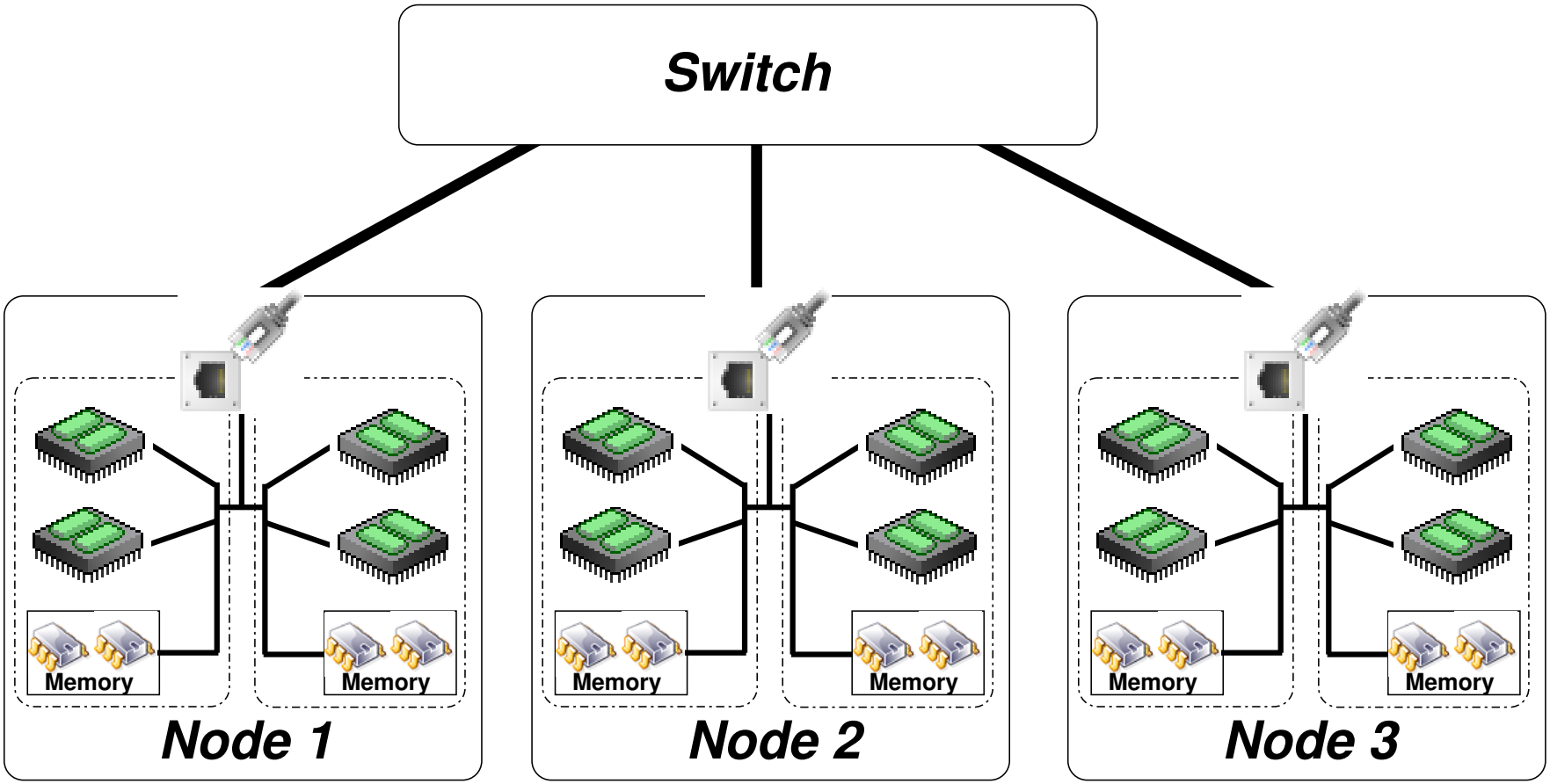
Predictive models for bandwidth sharing in high performance clusters

Jérôme VIENNE, Maxime MARTINASSO,
Jean-Marc VINCENT, Jean-François MEHAUT

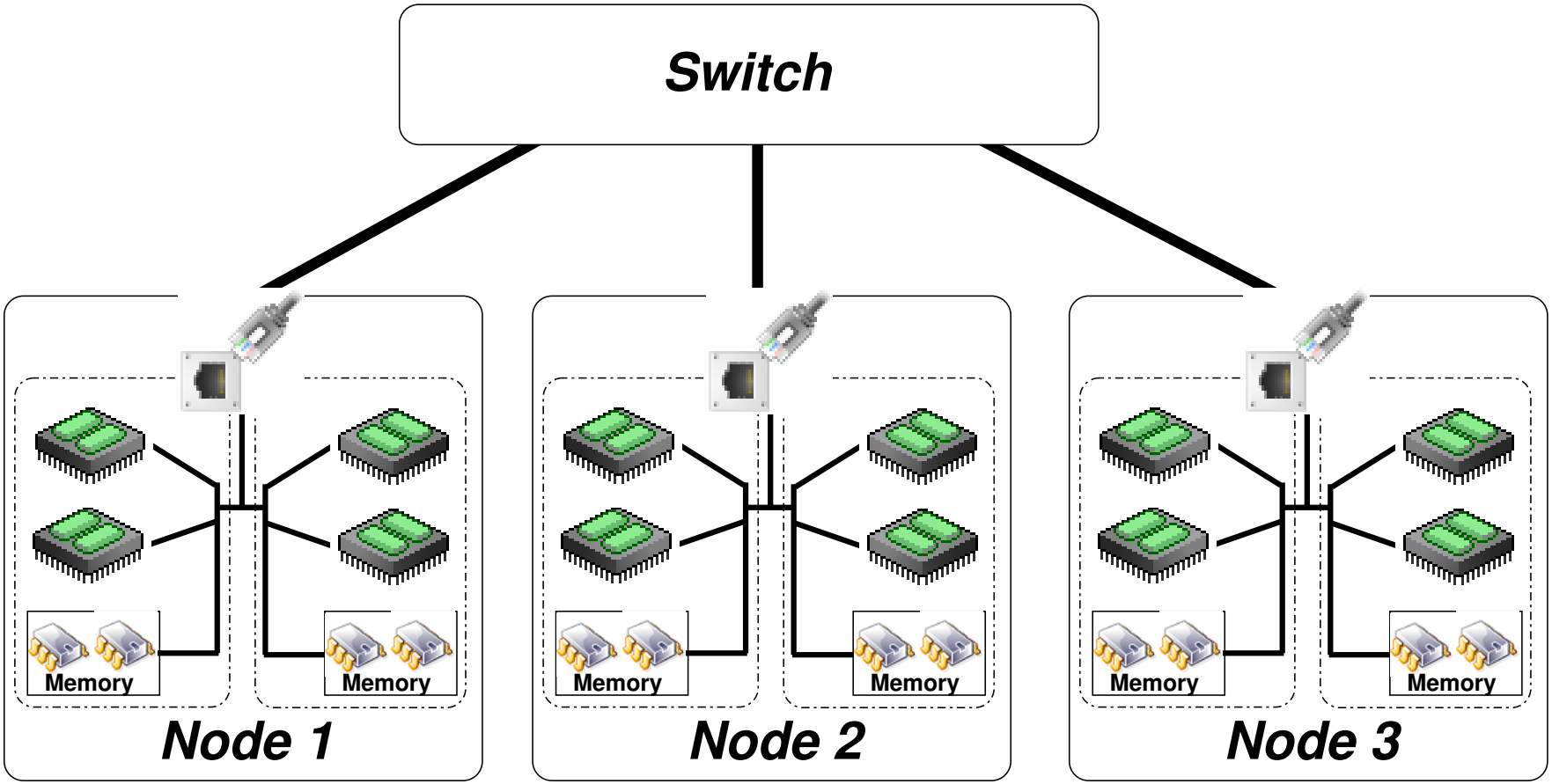
Laboratoire LIG, Mescal Team, Grenoble, France.
BULL - HPC, Benchmark Team, Échirolles, France.

30 September 2008

Composition of Clusters

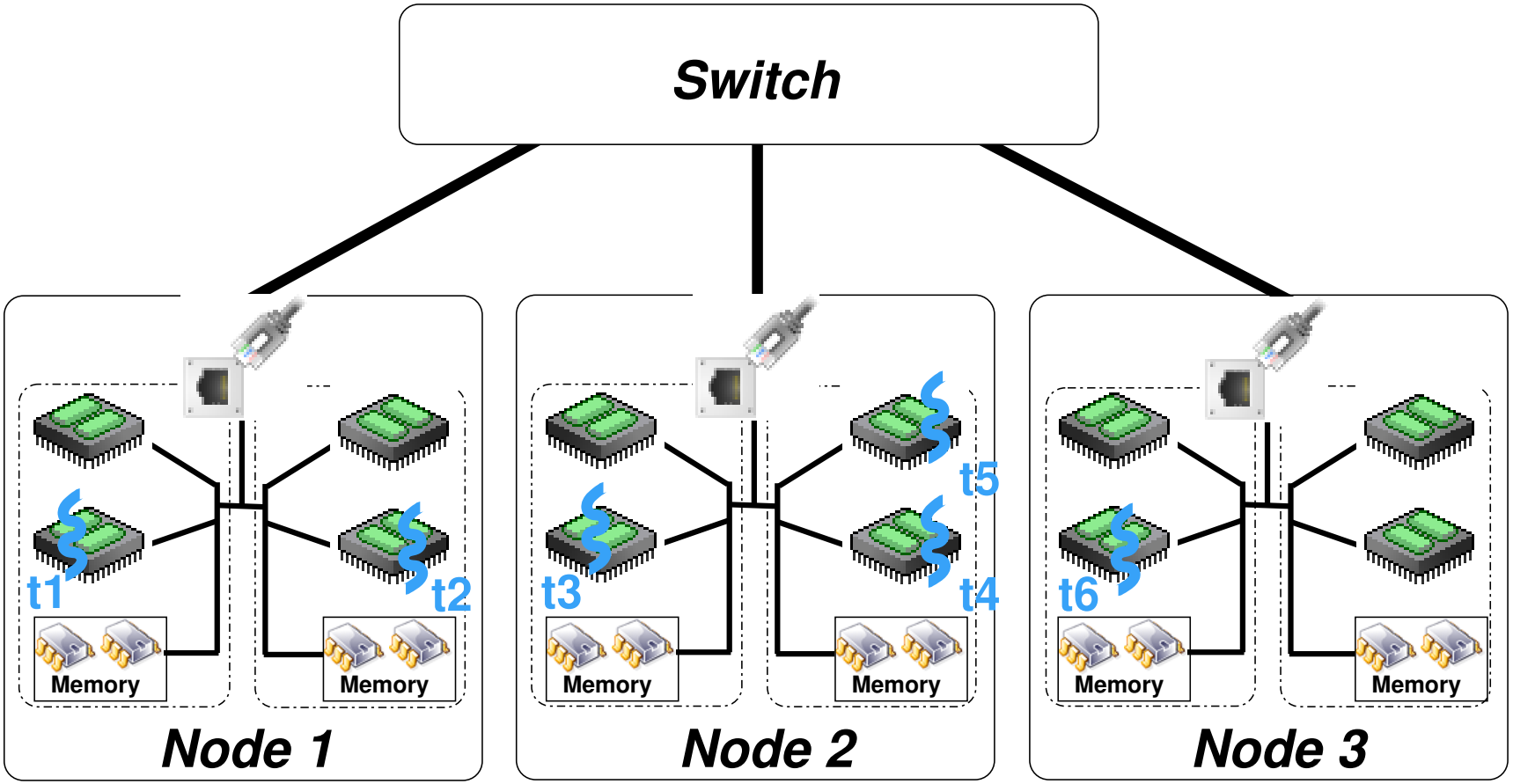


Network concurrency

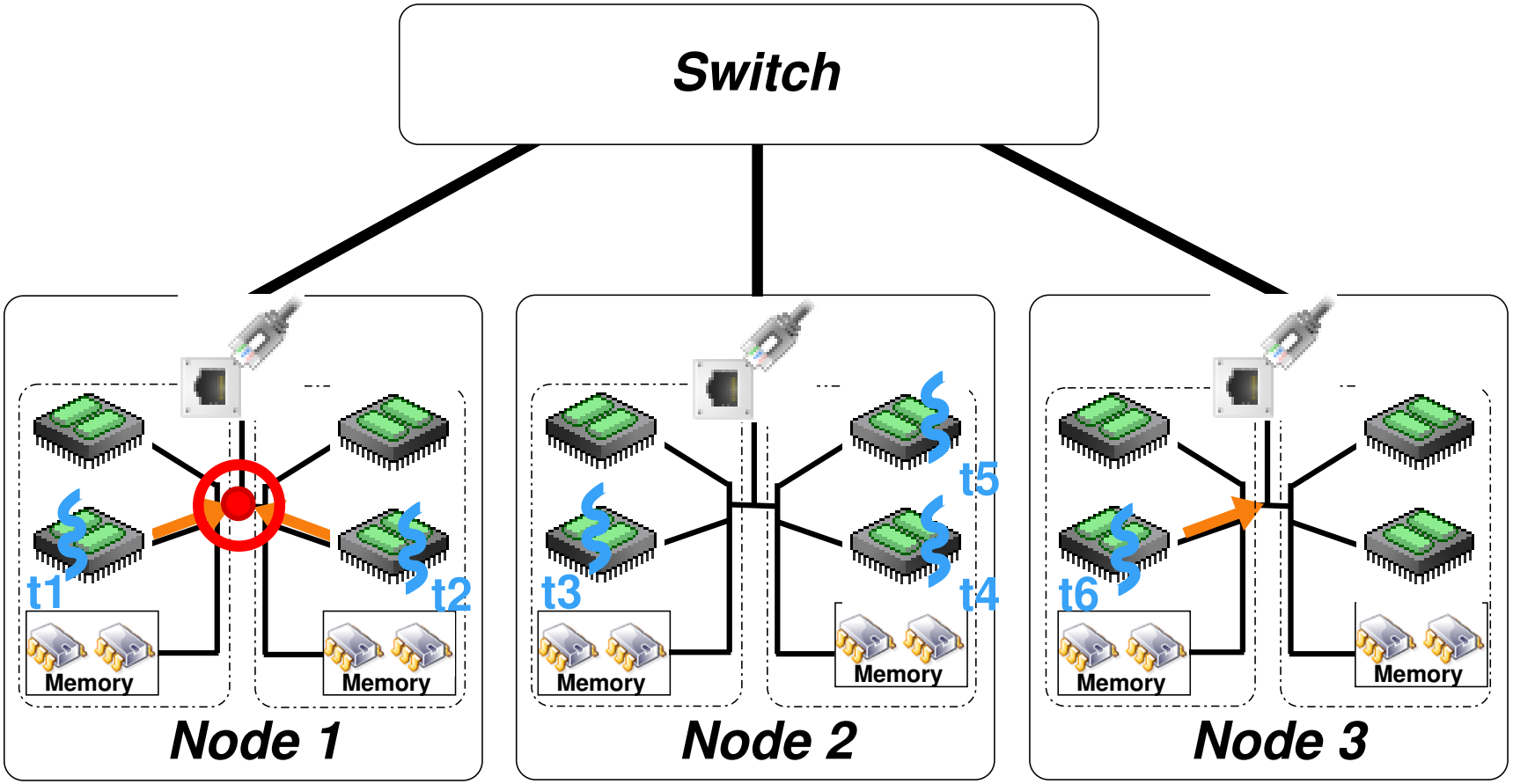


Sharing of the network resource

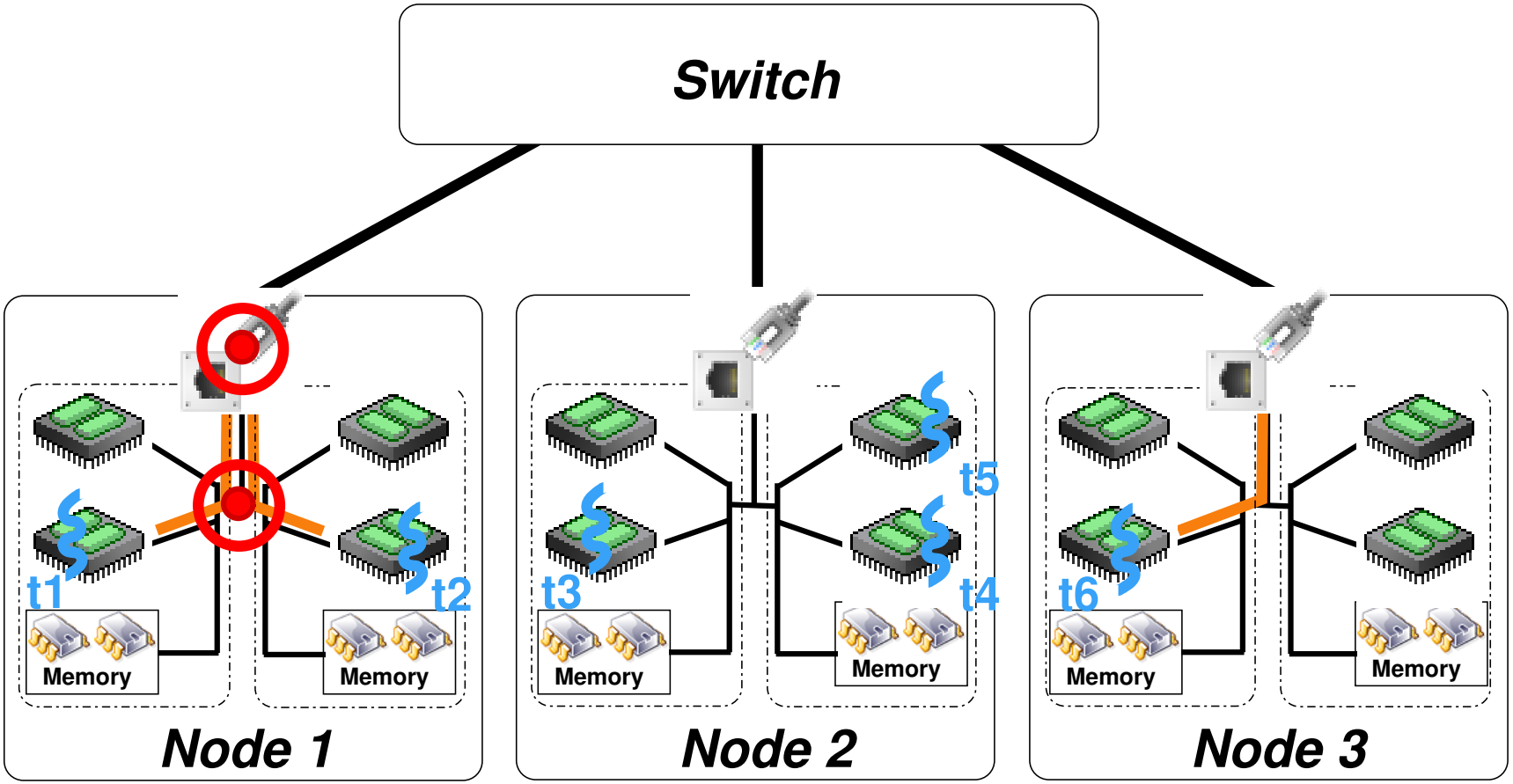
Network concurrency



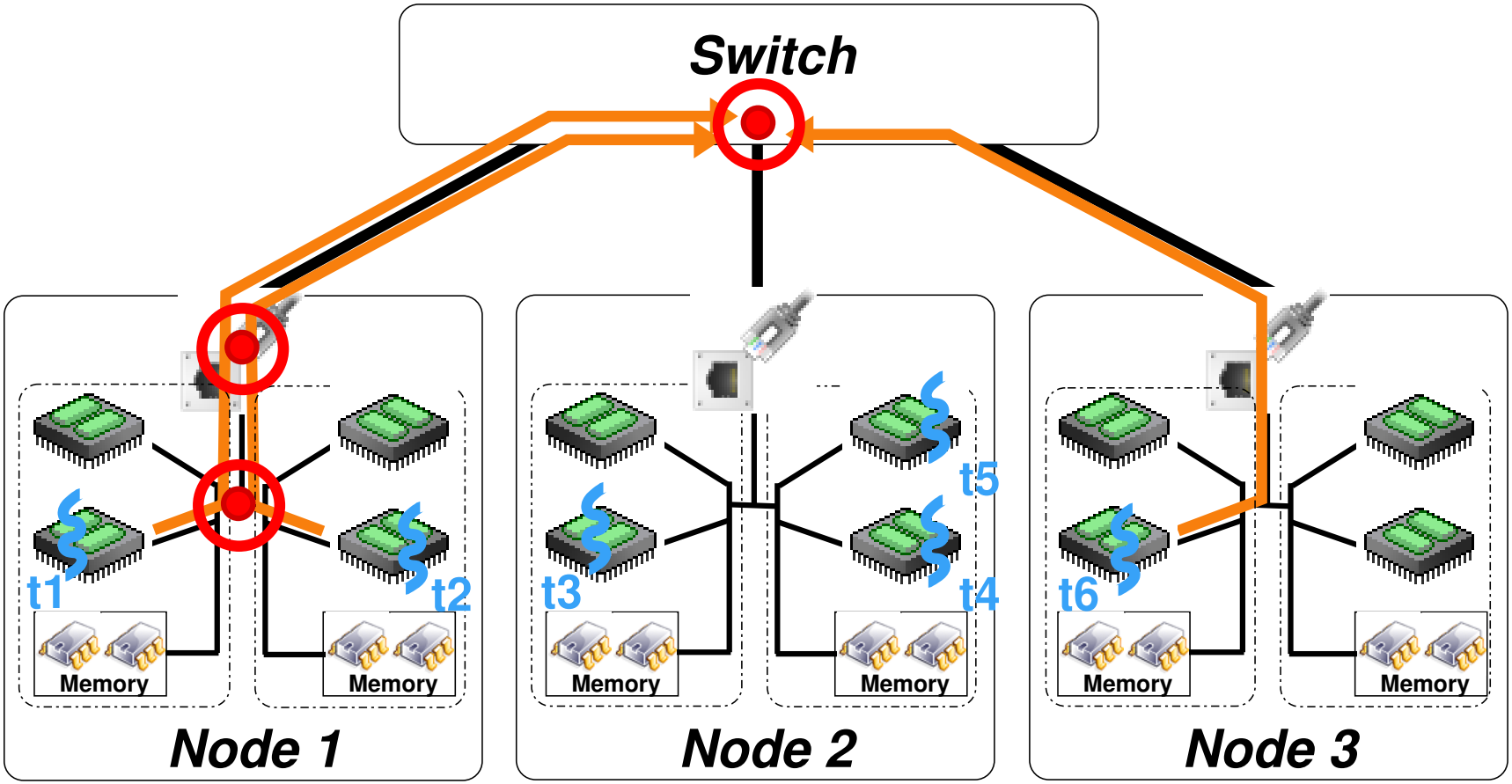
Network concurrency



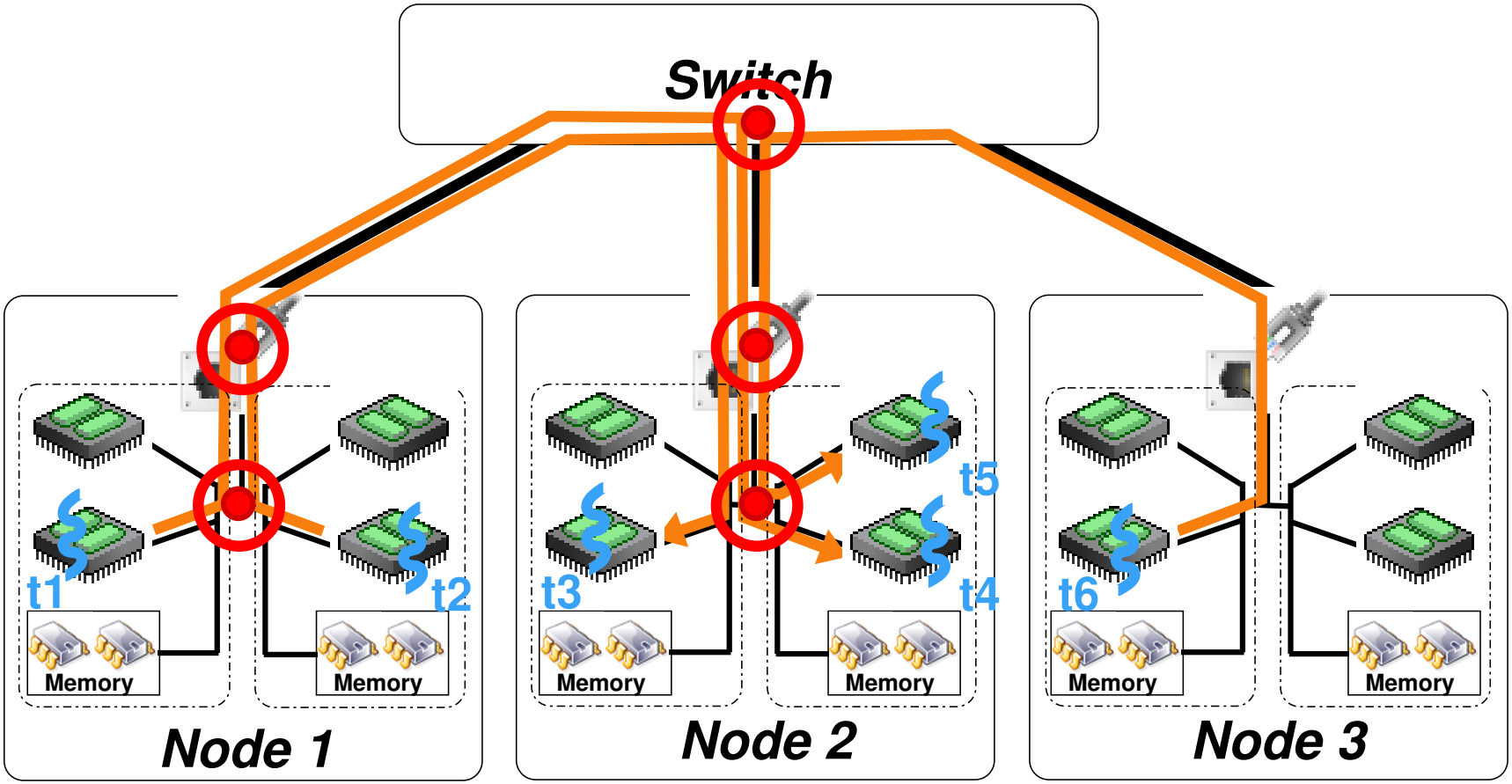
Network concurrency



Network concurrency



Network concurrency



Plan

Plan

Objectives

Study of network concurrency

- High Performance Networks (Myrinet, Gigabit Ethernet)
- Different kind of conflicts
- Analysis of performance

$\Rightarrow$ Observation of communication times

Experimental Method

Software Support

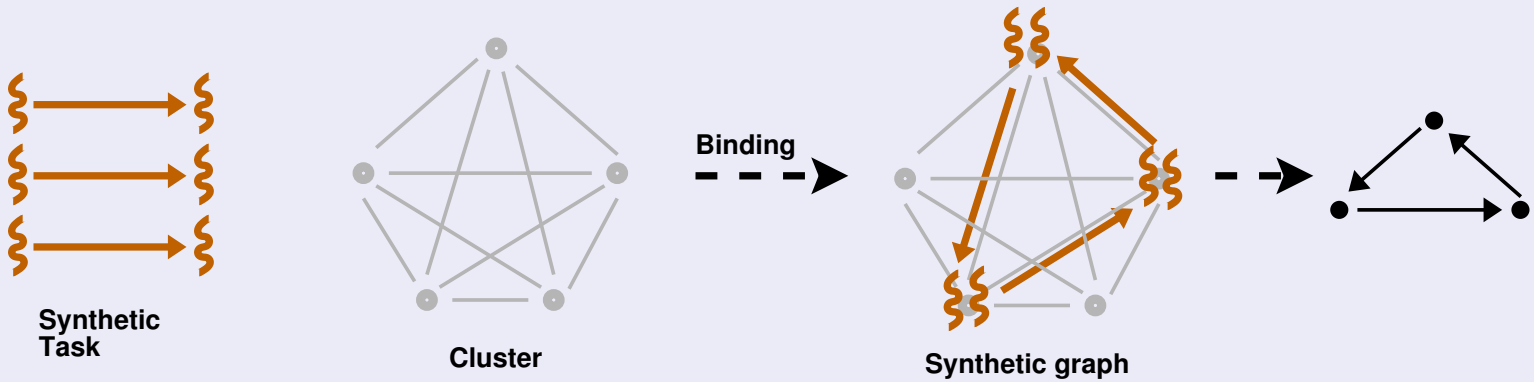
- Communication interface MPI : Mpich
- Synchronous communication : **MPI_Send/MPI_Recv**

Creation of a Benchmark

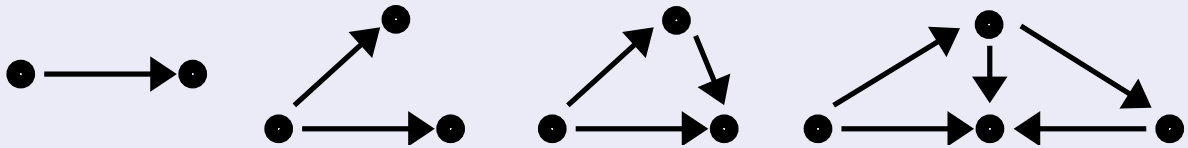
- A set of MPI tasks only sending or receiving.
- Communications : Same start time and size for identical datas
- Measurement : issue rate

Communication patterns

Synthetic graphs



Communication patterns : conflicts



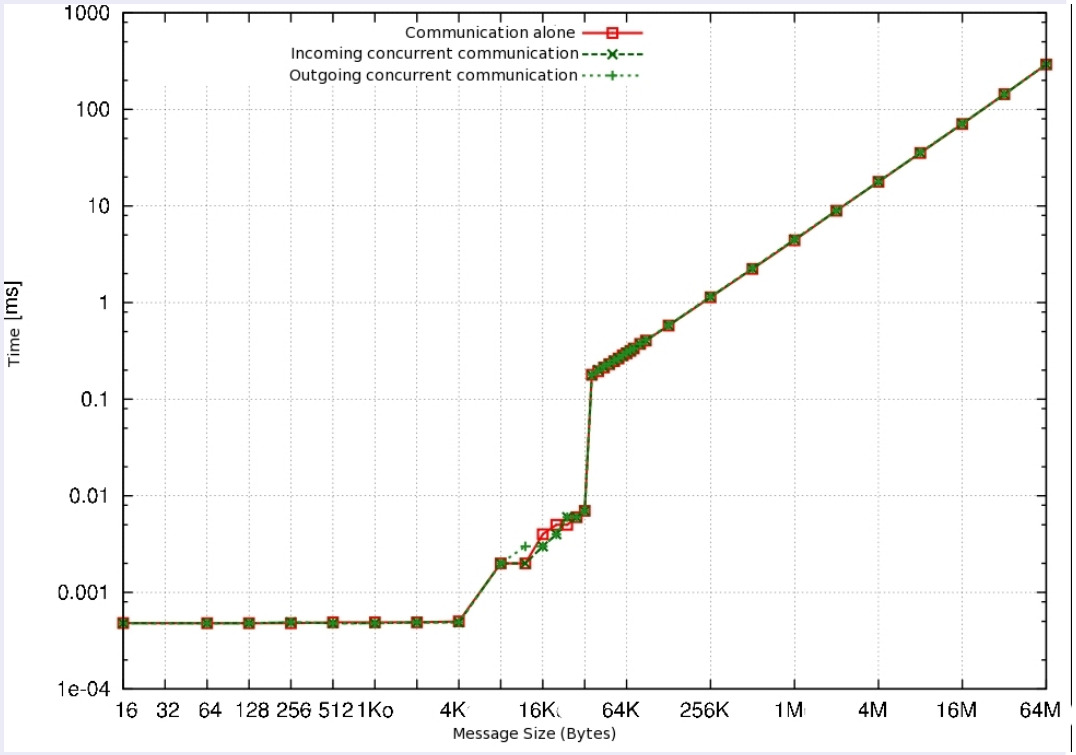
⇒ Comparison of times without conflict and with conflicts

⇒ Progressive augmentation of pattern complexity

Basic experiences

Outgoing/Incoming Conflict

Myrinet (loglog)



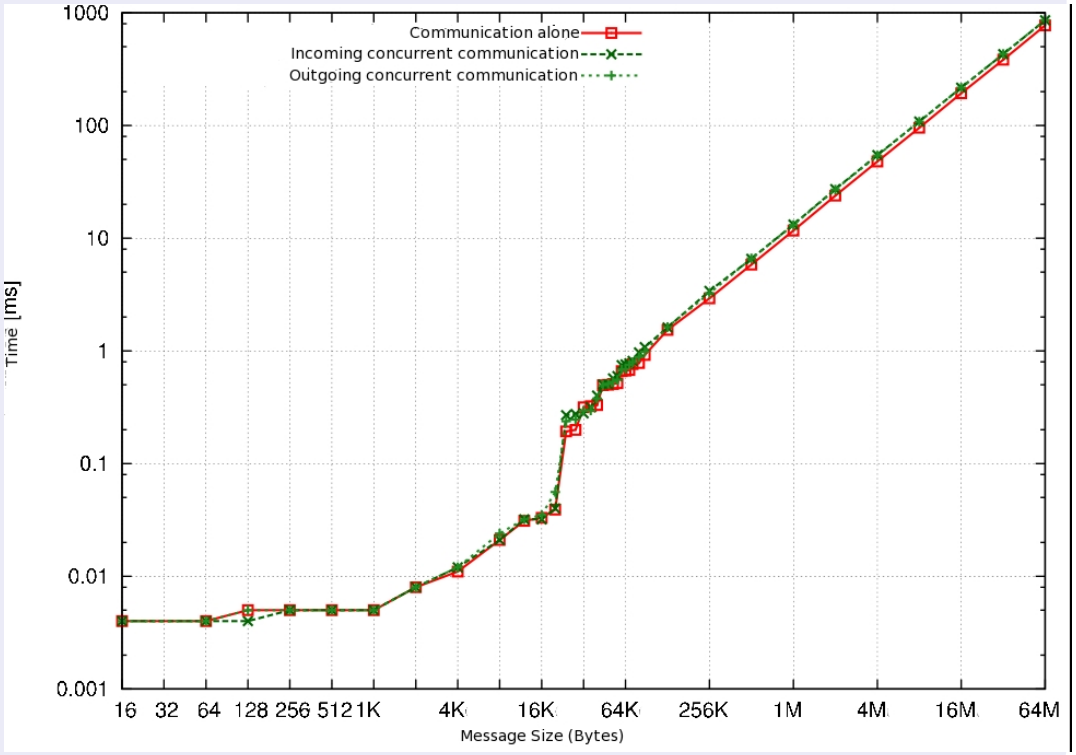
Myrinet

Size	Réf.	O/I conflict	
		Com 1	Com 2
16B [μs]	0.5	0.5	0.5
16KB [ms]	0.003	0.003	0.003
16MB [s]	0.071	0.071	0.071
Ratio	1	1	1

Basic experiences

Outgoing/Incoming Conflict

Gigabit Ethernet (loglog)

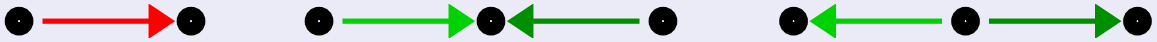


Gigabit Ethernet

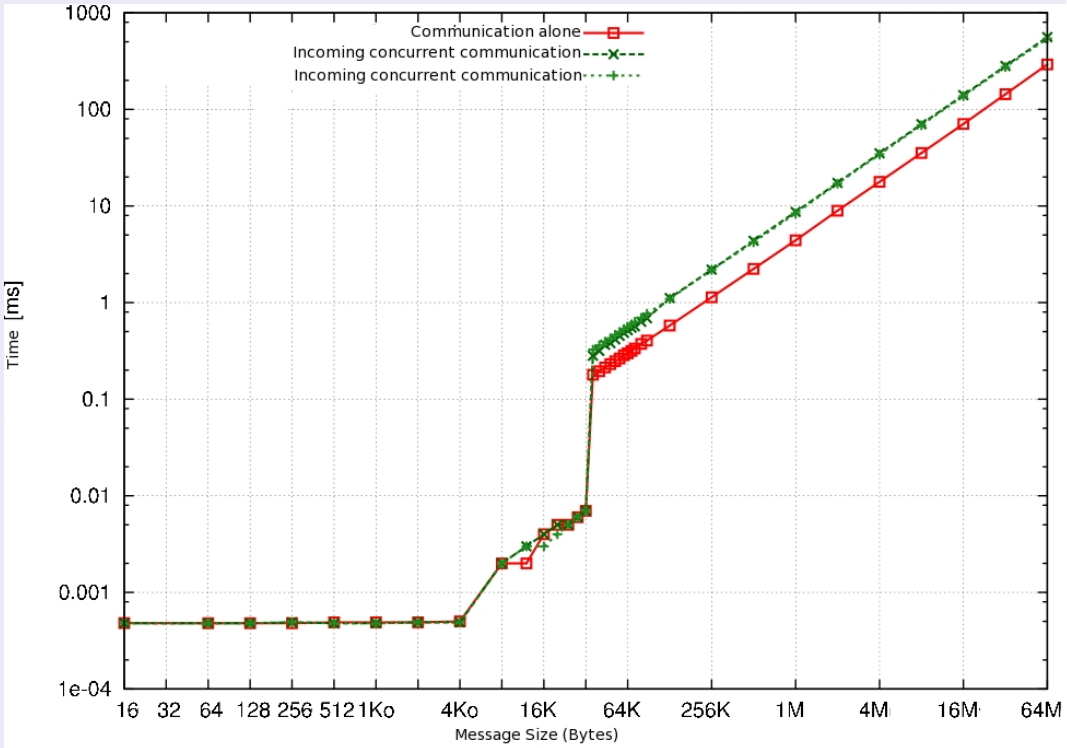
Size	Réf.	O/I conflict	
		Com 1	Com 2
16B [μs]	4.2	4.3	5.0
16KB [ms]	0.033	0.033	0.037
16MB [s]	0.194	0.218	0.207
Ratio	1	1.12	1.06

Basic experiences

Incoming/Incoming or Outgoing/Outgoing Conflicts



Myrinet (loglog)

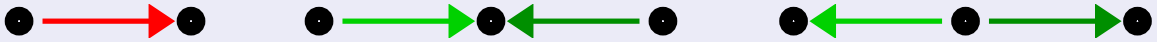


Myrinet

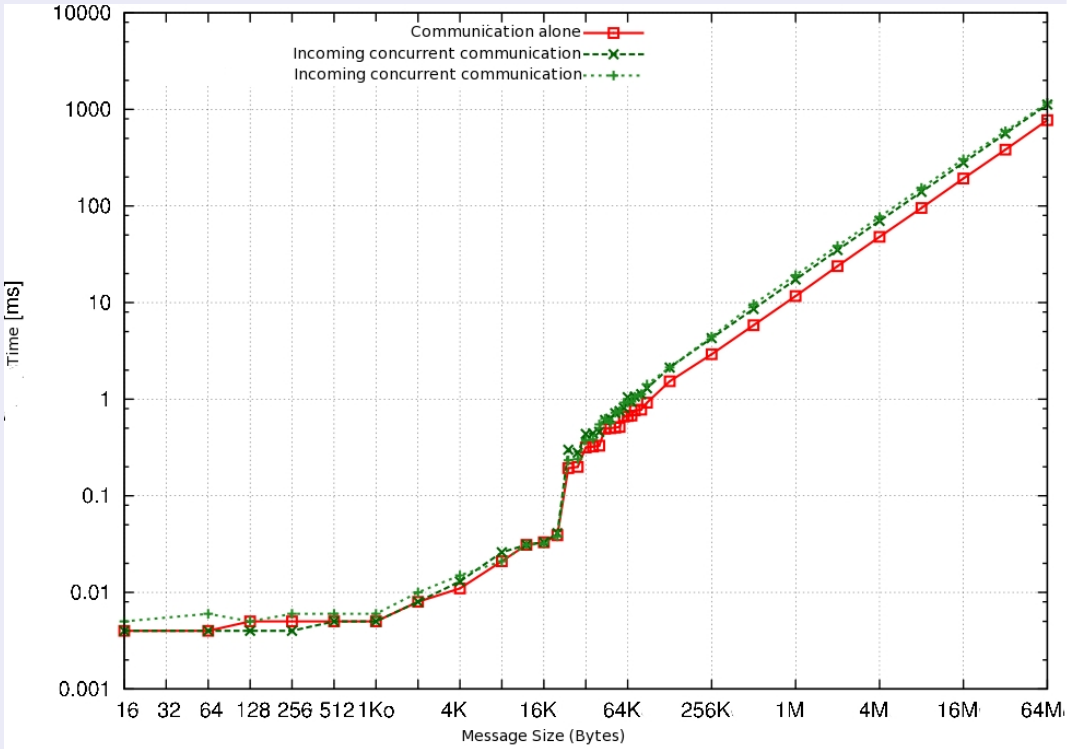
Size	Réf.	I/I conflict	
		Com 1	Com 2
16B [μs]	0.5	0.5	0.6
16KB [ms]	0.003	0.004	0.003
16MB [s]	0.071	0.138	0.138
Ratio	1	1.94	1.94

Basic experiences

Incoming/Incoming or Outgoing/Outgoing Conflicts



Gigabit Ethernet (loglog)

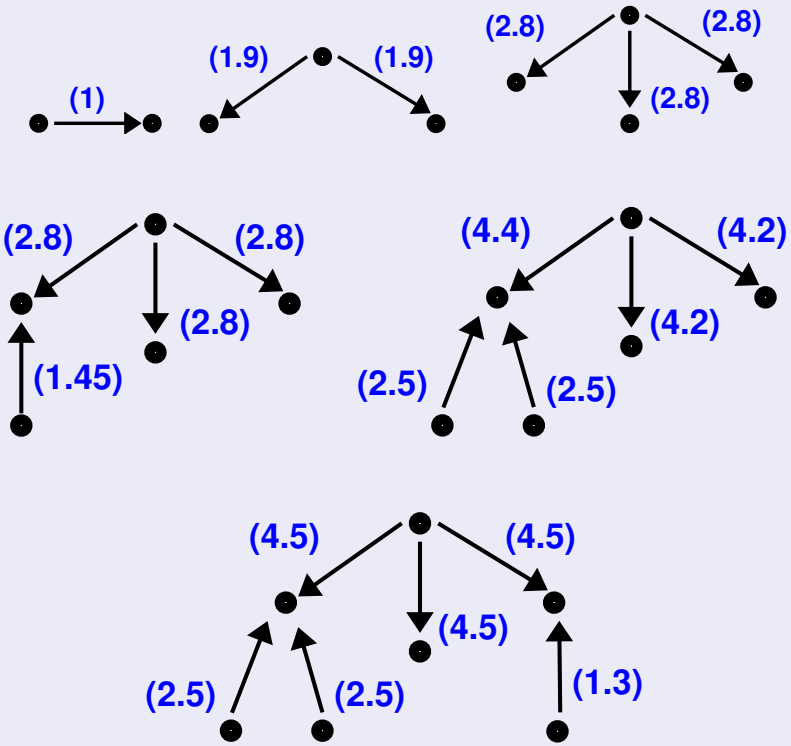


Gigabit Ethernet

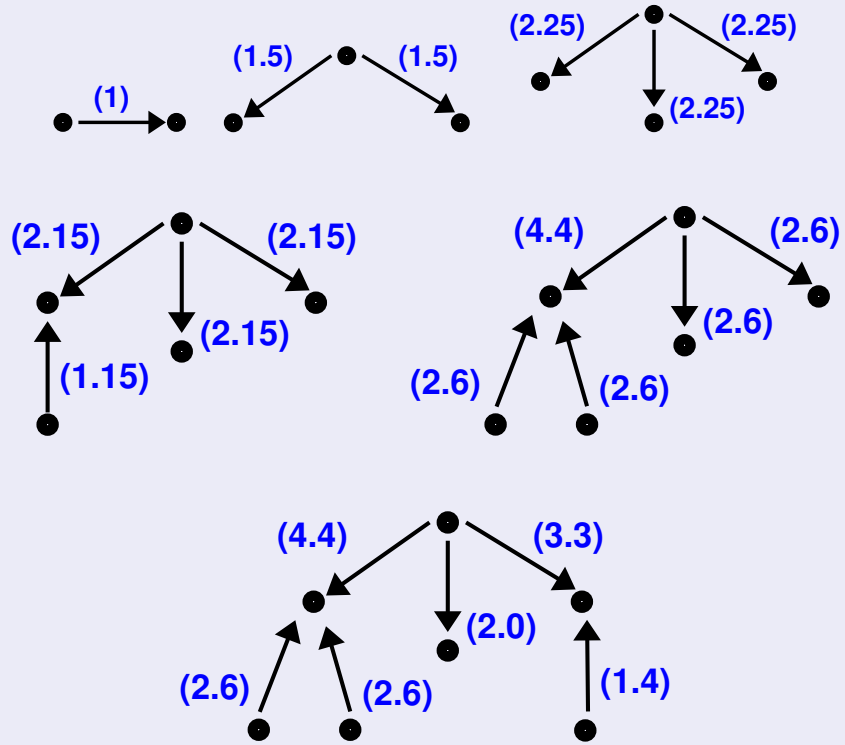
Size	Réf.	I/I conflict	
		Com 1	Com 2
16B [μs]	4.2	4.7	4.2
16KB [ms]	0.033	0.033	0.037
16MB [s]	0.194	0.290	0.291
Ratio	1	1.5	1.5

Elaborated experiences

Myrinet



Gigabit Ethernet



Gigabit Ethernet +TCP : Better concurrency management

Plan

Existing models of communication (without concurrency)

Parameters :

- L : Latency
- β : Bandwidth
- m : Message size

LogP, LogGP, pLogP

- LogP et LogGP :
  - $2o + L + (m-1) * G$
  - o overhead, $G = 1/\beta$
- Generalization : pLogP
  - $o(m)$, $g(m)$

Hockney Model (linear)

$$t(m) = m/\beta + L$$

Modelisation : goals

Complete existing models to modelise the concurrency
$\Rightarrow$ Predict communication times taking into account conflicts

Structure of the model

- Enriched linear model, penality p :

$$t(m) = p * m/\beta + L$$

- Evolution of communication patterns

Determine penalties for each network

Gigabit Ethernet

- Complex flow control : TCP
- Many constructors : heterogeneous components

$\Rightarrow$ Model : Quantitative approach (parameters + measures)

Myrinet

- Simple flow control
- Only one constructor : homogeneous components

$\Rightarrow$ Model : Descriptive approach

Plan

Evaluation method

Comparisons of predicted and measured values

- Synthetic Graphs : trees et complete graphs
- Application graph : benchmark Linpack (*HPL*)

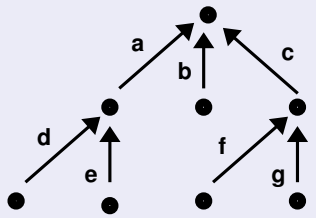
Formulas

$$E_{rel}(c_k) = \frac{T_p - T_m}{T_m} \times 100$$

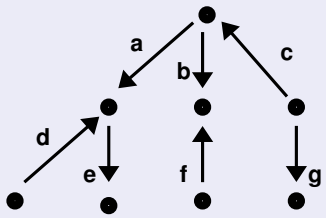
$$E_{abs}(G) = \frac{1}{N} \sum_{k=1}^{N} |E_{rel}(c_k)|$$

Synthetic graphs : trees

Trees



Ar1



Ar2

Myrinet

Ar1			
com.	T_m	T_p	E_{rel}
a	0.097	0.107	10.3
b	0.103	0.107	3.9
c	0.092	0.107	16.3
d	0.067	0.071	6.0
e	0.070	0.071	1.4
f	0.065	0.071	9.2
g	0.070	0.071	1.4
$E_{abs} = 6.9$			

Ar2			
com.	T_m	T_p	E_{rel}
a	0.087	0.089	2.3
b	0.087	0.089	2.3
c	0.070	0.071	1.4
d	0.052	0.053	1.9
e	0.037	0.035	-5.4
f	0.051	0.053	3.9
g	0.070	0.071	1.4
$E_{abs} = 2.6$			

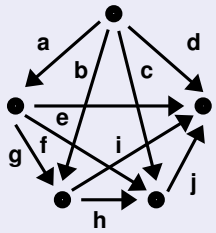
Gigabit Ethernet

Ar1			
com.	T_m	T_p	E_{rel}
a	0.223	0.214	-4.0
b	0.170	0.214	25.9
c	0.226	0.214	-4.0
d	0.147	0.143	-2.7
e	0.146	0.143	-2.0
f	0.147	0.143	-2.7
g	0.146	0.143	-2.0
$E_{abs} = 6.0$			

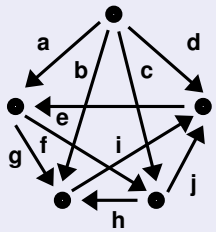
Ar2			
com.	T_m	T_p	E_{rel}
a	0.166	0.147	-11.4
b	0.164	0.147	-10.4
c	0.147	0.143	-2.7
d	0.140	0.137	-2.1
e	0.131	0.100	-23.7
f	0.131	0.137	4.6
g	0.145	0.143	-1.4
$E_{abs} = 8.0$			

Synthetic graphs : complete graphs

Graphs



Kn1



Kn2

Myrinet

Kn1			
com.	T_m	T_p	E_{rel}
a	0.171	0.191	11.7
b	0.171	0.191	11.7
c	0.171	0.191	11.7
d	0.171	0.191	11.7
e	0.149	0.144	-3.3
f	0.149	0.144	-3.3
g	0.149	0.144	-3.3
h	0.123	0.117	-4.9
i	0.123	0.117	-4.9
j	0.083	0.099	19.3
$E_{abs} = 8.6$			

Kn2			
com.	T_m	T_p	E_{rel}
a	0.164	0.177	7.9
b	0.164	0.177	7.9
c	0.164	0.177	7.9
d	0.164	0.177	7.9
e	0.043	0.053	23.2
f	0.086	0.085	-1.2
g	0.087	0.085	-2.3
h	0.108	0.101	-6.5
i	0.108	0.101	-6.5
j	0.059	0.073	23.7
$E_{abs} = 9.5$			

Gigabit Ethernet

Kn1			
com.	T_m	T_p	E_{rel}
a	0.257	0.252	-1.9
b	0.256	0.252	-1.6
c	0.270	0.252	-6.7
d	0.311	0.302	-2.9
e	0.181	0.189	4.4
f	0.207	0.206	-0.5
g	0.274	0.276	0.7
h	0.182	0.206	13.2
i	0.258	0.276	7.0
j	0.287	0.276	-3.8
$E_{abs} = 4.3$			

Kn2			
com.	T_m	T_p	E_{rel}
a	0.270	0.285	5.5
b	0.284	0.304	7.0
c	0.284	0.304	7.0
d	0.270	0.285	5.5
e	0.184	0.137	-25.5
f	0.167	0.137	-17.9
g	0.194	0.206	6.2
h	0.198	0.206	4.4
i	0.215	0.206	-4.2
j	0.188	0.206	9.6
$E_{abs} = 9.3$			

Comparison with simple models

Simple models

- $p = 1$, Hockney model without concurrency
- $p = \Delta_s$, proportionnal sharing between communications with the same origin
- $p = \Delta_e$, proportionnal sharing between communications with the same destination

Comparison

	E_{abs} on *Myrinet* [%]				E_{abs} on *Gigabit Ethernet* [%]			
Models	Ar1	Ar2	Kn1	Kn2	Ar1	Ar2	Kn1	Kn2
$p = 1$	54.2	40.7	74.8	62.8	42.3	33.7	60	55.7
$p = \Delta_s$	54.2	15.7	30.8	21.8	42.3	26.2	37.9	28.7
$p = \Delta_e$	6.9	30.7	33.4	31.3	37.3	31	36.4	27.4
Our models	6.9	2.6	8.6	9.5	6	8	4.3	9.3

Conclusion

Problem

- Bandwith sharing between concurrent communications
- High Performance Network : *Myrinet* et *Gigabit Ethernet*

Experimentations

- Study of communication patterns
- Concept of conflict and penalty

Conclusion

Modelization

- Model to get penalties :
  - Descriptive approach : *Myrinet*
  - Quantitative approach : *Gigabit Ethernet*
- Estimation of communication times
  - Dynamic evolution of patterns
  - Dicrete event simulation

Futur Work

- Study of contention on Infiniband network
- Estimation of computation time