IUD – PROJ731 – Projets

Le but des projets de PROJ731 est de vous faire développer un projet informatique dans lequel vous allez manipuler des flux de données dans un système distribué. Vous devez choisir un sujet et travailler dessus sur les 3 séances de 4 heures.
Lors de la dernière séance, vous devrez montrer votre travail à votre enseignant. Un démo est la bienvenue !
Vous devez donc : (i) manipuler des données, (ii) entre plusieurs machines.
Vous avez le choix du language : Java, Python, C, C++… Ici machine peut s'entendre au sens virtuel : par exemple vous pouvez faire du Java RMI entre plusieurs JVM sur la même machine physique…
Vous avez le choix du niveau d'implémentation : des sockets réseau en C à Java RMI… En fonction du niveau d'implémentation nous serons plus ou moins exigeants sur le niveau de fonctionnalités / finitions (e.g., prise en compte des fautes, ajout dynamique de participants…etc).

Sujets proposés (non exhaustif, vous pouvez proposer vos idées, celles-ci doivent toutefois être validées):

- **Répartiteur de charge**
  Il s'agit de s'inspirer du répartiteur de charge vu en TD. Des clients effectuent des lectures et des écritures sur des fichiers via des serveurs. Comme dans le TP, l'accès aux serveurs se fait via une machine « aiguilleur » qui va répartir la charge entre les serveurs. Le niveau de difficulté peut se « régler » :
  - En faisant l'hypothèse que tous les serveurs on accès au même disque contenant les fichiers ou au contraire que chaque serveur possède une copie de chaque fichier.
  - En proposant l'ajout / retrait dynamique de serveurs sur le répartiteur de charge.
  - En optimisant pour que le retour des lectures ne passe pas par le répartiteur de charge.

  Vous pourrez comparer différents algorithmes d'équilibrage de charge. Tester que votre système interdit bien les écritures concurrentes sur un même fichier…

- **Compteur de mots à la map-reduce**

  Il s'agit ici de reproduire le comportement d'Hadoop sur l'exemple de comptage de mots vu en cours. Vous ne devez pas utiliser la plateforme Hadoop, mais concevoir une plateforme distribuée de comptage de mots basée sur map-reduce (à programmer avec des Sockets ou du RMI par exemple).

  En entrée : un ensemble de fichiers texte contenant des mots ; en sortie : un dictionnaire comportant l'ensemble des mots associés au compteur de leur nombre d'occurrences. Dans une première phase, une tache par fichier va compter les nombres d'occurrences des mots du fichier qu'elle a reçue (MAP), dans un deuxième temps un nombre fixé de taches « reduce » vont récupérer chez chacunes des taches map le sous ensemble des mots dont elle est « responsable » et comptabiliser pour chacun des ces mots, le nombre total d'occurrences.

  Le niveau de difficulté peut se « régler » :

  - En fixant ou non de nombreux paramètres comme : le nombre de taches maps, la répartition des mots sur les taches reduces
  - En prenant en compte ou non les fautes de taches map et réduce… (monitoring + relance / exécution spéclative…)
  - Selon le niveau d'implémentation (socket/rmi/mono-processus multithreadé…)

  Aide : utilisez un nœud coordinateur pour lancer les taches map et reduce.

  Pour tester votre plateforme vous pouvez prendre des pages web quelconques. Testez votre plateforme en variant le nombre de machines travaillant en parallèle et observez (mesurez) les effets.