JAliEn status

Cristian Margineanu, Kalana Wijethunga, Thameera Hettiwatta, Adrian Negru

costin.grigoras@cern.ch

Outline

- JAliEn project overview
- Grid clients
- New monitoring tools
  - ELK for central services
  - Site Sonar
  - Firefox plugin
  - SE file crawler

Deployment status

One year since we have started the transition of Grid sites to JAliEn

70 VoBoxes are running the new service

  2 still to migrate

From them

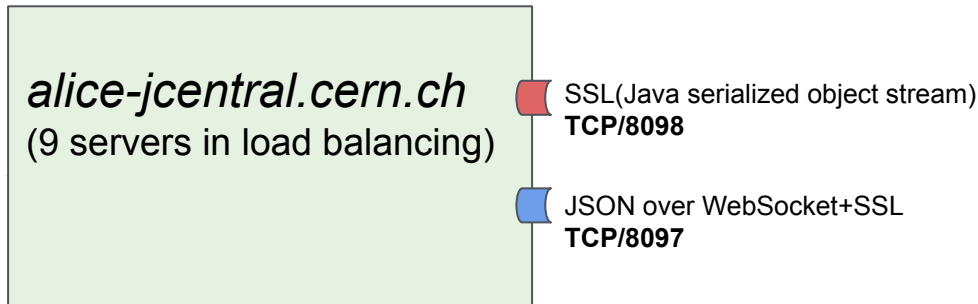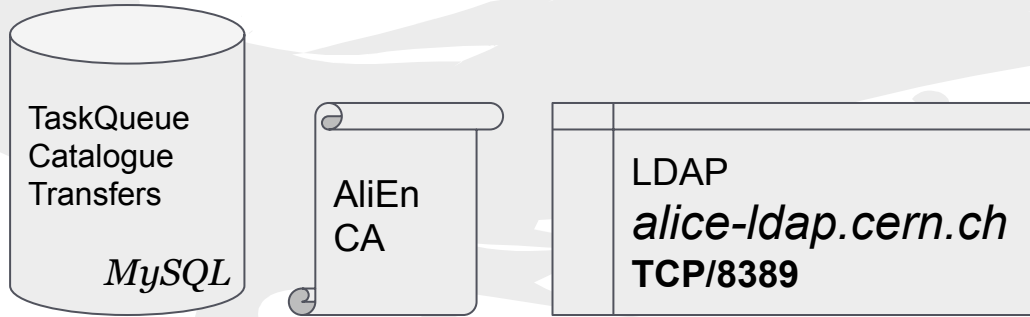 48 single core queues

  21 eight core queues

  1 whole node (HPCS)

~All CentOS7

  With a special CentOS8 deployment on EPNs for GPU support

Central services layout

TaskQueue
Catalogue
Transfers

MySQL

AliEn
CA

LDAP
alice-ldap.cern.ch
TCP/8389

alice-jcentral.cern.ch
(9 servers in load balancing)

SSL(Java serialized object stream)
TCP/8098

JSON over WebSocket+SSL
TCP/8097

Identical client-facing services

Background "optimizers"

Easily scalable

Same *.jar* from CVMFS can be used as server, agent or embedded as a library

Two connection options
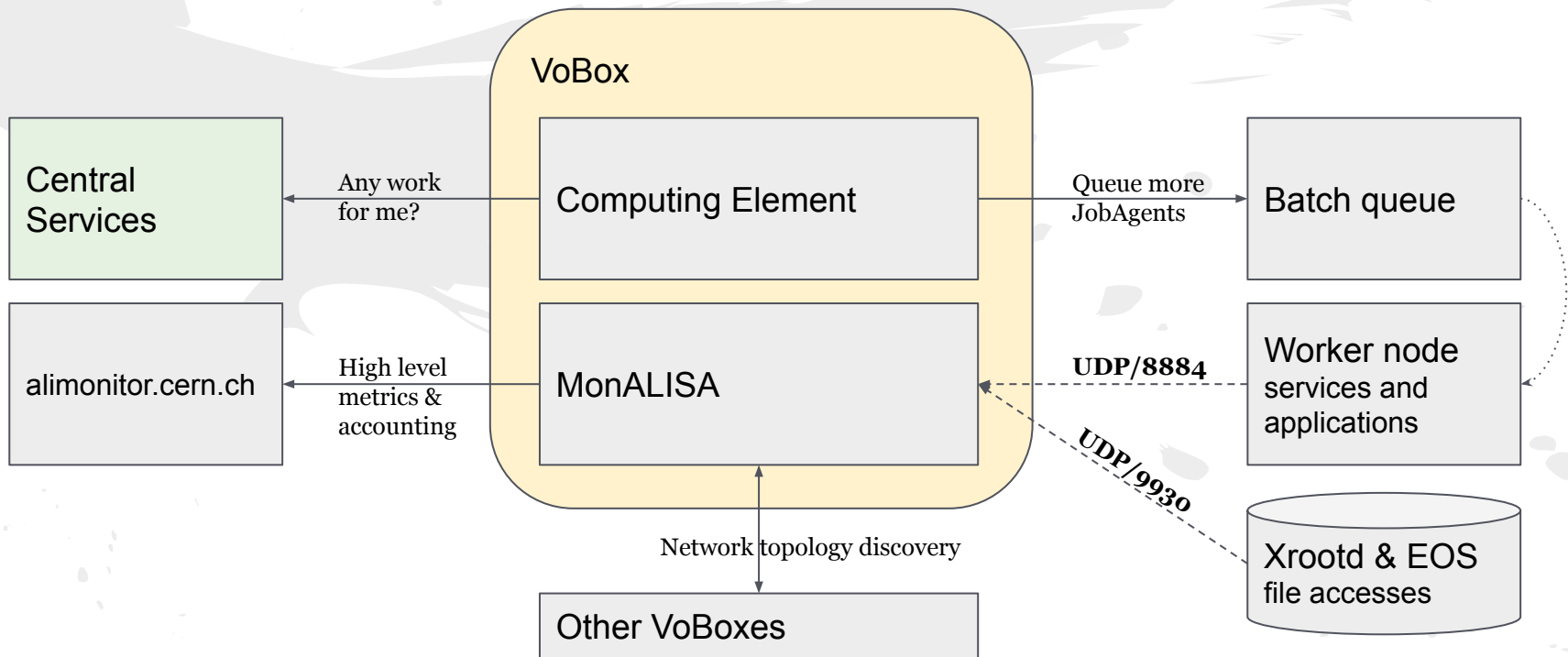- Native Java ObjectStream
- WebSockets + JSON

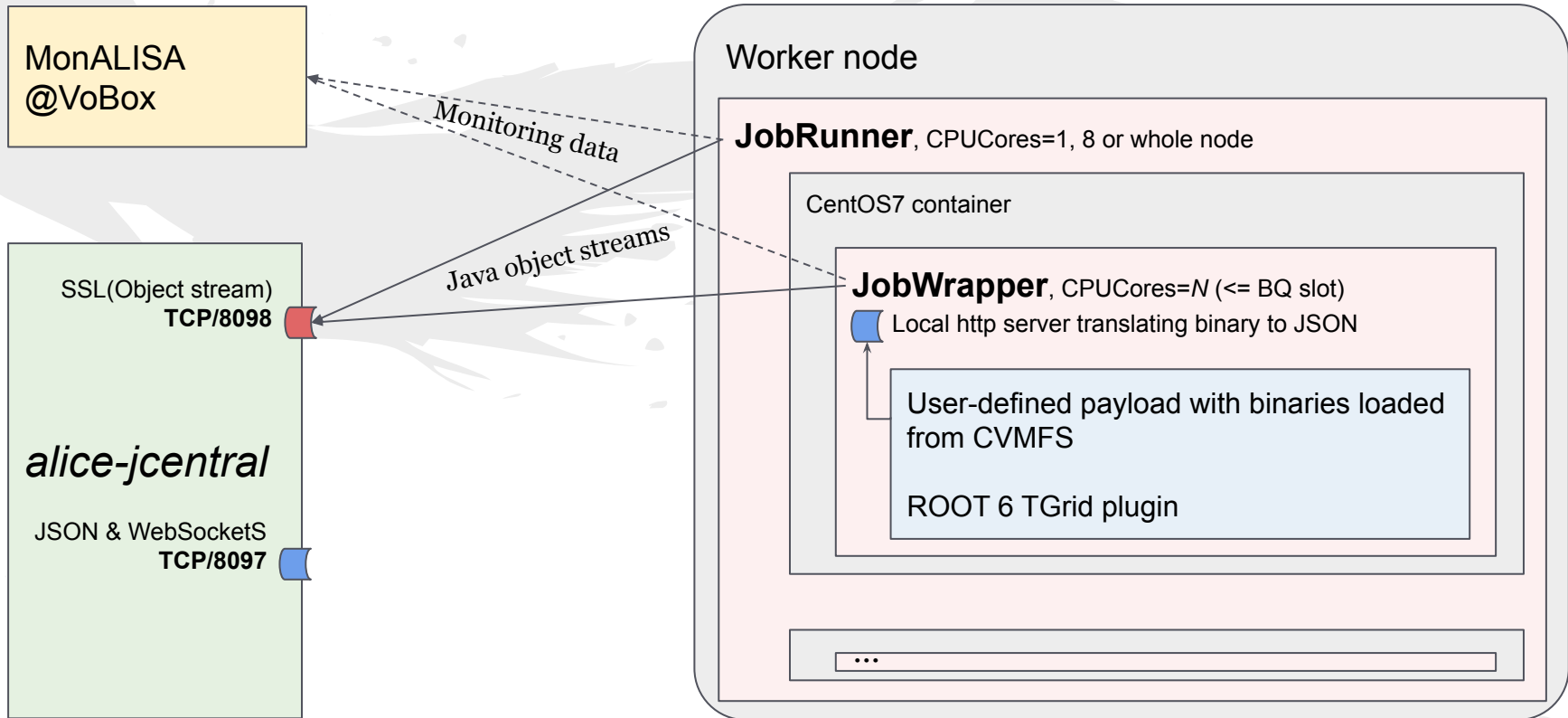Embedding an Apache Tomcat engine

Exclusive use of X.509 authentication

No proxies!

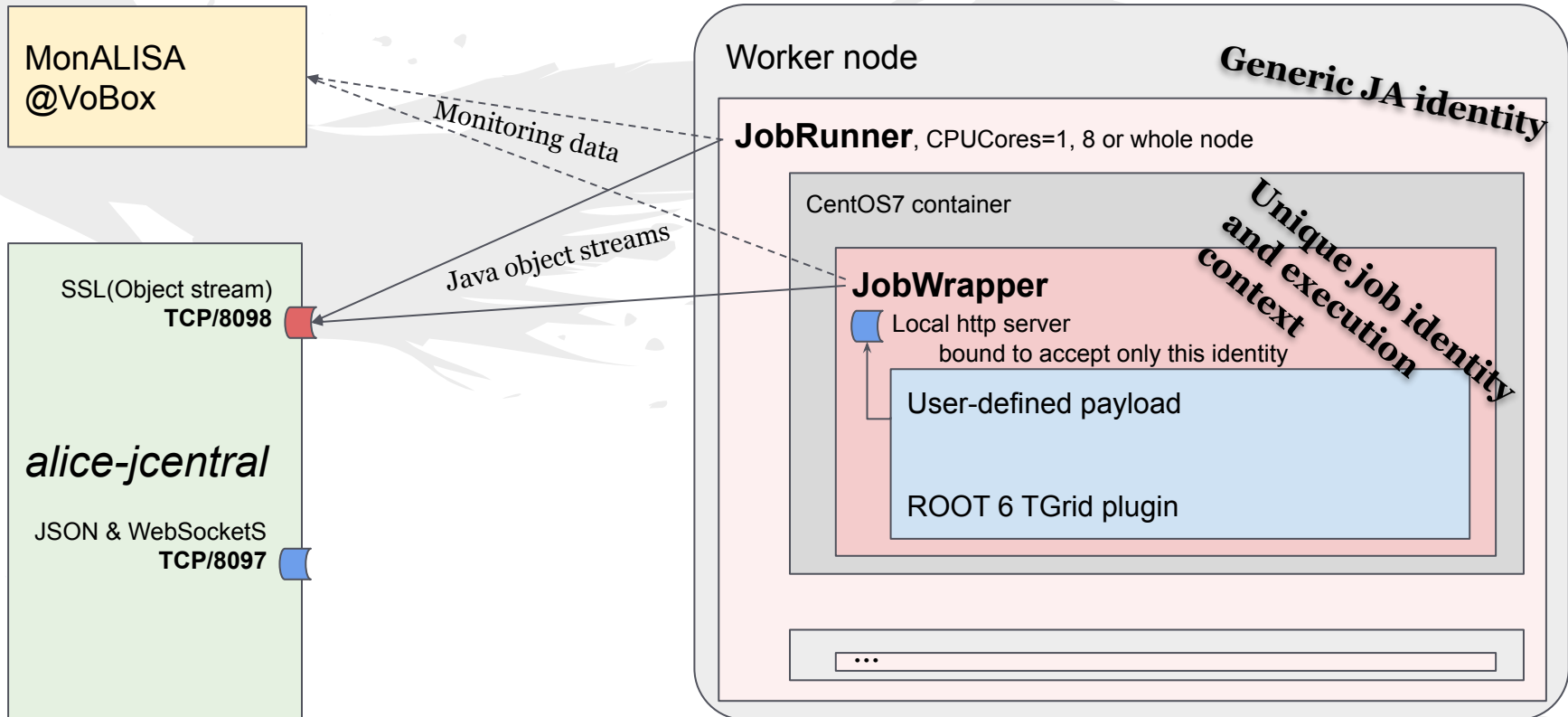Same old envelopes for per-file and per-operation SE access

Site services / VoBox

Central Services

Any work for me?

alimonitor.cern.ch

High level metrics & accounting

VoBox

Computing Element

MonALISA

Queue more JobAgents

Batch queue

UDP/8884

Worker node services and applications

UDP/9930

Xrootd & EOS file accesses

Network topology discovery

Other VoBoxes

Worker node components

MonALISA
@VoBox

alice-jcentral

SSL(Object stream)
TCP/8098

JSON & WebSocketS
TCP/8097

Monitoring data

Java object streams

Worker node

JobRunner, CPUCores=1, 8 or whole node

CentOS7 container

JobWrapper, CPUCores=*N* (<= BQ slot)

Local http server translating binary to JSON

User-defined payload with binaries loaded from CVMFS

ROOT 6 TGrid plugin

...

Worker node components

MonALISA
@VoBox

Monitoring data

Java object streams

SSL(Object stream)
TCP/8098

alice-jcentral

JSON & WebSocketS
TCP/8097

Worker node

Generic JA identity

JobRunner, CPUCores=1, 8 or whole node

CentOS7 container

JobWrapper

Local http server
bound to accept only this identity

Unique job identity
and execution
context

User-defined payload

ROOT 6 TGrid plugin

...

Worker node execution

Configurable slot size (no. of CPU cores)

> Automatically adapt to whole node resources where available

Free resources advertised to the JobBroker

A job can match any number of cores <= slot size

Isolated instances

> Container execution (CentOS7 from CVMFS atm.)
>
> Per job credentials (X.509 pair issued by our CA), with limited capabilities
>
> Constrained resources (Core pinning with *taskset -> cgroups2* when possible)

Continuous monitoring of payload resource usage

> Accounting and preempting them if running over boundaries

Grid clients

Java native: `alienv enter JAliEn`

> And use `$CLASSPATH` as your only dependency

Any other language: the WebSocketS + JSON endpoint

> One command line request -> JSON formatted reply
>
> Default for users: *alice-jcentral.cern.ch:8097*
>
> For jobs: localhost access on `$JALIEN_HOST:$JALIEN_WSPORT`

Use the full Grid certificate, user token or the job token indicated by

> `$JALIEN_TOKEN_KEY / $JALIEN_TOKEN_CERT`

Fully implemented for <u>ROOT</u> and in use since we moved to ROOT6

And as the Python+Xrootd bindings Grid <u>shell</u> (or library)

> `alienv enter xjalienfs`

API usage

ROOT6 payloads initially connected directly the central services

As sites were migrated to JAliEn, they started using the localhost JobAgent endpoint for edge conversion to JSON

Up to 6kHz of requests during IO-intensive analysis periods

All logged in ~real time via Logstash to an ElasticSearch cluster

10
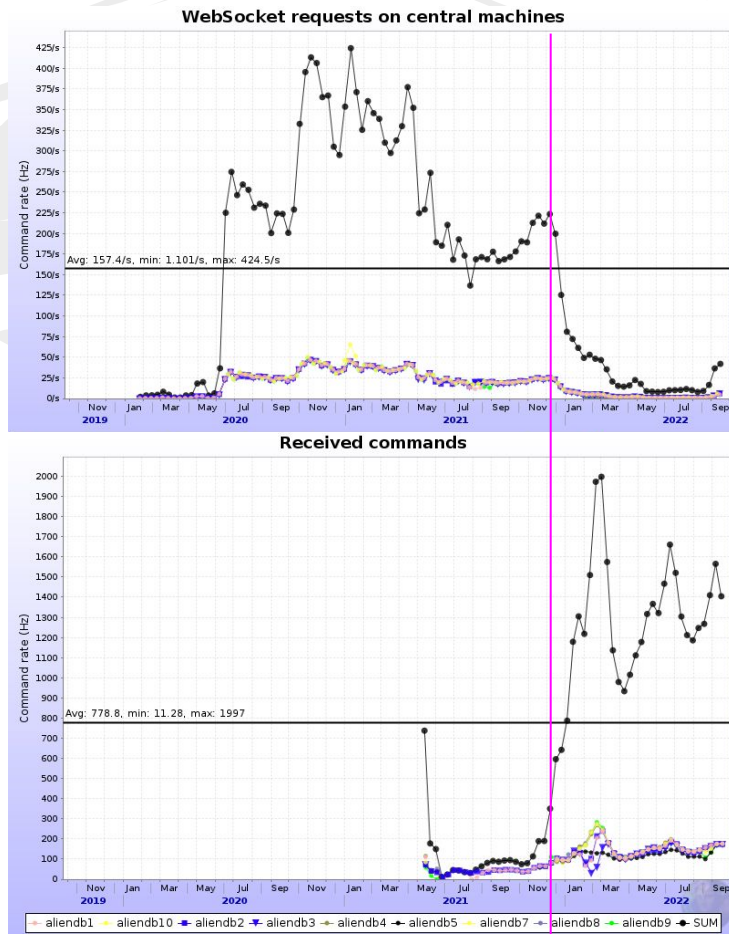
Client Request Analysis Tool for ALICE Grid Services (**Cristian Margineanu**)

Collect, parse and analyze service-level metrics extracted from central services logs

Solution centered around ELK (Elasticsearch, Logstash, Kibana)

Successfully integrated with AliEn, JAliEn and CCDB services

Alerting capabilities for service misuse

Daily report of hot files (i.e. hardcoded *.root* files in user jobs)

Aggregating file access as basis for the production <u>popularity</u> pages

Integration with Site Sonar (data and reports)

Integration with MonALISA (export aggregated metrics)

Schematic view of the relation between JAliEn components, MonALISA and ELK
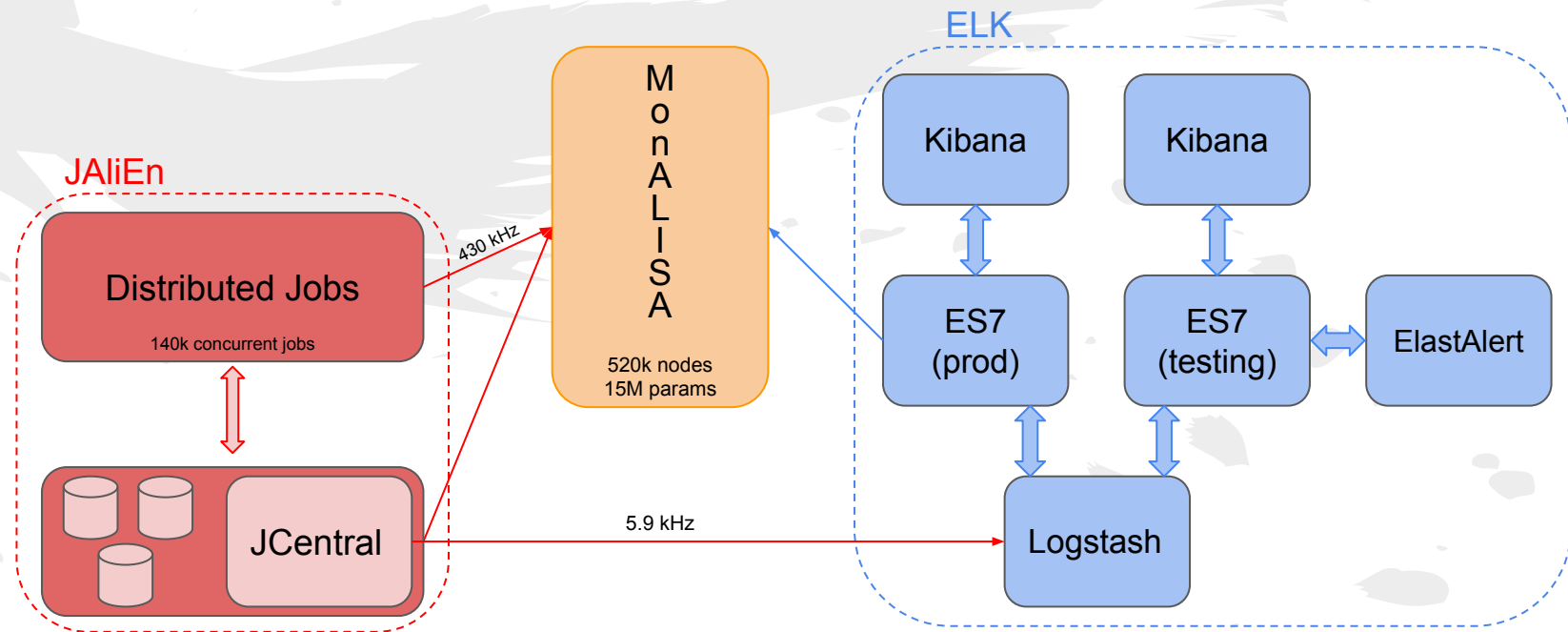
JAliEn Kibana Dashboard
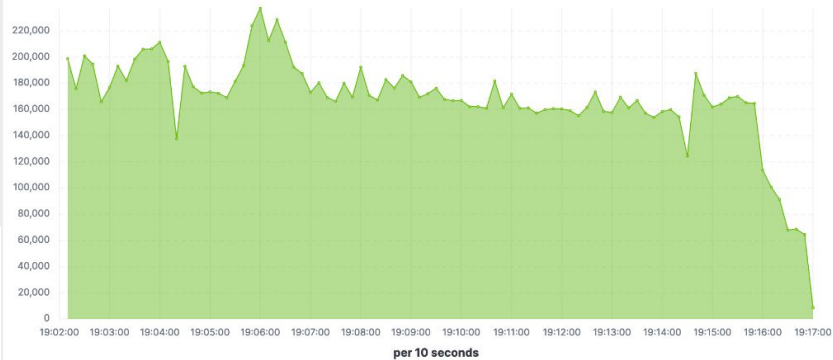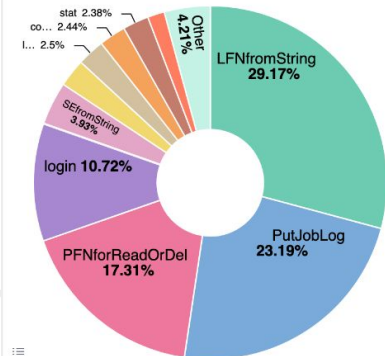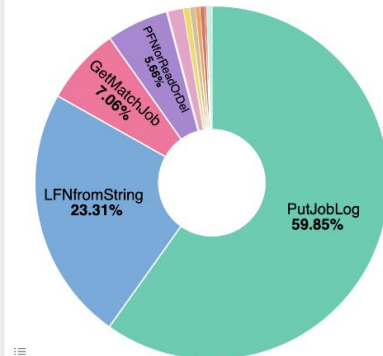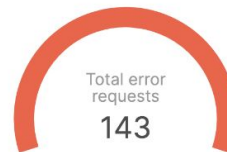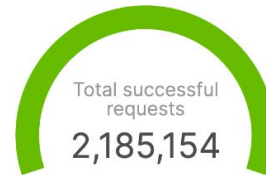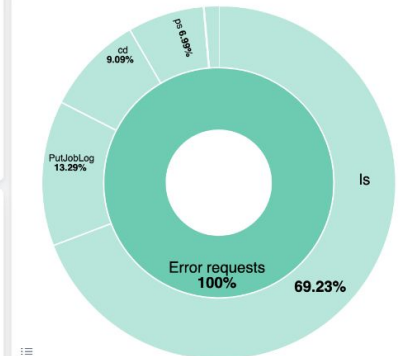
Site Sonar (**Kalana Wijethunga**)

- A flexible and extensible Grid infrastructure monitoring tool
- Reports data from ~10,000 Grid nodes daily
- Invoked at the beginning of the execution of the *JobRunner* to collect the information of the current node and report to Central Services
- Collected information is fed to an ELK stack daily
- Data is monitored, analyzed, and visualized using Elasticsearch and Kibana
- 29 probes currently defined

OS Distribution Dashboard

Architecture

Site Sonar Data Analysis Framework

User → Submit Job → **Central Services**

Grid Administrator ← **Monitoring Dashboards** ← Visualize Data ← **Elasticsearch**

Upload results for analysis

Check for probes to run

Publish results

Job Agent

Site Sonar Data Collection Framework

- Probe 1
- Probe 2
- Probe 3

Run probes → **Sonar**

Report results

Run the job

Job

Legend:
- Site Sonar Data Collection framework
- Site Sonar Data Analysis framework
- Job

16

Features

- Flexibility
  - Unstructured data collection
  - Probes can be updated any time

    `/cvmfs/alice.cern.ch/sitesonar/`

  - Can be used as a Grid debugging tool
- Extensibility
  - Easy to add one more script on the fly, results back in minutes
  - Probing frequency is customizable
- Powerful visualizations
  - No code visualizations with Kibana
  - Many visualization options out of the box

Discoveries

- Some sites were still using CentOS 6
  - Updated to CC 7+ by contacting the admins
- Some sites were reusing hostnames for nodes
  - Agreed to setup a environment variable unique to the node on those sites
- Some sites did not support `singularity`
  - Resolved by contacting the site admins

Few metrics collected

- ~97% of Grid nodes support ~~`singularity`~~ `apptainer`
- All Grid nodes use CC7 or above
  - CentOS Linux release 7.9.2009 (Core) - 90.74%
  - Scientific Linux release 7.9 (Nitrogen) - 6.78%
  - CentOS Linux release 7.7.1908 (Core) - 1.13%
- Only 3.2% of Grid nodes support *cgroups v2*

IPv6 support

SiteSonar WN probe

>7% can't resolve IPv6 and 45% cannot connect on IPv6
Less than half of the nodes are able to use it
Less than half of the sites have dual stacked VoBoxes

All components are IPv6 ready

>Java, Python, Xrootd 4+ (client and server)

We are still running some old ROOT5 jobs

>Only IPv4 for them, due to:

>Xrootd 3 client library, legacy API endpoints

94% of the storage volume is dual stacked

>14PB in 9 sites still to go

JAliEn CS network changes

If you have whitelisted the services, please update the rules:

IPv4:
    137.138.47.192/26   -> **128.141.25.192/26**
    137.138.99.128/26   -> **128.141.26.0/26**
    188.184.2.0/26

IPv6:
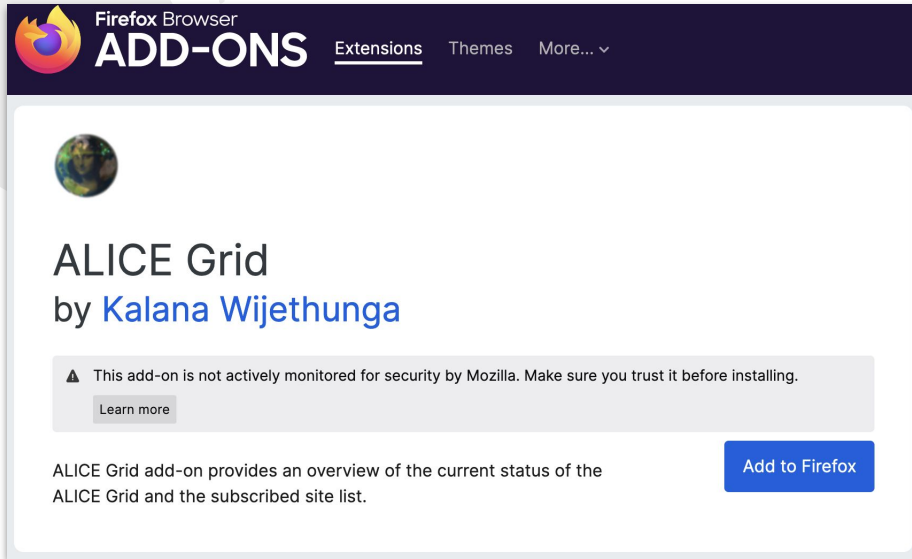
    2001:1458:201:22::/64
    2001:1458:201:b49f::/64
    2001:1458:201:b50e::/64

ALICE Grid Extension

- **Kalana Wijethunga & Thameera Hettiwatta**

- "ALICE Grid" browser extension installable via the Firefox Add-ons store

- Monitor Grid and Site parameters such as active jobs, disk usage, job efficiency and network usage

- Receive alert notifications on configured Sites

- Surveil active job composition on Grid

How to Install the Extension

Click "Add to Firefox" to install the extension

An ALICE Grid Certificate is required for the extension to function properly.

If the certificate is already installed in your Firefox browser, please ignore the pop-up and click "OK"
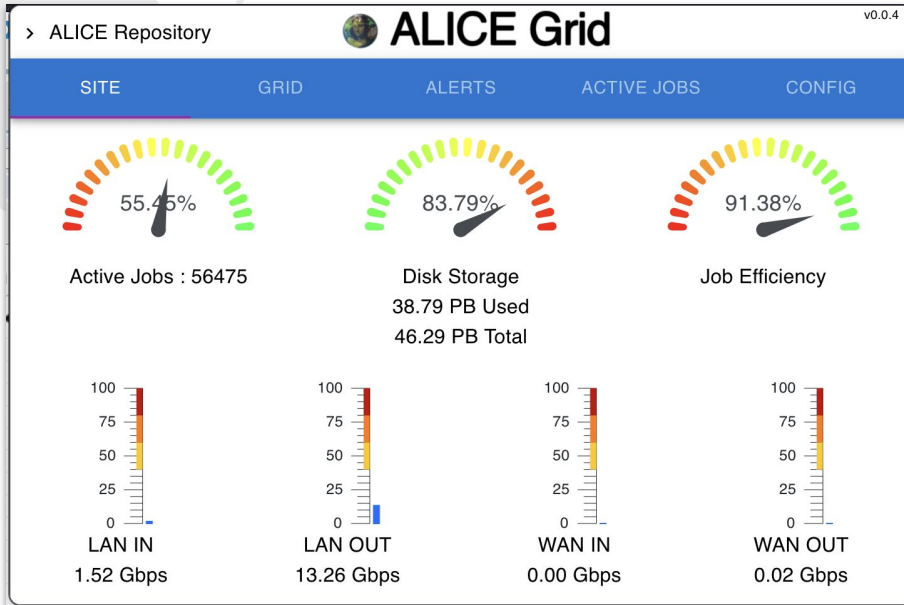
Configure the Extension

Comma separated list of sites in "Config" tab in the extension then "Save" to filter by sites
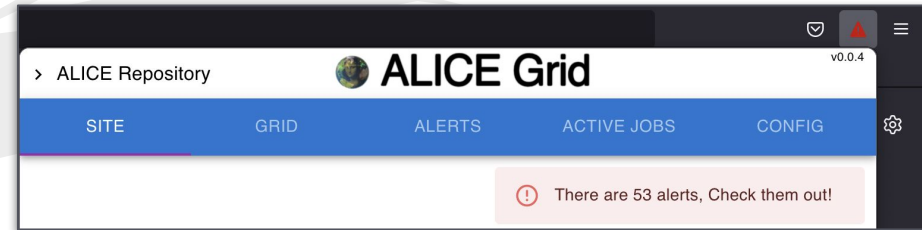
The extension icon will turn "Red" if alerts are available for the configured sites.

Features

Filter by site metrics available on "Site" tab
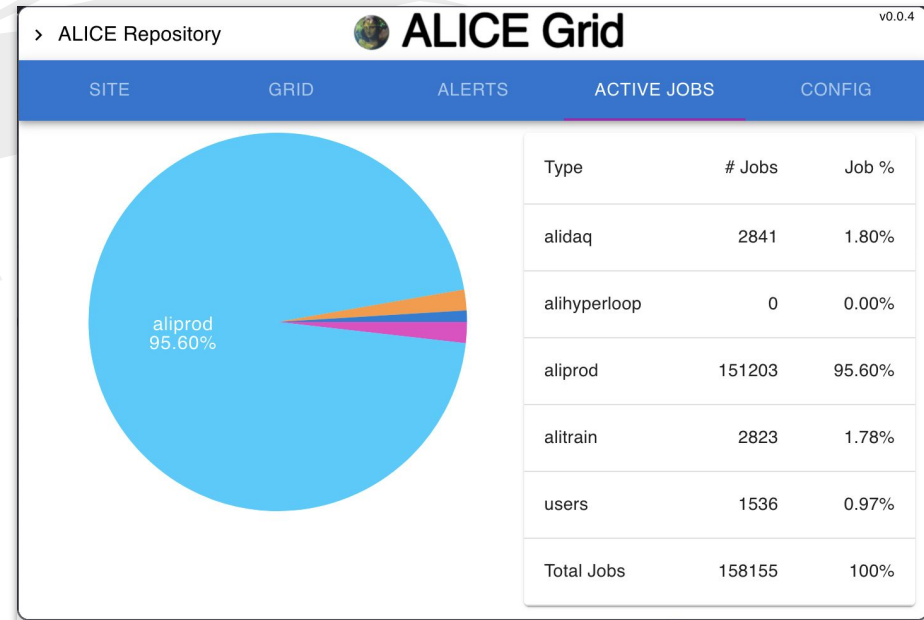

Overall grid metrics available on "Grid" tab

Features

Alerts filtered by configured sites are tabulated on "Alerts" tab

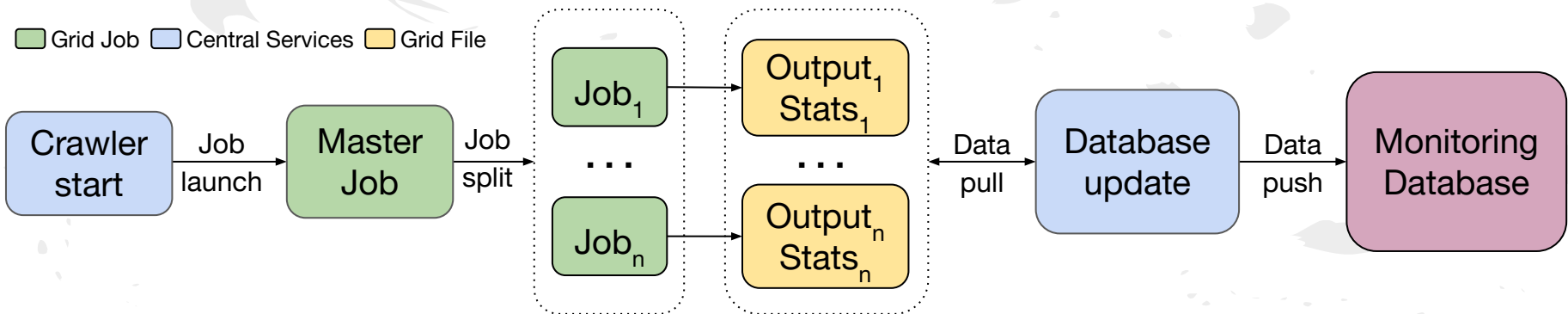Active job composition of the grid viewable on "Active Jobs" tab

File crawler (**Adrian Negru**)

File catalogue → Extract PFNs → File Crawler

File Crawler → Recompute MD5 and compare to catalogue → **Storage element**

File Crawler → File download => I/O performance → **Storage element**

File Crawler → **Collects fault causes and groups by category**

File Crawler → Output analysis Data upload → **MonALISA monitoring**
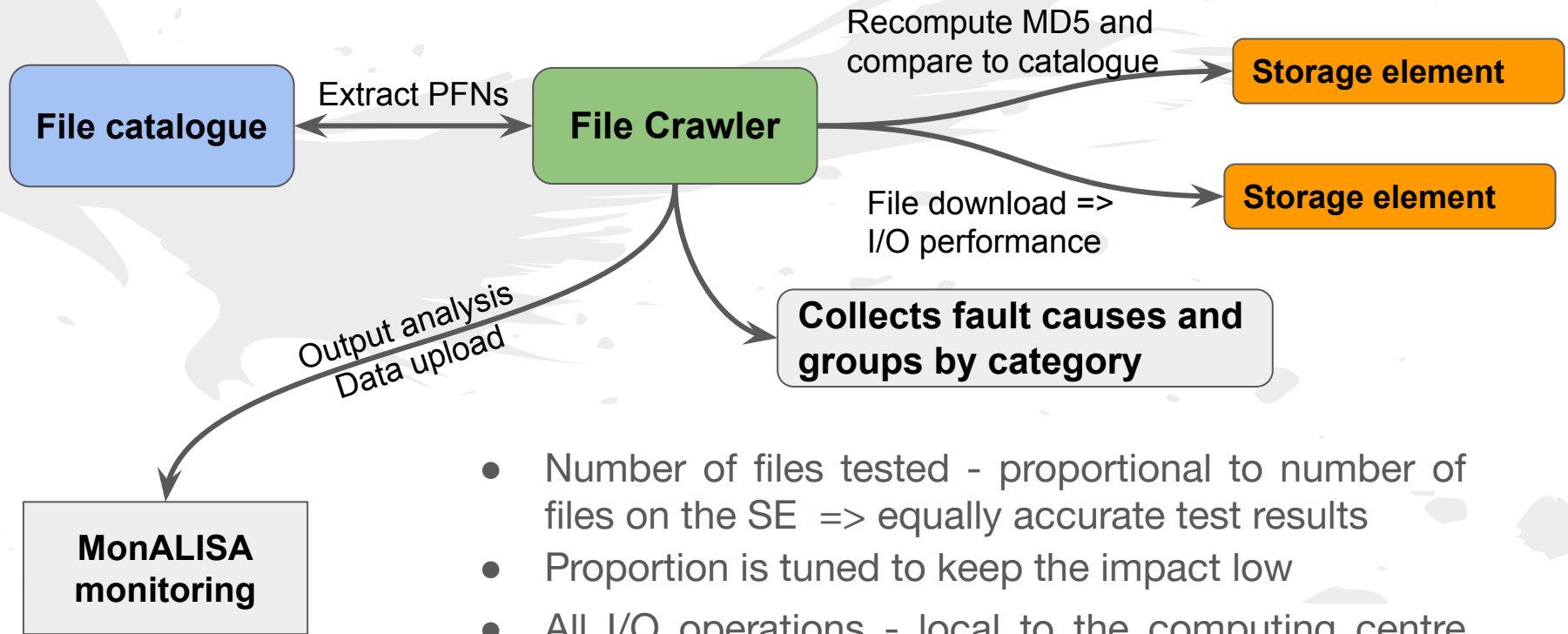
- Number of files tested - proportional to number of files on the SE => equally accurate test results
- Proportion is tuned to keep the impact low
- All I/O operations - local to the computing centre (regular Grid jobs, Java-based)

Architecture

- Periodically launch crawling jobs to the Grid

- Jobs are split according to data location (per SE)

- Completed crawling for an SE triggers analysis and subsequent corrective action + results are inserted into monitoring database

- The workflow below is applied in parallel for all SEs
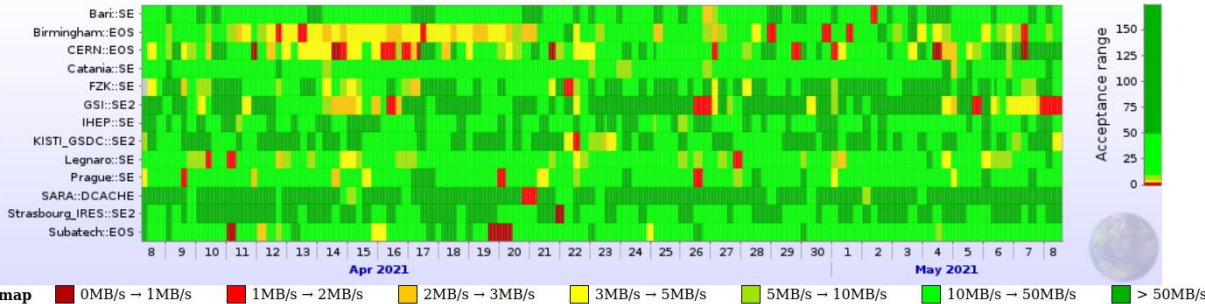
Crawler menu in ALIMonitor

Health metrics

Metrics are attached to a SE and time interval. Supported metrics:

- Success ratio = the number of files that passed all checks / all files analysed

- Inaccessible ratio = the number of files determined as inaccessible / all files analysed

- Corrupt ratio = the number of corrupted files / all files analysed

- Internal error ratio = the number of files with unknown status code / all files analysed

- Download throughput = the average download throughput for a downloaded PFN

- Xrdfs duration = the average `xrdfs stat` call duration for a PFN

Download throughput analysis

Download throughput

Color map: ■ 0MB/s → 1MB/s ■ 1MB/s → 2MB/s ■ 2MB/s → 3MB/s ■ 3MB/s → 5MB/s ■ 5MB/s → 10MB/s ■ 10MB/s → 50MB/s ■ > 50MB/s

Statistics

Link name	Data		Average
	Starts	Ends	Download throughput
Bari::SE	08 Apr 2021 05:14	08 May 2021 09:23	39.97 MB/s
Birmingham::EOS	08 Apr 2021 05:15	08 May 2021 09:21	14.68 MB/s
CERN::EOS	08 Apr 2021 05:19	08 May 2021 09:43	27.37 MB/s
Catania::SE	07 Apr 2021 23:03	08 May 2021 09:25	26.80 MB/s
FZK::SE	08 Apr 2021 11:25	08 May 2021 09:21	45.23 MB/s
GSI::SE2	08 Apr 2021 05:13	08 May 2021 09:14	58.79 MB/s
IHEP::SE	08 Apr 2021 05:15	08 May 2021 09:23	44.28 MB/s
KISTI_GSDC::SE2	08 Apr 2021 05:17	08 May 2021 09:08	46.00 MB/s
Legnaro::SE	08 Apr 2021 05:16	08 May 2021 09:06	23.10 MB/s
Prague::SE	08 Apr 2021 05:17	08 May 2021 09:18	26.75 MB/s
SARA::DCACHE	08 Apr 2021 05:18	08 May 2021 09:19	95.15 MB/s
Strasbourg_IRES::SE2	08 Apr 2021 05:16	08 May 2021 09:10	49.27 MB/s
Subatech::EOS	08 Apr 2021 05:15	08 May 2021 09:18	32.43 MB/s

- Average throughput per client for multiple crawler iterations (plot shows 1 month of data)
- ■ and ■ - below optimal (5MB/s) throughput => typical reason is heavy analysis load
- Collected data to be used to optimize analysis patterns

31

Success ratio color map

Xrdfs duration color map

Status code distribution for a SE

Status codes extracted from the crawler

SE Name: ALICE::BARI::SE   **Interval:** Last week

Status Type	Status Code	Status Count	Status Code Ratio	Download throughput
FILE_OK	S_FILE_CHECKSUM_MATCH	58729	99.23 %	79.88 Mb/s
	E_CATALOGUE_MD5_IS_BLANK	399	0.67 %	81.67 Mb/s
FILE_INACCESSIBLE	XROOTD_EXITED_WITH_CODE	22	0.04 %	
	XROOTD_TIMED_OUT	8	0.01 %	
FILE_CORRUPT	MD5_CHECKSUMS_DIFFER	29	0.05 %	
TOTAL				

PFNs analysed by the crawler

	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE		
PFN	**SE Name**	**Status Name**	**Status Type**	**Size (B)**
root://alicegrid2.ba.infn.it:1094//13/48298/c67d79c4-215b-11e6-8f31-5391d6a2eadd	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	18.83 MB
root://alicegrid2.recas.ba.infn.it:1094//09/33645/e9932dc4-870f-11e9-964a-37248bf6e634	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	8.627 MB
root://alicegrid2.recas.ba.infn.it:1094//09/17470/bc3c5c58-871c-11e9-9d67-f77a0479f22f	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	6.861 MB
root://alicegrid2.recas.ba.infn.it:1094//03/26474/af09ebf6-871a-11e9-9a5b-976bc5c74b72	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	8.576 MB
root://alicegrid2.recas.ba.infn.it:1094//14/15403/9f7a6f86-870f-11e9-a601-3f7c1b09855e	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	8.376 MB
root://alicegrid2.recas.ba.infn.it:1094//09/21791/82436150-8711-11e9-a6bc-2bf182dfc02a	ALICE::BARI::SE	XROOTD_EXITED_WITH_CODE	FILE_INACCESSIBLE	12.07 MB

Summary

The new Grid framework is in production

    Multicore support, execution isolation

Extended monitoring of payloads

    Worker node probes => flexibility in matching jobs

    TTL prediction for MC jobs

New tools to ensure smooth operations

    SE crawler, Firefox toolbar, hot file access alerts

Open communication protocols

    Embeddable Python and Java clients