RAID-контроллер: вчера, сегодня, завтра

Обзор развития RAID-контроллеров Adaptec.

Дмитрий Зотов – инженер компании PMC-Sierra, подразделение Adaptec by PMC.

Введение

В данной публикации рассматриваются характерные особенности, которые появились и продолжают появляться в области современных RAID-контроллеров, а также в сфере их применения.

Good bye RAID5. Новые уровни RAID

Несколько лет назад, в противовес "умиранию" решений, базирующихся на RAID5 (ниже даны объяснения, почему RAID5 теряет былую популярность), были предприняты меры (решения), направленные на то, чтобы сделать этот процесс менее болезненным (см. табл. 1 с комментариями относительно каждого решения).

Как некое усиление решения RAID5 (без HOT SPARE диска, далее – HS) существует решение RAID5 + HOT SPARE диск. Но, как видно из табл. 1, некоторые новые решения предлагают лучшую альтернативу, что делает RAID5 + HS заведомо менее привлекательным решением.

Что мы наблюдаем сегодня? Реальную потерю популярности RAID5 и его усиленной версии RAID5 + HS. Пока наиболее популярны замены RAID5EE, RAID 6. В новых контроллерах 7-й серии уже нет поддержки RAID5EE. Это отражает увеличение производительности современного контроллера. RAID6 просчитывается уже быстрее. Решение достаточно известное и популярное, и в цепочке между RAID5 + HS ---- RAID5EE ----- RAID6 уже нет необходимости в промежуточном звене.

Табл. 1. Характеристики новых уровней RAID.

Решение	Особенность при замене таким решением тома RAID5	Комментарий
RAID 1E и RAID 10	RAID 1E из трех дисков – 1,5 диска полезной емкости, и в плане емкости он немного проигрывает RAID5 из 3-х дисков – 2 диска полезной емкости.	**Само по себе решение 1E представляет из себя единственную возможную альтернативу RAID5 из 3-х дисков.** На старых стеках без поддержки 1E единственный подходящий уровень был только RAID5. Наибольший спрос на решение 1E находят как раз тома из 3-х дисков, хотя можно создавать тома и из 4-х, и из 5-ти, и больше. Но, на практике, 5 дисков в томе RAID 1E уже довольно большая редкость.
	RAID10 был и остается некоторой альтернативой RAID5 на четном количестве дисков, предлагающей более высокую надежность и производительность, но меньшую полезную емкость.	
RAID 5EE	Представляет из себя аналог RAID5+HS с более низким временем свертки в RAID5 при потере одного диска и с более высокой производительностью по сравнению с RAID5, которому назначен один HS диск.	**Является неплохой заменой любого тома типа RAID5 + HS с любым количеством дисков.** 5EE всегда более лучшее решение: такое же по цене, такое же по емкости, но лучшее по производительности и надежности!
RAID6	Представляет из себя аналог RAID5+HS с максимально улучшенной надежностью (**можно потерять два диска в любой момент времени**), но до сегодняшнего дня этот уровень RAID проигрывал в производительности решению RAID5EE в пределах 0-20% (в зависимости от шаблона трафика).	**Является неплохой заменой любого тома типа RAID5 + HS с любым количеством дисков.** Немного хуже 5EE по производительности, но лучше по надежности.
		Сравните надежность: RAID5EE не может потерять два диска в любой момент времени в отличие от RAID6, сначала теряется один диск, том сворачивается (в стеке Adaptec процесс называется COMPACTING) в RAID5, и только после этого может потерять еще один диск, что дает RAID5 degraded.
Гибридные тома, типа RAID1, 10	Гибридный RAID1,10.	Здесь пока немного сложно уравнять решения по цене, но, если предположить, что 2 SSD и 2 HDD для гибридного 10 стоят столько же, сколько RAID5 из 10-15 дисков, то мы получаем одинаковую цену, более высокую надежность, и примерно одинаковую производительность. Проблема в емкости (набрать емкость RAID5 на SSD дисках будет стоить дорого, ждем понижения цены на SSD решения). Если на RAID5 + HS будет возложена и запись, то гибридные тома применять будет сложно. **Применяются только для чтения**.
	Например, для 4 дисков (2 SSD диска и 2 HDD диска) может дать производительность RAID5 из 10-15 дисков и больше.	
SSD RAID1, RAID10	Мы вполне вправе ожидать, что при работе тома 1,1E,10 из SSD для достижения нужной производительности нам может понадобиться на порядок меньше дисков, чем на RAID5 + HS HDD.	Пока **только теоретически** может быть заменой для любого тома RAID5 или RAID5 + HS, но, в действительности, цена для SSD тома может быть выше при одинаковой производительности с HDD или гибридным томом, а надежность ниже при одинаковой производительности по сравнению с HDD или гибридным томом.
		В этом хороший задел под будущее, когда цена на SSD диски будет уменьшаться, а их надежность и емкость расти.
SSD кэширование	И чтение и запись кэшируется на SSD (возможность и чтения и записи существует только на 7Q серии контроллеров Adaptec, 6Q и 5Q только чтение).	Принцип работы можно посмотреть ниже. Уже на сегодняшний день ценовой расчет проекта показывает, что это вполне равное (и даже **более дешевое**) решение для RAID5 + HS. И этим можно пользоваться.
		Для SSD гибридных томов и томов, использующих SSD кэширование проявляется еще одно преимущество – меньшее количество дисков.
		Такая же или меньшая цена решения. Такая же производительность. Такая же или выше надежность. Значительно меньше дисков. Для площадок это означает меньше юнитов в серверах и стойках и меньше затрат на питание дисков и охлаждение. **Возможность эффективной замены RAID5 + HS для случайного трафика при наличии шаблона "горячих данных "(как правило, чем больше пользователей, тем больше вероятность появления шаблона ГОРЯЧИХ ДАННЫХ).**
BAD STRIPE поддержка	Некая защита тома и доступа к тому, если в работу запущен RAID5 или RAID5 без HS с низким показателем надежности.	Решение представляет из себя некую защиту в случае неправильного проектирования и создания/администрирования тома. Раньше, в случае неисправимого бэд блока в страйпе RAID тома, доступ к тому останавливался, поскольку данные считались искаженными. Сейчас доступ к тому остается, а отдельный страйп, в котором произошла проблема, помечается как искаженный.

Кроме того, популярность гибридных решений и зарождение чистых SSD-решений также вносит свой вклад в исчезновение RAID5 + HS. Как некий вывод – табл. 2.

Причины того, отчего RAID5 так сильно потерял в базовой надежности, представлены в табл. 3.

Цена за универсальность. Бойся десктопного диска, проблемы приносящего

Надежность десктопного диска выражается вероятностью потери диска или его части. Отметим, что, прежде, когда мы брали RAID5 на SCSI-дисках, создавая на них RAID5, при выходе из строя одного диска (при переходе тома в состояние DEGRADE), общая надежность такого виртуального диска (под виртуальным диском мы понимаем том RAID5 в режиме degrade – один диск потерян) была сравнима с надежностью одного десктопного диска, поскольку надежность одного SCSI-диска на порядок выше надежности десктопного диска.

Здесь мы говорим, что том в режиме RAID5 degrade боится появления на любом из оставшихся дисков хотя бы одного бэд-блока. И далее сравниваем вероятность появления бэд блока на одном диске SCSI серверного класса и на одном диске SATA десктопного класса.

Чтобы увидеть истинную разницу, надо правильно понимать значение MTBF для десктопных и серверных дисков. Их заявленная разница в 3–4 раза, как это выглядит на сайте производителя, заметно увеличится при использовании десктопных дисков в серверном режиме, т.е. 24 часа в сутки, 7 дней в неделю.

В наше время SAS контроллер легко поддерживает десктопные диски технологи SATA, причем даже такие (некоторая доля от десктопного класса), в которых один бэд-блок на диске заставляет контроллер потерять такой диск (*http://en.wikipedia.org/wiki/Error_recovery_control*).

Сравним вероятность перехода тома в RAID5 degrade на контроллере SCSI и на контроллере SAS, если предположить, что на последнем используются десктопные диски. Разница будет, буквально, в разы. Дескотпный диск на порядок ниже по надежности, следовательно, на порядок выше будет число томов, переходящих в DEGRADE режим. При этом ситуация, когда один бэд-блок убирает диск с контроллера (на SCSI-дисках такое просто невозможно) делает это отличие еще более существенным.

Если предположить, что при старом подходе на SCSI-контроллере типичный проект это – 4–10 дисков, то на SAS-контроллерах типичный проект – 4–36 дисков, а для производительных серверов разница в количестве дисков начинает достигать порядка 4 против 40 или 8 против 80. Что увеличивает вероятность перехода в degrade состояние еще на порядок.

Ну, а теперь – о емкости. Очевидно, чем выше емкость, тем выше вероятность бэд-блока. Сравниваем 400 ГБ HDD на SCSI RAID-контроллере и 4ТБ HDD на SAS RAID-контроллере. Увеличение еще на порядок. В табл. 4 представлены наихудшие сценарии.

При этом область влиянии всех трех факторов сужают с помощью специальных ограничений, иначе вероятность настолько большая, что ничто не может помочь. Большинство производителей просто запрещают использование desktop класса дисков на RAID-контроллерах. Но это правило слишком прямолинейно. Было бы крайне нелогично запретить RAID0 из desktop дисков для создания бэкап-тома. В этом нет ничего плохого.

Более профессиональный подход – не использовать desktop класс дисков для пользовательских данных (где данные имеют высокую ценность) на RAID5 томах при емкости дисков больше 150–200 ГБ. С учетом того, что на рынке практически нет desktop моделей такой емкости, правило звучит на практике так: не использовать desktop класса диски в RAID5 для пользовательских томов.

Для nearline-дисков граница лежит в районе 700ГБ – 1 ТБ. Диски большей емкости крайне нежелательно использовать в пользовательских томах RAID5. Простая система управления, которая контролирует переход тома в degrade и замечает потерю диска, устраняет необходимость в данном правиле. Но на практике приводит к исчезновению RAID5-решения.

Сама по себе система управления ничего не изменит, нужен запасной диск, который надо использовать для починки тома. Если в корзине есть свободный слот, такой диск лучше всего объявить как HOT SPARE для RAID5. Ура, правило, приведенное выше, уже можно нарушать. Но RAID5 + HS не имеет смысла, если контроллер поддерживает RAID5EE или RAID6. Получается, что RAID5 просто исчез из нашего проекта, если следовать данным правилам. Остается только RAID5 на SAS серверных и SSD-дисках. Отметим, что использование SSD-дисков – решение пока не очень популярное в силу ряда причин.

Табл. 2. Особенности уровней RAID с точки зрения предъявляемых требований.

Решение	Вчера	Сегодня
RAID5	Приемлемый уровень надежности.	Неприемлемый уровень надежности
RAID5 + HS	Норма проектирования.	Почти в 100% случаев, просто не имеет смысла, т.к. том может быть заменен на RAID5EE или RAID6, что значительно повышает надежность и производительность, не изменяя цены решения и емкости.
RAID1, 1E,10	RAID 1 как норма проектирования для служебных томов. RAID 10 - НОРМА ПРОЕКТИРОВАНИЯ ПЛЮС (под словом ПЛЮС понимаем дополнительную надежность). RAID 1E – такие решения не поддерживались.	1,1E Норма проектирования для небольшого количества дисков (2-5) или служебных томов. RAID10 как высоконадежное и высокопроизводительное решение, но по реализации значительно дороже RAID5.
RAID6	Нет поддержки.	Норма проектирования.
RAID6 + HS	Нет поддержки.	НОРМА ПРОЕКТИРОВАНИЯ ПЛЮС.
Гибридные тома, SSD кэширование	Нет поддержки.	Норма проектирования для ряда шаблонов трафика.
В чистом виде SSD тома без добавления HDD дисков	Нет поддержки.	Задел под будущее, начало использования.

Табл. 3. Особенности RAID5.

	Вчера:	Сегодня:
	RAID 5 на SCSI контроллере	RAID5 на SAS контроллере
Поддержка менее надежных дисков (класса Desktop, Nearline)	Нет.	Да.
	Поддерживаются только высший класс серверных дисков SCSI с высокой степенью базовой надежности.	Поддерживаются все классы.
Количество дисков на канале для достижения максимума производительности.	4 (15000 rpm) -6 (10000rpm) дисков на одном канале SCSI.	8 (15000 rpm) дисков на один канал SAS2 (6 GB/s), как результат для контроллера 72405 (24 порта SAS2 в систему контроллера) 8 x 24 = **192 диска.**
	Для 2-х канального контроллера и дисков 15000rpm – 2 x (4-6) = **8 -12 дисков.**	
Максимальная емкость диска	400ГБ	4 ТБ
		Например, в листах совместимости для 7 серии контроллеров Adaptec видно наличие 4ТВ дисков.
		www.adaptec.com/compatibility

Табл. 4. Особенности построения RAID5 на SCSI- и SAS-контроллерах.

Фактор, влияющий на вероятность перехода в Degrade для RAID5 из-за поломки (в отдельных случаях! Бэд блока) одного диска.	Вчера	Сегодня	Разница в вероятности
	SCSI RAID-контроллер	SAS RAID-контроллер	
Диски desktop класса	Неприменимо, только диски серверного класса	Да, причем в ряде случаев один бэд блок – приводит к потере диска контроллером.	На порядок и больше
Количество дисков	4-8	4-32	На порядок и больше. Вероятность увеличивается нелинейно от количества дисков!
Емкость дисков	400 ГБ	4ТБ	На порядок и больше. Вероятность увеличивается нелинейно от емкости!
Все факторы вместе	Вполне надежное решение	Крайне ненадежное решение	Мы имеем колоссальную разницу!

Как мы видим, пользовательские тома можно безболезненно заменить на 5EE, RAID6 или на гибридные, или SSD RAID1 и RAID10 и избежать использования RAID5. Более того, многие желаемые конфигурации на RAID5 будут либо запрещены, либо невозможны из-за ограничений контроллера. Отмечу, что ограничения и запреты распространяются и на RAID5 тома в составе RAID50! Пока для RAID5 остается только область бэкап-томов и томов, где хранятся данные низкой или нулевой стоимости, что ставит RAID5 вровень с таким решением, как RAID0.

С появлением SSD-дисков с высокими базовыми свойствами (на уровне физического диска) по надежности RAID5 может обрести второе рождение. RAID5 может находить спрос в SSD RAID-томах (только SSD-диски) и в виде специальных томов SSD RAID5 (только SSD диски) для создания надежного и скоростного пула SSD-кэширования. Такой пул кэширования используется для операций чтения (потребность в производительности – кэш пул работает очень быстро) или записи (потребность в надежности – потеря диска не приводит к потере данных в кэше).

Важно: здесь, в случае записи, при переходе такого тома в режим degrade, достаточно выключить кэш на запись, что происходит почти мгновенно. Сравните это с процессом встраивания нового диска размером 4ТБ в RAID5, который является обычным томом, а не областью кэширования, когда такой процесс может идти несколько дней, и выключать здесь ничего нельзя.

Новый элемент хранения – SSD-диск: готов ли он уже полностью заменить HDD в серверной области?

В последнее время находится много желающих создавать на контроллере "чистые" SSD RAID-тома (состоящие только из SSD-дисков). Если в предыдущем разделе мы выяснили, пора ли полностью "хоронить" RAID5, то здесь мы обсудим похожую тему: пора ли "хоронить" HDD диски? Выясняется, что совсем еще не пора.

Кажется, и эта иллюзия очень сильна, что SSD-диски – это что-то сверхсерверное, доминирующее над SAS HDD 15000 rpm. Ну как же, скорость чтения и записи, порой, на порядки больше (и в МБ в секунду и в количестве обслуживаемых операций ввода-вывода в секунду), но, к сожалению, на сегодняшний день только производительности мало, чтобы занять нишу "сверхсерверного продукта".

В рассмотрение приходится брать комплексный параметр, некую "серверность решения" (куда включаются производительность, емкость, цена, надежность и более специфичные факторы, например, такие как среднее время жизни, гарантия производителя, зависимость базовых характеристик от времени, устойчивость к температурным воздействиям, внешним и внутренним вибрациям, стартовые токи и т.д.).

В ряде случаев, например, внешние вибрации или потребляемая мощность в рабочем режиме, или отсутствие механических узлов наделяют SSD-решения суперсвойствами. Картина в целом представлена в табл. 5.

Табл. 5. Особенности разных классов дисков.

Фактор	Десктопный HDD	Nearline HDD (по свойствам нечто среднее между десктопным и серверным диском)	Серверный HDD	Серверный SSD
Базовая надежность	Низкая	Средняя	Высокая	Высокая, но сравнимая с серверным HDD. Для некоторых классов SSD ниже, чем у HDD.
Базовая производительность	Низкая	Средняя	Высокая	Очень высокая
Цена	Низкая	Средняя	Высокая	Очень высокая
Емкость	Большая	Большая	Низкая	Низкая, средняя
Защита от внешних вибраций	Плохая	Средняя	Хорошая	Очень хорошая. Нет влияния внешних вибраций.
Оптимизация алгоритмов входных очередей со стороны контроллера	Высокая	Высокая	Высокая	В разработке
Оптимизация алгоритмов кэширования со стороны контроллера	Высокая	Высокая	Высокая	В разработке
Оптимизация работы с ОС, команда TRIM или ее аналоги и ее поддержка RAID контроллером	-	-	-	В разработке

Первая ошибка, которую совершают многие проектировщики, – неумение свести к общему знаменателю, например, по цене, решение на HDD и SSD, а также выявление разницы в производительности и надежности. Если это делать умело, то получается, что HDD-тома и гибридные тома еще не утратили своих основных позиций в области проектирования систем хранения.

Вторая – неумение учитывать всех факторов вцелом или факторов, специфичных для проекта. Для некоторых проектов, например, сверхпроизводительные системы хранения высокой плотности с небольшими требованиями по емкости или системы, работающие в режиме внешних вибраций, время SSD-томов пришло. В других проектах SSD-тома уступают HDD или гибридным решениям.

С учетом того, что свойства SSD-дисков улучшаются, а цена уменьшается, операционные системы и RAID-контроллеры получают более совершенную поддержку SSD-решений. Ожидается, что в 2013–2014 годах будет значительный рост SSD-решений в серверной области. 7-я серия контроллеров Adaptec – это первая серия контроллеров, широко поддерживающих SSD-решения. На рубеже 2013–2014 годов индустрия получит RAID-контроллеры, полностью поддерживающие SSD-диски и тома на их основе.

Том на SSD и том на HDD. Совместное использование SSD и HDD – гибридные тома

Как уже упоминалось, в чистом виде SSD-тома пока не могут составить мощную конкуренцию HDD-решениям и за-

Табл. 6. Возможные типы гибридных томов.

Название	Краткое описание	Поддержка на RAID контроллерах ADAPTEC
Hybrid RAID1 (10)	В составе зеркала один диск SSD другой - HDD. Режим работы такого зеркала немного изменен по сравнению с обычным, созданным на HDD RAID1 (такие гибриные зеркала могут быть и в составе 10 томов).	Все контроллеры семейства 2,5,6,6E,6Q, 7,7E,7Q.
Hybrid RAID4 (40)	Все диски HDD, диск, который хранит parity (контрольные суммы – SSD), возможно использование комплексных RAID40 томов	Не поддерживается текущими моделями RAID контроллеров Adaptec.
SSD кэширование для HDD томов (чтение)	Особая работа между двумя томами. На томе HDD специальным алгоритмом помечаются данные, которые наиболее часто запрашиваются (ГОРЯЧИЕ ДАННЫЕ), и эти данные копируются на SSD том. При этом типы томов HDD и SSD буквально произвольные (например, RAID6 HDD кэшируется через RAID5 SSD, или SIMPLE VOLUME HDD кэшируется через SIMPLE VOLUME SSD).	Поддерживается только RAID контроллерами с индексами Q (5Q, 6Q, 7Q), при этом только для 7Q SSD кэш может быть и RAID томом (RAID0,1,1E, 5).
		Для более ранних моделей SSD том был вырожденным или SIMPLE VOLUME или CHAIN VOLUME. Здесь важно отметить, что, поскольку мы имеем копирование ГОРЯЧИХ ДАННЫХ, то надежность тому SSD особо не нужна.
SSD кэширование для HDD томов (запись)	Данные сначала записываются на SSD том (здесь избыточноть тома нужна для защиты данных кэша на запись), а после могут быть перемещены на HDD том.	Поддерживается только моделью 7Q (7805Q и 71605Q).
HDD-SSD layering (Tiring)	Еще более сложная схема, когда ГОРЯЧИЕ ДАННЫЕ не копируются при чтении, а перемещаются на другой, более быстрый, том. Применительно к Гибридным томам данные с HDD RAID тома (уровень один) перемещаются на SSD том (уровень два). Важно: поскольку SSD (и HDD) диски отличаются базовой производительностью, таких уровней может быть и больше, чем два. Самые невостребованные данные лежат на нижнем уровне, более востребованные - на средних, и самые востребованные на верхнем. Верхний уровень обладает наивысшей производительностью.	Не поддерживаются RAID контроллерами Adaptec.
SSD тома в составе шаблонных томов	Другие реализации комлексных томов, которые состоят из томов SSD и HDD, когда под определенные шаблоны трафика, шаблоны времени, шаблоны типов данных и пр. выделяются отдельные тома.	Пока такие тома не находят широкого применения в современных системах хранения, но их теоретические и практические возможности уже изучаются. Так что, использование таких решений дело ближайшего будущего.

метно отобрать популярность или, что еще хуже, полностью исключить использование HDD в области хранения данных для серверных систем. В то же время, наблюдается некий интересный симбиоз HDD и SSD, называемый гибридными томами.

В широком смысле гибридный том – это том, в составе которого работают HDD и SSD. И в этом смысле SSD-кэширование – это тоже частный вид гибридного тома.

Почти все семейство контроллеров Adaptec, начиная с 5-й серии SAS1-контроллеров поддерживает частный случай гибридного тома, не очень удачно названный Hybrid Volume (более правильным названием было бы RAID1 гибридный том). Возможные типы гибридных томов даны в табл. 6.

Давайте посмотрим внимательно на базовый механизм гибридного тома, вложенный в каждую модель SAS-контроллера Adaptec – гибридные RAID1. В чем их особенность?

"Зеркало" или RAID1 состоит из двух дисков (диск мастер /master/ и диск слейв /slave/), в обычной (негибридной) реализации запись идет на оба диска одновременно, а чтение – с мастера, пока количество запросов ВВОДА-ВЫВОДА не превысит некоторого порогового значения. Далее, события, запросы на чтение распределяются между мастер- и слейв-дисками.

При гибридной реализации важно убедиться, что мастером стал SSD-диск – на практике для этого требуется выбрать его первым и убедиться, что в составе тома он получил номер (0), а HDD, соответственно, номер (1). Иначе вы не получите гибридного RAID1 на RAID-контроллере Adaptec. Режим работы крайне прост: чтение ВСЕГДА будет с SSD, а запись – на оба диска. Вывод: с точки зрения производительности такой том на чтение ведет себя, как SSD, а на запись – как HDD. Такие же гибридные RAID1 могут попадать и в состав RAID10. Все SAS-контроллеры Adaptec поддерживают гибридные RAID1, 10 – hybrid.

Для проектировщика важно уметь смотреть на это и с точки зрения физического диска, и с точки зрения логического диска. Такой подход, например, позволяет, имея один SATA-диск большой емкости и три SAS SSD высокой производительности, получить 3 гибридных зеркала RAID1 для приложений на сервере, работающих на чтение. И для этого не надо покупать что-то специальное – любой контроллер с такой задачей справится (рис. 1).

Power Management

Совсем недавно в современных контроллерах появилась возможность управлять питанием дисков, речь, в первую очередь, идет о HDD. Как изделия с механическими узлами HDD способны давать больше всего экономии при управлении электропитанием. Более тонкие решения, такие как управление питанием процессора контроллера, его "засыпание" и пр. пока не практикуются.

Природа power management на современном контроллере Adaptec крайне проста: если система хранения не используется (а в качестве объекта, на котором настраиваются функции power management, выбирается RAID-том), то диски в составе системы хранения сначала переходят в режим более низкого энергопотребления, если и дальше система не используется, то диски совсем выключаются.

Промежуточный режим более низкого энергопотребления, называемый standby, поддерживается не всеми дисками. Если диск режим поддерживает после некоторого заданного промежутка времени в случае неактивности системы, диск в составе тома (например, в составе RAID5, на котором поднята схема power management) перейдет в режим standby. Если режим не поддерживается, то ничего не произойдет: те диски, которые поддерживают этот режим, перейдут в него, те, которые не поддерживают – останутся в рабочем режиме.

Все современные диски, как правило, поддерживают режим standby. Обычно

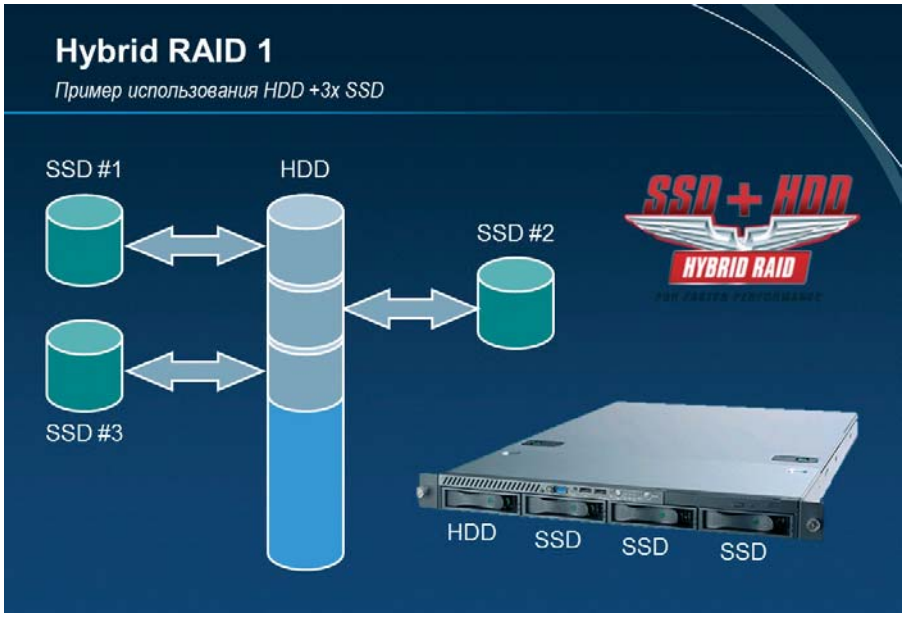

Рис. 1. Пример построения гибридного RAID1 на базе жестких дисков и трех SSD-дисков.

standby – это снижение скорости вращения в 2 раза, при этом потребляемая мощность уменьшается на 40–60%.

Следующий таймер настраивается для режима полного выключения питания – power off.

Приведем некоторые особенности power management механизмов, которые надо учитывать для профессионального проектирования и использования:

- в схеме есть не только прямые факторы экономии, но и косвенные;
- экономия достигается не только на потреблении электроэнергии, когда диски выключаются или снижают свою скорость, но и на уменьшении затрат на охлаждение – система охлаждения сервера уменьшает поток воздуха на диски или совсем останавливает его;
- выключенный и не нагретый HDD статистически имеет более высокий срок жизни, чем диск, который раскручен и, как следствие, более горячий;
- контроллер имеет внутреннюю процедуру, которую можно настроить для того, чтобы время от времени обесточенные диски включить и проверить не опасным для хранящихся данных способом, а затем выключить;
- при работе на включение после длительного отсутствия запросов, когда диски обесточены, отдельный запрос на запись, который приходит в систему, не приводит к раскручиванию дисков. Секрет в том, что он переносится в кэш. Такой подход с нужной степенью автоматизации не позволяет системе постоянно останавливать и снова раскручивать диски;
- если на томе настроены служебные функции, типа проверки целостности RAID-тома (background consistency check), отключение питания дисков или переход в standby происходить не будет, поскольку есть служебная задача, которая работает с дисками.

Несколько примеров использования power management.

Пример 1. В случае если есть основной сервер и сервер BACKUP, на который данные перемещаются раз в месяц, то на RAID-томе BACKUP-сервера использование POWER MANAGEMENT крайне эффективно. Если предположить, что полное копирование данных идет 2 дня, то оставшиеся дни месяца диски будут выключены, и не будет тратиться энергия на их вращение и охлаждение, кроме того, они будут меньше изнашиваться.

Пример 2. У вашей компании есть сервер на площадке в другом городе. Вы создаете тома по мере заполнения сервера приложениями и пользователями. У сервера есть плата удаленного управления. Если вы будете хранить диски у себя в офисе или на складе на удаленной площадке, добавление каждого нового диска потребует поездки к серверу, чтобы физически добавить диск (представьте, вы – в Москве, а сервер – во Владивостоке). Диск можно поставить в корзину (вы используете корзину сервера как склад, максимально набив ее дисками), но не вводить его в состав какого-либо тома. Тогда вы

всегда можете добавить его в том, либо получая доступ к BIOS-утилите RAID-контроллера удаленно, либо получая терминальный доступ в ОС и используя Adaptec Storage Manager. Но есть один недостаток: такие, не занятые никакой работой диски, будут крутиться, нагреваться и изнашиваться. Здесь на помощь приходит POWER MANAGEMENT. Вы берете такие диски и удаленно заводите их в любой том – это, пожалуй, единственный случай, когда совсем неважно, какой том вы используете, можно и RAID0. Далее активизируете для этого тома функцию POWER MANAGEMENT. Через несколько минут диски в составе этого тома остановятся. Какое красивое решение мы получили – том хранения дисков, а не данных.

Если диски будут нужны, получаете доступ к любой утилите управления, разбираете том и добавляете эти диски в рабочие тома. Все это можно сделать удаленно. Диски не изнашиваются, не надо "навещать" сервер.

Приближаемся к миллиону IOPSов

Темпы роста производительности ядра контроллера заметно ускоряются. Сравним: 6-я серия контроллеров Adaptec – 50 kiops, 7-я серия контроллеров –

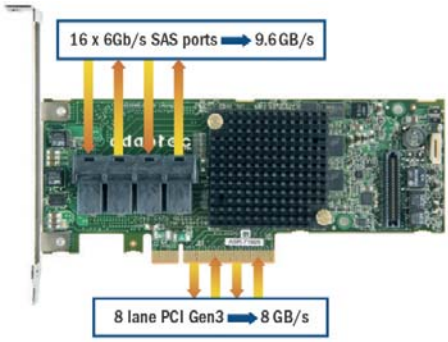

Рис. 2. Следующее поколение RAID-контроллеров будет поддерживать около миллиона IOPs.

450 kiops. Без сомнения, следующее поколение принесет поддержку, равную почти миллиону обращений к системе хранения в секунду. Понятно, что потребность в такого рода показателе обусловлена появлением и ожидаемым развитием SSD-дисков. Ну, а с точки зрения приложений – главный потребитель такой производительности это – виртуализация. На одном железном сервере фактически могут находиться все серверы большой компании (рис. 2).

Прощай, батарейка! Привет, суперконденсатор!

В этой главе мы расскажем о методах защиты кэша контроллера. Это методы кэширования, которые используют оперативную память контроллера. Почему появляется необходимость в защите?

Главная проблема – в кэшировании записи. Данные переносятся в оперативную память, затем – на том. Если по какой-либо причине пропадает питание, содержимое памяти теряется, что приводит к потере данных в системе хранения, а это недопустимо.

Рис. 3. Adaptec Flash Module (AFM) состоит из двух основных компонентов: суперконденсатора и флэш-памяти.

Большим заблуждением считается, что защиту можно обеспечить с помощью UPS. В случае с UPS любые неисправности на внутренних шинах питания приведут к потере данных.

Самый надежный способ защитить потерю питания напрямую на оперативной памяти контроллера. Эта роль раньше традиционно была закреплена за так называемыми батарейными модулями, или на жаргоне администраторов, – батареями, или, по-научному, BBU (battery back-up modules). Их задачей было – зарядиться и, в случае проблемы, питать как можно дольше оперативную память, сохраняя данные в ней. Но такой подход обладал рядом недостатков (табл. 7), и исправить эти недостатки был призван новый подход под названием "флэш-модуль" или по научном FM – flash module (для контроллеров Adaptec он имеет традиционное сокращение AFM – Adaptec Flash Module). Два основных компонента такого решения – это суперконденсатор и флэш-память. Суперконденсатор быстро заряжается на момент загрузки сервера, когда оперативная память контроллера еще не используется в рабочем режиме. И после, в случае падения питания, энергии суперконденсатора хватает на перенос данных из оперативной памяти на флэш-память, где данные могут храниться до нескольких лет, ожидая восстановления схем питания сервера (рис. 3).

SAS-контроллер и HBA в одном флаконе

На 7-й серии контроллеров мы видим добавление новых возможностей: контроллер может работать как HBA и как RAID контроллер. При этом часть портов может осуществлять один режим, а часть портов – другой. Это позволяет одновременно или по отдельности поместить на контроллер диски, которые образуют RAID-том, подключить недисковое SAS/SATA-устройство, подключить SAS/SATA HDD/SSD и тут же передать их диск менеджеру операционной системы в режиме HBA и последнее – подключить к портам контроллера как RAID внешнюю стойку, так и JBOD/EBOD внешнюю стойку, не обладающую функцией RAID. Это делает современный RAID-контроллер поистине универсальным устройством:

– *диски в составе RAID-тома*. Для поддержки этого решения на каждом диске создаются специальные служебные области для хранения информации о создаваемых на дисках виртуальных томах и др. служебной информации;

Табл. 7. Особенности различных батарейных модулей.

	Решения вчера с использованием BBU (батареи)	Первое поколение AFM модулей	Второе поколение AFM модулей, которое используется на 7 серии контроллеров	Будущее
Время зарядки	Порядка 24 часов	Несколько секунд на этапе загрузки	Несколько секунд на этапе загрузки	Несколько секунд на этапе загрузки
Геометрия решения	Дочерняя плата с логикой и батареей. Как опция, батарея на длинном проводе, подключаемая к дочерней плате	Дочерняя плата с логикой, флэш памятью и суперконденсатором на длинном проводе, подключаемом к дочерней плате	Логика в микросхеме контроллера, дочерняя плата с флэш памятью и суперконденсатором на длинном проводе, подключаемом к дочерней плате. Уменьшение размеров суперконденсатора	Логика в микросхеме, флэш на плате контроллера, суперконденсатор на длинном проводе, подключаемом к разъему на контроллере. Дальнейшее уменьшение размеров суперконденсатора
Время защиты кэша	24 часа в теории. В среднем, 8 часов с учетом деградации от срока эксплуатации и температуры	Несколько лет	Несколько лет	Несколько лет
Срок службы элементов питания	Батарея - в среднем, 2 года при правильном обеспечении термоусловий	Суперконденсатор – в среднем, 4 года при правильном обеспечении термоусловий	Суперконденсатор – в среднем, 4 - 6 лет при правильном обеспечении термоусловий	Дальнейшее увеличение срока эксплуатации суперконденсатора
Термоусловия	При неправильных теромоусловиях сокращение срока до 1 года	При неправильных теромоусловиях сокращение срока до 3 лет	При неправильных теромоусловиях сокращение срока до 5 лет	Дальнейшая защита сокращения срока эксплуатации от повышения температуры
Гарантия	1 год Нетипично для серверных компонентов	3 года	3 года и выше	-
Время тестирования до начала обеспечения защиты кэша	Порядка 24 часов	Несколько секунд	Несколько секунд	Несколько секунд
Утилизация	Специальная процедура утилизации из-за вредных компонентов	-	-	-

– *недисковые устройства*. Ленты, CD-ROM и др. недисковые устройства могут достаточно легко поддерживаться RAID-контроллером;

– SAS/SATA HDD/SSD диски в режиме HBA. В таком режиме на дисках не создаются специальные служебные области, и они сразу же передаются в работу операционной системе;

– *поддержка RAID внешних стоек*. Для полной поддержки требуется режим Multy LUN, т.е. способность видеть два и больше RAID-тома, созданных на внешней стойке. 7-я серия контроллеров поддерживает эту функцию.

Табл. 8. Преимущества использования MAX VIEW ADAPTEC STORAGE MANAGER.

Концепция	Комментарии
Концепция Сервера Управления (Management Server)	Позволяет в качестве клиента управления использовать любой WEB браузер на любой системе, имеющей доступ к сегменту LAN: админской машине, КПК, смартфоне и т.д.
Централизация управления и резервирование точки централизации (Centralized Management, Backup Management Server)	Администратор видит состояние всех контроллеров, всех серверов, подключенных к сегменту LAN, состояние всех томов и всех дисков на этих RAID контроллерах, при этом поддерживается вся текущая линейка контроллеров и контроллеры, уже снятые с производства. В случае потери системы, которая дает базовый IP, есть еще некоторый набор резервных систем, которые предоставляют такие же функции.
Вложенное управление (Layered Management)	Позволяет увидеть на общей схеме, что на LAN сегменте что-то случилось. Первый уровень показывает сервер, где есть проблема, следующий уровень – контроллер, на котором есть проблема, из общего числа контроллеров на данном сервере, следующий – том на контроллере, где есть проблема, и, наконец, последний уровень - физический диск, с которым что-то случилось.
Проактивное управление (Proactive Management)	Система управления показывает не саму проблему (потеря доступа и данных), а увеличение вероятности проблемы. Например, потерю перового диска в RAID6, затем, потерю второго, и далее - потерю третьего, которая приводит к остановке доступа к тому. Для проблем, которые повышают риск, используется уровень ПРЕДУПРЕЖДЕНИЙ, для реальных проблем - уровень ОШИБОК.
Обеспечение безопасности (Security Management)	Данная концепция решает все вопросы, связанные с обеспечением безопасности. Это крайне важно, поскольку, если кто-то получит интерфейс управления RAID контроллером, это может привести к безвозвратной потере данных. Обеспечение безопасности включает в себя: шифрование трафика управления, ведение журнала действий администраторов, авторизация в системе и раздача уровней прав на различные действия в системе хранения и ее настройках.
Агенты управления (Management Agents)	Позволяет настроить агентов (в случае системы управления Adaptec - это почтовые клиенты), которые при возникновении рисков или проблем шлют информацию администратору.
	Это позволяет в случае возникновения проблем получить сообщения в любое время, например, на мобильный телефон, и предпринять оперативные действия.
Удаленное управление (Remote Management)	Помимо удаленных возможностей управления, предоставляемых WEB браузером, функции удаленного управления можно получить и с помощью сторонних продуктов: терминалов карточек удаленного управления серверными платформами и терминалов операционных систем. Это позволяет использовать удаленно утилиту BIOS контроллера и CLI контроллера для настройки, мониторинга, восстановления данных и т.д.

Такая универсальность приводит к появлению новых функций. Несколько примеров: мы уже упомянули MultiLUN поддержку, кроме этого, наряду с функцией Initialize (которая создает служебные области) появляется функция Deinitialize (она убирает служебные области, чтобы обеспечить режим HBA). При управлении загрузкой в режиме HBA контроллер хранит WWN-адрес диска, с которого происходит загрузка, на своей флэш-памяти. Зарождаются и другие функции.

Лучше управляешь – крепче спишь

Со стороны производителя контроллеров существует некое базовое правило для администратора, которое гласит: для любого RAID-тома, где хранятся пользовательские данные, администратор обязан поднять систему управления.

Многие администраторы могут гневно воскликнуть: "Почему такой диктат? Это свободная страна". На самом деле, это правило защищает интересы самих администраторов, поскольку управление томом является вероятностным фактором, влияющим на возможность потерять доступ к тому и данным. Кроме того, утилиты управления совершенно бесплатные, их легко можно получить с сайта *www.adaptec.com/support*.

“Неверящие” люди могут провести такой (лучше мысленно) эксперимент: создаем RAID6 том, кладем на него сверхценные данные и НЕ ставим систему управления. Проходит несколько лет, один из дисков выходит из строя. Мы этого не видим. Еще несколько лет – еще один диск. Снова все работает. Ни админ, ни пользователи этого не замечают. Еще на одном диске начинаются массовые бэд-блоки, контроллер выдает много сообщений типа bad stripe, мы их по-прежнему не видим, и вот, наконец, доступ к тому обрывается. И мы не можем найти способ вернуть все 100% данных, если не было сделано резервных копий данных. При наличии системы управления мы увидели бы выход из строя первого диска, поняли бы, что том нуждается в ремонте, внедрили бы новый диск в наш том и вернули бы тому исходную надежность.

Некоторые базовые концепции – принципиально новые, реализованные только на 7-й серии контроллеров Adaptec – вложены в современные системы управления, такие как MAX VIEW ADAPTEC STORAGE MANAGER и представлены в табл. 8.

Заключение

RAID-контроллеры интенсивно развиваются. Можно с уверенностью сказать, что в самое ближайшее время в области систем хранения появятся и новые технологии и новые функции RAID-контроллеров, и новые подходы в проектировании. Все это требует хороших знаний старых, классических решений. Опыт и знания не только позволяют оптимизировать проектную стоимость, но и выполнить главную задачу – ХРАНИТЬ ДАННЫЕ НАДЕЖНО.

Дмитрий Зотов,
компания PMC-Sierra, подразделение Adaptec by PMC.