Using the GOVORUN supercomputer for the NICA megaproject

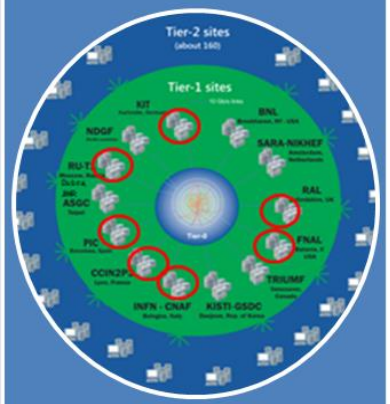
D.V. Podgainy

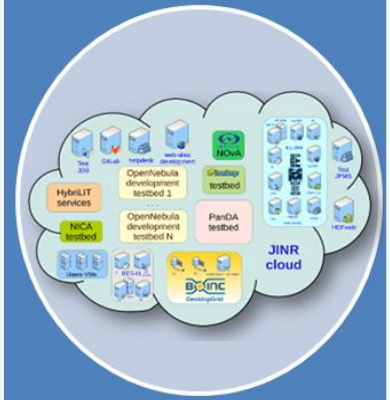
on behalf of HybriLIT team

Laboratory of Information Technologies

XXVII International Symposium on Nuclear Electronics and Computing
Montenegro, Budva, Becici
30 September – 4 October, 2019

Multifunctional Information and Computing Complex
Main components





Grid-Tier1:
10688 cores
8.3 PB disk
11 PB tape

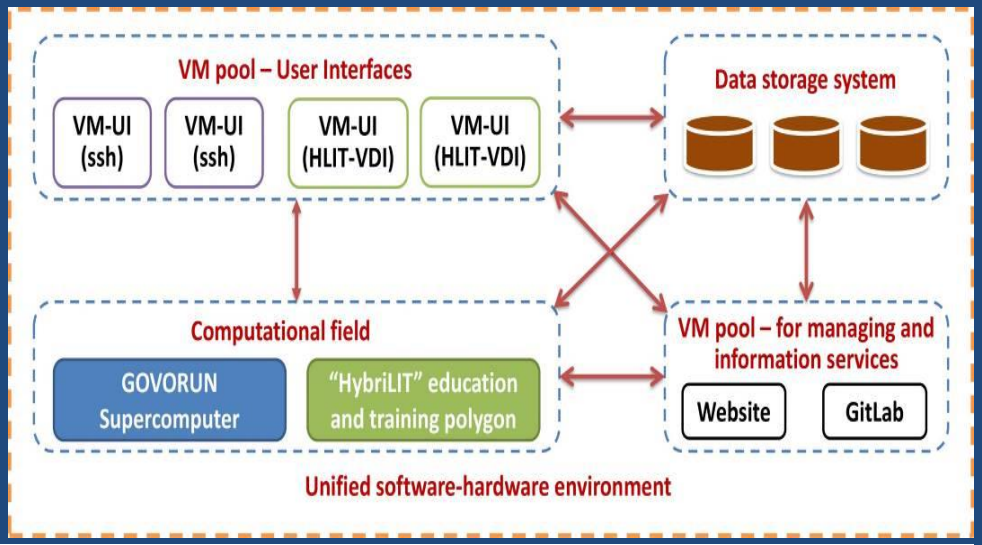
Grid-Tier2
CICC:
4128 cores
2.7 PB disk

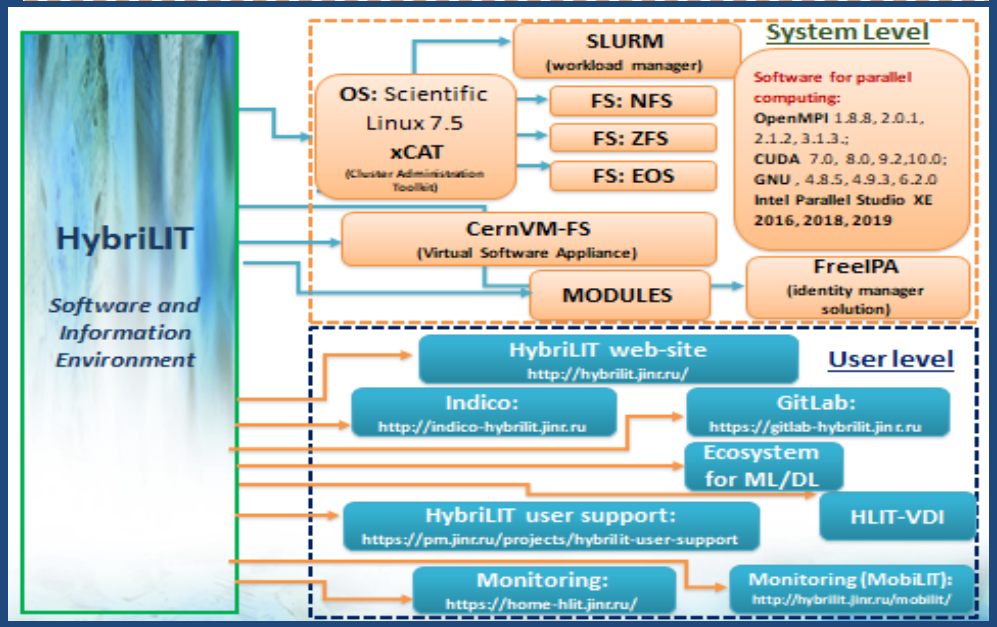
Cloud:
1572 CPU
8.142 TB RAM
1.1 PB disk

HPC Govorun
Peak ~0.5 Pflops
HybriLIT:
~70 Tflops

LIT IT-infrastructure is the one of JINR basic facilities

Software and information environment of the HybriLIT





The unified software and information environment of the HybriLIT platform allows users to use the education and testing polygon is aimed at exploring the possibilities of novel computing architectures, IT-solutions, to develop and debug their applications, furthermore, carry out calculations on the supercomputer, which allows them to effectively use the supercomputer resources. The unified software and information environment including the unified system level (the operation system, the job scheduler, file systems and software) as well as a set of services allowing users to quickly get the answers to their questions, jointly develop parallel applications, receive information about conferences, seminars and meetings dedicated to parallel programming technologies.

Presentation of Supercomputer Govorun

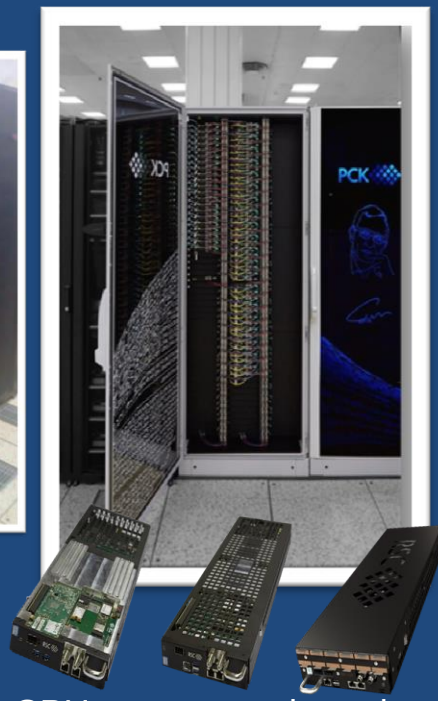
On **March 27, 2018**, a presentation of a new supercomputer named after Nikolai Nikolayevich Govorun, whose name is associated with the development of information technologies at JINR since 1966, took place in LIT in frames of a session of the Committee of Plenipotentiaries of the governments of the JINR Member States.

A seminar organized in frames of the presentation, gathered in the LIT conference-hall more than 200 guests from the different institutes and universities, employees of LIT and other JINR laboratories. The presentation received wide coverage in the Russian mass media (on TV, in print and online publications).

Supercomputer Govorun

100% liquid cooling in ’hot water’ mode engineering infrastructure for ultra high dense scalable and energy efficient cluster solution. Equipment racks, additional redundancy, automated monitoring and control system

GPU-component based on NVIDIA DGX-1 Volta.



CPU-component based on the newest Intel architectures (Intel Xeon Phi and Intel Skylake processors)

- Unique heterogeneous and hyper-converged system
- Multipurpose high-performance system with direct hot liquid cooling of all system components
- The most energy-efficient system in Russia (PUE = 1,03)
- First 100% hot liquid cooling of Intel® Omni-Path interconnect
- Record power density – up to 100 kW per 42U cabinet

NVIDIA DGX-1

The world's most powerful supercomputer for AI

8x Tesla V100 with NVLink interconnect

60 TFlops double precision

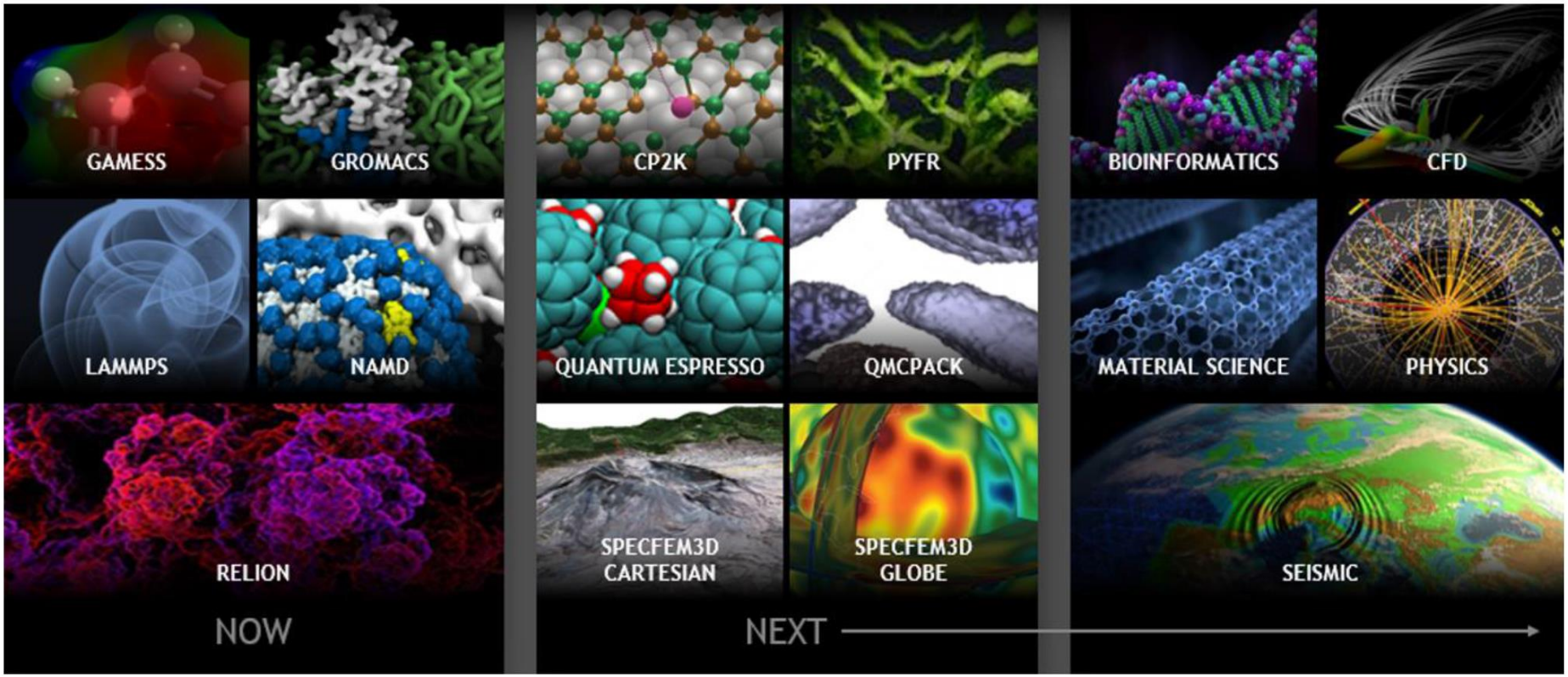
120 TFlops single precision

Unique energy efficiency 3.2 kW

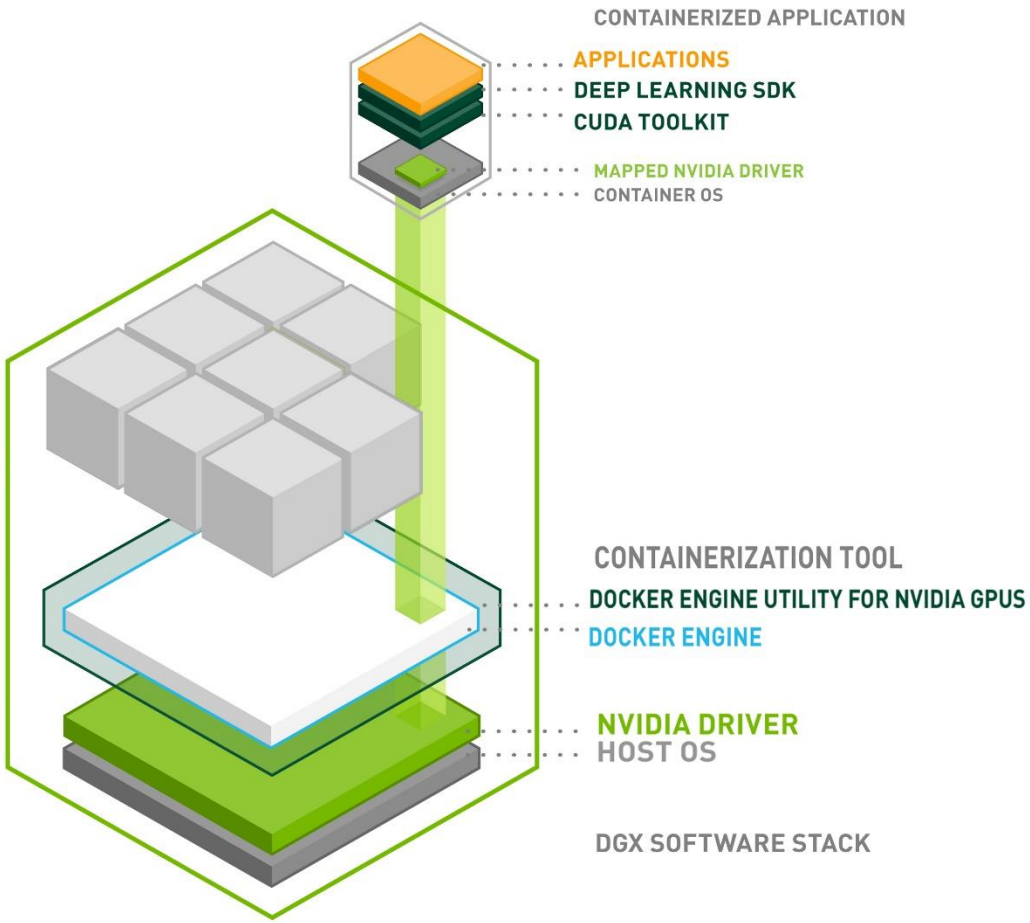
Full stack deep learning software preinstalled

Replaces 400 traditional dual CPU servers on DL applications

DGX-1: HPC APPS CONTAINERIZED



NVIDIA-dockers for DGX-1



Machine Learning frameworks optimized by NVIDIA

Caffe

TensorFlow

PYTORCH

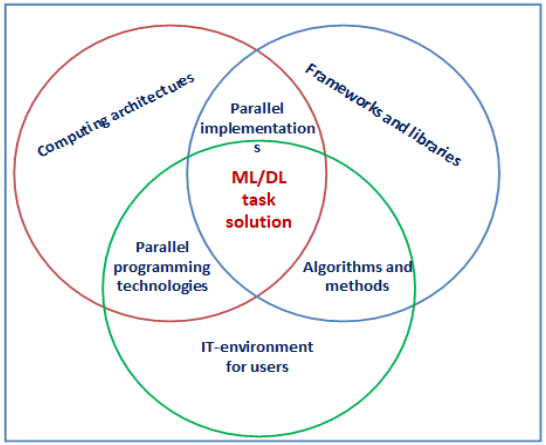
theano

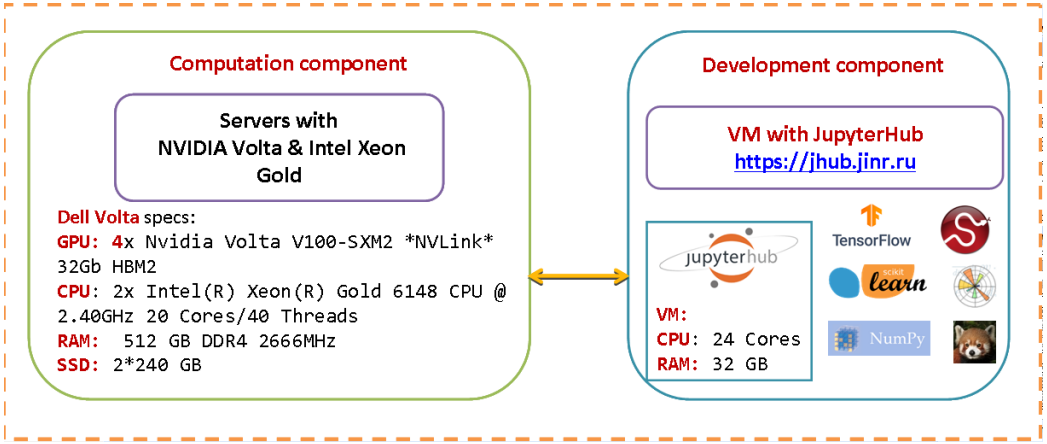
APACHE mxnet

The Microsoft Cognitive Toolkit

Implementation of the neural network approach, methods

To provide all the possibilities both for developing mathematical models and algorithms and carrying out resource-intensive calculations including graphics accelerators, which significantly reduce the calculation time, an ecosystem for tasks of machine learning and deep learning (ML/DL) and data analysis has been created and is actively developing for HybriLIT platform users.





The created ecosystem has two components:

- the computing component is aimed at carrying out resource-intensive, massive parallel tasks of neural network training using NVIDIA graphics accelerators;
- the component for the development of models and algorithms on the *JupyterHub* basis, i.e. a multi-user platform for working with *Jupyter Notebook* (known as *IPython* with the possibility to work in a web browser), including several libraries and frameworks.

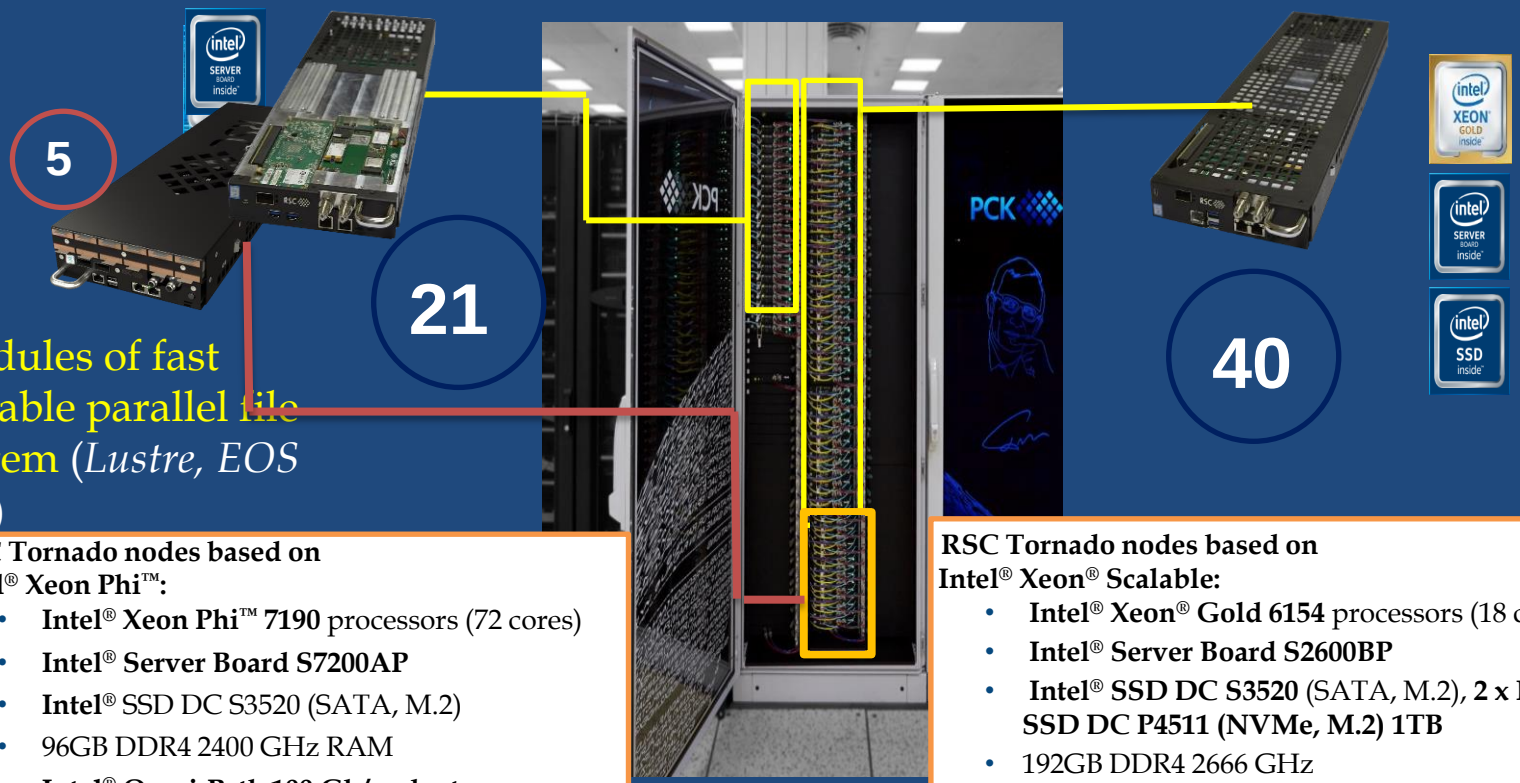
CPU-component of GOVORUN

- Unique heterogeneous and **hyper-converged** system
- Multipurpose high performance system with **direct hot liquid cooling of all system components**
- The most **energy-efficient** system in Russia (**PUE = 1,03**)
- First **100% hot liquid cooling of Intel® Omni-Path** interconnect
- Total peak performance – **210.816 TFLOPS**

System consists of:

- **«RSC Tornado» based on Intel® Xeon® Scalable:**
  - Peak performance – **138.24** TFLOPS
  - Intel® Xeon® Gold 6154 processors (18 cores)
  - Intel® Server Board S2600BP
  - Intel® SSD DC S3520 (SATA, M.2),
    2 x 1TB Intel® SSD DC P4511 (NVMe, M.2)
  - 192 GiB DDR4 2666GHz RAM
  - Intel® Omni-Path 100Gb/s adapter
  - 48-ports Intel® Omni-Path Edge Switch 100 Series with 100% direct hot liquid cooling
- **«RSC Tornado» based on Intel® Xeon Phi™:**
  - Peak performance – **72.576** TFLOPS
  - Intel® Xeon Phi™ 7190 processors (72 cores)
  - Intel® Server Board S7200AP
  - Intel® SSD DC S3520 (SATA, M.2)
  - 96 GiB DDR4 2400GHz RAM
  - Intel® Omni-Path 100Gb/s adapter
  - 48-ports Intel® Omni-Path Edge Switch 100 Series with 100% direct hot liquid cooling

CPU-component of Supercomputer Govorun



Modules of fast scalable parallel file system (*Lustre, EOS* etc.)

RSC Tornado nodes based on Intel® Xeon Phi™:

- **Intel® Xeon Phi™ 7190** processors (72 cores)
- **Intel® Server Board S7200AP**
- **Intel®** SSD DC S3520 (SATA, M.2)
- 96GB DDR4 2400 GHz RAM
- **Intel® Omni-Path 100 Gb/s adapter**

RSC Tornado nodes based on Intel® Xeon® Scalable:

- **Intel® Xeon® Gold 6154** processors (18 cores)
- **Intel® Server Board S2600BP**
- **Intel® SSD DC S3520** (SATA, M.2), **2 x Intel® SSD DC P4511 (NVMe, M.2) 1TB**
- 192GB DDR4 2666 GHz
- **Intel® Omni-Path 100 Gb/s adapter**

Hyper-converged system allows to use all Storage nodes as computing ones in parallel with store/retrieve data. This will add 230 TFLOPS to Govorun system, almost doubling the CPU part performance.

RSC “RSC Tornado” upgrade at JINR (2019)

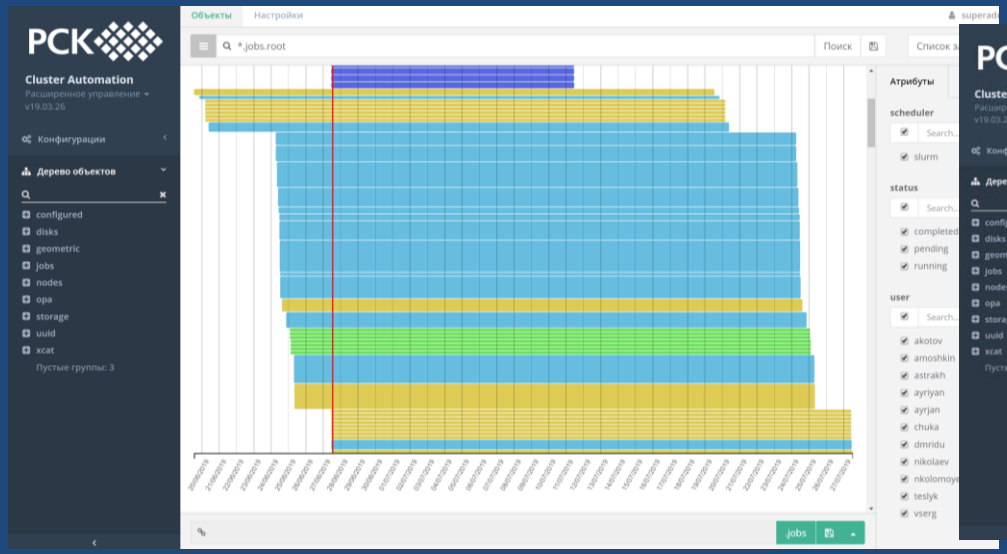
- Extended to **535TFLOPS** peak performance - **#10** Top50
- Software defined architecture
- **#1** in storage performance in Russia, **>300GB/s**
- **Storage-on-demand** scalable solution
- Multi tiered data storage for data management efficiency
- Hot liquid cooled (compute, storage, interconnect)
- Most energy efficient datacenter in Russia (**PUE = 1,027**)

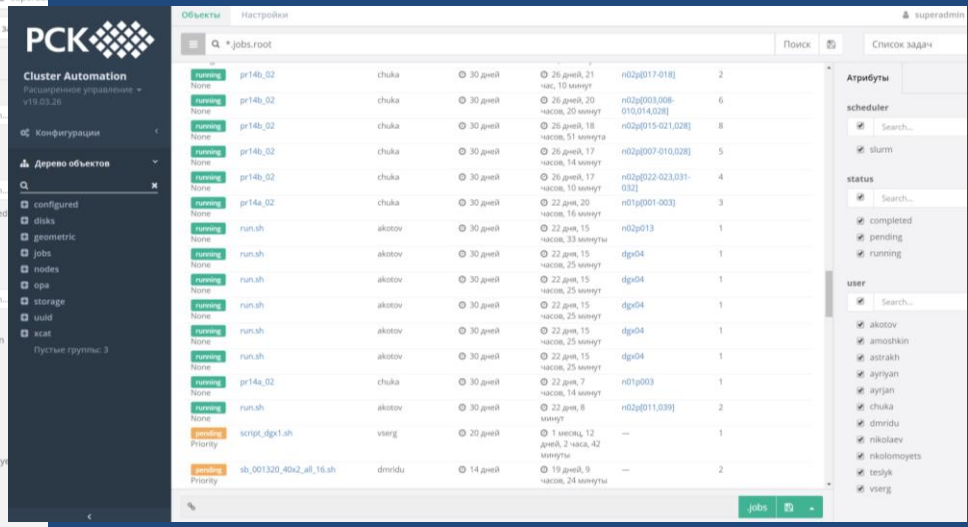
Supercomputer modules

- **Intel® Xeon® Scalable gen 2 nodes:**
  - Peak performance – **463**ТФЛОПС
  - Intel® Xeon® Platinum 8268 processors (24 cores)
  - Intel® Server Board S2600BP
  - Intel® SSD DC S4510 (SATA, M.2), 2 x Intel ® SSD DC P4511 (NVMe, M.2) 2 Tbytes
  - RAM – 192 GB DDR4 2933 ГГц
  - Intel® Omni-Path 100 Gbit/s
  - 48-port Intel® Omni-Path Edge Switch 100 Series co 100% liquid cooling
- **Hyperconverged:**
  - 18 nodes, 12 slots
  - 4 Optane node 3,4 TB IMDT
  - 12 SSD – **256 TB**
  - 2 MDS nodes **3,4 Optane TB**
  - Lustre filesystem as basic option
  - On-demand configuration with **RSC BasIS**
- **Intel® Xeon Phi™ nodes:**
  - Peak performance – **72,576** ТФЛОПС
  - Intel® Xeon Phi™ 7190 CPUs (72 cores)
  - Intel® Server Board S7200AP
  - Intel® SSD DC S3520 (SATA, M.2)
  - RAM – 96 GB DDR4 2400 ГГц
  - Intel® Omni-Path 100 Гбит/с
  - 48-port Intel® Omni-Path Edge Switch 100 Series 100% liquid cooling

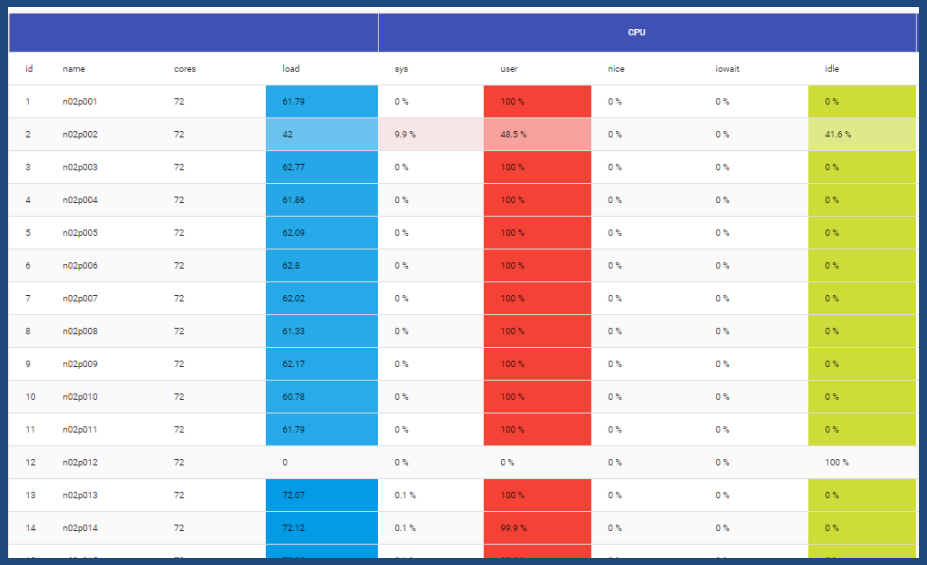
“RSC BasIS” monitoring software stack

Monitoring systems





RSC cluster automation: https://govorun.jinr.ru



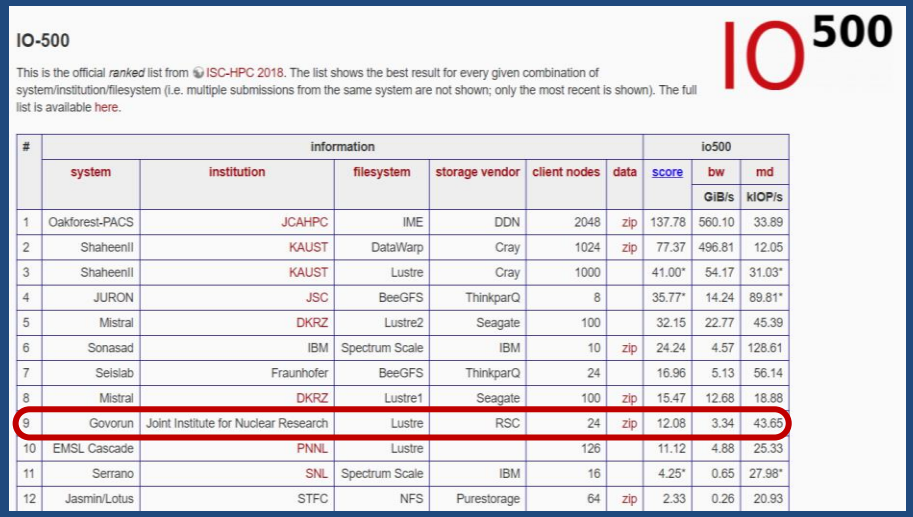
				CPU				
id	name	cores	load	sys	user	nice	iowait	idle
1	n02p001	72	61.79	0 %	100 %	0 %	0 %	0 %
2	n02p002	72	42	9.9 %	48.5 %	0 %	0 %	41.6 %
3	n02p003	72	62.77	0 %	100 %	0 %	0 %	0 %
4	n02p004	72	61.86	0 %	100 %	0 %	0 %	0 %
5	n02p005	72	62.09	0 %	100 %	0 %	0 %	0 %
6	n02p006	72	62.8	0 %	100 %	0 %	0 %	0 %
7	n02p007	72	62.02	0 %	100 %	0 %	0 %	0 %
8	n02p008	72	61.33	0 %	100 %	0 %	0 %	0 %
9	n02p009	72	62.17	0 %	100 %	0 %	0 %	0 %
10	n02p010	72	60.78	0 %	100 %	0 %	0 %	0 %
11	n02p011	72	61.79	0 %	100 %	0 %	0 %	0 %
12	n02p012	72	0	0 %	0 %	0 %	0 %	100 %
13	n02p013	72	72.07	0.1 %	100 %	0 %	0 %	0 %
14	n02p014	72	72.12	0.1 %	99.9 %	0 %	0 %	0 %

Stat-hlit: https://home-hlit.jinr.ru/#/

litMon: https://litmon.jinr.ru

Supercomputer Govorun in HPC ratings

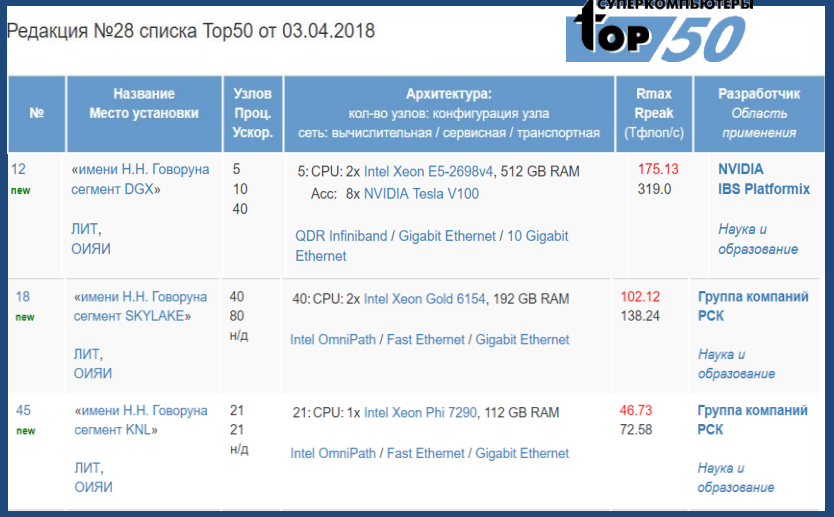
Hyper-converged Govorun system allows to all of its nodes with SSD drives to act as storage and compute nodes at the same time. RSC software stack BasIS and RSC Tornado hardware architecture support Software Defined Storage of different types (Luster, EOS, BeGFS etc.)



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



Редакция №28 списка Top50 от 03.04.2018

№	Название Место установки	Узлов Проц. Ускор.	Архитектура: кол-во узлов: конфигурация узла сеть: вычислительная / сервисная / транспортная	Rmax Rpeak (Тфлоп/с)	Разработчик *Область применения*
12 new	«имени Н.Н. Говоруна сегмент DGX» ЛИТ, ОИЯИ	5 10 40	5: CPU: 2x Intel Xeon E5-2698v4, 512 GB RAM Acc: 8x NVIDIA Tesla V100 QDR Infiniband / Gigabit Ethernet / 10 Gigabit Ethernet	175.13 319.0	NVIDIA IBS Platformix *Наука и образование*
18 new	«имени Н.Н. Говоруна сегмент SKYLAKE» ЛИТ, ОИЯИ	40 80 н/д	40: CPU: 2x Intel Xeon Gold 6154, 192 GB RAM Intel OmniPath / Fast Ethernet / Gigabit Ethernet	102.12 138.24	Группа компаний РСК *Наука и образование*
45 new	«имени Н.Н. Говоруна сегмент KNL» ЛИТ, ОИЯИ	21 21 н/д	21: CPU: 1x Intel Xeon Phi 7290, 112 GB RAM Intel OmniPath / Fast Ethernet / Gigabit Ethernet	46.73 72.58	Группа компаний РСК *Наука и образование*

Govorun is ranked on **9**th position in the edition on **26.06.2018** of **IO500** List a new industry benchmark for HPC storage systems:
https://www.vi4io.org/io500/list/18-06/start

All computing components of the supercomputer Govorun were included in the Top50 rating of the most powerful supercomputers in the CIS countries:
http://top50.supercomputers.ru/archive/2018/04

Hot NICA topics for HPC

- **Physical generators**
  Monte-Carlo simulations with different physics input

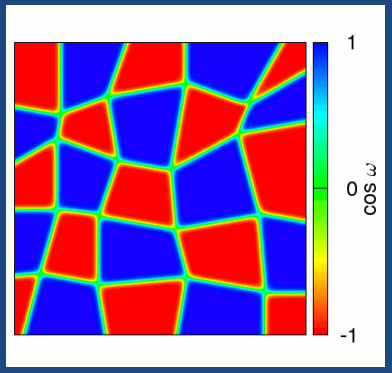
- **Detectors simulation**
  detailed detector description with realistic detector response
- **Tracks reconstruction**
  high efficiency for finding tracks with different methods (deep learning & etc.) ~ 1000 tracks in event

- **BigData analysis**
  > 10^{10} events, 1min/ev, ~2 years on our today resources
  Multicore & multithreads computing
  BigPanDA & GRID
  Clouds and cloud services

Hot Theoretical Physics Topics for HPC

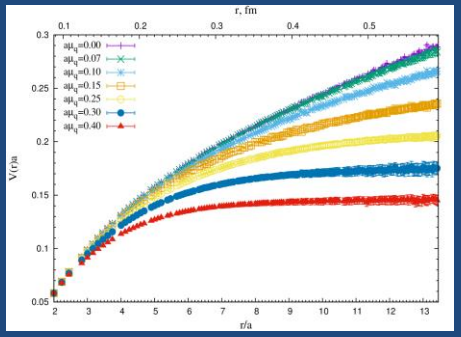
THEORY OF HADRONIC MATTER UNDER EXTREME CONDITIONS

(In theoretical support of NICA and other relativistic heavy-ion physics experiments)

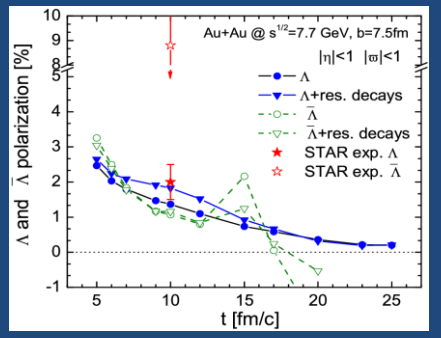


Parallel computing for Lattice QCD, functional RG, statistical and hydrodynamical models of HIC, sophisticated models of QCD vacuum, strongly correlated systems in condensed matter physics

Critical phenomena in hot dense hadronic matter in the presence of strong electromagnetic fields, deconfinement and chiral symmetry restoration:



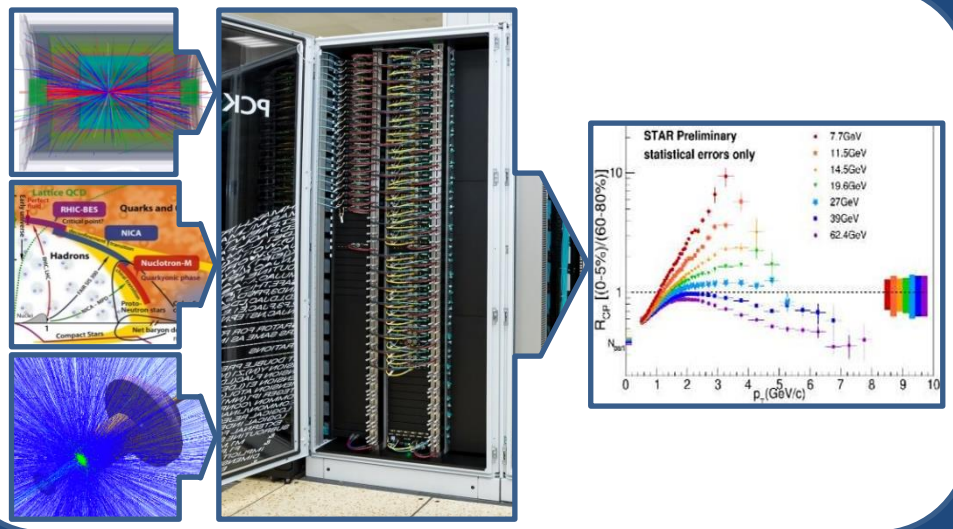
QCD Phase diagram
Thermodynamics of N_f=2+1+1 QCD
Real-time spectral properties of thermal QCD
Transport properties of hadronic matter
Properties of cold dense SU(2) QCD through lattice calculations
Anderson transition in the N_f=2+1+1 QCD
Z(N) symmetry & meta-stable states

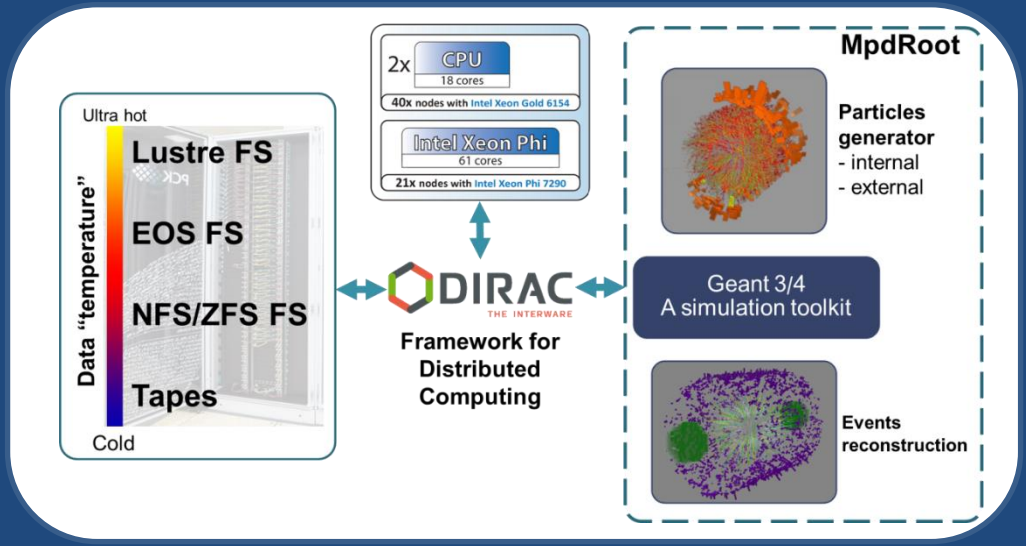


Upto the present time computations of BLTP group has been performed mostly at the external resourses: Russia: MSU - «Lomonosov», hybrid clusters at ITEP & IHEP; Japan: Osaka - SX-ACE, Kioto – CP-16000, Germany: hybrid clusters at Heidelberg & Hissen Uni; India: Annapurna (Inst. of Math. Sciences)

Supercomputer “GOVORUN” will tremendously increase efficency of theoretical investigations!

Supercomputer Govorun for the NICA megaproject



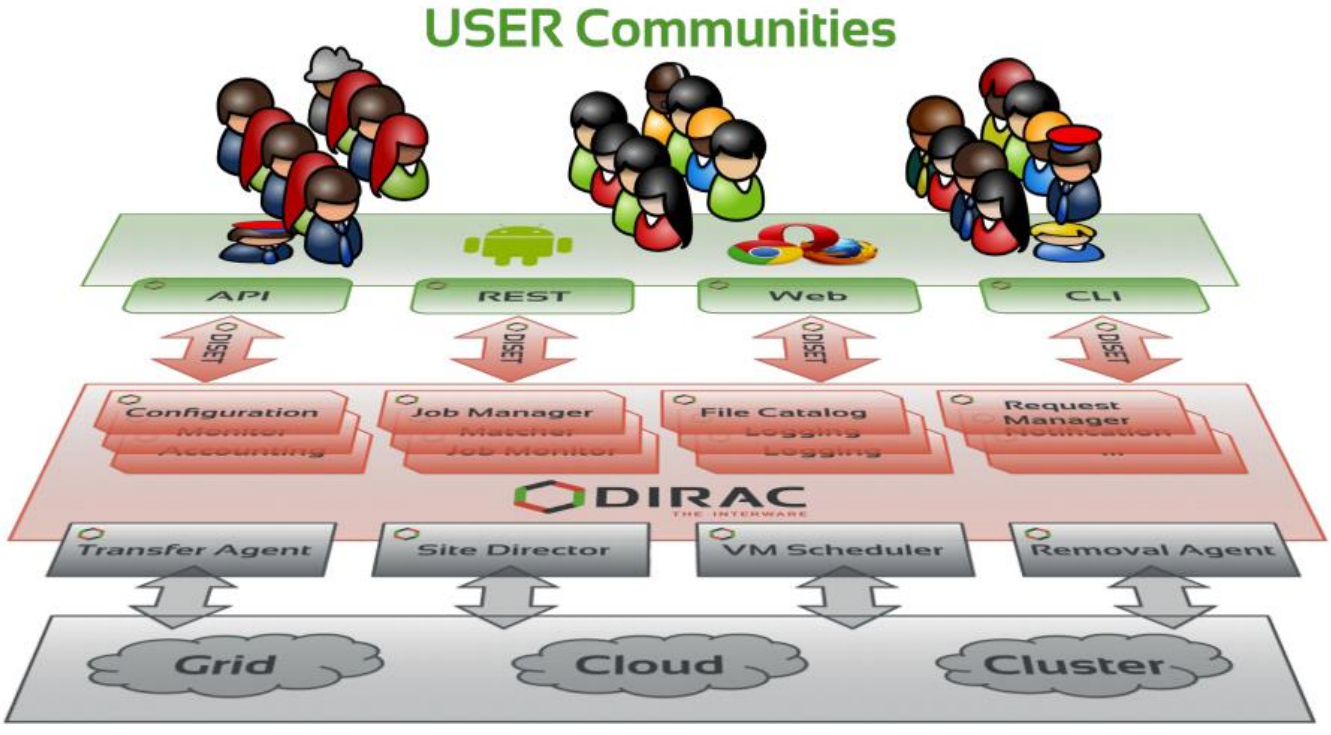


At present, the GOVORUN supercomputer is used for both theoretical studies and event simulation for the MPD experiment of the NICA megaproject. To generate simulated data of the MPD experiment, the CPU computing components of the GOROVUN supercomputer, i.e. Skylake (2880 cores) and KNL (6048 cores), are used; data are stored on the ultrafast data storage system (UDSS) under the management of the *Lustre file system* with a subsequent transfer to cold storages controlled by the *EOS* and *ZFS* file systems. UDSS currently has five storage servers with 12 SSD disks using the NVMe connection technology and a total capacity of 120 TB, which ensures low time of access to data and a data acquisition/output rate of 30 TB per second.

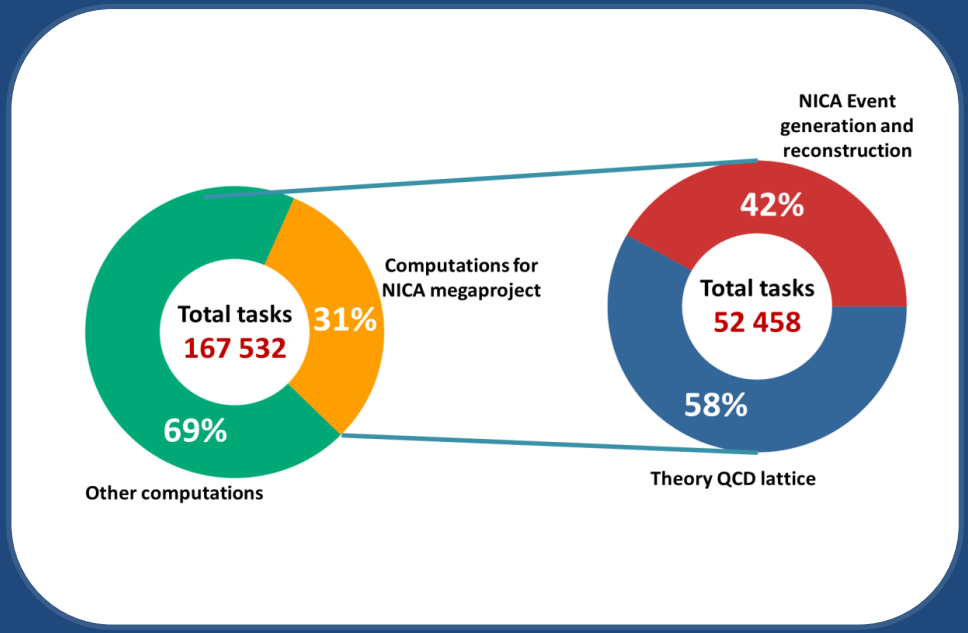
The *DIRAC* software is used for managing jobs and the process of reading out/recording/processing data from various types of storages and file systems.

DIRAC: the interware

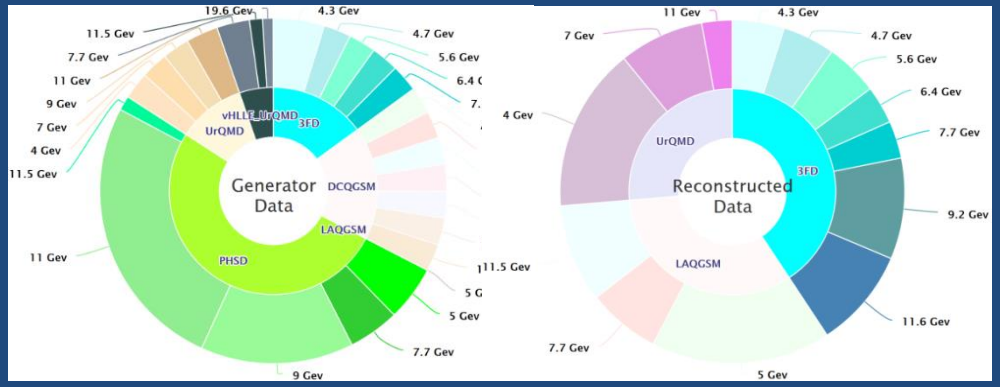
- A software framework for distributed computing
- A **complete** solution to one (or more) user community
- Builds a layer between users and resources



Statistics of using all components of the supercomputer for the NICA megaproject



Since commissioning, more than 164,000 tasks were performed on the GOVORUN supercomputer. One third of them refer to computing for the NICA megaproject. Moreover, more than half are theoretical calculations carried out on all computing components of the supercomputer. More than 40% directly relate to the event generation and reconstruction for the MPD experiment.

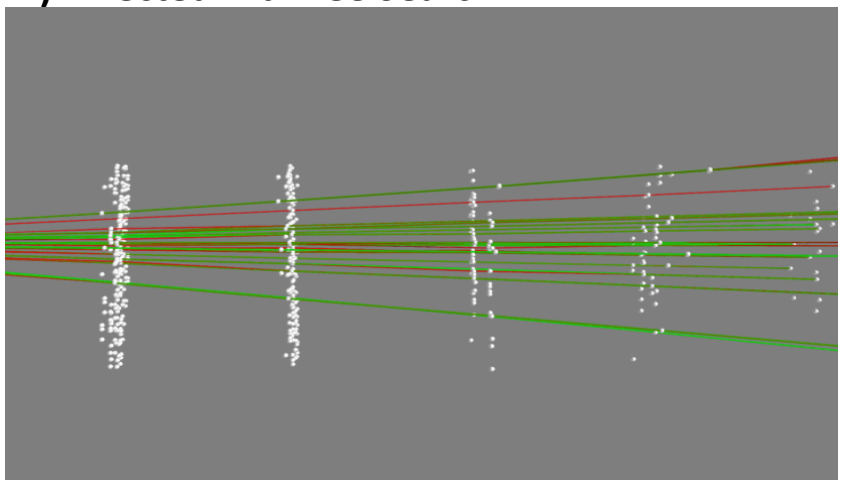


In 2019 over 120 million events for the MPD experiment have already been generated using the *UrQMD* generator. At the present time reconstructed almost 30 million events. The CPU component and the UDSS are used for the solution of these problems.

Track reconstruction based on deep learning methods

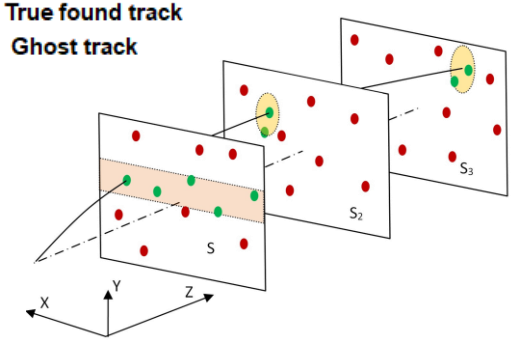
The tracking procedure is especially difficult for experiments like **BM@N with GEM detectors** due to the famous GEM shortcoming caused by appearance in them, **besides of true hits, much more fake hits**.

1) Directed K-d Tree Search



Real track
True found track
Ghost track

White dots are both hits and fakes



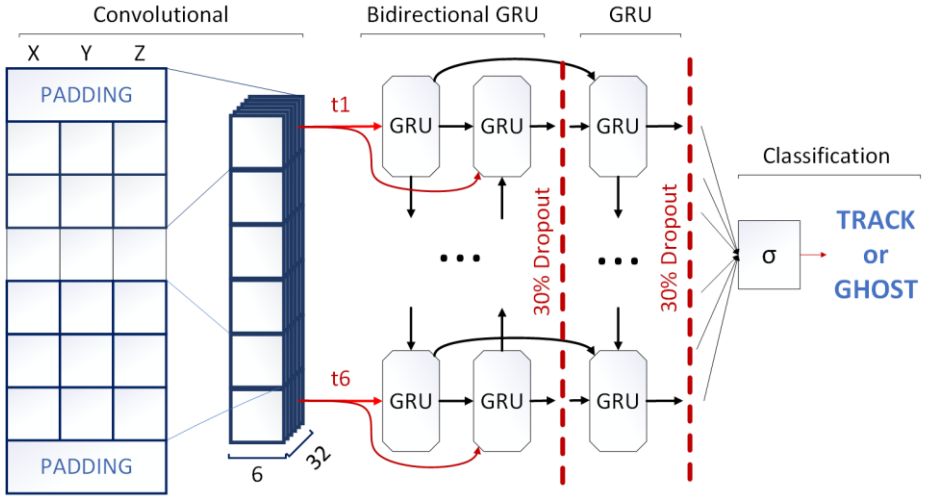
Bunch of track-candidates

Trained RNN can currently **process 6500 track-candidates in one second** on the single Nvidia Tesla M60 from HybriLIT cloud service.
http://ceur-ws.org/Vol-2023/37-45-paper-6.pdf

Our solution - two step tracking procedure:

1. **Preprocesing by directed K-d tree search** to find all possible track-candidates as clusters joining all hits from adjacent GEM stations lying on a smooth curve.
2. **Deep recurrent network** trained on the big simulated dataset with 82 677 real tracks and 695 887 ghosts **classifies track-candidates in two groups: true tracks and ghosts**.

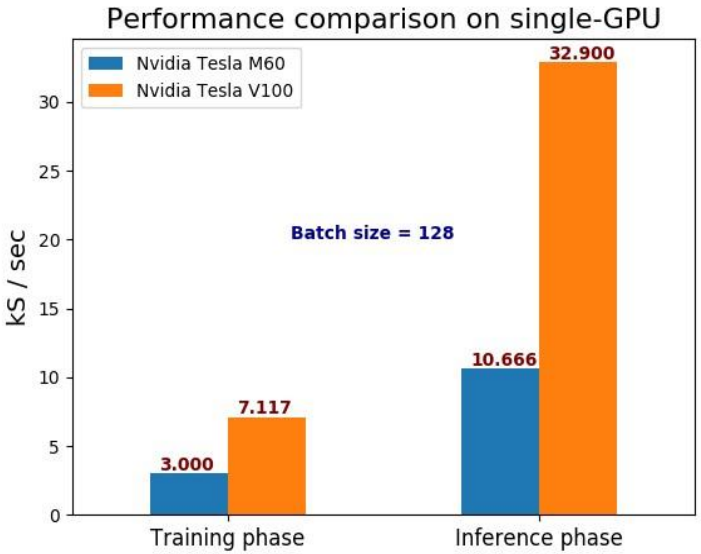
2) Deep Recurrent Neural Network Classifier



five hidden layers: convolutional network layer and two Gated Recurrent Unit (GRU) layers alternating with dropout layers

Results on simulated events: True track recognition efficiency is on the level of 97.5%.

Track reconstruction based on deep learning methods



	Training	Inference
CPU (track-candidate/sec)	11 122	528 018
GPU (track-candidate/sec)	7159	3 483 608

Results of the test run on GOVORUN allows also to evaluate approximately a processing speed for one event of a future HL-LHC or NICA detector with 10000 tracks on a reasonable level of 3 microseconds.

THANK YOU FOR YOUR ATTENTION !