# BIG FARMS AND THE GRID
# Job Submission
# and Monitoring issues

ATF Meeting, 20/06/03

Sergio Andreozzi

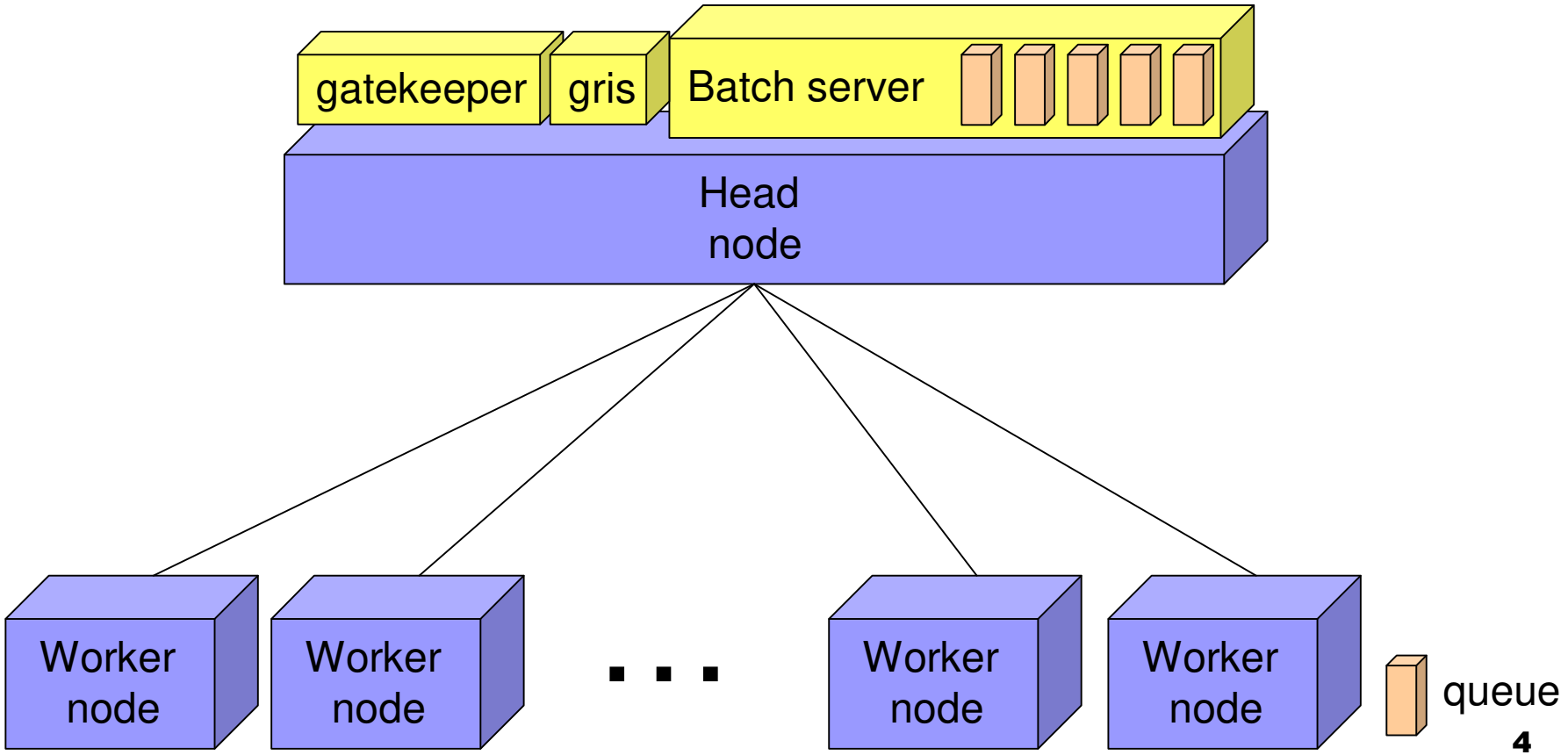
# OUTLINE

- Computing Resources in the Glue Schema
  - CE is access point to a queue (same as previous EDG)
  - only one access point envisioned per queue
  - E.g. edt001.cnaf.infn.it:2119/jobmanager-pbs-short
- Typical Cluster Configuration
- Broker Service and Cluster Services
  - Description of a the interaction
  - (goal: understand the submission process and how cluster services are stressed)
- LCG Proposed Cluster Configuration
- Job submission in this new scenario
- Discussion
- Proposal

# Typical cluster layout

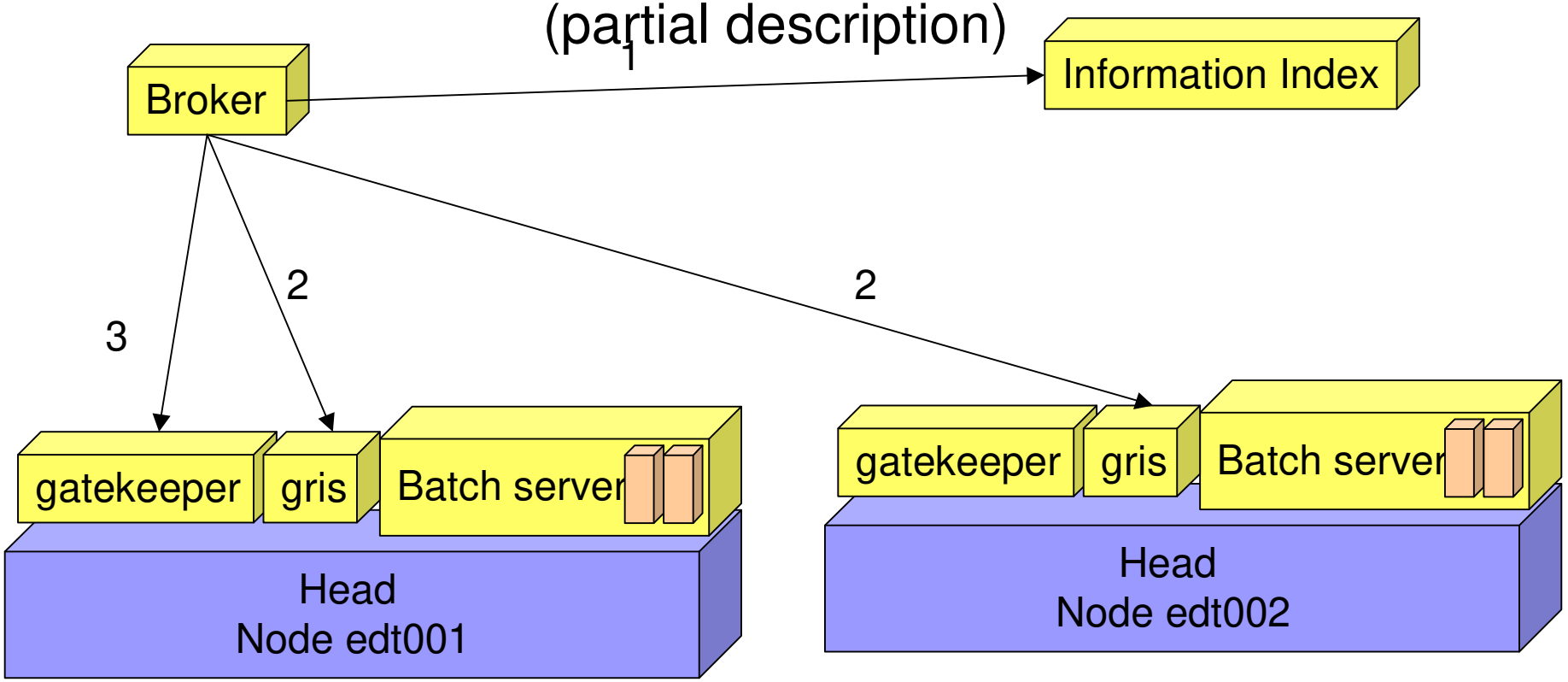
- In DataGrid, the typical cluster configuration is:
  - One node running:
    - Batch server (e.g. LSF, PBS)
    - MDS GRIS
    - Gatekeeper
  - Several Worker nodes driven by the batch server for job execution

# Typical cluster layout



# Typical job submission

(partial description)




Two different clusters

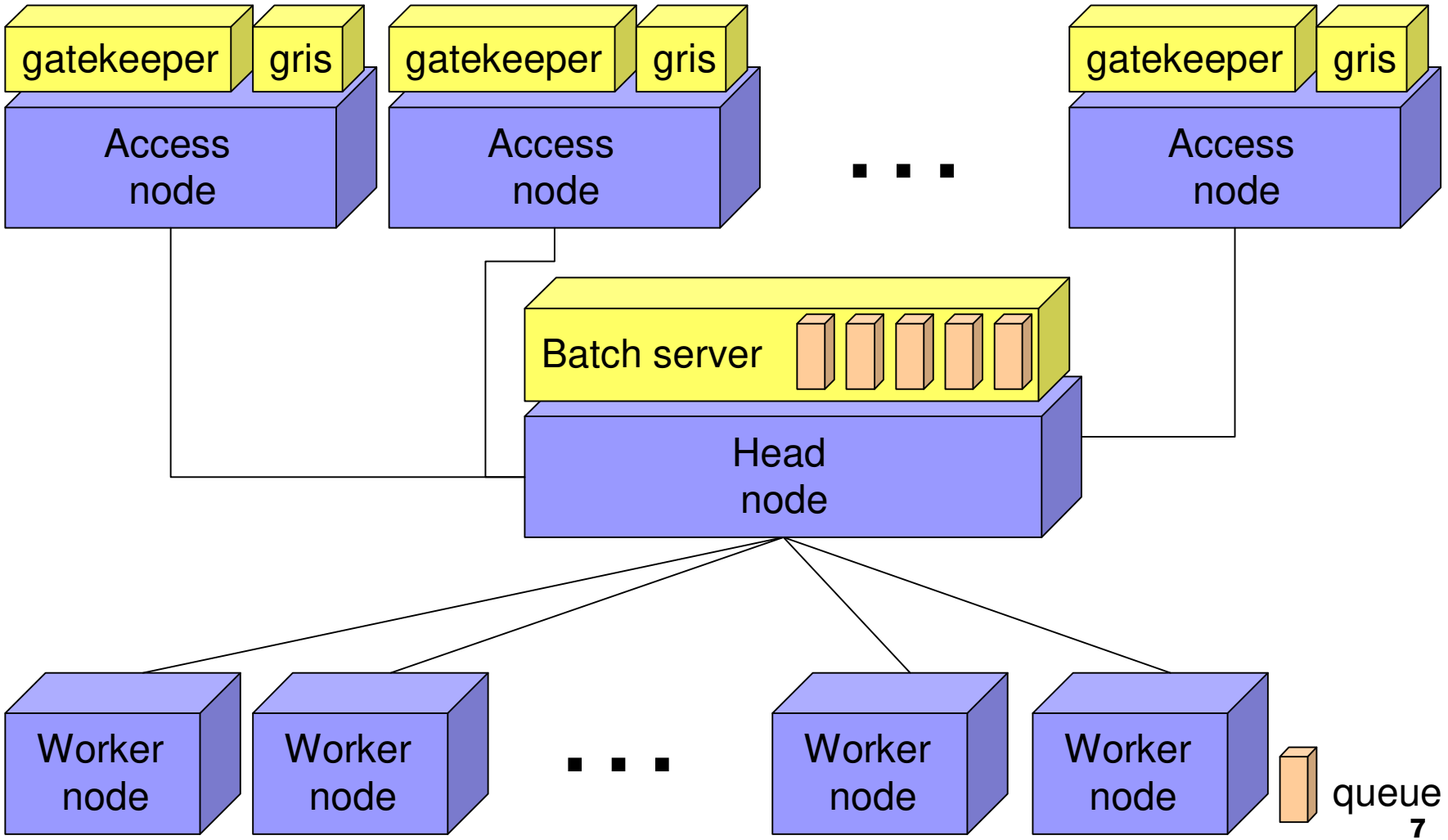
1. From II, broker gets list of CEs a user can access and that match JDL requirements
2. For each selected CE, the GRIS is contacted to get params in JDL rank option to order
3. The first CE in ordered list is used to run the job

(I don't remember if requirements are checked in 1 or 2)

# Cluster layout @ CERN

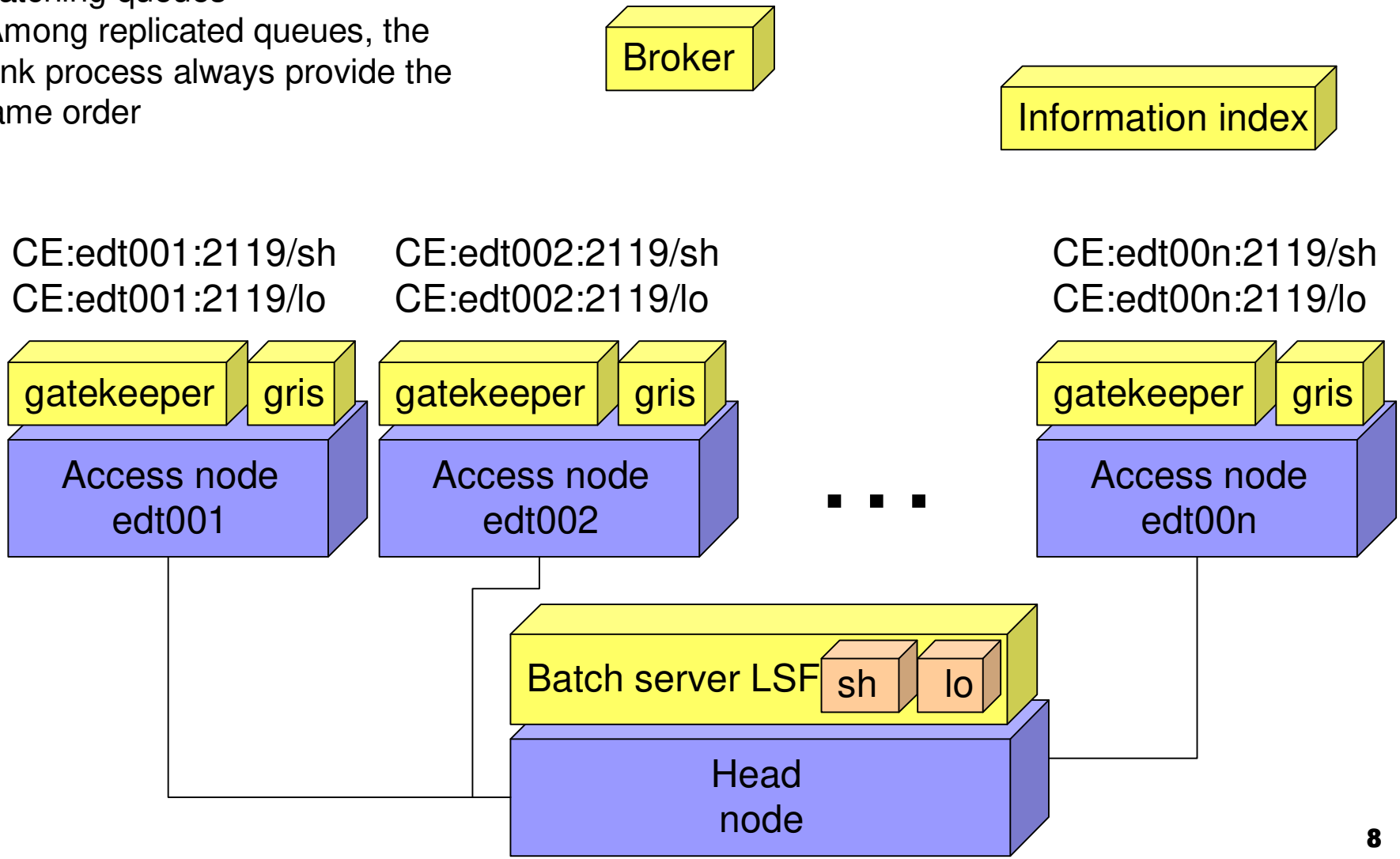
- People@CERN assert:
  - Gatekeeper service can be heavily loaded when managing several job submissions
    - e.g. gatekeeper design issue; one living process for each submitted job till the end of the computation
  - For scalability they deploy gatekeeper service in a different node than batch server
  - They can have several nodes running gatekeeper for the same batch server
  - They plan to set up a big farm with 400-1500 nodes with a set of O(10) access nodes and only one batch server
  - LSF can manage O(1000) nodes

# Cluster Layout @ CERN



- Each CE is a different queue for the broker
- Load balancing made among matching queues
- Among replicated queues, the rank process always provide the same order

# Example



# Cluster layout @ CERN
# Adv & Disadv

- +
  - Can scale to higher number of parallel job submission
  - From the site manager viewpoint, it provides great flexibility in managing/configuring the farm
- -
  - Not envisioned in both EDG schema and Glue schema -> duplication of info
    - E.g. given a queue/CE on a batch server, this will show in the GIS as many times as the number of configured gatekeepers for the batch server managing the queue
  - The LRMS is stresses from several info providers asking same info;
    - e.g. 10 gatekeepers/GRIS for an LSF server refreshing info each 30s
    - 20 req each min instead of 2… this might be a problem

# Discussion

- How does the broker deal with this scenario?
- What are the needed changes to support this?
  - GIS schema
    - introducing the concept of Access Point; a queue can have several Access Point
    - Defining a quality parameter for an access point (so that the broker can choose the less loaded one)
  - broker service
  - monitoring service

# GRIS AND GATEKEEPER SERVICES

- The important question is:
  - DO WE NEED TO REPLICATE THE GRIS AS WELL?

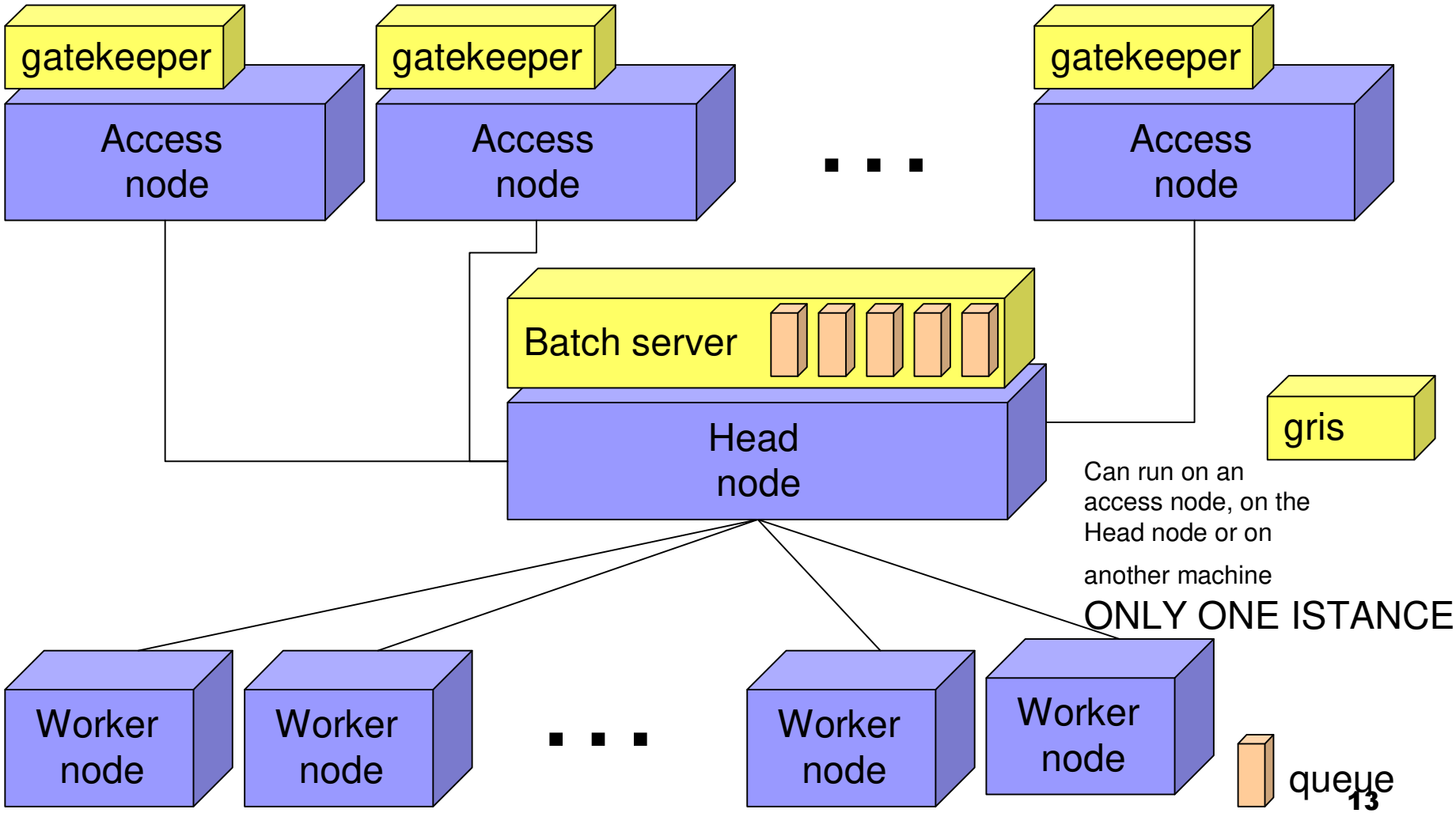
- The gakepeer does not need the GRIS
- The GRIS at the moment need some info from the gatekeeper
  - e.g. hostname, port, access to gridmap file

- If we don’t miss anything else, they can be easily decoupled on different machines

# PROPOSAL

- Decouple the CE Unique ID from the entry point
- CE ID should be just a global unique ID for the queue
- E.g.
  - <HEADNODE>/jobmanager-<LRMSType>-<QUEUENAME>
- Introduce a new attribute GlueCEAccessPoint
  - E.g. = current GlueCEUniqueID
- ONE GRIS PER BATCH SERVER

# Proposed scenario



# EXAMPLE CE REPRESENTATION WITHIN MODIFIED GLUE SCHEMA FOR LDAP

```
dn: GlueCEUniqueID=edt001.cnaf.infn.it/jobmanager-pbs-short,
   Mds-Vo-Name=local, o=grid
...
GlueCEUniqueID: edt001.cnaf.infn.it/jobmanager-pbs-short
GlueCEAccessPoint: edt002.cnaf.infn.it:2119/jobmanager-pbs-short
GlueCEAccessPoint: edt003.cnaf.infn.it:2119/jobmanager-pbs-short
GlueCEAccessPoint: edt004.cnaf.infn.it:2119/jobmanager-pbs-short
GlueCEStateFreeCPUs: 5
GlueCEPolicyMaxRunningJobs: 10
GlueCEAccessControlBaseRule: …
GlueCEAccessControlBaseRule:…
…
```

# Broker modification

- When querying the GRIS, the broker will maintain the several Access Points
- Once it selects the queue, it will submit to a random gatekepeeper

- Further improvement*:

- `GlueCEAccessPoint: edt004.cnaf.infn.it:2119/jobmanager-pbs-short:<CE_AP_LOAD>`
- Where CE_AP_LOAD is a metric that let the broker able to rank the access points

- BENEFITS:
  - the broker will do less work during matchmaking process (no duplicated GRIS to be queried to get info for ranking on EACH JOB SUBMISSION)
  - the broker will choose a random gatekeer among the listed as it does now (so no worse behaviour); this can be improved*

# Monitoring

- No need to deal with replicated info
- Able to show real number of queues, their state and access points
- Detailed host info can be aggregated and presented as access node loads