Mobilité, Cours Master 2, 2018, Introduction

Michel Habib
habib@irif.fr
http://www.irif.fr/~habib

Janvier 2018

Plan

Introduction

Plan du cours

Pour ou contre le parallélisme

Organisation du cours : Mobilité

Cours : Michel Habib habib@irif.fr
Contrôle = 50 % (projet + exposé) + 50 % CT
Projet par binômes = veille scientifique sur un sujet d'actualité (comprendre et donner un avis sur un point de vue technique, juridique)
exposé 20 mn + rapport.

Introduction

Plan du cours

Pour ou contre le parallélisme

Objectifs du module

- L'objectif principal de ce module est de présenter les évolutions actuelles en matière de mobilité (on parle d'ubiquité) et des techniques informatiques nécessaires à la mise en œuvre de cette ubiquité généralisée. En effet nous assistons à la convergence des appareils actuels téléphones portables, agendas électroniques, ordinateurs portables, voire des consoles de jeux, des appareils photos numériques ou encore des vidéos.
- Dans une première partie nous nous s'intéresserons à l'analyse de la puissance de calcul du parallélisme massif et de ses applications, puis nous étudierons les systèmes pair-à-pair en considérant les mécanismes (protocoles, routages, algorithmes) que leur développement nécessitent.


...Merci d'être venus
me rendre visite

Ce cours est complémentaire du cours Algorithmique Répartie et nous considérons que son programme est connu, en outre ce cours a des liens naturels avec le cours MAIN (Méthodes et Algorithmes pour l'accès à l'Information Numérique) et complète le module GRI (Grands Réseaux d'Interaction).

Introduction

Plan du cours

Pour ou contre le parallélisme

Plan du cours 2017

- Introduction
- Les modèles du parallélisme. Le modèle PRAM.
  Parallèle versus distribué.
  Parallélisme massif. Compilation pour machines parallèles.
  Langages pour machines parallèles. Machines vectorielles.
  Fermes de PC.
  PVM, MPI deux bibliothèques disponibles.
  Ordonnancements des calculs

- La plateforme Hadoop, le modèle de calcul Map - Reduce versus l'architecture SPARC.
- Hélas nous n'avons pas encore trouvé comment : concevoir, programmer, prouver et débugger des programmes parallèles malgré une intense recherche ces 20 dernières années.

Différences entre la communication et la réalité

- Le projet GRID. Public
  Un projet issu mené par les physiciens pour le CERN. Enjeux économiques et réalisations actuelles. Exécutions à distance, données réparties, autorisations (sécurité).
  Problèmes algorithmiques à résoudre : ordonnancements des tâches à la volée, localisation des codes.
  Les raisons de l'échec relatif de ce projet de ce projet.
- Cloud computing. Privé
  Problèmes de droit (la loi qui s'applique est celle du pays qui contient l'entrepot de données, Patriot Act ...).

- Autostabilité (algorithmes autostables)
- Réseaux ad hoc.
  Réseaux MANET (Mobile Ad-Hoc Networks)
  Discussion des modèles existants à base de graphes ;
  Affectation des fréquences.
  Exemple de protocole de routage, le protocole OLSR (Optimal Link State Routing Protocol).
  Synchronisation des horloges dans un réseau de téléphonie mobile.
  Exemples d'applications basées sur des réseaux de mobiles.
- Réseaux Pair-à-Pair
  Tables de hachage distribuées
  Problématique, protocoles existants, par exemple BitTorrente.
  Réseaux virtuels, routages vers celui qui possède la donnée.
  Structuration des réseaux virtuels.
- LT codes

Les outils de base

- Les protocoles.
- Les réseaux d'interconnexion qui interviennent à tous les niveaux (réseaux de tri, architecture parallèles, systèmes Pair-à-Pair ...)
- L'aléatoire. Ce cours permet de souligner l'apport des algorithmes aléatoires (randomisés ou probabilistes) pour résoudre efficacement des problèmes difficiles (ex : systèmes Pair-à-Pair, LT-codes, BitTorrente ...).

L'importance de l'algorithmique probabiliste

- Hasard viens de l'arabe "al-zahr" qui signifiait au début "jeux de dés". Puis il a pris le sens hasardeux (dangereux), ce qui a donné "hazardous" en anglais.
- Chance viens du latin cadencia qui signifiait le résultat d'un lancer de dés.

Introduction

Plan du cours

Pour ou contre le parallélisme

Calculs distribués versus parallélisme massif

Dans les deux domaines des calculs sont faits en parallèle, mais les objectifs (et donc les résultats) sont différents.

Parallélisme massif

Parallélisme On veut utiliser plusieurs machines pour calculer plus vite. Application type : analyse de millions de messages.
Priorité : Efficacité (i.e., temps d'exécution), éviter la perte en communications entre processeurs.
La validité des calculs est un objectif secondaire.

Calculs distribués

Distribué L'application est délocalisée par définition (terminaux bancaires ATM). Le parallélisme du calcul est imposé par l'application.
Priorité : Vérifier la cohérence des transactions (sureté des programmes) même en cas de pannes.
La rapidité des transactions est un objectif secondaire.

Premières applications du parallélisme

Lorsque la tâche à réaliser est bien partitionnable en sous-tâches identiques avec synchronisation limitée :

- Factorisation de nombres premiers et cryptographie
- Programme Sethi (analyse d'images, traitement du signal)
  sethi@home (1999)
  2 à 3 million de machines puissance de calcul en teraflop.
- folding@home
  repliement de protéïnes (playstation, cartes graphiques GPU)
- Gestion des données sur le WEB.
- Grands calculs pour physiciens
- Surveillance généralisée (NSA ...)

Nécessité du parallélisme

1. Les limites annoncées de la loi de Moore sur les circuits (limites physiques des composants). Nous arrivons aux limites des effets quantiques.
2. PC actuels sont sur des architecture multicœurs (on parle d'une centaine de cœurs bientôt sur une carte processeur de PC !)
3. Existence des GPU (Graphic Processeur Unit)
4. Application de type "cloud". L'offre de calcul d'Amazon.

1. Pour utiliser la pleine puissance de nos PCs (voire de nos téléphones portables) il sera nécessaire de paralléliser les applications.
2. Certaines applications liées à la mobilié sont distribuées. Routages distribués (RIP)
3. Utiliser la pleine puissance de machine en réseaux (stade de foot)

Conséquence de la parallélisation

Une meilleure compréhension de la localité d'un calcul. Cela peut amener des algorithmes séquentiels plus efficaces.
C.A.R. Hoare appelle cela : le paradigme du parallélisme.
Exemple du calcul de LCA en $O(1)$.

Le problème de la localisation des données

Même s'il a été fait des progrès énormes sur la puissance des processeurs et sur la taille des mémoires,
le transfert des données (même la simple lecture des données) reste un goulot d'étranglement.
80MB par seconde, cela donne 1Tera en 3,4 Heures.
Très lent au vu des bases de données existantes.
D'où l'émergence de la programmation avec des systèmes de type Hadoop, utilisant les fonctions Map et Reduce. Dans ce cadre les données sont statiques.

Exemples de questions de cours à l'examen écrit

1. Quelle différence(s) faites-vous entre programmation distribuée et parallélisme massif ?
2. Quelles sont les diffférences d'approches sur le problème de la synchronisation des horloges entre un physicien et un informaticien ?
3. Quel modèle de machine PRAM vous semble le plus réaliste (i.e. en fonction d'une implémentation sur une machine réelle) ?
4. Quelles caractéristiques ajouteriez-vous à ce modèle de machine PRAM, afin d'en faire un modèle vraiment réaliste ?
5. Quelles caractéristiques doit présenter une application distribuée pour être qualifiée de "Système pair-à-pair" ?

Exposé sur un sujet de veille scientifique

Protocole du projet

Il s'agit d'étudier les aspects techniques informatiques liés à ce cours concernant une question, un protocole, un logiciel voire d'un texte de loi.

La présentation du travail se fera à l'aide d'un rapport écrit et d'une soutenance (15 mn suivi d'une discussion), l'un des sujets de la liste jointe.

On s'intéressera tout particulièrement aux principes de fonctionnement de protocoles ou logiciels. La présentation portera sur une analyse critique. La notion de simulation d'un protocole devra aussi être envisagée lorsque c'est possible.

Le principe du projet correspond à celui d'une veille technologique en entreprise sur des questions d'actualité.

Bien sûr si vous tenez à étudier un sujet qui n'est pas dans la liste mais de même nature, il suffit de m'envoyer le sujet par mail, afin que je le valide

Liste des projets 2017

- Les principes des monnaies électroniques de type Bitcoin, issues du pair-à-pair. Comment cela marche ? Quel avenir ?
- Big Data. Faut-il de nouveaux paradigmes pour l'informatique ?
- Après les révélations de Snowden. Comment rebâtir une sécurité des transmissions avec Internet ?
- Sécurité d'un réseau national, quel matériels – logiciels utiliser (composants de réseau) ?
- Les brevets de 2008 des sociétés Microsoft et Apple, permettant l'arrêt à distance de certaines fonctionnalités des mobiles.
  Par exemple : interdire de téléphoner dans un cinéma, de faire des photos dans un musée ...
- Comparaison de BitTorrente et Avalanche (Microsoft).

- Comparaisons des Protocoles pour réseaux ad-hoc : AODV et OLSR
- Etude et essais du protocole de réseau ad-hoc BABEL
- Freenet et le projet Tor. Sont-ils vraiment sûrs ?
- Détection d'intrusion et sécurité dans les réseaux Ad-Hoc
- Aspects informatiques de la distribution de l'électricité (Smartgrids)
- Géolocalisation : comment et pourquoi. Il s'agit d'étudier les techniques à mettre en œuvre, donner les précisions de positionnement aujourd'hui possible. Mais il faudrait aussi répondre à la question : dans quel but Google piste tous nos déplacements ?
- Où en sommes nous légalement sur le streaming et le téléchargement ?
- Aspects informatiques de Wikileaks ou comment faire circuler de l'information secrètement.

- De quelles informations disposons-nous sur le réseau Internet (en tant que graphe) ?
- Etude en détail des performances d'une table de hachage distribuée.
- Cloud computing
- La plateforme Amazon de cloud computing
- Le point sur la légalité des trackers et des échanges P-2-P sans tracker
- Evolution du statut d'hébergeur (point de vue technique et juridique)
- La biométrie : atouts et inconvénients (avantages ? qui surveillent ? etc.)
- Les cartes GPU fonctionnement et applications dédiées. Architecture Multicore ou GPU ?
- Quel parallélisme dans nos PC, comment l'utiliser ?
- Le système de fichiers distribué de GOOGLE. Utilisation du framework Map and Reduce ?

- La plateforme Hadoop
- VANET un réseau ad-hoc pour le transport
- Réseaux de capteurs, gestion de l'énergie
- Blackphone -Anti-NSA smartphone