Network-aware Workload Management

Tommaso Coviello
INFN and Politecnico di Bari, IT

IT-CZ JRA1 Cluster

3rd EGEE Conference, Athens, Apr 20, 2005

Outline

Problem statement:

network-aware resource ranking

Requirements

WMS prototype status and deployment scenarios

Conclusions

Problem statement

- Accurate information about Grid resource status needed by the WMS in order to perform efficient workload distribution
- No network-awareness in the current WMS implementation

Extend the WMS

- resource discovery capability and
- ranking algorithm

to include *network status* information

CE/SE ranking algorithm: example

- ***Purpose***: for a given job, discover a CE and SE in order to minimize the job Minimum Completion Time (MCT) where:

  $CompletionTime(CE_i) =$

  $JobExecutionTime(CE_i) +$

  $+ \max \{InputDataTransferTime(SE_j \text{ --> } SE_k), QueueTime(CE_i)\}$

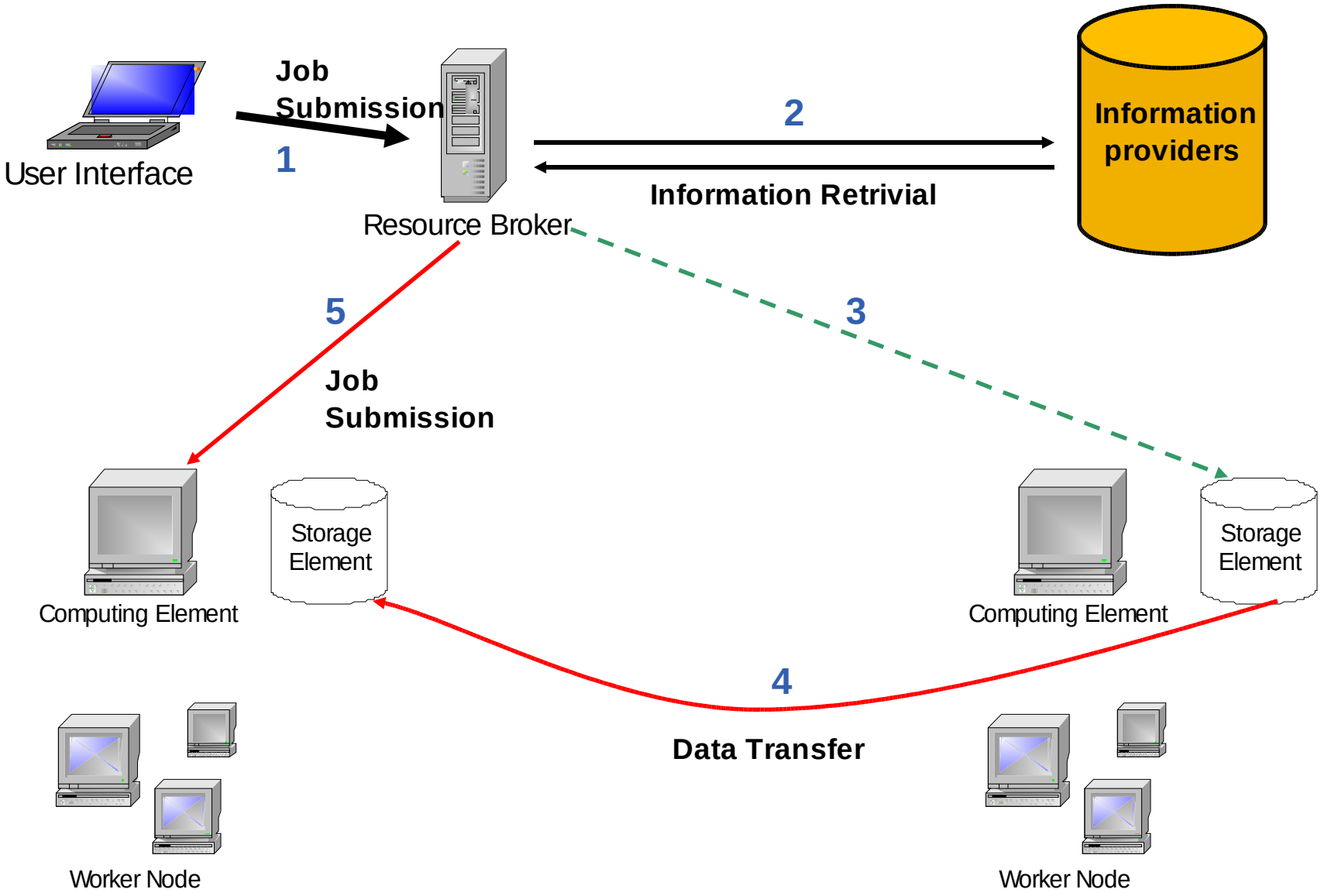
  (where SE_k is “close” to CE_i)

- WMS selects (CE_i,SE_j) which minimezes the job completion time:

$$MCT=\min\{CompletionTime(CE_i,SE_j)\}$$

$$\text{for } i=1..m,\ j=1..n$$

MCT at work


User Interface
Job Submission
1
Resource Broker
2
Information Retrivial
Information providers
5
Job Submission
3
Computing Element
Storage Element
Storage Element
Computing Element
4
Data Transfer
Worker Node
Worker Node

MCT computation

- QueueTime estimation on CE_i resource.
  - Information provided by the CE GLUE schema attribute: **GlueCEStateEstimatedResponseTime**
- JobExecutionTime on CE_i resource
  - depends on a number of factors, such as CPU architecture and hardware configuration
- InputData transfer time from the Storage Element SE_j to the target SE_k
  - depends on network status and file size

Problem statement:

network-aware resource ranking

Requirements

WMS prototype status and deployment scenarios

Conclusions

Requirements: Network metrics

How to estimate reliability and file transfer time on a given network path?

- Packet loss (one-way or two-way): reject paths that experienced some packet loss
- Available TCP throughput: for computation of file transfer time
- Instantaneous packet delay variation (if available): an indirect measure of network load and congestion

More requirements

Number of performance data points:
- A few per hour, for a maximum of 24 hours

Deployment of network sensors:
- One sensor for each main Grid site

Minimum query performance required to information provider:
- ~ 5 queries/s
- Use of caching if useful to decrease query latency

Discovery:
- For a path connecting a given couple of end-nodes, retrieve the most suitable performance results

Interoperability

Problem statement:

network-aware resource ranking

Requirements

WMS prototype status and deployment scenarios

Conclusions

WMS network-aware prototype status

Goal: ***demonstrate if usage of network status information actually improves workload management***

- A research branch of JRA1 IT-CZ
- Current prototype implementation based on the network sensor infrastructure (Gluedomains) currently deployed in INFN GRID
- Network performance data repository: the GridICE DB, the INFN GRID monitoring tool
- GridICE:
  - Collection of new data every 20 min
  - Query load is an issue

Gluedomains

Gluedomains:

- a network performance monitoring framework based on the partitioning of the Grid into Domains
- Typically one sensor per domain
- Measurement of domain-to-domain connectivity status (scalability)
- publishes data (LDIF format) into a site GRIS (LDAP directory service)
- deployment of LCG-2.4 on INFN GRID will bring in new Gluedomain sensors

GridICE:

- The central DB is populated by collecting network performance data from the various site GRIS
- Pull model
- Currently can return network metrics in the following formats:
  - LDIF
  - xml (non-standard)

Problem statement:

network-aware resource ranking

Requirements

WMS prototype status and deployment scenarios

Conclusions

Conclusions

- **Interoperability**:
  - interest in supporting an NM-WG interface-compliant client in the WMS in addition to the current LDAP-based one
- **Deployment**:
  - Population of the INFN GRID GridICE DB with EGEE performance data
- **Open issues**:
  - JRA4 publisher vs mediator (merge?)
  - Performance is crtiical --> caching? Where?
  - JRA4 information provider:

    option1: the JRA4 mediator (currently no internal cache, designed for human clients)

    option2: the RGMA secondary producer (collecting data from WP7 sensors, which are RGMA primary producers)

    option3: ask the Grid information service available