Grid services and relations
+
practical session and predominant problems

T1/T2 tutorial, May 26, 2009

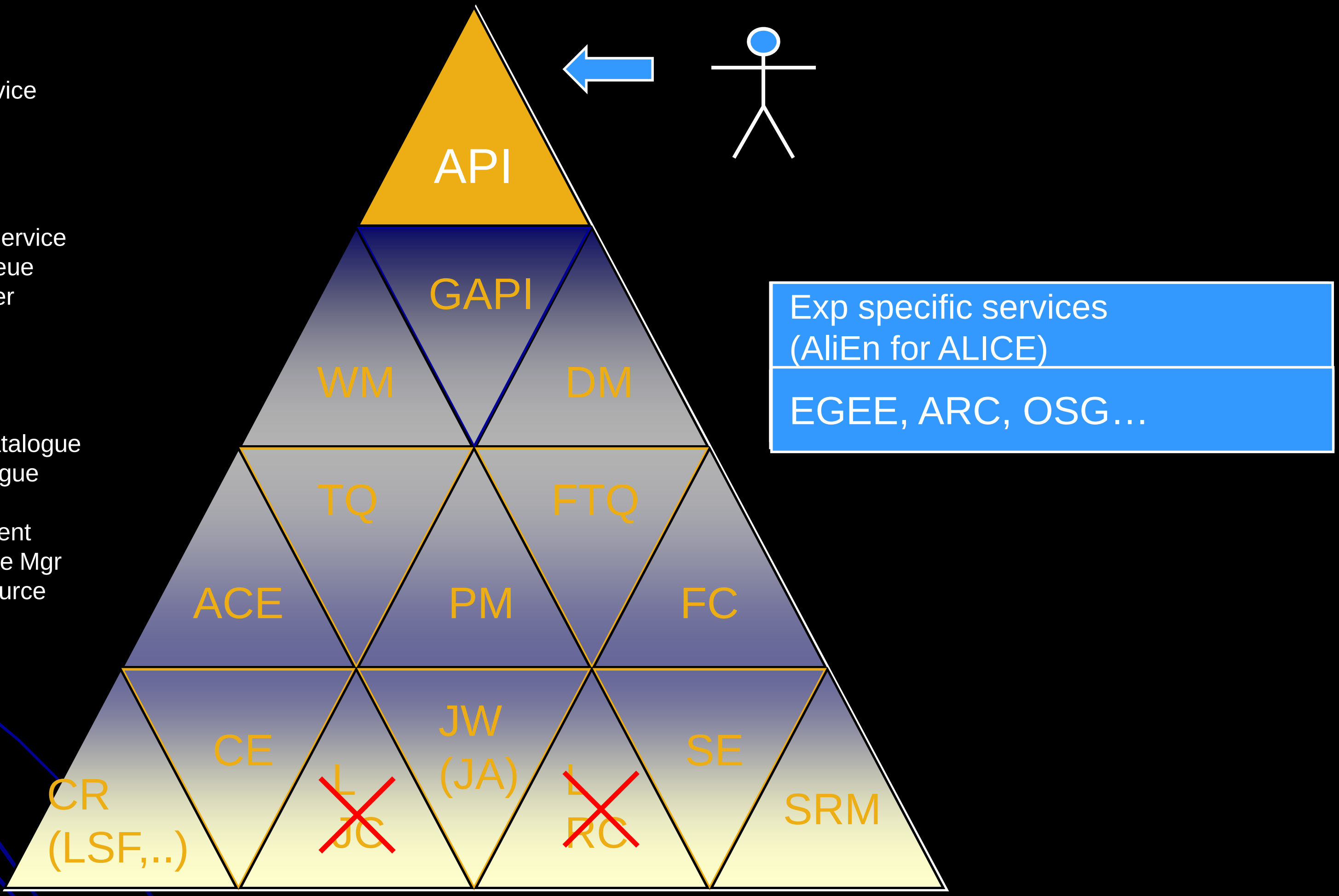
L. Betev

Outline

- General services structure
- Design
- Job management
- File catalogue
- Storage strategy
- Storage security
- Job flow – practical session
- Typical failures
- Summary

The services pyramid (how it was)

GAS	Grid Access Service
WM	Workload Mgmt
DM	Data Mgmt
RB	Resource Broker
TQ	Task Queue
FPS	File Placement Service
FTQ	File Transfer Queue
PM	Package Manager
ACE	AliEn CE (pull)
FC	File Catalogue
JW	Job Wrapper
JA	Job Agent
LRC	Local Replica Catalogue
?	Local Job Catalogue
SE	Storage Element
CE	Computing Element
SRM	Storage Resource Mgr
CR	Computing Resource (LSF, PBS,…)



Exp specific services (AliEn for ALICE)

EGEE, ARC, OSG…

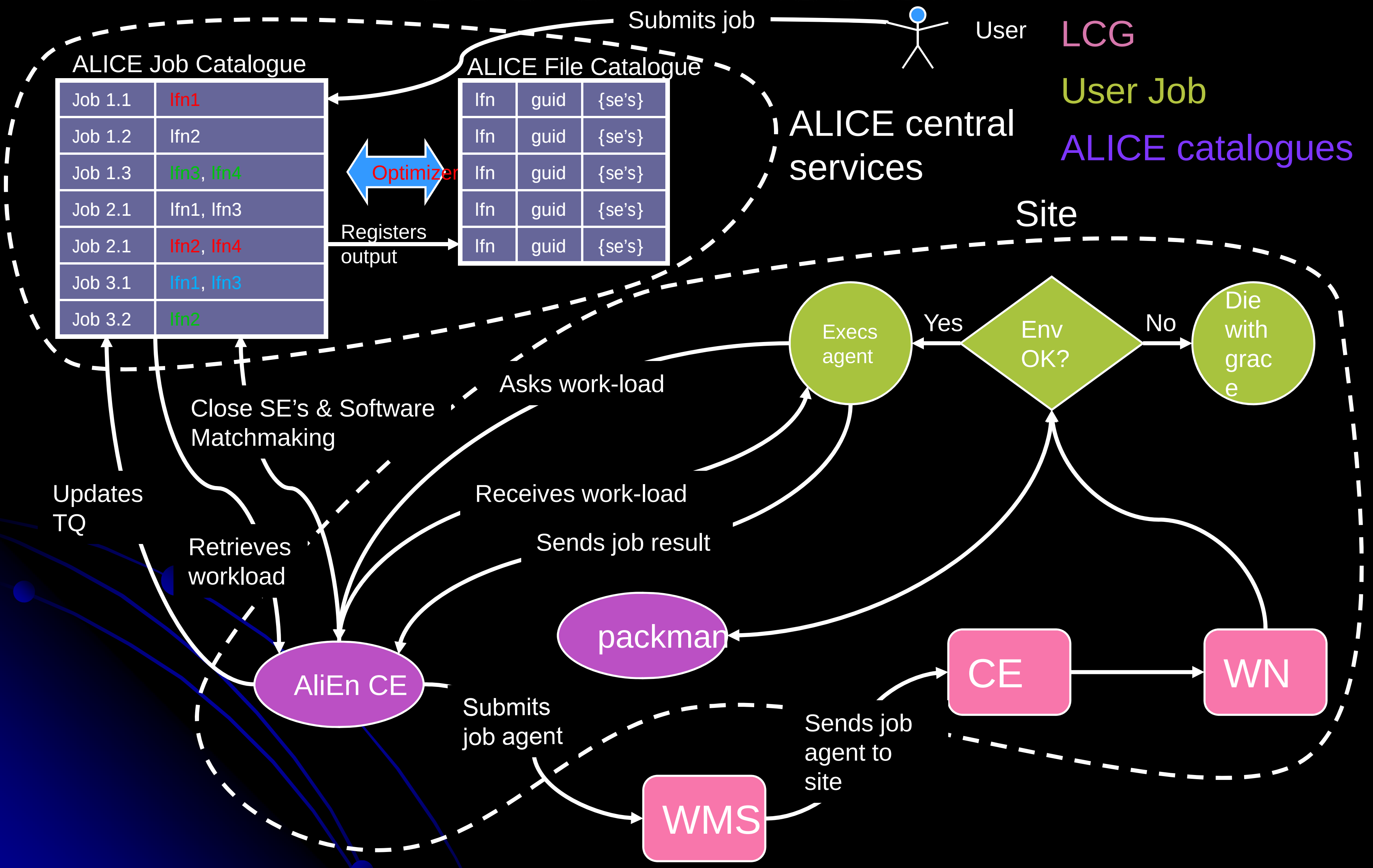
Design criteria

- Minimize intrusiveness
  - Limit the impact on the host computer centres
- Use delegation
  - Where possible acquire “capability” to perform operation, no need to verify operation mode at each step
- Centralise information
  - Minimise the need to “synchronise” information sources
- Decentralise decisions
  - Minimise interactions and avoid bottlenecks
- Virtualise resources
- Automatise operations
- Provide extensive monitoring

Job submission

- Minimize intrusiveness
  - Job submission is realised using existing Grid MW if possible or directly to CE otherwise
- Centralise information
  - Jobs are held in a single central queue handling priorities, and quotas
- Decentralise decisions
  - Sites decides which jobs to “pull”
- Virtualise resources
  - Job agents are run to providing a standard environment (job wrapper) across different systems
- Automatise operations
- Provide extensive monitoring

Job submission



ALICE Job Catalogue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

ALICE File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

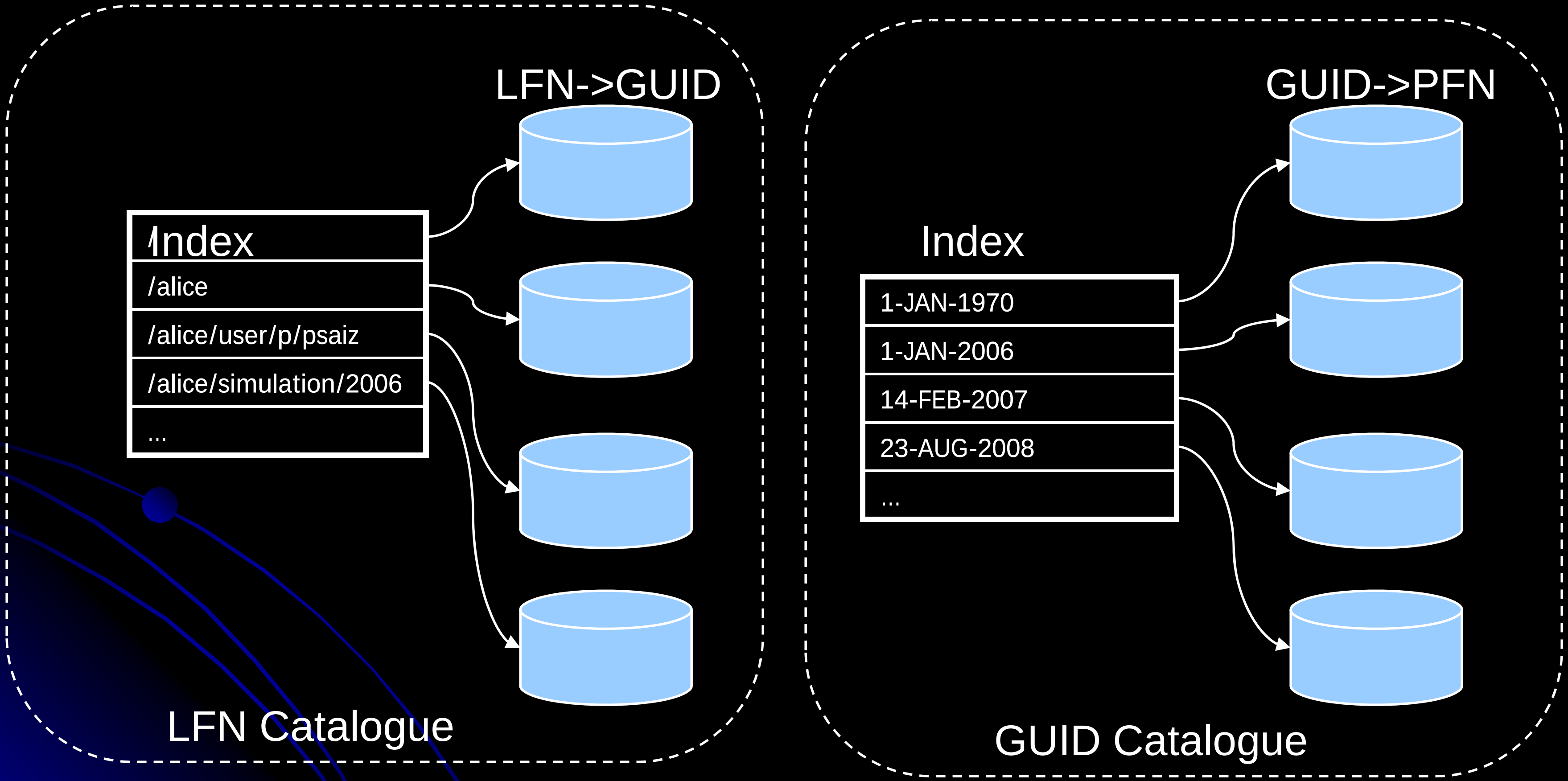
The AliEn FC

- Hierarchical structure (like a UNIX File system)
- Designed in 2001
  - Provides mapping from LFN to PFN
  - Built on top of several databases
    - Possible to add another database to expand the catalogue namespace
  - Possible to move directories to another table
    - Transparent for the end user
  - Metadata catalogue on the LFN
  - Triggers
  - GUID to PFN mapping in the central catalogue
    - No “local catalogue”
  - Possibility of automatic PFN construction (in use extensively now)
    - Store only the GUID and Storage Index and the SE builds the PFN from the GUID
  - Two independent catalogues: LFN->GUID and GUID->PFN
    - Possible to add databases to one or the other
    - We could drop LFN->GUID mapping if not used anymore

Other features

- Size
  - LFN tables: 130 bytes/entry
  - Binary log files: 1000 bytes/entry!
    - Needed for database replication (in use extensively now)
    - Automatically cleaned
  - The current database could contain 7.5 billion entries!
- Two QoS for SE
  - Custodial: File has low probability of disappearing
  - Replica: File has high probability of disappearing
  - User specifies QoS when registering a file
- Still to do: quotas: disk and job
- Entries in the LFN catalogue can have expiration time
  - The entry will disappear regardless of QoS of SE and is removed from storage
  - A GUID not referenced by any LFN will also disappear

File Catalogue



LFN Catalogue

LFN->GUID

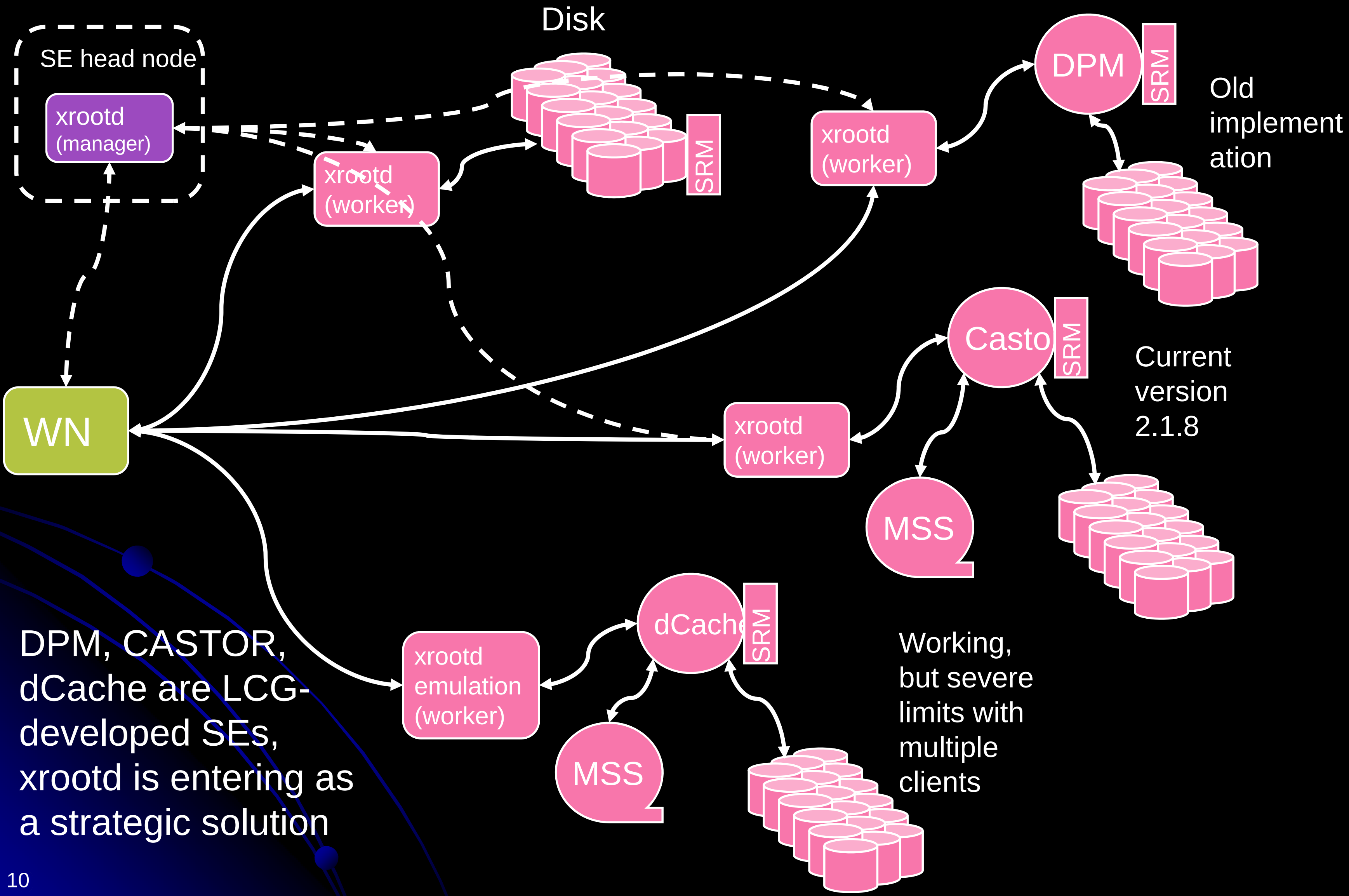
Index
/alice
/alice/user/p/psaiz
/alice/simulation/2006
...

GUID Catalogue

GUID->PFN

Index
1-JAN-1970
1-JAN-2006
14-FEB-2007
23-AUG-2008
...

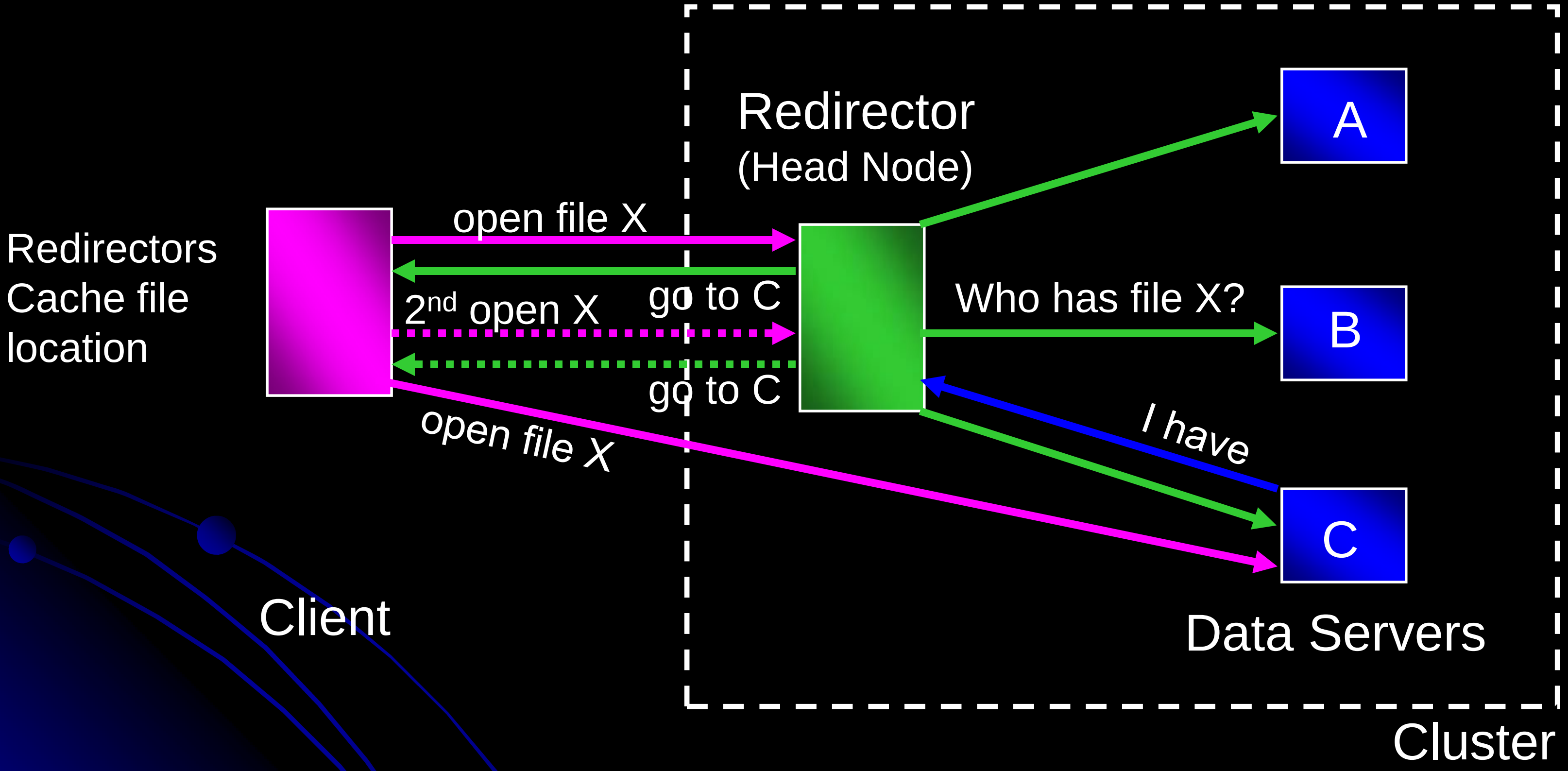
Storage strategy



DPM, CASTOR, dCache are LCG-developed SEs, xrootd is entering as a strategic solution

Xrootd architecture

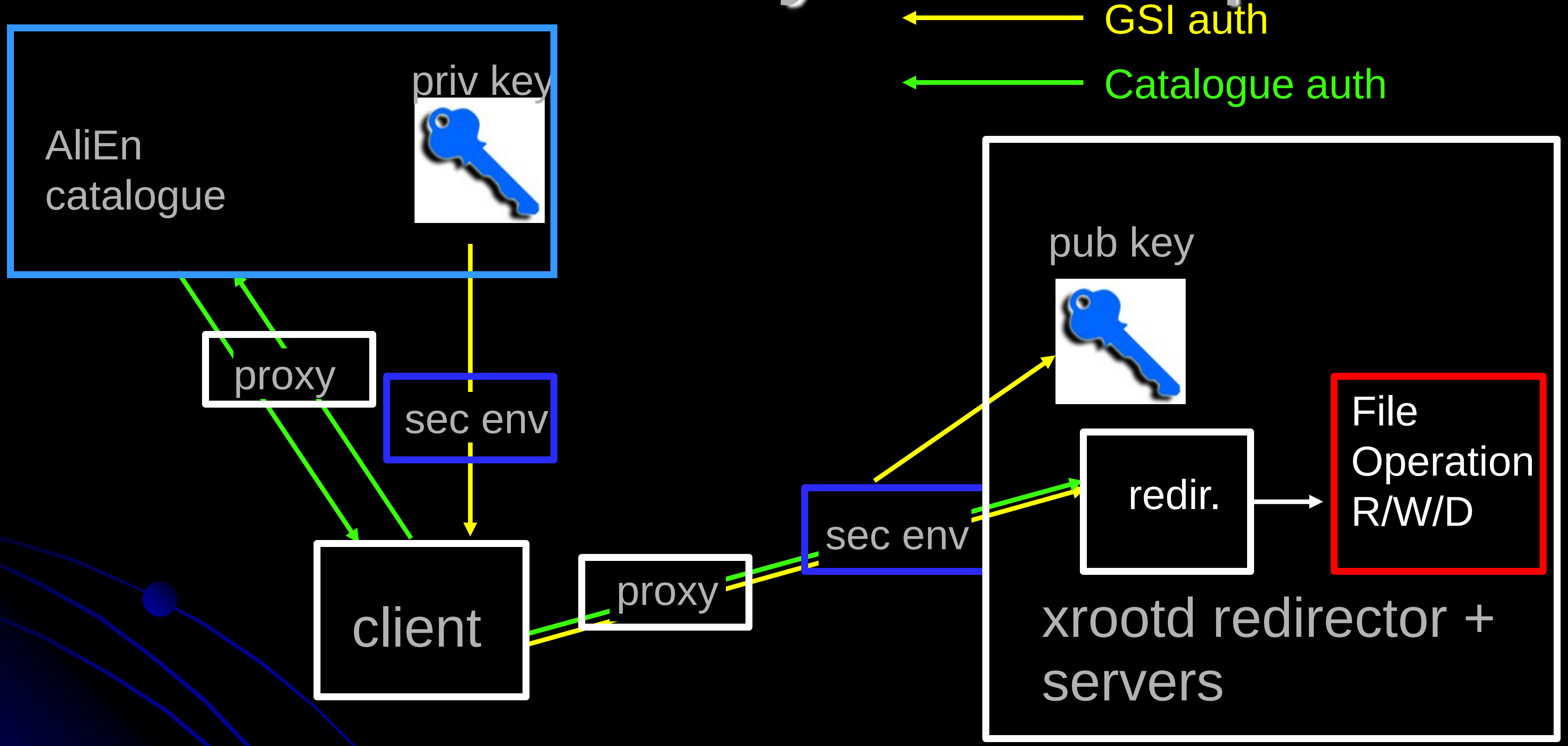
Global redirector (not in picture) – intra-site storage collaboration



Client sees all servers as xrootd data servers

All storages are on WAN

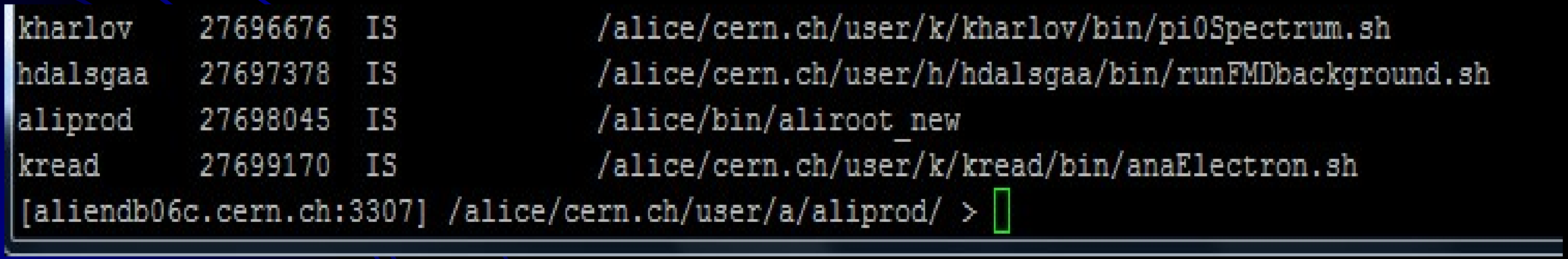
xrootd security - envelope



Services – practical session

Services in action – job flow

- Single task queue – general query with ‘ps –a’
  - list of all active ***master*** jobs (with one or more sub-jobs)
  - submitter, jobID, status, executable
  - more details with ‘masterJob <job ID>’

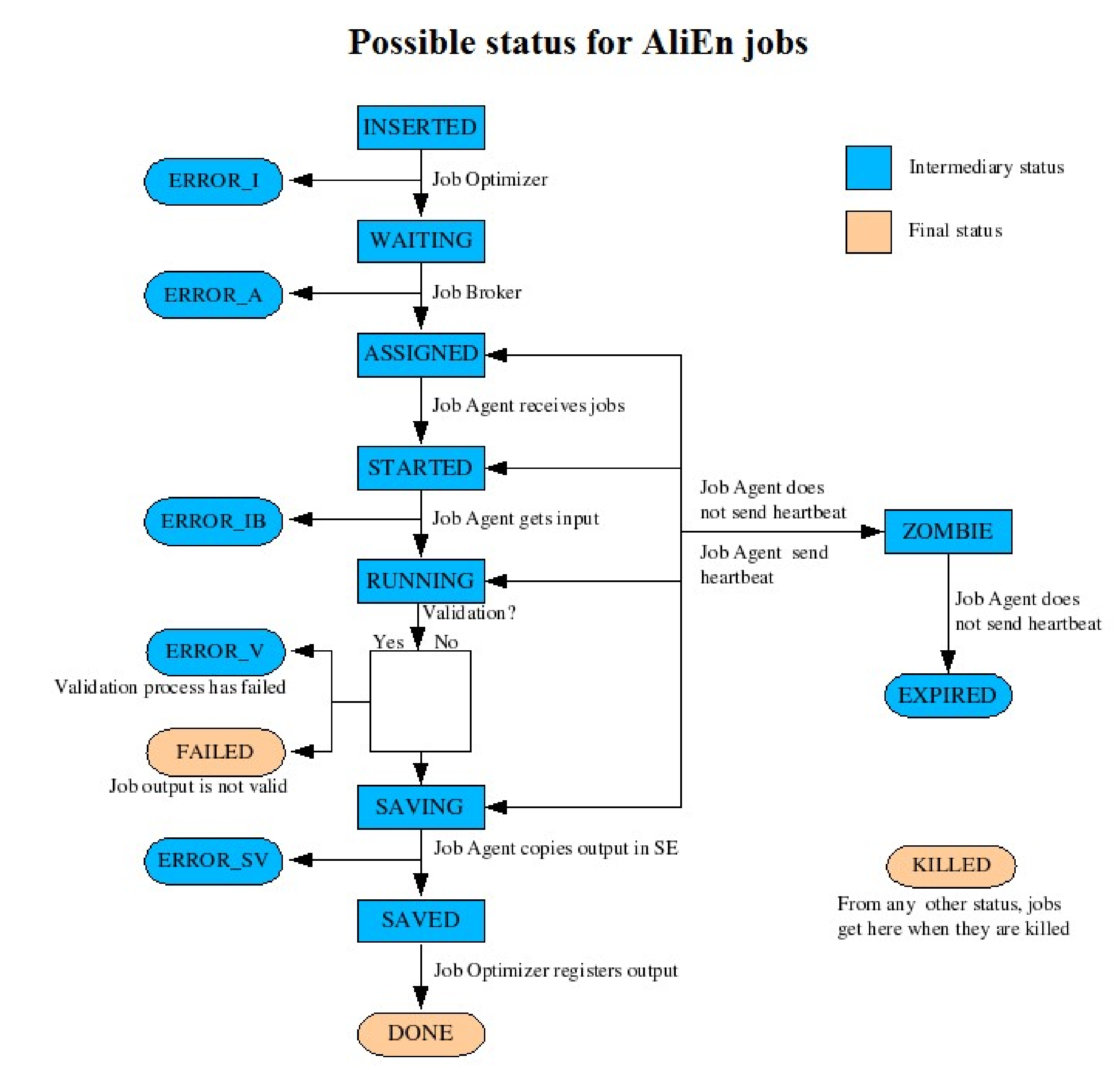


```
kharlov    27696676   IS            /alice/cern.ch/user/k/kharlov/bin/pi0Spectrum.sh
hdalsgaa   27697378   IS            /alice/cern.ch/user/h/hdalsgaa/bin/runFMDbackground.sh
aliprod    27698045   IS            /alice/bin/aliroot_new
kread      27699170   IS            /alice/cern.ch/user/k/kread/bin/anaElectron.sh
[aliendb06c.cern.ch:3307] /alice/cern.ch/user/a/aliprod/ >
```

Job flow for the site admin

- What is running on my site
  - top –status RUNNING –site <site name>
  - top –status SAVING –site <site name>
  - QUEUED, ERROR_V, etc.. (error codes on the next slide and here)
- The above command is listing all active sub-jobs (ps –a lists the master jobs)
- How to get the site name
  - Conveniently displayed in the hat of 'alien login' -> CE = <site name>

AliEn job status chart



Possible status for AliEn jobs

INSERTED

Job Optimizer

ERROR_I

WAITING

Job Broker

ERROR_A

ASSIGNED

Job Agent receives jobs

STARTED

Job Agent gets input

ERROR_IB

RUNNING

Validation?

Yes

No

ERROR_V

Validation process has failed

FAILED

Job output is not valid

SAVING

Job Agent copies output in SE

ERROR_SV

SAVED

Job Optimizer registers output

DONE

Job Agent does not send heartbeat

Job Agent send heartbeat

ZOMBIE

Job Agent does not send heartbeat

EXPIRED

Intermediary status

Final status

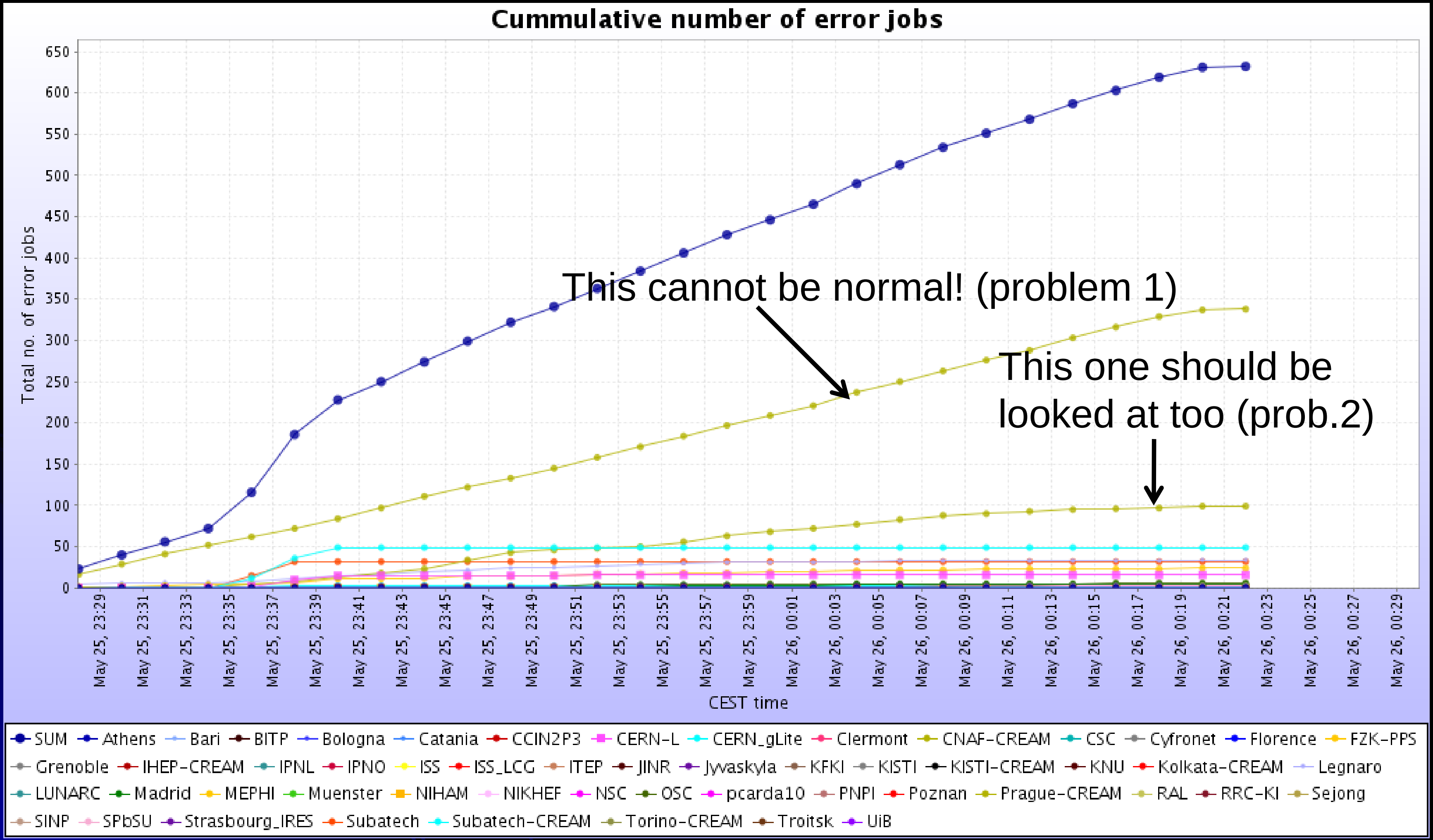
KILLED

From any other status, jobs get here when they are killed

Job flow – comparison methods

- The best method to diagnose local/global problem is to use MonALISA job monitoring information
  - Trace the errors!
  - More detailed view – see Costin’s presentation
  - Get a feeling for the predominant error on the site and compare with other sites
    - Taking into account the site size (CPUs)

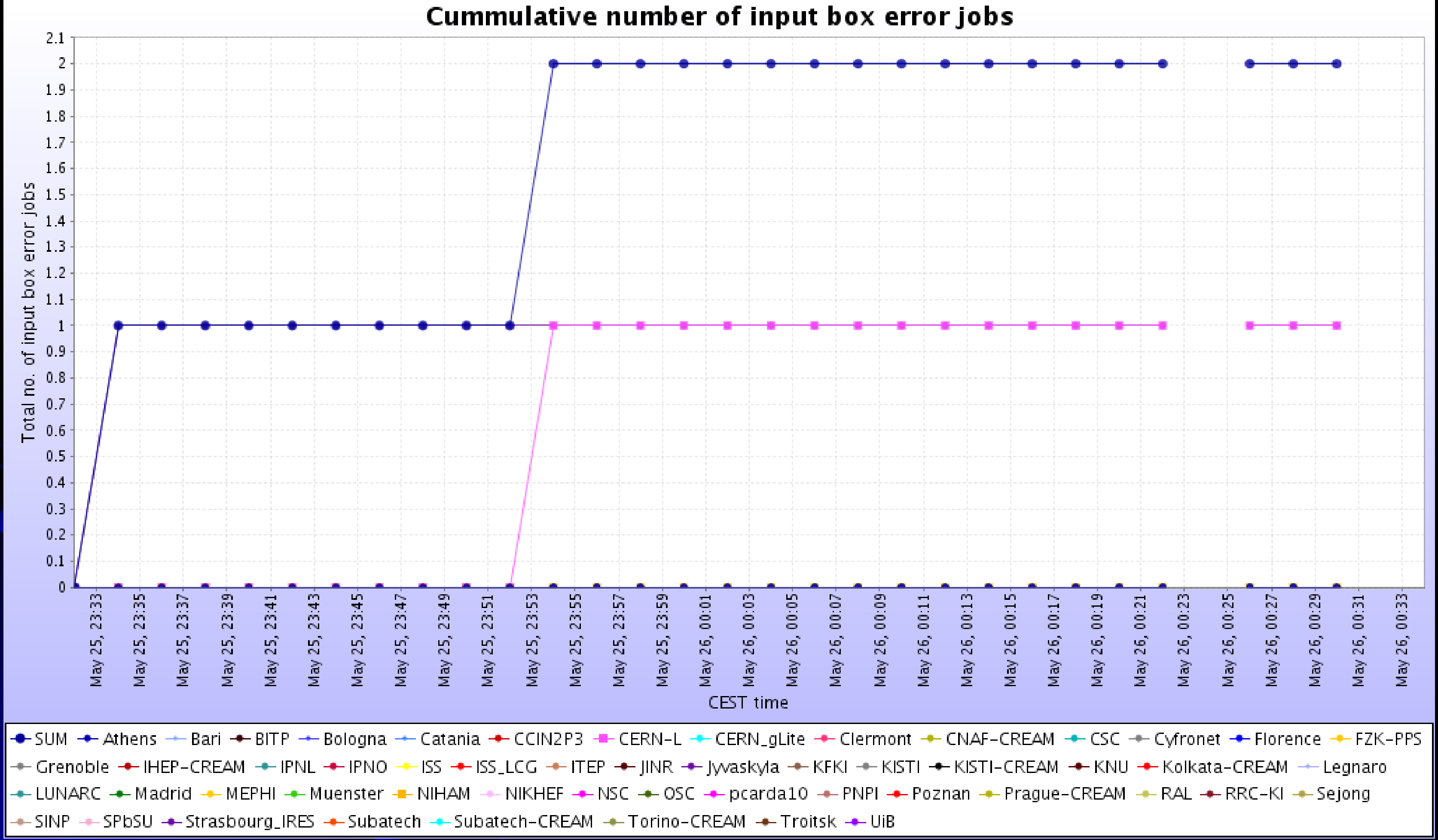
Job flow – all errors



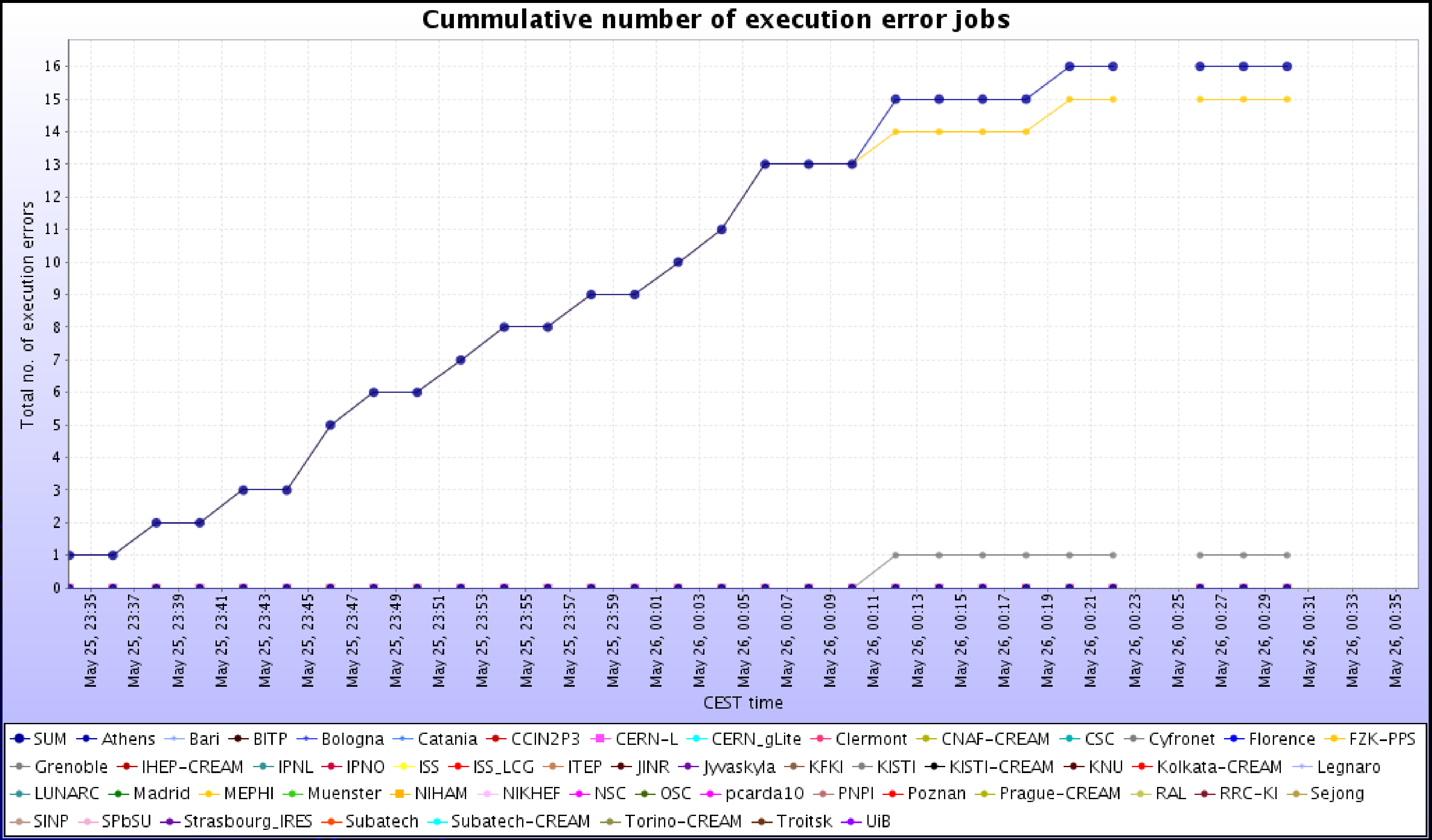
This cannot be normal! (problem 1)

This one should be looked at too (prob.2)

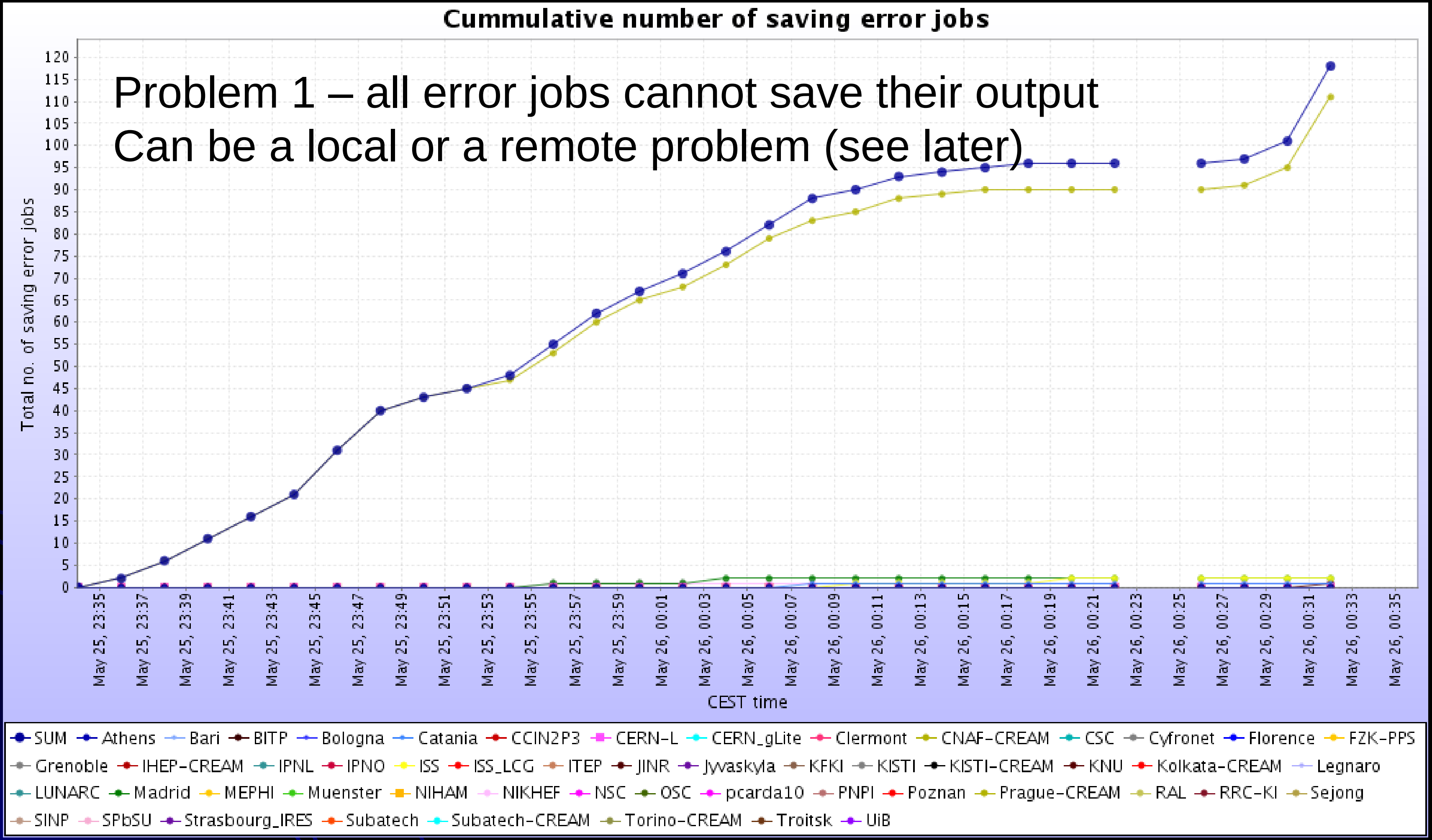
Job flow – what it is not



Job flow – what it is not

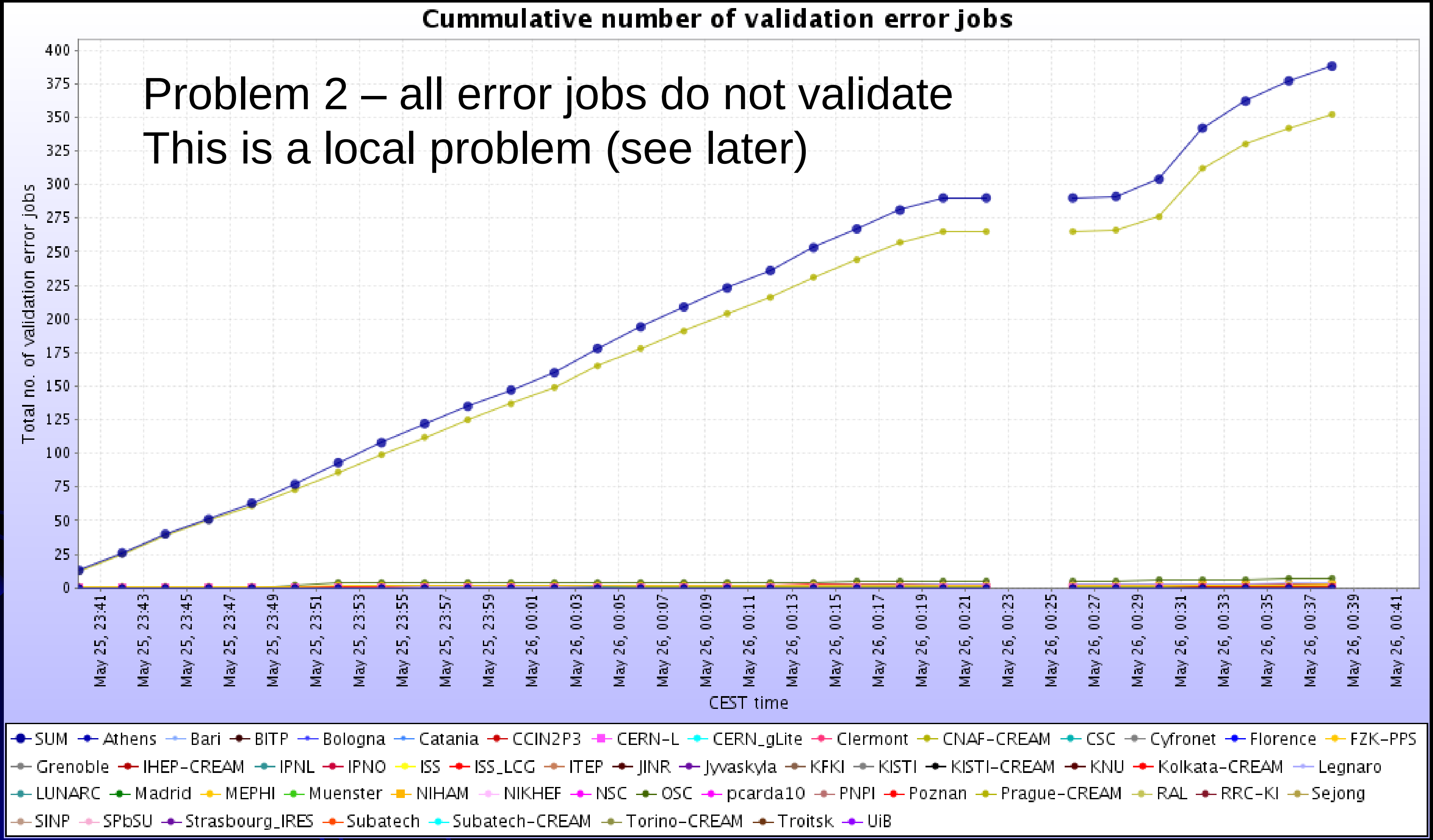


Job flow – what it is



Problem 1 – all error jobs cannot save their output
Can be a local or a remote problem (see later)

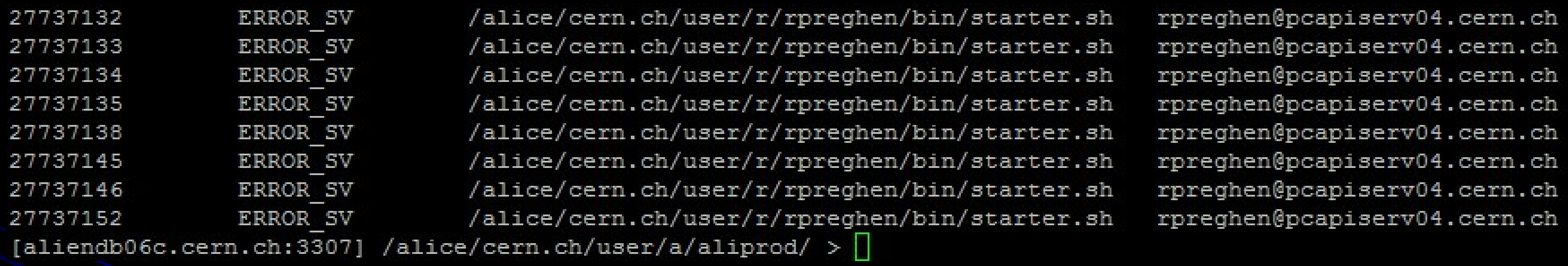
Job flow – what it is



Problem 2 – all error jobs do not validate
This is a local problem (see later)

Diagnosing the problem in detail

- Back to the trusted ‘alien login’
- Let’s first see problem 2
  - top –status ERROR_SV –site ALICE::CNAF::CNAF-CREAM

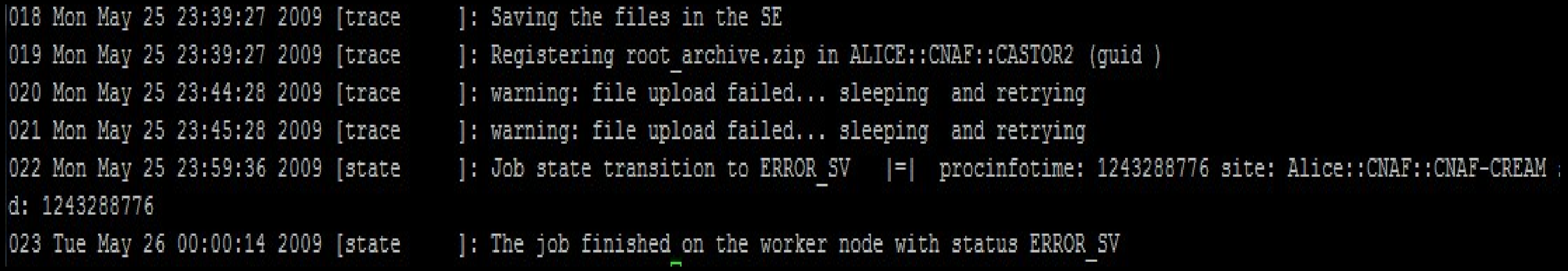


```
27737132        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737133        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737134        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737135        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737138        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737145        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737146        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
27737152        ERROR_SV        /alice/cern.ch/user/r/rpreghen/bin/starter.sh   rpreghen@pcapiserv04.cern.ch
[aliendb06c.cern.ch:3307] /alice/cern.ch/user/a/aliprod/ >
```

- All jobs (this is a partial list) seem to be from a single user. He must be unlucky
  - Let’s trace it in more details

Deeper trace

- ps trace 27737152 all

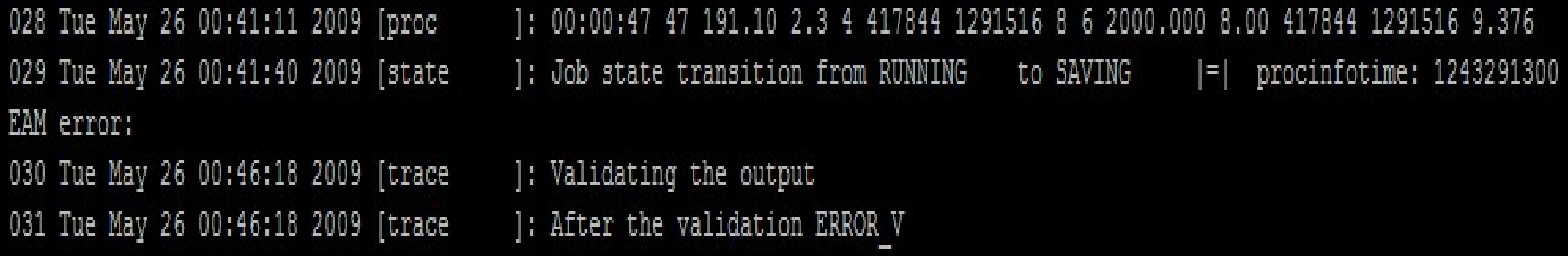


```
018 Mon May 25 23:39:27 2009 [trace     ]: Saving the files in the SE
019 Mon May 25 23:39:27 2009 [trace     ]: Registering root_archive.zip in ALICE::CNAF::CASTOR2 (guid )
020 Mon May 25 23:44:28 2009 [trace     ]: warning: file upload failed... sleeping  and retrying
021 Mon May 25 23:45:28 2009 [trace     ]: warning: file upload failed... sleeping  and retrying
022 Mon May 25 23:59:36 2009 [state     ]: Job state transition to ERROR_SV   |=|  procinfotime: 1243288776 site: Alice::CNAF::CNAF-CREAM :
d: 1243288776
023 Tue May 26 00:00:14 2009 [state     ]: The job finished on the worker node with status ERROR_SV
```

- Seems like SE ALICE::CNAF::CASTOR2 issue
- Let’s check the tests in MonALISA
  - http://pcalimonitor.cern.ch/stats?page=SE/table
  - The storage is down – fix it
  - The storage is OK (this case) – alert the central Grid team (through alice-lcg-taskforce@cern.ch)
  - In this user’s case – improper use of MSS to store small files, SE protected
    - This SE accepts only files larger than 10KB

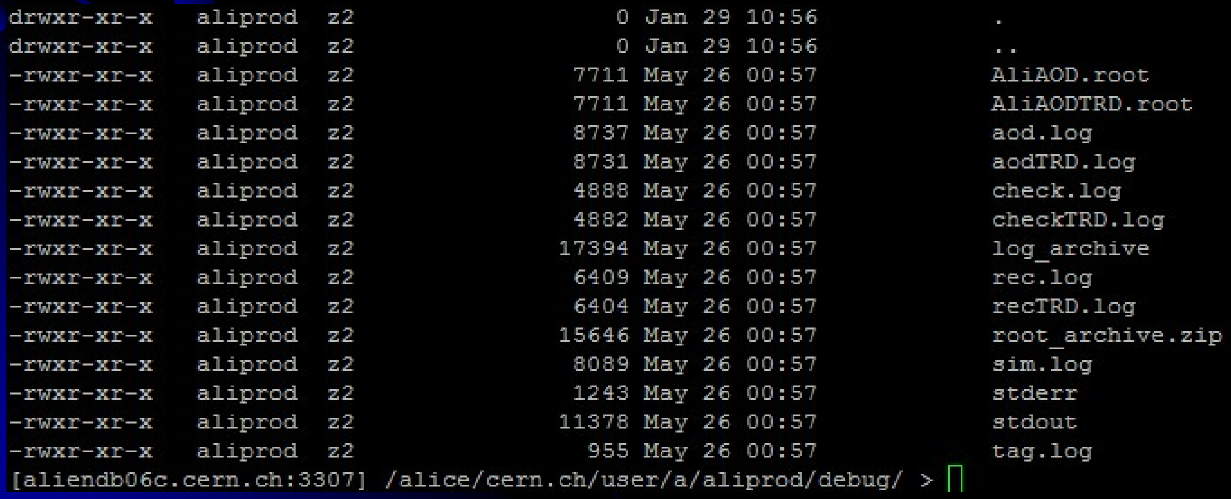
Deeper trace

- ps trace 27736677 all



```
028 Tue May 26 00:41:11 2009 [proc      ]: 00:00:47 47 191.10 2.3 4 417844 1291516 8 6 2000.000 8.00 417844 1291516 9.376
029 Tue May 26 00:41:40 2009 [state     ]: Job state transition from RUNNING    to SAVING     |=|  procinfotime: 1243291300
EAM error:
030 Tue May 26 00:46:18 2009 [trace     ]: Validating the output
031 Tue May 26 00:46:18 2009 [trace     ]: After the validation ERROR_V
```

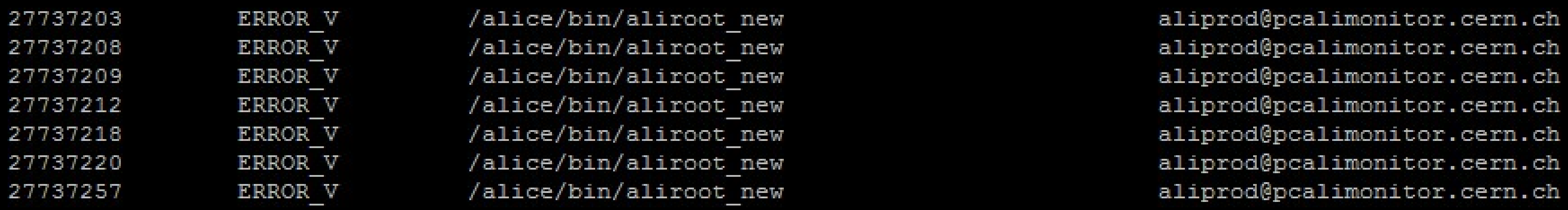
- The job ran only 47 seconds!
- Need output files
  - registerOutput 27736677



```
drwxr-xr-x   aliprod  z2                   0 Jan 29 10:56            .
drwxr-xr-x   aliprod  z2                   0 Jan 29 10:56            ..
-rwxr-xr-x   aliprod  z2                7711 May 26 00:57            AliAOD.root
-rwxr-xr-x   aliprod  z2                7711 May 26 00:57            AliAODTRD.root
-rwxr-xr-x   aliprod  z2                8737 May 26 00:57            aod.log
-rwxr-xr-x   aliprod  z2                8731 May 26 00:57            aodTRD.log
-rwxr-xr-x   aliprod  z2                4888 May 26 00:57            check.log
-rwxr-xr-x   aliprod  z2                4882 May 26 00:57            checkTRD.log
-rwxr-xr-x   aliprod  z2               17394 May 26 00:57            log_archive
-rwxr-xr-x   aliprod  z2                6409 May 26 00:57            rec.log
-rwxr-xr-x   aliprod  z2                6404 May 26 00:57            recTRD.log
-rwxr-xr-x   aliprod  z2               15646 May 26 00:57            root_archive.zip
-rwxr-xr-x   aliprod  z2                8089 May 26 00:57            sim.log
-rwxr-xr-x   aliprod  z2                1243 May 26 00:57            stderr
-rwxr-xr-x   aliprod  z2               11378 May 26 00:57            stdout
-rwxr-xr-x   aliprod  z2                 955 May 26 00:57            tag.log
[aliendb06c.cern.ch:3307] /alice/cern.ch/user/a/aliprod/debug/  >
```

Diagnosing the problem in detail

- Problem 1
  - top –status ERROR_V –site ALICE::Prague::Prague-CREAM

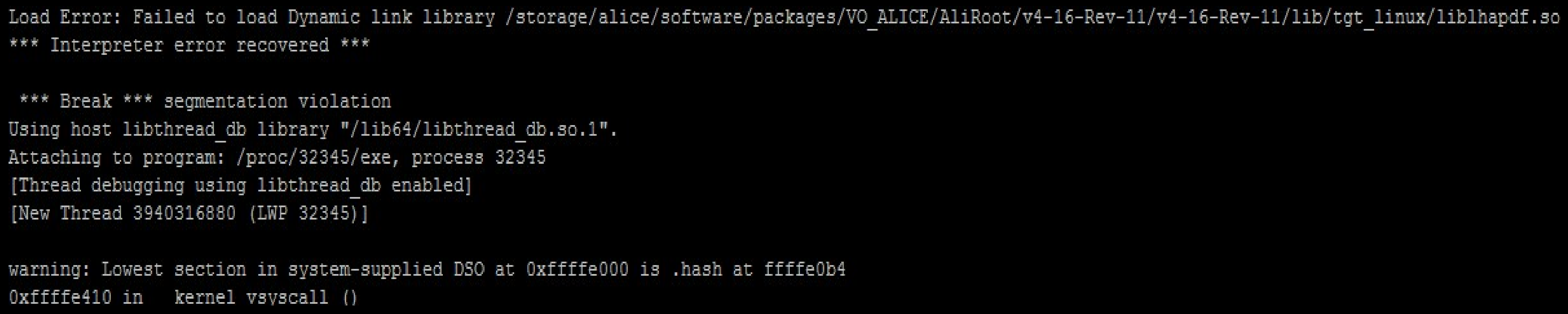


```
27737203        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737208        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737209        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737212        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737218        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737220        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
27737257        ERROR_V        /alice/bin/aliroot_new                        aliprod@pcalimonitor.cern.ch
```

- All jobs (this is a partial list) seem to be from a single user. This is production, must be serious
  - Let’s trace it in more details

Deeper trace (2)

- Let’s see the logs
  - cat sim.log



```
Load Error: Failed to load Dynamic link library /storage/alice/software/packages/VO_ALICE/AliRoot/v4-16-Rev-11/v4-16-Rev-11/lib/tgt_linux/liblhapdf.so
*** Interpreter error recovered ***

 *** Break *** segmentation violation
Using host libthread_db library "/lib64/libthread_db.so.1".
Attaching to program: /proc/32345/exe, process 32345
[Thread debugging using libthread_db enabled]
[New Thread 3940316880 (LWP 32345)]

warning: Lowest section in system-supplied DSO at 0xffffe000 is .hash at ffffe0b4
0xffffe410 in   kernel vsyscall ()
```

- Segfault – this cannot be good
  - alert the central team (Dagmar already did)
  - This will be a difficult one…

Diagnosing the problems - morale

- Two level of diagnosis
  - Minimal set – site services (VO-box and AliEn) through the MonALISA + SAM monitoring
    - This is the first task of the regional expert/site admin
  - Advanced set – job behaviour through alien shell and MonALISA
    - This is more challenging, ultimately increases site efficiency

Advanced set

- Problems are not always evident
  - Job errors have many origins, not surprisingly given the complexity of services interactions
- Diagnostic tools are fairly advances – job tracelogs and comparison studies are sufficient in 99% of the cases
  - More difficult is to ‘read’ the symptoms – the error messages are not always unambiguous
  - Experience comes with practice – some administrators are very skilled!
  - Do not hesitate to report your findings!

Statistics of services failures

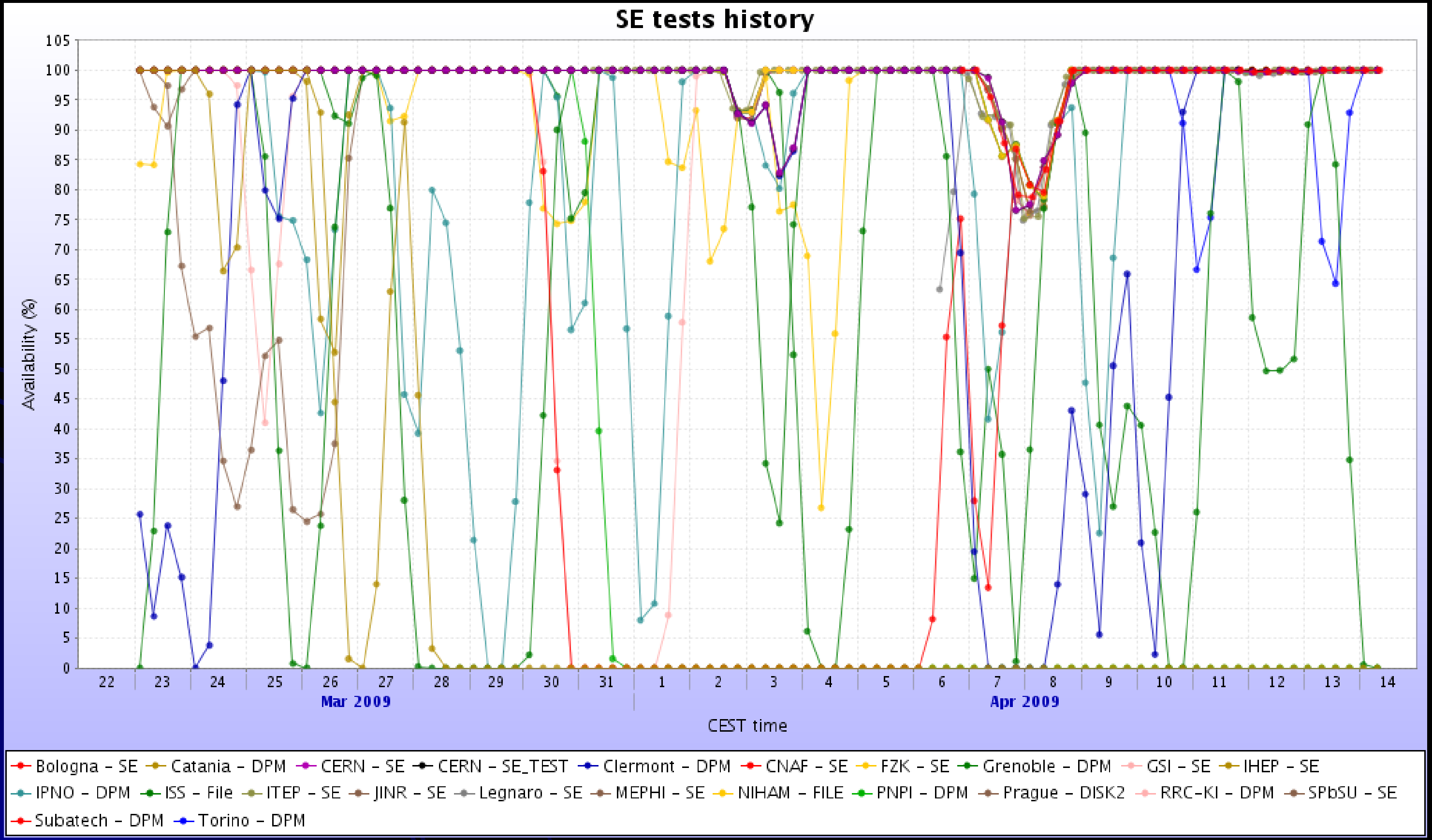
- The problems reported – typical errors, atypical causes
- ERROR_V – 90% of cases problem in user code or failed installation of application on the site shared software area
- ERROR_SV – 99% of the cases a non-working SE
- ERROR_E – 95% of the cases non-working SE

Software installation problems

- The weakest link – NFS
  - Errors of the installation occur while unpacking the package tarball – NFS stale file handle
  - Checking the integrity of the installation by software is far from trivial in this case
  - Blocks the site completely, all jobs end in error
  - Solution exists and will be reported tomorrow
  - The shared software area is an anachronism and should be eliminated!

Storage stability and availability

- T2 storage stability test under load - MonALISA test history



Storage availability scores

- Storage type 1 – average 73.9%
  - Probability of all three alive (3 replicas) = 41%
    - This defines the job waiting time and success rate, ALICE can keep only 3 replicas of ESDs/AODs
- xrootd native – average 92.8%
  - Probability of all three alive (3 replicas) = 87%

Contributing factors

- Hardware – every centre selects the best storage it can afford on price / performance basis
  - One element which is difficult to make better
  - In fact is rarely the case of failure (air conditioning is more problematic)
- Software – the selection is limited
  - Many of the current problems are overcome by inventive ‘local’ solutions – this helps, but is not a cure
  - There is always the hope of a new version, which will fix all present issues

Contributing factors (2)

- Software (contd)
  - The most advanced storage solution is xrootd – this has been demonstrated
- All other parameters being equal (protocol access speed and security): ALICE recommends wherever feasible a pure xrootd installation
  - Ancillary benefit from site admin point of view – no databases to worry about + storage cooperation through global redirector

Monitoring, monitoring…

- Even the best SEs fail
  - The key is to monitor closely the behavior and take corrective actions immediately
    - …Event few % unavailability has a dramatic effect on the job success rate
  - Rather effective testing methods and alert system (MonALISA) is in operation – all system administrators should subscribe to the alerts!

Summary

- The Grid (AliEn/gLite/other) services are many and quite complex
- Nonetheless, they are working together, allowing to manage thousands of CPUs and PBs of various storage types
- The ALICE choice of single Grid Catalogue, single Task Queue with internal prioritization and a single storage access protocol (xrootd) is benefitial from user and Grid management viewpoint

Summary (2)

- The elements and boundaries of the system are well established – for the sites the critical element is the VO-box
- Two additional elements, which need attention and improvement (in order of importance) are

1. Storage
2. Software distribution system

- Other elements are entering the picture (CREAM-CE, WMS), these are already in the AliEn system and in production

Summary (3)

- Regional experts/site admins – follow up on services status
  - Functionality tests and monitoring are performed by (two distinct) frameworks
    - SAM – gLite
    - MonALISA - AliEn
    - Services log files also help
- Site services support is a question of practice – for experienced sysadmin ~1/2 hour/day under normal circumstances