COMPOSABLE DISAGGREGATED ENVIRONMENTS FOR HPC WORKLOADS

A. Moskovsky, P. Lavrenko[a]

RSC Group, Moscow, Russia

E-mail: [a] p.lavrenko@rsc-tech.ru

The disaggregation of storage elements is just one of the first steps to changing the concepts of data handling. Until now, it was hard to manage methods of data handling, storage and processing on a computing cluster. Other aspects, except for storage volume restrictions (quotas) and computing time limits, remained and still largely remain hidden for the data center. What data access models are used? What will be the load on storage resources? How energy-efficient will a specific processing scenario be? How can we quickly and efficiently transfer the task with all its required data to a more advantageous (higher performance/lower cost) site? The industry thinks that soon we will need the significant redistribution of responsibility (separation of concerns) in data processing methods. Which alternative scenario of responsibility areas meets the current requirements?

Keywords: disaggregation, HPC workload, supercomputer “Govorun”

1. Introduction

The landscape of modern computing infrastructures constantly changes. Having been developing HPC solutions for more than 10 years, we see rapid changes in architectural approaches to data processing, from computing modules, storage systems and networks to new processing algorithms, industries and data sources.

According to many industry experts, we are going through a highly important historical period similar to the Cambrian explosion over 540 million years ago, which resulted in the rapid expansion of complex organisms on the Earth and in the current level of biodiversity. The world of general-purpose computers based on the CPU architecture and its capabilities is gradually fading away, leaving the door open for a multitude of special-purpose computing elements, such as GPU, FPGA, TPU, ASIC, Neural Engine, etc. To deal with the reasons for this transformation, we need to understand the primary source defining the transformation trends.

Is the ability to develop specific computing elements cheaper and faster the main driving force? Yes, but only partially. An active increase in the free oxygen content in the atmosphere during the Cambrian period provided extra energy for the development of creatures, and similarly a massive increase in performance and the emergence of new neural morphing, quantum, tensor, optical and other computing elements kick-started a rapid evolution of data handling principles.

The active development of super-heterogeneous computing scenarios (workflows) is one of the main trends and the subject of heated debates at major HPC industry events and conferences throughout the world. These workflows are based on one scenario of data processing with the addition of different computing elements in the required sequences. These scenarios must flexibly adapt to processed data, resulting in the emergence of Data-Driven Workflows.

The shift is easily explainable by the dramatic growth in the data volume and data types, as more and more information about the surrounding world becomes digitized. We are witnessing an industry transition from the Compute-Centric paradigm to the Data-Centric paradigm (Fig. 2).

As a result, we are moving away from the universal computing paradigm to the development of hybrid computing systems.

2. New paradigms

At present, we see many new architectures of computing elements, data storage and transfer elements. Moreover, the boundaries between the computing unit and the data storage or transfer device are now blurred.

Although we are still far from truly revolutionary methods, such as quantum computing, we already need to choose new methods of arranging semiconductor elements for the efficient solution of new computing tasks.

One of the key models of future computing platforms is the paradigm of Composable Disaggregated Infrastructure (CDI).

Composability and disaggregation are two different concepts that enhance each other, and their combination defines the principles of infrastructure systems based on high-speed low-latency data transfer networks that combine computing and storage components into dynamic resource pools available on demand.

Disaggregation is based on hardware solutions, and composability describes APIs for the infrastructure management of such objects. IDC, a global IT market research company, forecasts the CDI market growth from $700M in 2017 to $3.4B in 2022, which will take no more than 5% of the server market in 2023, so there will still be tremendous potential for further intensive growth in the coming years (Worldwide Composable/Disaggregated Infrastructure Forecast, 2018–2023, IDC, Aug 2018).

The Rack Scale Architecture is one of the first architectures to embrace the CDI concept and go beyond single-server platform limitations thus increasing component integration. This architecture is used by Intel, Facebook, Google, Yandex and many others.

3. RSC's strategy

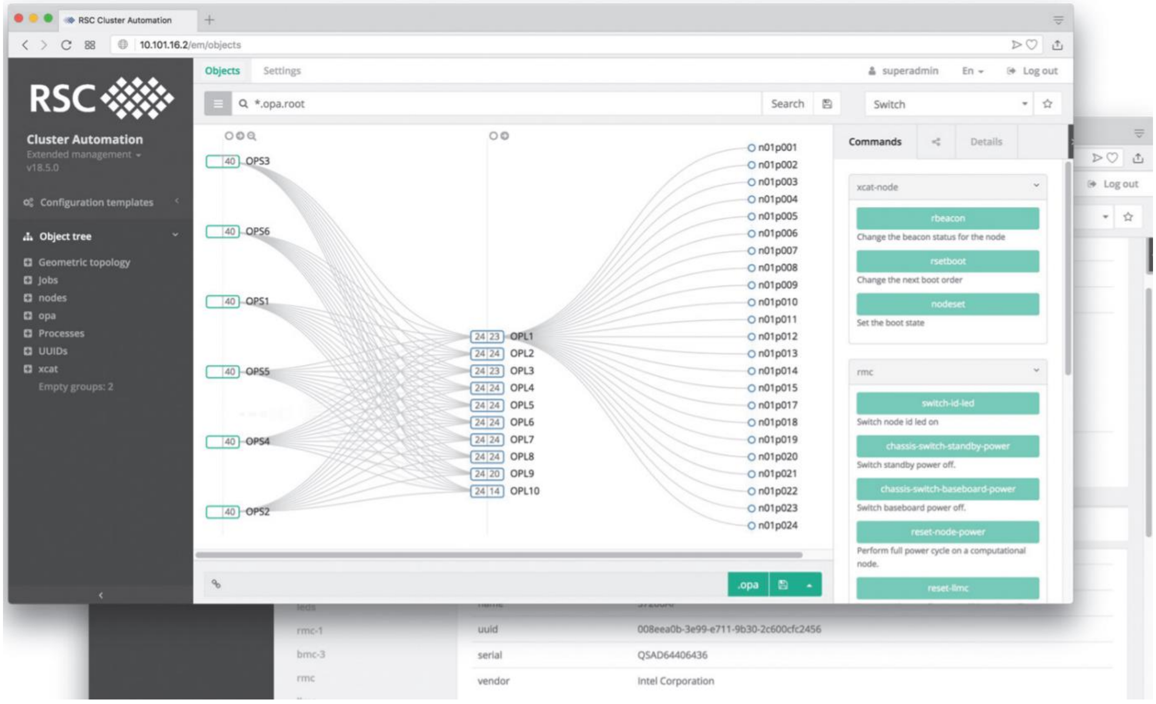
Since 2014, RSC has focused on improving the efficiency of developed computing systems in general, starting from the cooling and power infrastructure to managing computing elements, managing computing elements, storage systems and user application workloads. It became evident that optimization was impossible without a control system able to adapt to the current and future diversity of components. The Rack Scale Design initiative by Intel strengthened our belief that the focus was right.

Our approach resulted in the creation of an RSC BasIS software platform (Fig. 1-2), a microservice software environment based on the CNCF (Cloud Native Compute Foundation, cncf.io) infrastructure to manage all data center levels.

RSC BasIS provides full access and unified control methods for computing clusters, while preserving the values of an open and microservice structure making it extremely expandable to describe new system configurations.

It was developed based on knowledge and the automation of the HPC computing complex.

RSC BasIS made it possible to quickly adapt to new CDI challenges and concentrate on the Data-Driven Workflow methodology using the dynamic configuration of multiple data movement and processing pipelines[1].



Change the beacon status for the node

Change the next boot order

Set the boot state

Switch node id led on

Switch standby power off.

Switch baseboard power off.

Perform full power cycle on a computational node.

uuid 008eea0b-3e99-e711-9b30-2c600cfc2456
serial QSAD64406436
vendor Intel Corporation

Fig.1. Storage devices for hyperconverged scenarios are selected on the basis of connection topologies

[1] A pipeline is a more specific linear process of data movement and computing, several such processes can run simultaneously (e.g. building and configuring a storage system on demand for a specific computing task or applying object selection filters to video streams using DL methods).

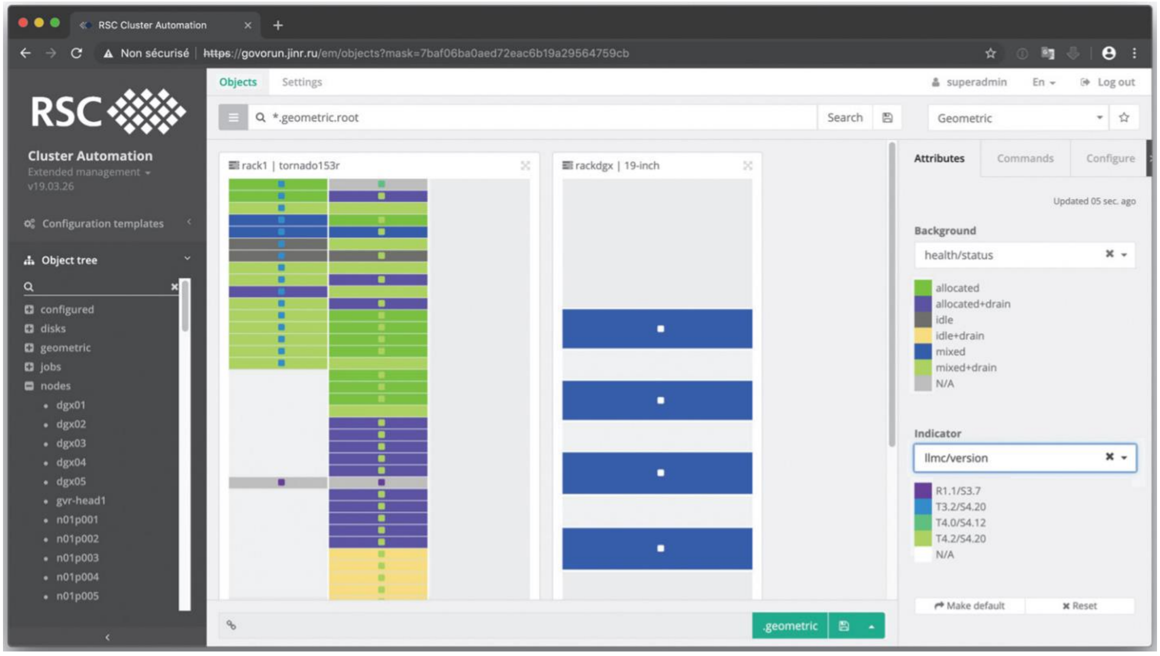


Fig.2. Data center drill-down tools enable “descending” troubleshooting methods

What qualities of such a system are the most important? We think that a reconfigurable platform for the creation and dynamic support of multiple pipelines must enable problem-oriented configurations, i.e. configurations intended for the optimal solution of a specific narrow class of tasks. The system must adapt to changes in user demand for resources and change its configuration, i.e. it must contain automated reconfiguration mechanisms.

4. Computing and storage convergence

Until now, we were talking about the architecture and open management protocols. But what kind of computing systems does RSC offer? What is the hardware platform based on?

The main RSC product is the high-density Tornado computing platform. It has been developed as a power-dense platform for data center systems (the portfolio of liquid-cooled computing racks based on the x86 and Elbrus architectures contains solutions with 100-400 kW power capacity per rack), advanced dynamic configuration and management features.

RSC’s knowledge of the advantages of the highly optimized solution composing architecture led to the launch of a new product class marketed under the same RSC Tornado trademark. The RSC hyperconverged platform is a confident step to a full-scale CDI platform (and despite significant progress in the disaggregation of hardware components, it is still impossible to achieve a full-scale implementation of CDI) enabling the creation of on-demand HPC systems.

The RSC hyperconverged solution is based on a unified platform with unique characteristics, such as 100% liquid cooling and extremely high computing density, i.e. up to 153 dual-socket (x86) or quad-socket (Elbrus) servers in a cabinet linked by a network based on the Intel Omni-Path or Mellanox Infiniband interconnect at speeds up to 100 Gb/s (Fig. 3).

The platform supports the installation of servers with fast processors and NVMe-based SSD drives. The NVMoF (NVMe over Fabric) protocol and platform management features in RSC BasIS give it the ability to dynamically select, aggregate and configure drives on demand over the entire computing framework, which is an implementation of CDI principles. Expanding on the CDI infrastructure, RSC group also provides computing node options with 12 hot-swappable M2 drives and multiple 100 Gb/s ports.

Fig. 3. RSC hyperconverged platform for HPC

On the basis of such a node, which is actually a POD container (Pool of Devices in the CDI paradigm), the system can both configure a node with up to 4TB available RAM (with the Intel Memory Drive Technology and Intel® Optane™ drives in the memory expansion mode) and provide internal storage resources for a dynamically distributed file system.

Surely, the RSC BasIS software platform is the key element for further orchestration of the system behavior.

What makes RSC solutions different? We plan to develop configurations for dynamic software-defined scenarios focused on user data processing workloads. It means that the system configuration logic is not defined by the administrator during the installation, but can dynamically change throughout the solution life cycle enabling deep integration with specific user workloads.

One of the important advantages of software-defined solutions is that they remove existing restrictions on storage element proportions in computing systems.

In Hardware-Centric systems, this proportion is based on the "optimal level", which primarily depends on the cost of technologies for each layer. The CDI paradigm enables a detailed identification of requirements of each user workload with the independent configuration and expansion of the required storage types throughout the entire system life cycle. With this approach, each layer is considered as a separate resource type that can be selected by the user (Fig. 4).

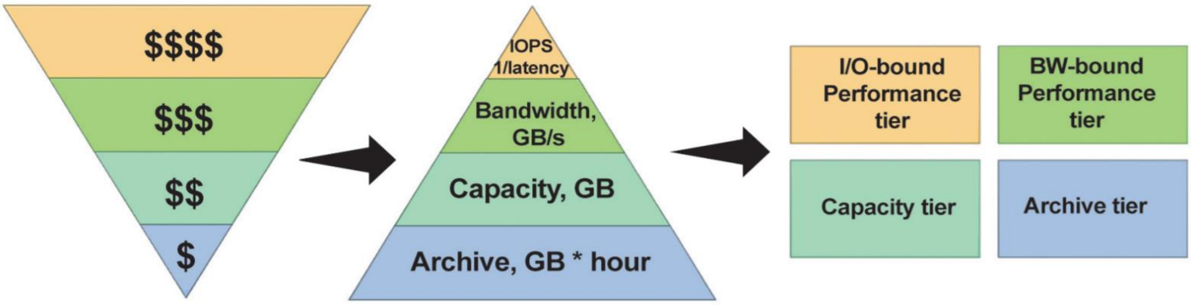


Fig.4. CDI paradigm, each layer can be considered as a separate resource type selected by the user

5. Capacity vs performance. Multi-layered storage systems

This is a classical scenario that is used in most shared data centers.

- the user accesses the entry node via ssh;
- the user loads data for processing;
- if necessary, the user assembles the design application on the compilation node optimizing it for the computing node architecture;
- the prepared task is placed in a queue with the specification of computing resource requirements;
- the task scheduler integrated into the software stack waits until the required resources are available;
- when the required resources are available, the task scheduler allocates them for exclusive access to the user and transfers the execution and control to the Pipeline entry point of the computing task;
- the computing nodes launch the application following the launch scenario;
- the running application performs read and write operations in a parallel file system accessible to all cluster nodes;
- the user task is completed, the results remain stored in the parallel storage system.

However, it has a number of assumptions regarding the storage system efficiency:

- its performance should be sufficient for the efficient task execution on the entire cluster simultaneously;
- this file system should be optimized as a hardware and software stack for some “average” workload requirements;
- with an increase in the number of simultaneously used nodes, the system should primarily support an increase in the number of simultaneous I/O operations, especially for MPI workloads with barrier synchronization, i.e. with time synchronization of group reading and group writing operations in the file system;
- expanding the cluster capacity, its owner should maintain the linear file system performance growth, otherwise increasing the computing capacity will become a waste of resources.

Storage system costs for a computing cluster can be very significant (up to 50%). One of the optimization methods is the Burst Buffer method using a relatively small (less than 100% of the storage volume) high-performance storage layer enabling the workload to process large sequences of I/O operations, which are later synchronized with a slower system.

The Burst Buffer approach has natural limitations, and here the industry starts moving from the Hardware-Centric paradigm (large parallel file system) to the paradigm of Data-Driven Workflows.

In this paradigm, the Burst Buffer as a file system accelerator and the file system itself are divided into:

- volume storage layer (Capacity Tier), which has a high volume, but does not meet the peak workload requirements;
- high-performance layer (Performance Tier), fast and supporting high workloads (Fig. 5).

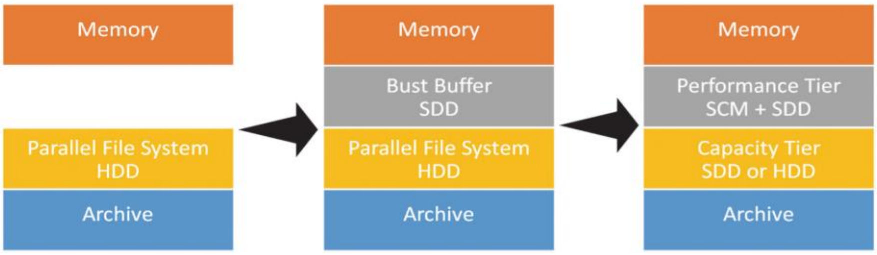


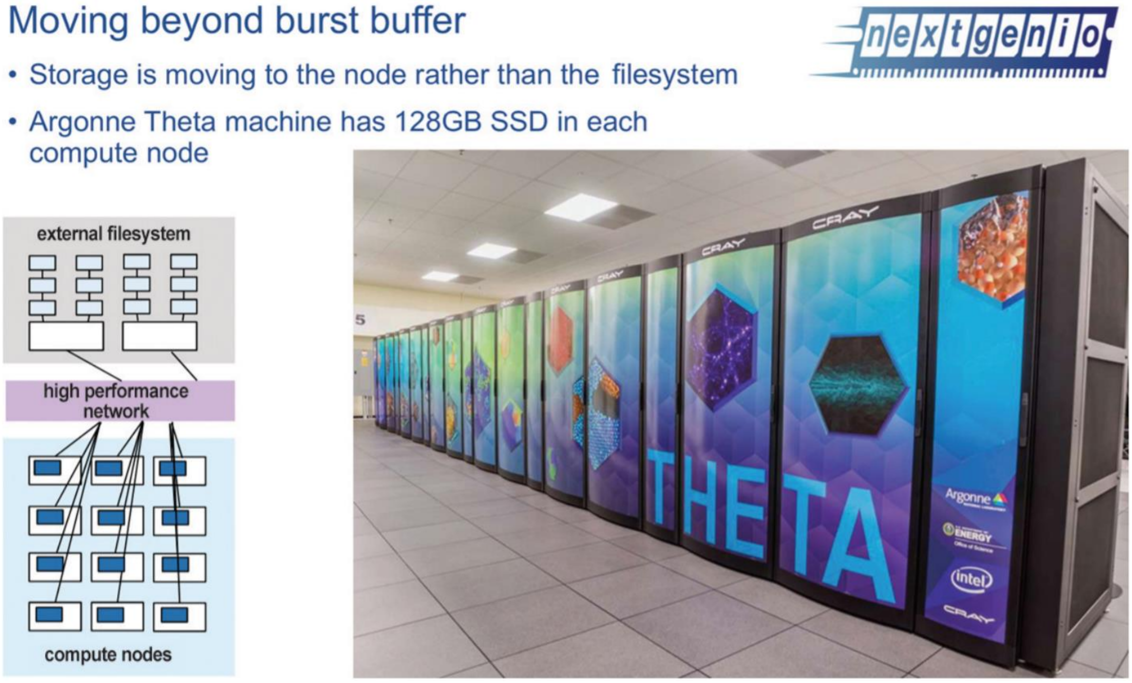
Fig.5. Moving from storage acceleration with the Burst Buffer to Performance and Capacity concepts

What is the key difference between the Performance Tier and the Burst Buffer? It can be configured and run on the fly during the workload launch and in accordance with its requirements, and stops when the workload is finished and data is moved to the Capacity Tier. In other words, the lifetime of the Performance Tier is synchronized with the logic of a specific pipeline user.

The Capacity Tier can have a longer lifetime. One of the main prerequisites for such a transformation was the ability to move storage devices inside computing servers. It linked the computing performance improvement with the capacity and performance of the cluster storage system and allowed the manufacturer to create highly integrated optimized solutions based on the CDI paradigm (Fig. 6).

Moreover, disk pools can be used in CDI not only for the Performance Tier, but also for:

- temporary increase in the RAM volume on the node;
- creating object storage systems between computing nodes in addition to file storage systems;
- use of disks for stateful Checkpoint/Restart operations;
- joint use with specialized I/O libraries for data processing
- and other scenarios.



Moving beyond burst buffer

- Storage is moving to the node rather than the filesystem
- Argonne Theta machine has 128GB SSD in each compute node

Fig.6. Moving towards CDI: transfer of storage devices to computing nodes using the example of the Theta system at Argonne National Laboratory, USA.

6. Usage experience at the Joint Institute for Nuclear Research

In 2018, RSC Group and the Joint Institute for Nuclear Research (JINR) launched the “Govorun” computing cluster in Dubna to cardinally accelerate complex theoretical and experimental research in the field of high energy physics, nuclear physics and condensed matter physics, as well as to implement the NICA (Nuclotron based Ion Collider fAcility) megaproject, i.e. a new JINR particle accelerator complex for studying the properties of dense baryonic matter. After the launch of the NICA collider, JINR scientists will be able to recreate in laboratory conditions a special state of matter in which our Universe was in the first moments after the Big Bang, i.e. quark-gluon plasma (QGP).

The RSC Tornado platform was chosen for this project. Due to the specifics of research underway at JINR, it was decided to use the hyperconverged approach with support of the Lustre parallel file system on demand, and the Performance Tier was selected for experimental data processing scenarios.

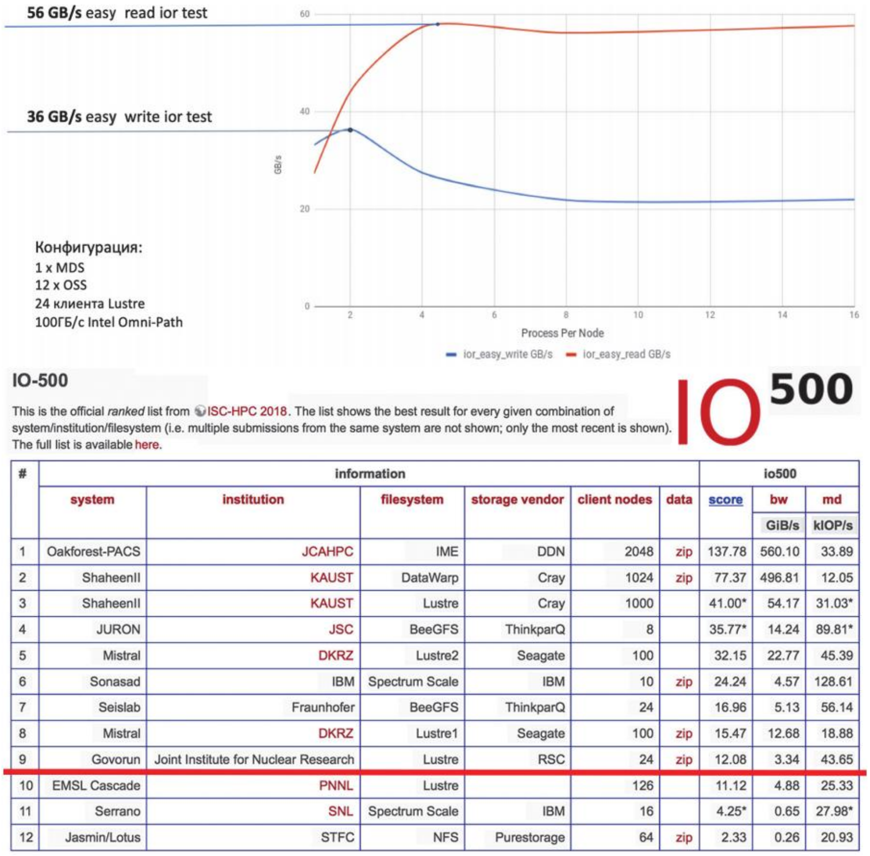
Each computing node based on two Intel® Xeon® processors has two NVMe solid-state drives in the high-density M.2 format. All computing nodes of the cluster are connected by the high-speed low-latency Intel Omni-Path interconnect with 100 Gb/s access speed, which forms the core of the hyperconverged approach. In the course of work, scientists’ computing workloads got access to the Lustre storage space through the task queue management system.

Data storage fault tolerance and reliability were ensured by the use of RAID technologies for hardware-level redundancy and Active/Passive schemes for software storage services. Due to the common NVMoF access scheme and the RSC BasIS software stack, the architecture, technology and redundancy level on each layer (disk, network, service, node) may vary for each launch scenario enabling the flexible setup of the redundancy-storage volume-performance ratio.

The efficiency of the system in operation was tested with the IO500 benchmark, i.e. a de facto standard for the performance assessment of storage systems for HPC.

In the course of the experiment, a standard task manager system was used to request 37 regular computing nodes, 1 of which acted as a metadata server (MDS) and 12 as storage servers (OSS). The remaining 24 nodes provided their storage devices for the Performance Tier and launched the IO500 test, performing intensive I/O operations on the Lustre file system in different modes.

The testing showed high performance values (up to 56 Gb/s for read and 36 Gb/s for write) and allowed the system to take the 9th place in the IO500 list in 2018 (Fig. 7) without any dedicated storage facility and consequently with greatly reduced costs. It proved that the approach was selected correctly and that the modern component base was ready for deeper support of the CDI paradigm.



IO-500

This is the official *ranked* list from ISC-HPC 2018. The list shows the best result for every given combination of system/institution/filesystem (i.e. multiple submissions from the same system are not shown; only the most recent is shown). The full list is available here.

#	information						io500		
	system	institution	filesystem	storage vendor	client nodes	data	score	bw	md
								GiB/s	kIOP/s
1	Oakforest-PACS	JCAHPC	IME	DDN	2048	zip	137.78	560.10	33.89
2	ShaheenII	KAUST	DataWarp	Cray	1024	zip	77.37	496.81	12.05
3	ShaheenII	KAUST	Lustre	Cray	1000		41.00*	54.17	31.03*
4	JURON	JSC	BeeGFS	ThinkparQ	8		35.77*	14.24	89.81*
5	Mistral	DKRZ	Lustre2	Seagate	100		32.15	22.77	45.39
6	Sonasad	IBM	Spectrum Scale	IBM	10	zip	24.24	4.57	128.61
7	Seislab	Fraunhofer	BeeGFS	ThinkparQ	24		16.96	5.13	56.14
8	Mistral	DKRZ	Lustre1	Seagate	100	zip	15.47	12.68	18.88
9	Govorun	Joint Institute for Nuclear Research	Lustre	RSC	24	zip	12.08	3.34	43.65
10	EMSL Cascade	PNNL	Lustre		126		11.12	4.88	25.33
11	Serrano	SNL	Spectrum Scale	IBM	16		4.25*	0.65	27.98*
12	Jasmin/Lotus	STFC	NFS	Purestorage	64	zip	2.33	0.26	20.93

Fig.7. Global performance benchmark of deployed storage systems for HPC

7. Conclusion

In the above JINR scenario, we still focused on providing low-level POSIX-compatible file systems for working with data. Although classic file systems with a unified storage space are clear to use, they are becoming a restrictive factor. Our conclusions are proved by the successful development of such projects as Intel DAOS, i.e. a revolutionary storage stack based on the NVMe and Storage Class Memory technologies.

Users should be able to concentrate on the development and creation of data handling models without being restricted by the data center. The data center should properly evaluate the task and plan for optimal task execution, and its control systems should support the launch and provide the required data movement.

Data processing specialists, end users, should provide:

- detailed scenario (workflow) of data processing independent from a specific computing cluster architecture – load path, processing, result generation and analysis;

- data processing plan for a life cycle, data types (processing logs, resulting data, analytics), including exchange formats and requirements for archive storage;
- directly pipelined applications for data processing;
- time and financial restrictions.

However, they should not specify when and where to store and process data.

Based on these components, the data center should:

- identify the best execution method creating an execution pipeline;
- simulate the pipeline execution to evaluate its convergence, power and money costs.

Data center management subsystems should use all the provided information to orchestrate I/O operations and define where to store and place data during computing, i.e. using the previously described Capacity and Performance Tiers directly during processing.

New computing components, the entire multitude of previously mentioned architectures and solutions, are a necessary catalyst for changing the data approach.

Today we are challenged to develop composing methods, and CDI enables us to respond to this challenge. However, to achieve a “bright future” in which the potential of new solutions is used to the full extent, the industry still has to formulate and standardize new interfaces for interaction between users and data centers. RSC is working on this in its products.