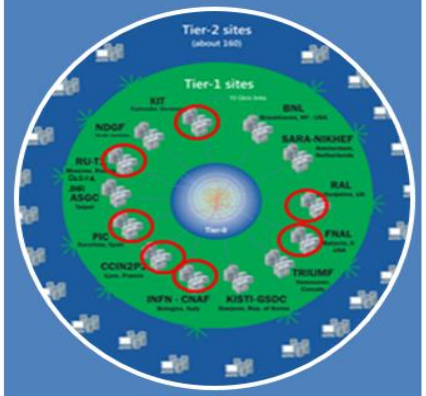
Modernized supercomputer “Govorun”
What's new for BM@N

K. Gertsenberger
VBLHEP, Joint Institute for Nuclear Research

Multifunctional Information and Computing Complex



Grid-Tier1:
9200 core,
8.3 PB disk,
11 PB tape

Grid-Tier2 CICC:
4128 core,
2.7 PB disk

Cloud:
1572 CPU,
8.142 TB RAM
1.1 PB disk

HPC Govorun:
Peak ~0.5 Pflops

HybriLIT:
Peak ~142 Tflops

Heterogeneous platform HybriLIT before Upgrade

Unified software and information environment

Education and testing polygon HybriLIT

- Intel Xeon
- Intel Xeon Phi
- NVIDIA Tesla K20
- NVIDIA Tesla K40
- NVIDIA Tesla K80
- 10 computation nodes

Supercomputer GOVORUN

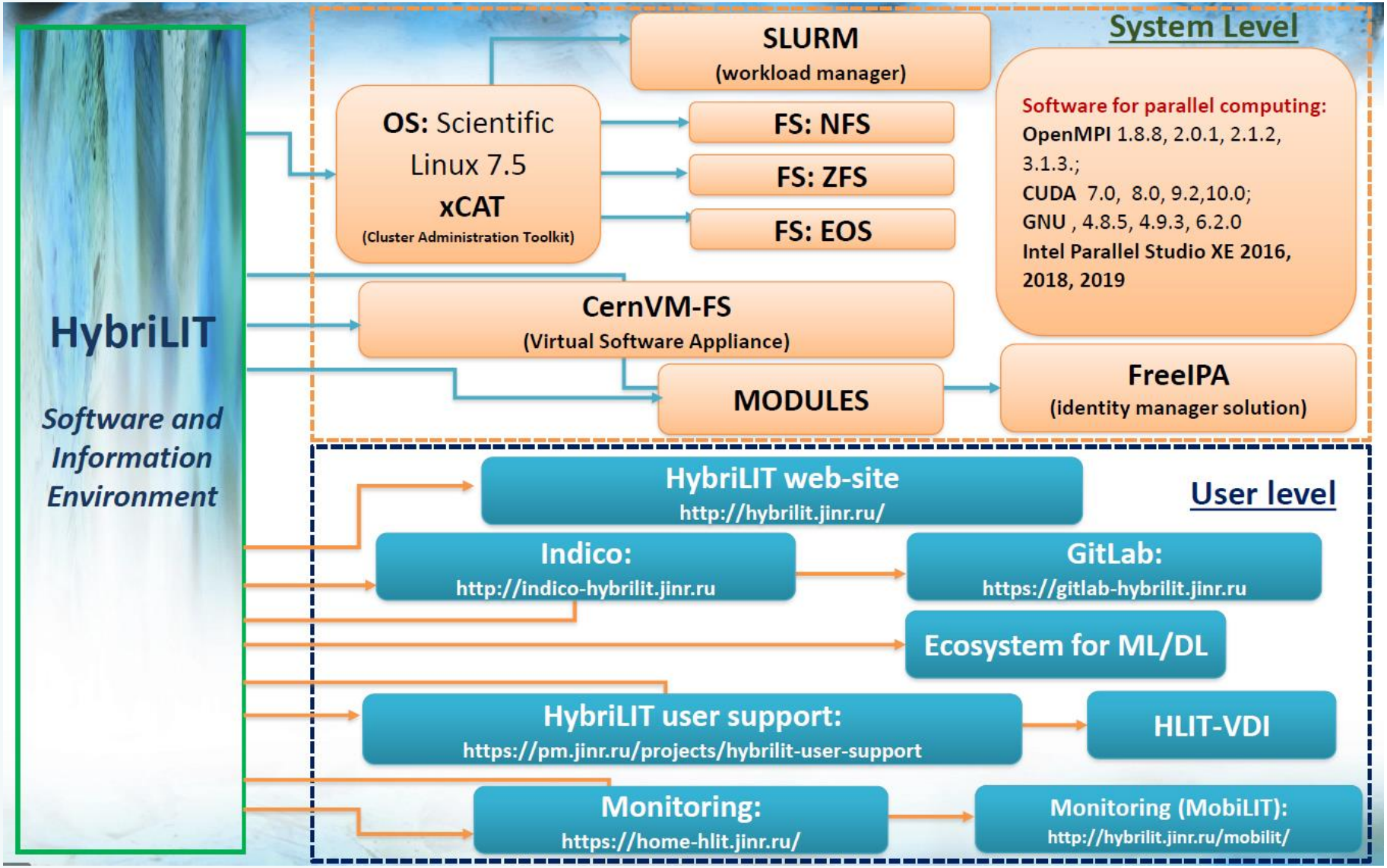
- 40 nodes — CPU-component — Intel Skylake
- 21 nodes — CPU-component — Intel Xeon Phi (KNL)
- 5 nodes — GPU-component — GPU DGX-1 Volta (NVIDIA Tesla V100)

Skylake
1440 cores
KNL
1512 cores

Volta
205 000 CUDA cores
26 000 Tensor cores
K40, K80
78 000 CUDA cores

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

HybriLIT ecoSystem



Govorun Monitoring Systems

				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 %		
11	n02p011	72	61.79	0 %	100 %	0 %		
12	n02p012	72	0	0 %	0 %	0 %		
13	n02p013	72	72.07	0.1 %	100 %	0 %		
14	n02p014	72	72.12	0.1 %	99.9 %	0 %		

Stat-hlit:

https://home-hlit.jinr.ru

litMon:

https://litmon.jinr.ru

Supercomputer Govorun Upgrade

- **535TFLOPS** пиковой производительности - **#10** в Top50
- Программно-определяема архитектура системы
- **#1** в производитель-ти систем хранения в России **>300GB/s**
- Масштабируемое решение **Storage-on-demand**
- Многоуровневая система хранения для максимальной эфф-ти
- Охлаждение горячей водой (compute, storage, interconnect)
- Наиболее энергоэффективный центр в России (**PUE = 1,027**)

Компоненты:

Узлы на Intel® Xeon® Scalable gen 2:

- Пиковая производительность – **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) 2TB
- RAM – 192 GB DDR4 2933 ГГц
- Intel® Omni-Path 100 Gbit/s
- 48-port Intel® Omni-Path Edge Switch 100 Series со 100% жидкостным охлаждением

Hyperconverged:

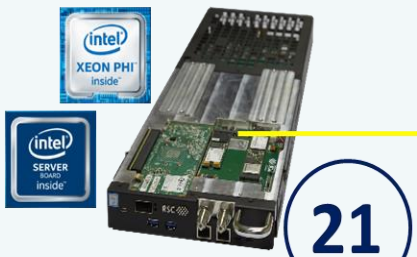
- 18 узлов с 12-ю NVMe SSD слотами
- 4 узла Optane с 3,4TB IMDT памяти
- 12 узлов OSS с NVMe SSD – **256TB**
- 2 узла MDS с 12-ю **Optane 375GB**
- ПФС Lustre как основная опция
- Storage-on-Demand с **RSC BasIS** на узлах кластера

Стек ПО “RSC БазИС”

Intel® Xeon Phi™ nodes:

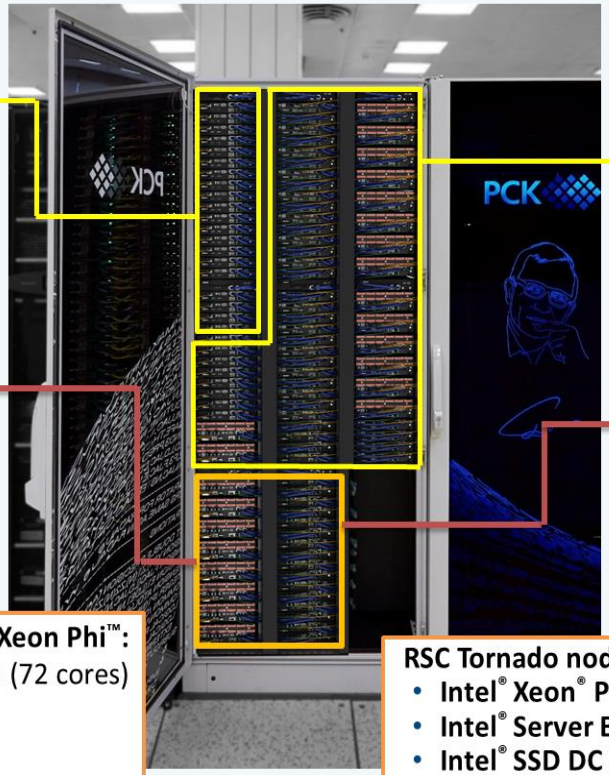
- Пиковая производительность – **72,576** ТФЛОПС
- Intel® Xeon Phi™ 7190 CPUs (72 cores)
- Intel® Server Board S7200AP
- Intel® SSD DC S3520 (SATA, M.2)
- RAM – 96 GB DDR4 2400 ГГц
- Intel® Omni-Path 100 Гбит/с
- 48-port Intel® Omni-Path Edge Switch 100 Series 100% liquid cooling

Supercomputer Govorun: CPU component upgrade



21

14



88

4

Узлы с большой памятью

intel OPTANE

Узлы для создания различных быстрых ПФС (Lustre, DAOS и др.)

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 gen 2:

- **Intel® Xeon® Platinum 8268** processors (24 cores)
- **Intel® Server Board S2600BP**
- **Intel® SSD DC S4510** (SATA, M.2),
  2x Intel® SSD DC P4511 (NVMe, M.2) 2TB
- 192GB DDR4 2933 GHz RAM
- **Intel® Omni-Path 100 Gb/s adapter**

Heterogeneous platform HybriLIT after Upgrade

Storage System	NFS/ZFS, EOS **ZFS, Lustre**
Software System	CVMFS *module* system
Batch System	SLURM
Processor Cores	**8 448 + 6 048**
Web-site	http://hybrilit.jinr.ru
Notifications	E-Mail
Monitoring	https://home-hlit.jinr.ru
Support System	https://pm.jinr.ru/projects/hybrilit-user-support

HybriLIT platform (HPC Govorun)
hydra.jinr.ru
(LIT, b.134)

OS: CERN CentOS 7
Exp. software: CVMFS, **Modules**

ZFS 200 TB,
Fast Storage on Lustre 352 TB_{ssd}

SLURM: 8 448 (Xeon cores) + **6 048** (Xeon Phi cores) + **40 GPU** NVidia Tesla V

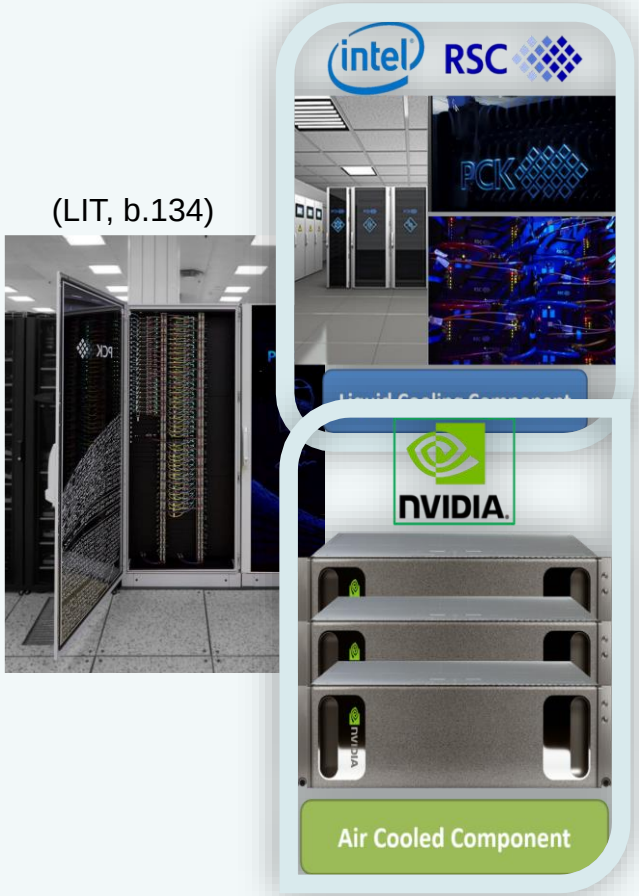
All external packages for BmnRoot were installed & configured.

Automatic BmnRoot deployment on CVMFS with GIT CI was implemented.

HybriLIT Platform for the BM@N experiment

OS: CERN CentOS 7

(LIT, b.134)



Cluster Administrator:
HybriLIT team

Storage

EOS:
for users: /eos/hybrilit.jinr.ru/user/
scratch: /eos/hybrilit.jinr.ru/scratch, /run/user/$UID
/eos/eos.jinr.ru → MICC EOS (…/nica/bmn/[sim.exp]

ZFS: /zfs/store7.hydra.local (200 TB, temporary)
Lustre: 30 TB SSD, ultra fast, temporary

Software

CVMFS: distributed software FS
export MODULEPATH="/cvmfs/hybrilit.jinr.ru/sw/slc7_x86-64/modulefiles:/cvmfs/hybrilit.jinr.ru/sw/slc7_x86-64/NICA/modulefiles"
module avail – print all modules
FairSoft & FairRoot: **module add FairRoot/v18.2.0**

Computing

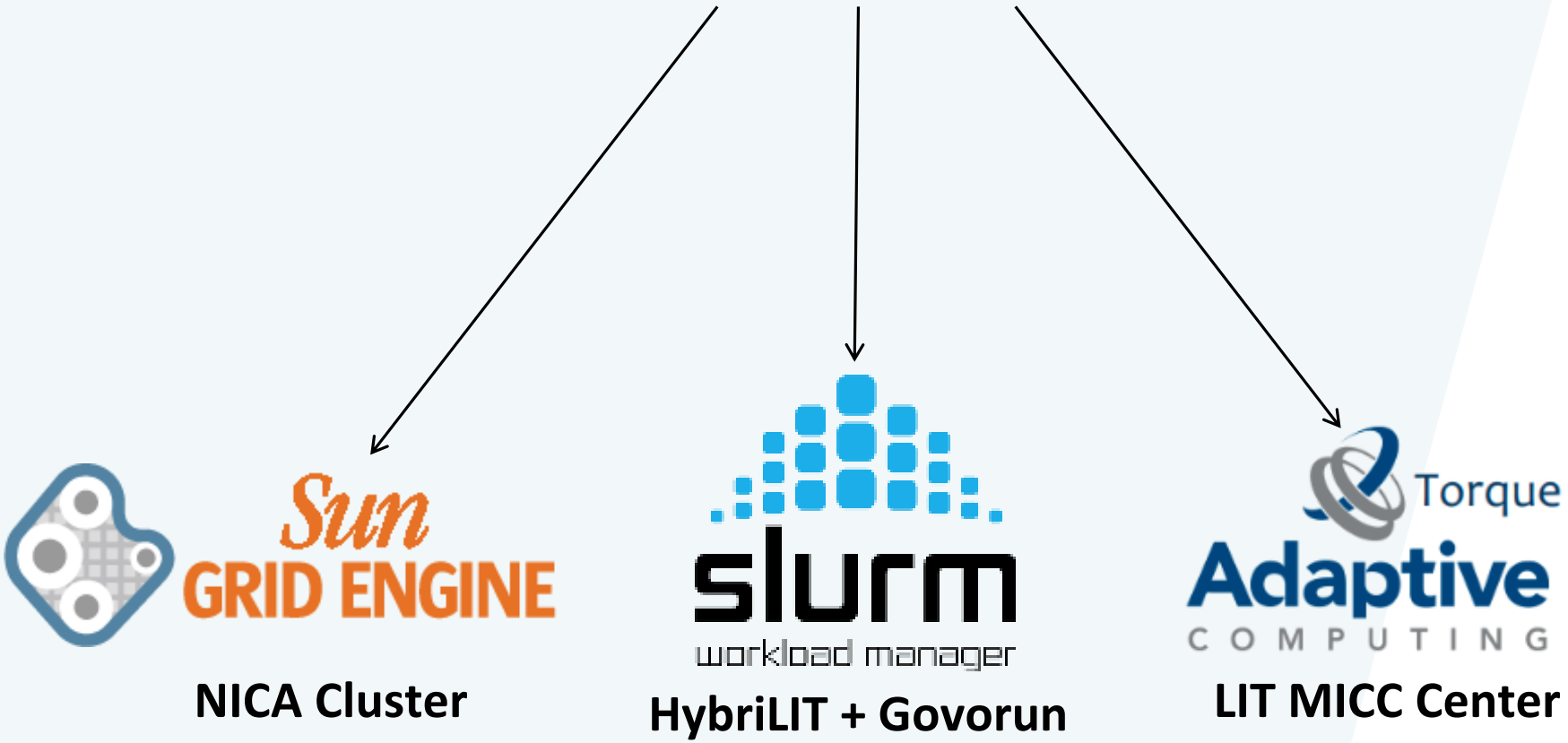
Batch System: **SLURM**
module add GVR/v1.0-1 → SuperComputer Govorun
Special queue (queue 'bmn'): 384 log. cores
Intel Xeon Platinum (queue 'cascadelake'): **5 664**
Intel Xeon Phi (queue 'knl'): **6048 log. cores**
NVidia Tesla V (queue 'dgx'): **40 GPU cards**

Registration http://hlit.jinr.ru/for_users/registration/

NICA-Scheduler: from SGE to SLURM

NICA-Scheduler

$ nica-scheduler my_job.xml



NICA Cluster | **HybriLIT + Govorun** | **LIT MICC Center**

The NICA-Scheduler Guide: http://bmn.jinr.ru/nica-scheduler/

Mass production for BM@N Run 7

MICC Complex
1 PB available
BM@N experimental data

raw_run7.data / run_digi.root
xrdcp
run_digi.root / bmndst.root

UltraFast Storage
200 TBssd available

Supercomputer GOVORUN

NICA-Scheduler
$ nica-scheduler
bmn_raw_run7_govorun.xml


```
<job name="convert_bmn_raw">
 <macro path="~/bmnroot/macro/raw/BmnDataToRoot.C">
  <file input="/eos/nica/bmn/exp/raw/run7/*">
   <put command="xrdcp" path="/lustre/stor/${file_name_with_ext}"/>
   <get command="xrdcp" path="/lustre/stor/bmn_run${last_number}_digi.root"
output="/eos/nica/bmn/exp/digi/run7/bmn_run${last_number}_digi.root"/>
  </file>
 </macro>
 <run mode="global" count="100" config="~/bmnroot/build/config.sh"
work_dir="/lustre/stor"/>
</job>
```

Intel Xeon Platinum (queue ‘bmn’): 384 cores

HybriLIT platform: Application and Report

http://hlit.jinr.ru/en/about_govorun_eng/registration-at-the-govorun-supercomputer/

1. Application Form from the BM@N Collaboration once per year
2. Reporting Form from the BM@N Collaboration before the Application

http://hlit.jinr.ru/en/heterogeneous-cluster-hybrilit/users_publications_eng/

Users Publications

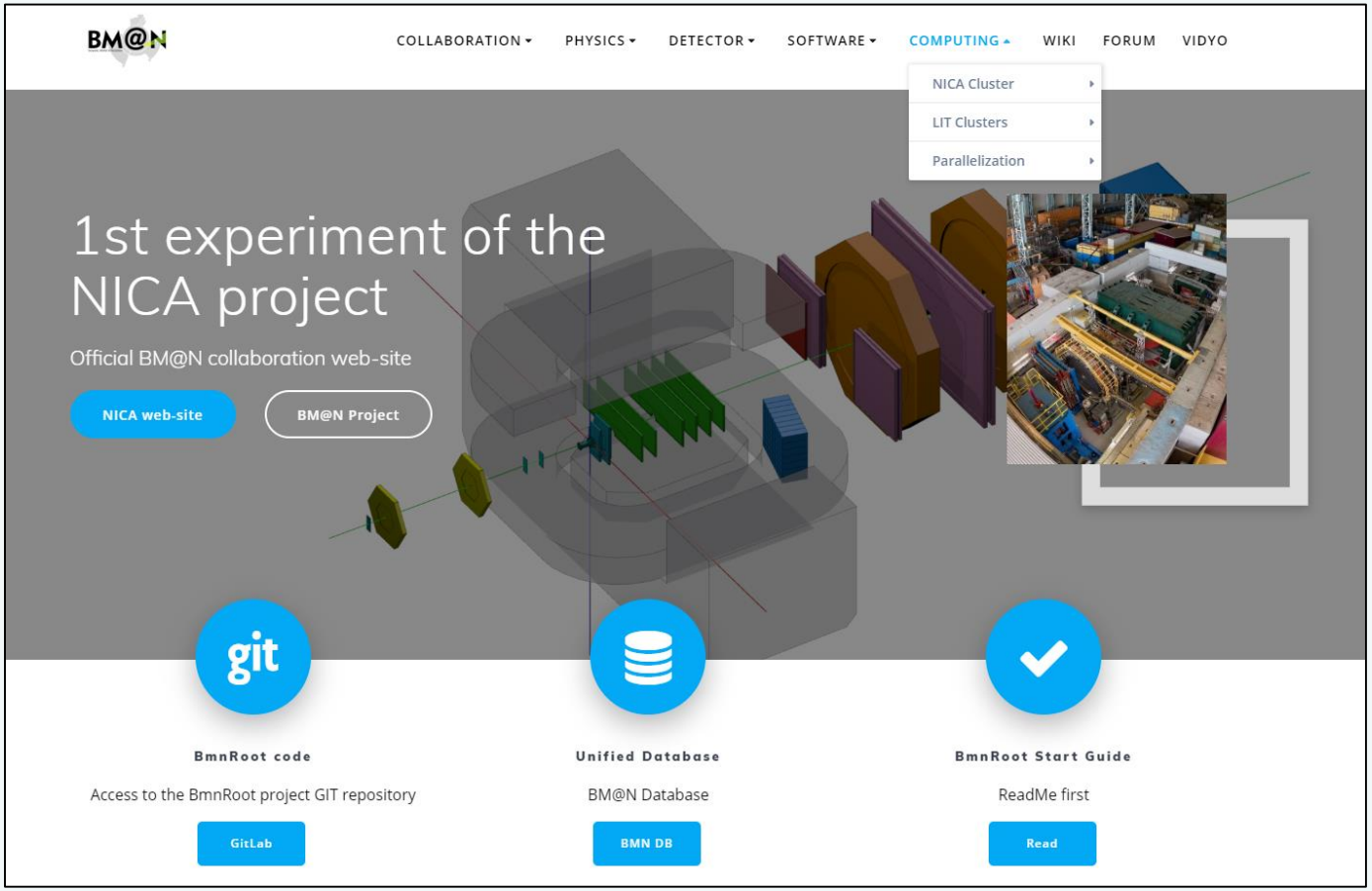
Authors should make references to the use of the resources of the heterogeneous platform in the following way:

Computations were held on the basis of the HybriLIT heterogeneous computing platform (LIT, JINR).

Please also use this link with the description of the heterogeneous platform:

Gh. Adam, M. Bashashin, D. Belyakov, M. Kirakosyan, M. Matveev, D. Podgainy, T. Sapozhnikova, O. Streltsova, Sh. Torosyan, M. Vala, L. Valova, A. Vorontsov, T. Zaikina, E. Zemlyanaya, M. Zuev. IT-ecosystem of the HybriLIT heterogeneous platform for high-performance computing and training of IT-specialists. Selected Papers of the 8th International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018), Dubna, Russia, September 10-14, 2018, CEUR-WS.org/Vol-2267"

Computing Section on the BM@N Web-site



- ✓ *Information*
- ✓ *Documents*
- ✓ *Software*
- ✓ *Databases*
- ✓ *Computing Section (NICA Cluster, MICC Complex, **HybriLIT & Govorun**)*
- ✓ *Tests dashboard*
- ✓ *Guides*
- ✓ *Forum*
- ✓ *Vidyo*
- ✓ *BM@N Mail-lists (updates, errors…)*
- ✓ *etc.*

Thank you for attention!

Registration on the HybriLIT platform

http://hybrilit.jinr.ru/en/registration_eng/

1. Fill the Registration Form
2. The Registration Form should be sent to me (gertsen@jinr.ru).
3. The printed and signed registration form should be brought to the room № 323, LIT.
4. The confirmation of registration will be sent to the e-mail address specified in the registration form; the letter will contain registration data, login and a temporary password that should be changed during 7 days.

Current Computing Clusters for NICA

NICA Cluster
ncx[101-106].jinr.ru
(LHEP, b.215, b.216)

OS: Scientific Linux 7
Exp. software: Local

EOS: 3.3 PB (replicated)
GlusterFS: 320 TB (*replicated*)

Sun Grid Engine: 3 096
(Intel Xeon cores)

MICC Tier1/2 Center
lx[pub,mpd-ui].jinr.ru
(LIT, b.134)

OS: Scientific Linux 6
Exp. software: CVMFS

EOS: 4 PB

Torque/Maui:
Tier2: **~300** (Xeon cores)
Tier1: ~**600** (Xeon cores)

HybriLIT platform (HPC Govorun)
hydra.jinr.ru
(LIT, b.134)

OS: CERN CentOS 7
Exp. software: CVMFS, **Modules**

ZFS 200 TB**,**
Fast Storage on Lustre 352 TB_{ssd}

SLURM: 8 448 (Xeon cores) + **6 048**
(Xeon Phi cores) + **40 GPU** NVidia Tesla V

All external packages for BmnRoot were installed & configured.

Automatic BmnRoot deployment on CVMFS with GIT CI was implemented.