DIGITAL LIBRARY services and their interoperability based on GRID technology

Amel BOUFENISSA
a.boufenissa@grid.arn.dz
e-library Project Manager
DZ e-Science GRID
ARN, CERIST, Algeria

Aouaoueche EL-MAOUHAB
elmaouhab@arn.dz
Algerian Research Network Manager
DZ e-Sceince GRID Manager
ARN, CERIST, Algeria

PKP 2019 International Scholarly Publishing Conference

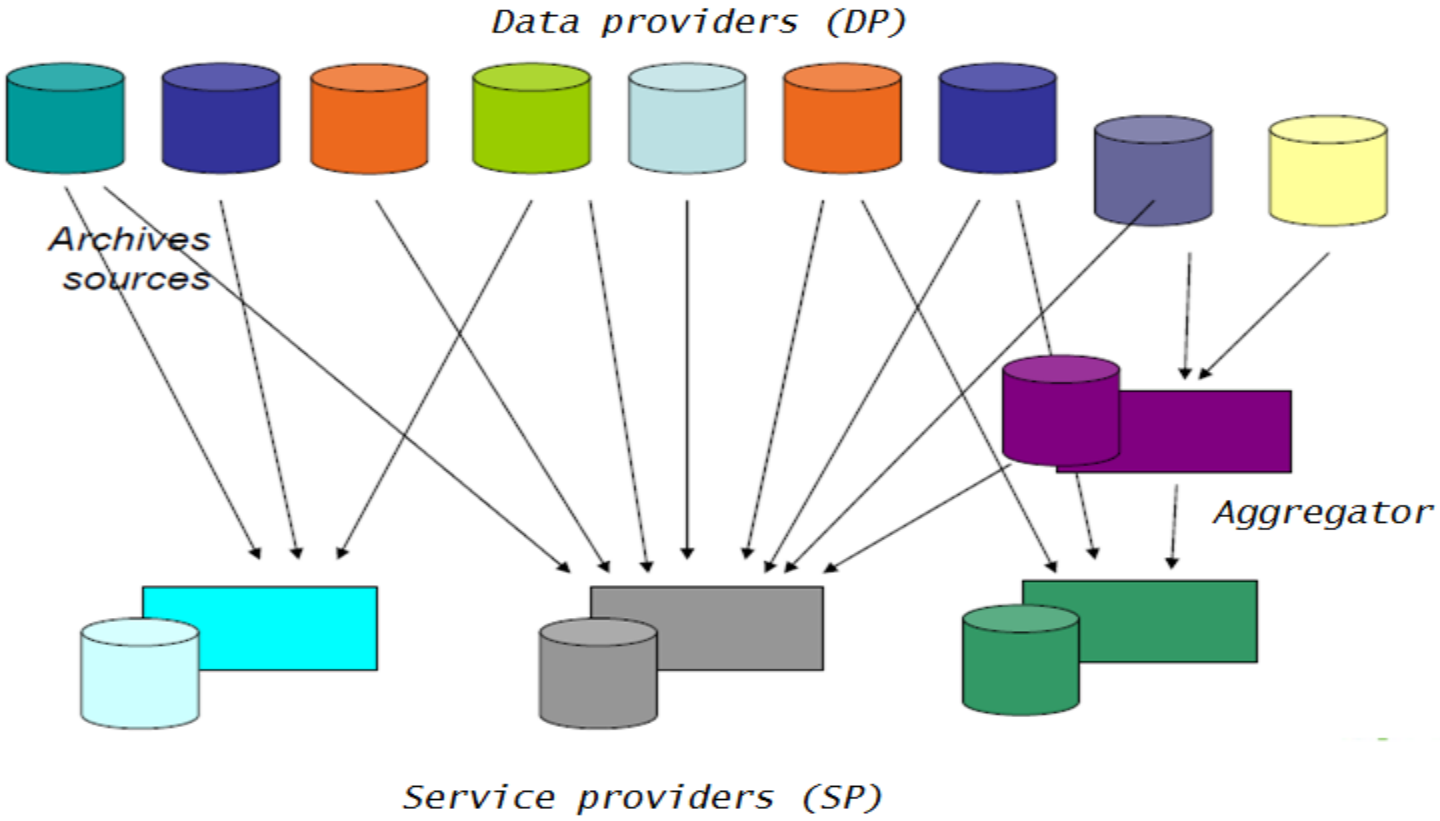
- **Digital Libraries** “DL” are organized collections of digital content
  - Aim to make cultural, audiovisual and scientific heritage accessible to all

- A key sector of DL is scientific information that represents scientific writings

- Birth of “open access movement" and emergence of open archives:
  - Promoting rapid and open access to the results of scientific research to global scientific community
  - Servers number of deposit and dissemination of the scientific production has increase.
- Multiplication of deposit servers induces the need to federate the search for information through the various member sites
- These different libraries are based on different technologies, platform, protocol and architecture depending on their different objectives and how they work.
- These factors create an imminent problem in the field of interoperations between these DLs !

- Several organizations that had an interest in interoperability between DLs, have come together to develop solutions that ensure interoperability:
  - Metadata collection technology, based on the OAI-PMH protocol
    - Most suitable solution
    - Simple and open protocol
    - Protocol based on HTTP and XML (Web standards)
    - Asynchronous polling of several bases
    - Easy to implement

OAI-PMH : Functional architecture



What's about its limits?!

A more efficient solution based on the use of new technology is to look for

Grid Computing

Computing Grid: Definition

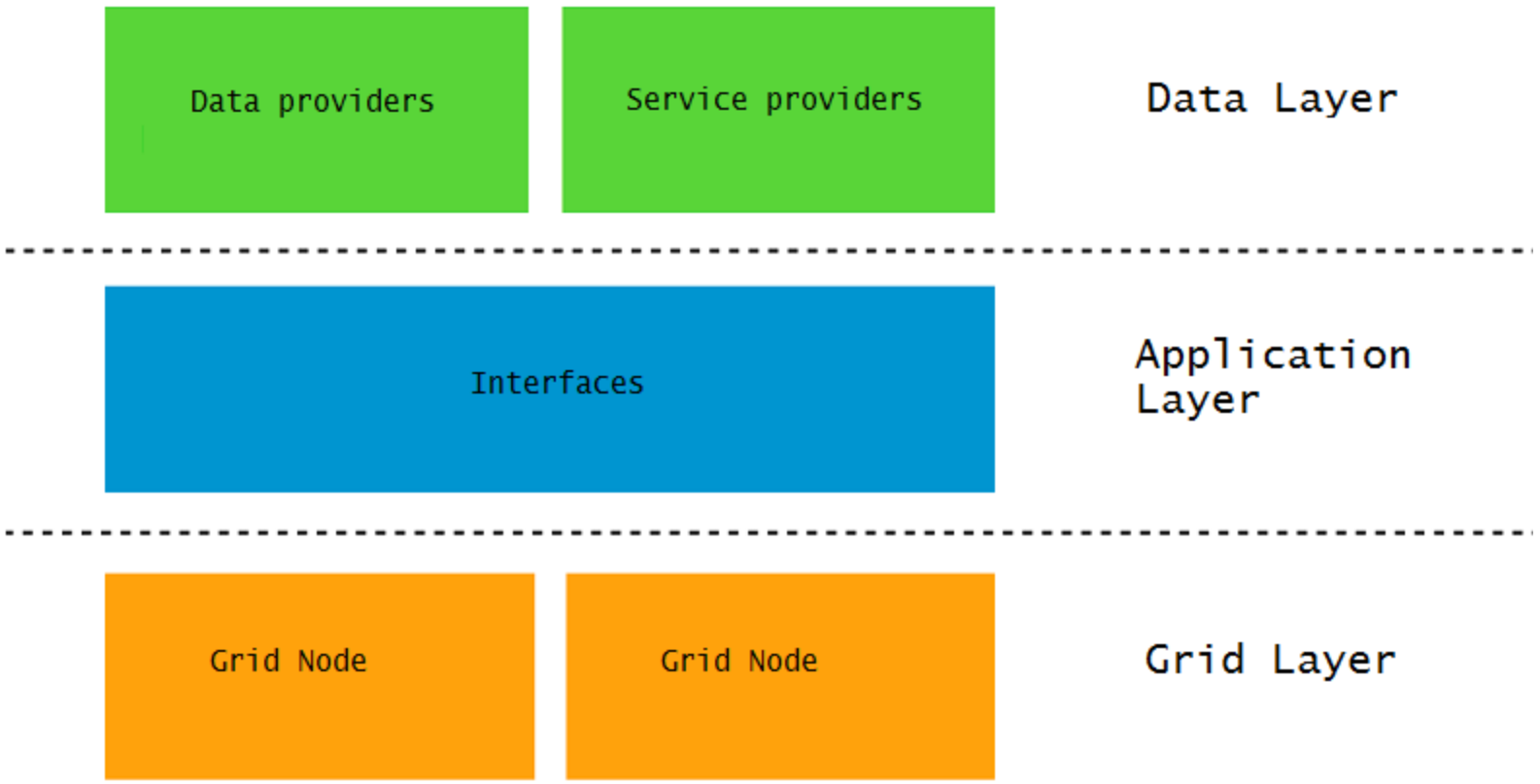
- Grid computing is a group of computers physically connected (over a network or with Internet) to perform a dedicated tasks together. Grids are a form of "super virtual computer" that solve a particular application.

- A computing grid is constructed with the help of grid middleware software that allows them to communicate. Middleware is used to translates one node information passed stored or processed information to another into a recognizable format. It is the form of "distributed computing".

Our objectives

- Offer a model of harvesting metadata based on the OAI-PMH protocol in a large scale environment using GRID architectures and with better performance than basic OAI-PMH model
  - ie: Implementing a solution that combines computing grid technology with that of the OAI-PMH
- Try to increase harvest performance by optimizing harvesting time and processing metadata

Layered architecture



Layered architecture

- Grid layer :
  - Responsible for the execution of metadata harvesting
  - Aims to improve the performance of the metadata harvesting process by enabling parallel harvesting of metadata that supports :
    - Different harvesting ways:
      - Parallel to a number of OAI URLs repositories ;
      - Parallel to all sets (or thematic) of a given OAI URL repository relative to a number of OAI URLs repositories ;
      - By combining the two first ways.
  - Mainly composed of:
    - Harvesting nodes
    - Harvest Planning Service of OAI-PMH compliant metadata

Layered architecture

- Grid layer
  - Harvesting nodes
    - Nodes of the grid whose number is not static
      - Computing Elements (CE) with their Compute Nodes (WN)
    - Concurrently run the metadata harvesting program for a number of OAI URLs repositories independently
  - Harvest Planning Service of OAI-PMH compliant metadata
    - Initiates the task of metadata harvesting
    - Keeps track of the harvesting task for the harvesting nodes
    - Dynamically allocates and distributes the harvesting load to the different harvesting nodes

Experimentation

- **Objectif** : Compare the sequential execution of the harvester on a single machine and the parallel execution using the grid

- **Uses** :
  - OAI URLs open access from the open archives initiative http://www.openarchives.org/Register/BrowseSites as inputs
  - PKP OHS Harvester as harvester program

- Performed a local run first and then on grid with several variants.

1st Test Bench

A sequential execution of harvesting OAI URLs repositories locally :

OAI Archive	OAI URL	Harvested records number	Harvest time per second	Rrecords number per second
Interuniversity Health Library (Paris)	http://web2.bium.univ-paris5.fr/oai/oai2.php	11670	6881	1.64
@rchiveSIC - ©HAL	http://archivesic.ccsd.cnrs.fr/oai/oai.php	1889	2095	0.88
Research Publications Base - Paris-Dauphine University	http://basepub.dauphine.fr/oai/request	10910	5867	1.81

Temps-total = 6881 +2095 + 5867 = 14843 secondes

2nd Test Bench

A parallel execution of harvesting OAI URLs repositories on the grid:

OAI Archive	OAI URL	Harvested records number	Harvest time per second	Rrecords number per second
Interuniversity Health Library (Paris)	http://web2.bium.univ-paris5.fr/oai/oai2.php	11670	4641	2.51
@rchiveSIC - ©HAL	http://archivesic.ccsd.cnrs.fr/oai/oai.php	1889	822	2.29
Research Publications Base - Paris-Dauphine University	http://basepub.dauphine.fr/oai/request	10910	3620	3.01

The execution was done on three free nodes for three tendered jobs

Temps-total = Max (4641 + 822 + 3620) = 4641 secondes

3rd Test Bench

A parallel execution of harvesting OAI URLs repositories by sets combined to sequential execution of harvesting OAI URLs repositories on the grid :

Archive OAI	URL OAI	Harvested records number	Total harvest time per sets per second	Sets Number
Interuniversity Health Library (Paris)	http://web2.bium.univ-paris5.fr/oai/oai2.php	11670	2647	100
@rchiveSIC - ©HAL	http://archivesic.ccsd.cnrs.fr/oai/oai.php	1889	800	31
Research Publications Base - Paris-Dauphine University	http://basepub.dauphine.fr/oai/request	10910	2220	30

The execution was done on 10 free nodes. The jobs tendered are 100 jobs, then 31 jobs, then 30 jobs sequentially.

Temps-total = 2647 +800 + 2220 = 5667 secondes

4th Test Bench

A parallel execution of harvesting OAI URLs repositories by sets combined to parallel execution of harvesting OAI URLs repositories on the grid :

OAI Archive	OAI URL	Harvested records number	Total harvest time per sets per second	Sets Number
Interuniversity Health Library (Paris)	http://web2.bium.univ-paris5.fr/oai/oai2.php	11670	7620	100
@rchiveSIC - ©HAL	http://archivesic.ccsd.cnrs.fr/oai/oai.php	1889	3720	31
Research Publications Base - Paris-Dauphine University	http://basepub.dauphine.fr/oai/request	10910	5760	30

The execution was done on 10 free nodes, and 161 jobs were tendered

Temps-total = Max (7620 + 3720 + 5760) = 7620 secondes

- Summary :

Total harvest time per second			
1st Test Bench	2nd Test Bench	3rd Test Bench	4th Test Bench
14843 secondes.	4641 secondes	5667	7620

- **Discussion**
  - Improved time to metadata harvest
  - Sequential harvesting combined to parallel harvesting by sets gives better results than sequential harvesting without sets
  - When the number of grid nodes approaches the number of jobs, the execution time will decrease significantly.
  - Distributing harvesting through several nodes is advantageous especially for archives with a large number of records which requiring a significant harvesting time

Synthesis

- We tried to prove that high performance for a harvesting service on a large number of DLs repositories can be achieved by using grid technology and parallel harvesting techniques.

- We successfully implemented a preliminary version which alows :
  - The parallelization by sets which gives better performances
  - Flexibility of the harvesting execution mode according to different cases that we had evaluate according to the available resources and the number of sets to treat.

THANK YOU !