Supercomputer "GOVORUN" — new prospects for heterogeneous computations at JINR

ADAM Gheorghe, KORENKOV Vladimir, PODGAINY Dmitry, STRELTSOVA Oksana

Laboratory of Information Technologies
Joint Institute for Nuclear Research

The 8th International Conference
"Distributed Computing and Grid-technologies in Science and Education"
10-14 September, 2018

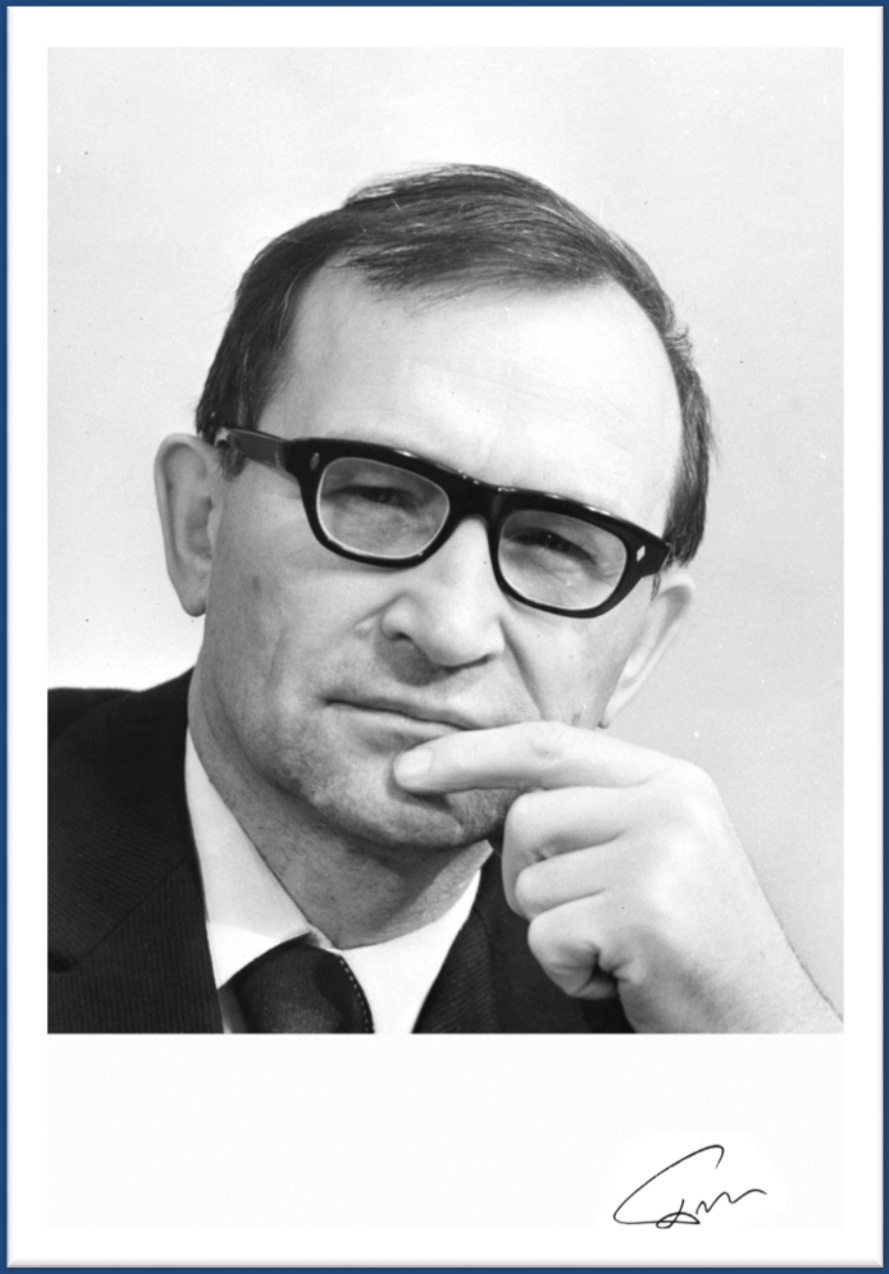
Supercomputer "GOVORUN" is a joint project of the *N.N. Bogolyubov Laboratory of Theoretical Physics* and the L*aboratory of Information Technologies* under support of *JINR Directorate*

The project is aimed to radically accelerate complex theoretical and experimental studies underway at JINR, including the NICA complex

Presentation of Supercomputer “GOVORUN”

On **March 27**, 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).



GOVORUN Nikolai Nikolayevich

Corresponding Member of the USSR Academy of Sciences,

1966-1988 – LCTA Deputy Director on research work,

1988-1989 – Director of LCTA, JINR

Since 1966, JINR has been involved with the overall development of a new scientific branch – informatics, the head of which became N.N.Govorun.

Under the guidance of Nikolai Nikolayevich, the Laboratory passed all the way of the computer science development beginning from the introduction of mathematical computations on the first computers and ending the creation of computer networks for scientific research, including both the issues of creating system mathematical software and the task of experimental data processing in off-line and on-line modes and real-time experiment management. Pioneering works in all these areas are associated with his name.

From HybriLIT cluster to HybriLIT platform

The supercomputer is a natural continuation of heterogeneous platform and leads to a significant increase in the performance of both CPU and GPU components.

2014

Total peak performance:
140 TFlops for single precision;
50 TFlops for double precision

2018

Total peak performance:
1000 TFlops for single precision;
500 TFlops for double precision

HYBRILIT HETEROGENEOUS COMPUTING PLATFORM

Unified software and information environment

HybriLIT education and testing cluster

- Intel Xeon
- Intel Xeon Phi
- Nvidia Tesla K20
- Nvidia Tesla K40
- Nvidia Tesla K80

10 computation nodes

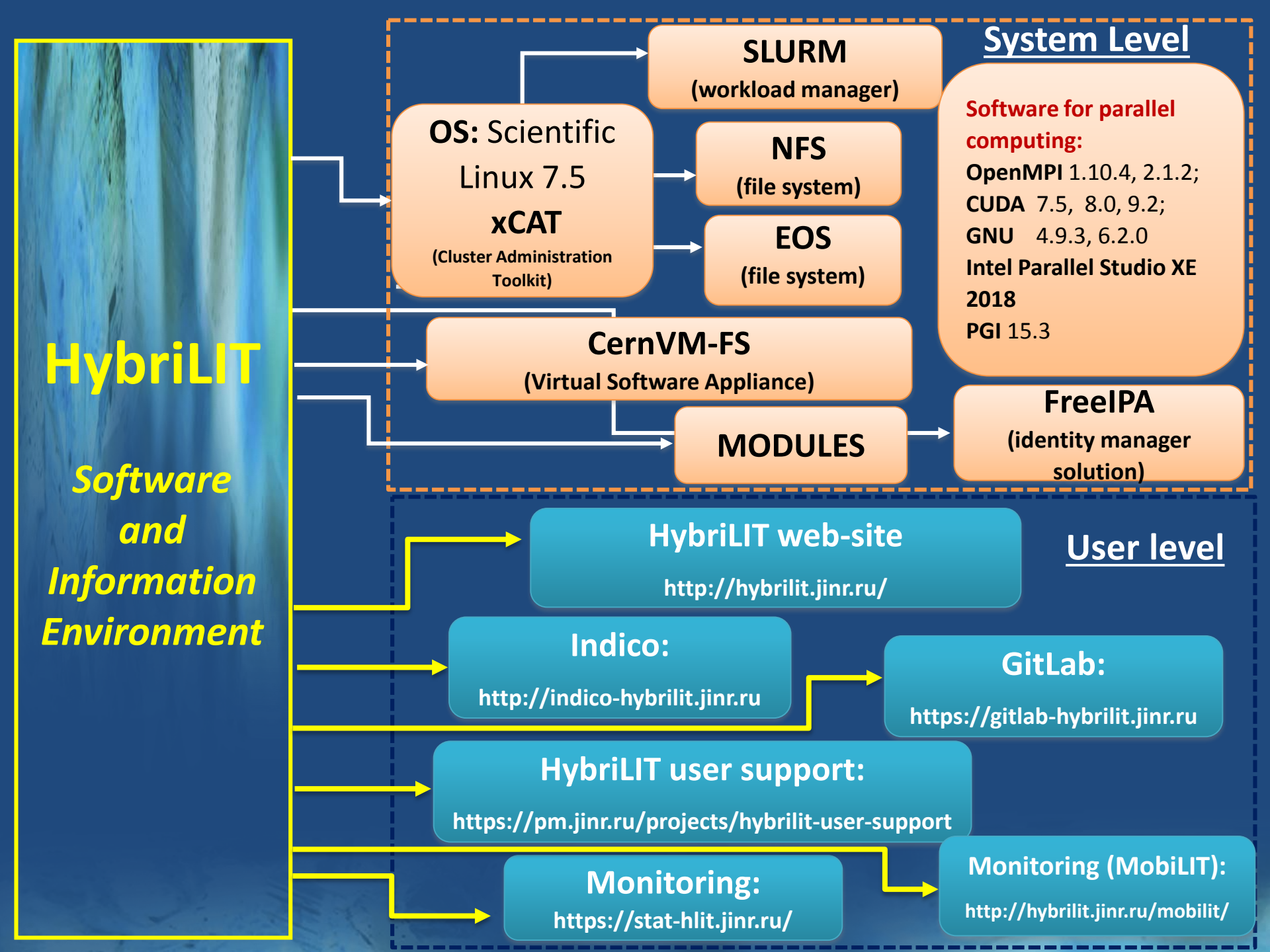
SUPERCOMPUTER «GOVORUN»

40 nodes	**CPU-component** Intel Skylake
21 nodes	**CPU-component** Intel Xeon Phi (KNL)
5 nodes	**GPU-component** GPU DGX-1 Volta (NVIDIA Tesla V100)

Total Peak Performance

Double precision **500** Tflops

Single precision **1000** Tflops


HybriLIT
Software and Information Environment
System Level
SLURM
(workload manager)
OS: Scientific Linux 7.5
xCAT
(Cluster Administration Toolkit)
NFS
(file system)
EOS
(file system)
Software for parallel computing:
OpenMPI 1.10.4, 2.1.2;
CUDA 7.5, 8.0, 9.2;
GNU 4.9.3, 6.2.0
Intel Parallel Studio XE 2018
PGI 15.3
CernVM-FS
(Virtual Software Appliance)
MODULES
FreeIPA
(identity manager solution)
User level
HybriLIT web-site
http://hybrilit.jinr.ru/
Indico:
http://indico-hybrilit.jinr.ru
GitLab:
https://gitlab-hybrilit.jinr.ru
HybriLIT user support:
https://pm.jinr.ru/projects/hybrilit-user-support
Monitoring:
https://stat-hlit.jinr.ru/
Monitoring (MobiLIT):
http://hybrilit.jinr.ru/mobilit/

NVIDIA DGX-1

The world's most powerful supercomputer for AI

8x Tesla V100 with NVLink interconnect

60 TFlops double precision

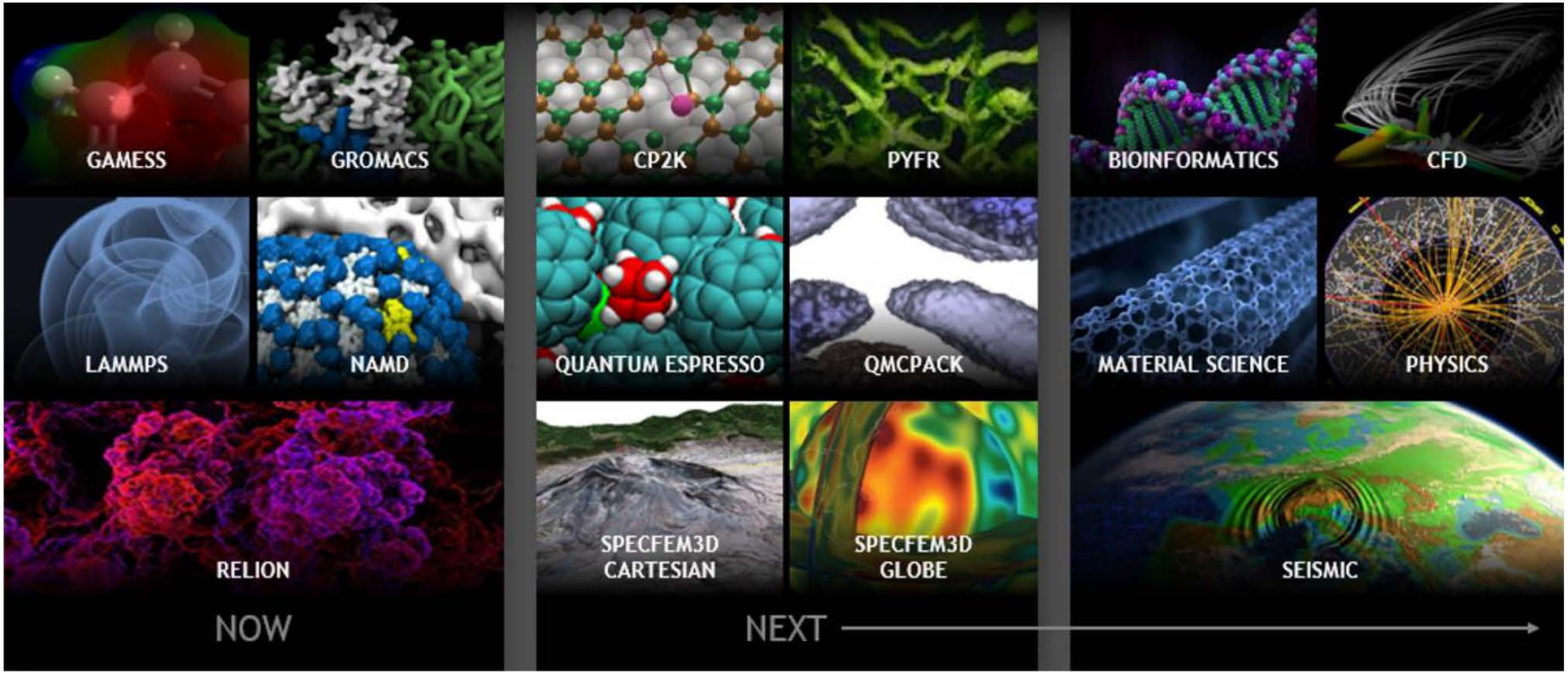
120 TFlops single precision

Unique energy efficiency 3.2 kW

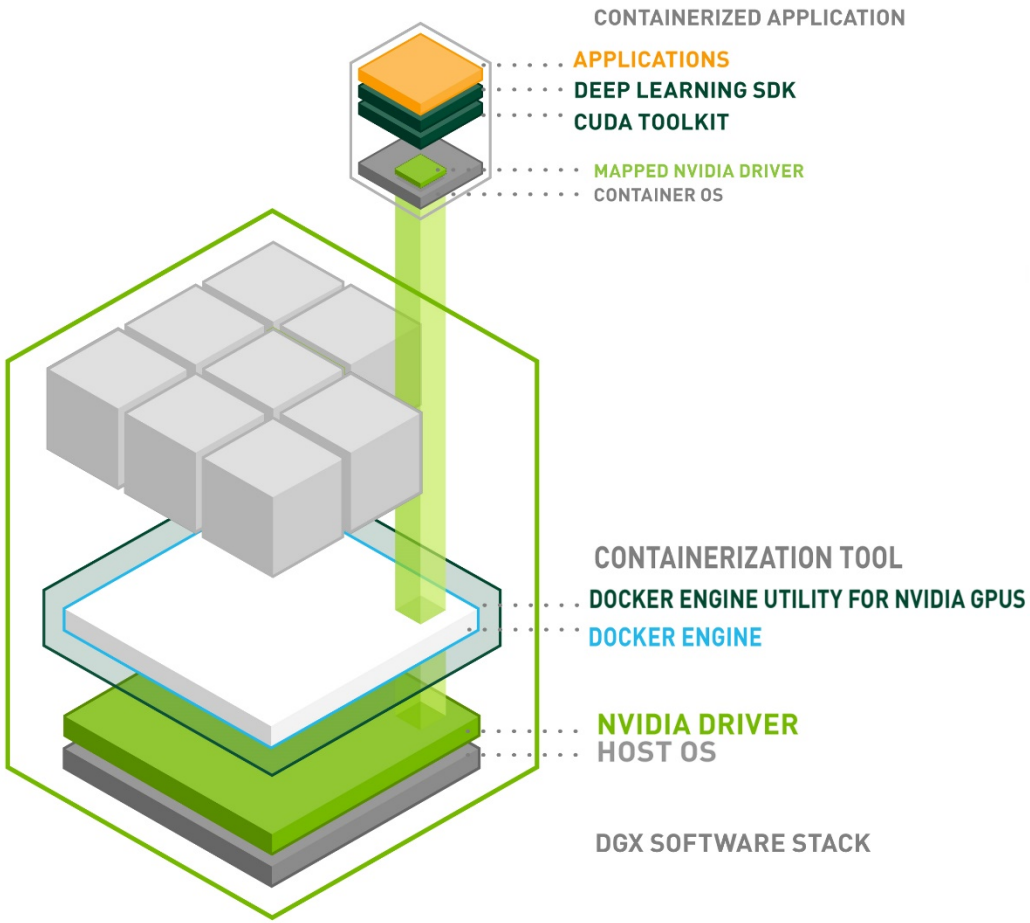
Full stack deep learning software preinstalled

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

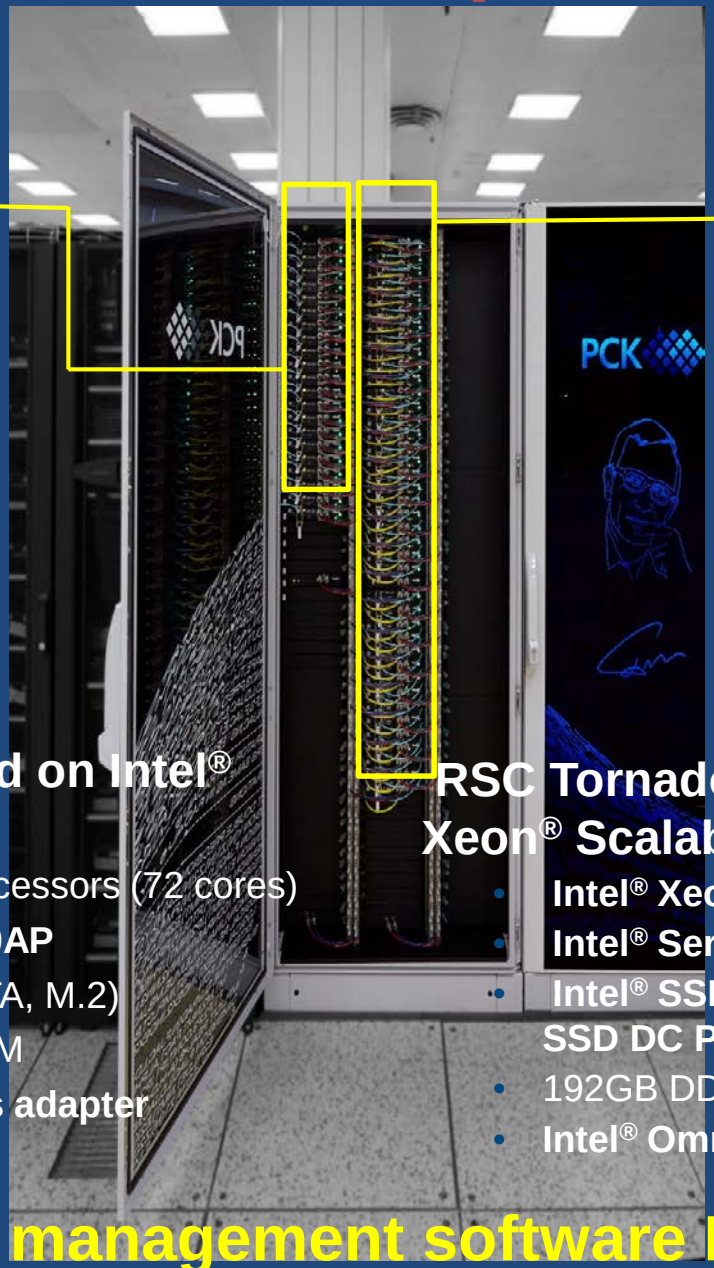
CPU-component of GOVORUN

- 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 = 1,02**)
- 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

Current CPU-componet ditails



21

40

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**

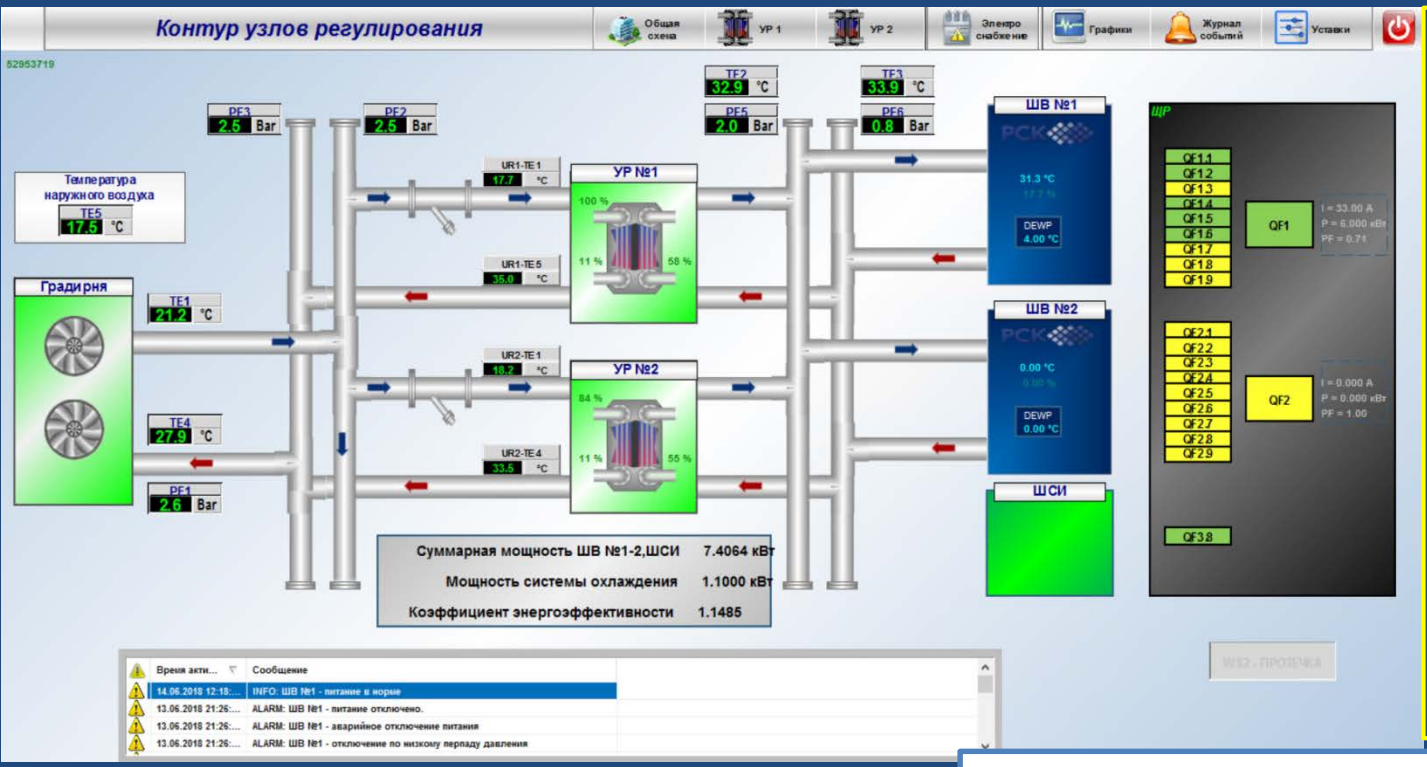
Multi-level management software RSC BasIS

Heterogeneous system at JINR: the decision of "RSK Tornado"

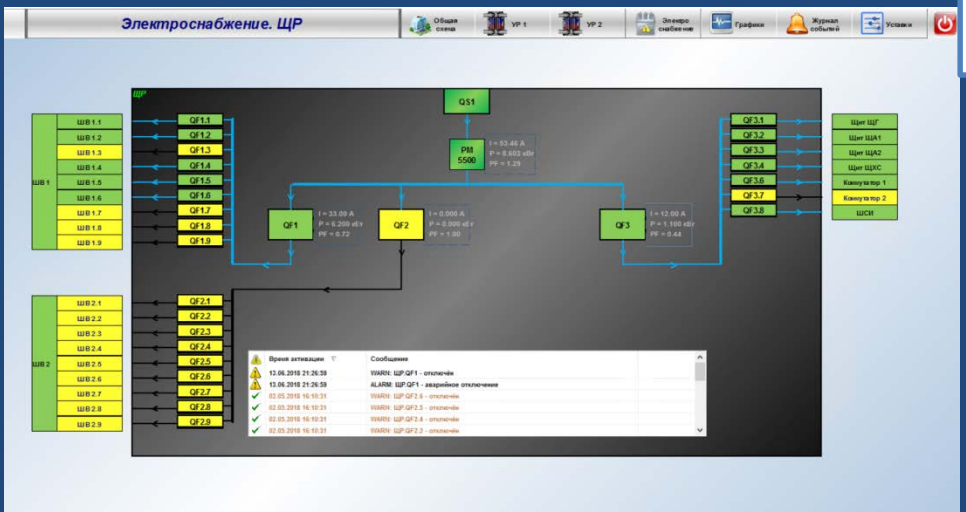
Record-breaking power density – up to **100 kW** per **42U** cabinet (800x800x200 mm) with proven high efficiency RSC liquid cooling technology providing record breaking power density and compactness

RSC

Monitoring system



High reliability: High availability and fail-safe operation are provided by an innovative system of control and monitoring of separate nodes and the entire cluster system



Software stack «RSC BasIS» for multi-level system management

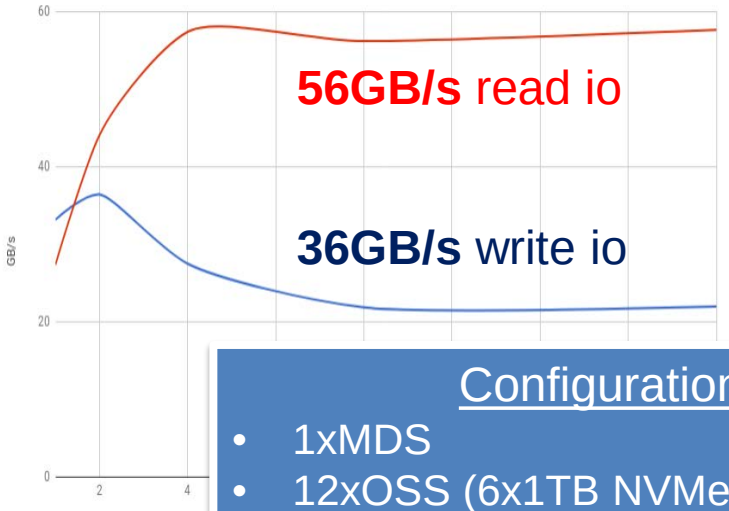
GOVORUN's File system-on-Demand

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



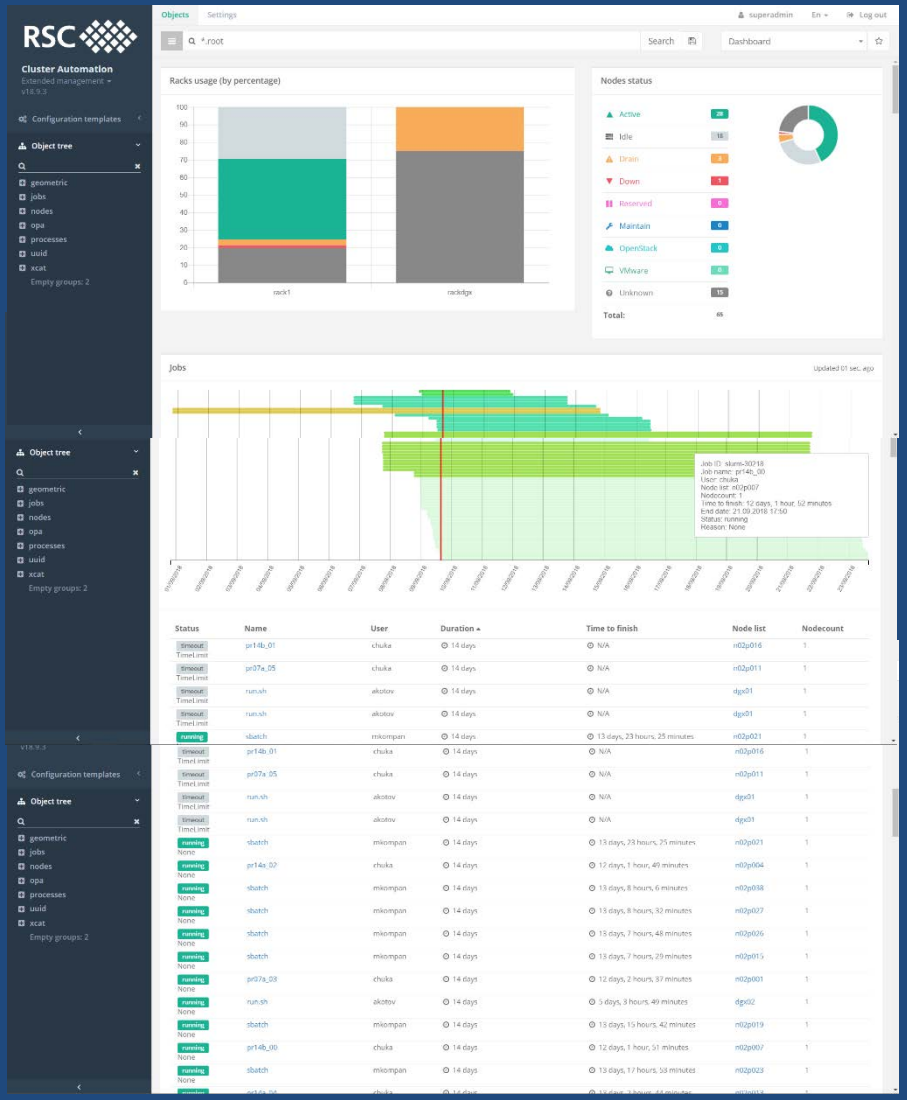
Configuration:

- 1xMDS
- 12xOSS (6x1TB NVMe SSDs each)
- 24xClients to load Lustre
- 100GBs Intel Omni-Path interconnect

GOVORUN is ranked on 9th position in the latest edition of IO500 List a new industry benchmark for HPC storage systems.

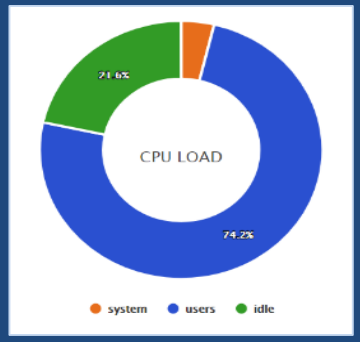
Supercomputer GOVORUN: user groups

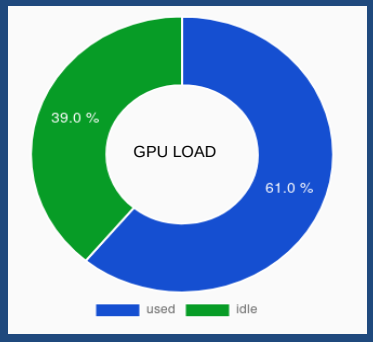
Current workload of supercomputer GOVORUN



GOVORUN user groups:

- BLTP (S.Nedelko group, M.Hnatich group)
- VBLHEP (O. Rogachevsky group)
- LIT (O. Chuluunbaatar, A.Ayryan group, E.V. Zemlanay group, I. Hristov)
- JUNO experiment,
- NOVA experiment

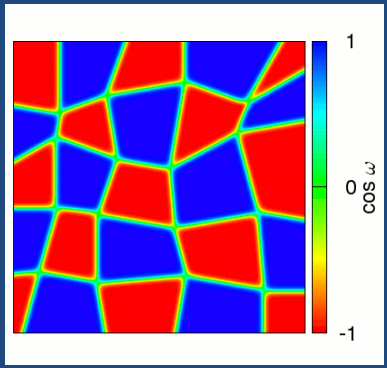




Hot Theoretical Physics Topics for HPC

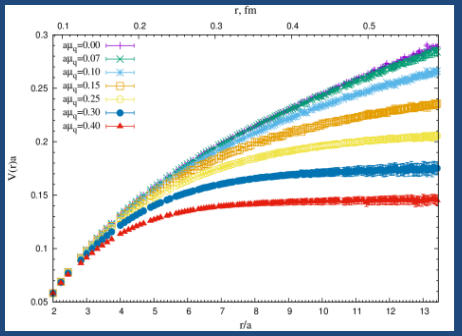
THEORY OF HADRONIC MATTER UNDER EXTREME CONDITIONS

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

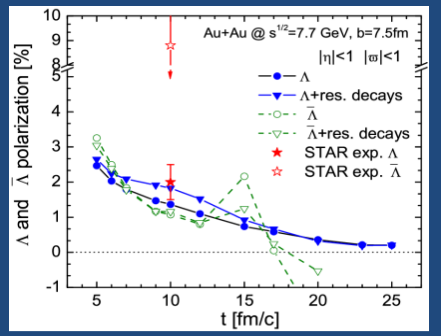


Parallel computing for Lattice QCD, functional RG, statistical and hydrodynamical models of HIC, sophisticated models of QCD vacuum, strongly correlated systems in condensed matter physics

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



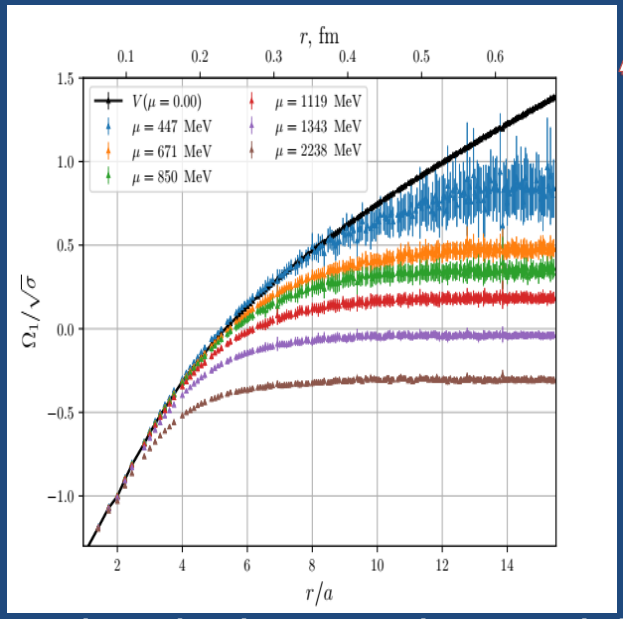
QCD Phase diagram
Thermodynamics of N_f=2+1+1 QCD
Real-time spectral properties of thermal QCD
Transport properties of hadronic matter
Properties of cold dense SU(2) QCD through lattice calculations
Anderson transition in the N_f=2+1+1 QCD
Z(N) symmetry & meta-stable states



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)

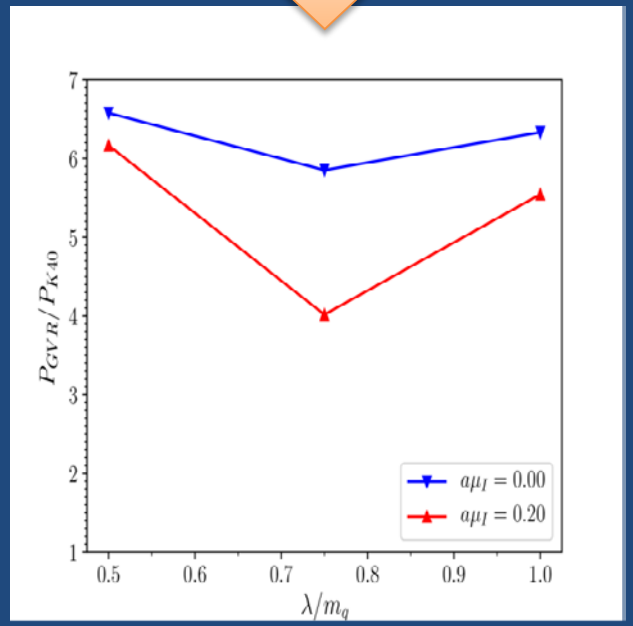
Supercomputer “GOVORUN” will tremendously increase efficency of theoretical investigations!

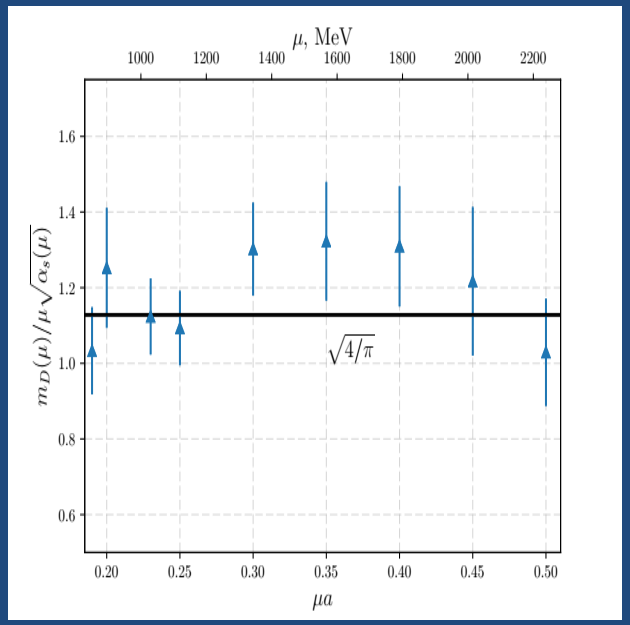
Lattice study of dense quark matter on GOVORUN



Color singlet grand potential of quark-antiquark pair in dense medium for several values of the chemical potential. The black curve is the potential of a static quark-antiquark pair at zero temperature and density.

Achieved acceleration using GOVORUN GPU V100: 6 times (GPU NVIDIA V100 vs K40)





Debye mass, normalized by $\mu\alpha_s^{1/2}(\mu)$ as a function of chemical potential. At the leading order this ratio equals the constant $(4/\pi)^{1/2}$, shown as a line.

N. Yu. Astrakhantsev, V. G. Bornyakov, V. V. Braguta,E.-M. Ilgenfritz, A. Yu. Kotov, A. V. Molochkov, A. A. Nikolaev and A. Rothkopf: Lattice study of static quark-antiquark interactions in dense quark matter. arXiv:1808.06466 [hep-lat]

GOVORUN Usage
Lattice Quantum Chromodynamics
Compiled by: Mridupawan Deka

- **CPU Only**

  1.tmLQCD Code Suite: (2+1+1)-flavor SU(3) Twisted Mass Fermions.
  Skylake: **2048 cores**, KNL: **864 cores** (on average).
  Total core hour: 2 million (approx.).
  Total data generated and saved: 400 GB (approx.).
  2.MILC Code Suite: 2-flavor SU(3) Staggered Fermions.
  Skylake: 1536 cores, KNL: 1024 cores (on average).
  Total core hour: 250 000 (approx.).
  Measurements are done online. Hence primary data are not saved.

- **GPU Only**

1.cuLGT Code Suite: Landau and Coulomb Gauge Fixing of tmLQCD Gauge configurations.
Single GPU application.
Number of GPUs used: 3 and Number of CPUs used: 3.
Total Gauge Fixed data generated and saved: **5 TB** (approx.).

GOVORUN Usage:
Lattice Quantum Chromodynamics
Compiled by: Mridupawan Deka

Immediate Planned Usage

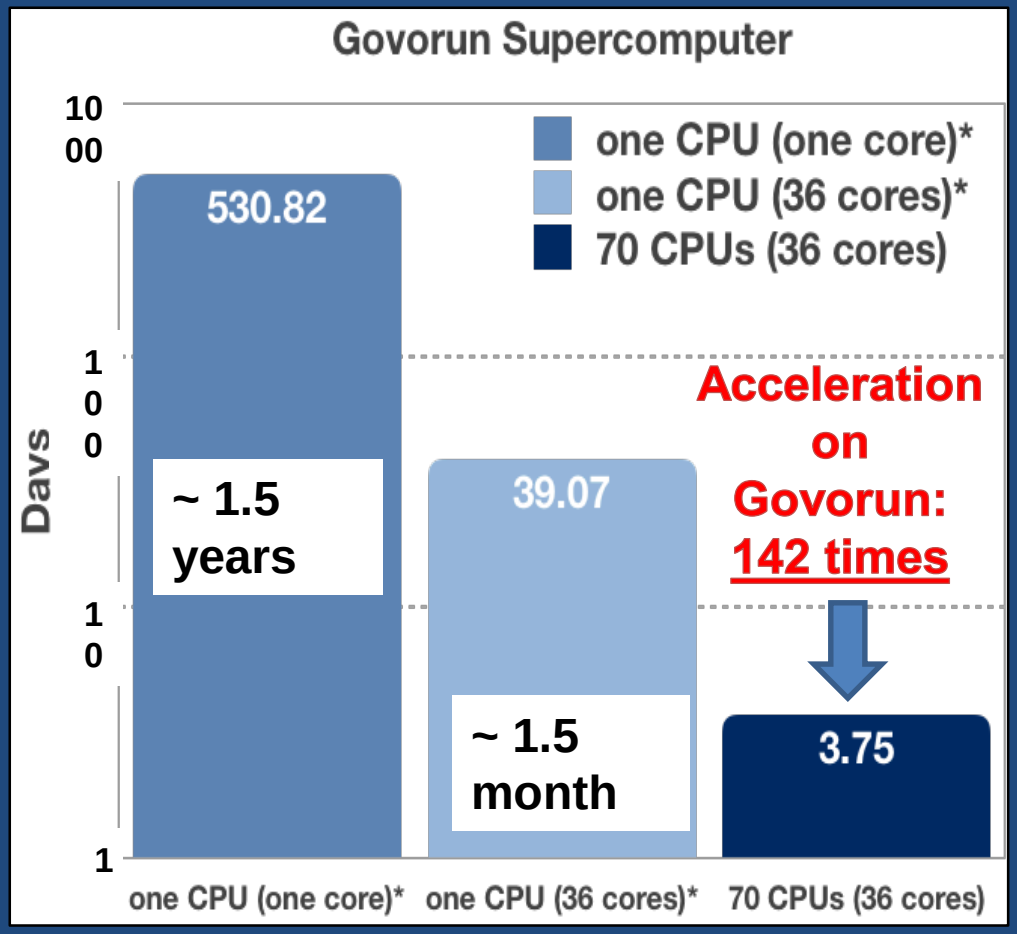
- **CPU + GPU**

1.MILC Code Suite: (2+1)-flavor SU(3) Staggered Fermions.
Multi-CPU and multi-GPU applications.
Planned usage:
1-2 GPU nodes. i.e. 8-16 GPU cards and 64-128 CPUS.
Current Status: In testing phase.

- **CPU**

1.SU(2) Gauge + Higgs Code Suite
Multi-node CPU or Core applications.
Planned usage: 2048 cores.
Current Status: The code is ready for data production and measurement.

Optimization problem for the heat equation towards improvement of the "temperature valves“ characteristics



A hybrid algorithm MPI+OpenMP has been developed for solving optimization problem for nonlinear unsteady heat equation. The optimization problem has been formulated in order to improve design of the so-called “temperature valves” technique for the pulse injection (in the millisecond range) of working gases into the multiply charged ion source ionization chamber.

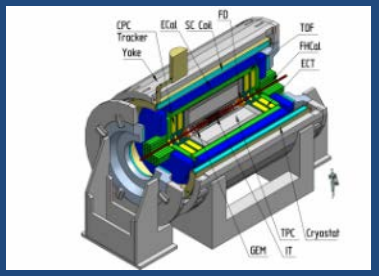
Achieved acceleration using GOVORUN: 142 times

The work has been done in collaboration with colleagues from VBLHEP and IEP SAS (Kosice, Slovakia)

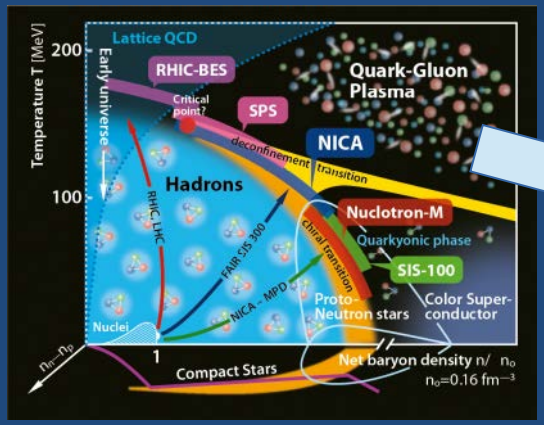
A. Ayriyan, J. Busa Jr., E. E. Donets, H. Grigorian, J. Pribis. Applied Thermal Engineering (2016), v. 94, pp. 151-158.

A. Ayriyan, J. Busa Jr. et al. Hybrid algorithm for optimization problem. (in preparation)`5

NICA computing challenge

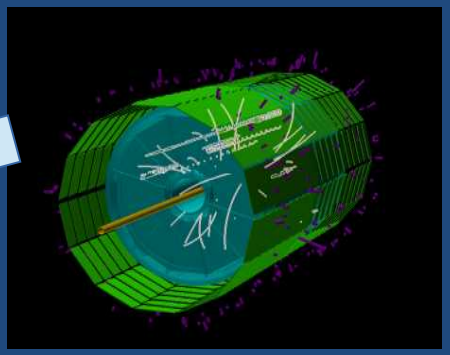


MPD experiment

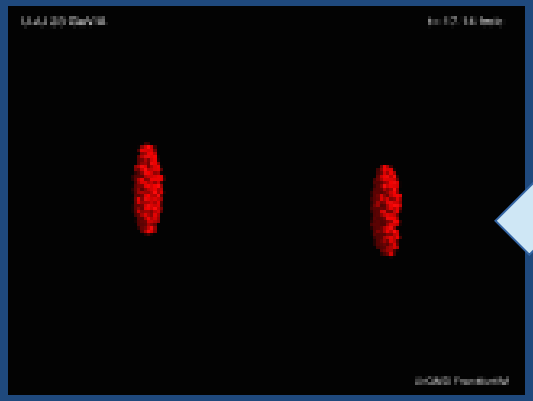


QCD phase diagram

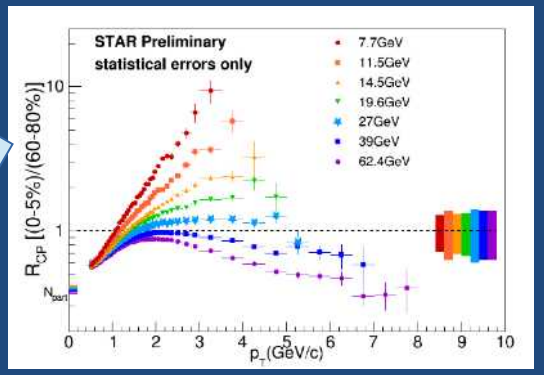
GOVORUN



Events reconstruction



Simulations



Physics analysis

Hot NICA topics for HPC

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

- **Detectors simulation**
  detailed detector description with realistic detector response

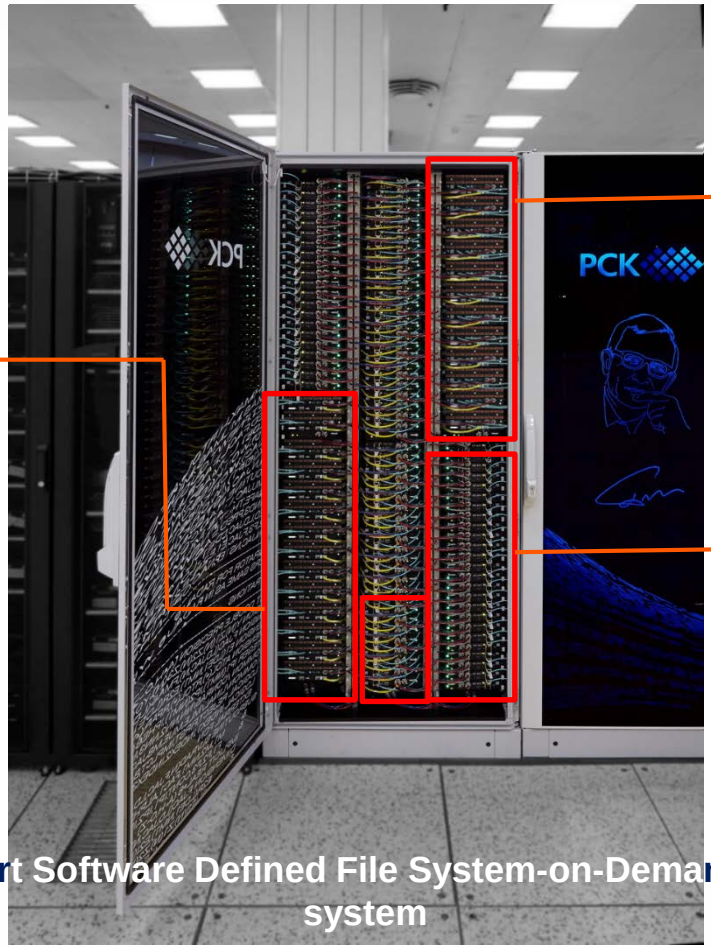
- **Tracks reconstruction**
  high efficiency for finding tracks with different methods (deep learning & etc.) ~ 1000 tracks in event

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

Proposed GOVORUN system extension

14

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



12

Modules of fast parallel data processing and analysis

40

Modules for additional workload and data analysis

RSC BasIS software to support Software Defined File System-on-Demand service for hyper-converged system

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

Conclusion

The supercomputer is a natural continuation of the heterogeneous platform. It leads to a significant increase in the performance of both the CPU and GPU components. Multi-level management system enables the creation of a software-configured platform based on GOVORUN

It will allow one

- to get usable access to computing and information HPC resources,
- implement the data processing workflows on HPC resources,
- to create an HPC-based hardware and software environment
- to accelerate the theoretical investigations in nuclear physics by conducting massively-parallel computations,
- development and adaptation of software for the NICA mega-project on the new computing architectures,
- to prepare IT specialists in all the required directions.

THANK YOU FOR YOUR ATTENTION !