Проектирование и внедрение системы хранения данных вычислительного кластера ИФВЭ на базе OpenAFS

Цель работы: увеличение доступного объема хранимых данных вычислительного кластера ИФВЭ, повышение надежности хранения данных, введение в эксплуатацию нового оборудования и программного обеспечения.

Описание: OpenAFS является распределенной файловой системой. Она обеспечивает защищенное и отказоустойчивое хранение файлов. Обладает широким спектром средств для управления доступом пользователей к данным и достаточной гибкостью в управлении. На вычислительном кластере ИФВЭ преимущественно используется для хранения данных пользователей, экспериментальных данных, программного обеспечения экспериментов ИФВЭ, сайта научно-технической библиотеки ИФВЭ.

Для достижения поставленной цели решено перенести существующую систему хранения данных OpenAFS на новую аппаратно-программную базу. С весны 2013 года по весну 2014 года мною выполнены или принято участие в выполнении (помечено (*)) следующих задач:

1. Проектирование отказоустойчивой системы OpenAFS ИФВЭ (*)
2. Настройка управляющих и дисковых серверов
3. Тестирование
   - Тестирование дисковых массивов (*)
   - Тестирование отказоустойчивости
   - Тестирование производительности
   - Запуск полнофункциональной системы в пилотном режиме
4. Ввод системы в эксплуатацию
   - Подготовка резервной копии и перенос пользовательских данных
   - Настройка счетных узлов кластера ИФВЭ для использования новой системы OpenAFS (*)
   - Вывод устаревшей системы OpenAFS из эксплуатации

В работе использовалось следующее оборудование:

- Дисковые сервера Dell PowerVault MD3200 емкостью 5Тб (2 ед.);
- Управляющие сервера Supermicro SYS-6026T-URF (2 ед.).

Результатом проектирования системы OpenAFS для вычислительного кластера ИФВЭ стала схема организации оборудования и программной части системы OpenAFS. Она представлена на Рис. 1.

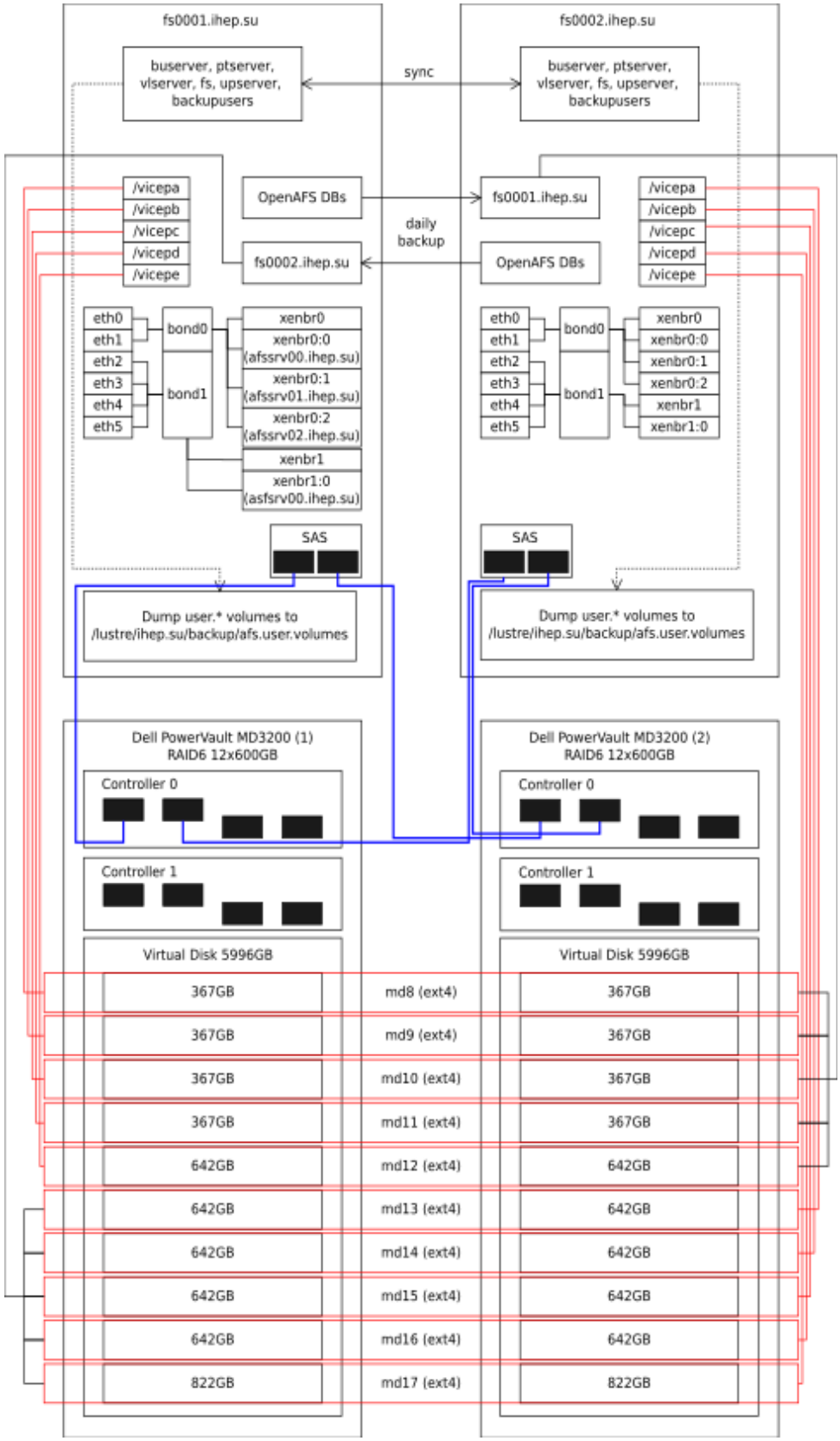

Рис. 1: Схема OpenAFS ИФВЭ. Внизу показана разметка дисковых массивов и подключение к управляющим серверам. Вверху представлена блок-схема управляющих серверов с указанием процессов, сетевых интерфейсов и настроек синхронизации.

Настройка дисковых массивов. Каждый дисковый сервер был разделен на 10 частей, которые были объединены на управляющих серверах fs0001.ihep.su и fs0002.ihep.su в программные массивы RAID-1. Впоследствии они были поровну разделены между управляющими серверами, что позволило добиться отказоустойчивого доступа ко всему предоставляемому объему данных. Неравное разделение дискового пространства позволило более гибко управлять выделяемыми квотами для пользователей OpenAFS. На устройствах RAID-1 создана файловая система ext4.

Настройка управляющих серверов. На управляющих серверах установлена операционная система Debian GNU/Linux 7 Wheezy, выполнена её основная настройка (автоматическое отключение при низком заряде батарей, настройки доступа, политики обновлений, синхронизации времени, …). Для обеспечения высокой пропускной способности и скорости доступа к данным 6 подключенных сетевых интерфейсов *eth#* объединены в иерархическую структуру логических сетевых интерфейсов *bond#* и *xenbr#*. Это позволило поддерживать несколько доменных имен и запущенных виртуальных машин одновременно. Канал с большей пропускной способностью предназначен для использования счетными узлами и пользователями ИФВЭ, с меньшей пропускной способностью – для нужд задач GRID.

Установлены и настроены программные компоненты OpenAFS, синхронизированы список пользователей, права доступа к данным, настроено создание кратко- и долгосрочных резервных копий пользовательских данных.

В качестве отказоустойчивого решения использовали виртуализацию Xen. Для этого созданы точные копии управляющих серверов, каждая из которых хранится на парном управляющем сервере. Настроена регулярная синхронизация каждого сервера с его копией. В случае аппаратного отказа одного из управляющих серверов происходит автоматический запуск его копии на другом управляющем сервере. В случае программного сбоя каждый управляющий сервер способен полностью обеспечить функциональность системы и доступ к данным.

Тестирование. Тестирование программно-аппаратной части дисковых массивов проводилось запуском непрерывных циклов записи-чтения случайных данных в течение одного месяца. Аппаратных и программных ошибок выявлено не было. Отказоустойчивость дисковых массивов не проверялась ввиду наличия у них сдвоенного электропитания.

Тестирование отказоустойчивости системы OpenAFS проводилось случайным отключением управляющих серверов и наблюдением за изменением доступности и скорости доступа к данным. Во всех случаях доступ к данным сохранялся, время простоя не превышало допустимого порога в две минуты. При тестировании отлажены сценарии запуска виртуальных машин управляющих серверов, минимизировано время простоя доступа к данным.

В дальнейшем система OpenAFS была запущена в пилотном режиме для проверки интеграции в существующую инфраструктуру вычислительного кластера ИФВЭ и определения ее рабочей производительности. На управляющих серверах настроены авторизация пользователей через службу Kerberos, система запуска задач PBS Torque, система архивирования Amanda. Два счетных узла кластера переведены на выполнение только задач GRID и перенастроены на использование серверов спроектированной системы OpenAFS. Проведены запуски задач GRID и тестовых задач пользователей ИФВЭ, показавшие работоспособность и функциональность системы в целом. Измеренная скорость доступа к данным пользователей составила 15-20 Мб/сек, что выше показателей старой системы OpenAFS.

Введение в эксплуатацию. Для ввода новой системы OpenAFS в эксплуатацию сделано архивирование всех пользовательских данных на файловую систему Lustre. Проведен перенос и проверка целостности пользовательских данных с устаревшей системы на новую

систему OpenAFS без потери пользователями доступа к данным («живая» миграция).

Для счетных узлов кластера ИФВЭ изменены настройки доступа к OpenAFS.

После ввода в эксплуатацию новой системы OpenAFS старые сервера выключены, сетевые адреса переданы новым управляющим серверам для поддержания обратной совместимости с клиентами.

Результаты:

Результатом выполненной работы является:

- Введение в эксплуатацию нового оборудования и программного обеспечения;
- Увеличение доступного дискового пространства системы OpenAFS;
- Увеличение скорости доступа к данным;
- Повышение надежности хранения данных за счет предупреждения аппаратных и программных сбоев;
- Возможность распределения нагрузки между управляющими и дисковыми серверами;
- Создание основы для расширения системы и увеличения дискового пространства в будущем.

Результаты работы представлены на 24 Международном симпозиуме по ядерной электронике и вычислительной технике (XXIV International Symposium on Nuclear Electronics & Computing), г. Варна, Болгария.