WN Working Group Status
WLCG GDB.
May 14th 2008

Steve Traylen, CERN, steve.traylen@cern.ch

Overview

- **Previous Slides:**
  - January 2008 GDB
    - § http://indico.cern.ch/conferenceDisplay.py?confId=20225
- **Motivation:**
  - Efficient use of Worker Nodes.
- **Information System Deployment**
  - The ClusterPublisher
- **Publishing Software Tags:**
  - Review of software tags w.r.t. experiments.
  - Resolution of WN to GlueSubCluster.
  - Resolution of GlueSubCluster to the ClusterPublisher.
- **Deployment Steps:**
  - Ordered deployment steps to make it all possible.

Recap of Motivation

- **Currently deployed EGEE Grid assumes :**
  - WNs behind a CE node are identical.
  - Clearly not the case at all but a tiny number of tiny sites.
  - Current advice has been to advertise smallest nodes available.
- **This results in:**
  - Large WNs being wasted by small memory jobs.
  - Large memory nodes cannot even be advertised.
  - Walltime and CPU.
    - § Hard for users to work out how long they will get.
    - § Hard for batch managers to allocate jobs efficiently.

Deployment of InfoSys

- **EDG and EGEE always published a generic CE node of:**
  - 1 GlueCluster (GClust). Should map to a Batch System
  - 1 GlueSubCluster (GSubClust). Should map to a set of WNs.
  - >= 1 GlueCE. Should be and is a batch queue end point on CENode.
- **This breaks down and is wrong when:**
  - More than 1 CENode on the same batch system.
    - § Second CENode should only add GlueCEs, not GClust or GSubClust.
    - § The causes of all the problematic CPU counting we see in gstat, gridmap, EGEE monthly reports, ...
    - § The extra GSubClust and GClust added are duplicates of physical resources.
- **Steps needed to correct this are at a YAIM and site manager level.**

YAIM Node Types

- **The lcg-CE node type is to be broken to:**
  - lcg-CE
    - § Configures the LCG CE and publishes the GlueCEs
  - glite-CLUSTER the ClusterPublisher.
    - § Publishes the GCluster and GSubCluster.
    - § Configures the software-tag area. i.e where the tags go.
      - */opt/edg/var/info -> /opt/edg/var/info/<GlueSubCluster>*
- **Typical Deployment scenarios:**
  - Small site = lcg-CE + glite-BDII + glite-CLUSTER on one node.
  - Large site = lcg-CE one node. glite-BDII + glite-CLUSTER on one node.
  - Huge site = lcg-CE, glite-BDII and glite-CLUSTER own nodes.
    - § I expect no sites are large enough to be huge inc CERN.
- **Current YAIM Status:**
  - YAIM configuration has been done but is not yet in the release path.
    - § https://twiki.cern.ch/twiki/bin/view/EGEE/YAIMInfosys

Software Tags.

- **Software tags currently managed with a CE hostname.**
- **Software tags currently appear in GlueSubCluster.**
  - Assumption has been made that GSubClust is on CE node.
- **In the future tags must be added to the GSubCluster in the first place.**
  - We may have multiple GSubClusters running not on the CEnode but the SiteBDII node for instance.
- **In short.**
  - lcg-ManageVO-Tags --host lcgce01.example.org --add “FastTrack”
  - is to become.
  - lcg-ManageVO-Tags --subclust RAL-xeon-bigmem --add “FastTrack”
- **What is the impact of this:**
  - Software must be managed at SubCluster and not CEnode level.
  - Quick survey of the 4 LHC & other VOs
    - How do you publish and use software tags?

Use of Software Tags by VOs(1)

- **ATLAS**
  - Use their LJSFi (?) https://atlas-install.roma1.infn.it/atlas_install/
  - Allows site admin or whoever to request installs and deletions of ATLAS software by CEName.
    - § Some CEs (e.g Tier1s) are auto-subscribed to new software.
  - lcg-tags is used.
  - ATLAS (A. De Salvo) comments he would be happy to move to cluster based submission and tagging.
    - § Would make CEnodes publishing multiple arch's via multiple GSubClusters easier. Exactly what we want to do.

Use of Software Tags by VOs(2)

- **ALICE**
  - Do not use software tags in anyway.
- **LHCb**
  - Regular SAM jobs validate and install software as necessary.
    - § Submission is as SAM, CE node based.
  - SAM job maintains the software tags list with lcg-ManageVOTags.
    - § Adds tags on sucessful installation and validation.
    - § Removes tags on a subsequent validation failure.
  - LHCb do not use the tags for matching sites within the WMS.
    - § They are only populated for convenience.
  - LHCb will continue to submit to CE nodes as SAM does.
    - § To be able publish a tag they must be able resolve WN -> GSubCluster -> ClusterPublisher.

Use of Software Tags by VOs(3)

- **CMS**
  - Perl script generates JDL to submit separate:
    - § installation jobs.
    - § validation jobs.
      - *Again based on CE node name.*
  - Actual tags are added async' depending on installation and validation results.
    - § lcg-tags command is used.
  - Moving to cluster based submission and tagging would need some work in the above perl scripts but not much.
- **DTeam and Geant4**
  - Both use lcg-ManageVOTags

Summary of Software Tag Use.

- **There are two commands in use.**
  - lcg-tags and lcg-ManageVOTags
  - Both maintained by the same section - EIS.
    - One should be dropped, else both must be updated.
- **Moving to real SubCluster tagging is okay for VOs**
  - For the LHCb and DTeam use cases a little more work must be undertaken to allow them do so.

Resolving Where to Publish Tags.

- **To add a tag the ClusterPublisher node must be located.**
  - Tags are added via GridFTP interface on ClusterPublisher
  - This is to be handled internally by lcg-tags/lcg-ManageVOTags.
- **Two use cases:**
  1. ***Remote Tag Publishers***
     Publish a tag to a known existing GlueSubCluster.
     - *CMS, ATLAS*
     - *lcg-tag --subclust lcgce01-bigmem.example.org --add "ATHENA-1.27"*
  2. ***Local (WN) Tag Publishers***
     Publish a tag from a WN not knowing the GlueSubCluster.
     - *LHCb, DTeam*
     - *lcg-tag --add "ATHENA-1.27"*
- **Reduces to two problems.**
  1. *Resolve a WN -> GlueSubCluster*
  2. *Resolve a GlueSubCluster -> ClusterPublisher.*

1 - Resolve WN -> GlueSubCluster

- **Running job must make this resolution.**
- **lcg-tags must be supplied this information.**
- **Proposals:**
  - SiteAdmin(YAIM) maintains config file on every WN. $GLITE_LOCATION/etc/glite-wn-configuration.conf
    - § YAIM could maintain this via addition to the nodes.conf file.
    - § Other things , yet to be determined could be put in this file.
      - *In the past LHCb have asked for batch scaling values to go here.*
  - SiteAdmin(YAIM) maintain a simple script.
    - § $GLITE_LOCATION/bin/glite-wn-subcluster --uniqueid
      - *Would returns SubCluster uniqueID.*
    - § Large sites can optionally write a script to query batch system.
    - § Other SubCluster entities could be returned, e.g SI2000 values.
- **TCG(?) site-representatives have be polled for their opinions.**

GSubClust to ClusterPublisher

- **Resolution must be performed anywhere, e.g UI, WN, installation framework, ...**
- **Again lcg-tags will handle this internally.**
- **Information obtained from BDII InformationSystem.**
  - No schema changes, just more information all from the ClusterPub
- **GSubClusters mapping to ClustPublisher as GlueServiceData objects.**

dn:
GlueServiceDataKey=GlueSubClusterUniqueID,GlueServiceUniqueID=host.name_org.glite.RTEPublisher_12345,mds_vo_name=resource,o=grid
GlueServiceDataKey: GlueSubClusterUniqueID
GlueServiceDataValue: sub.cluster.name
GlueChunkKey: GlueServiceUniqueID=host.name_org.glite.RTEPublisher_12345

- **Publish the ClusterPublisher (RTEPublisher) as a Service.**

dn: GlueServiceUniqueID=host.name_org.glite.RTEPublisher_12345,mds_vo_name=resource,o=grid
GlueServiceUniqueID: host.name_org.glite.RTEPublisher_12345
GlueServiceName: MySite-RTEPublisher
GlueServiceType: org.glite.RTEPublisher
GlueServiceEndpoint: gsiftp://host.name:2811/opt/edg/var/info
GlueServiceAccessControlBaseRule: VO:atlas
GlueServiceAccessControlBaseRule: VO:cms
GlueForeignKey: GlueSiteUniqueID=MySite

Deploying This

- **There are many things to change.**
  - YAIM to support multiple GClusters and GSubClusters.
    - § YAIM must publish these.
    - § YAIM must create per SubCluster directories in the software tags area.
    - § YAIM must optionally create per WN *glite-wn-subcluster* scripts.
    - § YAIM must split lcg-CE to lcg-CE and glite-CLUSTER
    - § The ServiceData relations of GlueSubCluster and Tag Locations must be published.
  - Software Tag Information Providers.
    - § The GIP publisher of tags must expect per SubCluster tags.
  - lcg-tags/lcg-ManageVOTags
    - § Must support adding/deleting tags to SubClusters.
- **These can be achieved in two steps.**
  1. Deploy all software with no visible changes in anything.
  2. Once updated lcg-tags command is established sites can start adding more than one SubCluster.

Deploy Now in Any Order

- **Updated lcg-tags can be deployed.**
  - Can be done so that users can use --subcluster or --host.
  - Both will work on old lcg-CEclassic or glite-CLUSTER and lcg-CE combination.
- **Split the lcg-CE into glite-CLUSTER and lcg-CE.**
  - This will support multiple Clusters and SubClusters.
  - This will support framework for tags per SubCluster
  - This will publish glite-CLUSTER service and relations.
  - We advise sites not to create more than one SubCluster for now.
    - § Doing so would break the old lcg-tags command.
- **Add per WN glite-wn-subcluster scripts.**
- **Update tag information provider to support per GSubCluster publishing.**

Point to Allow Multi SubClusters

- **We anticipate that everything deployed != installed at all sites.**
- **Migration happens once all previous steps are available.**
- **Remote tag installers**
  - must migrate to use
    - lcg-tags --subclust <GlueSubClustUniqueId> -add “Athena-1.02”
  - with a fall back of
    - lcg-tags --host <GlueCEHostname> --add “Athena-1.02”
- **Local tag installers**
  - must migrate to use
    - lcg-tags --add “Athena-1.02”
  - with a fall back of
    - lcg-tags --host <GlueCEHostname> -add “Athena-1.02”
- **Once taggers are migrated sites can start splitting up their SubClusters**

Deployment Conclusions

- **While there are many changes.**
  - We can deploy everything without making any changes.
  - At a future date we can allow site admins to reconfigure.
    - § This can be done site by site, .. a few at first.
    - § We expect sites to update between “now” and “never”
    - § The break that will happen when multiple subclusters appear:
      - *Remote tag installers using old lcg-tags or “--host” will be stopped in their path at multi SubCluster sites.*
      - *Local tag installers using “--host” within a multi SubCluster site will be stopped in their path.*
- **So many changes there is room for mistakes.**
  - This can and should go through PPS(?) first.
- **Resulting Improvements.**
  - Non-overlapping GlueSubClusters
  - Jobs can be matched to finer GlueSubClusters and can be distributed better within the batch farm.

Next and Future Steps

- **Next**
  - Agree how WN level cluster query scripts should be done.
  - Update SoftwareTag GIP Publisher
  - Preferably drop one of lcg-tags or lcg-ManageVOTags
  - Provide exact detail of what the lcg-tags/MVOTags should do.
  - Many updates to YAIM, this is started, hard bit done.
- **Future is CREAM.**
  - Everything here is still needed for the CREAM CE anyway to use its full potential.
    - § Passing job arguments only permits us to use one GlueCluster instead of one per GlueSubCluster
  - This needs doing anyway.