Enstore в ОИЯИ.

Состояние и перспективы

А.Н. Мойбенко

23 Апреля 2024

Enstore is a multi-Petabyte scale tape based Mass Storage System (MSS) for High Energy Physics (HEP) Experiments and other scientific endeavors. It has been designed to permit to scale to multiple petabytes of storage capacity, manage tens of terabytes per day in data transfers, support hundreds of users, and maintain data integrity. Enstore can be used for data storage needs of any scale, for different kinds of enterprises. The Enstore architecture allows easy addition and replacement of hardware and software components.

Designed at Fermi National Accelerator Laboratory (USA) initially for needs of Tevatron Run 2 experiments (~ 20 PB, 0.5GB/s aggregate throughput).

Integrated with dCache for files caching / buffering performance.

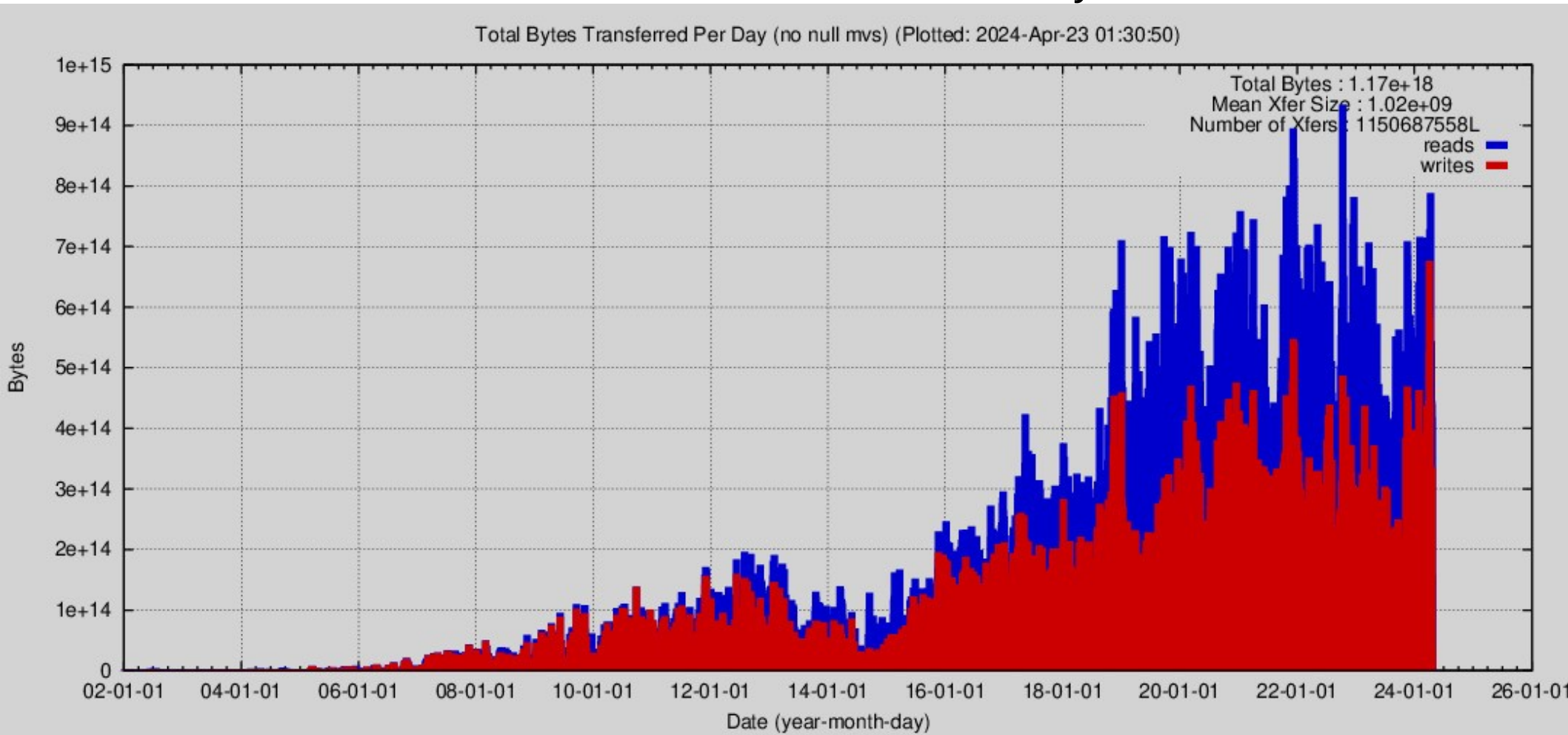
Substantial modifications for CMS experiments

- Performance
- Number of requests in the queue
- Total data capacity exceeding 100 PB and aggregate throughput approaching 5GB/s

Small Files Aggregation (SFA) feature to effectively store, access and transfer files < 500 MB

Currently largest installation has ~ 350 PB, ~200 Tape drives, last month max 800 TB/day (~9GB /s)

Enstore at Fermilab. Tolal activity



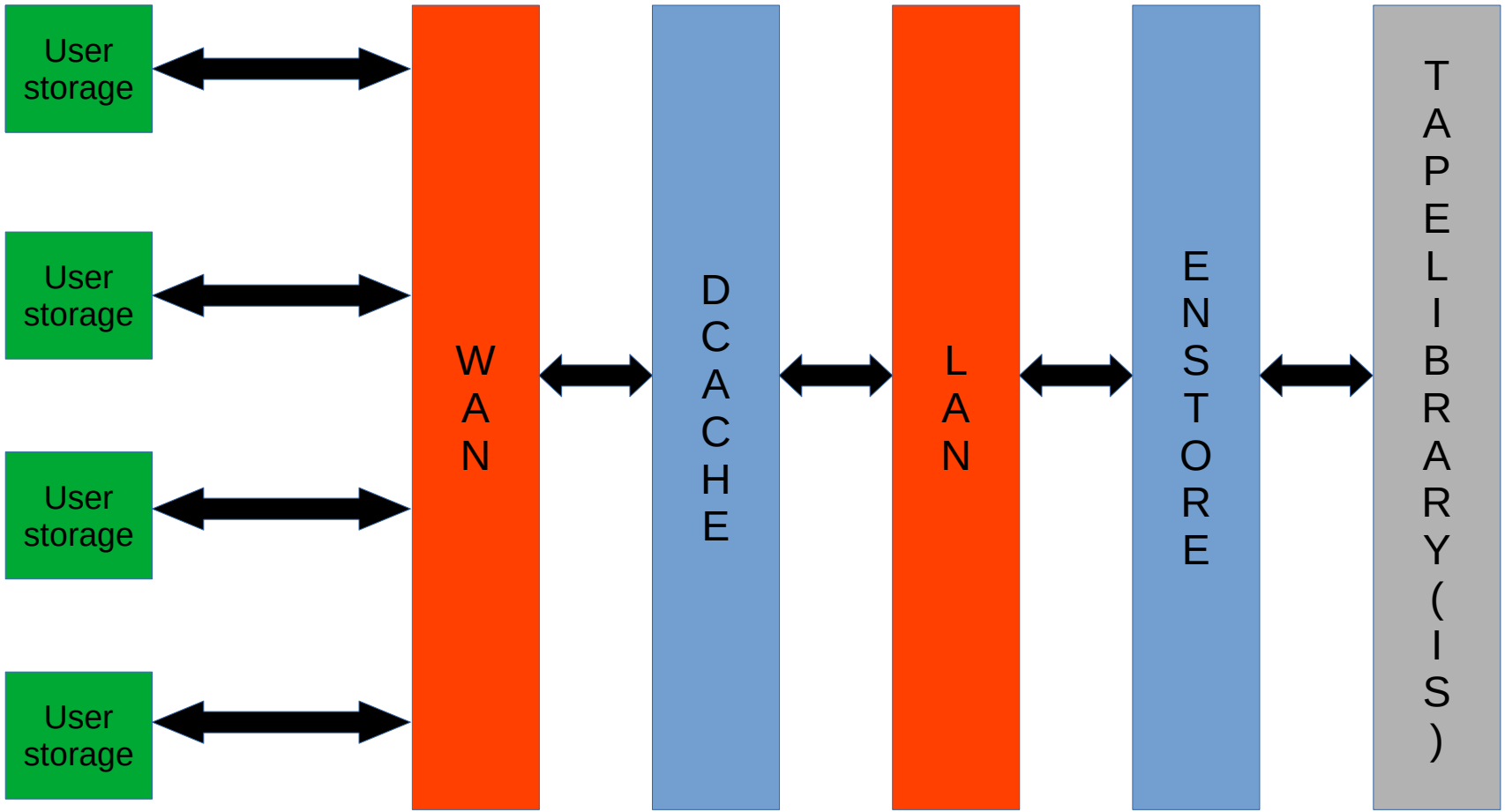
LICENSING

=========

Enstore:

(Open Source: AM)

Enstore в системе хранения и транспортировки данных



Используется в:

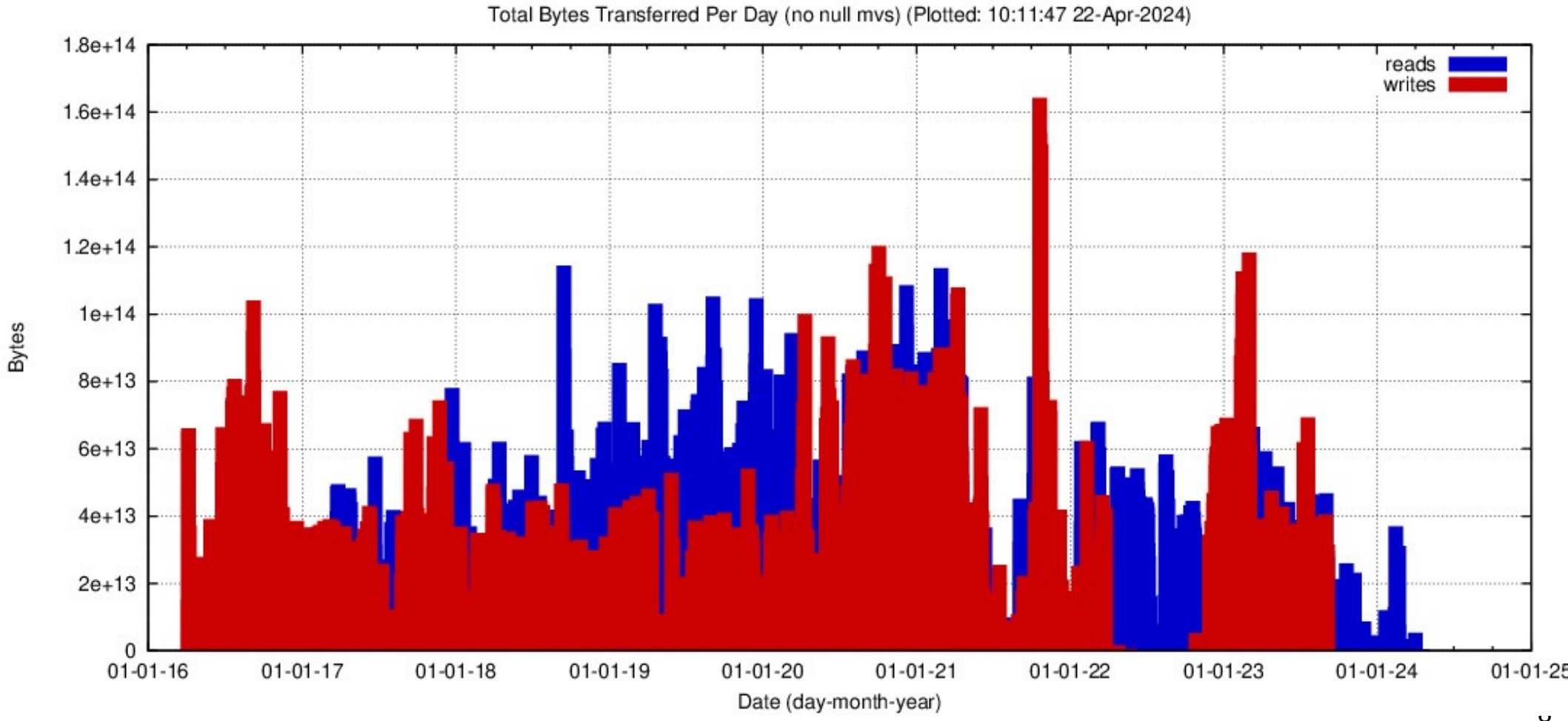
FNAL, USA (354.8 PB – всего, 307 PB - активных)

PIC, Spain (???)

ОИЯИ (11.7 PB – всего, 11.4 PB – активных, в основном CMS T1) – с 2016 года.

КИАЭ ???

Enstore в ОИЯИ: общая активность



Enstore в ОИЯИ
Изменения конфигурации

Изменена в декабре 2023

Задействованы control paths (CP) всех tape drives (TDr)

Операции монтирования / демонтирования через CP TDr

Media changer (новый тип) служит координатором

Запросы монтирования / демонтирования поступают в библиотеку асинхронно, где firmware определяет последовательность их исполнения

Library manager – 3 порта (было 2): control, encps, movers

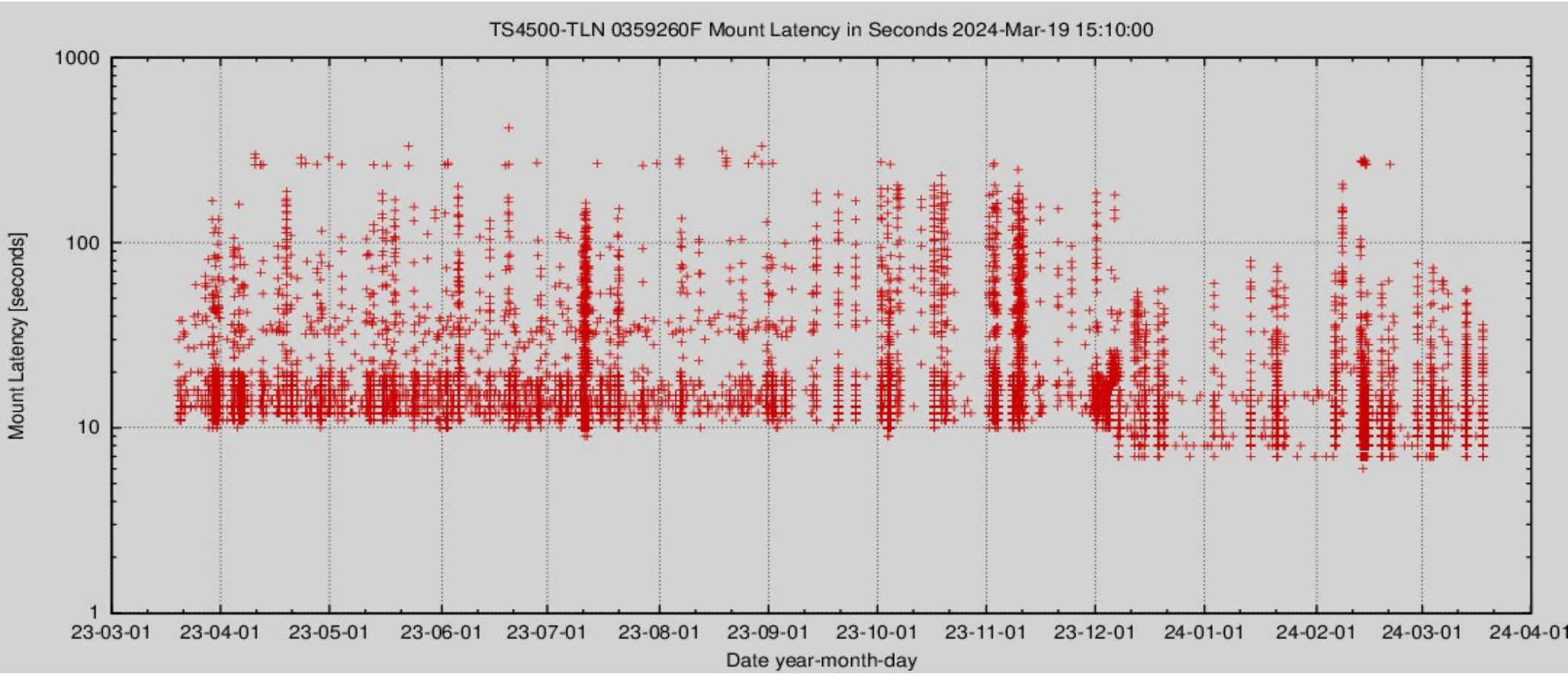
Использование raw udp – приемник / буферизатор сообщений, запускаемый сервисом как отдельный процесс

(Миграция на PostgreSQL 12. Изменения схем, продиктованные запретом OID)

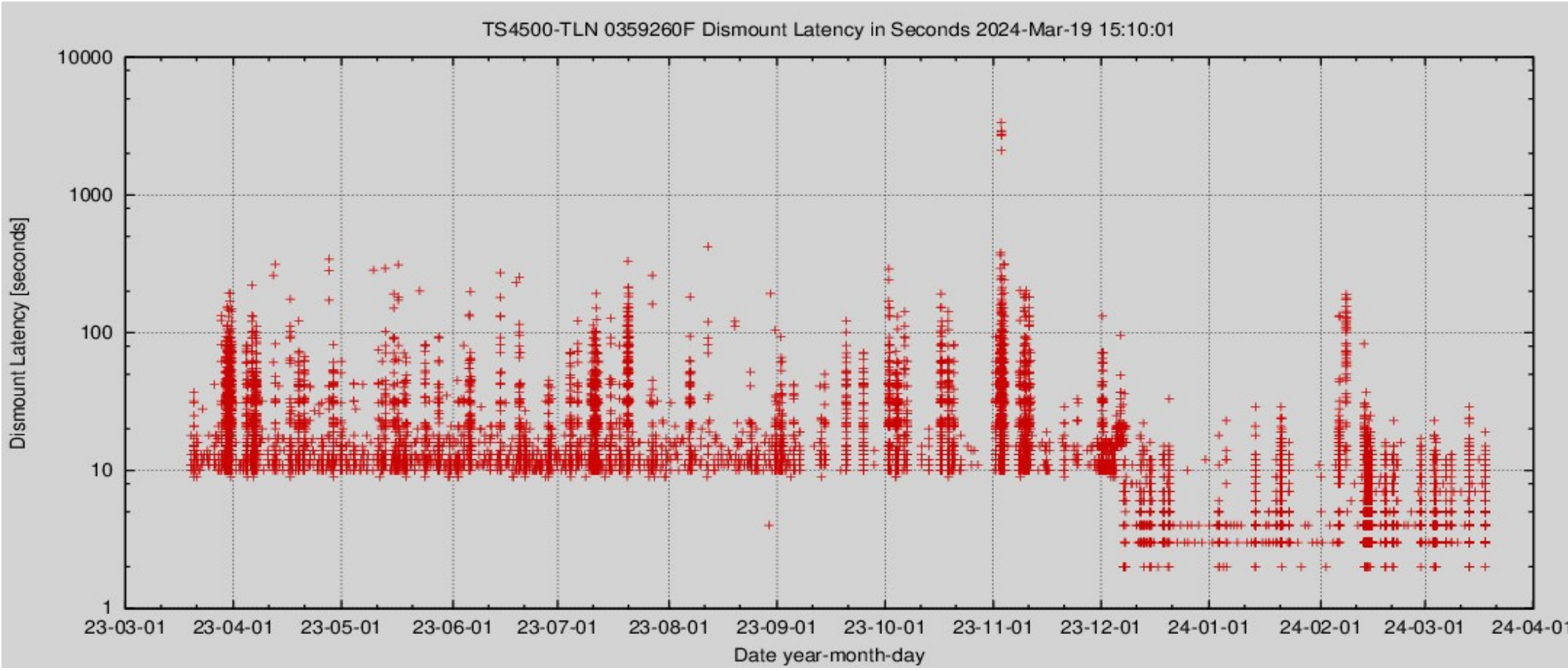
Web server

Backup всех баз данных Enstore, dCache, CTA скрипты для копирования на backup server (детальная информация сообщается через e-mail)

Время загрузки кассеты в tape drive (за год)



Время выгрузки кассеты из tape drive (за год)



Параметры обработки запросов

- Приоритет – определяет очередность
  - обычный – определяется в конфигурации или в запросе пользователя
  - Административный - определяется в конфигурации
- Дисциплина – ограничивает число активных запросов с определенных клиентских машин. Важна для обеспечения высокой скорости передачи и избежания bottlenecks при высокой загрузке Enstore
- Справедливая доля – ограничивает одновременное использование моверов (tape drives) для определенных групп пользователей

Все эти параметры в конфигурации Enstore ОИЯИ не используются.

Мониторирование с использованием веб сервера

Enstore web сервис предоставляет возможности мониторирования работы Enstore в целом и его компонент

Предоставляется информация:

О недавних передачах данных

Проблемах

Конфигурации

Графики

И т.д.

Другие средства монитирирования: cron jobs

Cron jobs задаются в кофигурации : configdict['crontabs']

Скрипт install_crons устанавливает cron jobs в */etc/cron.d*

Специальная программа – обертка запускает cron jobs и документирует результаты их работы в ~<user>/CRON

Другие средства монитирирования: email (1)

Адресаты задаются в конфугурации:

'crons': {'developer_email': 'moibenko@jinr.ru',

'email': 'tvv@jinr.ru',

...

Неработающий сервис:

Message from enstore_up_down.py:

Please check the full Enstore software system.

See the Status-at-a-Glance Web Page

Thu Mar 21 05:22:03 MSK 2024

TS4500-TLN.media_changer is not alive. Down counter 22

Thu Mar 21 05:22:03 MSK 2024

TS4500.library_manager is not alive. Down counter 22

Другие средства монитирирования: email (2)

Программный сбой:

Subject: Traceback found. Please investigate!

Date: Mon, 25 Mar 2024 11:01:02 +0300

From: enstore@enstore02.jinr-t1.ru

To: moibenko@jinr.ru

00:00:00 enstore01.jinr-t1.ru 044614 enstore E ACCSRV Traceback (most recent call last):

For more details check /diskb/enstore/enstore-log//LOG-2024-03-25

Другие средства монитирирования: email (3)

(Пере)запуск сервиса:

Subject: Output from your job 238

Date: Fri, 15 Mar 2024 16:18:13 +0300

From: root <root@dvl-es-mv01.jinr.ru>

To: root@dvl-es-mv01.jinr.ru

Checking mtx1.mover.

RTN {'status': ('TIMEDOUT', 'mtx1.mover')}

Starting mtx1.mover: 159.93.227.188:7540

Другие средства монитирирования: phone alerts

В ОИЯИ не задействован

Current Enstore production

IBM TS4500 tape library

9 0359260F (Jaguar) tape drives

2 System servers

3 Mover servers – 1 serving 3 tape drives

OS - Scientific Linux release 7.9 (Nitrogen)

Enstore rpm – enstore-6.3.4-19.8.el7.x86_64

Mtx rpm – mtx-1.3.12-19fnal_jinr.el7.x86_64 (fixes bug, seen at JINR only)

Конверсия Enstore

Среда и инструменты для конверсии Python 2 to 3 (local work station)

OS: Ubuntu 22.04.4 LTS

Environment: pyenv

Python 3.12.

Clean code: reindent (fix mixture of tabs and spaces)

PEP8 rules: autopep8

Конвертор: futurize

Автоматическая проверка ошибок: pylint

Исходный код

https://github.com/moibenko/enstore. Синхронизирован с https://github.com/Enstore-org/enstore 27 декабря 2023

Language	files	comment	code
Python	962	100487	317887
HTML	297	536	62117
C	141	8076	36851
Tcl/Tk	128	12335	30807
Bourne Shell	352	4947	25417
C/C++ Header	54	1377	2440
SQL	45	1053	2261
Bourne Again Shell	15	623	2159
make	26	392	1659
SWIG	10	239	1146
....			
SUM:	2154	130467	490663

Исходный код. Корневая директория

bin	crontabs	databases	DBUtils
dcache-deploy	doc	etc	external_distr
ftt	gadfly	helpDesk	HTMLgen
ingest	Makefile	modules	psycopg2
PyGreSQL	Python	README	release-notes
sbin	site_specific	spec	sphinx
src	SWIG	test	tools
ups	volume_import	www	xml2ddl

Последовательность для “pure” python фаза 1

Из исходного кода сделана фаза 1 (для каждой директории содержащей python code):

1. reindent *.py
2. futurize -w **--stage1** *.py > futurize.log 2>&1 # python 2
3. for f in `ls *.py`;do autopep8 --in-place --aggressive $f;done
4. pylint *.py > pylint.out
5. Проверка ошибок в pylint.out и их устранение в коде вручную

Отладка в Enstore development (python 2)

- Установлены необходимые пакеты rpm:
- gcc zlib-devel bzip2 bzip2-devel readline-devel sqlite sqlite-devel openssl-devel xz xz-devel libffi-devel libpq-dev, swig-2, qpid* (для SFA)
- Установлены pyenv, python 2.7.16, future, psycopg2-binary, PyGreSQL, qpid-python (для SFA).
- Реструктурирован код
- Отладка и устранение неисправностей
  - Особые трудности с C extensions

Корневая директория фаза 1 (python 2)

Удалены Python, psycopg2, PyGreSQL, SWIG. Используются соответствующие rpms и packages, установленные pip.

bin	crontabs	databases	DBUtils
dcache-deploy	doc	etc	external_distr
ftt	gadfly	helpDesk	HTMLgen
ingest	Makefile	modules	README
release-notes	sbin	site_specific	spec
sphinx	src	test	tools
ups	volume_import	www	xml2ddl

Код фаза 1

Language	files	comment	code
Python	624	65508	201112
HTML	289	455	58702
C	123	9377	52890
Tcl/Tk	128	12335	30807
Bourne Shell	351	4947	25384
SQL	45	1053	2261
Bourne Again Shell	15	623	2159
C/C++ Header	33	929	1556
make	23	371	1546
...			
SUM:	1679	95844	380582

Исходный код и последовательность для “pure” python фаза 2

Из кода полученного в фазе 1 сделана фаза 2 (для каждой директории, содержащей python code):

1. reindent *.py
2. futurize -w **--stage2** *.py > futurize.log 2>&1 # python 3
3. for f in `ls *.py`;do autopep8 --in-place --aggressive $f;done
4. pylint *.py > pylint.out
5. Проверка ошибок в pylint.out и их устранение в коде вручную

Отладка в Enstore development (python 3)

- Установлены python 3.9.18 (SL 7.9 не позволяет более позднюю), future, psycopg2-binary, PyGreSQL, tkinter.
- Отладка и устранение неисправностей
  - Особые трудности с C extensions – существенные изменения в python3

Код фаза 2

Language	files	comment	code
Python	654	80176	235176
HTML	289	455	58702
C	120	9150	52439
Bourne Shell	339	4419	23786
SQL	45	1053	2261
Bourne Again Shell	16	623	2189
XML	8	0	1615
make	24	348	1573
C/C++ Header	32	326	1429
Tcl/Tk	2	0	13
....			
SUM:	1569	96796	381733

Текущее состояние

- Работают практически все компоненты, но наблюдается нестабильность.
- Трудно идентифицировать некоторые наблюдаемые проблемы.
- Small Files Aggregation не работает. Необходимо заменить коммуникационный пакет. Использовавшийся Apache Qpid не работает с Python3. Кандидаты на замену Apache Proton Qpid, Rabbit MQ. Rabbit MQ представляется наиболее подходящим кандидатом, т.к. является довольно активным проектом и имеет хорошую документацию. В Enstore ОИЯИ SFA не используется.
- Удачная попытка генерации executables (python binaries) с использованием pyinstaller. Требует дополнительной работы по построению и тестированию.

Предстоящие работы

- Отладить CGI скрипты.
- Всеобъемлющее тестирование и устранение выявленных проблем.
- Ревизия кода, удаление или размещение в отдельных местах старых неиспользуемых частей.
- Ревизия схем баз данных.
- Выбрать способ(ы) упаковки продукта и реализовать их
- Ревизия web сервиса, с возможной заменой. Замена http на https
- Синхронизация с Enstore.org
- Написать документацию.
- Выбрать новый коммуникационный пакет и переделать коммуникационный интерфейс в Small Files Aggregation.
- Сконфигурировать Реплики Баз данных (может понадобиться дополнительный сервер)

Благодарности

Шматову С.В., Коренькову В.В., Стриж Т.А. - за инициацию проекта, полезные обсуждения и поддержку.

Мицыну В.В, Трофимову В.В. - за полезные обсуждения и помощь в работе.

Долбилову А.Г., Голунову А.О. - за помощь в организации работы.

Использованные материалы

Supporting Python 3 - http://python3porting.com/bookindex.html

Futurize - https://python-future.org/futurize.html

py3c reference - https://py3c.readthedocs.io/en/latest/reference.html

Subprocess management - https://docs.python.org/3/library/subprocess.html#