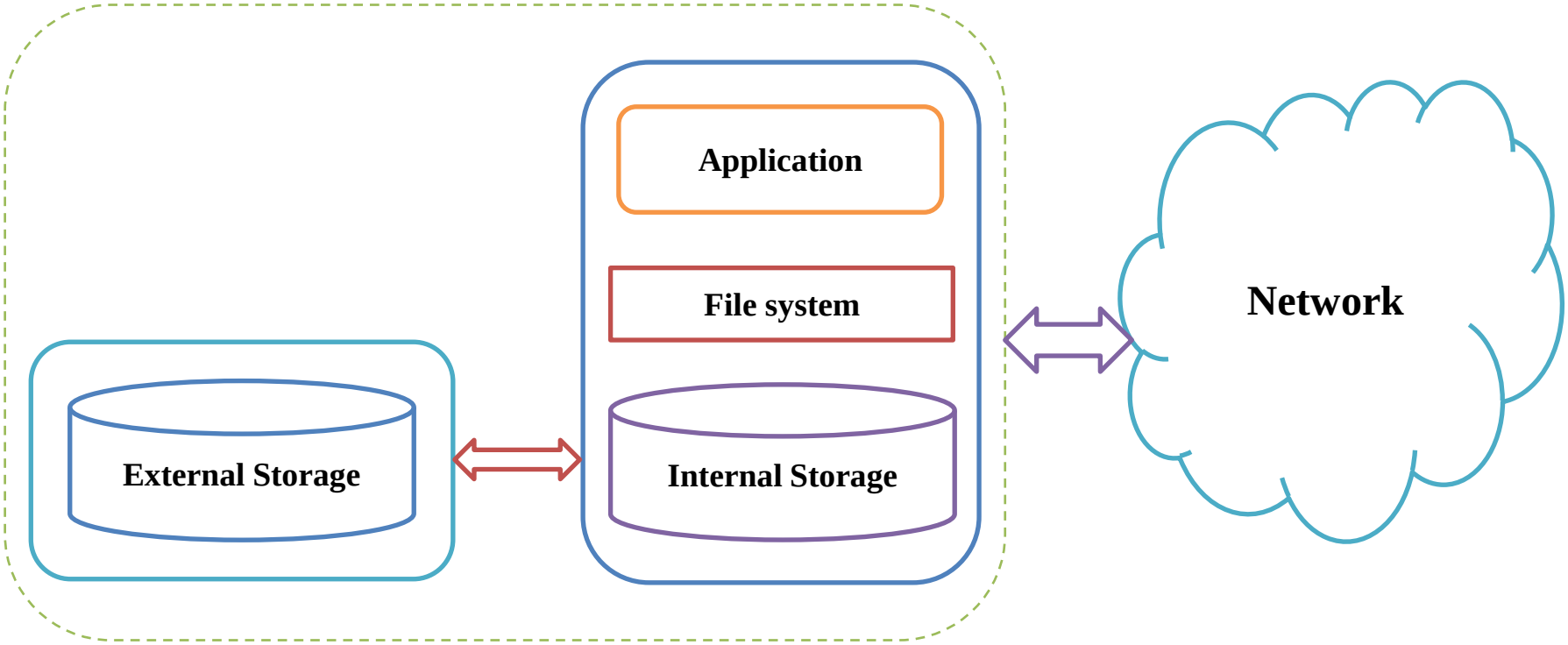
Обзор решений на рынке открытого ПО для создания СХД

Александр Клыга.

Минск, Беларусь.

DAS/SAS Model of Data Storage
DAS/SAS - Direct-Attached Storage/ Server-Attached Storage



Ключевой особенностью подключения блочных устройств в СХД модели DAS/SAS является непосредственное подключение к внутренним или внешним портам соответствующих интерфейсов с использованием необходимых кабелей. В этой модели все необходимое ПО для работы СХД устанавливается непосредственно на сервер, а блочные устройства форматируются в соответствии с требованиями к файловой системы.

ZFS
Zettabyte File System

- 128-битная масштабируемая файловая система.
- Максимальный размер файловой системы - 256 зеттабайт (10^{21} байт).
- Модель хранения данных – создание единого пула устройств хранения данных.
- Модель управления пулом – диспетчер томов (полный контроль над физическими и логическими томами).
- Модель управление данными – транзакционная файловая система (управление данными осуществляется с использованием семантики копирования при записи, данные никогда не перезаписываются, и любая последовательность операций либо полностью выполняется, либо полностью игнорируется).
- Модель репликации данных – RAIDZ (поддерживается динамический размер strip, есть возможность выбора уровня контроля четности).
- Поддерживается механизм самовосстановления данных, с различными уровнями избыточности данных, включая зеркальное отражение и варьирование.
- Поддерживается возможность создания «снимков» файловой системы или томов.
- Поддерживается облегченная модель администрирования.

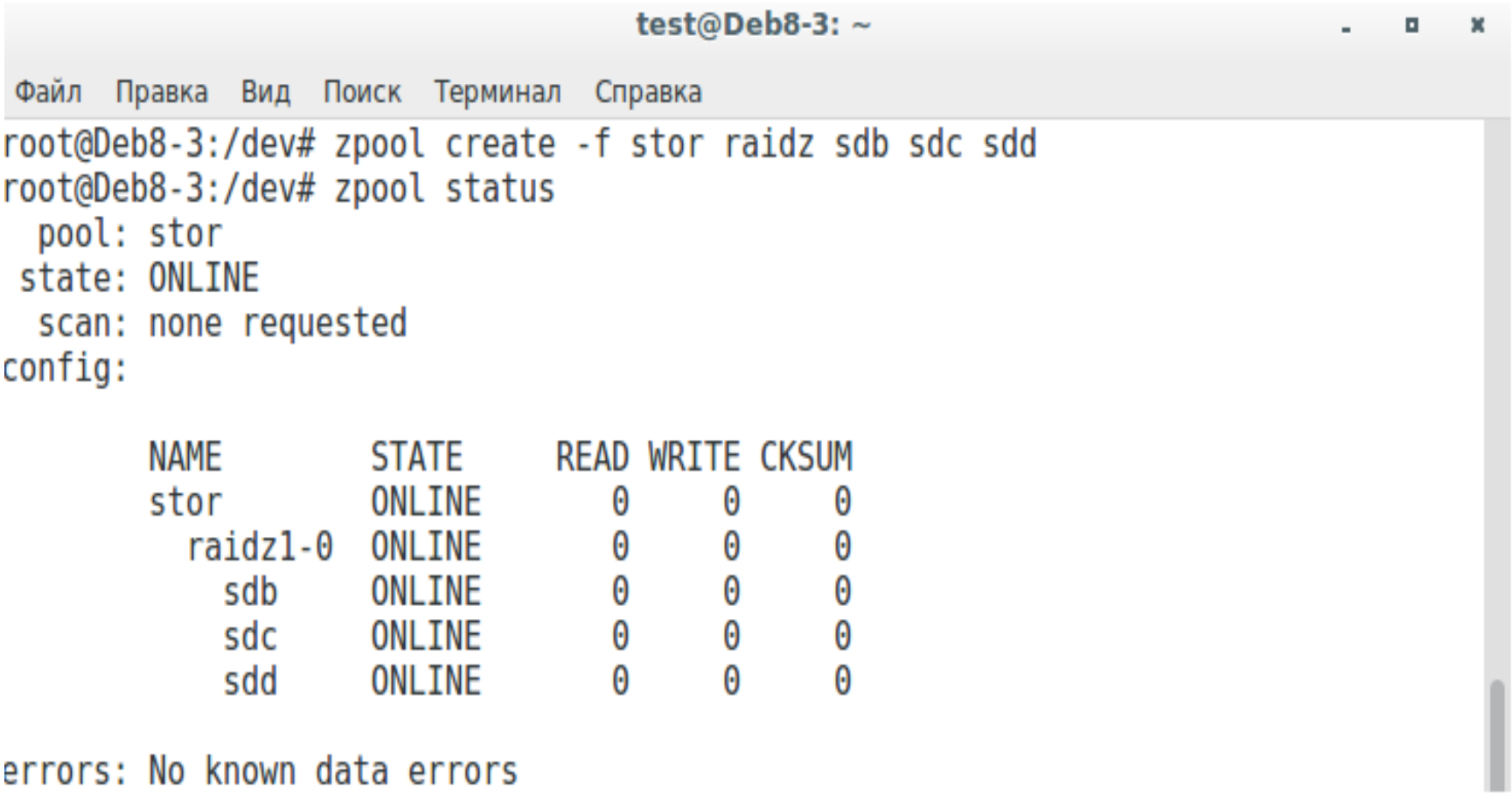
Open ZFS Project

- Проект по поддержки открытой спецификации файловой системы ZFS (Open source to the ZFS project).
- Дата официального анонса проекта – **Сентябрь 2013**.
- Объединяет разработчиков различных компаний включая llumos, FreeBSD, Linux, и OS X, использующих ZFS в своих разработках и продуктах.
- Официальная страница в сети Интернет – www.open-zfs.org
- Поддержка ZFS реализована в дистрибутивах Solaris и базе проекта OpenSolaris (IllumosOS, SmartOS, NexentaOS, OmniOS и другие.)
- В настоящие время ZFS портирована на ОС FreeBSD (https://wiki.freebsd.org/ZFS), Mac OS X (https://openzfsonosx.org/), Linux (http://zfsonlinux.org/).
- Данный проект активно развивается и используется для создания различных решений в области СХД, в том числе для облачной инфраструктуры.

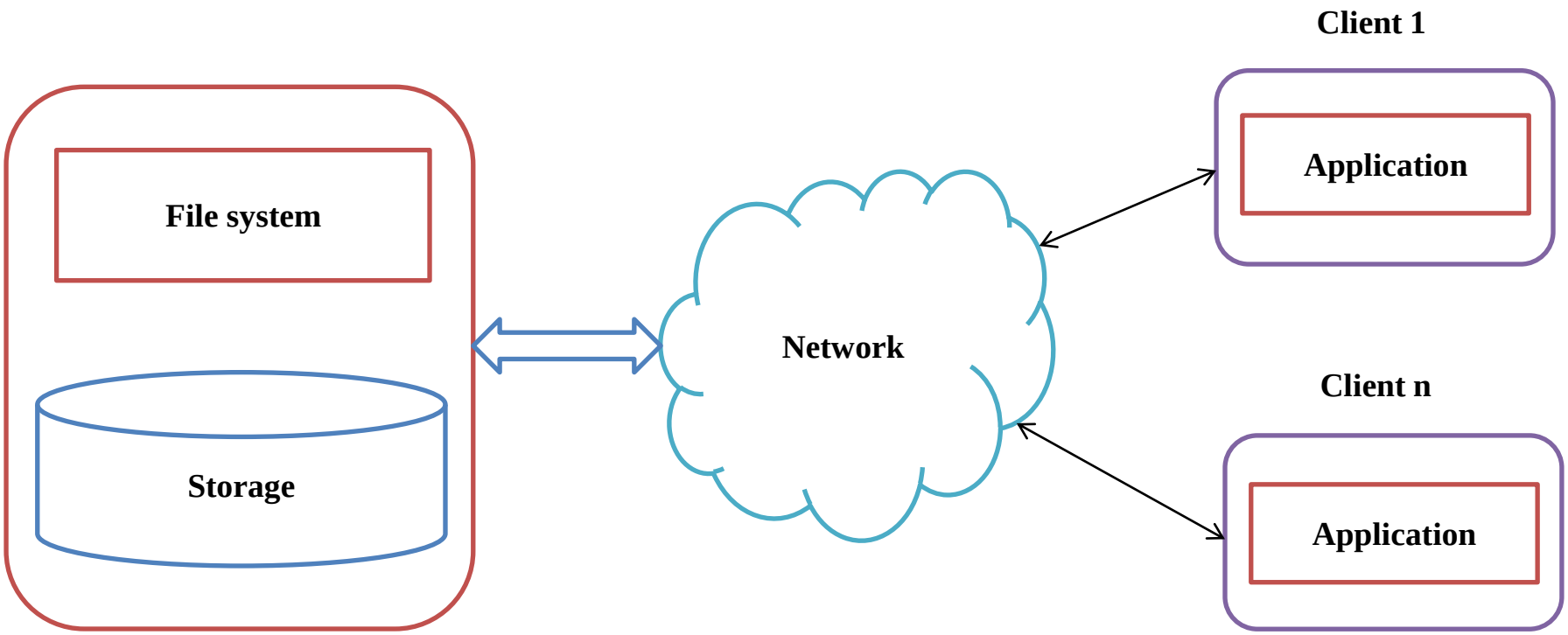
Преимущества использования ZFS в СХД

- Создание масштабируемого СХД большой емкости и высокой доступности.
- Обеспечение высокой отказоустойчивости в работе СХД, и оперативного мониторинга работы.
- Возможность реализации гибких механизмов резервного копирования.
- Возможности защиты данных (ACL списки).
- Возможность гибкого администрирования с делегированием полномочий.
- Поддержка протоколов NFS и SMB.
- Поддержка протокола iSCSI (проекты на базе OpenSolaris, FreeBSD).

Пример работы с zfs установленной на дистрибутив Debian 8.3



NAS Model of Data Storage
NAS - Network Attached Storage

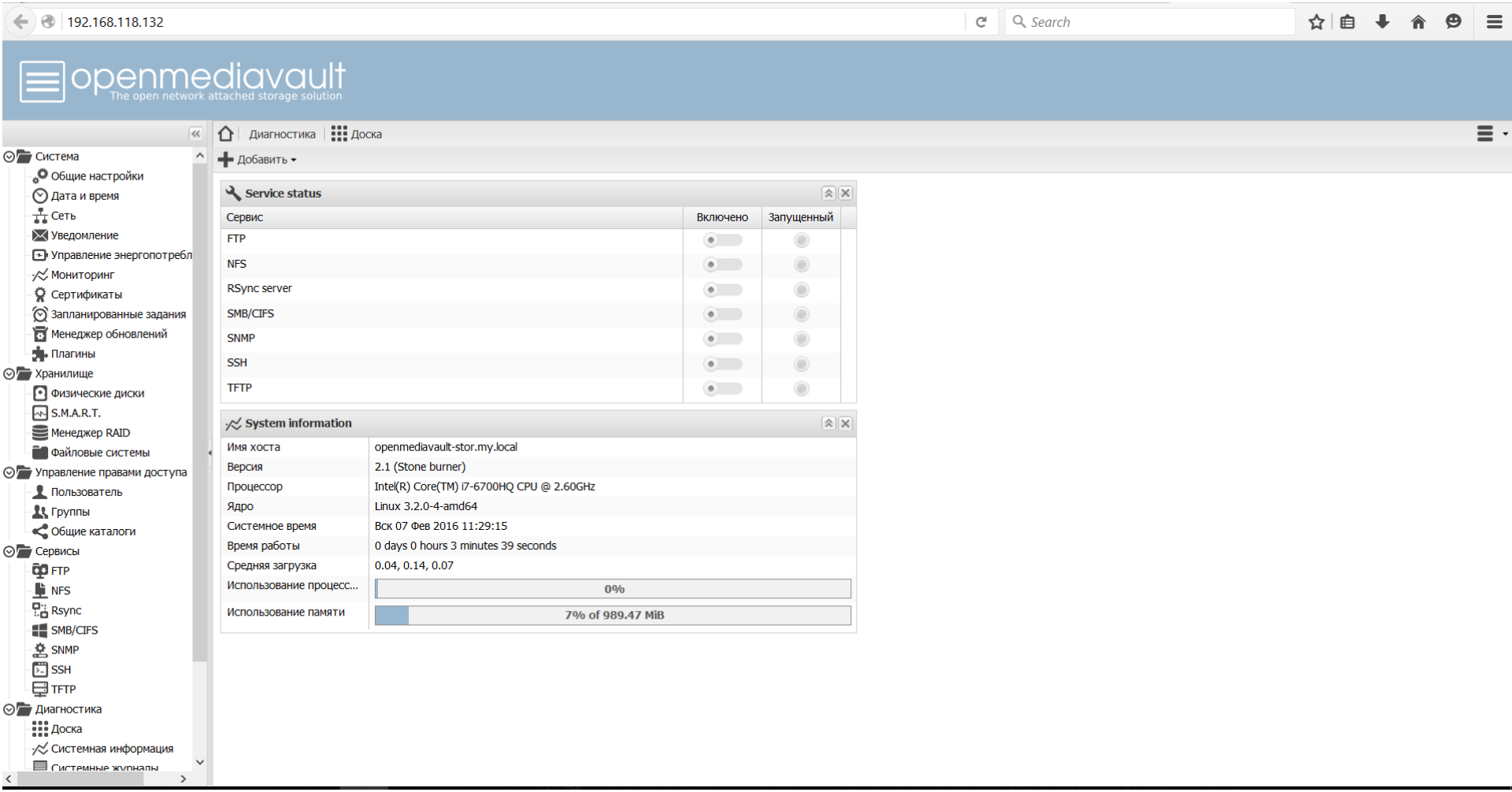


Ключевой особенностью модели СХД NAS является использование сетевой инфраструктуры с поддержкой протоколов NFS (Network File System), DAFS (Direct Access File System), SMB/CIFS (Server Message Block/Common Internet File System) или аналогичных для подключения клиентов к сетевому хранилищу данных. При этом клиентские узлы должны содержать необходимо ПО для взаимодействия с СХД (обеспечивать поддержку необходимых протоколов).

OpenMediaVault

- Проект **OpenMediaVault** с открытым программным кодом на базе дистрибутива Debian (официальный сайт проекта http://www.openmediavault.org/), ориентирован на создание СХД модели NAS.
- Включает в себя следующие компоненты: программный RAID (0,1,5,6,10,JBOD), почтовый клиент, средства управления пользователями, и мониторинга работы.
- Реализована поддержка протоколов SSH, FTP, SMB/CIFS, NFS, RSYNC.
- Возможности OMV могут быть расширены плагинами, например, такими как DAAP медиа-сервер, ISCSI, BitTorrent-клиент и другими.
- Управление NAS-сервером осуществляется через web-интерфейс, поддерживающий многоязычность (включая русский язык).
- Осуществляется поддержка диском MBR и GPT и файловых систем ext4/XFS/JFS
- В настоящий момент разработка и развитие OMV версии 2 приостановлено (последняя стабильная версия 2.1), и полным ходом идет работа над новой версией OMV 3.0 с кодовым названием «Erasmus».
- Ключевой особенностью 3 версии OMV станет адаптация проекта под новую версию Debian 8 и расширение возможностей за счет переработки старых и выпуска новых плагинов.
- Руководит проектом Волкер Тейле (Volker Theile) (бывший основной разработчик FreeNAS).

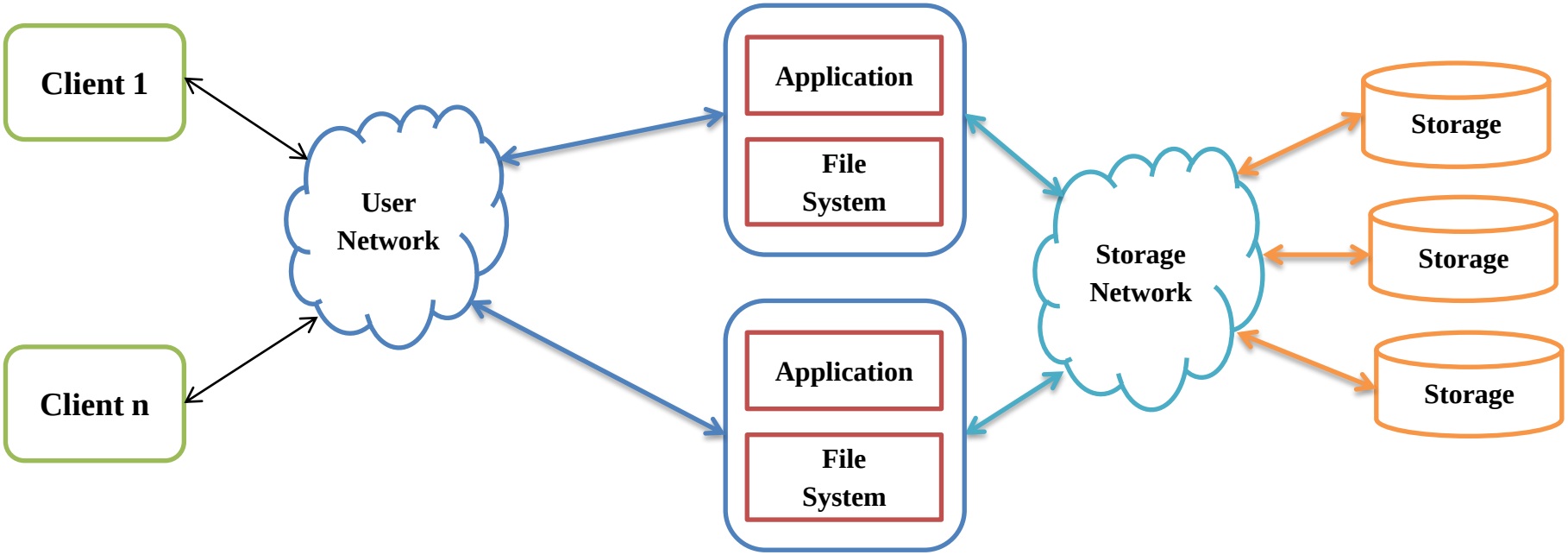
Пример web-интерфейса OpenMediaVault



Дистрибутивы на базе FreeBSD для построения NAS

- **FreeNAS** – наиболее известный дистрибутив для СХД модели NAS на дистрибутиве FreeBSD. Базовая файловая система zfs с поддержкой шифрования начиная с версии 8.3.1. Данное решение поддерживает протоколы: iSCSI, FTP/FTPS/TFTP, NFS, Samba, AFP (Apple Filing Protocol), SSH и RSYNC. Реализована поддержка программного RAID (0, 1, 5, 6, 10, 60), RAID-Z1/Z2/Z3, импорт дисков отформатированных в FAT, NTFS, EXT2/3, UFS RAID. Для авторизации клиентов используется LDAP / Active Directory.
- **NAS4Free** – еще один проект СХД модели NAS на дистрибутиве FreeBSD. Текущая версия 10.2.0.2 базируется на файловой системе zfs, с поддержкой шифрования и созданием RIADZ-массивов. Реализована поддержка протоколов SMB/CIFS, FTP, TFTP, NFS, AFP, iSCSI (initiator и target), SCP (SSH), BitTorrent, HAST, CARP, синхронизация данных посредством RSYNC (клиент/сервер) или Unison, UPnP (на базе Fuppes), CARP, HAST, VLAN и Wake On LAN. Поддерживается управление доступом на основе пользователей и групп, для аутентификации используется внутренняя база и средства Active Directory и LDAP.
- **ZFSGuru** – еще один интересный проект, в основе которого лежит дистрибутив FreeBSD и файловая система zfs. Так же реализовано создание программного RAID (0, 1, 5, JBOD, 5+0, 5+1, 0+1, 1+0), RAID-Z1/Z2 и поддержка файловых систем UFS и ext2/ext3. Доступ к данным реализуется посредством протоколов iSCSI (initiator и Target), SMB/CIFS, NFS, SSH, RSYNC (клиент и сервер) и AFP. Поддерживается возможность использования SSD в качестве кеширующего устройства (ZFS L2ARC). Предусмотрено применение резервных дисков, которые будут активированы автоматически в случае выхода из строя одного из дисков массива. Поддерживается управление учетными записями пользователей и групп, и аутентификация средствами Active Directory и LDAP.

SAN Model of Data Store
SAN - Storage Area Network

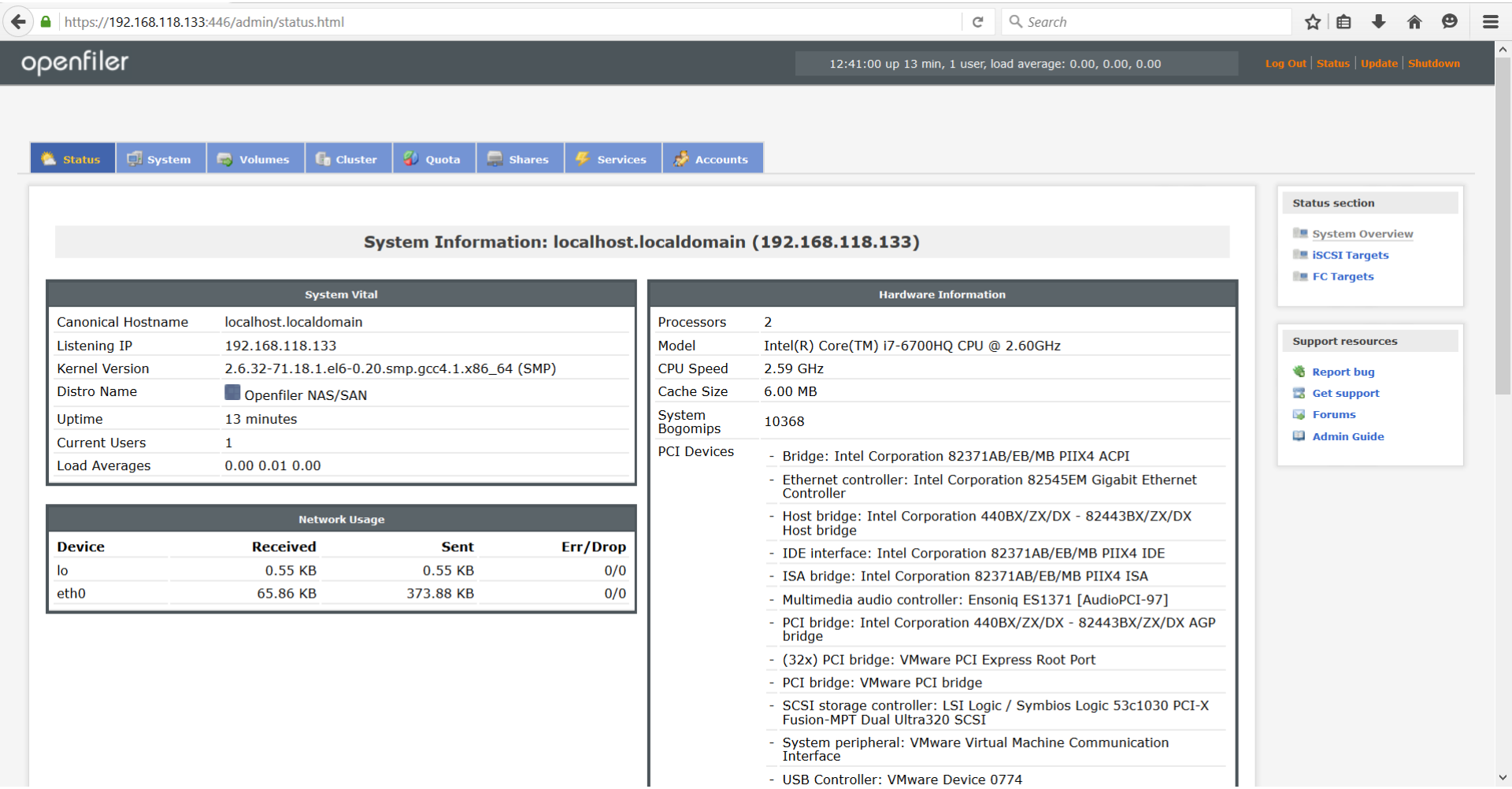


Ключевой особенностью модели СХД SAN является консолидированное использование блочных устройств хранения объединённых между собой сетевой инфраструктурой систем хранения для совместного использования группой серверов. Для построения инфраструктуры систем хранения используются различные протоколы и интерфейсы, например, протокол FC (Fibre Channel) и разновидности на его основе iFCP, FCoE, протокол SCSI и его разновидности на его основе iSCSI (Internet SCSI), SAS (Serial Attached SCSI) и другие.

Openfiler

- Проект **openfiler** (официальный сайт проекта http://www.openfiler.com/) основанный на интересном дистрибутиве rPATH Linux. В настоящий момент для скачивания доступна версия 2.99.
- Ключевой особенностью данного проекта является возможность создания СХД моделей NAS/SAN с поддержкой интерфейсов FC (Fibre Channel) и iSCSI большой емкости (более 60TB).
- Openfiler поддерживает создание и обслуживание программных RAID-массивов различных уровней (0,1,5,6 и 10), позволяет создавать и управлять кластерами.
- Данное решение предоставляет широкий круг возможностей для гибкого управления сетевыми сервисами с поддержкой протоколов NFS, CIFS, HTTP/DAV, FTP, rsync.
- Реализована поддержка сетевых каталогов: NIS, LDAP, Active Directory (в обычном и совмещенном режиме), контроллер домена Windows NT 4 и Hesio.
- Подключение по web-интерфейсу осуществляется по защищенному протоколу https. Поддержка русского языка отсутствуют.
- Все исходные кода проекта openfiler открыты и могут быть использованы для разработки и модификации собственных проектов.

Пример web-интерфейса openfiler

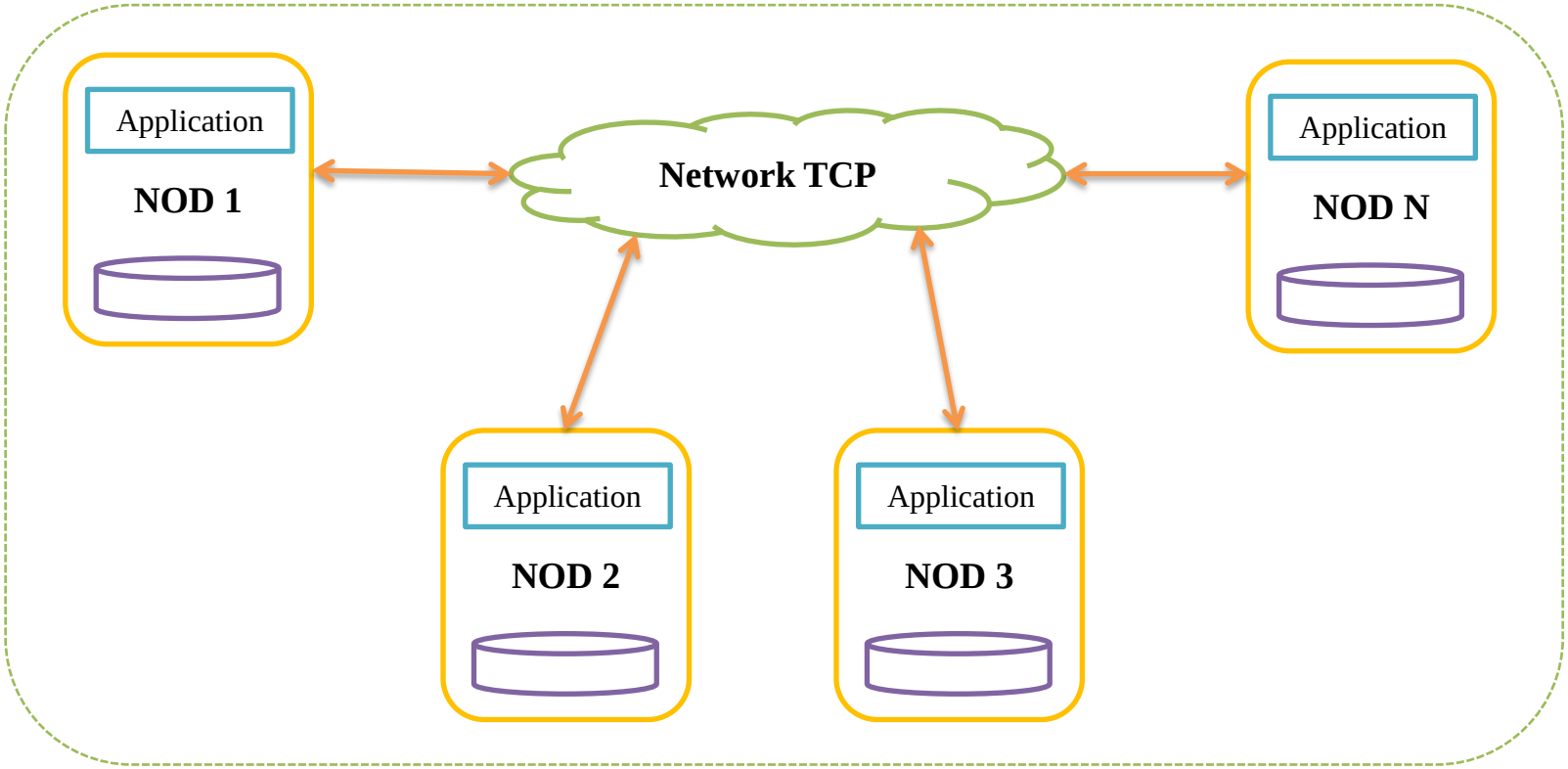


Решения для построения СХД модели SAN на базе дистрибутивов OpenSolaris.

Несмотря на то что проект OpenSolaris компанией Oracle был закрыт в 2010 году, он нашел своих последователей и на его основе был создан проект illumos (официальный сайт http://lllumos.org/). Создатели данного проекта активно поддерживают развитие файловой системы zfs, и по состоянию на сегодня в дистрибутивах на основе illumos наиболее полно реализована поддержка файловой системы zfs, и интерфейсов подключения к СХД модели SAN в частности FC, SCSI и iSCSI (initiator и Target). Наиболее популярными дистрибутивами для создания СХД являются OmniOS и NexentaOS.

- **OmniOS** классический дистрибутив с минимальным набором пакетов и с поддержкой ZFS, KVM, Crossbow, Dtrace и контейнеров (официальный сайт проекта http://omnios.omniti.com/).
- **NexentaOS** коммерческий дистрибутив для создания СХД, при объеме до 18TB предоставляется бесплатно. Базируется на файловой системе zfs, и имеет богатый функционал по управлению системой хранения, включая web-интерфейс.

Model of Object Storage Data

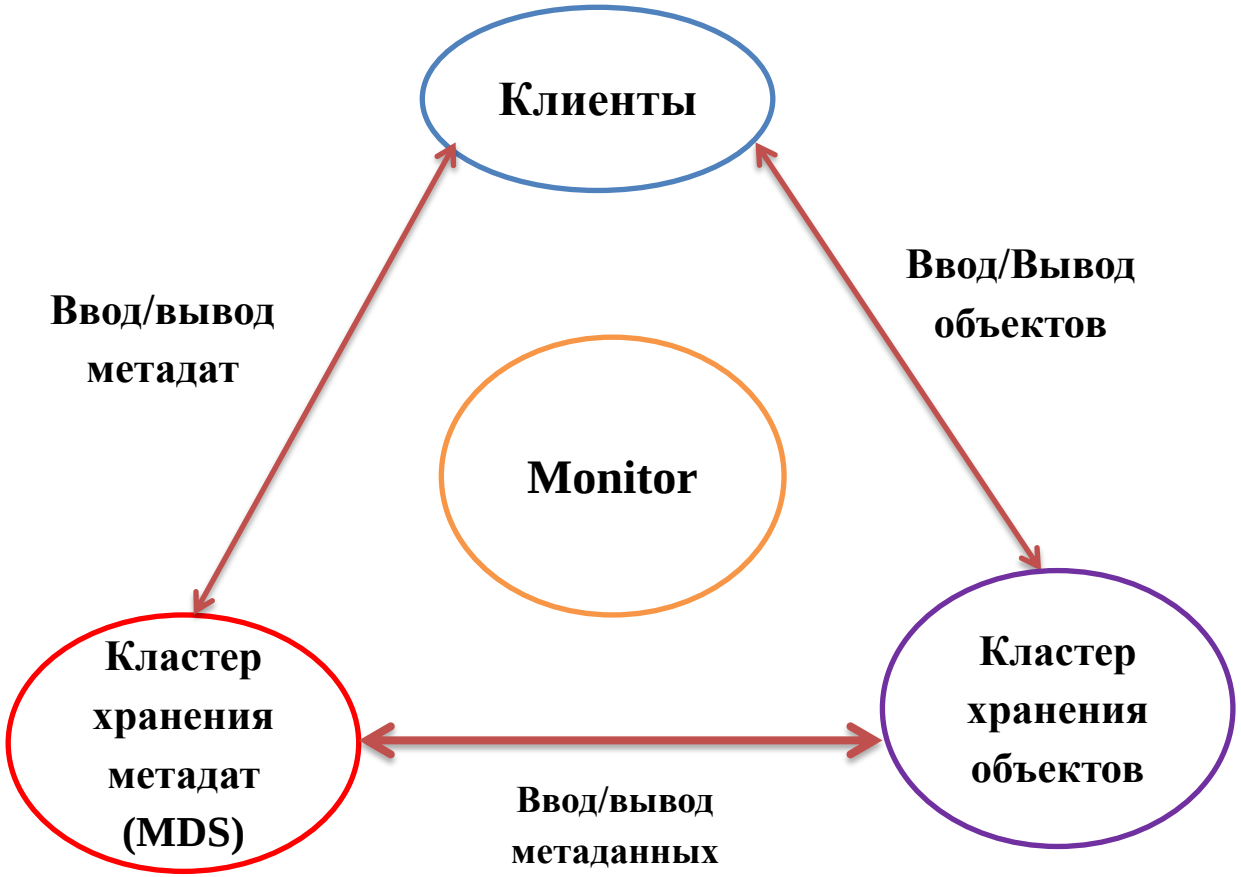


Ключевой особенностью СХД на базе распределенной файловой системы, например, Ceph, является объединение отдельных узлов (NOD-ов), в единое кластерное хранилище данных. СХД такой модели отличаются высокой емкостью и отказоустойчивостью, имеют возможность гибкой масштабируемости. При этом в качестве блочных устройств хранения могут использоваться как отдельные диски в каждом NOD, либо уже созданные хранилища данных.

Распределенная файловая система Ceph

- Ceph – распределенная файловая система, предназначенная для построения единого пространства СХД на базе отдельных узлов (официальный сайт проекта http://ceph.com/) через сеть передачи данных с поддержкой протокола TCP.
- Одно из базовых свойств Ceph — масштабируемость до петабайтных размеров.
- В Ceph реализован гибкий механизм кластерного управления, с широкими возможностями по администрированию.
- Концепция адаптивности к нагрузкам обеспечивает высокий уровень производительности. В рамках данной концепции система не проектировалась под определенные значения пиковых нагрузок.
- Высокая надежность хранилища на базе Ceph базируется на концепции постоянного единичного отказа. В рамках этой концепции считается, что постоянный единичный отказ является нормальным состояние в рабочей кластерной системы высокой емкости.
- Концепция абстрактности в Ceph обеспечивает гибкие архитектурные решения для построения и работы с хранилищем данных, связывая в единое ключевые понятия: объект (object), файл (file) и диск (disk). Для архитектора/администратора системы на выбор три различных абстракции для работы с хранилищем: абстракцию объектного хранилища (RADOS Gateway), блочного устройства (RADOS Block Device) или POSIX-совместимой файловой системы (CephFS).

Архитектура Ceph



Компонент Monitor является ядром распределенной файловой системы Ceph, клиенты выступают в роли инициаторов работы с объектами используя метадаты для управления ими. Чтение и запись объектов в хранилище идет через операции с метадатами. Для хранения метадат используются сервера метадат. Для хранения объектов используется сервера хранения объектов.

Основные компоненты Ceph

- **Metadata server (MDS)** — демон для обеспечения синхронного состояния файлов в точках монтирования CephFS. Работает по схеме активная копия + резервы, при этом активная копия в пределах кластера только одна.
- **Mon (Monitor)** — элемент инфраструктуры Ceph, который обеспечивает адресацию данных внутри кластера и хранит информацию о топологии, состоянии и распределении данных внутри хранилища.
- **Объект (Object)** — блок данных фиксированного размера (по умолчанию 4 Мб), используются для внутреннего представления данных в в Ceph.
- **OSD (object storage daemon)** — данный демон используется для управления данными находящимся в хранилище, основной элемент кластера Ceph. На одном физическом сервере может размещаться несколько OSD, каждая из которых имеет под собой отдельное физическое хранилище данных.
- **PG (Placement Group)** — логическая группа, объединяющая множество объектов, предназначенная для упрощения адресации и синхронизации объектов. Каждый объект может состоять только в одной PG.

Основные компоненты Ceph

- **OSD Map** — карта, ассоциирующая каждой группе PG набор из одной Primary OSD и одной или нескольких Replica OSD. Распределение PG по нодам хранилища OSD описывается выборкой из карты OSD map, в которой указаны положения всех PG и их реплик. Каждое изменение расположения PG в кластере сопровождается выпуском новой карты OSD, которая распространяется среди всех участников.
- **Primary OSD** — OSD, выбранная в качестве Primary для данной PG. Клиентское IO всегда обслуживается той OSD, которая является Primary для PG, в которой находится интересующий клиента блок данных (объект). Primary OSD в асинхронном режиме реплицирует все данные на Replica OSD.
- **RADOS Gateway (RGW)** — вспомогательный демон, исполняющий роль прокси для поддерживаемых API объектных хранилищ. Поддерживает географически разнесенные инсталляции (для разных пулов, или, в представлении Swift, регионов) и режим active-backup в пределах одного пула.
- **Replica OSD (Secondary)** — OSD, которая не является Primary для данной PG и используется для репликации. Клиент никогда не общается с ними напрямую.
- **RF (фактор репликации)** — избыточность хранения данных. Фактор репликации является целым числом и показывает, сколько копий одного и того же объекта хранит кластер.

Спасибо за внимание!

Questions!