Joint Institute for Nuclear Research

JINR Computing Infrastructure for the NOvA Experiment

N. Balashov

07 April 2018

NOvA Computing Overview

- General purpose Virtual Machines (GPVM) at FNAL: code development and quick test jobs only
- Fermigrid: dedicated resources to run resource intensive workloads
- Open Science Grid (OSG) infrastructure: globally distributed computing resources that supplement the main Fermigrid quota

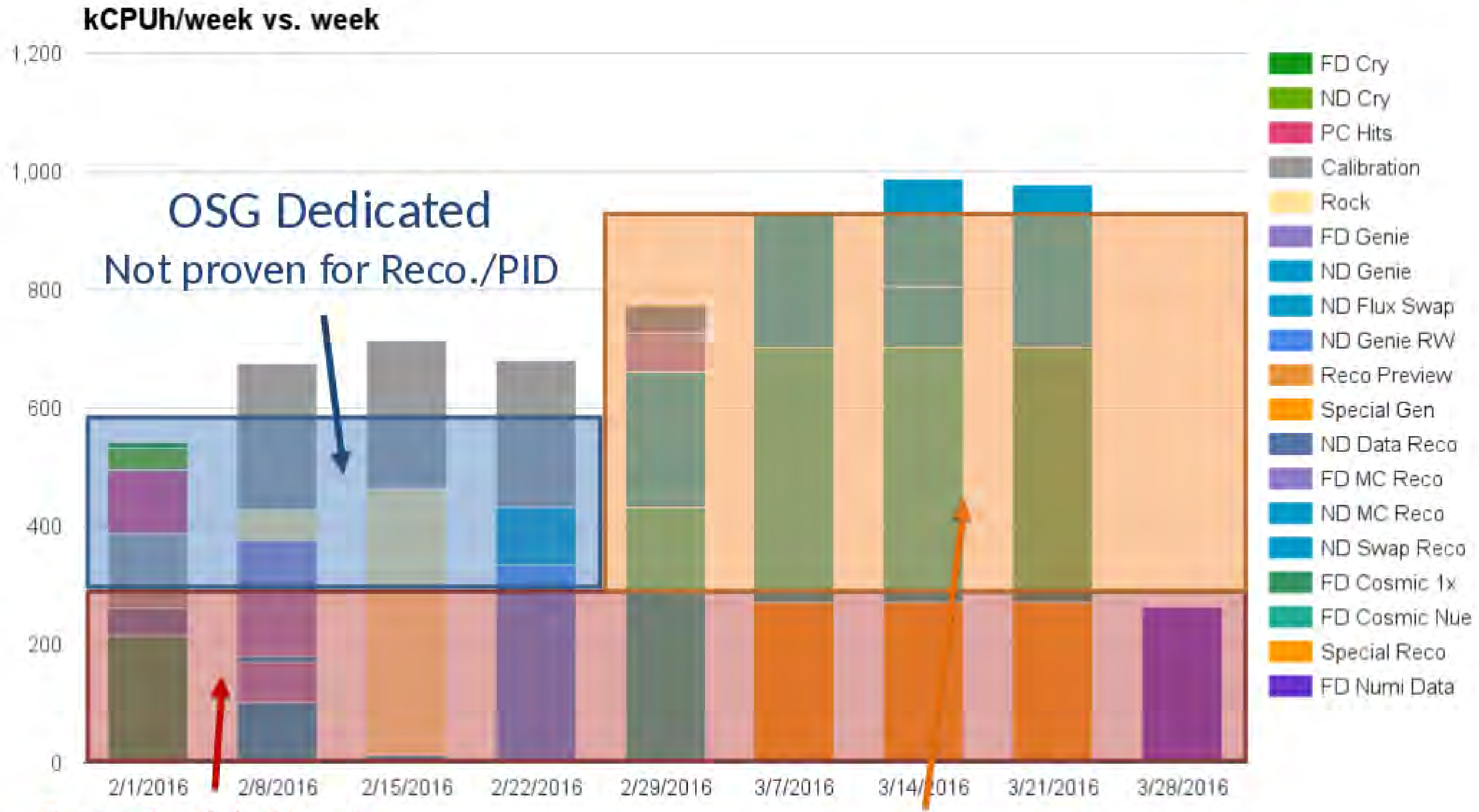
Interactive Virtual Machines

Initial Setup (2014)	Current Setup
• 4 machines: 4 cores, 8 GB of RAM and 100GB HDD each • OpenVZ virtualization • Manually installed software (Art Framwork, Cisco Anyconnect, etc.) • X2go to access GUI • JINR Kerberos authentication	• 6 machines: 4 cores, 8 GB of RAM and 210 GB HDD each • 20 TB NFS Storage • KVM virtualization • Software accessible via the CernVM-FS repository • Maximum similarity to the FNAL GPVMs
• Minimal virtualization overhead • No need to keep separate kernel for each container – less memory consumption	• No software limitations • Any OS, any kernel modules can be loaded from within the VM
• Modified Linux-kernel – linux-only, outdated version • Kernel modules need to be loaded on the host first • **No autofs** (cvmfs needs to be configured and mounted manually)	• Higher virtualization overhead

Interactive Virtual Machines

- Keep maximum similarity to the NOvA general purpose virtual machines (GPVM) at FNAL
- Interactive VMs are well-suited for the software development, evaluation and not too computationally intensive workloads
- Provide the most up-to-date NOvA environment
- Online 24/7 and accessible via ssh/X2go

NOvA CPU Request Spikes



OSG Dedicated

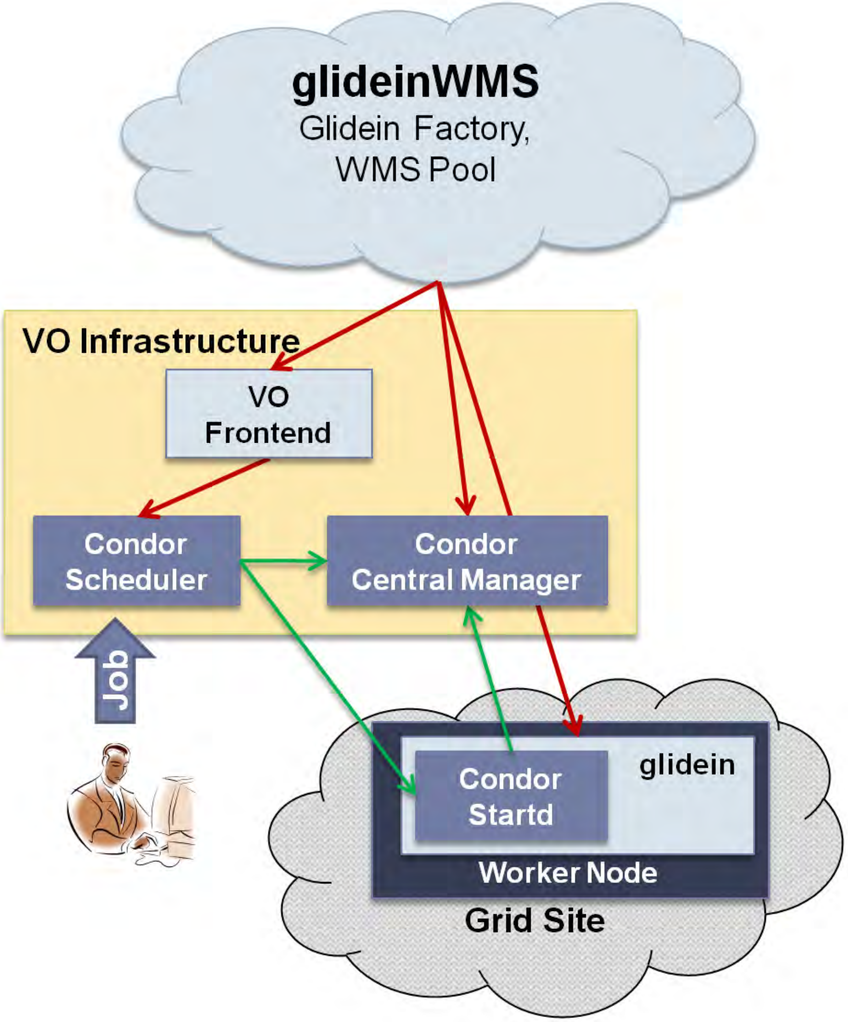
Not proven for Reco./PID

Fermigrid Quota
(2000 slots)

Additional request:
Triple allocation for ~1 month

NOvA Computing Infrastructure

- Extensive workloads are split in jobs which are processed in batches
- A Workload Management System controls the jobs – GlideinWMS at FNAL
- External resource providers are utilized via Open Science Grid (OSG) – American Grid infrastructure
- A Tier-2 batch cluster at JINR was connected to OSG to support NOvA
- A new virtual HTCondor-based cluster was created which was first dedicated to NOvA

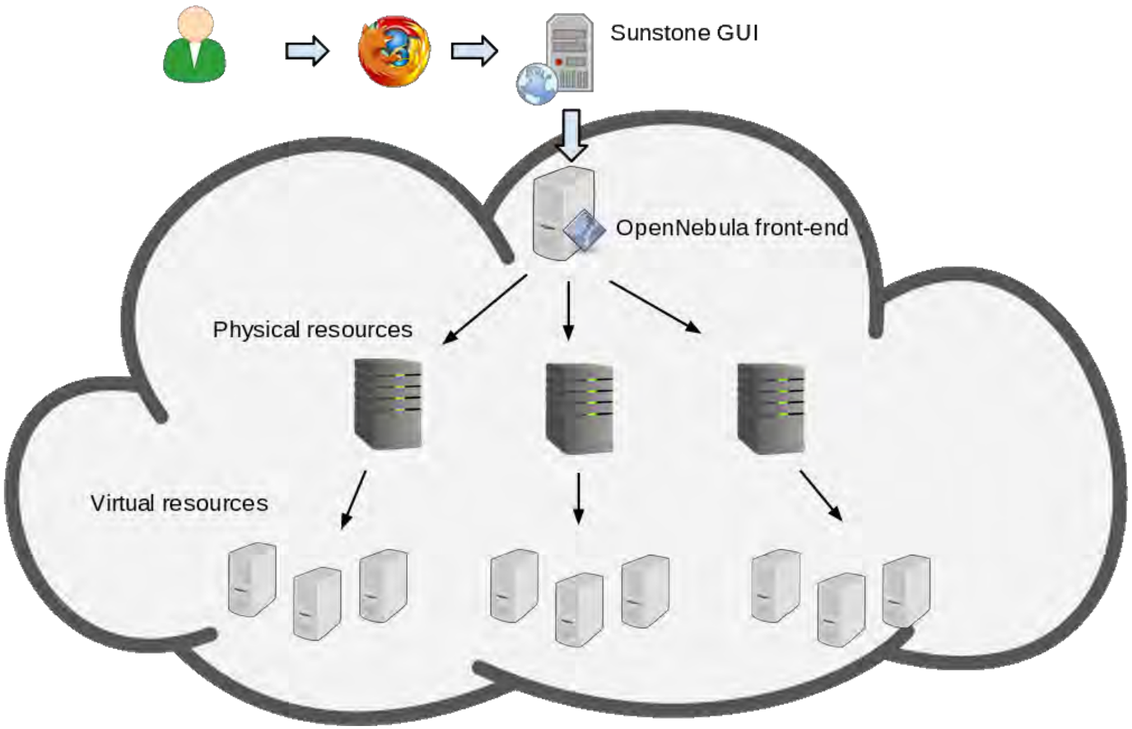


JINR Resources Extension

- We had two options at that moment:
  - add new servers to the existing Tier-2 cluster
  - put them in the Cloud service and install new virtualized infrastructure
- Opted out from extending Tier-2 in favor of the cloud for 3 main reasons:
  - Extremely slow configuration changes
  - Batch mode only
  - No enthusiasm from T2 team

Cloud Computing

- Technically clouds can be directly attached to the GlideinWMS, but that would require significant efforts from FNAL computing people
- We had to install a new fully virtualized batch cluster

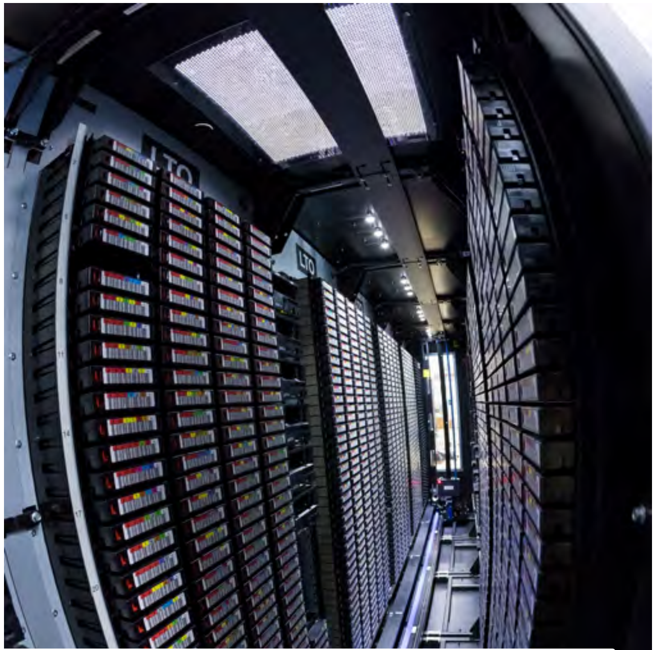


NOvA Computing Nodes

- 25 additional servers were purchased and added to the cloud in 2015-2020
- First servers had different configuration and were modernized: RAM and 10Gb network adapters were added

Platform	CPU	RAM	Disk	Network Interfaces
5 x Dell PowerEdge R430	2xE5-2650v3	128 GB	2x2TB NL SAS	4x1Gb and 2x10Gb Ethernet
4 x Dell PowerEdge R430	2xE5-2650v3	128 GB	4x4TB NL SAS	4x1Gb and 2x10Gb Ethernet
5 x Dell PowerEdge R430	2xE5-2650v4	128 GB	2x4TB NL SAS	4x1Gb and 2x10Gb Ethernet
5 x Dell PowerEdge R440	2xSilver 4116	128 GB	2x120GB SSD	4x1Gb and 2x10Gb Ethernet
5 x Dell PowerEdge R440	2xSilver 4214	128 GB	2x120GB SSD	4x1Gb and 2x10Gb Ethernet

JINR Datacenter

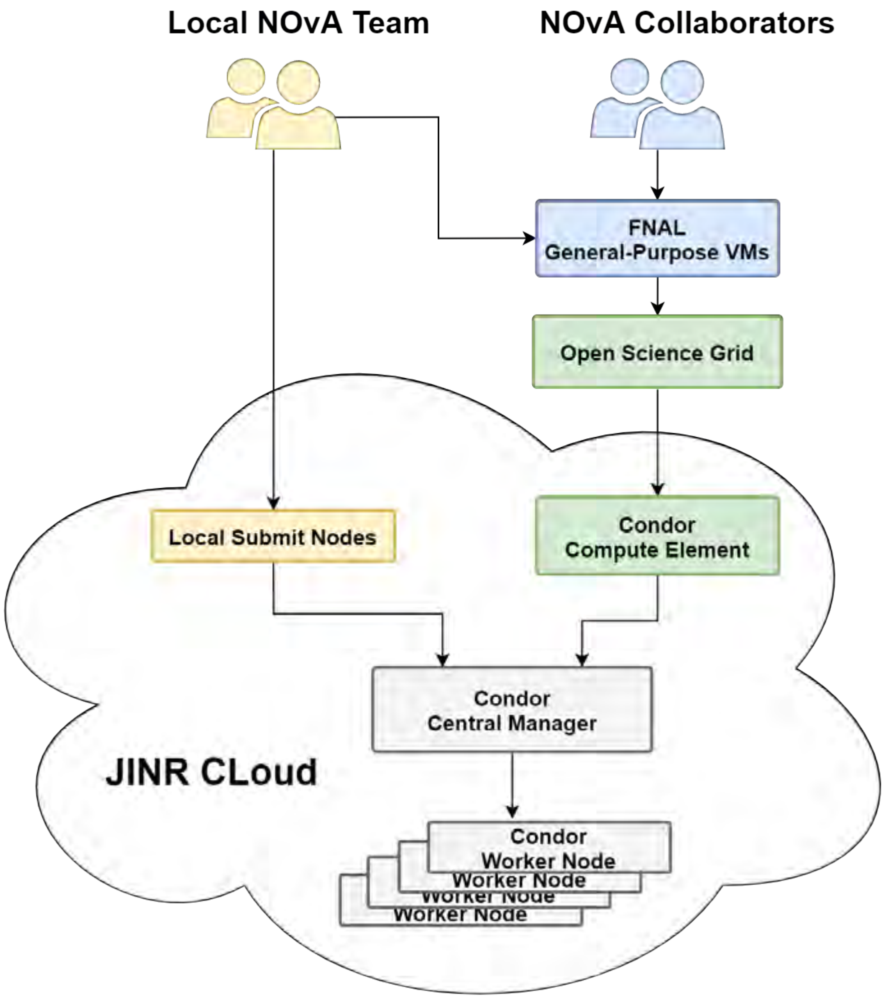


MICC JINR MULTIFUNCTIONAL INFORMATION AND COMPUTING COMPLEX

micc.jinr.ru

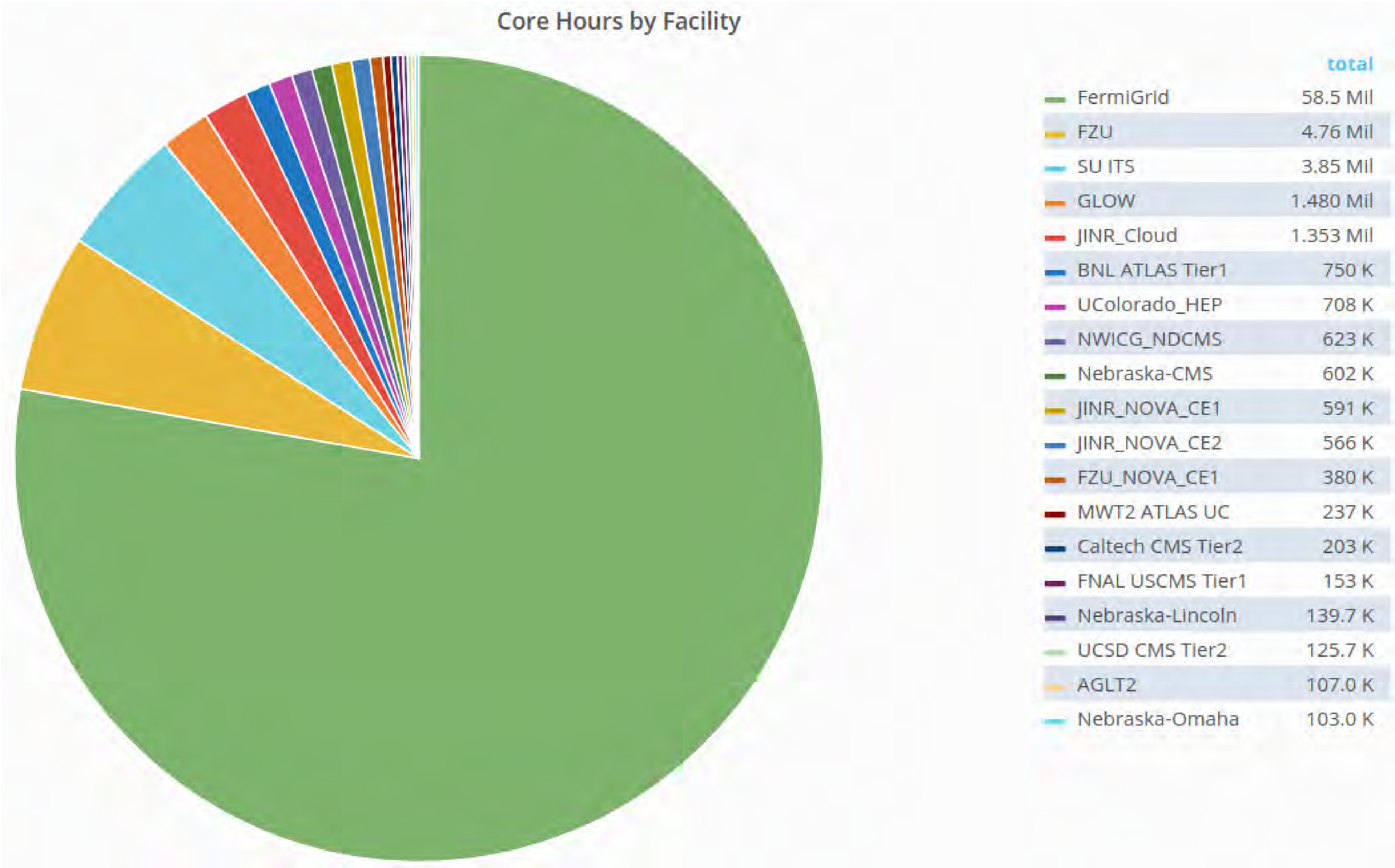
HTCondor Cloud Cluster

- Local JINR NOvA members can submit grid jobs and local jobs, all other NOvA members – grid jobs only
- Local submit nodes are similar to FNAL GPVMs with a few additions:
  - JINR Kerberos authentication
  - Additional 20 TB NFS storage
- Since everything in the cluster is a virtual machine, it can be easily scaled according to the needs
- Mu2e and DUNE VOs were provided access



SL6 everywhere...

NOvA Resource Usage

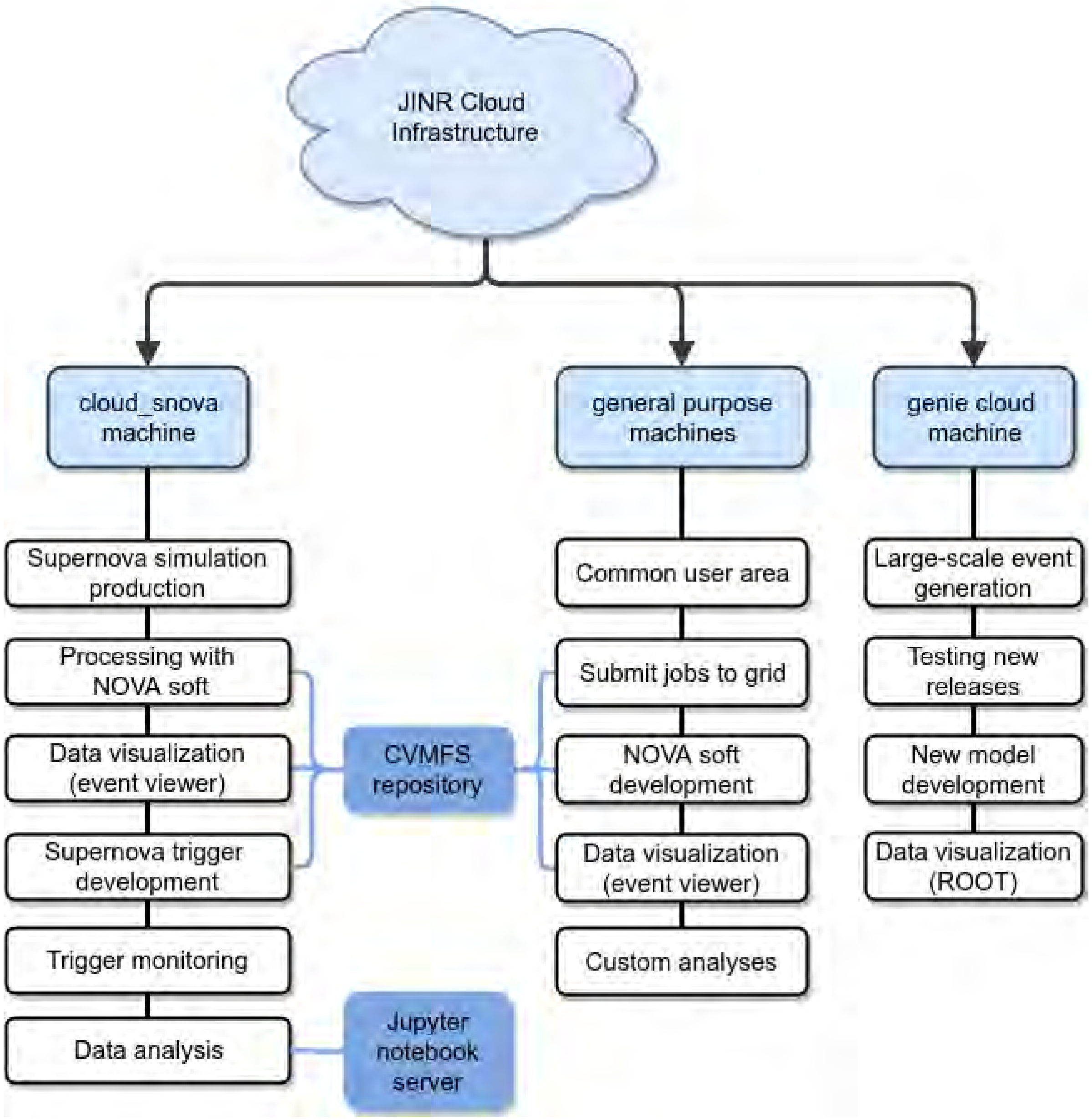


	total
FermiGrid	58.5 Mil
FZU	4.76 Mil
SU ITS	3.85 Mil
GLOW	1.480 Mil
JINR_Cloud	1.353 Mil
BNL ATLAS Tier1	750 K
UColorado_HEP	708 K
NWICG_NDCMS	623 K
Nebraska-CMS	602 K
JINR_NOVA_CE1	591 K
JINR_NOVA_CE2	566 K
FZU_NOVA_CE1	380 K
MWT2 ATLAS UC	237 K
Caltech CMS Tier2	203 K
FNAL USCMS Tier1	153 K
Nebraska-Lincoln	139.7 K
UCSD CMS Tier2	125.7 K
AGLT2	107.0 K
Nebraska-Omaha	103.0 K

Statistics for the period January, 1 2018 to March 26, 2020

Non-Batch Cloud Usage

- Cloud is universal
- Users can have personal Vms
- Interactive acces via ssh/X2go/VNC
- Any OS and software can be installed
- Can host web-services (like Jupyter notebooks)



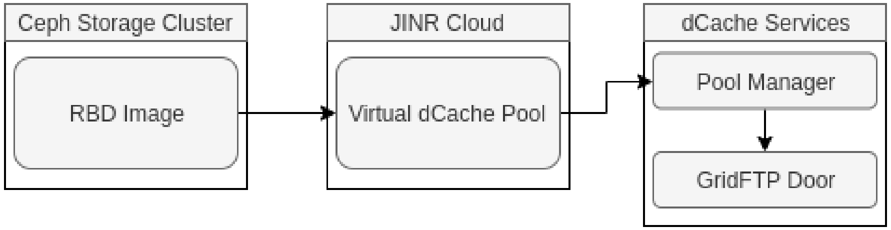
Storage Nodes

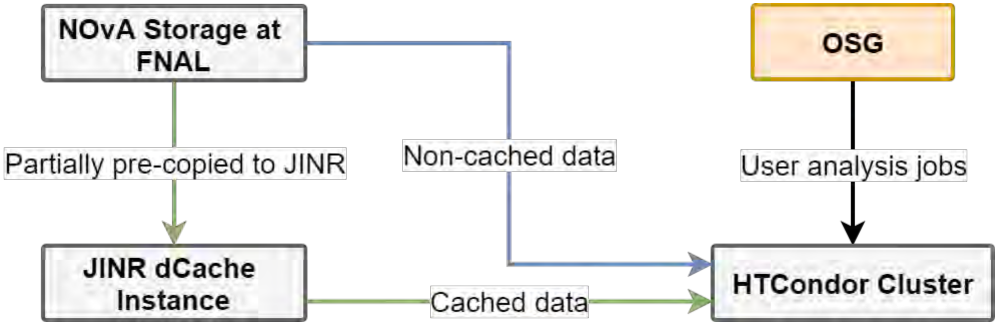
- 4 new storage servers were purchased and added to the cloud Ceph storage
- 448 TB of raw space with triple replication
- Provides virtual block devices (VBD) used as disks for all cloud VMs
- A dedicated VM exports 20 TB VBD as an NFS share to all NOvA machines in the cloud
- Has an S3 interface and potentially can be used as a CVMFS backend for storing containers

Platform	CPU	RAM	Disk	Network Interfaces
Dell PowerEdge R730xd	2xE5-2620v4	128 GB	2x400 GB SSD, 16x8 TB HDD	4x1Gb and 2x10Gb Ethernet
Dell PowerEdge R740xd	2xSilver 4114	128 GB	2x120 GB SSD, 12x10 TB HDD	4x1Gb and 2x10Gb Ethernet
Dell PowerEdge R740xd	2xSilver 4114	128 GB	2x400 GB SSD, 10x10 TB HDD	4x1Gb and 2x10Gb Ethernet
Dell PowerEdge R740xd	2xSilver 4114	128 GB	2x400 GB SSD, 10x10 TB HDD	4x1Gb and 2x10Gb Ethernet

Storage Element

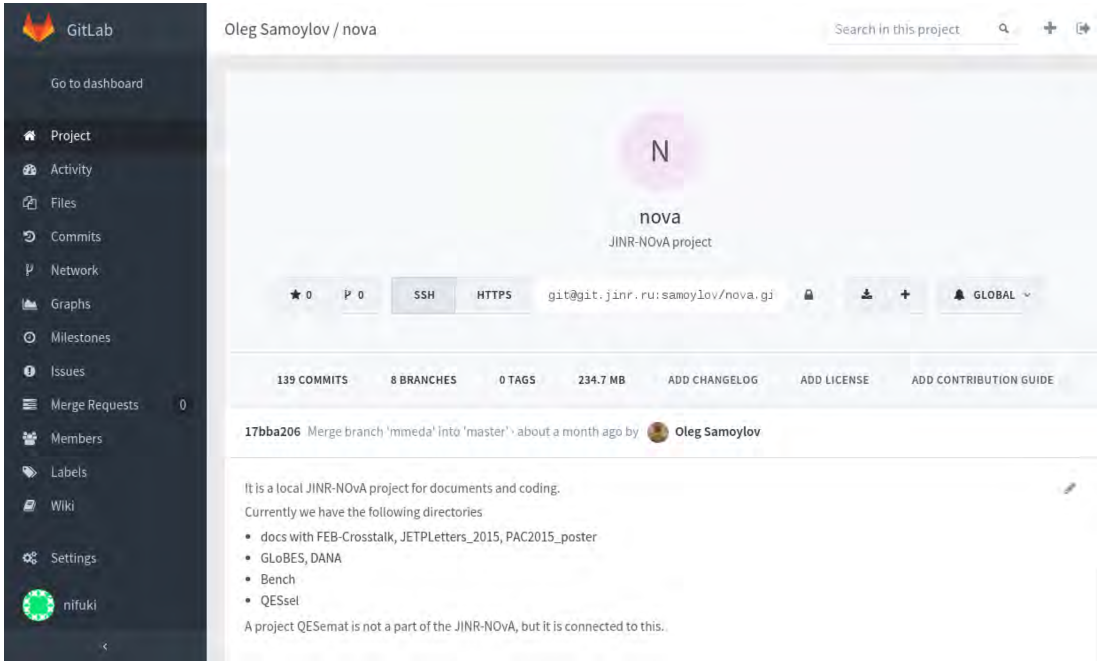
- We used a dCache instance already installed in LIT
- ~3 TB were provided by LIT
- 50 TB attached as a block device to a cloud VM which was configured as a dCache pool
- Permissions to clone datasets were acquired in NOvA SAM - Sequential Access via Metadata data handling system
- Waiting for the main NOvA SAM station at FNAL to be configured to automatically prefer JINR storage at our cluster





Side-projects: GitLab

- GitLab service at git.jinr.ru was first deployed by NOvA JINR team request
- It became a common JINR service
  - Almost 400 users
  - More than 400 projects



17bba206 Merge branch 'mmeda' into 'master' · about a month ago by Oleg Samoylov

It is a local JINR-NOvA project for documents and coding.

Currently we have the following directories

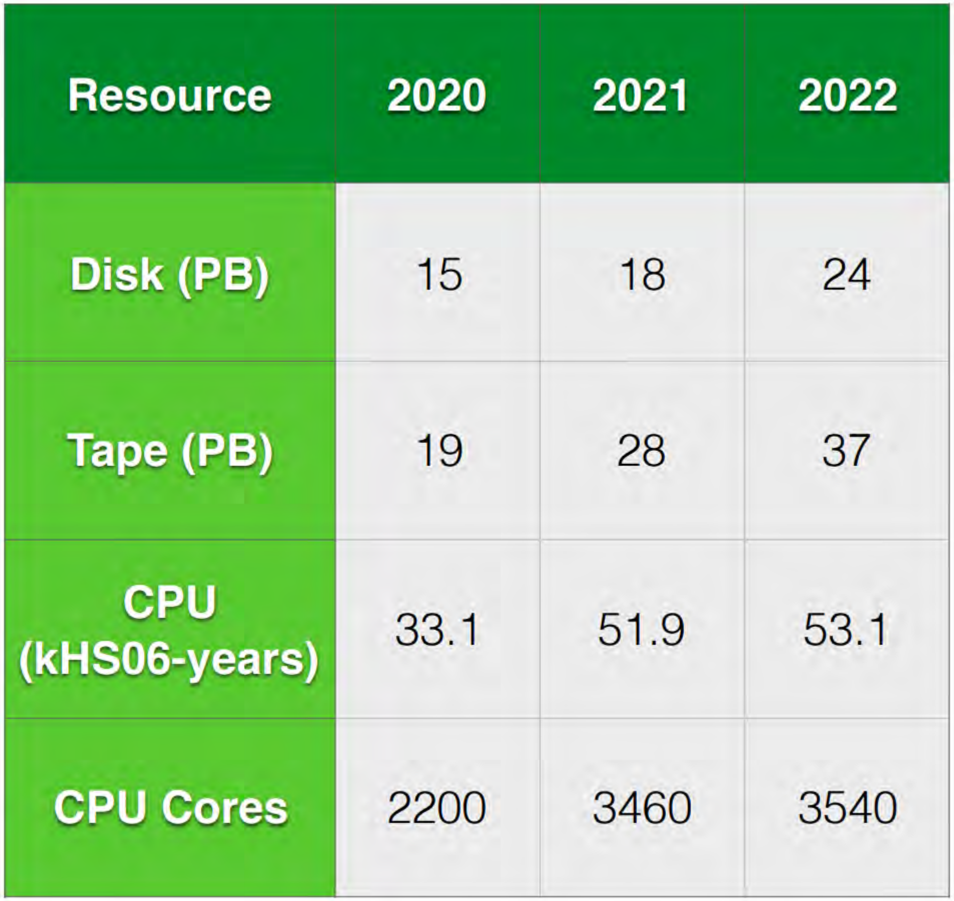
- docs with FEB-Crosstalk, JETPLetters_2015, PAC2015_poster
- GLoBES, DANA
- Bench
- QESsel

A project QESemat is not a part of the JINR-NOvA, but it is connected to this.

DUNE Resource Needs

- Simplified terms for current DUNE sites
  - Tape Site (Tier 1) - tape/staging
  - Disk Site (Tier 2) - disk + CPU
  - Compute Site (Tier 3) - CPU + cache
  - Analysis Site (Tier 3) - CPU + cache
  - HPC - (HPC, IaaS)
- Goal is to have resource split between FNAL and other institutions – 25% / 75%
- Notwithstanding the “Request” to any country wishing to make a serious contribution is
  - at least 5% of requirements
  - preferably 10%

Total Needs



Resource	2020	2021	2022
Disk (PB)	15	18	24
Tape (PB)	19	28	37
CPU (kHS06-years)	33.1	51.9	53.1
CPU Cores	2200	3460	3540

Plans on Computing

- NOvA computing resources are expected to be increased to ~1000 cores during 2020
- JUNO servers are being set up with ~2000 cores
- The idea is to unite all resources of the JINR neutrino projects into one shared system – **Neutrino Platform**
- CPU capacity of the combined Neutrino Platform should be enough to become the Tier-2 site for the DUNE and to represent Russian contribution into the experiment
- While CPU can be shared between the experiments, disk storage must be dedicated to DUNE: ~2.5 PB storage needs to be acquired

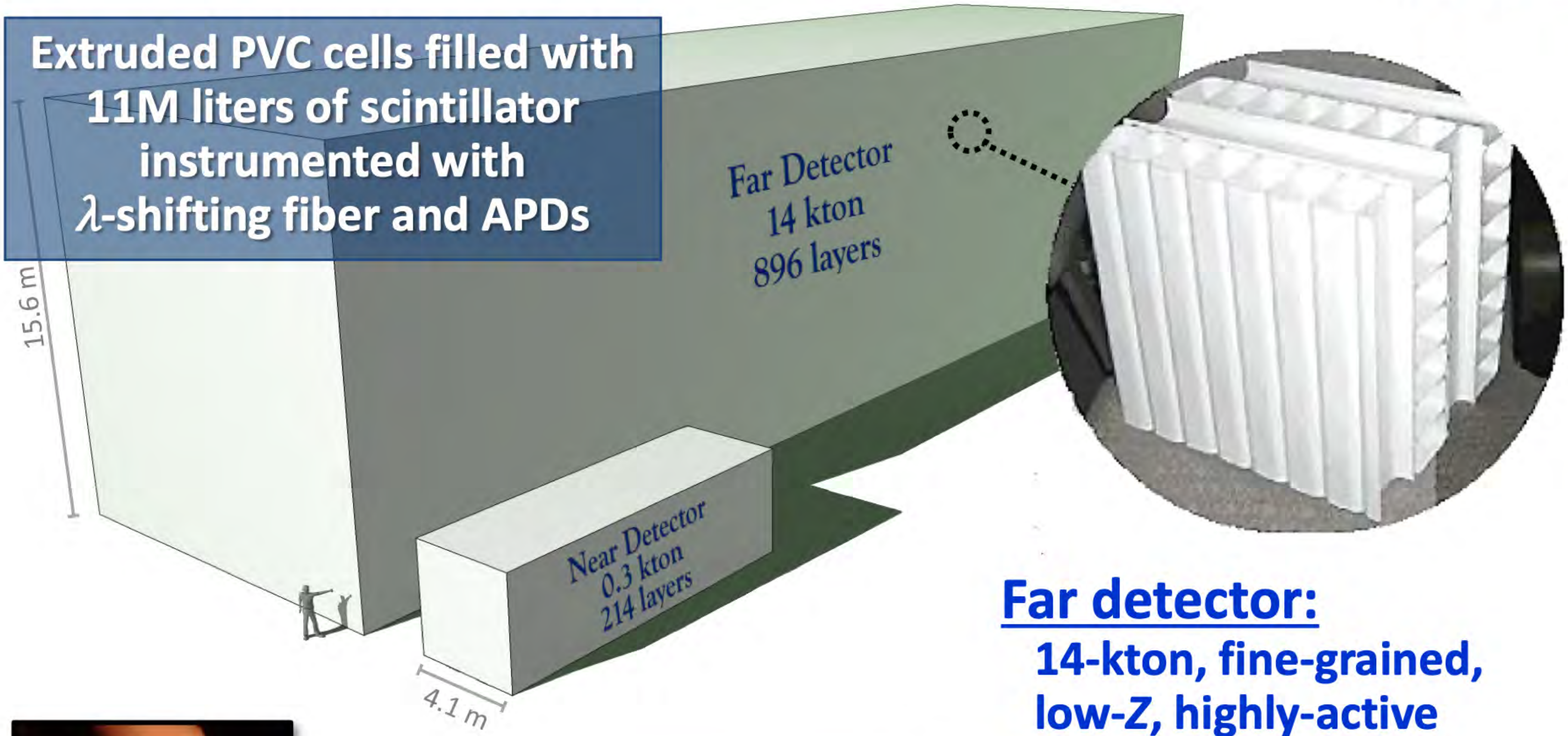
Computing Manpower

- NOvA Support:
  - **Nikita Balashov**: GPVM and HTCondor system administration, technology evaluation
  - **Evgeniy Kuznetsov**: HTCondor system administration
  - **Nikolay Kutovskiy**: procurements, cloud management
  - **Andrey Sheshukov**: GPVM software environment
- Same people are expected to support the DUNE experiment
  - **Oleg Samoylov**: computing consortium representative
  - **Nikita Balashov**: technical liaison

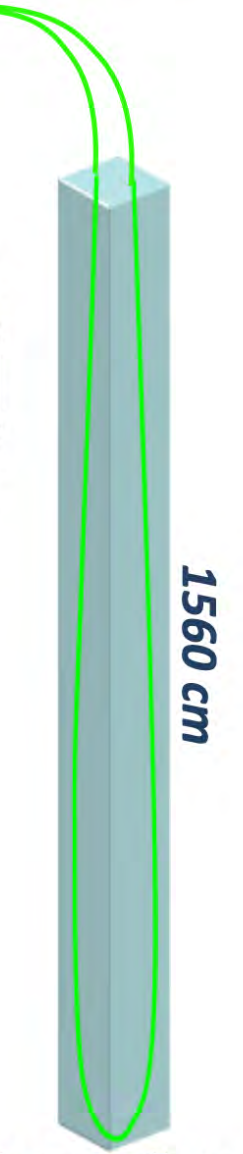
Remote Operations Center

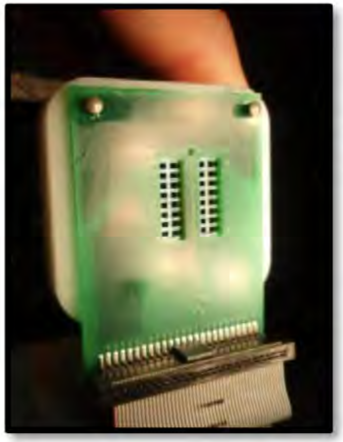
NOνA detectors

Extruded PVC cells filled with 11M liters of scintillator instrumented with λ-shifting fiber and APDs



A NOνA cell





32-pixel APD ←

Fiber pairs from 32 cells →

Far detector:

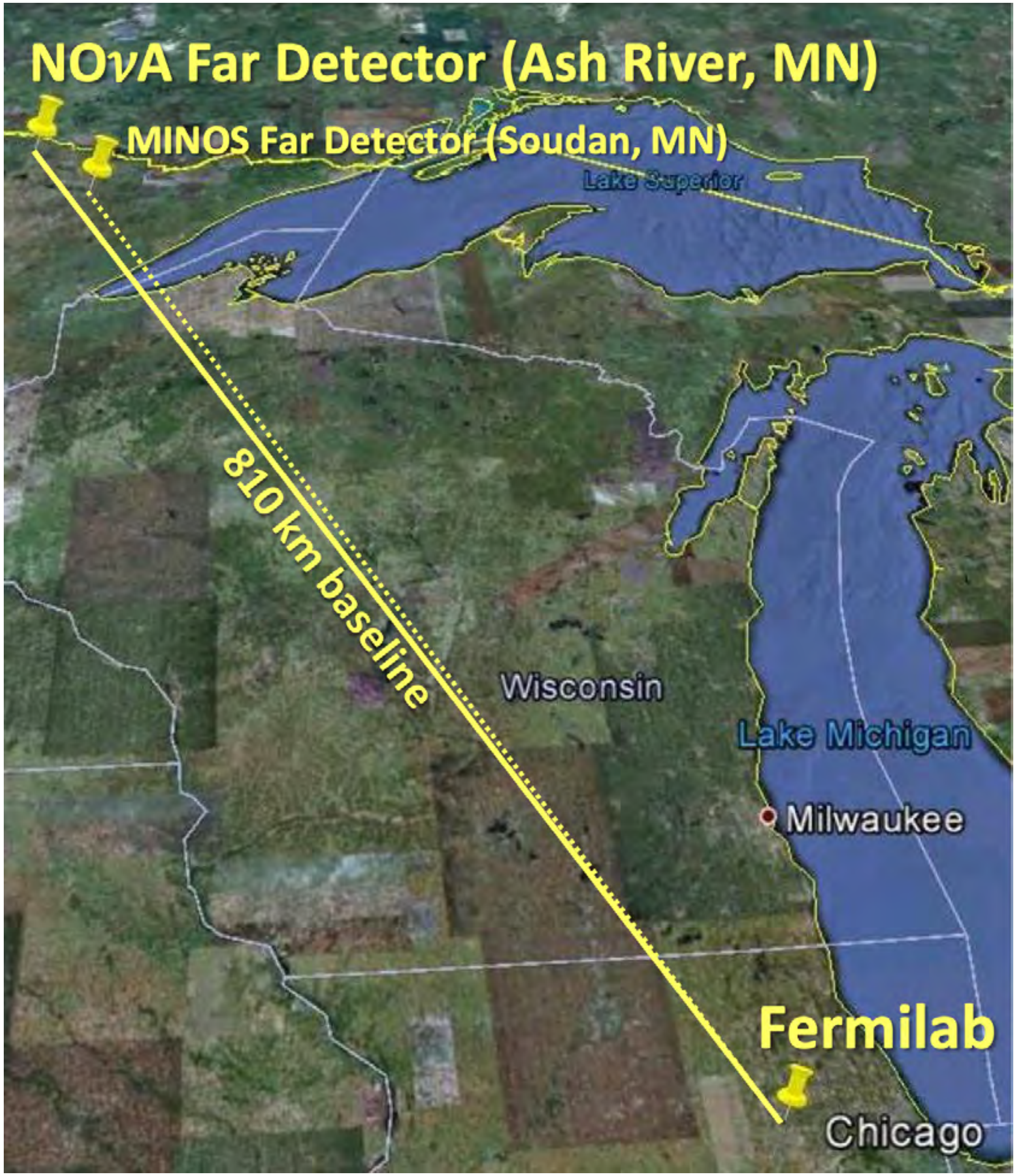
14-kton, fine-grained, low-*Z*, highly-active tracking calorimeter
→ ***344,000 channels***

Near detector:

0.3-kton version of the same
→ ***20,000 channels***

Remote Control

- Studying of the Neutrino Oscillation Phenomenon requires long baseline for Neutrino Source and (Far) Detector
- The data is first recorded to the local storage of each detector and then gets transferred to FNAL
- More efficient monitoring for the system is to use one (Remote) Operation Center for both places in the same time



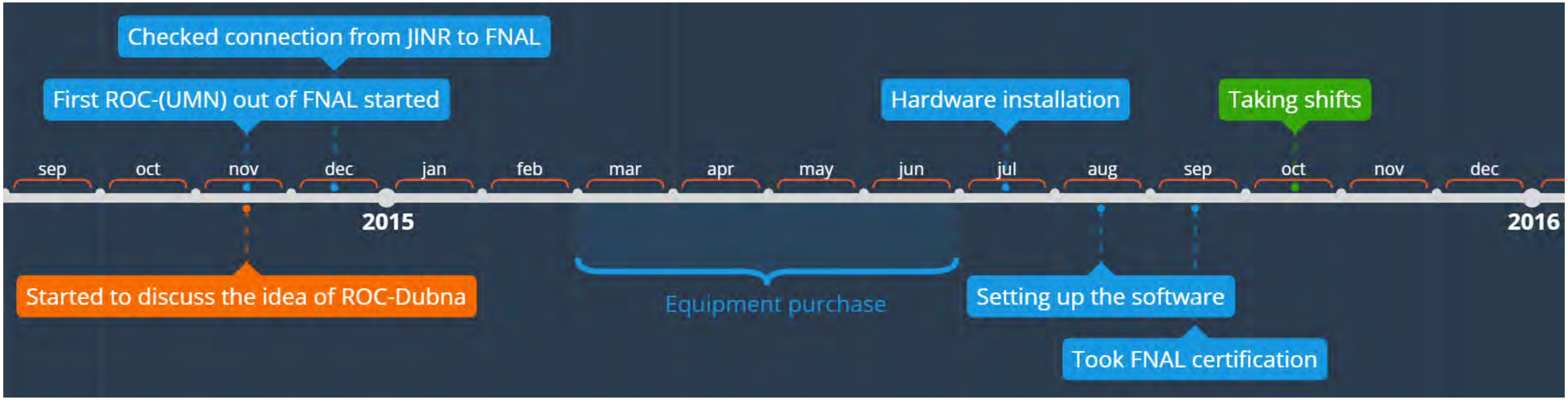
Remote Control

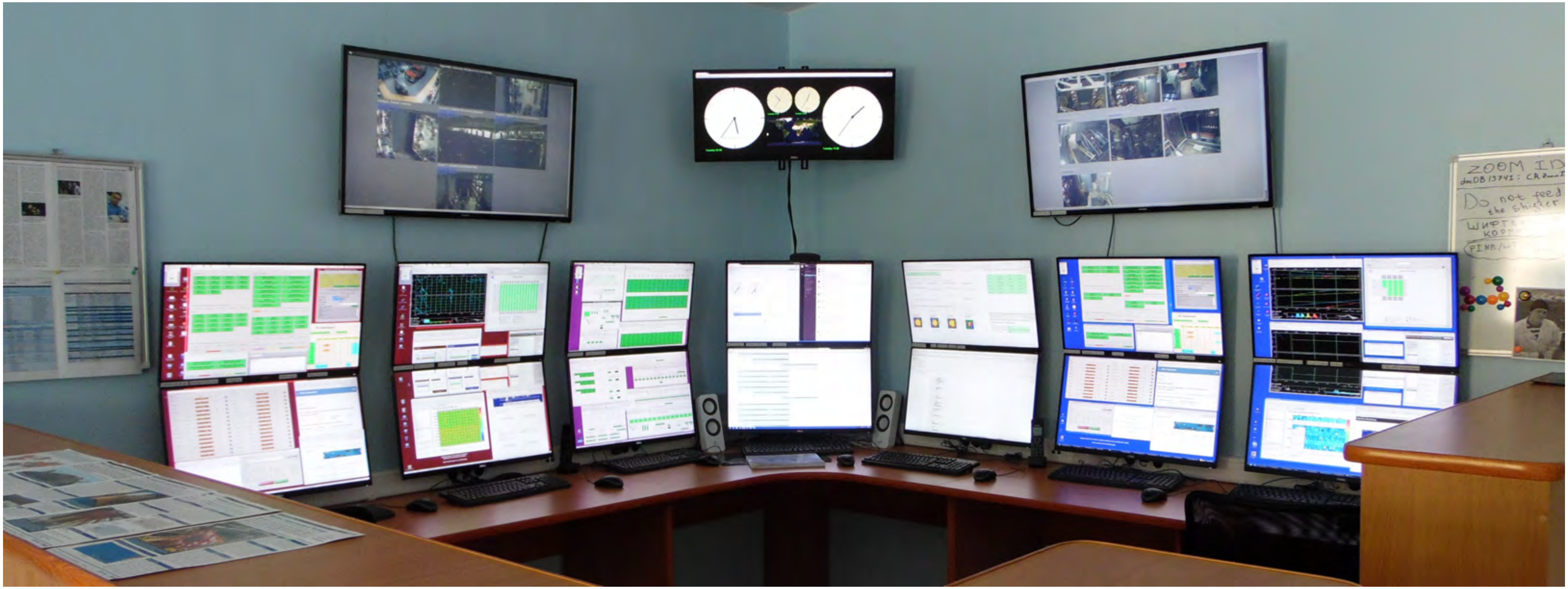
- This idea came to Fermilab after starting the LHC Era
- All the NuMI experiments develop Remote Operation Centers (ROC)
- Main Remote Operation Center ROC-West is placed in Wilson Hall at FNAL
- 25 NOvA ROCs are in operation by now

🇺🇸	🇬🇧	🇷🇺	🇧🇷	🇨🇿
20	2	1	1	1

ROC-Dubna Timeline

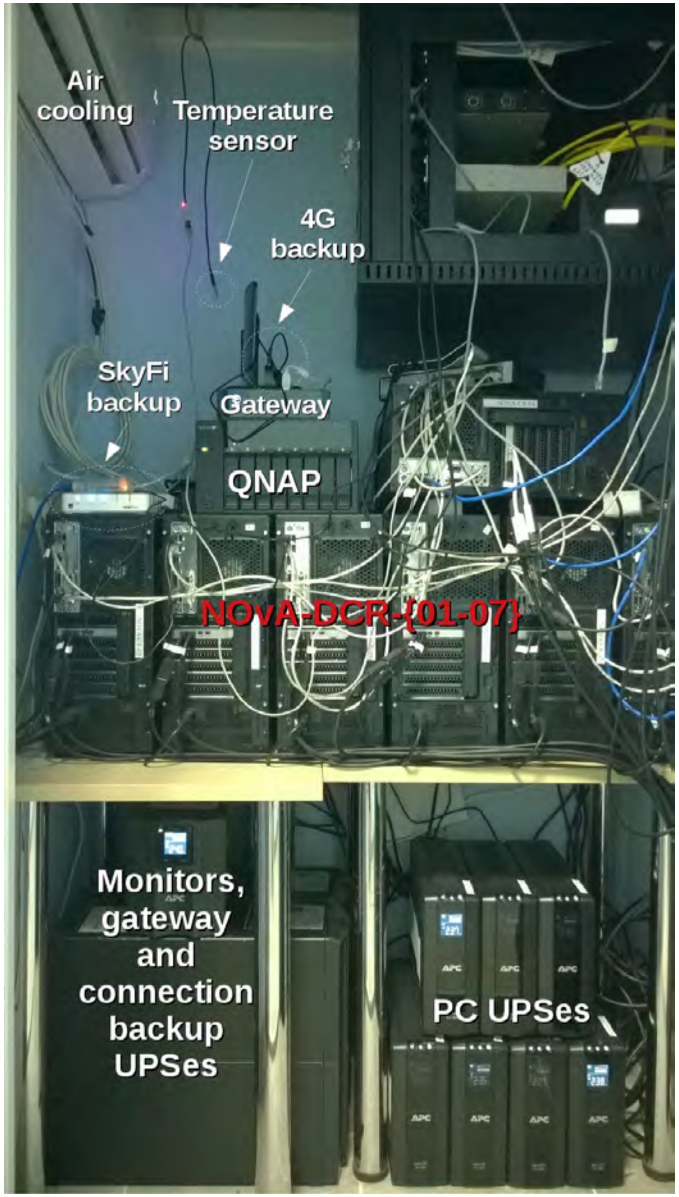
- The NOvA detectors are operated from ROC-West at FNAL since 2013
- It was the ROC-UMN at Minnesota University that first started its operation from the non-Fermilab area since November, 2014
- Right after the ROC-UMN was launched we started disscussing the ROC-Dubna idea
- ROC-Dubna hosted first shifters in October, 2015 and became the first ROC outside the US



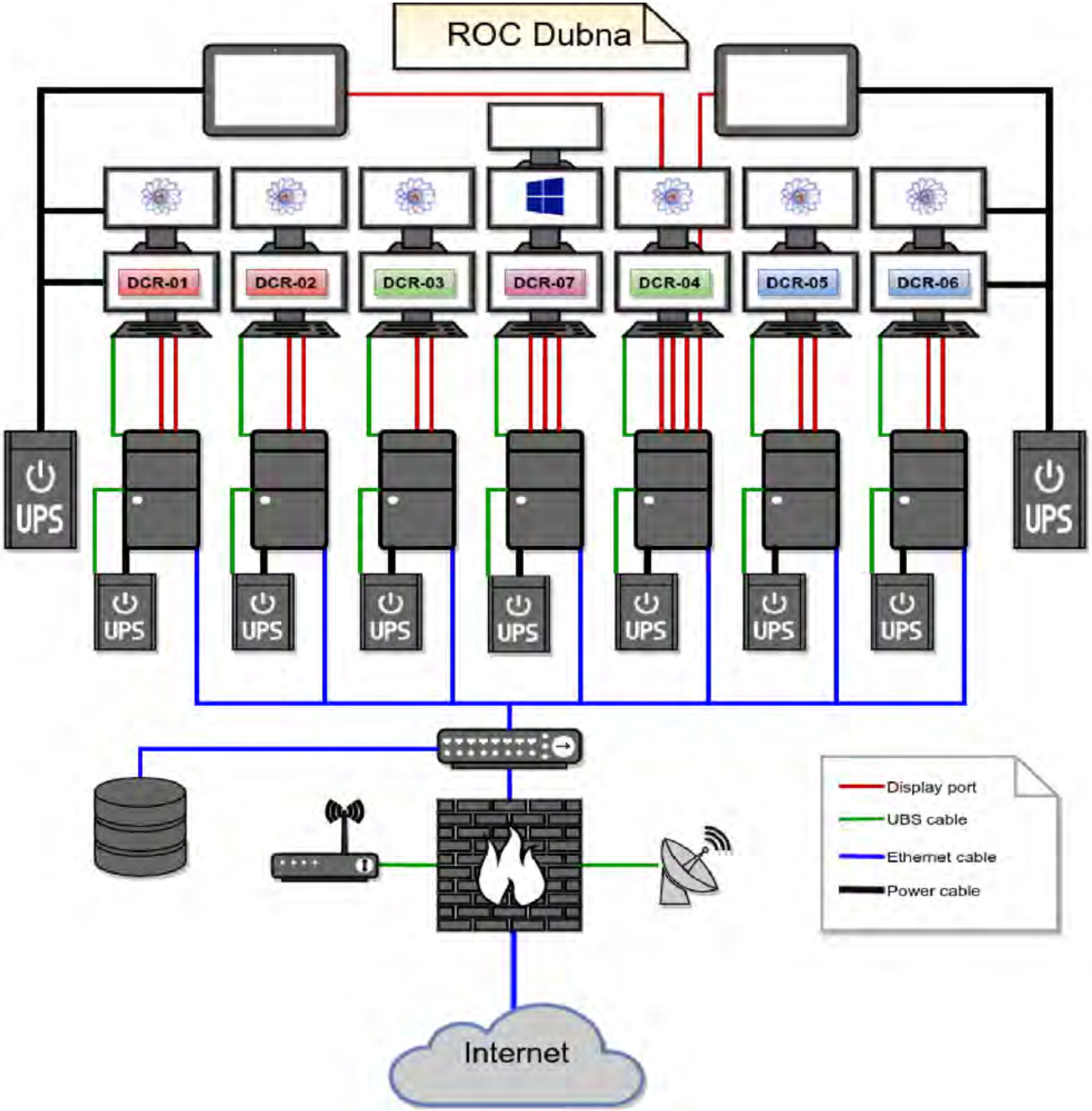

ZOOM ID
docDB 13741: CR
Do not feed the shifter
ШИФТ
КОР
FINB/

Backstage

- All the computing hardware is located in a dedicated server room bellow the ROC
- Air cooling system prevents overheating of such a dense setup
- UPSs should keep computers running for ~1 hour – enough to call the Run Coordinator and find a backup shifter
- 4G and SkyFi back up Internet channels

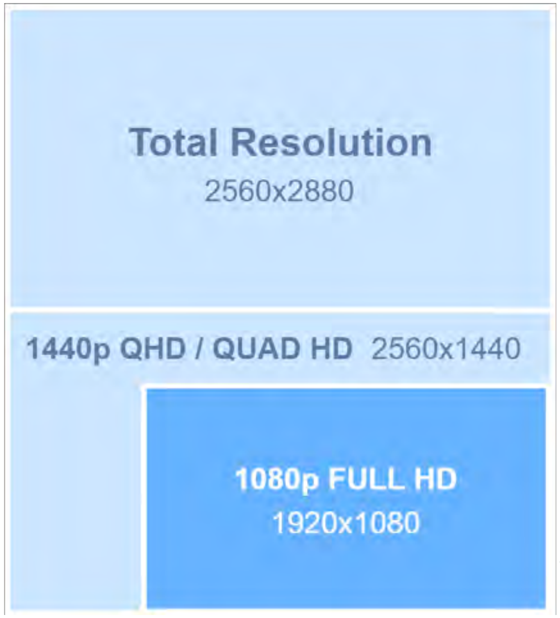


ROC-Dubna Scheme

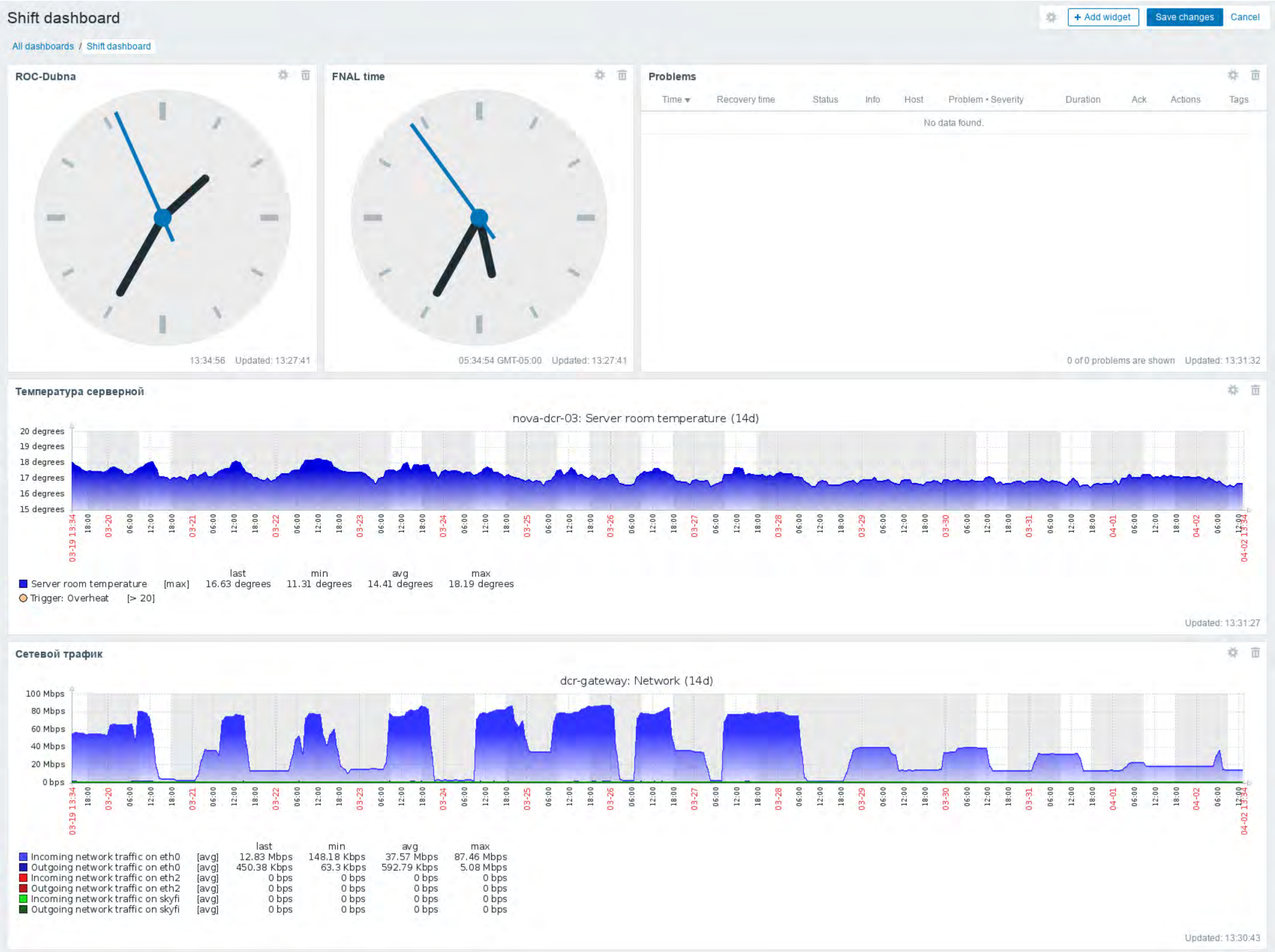


How ROCs work

- 5 active VNC-sessions are tunneled to Near and Far Detector Nodes at FNAL through the secure connection
- 1 Linux Node is used for Web-based monitoring (Beam, ND/FD Cameras, Data transfer control, Ganglia, Nearline)
- A Windows Node is used for Communication (NOvA electronic loogbook, latest version of Expert contact and Bulletin board, Polycom via Vidyo, Zoom, Slack-chat, Skype)
- Infrastructure is designed for the 8 hours long work (stable internet, international land-line, kitchen)



Monitoring



ROC-Dubna: 13:34:56 Updated: 13:27:41

FNAL time: 05:34:54 GMT-05:00 Updated: 13:27:41

Time ▾	Recovery time	Status	Info	Host	Problem • Severity	Duration	Ack	Actions	Tags
No data found.									

0 of 0 problems are shown Updated: 13:31:32

Температура серверной

nova-dcr-03: Server room temperature (14d)

		last	min	avg	max
🟦 Server room temperature	[max]	16.63 degrees	11.31 degrees	14.41 degrees	18.19 degrees
● Trigger: Overheat	[> 20]				

Updated: 13:31:27

Сетевой трафик

dcr-gateway: Network (14d)

		last	min	avg	max
🟦 Incoming network traffic on eth0	[avg]	12.83 Mbps	148.18 Kbps	37.57 Mbps	87.46 Mbps
🟦 Outgoing network traffic on eth0	[avg]	450.38 Kbps	63.3 Kbps	592.79 Kbps	5.08 Mbps
🟥 Incoming network traffic on eth2	[avg]	0 bps	0 bps	0 bps	0 bps
🟥 Outgoing network traffic on eth2	[avg]	0 bps	0 bps	0 bps	0 bps
🟩 Incoming network traffic on skyfi	[avg]	0 bps	0 bps	0 bps	0 bps
🟩 Outgoing network traffic on skyfi	[avg]	0 bps	0 bps	0 bps	0 bps

Updated: 13:30:43

ROC-Dubna Crew

- **Alexander Antoshkin**, Super-liaison, ROC-Dubna liaison, Hardware and Software expert
- **Oleg Samoylov**, ROC-manager, Software expert
- **Andrey Sheshukov**, Software expert
- **Nikolay Anfimov**, Hardware expert
- **Chris Kullenberg**, Super-Shifter
- **Nikita Balashov**, Software and IT-support
- **Andrey Dolbilov**, Internet and IT emergency

Advantages to have ROC in Dubna

- ROC-Dubna allows taking Shifts and save traveling budget and time scheduling / shifting / jet lag
- ROC-Dubna is an Operation and Communication Center of the NOvA experiment ~8000 km and 8/9 time zones away (night shifts during daytime)
- Our Russian colleagues (INR, Moscow and FIAN) have interest to visit us
- ROC-Dubna is a public place open for excursions and is visited by Scholars, Teachers, Students, Journalists and other JINR guests.

Plans

- NOvA was extended to operate till 2025
- Modern computer equipment lifetime is 5 years – it is time to prepare hardware upgrade
- Scientific Linux 6 will reach end of life in November of 2020 – migration of linux nodes to a modern OS is required
- Add visualization and notification system based on Grafana (with current Zabbix instance as a data provider) located out of ROC

Summary

- Cloud virtual computing infrastructure was created for the NOvA
  - Functions normally at current scale
  - Local and Grid jobs supported
  - Mu2e and DUNE VOs already supported
  - To have a more serious contribution to DUNE computing infrastructure needs to be extended
- Scientific Linux 6 will reach end of life in November of 2020 – migration of linux nodes to a modern OS is required
- ROC-Dubna operates normally, but needs to be upgraded