Project

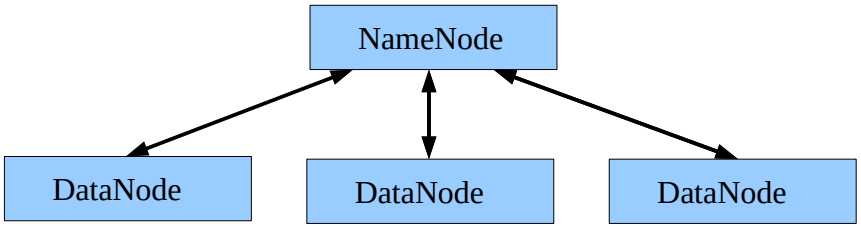
Daniel Hagimont - Daniel.Hagimont@enseeiht.fr

USTH – Teaching Unit MI 2.01

March 2021

Objectives

The objective of this project is to implement a tools which allows storing and downloading large files (such as video files) to and from a set of distributed servers.
This principle is to replicate each file to be stored in all servers, so that upon download, parts of the downloaded file can be downloaded in parallel.
The overall architecture of that tool is given below.



A NameNode is responsible for registering servers (DataNodes). The NameNode is started on one particular machine at startup. Then every DataNode registers at the NameNode, using an interface :

```
public interface Registration extends Remote {
        public void registerDataNode (Host h) throws RemoteException;
}
```

Host is a simple serializable class including 2 fields : host and port.

Each DataNode is then accepting TCP connections for receiving requests. A request may either send a file to be replicated on the DataNode or ask a part of a file for download.

When a client wants to store a file, he asks the NameNode the list of available DataNodes, using an interface :

```
public interface Consultation extends Remote {
        public Collection<Host> listDataNode () throws RemoteException;
}
```

Then the client can replicate the file on every DataNode (using TCP).

When a client wants to store or download a file, he uses the interface :

```
public interface ManageFile {
        public void store (String filename);
        public void download (String filename);
}
```

These methods are used by the client to store a file from the local file system onto the DataNodes (replicated on all DataNodes), and to download (in parallel) a file from the DataNodes to the local file system.

To implement these methods (store and download), it asks the NameNode the list of available DataNodes, and starts the store or download of the file. For the download, if the NameNode returned a list of 3 DataNodes, the client will download in parallel (with different threads) 1/3 of the file from each DataNode.

Expectations

You must implement this simple scheme.

You must write a report (less than 3 pages in total) describing :

- your achievements (less than 1 page)
- a tutorial for using you tool (less than 1/2 page). Your tool should be easy to use (I should be able to test it on my laptop (locally) in less than one minute).
- a performance evaluation of the benefit of parallel download (simply compare the download time with one DataNode, 2 DataNodes, 3 DataNodes). For the evaluation, one convenient environment is to run the NameNode and DataNodes in VM in AWS (t2.micro VMs with one vcpu) and to run the client on your laptop (benefiting from multiple cores). You can also compare this with a simple RCP or SCP copy. Of course, you must experiment with large file (at least 1Gb).

You must sent to me an email with the source code of your contributions and the report. The dead-line is April 2021, the 3rd (23:59 Hanoi time).