Performance analysis for clusters of symmetric multiprocesors

F. Almeida[1] J.A. Gómez[1] J.M. Badía[2]

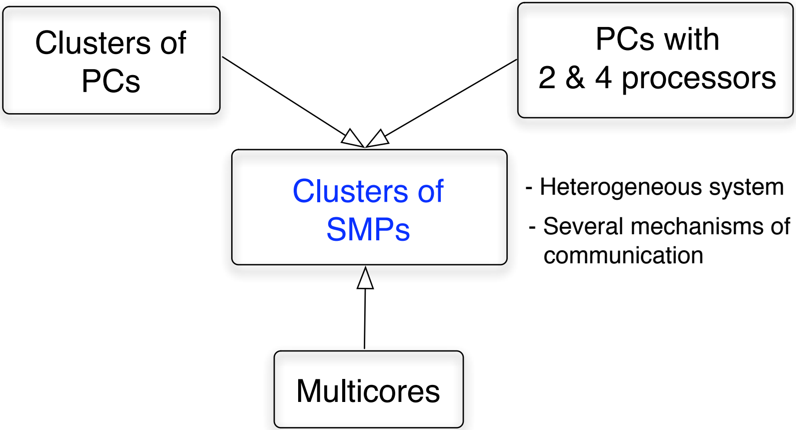
[1]Depto. Estadística, Investigación Operativa y Computación
Univ. La Laguna, España.

[2]Depto. Ingeniería y Ciencia de Computadores
Univ. Jaume I, España.

Parallel, Distributed and Network-based Processing, 2007

Outline

Widespread of parallel computers



- Heterogeneous system
- Several mechanisms of communication

Taking profit of the new architectures

- New programming models: Hybrid models (MPI + OpenMP).
- New debugging and profiling tools.
- New models of performance analysis.
- . . .

What have we done?

Analysis and modeling of the performance of clusters of SMPs

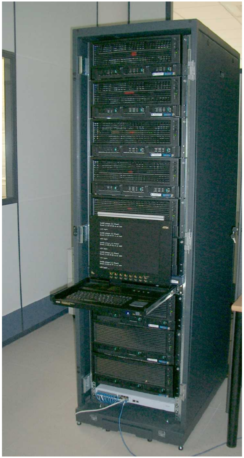
- Analysis of the point-to-point communications.
- Development of a communication model adapted to the architecture.
- Test of the model with three different applications.

Outline

Outline

Hardware environment: Cluster cat

- 9 nodes
- Each node:
  - 4 1.5GHz. Itanium-2 processors
  - 4 GBytes of RAM
  - 4 MBytes of L3 cache
- 2 networks
  - FastEthernet (control)
  - Infiniband (applications)



Software environment

- Linux Redhat 3, update 6. Kernel: 2.4.21-37 SMP.
- Intel `icc` compiler, version 9.
- MPICH v. 1.2.5
  - `VAPI` device.
  - Exploits the shared memory and the Infiniband network.

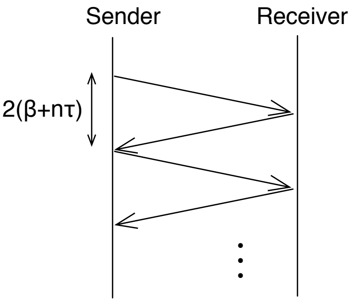
Outline

Basic communication model

$$\beta + n\tau$$

- β: Latency
- τ: time to send one byte
- n: size of the message

ping-pong method

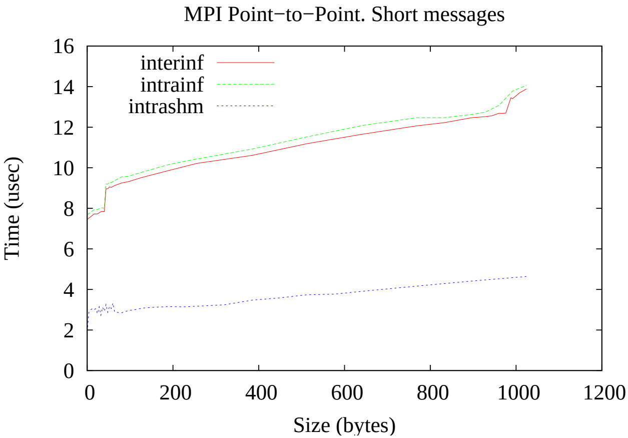


perftest [Gropp & Lusk,99]

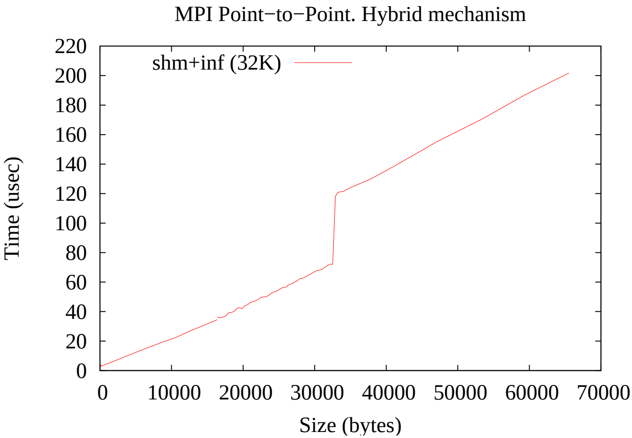
Parameters of the analysis

- Three mechanisms of communication:
  - `interinf`: Infiniband between different nodes.
  - `intrainf`: Infiniband into the nodes.
  - `intrashm`: Shared memory into the nodes.
- Two size ranges:
  - Short messages: $<$ 1Kbyte.
  - Long messages: 1 Kbyte - 1 Mbyte.

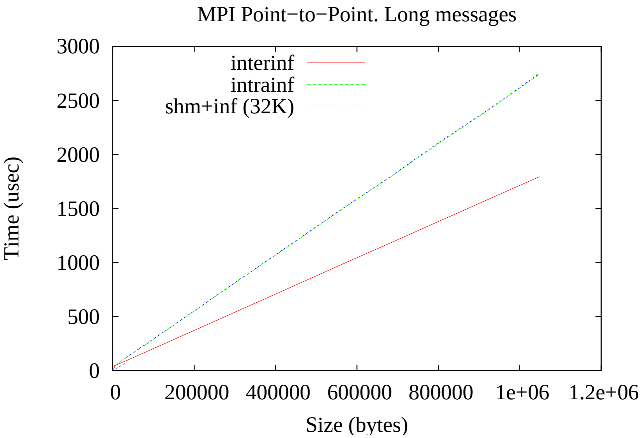
Short messages



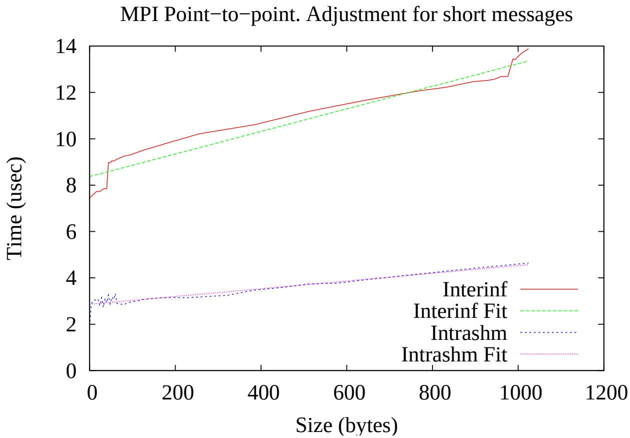
Switching of mechanism of communication



Long messages



Adjustment to the model



Values of the parameters of the model

Test	**Size**	**Adjustment β (τ)**	**Error**
Interinf	Short	8.3714 (0.0049)	4.49
	Long	35.363 (0.0017)	7.71
Intrainf	Short	8.5970 (0.0050)	5.08
	Long	26.2649 (0.0026)	7.15
Hybrid	Short	2.8622 (0.0017)	4.26
	Long (<32K)	0.9252 (0.0022)	5.90
	Long (>32K)	33.1205 (0.0026)	4.90

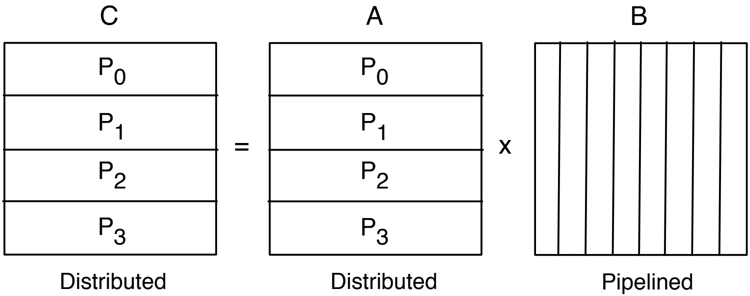
Table: Adjustments (μsec) and Errors (%).

Outline

Testing the model

- Two models:
  - `Homogeneous`: Cluster of uniprocessors connected with Infiniband.
  - `Heterogeneous`: τ and β depend on:
    - The location of the sender and receiver: mechanism of communication.
    - The size of the message.
- Three applications. Three communication schemes.
- Using only point-to-point communications.
- Applications:
  - Matrix product. Pipeline.
  - Matrix-vector product. Master-slave.
  - Heat diffusion. Synchronous iteration.

Pipeline matrix product



- Parallel scheme. Pipeline sequence of matrix-vector products:

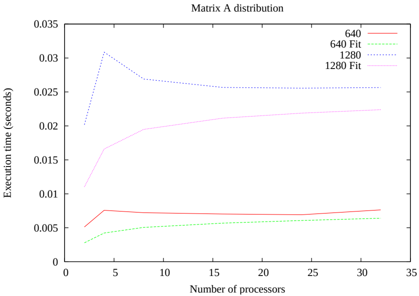
```
C(:,k) = A * B(:,k)
```

Parallel matrix product algorithm

```
if P_0
  for (pr=1; pr<numprocs; pr++)
    send block of A to P_pr
else
  Receive block of A to P_0
for (k=0; k<n; k++) {
  if P_0
    Send column k of B to P_1
  else if pr < numprocs - 1
    Receive column k from P_ant
    Send column k to P_sig
  else
    Receive column k from P_ant
  Calculate local block C(:,k)=A*B(:,k)
}
```

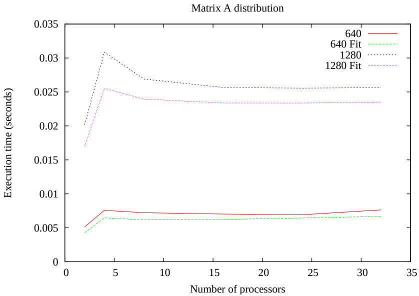
Matrix A distribution adjustment

Homogeneous model



$$T_A = P * T(n)$$

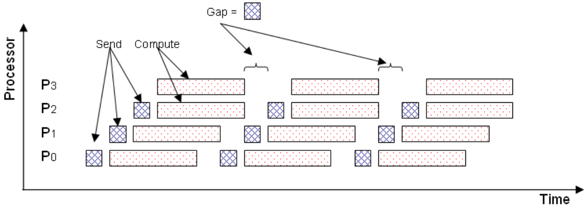
Heterogeneous model



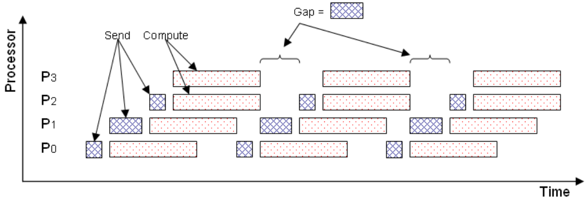
$$T_A = \sum_{i=1}^{P-1} T_{0,i}(n)$$

Pipeline computation modeling

Homogeneous model



Heterogeneous model



$$T_{startup} + N * T_{comput} + (N-1) * (gap + T_B)$$

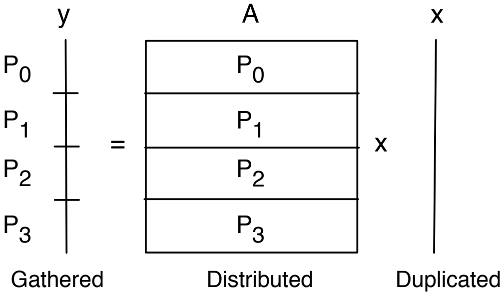
Global adjustment of the matrix product (n = 1280)

$$T = T_A + T_{startup} + N * T_{comput} + (N - 1) * (gap + T_B)$$

		Homogeneous		Heterogeneous	
Proc.	**Real time**	**Model time**	**Error**	**Model time**	**Error**
2	43.394	42.241	2.66	41.303	5.06
4	19.174	20.784	8.39	20.416	6.01
8	18.419	10.047	19.35	8.973	6.17
16	4.170	4.680	12.24	4.451	6.33
24	2.753	2.891	5.00	2.851	3.45
32	2.162	1.996	7.67	2.142	0.95

Table: Time (*sec*) and Errors (%).

Matrix-vector product

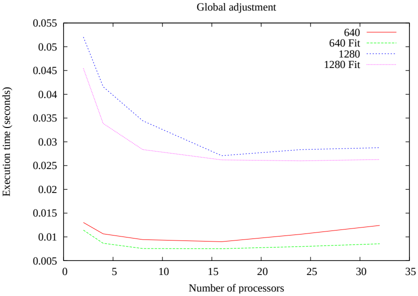


- Master-slave scheme by blocks of rows

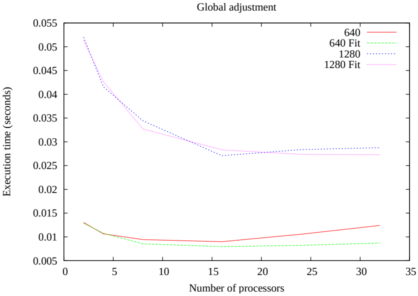
```
Process 0 broadcasts x and scatters blocks of A
Parallel computation of blocks of y
Process 0 gathers blocks of y
```

Global matrix-vector adjustment

Homogeneous model

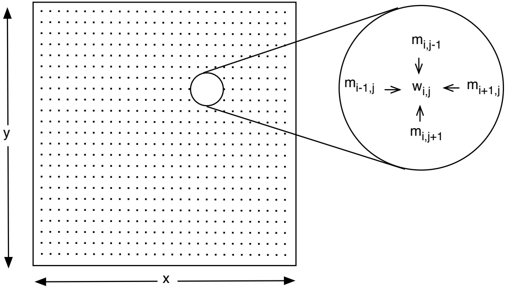


Heterogeneous model

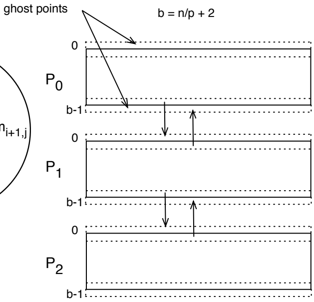


Heat diffusion

Problem



Local communications



Synchronous iteration

```
for (it=0; it<numiter; it++)
  Send local upper edge to P_ant
  Receive upper edge from P_sig
  Send local lower edge to P_sig
  Receive lower edge from P_ant
  Update the values of the local block
```

Heat diffusion global adjustment ($n = 480$, $it = 100$)

		Homogeneous		Heterogeneous	
Proc.	Real time	Model time	Error	Model time	Error
2	0.949	0.948	0.13	0.952	0.27
4	0.480	0.474	1.15	0.482	0.45
8	0.246	0.237	3.67	0.245	0.43
16	0.128	0.119	7.77	0.127	1.24
24	0.091	0.080	13.93	0.087	3.91
32	0.070	0.060	16.35	0.068	3.12

Table: Time (*sec*) and Errors (%).

Outline

Conclusions

- New performance models should have into account the heterogeneity of the architectures.
- Improved fitting of the theoretical communication model and the experimental results if we have into account:
  - The location of the sender and receiver, and then
  - the communication mechanism: shared memory or interconnection network.
- The heterogeneous model improves the adjustment of the theoretical and experimental results on different parallel applications.