Hyperconverged storage for high performance data analysis in high energy physics: a case of Intel DAOS deployment.

A.A. Moskovsky (RSC Group), A.T. Brekhov (RSC Group), D.V. Podgainy (JINR LIT), A.O. Kudryavtsev (Intel Corp.)

Joint Institute of Nuclear Research (JINR) is a word leading multinational research organization in high energy physics, it's a site for a novel experiment NICA (Nuclotron-based Ion Collider fAcility). The NICA experiment goal is to explore previously unknown properties of quark-gluon plasma, that should be formed after heavy atomic ion nuclei collisions. Data collection for NICA is due to start in 2022, but even preparation phase requires multi-petabyte storage with capability to rapidly process hundred-terabyte datasets for collision events simulation and reconstruction. Data flow during experiment is expected from tens to hundreds of GB/sec and more with several PB for one experimental run. Meshcheryakov Laboratory of Information Technology (MLIT) at JINR should provide IT support for NICA, i.e. MLIT should provide data collection, data processing in real time and data processing in off-line mode.

Data processing workflows at JINR are different in demands for storage access bandwidth (sequential read and/or write speed) and number of short operations on data or meta-data (IOPS). While some are mostly bandwidth-limited, a large and important fraction is mostly IOPS-limited. Additionally, MLIT should enable simulation that requires high-performance compute resources and vast amounts of main memory (tens of gigabytes per CPU core). Such combination of workloads makes fusion of compute and storage elements to be the most appealing option, with storage-class memory (SCM) devices to be the media for dual purpose: meta-data storage and RAM extension.

To address the issues described above, MLIT JINR deployed Govorun system, comprised of Intel Xeon Scalable 2nd generation dual-socket nodes equipped with a mixture of storage devices. With the help of NVMe-over-fabric technology, system administrator can create on-demand storage volumes out of SSDs physically installed on compute nodes (up to 2 2TB M.2 devices per node), specialized storage nodes (up to 12 M.2 2TB or 375 GB Intel Optane devices) or SCM devices on persistent memory nodes. Filesystem options include Lustre, ZFS, NFS and others.

Recently, experiments with Intel's Distributed Asynchronous Object Storage (DAOS) demonstrated substantial performance increase over alternative storage. IO500 benchmark results on meta-data performance grew substantially from Lustre 50-clients run to Intel DAOS 10-client run (see table below), with more profound advantage for more irregular ("hard") operations. In the both cases, NVM-over-fabric was used for device pooling and client-server communication, the fabric was non-blocking fat-tree topology Intel OmniPath interconnect at 100 Bit/s speed. Operating system was CentOS 7.7 build 1908.

Table 1 Lustre vs Intel DAOS metadata benchmark on Govorun cluster,

Parameters	Govorun in IO500 ISC20 submission, https://io500.org/submissions/view/55	Govorun IO500 ISC21, https://io500.org/submissions/view/536
Filesystem (storage)	**Lustre 2.13**	**Intel DAOS v1.2**
IO500 Meta-data Easy Write	492.38	1,102.72

IO500 Hard delete	67.45	714.16
Clients	50 nodes, each dual-socket Intel Xeon Scalable Gold 2nd gen 8268, 192 GB RAM,1x 100 Gbit/s adapter	10 nodes, each dual-socket Intel Xeon Scalable 1st gen 6154 192 GB RAM, 2x 100 Gbit/s
Servers	12 servers, each dual-socket Intel Xeon Scalable 1st gen 6154 192 GB RAM, 750 GB Intel Optane SSD, 20 TB NVMe SSD 2x 100 Gbit/s + 2 servers, each dual-socket Intel Xeon Scalable 1st gen 6154 192 GB RAM, 24T TB NVMe SSD? 2x 100 Gbit/s	8 nodes, each dual-socket Intel Xeon Scalable Gold 2nd gen 8268, 192 GB RAM, 0.5 TB Pmem Optane, 1x100 Gbit/s adapter