

IFT630
Processus concurrents et parallélisme

Projet final
Système distribué d'exécution
"Crawler Web"

Par:
Jimmy Duchesne

Pour:
Gabriel Girard

Date:
12 avril 2008

Université de Sherbrooke
Faculté des Sciences

Table des matières

Description du sujet.....	3
Système d'exécution distribué.....	3
“Crawler ” Web.....	3
Éléments technologiques.....	4
Éléments reliés aux notions du cours.....	4
Description de l'implantation.....	5
Paquetage ift630project.base.....	5
Paquetage ift630project.base.server.....	5
Paquetage ift630project.base.client.....	5
Paquetage ift630project.crawler.....	5
Comment tester.....	6
Description de la démonstration vidéo.....	7

Description du sujet

Système d'exécution distribué

Le projet consiste en fait en deux applications. Il y a tout d'abord le système d'exécution distribué. En résumé, il s'agit d'une application client-serveur pouvant s'exécuter sur un seul ou sur plusieurs machines. Au départ, le serveur et les clients ne font rien par eux-mêmes. Ils sont simplement interconnectés et ils attendent qu'on leur demande d'exécuter une commande. Pour exécuter une commande, un client doit explicitement annoncer au serveur qu'il souhaite traiter des commandes. On peut donc dire que le rôle des clients est l'exécution et que celui du serveur est de distribuer la charge

“Crawler ” Web

Pour le système énoncé précédemment, une commande est en fait un interface. En d'autres mots, c'est un conteneur qui contient du code inconnu. C'est à ce moment qu'une autre application est nécessaire. Il s'agit de l'application utilisatrice du système. Elle doit encapsuler le traitement qui peut être distribué à l'intérieur d'un objet implémentant l'interface de commande. Dans le cas particulier du “Crawler” Web, le traitement présent dans une commande consiste à visiter une page Web et récolter tous les liens s'y trouvant. Les clients retournent le résultat sur le serveur et l'application utilisatrice du système va récupérer le résultat sur ce même serveur.

Il faut noter que le visiteur de page a été limité à une petite quantité de commande, car le processus pourrait prendre un temps illimité.

Éléments technologiques

- Toutes les parties de l'application sont consoles seulement
- Le langage utilisé a été Java
- Le système a été distribué grâce à Java RMI
- L'accès aux pages Web a été fait grâce au paquetage java.net
- Aucune librairie externe n'a été nécessaire

Éléments reliés aux notions du cours

- Les trois parties du système peuvent être démarrées sur différents ordinateurs
- Chacun des éléments du système est son propre processus
- Le parcours du Web est multifils
 - Un nouveau fil est démarré sur l'utilisateur pour chacune des commandes envoyées au serveur
 - Lorsqu'un client s'enregistre comme travailleur, un nouveau fil est créé pour attendre une commande
 - Un nouveau fil est démarré sur le client lorsqu'il reçoit une commande à exécuter
- Le serveur utilise les fonctionnalités de moniteur afin de gérer les différents appels de l'utilisateur et des clients

Description de l'implantation

Paquetage ift630project.base

- Command – La classe de base représentant une commande
- Grid – L'interface que l'application utilisatrice utilise pour interagir avec le serveur
- IOManager – Classe de base pour gérer les entrées sorties
- Server – L'interface que les clients utilisent pour interagir avec le serveur

Paquetage ift630project.base.server

- ServerImpl – Implantation du serveur
- ServerIO – Classe gérant les entrées sorties pour la console du serveur
- Start – Classe permettant de lancer l'exécution du serveur

Paquetage ift630project.base.client

- ClientImpl – Implantation du client
- ClientIO – Classe gérant les entrées sorties pour la console d'un client
- Start – Classe permettant de lancer l'exécution d'un client

Paquetage ift630project.crawler

- Crawler – Implantation d'une commande pour visiter les pages Web
- CrawlerIO – Classe gérant les entrées sorties pour la console de l'application utilisatrice
- CrawlerManager – Classe implantant l'application utilisatrice
- Start – Classe permettant de lancer l'exécution de l'application utilisatrice
- WebpageRipper – Classe contenant des méthodes pour interagir avec les pages Web

Comment tester

Adapter les valeurs à votre système.

- Compiler
 - `javac -cp src -d bin src\ift630project\crawler*.java src\ift630project\base*.java src\ift630project\base\client*.java src\ift630project\base\server*.java`
- Lancer le registre RMI
 - Il s'agit d'exécuter `rmiregistry.exe` situé dans le dossier bin de votre jre
- Lancer le serveur
 - `java -cp bin -Djava.rmi.server.codebase=file:/E:/workspace/ift630Project/bin/ift630project.base.server.Start`
- Lancer les clients
 - `java -cp bin ift630project.base.client.Start 192.168.0.104`
 - Exécuter la commande "SubscribeAsWorker"
- Lancer le processus utilisateur
 - `java -cp bin ift630project.crawler.Start 192.168.0.104`
 - Exécuter la commande "Crawl <http://www.google.com>"
 - Exécuter la commande "PrintFoundLinks links.txt"

Pour chacune des applications lancées, appuyer sur entrer va afficher les commandes possibles.

Description de la démonstration vidéo

Ce qui suit est simplement une description de ce qui est montré dans la vidéo de façon chronologique.

- Démarrage du serveur local
- Démarrage d'un client local
- Démarrage du “crawler” local
- Démarrage d'un client à distance
- Inscription du client à distance en tant que travailleur
- Lancement de la commande “Crawl <http://www.digg.com>” sur le “crawler” local
 - À ce moment, des commandes sont envoyées au client à distance pour visiter les pages Web
- Affichage des pages trouvées à deux reprises pour montrer que ça peut prendre un certain temps pour visiter le Web
- Écriture des liens trouvés dans le fichier digg2.txt
- Inscription du client local en tant que travailleur
- Affichage du nombre de clients
 - Comprends même les clients qui ne sont pas enregistrés
 - Cette valeur est récupérée du serveur
- Affichage du nombre de commandes exécutées sur le client à distance
 - Le nombre de commandes montre que c'est bien le client à distance qui a fait le travail
 - Le pic dans le graphique montre que l'accès réseau a été sollicité
- Lancement de la commande “Crawl <http://www.google.com>” sur le “crawler” local
- Affichage du nombre de pages trouvées
 - Comprends celle trouvée pendant la recherche précédente
- Écriture des liens trouvés dans le fichier both2.txt
- Affichage du nombre de commandes exécutées sur le client local
- Fermeture du client local
- Affichage du nombre de commandes exécutées sur le client à distance
 - Les deux nombres affichés montrent que la dernière recherche a été partagée sur les deux clients
- Fermeture du client à distance
- Affichage du nombre de clients détectés.
 - Démontre que l'inventaire de clients sur le serveur a été mise à jour
- Fermeture du “crawler”
- Fermeture du serveur