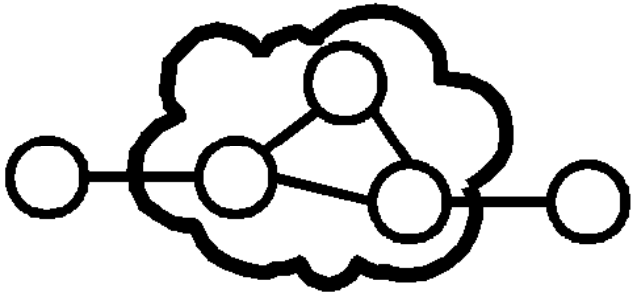


3PGCIC-2010

The Fifth International Conference on

P2P, Parallel, Grid, Cloud and Internet Computing

A fault-tolerant approach to storing objects in distributed systems

Ciprian Dobre, Florin Pop, Valentin Cristea

Emails: {ciprian.dobre, florin.pop, valentin.cristea}@cs.pub.ro

University "*Politehnica*" of Bucharest
Faculty of Automatic Control and Computer Science

Computer Science & Engineering Department

20.11.2010

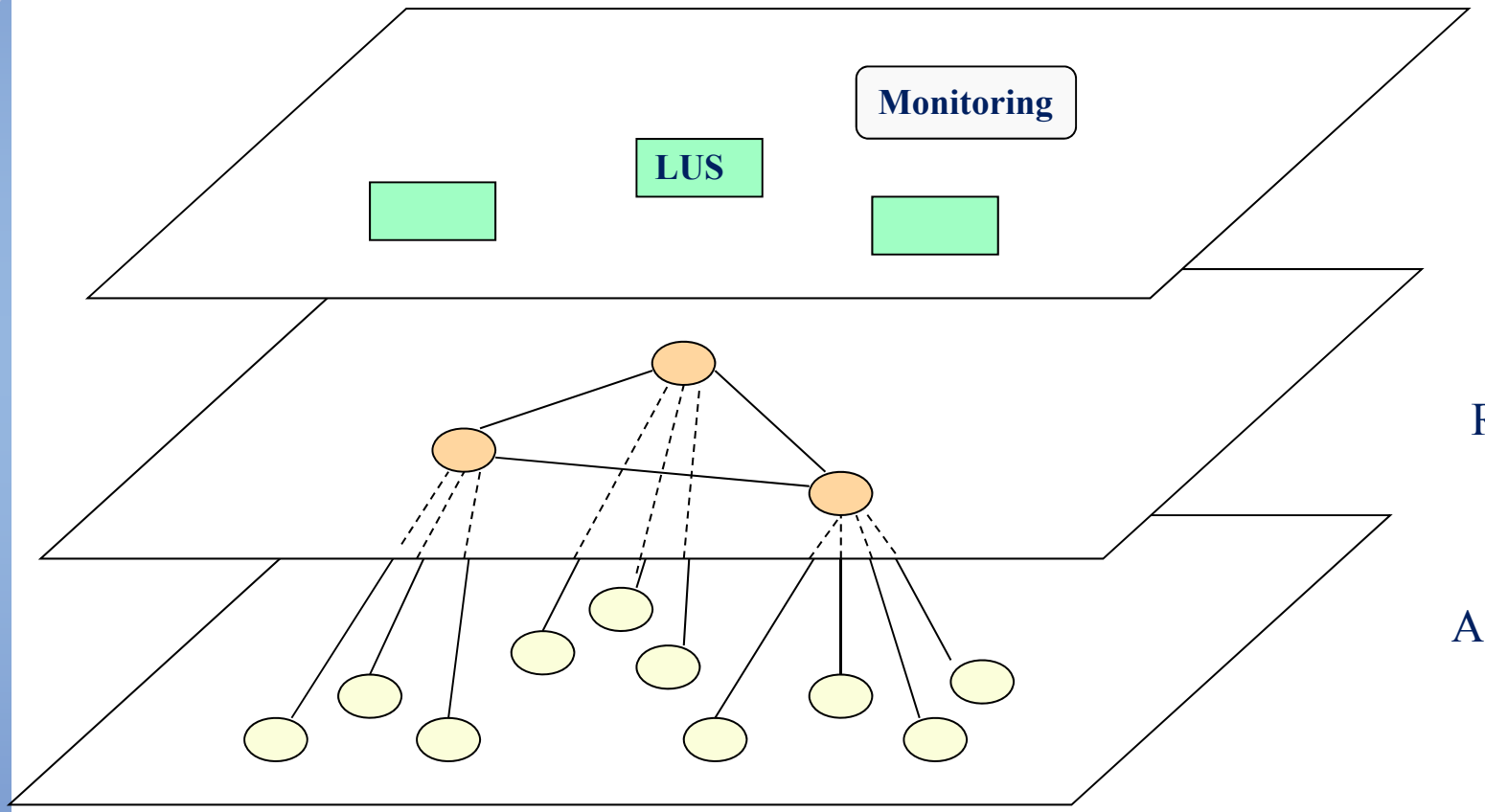
Outline

- Scope and motivation
- Architecture
- Data Management
- Fault Tolerance Management
- Conclusions

Scope and motivation

- Currently there has been much interest in P2P network overlays
- DHT-based systems support rapid development of a wide variety of Internet-scale applications
- We present an original DST-based P2P system, DistHash, to optimally share sets of distributed objects in highly dynamic large scale infrastructures.
- The system is based on the implementation of a DHT in which
  - a) peers do not equally participate in hosting published data objects;
  - b) peers can join or leave the network at any time, without prior knowledge;
  - c) the underlying network infrastructure can be different than the adopted communication scheme.

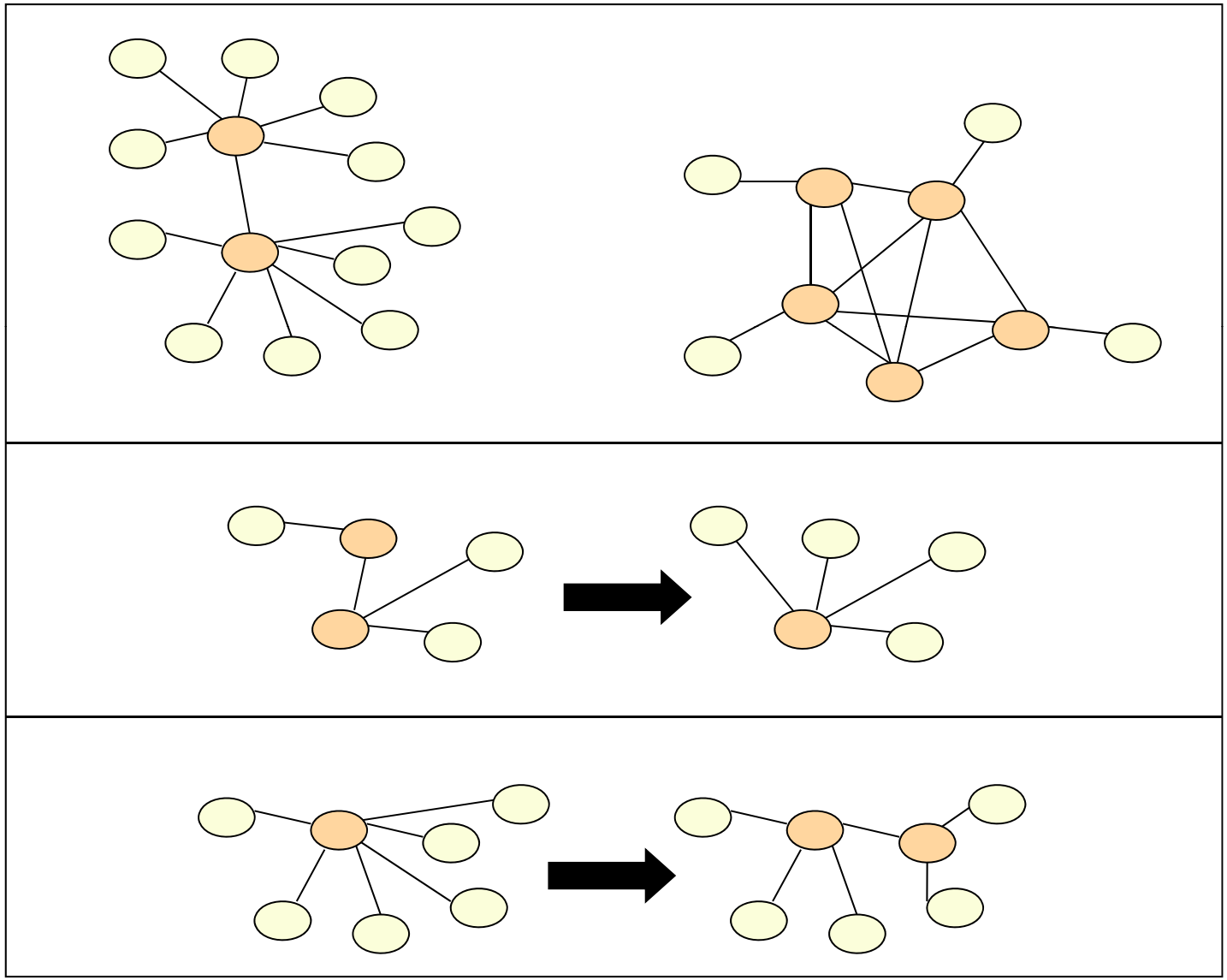
Architecture (1)



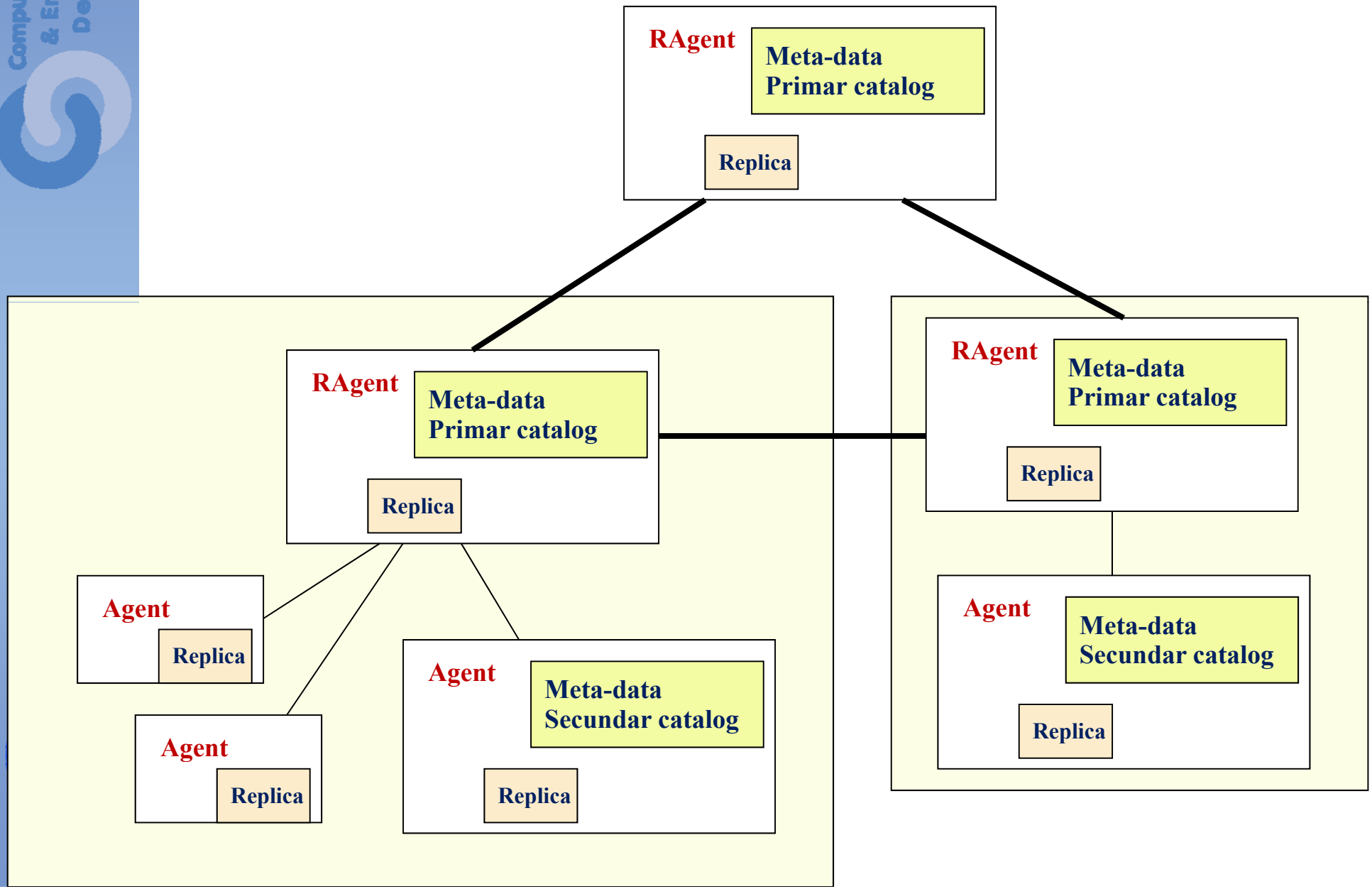
Architecture (2)

- The modality to connect the peers into clusters is based on the geographic position of each peer, as well as a set of environmental condition metrics.
- The RAgents connects several Agents (length of the bucket or bucketwidth).
- The determination of the number of Agents connected to a RAgent and of the number of RAgents is crucial to maintaining efficiency.
  - If the number of Agents is too great as compared to the number of RAgents, the meta-data catalogue will grow too big and the operations on it will require a longer time to complete.
  - On the other hand, if the number of Agents is too low as compared to the number of RAgents, the number of control messages required to access information from the system will be higher.

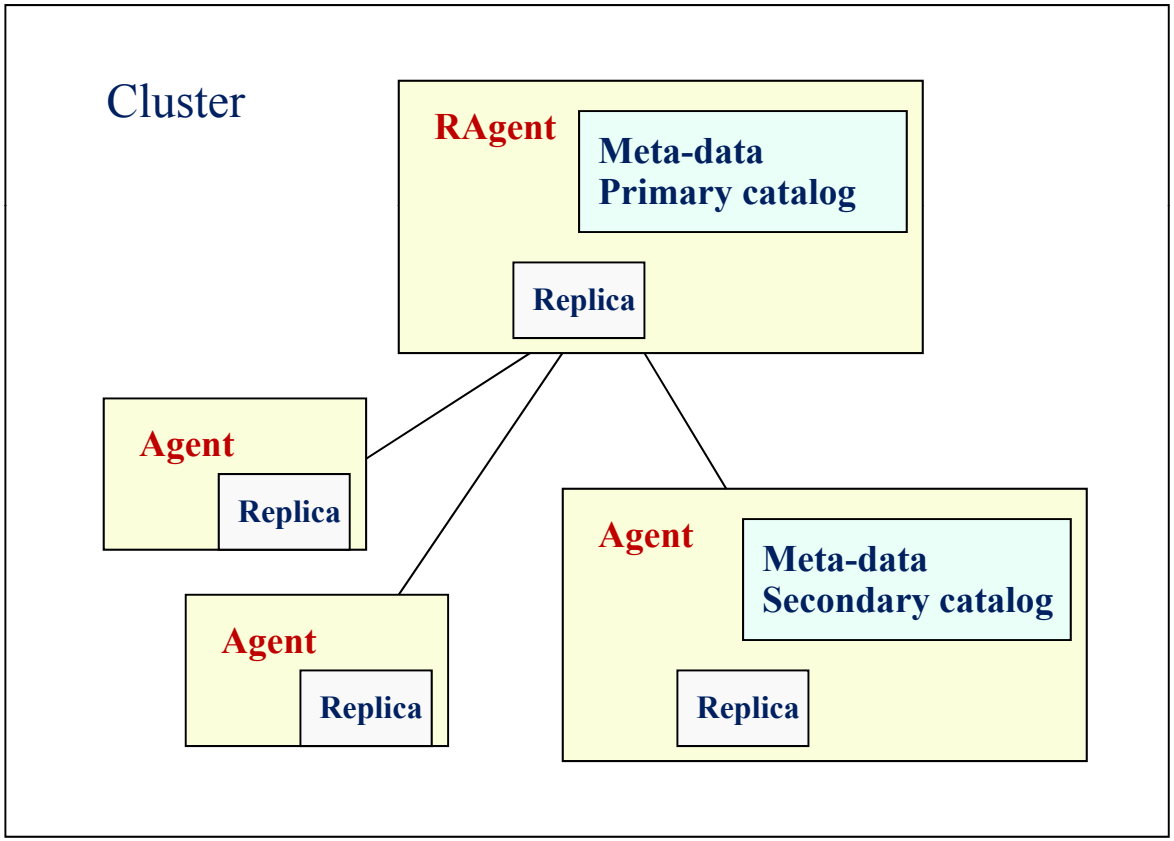
Clustering



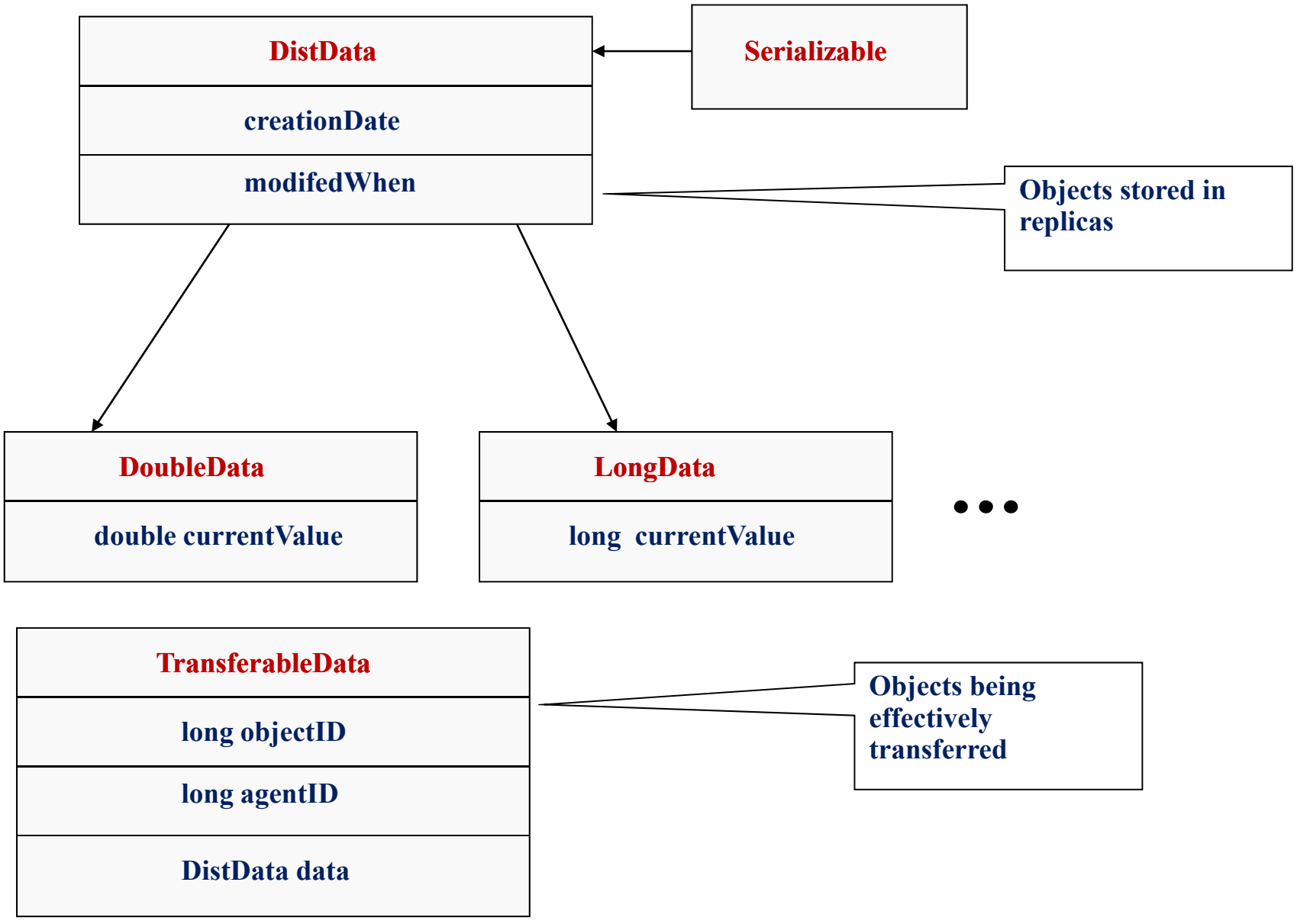
Agents


RAgent
Meta-data
Primar catalog
Replica
RAgent
Meta-data
Primar catalog
Replica
RAgent
Meta-data
Primar catalog
Replica
Agent
Replica
Agent
Meta-data
Secundar catalog
Replica
Agent
Replica
Agent
Meta-data
Secundar catalog
Replica

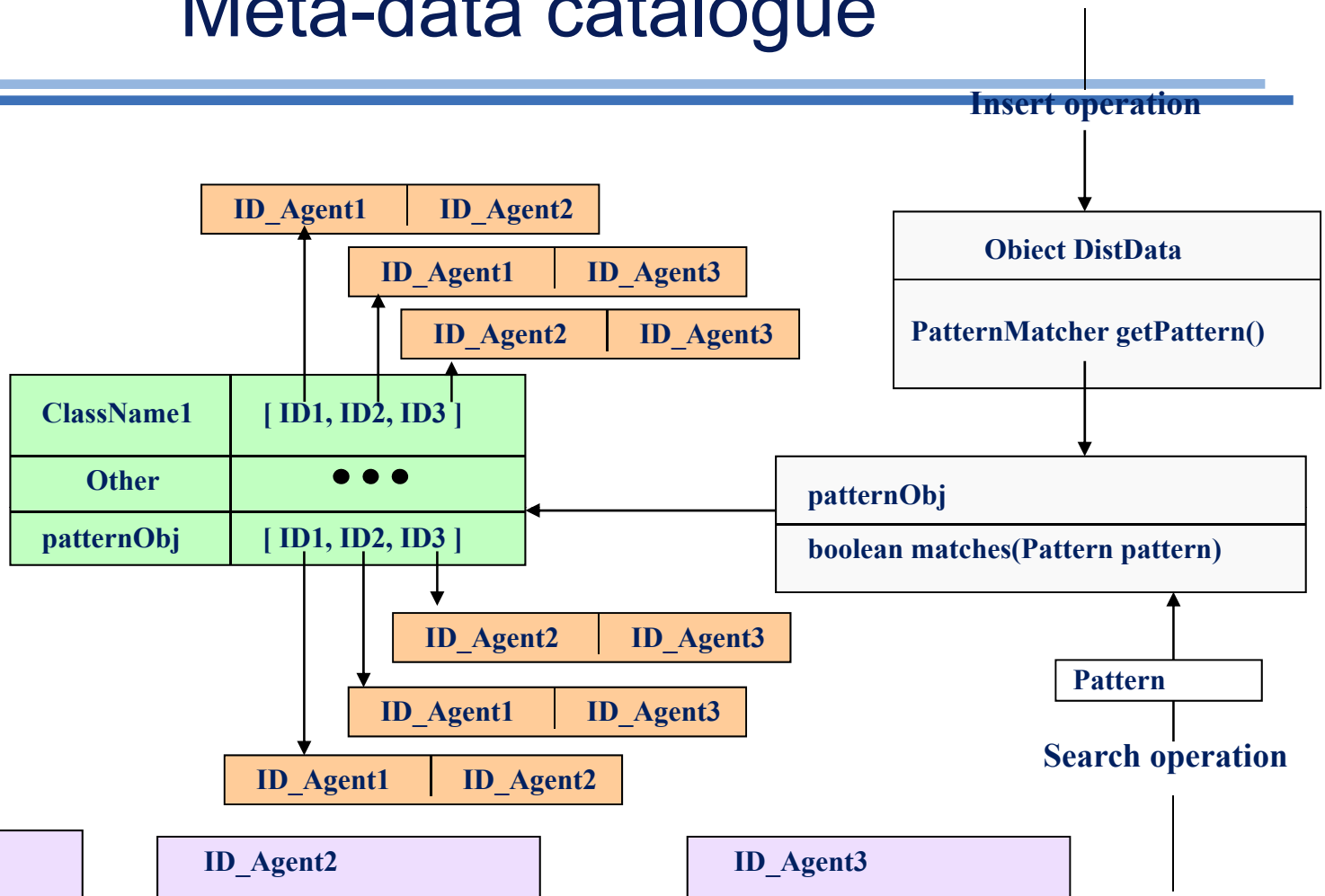
Roles of peers inside a cluster

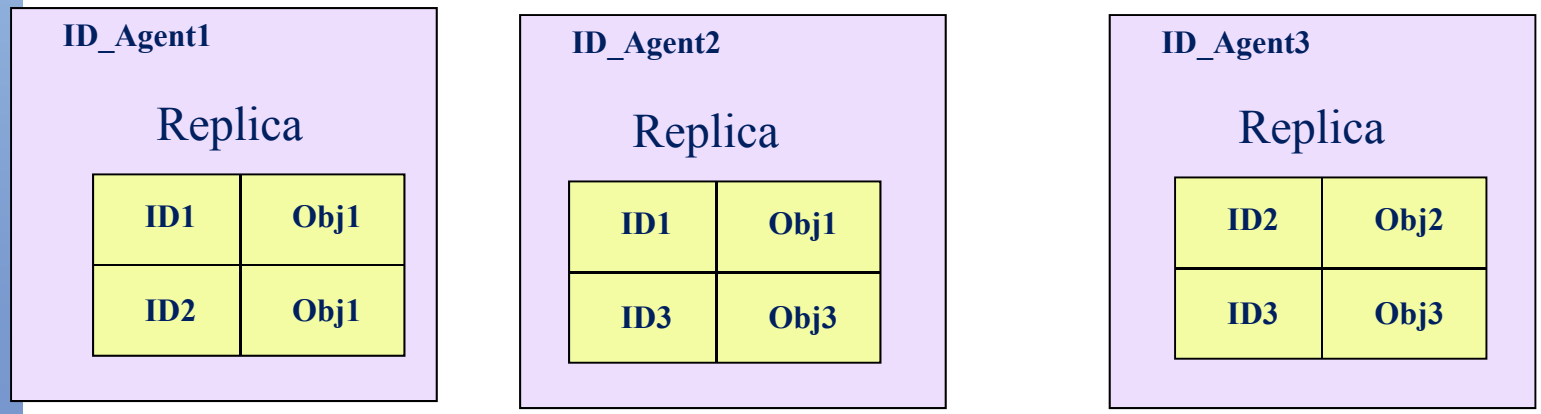


Distributed Objects

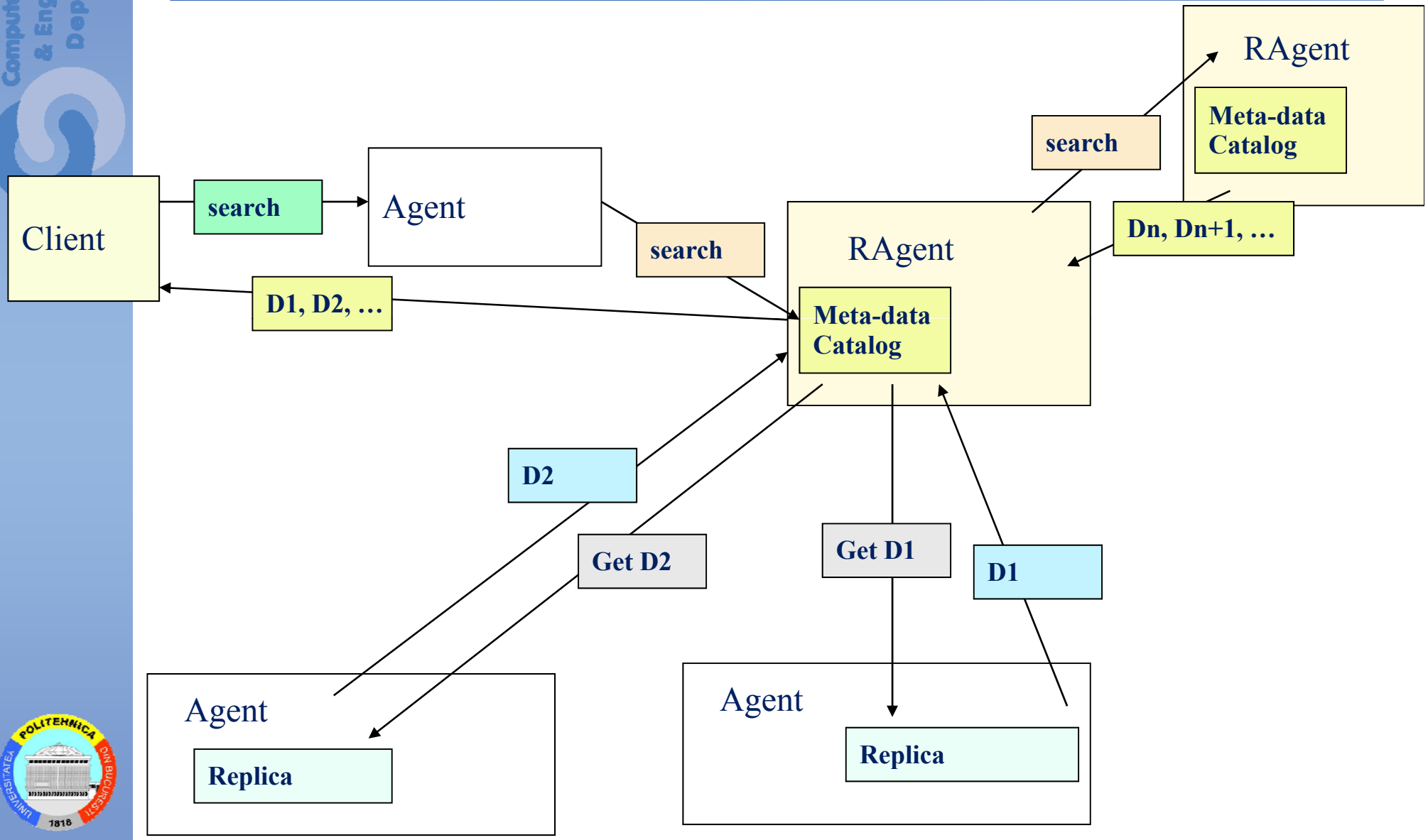


Meta-data catalogue

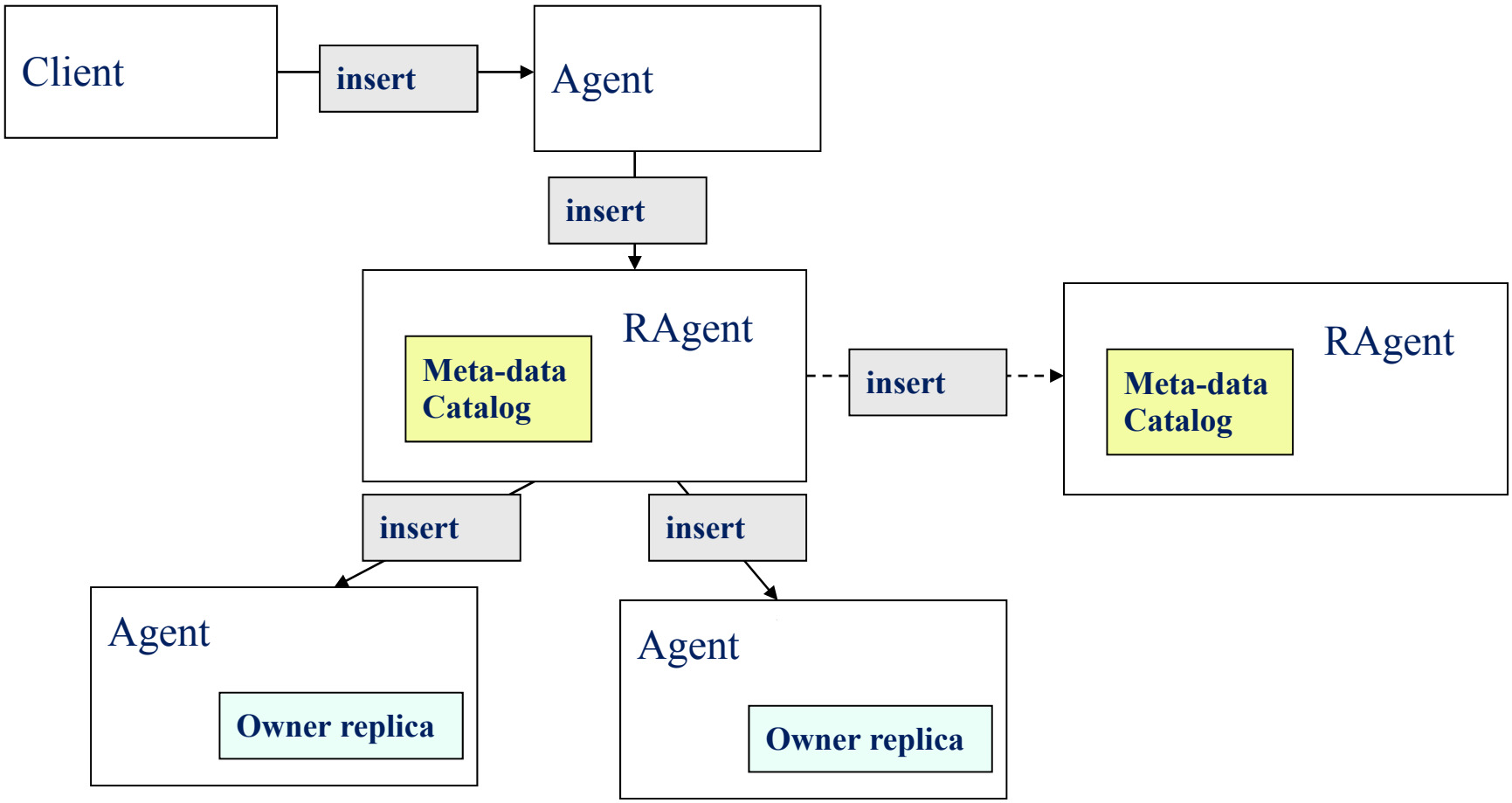




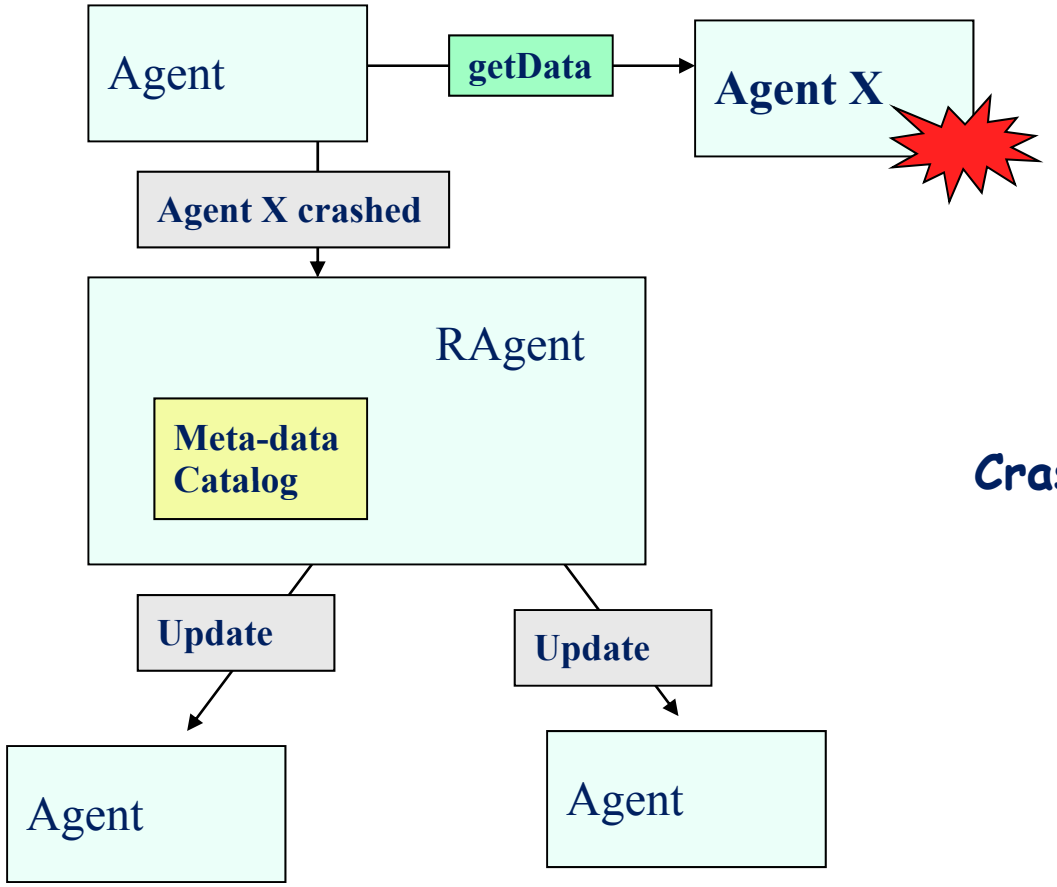
Search operation



Insert operation

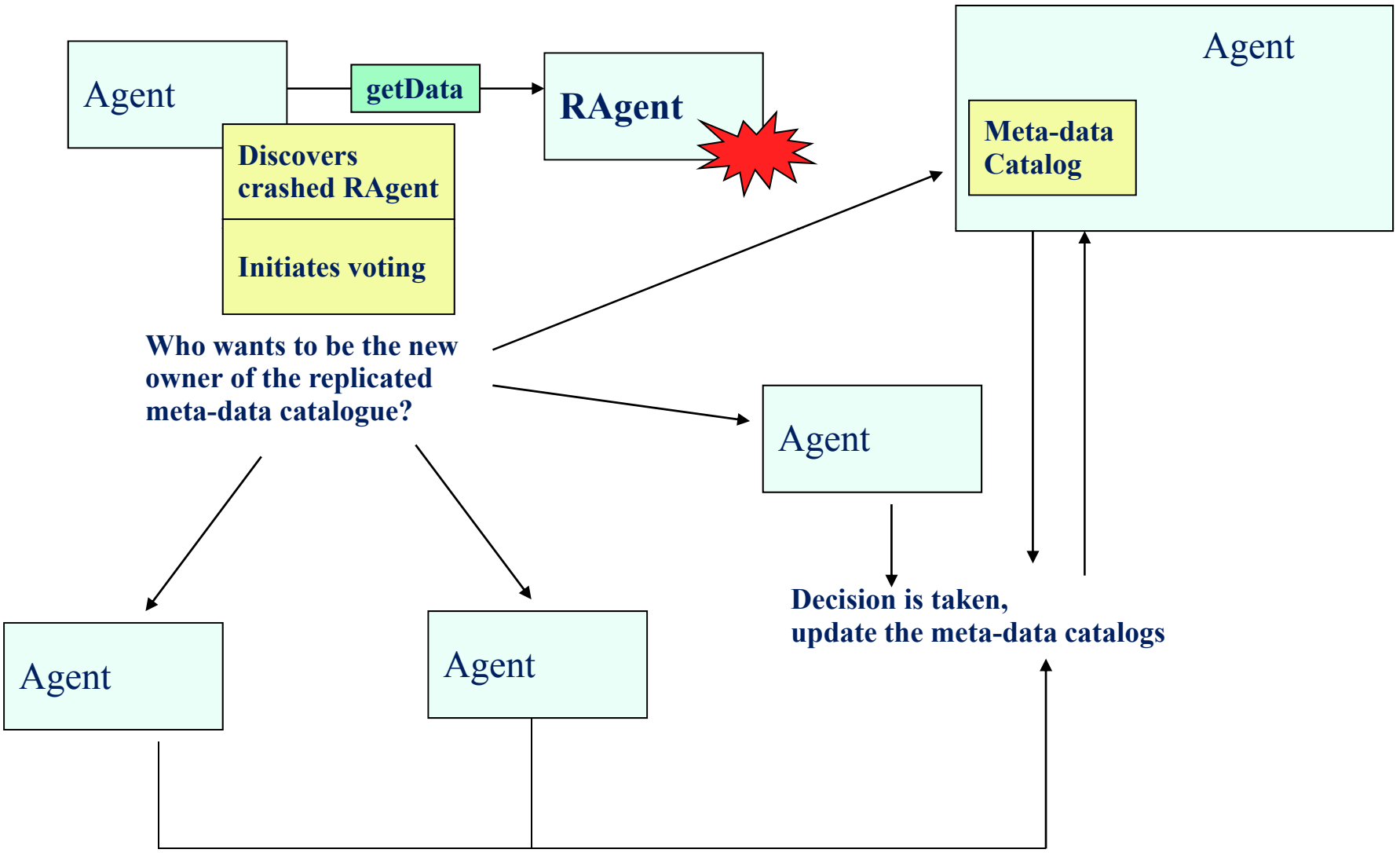


Fault tolerance (1)



Crash scenario

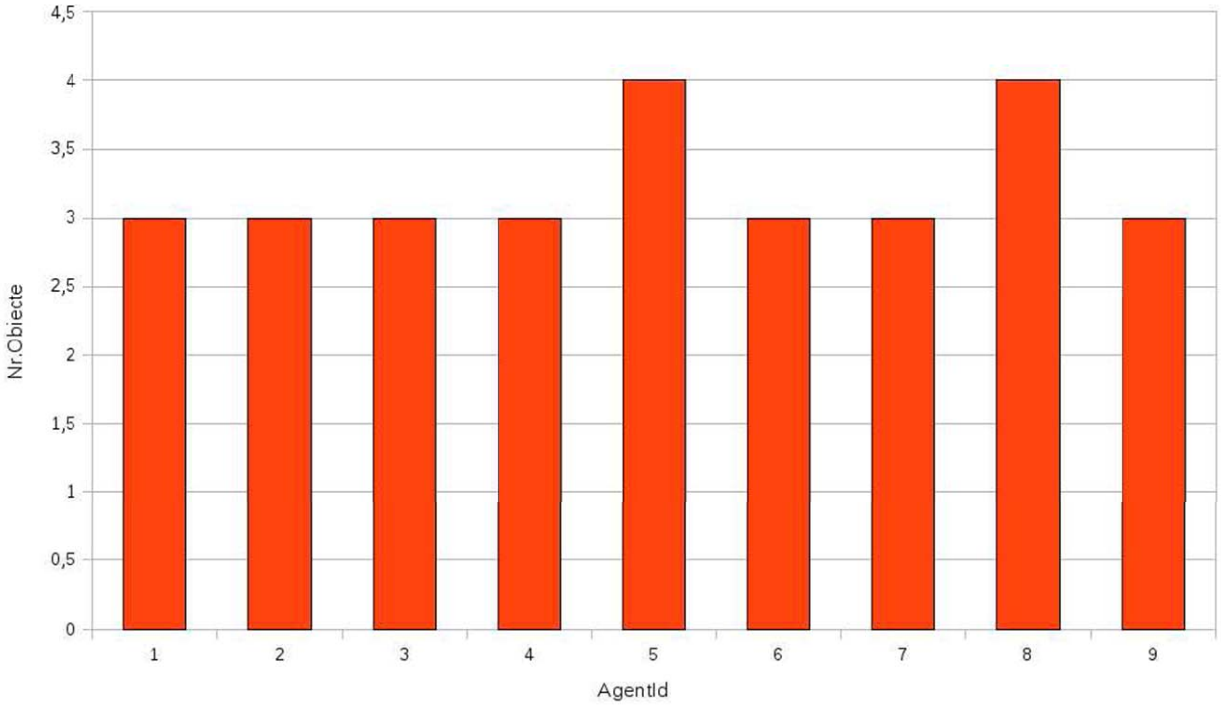
Fault tolerance (2)



Evaluation results

- A cluster of Intel Xeon E5405 quad-core stations, connected through Gigabit Ethernet.
- We used 9 Agents and inserted objects in various points in the systems.
- At some point we simulated a crash of different Agents in the system.

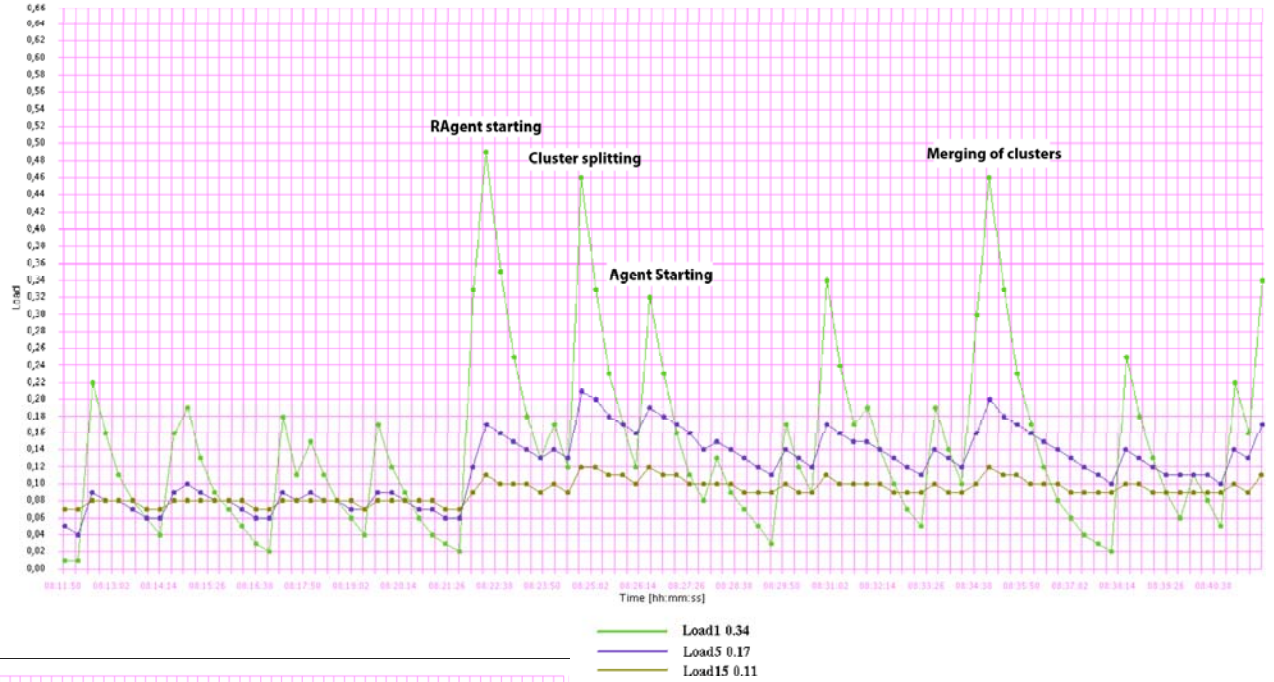
Evolution of the objects as a result of a crash

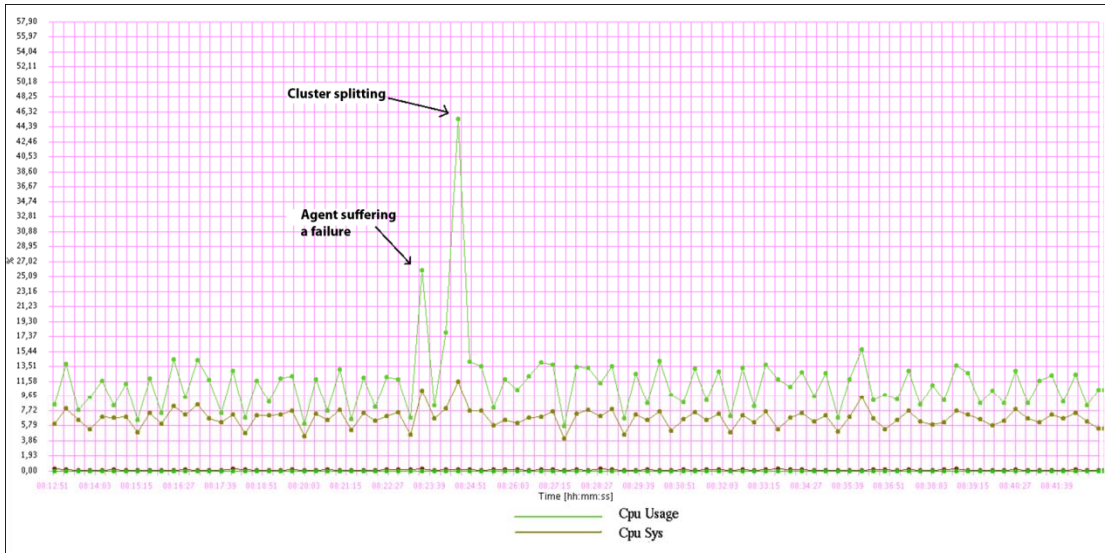


The system automatically replicate objects such that to preserve at least two copies of the same object.

Performance results

The load on a test machine

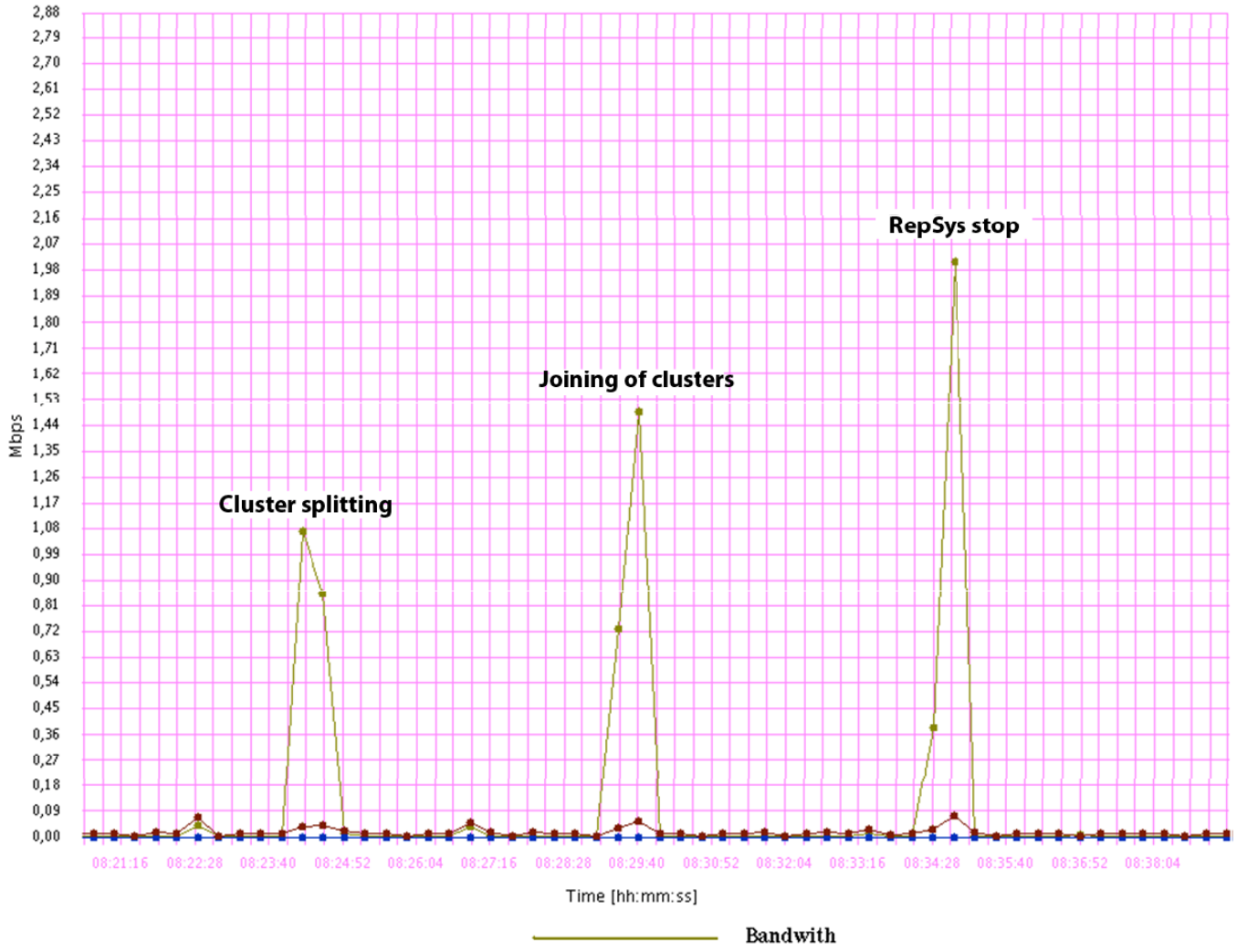




The CPU usage on a test machine

Performance evaluation

- The generated network traffic:



Conclusions

- DistHash is a P2P overlay network designed to share large sets of replicated distributed objects in the context of large-scale highly dynamic infrastructures.
- The system provides a middleware level functionality to users needing to use a shared memory concept system to be accessed by distributed application running on an Internet-like large-scale infrastructure.
- We achieve near-optimal message routing in hop-count and throughput, providing an adequate consistency approach among replicas, as well as provide a fault-tolerant substrate.
- In the next phase DistHash will be integrated together with a monitoring instrument in order to use the monitoring information to take decisions regarding the location of replicated objects in real-time

Acknowledgements

The research was supported by national project “DEPSYS – Models and Techniques for ensuring reliability, safety, availability and security of Large Scale Distributes Systems”, Project “CNCSIS-IDEI” ID: 1710.

The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

Q&A

Thank you!