MW & experimentware survey

- Service map
  - http://gridmap.cern.ch/ccrc08/servicemap.html
  - Tick List
    - Lots of red
    - Improve software, procedures, deployment
  - Test Status
    - Lots of blue (no test results), white (no data source)
    - Add tests (or run the ones that exist already)
  - What can we learn from the common middleware?

- WLCG Service Coordination Meeting Dec. 20, 2005 (!)
  - http://indico.cern.ch/conferenceDisplay.py?confId=a056628
  - "Availability Issues and Middleware Components"

“The song remains the same”

- How does the current middleware address service availability issues?
  - Or does not…
- Service availability from user perspective
  - Robustness
    - Maturity of single components
  - Redundancy
    - Avoid single points of failure
    - Avoid bottlenecks
  - Fail-over
    - Automatic
    - Or manual…
- Issues with basic components
  - Node types
  - Standard services
- Issues with high-level services

User Interface

- Runs no service daemons, but is a service itself, as grid entry point
- Its proper working may depend on peripheral services being up
  - CERN lxplus UI depends on AFS
- In case of problems the user often can switch to another UI

BDII (Berkeley DB Information Index)

- Without BDII jobs etc. only can refer to "hard-coded" gLite services
  - Can use "-r" option in job submission
    - No requirements, no matchmaking
  - Can set LFC_HOST explicitly
- LCG_GFAL_INFOSYS now is an ordered list of BDII endpoints
  - GFAL automatically fails over to the next BDII
- A WMS/RB is statically configured to use a certain BDII
  - Flexibility might reduce performance
- Information system not safe against pollution
  - Any site can define central LFC for any VO…

BDII (cont.)

- Deployment recommendations
  - Put the BDII hosts behind a (load-balanced) rotating alias
    - lcg-bdii.cern.ch is an alias for bdii101, bdii102, ...
  - Do not mix BDII with node types that can experience high loads
    - BDII response time displays a non-linear dependency on the load
- BDII improvements
  - Indices
  - Flexibility with GIP
- Further GFAL improvements
  - Query timeout has been increased to 60 seconds
  - Queries have been made smarter
- List of sites (certified/production/monitored) updated hourly from GOC DB
  - Depends on connectivity of CERN and RAL

WMS and LB

- WMS deployment recommendations
  - LB should be on different machine
  - Sandbox area should be put on its own file system
  - MySQL DB ditto, other work directories ditto
- Client selects random WMS out of a list
  - If the chosen WMS does not accept the request, another is tried
  - WMS will refuse new jobs e.g. when load too high or file system too full
- 3.1 WMS can sit behind a load-balanced alias, but…
  - Client needs smarter retry algorithm
- LB cannot yet sit behind a load-balanced alias
  - WMProxy code to be adapted

WMS and LB (cont.)

- User cannot indicate which BDII(s) the WMS should consult
  - Flexibility might reduce performance
- WMS should support multiple PX servers
  - MyProxy RFE
- For a user-defined configuration the user must ensure the chosen WMS is trusted by the chosen PX server
- If the WMS is rebooted, jobs in steady state will not be affected
  - Jobs in transit may be lost
- If the WMS is unreachable, jobs that are finishing may be lost
  - Job wrapper script will try for a while to deliver the output sandbox
    - By default up to 5h 15min
  - Not important for pilot jobs

WMS and LB (cont.)

- WMS does not yet support CEs behind load-balanced aliases
  - Considered for CREAM, but not the LCG-CE
- Proxy renewal only tries VOMS server found in original proxy
- Proxy renewal daemon may hang when VOMS server has a problem
  - To be fixed in VOMS client code
  - Workaround available: periodic restart
- Purger is unreliable
  - Workaround available
- Output sandbox size cannot be limited
  - Fixed in 3.1
- Collection nodes may end up with empty input sandbox
  - Fix scheduled for first 3.1 update
- Workload Manager request input file can grow to very large size
  - New code uses a directory instead, to be certified

WMS and LB (cont.)

- SAM test to be reenabled
  - Disabled because of:
    - Reference CE overload (fixed)
    - Globus bug – bad WMS would cause process pile-up on reference CEs (fixed)
- Test should also be run by experiments

LCG-CE

- The weakest link in the job submission chain
  - Relies on the grid_monitor to avoid high loads, but each job still needs a few processes at submission and cleanup
- Load spikes when:
  - Multiple users submit jobs destined to the same CE
    - Not an issue with multi-user pilot jobs
  - Many jobs finish at the same time
    - E.g. all exiting when some external service for the VO went down
  - Many jobs are canceled at the same time
- Patch 1752 (on PPS) provides significant performance improvements
  - Real scalability will come with CREAM (patch just entered certification)
- Deployment recommendations
  - Put site BDII on different node
  - Busy VOs should have dedicated CEs → fewer users per CE

LCG-CE (cont.)

- If CE is rebooted, jobs in steady state will not be affected
  - Jobs in transit may be lost
- CE cannot sit behind a load-balanced alias
  - WMS/RB code (Condor-G) would have to be fixed
  - Should become possible with CREAM
  - Site can advertize multiple equivalent CEs

Worker Node

- Job may fail if the WN does not provide enough disk space or memory
  - On a multi-core node one job may hinder another
    - By filling a file system
    - By causing excessive paging
      - Other job could run out of wall-clock time
    - By killing unrelated processes owned by the same user
    - By filling the open file or socket table
  - Can only be avoided by using a virtual machine per job
    - No known examples (?)
    - Virtual machines have their own deployment issues
- On failure the WMS may resubmit the job, to another site if possible
  - If the job requirements allow it
  - Deep (full) resubmission usually disabled by client for various reasons
  - Shallow resubmission should help
    - Job already failed before user payload started
  - Not relevant for pilot jobs

Worker Node (cont.)

- Environment sanity tested by SAM
  - Basic tests run by OPS
    - Test suite evolves
  - VOs to implement their own tests
    - E.g. check desired access to local SE

MyProxy Server (PX)

- PX software is stable
- Vital for long jobs using short proxies
  - WMS/RB jobs often submitted with proxies valid for a few days
    - Avoids problems with proxy renewal, at a small decrease of security
    - In the future services will refuse long proxies
  - FTS legacy mode obtains user proxy from PX
  - Downtime can cause many jobs to fail
- Jobs currently can only have a single PX server
  - Allowing a list → MyProxy RFE (no news since Oct.)
- User must ensure the WMS/FTS/VOBOX used is trusted by the chosen PX
  - Many popular WMS/FTS/VOBOX have been added to popular PX instances…
- Deployment recommendation
  - Linux HA (used @ CERN)
  - DNS load balancing being studied

FTS, SE

- See other sessions during this workshop!

LCG File Catalog

- LFC software and deployment documentation essentially stable
- Unavailability can cause many jobs to fail
  - In particular if the LFC is central
- SAM test for local (non-central) LFC instances to be added
  - LFC @ T1: ATLAS (rw), LHCb (ro)
- LFC_HOST ought to be an ordered list too

VOMS

- VOMS server should have read-only replicas off-site
  - ATLAS have replica at BNL
    - Except for the Roles ?!
  - CERN setup is robust
    - https://twiki.cern.ch/twiki/bin/view/LCG/VomsNodes
    - Recent problems due to middleware and sensor issues

- SAM test to be improved
  - Also run by experiments

VO Box

- VOBOX common software is essentially stable
  - gsiopenssh facility, proxy renewal service
- VO-specific software issues and requirements only known to VO !
  - Requirements may not be implementable by all sites
    - To be negotiated between VO and site
  - Downtimes could cause significant amounts of job failures etc.
  - Certain services might be replicated on another instance (on-site)
- Define VO-specific SAM tests

Monitoring

- See other sessions during this workshop!

SAM

- Critical for Freedom of Choice of Resources
  - Allows VOs to avoid sites in bad shape
  - When SAM service is down, the selection of sites is not updated
    - No automatic exclusion of sites that have gone bad
    - No automatic inclusion of sites that have recovered
- Critical for CIC-on-duty and site admins to discover and fix problems
- Only a single instance
  - Depends on connectivity of CERN
  - Hot spare nodes standby for front-end
    - DB uses RAC

GStat

- Monitors availability and sanity of site GIISes (site BDIIs) for SAM
- Vital tool for CIC-on-duty and site admins to discover and fix problems
- Supplies SAM with list of sites to be monitored
  - Obtained from GOC DB, cached
- Two instances
  - Primary @ ASGC
  - Secondary @ CNAF

Meta Middleware

- If my jobs fail, who will notice and do something about it ?

- If the failures are due to generic problems at a site:
  - CIC-on-duty might have things fixed unprompted by users
  - Two instances of CIC Operations Portal
    - Primary @ IN2P3
    - Secondary @ CNAF

- Generally a ticket should be opened with GGUS
  - Two instances
    - Primary @ FZK
    - Secondary near FZK (?)

Conclusions

- Middleware availability issues on various levels
  - Software readiness
  - Deployment procedures
  - Monitoring, by site and by VO
- Some fixes "easy" for significant gain
  - Better deployment documentation
  - More tests
- Others still require significant effort
  - Development, integration, certification
- VO applications will have to keep dealing with middleware failures
  - The grid is not a local batch/storage system