P2P-based Grid Architecture for Homology Searching

Hiroshi Sunaga, Toshiyuki Oka, Kiyoshi Ueda, and Hiroaki Matsumura
NTT Network Service Systems Laboratories, NTT Corporation
Background

Grid Technology:
- a promising way of attaining high processing capabilities, providing large amount of storage, and providing various types of distributed services.
- application areas are very wide:

- high energy physics
- nuclear physics
- astronomy
- geoscience
- fluid dynamics
- manufacturing
- bioinformatics
- electronic design
- financial analysis
- entertainment ...
Background

- types of Grid:
  - computing grid = high-performance computation
  - data grid = large-scale storage
  - service grid = Web Services
  - PC grid = e.g., SETI@Home

- standardisation:
  - GG : Global Grid Forum
  - EG : Enterprise Grid Alliance
  - ...
Grid for Bioinformatics

Our focus:
- Bioinformatics field:
  joint trial with the National Institute of Genetics
  and DNA Data Bank of Japan

What?
- homology searching application to find new genes or DNAs

How?
- A sample is compared with DNA-base patterns in the database.
- Both computing and data management are important.
- Integration of computing grid and data grid is promising.
Homology searching

Major tool for Homology searching:

- Basic Local Alignment Search Tool (BLAST)

- **BLAST** is a tool to be used for comparing gene and protein sequences against those already known in DNA databases by performing sequence similarity searches. The purpose of homology searching is to estimate functions of a newly found gene. If homology is observed in part of the sampled gene, these genes are expected to have the same kind of function, thus leading to progress in the development of vaccines.

- A BLAST computation program is small enough to be run on PCs. -> appropriate to use the Grid framework to obtain a huge amount of computing power
Problems

[existing procedure]
(1) The requester asks PCs to execute the BLAST program.
(2) Each PC has access to the database to obtain necessary data.
(3) The PC starts the BLAST
★ Database is frequently accessed. -- More bandwidth.

The turn around time increases. -- DB access & computation
[proposed solution]

(A) Avoids DB access. Instead, the PC asks necessary data to other PCs (peers) in the network.

(B) Searches a PC that has already had necessary data.

(A) integration method

(B) selective method
Field trial network configuration:

1. Pre-distribution of known DNA data
2. Job request for executing a BLAST application
3. Homology-searching job
4. Answer back results

DNA Data Bank of Japan

NTT's optical and ADSL subscribers

Nippon Telegraph and Telephone Corporation

NTT Proprietary

Network configuration

DNA data newly found

proxy

already-known genic data

DNA Data Bank of Japan

service soft.

resource providers

requester

broadband network

Japan National Institute of Genetics

(researchers)

Internet

job result

discovery overlay network SIONet

DNA

data retrieval from another peer

Blast-app.

Blast-app.

Blast-app.

Blast-app.

Blast-app.
Discovery mechanism

Discovery of appropriate resource providers:
- SIONet, which has a metadata routeing capability, is used.

SIONet relays a packet to peers that have registered their properties (metadata).
Results

Compared with the ordinary method, the total turn around time is shorter when the integration method is used, and much shorter when the selective method is used.

In this trial, P2P data exchange is via RPs. In future IPv6 environments, direct data exchange between PCs will reduce the data transfer time by half.
When the number of clients (PCs) increases, the turn around time becomes shorter.
Analyses and Considerations

(A) integration method:
takes longer than (B), but
is suitable when a specified data holder is busy.
Data-grid-only or computing-only participants may exist.

(B) selective method:
is effective if there are many PCs that have necessary DNA
data and are ready for computation.

In other words, it is not effective unless there are enough
PCs satisfying such conditions.
-- Most frequently used data should be distributed in
advance to client PCs.
Conclusion

• An advanced Grid architecture was discussed for homology searching in the bioinformatics field.

• Metadata-based resource searching and P2P-based integration of computing-grid and data-grid were proposed.

• Field trial results showed effectiveness of the proposals.
  - The total processing time was reduced.