

Распределенные иерархические системы
хранения данных (HSM)

А.Н. Мойбенко
МЛИТ, ОИЯИ

Требования пользователей

- Побольше места (лучше неограниченно)
- Побыстрее скорость передачи
- Минимализация времени доступа к данным
- Легкая доступность к данным
- Надежность хранения
- Защита
- Разрешение / ограничение доступа

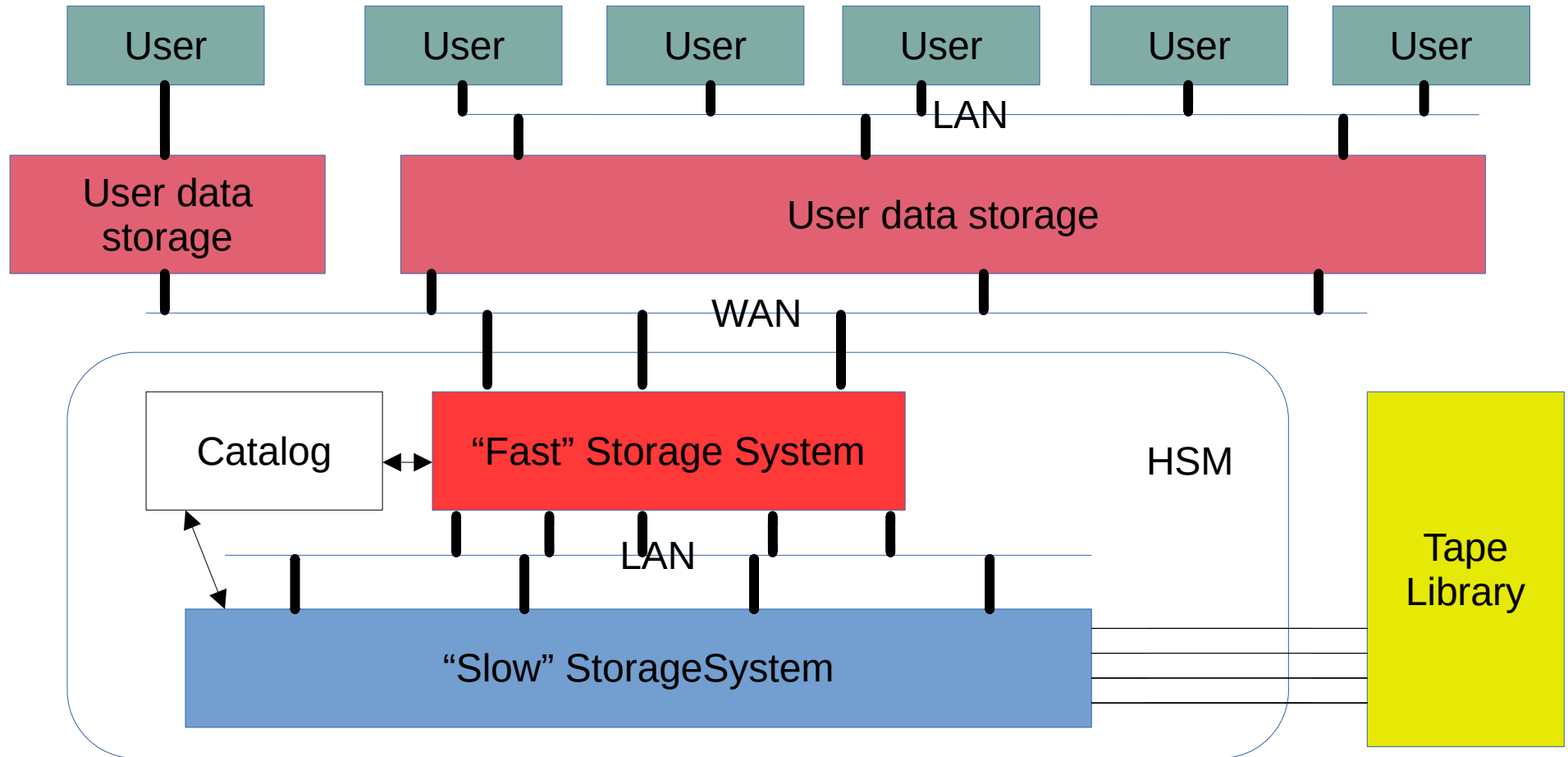
Немного истории

- ЭВМ коллективного пользования. (< ~1980)
- Пользователи используют одну ЭВМ
- Данные хранятся на диске (дисках) ЭВМ
- Данные архивируются на магнитных лентах (16 мм)
- ЭВМ индивидуального пользования (РС) (late 1980)
- Данные хранятся на диске (дисках) ЭВМ
- Данные архивируются на магнитных кассетах (8 мм)
- РС соединяются в кластеры для (> 1990):
- Увеличения скорости сбора и обработки данных
- Увеличения объема доступных данных (пользователю доступны данные с многих РС)
- Архивирование на магнитных кассетах через магнитофоны, расположенные в автоматических ленточных библиотеках

Продолжительность и стоимость хранения информации

- Hard Disk Drive (HDD) – 4 – 7 years
 - active (require power)
 - Max capacity 20 TB
 - Cost \$16/TB
 - Solid State Drives (SSD) – 5 – 10 years
 - Active, require power
 - Max capacity 4 TB
 - Cost \$60 / TB
 - Flash, passive - 10 years with average use
 - passive
 - Max capacity 1 TB
 - Cost \$90 / TB
 - Magnetic Tape – 15 – 30 years
 - Max capacity 20 TB
 - Passive, but require tape drive, tape library
 - 20000 load / unload cycles
 - \$5 / TB (\$2.5 / TB - compressed)
- Consider:
- Acquisition cost
 - Operational cost
 - Power and cooling
 - Warranty
- Tape system cost ~ 1/3 disk system cost per PB not considering lifespan

Структура иерархической системы хранения данных



Потребность в иерархических распределенных системах хранения данных

- Возрастание объемов данных (> 10 PB) и числа compute nodes.
- Распределение данных
- Распределение ресурсов хранения
- добавление элементов хранения
- Определение неисправных элементов и их замена
- Увеличение скорости передачи данных
- Автоматизация архивирования, освобождения /восстановления / и репликации “быстрого” хранилища данных в “медленном” хранилище
- Обеспечение надежного долговременного хранения

Способы доступа к данным в иерархических системах хранения

- Кэширование
- Если данных нет в быстром хранилище они копируются из более медленного.
- Если быстрое хранилище переполняется менее ценные данные удаляются
- Tiering (передача между уровнями) – данные копируются с более медленного уровня на более быстрый. После чтения данные не остаются на более быстром уровне (backup подход).

Иерархические системы в НЕР

- HPSS – High performance Storage System (DOE, NSF, IN2P3) (~1992)
- CASTOR – CERN Advanced STORage manager (CERN) (1998)
- Dcache + Enstore (FNAL) (1998)
- EOS + CTA – CERN Tape Archive (CERN) (2021)
- Dcache+HPSS – BNL (~2004 Atlas)

Как требования пользователей удовлетворяются в HSM

- Побольше места (лучше неограниченно)
- Высокоскоростная, расширяемая централизованная система управления метаданными, доступная на каждом уровне системы хранения и транспортировки
- Расширяемость за счет добавления элементов хранения (horizontal scalability)
- Побыстрее скорость передачи
- Распараллеливание процессов передачи данных
- Репликация данных на первом (user facing) уровне

Как требования пользователей удовлетворяются в HSM

- Минимализация времени доступа к данным
- Кэширование данных на “быстром” (1) уровне
- Massive data prestaging from “slow” (2nd) (Tape) to “fast” (1st) layer
- Легкая доступность к данным
- Каталог данных
- Средства доставки
- Надежность хранения
- Магнитные ленты (можно положить на полку)
- Автоматизация копирования по запросу на магнитных лентах

Как требования пользователей удовлетворяются в HSM

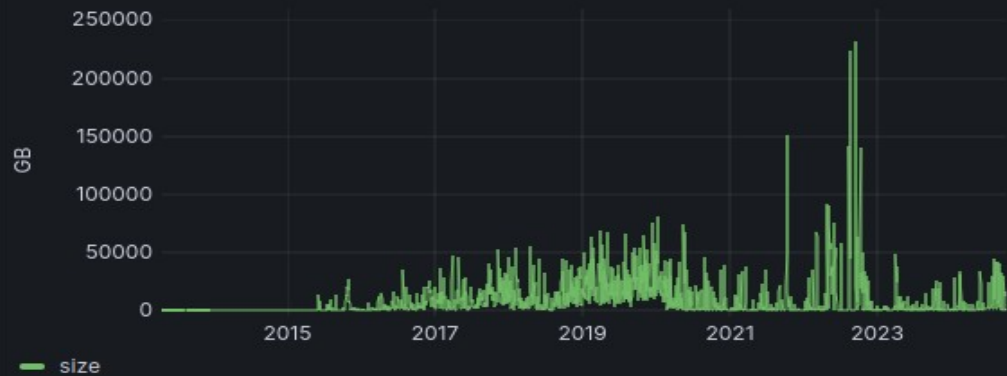
- Защита данных
- Аутентикация пользователя
- Данные на магнитной ленте трудно переписать
- Автоматизация переключения клавиши “запрет записи” на магнитной кассете
- Разрешение / ограничение доступа
- Авторизация доступа

Пример: JINR CMS T1 dCache + Enstore

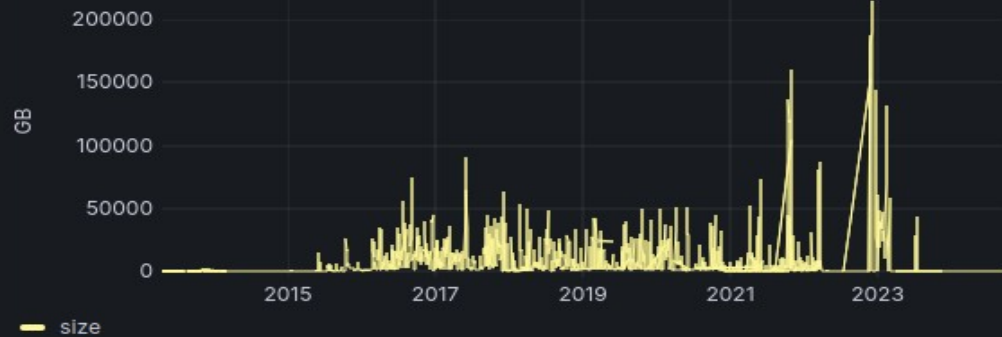
- 2 Layer HSM
- Dcache – “fast” data storage layer (distributed disk system)
- Current capacity – 2.4 PB
- Enstore – “slow” layer (distributed tape system)
- Current capacity ~ 5000 20 GB tapes
- Occupied – 11 PB
- In production since 2016

JINR CMS T1 dcache + enstore data transfers

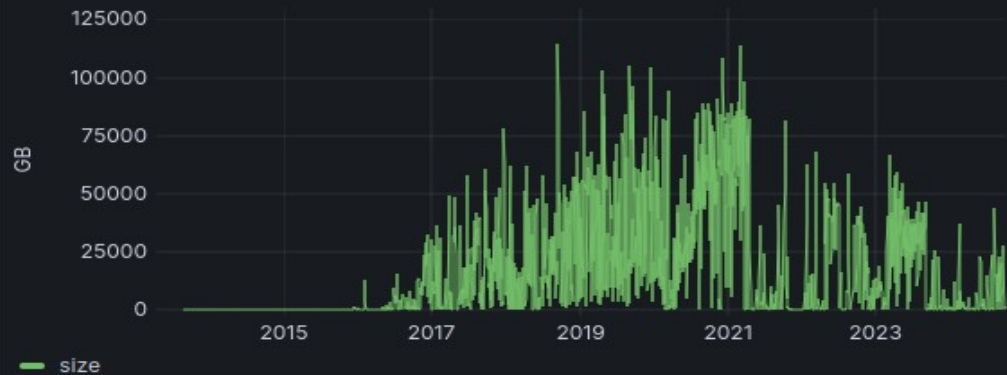
dcache reads



dcache writes



Enstore reads



enstore writes

