

Grid Middleware and Networking Technologies Track Summary

Aleš Křenek and Francesco Giacomini

Track contributions



- ▶ 44 oral presentations
- ▶ 76 posters

By category:

10	general
3	software distributions
14	operations
10	security
22	data
16	workload
28	monitoring
2	information systems
7	networking
4	interoperability
4	other

By origin:

5	ALICE
8	ATLAS
15	CMS
3	LHCb
4	CDF
15	EGEE
34	LCG
4	Nordugrid
13	OSG
19	other

- ▶ Monday 2pm: infrastructures view
- ▶ Monday 4pm: experiments view
- ▶ Tuesday: security & software distribution
- ▶ Thursday 2pm Panorama: data management
- ▶ Thursday 2pm Club B: monitoring
- ▶ Thursday 4pm Panorama: workload management
- ▶ Thursday 4pm Club B: networking and interoperability

- ▶ reports from EGEE, OSG, GridPP, GridKa
- ▶ move from building production infrastructures to running them
 - ▶ established production teams
 - ▶ staffing is one of key risks
- ▶ pledged resources delivered
- ▶ service reliability and availability critical
 - ▶ routine SAM and Nagios monitoring, dashboards
 - ▶ improving in general, SLA (EGEE) help
 - ▶ clustering core services, failover mechanisms
 - ▶ more critical for storage (5 files @ 90 % yield 60 % altogether)
- ▶ introduction of formal processes (ITIL in GridKa)
 - ▶ implemented Configuration Management, Incident Management, Problem Management and Change Management
 - ▶ more paperwork but general improvement

Grid Operations Center DB (GOCDB)

- ▶ what forms the infrastructure, contacts etc.

European Grid Initiative (EGI)

- ▶ general purpose grid infrastructure, uniform and secure sharing
- ▶ build on NGIs + small coordinating body
- ▶ structure, functions, and actors
- ▶ middleware is one of the functions
 - ▶ essential to enable the infrastructure
 - ▶ development by middleware consortia, coordination by a Middleware Unit inside EGI.org



Grid Middleware for WLCG

- ▶ serious critical view of big “customer”
- ▶ not all expectations met by the grid
- ▶ some things work: single sign-on, data transfer, simple catalogues, etc.
- ▶ disappointment in robustness and reliability
- ▶ generic services not easy to achieve
- ▶ decouple complexities and dependencies, leverage virtualization
- ▶ extensive discussion, most controversial talk
 - ▶ eagerly waiting for the paper for a more in depth analysis

CDF way to Grid

- ▶ “gridification” with minimal impact on users
- ▶ transparent use of tools, still Kerberos
- ▶ backends for Condor Glideins (OSG) and gLite WMS (EGEE)

CMS Distributed Computing System

- ▶ distributed production & analysis (CRAB + various gLite WMS's)
- ▶ 30–50 kjobs/day in single instance
- ▶ data transfer (PhedEx) critical – sustained 1-2 GB/s
- ▶ site readiness: overall metrics, SAM + commissioned links
- ▶ initial + sustained effort required



Atlas – software installation

- ▶ previous implementation (simple wget) saturated
- ▶ push software to near storage elements
- ▶ install via Panda pilot jobs (new type)

Atlas – migration of Panda servers

- ▶ execution system of all Atlas jobs
- ▶ developed and deployed in USA, migrated to CERN recently
- ▶ 35k simultaneous jobs, 150 kjobs/day
- ▶ careful planning and preparation, useful experience

Critical services in LHC

- ▶ “service readiness” questionnaire (eg. documentation, support process, automatic configuration)
- ▶ experiment ranking of services
- ▶ find out where effort is required

SVOPME: A Scalable Virtual Organization Privileges Management Environment

- ▶ many to many relationship between VOs and sites does not scale
- ▶ tools to express policies, verify, and help sites implementing
- ▶ similarities with the new gLite AuthZ service

New gLite Authorization System

- ▶ survey of gLite AuthZ mechanisms showed inconsistencies
- ▶ goal is to replace all currently used AuthZ solutions used in EGEE
- ▶ first key requirement: ban misbehaving user completely
- ▶ failure resistance, coherent policy handling, etc.
- ▶ deployment plan: central banning service, then CE, WMS
- ▶ DM is a bit special and requires further thinking
- ▶ similarities with SVOPME

An XACML profile and implementation for Authorization Interoperability between OSG and EGEE

- ▶ X509 certificates + attributes
- ▶ credentials and attributes pushed to resources
- ▶ local-granted privileges evaluated (PEP calls PDP)
- ▶ different AuthZ implementations in OSG and EGEE
- ▶ interoperation profile defined,
 - ▶ based on XACML and SAML
 - ▶ defines attributes names and meaning
- ▶ deployed in testbeds of both OSG and EGEE

VOMRS / VOMS Utilization Patterns And Convergence Plan

- ▶ VOMS-Admin: simple and intuitive, non JSPG compliant
- ▶ VOMRS: complex, on top of VOMS-Admin
- ▶ short term plans: JSPG compliance in VOMS-Admin, migration of essential VOMRS features



On the role of integrated distributions in grid computing

- ▶ known drawbacks of current approach
- ▶ technology as well as funding structure change
- ▶ service-oriented independent distributions proposed
- ▶ virtualization, more apparent vertical cut to service stacks
- ▶ hot topic – extensive discussion

CDF software distribution on Grid using Parrot

- ▶ complete kit is large
- ▶ NFS, AFS, GridFTP etc. evaluated
- ▶ Global Read Only Web FileSystem developed
- ▶ caching and integrity managed
- ▶ just-in-time file access with Parrot

Modern methods of application code distributions

- ▶ ALICE + required middleware: 200 packages, 300 MB to WN
- ▶ formerly in shared area: slow, security issues
- ▶ automatic installation within AliEn job agent
- ▶ use Bittorrent to retrieve files
- ▶ require open ports on the Worker Nodes within the site (CERN in the specific case)

PhEDEx Data Service

- ▶ data placement for CMS
- ▶ provide policies, reliability, etc.
- ▶ talk focus on monitoring (web data service)
- ▶ many types of data available: known datasets, their location, transfer requests, various statistics, monitoring
- ▶ integration of that information into other data management components, in particular for monitoring

Data Management @ CERN

- ▶ based on CASTOR
- ▶ CCRC'08 has validated the architecture currently in production
- ▶ DM ready for LHC startup
- ▶ areas of recent improvement
 - ▶ monitoring, security, SRM interface, tape efficiency, file access latency

Data Management in EGEE

- ▶ detailed roadmap for DPM (storage element), LFC (file catalog), FTS (file transfer), GFAL (posix-like access to SRM-based storage), lcg_util (command lines for the most common used data operations), Hydra (encrypted storage)
- ▶ recent and future improvements in terms of: stability, reliability, maintainability, performance, administration, monitoring, integration with other services

StoRM performance and scalability

- ▶ SRM implementation on top of parallel and cluster file systems
- ▶ test Front End, DataBase, Back End
- ▶ different deployments possible
- ▶ FTS use case @ 15HZ, <1s response time



dCache

- ▶ independent organization, sustainable funding, integrated product
- ▶ SRM interface, multiple access protocols, multiple mass storage systems, internal management transparent to users
- ▶ SRM, learning by doing, collaboration with the whole Grid DM crowd
- ▶ recent improvements, e.g. use of chimera
- ▶ standardization: GLUE, NFS 4.1 (pNFS) will give posix access for free (is there need for specific protocols?)
- ▶ ready for analysis? it doesn't look bad now, but room for improvement
- ▶ for fun: dCache on S3

- ▶ cannot afford to ignore major trend in the computing industry
- ▶ three-point checklist for viability of clouds:
 1. non trivial quality of service must be achieved (point included also in the three-point checklist for a Grid, by I. Foster)
 2. the scale of the test(s) must be meaningful for petascale computing
 3. data volumes, rates and access patterns representative of LHC data acquisition, (re-)processing and analysis
- ▶ and look at the cost (of entry; of ownership)
- ▶ from the discussion
 - ▶ the other two points on Foster's list should be considered, in particular open protocols and standards
 - ▶ maybe we haven't found yet the right abstractions we need; data is key

CREAM CE

- ▶ scalable and flexible architecture
- ▶ support for multiple batch systems
- ▶ full support for proper AuthN and AuthZ and accounting
- ▶ CREAM available in production since Oct 2008
- ▶ used in particular by ALICE
- ▶ submission through WMS available, but still some scalability issues
- ▶ defined criteria for the transition from LCG-CE to CREAM
- ▶ tests very satisfactory, almost no failures and better performance than LCG-CE
- ▶ focus on standardization, participation to OGF WGs, but implementation waiting for stable standards (BES/JSDL)
- ▶ interoperability: submission to CREAM from Condor-G and ARC

First experiences on using the CREAM CE

- ▶ All LHC VO experiments expressed their interest to use the CREAM-CE in direct submission mode
 - ▶ ALICE has been testing the direct submission since Summer 2008, with successful results
- ▶ the submission via the WMS (CMS) and Condor-G (CMS and ATLAS) are also required
 - ▶ CMS is currently testing the WMS setup with promising results
- ▶ ATLAS waiting until full support of CREAM in CondorG is available and the full deployment of CREAM is completed
- ▶ LHCb foresees the testing in about one month
- ▶ WLCG encourages all LHC sites to provide CREAM-CE services to all experiments in parallel mode to the LCG-CE
 - ▶ opportunity to check the system by the experiments
 - ▶ provide useful feedback to developers and site admins

ALICE WMS

- ▶ basic concepts: central task queue and optimizers, site VO-Box, job agent
- ▶ model mature, system stable and scalable
- ▶ experience with WMS disappointing
- ▶ CREAM matches better the model and is reliable and efficient

DIRAC3

- ▶ distributed data production and analysis system used in LHCb, it integrates both job and data management
- ▶ based on pilot jobs
- ▶ design and implementation reviewed by independent experts ⇒ DIRAC3
- ▶ addresses recommendations on security, instabilities of grid resources, comprehensive monitoring
- ▶ all analysis and production jobs in the same instance, VO-wide policies can be applied
- ▶ up to ~15Kjobs concurrently, further optimizations are possible

Use of the glite-WMS in CMS

- ▶ system based on BossLite, a common grid/batch interface
- ▶ interface to gLite based on full use of the gLite WMS and L&B
 - ▶ including bulk submission, bulk matchmaking, bulk status query
 - ▶ fruitful collaboration with the developers, addressing CMS needs
- ▶ both for production and analysis
- ▶ from may 2008 to march 2009 23Mjobs in total
- ▶ no scalability problem expected at the expected rates

GlideinWMS in CDF

- ▶ generic pilot-based workload management system
- ▶ CDF using glideins for the past 4 years
- ▶ focus on security, based on GSI
- ▶ improved cpu and memory footprints over glitekeeper (the previously used tool)
- ▶ 20-25 Kjobs per day



New strategy for job monitoring on the WLCG scope

- ▶ essential, resembles overall infrastructure quality
- ▶ large scale (200k CMS jobs/day)
- ▶ view levels: global, VO, site, user
- ▶ combination of middleware and application efficiency
- ▶ MSG (ActiveMQ) foreseen as common message delivery bus
- ▶ recent development of L&B to MSG publisher

Real Time Monitoring of Grid Job Executions

- ▶ graphical display of job flow on the grid
- ▶ L&B is primary information source
- ▶ over 70M LCG jobs seen, superlinear grow since 2005
- ▶ new development (L&B harvester) – streamline data acquisition
- ▶ reliable monitoring system, used also for dissemination

Evolution of SAM in an enhanced model for monitoring WLCG services

- ▶ drawback of current state
 - ▶ long time to alert admins, single point of failure, scaling issues, no history of topology
- ▶ new architecture
 - ▶ topology history, metrics description, results store
 - ▶ Nagios for actual monitoring (autoconfigured)
 - ▶ messaging system (ApacheMQ)

RSV: OSG Grid fabric monitoring and interoperation with WLCG

- ▶ initial goals: simple probes by experts, monitoring for local admins
- ▶ 106 of 131 services reporting to RSV
- ▶ probe output based on GMWG specification
- ▶ MyOSG presentation: consolidate data sources to give useful view
- ▶ use Universal Widget API (UWA) – customized views with iGoogle etc.



The impact and adoption of GLUE 2.0 in the LCG/EGEE production Grid

- ▶ GLUE 1.x
 - ▶ known limitations, too many embedded assumptions
 - ▶ improvements blocked by backward compatibility
- ▶ GLUE 2.0
 - ▶ generic concept of service
 - ▶ many-to-many relationships: complicated but flexible
- ▶ transition is starting, will take several years
- ▶ important positive case in point
 - ▶ convergence of de-facto and formal standard
 - ▶ there are reasons for breaking compatibility sometimes
 - ▶ enormous effort

Status and Outlook of HEP Network

- ▶ the major R&E networks serving HEP have progressed rapidly over the last few years
- ▶ our BW usage has kept pace
- ▶ groups in HEP have developed state-of-the-art methods to use these networks most effectively
- ▶ adapting the LHC computing models to fully exploit networks would have a profound positive impact on the LHC program
- ▶ urgent to close the digital divide

Where is the internet heading to

- ▶ IPv4 can't continue "as is" beyond 2011
- ▶ IPv6 looks "almost" unavoidable but is by no means "guaranteed" to happen!
- ▶ Last major architecture change was the introduction of MPLS
- ▶ The instability of the Internet routing system is preoccupying as well as the increasing lack of "network neutrality", copyright infringements, etc.

Monitoring and operational management in USLHCnet

- ▶ US Tier-1's
- ▶ using MonAlisa framework (fully distributed, no single point of failure)
- ▶ specific USLHCNet monitoring modules
- ▶ redundant sensors
- ▶ actions (alarms etc.) triggered locally and globally
- ▶ automatic path recovery

Deploying distributed network monitoring mesh for LHC Tier-1 and Tier-2 sites

- ▶ centralized monitoring does not scale
- ▶ perfSonar – collection of monitoring services
- ▶ live linux CD, self-contained, zero configuration
- ▶ information service: global lookup, topology service
- ▶ currently all Tier-1's, planned to all Tier-2's

WAN Dynamic Circuit Support at Fermilab

- ▶ isolate high-demanding traffic, guarantee bandwidth
- ▶ static configuration does not scale
- ▶ explicitly application driven too complicated
- ▶ flow-triggered configuration
 - ▶ on-line monitoring
 - ▶ history (traffic A-B lasts for 30 minutes typically)



Grid Interoperation with ARC middleware for CMS experiment

- ▶ gLite – ARC interoperation
- ▶ data transfers OK, both use SRM
- ▶ CRAB (CMS Remote Analysis Builder)
 - ▶ submit via gLite WMS, gateway to ARC
 - ▶ no job changes, gLite runtime at ARC WN
- ▶ ProdAgent
 - ▶ direct plugins for ARC
- ▶ discussion: similarities with Atlas

ARC middleware: evolution towards standards-based interoperability

- ▶ everything works, why change?
 - ▶ Grid is about sharing, standards must be in place
- ▶ different level of maturity (OGF context)
- ▶ new ARC client and hosting environment
 - ▶ breakdown to services, WS-based
- ▶ further evolution towards components in UMD

- ▶ production grids are there
- ▶ middleware is usable and used
 - ▶ job management, job management, monitoring, security
 - ▶ some expectations were not met, why?
- ▶ standards are emerging
 - ▶ long but unavoidable way to go
- ▶ networking
 - ▶ bandwidth use keeps pace with technology progress
 - ▶ urgent to close digital divide

Reviewers and session chairs

Mine Altunay, Simone Campana, Claudio Grandi, Michael Gronager, Josva Kleist, Daniel Mallmann, Mirek Ruda, Zdeněk Salvét, Andrea Sciabà, Jiří Sitera, Oxana Smirnova, Alain Roy, Frank Wuerthwein, Zdeněk Šustr, Oliver Keeble, John White