New technology for processing and storing data - DAOS

Matveev Mikhail

on behalf of the

Heterogeneous Computation Team HybriLIT

Meshcheryakov Laboratory of Information Technologies, JINR

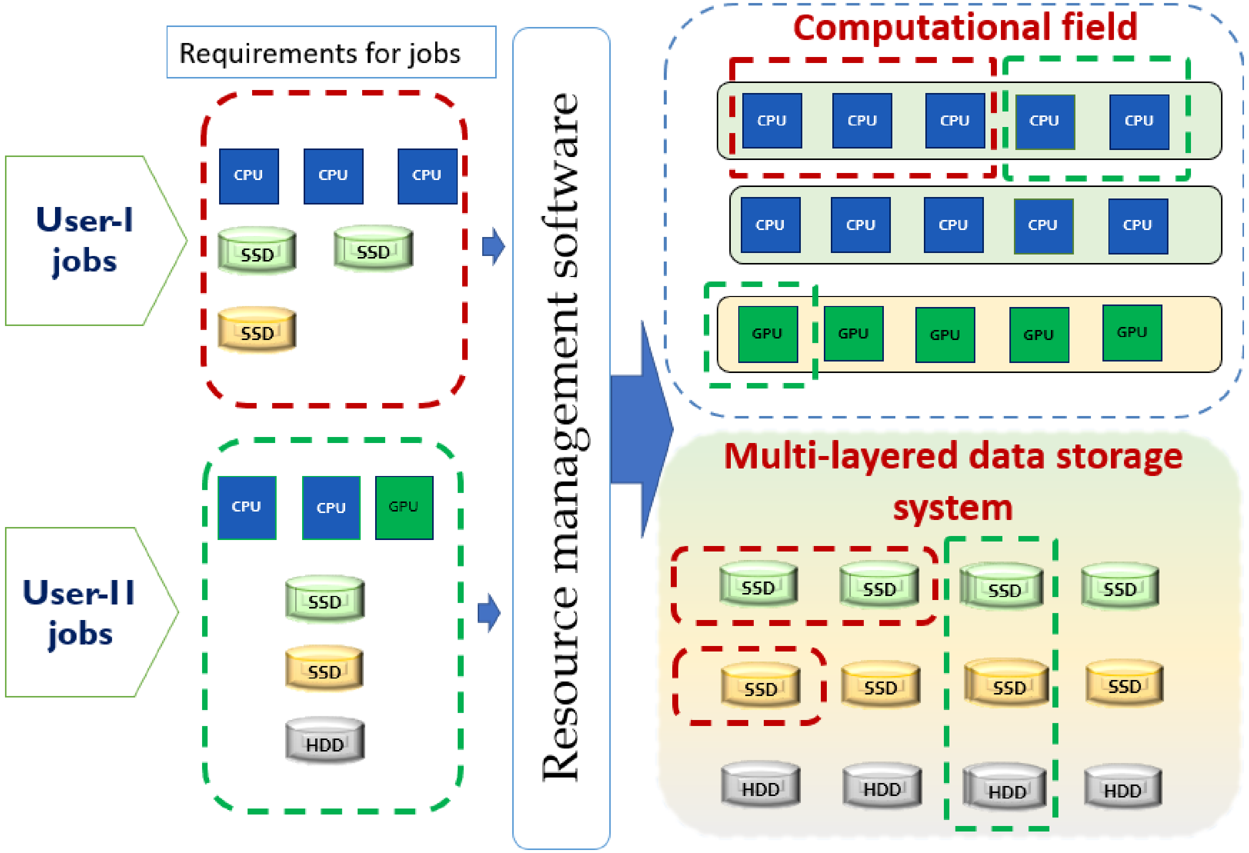
10.06.2021

matveevma@jinr.ru

THE MAIN PART OF HYBRILIT – GOVORUN SUPERCOMPUTER

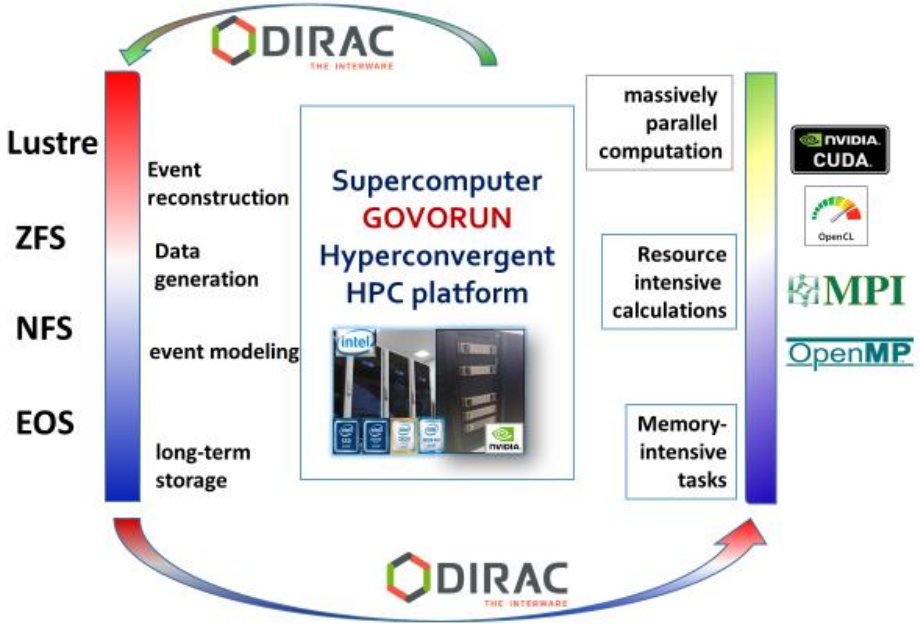
«**Govorun**» is a computing system for fast processing of big data, including the «**NICA**» project. One of the main tasks of the group in this direction is the introduction of new technologies to increase the efficiency of data processing.

One of the priority areas for the development of the supercomputer «**Govorun**» is to increase the volume of data storage and improve access parameters. **DAOS** (Distributed Asynchronous Object Storage) storage system is being implemented, which allows the use of NVMe non-volatile memory and also supports Optane DC Persistent Memory.

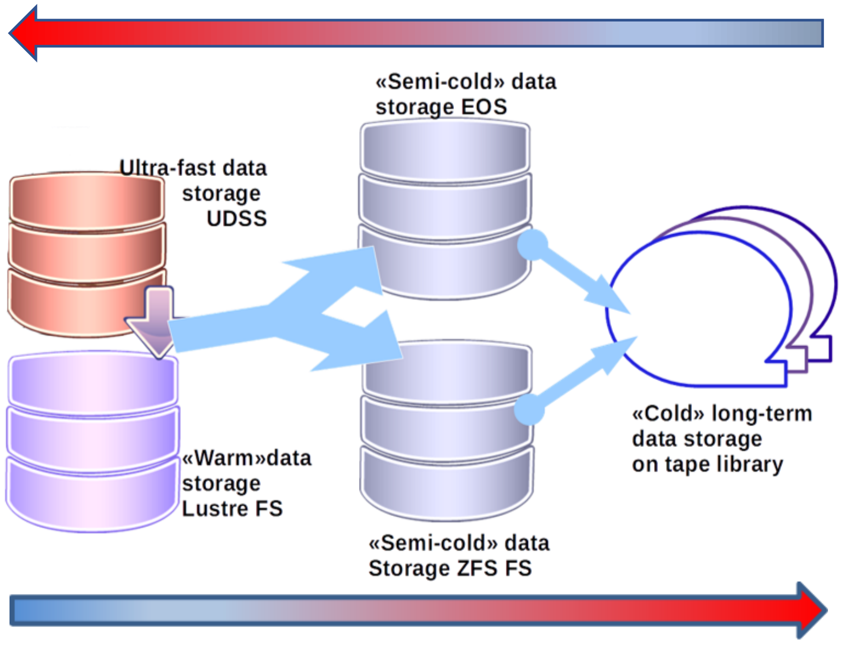


hlit.jinr.ru

TECHNOLOGIES IN THE NICA MEGASIENCE PROJECT

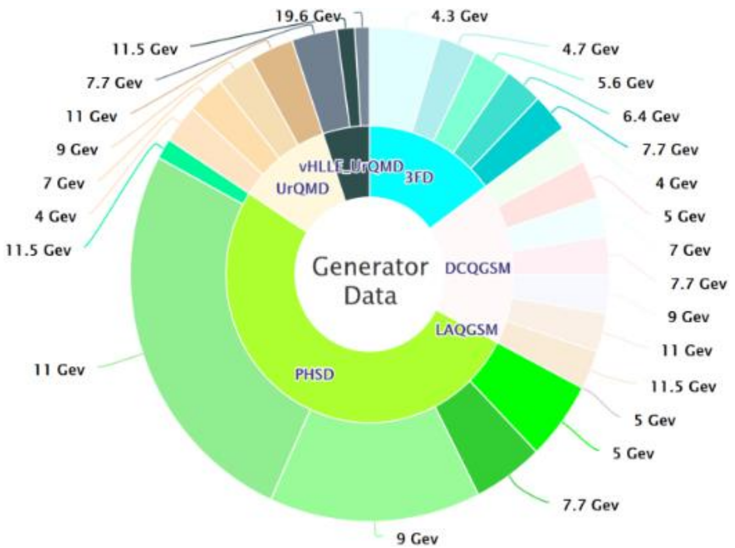


Velocity of data processing


«Semi-cold» data storage EOS
Ultra-fast data storage UDSS
«Warm»data storage Lustre FS
«Cold» long-term data storage on tape library
«Semi-cold» data Storage ZFS FS

Volume of data storage

Events of the MPD experiment are simulated and reconstructed on ultrafast data storage system under the **FS Lustre** management with a subsequent transfer to semi-cold storages (**FS ZFS, EOS**) and to the tape library for long-term storage.



About 50 million events were generated for the MPD experiment using the hierarchical structure of working with data. The unique composition of the “Govorun” supercomputer equipment, which includes a super-fast data access system and computing nodes with a large amount of RAM (3 TB per node), made it possible to process the same number of events on almost half the number of computing cores as on other available computing resources.

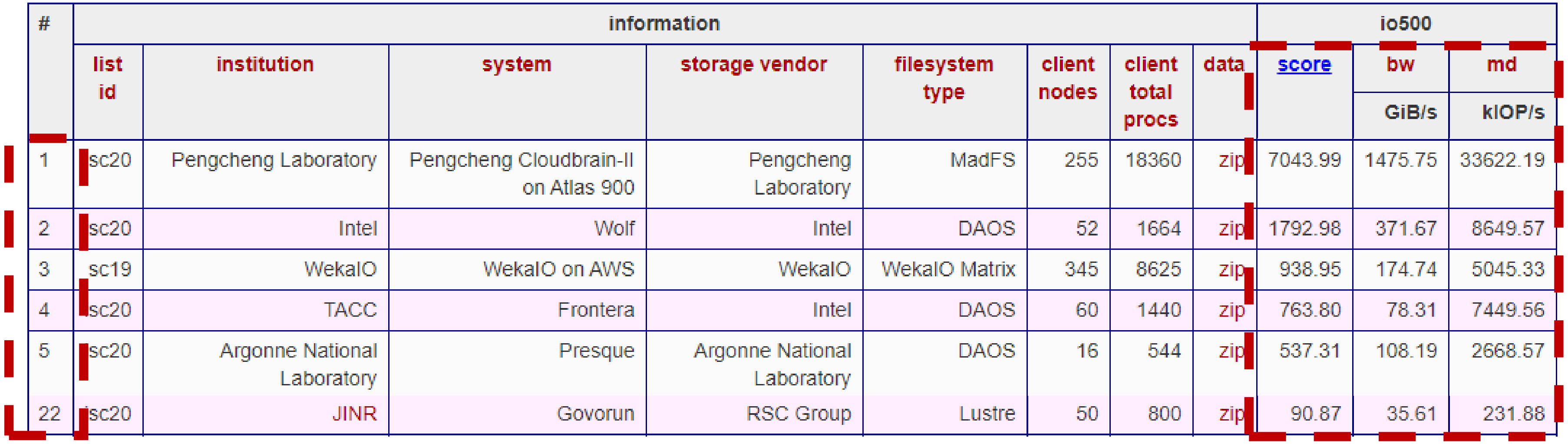
STORAGE SYSTEM - DAOS

Data flow from NICA experiment is expected from tens to hundreds of GB/sec and more with several PB for one experimental run. It looks very promising to use DAOS as a system for ultra-fast parallel access to complex hierarchical experiment data for:

- data collection
- data processing in real time
- data processing in off-line mode

The main advantages of DAOS for NICA are:

- The data access speed is comparable to the speed of server's RAM access
- Abstraction from the file storage system and data access process
- The ability to store additional metadata together with data for later easy storage, processing and subsequent data analysis
- Easy Object Storage extension during experiment

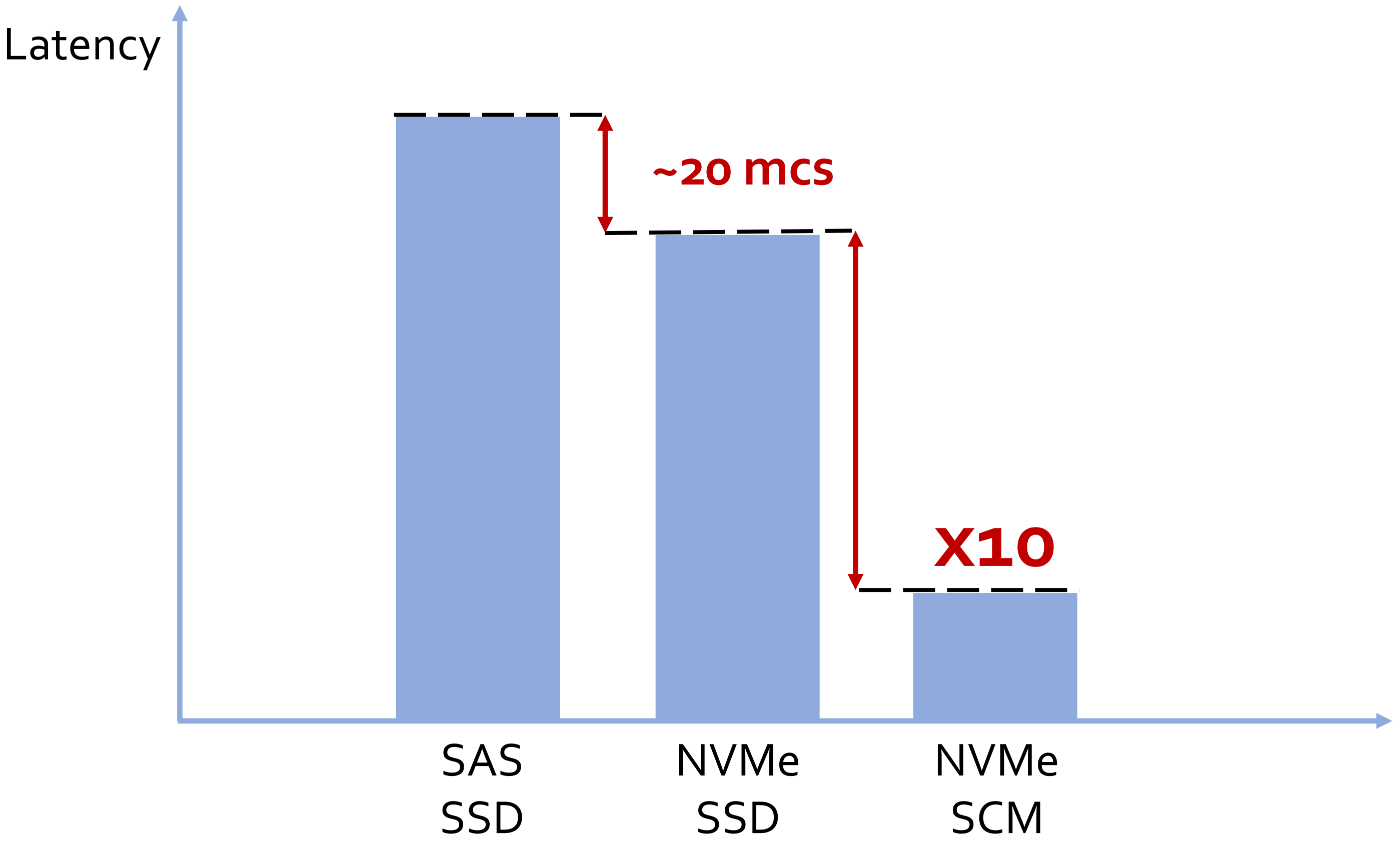


#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	sc20	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	255	18360	zip	7043.99	1475.75	33622.19
2	sc20	Intel	Wolf	Intel	DAOS	52	1664	zip	1792.98	371.67	8649.57
3	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
4	sc20	TACC	Frontera	Intel	DAOS	60	1440	zip	763.80	78.31	7449.56
5	sc20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	zip	537.31	108.19	2668.57
22	sc20	JINR	Govorun	RSC Group	Lustre	50	800	zip	90.87	35.61	231.88

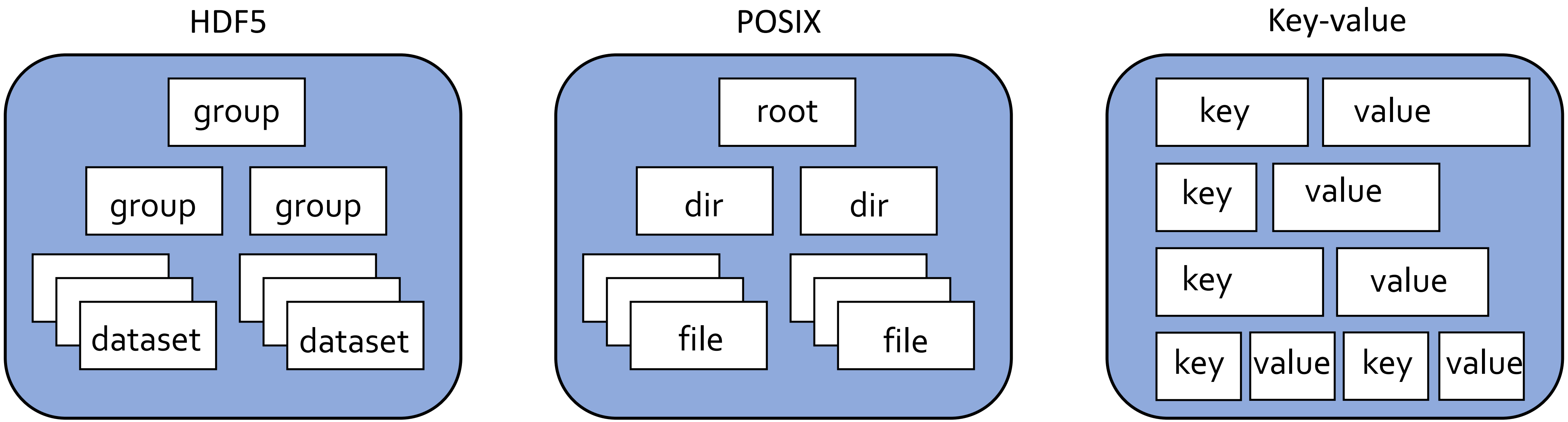
IO500 link

DATASTORAGE BY DAOS

Distributed Asynchronous Object Storage - designed for massively distributed Non Volatile Memory (NVM). DAOS takes advantage of next-generation NVM technology, like Storage Class Memory (SCM) and NVM express (NVMe)



DAOS MODES



HDF5

group

group group

dataset dataset

POSIX

root

dir dir

file file

Key-value

key value

key value

key value

key value key value

DAOS System - system of the DAOS servers

DAOS Target – virtual storage for data and metadate

DAOS Pool – pool of the virtual storages for data and metadata

DAOS Container - pool DAOS object for data management

DAOS Objects - container DAOS objects

DAOS REQUIREMENTS

HARDWARE

- Processors
  - Intel 64 bit
  - ARM 64 bit
- Network
  - Ethernet
  - InfiniBand
  - Intel OPA
- Storage
  - SSD NVMe
  - Optane DC Persistent Memory

SOFTWARE

- Compilers
  - C99, Go
- Build tool
  - scons
- Channel for administration
  - gRPC
- Persistent memory programming
  - PMDK
- NVMe device access
  - SPDK
- Discovering devices
  - hwloc
- Detecting fabric interfaces
  - libfabric

HYBRILIT SPECS

8 servers for storage with

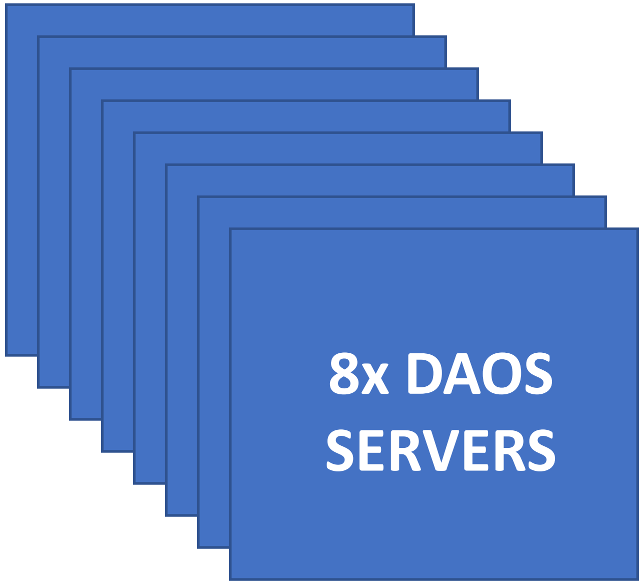
2x Intel Xeon Platinum 8268 24 cores

4x Intel Optane 512 Gb

2x Intel NVMe 2 Tb

RAM 192 Gb

Intel OPA 100 Gbit/s



BUILDING DAOS

Create Node-pool

Create data storage object with DAOS type

Create cluster. When creating, you must specify a group of nodes and a network provider

Create DAOS server with local storage, when creating, you must specify:

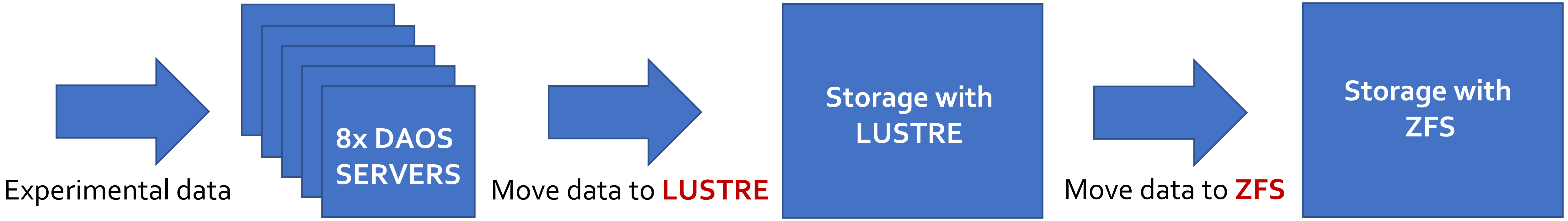
- number of targets per disk (2 targets per 1 disk for best performance)
- local or remote disks
- disc groups
- RAM size (if use RAM)

Run distribute resources

Run DAOS instances

HOW TO

1. Each server has **4x 256** Gb PMEM memory and receive **9-12** Gbyte/s (~100 Gbit/s) bandwidth
2. Transfer data from **DAOS** to **LUSTRE** servers and convert data from **Key-Value** to **POSIX** format
3. Transfer data from **Lustre** to **ZFS**



Experimental data

8x DAOS SERVERS

Move data to **LUSTRE**

Storage with LUSTRE

Move data to **ZFS**

Storage with ZFS

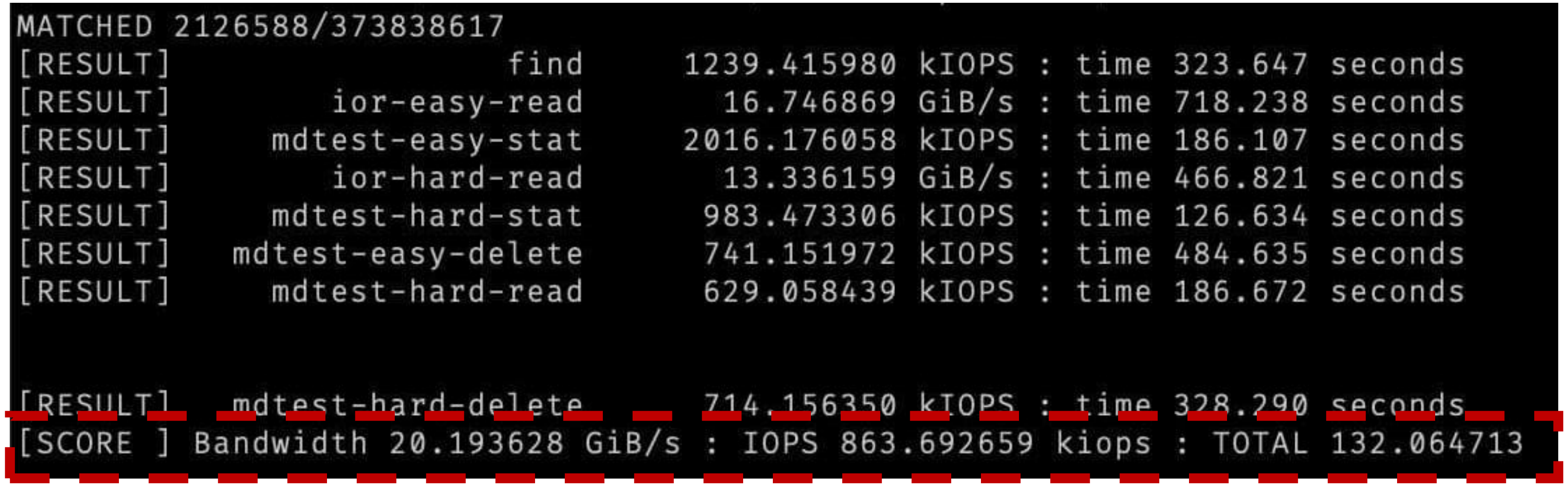
TESTS

4X servers

Bandwidth

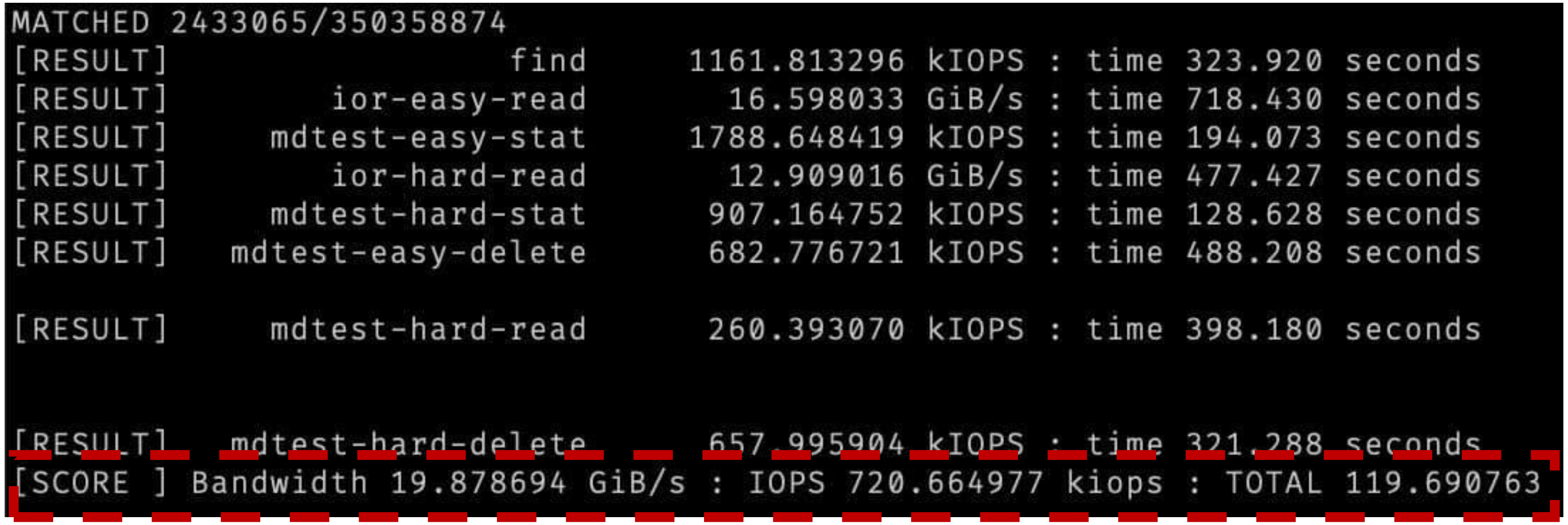
~20 Gb/s

Faster then our **LUSTRE**



```
MATCHED 2126588/373838617
[RESULT]                find     1239.415980 kIOPS : time 323.647 seconds
[RESULT]       ior-easy-read       16.746869 GiB/s : time 718.238 seconds
[RESULT]    mdtest-easy-stat     2016.176058 kIOPS : time 186.107 seconds
[RESULT]       ior-hard-read       13.336159 GiB/s : time 466.821 seconds
[RESULT]    mdtest-hard-stat      983.473306 kIOPS : time 126.634 seconds
[RESULT]  mdtest-easy-delete      741.151972 kIOPS : time 484.635 seconds
[RESULT]    mdtest-hard-read      629.058439 kIOPS : time 186.672 seconds


[RESULT]  mdtest-hard-delete      714.156350 kIOPS : time 328.290 seconds
[SCORE ] Bandwidth 20.193628 GiB/s : IOPS 863.692659 kiops : TOTAL 132.064713
```



```
MATCHED 2433065/350358874
[RESULT]                find     1161.813296 kIOPS : time 323.920 seconds
[RESULT]       ior-easy-read       16.598033 GiB/s : time 718.430 seconds
[RESULT]    mdtest-easy-stat     1788.648419 kIOPS : time 194.073 seconds
[RESULT]       ior-hard-read       12.909016 GiB/s : time 477.427 seconds
[RESULT]    mdtest-hard-stat      907.164752 kIOPS : time 128.628 seconds
[RESULT]  mdtest-easy-delete      682.776721 kIOPS : time 488.208 seconds

[RESULT]    mdtest-hard-read      260.393070 kIOPS : time 398.180 seconds


[RESULT]  mdtest-hard-delete      657.995904 kIOPS : time 321.288 seconds
[SCORE ] Bandwidth 19.878694 GiB/s : IOPS 720.664977 kiops : TOTAL 119.690763
```

10X servers

Bandwidth

~20 Gb/s

Faster then **MSC Russian Academy of Sciences**

THANK YOU FOR YOUR ATTENTION!