ALNeM status repport

Martin Quinson[1], Arnaud Legrand[2]

1: AlGorille, LORIA 2: MESCAL, ID Laboratory

31 mai 2006

Objectif de la présentation : État des lieux

Task 1.1 : Modeling and discovery mechanisms for stable platforms

Plan de la présentation

- (Re)situer les travaux
- Présenter notre compréhension actuelle du problème
- Présenter l'état d'avancement logiciel
- Donner des pistes pour l'avenir

Introduction

Motivation de ces travaux

- De nombreux services sont co-localisés
  (Content Distribution Networks, Application Service Providers)
- De clients/serveur ($N \times 1$) vers clients/serveurs ($N \times M$), voire le pair-à-pair
- Il y a des choses intelligentes à faire
  - Placement des repplicats à proximité des demandeurs
  - Exécution des calculs sur des serveurs proches des données
  - Sélection du bon serveur de stockage pour les données intermédiaires

$\Rightarrow$ Il s'agit de Distributed Network Aware Applications

$\Rightarrow$ Elles ont besoin de «connaître la topologie du réseau»

Objectif : «découvrir la topologie du réseau»

- Trouver un formalisme pour décrire le réseau
- Trouver un algorithme permettant d'instancier le modèle
  mesures + reconstruction

Que cherche-t-on ?

Données

- Un **ensemble de machines** $\mathcal{H}$
  - Des serveurs sur lesquels on peut faire des choses
  - Pas de machines du coeur du réseau
  - Typiquement, les desktops du labo + les frontales de cluster + ...

Résultat attendu

- Un **graphe** interconnectant les nœuds de $\mathcal{H}$
  - La représentation doit rester de taille raisonnable (gérable)
    (mais on s'autorise à ajouter des nœuds au besoin)
  - Le routage physique parfois «étonnant»
    $\Rightarrow$ Il faut adjoindre une fonction de routage
    $Route(a, b) = \{$ensemble des liens unitaires empruntés entre a et $b\}$
  - L'étiquetage des liens représente la métrique recherchée
    bande passante limitante, taux de perte, etc

Tomographie

Principe

- Imagerie médicale : On fait N images 2D et on les combine en une vue 3D
- Analogie réseau : N mesures des hôtes de $\mathcal{H}$, et on combine en vue globale

Inconvénient : approche bottom-up

Repose souvent sur une vision très bas niveau du réseau
(fusion de traceroute par exemple)

- Fusion difficile (problème d'aliasing)
- Graphes difficiles à étiqueter (latence, bande passante)
- Parties inutiles difficiles à identifier
- Graphes grande taille difficiles à utiliser

Metric Induced Network Topologies (1/2)

Métriques classiques

- Latence (RTT), Bande passante limitante, taux de perte
- Avantages : faciles à mesurer, sans privilège

Une approche plus haut niveau : top-down

- On cherche graphe vérifiant contraintes d'une matrice de distances (mesures de la métrique entre chaque couple)
- On ne rajoute des noeuds et des arêtes que si c'est nécessaire.

Inconvénient

- $lat(A, C) = lat(A, B) + lat(B, C)$ peu probable car mesures réelles
- Solution triviale : graphe complet annoté par la métrique
- On veut un graphe de taille «raisonnable»

Metric Induced Network Topologies (2/2)

Formalisation

- On se donne une métrique (latence, bande passante) et un opérateur d'agrégation ($+$, min) pour évaluer la valeur d'une route
- On se donne une certaine tolérence ϵ et on cherche un graphe avec le moins d'arêtes possible tel que pour tout A, B :

$$|valeur(A,B)_{graphe} - metrique(A,B)_{mesure}| < \epsilon$$

Pour résumer : MINT

- Notion de graphe routé et étiqueté vérifiant les contraintes induites par les mesures sur une métrique réseau, à une tolérance près
- Acronyme proposé par [Bestavros et Al. 2002],

2002 est fondamentalement bottom-up mais objectifs *similaires*

- Ils proposent également un framework à la fois théorique et pratique

ENV

Effective Network View [Shao, Berman, Wolski (UCSD) 1999]

Motivation

- Outil utilisable en pratique, pas de théorème
- Objectif : ordonnancement maître/esclaves

Méthode

- Mesures actives
- Raffinements successifs
  - Host to host : maître / esclave puis clustering
  - Pairwise : maitre avec deux esclaves à la fois pour couper les clusters
  - Internal : pour estimer les capacités de chaque cluster
  - Jammed : deux internes au cluster (hub ou switch)

Problèmes

- Outil sans preuve
- Vision en arbre uniquement
- Fusion reste difficile, il faut procéder différemment

Interference-centric approach

[Legrand, Mazoit, Quinson 2003]

Une nouvelle métrique

- Latence et BP pas suffisantes pour les communications collectives
- Notion d'interférences entre (AB) et (CD) de $\mathcal{H}$:

  Est ce que la BP sur (AB) varie selon que (CD) est saturé ou non ?
- Notation : non $\rightarrow (AB)//(CD)$; oui $\rightarrow (AB) \between (CD)$
- Dans le graphe : $(AB)//(CD) \iff (A \rightarrow B) \cap (A \rightarrow B) = \emptyset$

Un nouvel algorithme

- Prouvé optimal s'il existe un arbre ou une clique d'arbres
- Extensions pour les cycles

De nouveaux problèmes

- Matrice des mesures en N^4 au lieu de N^2 (+problèmes d'obtention)

Ce que l'on comprend

Le problème

- Plusieurs approches
- Travail de formalisation pas terminé
- Aucune idée de la complexité des problèmes d'optimisation associés

Les solutions

- Plusieurs solutions dans la litérature
- Pas de moyen de les comparer

C'est quoi «un bon algorithme de découverte de la topologie» ?

Ce que l'on comprend

Le problème

- Plusieurs approches
- Travail de formalisation pas terminé
- Aucune idée de la complexité des problèmes d'optimisation associés

Les solutions

- Plusieurs solutions dans la litérature
- Pas de moyen de les comparer

C'est quoi «un bon algorithme de découverte de la topologie» ?

$\Rightarrow$ évaluation sur simulateur (étonnant, non ?)

ALNeM status repport

- Introduction
- Le problème et les solutions existantes
- Application-Level Network Mapper
- Conclusion

Premier semestre 2006 : deux stages M2R

Darina Dimitrova (Grenoble, avec Arnaud)

- Implémenter les méthodes de reconstruction existantes dans SimGrid
  - Clustering (de latence et de bande passante)
  - Arbre recouvrant minimal (latence et bande passante)
  - ENV
  - ALNeM
- Évaluation «graphique» (on lance les découvreurs, et on regarde)
- Rapport à rendre lundi passé

Ahmed Harbaoui (Nancy, avec moi)

- Implémenter les mesures nécessaires à la reconstruction
- Mettre au point un banc de test quantitatif pour les méthodes
- Rapport à rendre jeudi prochain

Infrastructure logicielle

Vue d'ensemble

- Mesures sont réalisées a priori puis stockées dans un MySQL
- Les algos de reconstruction vont piocher dans la base

Un mot sur les mesures

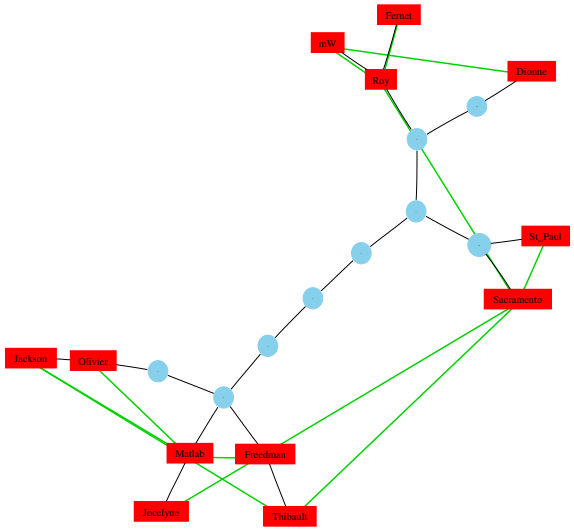
Implémentées avec GRAS : API de comm avec deux implémentations

- Le simulateur SimGrid
- *in situ* sur plate-forme réelle en utilisant les sockets BSD

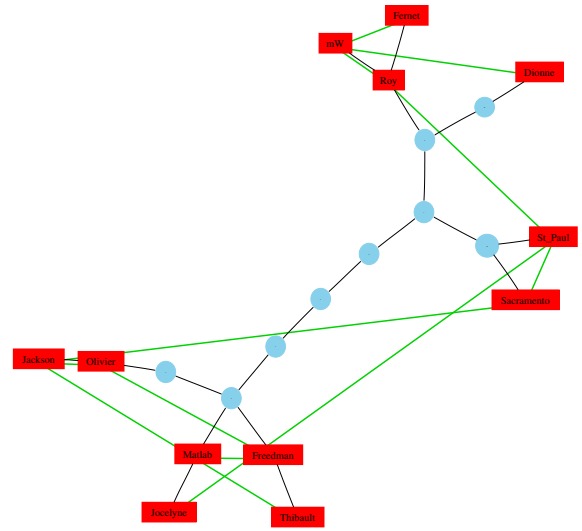
Objectif à terme

- Faire un outil utilisable «pour de vrai»
- Offrir différents algos de reconstruction

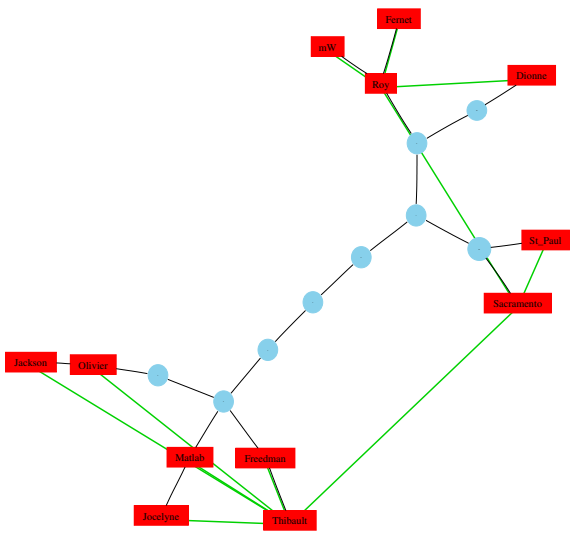
Résultats de Darina (1/3) : $|V| = 21$; $|\mathcal{H}| = 12$



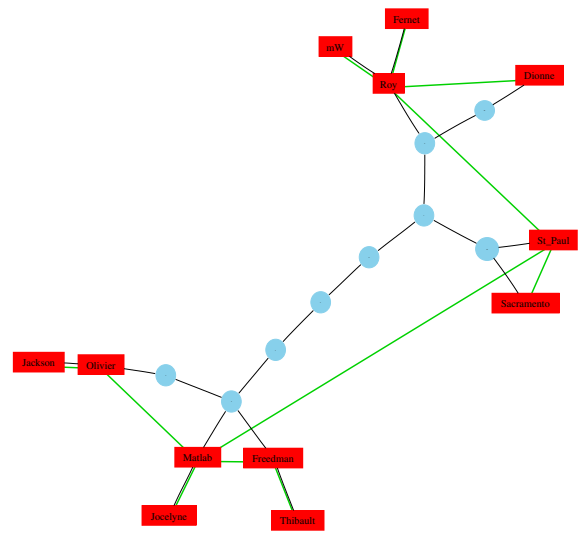
Clustering de bande passante



Arbre recouvrant maximal de BP

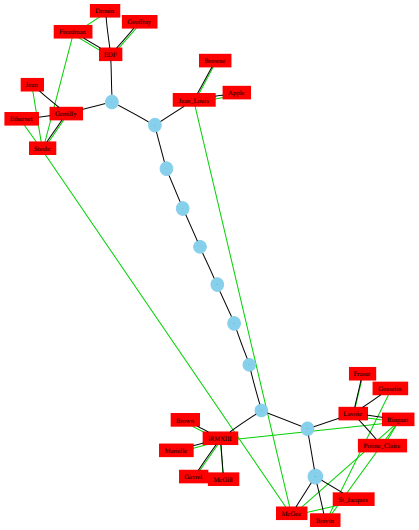


Clustering de latence

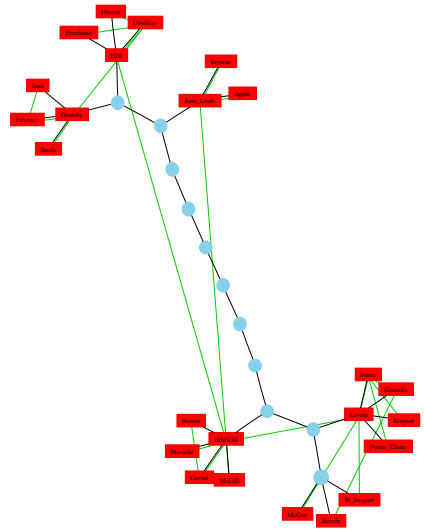


AR minimal de latence

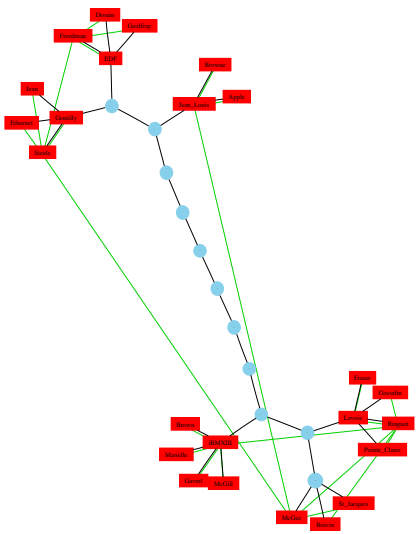
Résultats de Darina (2/3) : $|V| = 36$; $|\mathcal{H}| = 24$



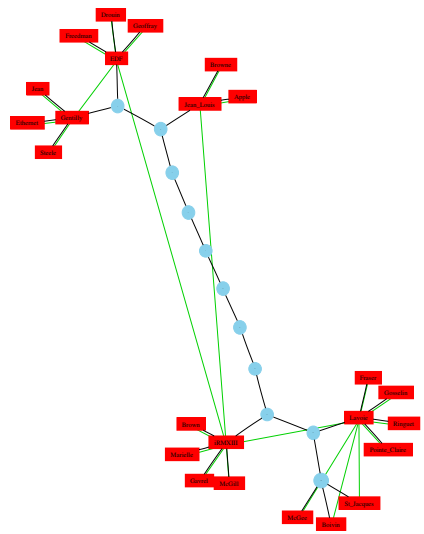
Clustering de bande passante



Arbre recouvrant maximal de BP

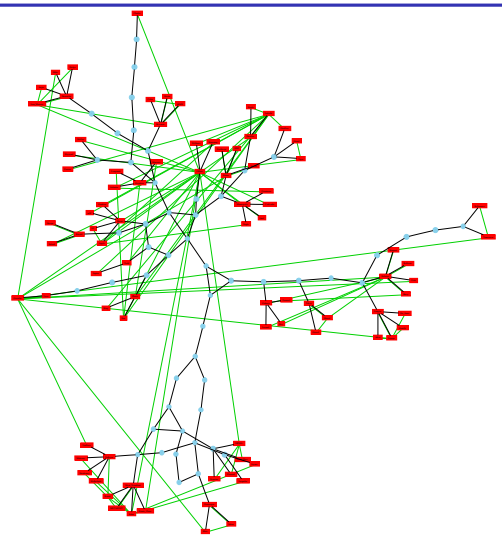


Clustering de latence

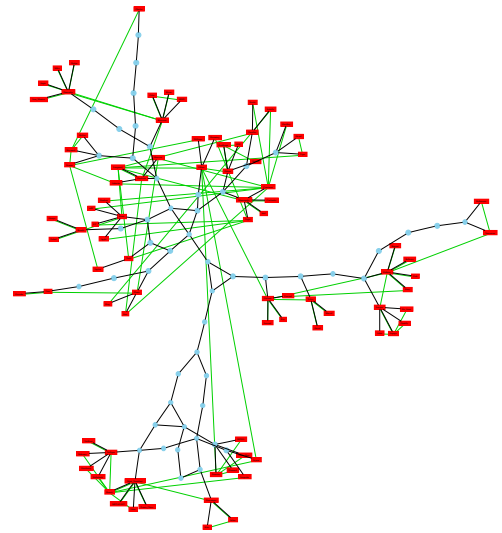


AR minimal de latence

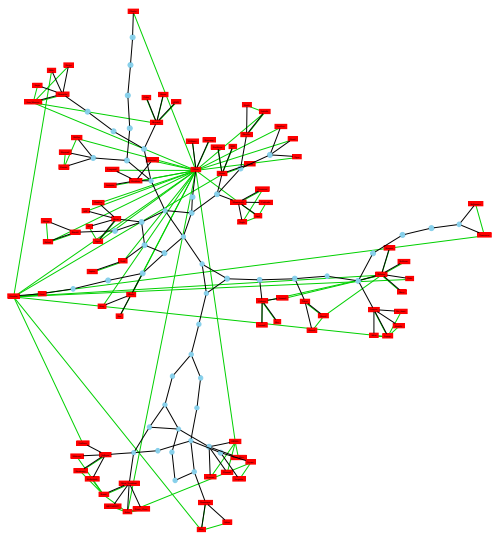
Résultats de Darina (3/3) : $|V| = 129$; $|\mathcal{H}| = 88$



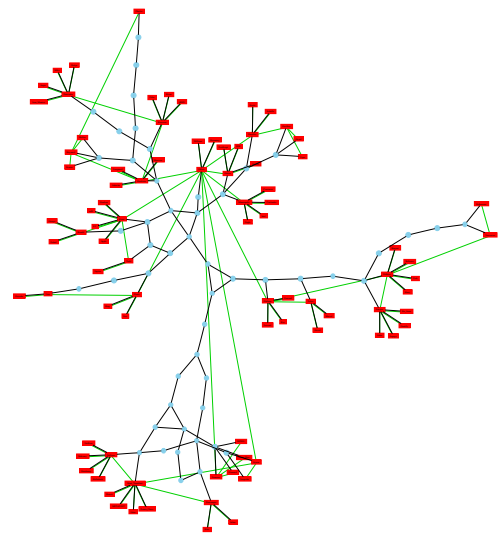
Clustering de bande passante



Arbre recouvrant maximal de BP

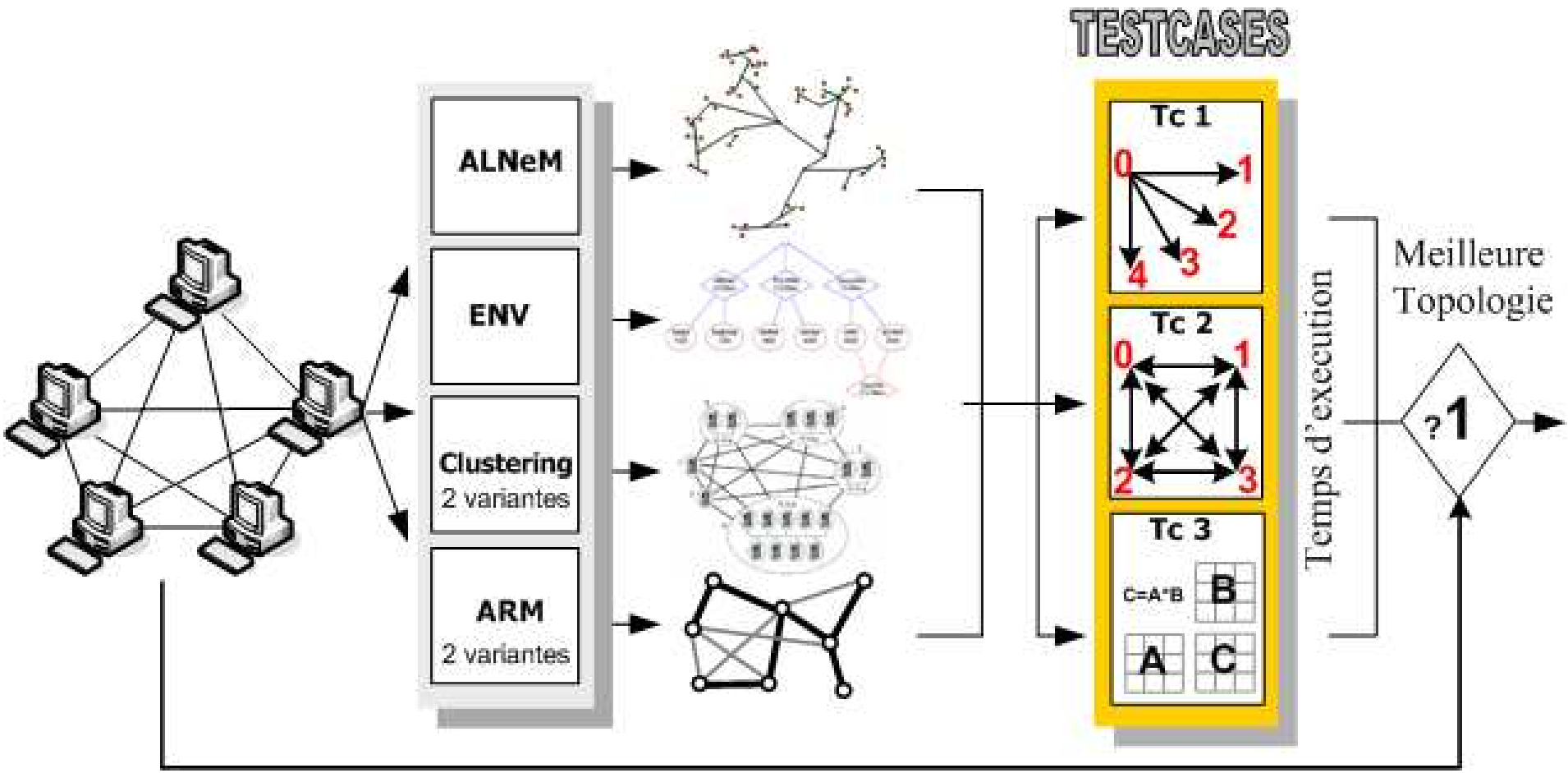


Clustering de latence

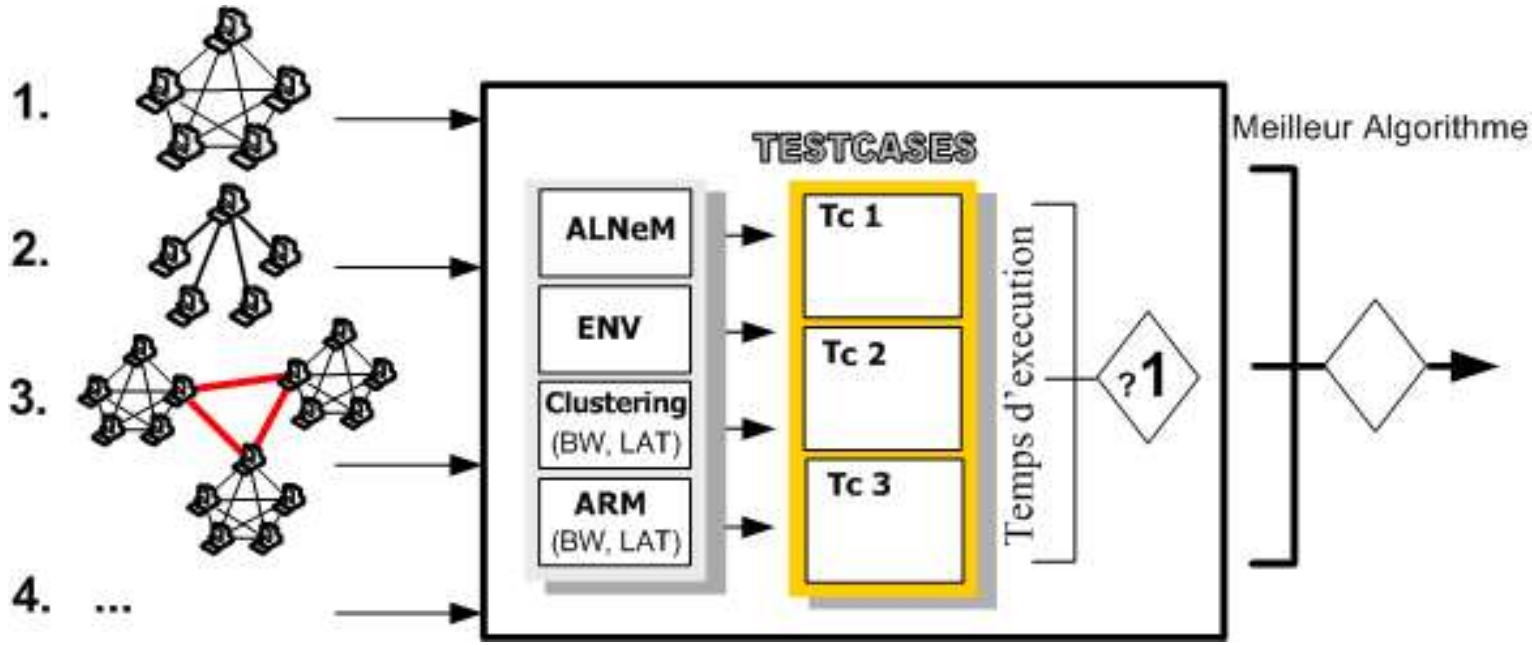


AR minimal de latence

Ahmed (1/3) : Banc de test quantitatif

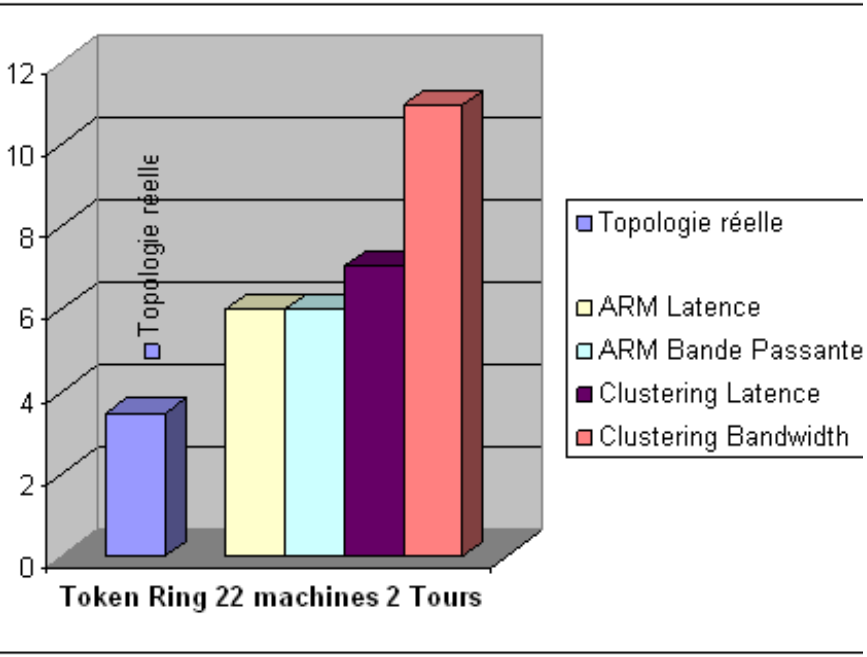


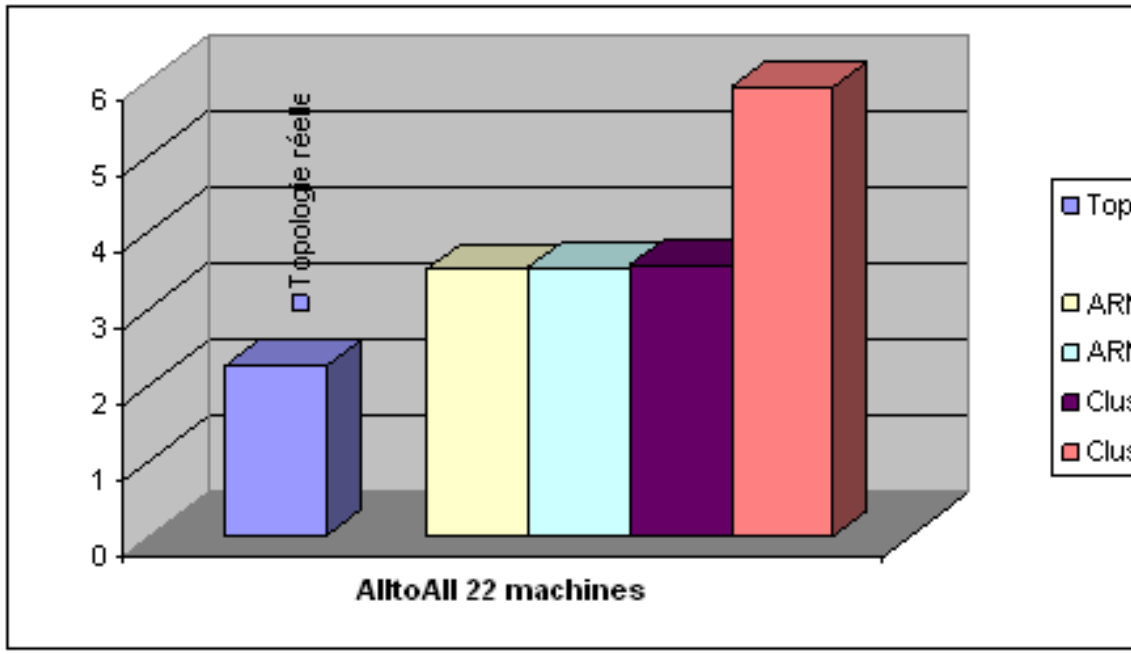
Ahmed (2/3) : Banc de test quantitatif

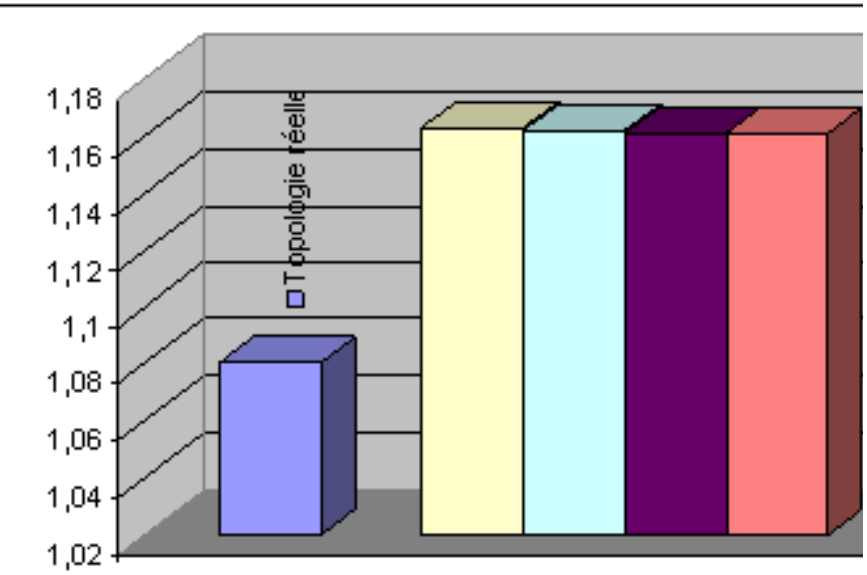


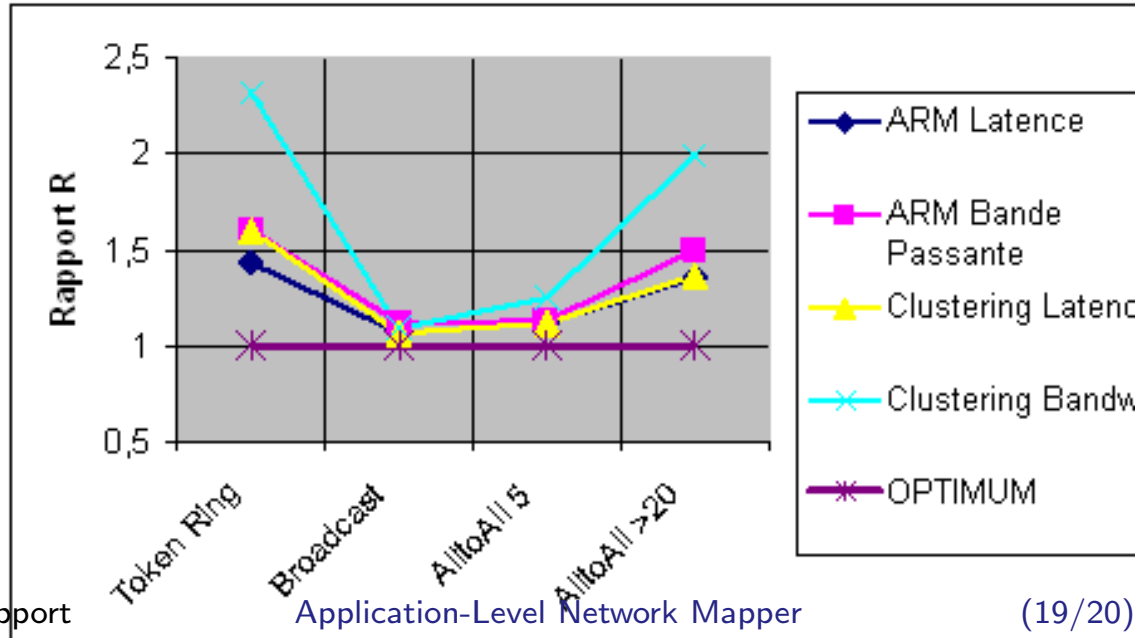
	TR	Bcast	All2All$_5$	All2All$_{100}$	PMM_{500}	PMM_{5000}
Topologie	Anneau	Arbre	Clique	Clique	Hypercube	
Transferts //	Non	Non	Oui	Oui	Oui	Oui
Contention	Non	Non	Non	Oui	Non	Oui
# Étapes	1	1	1	1	N	N

Ahmed (3/3) : les premiers résultats









Conclusion

Choses bien

- On a bossé à l'est ! (si ! si !)
- Le framework (mesures+comparaison résultats) est prêt
- Quelques cas d'écoles implémentés

Choses restant à faire

- Il manque quelques algos existant pour les comparaisons
- Vérifier les implémentations et les résultats
- Expérimentations sur plate-formes réelles (quantifier la qualité ?)

- **Il est temps de travailler à l'algo de découverte ultime**
  - On a toutes les billes pour jouer, maintenant

Conclusion

Choses bien

- On a bossé à l'est ! (si ! si !)
- Le framework (mesures+comparaison résultats) est prêt
- Quelques cas d'écoles implémentés

Choses restant à faire

- Il manque quelques algos existant pour les comparaisons
- Vérifier les implémentations et les résultats
- Expérimentations sur plate-formes réelles (quantifier la qualité ?)

- **Il est temps de travailler à l'algo de découverte ultime**
  On a toutes les billes pour jouer, maintenant

Conclusion

Choses bien

- On a bossé à l'est ! (si ! si !)
- Le framework (mesures+comparaison résultats) est prêt
- Quelques cas d'écoles implémentés

Choses restant à faire

- Il manque quelques algos existant pour les comparaisons
- Vérifier les implémentations et les résultats
- Expérimentations sur plate-formes réelles (quantifier la qualité ?)

- **Il est temps de travailler à l'algo de découverte ultime**
  On a toutes les billes pour jouer, maintenant

Conclusion

Choses bien

- On a bossé à l'est ! (si ! si !)
- Le framework (mesures+comparaison résultats) est prêt
- Quelques cas d'écoles implémentés

Choses restant à faire

- Il manque quelques algos existant pour les comparaisons
- Vérifier les implémentations et les résultats
- Expérimentations sur plate-formes réelles (quantifier la qualité ?)

- **Il est temps de travailler à l'algo de découverte ultime**
  On a toutes les billes pour jouer, maintenant

Lionel arrive bientôt en post-doc