199178, Санкт-Петербург, ВО, наб. реки Смоленки, д.33
Телефон: +7 (812) 622 16 80
www.raidix.ru
info@raidix.com

Описание продукта RAIDIX RAIN.Эльбрус

Распределенная программная система хранения данных

2018

Санкт-Петербург

Содержание документа

Общее представление о решении

RAIDIX RAIN.Эльбрус представляет собой распределенную программную СХД (SDS, software-defined-storage) для развертывания на стандартных (commodity) отечественных серверных платформах на базе процессоров Эльбрус 4С.

Решение позволяет объединить локальные носители серверов в единый отказоустойчивый кластер хранения, с централизованным управлением, возможностью горизонтального и вертикального масштабирования.

RAIN эффективно решает задачи хранения для предприятий среднего и крупного бизнеса, государственных учреждений и корпораций, обеспечивая лучшее в индустрии соотношение показателей производительности, отказоустойчивости и эффективного использования пространства.

Архитектура решения

RAIN.Эльбрус поддерживает стандартный выделенный вариант развертывания системы, в рамках которого выделяются серверные узлы с дисками для организации распределенной системы хранения.

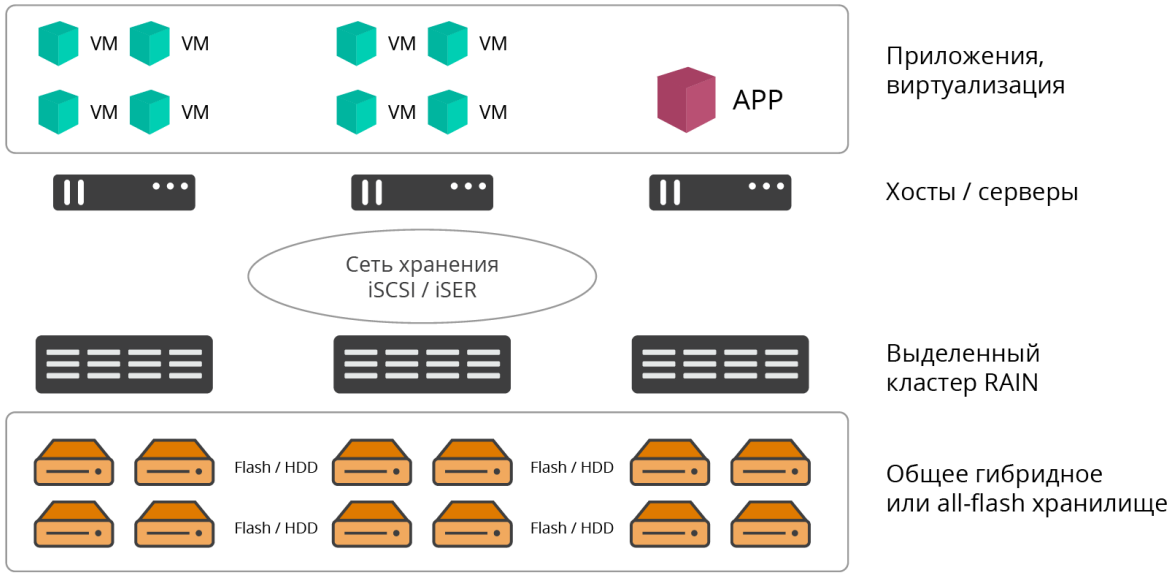

Рис. 1. Выделенный вариант развертывания

В выделенном варианте кластер RAIN представляет собой классическую программную СХД. Решение развертывается на необходимом количестве выделенных серверных узлов (не менее 3х, сверху количество практически не ограничено), ресурсы которых полностью утилизируются для задач хранения. При этом ПО RAIN устанавливается на голое железо. Приложения, сервисы, вычислительные ресурсы, использующие RAIN для хранения информации, размещаются на внешних хостах и подключают к нему по сети хранения данных (классическая архитектура ЦОД).

Взаимодействие узлов кластера RAIN между собой и с конечными потребителями ресурсов хранения (серверами, приложениями) осуществляется по протоколам iSCSI (IP, IPoIB) или iSER (RoCE, RDMA).

С точки зрения используемых носителей RAIN.Эльбрус поддерживает конфигурации с использованием SATA и SAS HDD.

Основные характеристики и функциональность решения

Операционные характеристики	Значение
Тип устройства хранения	Распределённая блочная СХД
Поддерживаемые типы узлов	Отечественные серверные платформах на базе процессоров Эльбрус 4С
Поддерживаемые типы носителей	SATA и SAS HDD
Максимальный объём хранения	16 ЭБ
Максимальный размер кластера	1024 узла
Функциональность хранилища	Горячее расширение томов Горячее добавление узлов в кластер, Ребалансировка кластера Отработка отказов без простоев
Технологии отказоустойчивости	Отработка отказов узлов, носителей, сети. Помехоустойчивое кодирование (Erasure Coding) с распределением по узлам кластера: сетевой raid-6,5,1,0. Коды коррекции на уровне локальных носителей узлов (локальный raid-6) Домены отказа

Основные возможности и преимущества решения

Отечественная разработка

ПО RAIN является на 100% отечественной разработкой. В случае развертывания на серверах на базе процессоров Эльбрус конечный заказчик получает современную отечественную платформу хранения, ключевые показатели которой не уступают и даже превосходят основных иностранных конкурентов, в т.ч. лидеров индустрии. Это особенно актуально для предприятий, имеющих ограничения на использование иностранных ИТ-систем, для инфраструктур, обрабатывающих гостайну.

Ключевые преимущества

Основная ценность RAIN — непревзойденное соотношение производительности, отказоустойчивости и эффективного использования ёмкости хранилища.
Решения конкурентов показывают высокую производительность, только при использовании зеркалирования. При этом полезная ёмкость хранилища сокращается в 2 раза и более: однократная репликация данных (зеркалировние) — 50% избыточность, двукратная репликация данных (двойное зеркалирование) — 66,6% избыточность. Использование технологий оптимизации хранения, таких как EC (erasure coding - помехоустойчивое кодирование), дедупликация и сжатие, реализованных в данных решениях, приводит к значительной деградации производительности хранилища, что неприемлемо для чувствительных к задержкам приложений.
Поэтому все продукты, представленные на рынке, в состав которых входят технологии оптимизации хранения, на практике вынуждены использоваться без данных технологий, либо включать их только для «холодных» данных.
Уникальные запатентованные технологии EC RAIDIX, используемые в RAIN, позволяют получить производительность сравнимую с зеркалированными конфигурациями. Это касается как случайной, так и последовательной нагрузки. При этом обеспечивается заданный уровень отказоустойчивости и значительно увеличивается полезная ёмкость, а накладные расходы составляют не более 30% сырой ёмкости хранилища.
Отдельного упоминания требует лучшая в индустрии производительность EC RAIDIX на последовательных операциях, в частности при использовании SATA-дисков большого объема.

Гибкость архитектуры

RAIN является программным продуктом и предполагает возможность установки на любое совместимое серверное оборудование.
Поддержка различных типов носителей позволяют исходя из бюджета и решаемых задач строить на базе RAIN:

- распределенные с высокой производительностью и гарантированным низким уровнем задержек;
- экономичные системы, удовлетворяющие большинству основных типов нагрузок.

Масштабирование решения

Традиционные СХД, выполненные в виде единого программно-аппаратного комплекса имеют общую проблему, связанную с масштабированием: производительность системы, упирается в контроллеры, их количество ограничено, наращивание ёмкости путем добавления полок расширения с носителями не даёт увеличения производительности. При таком подходе общая производительность СХД будет падать, поскольку с увеличением ёмкости прежнему количеству контроллеров необходимо обрабатывать больше операций доступа к возросшему объему данных.
RAIN поддерживает горизонтальное блочное масштабирование, в отличии от традиционных решений увеличение узлов (серверных блоков) системы приводит к линейному росту не только ёмкости, но и производительности системы. Это возможно поскольку каждый узел RAIN включает в себя не только носители, но и вычислительные ресурсы для ввода-вывода и обработки данных.
Процедура расширения кластера RAIN максимально проста и автоматизирована, система самостоятельно в фоновом процессе перераспределяет данные с учетом ёмкости новых узлов, нагрузка становится сбалансированной и равномерной, пропорционально повышается общая производительность и ёмкость хранилища. Процесс горизонтального масштабирования проходит «на горячую» без простоев, не требует остановки приложений и сервисов.

Обеспечение отказоустойчивости

Основным подходом обеспечения отказоустойчивости RAIN является использование уникальных кодов коррекции ошибок (EC — Erasure Coding или помехоустойчивое кодирование), которые являются собственной запатентованной технологией RAIDIX.
Все устройства кластера RAIN защищены от отказа электропитания подключением к бесперебойным источникам питания (UPS). Устройства, подключенные к одному UPS, называются группой отказа по питанию.
Требования к отказоустойчивости и доступности системы:

- Кластер должен переживать отказ не менее двух узлов, при количестве узлов строго больше 4. Для трех и четырех гарантируется отказ одной ноды.
- Узел должен переживать отказ не менее двух дисков в каждом узле при наличии не менее 5 дисков в узле.
- Уровень избыточности накопителей на типичном кластере (от 16 узлов) не должен превышать 30%
- Уровень доступности данных должен быть не ниже 99,999%

Для обеспечения данных требований кластер разбивается на независимые по подгруппы (субкластера). Количество узлов в одной подгруппе не более 20, что и обеспечивает требование по отказоустойчивости и доступности. Количество подгрупп неограниченно.
RAIN предлагает 3 варианта EC:

- для субкластера хранения от 5 до 20 узлов оптимальным подходом является использование сетевого raid-6;
- для 3 узлов оптимальным является использование raid- 1;
- для 4 узлов оптимальным является использование raid-5;

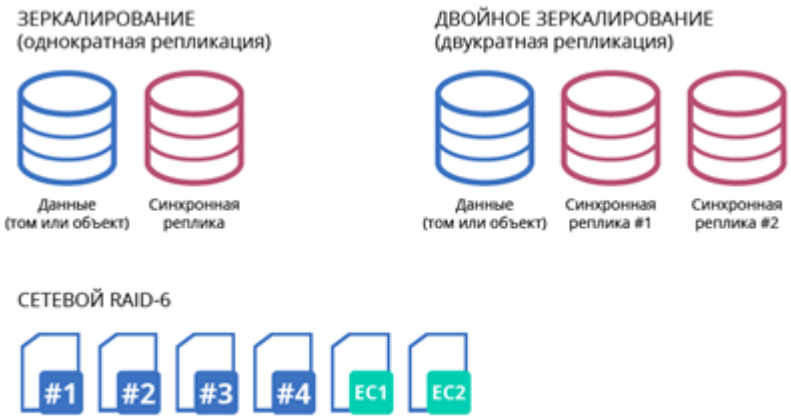

Рис. 3. Методы обеспечения отказоустойчивости

Все варианты предполагают равномерное распределение данных по всем узлам кластера с добавлением избыточности в виде контрольных сумм (или кодов коррекции). Данный подход позволяет провести параллели с кодами Рида-Соломона, используемыми в стандартных raid-массивах (raid-6) и позволяющими отработать отказ до 2х носителей. Сетевой raid-6 работает аналогично дисковому, однако распределяет данные по узлам кластера и позволяет отработать отказ 2х узлов.

RAIN предоставляет возможность объединения локальных носителей узлов в программный raid-0, raid-1, raid-5 или raid-6 на выбор. В последнем варианте при отказе 1-2 носителей внутри одного узла их восстановление происходит локально без использования распределённых контрольных сумм, минимизируя объем восстанавливаемых данных, нагрузку на сеть и общую деградацию системы.

Суммарная избыточность (накладные расходы на коды коррекции) с учетом локальных контрольных сумм узлов, порождаемая EC, в любом случае не превышает 30% сырой ёмкости кластера. Это позволяет значительно повысить полезную ёмкость хранилища по сравнению с традиционным подходом, использующим зеркалирование.

Зеркалирование также поддерживается в RAIN как альтернатива EC, в этом случае данные синхронно реплицируются между узлами (сетевой raid-1). Такой подход является единственным вариантом для небольших кластеров из 3-4 узлов и обеспечивает гарантированную отработку отказа 1 узла.

RAIN поддерживает концепцию доменов отказа (fault domain) или доменов доступности. Это позволяет отработать отказ не только отдельных узлов, но и целых серверных стоек или корзин, узлы которых логически группируются в домены отказа. Такая возможность достигается за счет распределению данных для обеспечения их отказоустойчивости не на уровне отдельных узлов, а на уровне доменов, что позволит пережить отказ всех сгруппированных в нем узлов (например, целой серверной стойки).

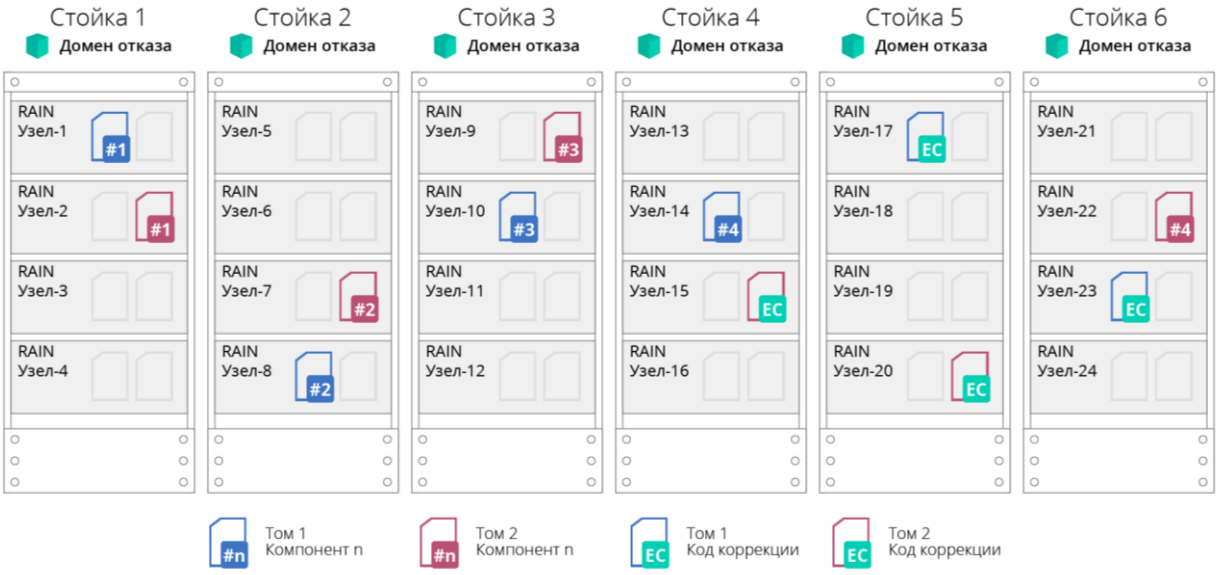

Рис. 4. Домены отказа

Отработка любых отказов (дисков, узлов или сети) осуществляется автоматически, без остановки работы системы.

Управление

RAIN обладает мощными встроенными средствами централизованного управления, предоставляет интерфейс командной строки.

Сценарии применения

- Корпоративные ИТ-инфраструктуры. Решение обеспечивает эластичное масштабирование, необходимое для развертывания инфраструктуры крупных предприятий: производительность, пропускная способность и объем хранения увеличиваются с каждым добавленным в систему узлом.
- Архивное хранение и системы видеонаблюдения. Хранение резервных копий, архивов, организация файлового хранилища, хранение данных видеонаблюдения, медийное хранение. RAIN позволяет строить надежные хранилища большого объема с быстрым последовательным доступом на базе экономичных емких SATA-дисков.