Многослойные и многоуровневые системы хранения данных.

Обзор архитектуры и решений на базе open source.

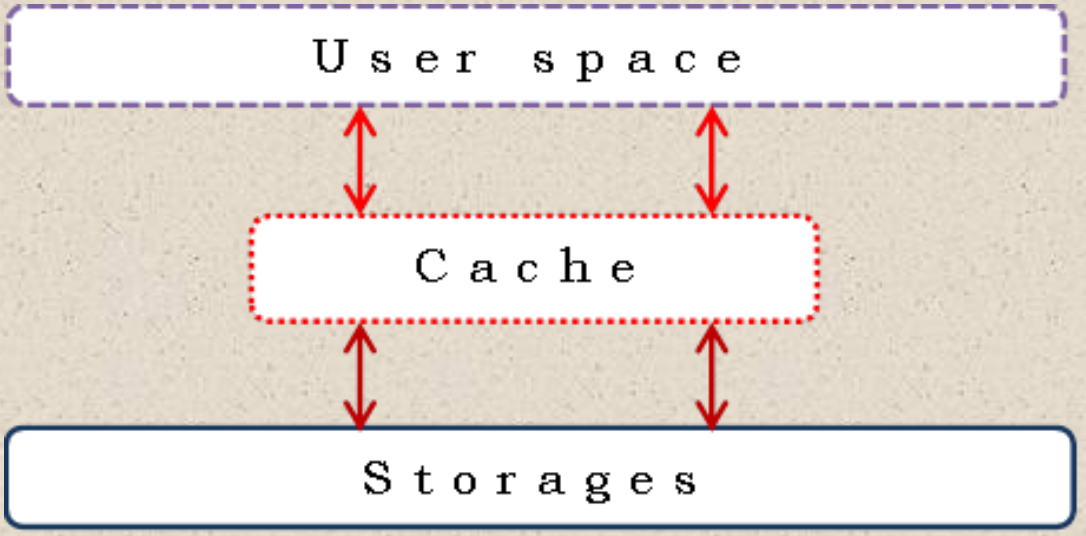
NVM Express

NVM Express (Non-Volatile Memory Host Controller Interface Specification) – спецификация протокола интерфейса доступа к SSD накопителям подключенных по шине PCI Express.

Ключевые характеристики:

- Низкая латентность отклика данных;
- Параллельный доступ к накопителям;
- Поддержка до 64K очередей ввода-вывода, причем каждая очередь ввода-вывода поддерживает до 64K команд.
- Приоритет с четко определенным механизмом арбитража ассоциирован с каждой очередью ввода/вывода;
- Поддержка MSI / MSI-X и агрегация прерываний;
- Поддержка нескольких пространств имен;
- Эффективная поддержка операций ввода / вывода для архитектур виртуализации, в том числе SR-IOV;
- Надежная отчетность об ошибках с возможностью управления;
- Enterprise: поддержка сквозной защиты данных (например, DIF / DIX).
- Enterprise: поддержка multi-path I/O, включая резервирование.

Кэширование в СХД



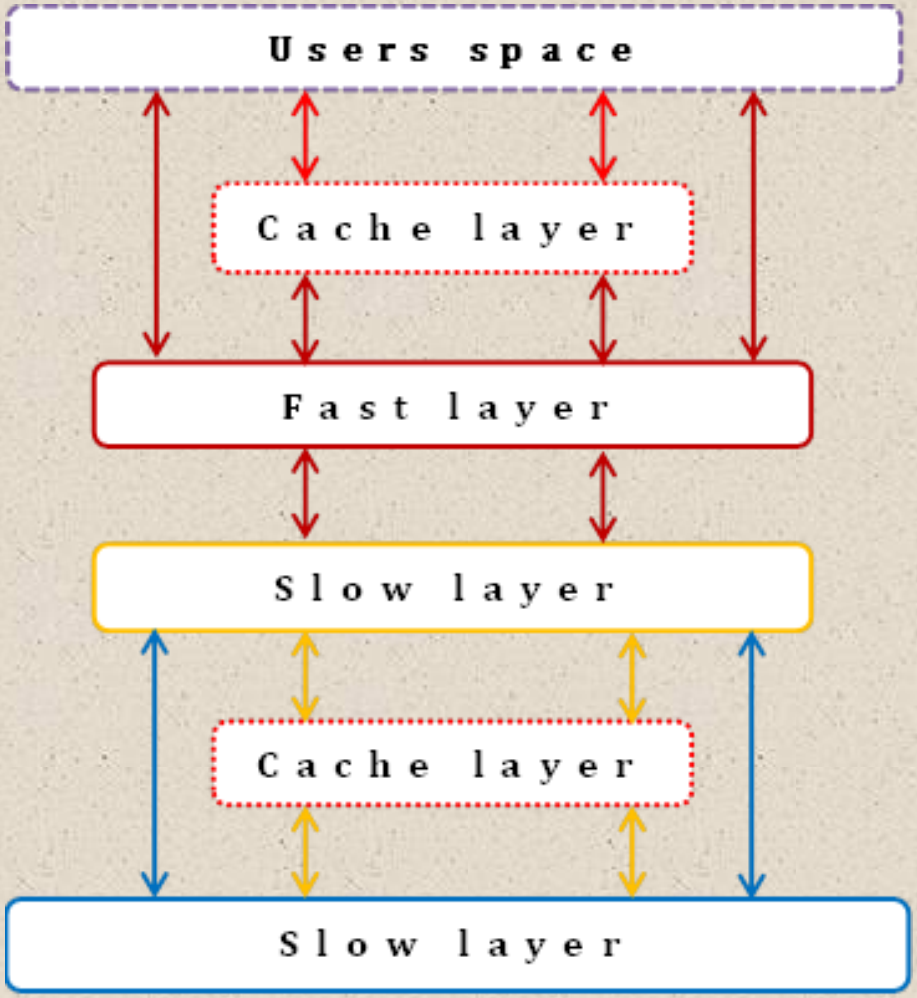
Первая концепция кэширования в СХД (cache only):

- Кэш не используется для постоянного хранения данных.
- В кэш помещаются наиболее востребованные блоки данных (создается избыточная копия данных).

Вторая концепция кэширования в СХД (cache-tiering):

- Распределение данных происходит по схеме: «горячие-холодные»;
- SSD накопители используются для хранения «горячих» данных;
- По мере «остывания» данные из «горячего» слоя мигрируют на «холодный» слой с последующем их удалением из «горячего» слоя.

Многослойные СХД



Распределение данных между слоями происходит по схеме:
«горячие-теплые-прохладные-холодные-архив»

Многослойные СХД

Типы данных размещаемых на слоях:

- Горячие (SSD или SSD NVMe)
- Теплые (SSD или HDD (15K или 10K))
- Прохладные (HDD (10K или 7,2K))
- Холодные (HDD (7,2K или 5K))
- Архив (HDD 5K или TAPE)

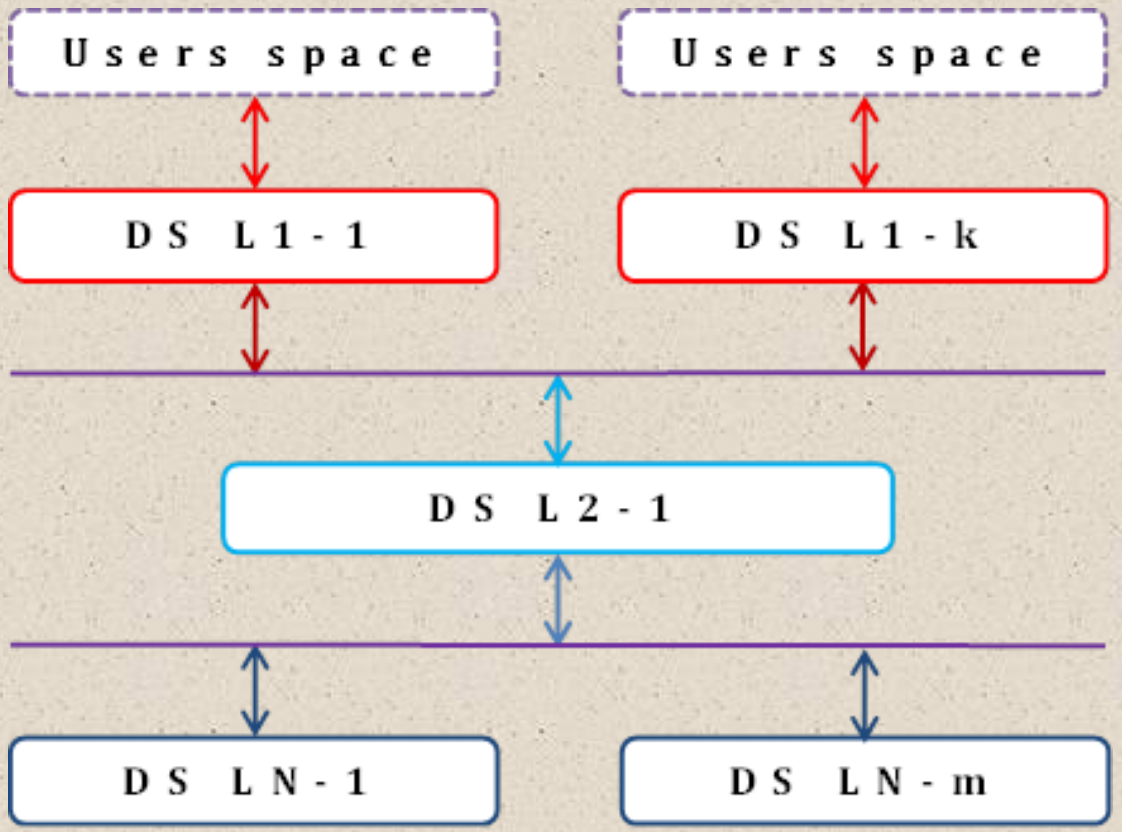
Дополнительные слои:

- Кэш (RAM)
- Внешние СХД (данная возможность реализуется только программными решениями)

Многослойные СХД

- На слоях не создаются избыточные копии данных
- Алгоритмы анализа «температуры» данных
- Использование различных механизмов копирования данных при миграции между слоями (Copy-on-Write (COW) и другие)
- Использование дополнительных кэш слоев для увеличения общей производительности СХД
- Для управления СХД используется аппаратный или программный контроллер
- Набор программных инструментов для мониторинга работы
- Резервирование части дискового пространства на слоях для операций миграции данных.

Многоуровневые СХД



При многоуровневом построении СХД отдельные хранилища данных группируются по уровням на основании их технических характеристик, требований доступности и безопасности.

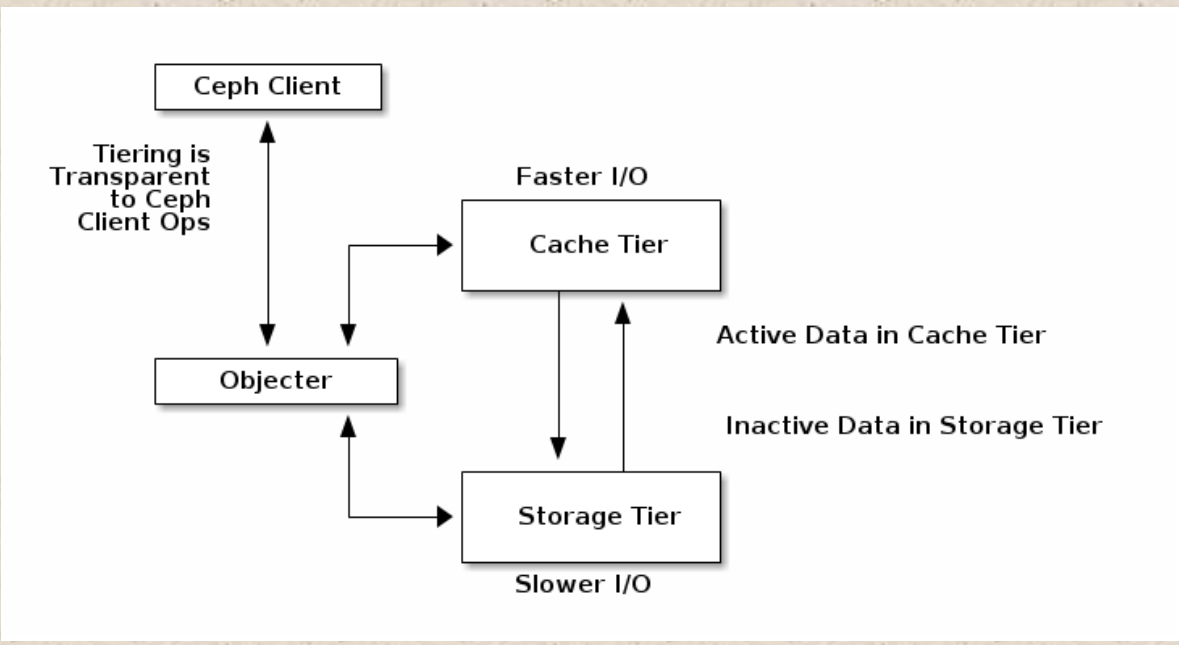
Многоуровневые СХД

- Архитектурное решение для программно-определяемых хранилищ данных (SDS) с единым центром управления и мониторинга на базе программного контроллера с платформой API
- Используется для объединения отдельных DS в единую сеть хранения данных
- Программный контроллер СХД не управляет отдельными DS, а обеспечивает взаимодействие между ними с учетом возможностей подключения
- DS подключенные к СХД на каждом уровне группируются в виде виртуальных накопителей хранения данных (блоки, пулы, массивы и др.)
- Для управления распределением данных по уровням используются технологии многослойных СХД (tiering, migration и т. д.)

ZettabyteFS

- Встроенный менеджер томов
- Виртуальные пулы хранения (zpool)
- Многоуровневая система кэширования с механизмами миграции данных:
  - «горячие» данные размещаются в RAM
  - «теплые» данные на SSD накопителях
  - «холодные» данные на HDD накопителях
- Адаптивный замещающий кэш (Adaptive Replacement Cache, ARC):
  - Первого уровня (L1ARC) хранит блоки данных и метаданных с дисковых накопителей с повышенной частотой обращения «горячие» данные (для этого уровня используется ОЗУ)
  - Второго уровня (L2ARC) используется для хранения таблиц дедупликации и «теплых» данных (как правило, для этого уровня используют SSD накопители)
- «Холодные» данные на HDD не кэшируются
- По мере «разогревания» данных, они с HDD переносятся на SSD или в ОЗУ (производится кэширование данных)
- В качестве второго уровня кэширования может использоваться запись в целевой журнал (ZFS Intent Log, ZIL)

Ceph



В OSD (Object Storage Device) обработчик объектов (Objecter) размещает объекты по уровням, а агент многоуровневого кэширования автоматически выполняет перенос данных между уровнем кэша и уровнем хранилища, используя один из двух сценариев :

- Writeback Mode: в режиме записи данные записываются на уровень кэша и постепенно мигрируют на уровень хранилища и удаляются из кэш. В режиме чтения в начале данные переносятся на уровень кэша, а затем передаются клиенту для работы;
- Read-proxy Mode: в этом режиме, если объект не находится в кэш, то запрос передается на уровень хранилища;

BcacheFS

- Технологии проекта Bcache
- Реализация на уровне блочных устройств
- Поддержка многослойного подключения накопителей
- Механизм Copy-on-Write (COW)
- Возможность подключения нескольких накопителей к одному разделу
- Репликация на уровне RAID
- Кэширование
- Шифрование
- Резервное копирование суперблоков
- Прозрачное сжатие и верификация целостности данных

GlusterFS

Базовые компоненты:

- Volume
- Brick (базовый юнит хранения)
- Server/Nodes (содержит в себе brick)

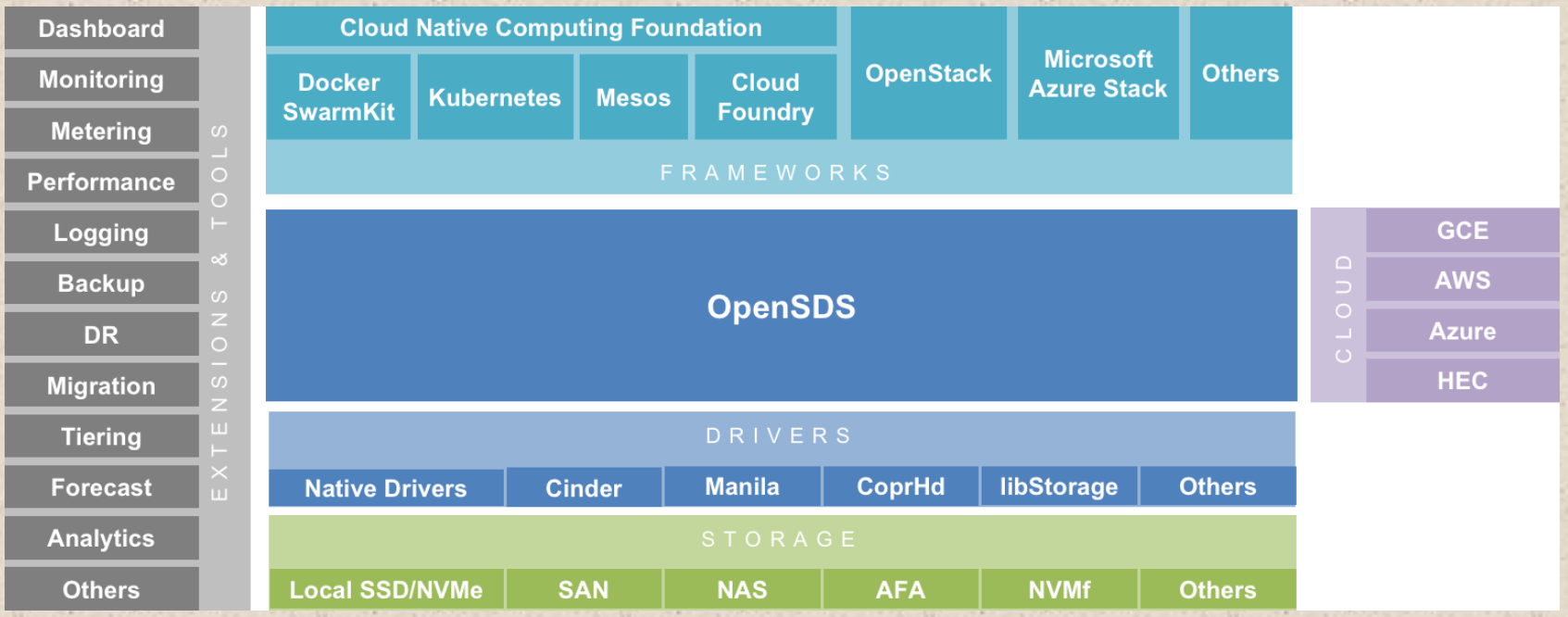
Концепция хранения данных:

- Организация хранения данных на уровне логических томов (volume)
- Поддержка единого пространства имен (namespace)
- Использование технологий кэширования и тиринга
- Поддержка технологий классификации и миграции данных между томами (разрешенными партнерами)
- Единая точка монтирования томов

Controller SDS CorpHD

- Программный котроллер SDS с открытым кодом и платформой API
- Возможность подключение СХД различных типов: блочных, объектовых или хранилищ файлов.
- Поддерживает виртуальные пулы и виртуальные массивы хранения данных
- Поддержка технологий платформ виртуализации VMWare и Vblock
- Интерфейс взаимодействия с объектным хранилищем данных Ceph
- Интерфейс взаимодействия с OpenStack на уровне интерфейса блочного устройства
- Текущая версия 3.0

Open Controller for SDS



Проект OpenSDS

Linux Foundation 8 ноября 2016 официально объявила о старте нового проекта по созданию открытой платформы контроллера программно-определяемого хранилища данных. «Основной целью нового проекта OpenSDS является радикальное упрощение системы хранения данных путем создания общедоступного открытого решения контроллера SDS для различных сред (облачной, контейнерной, виртуализации и т.д.)»

Open Controller for SDS

For Cloud-Native (Docker, Kubernetes, Mesos, CloudFoundry)

- Policy-Based Storage Control. Built-in policies for lifecycle management, data protection, data security, and orchestrated control for cloud-native apps
- Cloud-Native Storage. Integration with Kubernetes, Docker, and Mesos enables dynamic storage provisioning, responds to container events eg. support for container migration to another host
- Cloud-Native Deployment. Deploy and scale with container clusters
- Built-in Support For OpenStack Storage. Connect to all storage back ends supported by Cinder and Manila drivers
- Storage Discovery and Pooling. Support discovery of storage back ends and aggregation of storage resources into a seamless whole

Open Controller for SDS

For OpenStack

- Policy-Based Storage Control. Built-in policies for lifecycle management, data protection, data security, and orchestrated control for cloud-native apps
- Orchestrate Storage. Add orchestration to Cinder/Manila by automating operations such as snapshots, backups and lifecycle management
- Leverage Enterprise Storage Features. Advanced OpenSDS API's enables enterprise storage features to be fully utilized by OpenStack

NVM Express

Форм-факторы накопителей с интерфейсом NVM Express:

- M.2 (NGFF) — бескорпусные накопители в компактном форм-факторе;
- U.2 - накопители форм-фактора 2.5" высотой 15 мм с разъёмом SFF-8639
- Плат расширения для PCIe (Add-in PCIe card или AIC)

Поддержка в операционных системах.

- Chrome OS
- Linux (RHEL, SUSE, Ubuntu и др.)
- Windows (Windows 7, 8, 8.1, 10, Windows Server 2008 R2, 2012, 2012 R2, 2016)
- VMware (ESXi 5.5, 6.0 и выше)
- BSD (FreeBSD, OpenBSD и др.)
- Solaris 11.2 и выше
- UEFI