

SUSE® Enterprise Storage

программно-реализуемое хранилище
корпоративного уровня

Кирилл Степанов
Технический эксперт
kstepanov@suse.com



Содержание

Технология программно-реализуемых СХД

Архитектура и основные возможности Ceph

Современное состояние и перспективы развития
SUSE Enterprise Storage

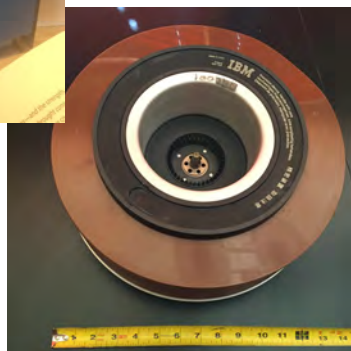
Вопросы и ответы

Программно-реализуемые системы хранения данных (Software Defined Storage, SDS)

Эволюция систем хранения данных

1970-1980

Закрытые системы
из нестандартных
компонентов



1990-2010



Расширяемые
системы из
стандартных
компонентов

(ключевые изобретения
– RAID и SAN)

2010-...



Программная
реализация на
кластере
“стандартных”
серверов

(ключевое изобретение
– 10/40GB Ethernet)



Похоже на развитие суперкомпьютеров

Cray X-MP
1982



“Ломоносов-2”
2014

Преимущества и недостатки SDS

Преимущества

- Удобство увеличения и *уменьшения* емкости СХД
- Простота развертывания и администрирования
- Низкая стоимость владения при минимальных первоначальных затартах
- Отсутствие привязки к технологиям того или иного изготовителя
- Совместимость с системами виртуализации и облачными платформами (Поддержка API Amazon S3 и OpenStack Swift)

Возможные недостатки

- Высокие требования к сетевой инфраструктуре
- Меньшая производительность по сравнению со старшими моделями традиционных СХД (при современном уровне технологии SDS)

Два подхода к реализации SDS

Репликация блоков данных

- “RAID из серверов”
- Примеры: DRBD и GlusterFS

Репликация объектов

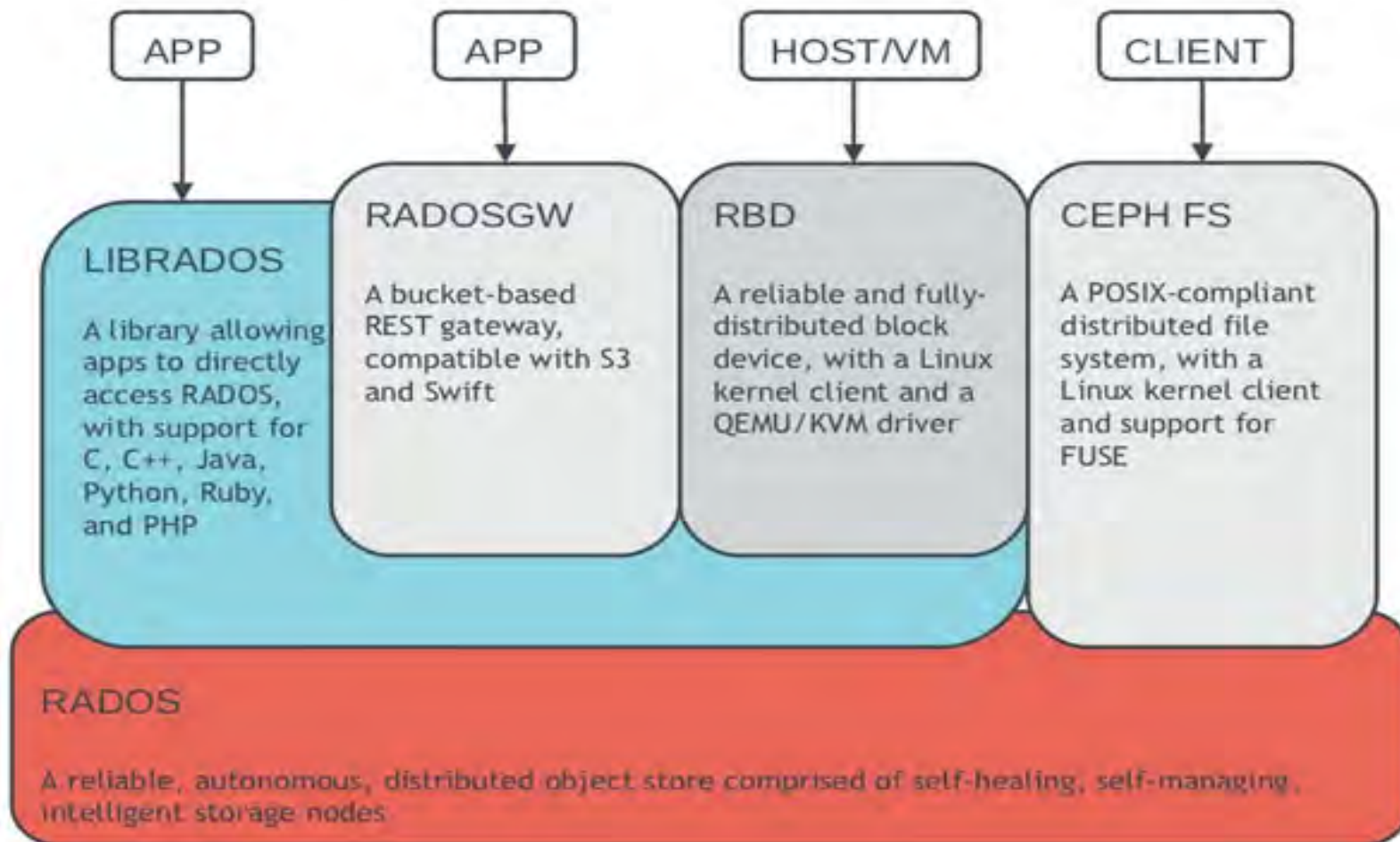
- Объектно-ориентированная система хранения поверх традиционных файловых систем заметно облегчает реализацию API облачных платформ
- Примеры: Swift и Ceph

SUSE Enterprise Storage

Краткий обзор архитектуры Ceph

SUSE Enterprise Storage

На основе технологии Ceph



SUSE Enterprise Storage

На основе технологии Ceph

Данные хранятся на кластере RADOS

- Reliable Autonomic Distributed Object Store (RADOS)

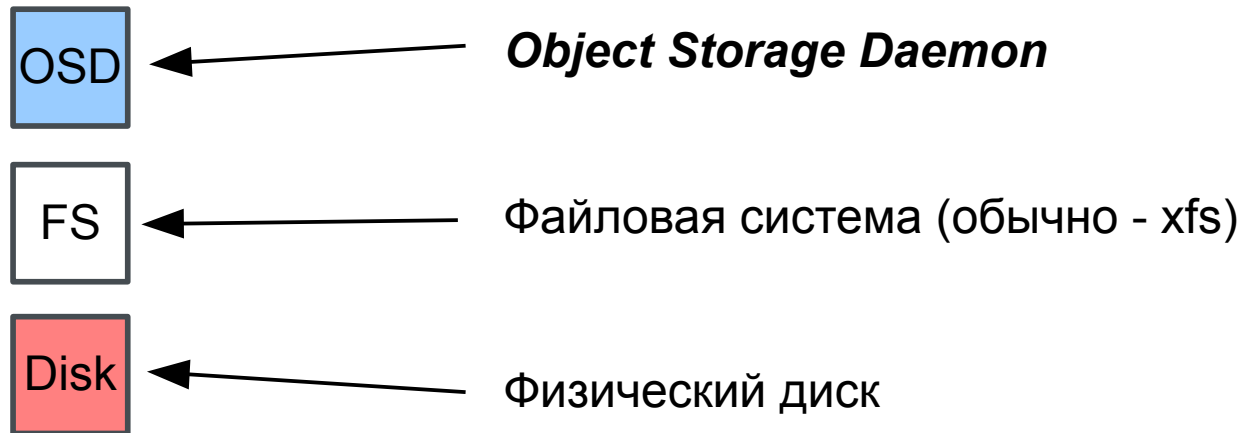
Доступ к ним возможен несколькими способами:

- Работа с объектами через API библиотеки librados для различных языков или командой rados (get, put, cp, rm etc.);
- Интерфейс REST для S3 и Swift через шлюз rgw;
- Работа с виртуальными отказоустойчивыми блочными устройствами (rbd) через клиент (модуль ядра) и драйвер для QEMU/KVM;
- Распределенная POSIX-совместимая файловая система, (работающая через FUSE) с собственным механизмом аутентификации cephx;
- Шлюз для доступа к rbd по iSCSI (формально не входит в проект Ceph).



SUSE Enterprise Storage

Object Storage Daemon (OSD) – основа Ceph

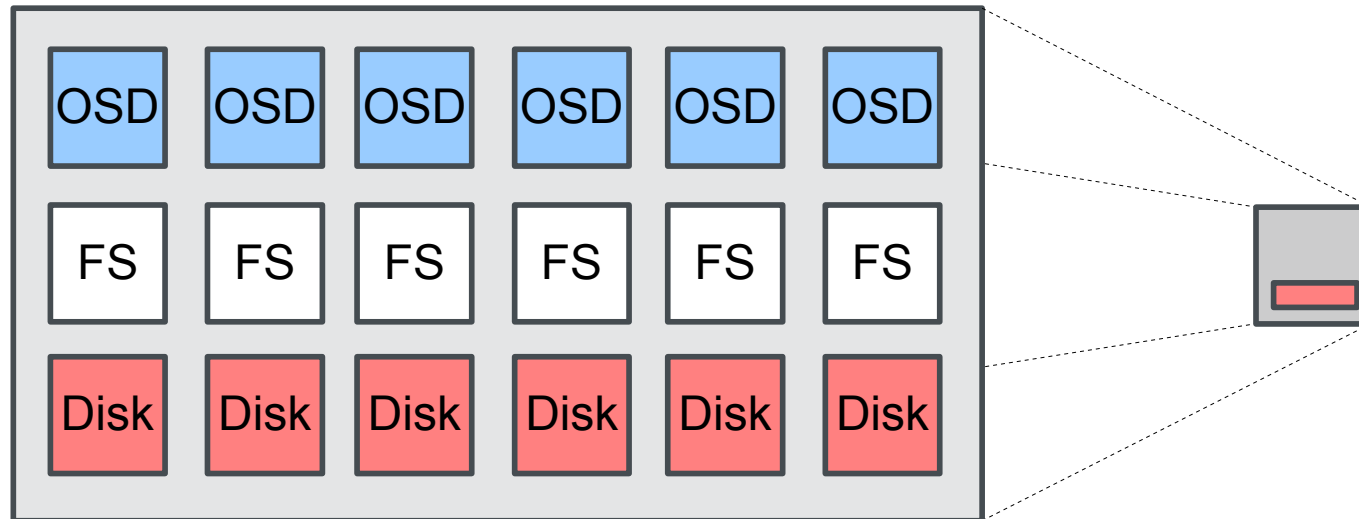


- OSD отвечают за хранение и репликацию объектов данных
- Обычно создается по одному OSD на физический диск

SUSE Enterprise Storage

Узлы хранения данных

На одном сервере может выполняться несколько OSD



SUSE Enterprise Storage

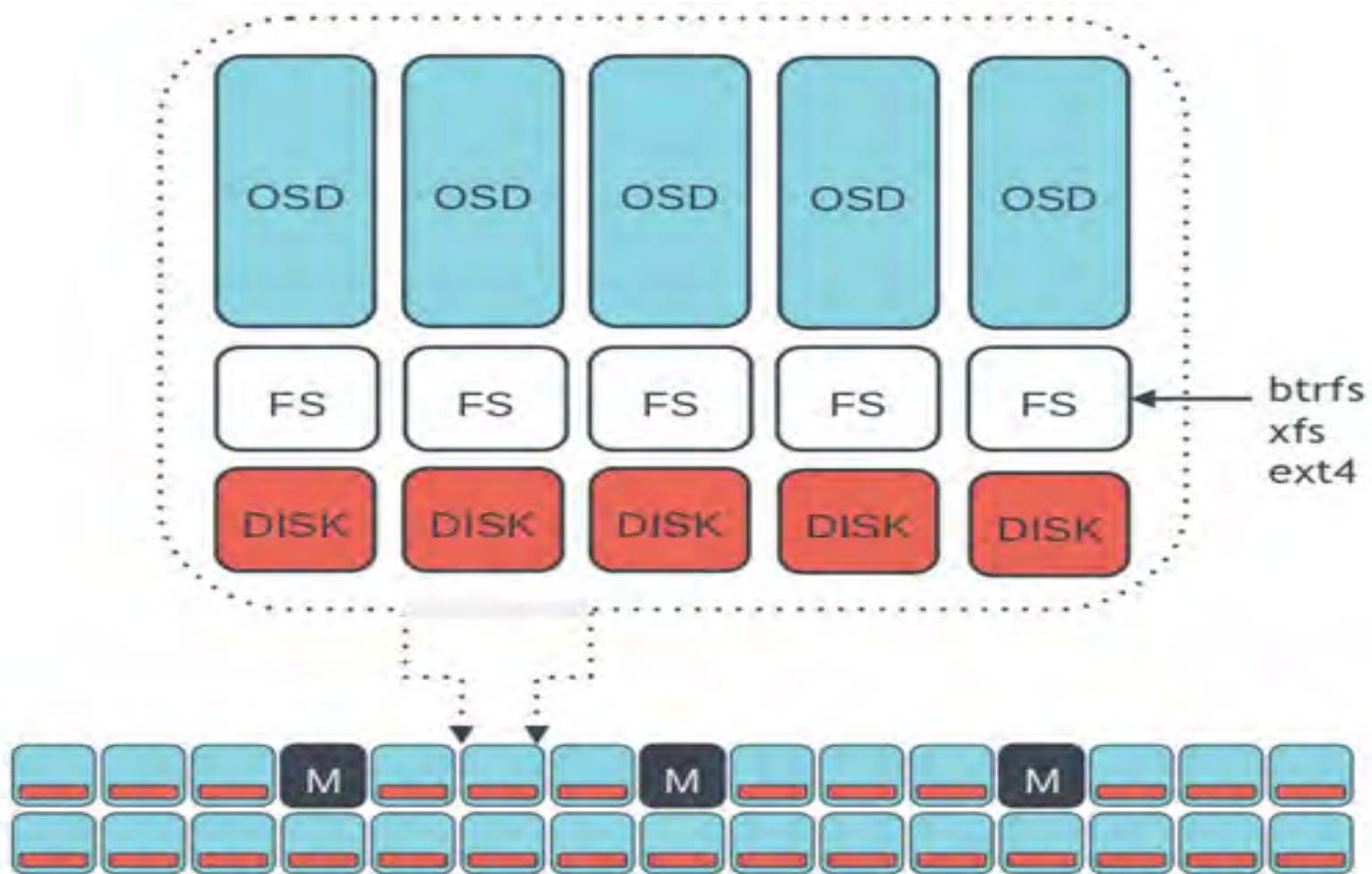
Узлы мониторинга кластера

M

- Управляют работой кластера
 - Обеспечивают подключение узлов
 - Управляют репликацией данных
- Не участвуют в операциях чтения/записи
 - Не хранят данные

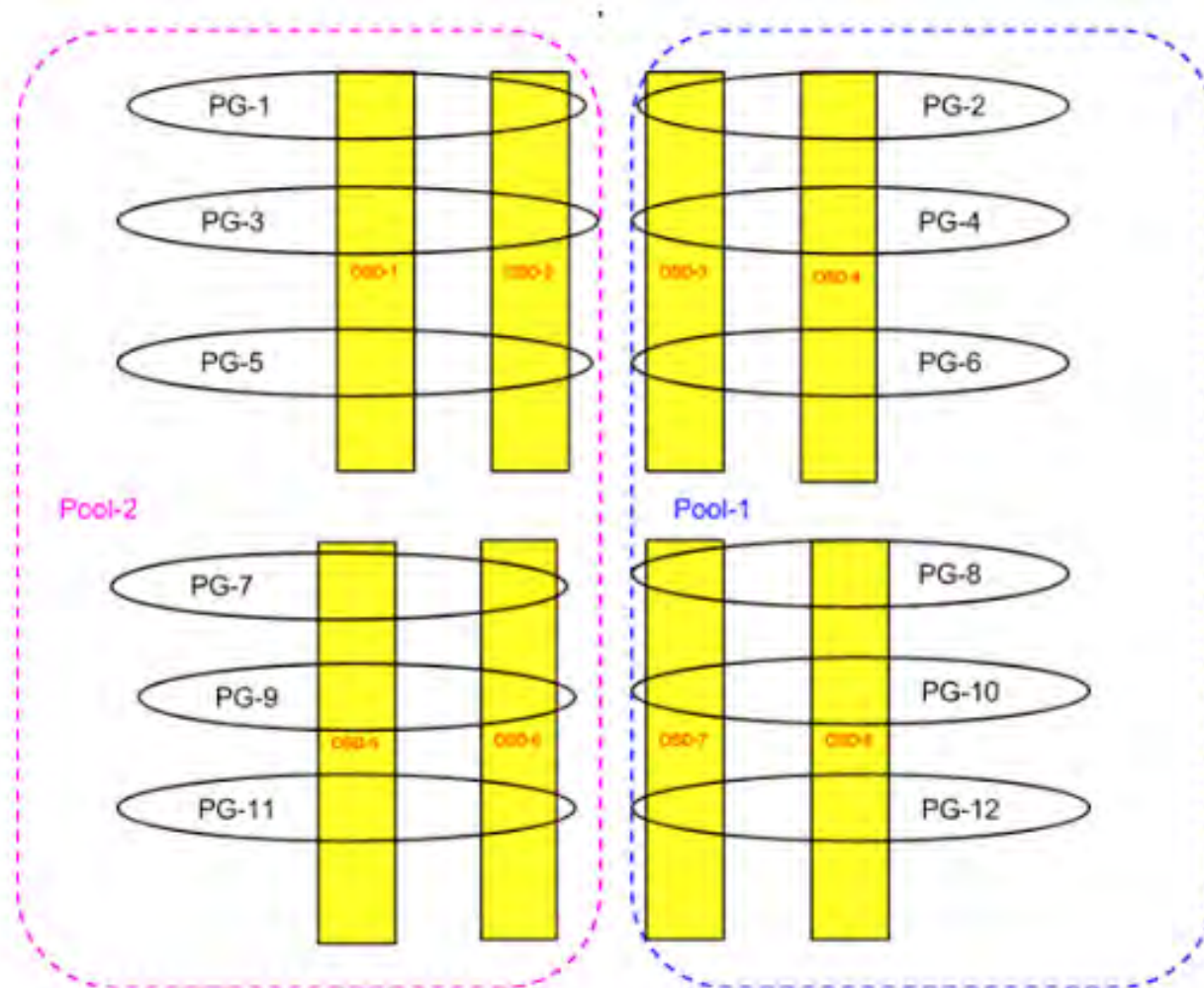
SUSE Enterprise Storage

Кластер Reliable Autonomic Distributed Object Store (RADOS)



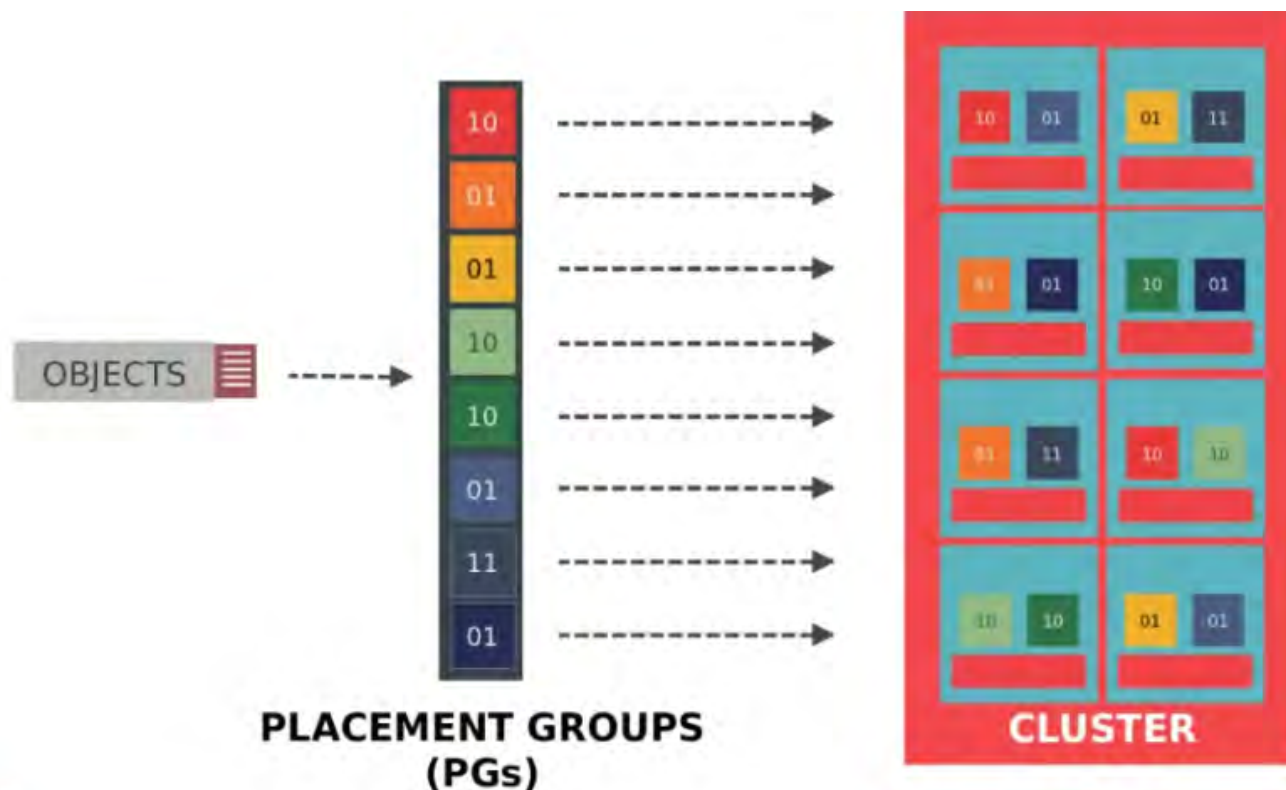
SUSE Enterprise Storage

Организация данных - placement group (PG) и пулы



SUSE Enterprise Storage

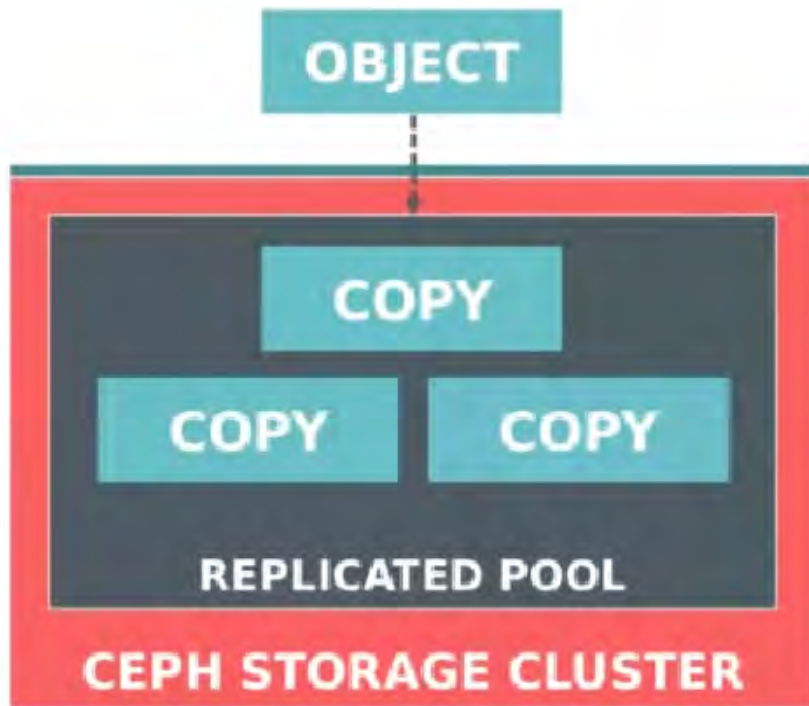
Хранение и репликация данных



Объект состоит из нескольких PG, которые распределяются по разным OSD псевдослучайным образом и реплицируются по алгоритму CRUSH (Controlled Replication Under Scalable Hashing). В случае аварии диска, PG автоматически копируется на исправный OSD.

SUSE Enterprise Storage

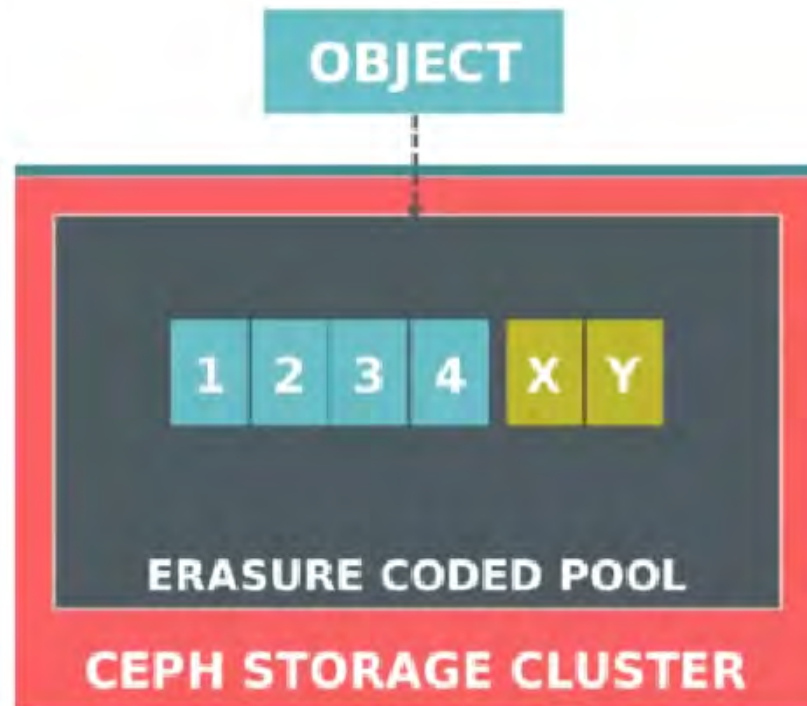
Способ репликации выбирается при создании пула



Объект реплицируется полностью и хранится в 3-х (по умолчанию) копиях.

Избыточность 200%.

Мгновенное восстановление.



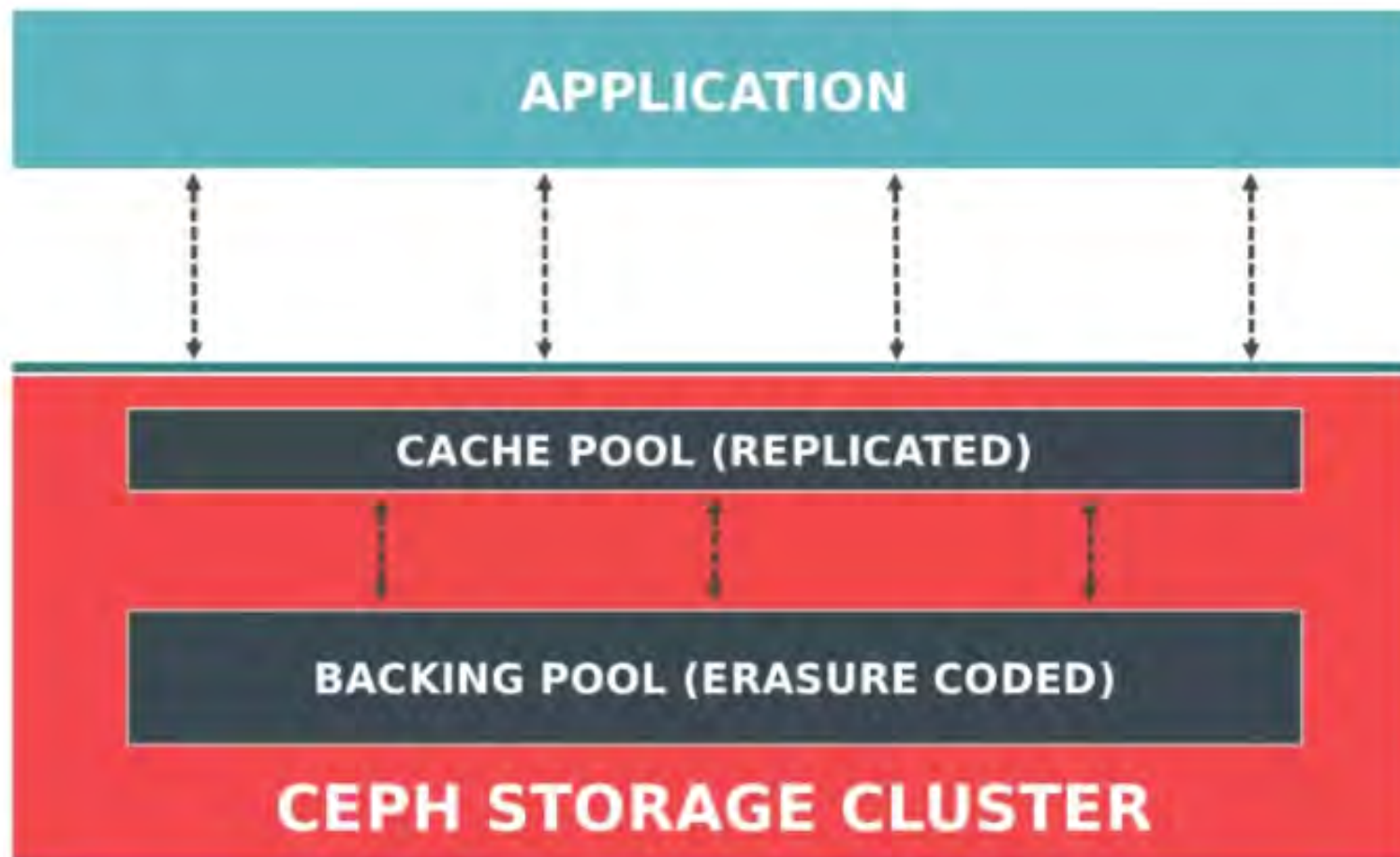
Объект хранится в одном экземпляре с блоками для восстановления.

Избыточность 50%.

Медленное восстановление.

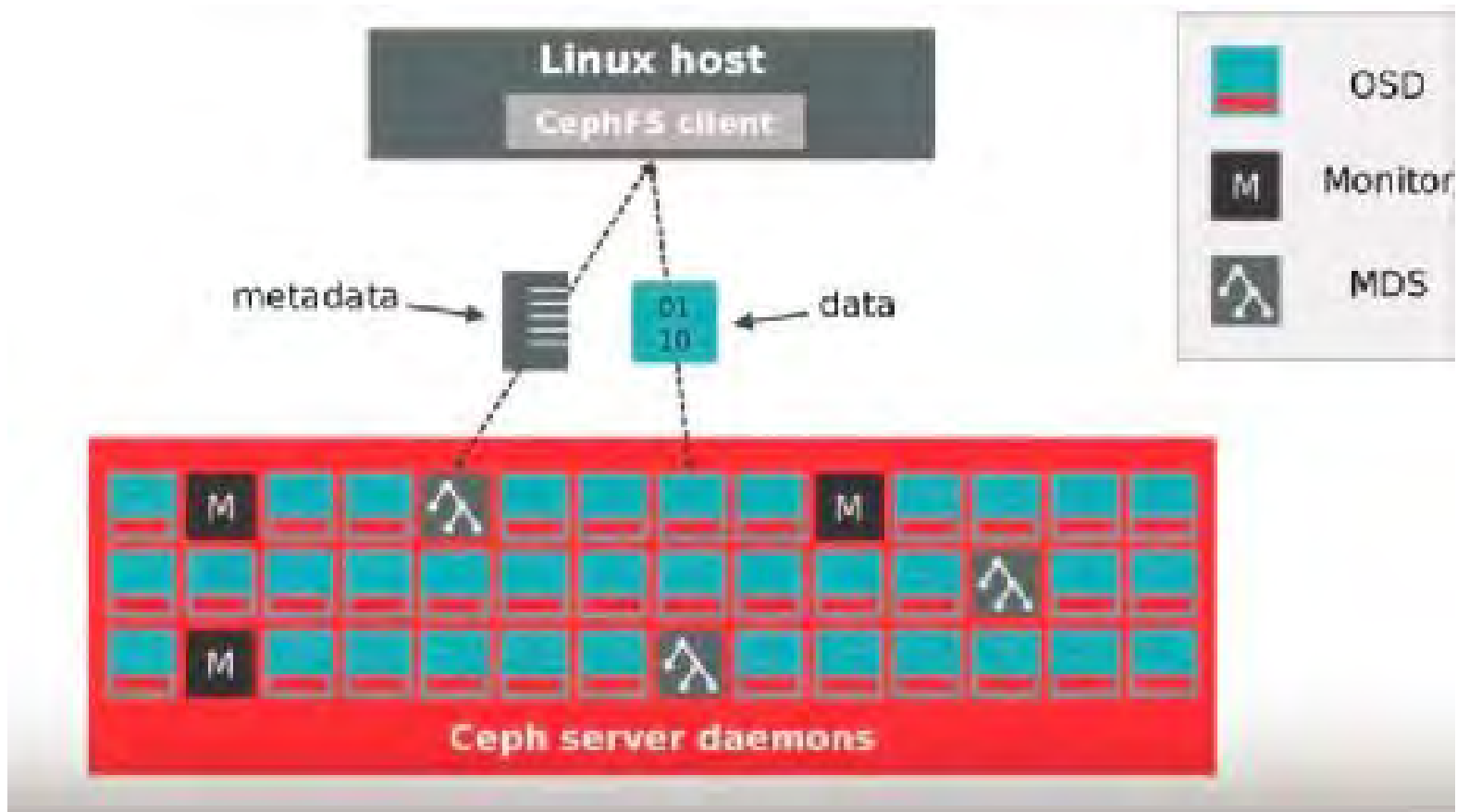
SUSE Enterprise Storage

Выделенные пулы для кэширования данных



SUSE Enterprise Storage

POSIX-совместимая файловая система CephFS



Technical preview в SUSE Enterprise Storage 3



SUSE Enterprise Storage

Основные возможности

- Неограниченное масштабирование
- Самовосстановление
- Работа с объектами и блоками данных
- Thin provisioning
- Клонирование и снимки объектов
- Возможность обновления ПО или добавления новых серверов без остановки кластера
- Двухуровневое кэширование
- Возможность выбора способа репликации



SUSE Enterprise Storage

На базе технологии Ceph

Ceph advisory board

- Компании и организации, участвующие в проекте Ceph:
 - Canonical
 - CERN
 - Cisco
 - Fujitsu
 - Intel
 - Red Hat
 - SanDisk
 - SUSE



SUSE Enterprise Storage сегодня и завтра

SUSE Enterprise Storage 3

Поддерживаемое решение на базе технологии Ceph

Дополнительный продукт (add-on) для SUSE Linux Enterprise Server 12 SP1

- Версия 3 вышла в июне 2016 года

Включает в себя

- Ceph версии v 10.2.x (Jewel)
- Шлюз iSCSI
- Веб-интерфейс мониторинга и управления Romana/Calamary

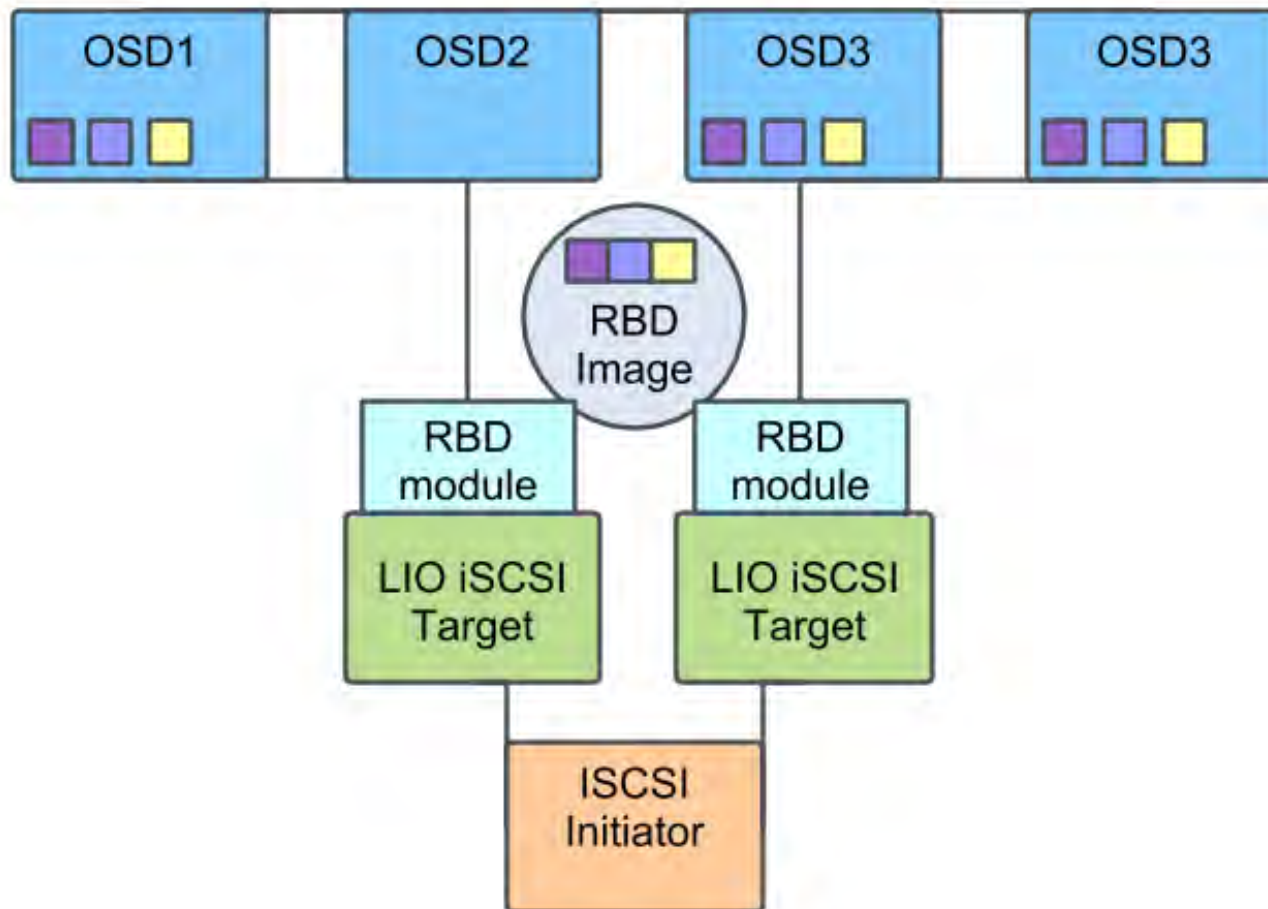
Способы развертывания кластера

- Ceph-deploy (штатный для проекта Ceph)
- Crowbar
- Salt



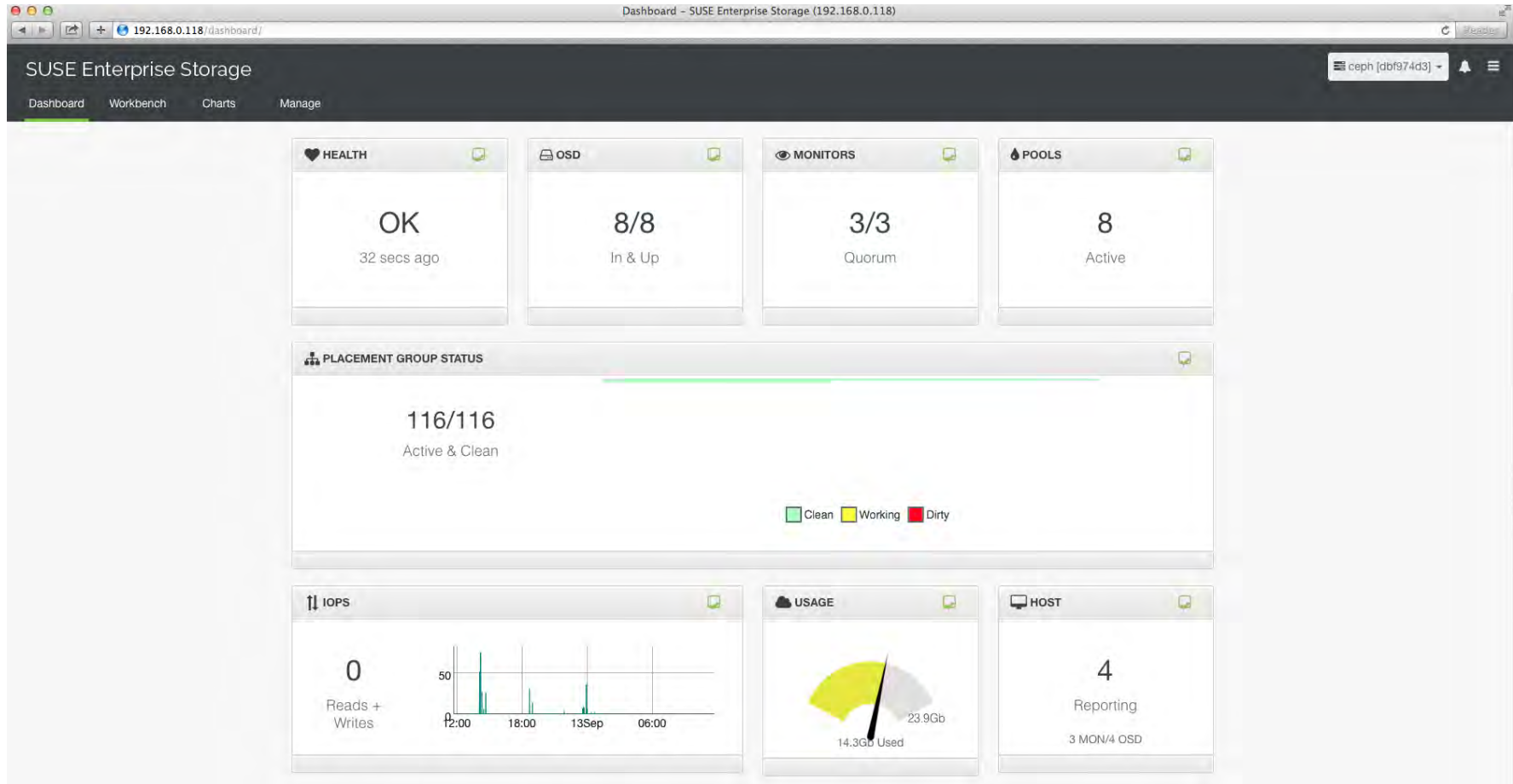
SUSE Enterprise Storage 3

Отказоустойчивый шлюз iSCSI для доступа к rbd



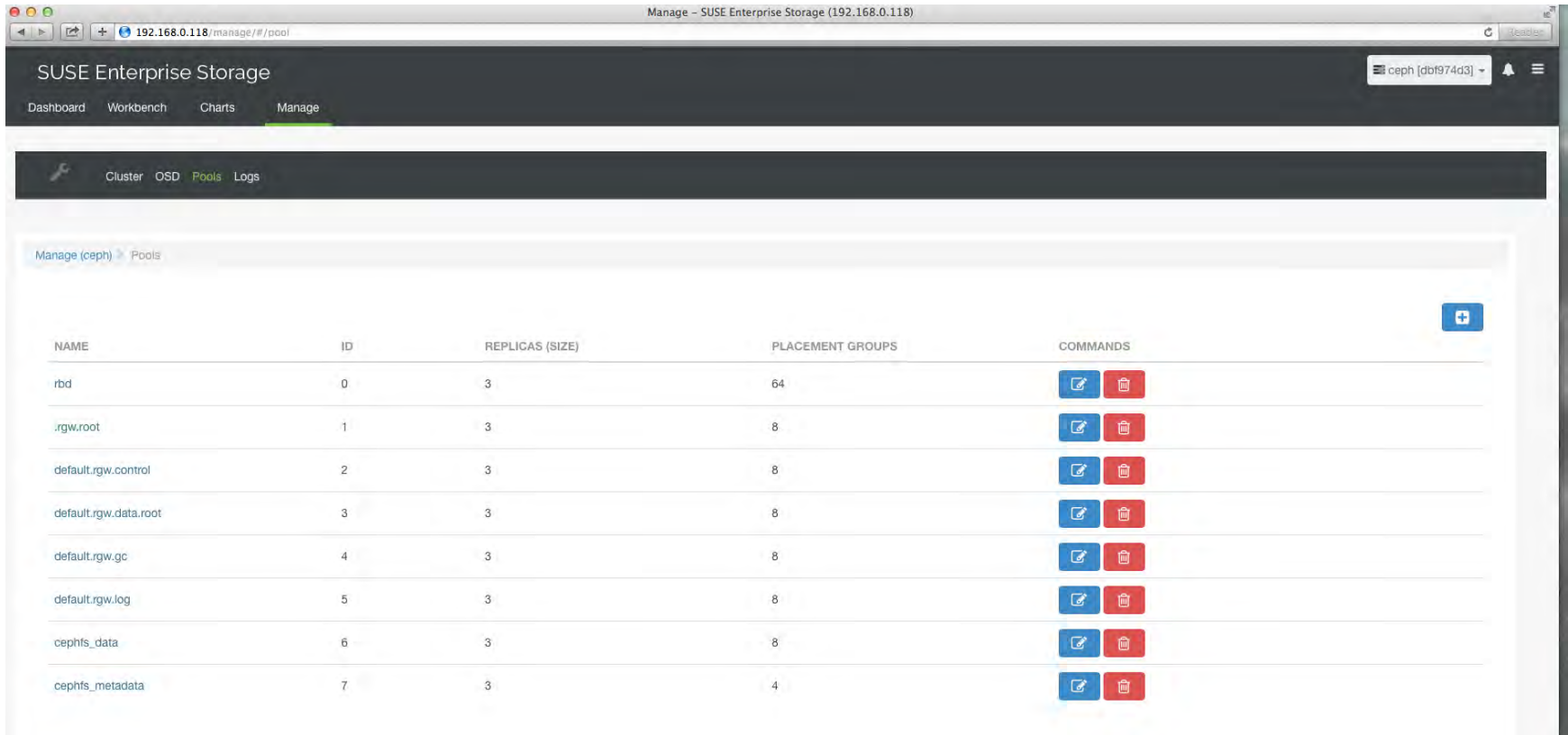
SUSE Enterprise Storage

Интерфейс мониторинга и управления



SUSE Enterprise Storage

Интерфейс мониторинга и управления



The screenshot displays the SUSE Enterprise Storage management interface in a web browser. The browser's address bar shows the URL `192.168.0.118/manage/#/pool`. The page title is "SUSE Enterprise Storage". The navigation bar includes links for "Dashboard", "Workbench", "Charts", and "Manage", with "Manage" being the active tab. A user profile dropdown shows "ceph [dbf974d3]". Below the navigation bar, a breadcrumb trail shows "Cluster" > "OSD" > "Pools" > "Logs", with "Pools" being the active section. The main content area is titled "Manage (ceph) > Pools" and features a table of storage pools. The table has five columns: "NAME", "ID", "REPLICAS (SIZE)", "PLACEMENT GROUPS", and "COMMANDS". There are eight rows of data, each representing a different pool. Each row has a blue edit icon and a red delete icon in the "COMMANDS" column. A blue plus icon in the top right corner of the table area indicates an option to add a new pool.

NAME	ID	REPLICAS (SIZE)	PLACEMENT GROUPS	COMMANDS
rbt	0	3	64	 
.rgw.root	1	3	8	 
default.rgw.control	2	3	8	 
default.rgw.data.root	3	3	8	 
default.rgw.gc	4	3	8	 
default.rgw.log	5	3	8	 
cephfs_data	6	3	8	 
cephfs_metadata	7	3	4	 

SUSE Enterprise Storage 3

Подписки и цены

Base Configuration - \$10000 (Priority Subscription на год)

- Подписка на SES и SLES для развертывания
 - 4 узлов хранения данных (1-2 sockets)
 - 6 инфраструктурных узлов
 - администрирование, мониторинг, хранение метаданных, шлюзы

Expansion Node - \$2300 (Priority Subscription на год)

- Подписка на SES и SLES для развертывания
 - одного дополнительного узла хранения (1-2 sockets) **или**
 - одного дополнительного инфраструктурного узла



SUSE Enterprise Storage 3

Минимальные системные требования

4 узла хранения данных (OSD storage nodes)

- 10 Gb Ethernet (2 интерфейса)
- До 32 OSD на узел (по одному на диск JBOD или LUN в RAID)
- 1ГБ RAM на 1ТБ дискового пространства OSD
- 1.5 ГГц на каждый OSD
- Отдельный диск для загрузки системы
- Инфраструктурные узлы могут выполняться на тех же серверах
- 3 узла мониторинга (рекомендуется выполнение на отдельных физических хостах)
- Шлюзы и серверы хранения метаданных рекомендуется дублировать
- Шлюзы iSCSI, шлюзы RGW и узлы MDS требуют дополнительно 4ГБ RAM и 4 CPU

Выделенный узел управления

- 4ГБ RAM, 4 процессорных ядра, 1ТБ диск

https://www.suse.com/docrep/documents/y13l1qoto9/SUSE_enterprise_storage_configurations.pdf



SUSE Enterprise Storage

Перспективы развития

Новые возможности кластера

- Специализированный драйвер для VMware ESXi
- Асинхронная репликация rbd между кластерами
- Дедупликация и шифрование данных

Интеграция с системами управления

- Intel Virtual Storage Manager (VSM)
- IT-Novum openATTIC
- InkScope
- и другие...

Некоторые из этих возможностей уже включены в SES 3 как technical preview



Вопросы?

Полная информация о
SUSE Enterprise Storage:
<http://suse.com/storage>

Спасибо за внимание





SUSE СНГ
101000, Москва,
Потаповский пер.,
д. 8/12, стр. 2

+7 (495) 623-11-55
www.suse.com

Сообщество разработчиков:
www.opensuse.org

Unpublished Work of SUSE LLC. All Rights Reserved.

This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE LLC. Access to this work is restricted to SUSE employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of SUSE. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.

