

Mémoire atomique auto-reconfigurable pour systèmes P2P

Vincent Gramoli

17 novembre 2005



Objectifs

- Partage *Pair-à-Pair* (P2P) des ressources.
- Read-Only → Read-Write Model.
- Équilibrage de charge.
- Auto-adaptation face au dynamisme.

Plan

- 1 Introduction
 - Motivations et Applications
 - Modèle
- 2 Procédure de démarrage
- 3 Problématique
 - Surcharge
 - Atomicité
- 4 Mémoire Atomique
- 5 Auto-Reconfiguration
- 6 Conclusion

Motivations et Applications



- Groupware (Collecticiel)
- Web Services
- Intelligence collective

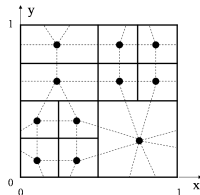
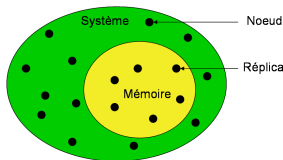
...en P2P, sans coût de maintenance, de stockage, etc.

Modèle

- Distribué
 - Réseau de n nœuds connectés
- Dynamique
 - Arrivées et départs de nœuds
 - Pannes crash
 - Détection de fautes ultime
- Asynchrone
 - Communication par voisinage
 - Délai de message arbitrairement long
 - Perte éventuelle de messages

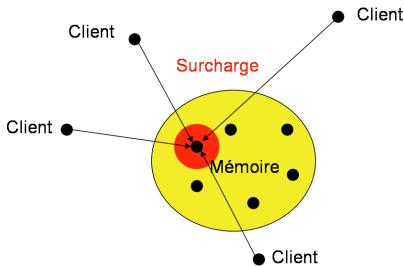
Procédure de démarrage

- Responsabilité de l'objet $X = \text{Plan fictif } [0, 1) \times [0, 1)$.
- Initialement un seul nœud R responsable maintient une copie de l'objet.
- R copie X sur $m - 1$ nœuds : les *réplicas*.
- Plan partagé au fur et à mesure des ajouts de réplicas.
- Réplicas forment un overlay en tore, la *mémoire* :
 - Les responsables des zones adjacentes sont *voisins*.



Surcharge

- Tout nœud est client potentiel.
- Un nœud envoie une requête sur un réplica de la mémoire (qu'il connait) pour effectuer une lecture/écriture.
- Il est possible qu'un réplica ne puisse traiter toutes les requêtes reçues. Il est alors surchargé.



Atomicité

Définition (Atomicité)

Une exécution vérifie la propriété d'atomicité [Lynch 96] ssi

- *les écritures sont totalement ordonnées,*
- *les lectures sont ordonnées par rapport aux écritures et renvoient la dernière valeur écrite, et*
- *cet ordre respecte la précedence de temps réel.*

Mémoire Atomique

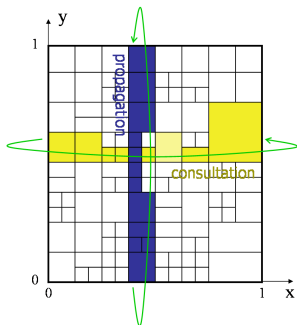
Définition ($\langle tag, val \rangle$)

Tout réplica de l'objet considéré conserve sa valeur, val , et un compteur d'écriture, tag , indiquant la version de cette valeur.

- Consultation
 - Les paires $\langle tag, val \rangle$ des responsables de tout une **ligne** du plan sont consultées :
 - La paire possédant le plus grand tag est retournée.
- Propagation
 - La paire $\langle tag, val \rangle$ est propagée à l'ensemble des responsables d'une **colonne**.

Les lignes et colonnes de notre grille torique sont des *quorums* [Giff79] : ces ensembles s'intersectent mutuellement.

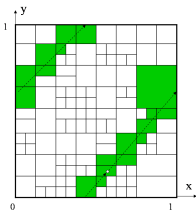
- $Lecture(v!) \Leftrightarrow Consultation(\langle t, v \rangle !) + Propagation(\langle t, v \rangle ?)$
- $Écriture(v?) \Leftrightarrow Consultation(\langle t', v' \rangle !) + Propagation(\langle t' ++, v' ?)$



- Atomicité : Le *tag* utilisé dans une opération définit son numéro d'ordre.

Mémoire Auto-Reconfigurable

- Surcharge :
 - En cas de réceptions quasi-simultanées, un réplica exécute plusieurs requêtes en une seule.
- Équilibrage de Charge :
 - Un réplica surchargé transfère la requête au voisin de son voisin (diagonale).
 - S'il récupère le même message, il approxime une surcharge du système.



- Extension
 - En cas de surcharge, une réplication est effectuée sur un nœud actif.
 - Le plan de responsabilité est partagé.
- Réduction
 - Un réplica sans sous-charge prévient et sort du système.
 - La responsabilité est rendue. (cf. changements dans les zones de [CAN01])

Conclusion

- Résumé
 - Réplication \Rightarrow Tolérance aux fautes.
 - Quorums \Rightarrow Atomicité.
 - Auto-adaptation \Rightarrow Compromis entre charge et complexité.
 - Connaissance locale \Rightarrow Passage à l'échelle.
- Résultats
 - Opérations et Reconfiguration passable à l'échelle.
 - Reconfiguration peu coûteuse (temps & #messages).
 - Opérations coûteuses en temps ($O(\sqrt{m})$ hops).
- Travaux en cours et futures :
 - Temps des opérations : $O(\log \sqrt{m})$ hops.
 - Modèle byzantin.
 - Opérations transactionnelles.

Références

Giff 79 Weighted Voting for Replicated data
D.K. Gifford - *SOSP 1979*

Lynch 96 Distributed Algorithms
Nancy Lynch - *Morgan Kaufman, 1996*

CAN01 A Scalable Content Addressable Network
S. Ratnasamy, P. Francis, M. Handley, R.M. Karp et S.
Shenker - *SIGCOM 2001*

RAMBO RAMBO, A Reconfigurable Atomic Memory for Basic Objects
Nancy Lynch, Alex Shvartsman - *DISC 2002*

GMS05 Operation Liveness and Gossip Management in a Dynamic
Distributed Atomic Data Service
Vincent Gramoli, Peter Musial, Alexander Shvartsman - *PDCS
2005*