Scalable indexing for large-scale distributed storage systems

Résumé du projet de recherche (Langue 1)

The initial research problem to be solved in this PhD is the design and implementation of a highly scalable, specialized indexing and search system, focused on queries on metadata. The system will be implemented as an extension to Scality’s storage system. Scality’s distributed storage system, which stores petabytes of data and is frequently updated, poses significant challenges to the implementation of an indexing and search subsystem:

- The primary challenge posed to the design of the indexing subsystem is to enable fast queries on billions of objects. Also, the system should maintain a small index size relative to the data size, despite indexing petabytes of data. Furthermore, the index design should support queries on multiple data types, since object metadata contain text (user defined attributes), integers (file size) as well as more complex data types (access control lists).
- Track updates incrementally as they occur. Data updates should provide low latency. Updating the index must not become a bottleneck for the storage system itself.
- Geo-distributed index. The indexing subsystem needs to receive concurrent updates and queries from a large number of clients, located in different geographic locations, and remain available in the presence of network partitions.