INFINIDAT

WHITE PAPER

АРХИТЕКТУРА СХД INFINIDAT

КРАТКИЙ ОБЗОР

Корпоративные системы хранения данных (СХД) INFINIDAT® основаны на уникальной запатентованной архитектуре хранения данных INFINIDAT: полностью независимый набор функций программно-определяемой СХД на лучшем среди аналогов стандартном «коробочном» железе. Поставляя программное обеспечение (ПО) на эталонной аппаратной платформе, прошедшей всесторонние испытания, INFINIDAT предлагает первую программно-определяемую СХД действительно корпоративного класса.

В этом документе описана технология, позволившая INFINIDAT стать единственным поставщиком сверхнадежной СХД с беспрецедентной доступностью (99,99999%), производительностью быстрее, чем all-flash (более 1 млн операций ввода-вывода в секунду с задержкой меньше одной миллисекунды), многопетабайтной емкостью, уместившейся в одной стойке 42U, и более низкой совокупной стоимостью владения.

ПРИНЦИПЫ ПОСТРОЕНИЯ

При проектировании архитектуры СХД современного ЦОД необходимо учитывать:

КАТЕГОРИЯ	ТРЕБОВАНИЕ
Надежность	Бизнес работает в режиме 24x7; простои не допустимы
Емкость	Экспоненциальный рост объемов данных, ускоряемый цифровой трансформацией, разрозненные архитектуры больших данных, искусственный интеллект и машинное обучение
Производительность	Производительность должна успевать за ростом объемов данных, чтобы обеспечивать прежние (или лучшие) результаты в те же (или более сжатые) сроки
Простота	Администраторам нужна простота эксплуатации, широкая интеграция с экосистемой решений и встроенных инструментов для перехода к моделям DevOps, чтобы тратить меньше времени на управление СХД, а больше времени уделять приложениям и бизнес-процессам
Консолидация	Точечные технологии уходят в прошлое; современная СХД должна обеспечивать максимальную эффективность, простоту и экономию затрат при любых вариантах использования
Стоимость	На увеличение емкости и производительности не закладывается адекватный бюджет; требуется кардинальное изменение архитектуры

В то же время Amazon, Google и Microsoft Azure заявляют, что их облака помогают заказчикам сократить расходы на весь ИТ-комплекс, и часто так и бывает в случае малого бизнеса, который не может содержать крупный ИТ-отдел и доверяет поддержку всех ИТ-ресурсов одному или двум специалистам широкого профиля. А вот крупные организации и региональные поставщики облачных услуг, внедрив более эффективный ИТ-комплекс, отвечающий их технологическим, финансовым и бизнес-требованиям, могут получить все преимущества облака внутри своей инфраструктуры, при этом сократив расходы и сохранив контроль над своими данными.

АРХИТЕКТУРА INFINIBOX

Флагманская технология INFINIDAT — InfiniBox® — была разработана на базе следующих основных принципов, помогающих находить все описанные ниже точки роста::

ПРИНЦИП	ОБОСНОВАНИЕ	ТОЧКИ РОСТА
Разработка инновационного ПО	Программное обеспечение, в отличие от аппаратного, со временем оптимизируется, повышая (а не понижая) производительность. В основе InfiniBox лежат более 80 выданных патентов на ПО — это настоящая программно-определяемая СХД	**Производительность** **Простота** **Надежность** **Стоимость**
Проектирование для отказоустойчивости	Проектируя масштабируемые решения, важно думать об отказоустойчивости. Система InfiniBox рассчитана на безотказную работу на уровне 99,99999% и тройную избыточность, когда все важные программные и аппаратные компоненты имеют по меньшей мере двойную избыточность N+2, что защищает от простоев и потерь данных.	**Отказоустойчивость** **Стоимость** **Простота** **Консолидация**
Построение масштабируемой архитектуры	Для получения нужной емкости и производительности и радикального снижения стоимости нужен масштаб. Система InfiniBox создавалась для крупных заказчиков с эффективной емкостью от 230 ТБ и ее масштабированием до 8,3 ПБ (с консервативным сокращением объема данных 2:1) в одной стойке высотой 42U.	**Консолидация** **Стоимость** **Простота** **Емкость**
Полная интеграция аппаратного и программного обеспечения	Высокая отказоустойчивость InfiniBox связана еще и с тем, что компания тестирует много «железа», но выпускает свои решения лишь на самом надежном (эталонная архитектура). Заказчики получают полностью интегрированное решение без расходов на администрирование и дополнительных усилий и затрат на его интеграцию на площадке.	**Надежность** **Простота** **Консолидация**
Стандартное «коробочное» аппаратное обеспечение	Стандартное «железо» и отсутствие длительных циклов разработки позволяют быстрее внедрять новые технологии, в том числе ЦП, память разных типов и новые типы носителей. Использование стандартного аппаратного обеспечения и сопутствующего ПО повышает стабильность работы, ведь на этом «железе» уже работают тысячи систем по всему миру.	**Стоимость** **Надежность** **Емкость** **Простота** **Производительность**

ПОВЫШЕНИЕ ПРОИЗВОДИТЕЛЬНОСТИ

InfiniBox® — это СХД с оптимизированным использованием флэш-технологий и комбинацией DRAM, SSD и дисков NL-SAS большой емкости для записи, чтения и хранения данных. Далее показано, за счет чего повышается скорость чтения и записи и достигается максимальная производительность при минимальном времени задержки. Размещение данных оптимизируется с помощью алгоритма Neural Cache. В этом разделе объясняется, как интеллектуальные программные алгоритмы Neural Cache обеспечивают меньшее время задержки по сравнению с all-flash СХД.

Важно помнить, что для большинства транзакционных приложений требуется по меньшей мере два отдельных устройства ввода/вывода (одно для записи транзакции в журналы и второе для записи данных в базу данных), из-за чего время задержки становится определяющим фактором как для удобства пользователей, так и для максимальной производительности приложений.

Уровень метаданных

Время отклика уровня метаданных напрямую влияет на задержку ввода/вывода. InfiniBox ускоряет операции с метаданными благодаря:

- **Хранению всех метаданных в DRAM** - что ускоряет как чтение, так и запись
- **Эффективной структуре метаданных** – любой ввод, изменение и удаление из структуры метаданных (т.н. “TRIE”) выполняются с одинаковой задержкой, что обеспечивает стабильную производительность

Ускорение записи

InfiniBox принимает в свою память DRAM все записи без предварительной обработки (например, удаление шаблона, сжатие, шифрование и т.п.) и перед тем, как отправить подтверждение хосту, делает вторую копию записи в DRAM другого узла посредством InfiniBand с малым временем задержки. Принимая записи не от внешнего флэш-устройства, а от памяти DRAM (связанной прямо с ЦП), InfiniBox выполняет запись с максимально низкой задержкой.

В отличие от многих других архитектур, где кэш записи разбит на небольшие сегменты (например, в матричных и двухконтроллерных архитектурах), InfiniBox использует единый большой пул памяти для приема записей. Это позволяет справляться с более крупными пакетами записей, перезаписывать часто изменяемые данные со скоростью DRAM и дает алгоритму Neural Cache время для оптимального распределения блоков данных между более быстрой DRAM и менее быстрыми SSD или HDD. Дольше храня данные в кэше записи, Neural Cache избегает ненужной нагрузки на ЦП и уровни сохраняемости на серверах.

Ускорение чтения

В отличие от традиционных массивов хранения, которые размещают наиболее активно используемые данные (т.н. «горячие данные») в кэш флэш-памяти, чтобы догнать all-flash СХД по производительности, InfiniBox использует инновационный алгоритм Neural Cache, который размещает все «горячие данные» в DRAM. InfiniBox Neural Cache позволяет выполнять большинство операций чтения со скоростью DRAM, которая почти в 1000 раз выше, чем у флэш-памяти.

По состоянию на 2017 год глобальная платформа данных INFINIDAT охватывает несколько экзабайт данных, а Neural Cache гарантированно выполняет почти все операции чтения из DRAM, что позволяет заказчикам пользоваться технологией All-DRAM-Array по цене ниже, чем all-flash СХД.

Neural Cache – это самообучающийся алгоритм, который со временем оптимизирует свою работу. InfiniBox использует большой флэш-слой SSD, который служит “подушкой” для промахов DRAM. По мере того, как Neural Cache обучается шаблонам ввода/вывода и оптимизирует размещение данных в DRAM, флэш-слой переключается с обработки промахов DRAM на обработку изменений шаблонов ввода/вывода, которые алгоритм не может прогнозировать (например, периодический аудит, для которого требуются данные, отсутствующие в DRAM).

ПРОГРАММНАЯ АРХИТЕКТУРА

Проектируя СХД InfiniBox с надежностью хранения на уровне 99,99999%, компания INFINIDAT сумела справиться с непредсказуемостью аппаратных сбоев с помощью ПО. Программная архитектура InfiniBox с тремя активными узлами и избыточностью N+2 обеспечивает постоянный мониторинг и самовосстановление, а также надежное восстановление после аппаратных сбоев на всех уровнях.

Все компоненты, от RAID-массивов и до кластерных служб, реализованы программными средствами, что позволяет оптимизировать решение с выходом каждой новой версии. За первые четыре года с момента выпуска первой общедоступной версии максимальная производительность InfiniBox выросла в 4 раза только благодаря обновлению ПО без вмешательства в работу системы. В этом сила настоящего программно-определяемого решения.

Кластерные службы

Все службы передачи данных работают на всех узлах (согласно архитектуре N+2) и активны на всех узлах (в кластере нет пассивных узлов). Службы передачи данных предназначены для работы в пользовательском пространстве, включая компоненты низкого уровня, такие как драйверы оптоволоконного канала.

Поскольку в ядре нет служб передачи данных, сбой не может повлиять ни на другие службы в системе, ни на доступность узла. Такие принципы построения применимы как к клиентским (например, протоколы данных NFS, iSCSI, FC, FICON), так и к серверным службам передачи данных (Neural Cache, InfiniRAID® и InfiniSnap®).

Службы передачи данных запускаются и контролируются диспетчером кластера, который обнаруживает проблемы в работе служб и может при необходимости перезапускать их. Служба, в которой произошел сбой, будет перезапущена и выполнит самопроверку перед переподключением к кластеру. Подключение к кластеру некорректно запустившейся службы исключается во избежание сбоя внутри кластера (византийская ошибка). Если диспетчер кластера обнаруживает службу, которая несколько раз безуспешно пыталась перезапуститься на определенном узле, он останавливает перезапуск и уведомляет службу поддержки INFINIDAT.

Отчет о каждом сбое службы (независимо от успешности автоматического восстановления) направляется в систему аналитики данных INFINIDAT для выявления программных проблем и дальнейшего повышения качества кода.

Структура диска

Структурой диска InfiniBox управляет запатентованное инновационное ПО InfinRAID. Этот программно-определяемый RAID-массив контролирует размещение всех данных, их защиту и аварийное восстановление по определенным сценариям.

InfiniRAID представляет собой рассредоточенный RAID-массив, то есть массив такого типа, который отделяет размещение данных от физического слоя и задействует тысячи виртуальных RAID-групп, распределяя данные по всем дискам и предотвращая образование “горячих точек”. InfiniRAID создает RAID-группы таким образом, чтобы каждая пара дисков в них не содержала больше 2,5% информации их RAID-групп.

Низкий процент пересечений RAID-групп дает множество преимуществ:

- **Равномерное распределение** – все наборы данных, даже самые мелкие, распределяются по всем дискам в системе, что обеспечивает максимальную пропускную способность для каждого приложения.
- **Самовосстановление** – потенциальные “горячие точки” автоматически нивелируются благодаря оптимизации размещения данных.
- **Виртуальные резервы** – емкости равномерно распределяются по всем дискам в системе. Физических дисков горячего резерва нет, и это позволяет процессу восстановления оптимально распределять данные и сводит к минимуму необязательные расходы. В системе достаточно резервной емкости на случай сбоя 12 дисков в F6000.
- **Защита производительности** – сбой одного диска (данные по-прежнему защищены) запустит восстановление RAID лишь с низким приоритетом (“Rebuild-1”), когда во главу угла ставится производительность приложения и используются только резервные ресурсы системы.

- **Быстрое восстановление** – в случае сбоя второго диска система ускорит восстановление 2,5% информации RAID-групп, общей для двух неисправных дисков ("Rebuild-2"), а затем вернется к имеющему низкий приоритет процессу Rebuild-1 (т.к. незащищенных RAID-групп больше нет).
- **InfiniSpares** – помимо гарантированной емкости 12 резервных дисков, InfiniBox может при необходимости использовать свободную емкость как резервную, т.е. данные останутся под защитой даже при сбое 100 дисков.

Службы защиты данных

InfiniBox предлагает заказчикам целый ряд служб защиты данных:

- **Моментальные снимки** – механизм создания моментальных снимков в InfiniBox называется InfiniSnap и представляет собой неблокирующие моментальные снимки с отложенной записью (redirect-on-write), обеспечивая устойчивую производительность как с моментальными снимками, так и без них. Для каждого набора данных предусмотрено до 1000 моментальных снимков, каждый либо только для чтения (защита данных), либо с возможностью перезаписи (для сред тестирования и разработки). InfiniSnap выполняет моментальные снимки в DRAM, не требуя записи на уровне сохраняемости данных.
- **Асинхронная репликация с низкими RPO** – механизм асинхронной репликации может поддерживать целевой показатель по точкам восстановления (RPO) в 4 секунды, используя при этом IP-инфраструктуру для снижения затрат и упрощения процессов.
- **Синхронная репликация** – механизм синхронной репликации обеспечивает синхронную защиту данных с нулевым RPO, а время задержки СХД не превышает 400 микросекунд. В случае проблем с глобальной вычислительной сетью (ГВС) (большое время задержки, потеря подключения), механизм синхронной репликации InfiniBox автоматически возвращается в асинхронный режим. После восстановления ГВС автоматически воспроизводятся все недостающие данные и возобновляется синхронная репликация без нарушения операций ввода/вывода.

Сокращение объема данных

InfiniBox использует несколько способов сокращения объема данных, еще больше снижая стоимость хранения за счет:

- **Динамического выделения ресурсов по умолчанию** – по умолчанию ресурсы всех томов выделяются динамически. Благодаря интеллектуальным пулам емкости InfiniBox риск избыточного выделения можно легко снизить путем настройки предупреждений о пороговых значениях и аварийных буферов в пуле и тем самым сохранить доступность приложения.
- **"Нулевого" возврата** – при очистке хост-машинами (физическими и виртуальными) пространства на диске (LUN) они записывают туда нули, выполняя операцию повторной записи write-same (более эффективный вариант), либо просто записывая отдельные нули в это пространство. InfiniBox обнаруживает оба варианта и удаляет такое пространство, как будто в нем никогда ничего не записывалось, тем самым улучшая динамическое выделение ресурсов.
- **Сжатия** – InfiniBox сжимает данные только после того, как начался их перенос из кэша записи (DRAM) на диск. Это ускоряет запись (не добавляется задержка из-за сокращения объема данных) и предотвращает сжатие передаваемых данных, которые будут перезаписаны через несколько секунд (экономия ресурсов ЦП). Технология сжатия InfiniBox использует алгоритм LZ4 с размером блока 64 КиБ, позволяя получить более высокий коэффициент сжатия по сравнению с традиционным сжатием с малыми блоками (используется во всех all-flash СХД).
- **Моментальные снимки** – моментальные снимки InfiniBox по умолчанию являются динамическими, что позволяет заказчикам избежать штрафов за дефицит емкости, обусловленный созданием полной копии.

Сетевая архитектура

Доступность сети важна для работы всех сетевых служб. В частности, когда речь заходит об IP-службах (iSCSI, NFS, асинхронная и синхронная репликация), ИТ-администраторы обычно ожидают, что в случае сбоя СХД переключится на другой IP-адрес и быстро решит проблемы конфигурации. InfiniBox привносит в этот процесс новаторские идеи — мгновенное переключение на другой IP-адрес в случае проблем с подключением, при этом IP-адреса переносятся в те сетевые интерфейсы, которые могут предоставить соответствующие службы.

Мгновенное переключение на другой IP-адрес происходит во всех сценариях отказов, включая аппаратные (отказ узла, порта/сетевой карты Ethernet) и программные (отказ службы на отдельном узле). Чтобы свести к минимуму воздействие на другие службы, InfiniBox перемещает минимальное количество IP-адресов, так что IP-адреса другой службы на том же узле, равно как и IP-адреса на других узлах не будут перемещены.

InfiniBox также использует виртуальные MAC-адреса и назначает каждый IP-адрес виртуальному MAC-адресу. При перемещении IP-адресов виртуальные MAC-адреса перемещаются вместе с ними, т.е. можно не тратить время на переключение, менять конфигурацию только на коммутаторе (не “накатывая” изменение на каждую хост-машину), а также избежать проблемы самопроизвольного ARP и повысить доступность.

InfiniBox с помощью интеллектуального мониторинга сети (команда IPv6) обнаруживает потенциальные ошибки конфигурации, например, когда интерфейсу сети хранения случайно заблокирован доступ к виртуальной локальной сети, используемой для служб передачи данных. Постоянный мониторинг каждой сети, настроенной в InfiniBox, помогает администраторам СХД понять, почему приложение потеряло доступ к СХД, зачастую задолго до того, как они сами задались бы этим вопросом.

АППАРАТНАЯ АРХИТЕКТУРА

InfiniBox — это программно-определяемая СХД на доступном «коробочном» железе. На этапе разработки компания INFINIDAT инвестировала в ПО, которое делает аппаратное обеспечение более надежным, экономичным, а также простым в администрировании и обслуживании. Самый главный принцип разработки “N+2” предусматривает для всех компонентов по меньшей мере тройную избыточность и обеспечивает надежность на уровне 99,99999%. Система InfiniBox поставляется предварительно собранной в стойке (см. рис. 1):

Узлы

Узлы — это контроллеры СХД в InfiniBox. В кластере работают три активных узла с полной избыточностью, и поэтому операции ввода/вывода легко выполняются на всех трех узлах. Узлы напрямую связаны с модулем быстрой сети InfiniBand для прямого доступа в память с помощью RDMA, что позволяет выполнять быструю репликацию новых записей между узлами с минимально возможной задержкой.

В случае отказа одного узла два оставшихся принимают на себя его функции и выполняют ресинхронизацию любой части кэша записей, которая больше не реплицируется, восстанавливая полную защиту данных и обслуживание без вмешательства в работу системы. Архитектура узлов N+2 также упрощает обслуживание на отдельном узле (например, замену компонента), так как в системе остаются 2 активных узла, которые продолжают работать и защищать данные.

РИС. 1 *Внешний вид InfiniBox*

Физические соединения

Интерфейсы для соединения клиентских узлов с платформой данных заказчика:

- **Оптоволоконный канал** – восемь портов на узел, всего 24 порта. Все порты активны, т.е. каждый хост видит несколько путей (минимум по одному на узел; рекомендуется по два на узел). Благодаря наличию нескольких путей, отказ порта или адаптера главной шины повлияет только на отдельный путь, но не на работу приложения.
- **Порты Ethernet (Eth)** – четыре порта на узел, всего 12 портов с соединением по медному или оптоволоконному кабелю, с поддержкой iSCSI, NFS, протоколов синхронной и асинхронной репликации, а также интеграцией с InfiniSync (уникальным решением INFINIDAT с нулевыми RPO на любые расстояния). В случае отказа порты переключаются на другой IP-адрес, не позволяя аппаратному сбою повлиять на доступность системы.

Для узлов предусмотрена и избыточная внутренняя коммутация:

- **Порты InfiniBand (IB)** – предназначены для соединения кластеров. Если отказ InfiniBand привел к разрыву соединения между двумя узлами, то они будут связываться через третий узел. В случае разрыва соединения между одним узлом и двумя оставшимися первый корректно выводится из кластера на время, пока соединение не будет восстановлено.
- **Порты SAS** – соединяют узлы со всеми дисковыми модулями. После отказа SAS, в результате которого один узел потерял доступ к некоторым дискам, InfiniBand получает удаленный доступ к этим дискам через другой узел.

Узлы имеют избыточные источники питания и питаются от разных резервных аккумуляторов, которые в свою очередь питаются от нескольких силовых вводов, что позволяет обеспечить непрерывность операций на случай сбоев электроснабжения.

Автоматические тестовые переключатели

Автоматические тестовые переключатели (АТП) регулируют электропитание резервных аккумуляторов, гарантируя для них входной ток даже в случае отказа одного из источников питания. АТП мгновенно переключается с одного источника питания на другой в случае отказа первого, обеспечивая непрерывность энергоснабжения резервных аккумуляторов.

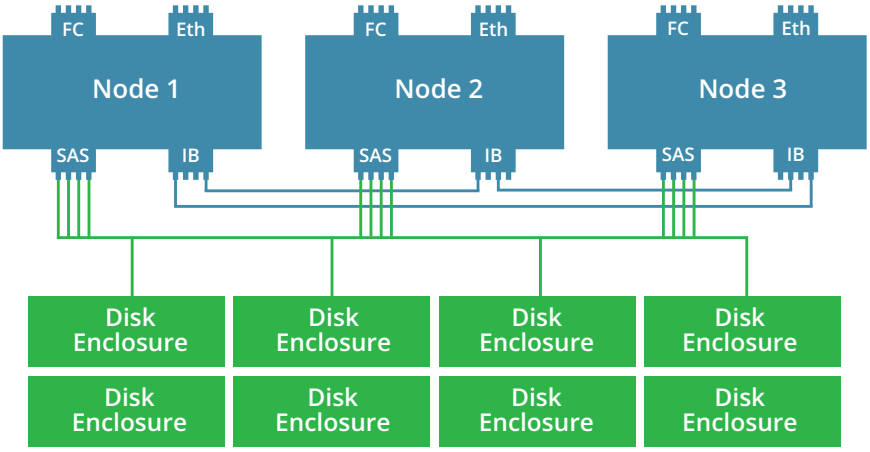

РИС. 2 *InfiniBox: интерфейсы для подключения серверов и внутренняя коммутация*

Резервные аккумуляторы

Резервные аккумуляторы обеспечивают питание узлов InfiniBox, чтобы можно было не выключать систему на время короткого сбоя в электроснабжении (например, до выхода генераторов на полную мощность) или в случае длительного сбоя корректно завершить работу InfiniBox, надлежащим образом удалив (переместив) данные из кэша DRAM.

Все резервные аккумуляторы находятся под контролем и раз в неделю автоматически проверяются на предмет исправности и готовности защитить систему в случае реального сбоя в энергоснабжении.

ЗАКЛЮЧЕНИЕ

Благодаря уникальной архитектуре InfiniBox больше не нужно идти на компромисс между производительностью, устойчивостью, емкостью и стоимостью. Впервые ИТ-департаменты дают возможность бизнесу хранить все критически важные данные, не раздувая бюджеты на ИТ и не снижая прибыль. С InfiniBox бизнес может уверенно приступать к цифровой трансформации и проектам с использованием технологии больших данных.