Services for Tracking and Archival of Grid Job Information


CESNET, Czech Republic
Logging and Bookkeeping

- functionality overview, main features
- recent development
- deployment

Job Provenance

- motivation
- interaction with gLite WMS and L&B
- architecture and usage overview
Purpose

- track **Grid jobs** during their life
- capture passing job control between Grid components
- provide user with high-level view on **job state**
- short-term post-mortem analysis
Purpose

• track **Grid jobs** during their life
• capture passing job control between Grid components
• provide user with high-level view on **job state**
• short-term post-mortem analysis

Main features

• important points in job life gathered as **L&B events**
  – transfer of job between grid components
  – finding suitable computing element
  – starting/terminating execution
• events delivered to L&B server **reliably** but in **non-blocking** way
• job state computed by fault-tolerant state machine
• user can query job state or register for receiving notifications
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

CE2

RB

LB

Register

Waiting

Ready

Match

Transfer

Running

Run

Waiting

Running

Run

Transfer

Done

Done

Job control

LB events

LB job state
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

CE2

LB events

LB job state

Cracow Grid Workshop, November 20–23, 2005
L&B overview

Enabling Grids for E-sciencE

Job control
LB events
LB job state

CE1
CE2
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

CE2

Register

RB

Waiting

LB

LB events

LB job state

Cracow Grid Workshop, November 20–23, 2005
L&B overview

Enabling Grids for E-sciencE

Job control
- CE1
- Register
- Match
- Waiting
- Ready

LB events
- CE2
- Match
- Waiting
- Ready

LB job state
L&B overview

Enabling Grids for E-sciencE

Job control
CE1
LB job state
CE2
RB
Match
Waiting
Ready
Transfer
Running
Run
Done

Cracow Grid Workshop, November 20–23, 2005 4
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

RB

Register

Match

LB

Waiting

Ready

CE2

Transfer

Done

Done

Running

Running

Job control

LB events

LB job state
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

LB job state

CE2

Register

Match

Transfer

Waiting

Ready

Running

Waiting

Cracow Grid Workshop, November 20–23, 2005
L&B overview

Enabling Grids for E-sciencE

Job control
- CE1
- CE2
- LB

RB
- Register
- Match

LB
- Waiting
- Ready
- Running
- Waiting

Transfer

Cracow Grid Workshop, November 20–23, 2005
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

Register

Match

RB

Transfer

CE2

Running

LB

Waiting

Ready

Running

LB job state

Done

Done

Cracow Grid Workshop, November 20–23, 2005
L&B overview

Enabling Grids for E-sciencE

Job control

CE1

LB job state

RB

Register

CE2

Match

Waiting

Ready

Running

Transfer

Transfer

Done

Done

Running

Job control

LB events

LB job state
L&B Proxy

- gLite Workload Manager processing depends on job state
  - consistency checks
  - original job description retrieval on job resubmission
- non-blocking, asynchronous L&B event delivery is a problem
  - query following a logged event may not see it
L&B Proxy

- gLite Workload Manager processing depends on job state
  - consistency checks
  - original job description retrieval on job resubmission
- non-blocking, asynchronous L&B event delivery is a problem
  - query following a logged event may not see it
- addressed by L&BProxy
  - lightweight L&B server, runs on WM node
  - only events coming from this node gathered
  - partial, local view on job state
  - all communication is local, synchronous
  - no SSL authentication and encryption – better performance
  - all events forwarded to full L&B server
- WMS daemons being converted to use L&B Proxy
Job statistics

- EGEE JRA2 defined schema of job record
- most of the information available in L&B
- currently dug from MySQL database of L&B server
  - inaccurate
  - too heavy-weight
- use **L&B dumps**
  - files generated on purging expired data from L&B servers
- uploaded to statistics server
- processed (re-compute terminal job state) to give job record
- compatible with older (EDG, LCG, . . .) L&B servers
- L&B code is ready and tested, deployment pending
Computing Element reputability ranking

- “black hole” problem
  - CE accepts jobs but they fail there at high rate
  - not visible in Grid information services (the CE is always free)
Computing Element reputability ranking

• “black hole” problem
  – CE accepts jobs but they fail there at high rate
  – not visible in Grid information services (the CE is always free)

• auxiliary on-line statistics computed by L&B server
  – rate of incoming jobs
  – rate of job failure
  – duration of job execution
  – . . .

• made available as ClassAd function
  – can be included in job description
  – affects overall CE ranking

• implementation optimised for high query rate (no disk access)
• currently being tested with WMS
**EGEE**

- approx. 50 production installations
- over 20,000 jobs per day in average
- over 60 GB of data since January 2005

**Other projects using EDG or EGEE software**

- LCG
- CrossGrid
- ...
Recent requirements

- Condor jobs
- tracking other entities
  - data transfer jobs
  - resource reservations
Recent requirements

- Condor jobs
- tracking other entities
  - data transfer jobs
  - resource reservations

Generalised L&B design

- distinguish between core L&B “skeleton” . . .
  - principal data entities are abstract jobs and events
  - events of a single job are gathered at one server
  - server computes job state
  - users pose queries or receive notifications
Recent requirements

- Condor jobs
- tracking other entities
  - data transfer jobs
  - resource reservations

Generalised L&B design

- distinguish between core L&B “skeleton” . . .
  - principal data entities are abstract jobs and events
  - events of a single job are gathered at one server
  - server computes job state
  - users pose queries or receive notifications
- . . . and application specific “flesh”
  - concrete event and job state datatypes
  - plugins for L&B components, namely job state computation
Motivation

• preparing job submission requires a lot of work
• the work is not completely reflected in job results
• **preserve information on Grid jobs**
  – what were the executed jobs
  – job execution environment (installed software etc.)
  – track of execution (e.g. number of failures and resubmission)
• allow data-mining in this information and assisted job re-running
  – “What were jobs of this VO, run on input data X, using (faulty) software Y?”
Gathered data

- scalability issues
  - strict limits on reasonable JP record size
  - record volatile data only
Gathered data

- scalability issues
  - strict limits on reasonable JP record size
  - record volatile data only
- job inputs
  - job description (JDL) as submitted to RB
  - miscellaneous input files (input sandbox)
  - do not copy input files from remote storage elements
Job Provenance

Gathered data

- scalability issues
  - strict limits on reasonable JP record size
  - record volatile data only

- job inputs
  - job description (JDL) as submitted to RB
  - miscellaneous input files (input sandbox)
  - do not copy input files from remote storage elements

- job execution track
  - L&B data (when and where was the job planned and executed etc.)
  - “measurements” on CE (installed software, environment)
  - accounting data (DGAS)
Gathered data

- scalability issues
  - strict limits on reasonable JP record size
  - record volatile data only

- job inputs
  - job description (JDL) as submitted to RB
  - miscellaneous input files (input sandbox)
  - do not copy input files from remote storage elements

- job execution track
  - L&B data (when and where was the job planned and executed etc.)
  - “measurements” on CE (installed software, environment)
  - accounting data (DGAS)

- user annotations (at run-time or afterwards)
JP interaction with WMS and L&B

Enabling Grids for E-sciencE

RB

LB

JP

CE

job control
record value
upload file
retrieve data
JP interaction with WMS and L&B

- Job control
- Record value
- Upload file
- Retrieve data
JP interaction with WMS and L&B

Enabling Grids for E-sciencE

- job control
- record value
- upload file
- retrieve data
JP interaction with WMS and L&B

Enabling Grids for E-sciencE

job control
record value
upload file
retrieve data

JP

RB

LB

CE

tag

register

register

tag
JP interaction with WMS and L&B

Enabling Grids for E-sciencE

- job control
- record value
- upload file
- retrieve data

- register
- purge
- tag
Primary data

- job is the principal entity
- minimal set of core attributes: jobid, owner, registration time
- short data items: **tags** – “key = value” pairs
- bulk data: uploaded **files**
Primary data

- job is the principal entity
- minimal set of core attributes: jobid, owner, registration time
- short data items: tags – “key = value” pairs
- bulk data: uploaded files

JP job attributes

- generic unified view on any job data
- “namespace:key = value” format
- can be multi-valued
- namespaces may have defined schema
- used for both internal handling and user queries
- JP tags mapped directly
- bulk files processed by file-type specific plugins
Primary storage

- gather data from their sources and store them “forever”
- process bulk files to extract JP attributes – on demand
- user queries
  - retrieve job attributes, download files
  - keyed by jobid only for performance reasons
- serve Index server queries
- WS control interface, gsiftp for file transfer
Index server

- created and configured semi-dynamically for particular purpose
  - list of Primary storages to register with
  - conditions on jobs to retrieve (specified via attributes)
    - e.g. jobs of VO X, submitted in 2005
  - list of job attributes to gather

- contain only fraction of data from Primary storage(s)
Index server

- created and configured semi-dynamically for particular purpose
  - list of Primary storages to register with
  - conditions on jobs to retrieve (specified via attributes)
    - e.g. jobs of VO X, submitted in 2005
  - list of job attributes to gather
- contain only fraction of data from Primary storage(s)
- two mode of communication with Primary storage
  (may be combined)
  - batch feed – retrieve all jobs matching the query
  - incremental feed – register for receiving updates on matching jobs
Index server

- created and configured semi-dynamically for particular purpose
  - list of Primary storages to register with
  - conditions on jobs to retrieve (specified via attributes)
    - e.g. jobs of VO X, submitted in 2005
  - list of job attributes to gather

- contain only fraction of data from Primary storage(s)

- two mode of communication with Primary storage (may be combined)
  - batch feed – retrieve all jobs matching the query
  - incremental feed – register for receiving updates on matching jobs

- serve user queries
  - may be quite complex (two-level, and-or structure)
  - unlike primary storage, jobid is not required
  - may refer only to IS configured attributes
  - return list of jobid’s and PS contacts
Current status

- Implementation done, included in gLite 1.5 RC
  - volatile PS → IS communication
  - limited flexibility of IS configuration
- Supported file types: L&B and input sandboxes
- Deployed at development testbed, receiving first real jobs
Current status

- implementation done, included in gLite 1.5 RC
  - volatile PS → IS communication
  - limited flexibility of IS configuration
- supported file types: L&B and input sandboxes
- deployed at development testbed, receiving first real jobs

Immediate plans

- deployment in larger scale
- user-side CLI and integration in gLite WMS GUI to support re-running jobs
- more complex authorisation
Status and plans

Current status

- implementation done, included in gLite 1.5 RC
  - volatile PS → IS communication
  - limited flexibility of IS configuration
- supported file types: L&B and input sandboxes
- deployed at development testbed, receiving first real jobs

Immediate plans

- deployment in larger scale
- user-side CLI and integration in gLite WMS GUI to support re-running jobs
- more complex authorisation

Longer-term plans

- integration with Grid accounting (DGAS)
- support for non-gLite-WMS jobs (CREAM CE, Condor)
- interface to gLite Storage Element
Conclusions

Job-centric monitoring approach

- users are interested in their jobs
- data from different sources form the overall job state

Logging and Bookkeeping

- track job during its life
- developed in EDG, continued in EGEE
- production quality, widely deployed

Job Provenance

- archive job data for long time
- allow data-mining, help with re-running jobs
- prototype available, wider deployment expected