Виртуализация хранения

Публикация дает обзор наиболее популярных в России решений по виртуализации хранения (storage virtualization). Рассматриваются преимущества, ограничения и позиционирование каждого из решений.

Введение

Первые решения по виртуализации хранения стали доступны еще в начале 2000 г., но наиболее полная таксономия и стандартизация этих решений была сделана SNIA (Storage Networking Industry Association) лишь в 2003 г. Среди основных целей виртуализации можно выделить следующие: 1 – абстрагирование от деталей на различных физических уровнях при управлении ими через их единообразное логическое представление; 2 – объединение в общий пул гетерогенных ресурсов, прозрачных для приложений, клиентов и др.; 3 – упрощение и оптимизация управления ресурсами хранения, а во многом – автоматизация управления ими и, как следствие, – снижение затрат на администрирование и повышение эффективности. Помимо этого, также можно отметить, что виртуализация позволяет:

- *значительно упростить масштабирование IT-инфраструктуры (добавления/изменения типов/состава серверов/сетевых компонент/систем хранения);*
- *значительно повысить доступность системы (снизить плановые и незапланированные простои);*
- *повысить производительность системы;*
- *упростить определение storage-политик/процедур;*
- *улучшить качество storage-сервисов и др.*

Все множество решений storage-виртуализации SNIA классифицирует в соответствии с их функциональностью на трех уровнях (рис. 1). Первый уровень дает представление о том, "что виртуализируется", например:

- *"блоковая виртуализация" (block virtualization) дает возможность объединить в общий пул блоки данных от различных систем хранения;*
- *"виртуализация дисков" (disk virtualization) подразумевает, что цилиндры, головки, сектора HDD виртуализируются в логический блок адресов;*
- *ленточная виртуализация позволяет представить множество различных ленточных приводов и ленточных систем как одно устройство или, наоборот, одно ленточное устройство – как множество приводов;*
- *file-system-виртуализация дает возможность разделения файловой системы множеством ОС;*
- *file/record-виртуализация дает возможность "представления" файлов/записей на различных томах.*

Второй уровень дает представление о том, где физически реализуется механизм виртуализации: на хостах, на системах хранения или в storage-сети (в SAN). Третий уровень дает представление о том, как реализуется виртуализация. При "in-band"-виртуализации потоки управления и данных не разделяются, при "out-band" – разделяются.

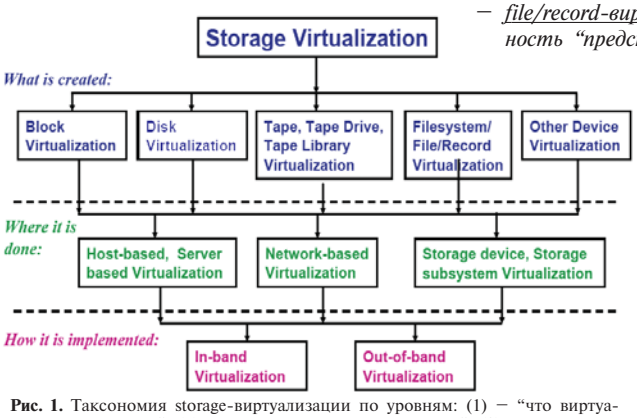

Рис. 1. Таксономия storage-виртуализации по уровням: (1) – "что виртуализируется"; (2) – "где виртуализация "работает"; (3) – "как реализуется".

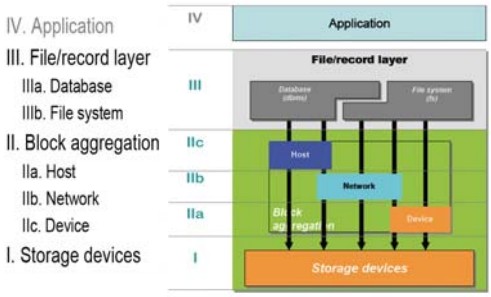

Рис. 2. SNIA классифицирует работу с данными на четырех уровнях (в соответствии с т.н. "SNIA Shared Storage Model"): уровне приложений, файловом уровне (файлов/записей), блоковом уровне и отдельно на уровне устройств хранения.

Другую классификацию решений виртуализации SNIA дает на основе уровней работы с данными (в соответствии с т.н. "SNIA Shared Storage Model"): приложений, файловом (файлов/записей), блоковом и отдельно на уровне устройств хранения (рис. 2). Уровни файловый и блоковый определяют способ доступа к файлам, компонента "block virtualization" (см. рис. 1.) эквивалентна "block aggregation"[1] (см. рис. 2), а уровни IIa, IIb, IIc эквивалентны второму уровню на рис. 1.

Тема решений по storage-виртуализации рассматривалась SN неоднократно[2], цель данной публикации – дать сравнительный обзор и позиционирование наиболее популярных решений данного класса, используемых в России на основе экспертизы, накопленной компанией ЛАНИТ за последние несколько лет их внедрения. Необходимо отметить, что часть решений по виртуализации, также распространенных в России,

1) В SNIA Shared Storage Model используется термин aggregation вместо virtualization.

2) "HP и ILM" *(http://www.storagenews.ru/27/InlGR-HP_27-05-last.pdf)*, "EMC Rainfinity – NAS-виртуализация" *(http://www.storagenews.ru/26/Raifinity_26.pdf)*, "Invista – сетевая виртуализация по EMC" *(http://www.storagenews.ru/24/Invista_24.pdf)*, "Виртуальное управление корпоративным хранилищем = SVC+SFS" *(http://www.storagenews.ru/18/ibm_18a.pdf)*, "MDS-коммутаторы: архитектурные особенности" *(http://www.storagenews.ru/18/cisco_mds_18.pdf)* и др.

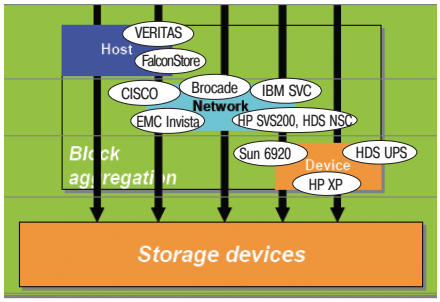

Рис. 3. Классификация решений блок-виртуализации по подуровням.

относящихся к файловой виртуализации (NAS-виртуализация) и виртуальным библиотекам (VTL-решения – Virtual Tape Library), в данный обзор не вошла.

Обзор решений

Основные решения, которые будут рассмотрены в данной статье, относятся к "block virtualization", а их непосредственная классификация по подуровням дана на рис. 3.

Виртуализация на уровне устройств хранения

1. HP XP/HDS USP/Sun 99xx

Данный тип виртуализации основывается на дисковом массиве и его ПО. Основным поставщиком подобных решений является компания HDS и ее OEM-партнеры – HP и Sun. Идея данной технологии (Hitachi TagmaStore Universal Storage Platform – USP) в том, чтобы предоставить потребителю универсальную платформу для консолидации всех ресурсов хранения, подключаемых к USP. Отличительной особенностью этой технологии является то, что, при подключении к USP какую-либо системы хранения на нее бесплатно распространяются возможности функционала USP (сервисы данных, общая консоль управления) и при этом обеспечивается сквозное управление всеми ресурсами. При подключении, вне зависимости от способа, внешние системы хранения представляются полностью интегрированными с внутренним массивом дисков USP в виде единого пула LUN. Следует учесть, что матрица совместимости оборудования у всех поставщиков разная и есть особенности в реализации.

Еще одна важная особенность USP – возможность деления системы хранения на логические разделы, с помощью которой можно распределять внешние и внутренние физические ресурсы хранения (включая порты, кэш и диски) по независимо управляемым "виртуальным машинам" (Private Virtual Storage Machine – PVSM, макс. кол. PVSM – 32). PVSM можно динамически изменять и физически она выглядит как собственная, отдельная система хранения. Данное решение поставляется в трех исполнениях – начальном, расширенном и high-performance.

Данное решение не относится к классу дешевых, поскольку для организации объединения дисковых массивов на основе одного дискового массива0 необходимо сначала приобрести этот дисковый массив вместе с дисками. Как правило, такие дисковые массивы относятся к high-end классу (HP XP 10000/12000, HDS USP, Sun 9980/9990) и сами по себе являются очень дорогими решениями. Начальная стоимость, например, HP XP10000 от 400k$, т.е. для создания консолидации всех существующих дисковых подсистем на базе такого массива нужно потратить минимум 400k$. Если использовать для виртуализации такое устройство, то для него рекомендуется добавить дополнительную кэш-память и дополнительные порты, что ведет к дополнительным затратам. К тому же софт для дискового массива лицензируется чаще всего по терабайтам.

2. Sun StoreEdge 6920

Sun StoreEdge 6920 – решение, призванное значительно расширить масштабируемость и функциональность модульных систем хранения и сократить различие между модульными (в основе – несколько контроллеров, число портов обычно 4–8) и монолитными (в основе – многопортовый коммутатор или кроссбар, число портов – 32–128 и более) системами хранения. По сути, это заявка Sun Microsystems предложить свою основу для построения целой серии решений для среднего сектора рынка и первая реализация Sun своего направления в SAN-виртуализации. 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 (т.н. платформа сервисов данных – Data Services Platform, DSP).

Первая реализация DSP – продукт SE6920 – состоит (рис. 4) из основных двух компонент: самого DSP и непосредственно дисковых блоков (в настоящее время поддерживаются дисковые полки Sun и некоторые продукты EMC, HP и др.). Спектр поддерживаемых в настоящее время операционных серверных платформ более широк: Solaris, Microsoft Windows, IBM-AIX, HP-UX, Red Hat Linux и др.

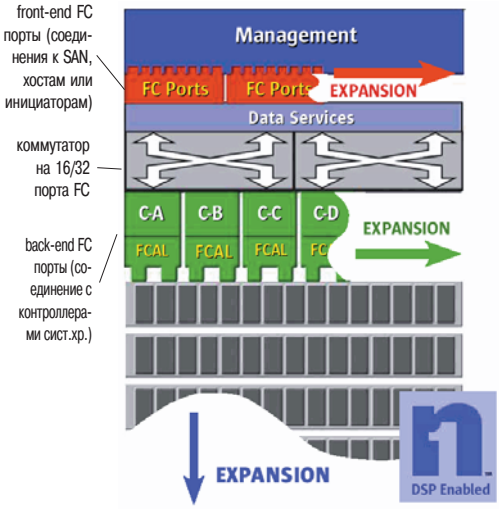

Рис. 4. Архитектура Sun StorEdge 6920.

В отличие от традиционных модульных систем, при построении решений на основе DSP интеллектуальная часть (сама DSP) полностью независима от самого массива дисков, за счет чего появляется возможность строить достаточно мощные по функциональности и производительности специализированные системы на базе недорогих дисков. При этом система имеет гораздо бóльшую масштабируемость, чем традиционные модульные системы. Безусловно, эта функциональность может быть достигнута на существующих системах, но это требует их интеграции с другими продуктами и значительных дополнительных вложений средств.

В состав DSP входят: масштабируемая процессорная система, коммутатор (доступен с 16 и 32 FC 2Gbit портами) и специализированное ПО, которое тоже может масштабироваться и развиваться.

SE6920 имеет два встроенных сервиса данных в своем составе. *Во-первых*, 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. *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.

В целом, сдерживающим фактором использования SE6920 является ограниченная матрица поддерживаемого оборудования и ориентация на решения Sun.

Сетевая виртуализация

Это наиболее обширный класс решений по виртуализации, представленных сегодня на рынке.

1. HP SVS200/HDS NSC55

Данное решение – HP SVS200/HDS NSC55 – совместная разработка HDS и HP, представляющее собой "урезанный" вариант (только "интеллектуальную" часть без дисковых полок) USP (или HP XP12000/10000).

По сути, SVS200/NSC55 это коммутационная матрица с механизмом виртуализации, взятая от XP/USP, но с сокращенным числом портов и, соответственно, с меньшей производительностью – в 4 раза. Основные параметры масштабируемости SVS200 представлены в табл. 1.

Табл. 1. Показатели масштабируемости SVS200/NSC55

	Min	Increment	Max
External Capacity		-	16 PB
Cache	4 GB	4 GB	64 GB
Shared Memory	1 GB	1 GB	6 GB
Host Ports	16	32	48
LDEVs	1	1	16,834

Основное назначение SVS200/NSC55 – предоставить базовую функциональность (консолидированное централизованное управление гетерогенной инфраструктурой хранения до 16 Пбайт) дисковых массивов уровня high-end "сред-

ним" потребителям (базовая комплектация поставляется от ~$120 тыс.). Поскольку SVS200 имеет пока ограниченную масштабируемость, для сбалансированности потока ввода-вывода от серверов с пропускной способностью систем хранения в предположении, что доступные 48 портов SVS200 делятся поровну между хостами и хранением (24 FC-порта – для подключения хостов), следует руководствоваться данными, приведенными в табл. 2, по максимально возможному числу дисковых систем, подключаемых к SVS200. При этом SVS200 не рекомендован для использования с приложениями, имеющими высокую случайную нагрузку ввода-вывода.

С помощью, например, HP StorageWorks XP Tiered Storage Manager данные в онлайновом режиме могут перемещаться между уровнями, сохраняя адрес LUN для хоста (и SVS200 LDEV). Дополняя такие решения функциональностью, поставляемой по OEM-соглашениям, можно строить "активные" многоуровневые хранилища.

SVS200 поддерживает удаленную репликацию (синхронную и асинхронную) с другим SVS200, а также многочисленные топологии соединения хостов-SVS200-массивов.

Табл. 2. Рекомендованное максимальное число подключаемых дисковых массивов к SVS200 (для конфигурации 48=24x24 портов) при разных типах нагрузки

Disk Array	Random Workload	Sequential Workload
HP EVA3000/EVA5000	3	6
HP EVA4000/6000/8000	1.5	6
HP XP	1	1
EMC CX300	3	6
EMC CX500	3	6
EMC CX600	1.5	6
EMC CX700	1.5	6
EMC Symmetrix	**1**	**1**
EMC DMX	1	1
IBM FastT	1.5	6

Поскольку базовое ПО, поставляемое с SVS200, является отнюдь не дешевым, то цена на SVS200 меньше, чем на XP 10000, только за счет отсутствия дисков, т.е. выгода покупки SVS 200 при организации виртуализации не столь существенна. Количество поддерживаемых дисковых массивов тоже существенно мало.

2. IBM SAN Volume Controller (SVC)

IBM SVC – одно из первых сетевых решений виртуализации на рынке, появившееся весной 2003 г., соответственно – одно из наиболее отработанных.

SVC было создано в соответствии с программой виртуализации IBM для продуктов хранения, в рамках которой разрабатывались два основных продукта, обеспечивающих ее реализацию – SAN Volume Controller (SVC) и SAN File System (SFS). SVC – базовый продукт, обеспечивающий доступ к данным на блоковом уровне в SAN-окружении, который может использоваться как самостоятельно, так и совместно с SFS. SFS – обеспечение корпоративного доступа к файлам в гетерогенной среде хостов. Главная идея, лежащая в основе семейства продуктов IBM TotalStorage Virtualization Family – перенесение части функциональности на базе принципов виртуализации с уровней сервера и систем хранения на SAN-уровень

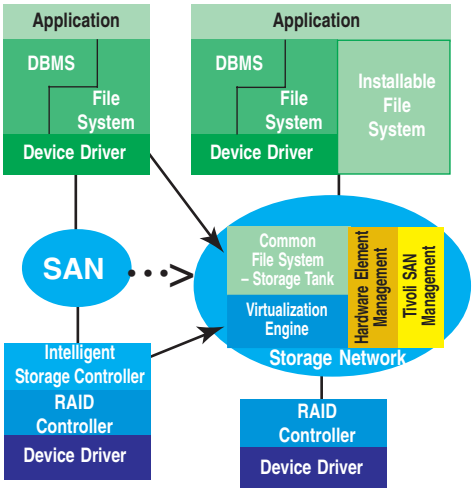

Рис. 5. Идея, лежащая в основе семейства продуктов IBM TotalStorage Virtualization Family – перенесение части функциональности с уровней сервера и систем хранения на SAN-уровень.

(рис. 5). SVC использует in-band модель реализации, SFS – out-of-band.

SVC представляет на общей схеме (см. рис. 5) компоненту Virtualization Engine и архитектурно относится к классу симметричных решений виртуализации. SVC реализован на базе IBM xSeries 336 серверов и умеет масштабироваться от 2-узлового кластера до 8-узлового. Внутри стоит урезанная версия Linux и ПО для управления SVC, написанное на Java, т.е. доступ для управления используется через веб-браузер. Виртуализация на базе SVC отличается тем, что не нужно для каждого дискового массива использовать свое ПО multipath, оно для всех одинаковое, кроме нескольких случаев: ОС OpenVMS, Vmware ESX и др.

Физически SVC подключается только к коммутатору, и поддержание всей логики управления хостами и системами хранения, а также потоками данных, осуществляется только через этот интерфейс (рис. 6).

SVC это по сути дополнительная прослойка на SAN-уровне между дисковыми массивами и серверами, позволяющая транслировать LUN с дискового массива и отдавать их серверу, не изменяя уровень RAID. С помощью SVC можно зеркалировать LUNы с одного массива на другой, делать клоны и т.д., не используя дорогое ПО дискового массива. Стоимость SVC в конфигурации из двух узлов – от 70k$.

Другие плюсы использования SVC:

– более гибкое решение (возможность динамического изменения размеров виртуального диска и динамического переноса данных с одного массива на другой без останова приложения);
– "добавление" производительности;
– лучшее управление (централизованное управление общим пулом томов на базе гетерогенных систем хранения);
– единое ПО управления множеством путей;
– хорошее соотношение цена/качество.

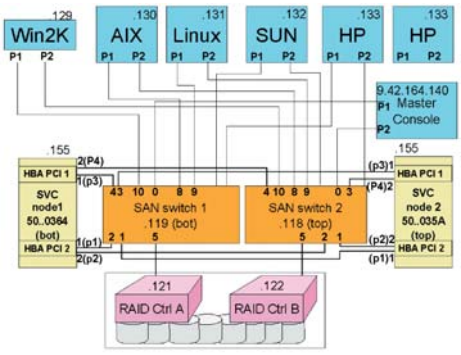

Рис. 6. Пример конфигурирования 2-узлового SVC для высокодоступной SAN.

В целом, это удачное, масштабируемое, недорогое решение с часто обновляющимся списком совместимости. Минусы использования SVC: архитектура на базе x86 серверов, небольшое количество портов.

3. Решения виртуализации на базе "интеллектуальных коммутаторов"

Это достаточно большой класс решений, развиваемый как глобальными вендорами, так и множеством поставщиков независимых решений. Все эти решения строятся на т.н. "интеллектуальных коммутаторах" (ИК). В России ИК поставляются от компаний Cisco и Brocade. Сам ИК представляет из себя модуль (или appliance), устанавливаемый в FC-коммутатор, поддерживающий его. У Brocade ИК называется SilkWorm Multiprotokol Router (ранее – SilkWorm Application Platform), у Cisco – MDS 9000 Storage Services Module.

Суть ИК – в том, что за счет встроенной процессорной мощности (на базе специализированного ASIC), доступной на каждом порту, ИК способен "на лету" "разбирать" сетевой трафик на управляющий поток (~5%) и поток данных (~95% объема), а также модифицировать фрейм данных, почти не внося задержки, основываясь на управляющих инструкциях, поступающих от управляющего (внешнего или внутреннего) устройства.

К преимуществам данного подхода виртуализации можно отнести: 1) не требуется установка никакого дополнительного ПО на хостах; 2) отсутствует необходимость замены традиционных HBA на специализированные, разработанные для реализации функций, виртуализации; 3) не требуется, чтобы весь I/O трафик был виртуализирован, например, при необходимости SilkWorm AP может быть сконфигурирован так, что он будет виден, как стандартный Brocade FC коммутатор, позволяя виртуализировать только часть SAN (или отдельные массивы и/или серверы), оставляя другую часть SAN нетронутой.

Среди компаний, продвигающих свои решения на основе технологии ИК: EMC, Symantec, IBM, Incipient, Xiotech, FalconStor, Topio, Kashya, Alacritus, Cloverleaf. Помимо виртуализации томов, на базе ИК активно развиваются решения поддержки различных сервисов данных – Continuous Data Protection, Snapshots, Replication, "Data Migration, Perf Mgmt (компании Xiotech, FalconStor, Topio, EMC, Alacritus, Cloverleaf и др.).

Решения на основе ИК могут строиться с внутренним и внешним управляющим процессором. Пример организации систем хранения на базе Cisco SSM с использованием технологии FAIS дан на рис. 7.

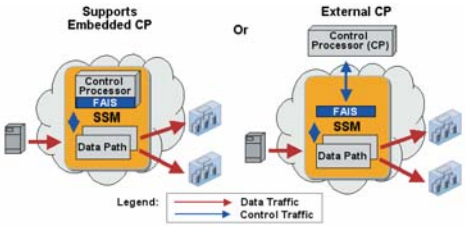

Рис. 7. Возможные варианты организации систем хранения на базе SSM с использованием технологии FAIS.

Гораздо бóльшее распространение получили решения по предоставлению сервисов данных на основе ИК. Пример такого решения – EMC Invista, являющаяся в контексте выше сказанного внешним управляющим процессором. Со стороны хостов (или front end), Invista "видится" точно так же, как дисковый массив. Хосты, которые могут работать с SCSI-дисками по FC, могут работать и с Invista без каких-либо дополнительных драйверов или другого специализированного ПО. Хосты соединяются с ИК с помощью FC через стандартные N_ports (соединение может осуществляться и с обычными портами).

Хост-ориентированная виртуализация

Данный класс решений "на местном уровне" в основном представлен двумя вендорами – Symantec (Veritas Storage Foundation и Veritas CommandCentral Storage) и FalconStor.

Вследствие того, что это программные решения, за счет них нельзя повысить производительность уже существующего оборудования. Из плюсов можно отметить большую матрицу совместимости, частое использование в кластерных решениях.

Общее сравнение различных решений виртуализации хранения приведено в табл. 3.

Заключение

По мнению специалистов ЛАНИТ, интерес к решениям виртуализации в ближайшей перспективе будет только возрастать. Это вызвано объективными причинами и, прежде всего, стремлением упростить администрирование систем хранения во все возрастающих по сложности IT-инфраструктурах. И проблемы управления корпоративными хранилищами, со всей остротой проявляющиеся в последнее время при организации высокоэффективных информационных систем, постепенно становятся приоритетными не только во всем мире, но и в России. Они в полной мере показывают, что затраты на управление (вместе с потерями от запланированных простоев и недоиспользованием ресурсов) могут в самый короткий срок превзойти непосредственно первоначальную стоимость самих технических средств. Эффективность управления и снижение затрат на управление становятся ключевыми задачами большинства проектов при построении центров данных. Ориентируясь в своем развитии на тенденции рынка, ЛАНИТ сегодня готов осуществить внедрение любого из всех вышеперечисленных решений. Огромный опыт ЛАНИТ, высокая квалификация специалистов и налаженные партнерские отношения с ведущими производителями на IT-рынке позволяют гарантировать высокое качество работы.

Павел Ерофеев,
инженер системно-технического отдела, ДСИ ЛАНИТ

Табл. 3. Сравнительные характеристики решений виртуализации хранения

Наименование	Device based			Storage network based				Host based
	XP 12000	HDS USP	Sun 6920	SVS 200	SVC	Cisco MDS	EMC Invista	Veritas
Пропускная способность по портам	128x 4GB/s	128x 4GB/s	28x 2GB/s	48x 2GB/s	16x 4GB/s	>128x 4GB/s	16 2Gb/s	Нельзя сказать т.к. это софт
Количество IOPS, тыс.	120 (datasheet)	120 (datasheet)	19 (SPC-1)	16 (datasheet)	155 (SPC-1)	130 (datasheet)	80	Нельзя повысить IOPS существующего оборудования
Подключение массивов разных производителей	+/-	+/-	+/-	+/-	+	+	+	+
Поддержка опций (mirroring, snap clone и т.д.) между разными массивами	+	+	+/-	+	+	+	+	+
Количество кэш памяти	256GB	256GB	28GB	32GB	64GB	-	-	-
Ограничения	Нужно покупать диски и ПО для этого массива	Нужно покупать диски и ПО для этого массива	Маленькая матрица поддерживаемого оборудования	Маленькая матрица поддерживаемого оборудования	Нет ограничений	Нужно покупать director	Нужны интеллектуальные коммутаторы	Т.к. это ПО, то нужно смотреть совместимость со всеми устройствами которые будут использоваться в процессе виртуализ.
Архитектура	crossbar	crossbar	коммутатор	crossbar	cluster	-	cluster	-
Приблизительная цена решения в тыс.	<300$	<300$	<150$	<200$	<70$	<200$	<150$	<20$