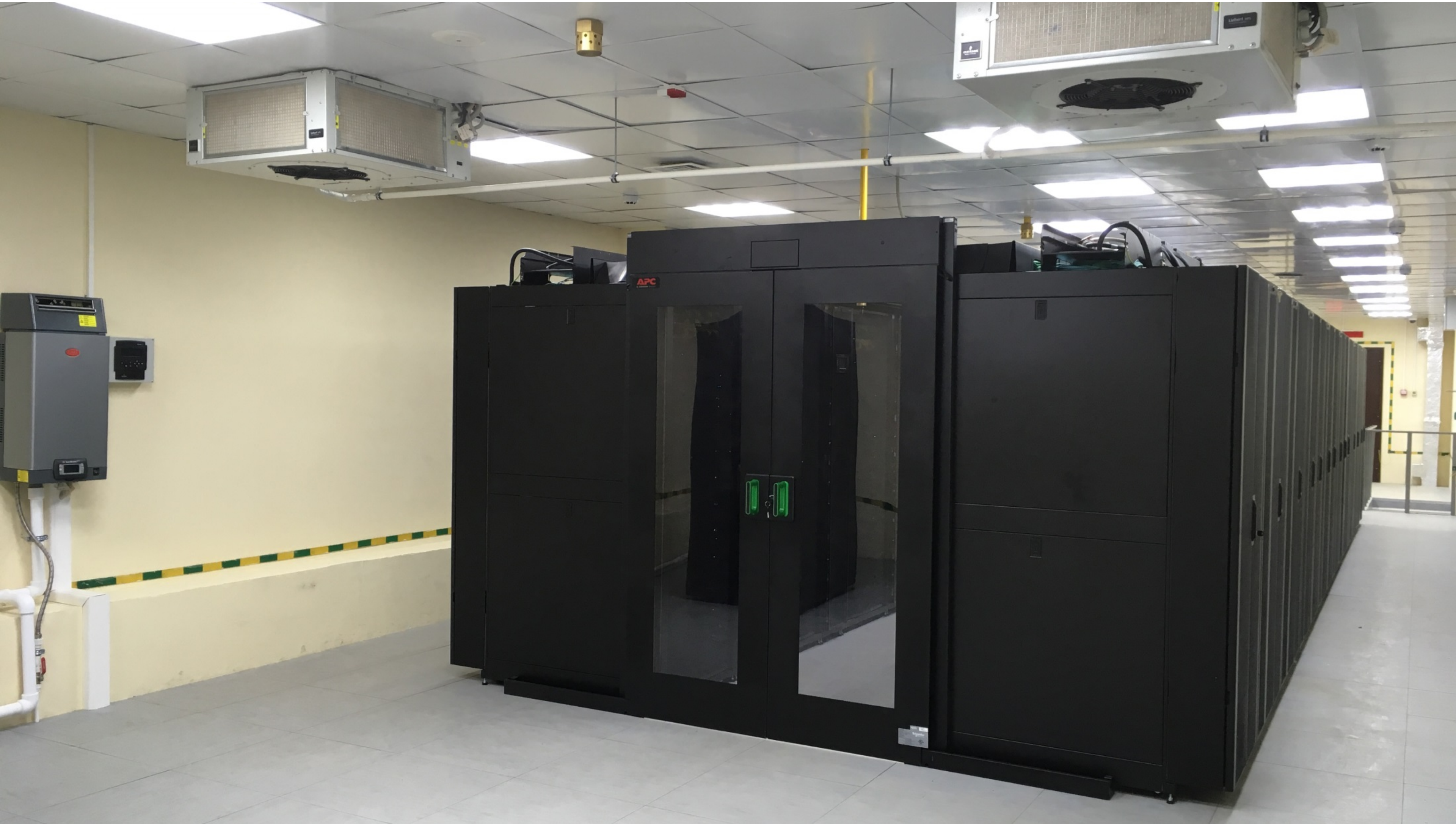
NATIONAL RESEARCH CENTRE
«KURCHATOV INSTITUTE»

PETERSBURG NUCLEAR PHYSICS INSTITUTE

PIK Computing Centre

Andrey Kiryanov

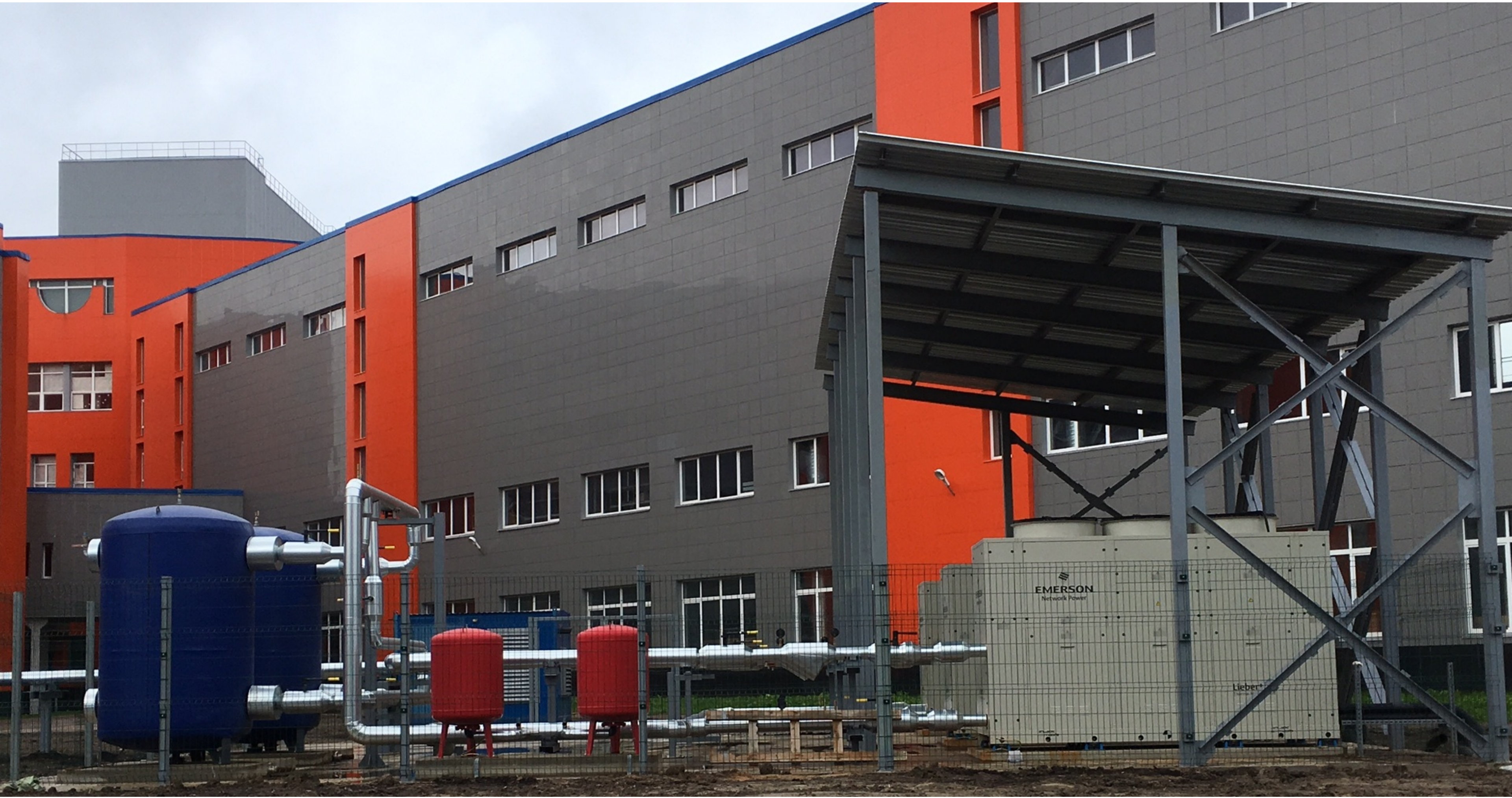
CACS with HACS



Power and cooling infrastructure

- Both main power and cooling can sustain up to 300 kVA of load
  - With maximum load the computing equipment can run on batteries for about 15 minutes, which is enough for a graceful shutdown
  - Cooling system has its own UPS and runs on batteries for about 30 minutes
  - Diesel powers the cooling system pumps if everything else fails
  - No guaranteed power in case of a major failure, but power line is redundant
- Currently the Computing Centre equipment is worth 120 kVA of load
  - Roughly 40 minutes of runtime on batteries
  - Over 50% of rack space is unused (12 racks populated out of 28)
  - Network infrastructure is designed for full capacity (except InfiniBand)
  - We can move in new servers without any modifications to the infrastructure

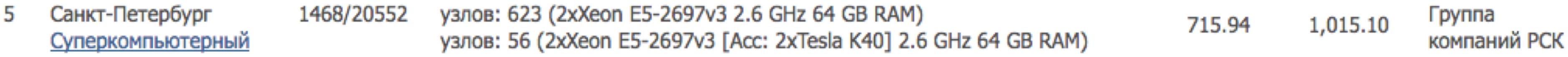
Chillers and refrigerant tanks



Computing equipment

- Peak theoretical performance is ~362 Tflops
- Real LINPACK results:
  - ~200 Tflops on Xeon CPUs (no AVX-512), effectiveness ~80%
  - ~68 Tflops on Xeon Phi (KNL) CPUs (AVX-512), effectiveness ~50%
- Computing equipment:
  - 160 nodes with Xeon CPUs: 2.4 GHz, 28 cores, 128 GB RAM per node (4.5 GB RAM per core) – **4 480 cores**
  - 40 nodes with Xeon Phi (KNL) CPUs: 1.4 GHz, 68 cores (272 virtual), 96 GB RAM per node – **10 880 virtual cores**
  - 16 nodes with Xeon CPUs: 2.4 GHz, 28 cores, 1 TB RAM + 1.6 TB NVMe SSD – **448 cores**
  - 2 nodes with Xeon CPUs: 2.4 GHz, 28 cores, 1.5 TB RAM – **56 cores**

http://top50.supercomputers.ru/



5	Санкт-Петербург Суперкомпьютерный	1468/20552	узлов: 623 (2xXeon E5-2697v3 2.6 GHz 64 GB RAM) узлов: 56 (2xXeon E5-2697v3 [Acc: 2xTesla K40] 2.6 GHz 64 GB RAM)	715.94	1,015.10	Группа компаний РСК

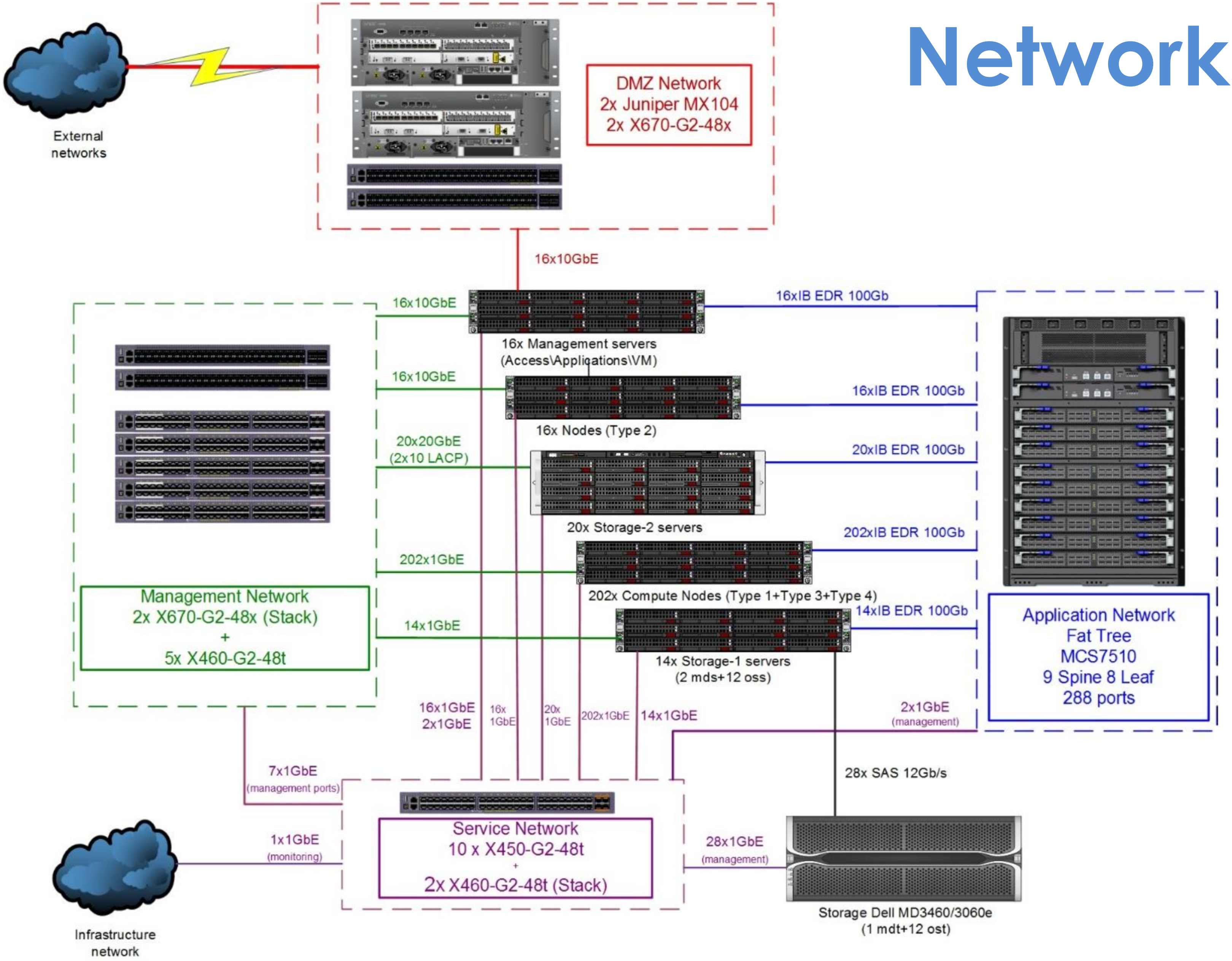
Storage

- Lustre with 2.9 PB of raw disks (~2.3 PB of visible storage + 29 TB of metadata)
  - 2.10.4 LTS release + Mellanox OFED
  - Dense DELL disk shelves connected via SAS
  - Connection through 100 Gbps InfiniBand
- Ceph with 2.5 PB of raw disks (two full racks)
  - 13.2.1 Mimic release
  - Standard Supermicro disk servers
  - Connection through 2x10 Gbps Ethernet
  - InfiniBand is also available, currently used for replication

Lustre disk shelves and servers

InfiniBand switch and Ceph servers

Networks



External networks

DMZ Network
2x Juniper MX104
2x X670-G2-48x

16x10GbE

16x10GbE

16xIB EDR 100Gb

16x Management servers
(Access\Applications\VM)

16x10GbE

16xIB EDR 100Gb

16x Nodes (Type 2)

20x20GbE
(2x10 LACP)

20xIB EDR 100Gb

20x Storage-2 servers

202xIB EDR 100Gb

202x1GbE

202x Compute Nodes (Type 1+Type 3+Type 4)

14xIB EDR 100Gb

14x1GbE

14x Storage-1 servers
(2 mds+12 oss)

Management Network
2x X670-G2-48x (Stack)
+
5x X460-G2-48t

Application Network
Fat Tree
MCS7510
9 Spine 8 Leaf
288 ports

16x1GbE
2x1GbE

16x
1GbE

20x
1GbE

202x1GbE

14x1GbE

2x1GbE
(management)

7x1GbE
(management ports)

28x SAS 12Gb/s

1x1GbE
(monitoring)

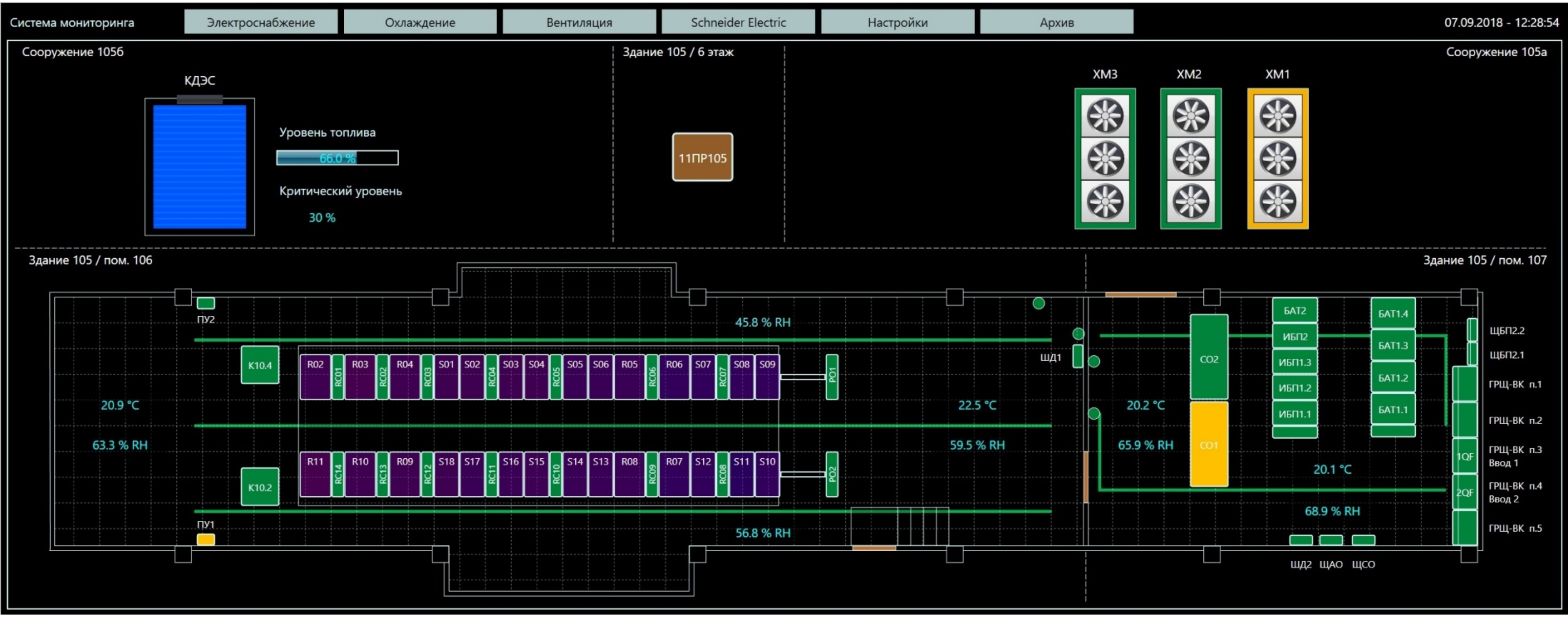
Service Network
10 x X450-G2-48t
+
2x X460-G2-48t (Stack)

28x1GbE
(management)

Storage Dell MD3460/3060e
(1 mdt+12 ost)

Infrastructure network

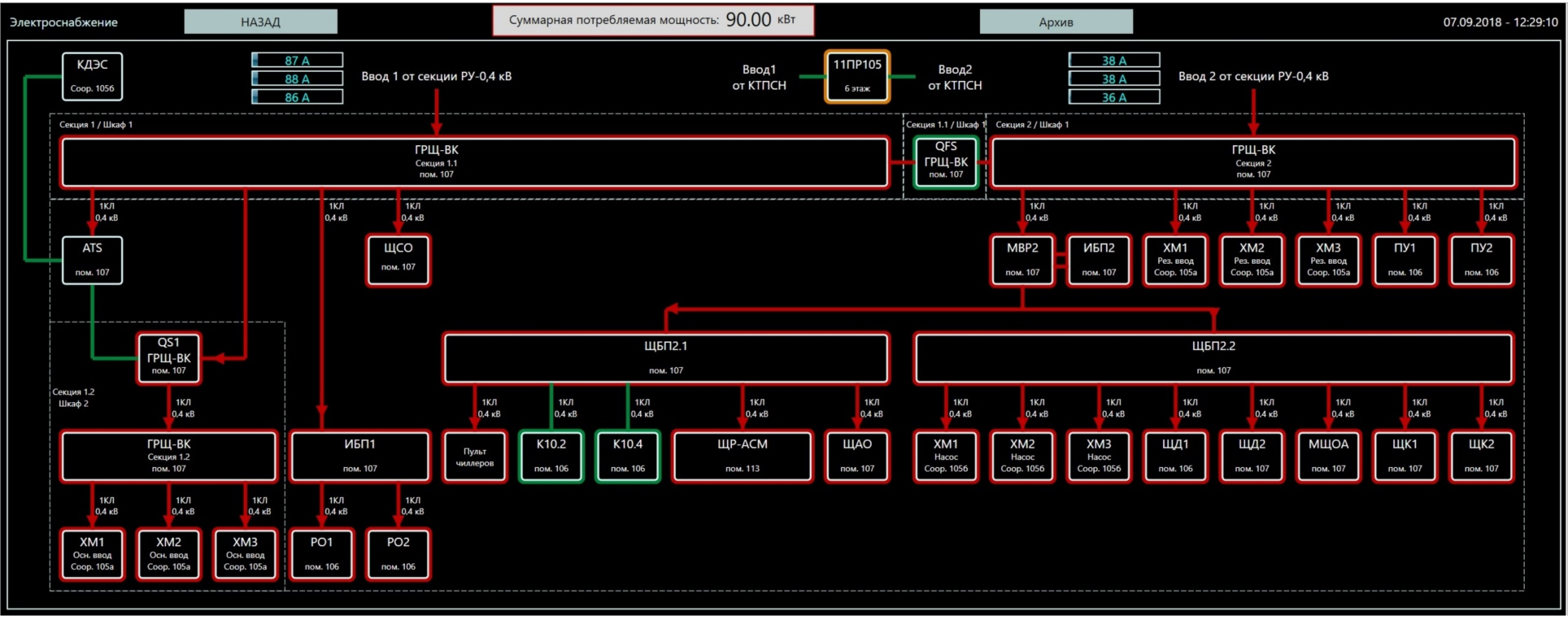
Engineering Systems Monitoring



Integral system covering mains, UPS, cooling, climate, leak sensors, chillers and diesel

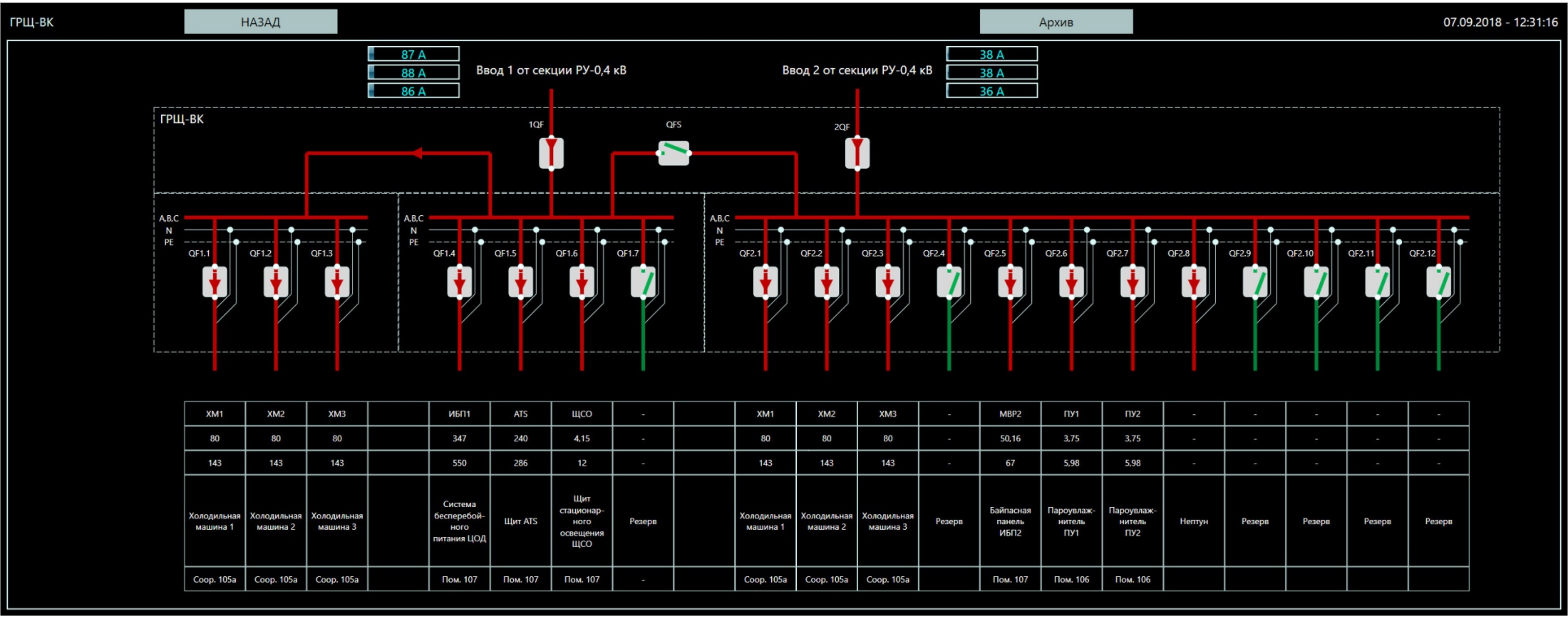
Main screen with a bird's-eye view

Engineering Systems Monitoring



Detailed mains monitoring

Engineering Systems Monitoring



ХМ1	ХМ2	ХМ3		ИБП1	ATS	ЩСО	-		ХМ1	ХМ2	ХМ3	-	МВР2	ПУ1	ПУ2	-	-	-	-	-
80	80	80		347	240	4,15	-		80	80	80	-	50,16	3,75	3,75	-	-	-	-	-
143	143	143		550	286	12	-		143	143	143	-	67	5,98	5,98	-	-	-	-	-
Холодильная машина 1	Холодильная машина 2	Холодильная машина 3		Система бесперебойного питания ЦОД	Щит ATS	Щит стационарного освещения ЩСО	Резерв		Холодильная машина 1	Холодильная машина 2	Холодильная машина 3	Резерв	Байпасная панель ИБП2	Пароувлажнитель ПУ1	Пароувлажнитель ПУ2	Нептун	Резерв	Резерв	Резерв	Резерв
Соор. 105а	Соор. 105а	Соор. 105а		Пом. 107	Пом. 107	Пом. 107	-		Соор. 105а	Соор. 105а	Соор. 105а		Пом. 107	Пом. 106	Пом. 106					

Even more detailed monitoring of a single mains switchboard

Software

- OS: CentOS 7, 64-bit
- High availability: pacemaker
- Management: xCAT
- Auth: FreeIPA
- Batch system: Slurm
- Cluster monitoring: Ganglia, Nagios, Zabbix
- MPI: Open MPI, Intel MPI, Platform MPI
- Compilers: GCC, Intel Compiler
- Intrusion prevention: fail2ban
- Virtualization: KVM

User perspective

- Heterogeneous resources
- Different types of nodes are organized in distinct queues
  - ”Standard” CPUs
  - Big/Huge memory
  - KNL
- User data reside on Lustre
  - Home directory is accessible from all nodes
  - Common software and shared data in dedicated areas
  - Environment modules for switching between compilers and MPI flavours

High Availability

- Most critical services are running in HA mode
  - xCAT: two servers in master-slave mode
  - Lustre: every disk shelf is connected to two servers
    - Resources can be reallocated in real time, allowing for almost transparent Lustre server maintenance
  - Slurm: two servers in master-slave mode
  - MariaDB: Galera cluster with two servers
  - FreeIPA: two servers with replication
  - Ceph: multiple MON, MGR, MDS
- Automatic failover for all services except xCAT

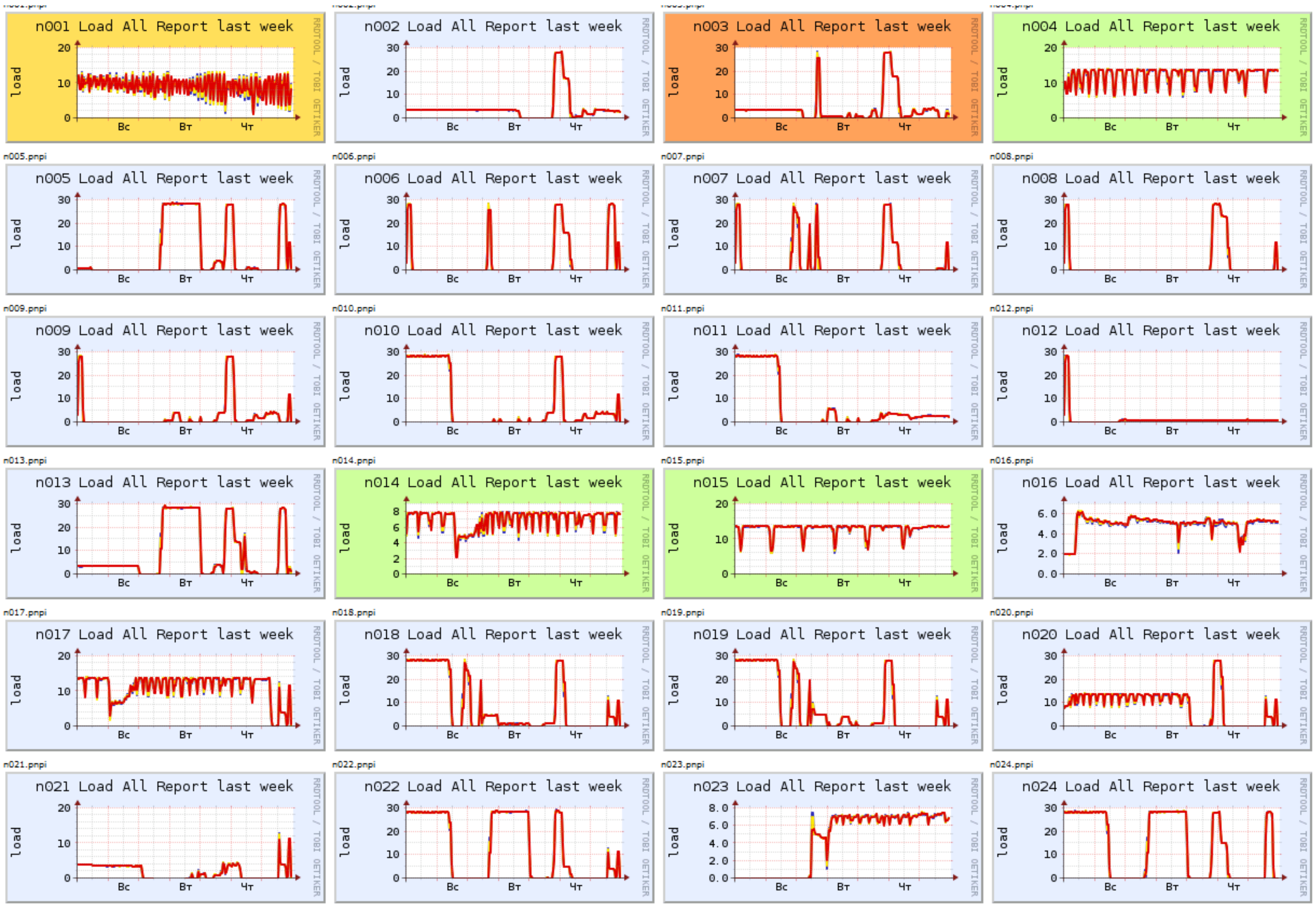
Intrusion prevention

- We use a pool of login nodes for user access
- Key-based SSH, no password authentication
- Primary firewall on a main router enforces static access policy
- Individual firewalls on all hosts with global addresses
- Custom-compiled version of fail2ban with IPv6 support
- Ban log on a shared filesystem
  - If two nodes ban the same IP it will propagate to the whole pool of nodes

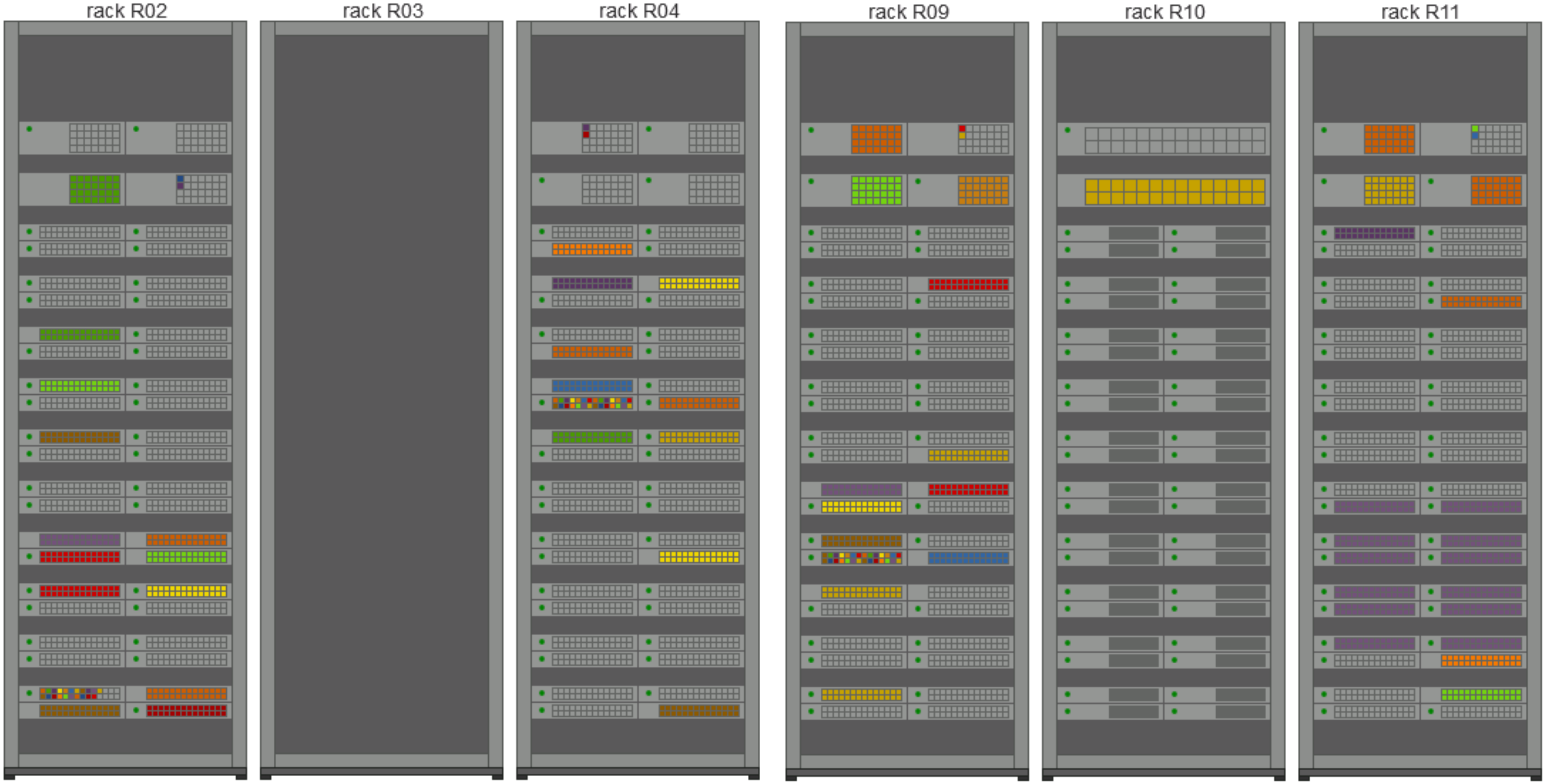
Virtual Infrastructure

- 14 servers with 256 GB RAM as hypervisors
- KVM-based VMs managed by xCAT
- Disk volumes on Ceph
- SR-IOV for InfiniBand and 10 Gbps Ethernet
  - Zero VM I/O overhead
  - Lustre over IB works seamlessly
  - Needed some xCAT modifications
  - Libvirtd allocates ethernet VFs automatically, but fails miserably with IB because of a longer MAC

Ganglia Monitoring



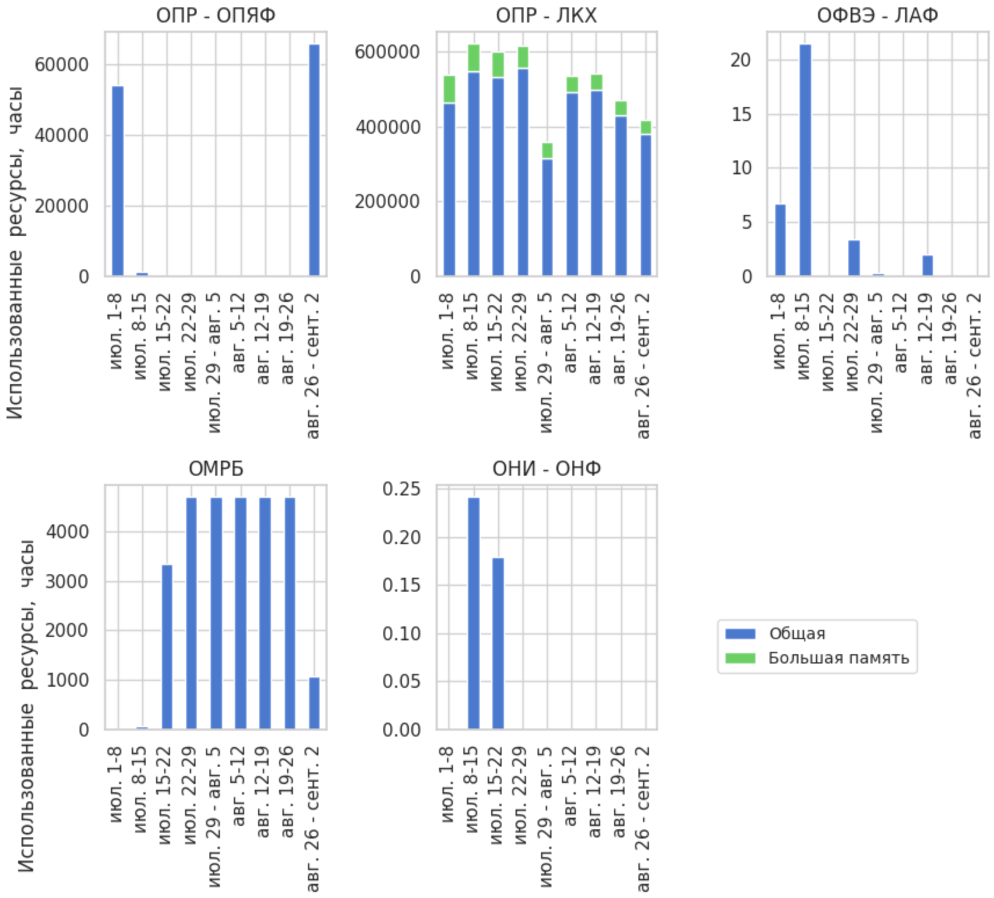
Slurm Monitoring

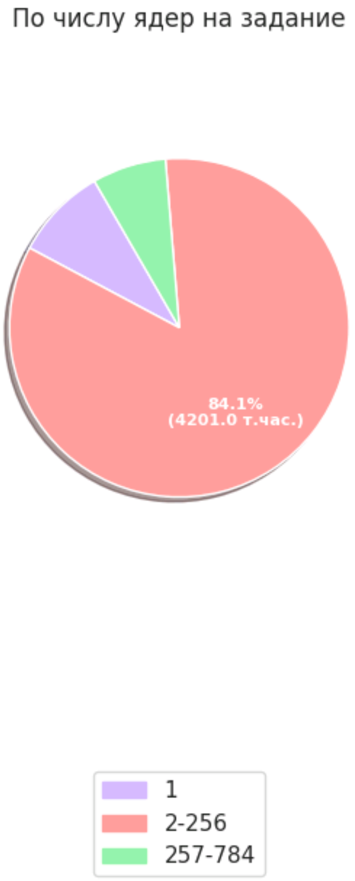


Per-core job distribution

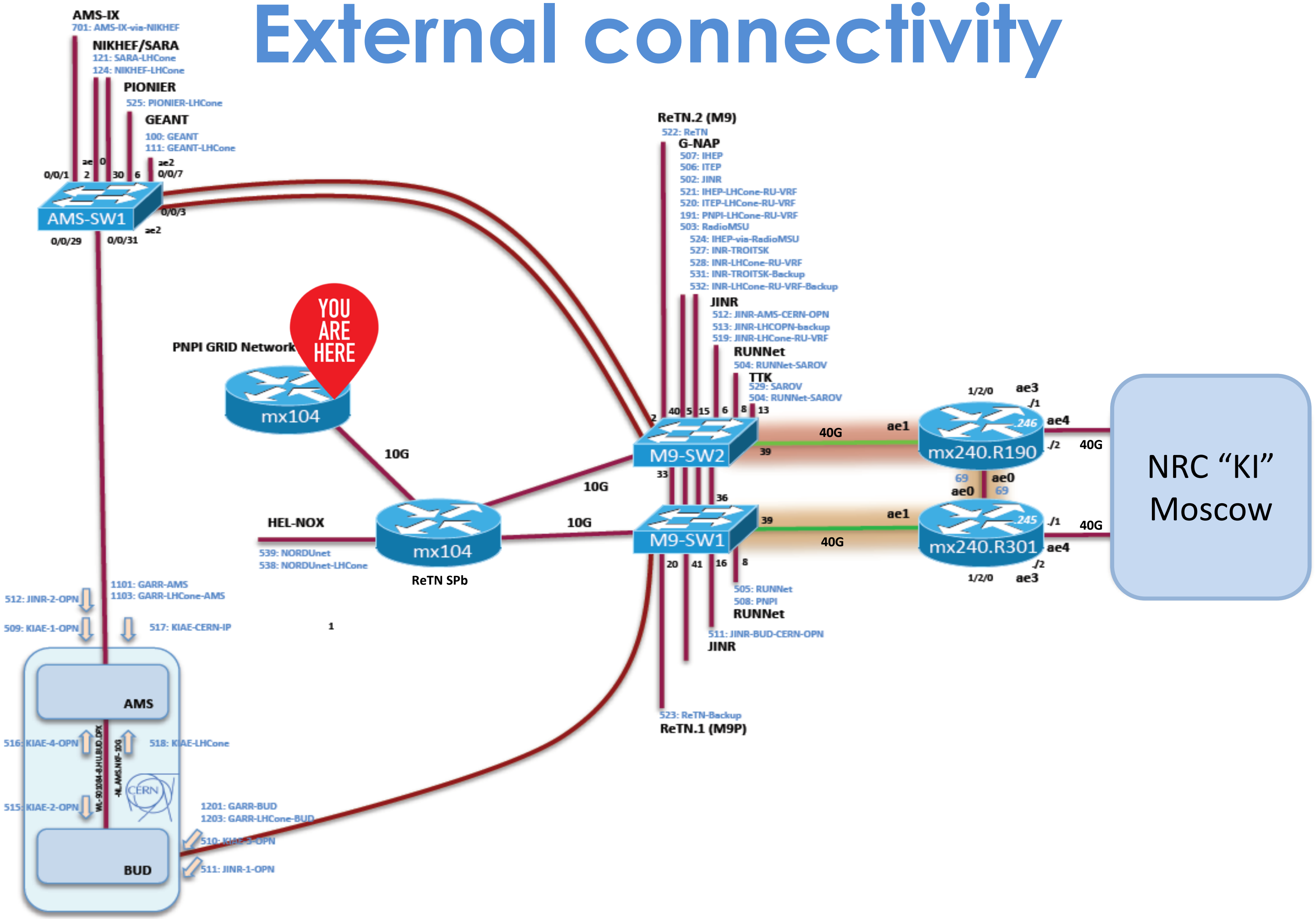
Reports

Динамика использования ресурсов коллективами





External connectivity



Challenges

- Different default fan modes on different servers
  - Firmware upgrades and manual tuning
  - IB card may overheat with low fan speed
- KNL performance issues
  - Firmware upgrades and memory mode tuning
- Mellanox OFED memory allocation problems
  - PR is still open with Mellanox
  - Seems to be fixed in 4.4-2 release
- Issues with Ceph EC pools
  - Solved by moving to Mimic and RHEL 7.5 kernel

Thank you!

APC