Решение для хранения данных RAIDIX

СХД на базе ПО RAIDIX и распределенной файловой системы Gfarm

2017

Оглавление

Резюме

ПО для систем хранения данных RAIDIX — отечественное решение для управления отдельными серверами СХД и построения масштабируемых высокопроизводительных кластеров из множества узлов хранения. Для обеспечения максимальной горизонтальной масштабируемости RAIDIX интегрируется с распределёнными файловыми системами (HyperFS, Intel Lustre, GPFS, GFarm и др.). Такие системы обеспечивают дополнительную защиту от сбоев, разделяют и реплицируют данные на многие сервера для сохранения целостности данных и высокой производительности. Кластерные файловые системы, как правило, используются в высокоскоростных вычислениях (HPC) и других отраслях, требующих максимальной доступности данных 24/7.

В данном документе описывается техническое взаимодействие ПО RAIDIX и файловой системы Gfarm. Gfarm использует OpenLDAP или PostgreSQL для метаданных и FUSE или LUFS для монтирования. Доступна в Linux, FreeBSD, NetBSD и Solaris под лицензией X11.

Далее рассматриваются особенности архитектуры и конфигурации, а также функциональные характеристики Gfarm и RAIDIX.

Введение

RAIDIX — программно-определяемая технология для организации высокопроизводительных и отказоустойчивых систем хранения, работающих с высокими нагрузками в видеонаблюдении, медиаиндустрии, корпоративном секторе, HPC и других отраслях. RAIDIX обеспечивает вертикальную масштабируемость до 1800 ТБ сырого объема и горизонтальную масштабируемость до 64 ЭБ. СХД на базе RAIDIX поддерживают работу в режиме кластера Active-Active (двухконтроллерная конфигурация) прямо «из коробки» — без использования внешних компонентов.

Одним из способов построения кластерной системы является интеграция ПО RAIDIX и распределенной файловой системы (ФС) с открытым кодом Gfarm. Данная ФС используется для крупномасштабных кластерных вычислений и обмена данными и предоставляет функционал для прозрачного управления копиями.

Название системы отсылает к архитектуре Grid Data Farm, разработанной в Японии в рамках ресурсоемкого проекта для хранения данных петабайтного уровня. Проект является плодом сотрудничества «Организации по изучению высокоэнергетических ускорителей» (KEK), «Национального института передовой промышленной науки» (AIST), Токийского университета, Технологического института Токио и Цукубского университета. Задачей разработчиков Grid Data Farm и GFarm стало построение петафлопсной и экзафлопсной параллельной файловой системы, использующей локальное хранилище ПК, распределенных по глобальной grid-системе.

Файловая система Gfarm решает проблемы производительности и надежности в NFS и AFS путем создания многих файловых копий. Система не только предотвращает деградацию производительности благодаря централизации доступа, но и поддерживает отказоустойчивость и защиту от катастроф. Уникальность Gfarm состоит в том, что каждый узел файловой системы также является клиентом Gfarm. Распределенный доступ с каждого узла файловой системы способствует высокомасштабируемой производительности I/O.

В связке с распределенной системой Gfarm ПО RAIDIX поможет системному администратору организовать кластер хранения из множества узлов на базе стандартного серверного оборудования.

Описание системы Gfarm

Gfarm — кластерная файловая система с открытым кодом, которая устанавливается на локальном диске узлов кластера или grid-структуры и обеспечивает совместный доступ к данным со всех узлов и клиентов кластера. Gfarm предоставляет масштабируемую производительность ввода-вывода для множественных параллельных процессов и пользователей. Файлы могут распределяться между всеми узлами и клиентами и физически реплицироваться и храниться на любом узле файловой системы. При этом узлы могут быть географически распределены.

Gfarm обеспечивает децентрализацию дискового доступа, отдавая приоритет локальному диску. При создании нового файла выбирается локальный диск при наличии свободного пространства. В ином случае выбирается ближайший и наименее загруженный узел. Аналогично при доступе к файлу выбирается локальный диск, если он содержит одну из файловых копий. В обратном случае выбирается ближайший и наименее загруженный узел, содержащий одну из файловых копий.

Доступ к Gfarm

Существует несколько способов получения доступа к файловой системе Gfarm:

- **Команды Gfarm и нативные API Gfarm для файловых операций ввода-вывода.** С их помощью возможно использовать специальные функции Gfarm, такие как файловая репликация, управление узлами файловой системы и др.
- **GfarmFS-FUSE (gfarm2fs).** Вы можете смонтировать файловую систему Gfarm с клиентских машин Linux, используя технологию FUSE (http://fuse.sourceforge.net/), совершенно прозрачно для клиента.
- **Плагин Gfarm Samba.** Данный плагин позволяет серверу Samba получить доступ к файловой системе Gfarm. Используя плагин, клиенты Windows могут получить доступ к Gfarm через сервис обмена файлами Windows.

- **Плагин Gfarm Hadoop.** Используя этот плагин, приложения Hadoop MapReduce могут получить доступ к файловой системе Gfarm с помощью Gfarm URL.
- **Gfarm GridFTP DSI.** Это плагин позволяет серверу Globus GridFTP и клиентам GridFTP получить доступ к файловой системе Gfarm.

Типы хостов Gfarm

- **Клиентский узел**. Терминальный узел для пользователей.
- **Узел файловой системы.** Gfarm использует хранилище данных и процессоры на узле файловой системы. Для упрощения удаленных файловых операций и контроля доступа в файловой системе на каждом узле работает «демон» ФС Gfarm (gfsd). Gfsd выполняет функции авторизации пользователя, файловой репликации, мониторинга ресурсов узла и управления системой.
- **Узел сервера метаданных.** Данный узел управляет метаданными файловой системы Gfarm. На узле запущены метасервер файловой системы Gfarm (gfmd) и бек-энд сервер базы данных, например, сервер LDAP (slapd) или PostgreSQL (postmaster).

В случае если количество хостов ограничено, возможно использовать один и тот же хост для различных целей. Физически файлы реплицируются и распределяются по дискам узлов файловой системы, обеспечивая параллельную доступность с различных рабочих машин.

Структура ПО Gfarm

Файловая система Gfarm состоит из следующих компонентов:

- **Библиотека libgfarm.a**

  Библиотека с API Gfarm, включая файловый доступ Gfarm, файловую репликацию, планирование процессов в привязке к файлам.

- **gfmd — сервер метаданных файловой системы Gfarm**

  Сервер метаданных файловой системы Gfarm (запущен на узле сервера метаданных). Управляет структурой каталогов, файловой информацией, каталогом копий, информацией о пользователях/группах пользователей, информацией о хостах. Gfmd сохраняет метаданные в памяти, при этом бекэнд базы данных (PostgreSQL или OpenLDAP) остаются в фоновом режиме.

- **gfsd — «демон» файловой системы Gfarm**

  «Демон» I/O для файловой системы Gfarm, запущенный на каждом узле файловой системы, предоставляет возможность осуществлять удаленные файловые операции с контролем доступа, а также авторизацию пользователей, репликацию файлов и проверку статуса ресурсов узла.

- **Командные инструменты Gfarm**

  Инструменты включают в себя команды файловой системы, такие как gfls, gfrm, gfwhere и gfrep; элемент управления узлами файловой системы gfhost; средства управления файлами, такие как gfreg и gfexport; средства управления ключом сеанса, такие как gfkey.

Авторизация Gfarm

gfmd и gfsd поддерживают следующие методы авторизации:

1. **sharedsecret**

   ПО Gfarm автоматически генерирует совместно используемый ключ в файле ~/.gfarm_shared_key. Данный метод авторизации легко использовать в среде, которая предоставляет общий доступ к домашним каталогам пользователей через NFS. Также используется в средах, защищенных брандмауэром.

2. **gsi**

   Метод GSI — Grid Security Infrastructure — использует публичный ключ авторизации, основанный на сертификате класса ИОК (инфраструктура открытых сетей). Этот метод позволяет зашифровать сетевую коммуникацию и подходит для использования через Интернет.

3. **gsi_auth**

   Этот способ использует GSI для авторизации, но позволяет переключиться на обычное соединение TCP по завершении авторизации. Подходит для среды, защищенной брандмауэром.

Далее мы рассмотрим архитектуру СХД под управлением RAIDIX с использованием файловой системы Gfarm.

Решение RAIDIX

В условиях лавинообразного роста объемов данных многие исследовательские центры, суперкомпьютерные кластеры и компании, работающие с большими данными, нуждаются в высокомасштабируемых решениях. Зачастую речь идет о петабайтах информации и необходимости создания кластера хранения из нескольких узлов. В этом случае ПО RAIDIX может быть использовано как платформа для развертывания распределенной кластерной системы Gfarm.

Пример двухконтроллерной архитектуры RAIDIX

RAIDIX позволяет строить высокопроизводительные системы хранения данных с использованием широко распространенных аппаратных платформ на базе стандартного серверного оборудования. Для обеспечения полной отказоустойчивости RAIDIX может функционировать в режиме двухконтроллерного кластера (Active-Active), см. рис. 1.

Рекомендуется использовать два отдельных узла для двухточечных соединений:

- 1 GbE для контрольного сигнала (HeartBeat)
- 6 x SAS 12G для синхронизации кэша на запись (CacheSync).

Оба узла соединяются с дисковыми полками JBODs (60 дисков по 10ТБ каждый) через 20 портов SAS 12G для каждого узла. Подключение к серверам Gfarm осуществляется посредством 8 x SAS 12G.

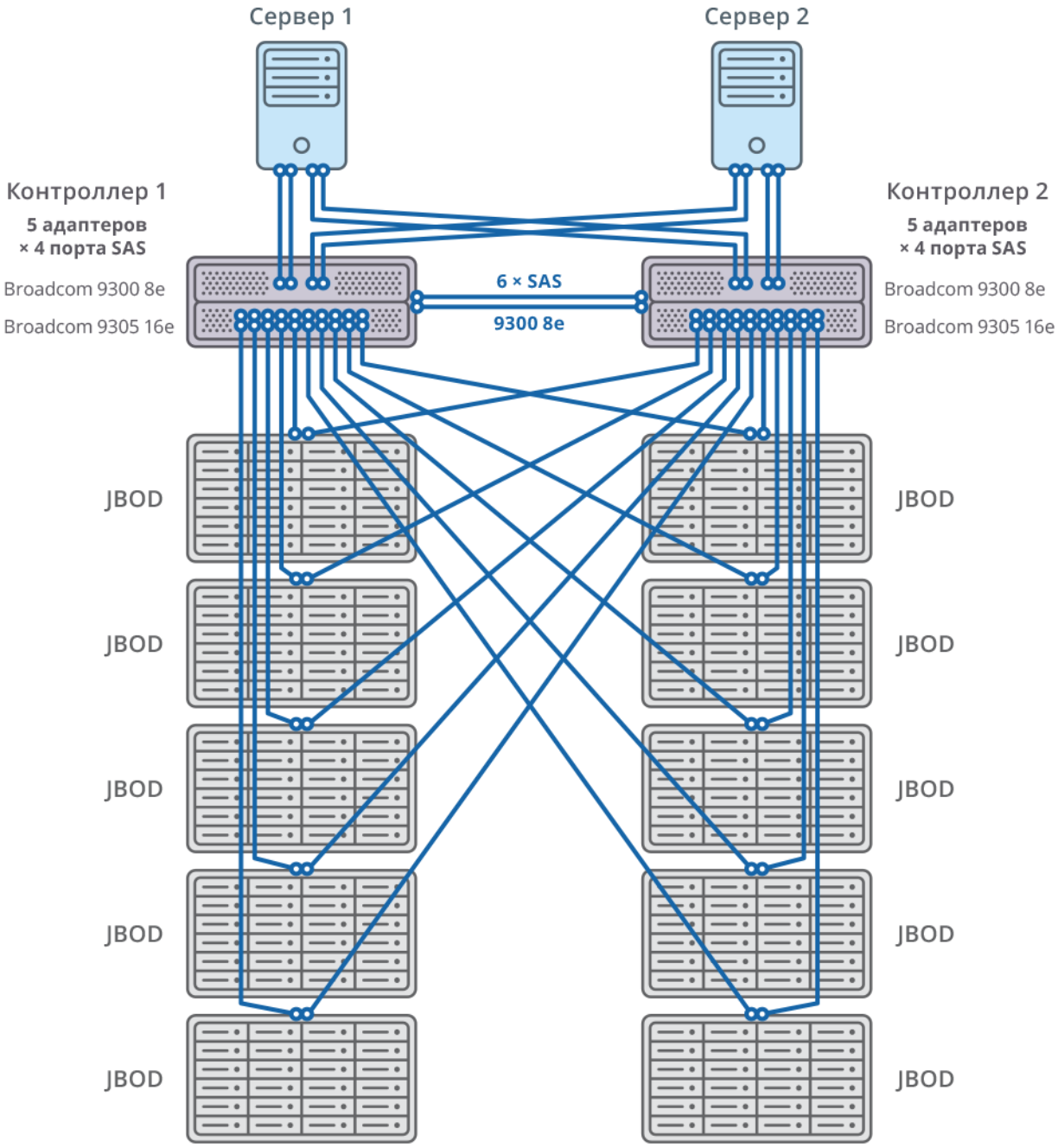

Рис. 1. Пример двухконтроллерной архитектуры RAIDIX

Общие требования для двухконтроллерной платформы RAIDIX:

CPU	Процессоры Intel Xeon E5-2637
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 256 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 9305 16e, 9300 8e
HBA (контроллер для синхронизации кэша)	Рекомендуется Broadcom 9300 8e
HDD (жесткие диски)	Для двухконтроллерной архитектуры необходимы диски SAS
Сеть Gfarm	InfiniBand QDR/FDR/EDR, Ethernet 10GbE/40GbE/100GbE
Управляющая сеть	Ethernet 1GbE

Пример проекта

Далее приводится схема проекта для крупного японского исследовательского института — масштабируемой системы хранения для совместного доступа к данным инфраструктуры HPCI (High-Performance Computing Infrastructure).

Архитектура

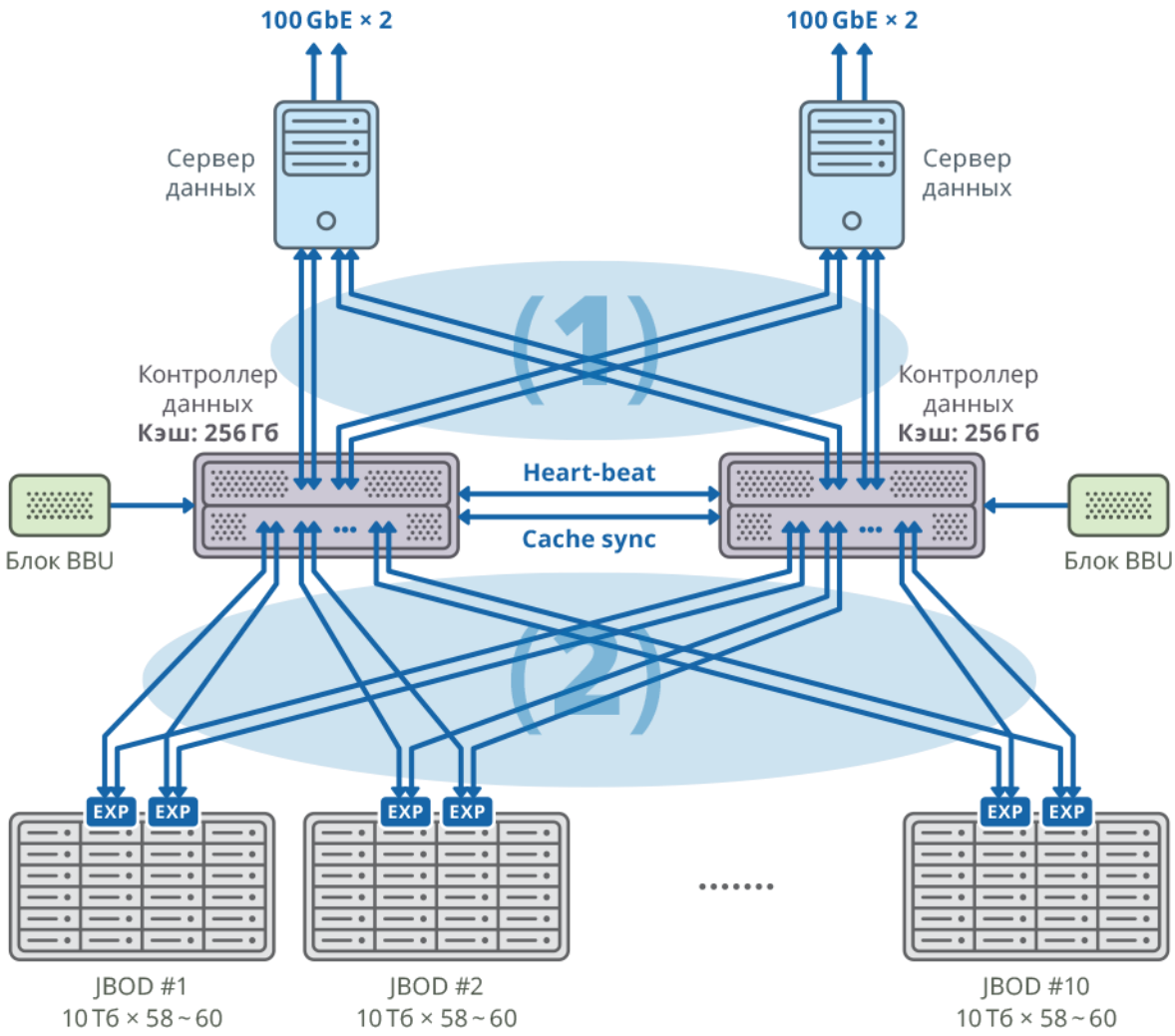

Рис. 2. Конфигурация дискового хранения с отдельным сервером данных для каждого кластера

- **(1)** 48Gbps × 8 соединений SAN Mesh connection; пропускная способность: 384Gbps
- **(2)** 48Gbps × 40 соединений Mesh FABRIC; пропускная способность: 1920Gbps

- 10 JBOD => 58 × RAID6 (8 дисков с данными (D) + 2 диска четности (P)), LUN из 580 HDD + 12 HDD для «горячей замены» (2.06% всего объема)
- 592 HDD (10ТБ SAS/7.2k HDD) на кластер * HDD: HGST (MTBF: 2 500 000 часов)
- Полезная емкость на кластер: 4,64 ПБ ((RAID6 / 8D+2P) LUN × 58) на кластер
- **Суммарная емкость всей системы: 51,04 ПБ** (4,64ПБ × **11 кластеров** => 51,04ПБ).

Производительность системы — **15 ГБ/с** на чтение/запись.

Общая схема системы и соединений

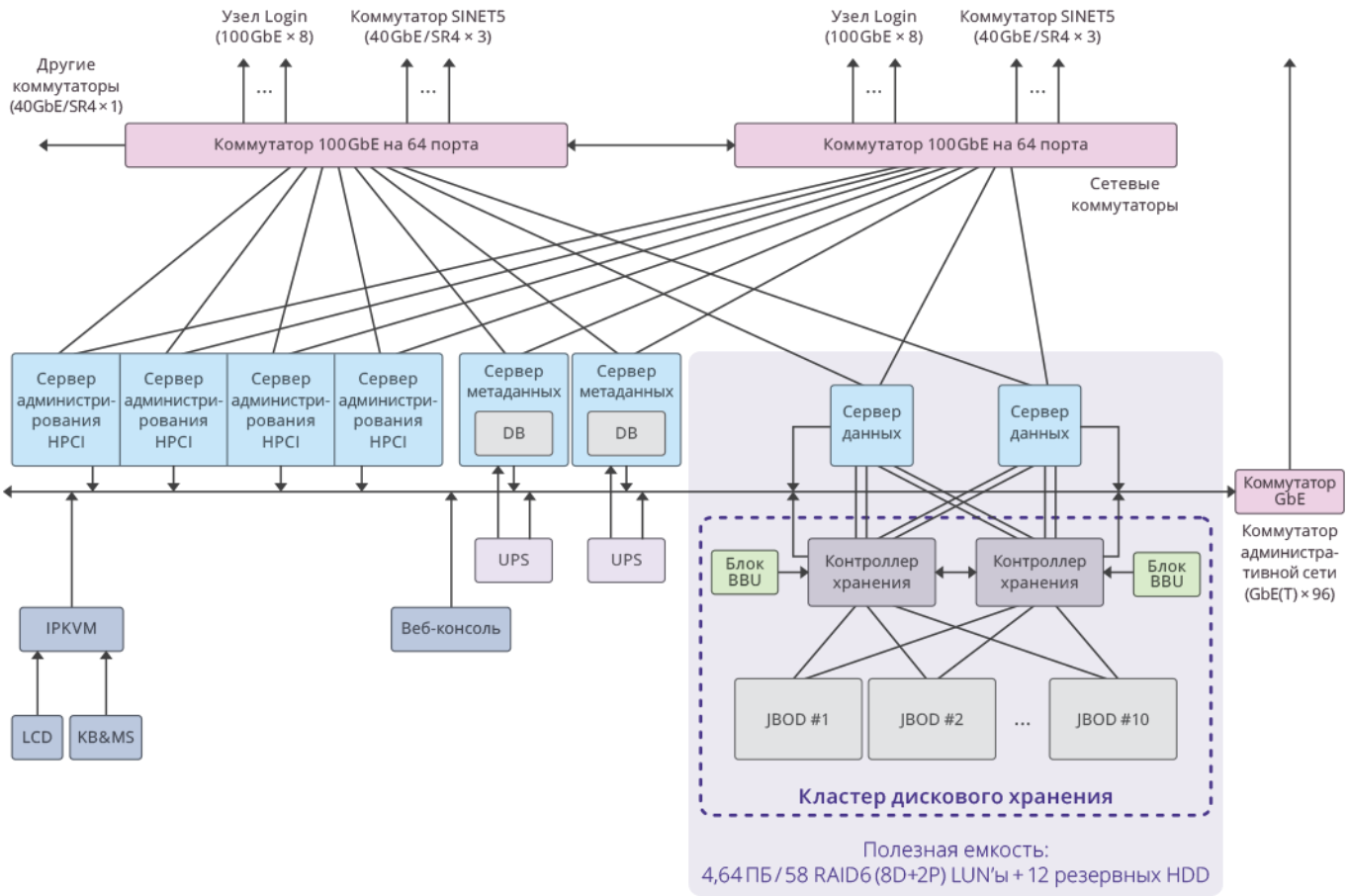

Рис. 3. Кластер дискового хранения в рамках целой системы

Результаты для бизнеса

Интегрированное решение на базе RAIDIX и Gfarm позволяет использовать множество узлов хранения (СХД), динамически распределяя информацию между ними и балансируя нагрузку. Архитектура решения позволяет добавлять к системе новые узлы хранения по требованию — без необходимости переносить данные и менять конфигурацию системы.

Технология RAIDIX в сочетании с файловой системой Gfarm удовлетворяет высочайшим требованиям по скорости и отказоустойчивости, обеспечивает одновременную работу с данными с нескольких рабочих станций. Использование RAIDIX позволяет горизонтально масштабировать действующую инфраструктуру без простоев и снижения производительности и минимизировать расходы на апгрейд оборудования при создании кластеров хранения.

О компании «Рэйдикс»

Компания «Рэйдикс» (www.raidix.ru) (осн. в 2009 году) — 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.