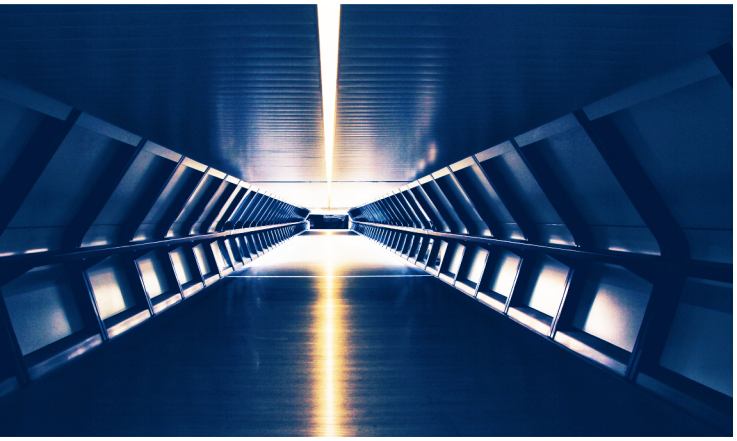
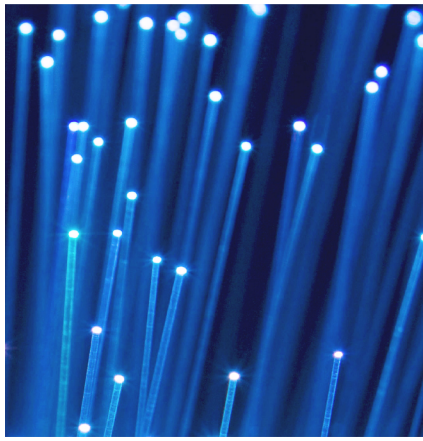
МУЛЬТИПЕТАБАЙТНАЯ ИНСТАЛЛЯЦИЯ RAIDIX В ЯПОНСКОМ ИССЛЕДОВАТЕЛЬСКОМ ЦЕНТРЕ RIKEN





О проекте

В прошлом году была реализована самая крупная на данный момент инсталляция СХД на базе RAIDIX. Система из 11 отказоустойчивых кластеров была развернута в Центре вычислительных наук института RIKEN (Япония). Основное назначение системы — хранилище для инфраструктуры HPC (HPCI), которая реализована в рамках масштабного национального проекта по обмену академической информацией Academic Cloud (на базе сети SINET).

Знаковой характеристикой этого проекта является его суммарный объем — 65 ПБ, из которых полезный объем системы составляет 51,4 ПБ. Чтобы точнее понять эту величину, добавим, что это 6512 дисков по 10 ТБ (самых современных на момент установки). Это много.

Работы над проектом шли на протяжении года, после этого еще около года длился мониторинг стабильности работы системы. Полученные показатели удовлетворили заявленным требованиям, и сейчас мы можем говорить об успешности этого рекордного и значимого для нас проекта.

Ключевые показатели проекта

- Полезная емкость на один кластер: **4.64 ПБ** ((RAID6 / 8D+2P) LUN × 58)
- Суммарная полезная емкость всей системы: **51,04 ПБ** (4,64ПБ × 11 кластеров)
- Общая емкость системы: **65 ПБ**
- Производительность системы составила: **17 ГБ/с** на запись, **22 ГБ/с** на чтение
- Суммарная производительность дисковой подсистемы кластера Gfarm на 11 СХД RAIDIX: **250 ГБ/с**

Окружение проекта

Суперкомпьютер в Центре вычислительных наук института RIKEN

Суперкомпьютер помогает Центру вычислительных наук в реализации сложнейших масштабных исследований: он позволяет осуществлять моделирование климата, погодных условий и молекулярного поведения, расчет и анализ реакций в ядерной физике, прогнозирование землетрясений и многое другое. Мощности суперкомпьютера также используются для более «повседневных» и прикладных исследований — для поиска месторождений нефти и прогнозирования трендов на фондовых рынках.

Подобные расчеты и эксперименты генерируют огромное количество данных, ценность и значимость которых нельзя переоценить. Чтобы извлечь из этого максимальную пользу, японскими учеными была разработана концепция единого информационного пространства, в котором профессионалы HPC из разных исследовательских центров будут иметь доступ к полученным HPC-ресурсам.

High Performance Computing Infrastructure (HPCI)

HPCI работает на базе SINET (The Science Information Network) — магистральной сети для обмена научными данными между Японскими университетами и научными центрами. В настоящий момент SINET объединяет около 850 институтов и университетов, создавая огромные возможности для информационного обмена в исследованиях, которые затрагивают ядерную физику, астрономию, геодезию, сейсмологию и компьютерные науки.

HPCI представляет собой уникальный инфраструктурный проект, который формирует единую систему обмена информацией в сфере высокопроизводительных вычислений между университетами и научными центрами Японии.

Объединив возможности суперкомпьютера “K” и других научных центров в доступную форму, научное сообщество получает очевидные выгоды для совместной работы с ценным данным, создаваемым суперкомпьютерными вычислениями.

Для того, чтобы обеспечить эффективный совместный доступ пользователей к среде HPCI, к хранилищу предъявлялись высокие требования по скорости доступа. А благодаря «гиперпродуктивности» K-компьютера, кластер хранения в Центре вычислительных наук института RIKEN рассчитывалось создать с рабочим объемом не менее 50 ПБ.

Инфраструктура проекта HPCI строилась на основе файловой системы Gfarm, которая позволила обеспечить высокий уровень производительности и объединять разрозненные кластеры хранения в единое пространство для совместного доступа.

Файловая система Gfarm

Gfarm — разработанная японскими инженерами распределенная файловая система с открытым кодом. Gfarm является плодом разработки Института передовой индустриальной науки и технологий (AIST), а название системы отсылает к используемой архитектуре Grid Data Farm.

Эта файловая система сочетает в себя ряд, казалось бы, несочетаемых свойств:

- Высокая масштабируемость по объему и производительности
- Распределенность по сети на большие расстояния с поддержкой единого пространства имен для нескольких разнесенных научных центров
- Поддержка POSIX API
- Высокий уровень производительности, необходимый для параллельных вычислений
- Обеспечение безопасности хранения данных

Gfarm создает виртуальную файловую систему, используя ресурсы хранения множества серверов. Данные распределяются сервером метаданных, а сама схема распределения скрыта от пользователей. Надо сказать, что Gfarm состоит не только из кластера хранения, но и вычислительного грида, использующего ресурсы тех же серверов. Принцип работы системы напоминает Hadoop: отправленная работа “опускается” на узел, где лежат данные.

Архитектура файловой системы ассиметричная. Явно выделены роли: Сервер хранения, Сервер метаданных, Клиент. Но в тоже время все три роли могут выполняться одной и той же машиной. Сервера хранения хранят множество копий файлов, а сервера метаданных работают в режиме master-slave.

Работа над проектом

Внедрением в Центре вычислительных наук института RIKEN занималась компания Core Micro Systems — эксклюзивный поставщик RAIDIX в Японии. Для реализации проекта потребовалось около 12 месяцев кропотливой работы, в которой принимали активное участие не только сотрудники Core Micro Systems, но и технические специалисты команды «Рэйдикс».

В ходе длительных тестов, проверок и доработок RAIDIX продемонстрировал стабильно высокую производительность и эффективность при работе с такими объемами данных.

Про доработки стоит рассказать немного подробнее. Требовалось не только создать интеграцию СХД с файловой системой Gfarm, но расширить некоторые функциональные характеристики программного обеспечения. Например, для соответствия установленным требованиям технического задания пришлось в кратчайшие сроки разрабатывать и внедрять технологию автоматической сквозной записи (Automatic Write-Through).

Само развертывание системы проходило и планомерно. инженеры из Core Micro Systems очень внимательно и аккуратно проводили каждый этап испытаний, постепенно увеличивая масштаб системы.

В августе 2017 года была завершена первая фаза развертывания, когда объем системы достиг 18 ПБ. В октябре этого же года была реализована вторая фаза, при которой объем поднялся до рекордных 51 ПБ.

Архитектура решения

Конфигурация двухконтроллерной платформы:

CPU	Intel Xeon E5-2637 - 4шт
Материнская плата	Совместима с моделью процессора с поддержкой PCI Express 3.0 x8/x16
Внутренняя кэш-память	256 ГБ для каждого узла
Шасси	2U
SAS-контроллеры для подключения дисковых полок, серверов и синхронизации кэша на запись	Broadcom 9305 16e, 9300 8e
HDD	HGST Helium 10TB SAS HDD
Синхронизация HeartBeat	Ethernet 1 GbE
Синхронизация CacheSync	6 x SAS 12G

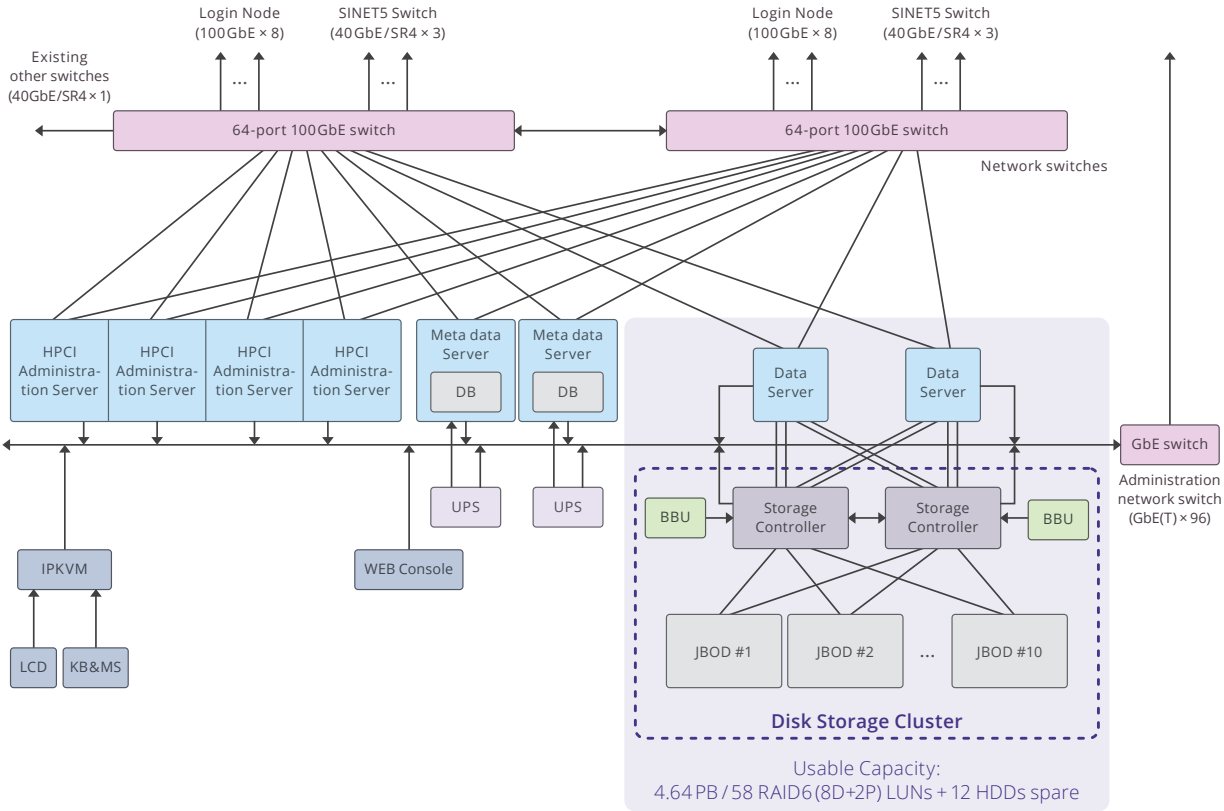


Рис. 1. Изображение одного кластера в рамках системы HPCI

Решение было создано на основе интеграции СХД RAIDIX и распределенной файловой системы Gfarm. В связке с Gfarm удалось создать масштабируемое хранилище с использованием 11 двухконтроллерных систем RAIDIX.

Подключение к серверам Gfarm осуществляется посредством 8 x SAS 12G.

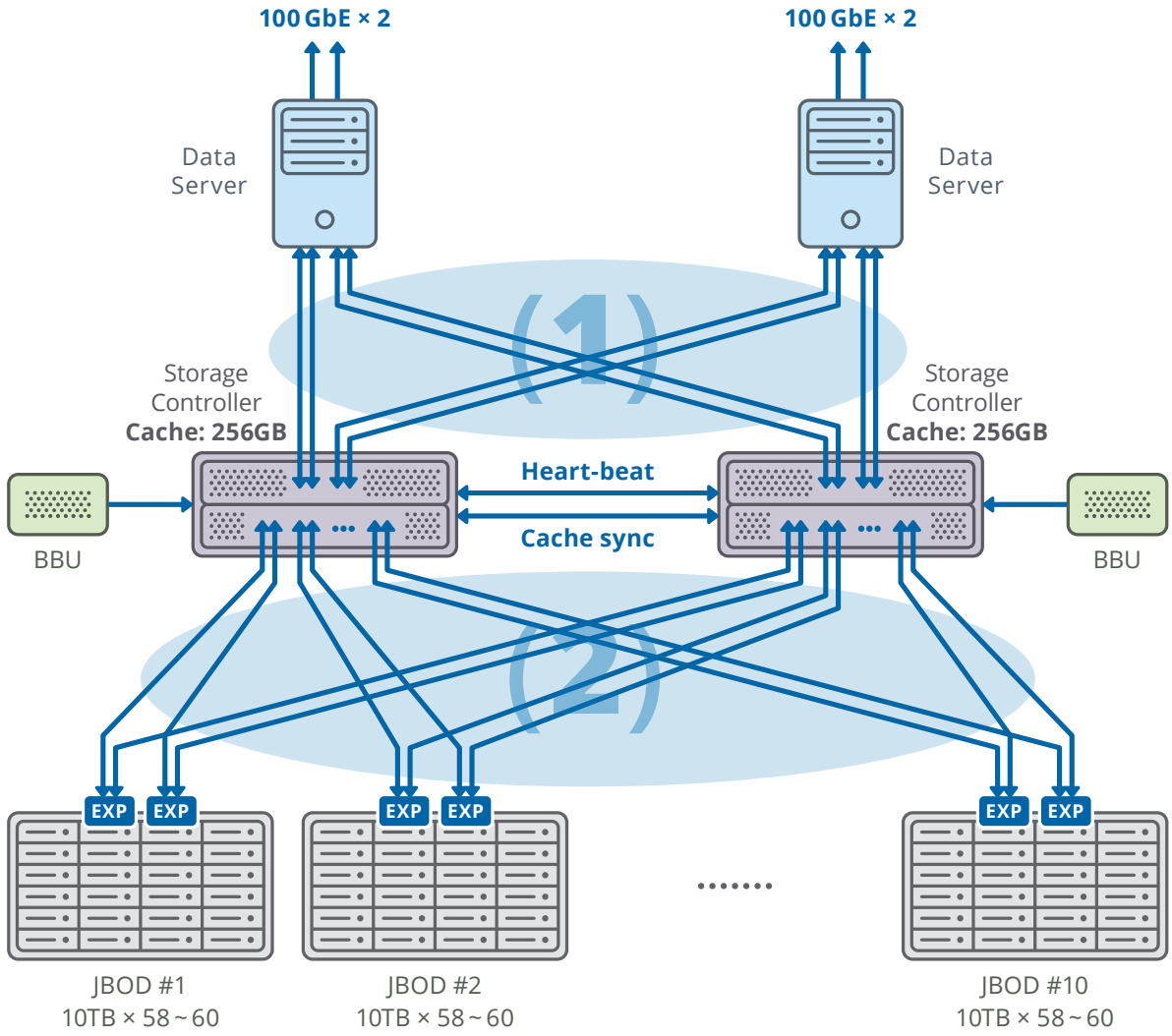


Рис. 2. Изображение кластера с отдельным сервером данных для каждого узла

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

Оба узла отказоустойчивого кластера соединяются с 10 JBOD (60 дисков по 10ТБ каждый) через 20 портов SAS 12G для каждого узла.

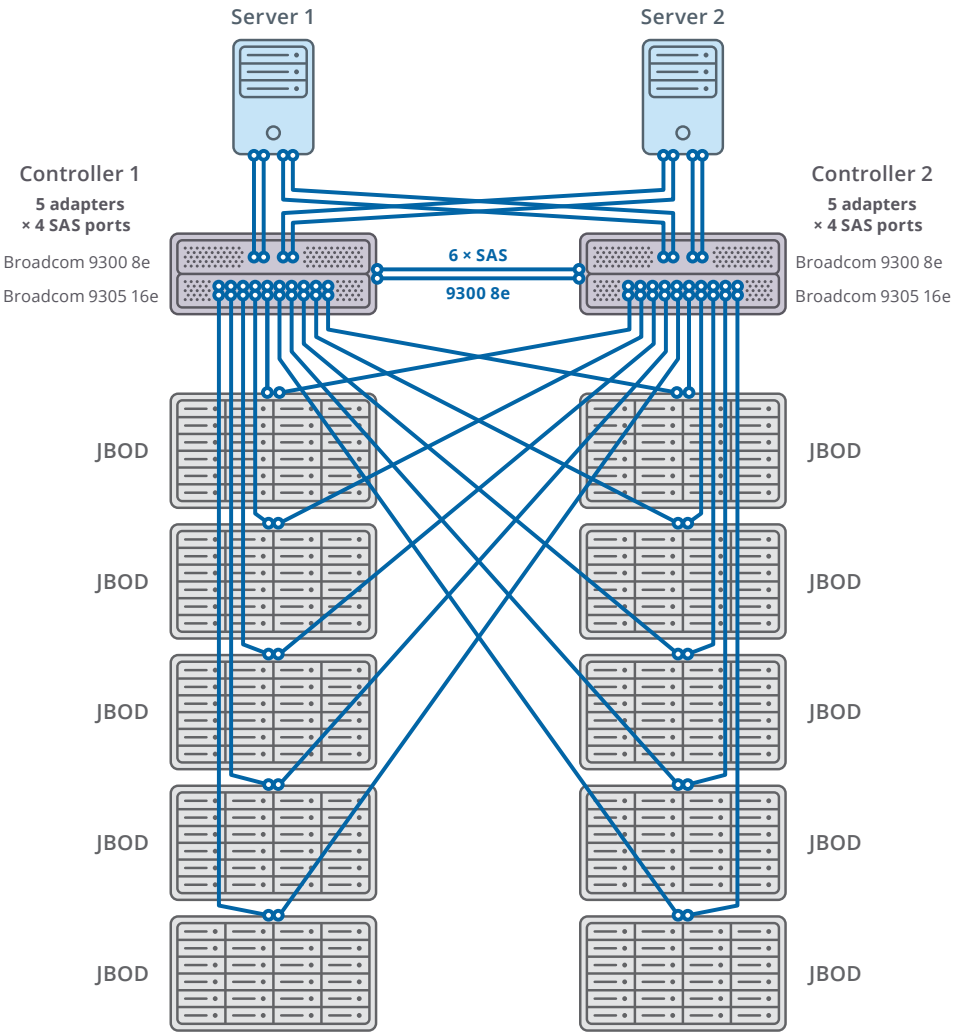


Рис. 3. Отказоустойчивый кластер со схемой подключения 10 JBOD

На этих дисковых полках было создано 58 массивов RAID6 по 10ТБ (8 дисков с данными (D) + 2 диска четности (P)) и 12 дисков было выделено под «горячую замену».

«Рэйдикс» — компания-разработчик программного обеспечения для управления высокопроизводительными системами хранения данных. Системы на базе RAIDIX отличаются высокой скоростью обработки последовательных нагрузок и востребованы в сфере видеопроизводства, в проектах видеонаблюдения, в инфраструктурах суперкомпьютерных вычислений и других высоконагруженных отраслях.

E-mail: request@raidix.ru
www.raidix.ru