PanDA @ NRC KI

R. Mashinistov
National Research Centre “Kurchatov Institute”, Moscow, Russia

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "КУРЧАТОВСКИЙ ИНСТИТУТ"

CERN

PanDA

The University of Texas ARLINGTON™

BROOKHAVEN NATIONAL LABORATORY

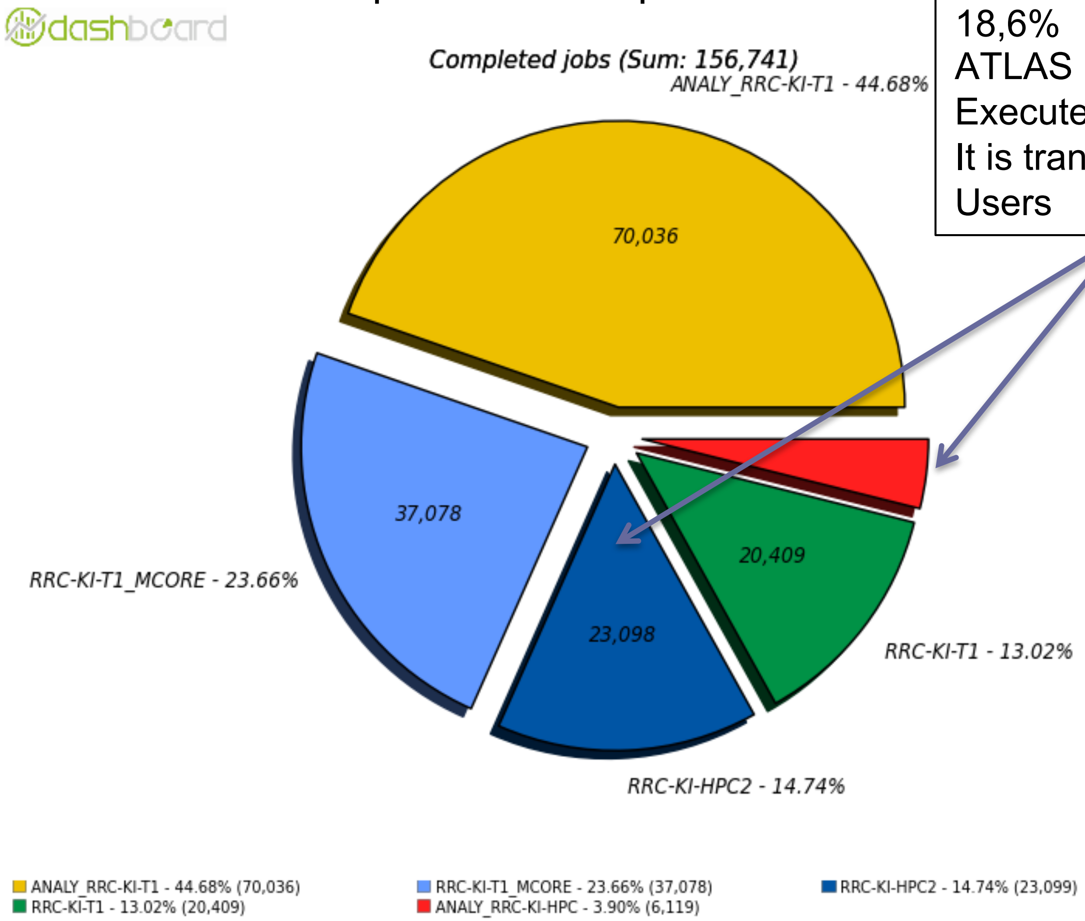
HPC2 supercomputer at NRC KI

High Performance Cluster - HPC2 second generation HPC with peak performance 122,9 TFLOPS (commissioned 2011).
#2 in 15-th issue of Russian top50 Supercomputers

- 10240 CPU cores = 1280 nodes 2x Intel Xeon E5450 3,00GHz 4 core 16 Gb RAM;
- UI node only allows to run jobs in batch system (SLURM) or to compile the code
- Shared FS Lustre for WN’s and UI
- WN’s has an access to WAN
- CVMFS connected to WN’s
- Broadband to Tier-1 Storage Element (ANALY_RRC-KI grid site)

First steps

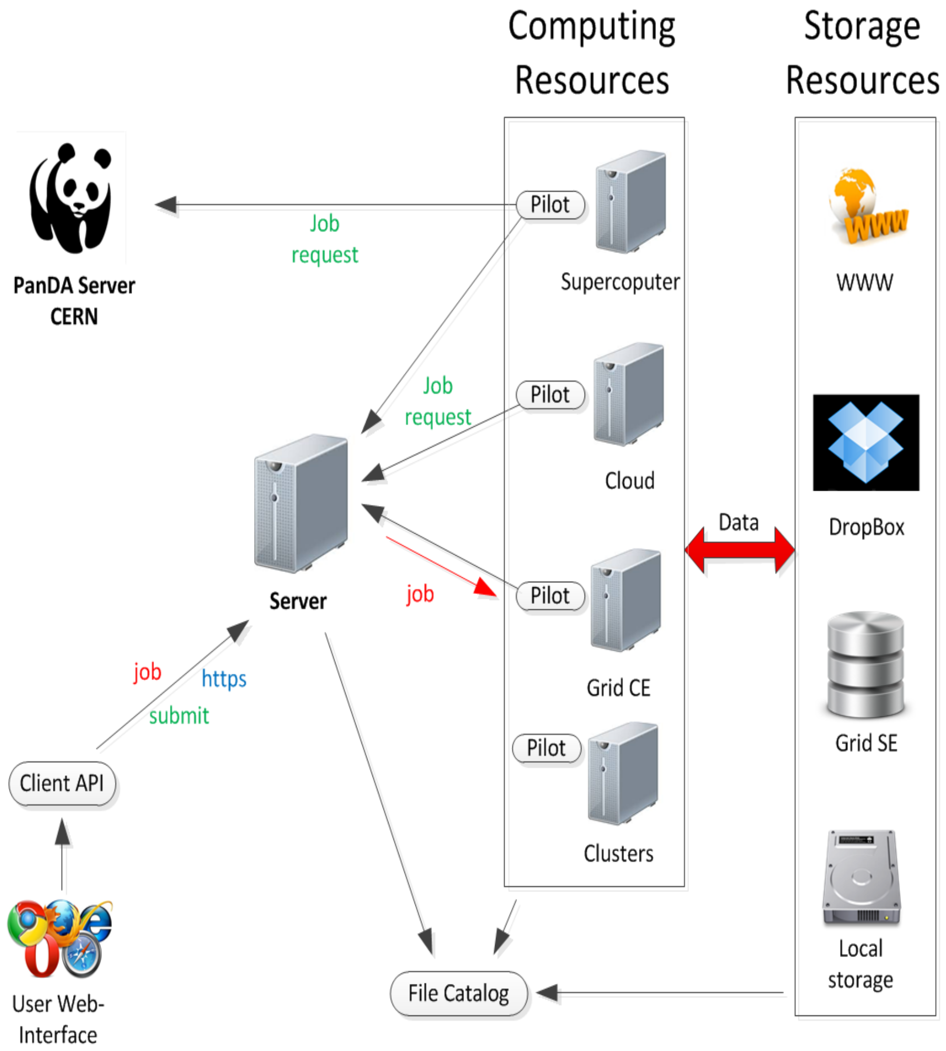
- PanDA@NRC KI
    - server, auto pilots factory, monitor and database server (MySQL)
- After APF was installed in 2014 we immediately set up Analysis queue
    - Condor-SLURM connector
- In 2015 we set up Production queue



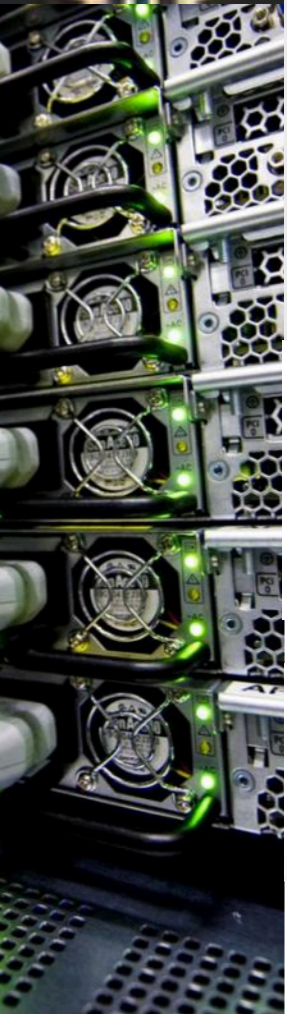
18,6%
ATLAS MC Production and Analysis
Executed on SC@NRC-KI
It is transparent for Prodsys and Users

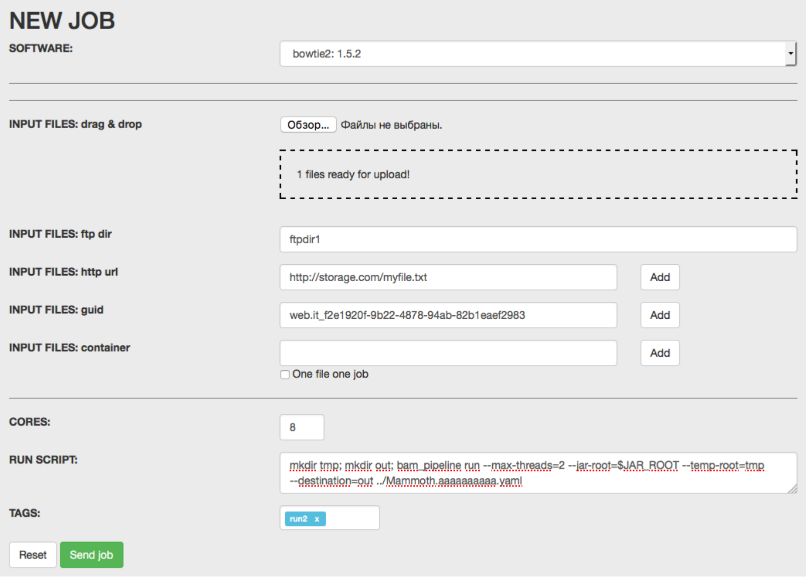
Computing portal @ NRC KI

- Web-user interface (FLUSK)
- File catalog & data transfer system
  - Plugin based (ftp, http, grid and etc.)
- Authorization OAuth 2.0
- API for external applications
  - Token authentication
  - Upload/fetch files
  - Send a job
  - Get job status/statistics
- Private FTP storage for each user



User friendly jobs running and monitoring interface
Submit jobs and obtain results





NEW JOB

SOFTWARE: bowtie2: 1.5.2

INPUT FILES: drag & drop: Файлы не выбраны.

1 files ready for upload!

INPUT FILES: ftp dir: ftpdir1

INPUT FILES: http url: http://storage.com/myfile.txt

INPUT FILES: guid: web.it_f2e1920f-9b22-4878-94ab-82b1eaef2983

INPUT FILES: container:

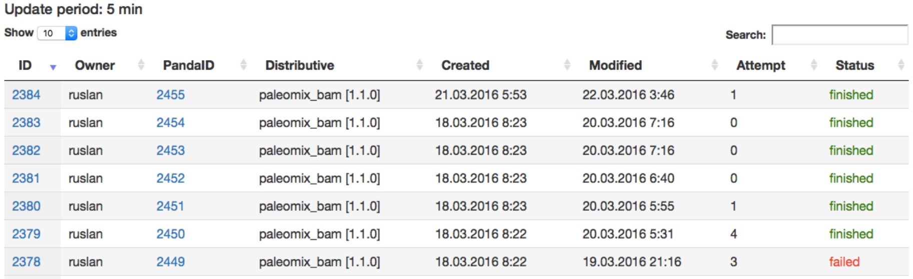
□ One file one job

CORES: 8

RUN SCRIPT:

```
mkdir tmp; mkdir out; bam_pipeline run --max-threads=2 --jar-root=$JAR_ROOT --temp-root=tmp
--destination=out ../Mammoth.aaaaaaaaa.yaml
```

TAGS: run2 x

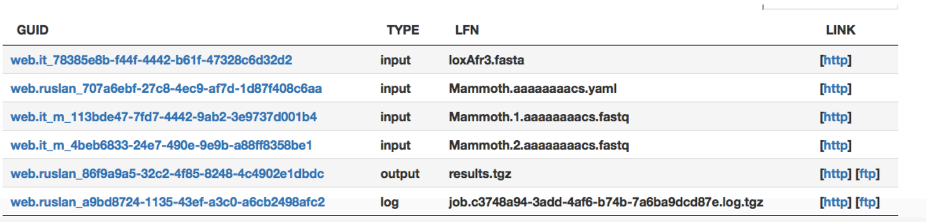


Update period: 5 min

Show 10 entries

Search:

ID	Owner	PandaID	Distributive	Created	Modified	Attempt	Status
2384	ruslan	2455	paleomix_bam [1.1.0]	21.03.2016 5:53	22.03.2016 3:46	1	finished
2383	ruslan	2454	paleomix_bam [1.1.0]	18.03.2016 8:23	20.03.2016 7:16	0	finished
2382	ruslan	2453	paleomix_bam [1.1.0]	18.03.2016 8:23	20.03.2016 7:16	0	finished
2381	ruslan	2452	paleomix_bam [1.1.0]	18.03.2016 8:23	20.03.2016 6:40	0	finished
2380	ruslan	2451	paleomix_bam [1.1.0]	18.03.2016 8:23	20.03.2016 5:55	1	finished
2379	ruslan	2450	paleomix_bam [1.1.0]	18.03.2016 8:22	20.03.2016 5:31	4	finished
2378	ruslan	2449	paleomix_bam [1.1.0]	18.03.2016 8:22	19.03.2016 21:16	3	failed



GUID	TYPE	LFN	LINK
web.it_78385e8b-f44f-4442-b61f-47328c6d32d2	input	loxAfr3.fasta	[http]
web.ruslan_707a6ebf-27c8-4ec9-af7d-1d87f408c6aa	input	Mammoth.aaaaaaaacs.yaml	[http]
web.it_m_113bde47-7fd7-4442-9ab2-3e9737d001b4	input	Mammoth.1.aaaaaaaacs.fastq	[http]
web.it_m_4beb6833-24e7-490e-9e9b-a88ff8358be1	input	Mammoth.2.aaaaaaaacs.fastq	[http]
web.ruslan_86f9a9a5-32c2-4f85-8248-4c4902e1dbdc	output	results.tgz	[http] [ftp]
web.ruslan_a9bd8724-1135-43ef-a3c0-a6cb2498afc2	log	job.c3748a94-3add-4af6-b74b-7a6ba9dcd87e.log.tgz	[http] [ftp]

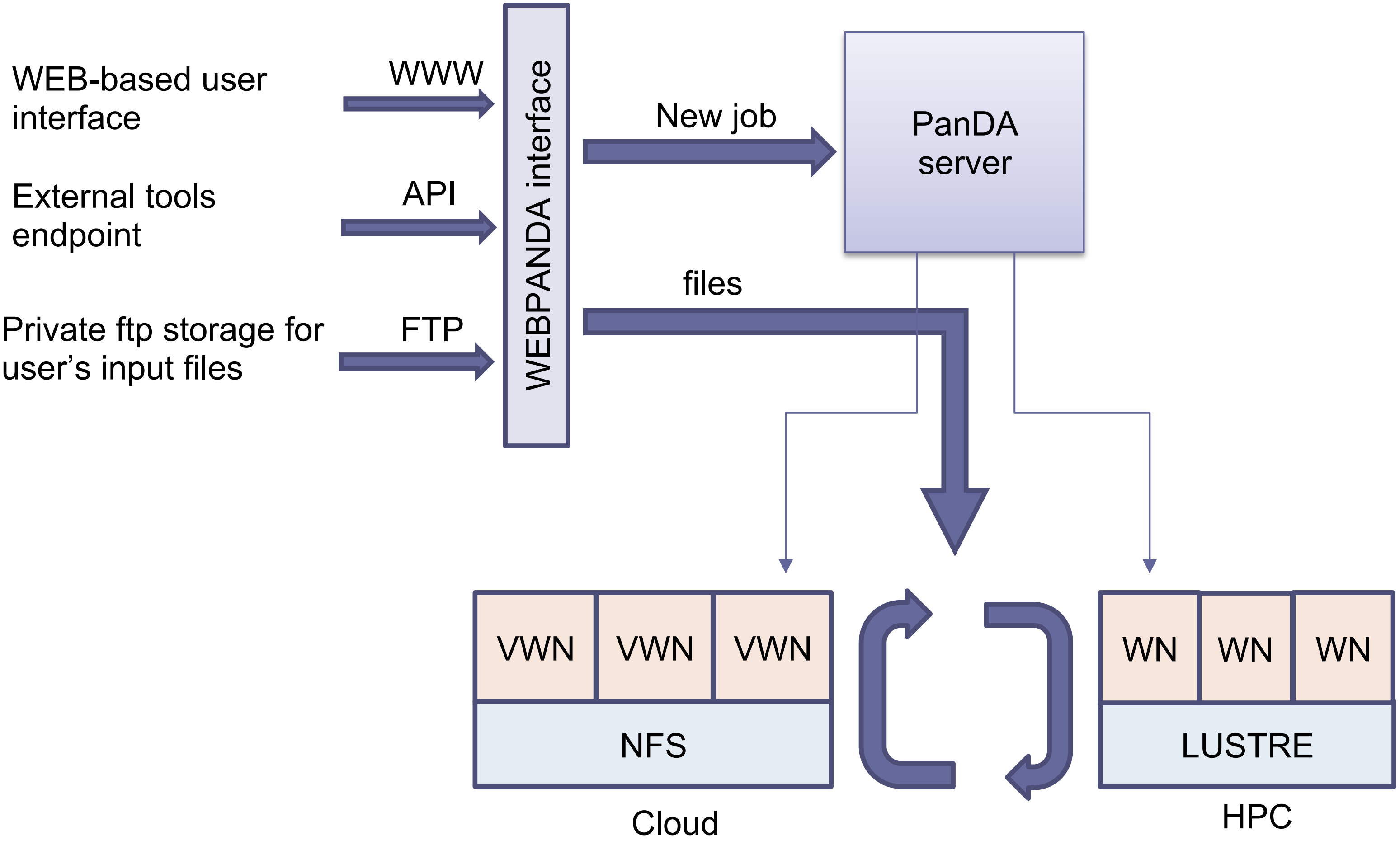
- Job setup interface
  - Easy setup distributive, input files, parameters and output file names
  - Local authentication (users don’t need a certificate)

- Jobs monitor

- After submitting a job you can monitor its status in the "Job list" tab. When the job finishes, you can get your results from the detailed info page
  - Links to the input/output files

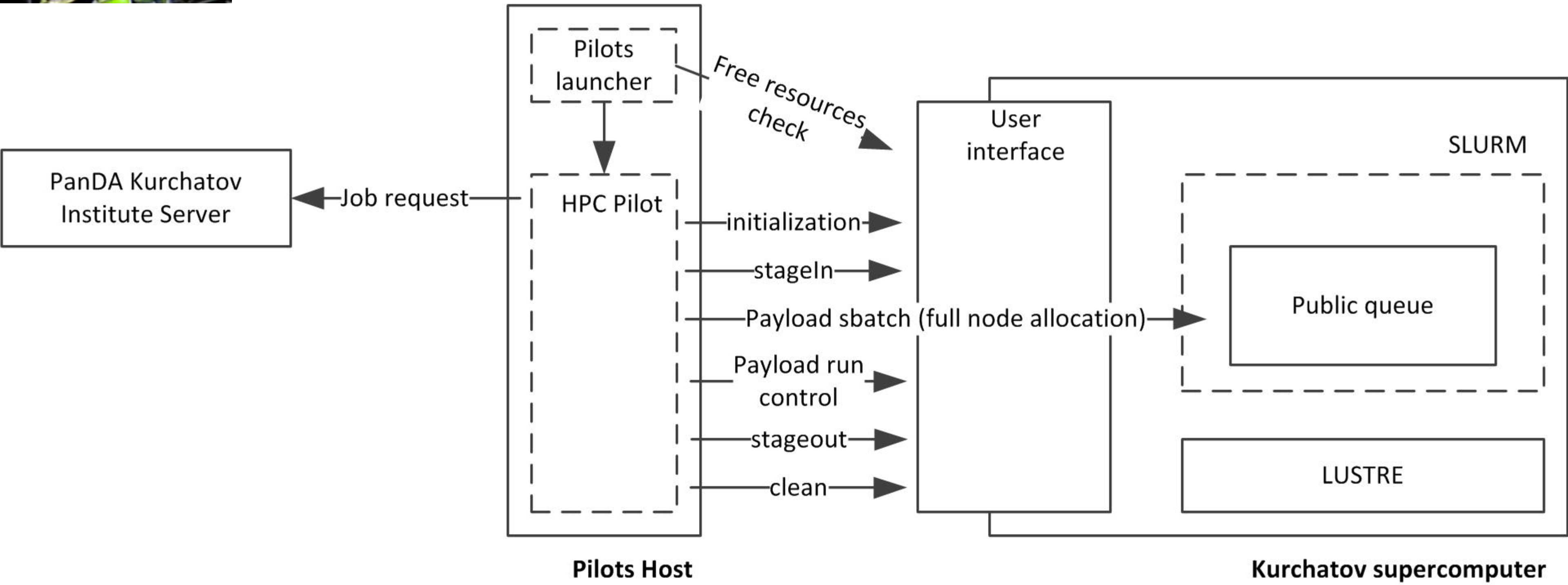
Portal operation

- OAuth 2.0 authentication
- Asynchronous operations – Celery
- API for external applications



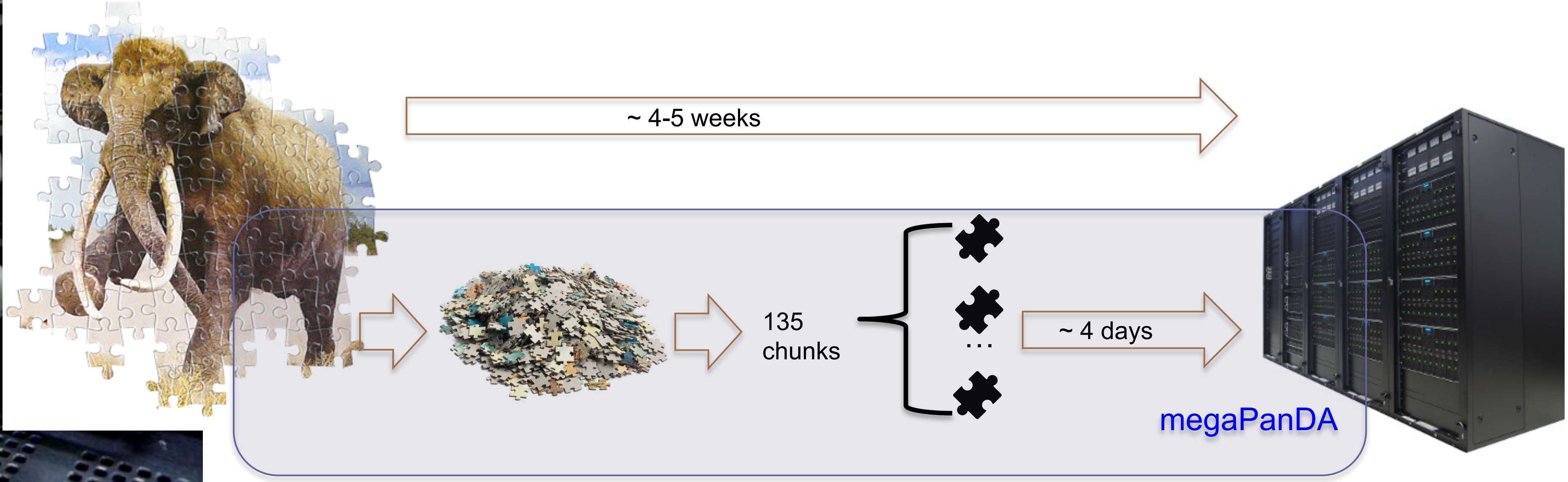
HPC Pilot for biology jobs

- Pilots are running on auxilliar node
- Pilot runs job on the WN’s via SLURM



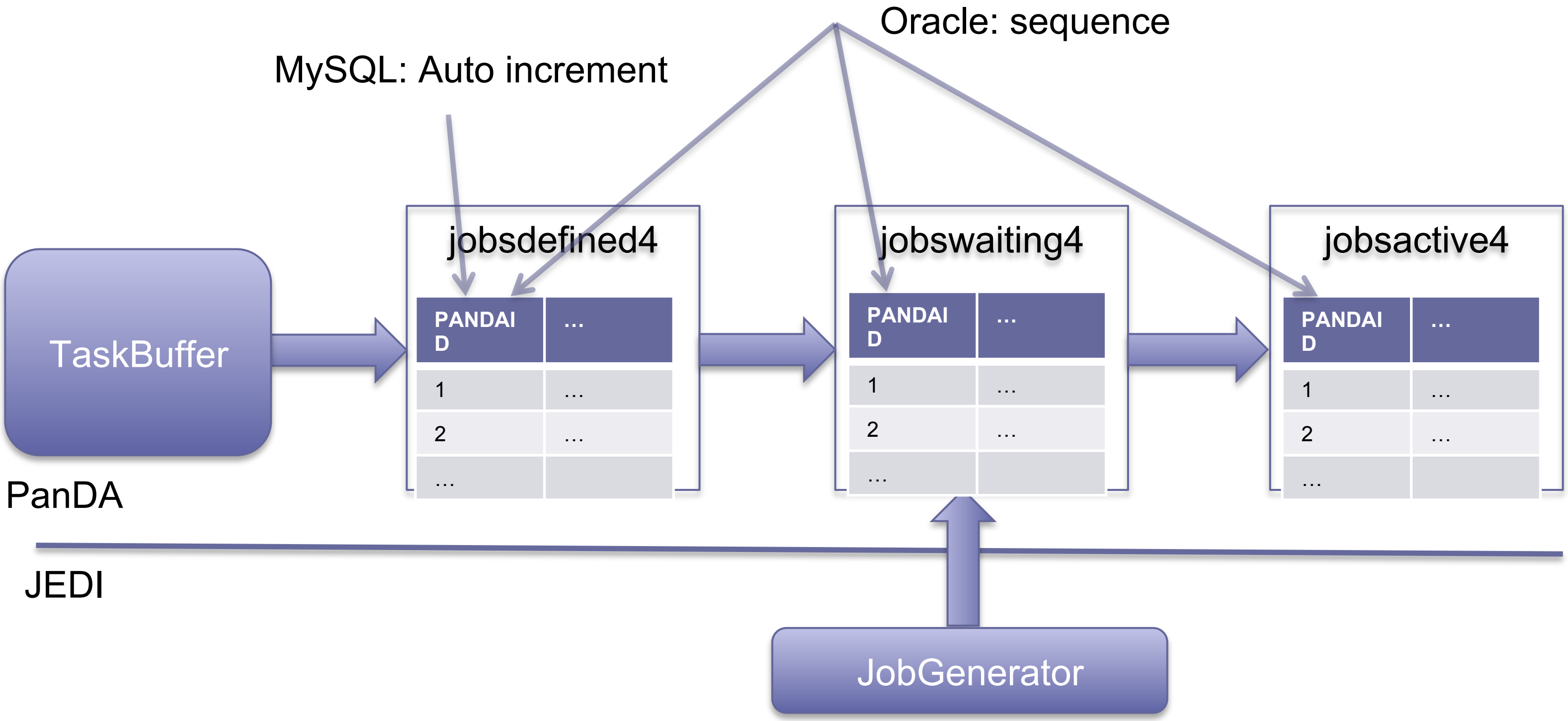
Biology application

- Next Generation Genome Sequencing (NGS)
- Analysis of ancient genomes sequencing data (Mammoths DNA) using popular software pipeline PALEOMIX can take a month even running it on the powerful computer resource. PALEOMIX include typical set of software used to process NGS data
- We adapted the PALEOMIX pipeline to run it on a distributed computing environment powered by PanDA.
- To run pipeline we split input files into chunks which are run separately on different nodes as separate inputs for PALEOMIX and finally merge output file, it is very similar to what it done by ATLAS to process and to simulate data.
- Using software tools developed initially for HEP and Grid can reduce payload execution time for Mammoths DNA samples from weeks to days.



JEDI Oracle->MySQL migration

- Originally all ID’s were maintained using Oracle sequences
- While migration PanDA to MySQL this concept was replaced by auto increment columns.
- While migrate JEDI to MySQL we met the inconsistency issue



PANDAID	...
1	...
2	...
...	

PANDAID	...
1	...
2	...
...	

PANDAID	...
1	...
2	...
...	

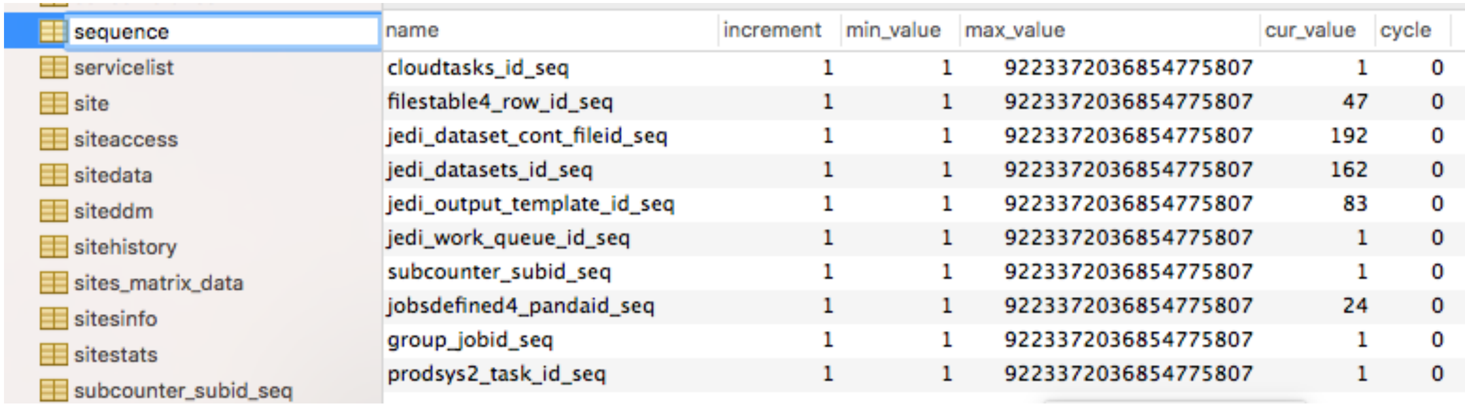
JEDI Oracle->MySQL migration

- Back to the sequences concept
- But as MySQL don't have Sequences we implemented it
- We add new table "sequence" and implemented 2 functions (nextval() and curval()). This allowed to handle sequences in a very similar way how it's done in Oracle.
- All the changes in the code a localized in the WrappedCursor
  - ▫ 1st WrappedCursor realisation:

    Schema_name.SEQUENCE_NAME.nextval -> NULL

  - ▫ WrappedCursor update:

    Schema_name.SEQUENCE_NAME.nextval -> Schema_name.nextval(SEQUENCE_NAME)



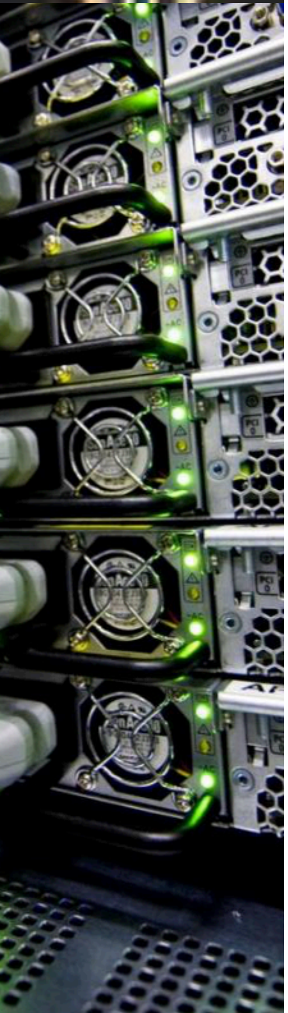
sequence
servicelist
site
siteaccess
sitedata
siteddm
sitehistory
sites_matrix_data
sitesinfo
sitestats
subcounter_subid_seq

name	increment	min_value	max_value	cur_value	cycle
cloudtasks_id_seq	1	1	9223372036854775807	1	0
filestable4_row_id_seq	1	1	9223372036854775807	47	0
jedi_dataset_cont_fileid_seq	1	1	9223372036854775807	192	0
jedi_datasets_id_seq	1	1	9223372036854775807	162	0
jedi_output_template_id_seq	1	1	9223372036854775807	83	0
jedi_work_queue_id_seq	1	1	9223372036854775807	1	0
subcounter_subid_seq	1	1	9223372036854775807	1	0
jobsdefined4_pandaid_seq	1	1	9223372036854775807	24	0
group_jobid_seq	1	1	9223372036854775807	1	0
prodsys2_task_id_seq	1	1	9223372036854775807	1	0

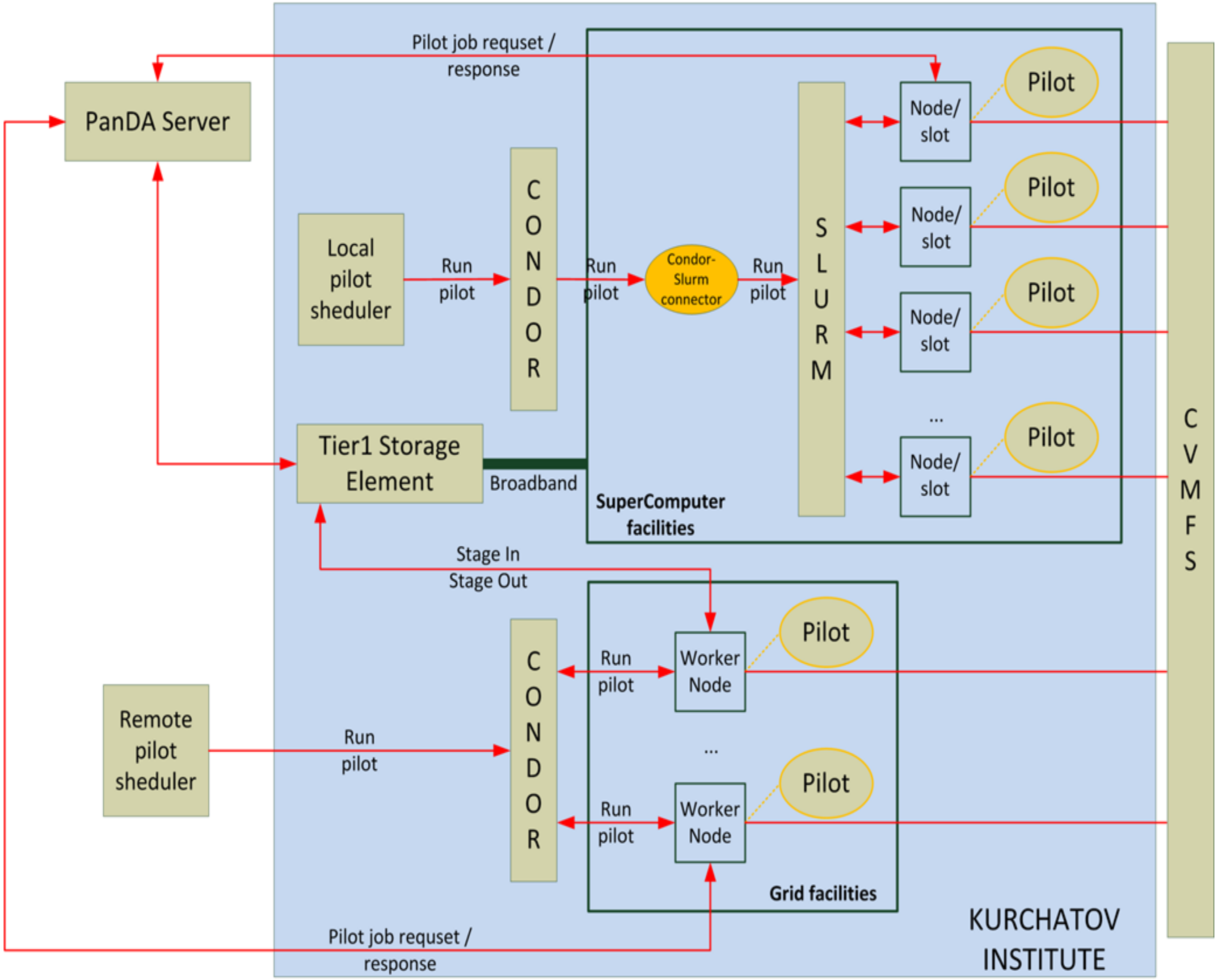
Summary

- Migration to MySQL
  - panda-server **DONE**
  - jedi-server **PROGRESS**
- Update the JEDI installation to make it standard PROGRESS
- Twiki for PanDA@NRC-KI
  - https://twiki.cern.ch/twiki/bin/view/PanDA/BigPanDAforNRCKI
  - Needed to be reviewed and updated
- Setupper & Adder plugins
  - Implemented for NRC KI but not documented
  - Class Setupper – called while receiving a job
    - Checks input files, it weights and checksum.
    - Gets srm-links to the files in given datasets
  - Class Adder – called while pilot sends job ‘finished’(‘failed’) status
    - Files already in grid and we have srm-links
    - Registers output data in Rucio
  - Plugins could be set in panda_server.cfg
- PanDA queue definition in schedconfig table & "cached" json files for pilot
  - AGIS handle it for central ATLAS server
  - Not documented for others

Backup slides



Integration scheme of Russian Tier-1 Grid Center with High Performance Computers at NRC-KI (CHEP15)


Pilot job requset / response
PanDA Server
Local pilot sheduler
Run pilot
CONDOR
Run pilot
Condor-Slurm connector
Run pilot
SLURM
Node/ slot
Pilot
Node/ slot
Pilot
Node/ slot
Pilot
...
Node/ slot
Pilot
Tier1 Storage Element
Broadband
SuperComputer facilities
CVMFS
Stage In
Stage Out
Remote pilot sheduler
Run pilot
CONDOR
Run pilot
Worker Node
Pilot
...
Run pilot
Worker Node
Pilot
Grid facilities
Pilot job requset / response
KURCHATOV INSTITUTE