WN Working Group

Steve Traylen

LCG GDB Meeting - January 9th 2008

Outline

- **Group Profile**
  - Mandate
  - Outputs
  - Members
- **WN resources to be considered.**
- **Current deployed situation**
  - GlueClusters, GlueSubClusters and GlueCEs.
- **Proposed future deployment.**
  - Introduction of Gatekeeper and Cluster Node types.
  - Problems with this and proposed solutions.
- **Summary of Next Steps.**

Mandate

- **Investigating:**
  - Consider hard limits, e.g memory, hard disk space, ...
  - Subsequent matchmaking of those resources.
  - The efficient choice of worker nodes within the EGEE grid.
    - e.g low memory jobs on low memory nodes.
  - Batch farms contain heterogeneous resources.
    - Never really represented before within the deployment.
- **Outputs:**
  - Provide deployment process with details to advertise heterogeneous WNs.
  - Provide users with examples for matching a particular set WNs.
  - Provide middleware development with shortfalls which make efficient utilization of WN resources difficult or impossible.

Group Details

- **Home Page**
  - https://twiki.cern.ch/twiki/bin/view/EGEE/WNWorkingGroup
- **Membership**
  - Chair - Steve Traylen.
  - VO - Roberto Santinelli & Simone Campana.
  - Site - Alessandra Forti & Ulrich Schwickerath.
  - Middleware - Francesco Prelz
  - Glue Schema - Stephen Burke
- **Contact: egee-wn-resources-wg@cern.ch**
- **Discussion on mailing list and at last EGEE conference. Just one formal meeting.**
  - https://twiki.cern.ch/twiki/bin/view/EGEE/WNWorkingGroupMinutes20070920

WN Resources and Batch

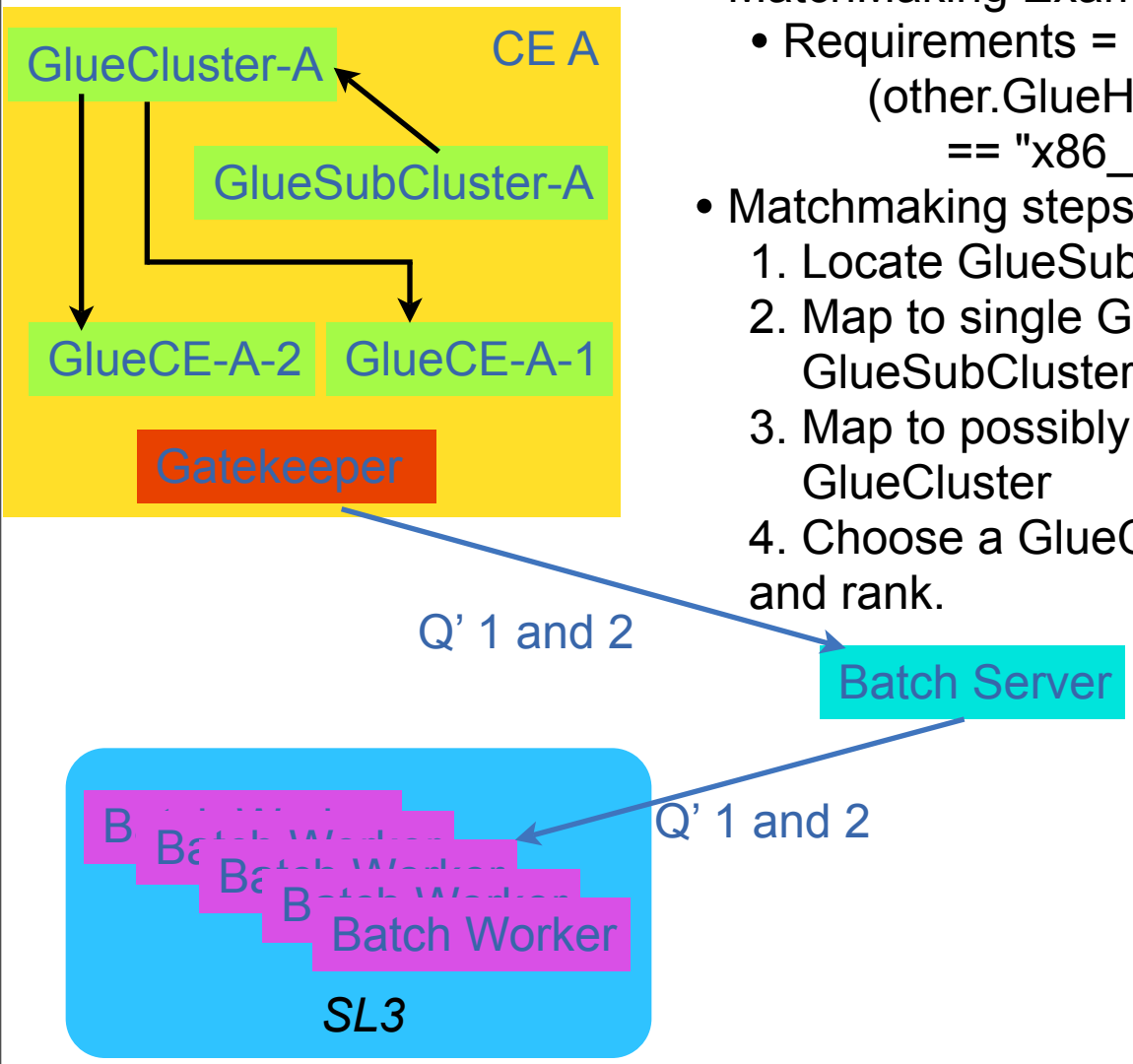
- **Potential WN Resources needed by a Job**
  - Operating System
  - 32/64 bit environment of the OS.
  - Local disk space for job.
  - Memory Constraints for Job.
- **Torque and LSF can support requests for all of these.**
  - SGE and condor unknown but to a large extent their problem.
  - We don't expect the middleware or grid to cope with short falls in the batch system.
    - If the batch system can't support heterogeneous clusters then why should the grid interfaces.
- **The aim in GLUE was that these WN varieties within a single GlueCluster should be represented by multiple GlueSubClusters.**
  - This has never been deployed.

WN Resources and Glue 1.3

- **The GlueSubCluster has attributes:**
  - GlueHostOperatingSystem{Name||Release}
    - Well defined and well deployed.
  - GlueHostArchitecturePlatformType
    - Well defined , deployment done but next release will enforce it.
  - GlueHostMainMemoryRAMSize
  - GlueHostMainMemoryVirtualSize
    - Not well defined, but deployed.
    - Is it per box, per core or per job slot
      - *- per core has been the rough recommendation till now.*
- **A GlueSubCluster contains nothing about WN space.**
  - However JSDL 1.0 contains DiskSpace elements and this is a basis for Glue 2.0 computing elements.
  - So for disk space just watch GLUE 2.0 development for now.

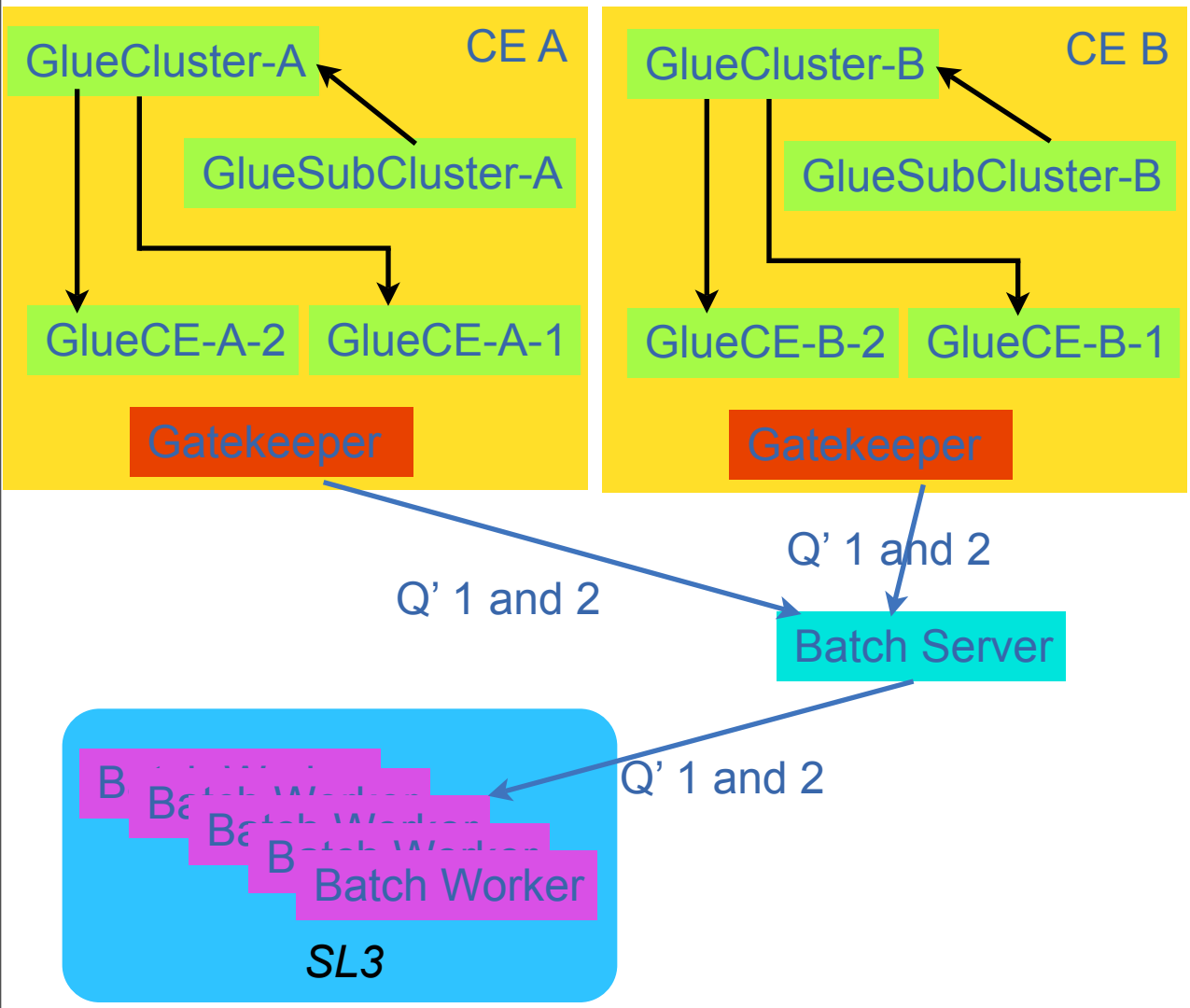
GlueCE/Cluster/SubCluster Today

GlueCE/Cluster/SubCluster Today

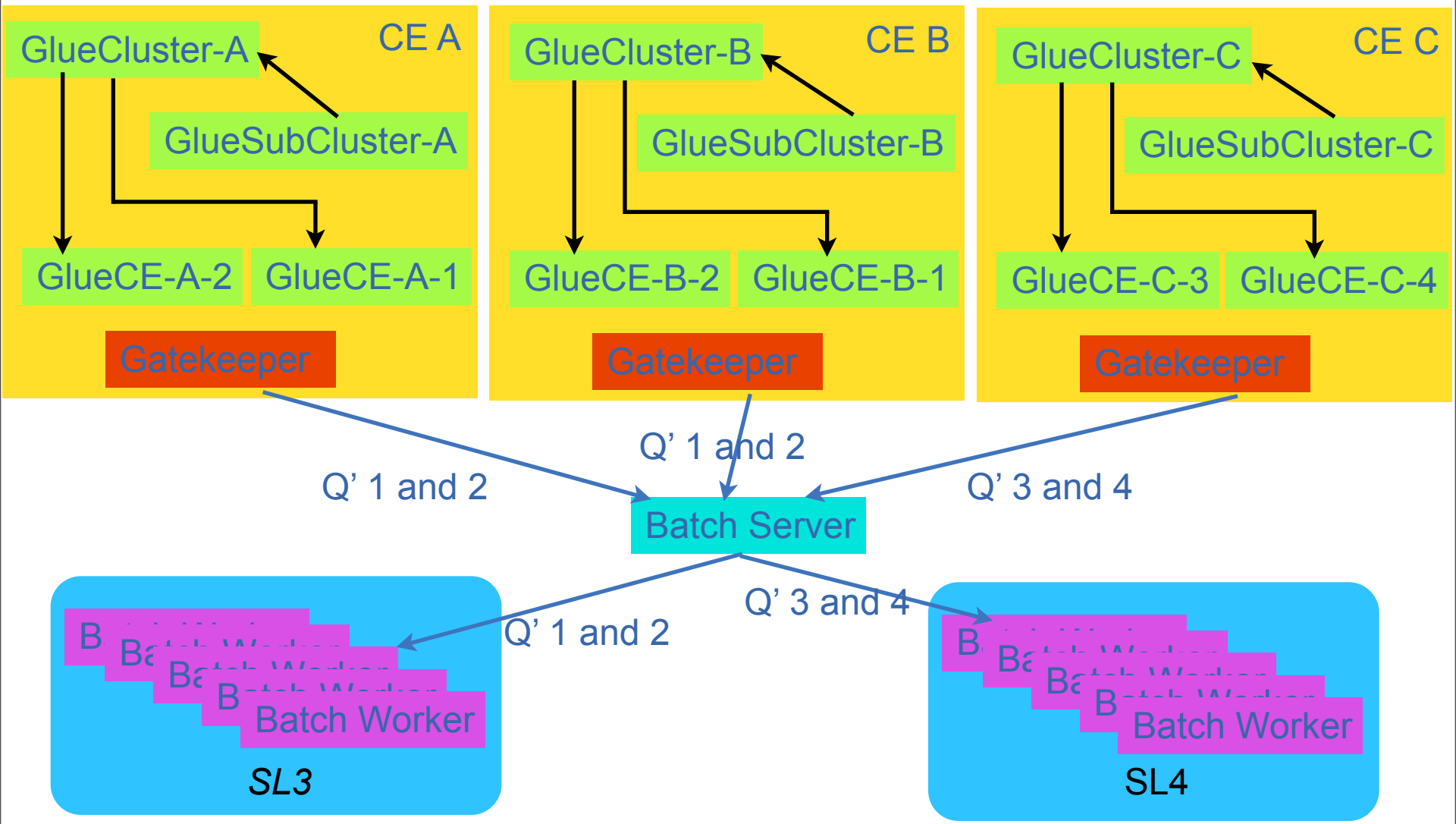


- MatchMaking Example with JDL
  - Requirements = (other.GlueHostArchitecturePlatformType == "x86_64") ;
- Matchmaking steps are
  1. Locate GlueSubClusters that match PlatformType.
  2. Map to single GlueCluster for each GlueSubCluster.
  3. Map to possibly multiple GlueCEs for each GlueCluster
  4. Choose a GlueCE based on CE properties (length) and rank.

GlueCE/Cluster/SubCluster Today



GlueCE/Cluster/SubCluster Today



Issues with Current Deployment

- **Adding new hardware types or OSes requires sites:**
  - add complete new CE nodes.
  - unique queues to distinguish resources.
- **SubClusters can currently contain intersecting WN sets.**
  - This has always given gstat, the quarterly reports and now gridmap an impossible task when CPU counting.
- **More information is published than is needed.**
  - e.g CERN publishes 21 GlueClusters and 21 GlueSubClusters
  - Only one GlueCluster and 2 GlueSubClusters representing SL4-i686 and SL4-x86_64 is needed.

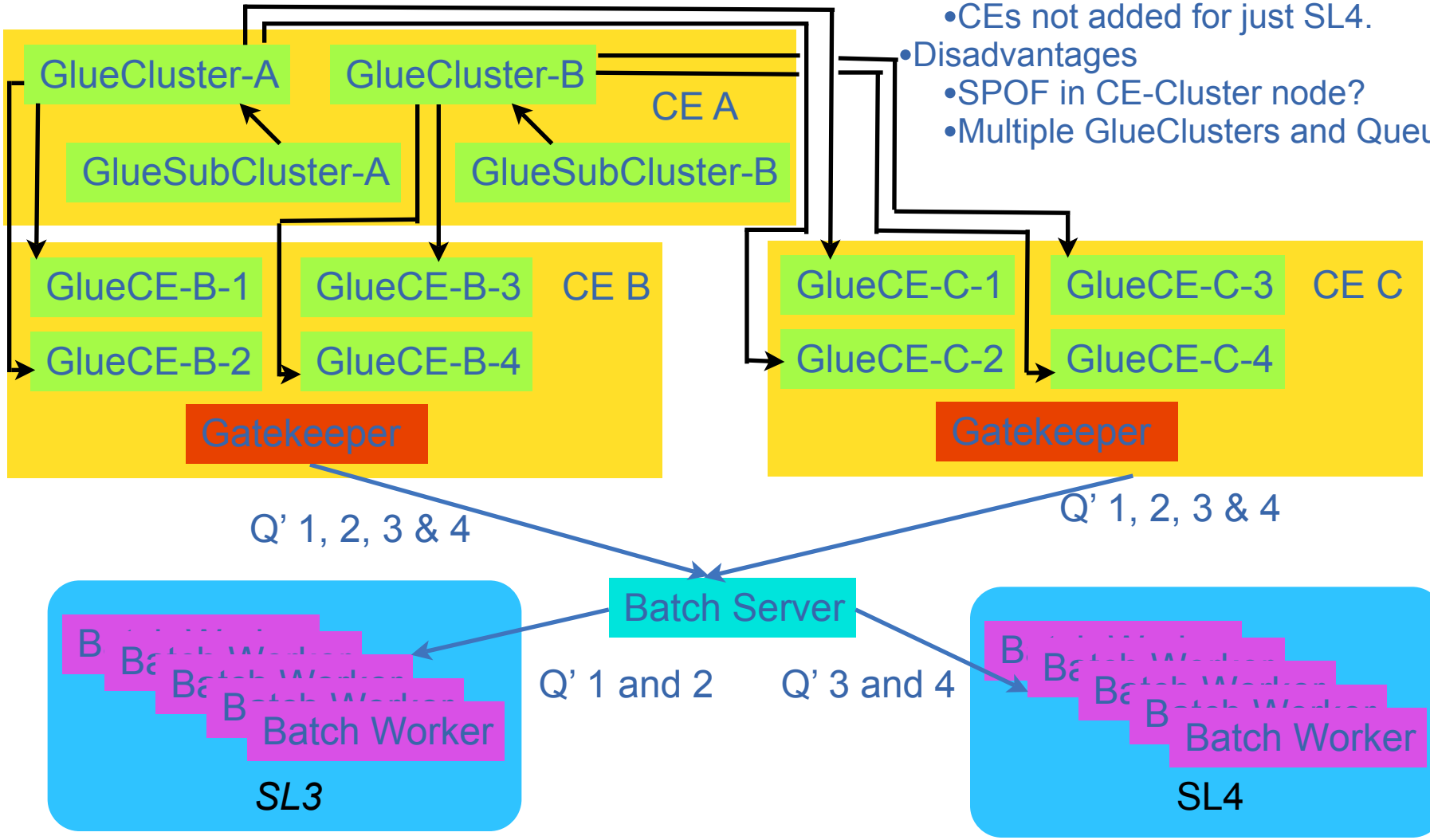
Short Term Fix at an lcg-CE Site

1. **Split the CE node type into three (yaim) node types.**
    1. **CE-ClusterPublisher** - Publishes GlueCluster and GlueSubCluster
    2. **CE-GateKeeper** - Configuration of lcg-CE (or creamCE).
    3. **CE-CePublisher** - Publishes the GlueCE (and VOView) objects.
   - These may well run on the same physical node of course.
   - None of these components interact with one another.
   - Only a case of detangling their configuration.
     - One YAIM function - *config_gip_ce* - needs to be split.
2. **Extend the ClusterPublisher so GlueCEs can be joined to named GlueClusters.**
3. **Extend the ClusterPublisher to support multiple GlueSubClusters joined to GlueClusters.**
4. **Extend the ClusterPublisher to support multiple GlueClusters.**
   - Only needed for an lcg-CE world.

Short Term Fix at lcg-CE Sites

- Addition of one CE-ClusterPublisher Node

- Advantages
  - All CEs identical supporting all jobs.
  - CEs not added for just SL4.
- Disadvantages
  - SPOF in CE-Cluster node?
  - Multiple GlueClusters and Queues.

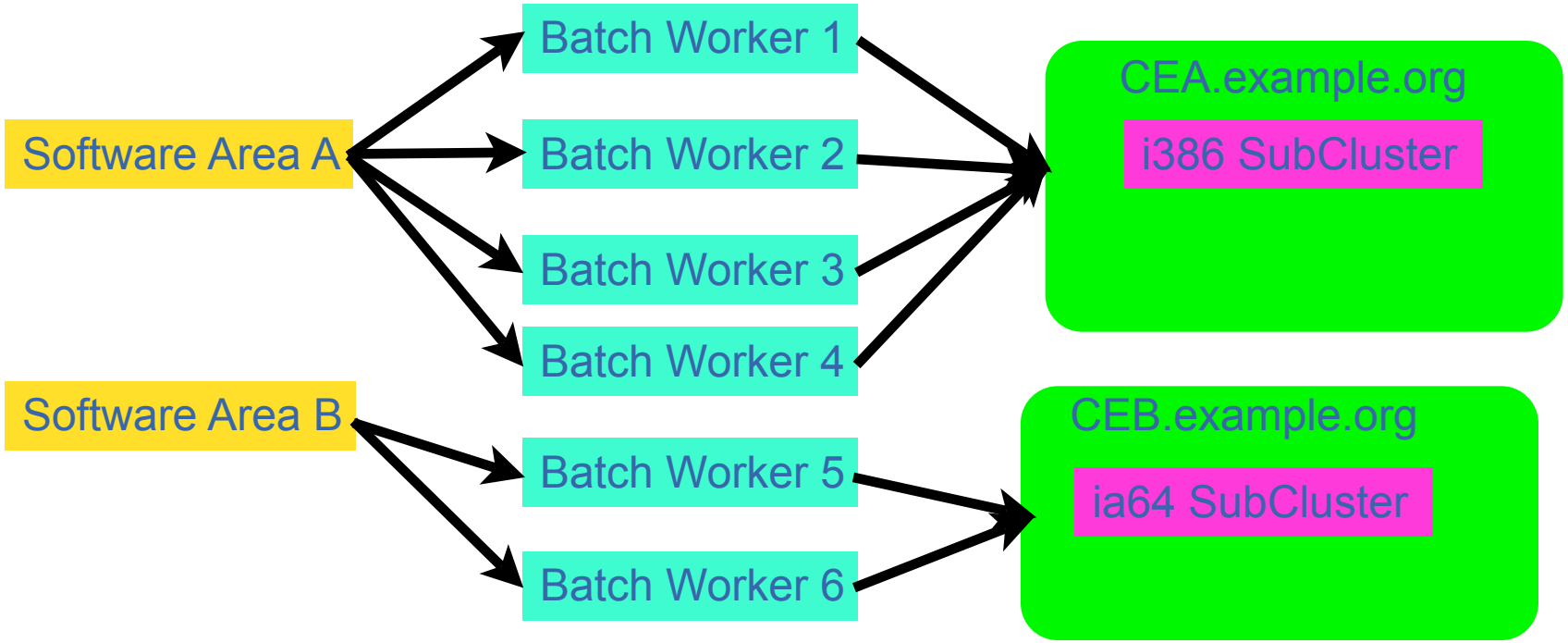


Deployment of Cluster Node

- **Avoiding SPOF when creating a new ClusterPublisher component.**
  - **Tiny - Small Site**
    - siteBDII, GlueCE, Gatekeeper, ClusterPublisher on one node.
  - **Medium - Large Site**
    - Move ClusterPubliser component to existing stand alone siteBDII
  - **Large - Huge Site**
    - Move ClusterPublisher component to existing load balanced siteBDII.
- **The GlueCluster and SubCluster are easy to locate anywhere.**
  - They don't require access to batch system.
  - They are completely static except for:
    - Significant hardware changes, i.e purchases, disposals a new OS.
    - Software Tags as published by the VO are in the GlueSubCluster.
      - *A rethink of lcg-ManageVOTag is consequently needed to proceed.*

ManageVOTags Today

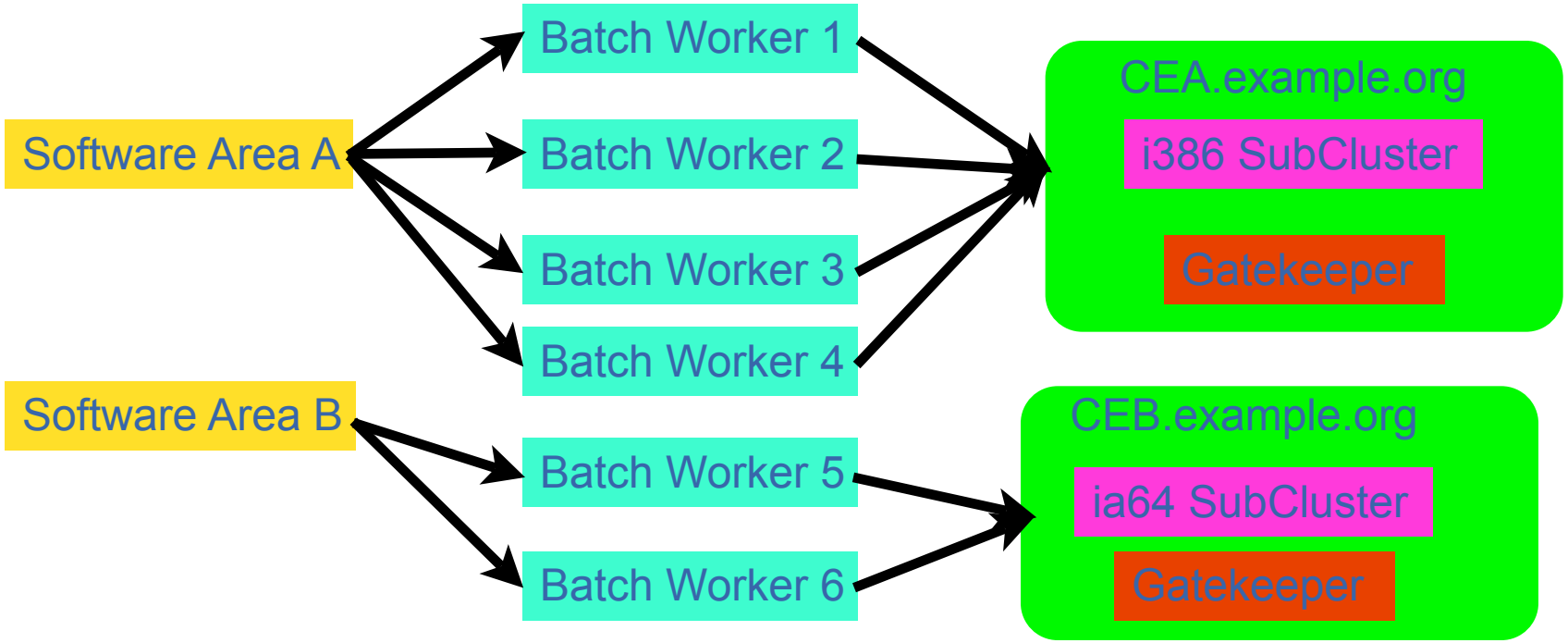
- The site setups the following mappings.



- A software area is mapped to a SubCluster.
- The SubCluster is hosted on the same physical host as the GateKeeper

ManageVOTags Today

- The site setups the following mappings.



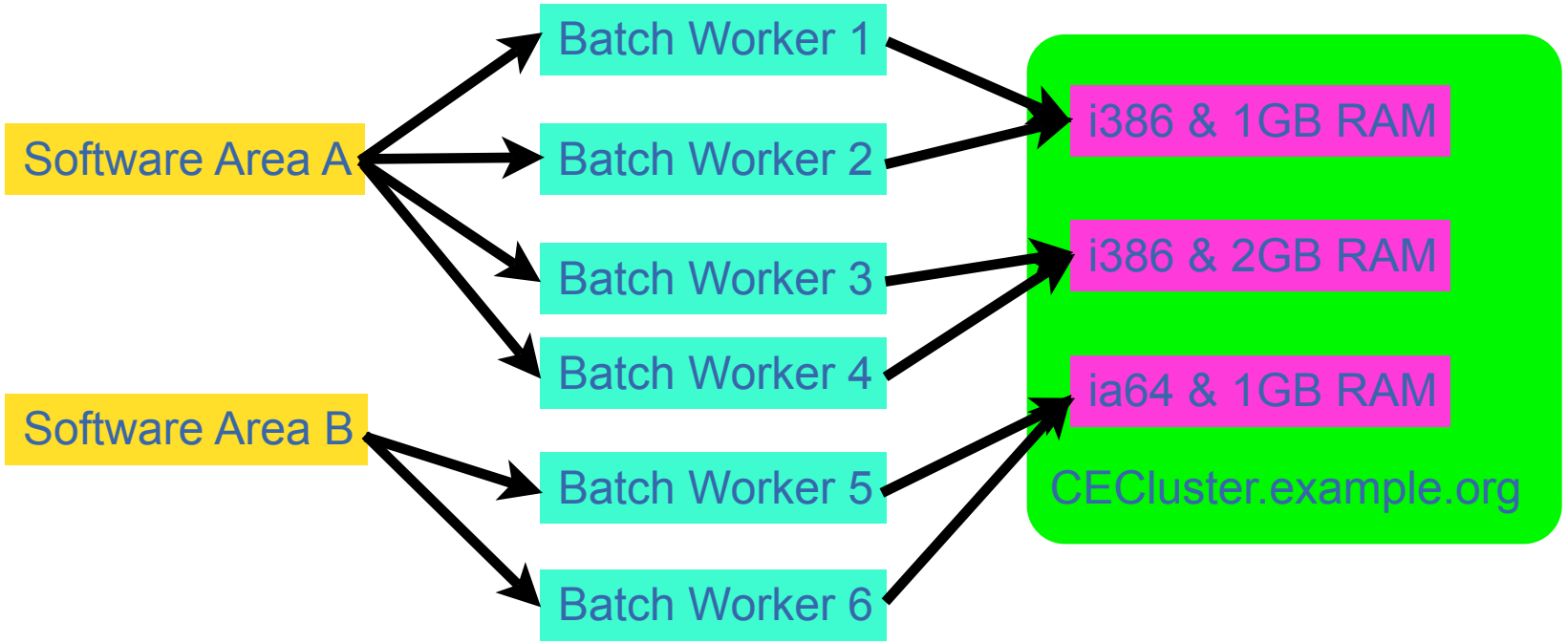
- A software area is mapped to a SubCluster.
- The SubCluster is hosted on the same physical host as the GateKeeper

ManageVOTags Summary

- **VO Tag Use Case**
  - A running job installs and/or validates software.
  - A GlueHostApplicationSoftware..... tag is added GlueSubCluster.
- **Current Solution is Two Part**
  1. From WN to CEnode (= 1 SubCluster) user runs:
     - *lcg-ManageVOTag -host <host> -vo <voname> --add <tag>*
     - This commands reads a config file on the CE <host> and drops the tag in a directory (typically /opt/edg/var/info).
  2. From CE to Information system a GIP plugin runs.
     - *lcg-info-provider-software -p /opt/edg/var/info -c <SubClusterId>*
- **Issues with Solution**
  - Till now all GateKeepers submit to exactly one SubCluster.
  - Different SubClusters may share a software area.
    - They obviously would for i686 and ia64.
    - They might not for 1 GB RAM or 2 GB RAM.

ManageVOTags Proposal

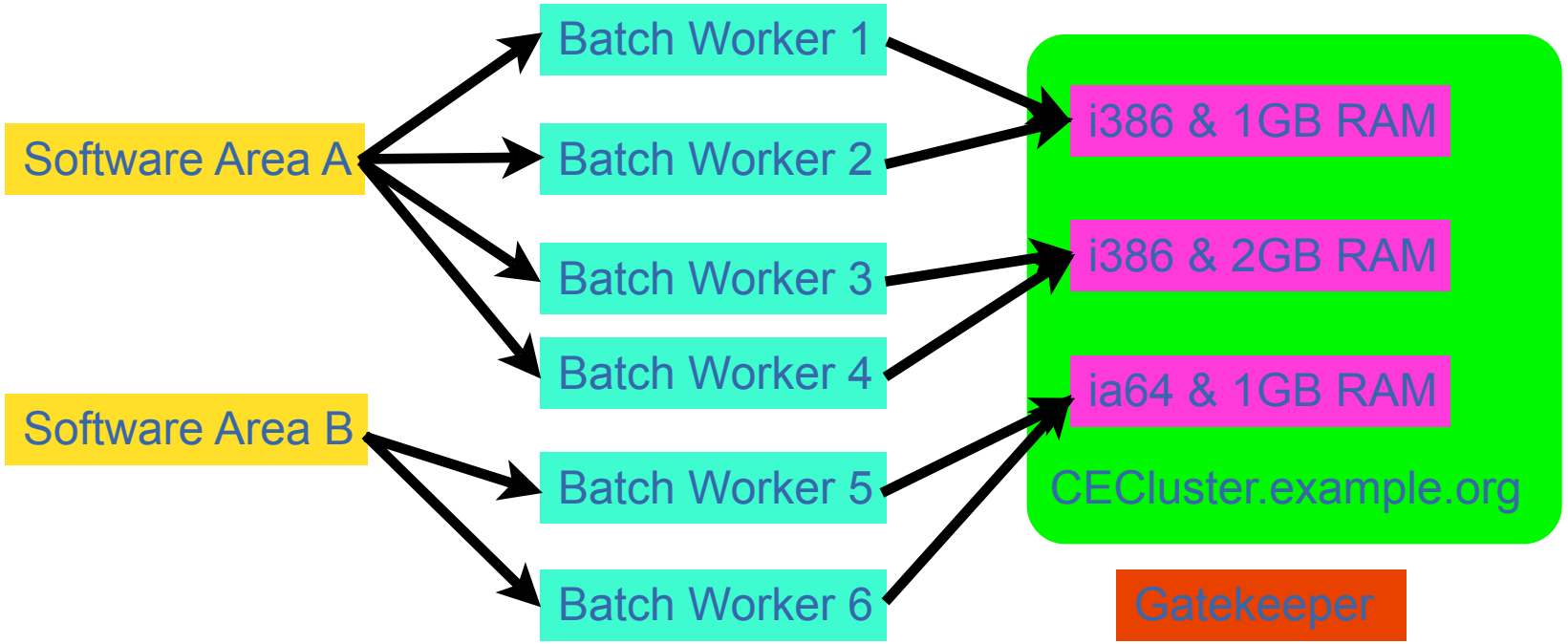
- The site setups the following mappings.



- A software area is mapped to at least one GlueSubCluster
- The sysadmin must maintain this mapping in configuration for any new lcg-ManageVOTags.

ManageVOTags Proposal

- The site setups the following mappings.



- A software area is mapped to at least one GlueSubCluster
- The sysadmin must maintain this mapping in configuration for any new lcg-ManageVOTags.

Software Area 2 SubCluster Map

1. **On every batchworker a configuration file.**

```
[batch100]
clusterID=IronCluster
clusterPath=/opt/edg/var/info/IronCluster
clusterHostPublisher=chp1.example.org,
                     chp2.example.org
[batch101]
...
```

   - This could be identical on every node using multiple entries.
     - Only the WNs own details are needed though.
   - Keeping your batch workers identical is desirable.
   - Support for two hosts, chp1 and chp2, is needed for load balanced ClusterPublishers.
2. **Place a root owned config file in the software area containing the clusterID, clusterPath and clusterHostPublisher(s).**
   - *$VO_ATLAS_SW_DIR/.swdir2subcluster.conf*
3. **Other suggestions welcome...**

New ManageVOTags Command

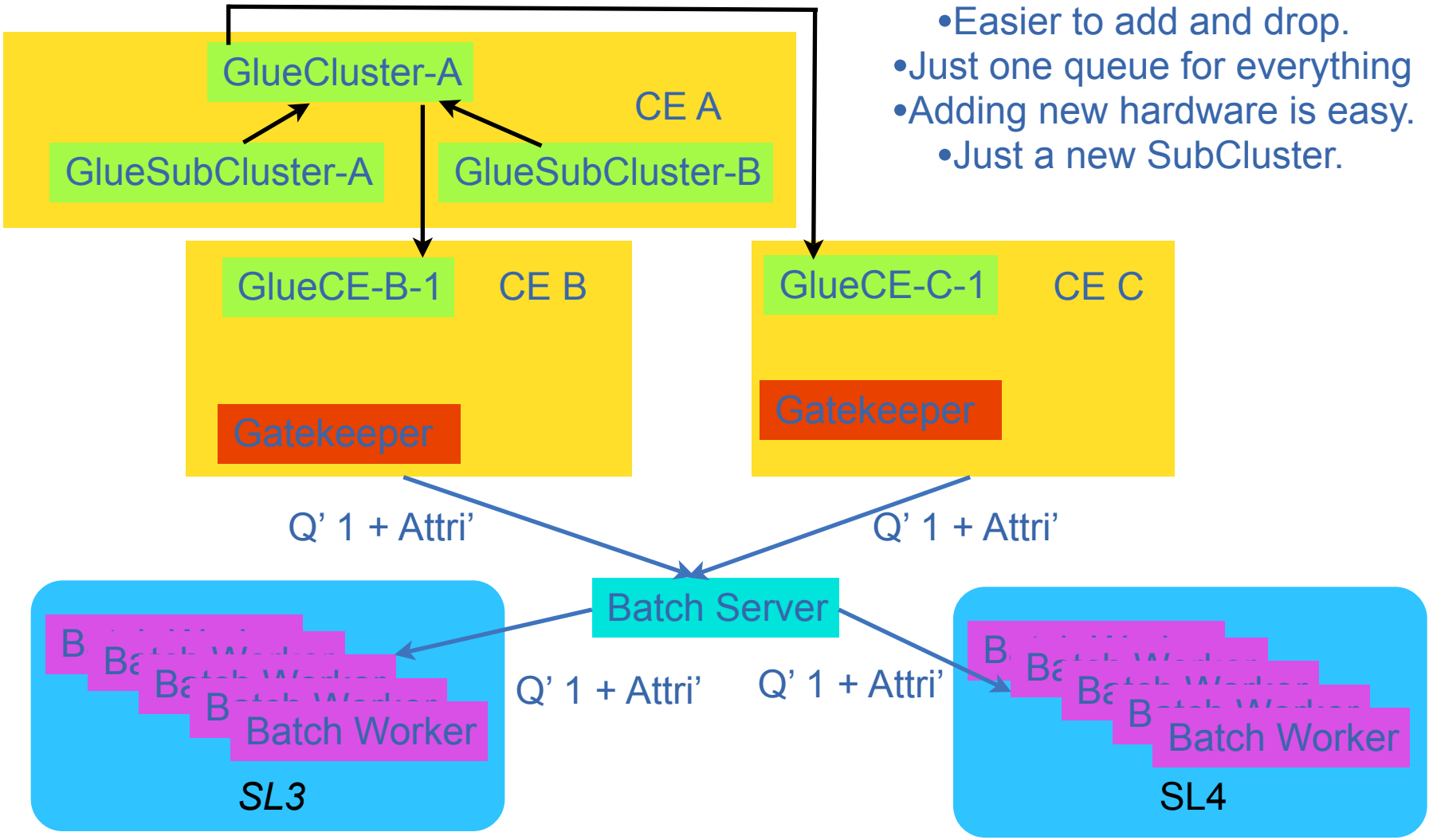
- **The ManageVOTags command when run within a grid job would reduce to**
  - `lcg-ManageVOTags [--vo <VO>] --add <MyTag>`
  - All the other options will come from the config file.
- **The new command can be backward compatible with the extra options being ignored if a configuration file is present.**
- **This migration is smooth and can happen staggered across the grid.**

Longer Term with a CreamCE

- **A CreamCE supports the passing of JobRequirements to the batch system.**
  - Verification has been done already within HEPiX by Ulrich (LSF) and Alessandra (torque) using the Blah+gLiteCE.
  - All this work must be reused of course.
- **When batch requirements can be passed:**
  - There is no need for queue sets per GlueSubCluster
  - We can get rid of multiple GlueClusters and replace with one as the schema intended.
- **It is only with the CreamCE that things can change a lot for sites or users.**
  - Small memory jobs be targeted at small memory machines.
  - Sites can add and remove new SubClusters trivially.
- **When moving from the lcg-CE to the BlahCE no more information system changes are needed.**

Longer Term with CREAM-CE

- All CEnodes are identical:
  - Easier to add and drop.
- Just one queue for everything
- Adding new hardware is easy.
  - Just a new SubCluster.



Unsolved Problems

- **Estimated Traversal Time - Again.**
  - Suspose a site has:
    - One SubCluster of 2 nodes with 4 GB RAM JobSlots
    - One SubCluster of 200 nodes with 2 GB RAM JobSlots
  - Submission of 4 * 3 GB Jobs will cause queuing and so ETT will deflect all jobs even the 1 GB ones.
  - Exactly the same problem as that which was “solved” by VOViews.
    - Is the same view solution applicable?
      - *VO Role * Memory * OS * Architecture = Lot of views.*
      - *VO Role * SubClusters = Not so many.*
        - *Is the last enough?*
        - *Can it even be calculated?*
- **Accounting.**
  - Impact if any on accounting should be checked.
    - Normalization may be easier due to homogenous SubClusters.

Small Test Bed

- **CERN PPS now contains a small 2 WN cluster.**
  - A CECluster with one SL4 and one SLC4 SubCluster. (it is a lie).
  - A GateKeeper and CEPublisher
- **This is an almost standard configuration with some manual YAIM alterations.**
- **Early testing shows it is working as expected.**
  - An updated ManageTags must be applied however.
- **A CreamCE will be taken at the earliest opportunity.**

Summary

- **Short Term**
  - Rewrite lcg-ManageVOTag to support multiple SubClusters
  - Break lcg-CE node apart at the yaim level to form:
    - CePublisher
    - ClusterPublisher
    - Gatekeeper
  - The ManageVOTag must be deployed first but only within a site. Not across the grid.
  - Release all of this through standard release process.
- **Longer Term**
  - Take the CreamCE.
  - Repeat the gLiteCE tests on argument passing completed under HEPiX.
  - Confirm that the rearranged CE information will support a CreamCE, matchmaking and argument passing.
  - Follow Glue 2.0 progress on DiskSpace.