
# Решение для хранения данных RAIDIX

СХД для кластера Intel Lustre

2017

# Оглавление



# Резюме

Решения для хранения данных HPC должны обеспечивать защиту данных, доступность информации, масштабируемость и гарантированно высокую производительность системы. ПО для СХД RAIDIX в связке с Intel® Enterprise Edition for Lustre* обеспечивает необходимую функциональность и позволяет создать эффективный кластер хранения на базе стандартного оборудования.

В данном документе приведены технические описания решений RAIDIX и Lustre, рекомендуемая аппаратная архитектура и схема развертывания СХД для высокопроизводительных вычислений.

# Введение

На сегодняшний день технология HPC — это не только и столько ИТ-инструмент исследователей. Все больше компаний открывают для себя конкурентные преимущества HPC в рамках тех или иных бизнес моделей. Предприятия генерируют большие объемы данных и используют высокопроизводительные приложения для анализа и обработки информации. Для корпоративного сектора критичными становятся не только непрерывность бизнес-процессов, но и доступность данных и производительность доступа. Компании нуждаются в инфраструктуре хранения с возможностью гибкого горизонтального масштабирования и высокими показателями пропускной способности и отказоустойчивости без потери данных.

Коммерческий продукт Intel® Enterprise Edition for Lustre* включает в себя функционал ПО Lustre, оптимизированный под задачи надежного хранения и обеспечения максимальной пропускной способности в среде HPC (High Performance Computing — высокопроизводительные/суперкомпьютерные вычисления). Основные преимущества Intel® Enterprise Edition for Lustre* — высокая производительность, гибко масштабируемая емкость, собственные управляющие компоненты и поддержка 24/7.

Для решения задач индустрии HPC компания «Рэйдикс» создала комплексное совместное решение на базе технологии «готового кластера» (cluster-in-a-box) RAIDIX HPC и программного обеспечения Intel Lustre. В составе решения — управляющее ПО RAIDIX для систем хранения данных, работающее на стандартном серверном оборудовании с Lustre OSS (object storage server — сервер объектного хранения) / OST (object storage target — таргет объектного хранения) или MDS (metadata server — сервер метаданных) / MDT (metadata storage target — таргет хранения метаданных) как конструктивный блок для инфраструктуры хранения Lustre HPC.

Такие конструктивные блоки могут содержать от 8 до 128 дисков в шасси высокой плотности с производительностью до 12 ГБ/с. Отдельные узлы хранения объеди-

няются в горизонтально-масштабируемую систему, использующую Intel® Enterprise Edition for Lustre*.

СХД RAIDIX соответствует высоким требованиям по производительности, отказоустойчивости и целостности рабочих процессов, обеспечивает высокую пропускную способность, низкие задержки и надежность хранения благодаря использованию параллельных вычислений и патентованных алгоритмов в RAID 6 и RAID 7.3. Уникальные алгоритмы обеспечивают скорость вычислений на уровне 37 ГБ/с (в RAID 6) и 25 ГБ/с (в RAID 7.3) на ядро процессора.

В отличие от традиционных методов конфигурации серверов Lustre OSS и MDS с помощью дополнительного оборудования и отдельной настройки каждого сервера, RAIDIX позволяет построить инфраструктуру хранения HPC из интегрированных блоков и сократить стоимость владения системой благодаря универсальной совместимости со стандартным аппаратным обеспечением и протоколами SAN и NAS.

Далее приводятся аппаратные требования к двухконтроллерной конфигурации RAIDIX и описывается функционал комплексного решения.

# Двухконтроллерная СХД RAIDIX

RAIDIX позволяет строить высокопроизводительные системы хранения данных с использованием широко распространенных аппаратных платформ на базе процессоров Intel. Для обеспечения полной отказоустойчивости решения RAIDIX может функционировать в режиме двухконтроллерного кластера (Active-Active).

Для двухконтроллерных конфигураций наиболее подходящими являются платформы, совместимые со Storage Bridge Bay (SBB), которые уже содержат компоненты, необходимые для организации хранилища высокой доступности.

Общие требования для двухконтроллерной платформы RAIDIX:

| | |
|---|---|
| CPU | Процессоры Intel Xeon E5-2637 v4/E5-2667 v4 |
| Материнская плата | Должна быть совместима с моделью процессора и поддерживать PCI Express 3.0 x8/x16 |
| Внутренняя кэш-память | Должна быть совместима с соответствующей материнской платой, от 64 ГБ для каждого узла |
| Шасси | Рекомендуется двойной блок энергопитания и двойная материнская плата |
| SAS-контроллер (могут быть использованы дополнительные порты для подсоединения внешних JBOD) | Рекомендуется Broadcom 93xx |
| HBA (контроллер для синхронизации кэша) | Рекомендуется Mellanox ConnectX-3 VPI и выше |
| HBA (контроллер для соединения с Lustre по сети) | Рекомендуется Mellanox ConnectX-3 VPI и выше |
| HDD (жесткие диски) | Для двухконтроллерной архитектуры необходимы диски SAS |
| Устройства для кэша 2-го уровня | HGST SSD SS200 |
| Сеть Lustre | InfiniBand* QDR/FDR/EDR, Ethernet |

| | |
|---|---|
| | 10GbE/40GbE/100GbE |
| Управляющая сеть | Ethernet 1GbE |

## Комплексное решение

RAIDIX позволяет организовать хранилище с возможностью быстрого и надежного аварийного переключения (failover), высокопроизводительной обработкой данных, широкой функциональностью для обеспечения целостности информации и мониторинга системы. ПО RAIDIX, интегрированное с Intel® Enterprise Edition for Lustre*, включает в себя пакет для установки на системы на базе процессоров Intel Xeon. Алгоритмы помехоустойчивого кодирования RAIDIX, настроенные для работы с процессорами Intel, обеспечивают высокую скорость производимых операций.

Что касается горизонтально-масштабируемого кластера на базе Intel Lustre, то данная технология представляет целый ряд преимуществ:

- высокая управляемость с Intel Manager for Lustre;
- высокая производительность операций ввода-вывода для корпоративных приложений, таких как MapReduce;
- поддержка клиента Intel Xeon Phi;
- коннектор Hadoop, который позволяет использовать кластер Lustre для приложений Hadoop
- полное управление иерархической структурой хранения данных;
- специальный патч для улучшения обработки однопоточных запросов.

### Управление хранилищем

СХД на базе RAIDIX имеет удобный веб-интерфейс, который позволяет конфигурировать тома хранения и осуществлять мониторинг производительности системы.

## Управление кластером Lustre

Кластер Lustre управляется через Intel Manager for Lustre — веб-приложение, построенное на REST API и полноценном CLI. Приложение имеет следующую функциональность:

- формирование и мониторинг файловых систем Lustre;
- конфигурация серверов и томов;
- средства мониторинга производительности и использования ресурсов.

## Защита данных томов

ПО RAIDIX использует помехоустойчивое кодирование на базе патентованных алгоритмов, оптимизированных для высокопроизводительных задач. RAIDIX поддерживает различные уровни RAID (RAID 0, RAID 5, RAID 6, RAID 7.3, RAID N+M и RAID 10) и позволяет системным администраторам достичь нужного уровня защиты данных.

Патентованные компанией «Рэйдикс» уровни RAID:

- **RAID 6.** Уровень чередования блоков с двойным распределением четности, основанный на математических алгоритмах собственной разработки «Рэйдикс». Данные и служебная информация распределяются по всем дискам RAID-группы. RAID 6 обеспечивает повышенную производительность, так как каждый диск обрабатывает I/O запросы самостоятельно, позволяя осуществлять доступ к данным в параллельном режиме. RAID 6 может выдержать полный отказ двух дисков в одной группе.

- **RAID 7.3.** Уровень чередования блоков с тройным распределением четности, который позволяет восстанавливать данные при отказе до 3-х дисков. В основе RAID 7.3 заложен собственный уникальный алгоритм RAIDIX, позволяющий достигать высоких показателей производительности без дополнительной нагрузки на процессор.

  RAID 7.3 является аналогом RAID 6, но имеет более высокую степень надёжности — рассчитываются 3 контрольные суммы по разным алгоритмам, под контрольные суммы выделяется ёмкость 3-х дисков.

Для массивов более 32 ТБ рекомендуется использовать именно RAID 7.3, который существенно снижает вероятность отказа дисков без потерь в производительности и стоимости.

- **RAID N+M.** Уровень чередования блоков с M распределением четности, позволяющий пользователю самостоятельно определить количество дисков, выделяемых под хранение контрольных сумм. Уникальная технология RAIDIX позволяет восстановить данные при отказе до 32 дисков (в зависимости от количества дисков, выделяемых под контрольные суммы).

### Защита от скрытого повреждения данных

Скрытое повреждение данных (Silent Data Corruption), как правило, возникает из-за ошибок в работе драйверов, прошивки диска, памяти, повреждений поверхности диска и аналогичных программных и аппаратных сбоев. Скрытые ошибки не распознаются контроллерами жестких дисков и операционной системой до тех пор, пока не приведут к повреждению структуры данных.

Используемый в RAIDIX уникальный алгоритм позволяет обнаружить и исправить скрытые ошибки во время выполнения обычных дисковых операций путем анализа RAID-метаданных, без потери производительности. Сканирование и исправление скрытых ошибок выполняется RAIDIX в фоновом режиме в периоды низкой активности СХД.

### Частичная реконструкция

В ПО RAIDIX реализован уникальный механизм частичной реконструкции, позволяющий восстанавливать конкретную область жесткого диска, тем самым сокращая общее время восстановления массива. Частичная реконструкция эффективна для массивов больших объемов.

Пространство массива разбито на 2048 частей, по которым ведется отслеживание изменений. Восстановление данных происходит только в тех зонах, где было зафиксировано изменение блоков данных.

## Гарантированно высокая производительность

Все RAID-алгоритмы рассчитываются на стандартных процессорах Intel Xeon с высокой производительностью и высоким уровнем параллелизации вычислений. В составе ПО RAIDIX функционирует механизм упреждающей реконструкции, позволяющий оптимизировать скорость чтения в процессе восстановления данных на дисках за счет исключения из процесса дисков, скорость чтения с которых ниже, чем у остальных.

Упреждающая реконструкция позволяет восстанавливать данные с помощью RAID-вычислений быстрее, чем физически считывать данные с диска, — на уровне 25 ГБ/с. Данный функционал обеспечивает высокую производительность системы даже в режиме деградации и при отказе нескольких дисков.

## Высокая доступность данных

Кластерная система RAIDIX создает отказоустойчивый, высокопроизводительный кластер (в двухконтроллерном режиме) и размещает RAID'ы ассиметрично на узлах. Каждый RAID может быть доступен через другой узел. При этом параллельная файловая система Lustre позволяет клиенту осуществлять чтение и запись на множественные тома OST одновременно, увеличивая общую производительность.

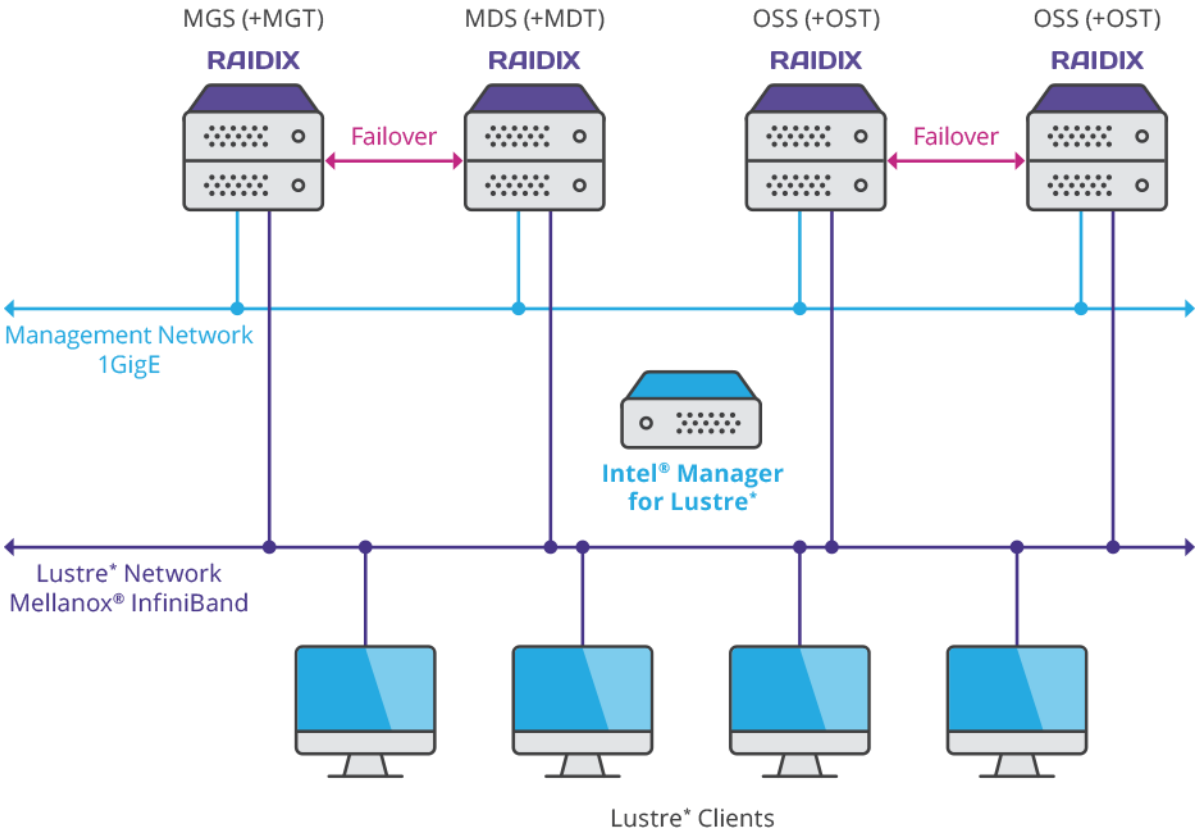
Функции автоматического и ручного аварийного переключения в RAIDIX помогают увеличить отказоустойчивость системы. Кроме того, RAIDIX обеспечивает высокую сбалансированную производительность благодаря возможности мигрировать RAID'ы с любого узла кластера.

Интеграция Lustre в двухконтроллерный RAIDIX позволяет пользователю:

- асимметрично разместить несколько Lustre OST на каждом узле кластера RAIDIX и сбалансировать нагрузку на каждый узел;
- обеспечить высокую доступность данных, хранимых на OST и MDT: в случае если узел выйдет из строя, данные останутся доступными на другом узле;

- интегрировать механизм отказоустойчивости Lustre OST и MDT в процесс аварийного переключения для всего узла. В этом случае нет необходимости использовать дополнительные сервисы, такие как Corosync и Pacemaker, поскольку кластер RAIDIX полностью берет на себя аварийное переключение Lustre.

# Схема развертывания




Приведенная схема развертывания системы рекомендована для типичного приложения HPC:

- Для более высокой доступности каждого OST используется двухконтроллерная (DC) архитектура RAIDIX
- На каждом контроллере в RAIDIX DC, используемом для OST, установлена Lustre OSS в конфигурации Active-Active.
- Каждый OST в кластере RAIDIX регистрируется на обоих OSS-серверах, установленных на узлах кластера. Конфигурируется «нативное» аварийное переключение (failover) RAIDIX: в случае выхода из строя одного OSS, отказоустойчивый механизм RAIDIX передает контроль над OST второму, функционирующему OSS.
- MGS (management server — управляющий сервер) и MDS (сервер метаданных) Lustre также должны быть сконфигурированы в отказоустойчивом ре-

жиме в рамках RAIDIX DC, чтобы достичь более высокой доступности таргетов MGT и MDT.

- Для обеспечения расширенной функциональности по управлению и мониторингу системы устанавливается Intel Manager for Lustre.
- Для управления сетевыми соединениями используется 1GbE Ethernet
- Для соединения с Lustre используется InfiniBand 56Gb.
- На каждой клиентской машине установлена система Lustre.

Выполнение данных рекомендаций позволяет создать инфраструктуру хранения HPC высокой доступности.

# Предлагаемая архитектура

В качестве аппаратной платформы «Рэйдикс» рекомендует использовать узлы кластера в рамках одних шасси и идентичные устройства SBB. Платформа должна масштабироваться с помощью дополнительных дисковых полок JBOD для увеличения емкости и производительности.

AIC HA201-TP — это 2U решение высокой доступности формата cluster-in-a-box («готового кластера») с использованием широко доступных компонентов. Двухконтроллерная конфигурация строится из двух серверов Intel (S26xxTP). Каждый узел поддерживает двойной процессор Intel Xeon серий E5-2600 v4.

Решение HA201-TP обеспечивает высокую доступность данных в режиме Active-Active и включает в себя отказоустойчивые, заменяемые в «горячем» режиме вычислительные узлы, 24 отсека для жестких дисков и 5 слотов PCIe Gen3 на узел.

| | |
|---|---|
| Платформа | AIC HA201-TP SBB |
| CPU | Двойной процессор Intel Xeon E5-26xx v4 для каждой материнской платы |
| Материнская плата | Intel Server Board S2600TP |
| Внутренняя кэш-память | 64 ГБ на узел |
| Шасси | AIC HA201-TP, двойная материнская плата, двойной блок энергопитания, 24 отсека для HDD с возможностью горячей замены |
| SAS-контроллер (соединение через внутреннюю объединительную плату) | Broadcom 9300 8-i |
| HBA (для синхронизации кэша) | Двухпортовый адаптер от Mellanox ConnectX-3 и выше |
| HBA (соединение с Lustre по сети) | Mellanox ConnectX-3 и выше |
| HDD | 24x NL-SAS 7.2K |

| | |
|---|---|
| ПО RAIDIX | v. 4.5 |
| Intel® Enterprise Edition for Lustre* | v. 2.x/3.x |

Модуль AIC HA201-TP SBB — передняя и задняя панель.

## Результаты для бизнеса

Интегрированное решение на базе RAIDIX HPC и Intel Enterprise Edition for Lustre – надежный конструктивный блок для построения инфраструктуры HPC. Решение отвечает требованиям высокой производительности, отказоустойчивости и целостности данных, обеспечивает высокую пропускную способность, низкие задержки и высокую надежность. Среди преимущества RAIDIX и Lustre:

- сокращение расходов на оборудование;
- сокращение расходов на средства соединения;
- гибкая конфигурация и простота внедрения и сопровождения;
- быстрое аварийное переключение (failover) и высокая доступность данных.

## О компании «Рэйдикс»

Компания «Рэйдикс» (www.raidix.ru) — ведущий поставщик систем хранения данных. Системы RAIDIX поставляются во многие страны мира. Используя собственную, запатентованную в России и США, технологию помехоустойчивого кодирования и обширную научную базу, компания предлагает отечественное решение для управления отдельными серверами СХД и построения масштабируемых высокопроизводительных кластеров из множества узлов хранения.