода. Данный тип бизнес-приложений характеризуется очень большим количеством мелких, случайных, непоследовательных транзакций чтения и записи в течение ограниченных периодов времени. Эти требования в SE6920 поддерживаются подтвержденной производительностью ввода/вывода, высокомасштабируемым зеркалируемым кэшем (до 2 Гбайт/контроллер, до 16 Гбайт/том, до 28 Гбайт/устройство) и отсутствием недублированных точек отказа.

Почтовые приложения комбинируют нагрузку OLTP и DSS баз банных в виде интенсивных запросов ввода/вывода как с короткими, так и очень большими вложенными файлами. Данные требования поддерживаются высокой масштабируемостью и гибкостью SE6920.

Поддержка высокодоступных приложений, требующих работоспособности в течение 24 час/сут обеспечивается всей архитектурой и конструктивом SE6920 – сдвоенными RAID-контроллерами/дисками/кэшем/путями доступа/блоками питания/блоками вентиляторов/картами внутренних межсоединений, с возможностью их горячей замены без останова приложений, а также возмож-

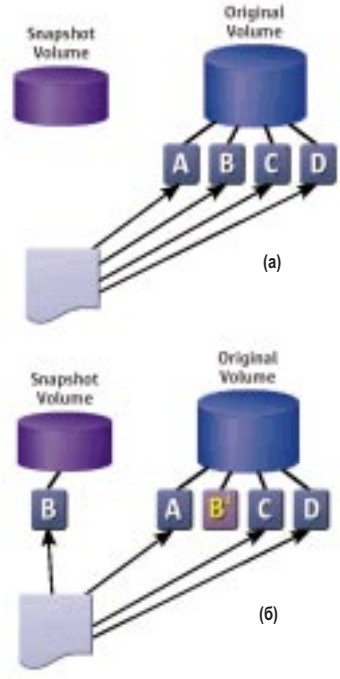

Рис. 7. Когда мгновенный снимок создается, все указатели метаданных указывают на данные оригинального тома (а). Когда данные в оригинальном томе модифицируются, данные копируются в snapshot и указатели метаданных заменяются (б).

ностью дублирования путей передачи данных между хостом и SE6920 под управлением Sun StorEdge Traffic Manager ПО.

Функциональные особенности SE6920

Важной встроенной функциональной особенностью SE6920 является возможность создания мгновенных снимков (snapshot) тома (до 8 снимков/том). Мгновенный снимок в SE6920 становится доступным для использования в течение нескольких секунд. Наличие этой функциональности в значительной степени облегчает выполнение множественных операций над одними и теми же данными, например, резервное копирование, тестирование, анализ данных, подготовка отчетов и др. Однако поддержание этой функциональности в ряде систем приводит к значительному снижению производительности и уменьшению доступных ресурсов для основных производственных задач. В SE6920 это влияние сведено к минимуму. *Во-первых*, за счет того, что в SE6920 создаются не реальные тома, а их образы и в snapshot'е фиксируются только изменения первичного тома (рис. 7). *Во-вторых,* поддержание snapshot'ов осуществляется архитектурно на уровне специализированных аппаратных средств (которые могут масштабироваться).

SE6920 поддерживает до 1024 томов (включая мгновенные снимки томов). Размер тома – от 16 Мбайт до 2 Тбайт.

Другой интересной особенностью SE6920 является профилирование томов по шаблону. Каждое приложение имеет свои характеристики, которые необходимо учитывать для оптимизации доступа к данным при развертывании приложений. Таких параметров при инициализации тома достаточно много: Sequential Access, Segment Size, Queue Depth, Block Size, Chunk Size (квант, на который делится блок данных при работе с диском), Read-ahead, тип RAID. Они меняются в зависимости от требований и типа приложения (*SN № 7, апрель-май, 2001*) и требуют достаточно много времени и бóльших усилий по анализу данных. В SE6920 имеется 14 предустановленных профилей нагрузки: Oracle OLTP, Oracle DSS, Mail spooling, File serving, HPC и др., значительно уменьшающих сложность, повышающих эффективность использования массива, снижающих вероятность ошибок, а также увеличивающих легкость переноса приложений на другие системы.

Конструктивная реализация SE6920

Конструктивно (рис. 8) Sun StorEdge 6920 поставляется или в виде одной базовой стойки (2 SRC системы –Storage Resource Card; четыре 2x2 Sun StorEdge 6020 с 7 дисками на трэй), или в виде трех стоек (базовая в комплектации с 2 стойками расширения – 4 SRC системы; восемь 2x4 Sun StorEdge 6020 с 7 дисками на трэй).

Максимальная емкость – от 112 (базовая стойка) до 448 накопителей 4 типов (36 Гбайт,15K rpm; 73 Гбайт,10K rpm; 73 Гбайт,15K rpm; 146 Гбайт,10K rpm) или до 65,4 Тбайт.

Направления развития DSP

Первая версия SE6920 имеет ограниченную интегрируемость с другими системами хранения. Не реализована как совместная работа с продуктами от других вендоров, так и со всем рядом платформам Sun Microsystems. На данный момент виртуализационное ядро SE 6920 (DSP-1000) поддерживает только массивы SE6020 и T3+. Гетерогенность планируется ввести в следующей версии DSP (обновленный SE6920 с новым DSP скорее всего будет иметь другое маркетинговое название). Выпуск нового SE6920 планируется на первую половину 2005 г. и, возможно, будет совместим со следующими дисковыми массивами: Sun StorEdge 3x00, 6x00, 9x00; EMC CXs; HP EVA, HP EMA; HDS 9500 и LSI.

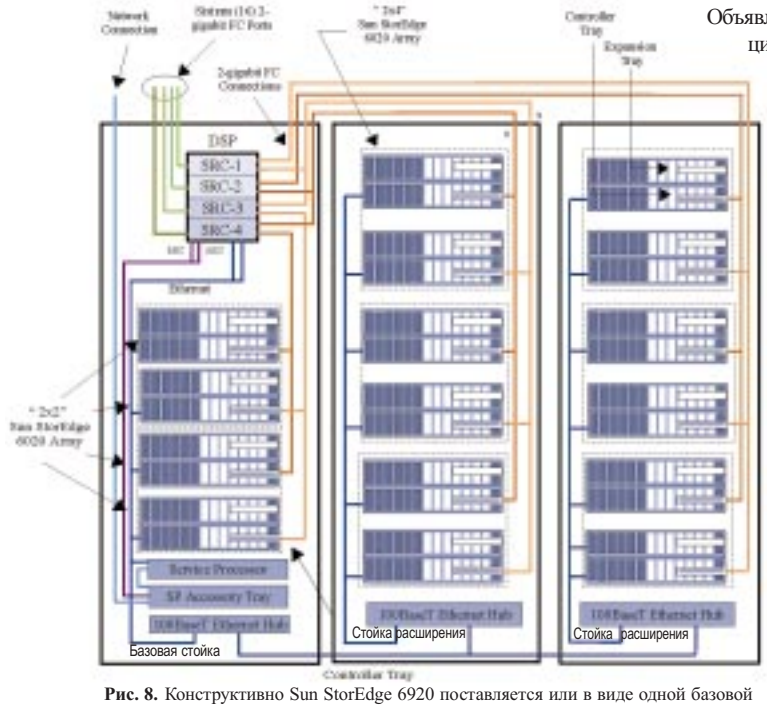

Рис. 8. Конструктивно Sun StorEdge 6920 поставляется или в виде одной базовой стойки, или в комплектации с двумя стойками расширения.

Объявлено также об интеграции DSP в линейку SE6130. В целом, гетерогенность SE6920 будет развиваться на основе стандартов SMI-S и иметь единую точку управления для всей среды.

Планируется, что в 1 кв. 2005 г. к сервисам данных SE6920 будут добавлены: Volume Copy, Data Mirror, Remote Replication. В текущей версии удаленное и локальное зеркалирование томов можно осуществлять средствами Sun StorEdge Availability Suite (по TCP/IP). Также в течение первого квартала следующего года в SE6920 будет включена поддержка SATA-дисков.

Заключение

SE6920 – "первая ласточка" семейства N1 Grid DSP ориентированного на повышение общего уровня интеллектуализации систем хранения среднего класса, снижение их стоимости, увеличение гибкости и функциональности, а также достижение построения более специализированных систем среднего уровня, например, контентно-адресуемых. Данное направление лежит в рамках общих мировых тенденций и представляет большой интерес для рынка.

Sun Microsystems представляет UltraSPARC IV+

Октябрь 2004 г. – Корпорация Sun Microsystems представила процессор нового поколения UltraSPARC IV+. Этот процессор на базе технологии Chip Multithreading (многопотоковость на кристалле) – очередной шаг в реализации стратегии Throughput Computing и позволяет одновременно выполнять несколько инструкций, или потоков команд, что значительно увеличивает производительность системы.

"Процессор UltraSPARC IV+ – это второе поколение развивающейся двухъядерной архитектуры Sun Microsystems. Мы с нетерпением ждем появления революционного процессора под кодовым названием Niagara, выход которого на рынок планируется в 2006 г. Он будет существенно отличаться от традиционных процессоров и коренным образом изменит наши представления о сетевых вычислениях", – прокомментировал выпуск новинки Кевин Кревелл (Kevin Krewell), главный аналитик Microprocessor Report.

UltraSPARC IV+ разработан на базе 90-нанометровой производственной технологии Texas Instruments. По сравнению с процессором UltraSPARC IV он позволяет удвоить производительность приложений благодаря увеличению объема кэш-памяти и буферов, улучшенному механизму прогнозирования ветвления, расширенным возможностям упреждающей выборки из памяти и новым вычислительным возможностям. Кроме этого, в UltraSPARC IV+ применяется новая трехуровневая иерархия кэш-памяти, включающая интегрированную на кристалле кэш-память второго уровня объемом 2 Мбайт и внешнюю кэш-память третьего уровня объемом 32 Мбайт.

Кроме новых функций увеличения производительности, процессор UltraSPARC IV+ имеет значительно более высокую тактовую частоту (первоначально 1,8 ГГц), что обеспечивает наивысшую пропускную способность по сравнению с другими моделями процессоров UltraSPARC. По сравнению с процессором UltraSPARC IV производительность каждого потока выросла примерно вдвое.

Новый процессор UltraSPARC IV+ обеспечивает бинарную совместимость с предыдущими поколениями процессоров архитуктуры SPARC и позволяет пользователям сохранить инвестиции в средства разработки и в прикладное ПО.