Scatter

Definition

- One process sends different messages to each other process
  - Data flow is identical to that of MPI_Gather

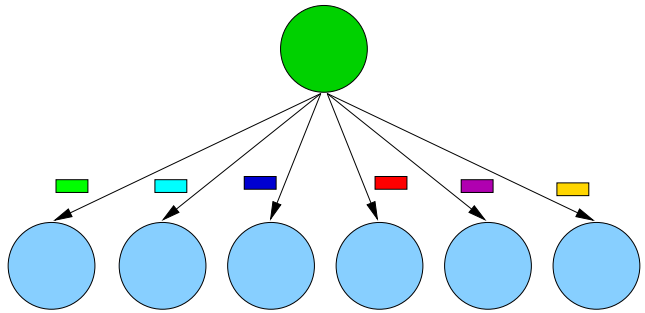
Scatter

Definition

- One process sends different messages to each other process
  - Data flow is identical to that of MPI_Gather

Strategies

- Flat Trees - default
- Binomial Trees
- Chains



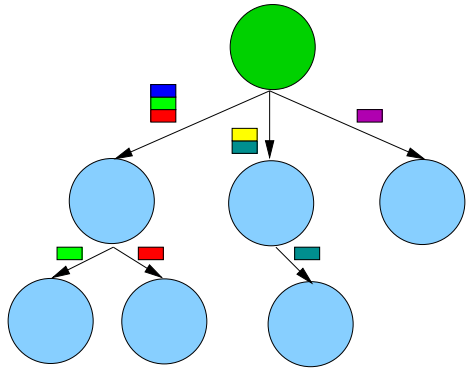
Scatter

Definition

- One process sends different messages to each other process
  - Data flow is identical to that of MPI_Gather

Strategies

- Flat Trees - default
- Binomial Trees
- Chains



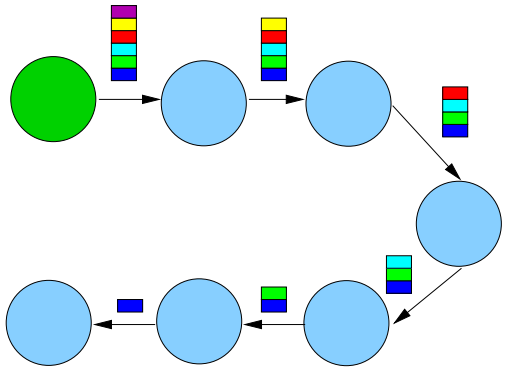
Scatter

Definition

- One process sends different messages to each other process
  - Data flow is identical to that of MPI_Gather

Strategies

- Flat Trees - default
- Binomial Trees
- Chains



Modelling the Scatter Operation

Modelling Scatter

- Flat Tree - every process is directly contacted by the *root*
- Other strategies - messages are relayed through auxiliary nodes

Implementation Technique	**Communication model**
Flat Tree	$(P-1) \times g(m) + L$
Chain	$\sum_{j=1}^{P-1} g(j \times m) + (P-1) \times L$
Binomial Tree	$\leq \lceil log_2 P \rceil \times L + \sum_{j=0}^{\lfloor log_2 P \rfloor - 1} g(2^j \times m)$ $\approx \lceil log_2 P \rceil \times L + (P-1) \times g(m)$

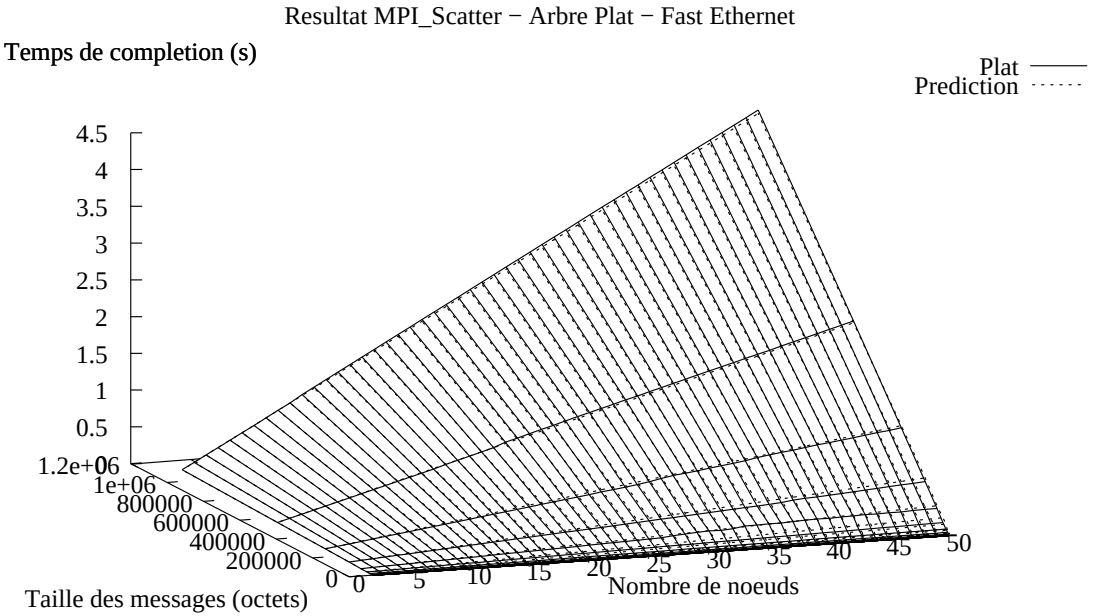
Modelling the Scatter Operation

Modelling Scatter

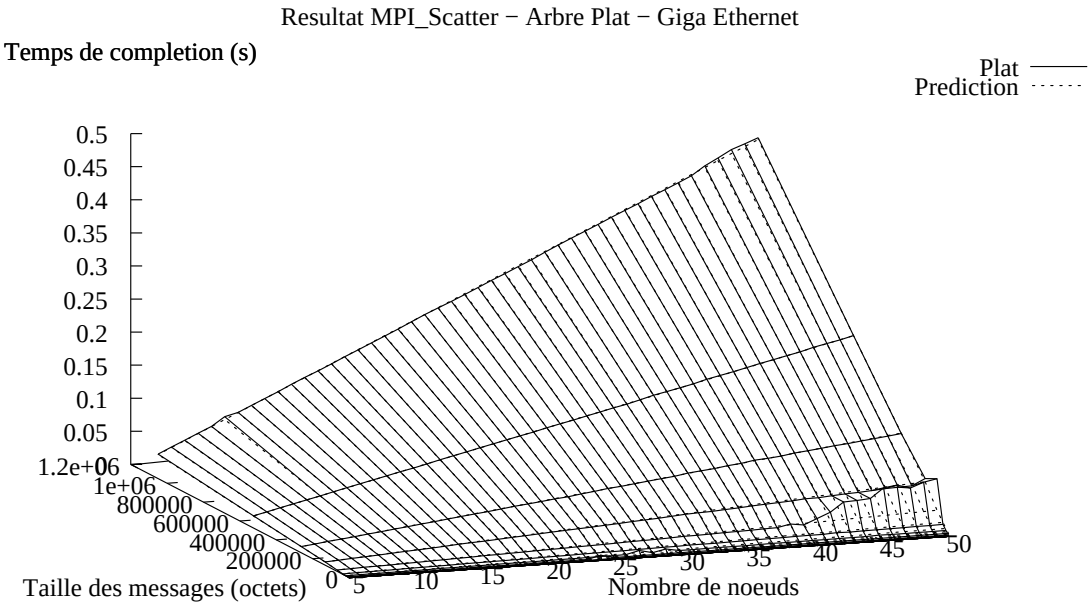
- Flat Tree - every process is directly contacted by the *root*
- Other strategies - messages are relayed through auxiliary nodes

Implementation Technique	Communication model
Flat Tree	$(P-1) \times g(m) + L$
Chain	$\sum_{j=1}^{P-1} g(j \times m) + (P-1) \times L$
Binomial Tree	$\leq \lceil log_2 P \rceil \times L + \sum_{j=0}^{\lfloor log_2 P \rfloor - 1} g(2^j \times m)$ $\approx \lceil log_2 P \rceil \times L + (P-1) \times g(m)$

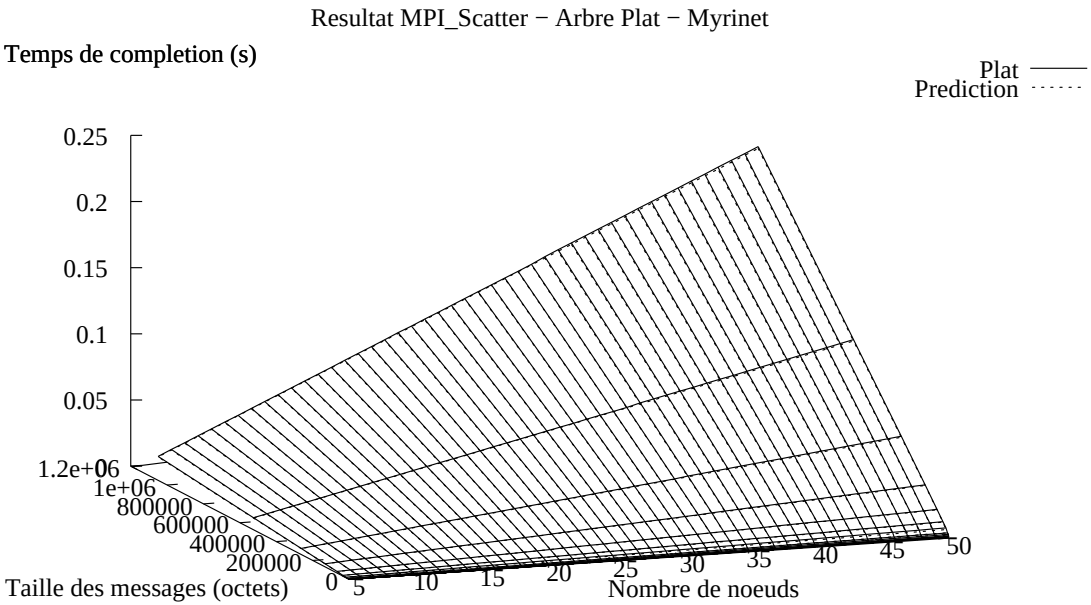
Validating the model - Fast Ethernet



Validating the model - Giga Ethernet



Validating the model - Myrinet



Scatter for Grid Environments

Problem

- Optimisation is harder than Broadcast
- Multi-layered scheduling implies extra efforts

Scatter for Grid Environments

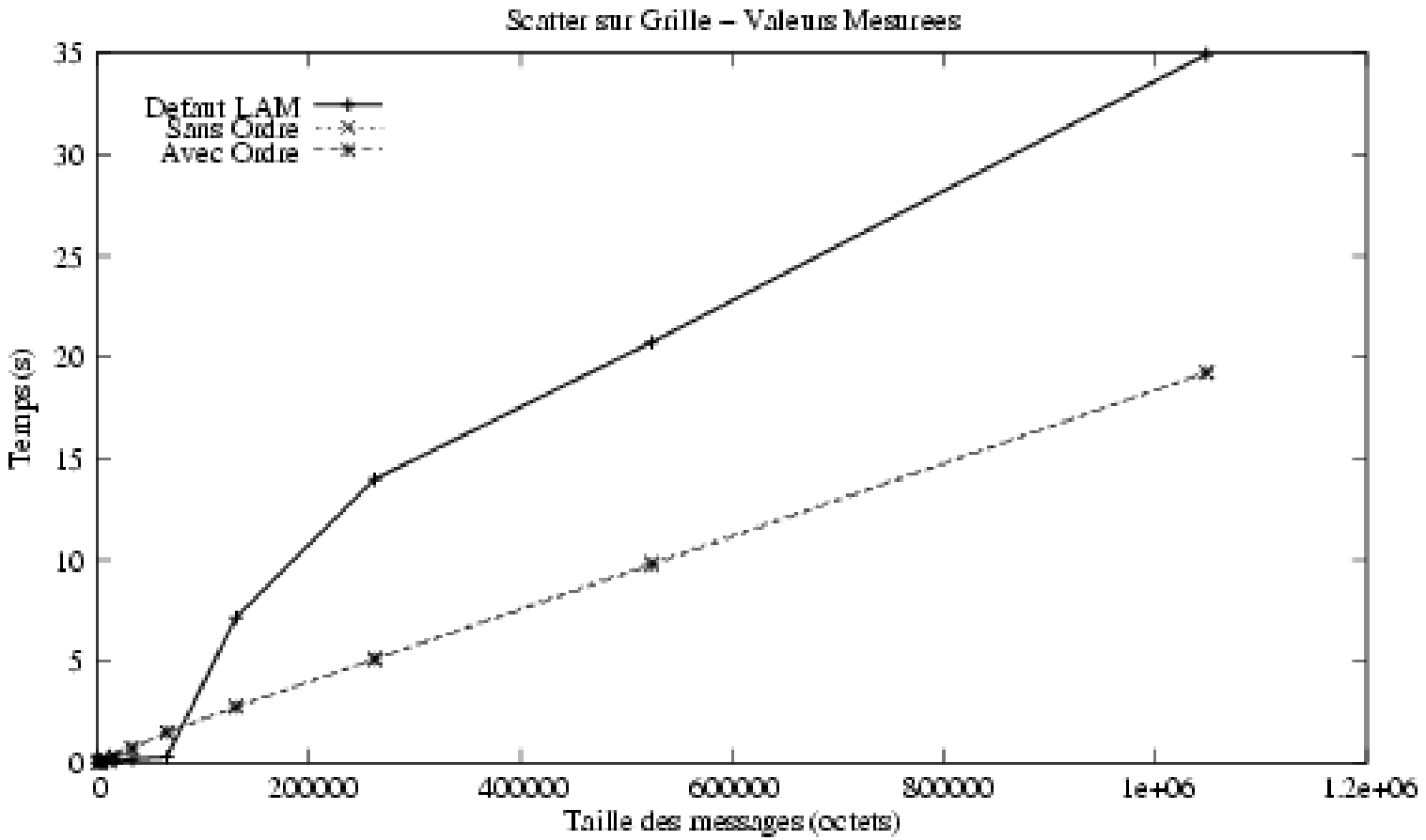
Problem

- Optimisation is harder than Broadcast
- Multi-layered scheduling implies extra efforts

Simple Strategies

- Direct Connexion - standard MPI approach
- Flat Trees with message packing (two layers)
- Flat Trees with order scheduling

Comparison


Scatter sur Grille – Valeurs Mesurees
Defaut LAM
Sans Ordre
Avec Ordre
Temps (s)
Taille des messages (octets)
0
5
10
15
20
25
30
35
0
200000
400000
600000
800000
1e+06
1.2e+06

Summary

Efficient with large messages

- minimises the startup cost on a grid connexion

Future works

- evaluate the impact of memory swap
- improve local area data multiplication (parallel senders)

All-to-All

Personalised Many-to-Many communication pattern

- Every process send a different message to each other

All-to-All

Personalised Many-to-Many communication pattern

- Every process send a different message to each other

Implementation Strategies

- By default, each message is directly sent to the destinations

All-to-All

Personalised Many-to-Many communication pattern

- Every process send a different message to each other

Implementation Strategies

- By default, each message is directly sent to the destinations
  - may saturate the network

All-to-All

Personalised Many-to-Many communication pattern

- Every process send a different message to each other

Implementation Strategies

- By default, each message is directly sent to the destinations
  - may saturate the network
  - contention effects are difficult to model

Modelling All-to-All

Traditional Approach

- Most authors use a model based on the Scatter operation
- Hypothesis: one All-to-All is equivalent to simultaneous Scatters

Modelling All-to-All

Traditional Approach

- Most authors use a model based on the Scatter operation
- Hypothesis: one All-to-All is equivalent to simultaneous Scatters

Drawbacks

- does not supposes network contention
- all processes start communication simultaneously

Our Approach

Augmenting the communication model

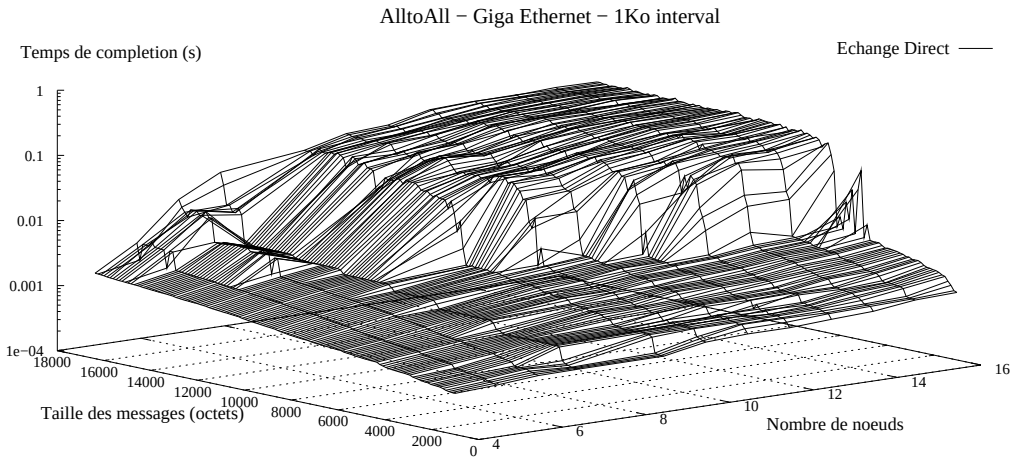
- Hypothesis: Contention is a linear factor that characterises each network
  - Objective: identify a *Contention Signature* - γ
    - Augment the Scatter model with this signature

$$T = \mathit{Free} \times \gamma$$

$$= ((P-1) \times g(m) + L) \times \gamma$$

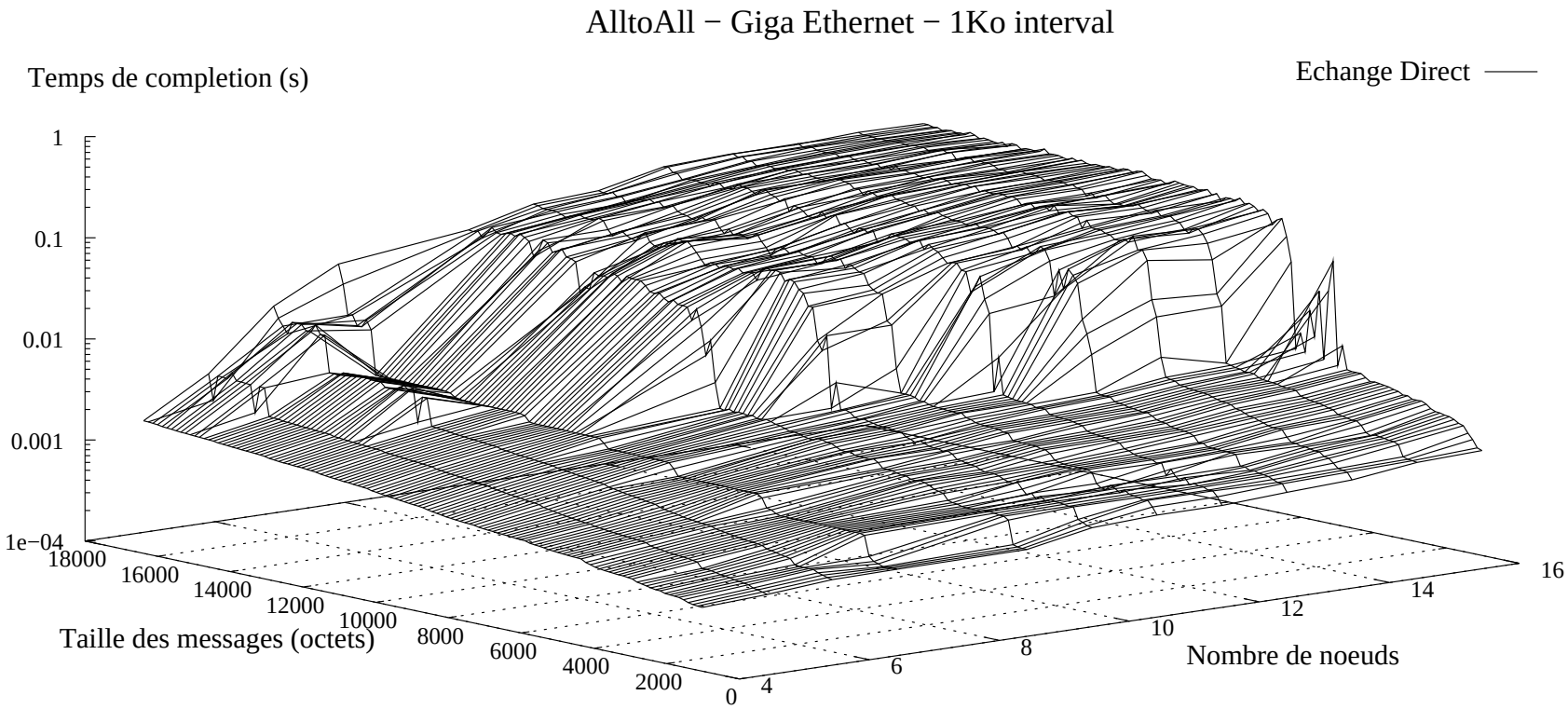
Another Factor

- Buffering, segmentation and memory constraints may impact on the performance
  - Overhead factor δ_P depending on P and m



$$T = ((P-1) \times g(m) + L) \times \gamma + (P-1) \times \delta_P$$

Buffering


AlltoAll − Giga Ethernet − 1Ko interval
Temps de completion (s)
Echange Direct
1
0.1
0.01
0.001
1e−04
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Taille des messages (octets)
4
6
8
10
12
14
16
Nombre de noeuds

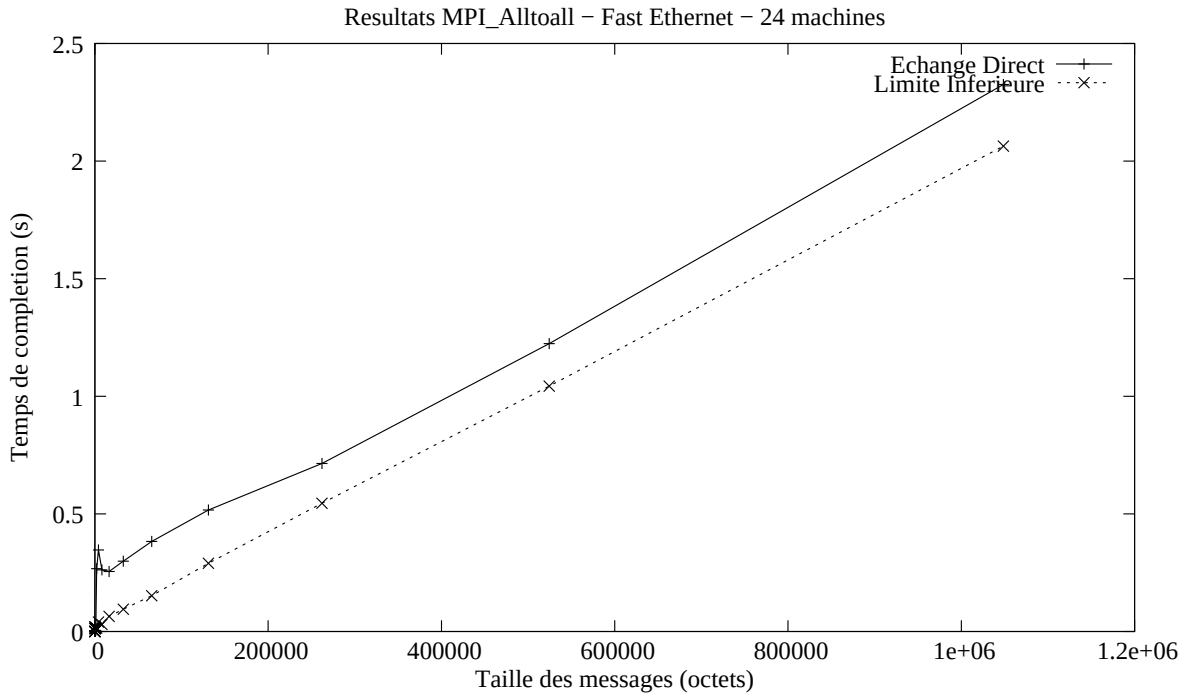
Defining Parameters

To obtain γ et δ_P

- Performance prediction (Scatter model)
- Execution sample
- Parameter fitting with Linear Regression
  - Least Squares method
- Once identified, γ et δ_P are used to correct the predictions

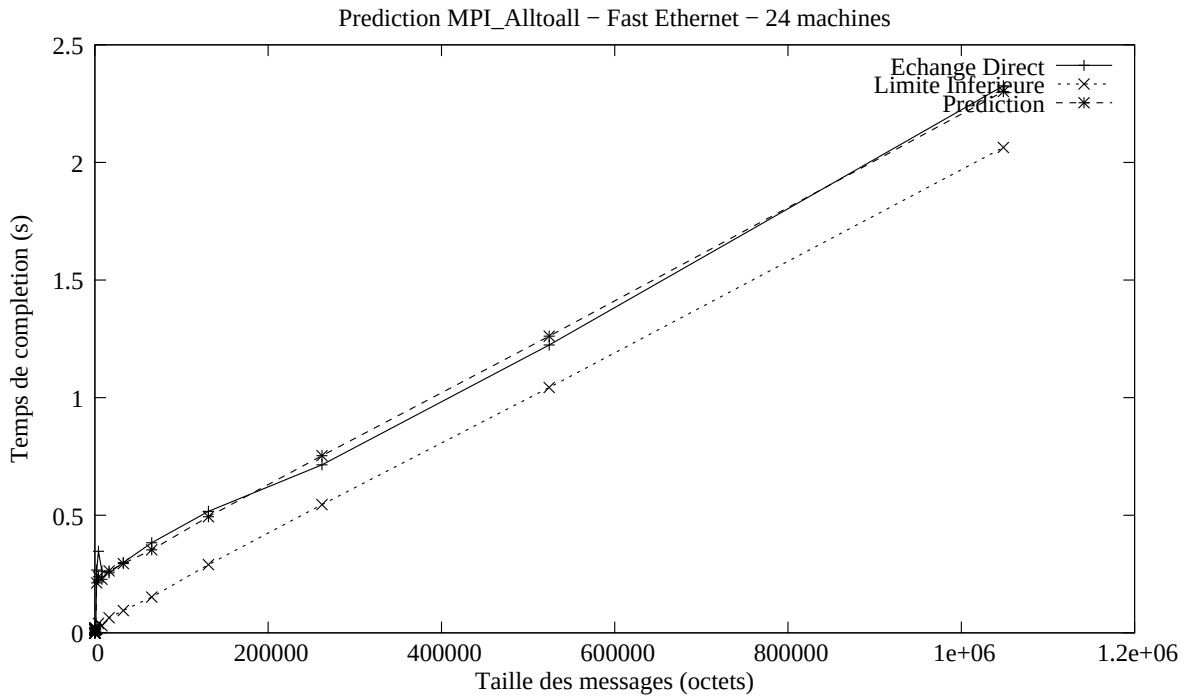
Fast Ethernet Network

- Reduced contention

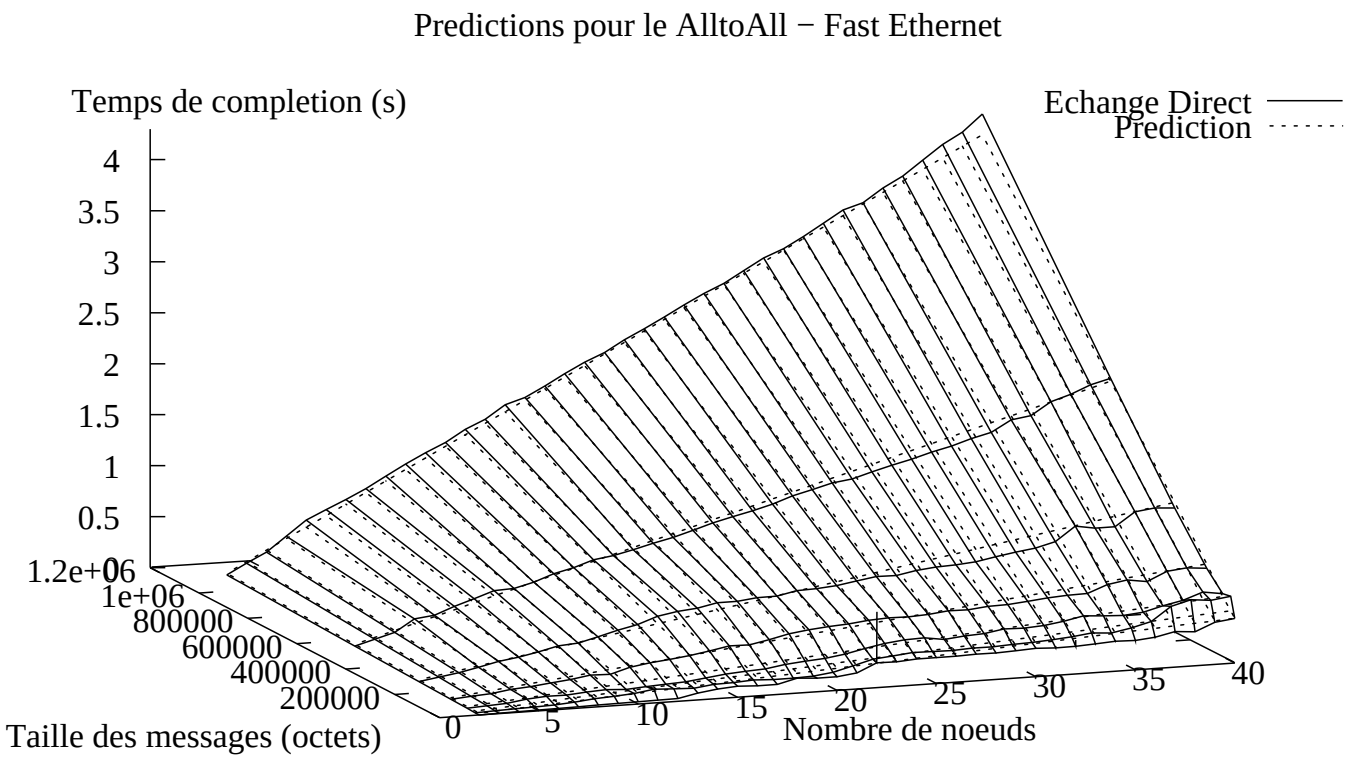


Fast Ethernet Network

- γ = 1.0195
- δ = 8.23 ms (for messages larger than 2 KB)

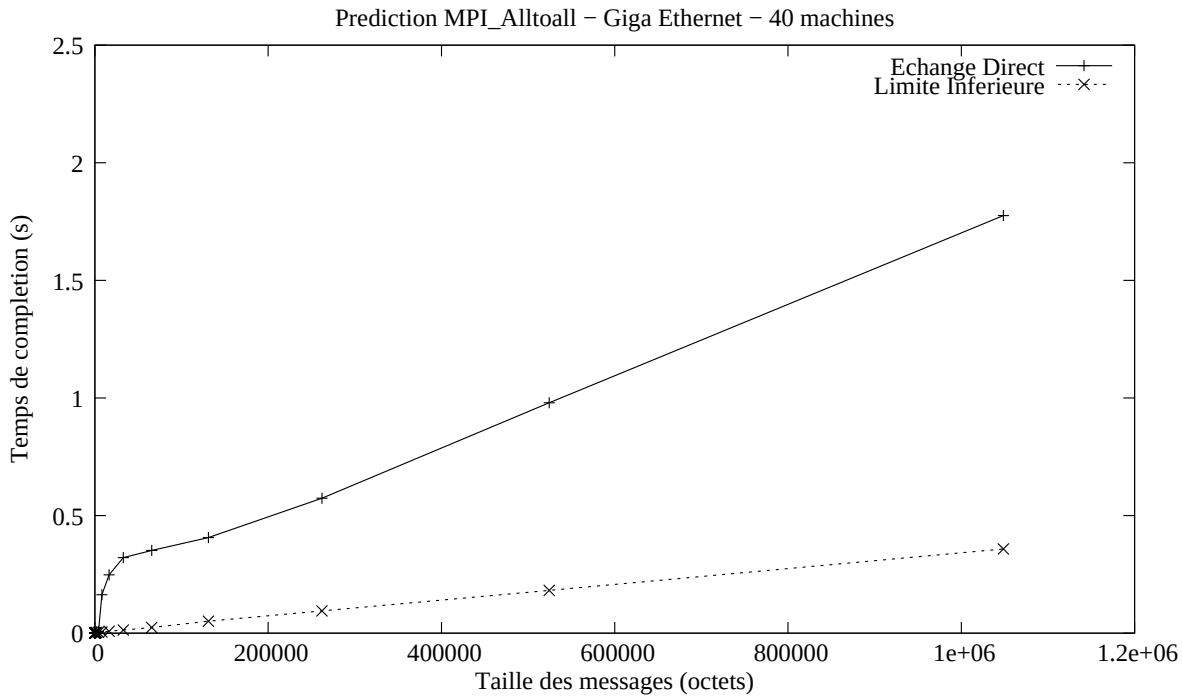


Applying the Model



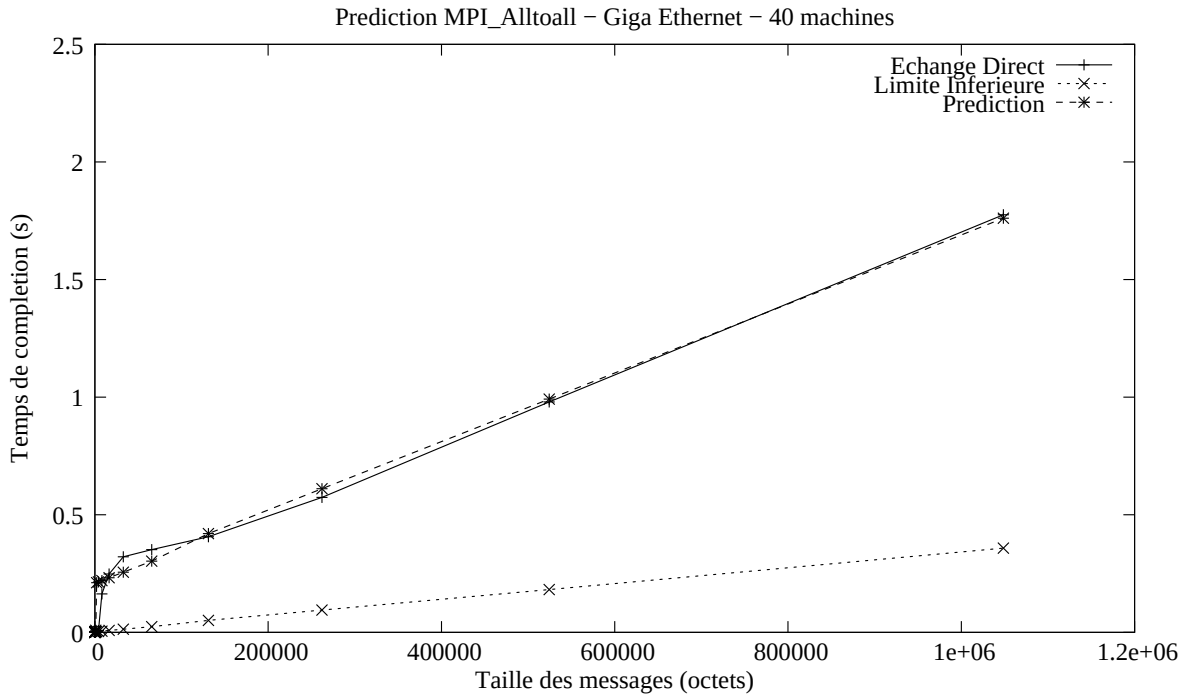
Giga Ethernet Network

- Impact of contention is important
- Buffering plays an important role

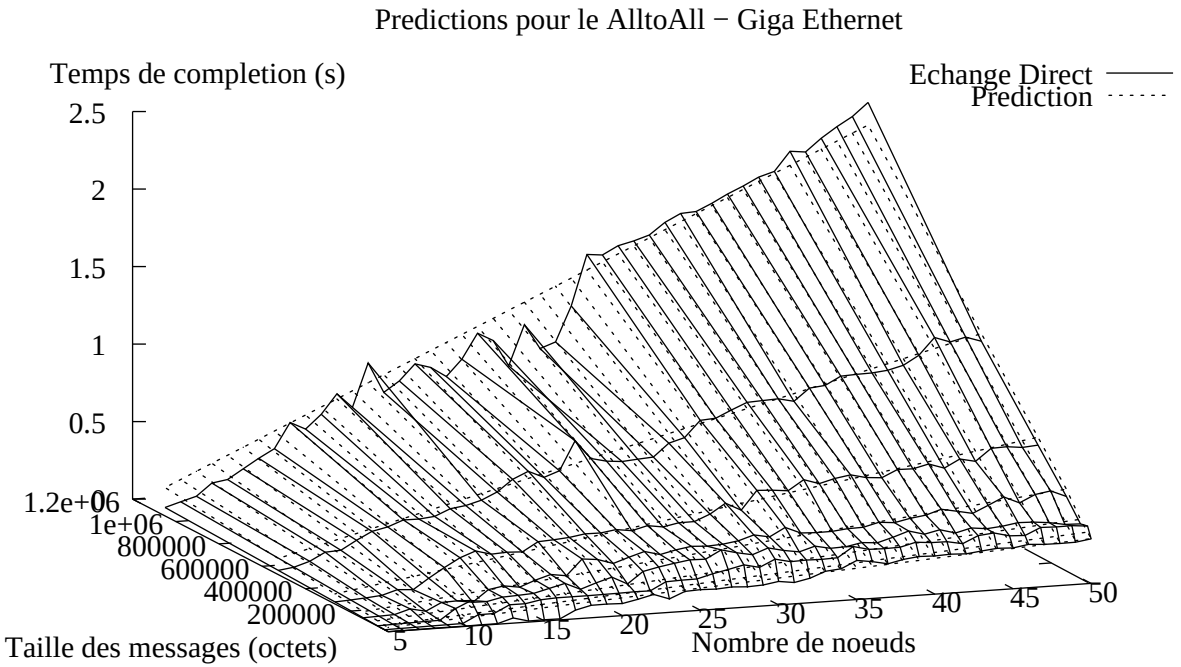


Giga Ethernet Network

- γ = 4.3628
- δ = 4.93 ms (for messages larger than 8KB)

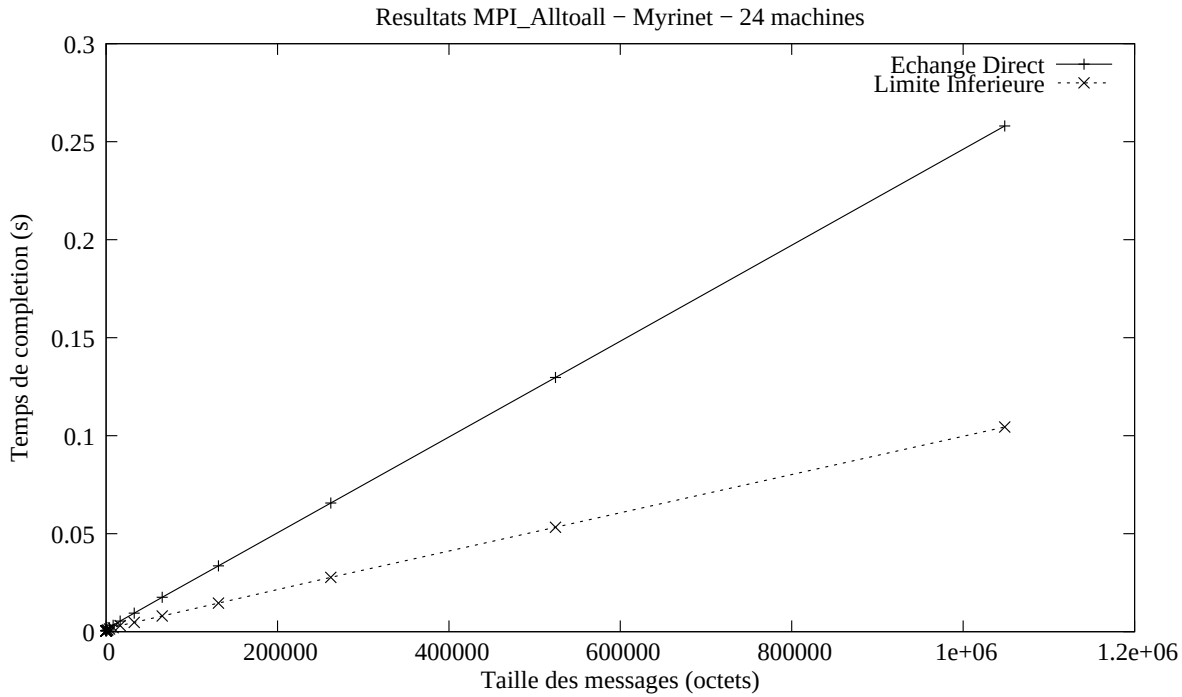


Applying the Model



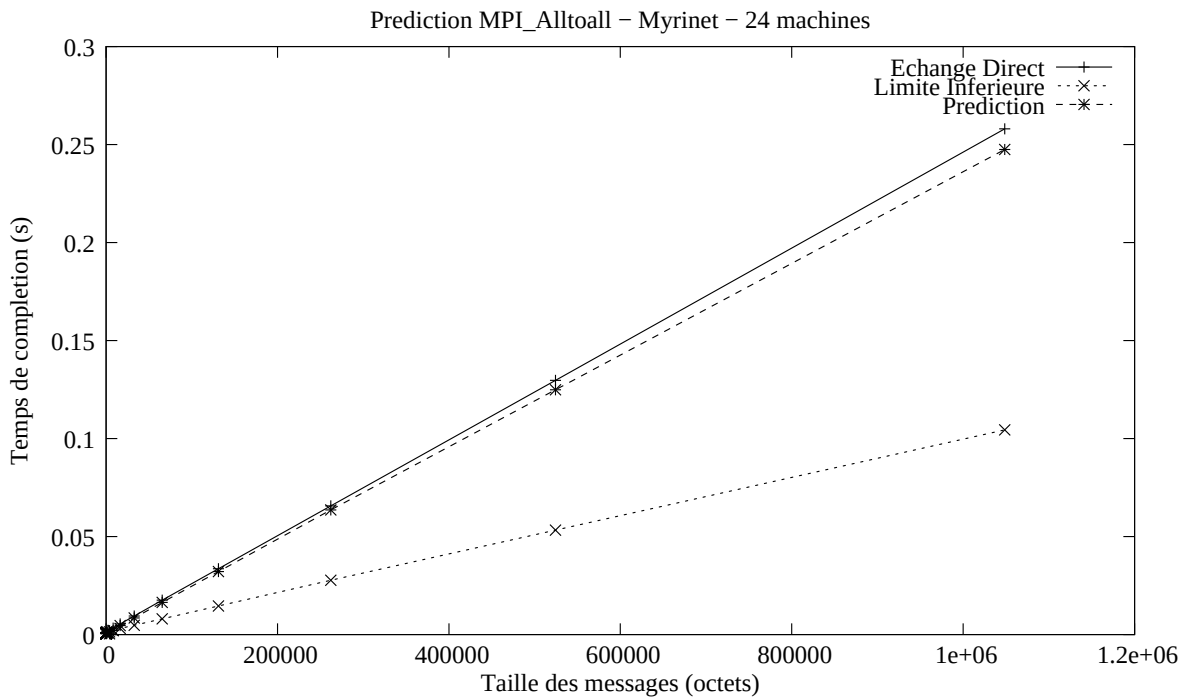
Myrinet Network

- Contention is important
- No buffering side-effects

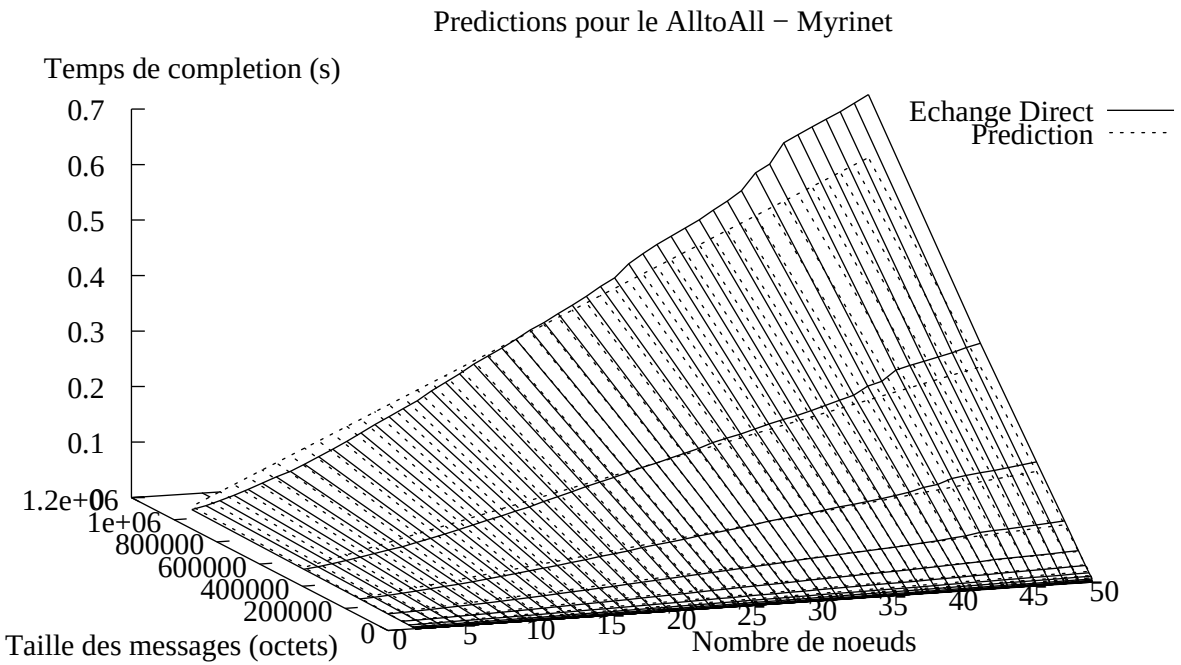


Myrinet Network

- $\gamma = 2.4975$
- $\delta = 0$



Applying the model



AlltoAll

THE hardest problem

- How to explore heterogeneity?

AlltoAll

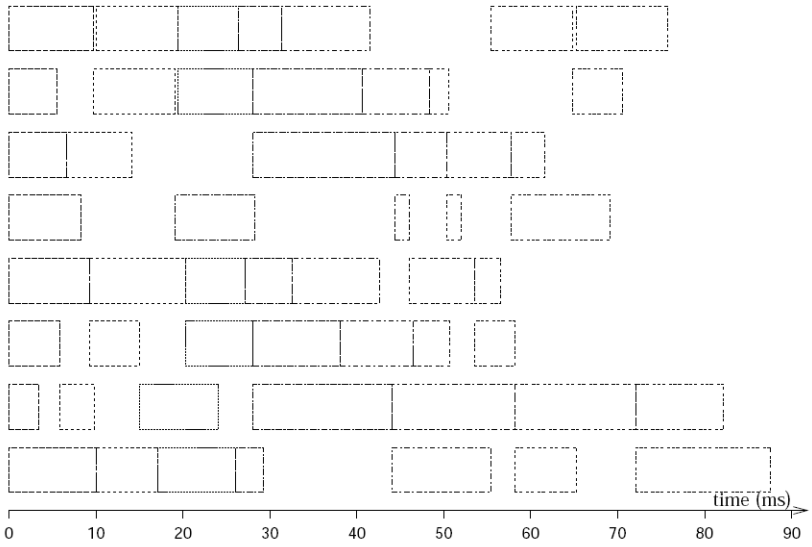
THE hardest problem

- How to explore heterogeneity?

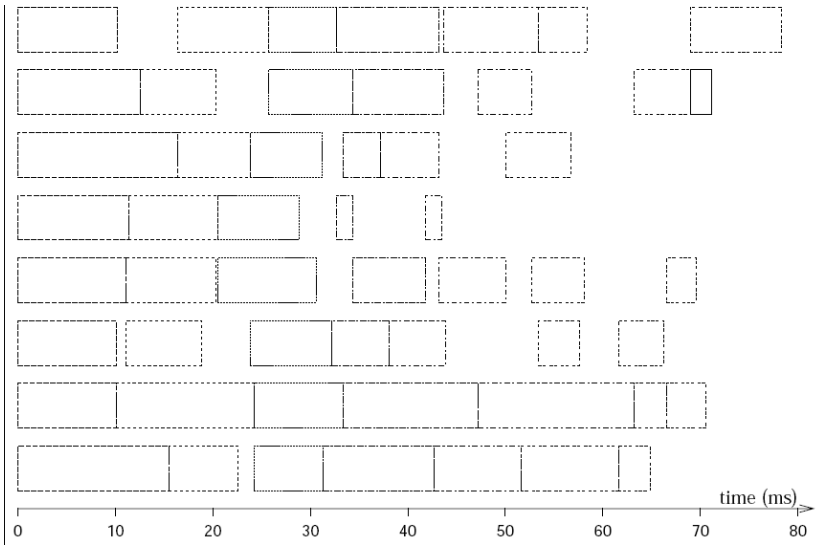
Works on the AlltoAll-v operation

- Messages with different sizes
  - different sizes = different communication times
- Matching heuristics

Motivation



Fixed Pattern



Max-min

Heuristics for the AlltoAll

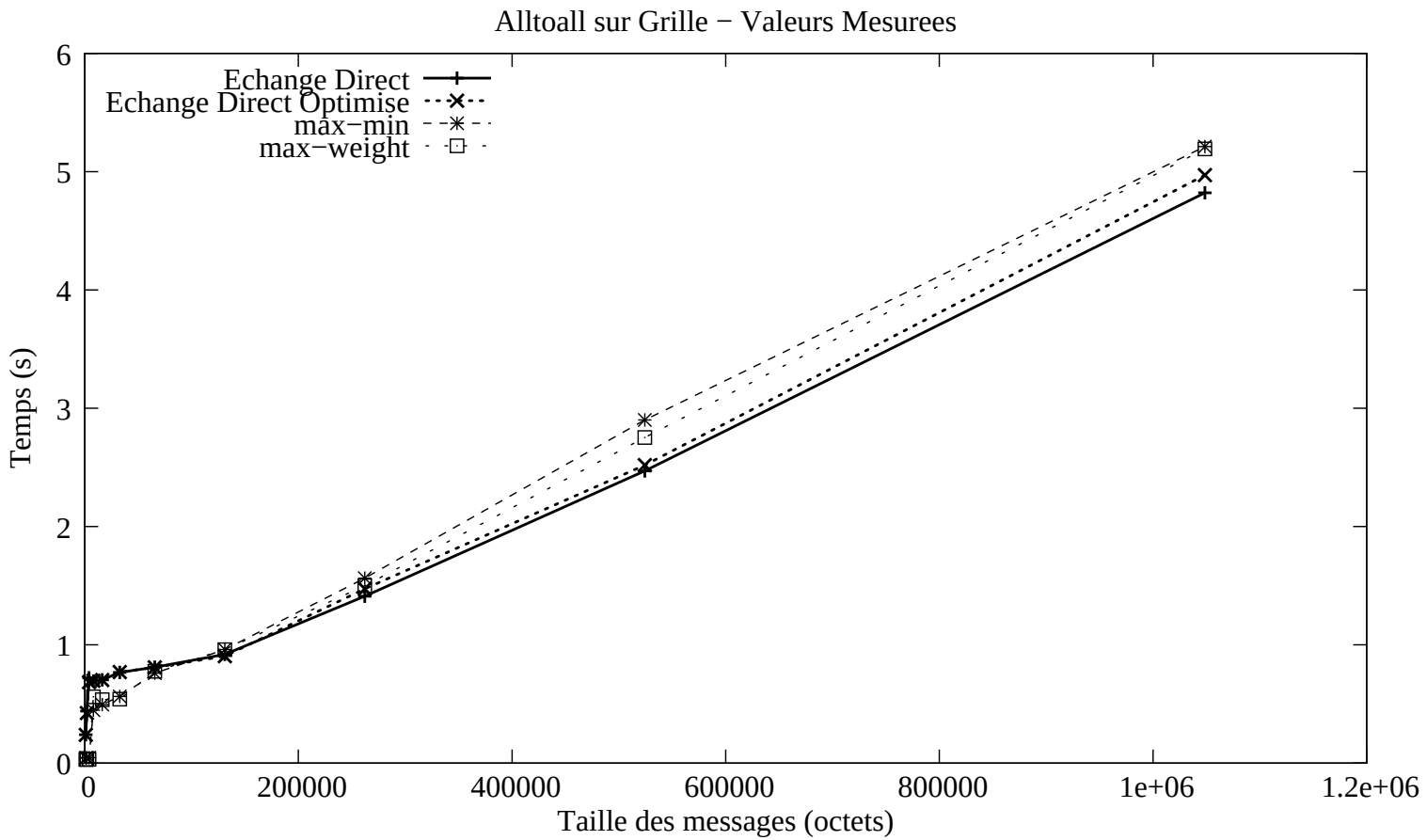
Max-min heuristic

- The minimum communication weight is maximised

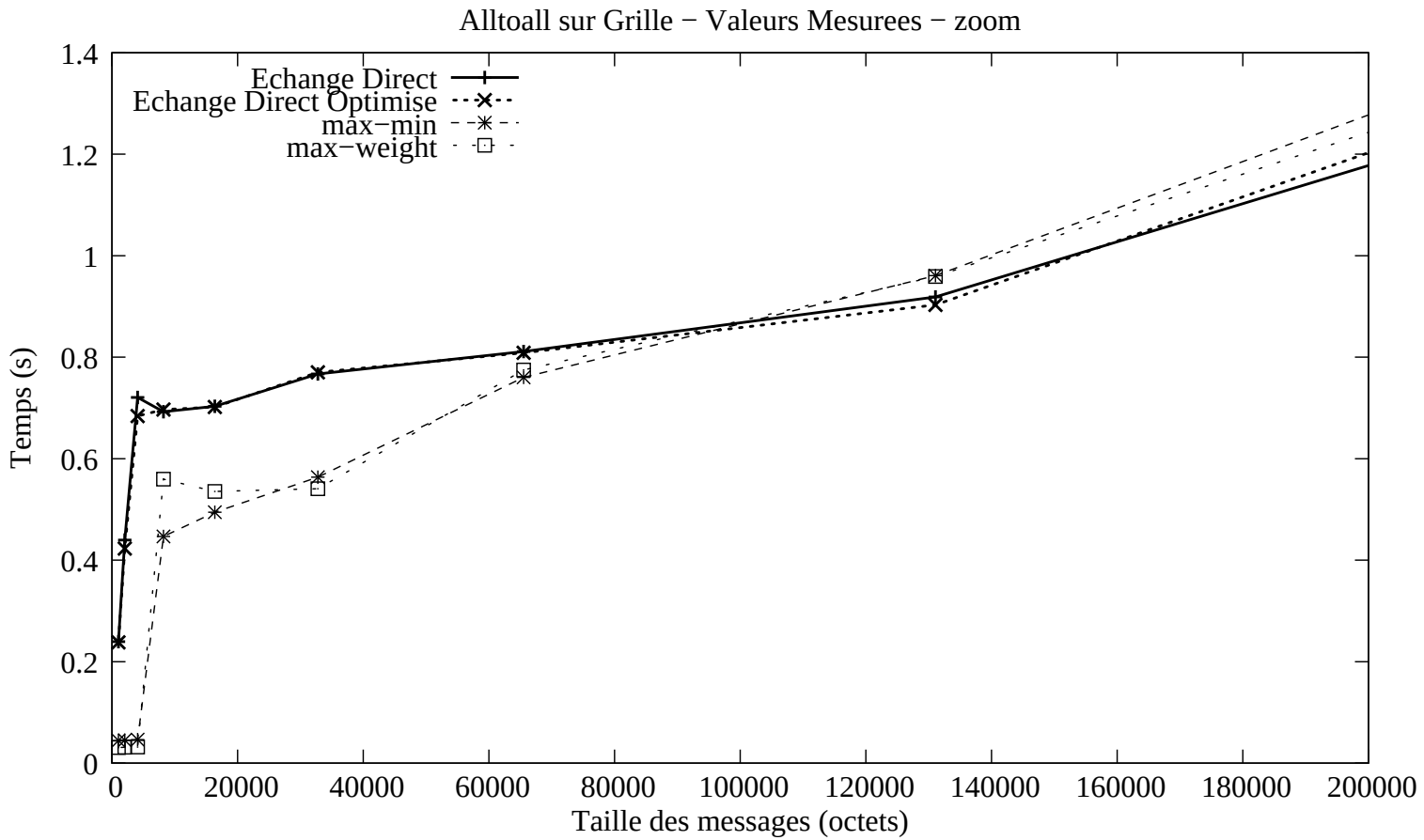
Max-weight heuristic

- The sum of the communication weights is maximised
  - Maximises the data transfer in a round

Comparison


Alltoall sur Grille − Valeurs Mesurees
Echange Direct
Echange Direct Optimise
max−min
max−weight
Temps (s)
Taille des messages (octets)
0
1
2
3
4
5
6
0
200000
400000
600000
800000
1e+06
1.2e+06

Comparison


Alltoall sur Grille − Valeurs Mesurees − zoom
Echange Direct
Echange Direct Optimise
max−min
max−weight
Temps (s)
Taille des messages (octets)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
20000
40000
60000
80000
100000
120000
140000
160000
180000
200000

Summary

- Limited efficiency
  - Contention may prevents any performance improvement
- Suited for small messages

Future Works

- Extend the study with the *All-to-All-v* operation

Available Bandwidth on Wide-Area Connections

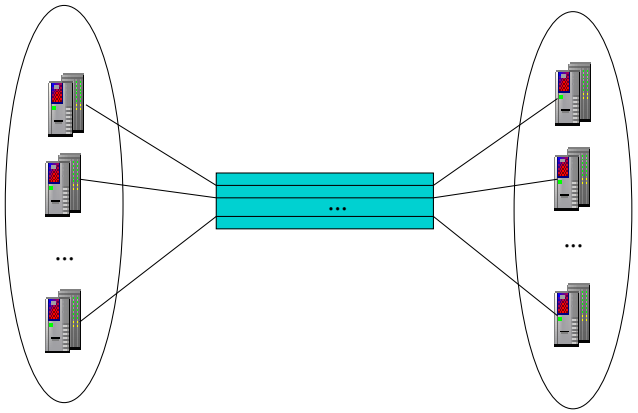
In collaboration with Carlos Barrios (June 2005)

- Modelling massive data transfer rates
- Identify available bandwidth on broadband connexions
- Help to evaluate different tuning parameters

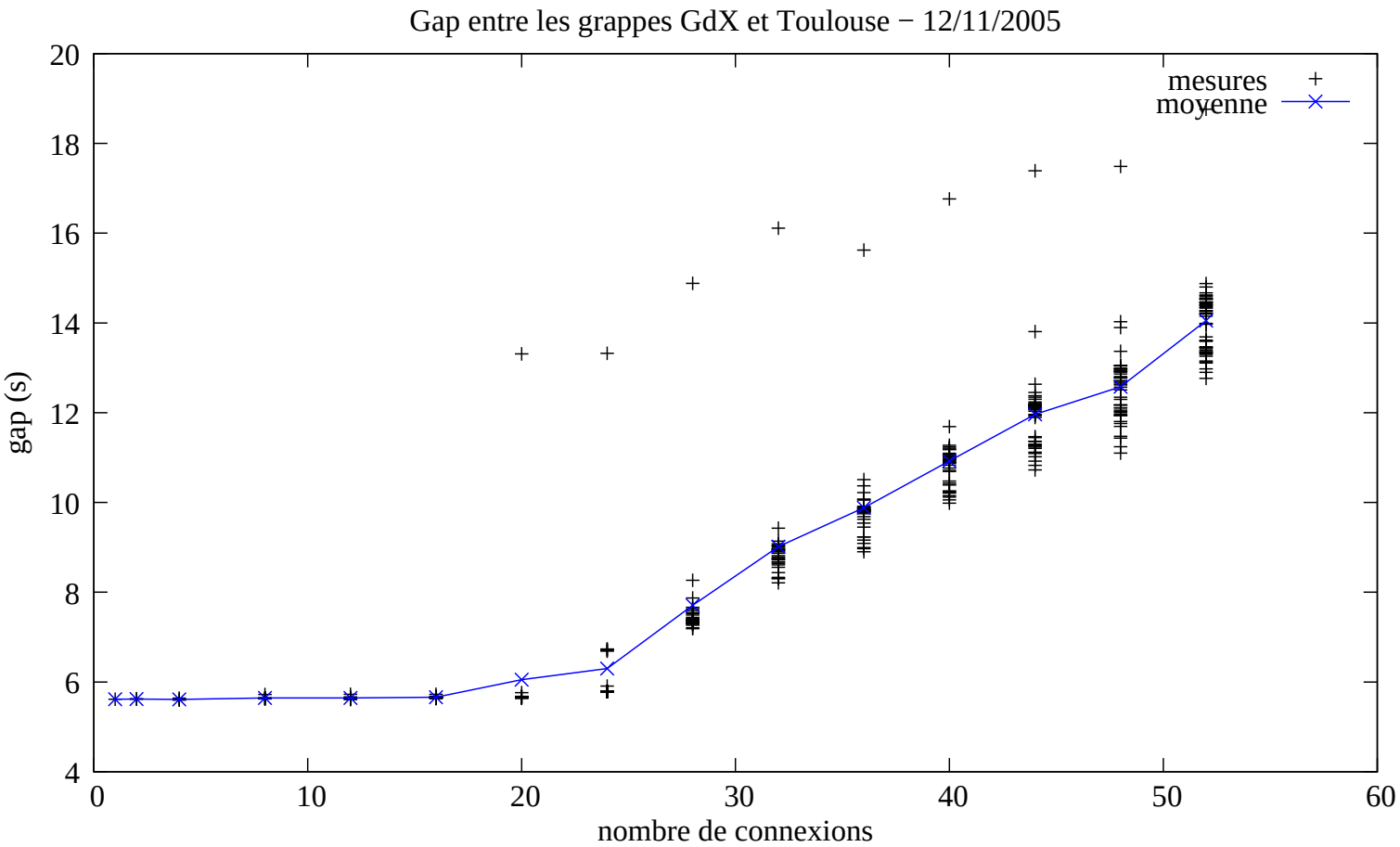
How to saturate a broadband link

Problem: a single connexion is not enough

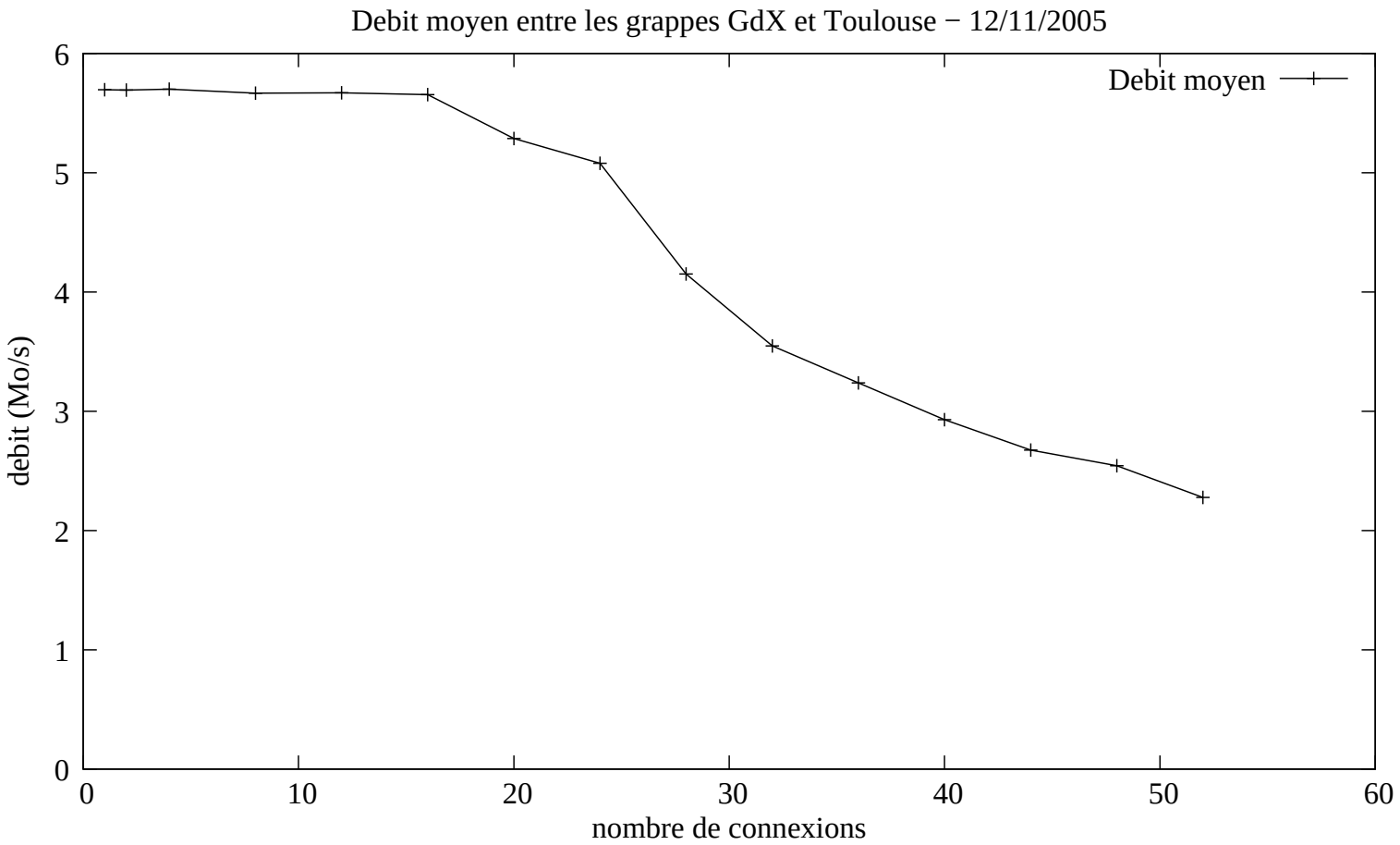
- Simultaneous connexion of several nodes
  - TCP fairness



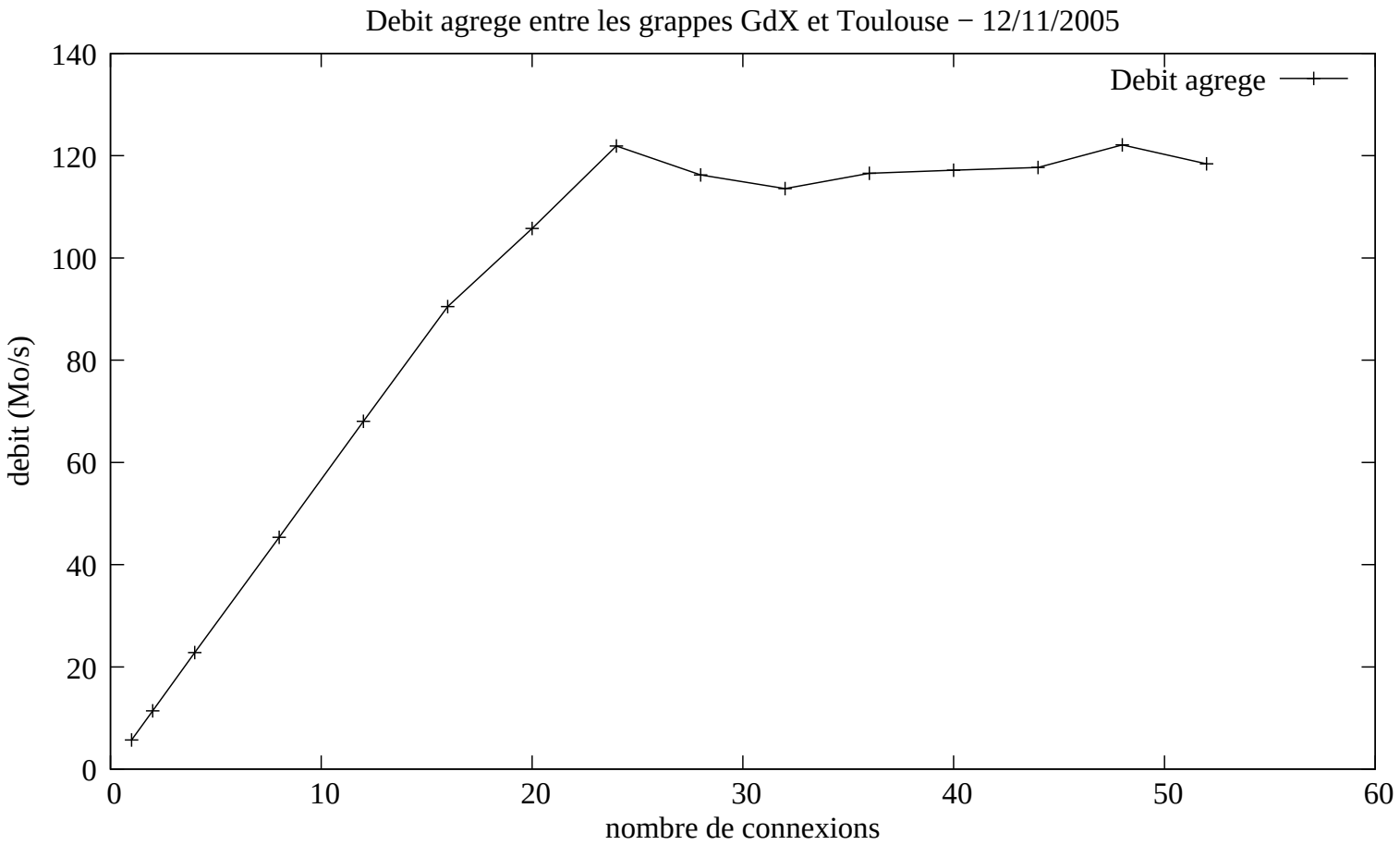
Gap


Gap entre les grappes GdX et Toulouse − 12/11/2005
mesures
moyenne
gap (s)
nombre de connexions
20
18
16
14
12
10
8
6
4
0
10
20
30
40
50
60

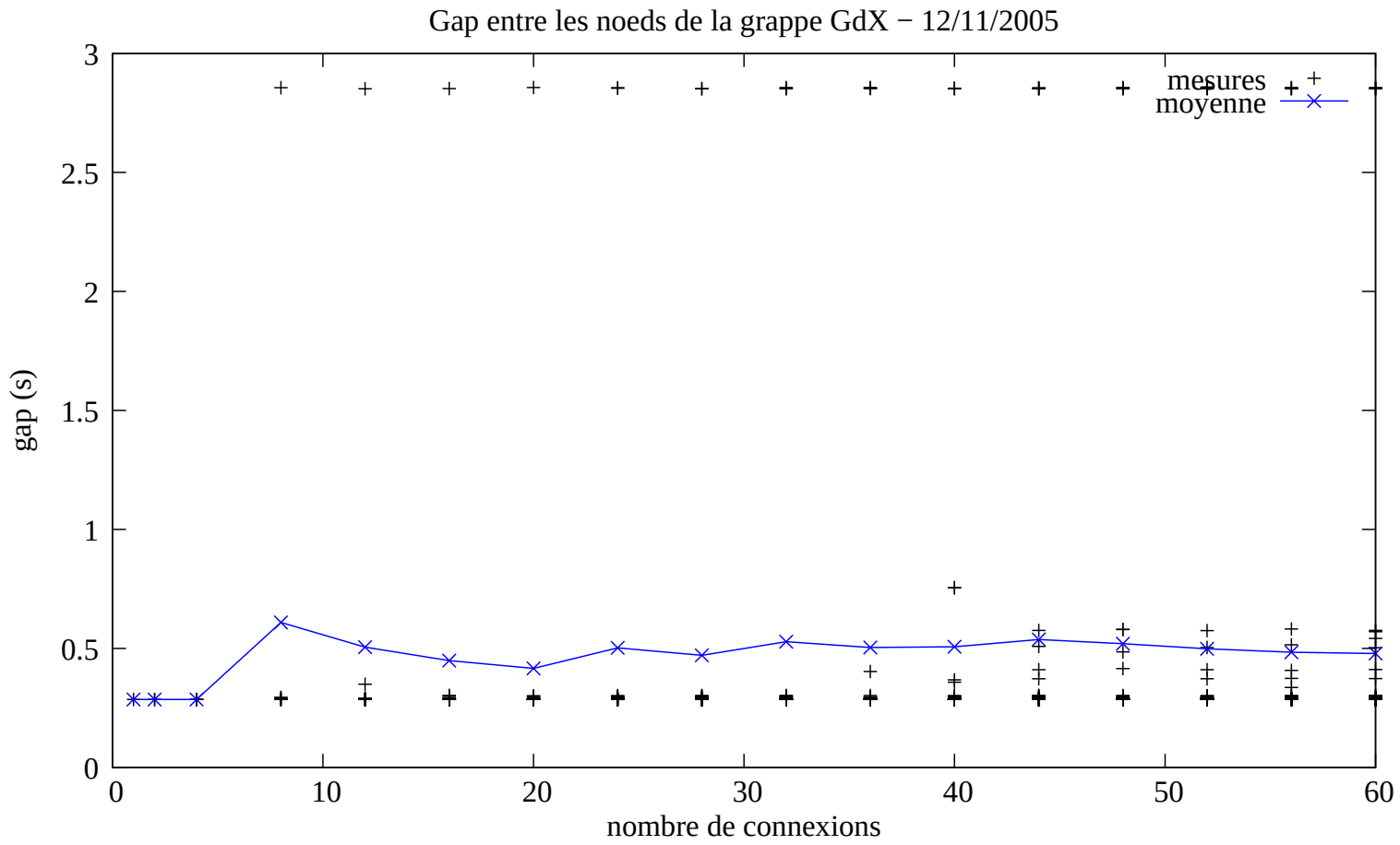
Bandwidth


Debit moyen entre les grappes GdX et Toulouse − 12/11/2005
Debit moyen
debit (Mo/s)
nombre de connexions
0
1
2
3
4
5
6
0
10
20
30
40
50
60

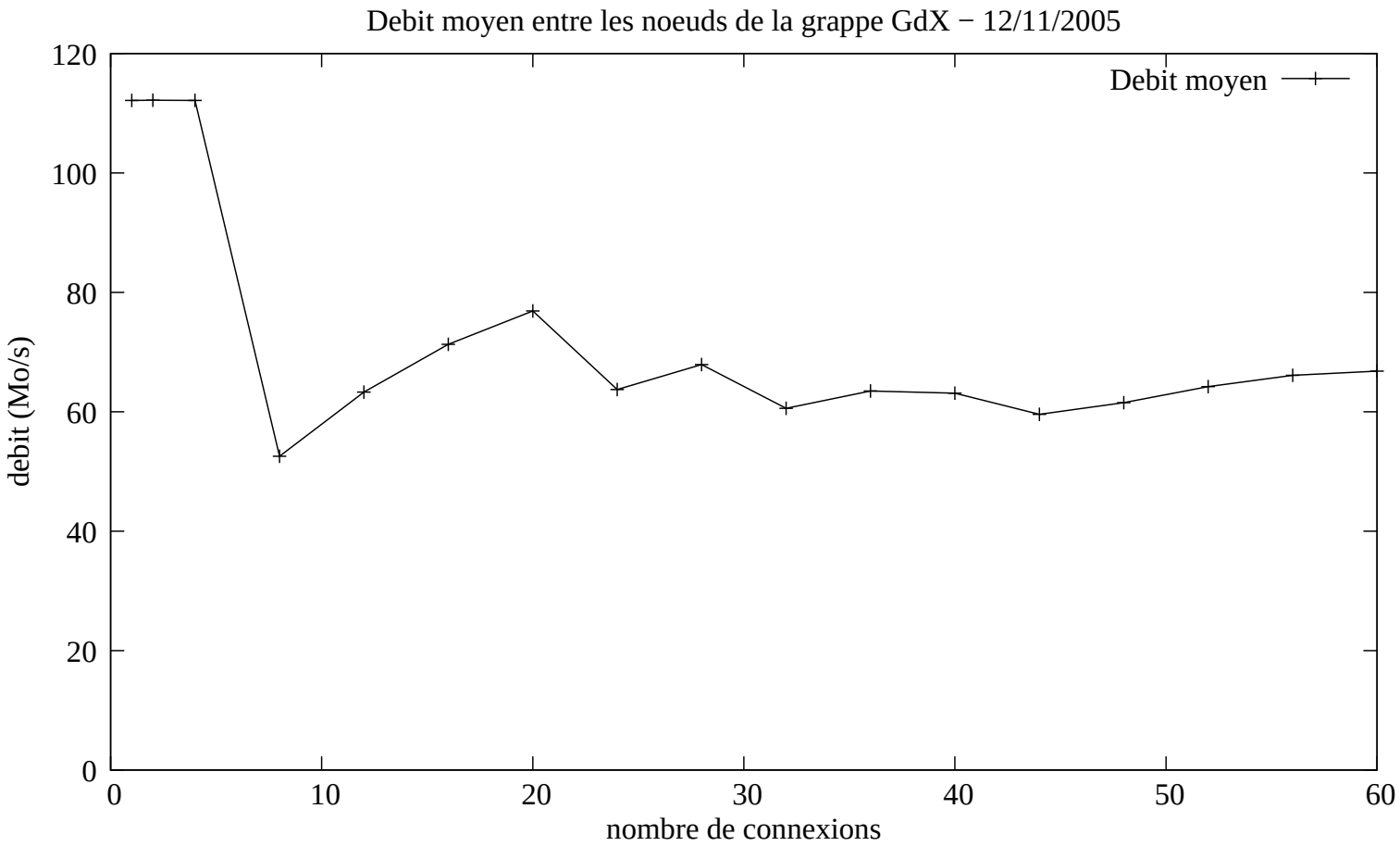
Aggregate Bandwidth



Quantifying contention on LANs



Bandwidth


Debit moyen entre les noeuds de la grappe GdX − 12/11/2005
Debit moyen
debit (Mo/s)
nombre de connexions
0
20
40
60
80
100
120
0
10
20
30
40
50
60