GSI Analysis Facility

Anna Kreshuk
Victor Penso

ALICE-FAIR Computing Meeting
29 April 2008

Principle: ALICE computing is integrated with GSI computing to share resources and the infrastructure

GSI Analysis Facility

Coexists with GSIs batch system on the same nodes

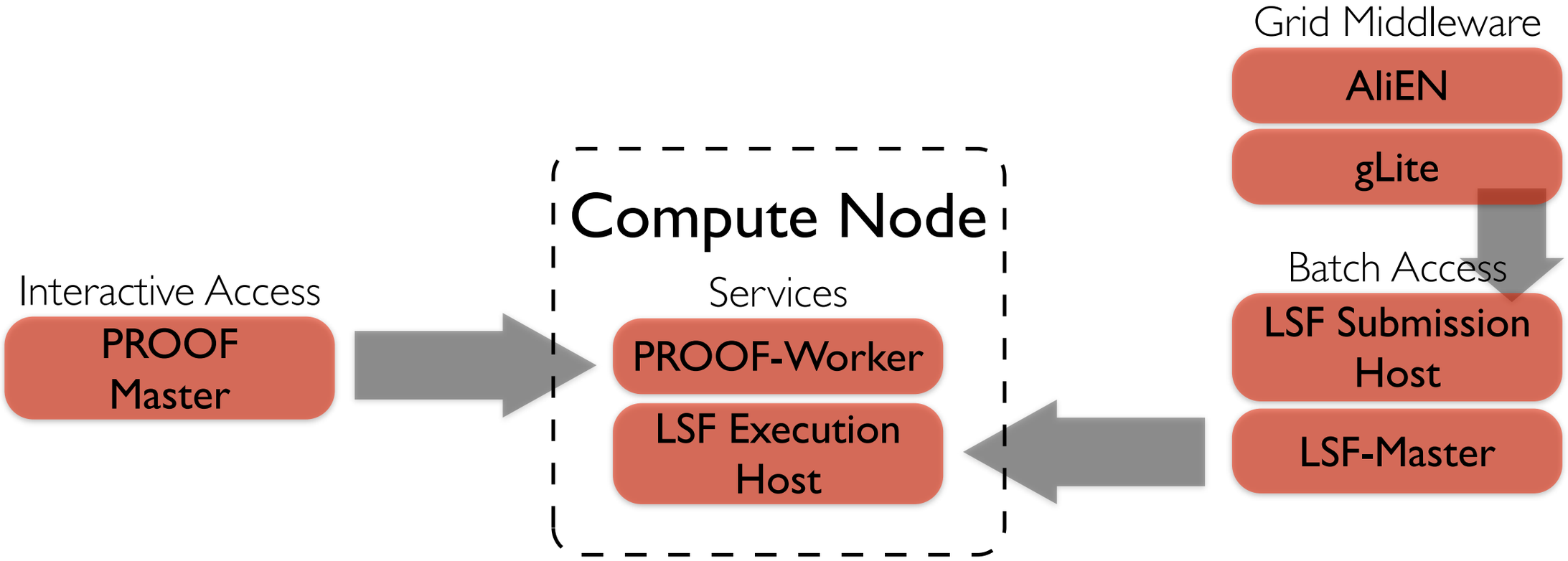
	Status			LSF	PROOF	CPU		Memory		Swap		Network		Storage	
Machine	online	xrootd	olbd	jobs	users	load	idle	total	free	total	free	in	out	total	free
lxb284.gsi.de				7	3	9.85	50.33	15.7 GB	7.833 GB	29.81 GB	18.4 GB	15.37 KB/s	6.673 KB/s	1.121 TB	782.2 GB
lxb285.gsi.de				5	3	6.72	59.53	15.7 GB	8.86 GB	29.81 GB	17.94 GB	15.73 KB/s	7.036 KB/s	1.121 TB	1.076 TB
lxb286.gsi.de				6	3	8.42	53.68	15.7 GB	10.66 GB	29.81 GB	16.21 GB	16.73 KB/s	8.354 KB/s	1.069 TB	1.024 TB
lxb289.gsi.de				5	3	6.56	46.61	15.7 GB	8.944 GB	29.81 GB	18.74 GB	13.46 KB/s	5.815 KB/s	1.121 TB	1.086 TB
lxb291.gsi.de				7	3	8.72	44.53	15.7 GB	10.62 GB	29.81 GB	29.73 GB	12.59 KB/s	4.336 KB/s	1.121 TB	1.1 TB
lxb293.gsi.de				4	3	6.14	78.02	15.7 GB	11.61 GB	29.81 GB	29.54 GB	13.08 KB/s	4.652 KB/s	1.121 TB	1.084 TB
lxb294.gsi.de				6	3	7.82	29.02	15.7 GB	11.44 GB	29.81 GB	29.6 GB	13.75 KB/s	5.52 KB/s	1.069 TB	1.025 TB
lxb295.gsi.de				5	3	7.28	61.74	15.7 GB	10.9 GB	29.81 GB	29.66 GB	13.64 KB/s	5.373 KB/s	1.069 TB	1.025 TB
lxb296.gsi.de				6	3	6.75	23.21	15.7 GB	10.64 GB	29.81 GB	29.65 GB	12.25 KB/s	3.716 KB/s	1.069 TB	1.039 TB
lxb297.gsi.de				6	3	8.05	49.71	15.7 GB	10.39 GB	29.81 GB	29.69 GB	14.12 KB/s	5.734 KB/s	1.069 TB	1.03 TB
lxgrid2.gsi.de				0	3	0	99.96	7.801 GB	7.089 GB	1.953 GB	1.851 GB	8.605 KB/s	0.163 KB/s	931.2 GB	920.3 GB
Total	**11**	**11**	**11**	**57**								**149.3 KB/s**	**57.37 KB/s**	**11.86 TB**	**11.15 TB**
Average						**6.937**	**54.21**					**13.58 KB/s**	**5.216 KB/s**		

Outline

1. Deployment
2. Operation
3. Data Access

GSI Analysis Facility

From the users point of view




```
root [1] TProof::Open("gsiaf.gsi.de")
root [15] chain->Process("analyzer.C")
aliensh> submit analyzer.jdl
glite-job-submit --vo alice analyzer.jdl
globusrun-ws -submit -c analyzer.sh
bsub -q alice-t3 analyzer.sh
```

Deployment

Adapting GSIs environment

Host system Debian needs:
- specific start/stop scripts
- follow conventions for: system accounts, log files…
- software distribution via: shared filesystem, (packages)

Coexistence with LSF needs to be understood better:
- Potential interference between the systems
- Intelligent scheduling of batch jobs; higher priority for PROOF

Deployment

Adapting GSIs configuration management

All Linux boxes are maintained using
http://www.cfengine.org/

We have developed configuration recipes for XROOTD and MonALISA which allows resizing of GSIAF.

```
xrootd          = ( HostRange(lxb,283-304) )
monalisa        = ( HostRange(lxb,283-304) )

xrootd::cf.xrootd
monalisa::cf.monalisa
```

Operation

On demand monitoring and management

Problem: Clusters with more then 10 nodes aren't manually manageable. GSIs cfengine has an delay of 60 minutes.

We need a tool…

- to react fast on technical incidents
- to start fine grained monitoring for a short period
- for interactive workflow to simplify problem detection

Operation

On demand monitoring and management

Implementation with http://capify.org/
Parallel command execution via SSH

Pre-configured tasks:

```
cap cluster:check:tmp_space      # disk space in /tmp
cap cluster:lustre:mount         # mount to /lustre_alpha
cap cluster:monitor:restart      # restart MonALISA sensors
cap gsiaf:check:data_space       # disk space in /data.local2
cap gsiaf:check:memory           # detailed statistics
cap gsiaf:check:services         # xrootd/olbd daemons running?
cap gsiaf:check:uptime           # uptime and idle time
cap gsiaf:proof:count_workers    # proof workers on the nodes
cap gsiaf:xrootd:start           # start xrootd/olbd
cap gsiaf:xrootd:stop            # stop xrootd/olbd
```

Operation

On demand monitoring and management

Using a parallel shell:

```
cap shell
cap> on lxb284,lxb285,lxb286
cap> ps -e -o rss=,args= | sort -b -k1,1n | tail -1
 ** [out :: lxb284] 829084 aliroot -b -q sim.C
 ** [out :: lxb285] 691184 aliroot -b -q rec.C
 ** [out :: lxb286] 840064 aliroot -b -q sim.C
```

On the command line or from scripts:

```
export HOSTS="lxb001","lxb002"
cap -vXx invoke COMMAND="df -h /data.local2 | tail -1"
cap -vXx invoke SUDO=1 COMMAND="netstat -punta"
```

Operation

Keeping the Daemons running

Problem: Master killed by PROOF-Users

```
[2008-04-22 09:46:28 #14757]  INFO -- : xrootd process is not running
[2008-04-22 09:48:38 #14757]  INFO -- : xrootd process is not running
[2008-04-22 09:58:48 #14757]  INFO -- : xrootd process is not running
[2008-04-22 10:13:58 #14757]  INFO -- : xrootd process is not running
[2008-04-22 10:18:08 #14757]  INFO -- : xrootd process is not running
```

Solution: Watchdog for automatic restart

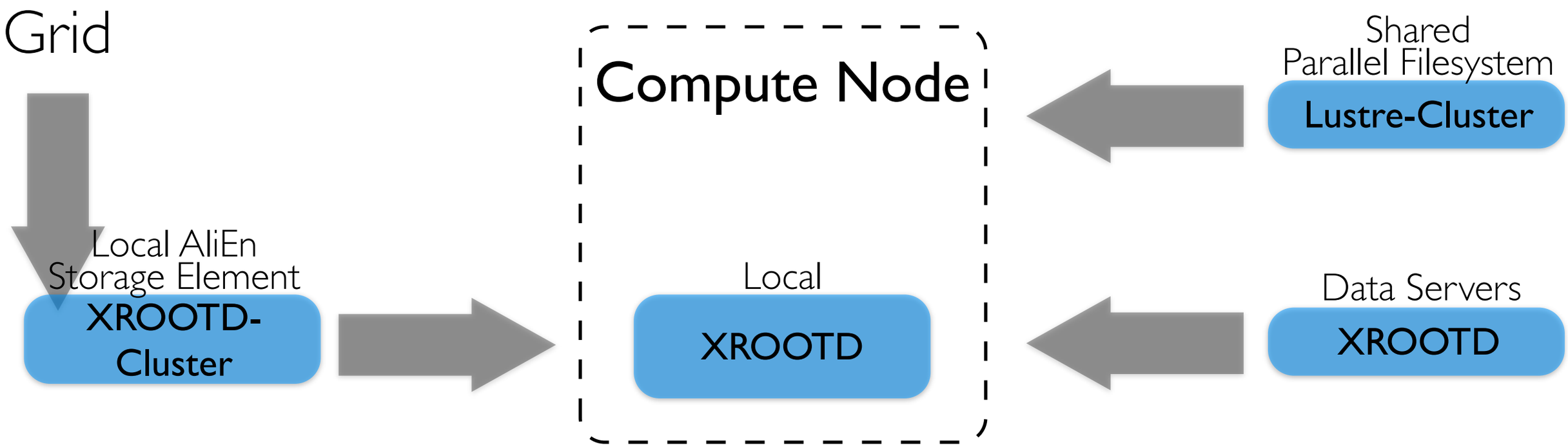
```
> god -c xrootd.god -l /var/log/god.log
> god status
xrootd: up
```

Advantage: Downtime less then 60 seconds

Logbook for statistics

GSI Analysis Facility

From the data point of view



Jobs, running on the cluster, can read data from…

- local disk on the machine
- file servers with local production data or Lustre
- the AliEn storage element

Data Access

Today

All PROOF user are allowed to copy the needed data from anywhere to anywhere with plain xrdcp and AliEn tools.

Problems:

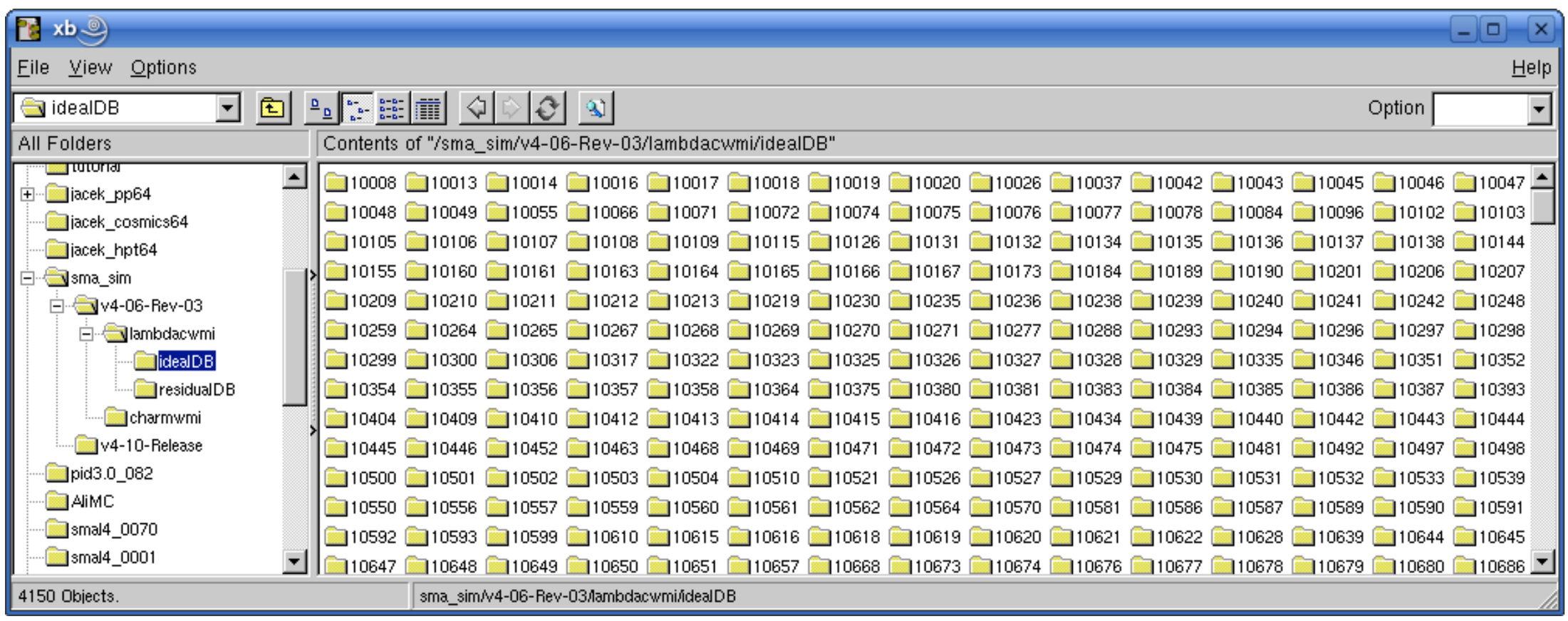
- Users can't get an overview of existing data
- In case if local disk failing it's unclear what data is missing
- Data transfer can slow down jobs reading remote data

Data Access

Listing files in the cluster

We have a solution for listing files on the cluster, using XROOTD client-admin-tools and regular ROOT TBrowser.

```
root [0] .L TDirXroot.h+
root [1] .L sysdir.C+
root [2] Browse("/", "hosts.txt")
```



Data Access

Retrieve Missing Data

Fabrizio's solution to automatically copy lost files from remote sources:

- Concept works on our test system, will be implemented as a general solution as soon as the new PROOF with new XROOTD becomes stable on CAF
- Should it be possible to switch off this mechanism by the user?
- Pulling data from more then one source?

Data Access

Data Management for Users

PROOF dataset management tools:

- Currently investigating the changes needed in our case
- In case of big amounts of data to be transfered we need to make sure not to disconnect other applications from their data by blocking bandwidth of the single line to local file servers or AliEn SE

Conclusion

Cluster administration is clear for us

Number of our users is expected to grow in the coming months

Coexistence of LSF and PROOF requires more investigations:
- Scheduling/PROOF priority
- Network Bandwidth sharing