

<p style="text-align: center;"><b>Campaign 2010</b> <b>Description of the PHD</b></p>
---

**Orange Labs Supervisor: Ruby Krishnaswamy**

**Supervisor email: ruby.krishnaswamy@orange-ftgroup.com**

**Location: Issy Les Moulineaux**

**PHD title: *Adaptive replica management in federated cloud storage systems***

### **Global context and state of the art**

On line (cloud) data storage as a service is rapidly gaining importance for different kinds and categories of data. While currently cloud storage services are offered by large-sized data centers, we foresee emergence of hybrid and massively distributed infrastructures where data centers are combined with consumer-end edge devices such as personal computers, Settop Boxes, owned by the cloud users themselves. One of the advantages in combining these different storage areas is to expand the storage capacities of the cloud provider's data center by exploiting such less-expensive edge resources. However this idea needs to be accompanied by smart data placement and replication techniques to overcome the dis-advantages of edge resources, i.e. volatility, heterogeneity and limited capacities.

Peer-to-Peer (P2P) system concepts and scalable functions are being incorporated into the domain of file systems to build scalable, self-managing (autonomic) and decentralized file systems and file sharing services. Every node in the system is equivalent and in general peer-to-peer systems assume a symmetric network of machines. Such storage systems are being used as backup systems, as file sharing systems and also content distribution systems.

The down side of exploiting edge storage nodes is that average reliability of these is low, either due to unreliable networks or due to node departures. Performance offered to clients may differ according to popularity of the data item requested or the bandwidth available and finally the assumption of perfect symmetry of peers is not realist. In all this, an important design issue is the method used for localisation of replicas and replica placement. As resources are distributed, an efficient mechanism is required for object location. Distributed Hash Tables are used as a mechanism to address the location problem and is based on hashing of content, name and other unique attributes. Files (and data) are placed on nodes that are close to the result of the hashing.

Replication is a technique employed to improving availability and performance on unreliable systems. However replication has a cost and many of the current systems proactively and automatically replicate stored objects without considering the popularity of files, the availability of nodes current storing a copy of a stored object. This is the case of systems such as PAST that automatically replicates files and tries to maintain a constant number of replicas for every file. Replication is initiated systematically when nodes fail or when new nodes join and selection of nodes replicating a file is done independent of criteria specific to the file. However all files may not be equivalently popular and all nodes may not have same level of availability. Similarly selection of a node where a new replica is to be placed should take into account node behaviour and capacity.

In the context of this thesis, the target storage system that we consider will be composed of storage nodes of different classes, expressed in terms of availability, robustness, capacity (storage and

servicing) and cost. This characterisation may be used in the design of replication and node selection algorithms.

## **PHD objectives / Expected results / Scientific challenges / Key Issues**

The objective of the thesis is to propose dynamic and adaptive replica management i.e., guide replica creation and replica placement so as to propose solutions to the followings questions: (a) Optimal number of replicas for a stored object and when replication should be triggered and (b) where to place a replica, i.e. the best target node to host a replica of a file. These two questions are also related to decision on which files should be replicated.

Strategies for replication decision and replication placement selection may be based on analytical models and implemented based on measuring and predicting peer availability (leaving and returning) and data popularity. Number of replicas required are adjusted to maintain optimal (across the set of stored objects), availability and performance levels.

We plan to propose dynamic and adaptive data placement and replication strategies on an infrastructure composed of storage nodes of different characteristics (availability, storing and serving capacity, cost ...). While not restricted to, the scope of this system is to address storing, sharing and delivering of personal content such as video, audio, images and personal data.

The approach could also envisage utility and/or market-based control mechanisms to provide mappings between resource distributions (storage space, network bandwidth, processing capacity for stored objects) and the quality of service agreements. Market-based mechanisms could be the basis for the autonomic behaviour that we require on the system.

## **Methodological approach proposed by the supervisor**

The first step is to study the state of the art of replication and placement algorithms, keeping in mind requirements of scalability and decentralization. While many large scale storage systems are based on using techniques of Distributed Hash Tables (DHT), limitations of these approaches are also identified (lack of locality, assumption of homogeneity). It will be necessary to model the architecture before proposing new tailored placement and replication protocols taking into account file/content properties popularity on the one hand and node characteristics such as stability and capacity.

First ideas may be evaluated through simulations, but then, a proof-of-concept (POC) prototype will be evaluated on real architectures, namely, Grid'5000 and/or PlanetLab.

This approach requires a highly motivated by research PhD student who should have a good knowledge of distributed systems and algorithms. He should have good theoretical and implementation skills. Good system building skills and knowledge of storage and file systems is an advantage.

## **Global schedule**

The thesis will

- Study the state of the art concerning data replication in large scale systems (P2P, cluster file systems).

- Propose an analytical framework for availability and performance driving the replication decision strategies and propose a control model (utility, market, fairness etc) that drives replication decisions.
- Architecture and infrastructure platform to design dynamic replica management including replica placement policies.
- Propose new dynamic and adaptive replication protocols taking into account data popularity, node stability and capacities.
- Implement a prototype and evaluate it on large-scale distributed platforms such as Grid'5000 or PlanetLab.

### **Additional contributions**

We are currently in the process of submitting a collaborative project (FUI-10 Odisea) to design and develop a large scale and federated storage system. We are also planning to submit a proposition in response to the Call 8 of the FP7 (to be initiated in 2011). The Call 8 is expected to focus on Cloud Computing and more particularly on data management solutions for Clouds. Should these propositions be accepted, this thesis work will be executed within the scope of these collaborative project(s).

The thesis will be directed by Prof. Pierre Sens of Université of Pierre and Marie Curie, Paris.