Enstore at JINR.

Short description and current status

A.N. Moibenko

July 8 2025

Enstore is 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) for Tevatron Run 2 experiments (~ 20 PB, 0.5GB/s aggregate throughput).

Integrated with dCache (DESY, Germany) for files bufferization and caching to achieve high throughput.

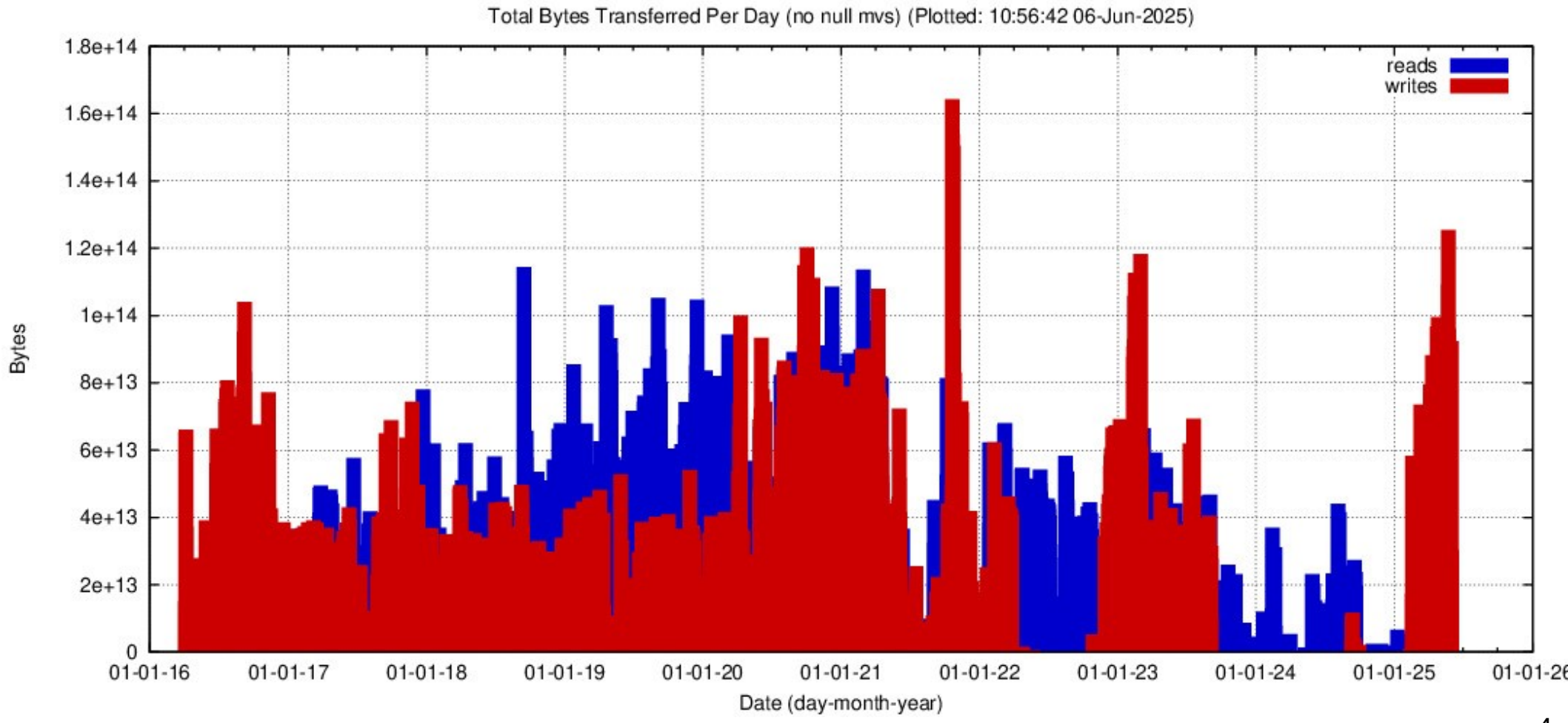
Substantial modifications for US CMS T1 data storage and transfers.

Small files Aggregation for effective storage, access and transfer of files < 500 MB essentially for neutrino program.

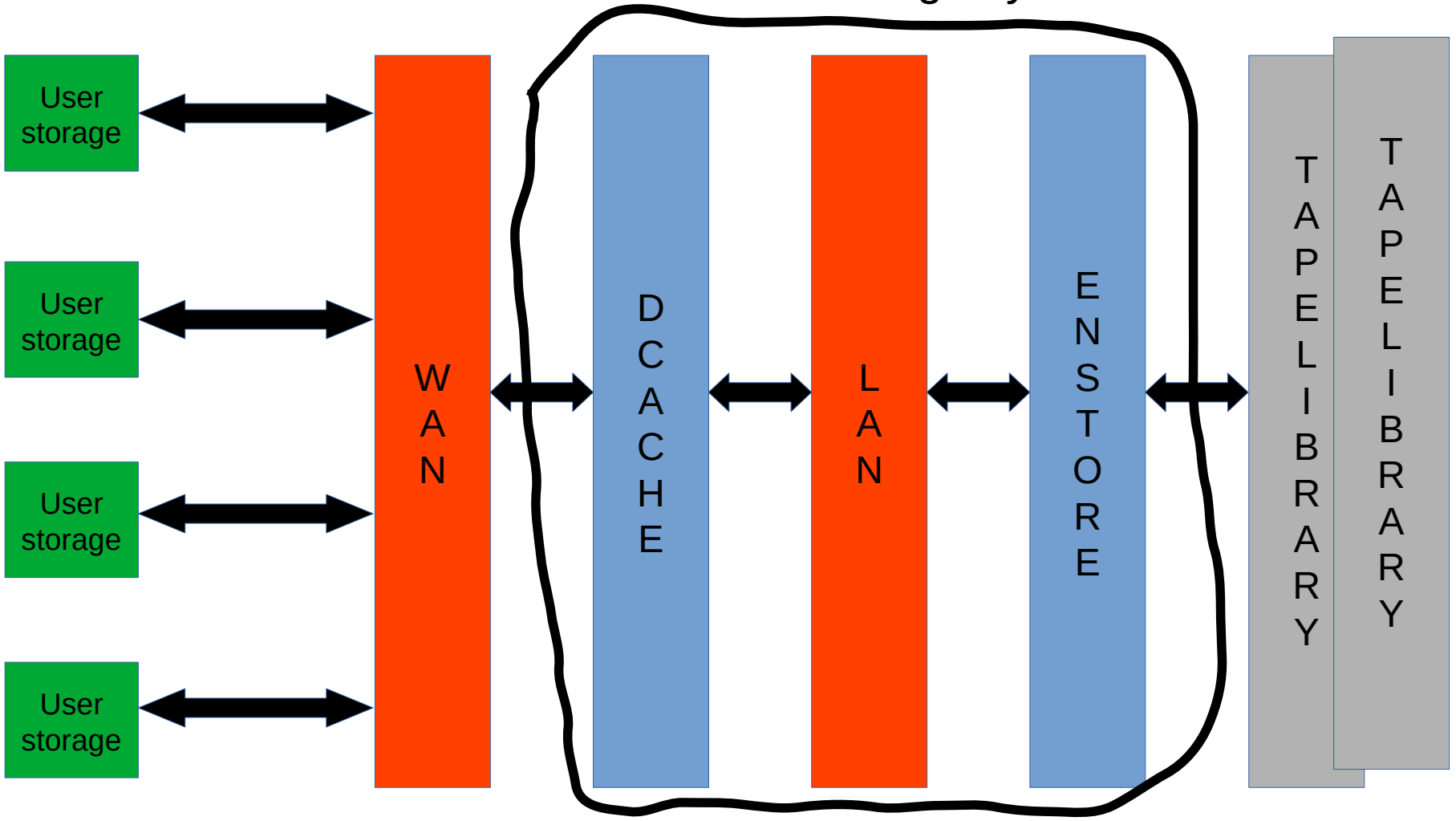
Scalable from one work station and one tape drive to hundreds of worker nodes, several tape libraries of different types and hundreds of tape drives. The biggest working configuration:

- 8 Tape Libraries
- > 190 Tape drives
- > 120 worker nodes
- 350 PB stored data
- 800 TB/day (~9GB /s)

Enstore at JINR (CMS T1 tape storage): general activity (15.8 PB – total, 15.4 PB - active)



Enstore in data transfer and storage system



General Features

End-to-end checksum

Optimized access to user data by utilizing steering

information stored in PNFS directory tags:

- – Enstore puts files on tape in the order the files were submitted.
- – Files are grouped on tape using file family and file family width scheme.

Utilities to query tape content.

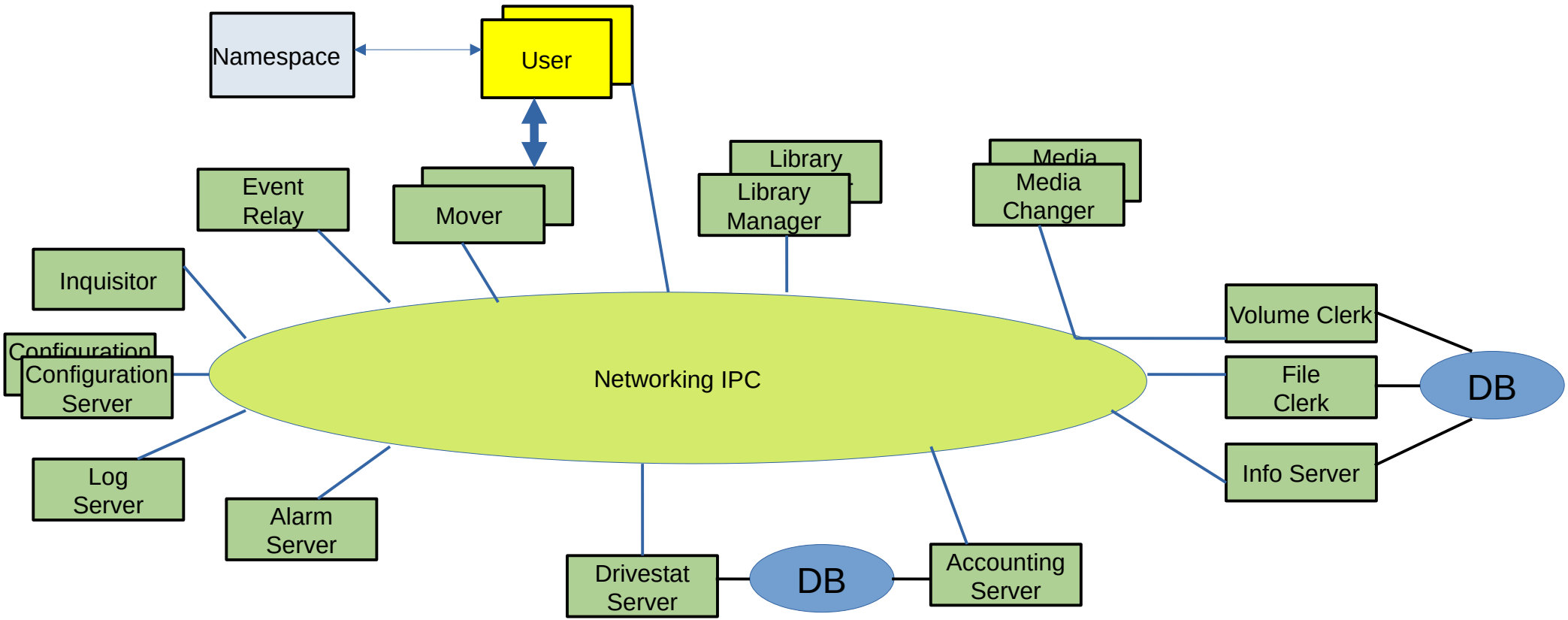
Filesystem-like view of user stored data (thanks to PNFS/Chimera).

Policy driven small file aggregation.

Design considerations

Client/Server Architecture
– Reuse python server framework
Networked distributed drive access (crucial for meeting scalability requirements)
Portability
– Python
– Time critical code in C
Communication Protocols
– UDP for control request/response
• Messages fit in 1 datagram
• Retries, unique ID
• Clients can not hang servers
– TCP for data transfers
Products reuse
– Fermilab FTT − portable hooks for handling tape drives
– PNFS (later chimera) namespace
– Apache web server for web based monitoring and admin interface
– PostgreSQL DB
– Gnuplot utility for monitoring and performance metrics
– crond for scheduling various ancillary tasks

Enstore Structure



Configuration Server (CS)

- Configuration Server maintains and distributes all information about
- system configuration such as host, port and other parameters of
- each server.
- On startup, each Enstore server queries the Configuration Server
- for :
- – Information on how to setup itself.
- – Locations of other servers it needs to communicate.
- A new configuration can be loaded into the Configuration Server
- from a file without disrupting the running system.
- Configuration is stored in a file in a form of python dictionary

Library Manager (LM)

- Receives write / read requests from clients to the virtual library of tapes, keeps them and directs to movers.
- Virtual library – set of tapes of one type in the physical tape library and muvers, performing operations on these volumes.
- LM receives requests from clients, sorts them according priorities and positionf on tape.
- Waits on mover request for work and directs to it the client request.
- Read requests addressed to one mover get sent to mover according to their position on tape.
- There are flexible creteria of prioritization and restriction of simultaneously writable tapes.
- There is a way to restrict the number of simultaneous transfers from the same client host to aviod network overloading.

Volume Clerk (VC)

- Keeps information about all volumes (tapes, e.t.c.) in the system.
- Refreshes information according to requests, coming from administrator and servers.
- Data for each volume is kept in DB (volume table).
- For each volume there is a single record with unique key on volume id.
- Volume Clerk responsibility is to:
  - Assign new volumes
  - Draw volumes for write on request from Library Manager
  - Provide interface to query and modify volume information
  - Maintain tape quotas by storage_group and library

File Clerk (FC)

- Keeps information about all files in the system.
- Data is kept in Data Base in records keyed by unique BFID.
- File Clerk provides:
  - Generation of unique BFIDs
  - DB connection handling (multiple threads).
  - Pool of threads for request handling.

Info Server

- Provides information about files and volumes.
- Essentially duplicates read-only functionality provided by File, Volume Clerks.
- Reduces load on File, Volume Clerks
- File, Volume Clerks and Info Server have similar structure.

Media Changer (MC)

- Mounts / dismount tapes into tape drives.
- In last versions serves a coordinator of requests, performed by movers.

Mover

- Asks Library Manager for read / write request from client and makes connection to client.
- Receives data from client and writes it to tape.
- Reads data from tape and send it to client.
- For data transfer TCP/IP transfer protocol is used.
- Data transfer is accompanied with calculation of checksum on client mover sides.
- Data bufferization is used for transfer rates adaptation.

Log Server (MC)

- Receives messages from system components and writes then into formatted log file.
- At 00:00 a new log file gets open.
- UDP protocol is used for message transfers.

Inquisitor (Inq)

- Monitors status of the whole system and each server.
- Results get periodically written to web-pages, files, plots, and e-mail messages.
  - Example: http://enstore02.jinr-t1.ru/enstore/status_enstore_system.html

Alarm Server (AS)

- Receives and maintains alarm messages sent by Enstore components.
  - Example:
  - http://enstore02.jinr-t1.ru/enstore/enstore_alarms.html

Event Relay

- Receives events and directs them to subscribers.
- Example: reload of configuration changes into configuration server.

Name Space (Chimera, DESY Germany)

- Used to:
  - Present user data as familiar hierarchical filesystem
  - Store information related to Enstore in associated “layers”
  - File:
    - Layer 1: BFID
    - Layer 4: additional information
  - Directory (tags):
    - storage_group (usually name of an experiment)
    - file_family - name of a dataset to be grouped on the same tape/set of
    - tapes
    - file_family_width – how many movers can be used simultaneously
    - file_family_wrapper – file wrapper type
    - library – name of virtual library

Directory tag example

- enstore01:~ > enstore sfs --tags /pnfs/jinr-t1.ru/data/cms/store/mc
  - .(tag)(file_family) = dcache
  - .(tag)(file_family_width) = 8
  - .(tag)(file_family_wrapper) = cpio_odc
  - .(tag)(library) = TS4500
  - .(tag)(OSMTemplate) = StoreName cms
  - .(tag)(sGroup) = STATIC
  - .(tag)(storage_group) = cms

File layers example

- Layer 1
- enstore01:~ > enstore sfs --layer /pnfs/jinr-t1.ru/data/cms/store/data/Run2017D/Charmonium/AOD/09Aug2019_UL2017-v1/30000/1845E4EB-C67C-9847-96FE-BB034B468F03.root 1
- DEMO167435604600000 - BFID
- Layer 4
- enstore01:~ > enstore sfs --layer /pnfs/jinr-t1.ru/data/cms/store/data/Run2017D/Charmonium/AOD/09Aug2019_UL2017-v1/30000/1845E4EB-C67C-9847-96FE-BB034B468F03.root 4
- 000297JE - Volume
- 0000_000000000_0004223 – File position on tape
- 3328010436 – file size
- dcache – file family
- /pnfs/jinr-t1.ru/data/cms/store/data/Run2017D/Charmonium/AOD/09Aug2019_UL2017-v1/30000/1845E4EB-C67C-9847-96FE-BB034B468F03.root – full file name
- 0000D18A57D1F7FC4C5D9FC37F0585EF8E2A – pnfs ID
- DEMO167435604600000 - BFID
- enst-rmt01.jinr-t1.ru:/dev/rmt/tps17d0:00000781608A – device, file written by
- 1261236215 - CRC

Enstore user client (encp)

- Copies use data (dcache is enstore user too) to / from Enstore.

- enstore01:~ > encp --help
- Usage:
- encp [OPTIONS]... <source file> <destination file>
- encp [OPTIONS]... <source file> [source file [...]] <destination directory>
- encp [OPTIONS]... --get-bfid <bfid> <destination file>
- encp [OPTIONS]... --get-cache <pnfs|chimera id> <destination file>
- encp [OPTIONS]... --put-cache <pnfs|chimera id> <source file>

Enstore user client (encp)

- Write: destination file/directory is namespace.
- Read: source file is namespace.
- Distributes as statically linked executable produced with Python freeze tool
- => Requires no dependencies. - old version
- Produced with pyinstaller – new version
- Control communicatios are done using UDP to avoid hangs of Enstore servers.
- For data transfers TCP / IP is used.
- Provides end-to-end checksum.

Request processing parameters

- Priority
  - regular – defined in configuration of user request
  - administrative – defined in configuration
- Discipline – restricts the number of active requests from certain client hosts. Used to provide high data transfer rates and avoid bottlenecks during high load on Enstore
- Fair share – restricts simultaneous use of movers (tape drives) for certain groups of users.

Web monitoring

Enstore web provides monitoring of Enstore as whole or its components

Information is provided about:

Recent data transfers

Problems

Configuration

Plots

Main page

CMS

Enstore System Status

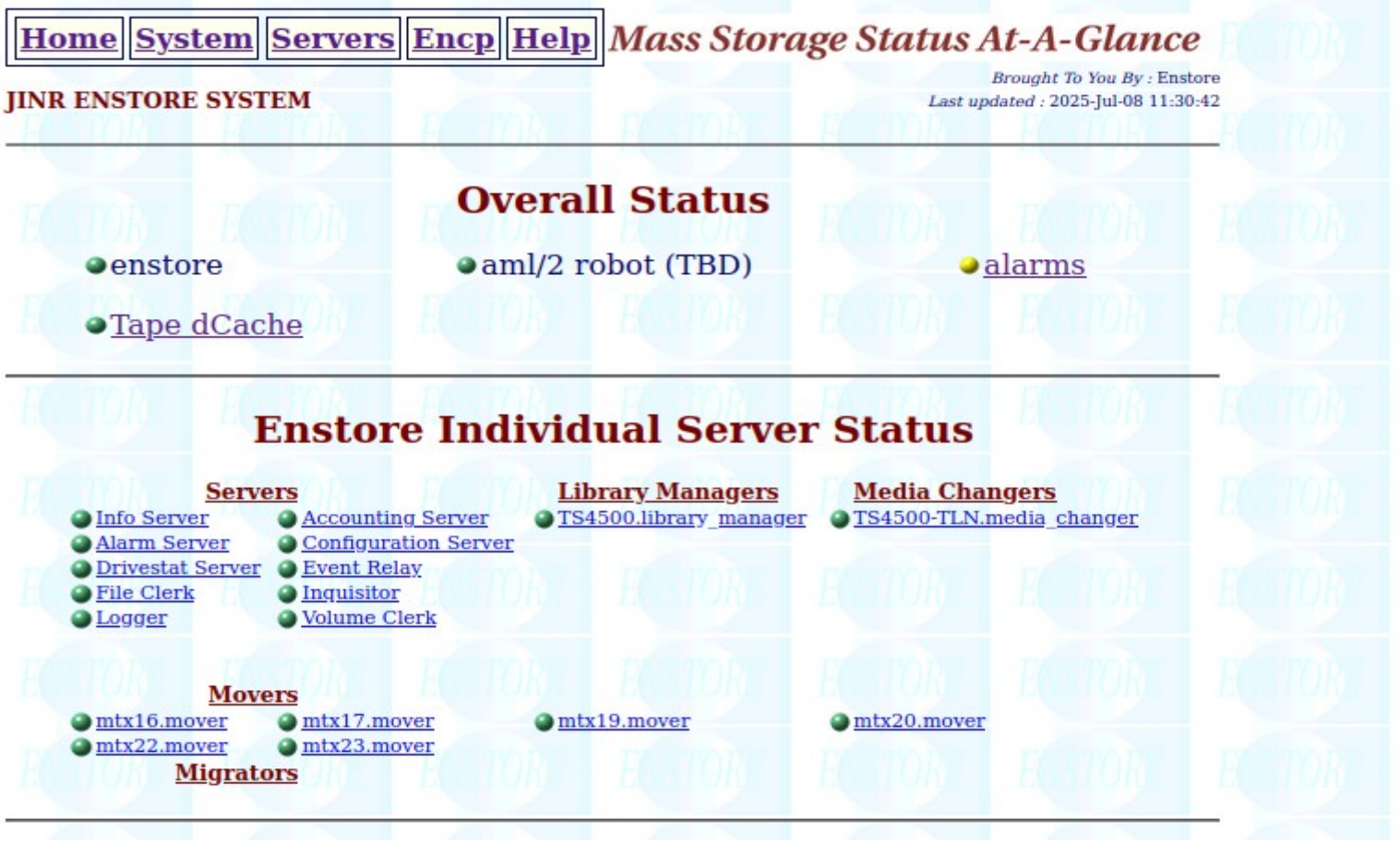
User Data on Tape [TiB]	
active	total
15413.15	15908.83

Enstore System Summary	Enstore System Status-At-A-Glance
Enstore Server Status	Current status of the Enstore servers
encp History	History of recent encp requests
Configuration	Current Enstore System Configuration
Alarms	Active alarms and alarm history
Log Files	Hardware and software log files
Quota and Usage	How tapes are allocated and being used
Plots	Enstore Plots
Web Pages	Enstore Web Pages
Active Volumes	Currently Active Volumes per Library
Tape Inventory Summary	Summary of inventory results
Tape Inventory	Detailed list of tapes and their contents
Cronjob Status	Lots of cronjob exit status for past week

Information

Mass Storage System Documentation Page	Documentation, reports, talks for Enstore and dCache

Overall status



Server status

Name	Status	Host	Date/ Time	Last Time Alive
accounting_server	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:36:59	
alarm_server	alive	enstore02.jinr-t1.ru	2025-Jul-08 11:36:33	
drivestat_server	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:36:57	
event_relay	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:37:02	
file_clerk	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:36:42	
info_server	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:36:42	
inquisitor	alive	enstore02.jinr-t1.ru	2025-Jul-08 11:37:02	
log_server	alive	enstore02.jinr-t1.ru	2025-Jul-08 11:36:54	
volume_clerk	alive	enstore01.jinr-t1.ru	2025-Jul-08 11:36:42	
TS4500-TLN.media_changer	alive	enst-rmt03.jinr-t1.ru	2025-Jul-08 11:36:34	
TS4500.library_manager	alive : unlocked	enst-rmt03.jinr-t1.ru	2025-Jul-08 11:36:42	
	Ongoing Transfers 2 Pending Transfers 4 Full Queue Elements Writing 000683JE(cms dcache-MIGRATION) using mtx19.mover from enstore02.jinr-t1.ru by root Reading 000935JE(cms dcache) using mtx22.mover from enstore02.jinr-t1.ru by root Pending read of 000935JE from enstore02.jinr-t1.ru by root [] Pending read of 000935JE from enstore02.jinr-t1.ru by root [] Pending write for cms dcache-MIGRATION from enstore02.jinr-t1.ru by root [VOLS_IN_WORK] Pending write for cms dcache-MIGRATION from enstore02.jinr-t1.ru by root []			
TS4500_tst.library_manager	alive : unlocked	enst-rmt03.jinr-t1.ru	2025-Jul-08 11:36:22	
	Ongoing Transfers 0 Pending Transfers 0 Full Queue Elements			
mtx13.mover	alive : IDLE	enst-rmt01.jinr-t1.ru	2025-Jul-08 11:36:42	
mtx14.mover	alive : IDLE	enst-rmt01.jinr-t1.ru	2025-Jul-08 11:37:02	
mtx16.mover	alive : IDLE	enst-rmt02.jinr-t1.ru	2025-Jul-08 11:37:02	
mtx17.mover	alive : IDLE	enst-rmt02.jinr-t1.ru	2025-Jul-08 11:37:02	
mtx19.mover	alive : busy writing 2,674,409,684 bytes to 000683JE	enst-rmt03.jinr-t1.ru	2025-Jul-08 11:36:42	
mtx20.mover	alive : IDLE	enst-rmt03.jinr-t1.ru	2025-Jul-08 11:36:42	
mtx22.mover	alive : busy reading 4,210,634,692 bytes from 000935JE	enst-rmt04.jinr-t1.ru	2025-Jul-08 11:37:02	
mtx23.mover	alive : IDLE	enst-rmt04.jinr-t1.ru	2025-Jul-08 11:36:42	

Latest data transfers

Home | System | Servers | Encp | Help

JINR ENSTORE SYSTEM

Encp History

Brought To You By : The Inquisitor
Last updated : 2025-Jul-08 11:39:28

Time	Node	User/ Storage Group	Mover Interface	Bytes	Volume	Network Rate (MB/ S)	Transfer Rate (MB/ S)	Drive Rate (MB/S)	Disk Rate (MB/S)	Overall Rate (MB/ S)
2025-07-08 11:39:20	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	2314814342 (1)	from 000935JE	1147.61	229.73	378.90	231.34	227.11
2025-07-08 11:39:12	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	1144710 (2)	to 000683JE	1514.17	8.86	36.05	140.46	1.65
2025-07-08 11:39:11	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	2295526894 (3)	to 000683JE	2912.38	129.02	1021.47	129.73	125.69
2025-07-08 11:39:10	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	3950565334 (4)	from 000935JE	1063.49	230.06	368.41	232.38	189.24
2025-07-08 11:38:54	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	2303076182 (5)	to 000683JE	3098.77	202.66	394.27	204.23	193.49
2025-07-08 11:38:51	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	2298557557 (6)	from 000935JE	875.51	255.74	366.26	271.82	153.98
2025-07-08 11:38:42	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	2610927304 (7)	to 000683JE	2710.21	130.06	409.67	153.80	126.97
2025-07-08 11:38:36	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	4081111728 (8)	from 000935JE	1051.37	229.47	429.20	230.32	227.64
2025-07-08 11:38:22	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	3664483588 (9)	to 000683JE	2934.74	127.88	1026.55	128.35	125.28
2025-07-08 11:38:19	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	38426552 (10)	from 000935JE	476.87	195.20	526.07	270.96	115.40
2025-07-08 11:38:18	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	2296989064 (11)	from 000935JE	981.01	226.65	382.62	228.03	223.40
2025-07-08 11:38:08	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	3820836806 (12)	from 000935JE	1149.71	232.86	367.92	234.65	152.13
2025-07-08 11:37:54	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	2321957589 (13)	to 000683JE	3551.37	177.97	323.80	269.36	171.23
2025-07-08 11:37:45	enstore02.jinr-t1.ru	root/cms	enst-rmt04.jinr-t1.ru	3013292065 (14)	from 000935JE	986.13	234.43	378.18	235.76	232.09
2025-07-08 11:37:41	enstore02.jinr-t1.ru	root/cms	enst-rmt03.jinr-t1.ru	2313408760 (15)	to 000683JE	2453.81	114.53	1015.05	114.99	111.88

Alarms

Home | System | Servers | Encp | Help

Enstore Active Alarms

Brought To You By : The Alarm Server
Last updated : 2025-Jun-23 16:33:55

JINR ENSTORE SYSTEM

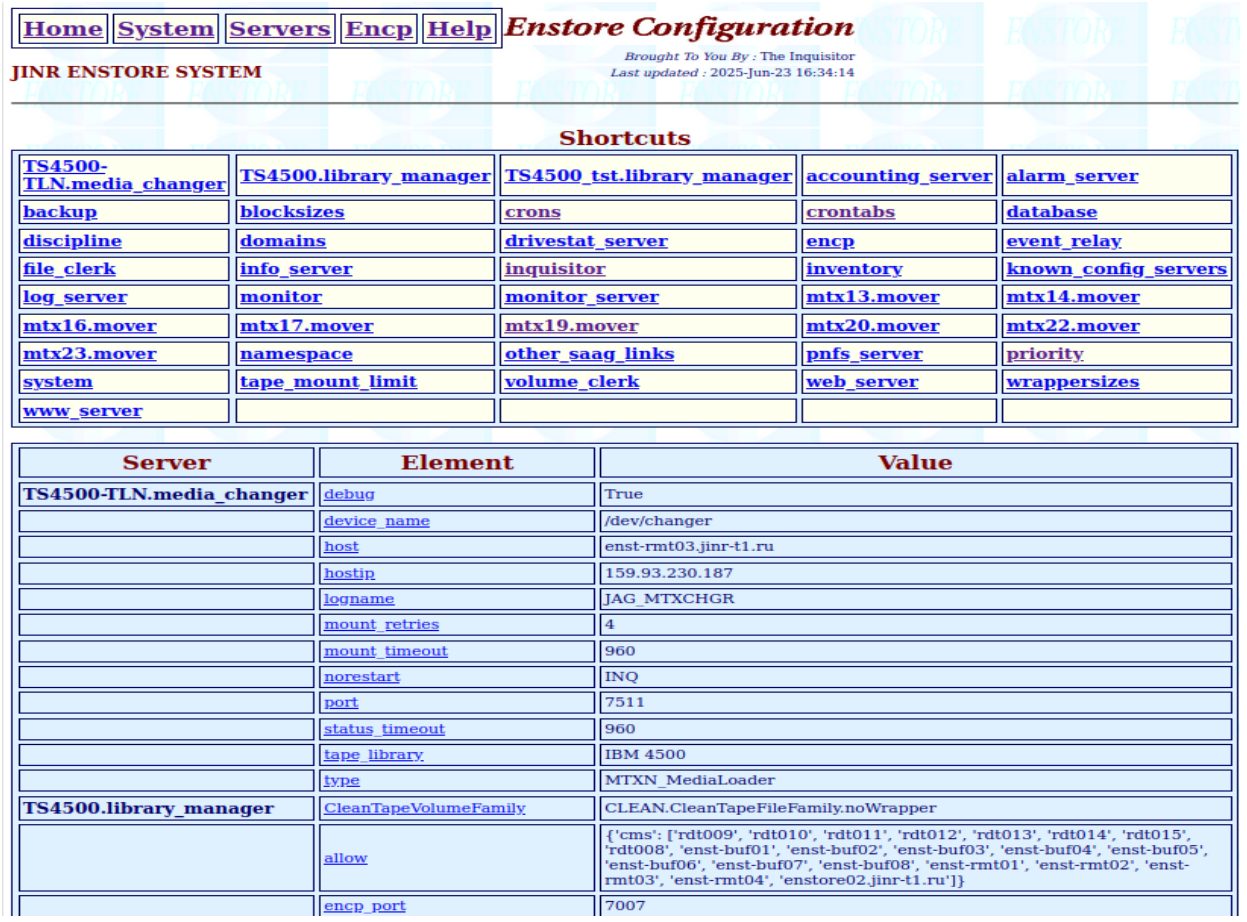
Previous alarms may also be displayed.
And a **volume audit**.

Resolve Selected | Resolve All | Reset

Alarms may be cancelled by selecting the alarm(s), pressing the **Resolve Selected** button and then reloading the page. All alarms may be cancelled by pressing the **Resolve All** button.

Key	Time (last)	Node	PID	User	Severity	Process	Error	Ticket Generated (Condition/ Type)	Additional Information
☐ 1741097816.07	2025-Mar-04 17:16:56 *(2025-Jun-23 16:33:55)*	enstore01.jinr-t1.ru	22548	5744	I (15)	configuration_server	Configuration reloaded from ::ffff:159.93.229.103		{'text': {'configfile': '/home/enstore/site_specific/config/enstore_system.conf', 'user': 'enstore'}}
☐ 1745842468.29	2025-Apr-28 15:14:28 *(2025-Jun-18 16:44:14)*	enstore02.jinr-t1.ru	7314	enstore	E (2)	INQSRV	SERVERDIED		{'text': {'server': 'mtx19.mover'}}
☐ 1750254187.3	2025-Jun-18 16:43:07 *(2025-Jun-18 16:43:07)*	enst-rmt03.jinr-t1.ru	281365	root	E (1)	MVRRDT019MC	mtx server not responded in 960 s. Exiting		{'text': {}}
☐ 1747641375.33	2025-May-19 10:56:15 *(2025-Jun-18 16:40:33)*	enstore02.jinr-t1.ru	7314	enstore	E (2)	INQSRV	SERVERDIED		{'text': {'server': 'mtx23.mover'}}

Configuration



Home | System | Servers | Encp | Help

Enstore Configuration

Brought To You By : The Inquisitor
Last updated : 2025-Jun-23 16:34:14

JINR ENSTORE SYSTEM

Shortcuts

TS4500-TLN.media_changer	TS4500.library_manager	TS4500_tst.library_manager	accounting_server	alarm_server
backup	blocksizes	crons	crontabs	database
discipline	domains	drivestat_server	encp	event_relay
file_clerk	info_server	inquisitor	inventory	known_config_servers
log_server	monitor	monitor_server	mtx13.mover	mtx14.mover
mtx16.mover	mtx17.mover	mtx19.mover	mtx20.mover	mtx22.mover
mtx23.mover	namespace	other_saag_links	pnfs_server	priority
system	tape_mount_limit	volume_clerk	web_server	wrappersizes
www_server				

Server	Element	Value
TS4500-TLN.media_changer	debug	True
	device_name	/dev/changer
	host	enst-rmt03.jinr-t1.ru
	hostip	159.93.230.187
	logname	JAG_MTXCHGR
	mount_retries	4
	mount_timeout	960
	norestart	INQ
	port	7511
	status_timeout	960
	tape_library	IBM 4500
	type	MTXN_MediaLoader
TS4500.library_manager	CleanTapeVolumeFamily	CLEAN.CleanTapeFileFamily.noWrapper
	allow	{'cms': ['rdt009', 'rdt010', 'rdt011', 'rdt012', 'rdt013', 'rdt014', 'rdt015', 'rdt008', 'enst-buf01', 'enst-buf02', 'enst-buf03', 'enst-buf04', 'enst-buf05', 'enst-buf06', 'enst-buf07', 'enst-buf08', 'enst-rmt01', 'enst-rmt02', 'enst-rmt03', 'enst-rmt04', 'enstore02.jinr-t1.ru']}
	encp_port	7007

Plots

Home | System | Servers | Encp | Help

Enstore Plots

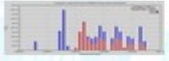
Brought To You By : The Inquisitor
Last updated : 2025-Jul-08 11:45:02

Enstore Plots

Tape Drive usage hours per drive type, stacked by storage group
Tape Drive usage hours per drive type, separately for each storage group
Tape occupancies per Storage Group Plots
Files read and written per mount per drive type, stacked by storage group
Files read and written per mount per drive type, separately for each storage group
Encp rates per Storage Group Plots
Quota per Storage Group Plots
Bytes Written per Storage Group Plots
Bytes/Day per Mover Plots
Xfer size per Storage Group Plots
Migration/Duplication Summary Plots per Media Type
Mover Plots
Mount Latency plots
Mounts/day per tape library
Small Files Aggregation Statistics

0359260F Mounts/Day (no null mvs as of 8/1/01) (2025-Jul-08 06:30:32)
(postscript)

IBM-4500_0_all (2025-Jul-08 11:31:02)
(postscript)



0359260F Bytes/Day (2025-Jul-08 06:30:05)
(postscript)

Other monitoring tools: cron jobs

Cron jobs defined in configuration: configdict['crontabs']

install_crons script installs cron jobs в */etc/cron.d*

Special script – wrapper starts cron jobs and documents results of their work in ~<user>/CRON

Other monitoring tools: email (1)

- Addresses are defined in configuration:

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

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

...

Not working service:

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

Other monitoring tools: email (2)

Program fault:

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

Other monitoring tools: email (3)

Service (re)start:

Subject: Output from your job 238

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

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

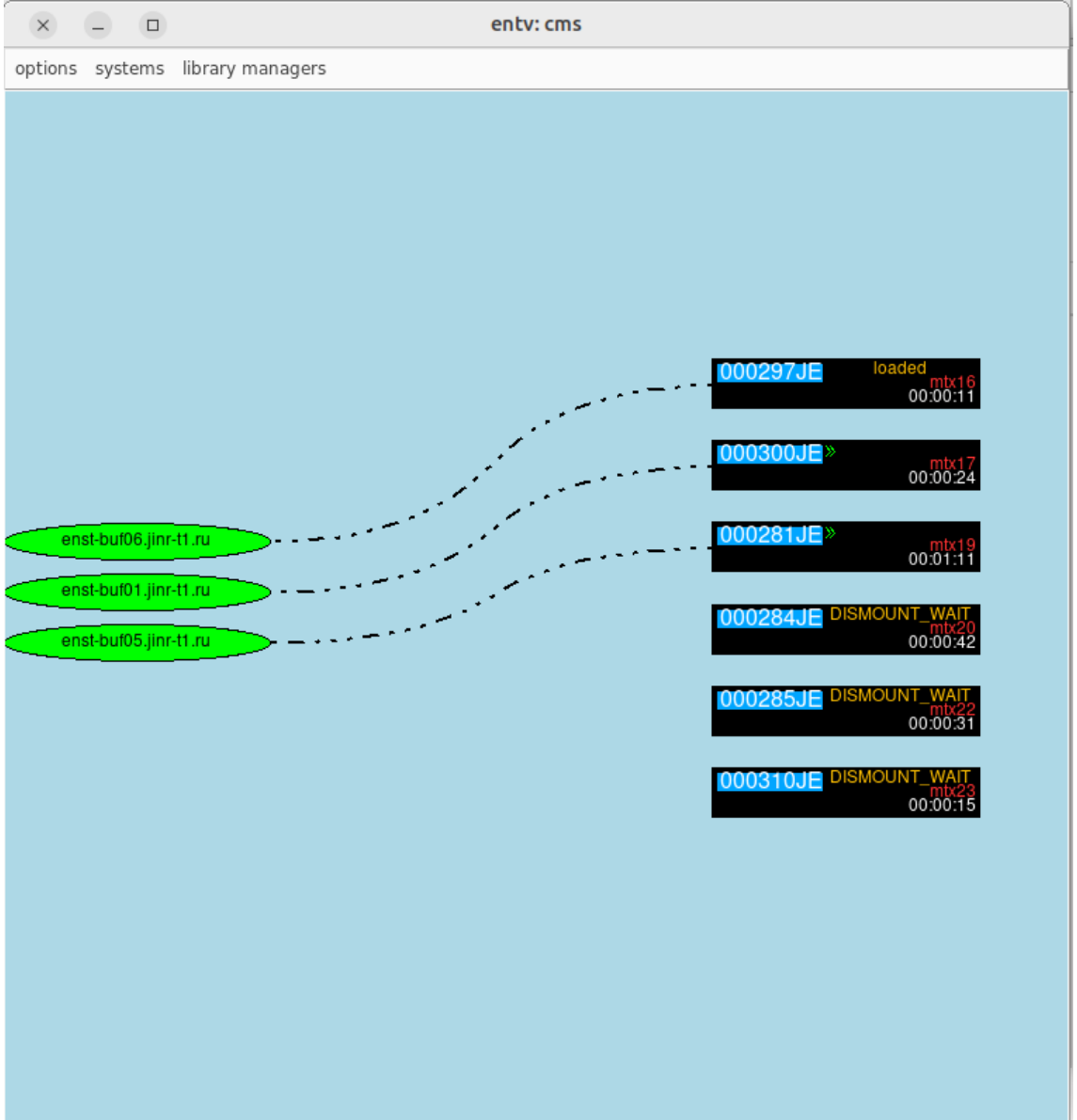
To: root@dvl-es-mv01.jinr.ru

Checking mtx1.mover.

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

Starting mtx1.mover: 159.93.227.188:7540

entv


entv: cms
options
systems
library managers
000297JE
loaded
mtx16
00:00:11
000300JE
mtx17
00:00:24
000281JE
mtx19
00:01:11
000284JE
DISMOUNT_WAIT
mtx20
00:00:42
000285JE
DISMOUNT_WAIT
mtx22
00:00:31
000310JE
DISMOUNT_WAIT
mtx23
00:00:15
enst-buf06.jinr-t1.ru
enst-buf01.jinr-t1.ru
enst-buf05.jinr-t1.ru

Enstore production (CMS T1 Tape storage)

Tape library - IBM TS4500

8 0359260F (Jaguar) tape drives

2 system nodes

4 mover nodes – 1 serves 2 tape drives

OS - Scientific Linux release 7.9 (Nitrogen)

Enstore clients (encp) were built for dcache pool nodes under OS Alma Linux 9 with python2

Enstore-Python3 under OS Alma Linux 9 is tested and ready

Thank you for attention