Primary actor
Goal in context
Scope
Level
Stakeholders and interests
Precondition
Minimal guarantees
Success guarantees
Trigger
Main success scenario
Extensions
Technology and data variations
Priority
Releases
Response time
Frequency of use
Channel to primary actor
Secondary vendors

HEP replica management

Overview

This collection of use-cases is taken from the proposed (?) EGEE computing architecture, plus scenarios predicted for its use. [hopefully DJRA1.1]
The collection includes a number of high-level uses of the transfer-management middleware proposed for use in EGEE. These use-cases are: A Tier 1 Site retrieves raw data from the LHC, Data reprocessing? and User submits a job to run on the Grid. These use-cases provide a context for the use-cases involving requests to transfer a file, specifically X requests a logical file to be transferred to an SE+ and X requests that a physical file be transferred to an SE+. These use-cases make use of the remaining use-cases, which represent the functioning of what I call the Data Scheduler architecture, a description of which is given below.

The Data Scheduler Architecture

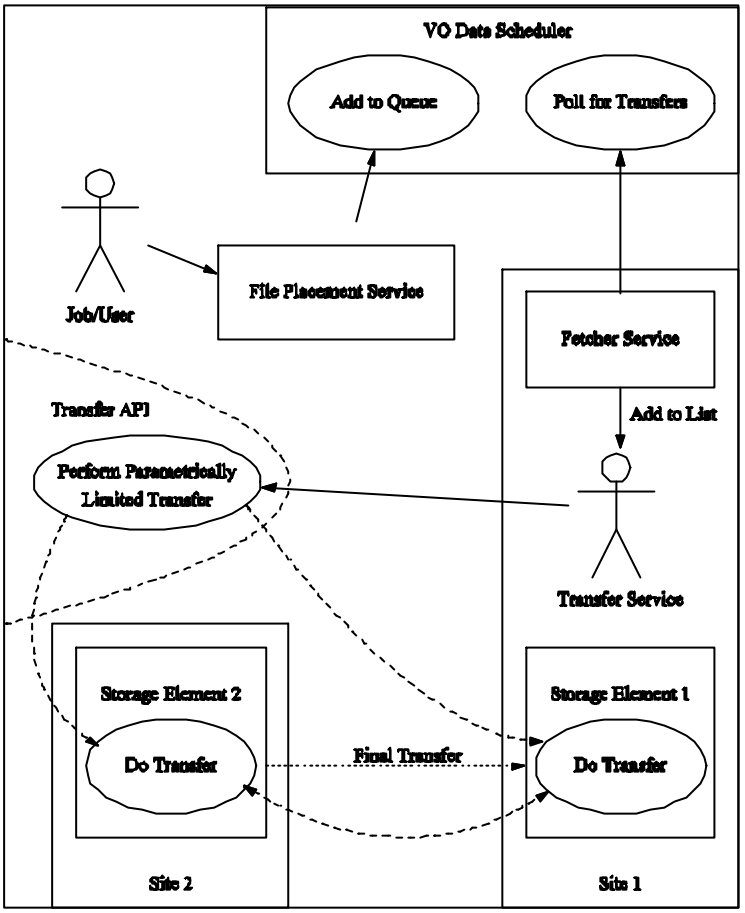


Figure 1: The entities involved in a transfer

For each Virtual Organisation (VO), a top-level Data Scheduler maintains a list of transfers required for that VO that cannot be handled locally. At each Site, a VO Fetcher Service polls the VO Data Scheduler for new transfers *to* Storage Elements at that Site. Any transfers it finds are added to a list maintained by the Transfer Service. The Transfer Service maintains a list of all transfers to that Site, which includes transfers routed via the Fetcher Service, or from a local File Placement Service. There will be one Fetcher Service per VO at each Site.
The Site Transfer Service later makes use of a generic service whose physical location and implementation is at present undecided to perform the transfers according to per-VO usage rules laid down by the Transfer Service. Ultimately, however, a file will be transferred from Storage Element 2 to Storage Element 1.

Notes

Where a step constitutes a complete use-case in itself, the name of that use case is given underlined.

Use cases

A Tier 1 Site retrieves raw data from the LHC

CERN will produce a lot of data when the LHC is up and running. The data is streamed from the detectors and resides on disk for a while before being translated to tape. While the files are on disk, Tier 1 sites can pull over any files they are interested in. After the files are migrated to tape, retrieving the data will involve staging them to disk first, and this is time-consuming and inefficient. Each Tier 1 site will be interested in different pieces of the data, partitioned by experiment but also on activities within each experiment. [PeterKunsztEmail]

Actors

Data Manager

Stakeholders and interests

The Tier 1 Site: requires that any interesting data is transferred in to an SE at that site, before it is migrated to tape at CERN.
The Tier 1 Site: also requires that any bandwidth reservations it makes are made use of.
CERN: must not exceed its physical bandwidth capability.

Main success scenario

1. The LHC experiments stream data to disk at CERN as it is recorded.
2. A Site Data Distribution Tsar* at the Tier 1 Site is notified as to what data is available, and decides which files are of interest to the site.
3. The Tzar might reserve bandwidth from CERN to the Site to ensure that all the data files are transferred in time. It depends on whether bandwidth reservation can be done efficiently on a per-transfer basis.
4. For each file, the Tzar then requests that that file is transferred to an SE at the Tier 1 Site, before the files are migrated to tape. This request process may be done as a batch, but otherwise the mechanism is probably the same.
5. The data on disk at CERN is then migrated to tape.

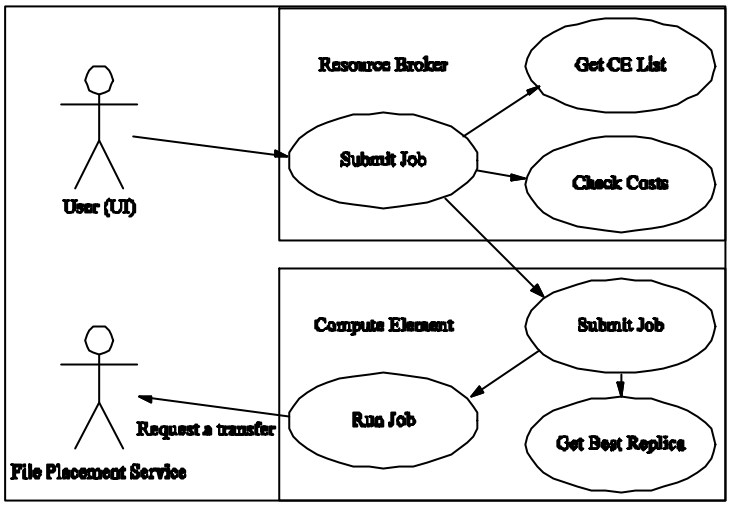
Extensions

4a. Step 5 occurs before all the files are retrieved

* my term.

Data reprocessing?

User submits a job to run on the Grid



Actors

The User

Stakeholders and interests

The Compute Element administration:

Main success scenario

1. The User submits a job via his user interface to the Resource Broker. The job specification contains all the restrictions and requirements on the CE that must run it, as well as a list of required logical files, plus the physical filenames of any data produced by it.
2. The Resource Broker obtains a list of CEs that are capable of running the job.
3. The Resource Broker locates all the physical locations of the required logical files.
4. For each CE and for each required logical file, the Resource Broker calculates the cost of retrieving all the replicas of that logical file, and of placing any output files onto their target SE. It picks the CE for which the total cost would be lowest. This cost includes the cost of transferring the files over the network.
5. The Resource Broker passes the job on to the Compute Element, where it is queued.
6. The Compute Element eventually runs the job.
7. The job requests each logical file should be transferred to the local SE (at least, it requests that each logical file should be made available to it, and local middleware is configured with the SE to use).
8. The job runs.
9. The job transfers any output files to the designated SE (should this be a link to X requests a physical file should be transferred to an SE?).

Extensions

X requests a logical file to be transferred to an SE+

Actors

The primary actor in this use case is the X. The X represents any entity which can request a file transfer, including computational jobs, replica management middleware and so on, as well as a real user.

Stakeholders and interests

The User: requires that a particular logical file be made available on an SE.
?????: requires that best use is made of all resources for all transfers.

Main success scenario

In this scenario, we assume that the replica is not already available on the target Storage Element.

1. The X calls the getBestReplica function, passing the logical name of the file he requires, and the target SE (FIXME: actually, does the middleware work out where he wants to place the file?).
2. getBestReplica locates all the replicas of the logical file.
3. getBestReplica calculates the total cost of making each replica available on the target Storage Element, including transfer times.
4. getBestReplica requests that the best replica be transferred to the target SE.

Extensions

1a. The logical file already exists on the target SE.
1a1. getBestReplica need do nothing, and can return immediately.

X requests that a physical file be transferred to an SE+

Actors

The primary actor in this use case is X. X represents any entity which can request a file transfer, including computational jobs, replica management middleware and so on, as well as a real user. Specifically, the X might in this case might also represent the getBestReplica function of user requires a logical file to be transferred to an SE.

Stakeholders and interests

The Site: the site has agreements with the VOs to provide a certain level of service for file transfers. It wants to avoid allowing the VOs to exceed their agreements.
The VOs: each VO wants to ensure that it makes best use of its agreements with each Site. The VOs also want to ensure that the users do not violate policy.
The User: wants to transfer a file to the Target Storage Element.

Main success scenario

1. The User requests of his local File Placement Service that a file be transferred between SEs.
2. The File Placement Service requests of the VO Data Scheduler that that particular replica should be transferred to the Target Storage Element.

3. After some time, the VO Fetcher Service at the Target Site polls the VO Data Scheduler for any new transfers to Storage Elements at that Site.
4. The Fetcher Service adds any new transfers to the Transfer Service at the Target Site.
5. For each file it has listed, the Transfer Service transfers the file.

Comment [PDM1]: How complicated is this process? If some scheduling goes on when the file is added, possibly there will be some steps that involve network services.

Extensions

1a. The request violates VO policy at Site level.
2a. The request violates VO policy at Data Scheduler Level in a fatal manner.
2b. The request must be transformed or delayed so as to avoid breaking VO policy.
2c. The Target Storage Element is at the same site as the File Placement Service.
2c1. The File Placement Service adds the transfer to the Site Transfer Service.
4a. The Transfer Service reports that the transfer breaks Site policy.

Site Transfer Service transfers a file

Trigger

The Site Transfer Service

Main success scenario

1. The Site Transfer Service schedules a window in which the transfer should take place, ensuring that the VO to which the transfer belongs does not exceed any limits set on it.
2. The Transfer Service then passes the file source and destination information, the latest arrival time, the earliest pickup time and any bandwidth limitations to an as-yet undecided service S1.
3. S1 ensures that it can complete the transfer according to the limitations.
4. S1 transfers the file

Extensions

These extensions are taken from Peter Clarke's use case shown below.

3a. The file cannot be transferred according to the limitations without requesting other services.
3a1. S1 queries the information services to discover any services that it can use to complete the transfer in time
3a2. S1 subscribes to any such services, and uses them to transfer the file (this replaces step 4).

Compare with Pete Clarke's use case below. The two are very similar, except for the nomenclature, and the fact that Peter Clarke's specifically includes the bandwidth allocation sections, while Peter Kunszt's does not.

High Energy Physics replication of re-processed data from a central point to several data centres

Use Case Summary

The High Energy Physics experiments will record data sets of several PetaBytes per year. This is re-processed from time to time at a subset of data centres. After re-processing it must be delivered to all other data centres according to some strict delivery parameters.
Data centres are divided into three Tiers: Tier 1 sites each have all the raw data; Tier 2 and 3 sites have a subsets of raw and processed data.
Tier 1 sites are distributed so that large geographical areas (on the scale of a country) each have one in.

Background Scenario

- Following a period of data-taking, each of the Tier 1 sites has a copy of the raw data set for an experiment. This data set is approximately 2 petabytes in size. The Tier 1 sites are distributed throughout the world as described above. These data have already been processed once (as soon as they were recorded). This process takes the data from the Raw form to ESD form, with a volume reduction of a factor of 10. Raw and ESD data have a simple file format.

- The experiment management decides that enough further detector calibration has been performed for a complete data re-processing cycle to be performed. This requires that the complete data set is passed through and reprocessed again from Raw to ESD data. The responsibility for this is handed to the data re-processing Tzar.

- As the data sets are replicated at several sites, a process occurs (not described here) to select three sites which will each run 1/3 of the reprocessing. Assume these are widely geographically separated (for example, one in the Asia-Pacific area, one in the EU and one in the US)

- The Tzar requires that the re-processing is completed within 2 weeks of commencement (again assume the enough CPU has been identified for this). During this process the re-processed ESD data is produced pseudo-continuously, and results in the ability to produce a set of interim ESD files on a daily basis.

- The Tzar requires that the re-processed ESD data is distributed to all Tier 1 sites throughout the World and that this should be complete within 1 week of the finish of re-processing.

- The Tzar will use some Data Distribution Service which will take responsibility for arranging all logistics of data delivery. The Tzar will want to hand off responsibility for completion, and merely be notified when the job is done.

Customers

In this case the customer is the management of a high energy physics experiment and, by delegation, the data re-processing Tzar.

Scenarios

Data distribution by a DDS

- This data distribution is undertaken by a Data Distribution Service (DDS). This is the principle client of network services.
- The DDS is handed the relevant information, which includes
    - Identifiers of data to be delivered
    - Pick up address(es) for some services able to serve data to be delivered
    - Destination address(es) of some agent able to negotiate accepting delivery
    - Latest time for delivery
    - ???
- The DDS is able to contact each of the remote sites and query for a delivery endpoint. The remote site also specifies a profile during which it can accept the data. This profile specifies
    - earliest start,
    - latest end,
    - maximum receive rate
    - ???
- The DDS queries the information services to check if the predicted data rate for “normal” streams is high and stable enough to allow the data to be delivered by the latest end time. In this scenario, we assume that the data rate is not predictably high and stable enough.
- The DDS therefore decides to request (how) a low level network service which can accept as parameters:
    - Source and destination information (need to be more specific)
    - earliest start,
    - latest finish,
    - payload volume,

    - maximum delivery rate at destination,
    - some information specifying how and when the data can be picked up
    - ????

- The DDS hands responsibility for delivery to the network service (NS1) (note: this implies the service must actively come and pick up the data at some point in the future).
- The NS1 must now decide how best to honour the agreement. Firstly it must pick up the data within the parameters specified. In general this will required reading the data at some point from the source within a finite period, probably at some minimum rate in order to ensure the pick up doesn't take longer than some specified time (this is all to be defined in the "pick up information")
- Independently the NS1 must decide if/when to store and forward the data. During any forward process NS1 will likely require some minimum transfer rate averaged over a specified time.

Main success scenario

Data Processing Tzar hands all file information to DDS

DDS checks that each file can be transferred in time

DDS transfers file

Extensions

1a. Tzar hands incomplete information to DDS

2a. File cannot be transferred in time

2a1. DDS hands information to NS1

2a2. NS1 transfers the file

Functional requirements

- The DDS must be able find out the expected time for a given volume of data to be transferred between two endpoints, if it is transferred without any special provisioning; more specifically, the DDS must be able to calculate whether the probability of a particular transfer taking place by a given time is acceptable.
- The DDS must be able to discover the bandwidth-allocation resources available between two endpoints.
- NS1 must be able to reserve a service which guarantees that a volume of data will reach its destination in a given time, given these parameters:
    - Source and destination information,
    - earliest start,

- latest finish,
- payload volume,
- maximum delivery rate at destination.

Service utilization

In these scenarios, little use is made of standard Grid services such as traditional resource brokering (i.e. meta-scheduling) or fine-grained replica management between large numbers of separate storage elements.
The DDS or a service utilised by the DDS requires a prediction of the average bandwidth over a period of time, and the expected error on that bandwidth. Use of this service by the DDS might be via a network cost estimation service or some other service.

Security considerations

The guaranteed delivery time service will make use of a finite resource (in particular, the reservable bandwidth of a link or series of links). Some applications may require the guarantee more than others. Some sort of conflict resolution system must be in place for this eventuality.

Performance considerations

A very large number of files containing a very large total volume of data will be reprocessed and replicated in a relatively short period of time. Replication between the Tier 1 and 2 sites means that generally files will be transferred in large groups between a limited number of endpoints, so the number of requests for services may not be that great. Services must be reasonably robust, in that agreements will last on the scale of days.

Use case situation analysis

References