Consistency of Accounting Information with DGAS

Rosario Piro, Andrea Guarise,

Riccardo Brunetti, Luciano Gaido,

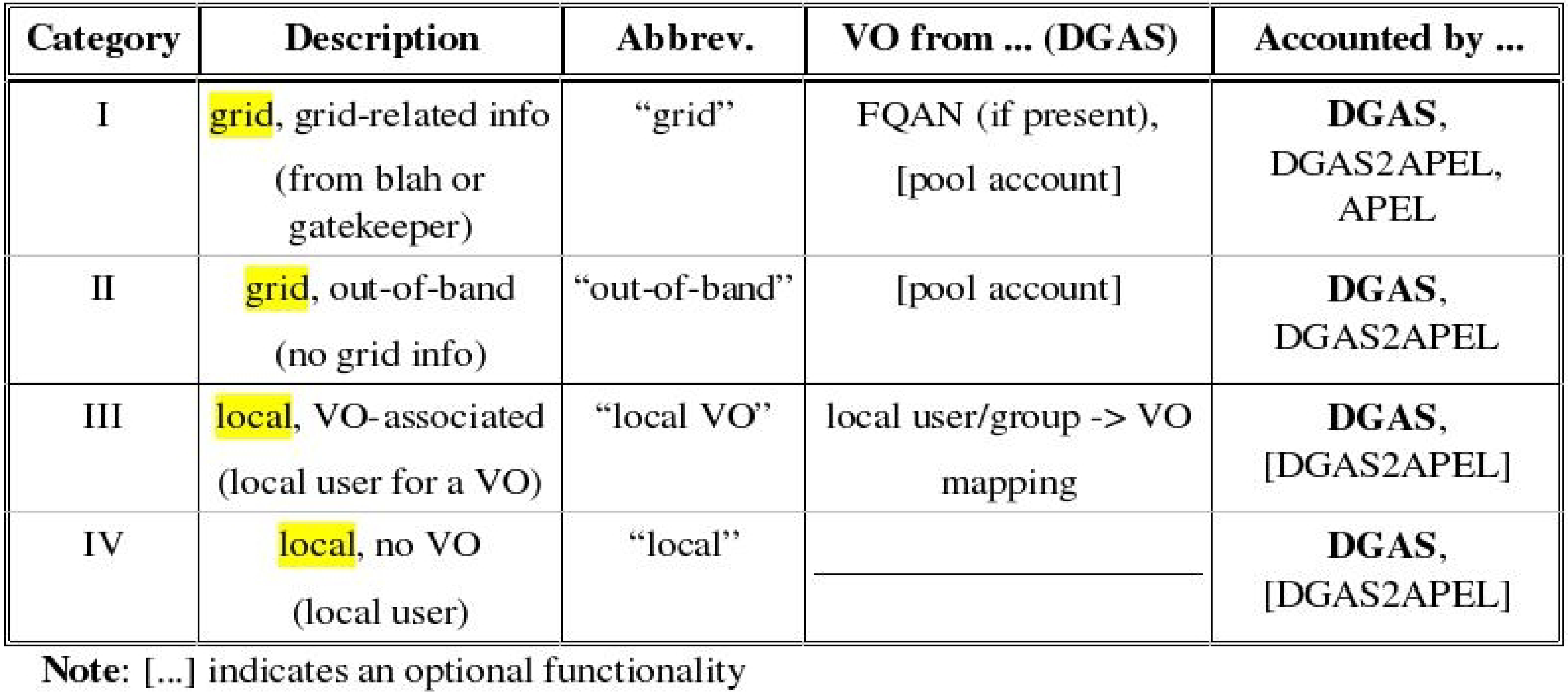
Giuseppe Patania, Paolo Veronesi

JRA1 All Hands Meeting, Espoo, June 18-20, 2007.

www.eu-egee.org

Job Classification

- From the accounting point of view, executed jobs can be classified in the following categories:



Category	Description	Abbrev.	VO from ... (DGAS)	Accounted by ...
I	grid, grid-related info (from blah or gatekeeper)	"grid"	FQAN (if present), [pool account]	**DGAS**, DGAS2APEL, APEL
II	grid, out-of-band (no grid info)	"out-of-band"	[pool account]	**DGAS**, DGAS2APEL
III	local, VO-associated (local user for a VO)	"local VO"	local user/group -> VO mapping	**DGAS**, [DGAS2APEL]
IV	local, no VO (local user)	"local"		**DGAS**, [DGAS2APEL]

Note: [...] indicates an optional functionality

- We need to account at least categories I – III !

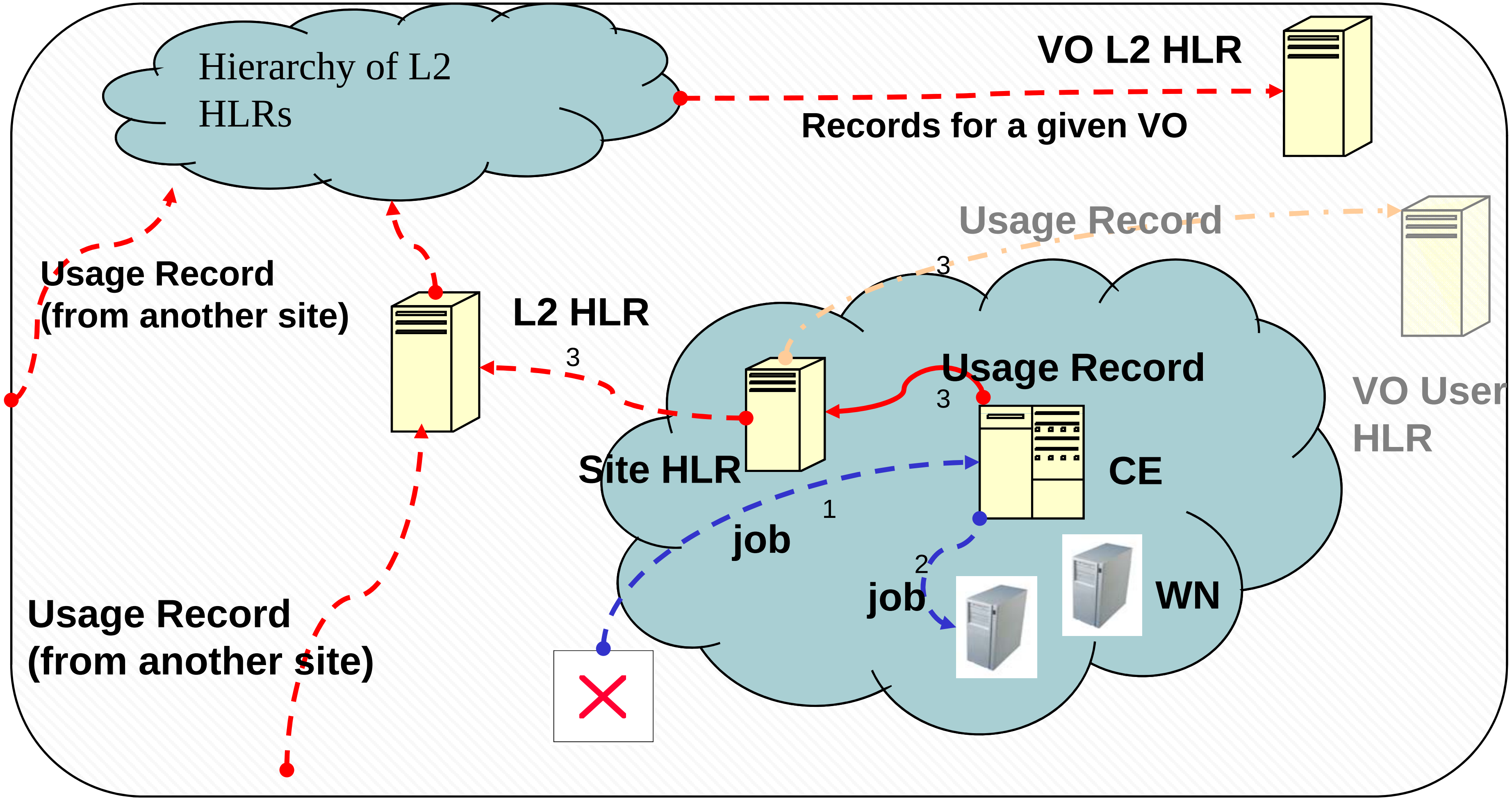
Problems

- **Accounting "grid" jobs (with grid-related info; cat. I) is mostly straight forward** (Some 'features' of the job submission chain and of the underlying services, makes it difficult to perform proper accounting also in the trivial cases).
- **Accounting "out-of-band", "local VO" and "local" jobs (cat. II-IV) is a non trivial task**
  - **risk of record duplication** for certain site configurations
    - e.g. one LRMS head node for multiple CEs (sensors on the CEs read *the same* LRMS log file to get usage information)
    - The DGAS HLR server checks incoming records for possible duplications!
      - *There are many possible circumstances possibly resulting in record duplication, each of them must be taken into consideration before accepting the insertion request for a Usage Record.*
  - **use of pool accounts to determine the VO is risky**
    - e.g. wrong mapping of credentials to pool accounts can occur
      - *real case: "/biomed/..." -> "cms003" >.This is not a problem if FQAN is available. Unfortunately many jobs are still submitted with the use of plain, no VOMS, credentials. This should be highly deprecated.*
    - DGAS now allows to consider pool accounts *optionally*.
  - **use of a mapping from local user and group accounts to VOs** requires an appropriate and up-to-date configuration
    - DGAS allows site administrators to map their local users and/or groups to specific VOs. This can be done separately per each CE of the site.

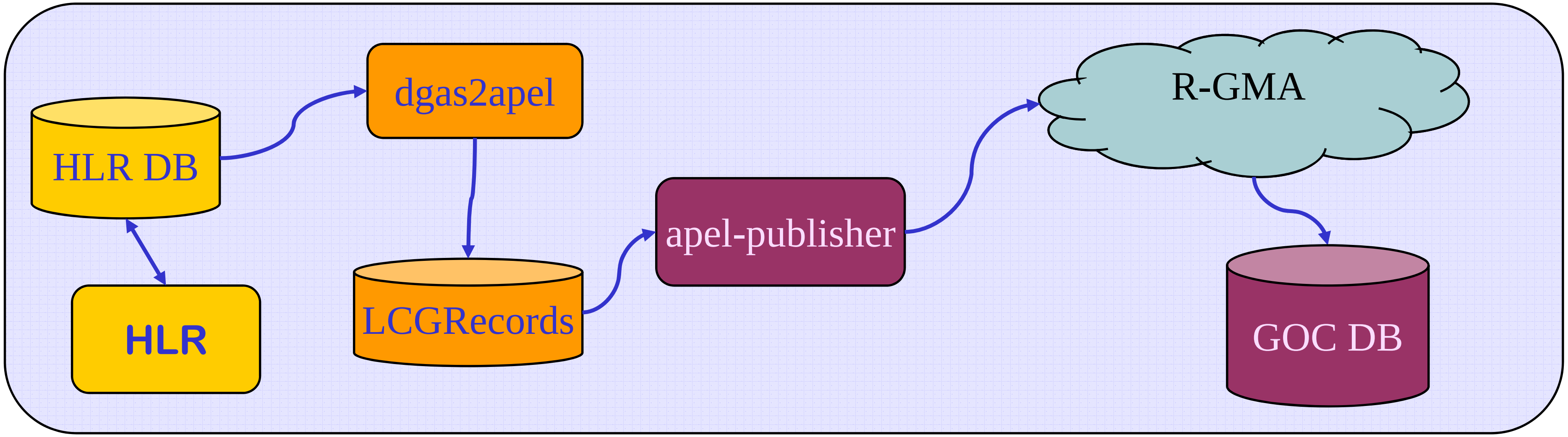
Important

- **A thorough and pedantic verification of accounting information is ESSENTIAL!**
  - **cross-check of accounting records with LRMS log files**! How much information do we lose?
  - **cross-check of local accounting records (on sites) with what ends up in the GOC DB!**
  - make sure **only true accounting information** can end up in the GOC DB (can normal users publish fake accounting records in RGMA?)

DGAS (simplified) Workflow



DGAS2APEL workflow

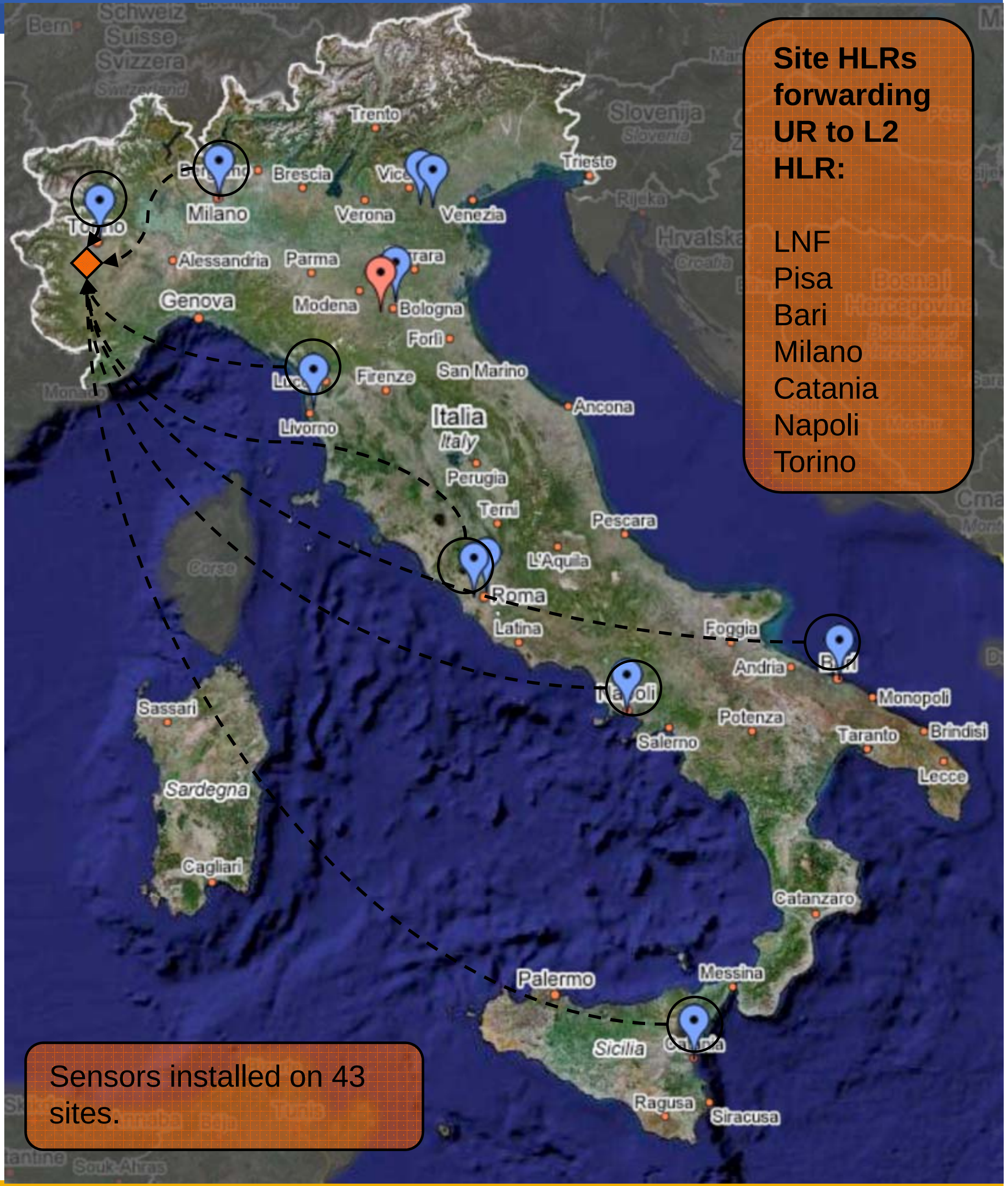


- ***DGAS2APEL*** **is a process that** ***converts*** **the Usage Records from the format adopted by DGAS to the one adopted by APEL. Converted records are then inserted in an RDBMS table known as LcgRecords.**
- **Such records are then** ***forwarded*** **to the GOC** ***by APEL*** **itself via its 'apel-publisher' process, which uses** ***RGMA*** **as a high level transport service toward the GOC.**

Enabling Grids for E-sciencE

Deployment status

	HLR	Hlr-t1,cr.cnaf.infn.it	T1: INFN-T1
	HLR	Prod-hlr-01.pd.infn.it	T2: INFN-PADOVA Reference for central-northern areasites: CNR-ILC-PISA INAF-TRIESTE INFN-CNAF INFN-BOLOGNA INFN-BOLOGNA-CMS INFN-FERRARA INFN-FIRENZE INFN-GENOVA INFN-PARMA INFN-PERUGIA INFN-TRIESTE SNS-PISA UNIV-PERUGIA
	HLR	Prod-hlr-01.ct.infn.it	Reference for central-southern area sites: ENEA-INFO ESA-ESRIN INFN-CAGLIARI INFN-LECCE INFN-LNS INFN-NAPOLI-CMS INFN-NAPOLI-VIRGO INFN-ROMA2 INFN-ROMA3 ITB_BARI SPACI-COSENZA SPACI-LECCE-IA64 SPACI-NAPOLI SPACi-NAPOILI-IA64
	HLR	prod-hlr-02.ct.infn.it	T2:INFN-CATANIA
	HLR	**Prod-hlr-01.ba.infn.it**	**T2:INFN-BARI**
	HLR	**Atlashlr.lnf.infn.it**	**T2:INFN-FRASCATI**
	HLR	**T2-hlr-01.lnl.infn.it**	**T2:INFN-LEGNARO**
	HLR	**Prod-hlr-01.mi.infn.it**	**T2:INFN-MILANO**
	HLR	**T2-hlr-01.na.infn.it**	**T2:INFN-NAPOLI ,INFN-NAPOLI-ATLAS**
	HLR	**Gridhlr.pi.infn.it**	**T2:INFN-Pisa,INFN-PISA2**
	HLR	**T2-hlr-01.roma1.infn.it**	**T2:INFN-ROMA1,INFN-ROMA1-CMS,INFN ROMA1-VIRGO**
	HLR	**T2-hlr-01.to.infn.it**	**T2:INFN-TORINO**
◆	**L2HLR**	**Hlr2-test-26.to.infn.it**	**Second level HLR.**



Site HLRs forwarding UR to L2 HLR:

LNF
Pisa
Bari
Milano
Catania
Napoli
Torino

Sensors installed on 43 sites.

Consistency Checks for DGAS

- **For INFN-Grid we have monitored the consistency of accounting information in DGAS**
  - Helped to realize and solve problems we didn't even imagine ...
  - Helped to end up with a more complete picture of resource usage by VOs
  - In our opinion the following set of checks are needed:
    - Comparison between data in ***LRMS logs and the Site HLR server***
      - To check if DGAS is correctly collecting usage records.
    - Comparison between data on ***Site HLR and converted by DGAS2APEL***
      - To check if DGAS2APEL is correctly translating into the LcgRecord format all the records (and only them) that we plan to forward to GOC.
    - Comparison between data on ***Site HLR and published via DGAS2APEL (conversion) + APEL Publisher (forwarding to GOC)***
      - To check that the information are correctly forwarded to GOC by APEL Publisher and RGMA
    - Comparison between data on ***LRMS logs and APEL Parser + Publisher*** (without forwarding to GOC)
      - *Not strictly related to DGAS operation: to check if APEL sensors are correctly collecting usage records.*

LRMS logs vs. Site HLR (1)

In order to cross-check the information available in the LRMS plain log files with the filtered Usage Records on the Site HLR the following *methodology* was adopted:

- A script *parses the LRMS logs* and insert the information needed for the checks *in a relational database*, trying to reflect the way some of this information are filtered by the DGAS algorithms (for example the start date of the job is not straightforward to determine, and this should be taken into consideration performing the checks).
- A set of *aggregates representing the same quantities, are derived from both datasets (HLR and LRMS) and compared*.
- If the cross-check script and *queries are properly tuned the results should match*, a part form minor differences due to little (but unavoidable) differences in the aggregation process of the single records (roundings, boundary conditions, slight differences in time partitioning of the datasets…).
- *When significant differences are found an in-depth analysis is performed to highlight its causes*. When a bug is found in DGAS it is fixed, otherwise if the problem is in the site configuration, the latter is changed and checks performed again when new information are available.

LRMS logs vs. Site HLR (2)

Sites HLR/LRMS logs

Green: x <= 0.25 %

Yellow: 0.25<x<=1%

Red: x> 1%

	Torino	Job			Cpu time			Wall time		
	Month	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs
Torino	Sep	16017	16017	0,00%	23306,88	23306,88	0,00%	40604,04	40604,04	0,00%
	Oct	37122	37122	0,00%	29347,93	29347,93	0,00%	42572,05	42572,05	0,00%
	Nov	25990	25990	0,00%	19731,09	19731,09	0,00%	38066,51	38066,51	0,00%
	Dec	7986	8042	0,70%	32083,11	32475,44	1,22%	38292,4	38772,22	1,25%
	Jan	7450	7449	-0,01%	27008,01	27008,01	0,00%	39707,03	39707,03	0,00%
	Total:	94565	94620	0,06%	131477,02	131869,35	0,30%	199242,03	199721,85	0,24%
	Pisa	Job			Cpu time			Wall time		
	Month	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs
Pisa	Sep	8176	8179	0,04%	31795,93	31830,9	0,11%	37748,99	37783,8	0,09%
	Oct	13925	13928	0,02%	20424,45	20425,04	0,00%	37070,06	37075,59	0,01%
	Nov	10166	10166	0,00%	21609,51	21609,51	0,00%	31640,29	31640,29	0,00%
	Dec	5604	5608	0,07%	27778,97	27813,13	0,12%	33425,88	33460,42	0,10%
	Jan	6921	6919	-0,03%	25828,53	25819,89	-0,03%	33312,53	33303,65	-0,03%
	Total:	44792	44800	0,02%	127437,39	127498,47	0,05%	173197,75	173263,75	0,04%
	Milano	Job			Cpu time			Wall time		
	Month	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs
Milano	Sep	3279	3284	0,15%	25737,42	25737,62	0,00%	28352,93	28353,39	0,00%
	Oct	5342	5384	0,79%	23274,14	23328,49	0,23%	38447,13	38556,57	0,28%
	Nov	3164	3171	0,22%	14906,89	14985,44	0,53%	30183,28	30496,75	1,04%
	Dec	8631	8677	0,53%	21597,16	22290,97	3,21%	29769,61	30857,86	3,66%
	Jan	14256	14258	0,01%	14347	14410	0,44%	26589,02	26656	0,25%
	Total:	34672	34774	0,29%	99862,61	100752,52	0,89%	153341,97	154920,57	1,03%
	Bari	Job			Cpu time			Wall time		
	Month	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs
Bari	Sep	22530	22200	-1,46%	55368,36	57581,24	4,00%	68366,87	71152,01	4,07%
	Oct	27354	27305	-0,18%	45333,97	47773,6	5,38%	55028,9	57989,41	5,38%
	Nov	18750	18740	-0,05%	53466,76	55635,85	4,06%	59922,37	62358,66	4,07%
	Dec	12135	12159	0,20%	38693,37	42539,11	9,94%	40727,92	44751,58	9,88%
	Jan	12835	12842	0,05%	62448	62739	0,47%	67257	66605	-0,97%
	Total:	93604	93246	-0,38%	255310,46	266268,8	4,29%	291303,06	302856,66	3,97%
	Napoii	Job			Cpu time			Wall time		
	Month	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs	pbs	hlr	diff hlr-pbs
Napoli	Sep	28949	28956	0,02%	17644	17647	0,02%	28499	28504	0,02%
	Oct	19855	19857	0,01%	20345	20345	0,00%	31858	31859	0,00%
	Nov	5188	5202	0,27%	22242	22505	1,18%	36170	36439	0,74%
	Dec	6399	6401	0,03%	20338	20343	0,02%	26999	27005	0,02%
	Jan	11933	11930	-0,03%	14463	14461	-0,01%	27283	27279	-0,01%
	Total:	72324	72346	0,03%	95032	95301	0,28%	150809	151086	0,18%
	Frascati	Job			Cpu time			Wall time		
	Month	Lsf	hlr	diff hlr-pbs	Lsf	hlr	diff hlr-pbs	Lsf	hlr	diff hlr-lsf
Frascati	Sep	6498	6502	0,06%	15915,123	15915,13	0,00%	21421,85975	21421,88	0,00%
	Oct	7321	7321	0,00%	14947,921	14947,931	0,00%	22429,5	22429,5	0,00%
	Nov	4089	4089	0,00%	17035,363	17035,363	0,00%	24724,72443	24724,724	0,00%
	Dec	5109	5112	0,06%	14107,148	14130,226	0,16%	21189,17873	21394,055	0,97%
	Jan	12188	12189	0,01%	8588,6073	8588,6173	0,00%	9448,741807	9449,1918	0,00%
	Total:	35205	35213	0,02%	70594,16271	70617,26777	0,03%	99214,00472	99419,35134	0,21%
		Job			Cpu time			Wall time		
Avg				0,04%			1,05%			1,04%

LRMS logs vs. Site HLR (3)

Green: x <= 0.25 %

Yellow: 0.25<x<=1%

Red: x> 1%

Site HLR/LRMS logs T1.

SOURCE	VO	NJOBS	NJOBS DIFF%	CPUTIME(h)	CPUTIME(h) DIFF%	WALLTIME(h)	WALLTIME(h) DIFF%
HLR	alice	10000	0,00%	22709,94	0,01%	30567,09	0,01%
LSF	alice	10000		22707,14		30563,95	
HLR	atlas	12932	0,00%	105340,45	0,00%	140501,21	0,00%
LSF	atlas	12932		105340,45		140501,21	
HLR	babar	1414	0,00%	19168,79	0,00%	22627,63	0,00%
LSF	babar	1414		19168,79		22627,63	
HLR	cdf	5179	0,00%	46374,48	0,00%	110704,15	0,00%
LSF	cdf	5179		46374,48		110704,15	
HLR	cms	32030	0,00%	1783,42	0,00%	12912,30	0,00%
LSF	cms	32030		1783,42		12912,27	
HLR	dteam	915	0,00%	5,15	0,01%	147,79	0,00%
LSF	dteam	915		5,15		147,79	
HLR	infngrid	4013	0,00%	3,36	-0,02%	488,21	0,00%
LSF	infngrid	4013		3,36		488,21	
HLR	lhcb	4461	0,00%	6118,25	0,00%	10602,03	0,00%
LSF	lhcb	4461		6118,25		10602,03	
HLR	ops	716	0,00%	19,47	0,01%	140,28	0,00%
LSF	ops	716		19,47		140,28	
HLR	theophys	467	0,00%	9127,12	0,00%	9848,65	0,00%
LSF	theophys	467		9127,12		9848,65	
Average:			0,0000%		0,0028%		0,0012%

This cross-check has been ***performed after the latest DGAS upgrade*** at the T1 site and covers the period from 2007-05-23 to 2007-06-03 (boundaries included).

In this view the cross-checks have been done for each of the major VOs.

There’s no need for comments.

Site HLR vs. DGAS2APEL (1)

Green: x <= 0.25 %

Yellow: 0.25<x<=1%

Red: x> 1%

HLR/DGAS2APEL consistency check in 'Torino'

SOURCE	VO	NJOBS	NJOBS DIFF%	CPUTIME(h)	CPUTIME(h) DIFF%	WALLTIME(h)	WALLTIME(h) DIFF%
HLR	alice	41377	0,03%	41867,89	0,11%	98433,83	0,12%
dgas2apel	alice	41366		41821,55		98311,49	
HLR	atlas	1519	0,20%	51,06	0,02%	284,27	0,02%
dgas2apel	atlas	1516		51,05		284,21	
HLR	biomed	1528	0,07%	2487,26	0,00%	6331,51	0,00%
dgas2apel	biomed	1527		2487,25		6331,50	
HLR	cms	3	0,00%	0,08	0,00%	1,16	0,00%
dgas2apel	cms	3		0,08		1,16	
HLR	dteam	1113	0,00%	3,42	0,00%	273,28	0,00%
dgas2apel	dteam	1113		3,42		273,28	
HLR	infngrid	31	0,00%	0,04	0,00%	11,76	0,00%
dgas2apel	infngrid	31		0,04		11,76	
HLR	lhcb	686	1,31%	2641,27	0,00%	2926,29	0,01%
dgas2apel	lhcb	677		2641,23		2926,03	
HLR	ops	1388	0,22%	3,94	0,25%	243,83	0,03%
dgas2apel	ops	1385		3,93		243,75	
HLR	zeus	9034	0,45%	0,00	0,00%	2,65	0,38%
dgas2apel	zeus	8993		0,00		2,64	
Average:			0,1262%		0,0214%		0,0627%

This is cross-check in the period 01/05/2007 - 14/06/2007 of the information available in the HLR database and the LcgRecords table generated by DGAS2APEL local to the 'Torino' site.

What we have learnt

- The ***cross-checks for the sites*** have been performed on the period **September'06 - January'07**. As it can be seen, although results where almost good (the ***average discrepancies where around 1%***, and mainly concentrated just on some sites), we started from these results to analyse the records and ***found the source of those errors***. A certain number of bugs where found and ***fixed*** in two subsequent releases of DGAS.
- Not all the sites where affected by the bugs, since these usually involved just ***sites with complex configurations*** (or as in the case of 'Bari', mainly running long-lasting jobs).
- The ***latest available release*** of DGAS is that deployed at CNAF-T1 (using LSF), and being deployed all over INFNGrid, whose consistency checks are illustrated in the previous slide.
- Note that the ***checks do require a huge amount of work*** and are very time consuming. During the period of the checks form September'06 till January'07, one of the DGAS developers was full time dedicated to these checks. And all the involved sites also spent a non negligible amount of time on it.
- For this reason ***further checks are no more performed systematically*** but just on some sites after new release deployment (as for example the T1 checks illustrated in this talk), or when it is needed (as in case of major changes in the site configuration).

Consistency Checks for APEL?

- **In order to perform some consistency check also for APEL we tried to set up the apel-pbs-log-parser and the apel-publisher on one production CE, in order to compare APEL accounting data with the LRMS and DGAS.**
  - We configured the ***apel-pbs-log-parser*** and run it manually.
  - We configured the ***apel-publisher*** in order to ***avoid sending data to GOC***.
    - `<Republish>nothing</Republish>`
  - However we ***didn't manage to fill the LcgRecords table***, since we continuously hit some problem, such as the following:
    - Unable to locate an available Registry Service
    - Read timed out to: https://grid009.to.infn.it:8443/RGMA/PrimaryProducerServlet/declareTable?connectionId=783683734&tableName=LcgRecords&predicate=&hrpSec=600&lrpSec=3600
    - No records joined (apparently failed to merge with the gatekeeper log files ??)
- **In nearly one month of tests this made it impossible to compare the two systems.**
- **However, it is even worse, that the *same errors are found many times when trying to publish data from DGAS2APEL LcgRecords local table to GOC* (GGUS ticket 21637). Trying to track and fix these failures is frustrating and time consuming.**
- **The source for these errors seems to be *RGMA*, its standard configuration on the sites, or the way apel-publisher uses it. As far as we know it is foreseen the possibility for APEL to send LcgRecords to GOC using different transport mechanisms other than RGMA. Is it eventually possible to agree on *another transport mechanism* and switch to this? (directly use MySQL? L2HLR at GOC?)**

gLite Restructuring

Concerning the status of the code restructuring, the *main activities* are:

- ***Restructuring of the sensors (pushd/urcollector):***
  - to achieve a better decoupling between the production of the UR on the CE (needed also for interoperability with OSG), and their forwarding to the HLR.
- ***Rewrite DGAS2APEL in order to:***
  - Drop dependencies over perl-DBD,perl-DBI (in the past source of portability problems).
  - Be able to run DGAS2APEL also on Second Level HLRs (L2HLRs) and not just on Site HLRs.
  - C++ implementation allows for better performance and reuse of code already developed for the HLR, achieving easier maintenance of the code. (Work 50% Done.)
- ***Adoption of common logging format:***
  - Production release already able to log via SYSLOG facility.
  - Waiting for proper definition of the logging format to complete the task.

Future plans

Once these activities are over. Including the full support via ETICS for the reference platforms, we plan to freeze the code as much as possible (I.e. just critical bug fixes) and proceed with a deeper restructuring, focusing on:

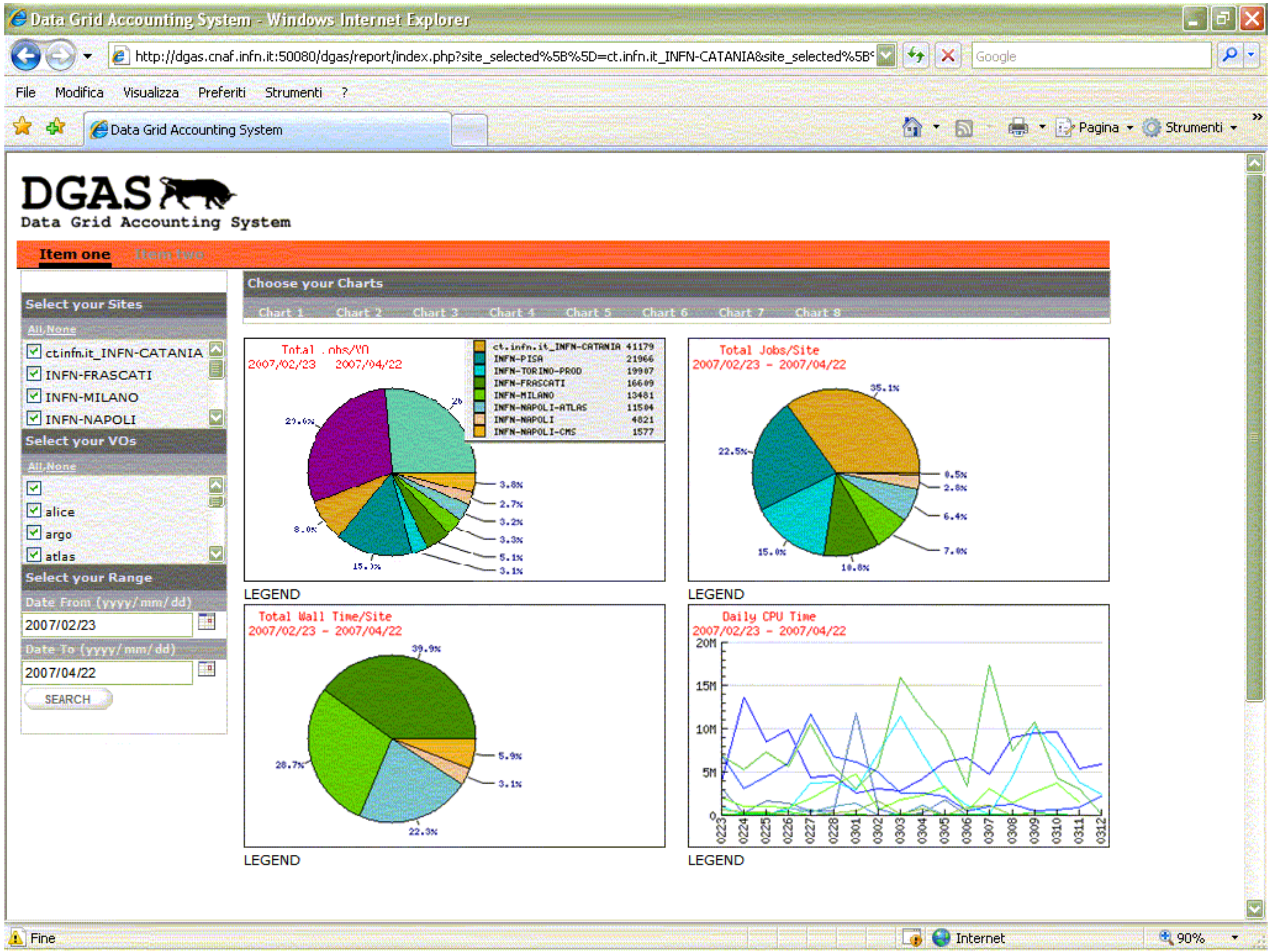
- ***Easier configuration:*** Introduce as much automatic tuning of the configuration parameters as possible, in order to reduce the effort required to system managers.
- ***Code clean-up:*** Remove obsolete and unused code to allow for better understanding of the code itself to new (and also old) developers.
- ***Database schema clean-up:*** Many years of on-demand new features without proper general planning result in a complex database schema that needs to be revised.
- ***Code profiling and optimization:*** In order to tune the (already good) performances, mainly in the query engines.

HLRMon (1)

Web Interface to DGAS HLR: HLRMon (Work in progress)

HLRMon, the web interface to DGAS is being developed by:

F. Pescarmona
S. Dalpra
F. Rosso
G. Misurelli
E. Fattibene
G. Patania



Data Grid Accounting System - Windows Internet Explorer

http://dgas.cnaf.infn.it:50080/dgas/report/index.php?site_selected%5B%5D=ct.infn.it_INFN-CATANIA&site_selected%5B

DGAS
Data Grid Accounting System

Select your Sites
All,None
☑ ct.infn.it_INFN-CATANIA
☑ INFN-FRASCATI
☑ INFN-MILANO
☑ INFN-NAPOLI

Select your VOs
All,None
☑
☑ alice
☑ argo
☑ atlas

Select your Range
Date From (yyyy/mm/dd)
2007/02/23
Date To (yyyy/mm/dd)
2007/04/22
SEARCH

Choose your Charts

🟧	ct.infn.it_INFN-CATANIA	41179
🟦	INFN-PISA	21966
🟦	INFN-TORINO-PROD	19907
🟩	INFN-FRASCATI	16609
🟩	INFN-MILANO	13481
🟦	INFN-NAPOLI-ATLAS	11504
🟧	INFN-NAPOLI	4821
🟧	INFN-NAPOLI-CMS	1577

HLRMon (2)

- **Shows accounting data in aggregate form**
- **A set of predefined aggregates are built using data available on DGAS HLRs.**
- **It is mainly intended as an interface toward Second Level HLRs.**
- **User is identified by means of his certificate and is allowed to plot charts according to his own VO role.**
- **These pre-defined roles are actually available:**
  - Normal User
  - VO Manager
  - Site manager
  - ROC Manager
- **Capability to completely customize the queries, as for the CLI interface, is foreseen (but need to pay special attention with authorizations).**

Conclusions

- DGAS is deployed on the Italian Production Grid. During the last year it has been thoughtfully evaluated and was subject to a fast turnaround cycle of user-driven improvements.
- Our experience demonstrated that *it is crucial to pedantically cross-check the information available in the relational databases with the raw source for these information*. This allows for immediate discovery of configuration problems, bugs or undesired behaviours.
- However this task is very difficult and time consuming.
- DGAS sensors and HLR server infrastructure is proven to be able to account job usage metrics with *good levels of reliability and precision*, up to the scale of the average output of a T1.
- We had many problems (with R-GMA??) using ‘apel-publisher’ to send the Usage Records produced by DGAS2APEL to the GOC repository. Are *alternative transport mechanisms* available? (directly use MySQL? L2HLR at GOC?)
- A full featured *web interface is in development*, and a first public version will be presented shortly.
- Now that the *core system is proven to be stable enough and presents all the required functionalities*, we plan to *freeze the development of new features* and concentrate on cleaning up the code and improve the overall user friendliness.

References

- **Information on DGAS can be found at:**
  - http://www.to.infn.it/grid/accounting
- **Problems with DGAS can be signalled to dgas-support[AT]to.infn.it**