Availability issues & MW components

- 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
  - Caveat: globus-job-run does start a temporary server
    - Only supported for debugging and tests, but still...
    - Needs inbound connectivity in GLOBUS_TCP_PORT_RANGE
- Its proper working may depend on peripheral services being up
  - Even if they are not used by the middleware
    - AFS, NFS, DNS (!)
- In case of problems the user often can switch to another UI
  - At the same site (cf. lxplus.cern.ch)
  - Many users also have accounts at CERN
  - Most users depend on centrally run UI
    - Tar ball distribution has made it easier to install UI on PC
      - CAs, CRLs still not completely in user land

BDII (Berkeley DB Information Index)

- Without an II jobs etc. only can refer to hard-coded services
  - Can use "-r" option in job submission
    - No requirements, no matchmaking
  - Can set LFC_HOST explicitly
- If a particular BDII has problems:
  - Can set LCG_GFAL_INFOSYS to point to another BDII
  - An RB is statically configured to use a certain BDII
    - Cannot yet be controlled by the user
      - Misfeature to be addressed
- Services and clients should allow for a list of equivalent BDIIs
  - A random element would be picked at any time the system can change efficiently
    - E.g. per job or group of jobs, or periodically
    - Automatic fail-over
    - Local BDII can be tried first

BDII (cont.)

- The best a site can do at this time:
  - Have the BDII hosts sit behind a (load-balanced) rotating alias
    - lcg-bdii.cern.ch is an alias for bdii001, bdii002, …
- BDII software is stable, but the service does not mix well with other contributions to the load, e.g. gridftp transfers
  - Under high load the BDII processes often get less of the CPU than expected
  - Could be suffering from current implementation
    - 3 processes involved per connection
  - lcg_utils have a 15s timeout on any LDAP search
    - Could be increased
- List of sites (certified/production/monitored) updated hourly from GOC DB
  - Depends on connectivity of RAL

Resource Broker (LCG-2)

- RB software is stable, but not robust against full file systems
  - If /var/edgwl gets full, all jobs managed by the RB may be lost
  - Sandbox area should be put on different file system
  - Production RB needs >~ 200 GB for the sandbox area, >~ 10 GB for the rest of /var/edgwl, >~ 30 GB for /var/lib/mysql
  - Sandbox cleanup script is available, cron job will be part of LCG-2_7_0
  - DB cleanup still does not work
- RB cannot sit behind a load-balanced alias yet
  - Client code to be fixed, not difficult
- Client configuration allows RB to be randomly selected out of a list
  - If the chosen RB does not respond, another is tried
- User configuration of the client overrides system configuration
  - Allows the user to react to failure of the system RB

Resource Broker (cont.)

- User cannot indicate which BDII(s) the RB should consult
  - Code changes should not be difficult
- For a user-defined configuration the user must ensure the chosen RB is trusted by the chosen PX server
- If the RB is rebooted, jobs in steady state will not be affected
  - Jobs in transit may be lost
- If the RB is unreachable, jobs that are finishing may be lost
  - Job wrapper script will try up to 5 hours to deliver the output sandbox

Computing Element (LCG-2)

- 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
  - 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

- Also runs site BDII (GIIS) and GRIS, both may suffer from high load
  - Site can disappear from top-level BDII
  - CE may be advertized with default values for number of jobs etc.
  - Site BDII can be easily moved to another node, GRIS not so easily

Computing Element (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
  - RB code (Condor-G) would have to be fixed
  - Site can advertize multiple equivalent CEs, also for fail-over

Worker Node

- Job may fail if the WN does not provide enough disk space or memory
  - On a dual-CPU 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 fixed by using virtual machines
    - First usage expected in a few months, currently being tested

- On failure the RB may resubmit the job, to another site if possible
  - If the job requirements allow it
  - Most experiment job/data management systems preclude resubmission
    - Current middleware cannot guarantee that a job runs at most once
  - Shallow resubmission would help, code mostly available

MyProxy Server (PX)

- PX software is stable

- Vital for long jobs using short proxies
  - RB jobs often submitted with proxies valid for a few days
    - Avoids risk of proxy renewal failure, at a small decrease of security
    - In the future services will refuse long proxies
  - FTS currently obtains user proxy from PX (to be changed)
  - Downtime can cause many jobs to fail

- Jobs currently can only have a single PX server
  - Allowing a list should not be hard

- PX can have a master node for writes, with slave nodes for reads
  - All on the same site, but allowing for load-balancing and fail-over

- User must ensure the RB/FTS used is trusted by the chosen PX

File Transfer Service

- FTS software is fairly new, but should be stable by SC4
  - FTS is the data management workhorse of the SC !
  - Significant enhancements still to be implemented
    - Multi-hop routing
    - Data placement
- By its nature there can be only a single FTS per channel
  - If the channel goes down, an automatic detour may not be desired
    - Might overload other channels
- Currently depends critically on PX servers specified in transfer jobs

LCG File Catalog

- LFC software essentially stable

- An LFC instance may have read-only replicas
  - At the same site already allowing for load-balancing and fail-over
  - At other sites possible for Oracle back-ends
    - To be tested
    - MySQL?
  - Client code does not yet allow for a list of equivalent LFCs
    - Load-balanced rotating alias should work

- Downtime can cause many jobs to fail
  - In particular if the LFC is central

Storage Element

- SE software is in flux
  - CASTOR/dCache/DPM/… have varying degrees of stability in various areas
  - Data losses are very rare, but some services may be unavailable at times

- Some experiments fail over to other SEs on writes

- Fail-over on reads only possible for replicated data sets
  - Might cause chaotic data transfers, bad usage of network and CPU

MON Box / R-GMA

- MON software (R-GMA) is in flux
  - Downtimes are frequent, but should be much better by SC4
- Clients can handle only a single instance
  - Fail-over list ?
- Server and client critically depend on single registry
  - Depends on connectivity of RAL
  - To be fixed

- Currently critical for Site Functional Tests ?
  - Lists of sites communicated by GIIS Monitor via R-GMA
  - To be decoupled
- Critical for GridView
  - Critical for SC4 monitoring !
  - To be decoupled ?
- Used by APEL to transport accounting records
  - Downtimes will only create a backlog

VOMS

- VOMS software is partly stable, partly in flux
  - VOMS core fairly stable
  - VOMS admin has issues (possibly due to Tomcat etc.)
  - VOMS servers not production-ready yet, but should be by SC4
- VOMS functionality critical for SC4
  - New analysis and production groups and roles to be exercised
- VOMS server should have read-only replicas
  - Not clear how ready the code is
  - Not clear if the replica can be off-site
  - There must at least be a hot spare !

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)

New node types

- Workload Management System + Logging & Bookkeeping Server
  - To replace Resource Broker
    - Similar issues
    - Many enhancements, e.g. bulk submission
    - Should be stable by SC4
- g-PBox
  - To become critical for implementing VO and site policies for job management
- DataGrid Accounting Server
  - To become vital for user-level accounting
    - Local agent can handle retries
    - Complementary to APEL
- gLite Computing Element
  - To become a robust replacement for the fragile LCG-2 CE
    - Should be stable by SC4
- gLite I/O Server
  - Front-end for “dumb” SE
    - Potential bottleneck
- FiReMan catalog
  - Alternative to LFC
    - Similar issues

Site Functional Tests

- Critical for Freedom of Choice of Resources
  - Allows VOs to avoid sites in bad shape
  - When SFT 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 for the time being
  - Depends on connectivity of CERN
  - Plans for multiple instances
  - Hot spare node ready

Gstat / GIIS Monitor

- Monitors availability and sanity of site GIISes
- Vital tool for CIC-on-duty and site admins to discover and fix problems
- Supplies SFT service with list of sites to be monitored
  - Obtained from GOC DB, cached
- Only a single instance for now
  - Depends on connectivity of Taipei
  - Plans for replication

GridICE

- Monitors site occupancy per VO
  - Numbers of active/waiting jobs
  - Used/available storage
- Tool for site admins and VOs to see if and how resources are being used
  - A site that passes the SFTs may still be left idle:
    - Black-listed by VO
    - Does not meet some requirement (too little memory, wrong OS version, …)
    - Middleware bug
- Only a single instance for now
  - Depends on connectivity of CNAF

Archiver

- R-GMA client recording monitoring history and serving it to other clients
- Consulted by various monitoring facilities
- If it is down, only a small amount of history is available
- Currently only a single instance per table

GridView

- Vital tool for monitoring SC activity !
  - FTS traffic statistics
  - Job statistics foreseen
- Depends critically on R-GMA, MON boxes
  - Could be decoupled ?
- Currently only one instance

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
    - Depend on GIIS Monitor, SFTs, …
    - Depend on CIC-on-duty connectivity
    - Depend on IN2P3 dashboard ?
- Generally a ticket should be opened with GGUS
  - Only a single GGUS instance ?
    - Depend on FZK connectivity
  - Depend on GGUS “middleware”
  - Depend on ROC/CIC “middleware”
    - E.g. secondary ticketing system

Conclusions

- Middleware availability issues on various levels
- Some fixes “easy” for significant gain
  - Can still be non-trivial amount of work !
  - BDII list in LCG_GFAL_INFOSYS
    - Also to be used by RB
  - List of PX servers for RB, FTS
  - Shallow resubmission
- Others still require significant effort
  - Development, integration, certification
    - New WMS, CE, g-PBox, DGAS, ...
  - Some not yet vital at start of SC4 ?
- VO applications should deal with middleware failures
  - The grid is not a local batch/storage system
- Users can often override default configurations