

Accounting Facilities and Policies used by DEISA

Johannes Reetz

RZG, johannes.reetz@rzg.mpg.de

MLA Workshop, HPDC 2009,

Munich, Germany

June 10, 2009













Agenda



- ⇒ DEISA HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements

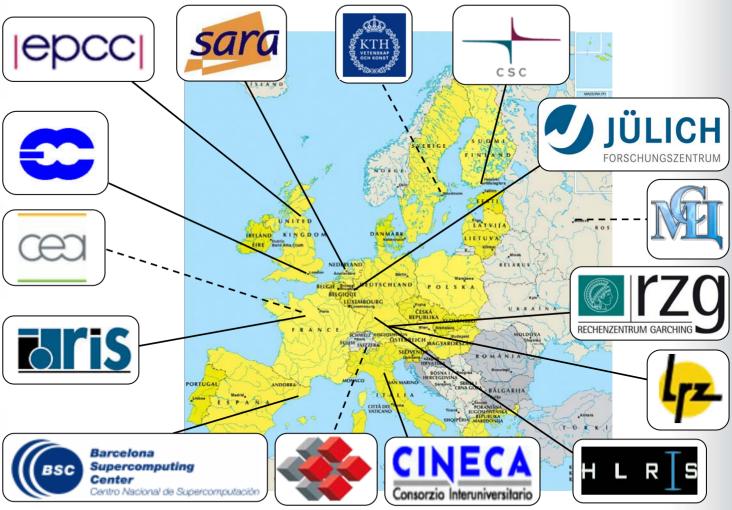
Agenda



- ⇒ **DEISA** HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements

DEISA consortium





Principal padræssociate partners



DEISA HPC compute resources

•	Distributed
	European
	Infrastructure for
	Supercomputing
-	Applications

Site	Architecture	CPU@Clock	Tf/s	#Cores	Memory
BSC	IBM PowerPC	PPC@2.3	94.2	10240	20.0
CINECA	IBM Blade Center	Opt. DC@2.6	26.6	5120	10.0
	IBM Power5-575	P5@1.9	3.9	512	1.0
CSC	Cray XT4 AMD 4-core	QC@2.3	32.3	4048	4.4
	Cray XT5 AMD 8-core		54.4	6816	6.1
ECMWF	IBM P5-575+	P4+@1.9	33.0	4552	
	Cray XT4	DC@2.8	59.0	11328	33.2
EPCC	Cray X2	X2	2.9	112	0.9
	IBM Power5-575	P5@1.9	15.4	2560	5.2
FZJ	BlueGene/P	PPC450+@0.85	223.0	65536	32.0
	IBM Power6-575	P6@4.7	8.4	448	1.8
HLRS	NEC-SX8	SX-8@2.0	12.7	576	9.2
IDRIS	BlueGene/P	PPC450+@0.85	139.0	40960	20.0
IDNIS	IBM Power6-575	P6@4.7	67.3	3584	10.8
LRZ	SGI-Altix Itanium2	I2 DC@1.6	62.3	9728	39.0
RZG	BlueGene/P	PPC450+@0.85	54.0	16384	4.0
KZG	IBM Power6-575	P6@4.7	120.0	6560	18.5
SARA	IBM Power6-575	P6@4.7	60.2	3328	15.6

Status: Apr 2009

Projects and Science Communities



DECI call 2005

29 proposals accepted 12 mio core-h granted*

DECI call 2006

23 proposals accepted 12 mio core-h granted*

DECI call 2007

45 proposals accepted 30 mio core-h granted*

DECI call and Science Communities 2008

42 proposals accepted 50 mio core-h granted* 3 virtual communities 7 mio core-h granted*

DECI call and Science Communities 2009

75 proposals >200 mio core-h requested*

>5 virtual communities

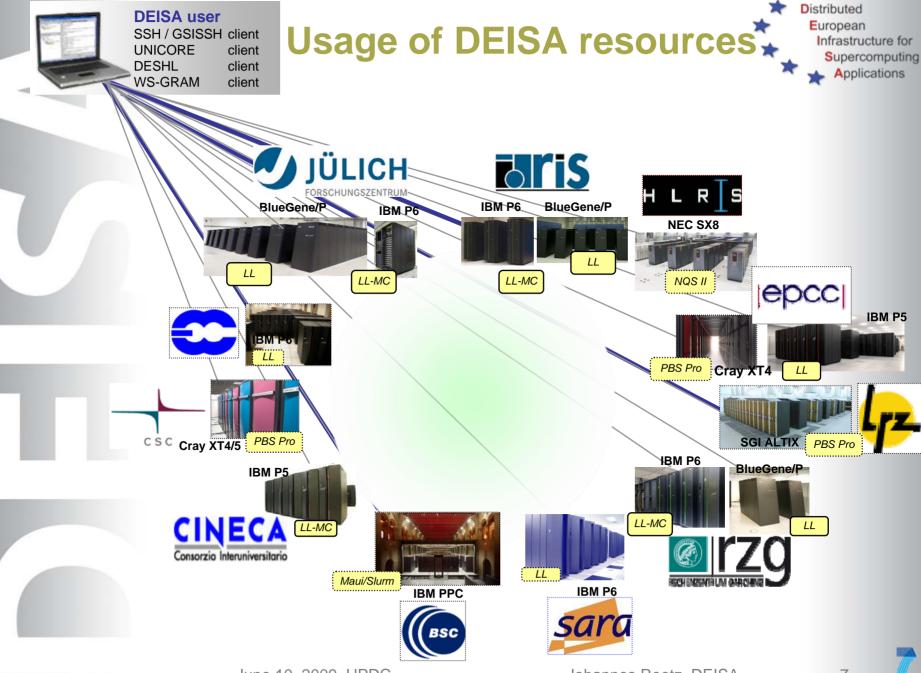
*) Core-h normalized to IBM P4+@1.6GHz

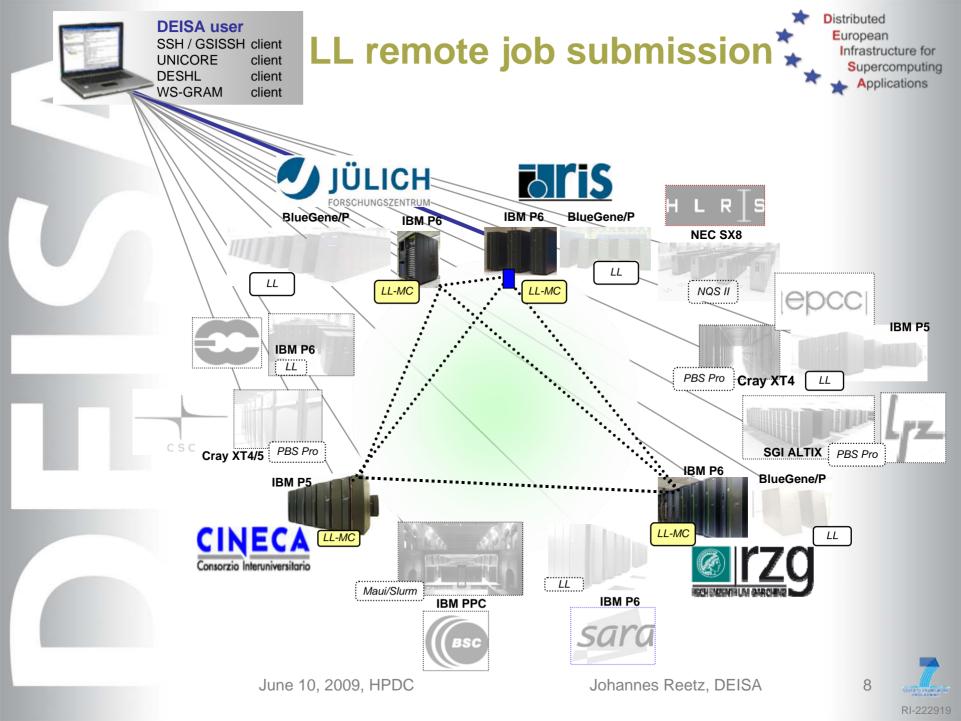
DECI: DEISA Extreme Computing Initiative

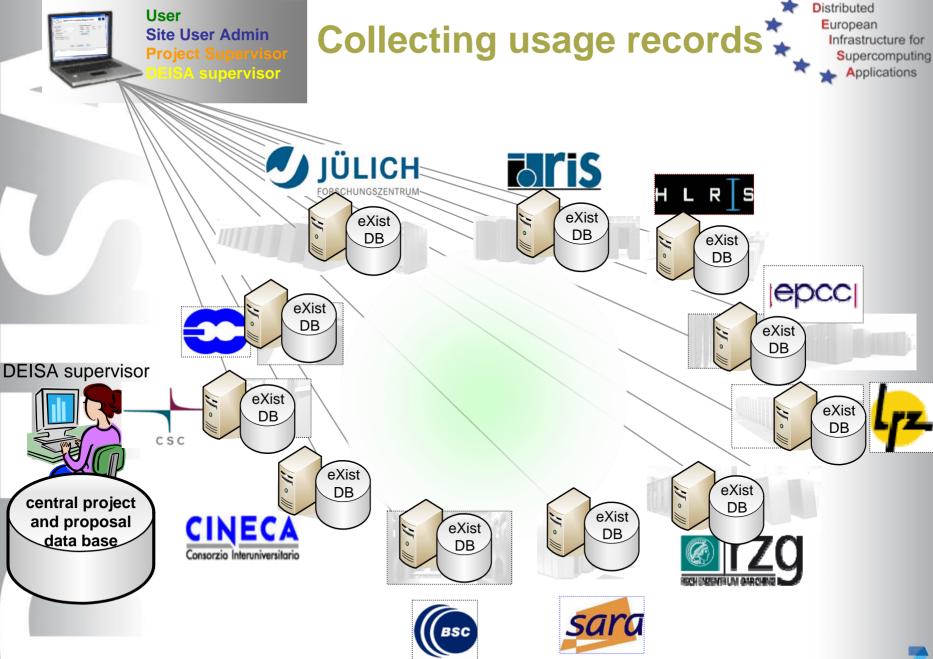
Yearly call for proposals

Communites: Virtual Scientific Communitities





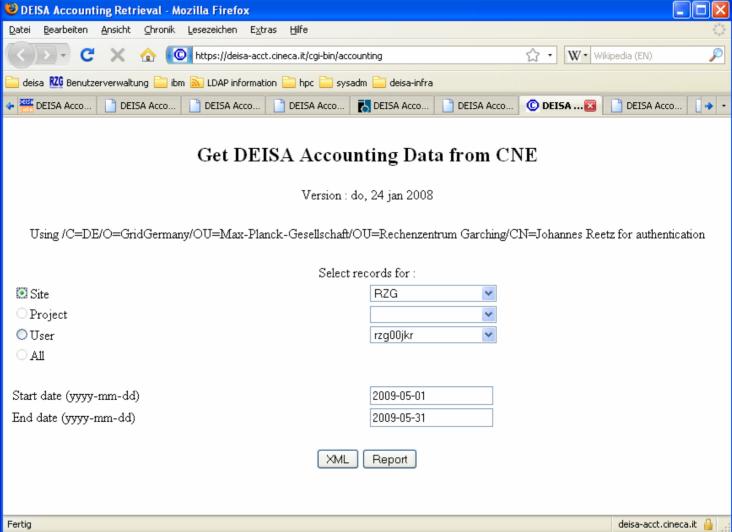






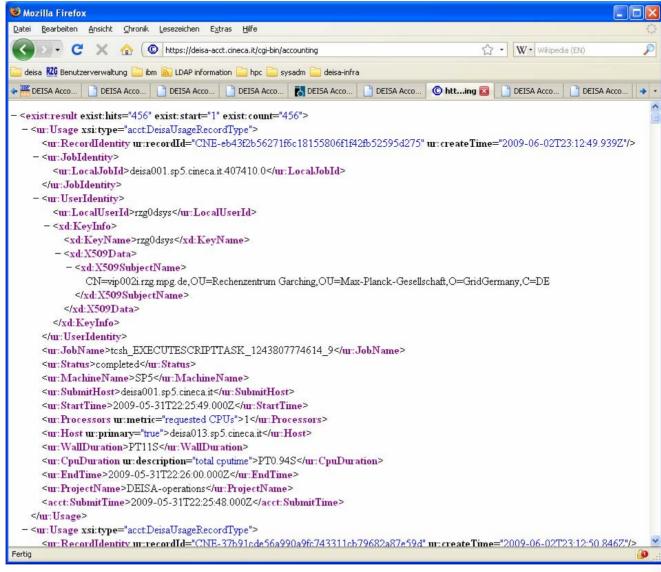
Collecting URs via Web interface





Results obtained via Web interface





RI-222919

Agenda



- ⇒ DEISA HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements

User and Project Administration Concept



- 1. Each user has an administrative DEISA Home Site
- 2. Each Project (or community) is administrated by an Home Site
- 3. A project is mapped to one more Execution Sites
- 4. The budget of compute cycles granted to a project is assigned to one or more **Execution Sites** for a given period of time.

 → the Execution Sites know about the budget of allocated core
 - → the Execution Sites know about the budget of allocated core cycles
- 5. The consumed computing resources are collected for each project monthly to be registered into the centralized project and proposal data base for project controlling purposes.

Projects and Proposal Data Base



U DECI Database - Mozilla Firefox	
Detei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe	9.0
C X 🟠 📳 https://www.rzg.mpg.de:8443/cgi-bin/deci.pl	P
eisa R26 Benutzerverwaltung 🔤 ibm 🔊 LDAP information 🗀 hpc 🔛 sysadm 🦲 deisa-infra	
*	
DECI Database Welcome, Johannes Reetz Database: DECI-5 (2009) 🔻 access (DEISA@RZG, DARD) Ga	
Problems ? Contact andreas.schmidt	at rzg.mpg.de
3.1 Project list views	[hide]
3.1.1 Simple list of projects (without site): standardized Core hours show only PIs split int./ext. show full affiliation	show
3.1.2 Complete survey sorted by projects: standardized Core hours on o enabling of show homesite show suitability on dates	show
3.1.3 Complete/partial survey sorted by execsites: standardized Core hours v no enabling v show homesite show suitability no dates v all sites v	show
3.1.4 Complete/partial survey sorted by homesites: standardized Core hours v no enabling v show execsites show suitability no dates v all sites v	show
3.1.5 Complete/partial survey sorted by discipline/project: standardized Core hours V no enabling V show homesites show suitability no dates V all disciplines V	show
3.1.6 Committed/assigned survey sorted by execsites: standardized Core hours v no enabling v show homesite show machines no dates v all sites v	show
3.1.7 Monthly survey sorted by projects: standardized Core hours V show homesite 01 V / 09 V until 05 V / 09 V	show
3.1.8 Monthly survey sorted by execsites: standardized Core hours V show homesite 01 V / 09 V until 05 V / 09 V all sites V	show
3.1.9 Monthly survey sorted by homesites: standardized Core hours 🔻 🗆 show machines 01 💌 / 09 💌 until 05 💌 / 09 💌 all sites 💌	show
3.2 Additional project views	[hide]
3.2.1 Complete dump for project: ACES-X 🔻 show 3.2.6 Display all sites standard. Core hours (& committed) 🔻 🗌 show machines show	
3.2.2 Display all resources of all projects show 3.2.7 Display all Core factors show	
3.2.3 Display all used codes sorted by discipline ▼ show 3.2.8 Display all machines (original Core hours) show	
3.2.4 Display affiliations sorted by country ♥ show 3.2.9 Display all aggregations show	
3.2.5 Display investigators sorted alphabetically	
4. Additional views/tools	[hide]
* Additional views/tools	inac
4.1 Full DEISA survey sorted by execsites/projects standardized Core hours (& committed) 🔻 no enabling 🔻 🗆 split int./ext. 🗕 💌 / 09 until 🗕 💌 / 09 🔻 show	
4.2 Complete survey for virtual communities sorted by communities standardized Core hours v summary v no enabling v - v / 09 until - v / 09 v show	
4.3 Complete survey for virtual communities sorted by execsites standardized Core hours v summary v no enabling v - v / 09 until - v / 09 v coming soon !	
4.4 Complete dump for virtual community EFDA ✓ show	
4.5 Complete dump for virtual community project F-CEA-09 ▼ show	
4.6 DEISA-internal projects usage standardized Core hours 🔻 all projects 🔻 🗆 split origin 🗆 show machines - 💌 / 09 until - 💌 / 09 show	
4.7 DEISA-internal projects usage by execsites standardized Core hours 🔻 all sites 🔻 📗 split origin 📗 show machines 🗕 🔻 / 09 until 🛑 🔻 / 09 show	
4.8 DEISA-internal projects usage by origin standardized Core hours vall sites value split executes show machines value	
4.9 Project assignment-support standardized Core hours 🔻 all sites 💌 🗆 modification-mode 🗀 show suitability show	
× Suchen: star ♣ Abwärts ♠ Aufwärts № Hervorheben ☑ Groß-(Kleinschreibung 1 Ausdruck nicht) gefunden	
Fertig	(1)
-	

Projects and Proposal Data Base (2)



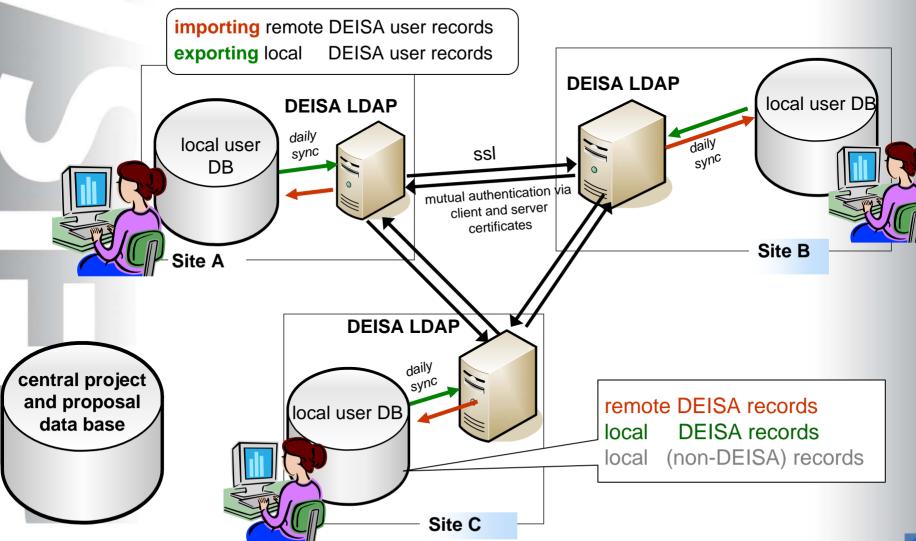
Contains information about

- project acronym and unique project identifier
- project membership
- project supervisors
- project execution machines and corresponding budgets
- core cycles (job time) per project assigned to execution machine
- CPU normalization factors



DEISA User Administration System (UAS)





Agenda

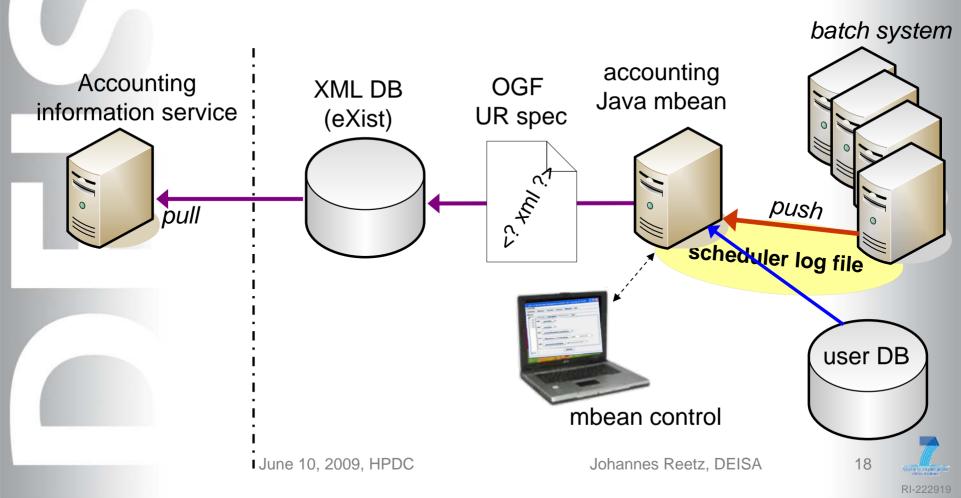


- ⇒ DEISA HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements

Accounting Data Provider Tool (1)

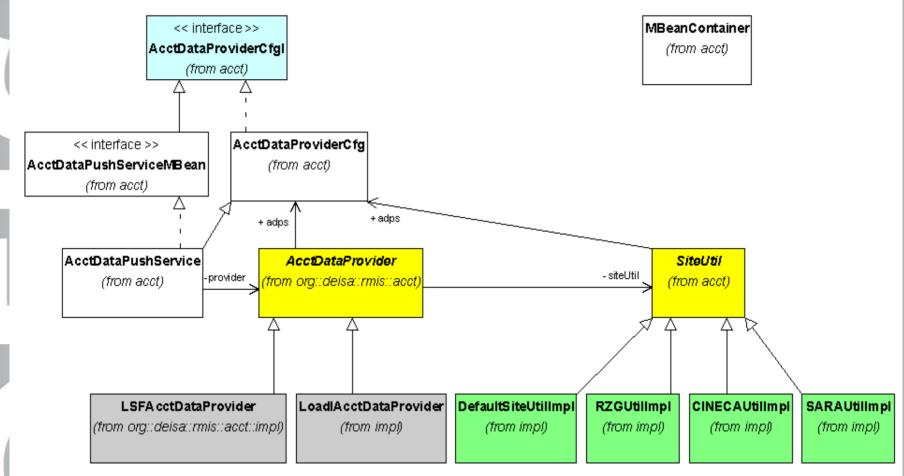
Distributed
European
Infrastructure for
Supercomputing
Applications

- implemented in Java
- analyses the batch scheduler log files (e.g. LoadLeveler history file)
- adds missing information from local user DB or directly from DEISA Idap)
- the result are XML records which are stored in eXist DB



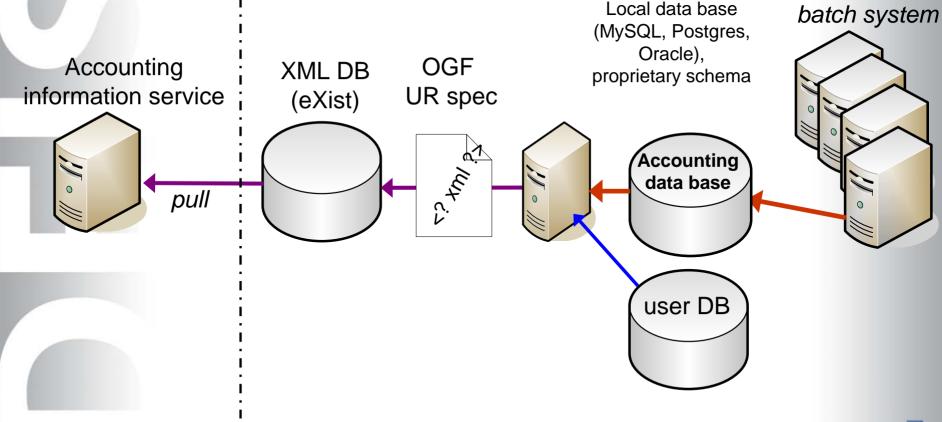
Accounting Data Provider Mbean design



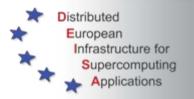


Accounting Data Provider Tool (2)

- Distributed
 European
 Infrastructure for
 Supercomputing
 Applications
- local accounting data base, stored information from batch system log file
- a second procedure (e.g. Perl script) creates XML usage records, missing information comes from user DB (or from DEISA Idap directly)
- XML records are stored in eXist DB



UR Properties used by DEISA



UR-WG Element name	Description				
RecordIdentity	Identifies uniquely the usage record				
JobIdentity	Contains local job identifier (LocalJobId) as assigned by the batch queue and a GlobalJobId (may be LocalJobId with a sitename prefixed)				
UserIdentity	The username the job has run under (LocalUserId) and the Subject name of the X.509 cert (Keyinfo)				
JobName	The global job name				
Status	Completion status of job, e.g. completed, aborted.				
WallDuration	Total wall clock time that elapsed while the job was running.				
CpuDuration	Total CPU time used, summed over all processes of the job				
MachineName	A descriptive name of the system on which the job ran				
Host	The system hostname on which the job ran				
SubmitHost	The system hostname from which the job was submitted				
ProjectName	The name of the project that the job was run under				
Processors	The number of processors used or requested (reserved)				
EndTime	The time at which the usage ended				
StartTime	The time at which usage started				
NodeCount	The number of nodes used				
SubmitTime	Not a UR-WG defined property. It gives the time the job is submitted to the system the job has run on.				

UR Properties used by DEISA



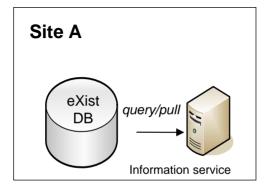
Not all attributes come from log files of the batch system

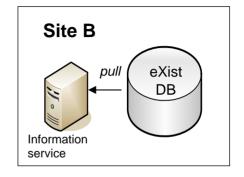
- Key info (X.509 Subject name)
- Project name
- Machine name
- → additional information from local user DB or DEISA LDAP
 - Key info and project name can be retrieved from user DB
 - Machine name from local file

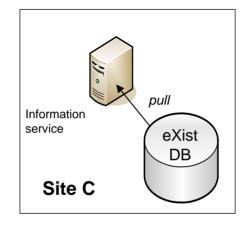


Accounting Information Services *









Distributed European **Information Requestors** Infrastructure for Supercomputing Applications **DEISA Supervisor** gets all usage records of user from site A **Site Admin** gets all usage records of user from site A Site A Site B eXist eXist DB query/pull DB Information Information service service **Project Supervisor User** obtains usage records obtains his own from users of his project. pull Information usage records service eXist DB Site C

Accounting Information Service



- CGI script in Apache HTTPS server (REST like interface)
- Protected with X.509 certificates
- 4 roles: User, Site administrator, Project supervisor, and DEISA supervisor
- Access control via accounting-grid-mapfile
- accounting-grid-mapfile can be automatically generated from information of DEISA UAS

```
"/O=dutchgrid/O=users/O=sara/CN=Jules Wolfrat" site-SARA
"/O=dutchgrid/O=users/O=sara/CN=Jules Wolfrat" user-sar00005
"/O=dutchgrid/O=users/O=sara/CN=Bart Heupers" project-xyz
```



DART Client Tool



- DART = DEISA Accounting Reporting Tool
- Java 1.5 program → can be used on most platforms
- Supports PKCS12 and Java Keystore for authentication
- Connects and processes accounting data from all sites simultaneously
- Java WebStart application. No local configuration needed.
 URLs of accounting information servers and processors normalization factors always up to date
- User friendly GUI



Obtaining Usage Reports



<u> </u>				
Distributed European Infrastructure for Supercomputing Applications DEISA Account	ing Rep	ort To	ool	i ?
Role selection DEISA user DEISA project DEISA site DEISA supervisor Date selection Starttime (YYYY-MM-DD): 2009-03-01 Endtime (YYYY-MM-DD): 2009-06-07 BACK Save	AUTO CARROTTO	✓ LRZ ✓ EPCC ✓ HLRS tion	✓ RZG ✓ IDRIS	⊮ FZJ ⊮ CSC

Usage Reporting (2)



### User User User User User User User User	Site / Machine SUMMARY SUMMARY RZG / VIP	Jobs 13118	Cpu Time (norm) [h]			
LL rg0joec (CN= LL g0ddsys (CN=ţ ^^ je)	SUMMARY SUMMARY RZG / VIP	13118				
LL g0joec (CN=	SUMMARY RZG / VIP			Cpu Time [h]	Job Time (norm) [h]	Job Time [h]
rgOjoec (CN= LL rgOdsys (CN=ţ ^^	RZG / VIP		2093293,6493		2186898,1254	5.00
LL rg0dsys (CN=k ^^		6	4732,5742		11264,8000	
rg0dsys (CN=r '^* le)		6	4732,5742	1577,5247	11264,8000	3754,93
	SUMMARY	11434	78,0243		230,9724	
	EPCC / HPCx	770	0,0311	0,0311	0,6069	0,60
	HLRS / NEC SX-8	745	0,0000	0,0000	2,3600	0,39
rg0dsys (CN=vij 3)	BSC / MareNostrum	1413	4,5838	5,7297	4,5838	5,7:
	CNE/SP5 CSC/Louhi QC	1432 377	0,2762 1,5423	0,1842 1,2853	13,3771 1,5423	8,9° 1,21
		272	0,0008	0,0008	0,1903	0,11
	EPCC/HPCx FZJ/JUMP-P6	1012	0,0008	0,0008	20,6800	6,8
	LRZ/HLRB II	1012	0,2050	0,2769	4,1328	3,7
	RZG / VIP	1586	13,0708	4,3569	6,9075	2,3
	SARA/HUYGENS	1398	55,8008	18,6003	117,7317	39,2
g0dsys (UNKNOWN)	IDR / VARGAS	1334	1,6767	0,5589	58,8600	19,6
LL	SUMMARY	569	84589,6036	0,0000	84574,4516	15,0
rg16tsk (CN=:,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	BSC / MareNostrum	569	84589,6036	105737,0044	84574,4516	105718,0
LL	SUMMARY	409	406227,2892		435404,4750	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
rg17boe (UNKNOWN)	RZG / VIP	263	86535,7408	28845,2469	106994,5817	35664,8
g17fre (UNKNOWN)	RZG / VIP	146	319691,5483	106563,8494	328409,8933	109469,9
ĹĹ	SUMMARY	104	232800,9849		235907,8693	
g18bds (CN	BSC / MareNostrum	15	692,5789	865,7236	692,4869	865,6
	RZG / VIP	87	232108,0933	77369,3644	235215,0767	78405,0
g18ttr (N/A)	BSC / MareNostrum	2	0,3127	0,3908	0,3058	0,3
LL	SUMMARY	34	90401,5360		90401,5360	
g13aub (UNKNOWN)	EPCC/HECToR	34	90401,5360	75334,6133	90401,5360	75334,6
LL	SUMMARY	13	300,2942		392,6933	
g26dum (UNKNOWN)	RZG / VIP	13	300,2942	100,0981	392,6933	130,8
LL	SUMMARY	1	13,8525		85,3333	
g14huy (UNKNOWN)				4,6175		28,4
g12bds (CN						0,0
				159623,8631		164888,2
				164 6000		220.0
		10000		161,9283		279,3
				47070 1544		48168,91
				4/0/0,1544		40108,91
rg21zhu (UNKNOWN)				4 0547		5,21
LL g12bds LL g24ser LL g11vla LL	s (Ch (UNKNOWN)	SUMMARY S (CN EPCC / HECTOR RZG / VIP SUMMARY RZG / VIP SUMMARY (UNKNOWN) RZG / VIP SUMMARY SUMMARY	SUMMARY 52	SUMMARY 52 478871,5892	SUMMARY 52 478871,5892	SUMMARY 52 478871,5892 49464,7508 S(CN EFPC/HECTOR 12 0,0000 0,0000 0,0000 RZG/VIP 40 478871,5892 159623,8631 494664,7508 SUMMARY 44 485,7850 837,9933 RZG/VIP 44 485,7850 161,9283 837,9933 SUMMARY 83 143610,4633 144506,7200 (UNKNOWN) RZG/VIP 83 143610,4633 47870,1544 144506,7200 SUMMARY 7 12,1642 15,7867

INFO: Retrieve data from https://ldeisap.sara.nl/cgi-bin/accounting

INFO: Retrieve data from https://deisacc.zam.kfa-juelich.de/cgi-bin/accounting

INFO: Retrieve data from https://gridacct.lrz-muenchen.de/cgi-bin/accounting

Agenda



- ⇒ DEISA HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements

Procedures and Policies



- LDAP server and user DB of the DEISA UAS synchronize in less than 24h
 - → changes in a local DB propagate within 24h
 - → this includes modifications of user roles
- Accounting information service (eXistDB) is up-to-date within 24h
 - → local accounting data provider must sychronize with 24h

- eXist DB data remains persistent for at least one year at all sites



Agenda



- ⇒ DEISA HPC Resources, Projects, Accounting Requirements
- ⇒ User administration: Projects, Users, Budgets
- ⇒ Accounting facilities
 - Site-local accounting tools (Accounting Data Provider)
 - Utilization of the OGF Usage Record format v1.0
 - Accounting Information Service
 - Accounting Reporting Tool (DART)
- ⇒ Procedures and Policies
- ⇒ Lessons learned and future Improvements



Lessons learned



Advantages:

- Standardized protocols and XML schema
- No central database: distributed XML databases, information services
- X.509 based authentication, authorized access
- Simple system architecture, easy to used by users, supervisors, ...
- Open Source (Apache HTTP Server, eXist DB, Java, PERL)

Problems, open topics:

- All the requested usage records must be transferred from the eXistDB to the client for report generation
 (1UR ~ 1,5 kB ⇒ 10.000 URs ~ 14,7 MB ⇒ 1.000.000 URs ~ 1,4 GB)
 large numbers of records → high performance load of the eXistDB
- Risk of incomplete reports, if a site is not reachable (1 site: 99% availability => 11 sites: 89,5% availability)
- Risk of incomplete reports if eXist data is complete due to failures of the local accounting data provider.

Further Improvements



- Allow Supervisors to see only summaries over a longer period
 - → less details
 - → more user data privacy
- Aggregation of UR Properties
- Introduce Budgeting using the LDAP infrastructure to propagate execution site specific project budgets
- Report generation, e.g. summation over URs on the server side
 Summarized values (e.g., jobtime within a given time frame) sufficient for client
 → transfer of less amount of data
- Improve availability of service (Data Replication? Central Repository?)
 - Possible strategies:
 - save remote accounting data into databases at the homesite of the user and let users obtain their data from their homesite only
 - central database for aggregated accounting information

Summary



- Accounting data is gathered
 - → directly from **batch system logs** (Loadleveler, PBS, NQS, Slurm)
 - → or from an **accounting database** (e.g. mysql, postgres,...) that was previously fed by information from the batch systems.

Proper project assignment requires information from the UAS.

- Accounting data is stored in eXist XML database using OGF UR-WG format
- Distributed eXist DBs and information services provide accounting information.
 Information server realized as HTTPS server with CGI script.
 Each site controls who gets access to which usage records
- Authorized access based on X.509 certificates, using secure communication (https),
 different roles: user, site admin, supervisors
- **DART Client tool** collects all data simultaneously and creates reports (e.g. taking processor performance factors into acount).
- Jobtimes (#cores * elapsed time) are normalized before summing up across different computer platforms
- Accounting information from all sites is regularly collected by the DEISA supervisor and fed back into the projects and proposal database for project progress controlling purpose.