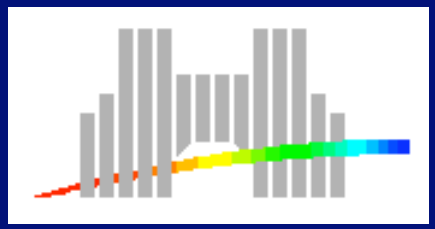
Conception, analyse et validation de protocoles de multicast fiable avec assistance des routeurs

Soutenance de thèse présentée par:

Moufida MAIMOUR

mardi 25 novembre 2003



Introduction

- L'Internet haut débit et les application émergentes
- Applications multi-points:
  - distribution de logiciels
  - mise à jour et réplication de base de données
  - simulation interactive distribuée
  - calcul distribué
  - conférences multimédia
  - jeux distribués
- Il est nécessaire d'avoir un mécanisme efficace de communication multipoint

Introduction

- Multicast IP **[Deering:RFC1112]** permet une délivrance efficace des données à un ensemble de récepteurs
- Un service best effort qui ne garantit pas la réception correcte des paquets par les récepteurs
- Le but de ce travail est de fournir des services de niveau transport aux applications multicast fiables

Introduction

- Les problèmes abordés :
  - le recouvrement des erreurs
  - le contrôle de congestion
- Objectifs :
  - passage à l'échelle
  - adaptabilité aux changements dynamiques d'un arbre de multicast
  - support de l'hétérogénéité
- L'approche adoptée consiste en l'implication des routeurs pour permettre des solutions plus efficaces
- La technologie choisie : Réseaux actifs

Contenu de la thèse

- Une analyse de débit où un certain nombre de protocoles actifs sont comparés
- Une analyse de délai afin d'évaluer le bénéfice d'un service de détection de pertes par les routeurs
- Un protocole de multicast fiable : DyRAM
- Un protocole d'évitement de congestion : AMCA
- Le support de la présence de multiples récepteurs hétérogènes

Plan de la présentation

- Un protocole de multicast fiable DyRAM
- Un protocole d'évitement de congestion AMCA
- Support de l'hétérogénéité : une réplication par les récepteurs
- Conclusion et perspectives

Un protocole de multicast fiable DyRAM

Dynamic Replier Active Reliable Multicast

Protocoles de multicast fiable

- Approches de bout en bout :
  - avec recouvrement local :
    - Approche probabiliste **[SRM]**
    - Approches hiérarchiques statiques **[RMTP]** ou dynamiques **[TMTP, TRAM]**
- Approches avec assistance de routeurs
  - un arbre de recouvrement identique à l'arbre physique du multicast avec cache de données au niveau de nœuds intermédiaires **[ARM, RMANP, AER]**
  - un arbre de recouvrement logique congruent construit avec l'assistance des routeurs **[LMS, PGM, AIM]**

Le protocole DyRAM

- La construction implicite via l'élection de retransmetteurs d'un arbre de recouvrement logique dynamique et congruent
- L'avantage de DyRAM est que l'élection se fait par paquet perdu :
  - pas d'arbre de recouvrement préalablement construit
  - possibilité de faire un équilibrage des charges

Description de DyRAM

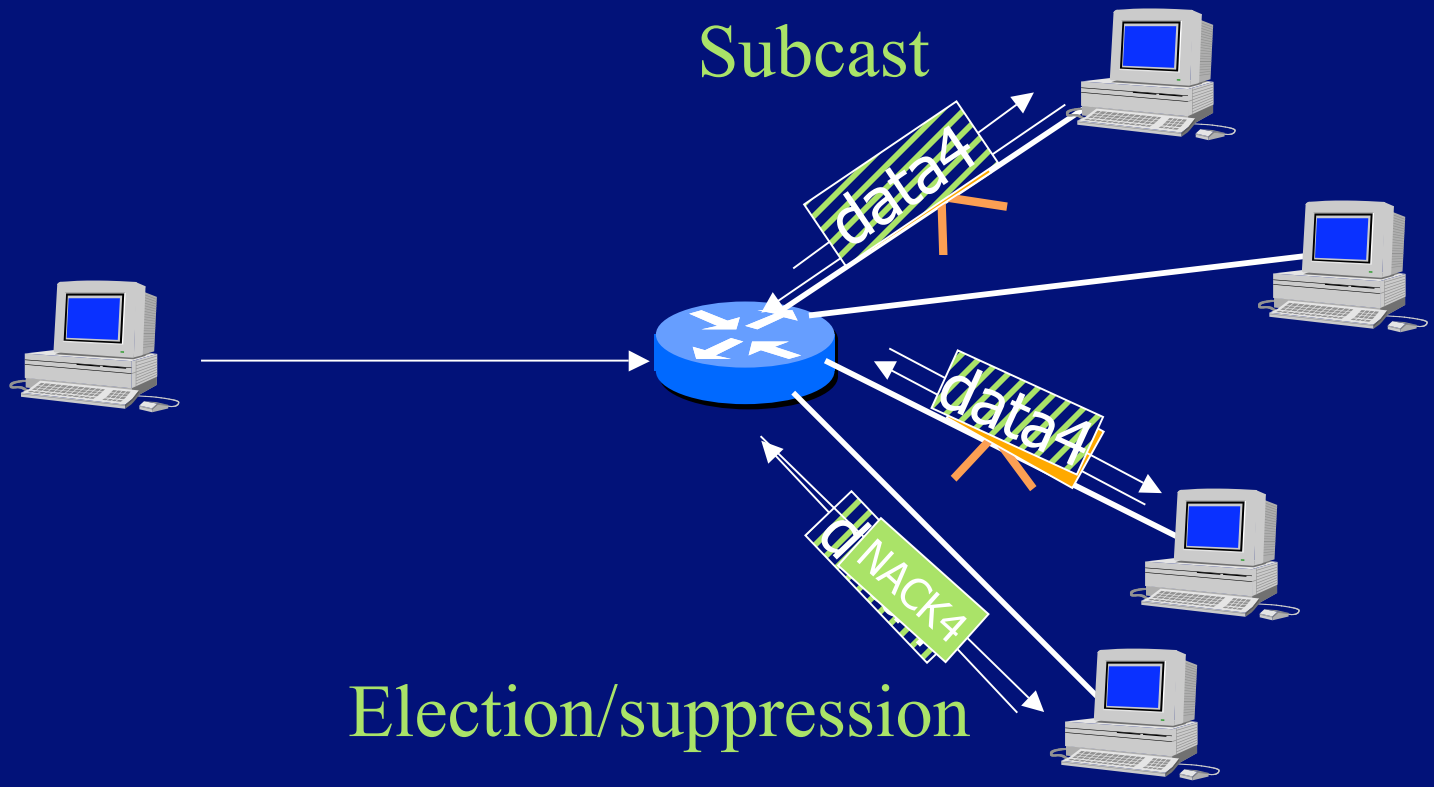
- Une perte est signalée par un récepteur par l'envoi immédiat d'un acquittement négatif (NACK).
- En l'absence de pertes, des acquittements positifs (ACKs) sont périodiquement portés par des messages spéciaux (CRs, Congestion Report)
  - libération de mémoire
  - contrôle de flux
  - contrôle de congestion
- Les différents types de paquets sont traités d'une façon personnalisée par les routeurs.

Les services actifs de DyRAM

- La suppression des NACKs et agrégation des CRs
- Le subcast
- L'élection de retransmetteurs
- La détection des pertes
- Le calcul des RTTs

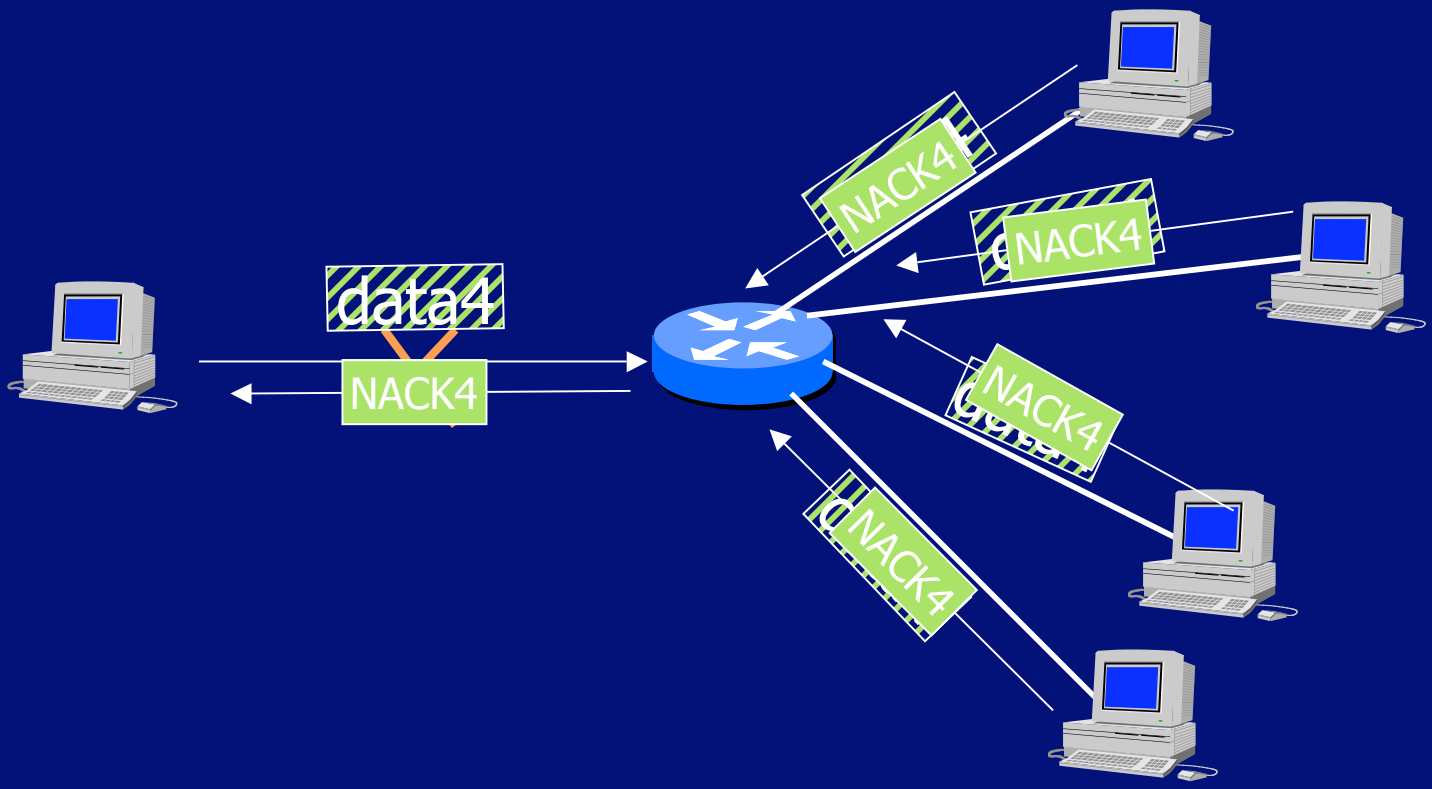
Les services actifs de DyRAM

Suppression, élection et subcast



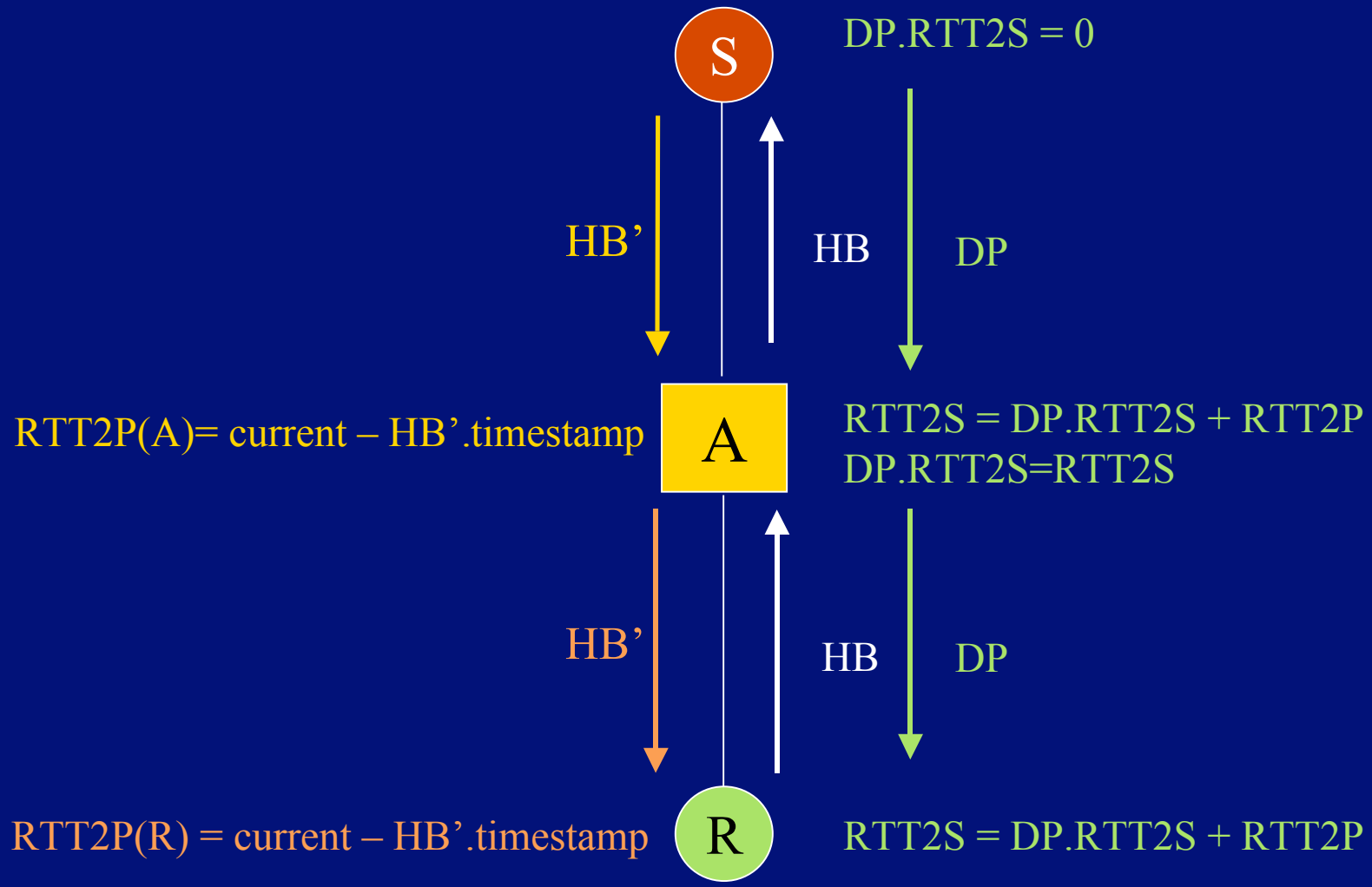
Les services actifs de DyRAM

Détection des pertes par le routeur



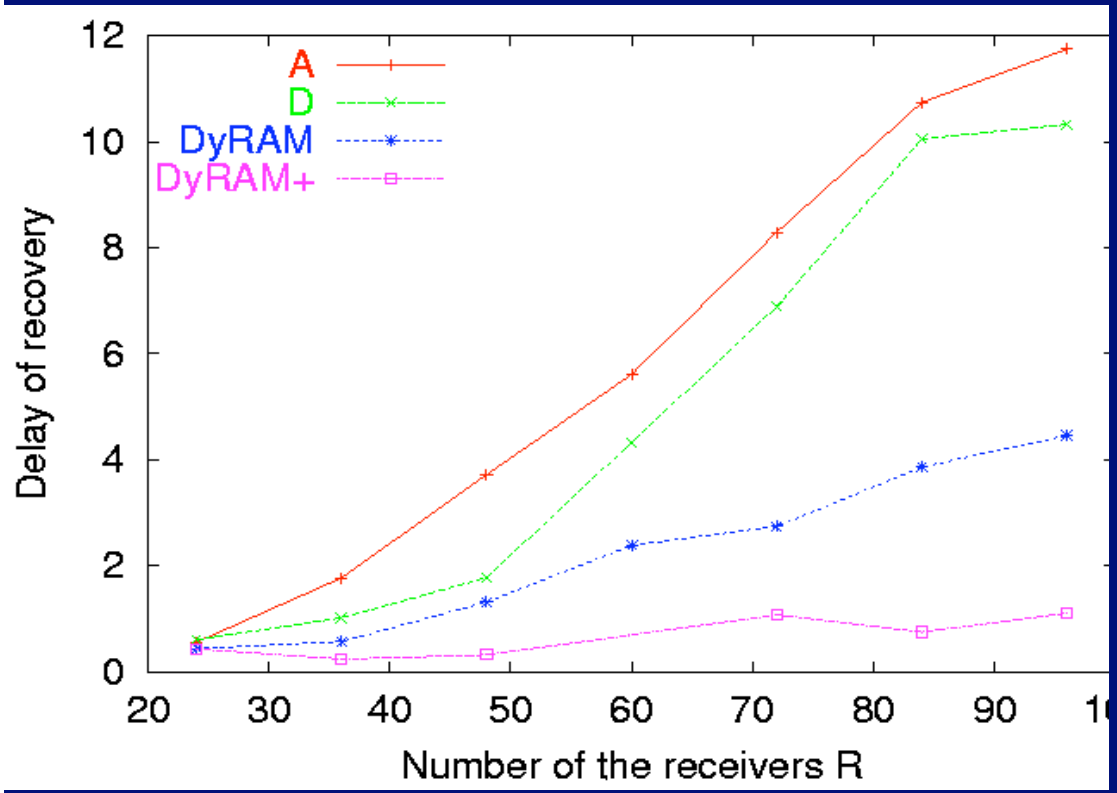
Les services actifs de DyRAM

Estimation des RTTs



Résultats de simulations

Détection des pertes et recouvrement local



A : suppression des NACKs

D : A + détection des pertes

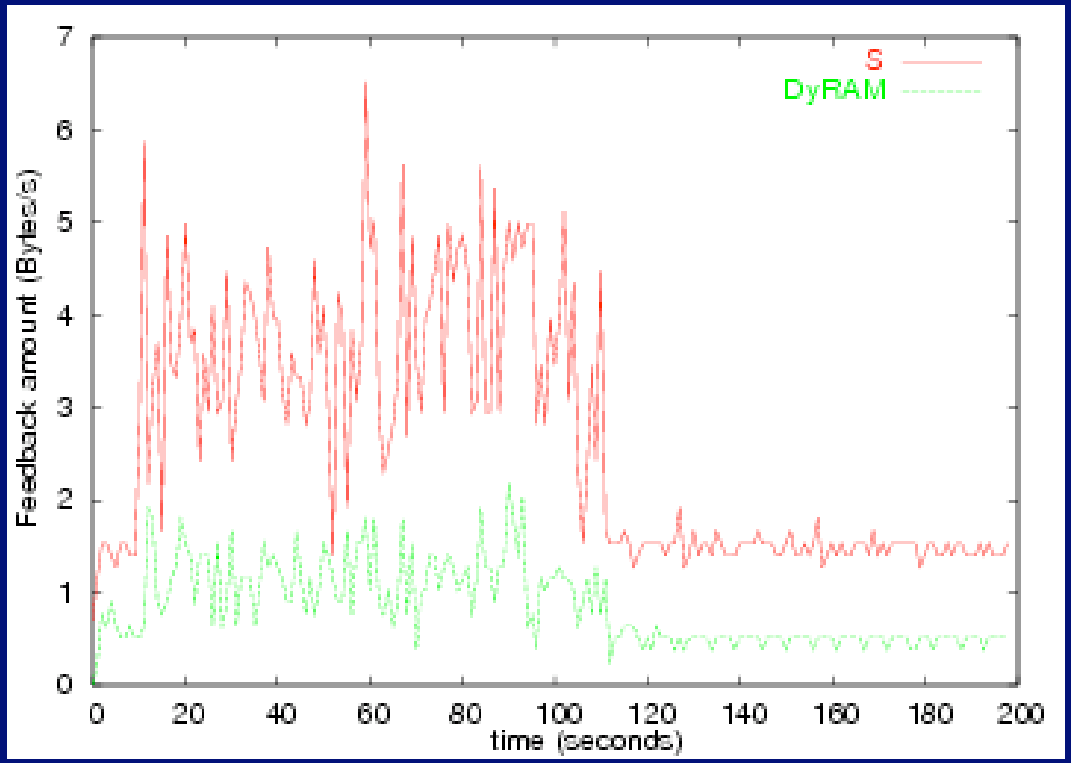
DyRAM : A + élection de retransmetteur

DyRAM+ : DyRAM + détection des pertes

La latence normalisée en fonction du nombre des récepteurs

Résultats de simulations

Suppression des messages de contrôle

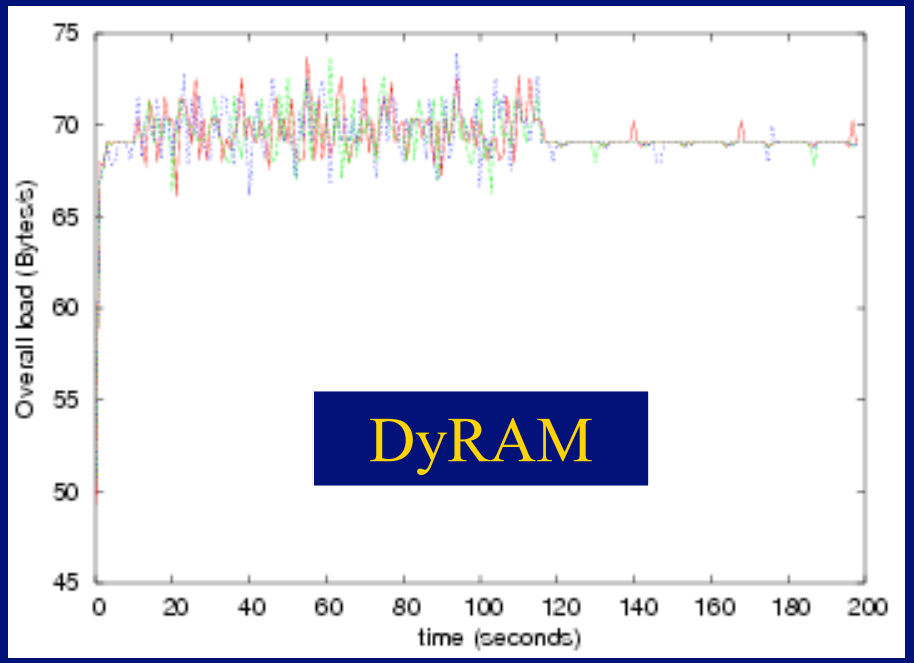


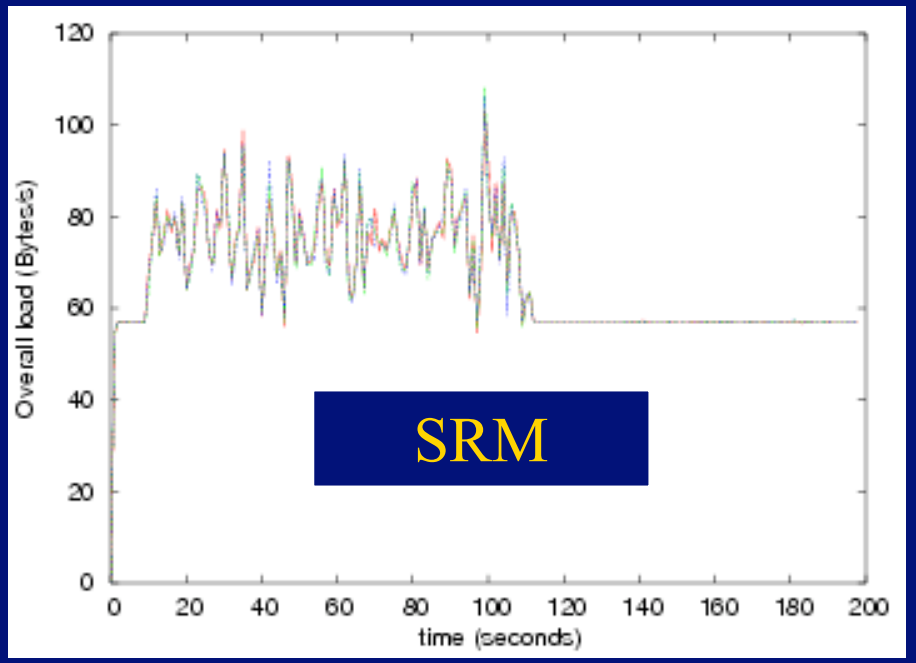
La charge de la source en réception
taux de pertes 10%

Résultats de simulations

Equilibrage des charges

La charge de 3 récepteurs différents pris au hasard avec 1% de pertes



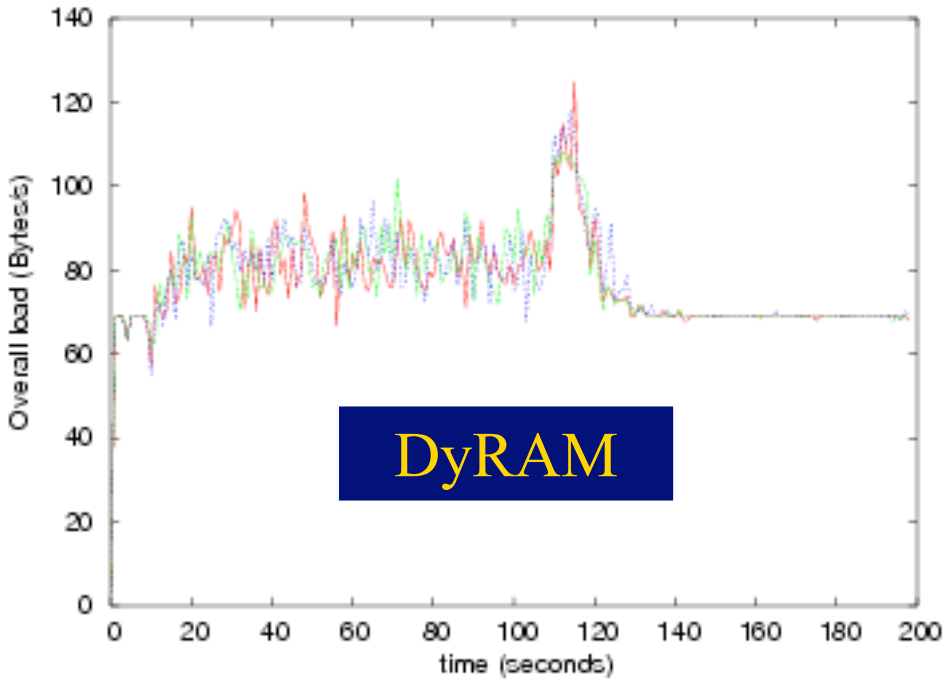


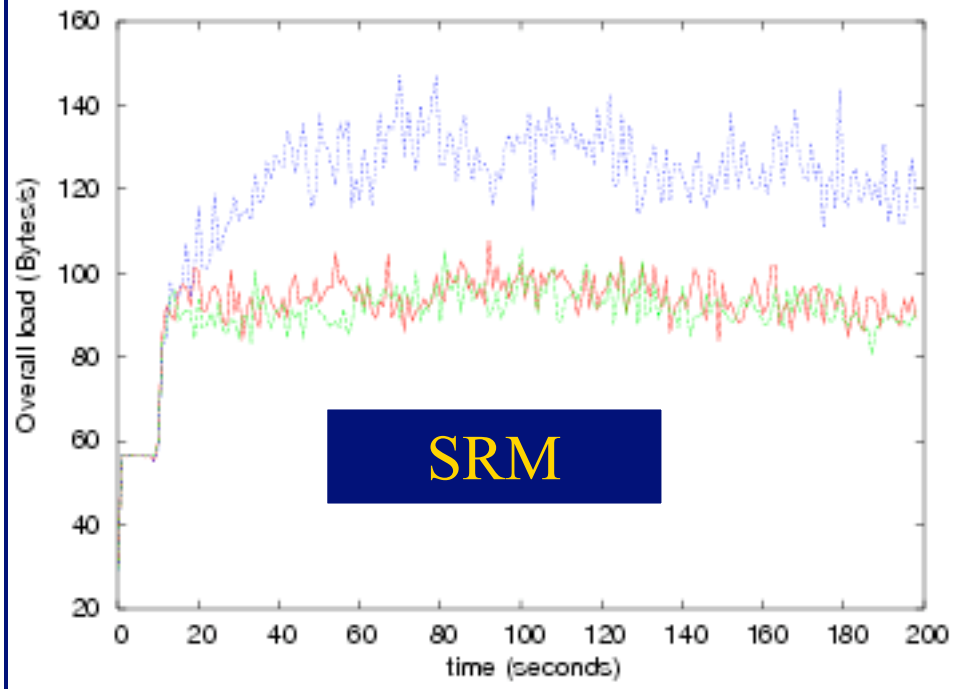
La charge est la même au cas d'un taux de pertes faible

Résultats de simulations

Equilibrage des charges

La charge de 3 récepteurs différents pris au hasard avec 25% de pertes



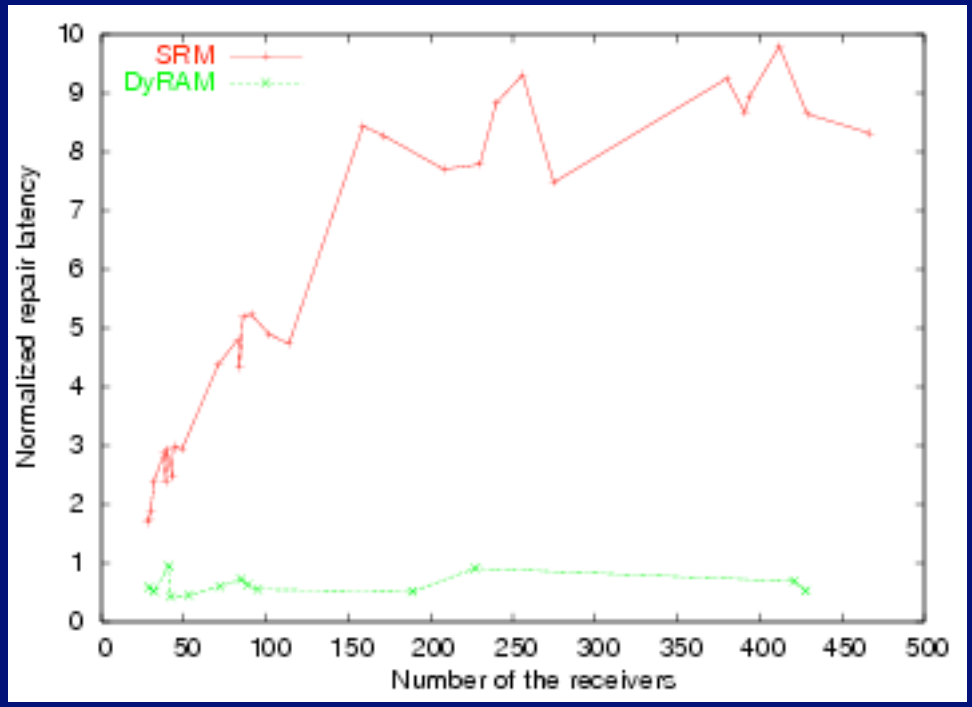


Il y a plus d'équilibrage de charge dans DyRAM contrairement à SRM au cas d'un taux de pertes élevé

Résultats de simulations

Passage à l'échelle

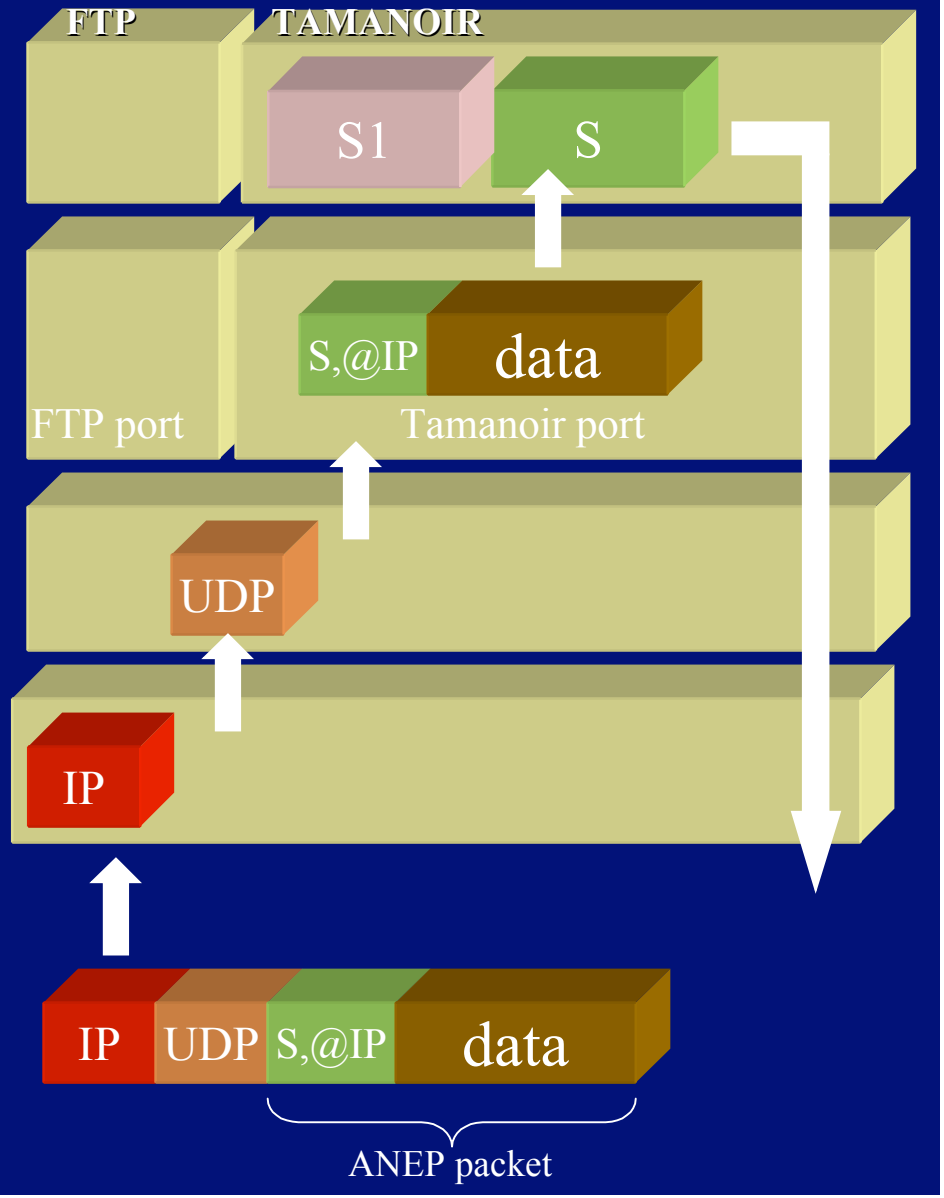
Le temps de recouvrement moyen normalisé au RTT à la source en fonction du nombre des récepteurs avec 5% de pertes



- DyRAM a généralement un temps de recouvrement inférieur à 1 RTT
- Dans SRM, cette latence croît avec le nombre des récepteurs

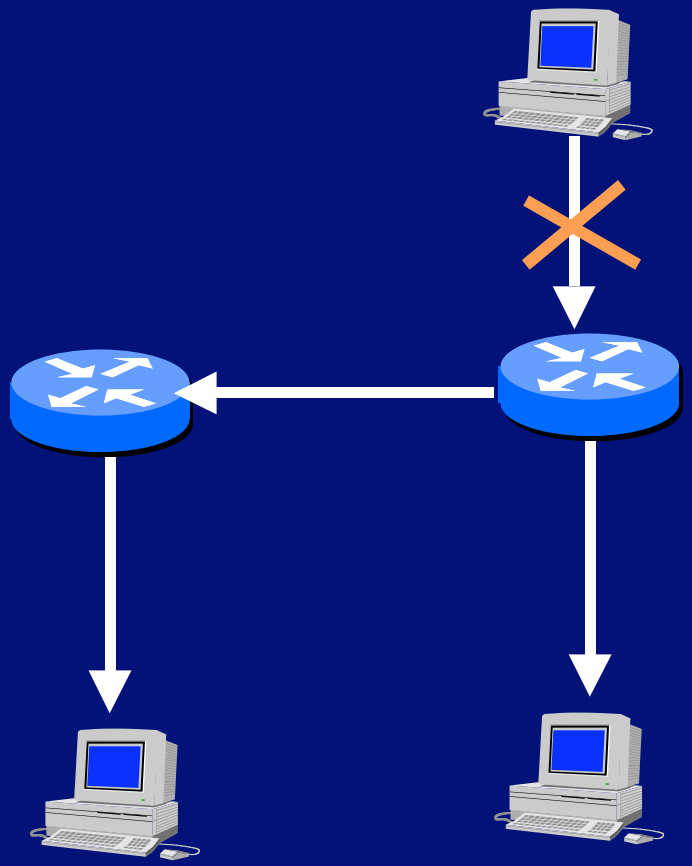
Implémentation

- Java 1.3.1 et noyau linux 2.4
- Tamanoir
- Format ANEP



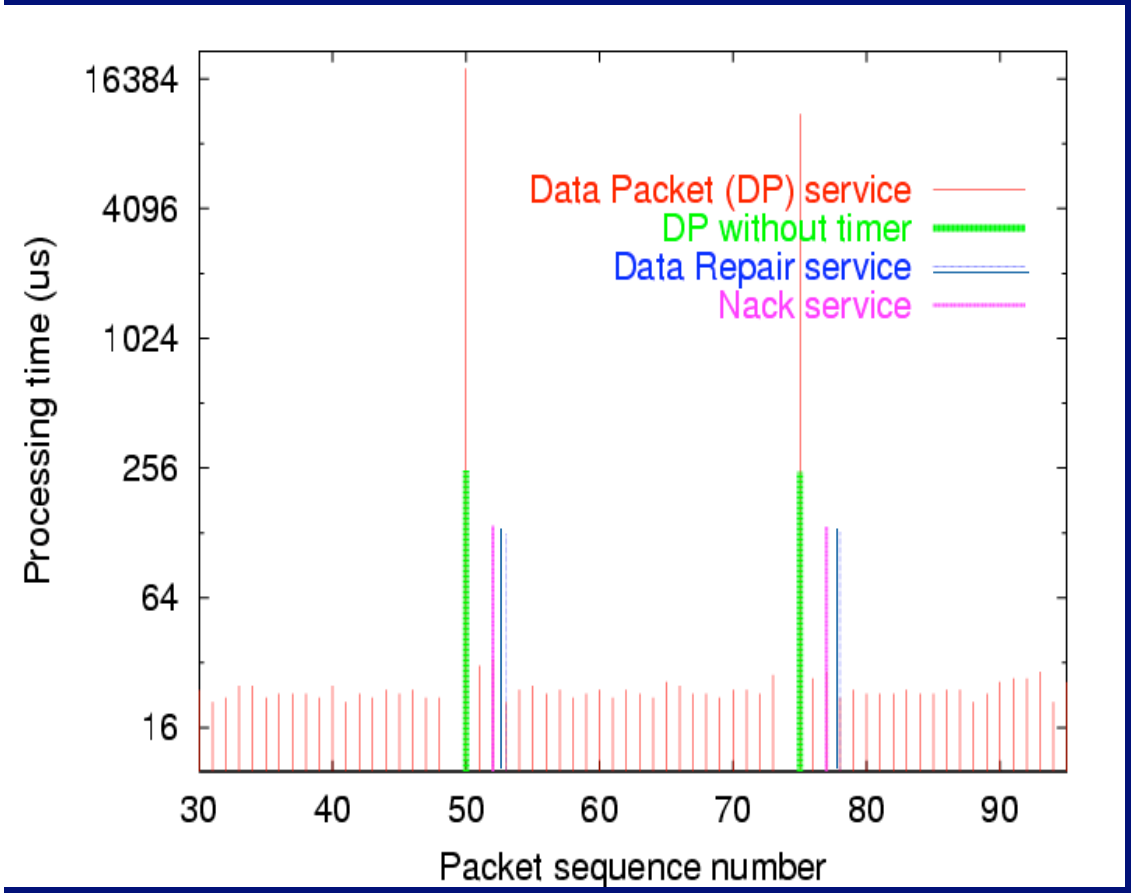
Expérimentations

- Une plate-forme de test local avec un ensemble de PC
- Expérimentation sur le réseau vthd dans le cadre du projet e-toile



Résultats expérimentaux

Coût des services actifs



NACK : 135_s

Retransmission : 123_s

DP : 20_s si pas de pertes
sinon 12-17ms
256_s sans la m-à-j du timer

Pentium II 400 MHz, 512KB de cache et 128 MB de RAM

Un protocole d'évitement de congestion : AMCA

Active-based Multicast Congestion Avoidance

Le problème

- Co-existence de plusieurs récepteurs:
  - plusieurs chemins avec des caractéristiques différentes
  - implosion des messages de contôle
- La suppression des messages de contrôle permet d'éviter l'implosion
- mais elle doit préserver l'information pour éviter d'affecter la vitesse de réaction à la source
- En plus de la stabilité et la vitesse de réaction, en multicast, un contrôle de congestion doit permettre le passage à l'échelle et l'équité avec TCP et les autres flux multicast

Solutions existantes

- Suppression des messages de contrôle une fois arrivés à la source :
  - suppression probabiliste **[RLA]**
  - suppression à partir d'un seuil **[LTRC]**
- Suppression des messages de contrôle avant leur arrivée à la source :
  - représentants **[Representatives:DeLucia] [PGMcc]**
  - hiérarchie **[TRAM] [MTCP]**
- L'équité avec TCP :
  - émulation de TCP : fenêtre de congestion **[MTCP]**
  - équation de TCP : régulation de débit [TFMCC]

Solutions actives

- L'utilisation de l'arbre de multicast physique pour :
  - la suppression des messages de contrôle **[NCA] [RMANP]**
  - l'agrégation d'informations afin d'élire un représentant **[NCA]**
- Le cache des paquets de données pour une régulation de débit par les routeurs **[RMANP]**
- L'utilisation des réseaux actifs pour implémenter une approche multi-couches « layered approach» **[ALMA]**

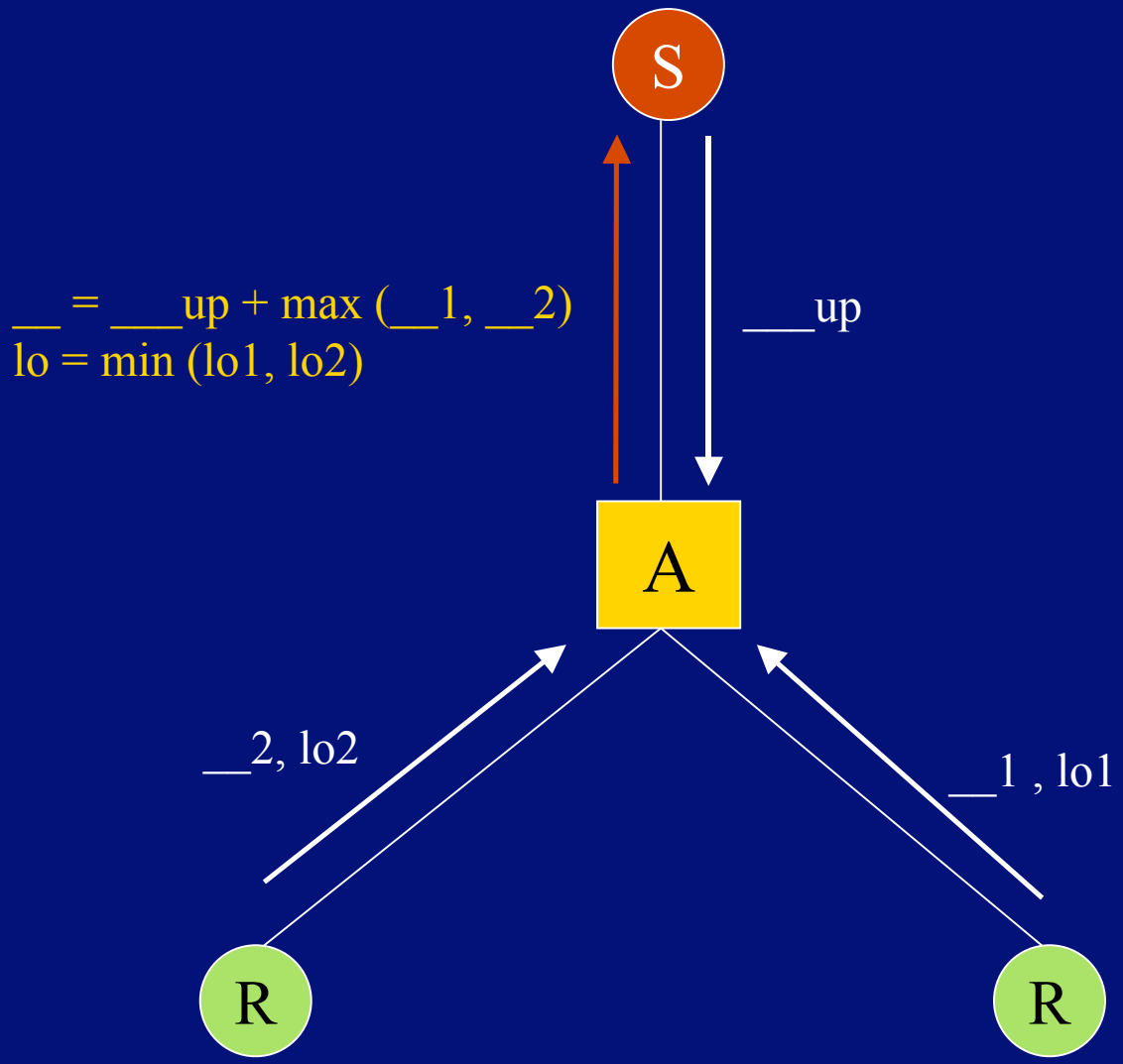
La solution AMCA

- La bande passante disponible est estimée grâce à des mesures de RTTs.
- L'estimation est faite par saut, ainsi :
  - Le passage à l'échelle est assuré,
  - et le problème de changement de chemins dans l'arbre de multicast est évité
- Les messages de contrôle sont agrégés de telle sorte que la source finit par recevoir le message de contrôle du récepteur le plus faible :
  - une élection « scalable » (implicite) d'un représentant
  - résout le problème de l'implosion des messages de contrôle

Messages de contrôle

- Un NACK est envoyé immédiatement lors de la détection d'une perte
- Un récepteur envoie un CR (Congestion Report) à la source pour N paquets reçus
- Un CR contient en particulier :
  - la variation de RTT (__) et sa période (T)
  - le numéro de séquence du dernier paquet reçu dans l'ordre (lo)

Agrégation des CRs



Régulation de débit

- La régulation de débit se fait par la source en calculant :
  - La variation de la taille de file par paquet émis :

$$\_q\_\_ = \_\_/(\_\_+T)$$

à maintenir $< \varepsilon < 1$

  - Le nombre des paquets pas encore acquités :

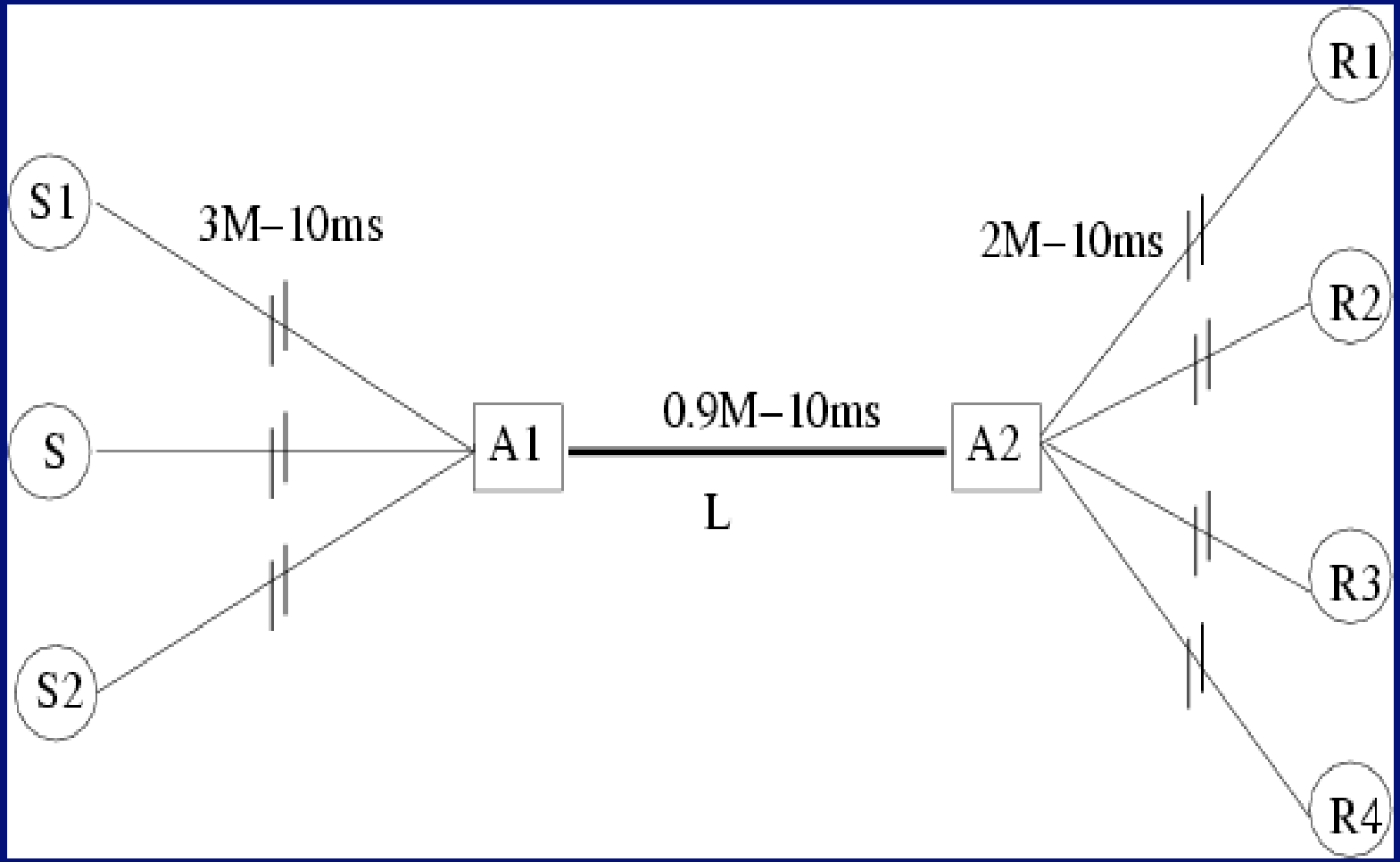
$$\_q\_p = \max(0, N - (loi+1 - loi))$$

à maintenir entre α et β (similaire à TCP-Vegas)

- Le débit est à maintenir entre rate_min et rate_max
- Phases de slow-start et d'évitement de congestion

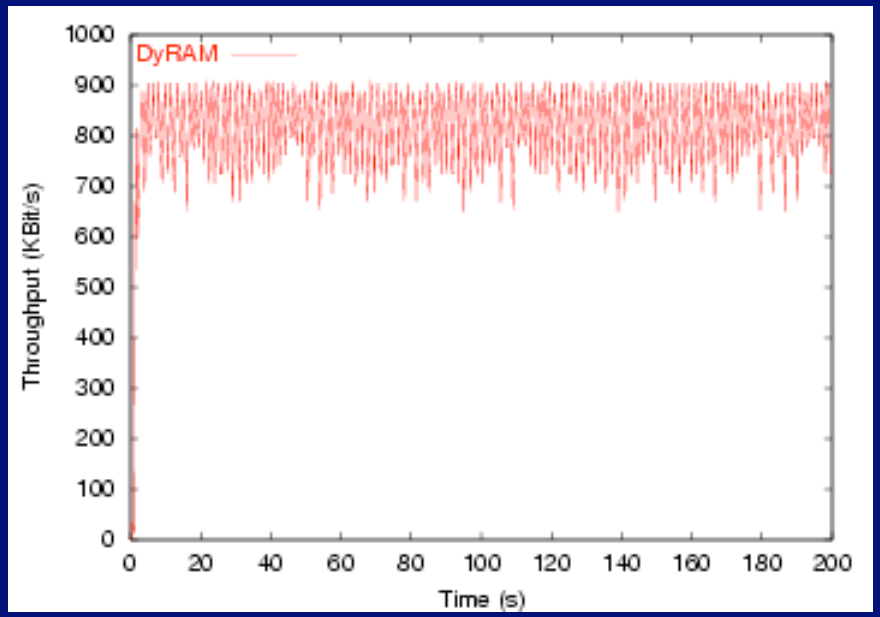
Résultats de simulation

Topologie d'évaluation

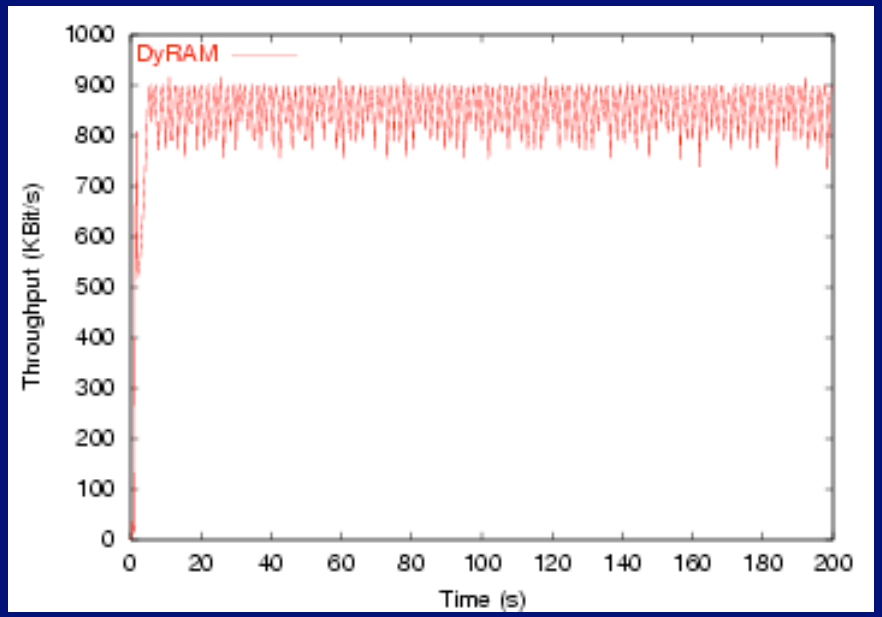


Résultats de simulation

Convergence et temps de réponse



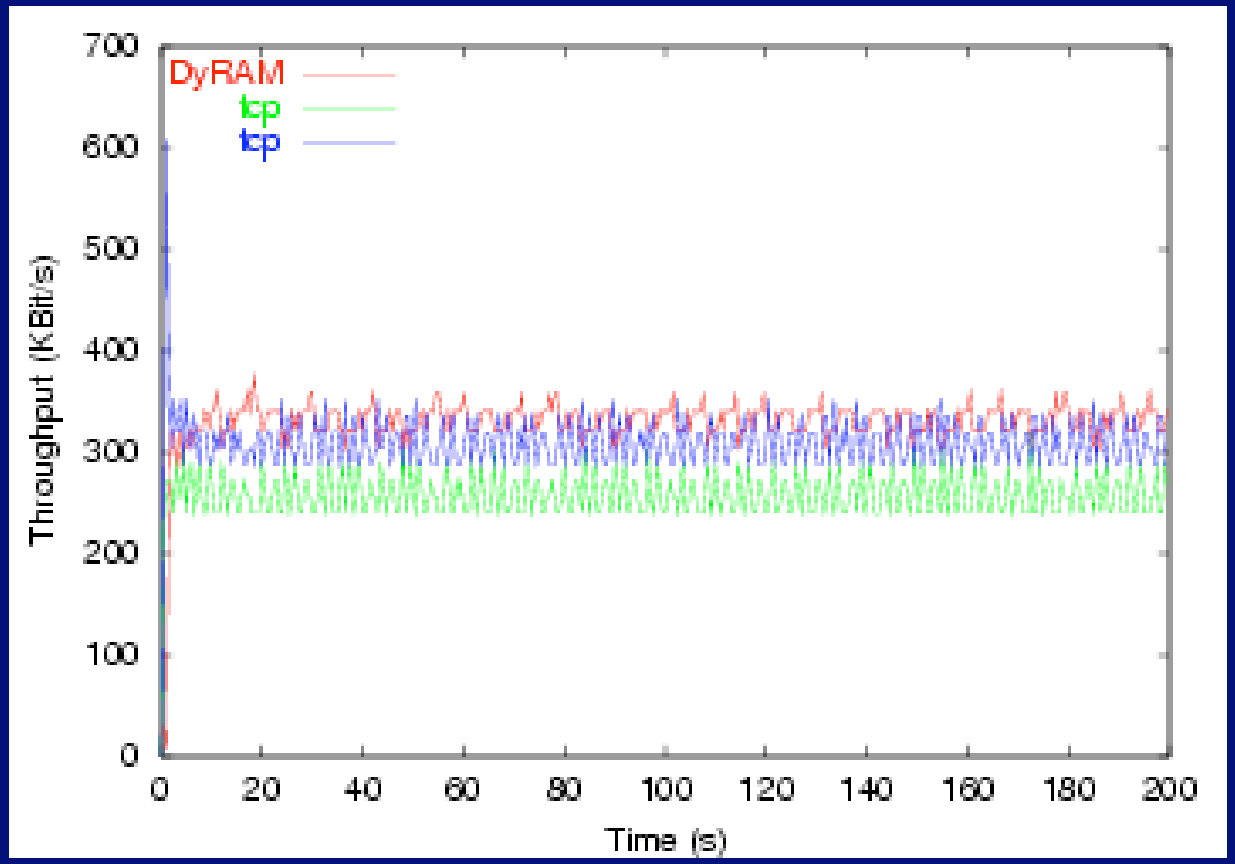
N=16



N=32

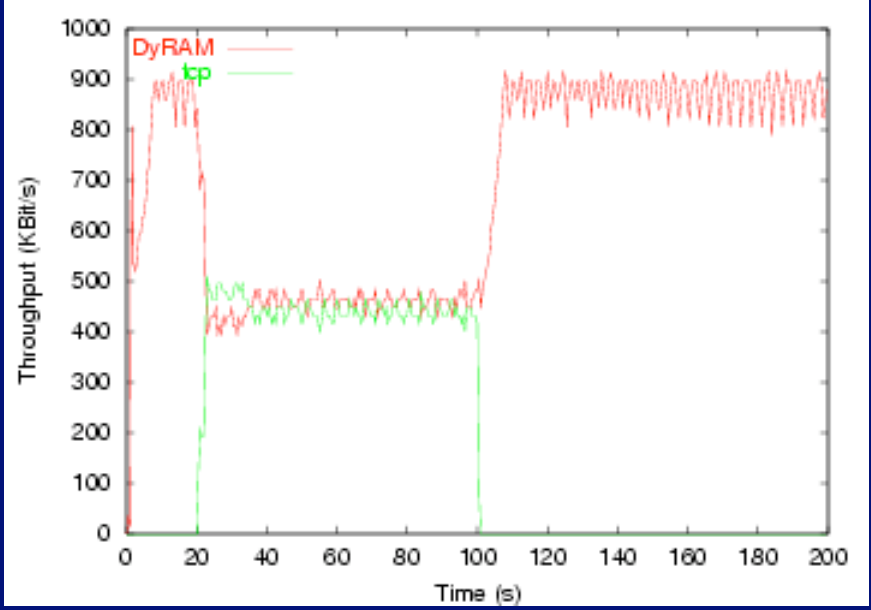
Résultats de simulation

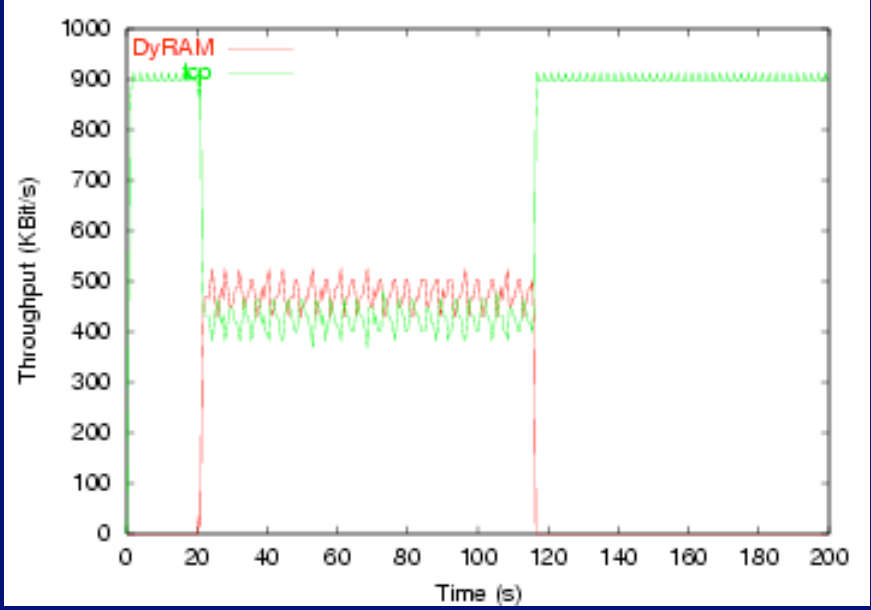
L’équité avec 2 flux TCP



Résultats de simulation

L’équité avec un flux TCP





La prise en charge de l'hétérogénéité

Une approche multi-débits avec réplication par les récepteurs

La problématique

- AMCA ajuste son débit en réponse au récepteur le plus faible
- Pour améliorer la satisfaction (ou l'équité inter-récepteur) des récepteurs, une approche multi-débits est plus appropriée :
  - approches multi-couches (layered) **[RLM][RLC][FLID-DL]**
  - approches avec réplication de flux [DSG]
- Le débit de transmission pour chaque groupe doit être finement ajusté en fonction des changements dynamiques des capacités des récepteurs

➢ Adapter les débits des différentes couches (layering) ou sous-groupes (réplication)

- Un algorithme de partitionnement est nécessaire

Algorithme de partitionnement

- L'algorithme proposé s'exécute à la volée en utilisant les variations de RTT reportés par les CRs envoyés périodiquement par les récepteurs
- Il n'a pas besoin d'une estimation préalable de la capacité des récepteurs
- Il est simple mais atteint ou au moins approche la solution optimale
- Il assure un minimum de satisfaction en fonction de ses paramètres et le degré d'hétérogénéité des récepteurs

Algorithme de partitionnement

Input: $P_0 \leftarrow \{R_j, j = 1, \ldots, N\}$, a set of receivers.

Output: a partition $\{P_0, P_{K-1}, \ldots, P_2, P_1\}$ where $K \leq G$

Require: $N > 1$ and $a < b < \dot{\epsilon}$

Initially the source rate $r \leftarrow r_{min}$ and $i \leftarrow 1$

Periodically,

if $\exists j, R_j \in P_0$ such that $\Delta\dot{\tau}_j > b$ **then**

$P_i \leftarrow \{R_j \in P_0, \Delta\dot{\tau}_j > a\}$ and $P_0 \leftarrow P_0 - P_i$

$i \leftarrow i + 1$

if $\exists j, R_j \in P_0$ such that $\Delta\dot{\tau}_j > \dot{\epsilon}$ **then** $dec(r)$ **else** $inc(r)$

end if

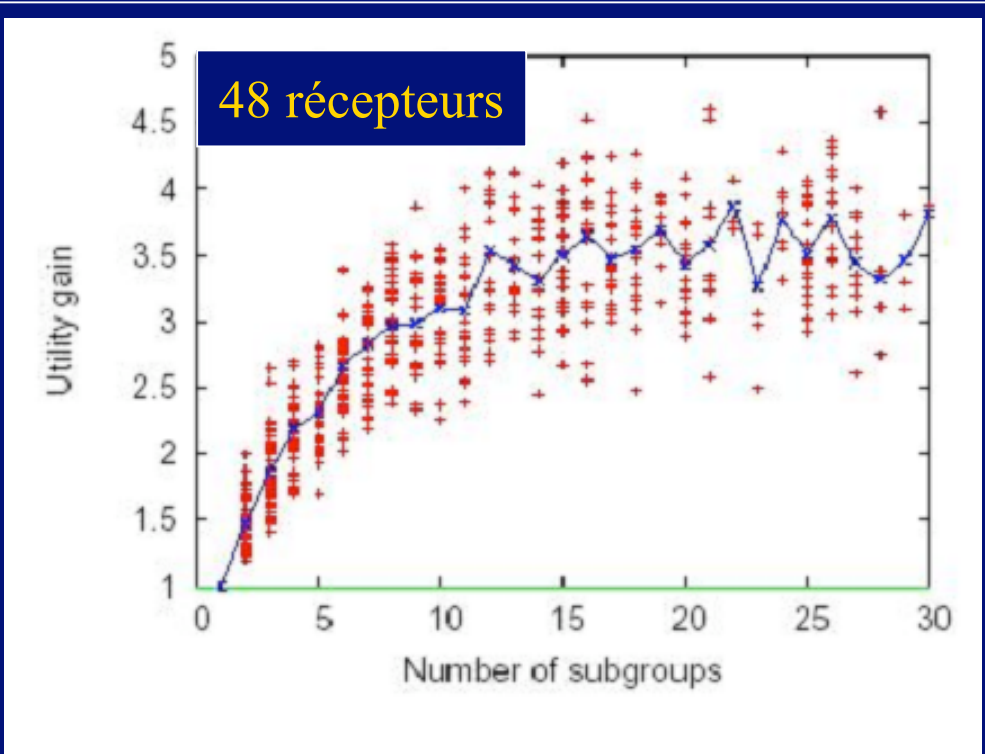
until $i = G$ or $|P_0| = 1$ or $\forall j < N, R_j \in P_0, \frac{1+\Delta\dot{\tau}_{j+1}}{1+\Delta\dot{\tau}_j} \geq \rho$ where

$\rho = \frac{a+1}{b+1}$

Algorithme de partitionnement

Le gain en satisfaction

Une distribution uniforme entre 5 et 55 des débits isolés des récepteurs



• Le gain n'augmente pas considérablement avec le nombre de groupes à partir d'un certain seuil.

Un schéma de réplication par les récepteurs

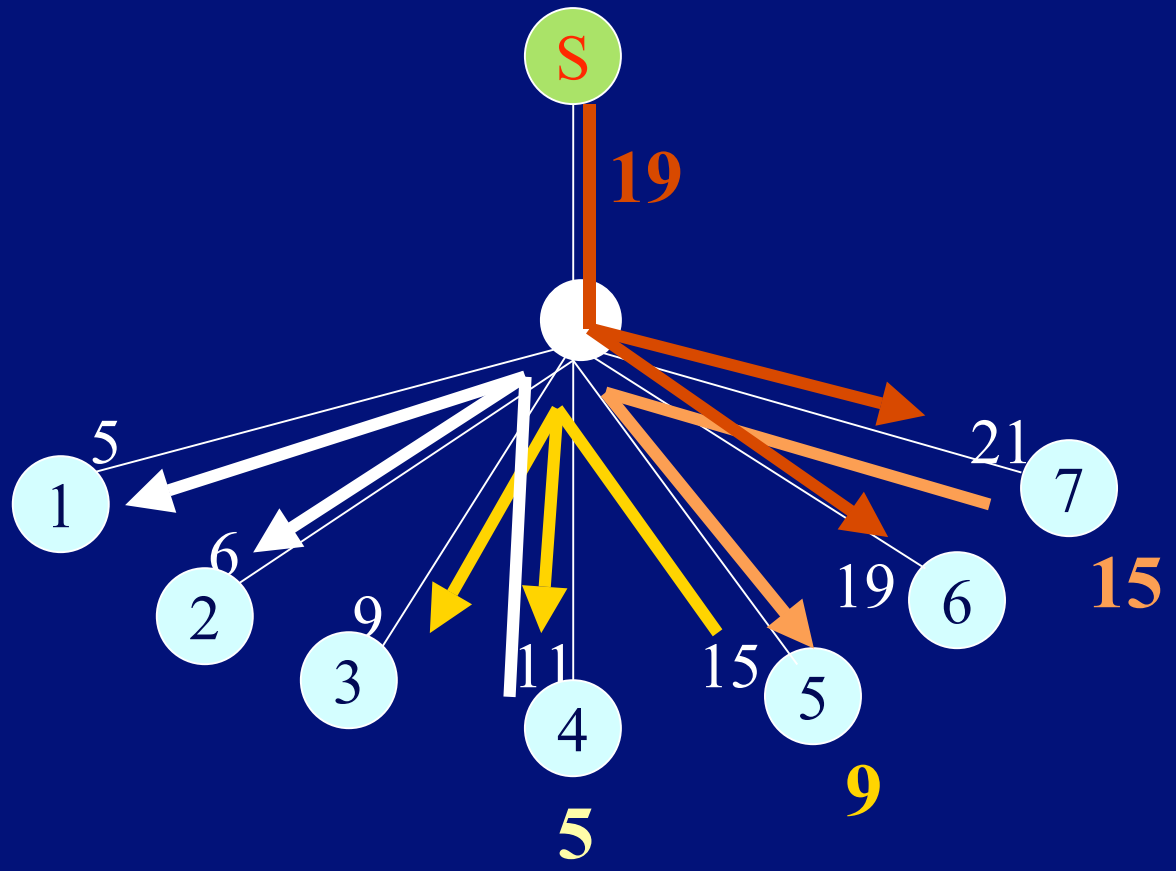
- Une approche multi-débit avec réplication par les récepteurs (RbR : receiver-based replication) par opposition à la réplication par la source (SbR : source-based replication) :
  - La réplication est adoptée pour éviter la complexité des approches multi-couches
  - La réplication ne se fait plus par la source
    - pour éviter la surcharge de la source et ses liens
    - diminuer l'agressivité par rapport aux autres flux

Un schéma de réplication par les récepteurs

- Un nombre de récepteurs (réplicateurs) sont désignés pour répliquer le flux à des récepteurs de capacité inférieure
- A travers l'exécution d'un algorithme de partitionnement, un réplicateur est choisi pour chaque sous-groupe formé
- ainsi un arbre de régulation est construit avec la source à la racine et les réplicateurs comme nœuds intermédiaires
- Un réplicateur régule son flux finement en fonction de la capacité des récepteurs du sous-groupe qui lui est associé à travers un protocole de contrôle de congestion (AMCA par exemple)
- L'algorithme de partitionnement est exécuté par les routeurs:
  - plus de passage à l'échelle
  - l'arbre de régulation est plus proche de l'arbre de multicast

Un schéma de réplication par les récepteurs

Exemple avec une topologie en étoile



Un schéma de réplication par les récepteurs

- L'approche RbR a été comparé à travers des analyses et simulations à une approche SbR ainsi qu'une approche à un seul débit :
  - satisfaction améliorée (équité inter-récepteurs)
  - plus d'équité avec les autres sessions (équité inter-session)

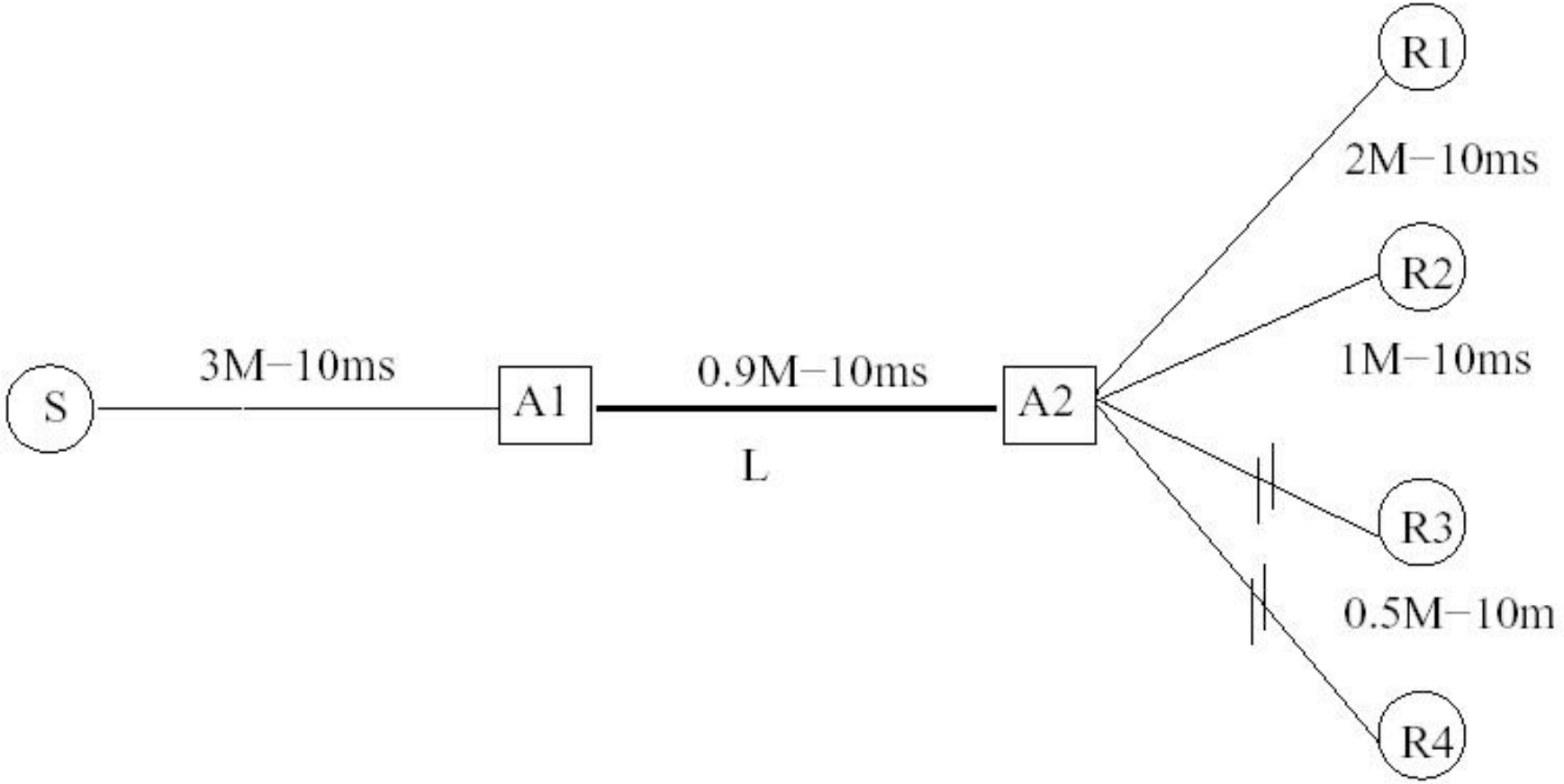
Un schéma de réplication par les récepteurs

Extension de AMCA : choix pratiques

- Un routeur actif maintient des informations locales sur la structure de l'arbre de régulation
- Il notifie les réplicateurs par leur désignation
- Un réplicateur envoie son flux à son sous-groupe via son routeur actif qui se charge d'aiguiller le flux à l'ensemble des récepteurs correspondants
- AMCA : choix de 2 sous-groupes par routeur actif afin de limiter l'*overhead* dû à la gestion de l'arbre de régulation

Un schéma de réplication par les récepteurs

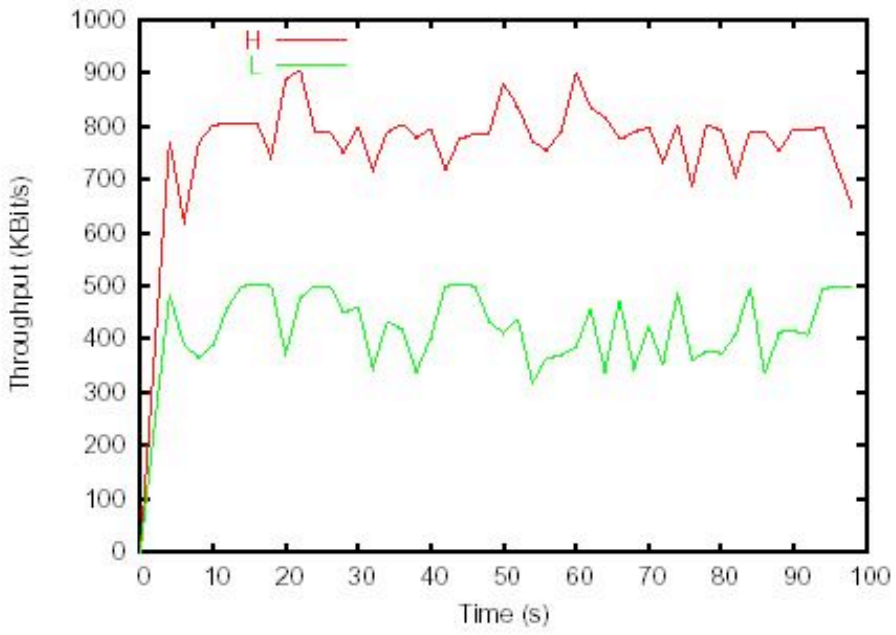
Extension de AMCA : topologie de simulation



Un schéma de réplication par les récepteurs

Extension de AMCA : Résultats de simulation

Débits de réception



Conclusion et perspectives

Conclusions

- Ce travail portait sur les problèmes de niveau transport pour le multicast fiable :
  - Le problème de recouvrement de pertes : DyRAM
  - Le problème de contrôle de congestion : AMCA
  - Le problème de l'hétérogénéité : RbR

Conclusions

- Des services légers au niveau des routeurs ont permis l'amélioration des performances
- Les récepteurs qui font partie d'un même groupe peuvent coopérer pour améliorer les performances des applications :
  - DyRAM : un récepteur peut retransmettre un paquet perdu à un autre
  - RbR : un récepteur peut répliquer et réguler un flux de données à d'autres récepteurs de capacité inférieure

Conclusions

- Les solutions proposées ont fait l'objet de validation à travers des analyses, simulations et implémentations.
- Expérimentations sur une plate-forme locale.
- Expérimentations dans le cadre d'un projet de grille de calcul.

Perspectives

- Implémentation complète de DyRAM/AMCA
  - optimisation de l'implémentation des timers
  - estimation des RTTs
  - contrôle de congestion et de flux
- AMCA
  - à évaluer sur des topologies plus complexes
  - étudier l'impact exact des différents paramètres de l'algorithme

Perspectives

- L'approche RbR :
  - des simulations sur des topologies plus larges
  - évaluer l'impact de la dynamicité
  - considérer les moyens de stockage nécessaires au niveau des réplicateurs
  - la comparer avec une approche multi-couches

Perspectives

- Les problèmes de déploiement :
  - solutions proposées
  - multicast
  - réseaux actifs
- Les problèmes de sécurité
  - inhérents à l'implication des récepteurs
  - réseaux actifs
  - multicast en général

Questions ?