Etudes de performance sur une machine parallèle de type « Grappe de PCs »

Encadrants :

Pr. Alain Greiner (LIP6 - ASIM)

Daniel Millot, Philippe Lalevee (INT)

Introduction / Sujet

- **Lieu** : LIP6
- **Cadre** : Projet Multi-PC (MPC - 1995)
- **Machine parallèle à faible coût**
  - Grappe de PCs reliés par un réseau haut-débit
  - Technologie HSL développée au laboratoire
- **Objet** : études de performance à différents niveaux des couches logicielles
  - couches basses MPC, couches hautes PVM
  - parallélisation de la résolution de l'équation de Laplace sur un domaine 2D
  - comparaisons avec ETHERNET 100

Plan de la Présentation

- Introduction / Sujet
- **La machine MPC**
- L’environnement PVM-MPC
- L’équation de Laplace
- Fast-PVM
- Conclusions

La machine MPC

http://mpc.lip6.fr/

- Architecture matérielle
- Architecture logicielle
  - L'écriture distante
  - Les couches basses MPC

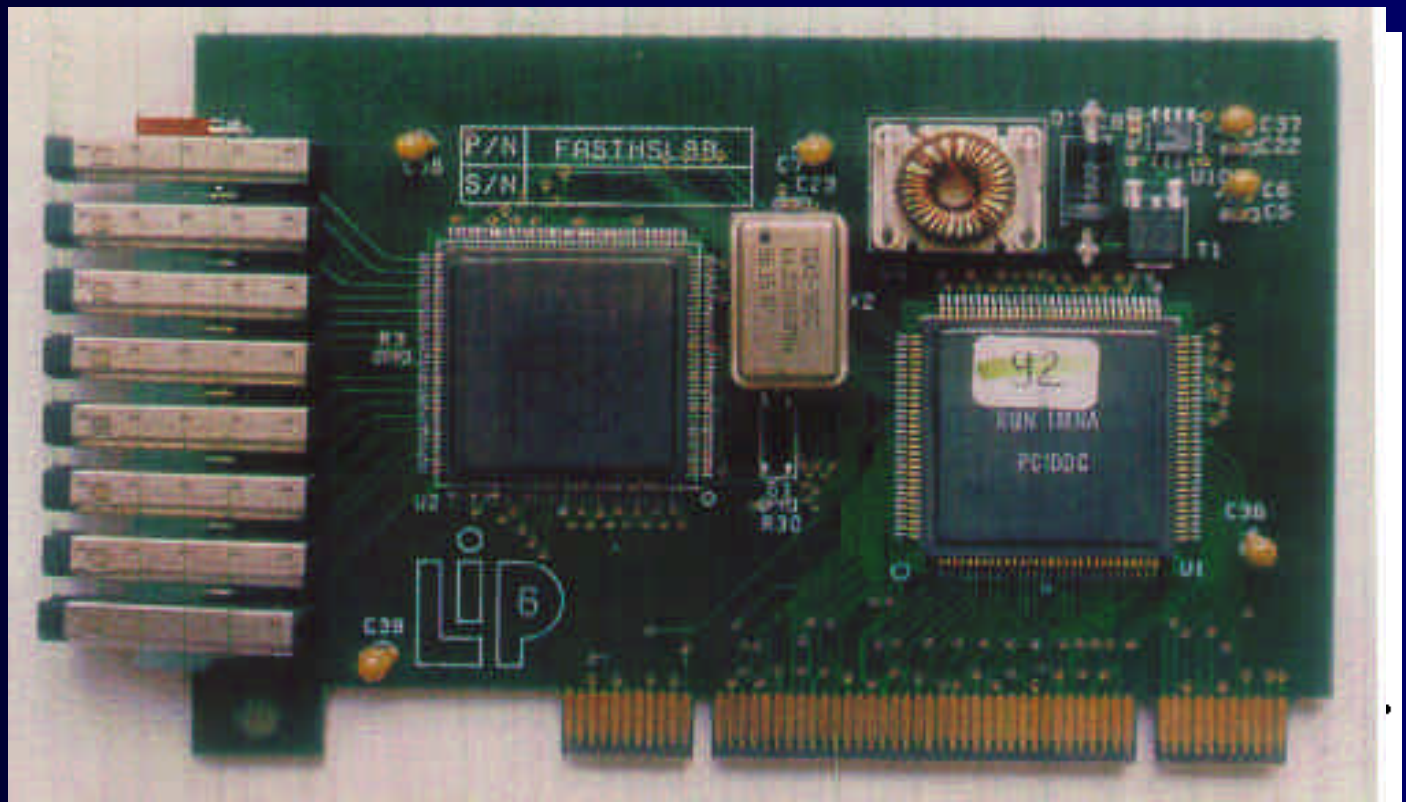
Architecture matérielle (1/2)

- Nœuds de calcul = PCs (Bi-pentium)
  - Machine MPC prévue pour plus de 250 nœuds
  - Réseau HSL
  - Réseau de contrôle : ETHERNET
- La technologie HSL (IEEE 1355 - 1993)
  - Deux composants VLSI
    - **RCUBE** : routeur rapide (8 liens HSL)
    - **PCI-DDC** : interface avec le bus PCI - réalise l'écriture distante (accès DMA)
    - bus PCI : 130 Mo/s - PCI-DDC : 160 Mo/s

Architecture matérielle (2/2)

– Le lien HSL

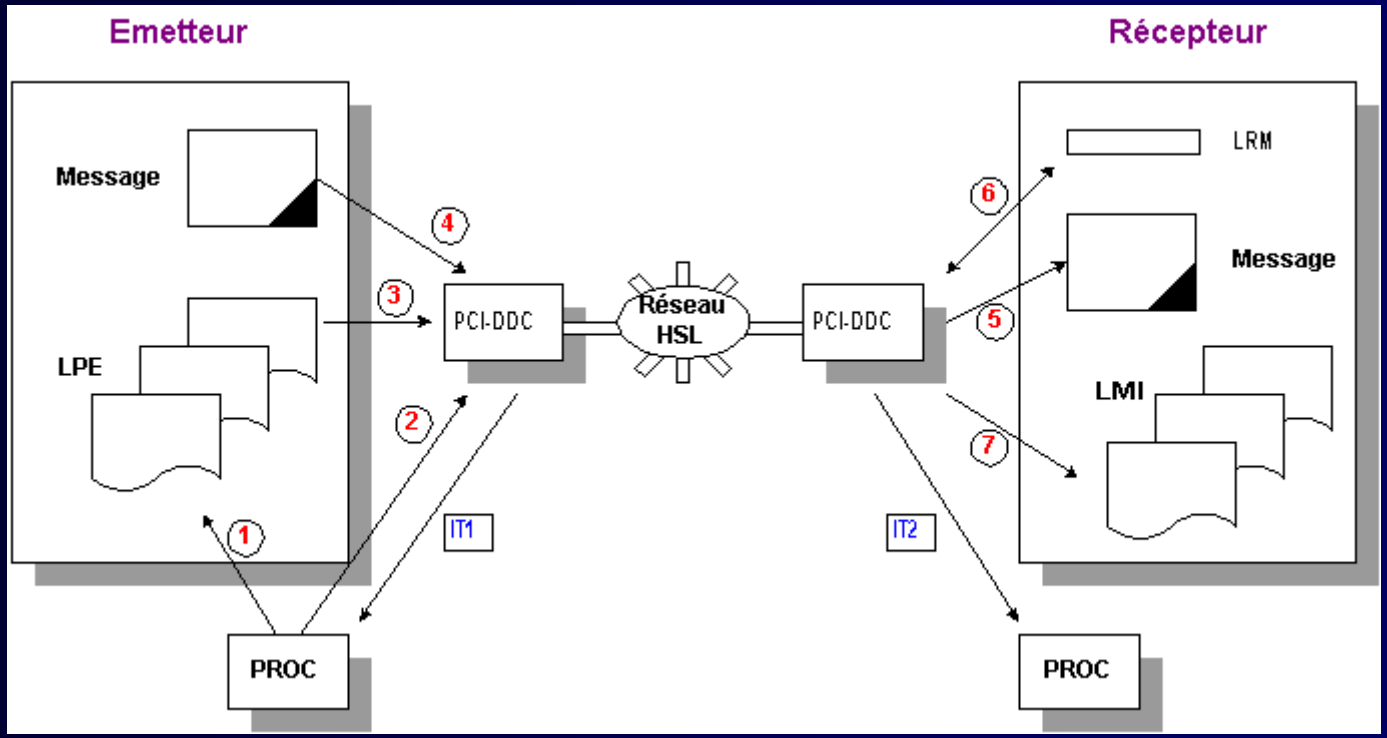
- lien série, point à point, bidirectionnel
- débit maximum : 1 Gigabit/s



L’écriture distante (1/2)

- Descripteur DMA (entrée de LPE) :

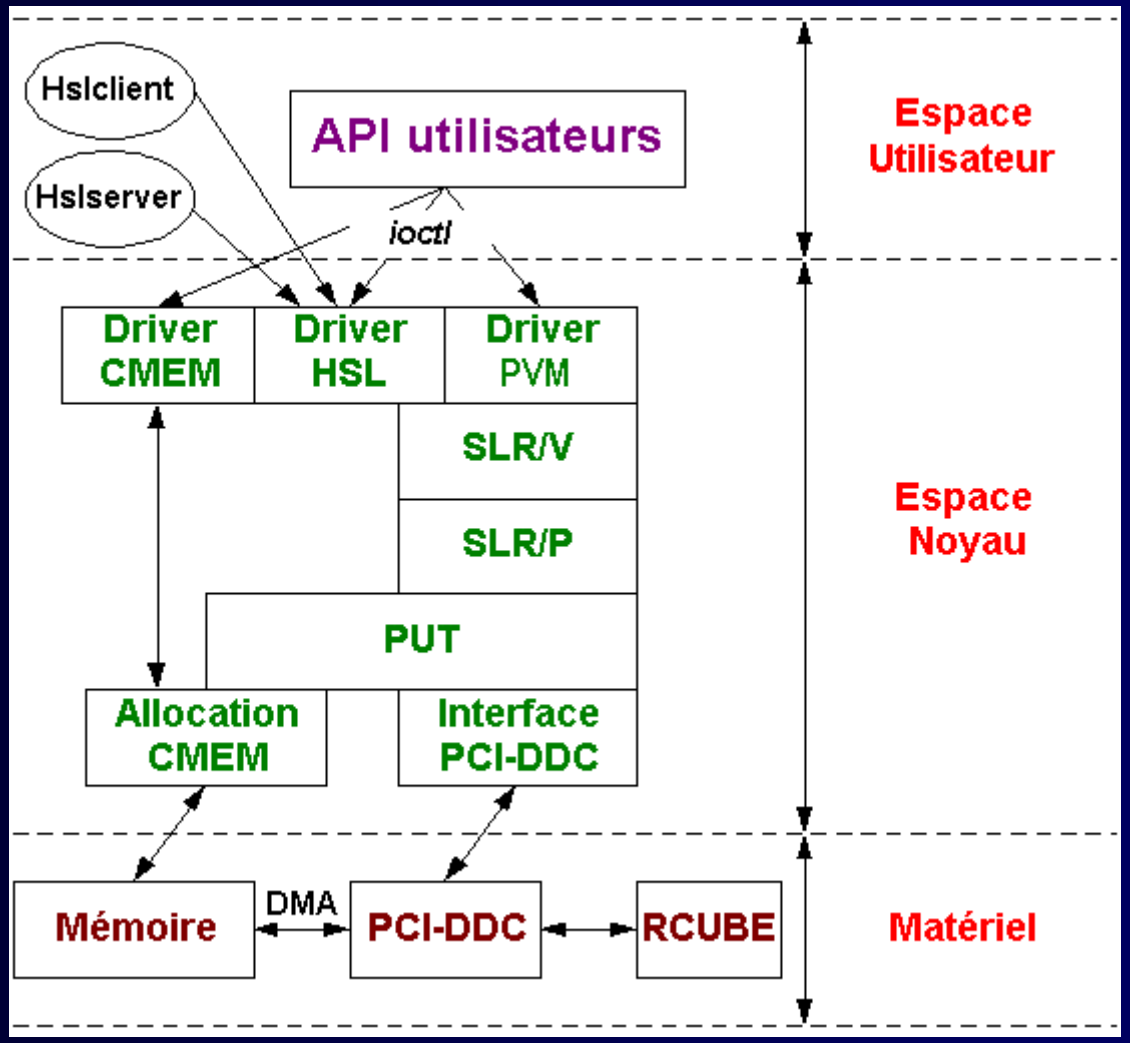
  nœud destinataire, adresse physique locale, nb d ’octets, adresse physique distante, ...



L’écriture distante (2/2)

- Caractéristiques du « Remote Write » :
  - Zéro copie (faible latence)
  - Récepteur actif (RDV préalable)
  - Le récepteur ne connaît pas la taille des messages qu’il reçoit
  - DMA → verrouillage mémoire

Les couches de communication MPC



Plan de la Présentation

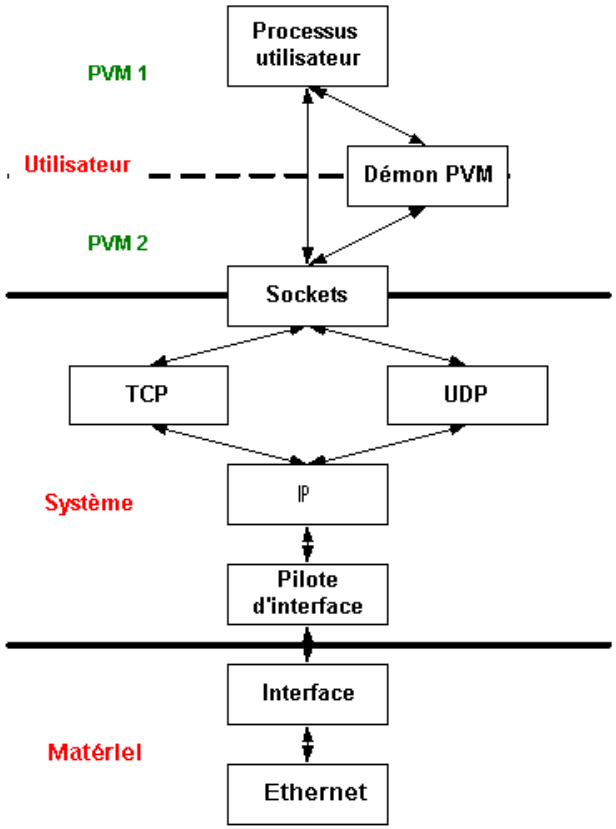
- Introduction / Sujet
- La machine MPC
- **L’environnement PVM-MPC**
- L’équation de Laplace
- Fast-PVM
- Conclusions

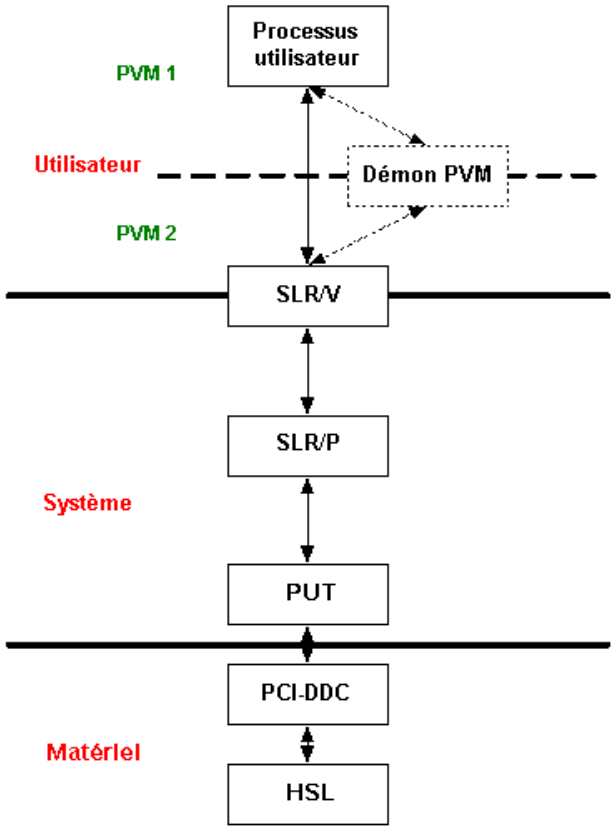
L’environnement PVM-MPC

- Les couches PVM
- PVM sur la machine MPC
  - Architecture de PVM-MPC
  - Les communications
  - Mesures de performances

PVM : Les différentes couches

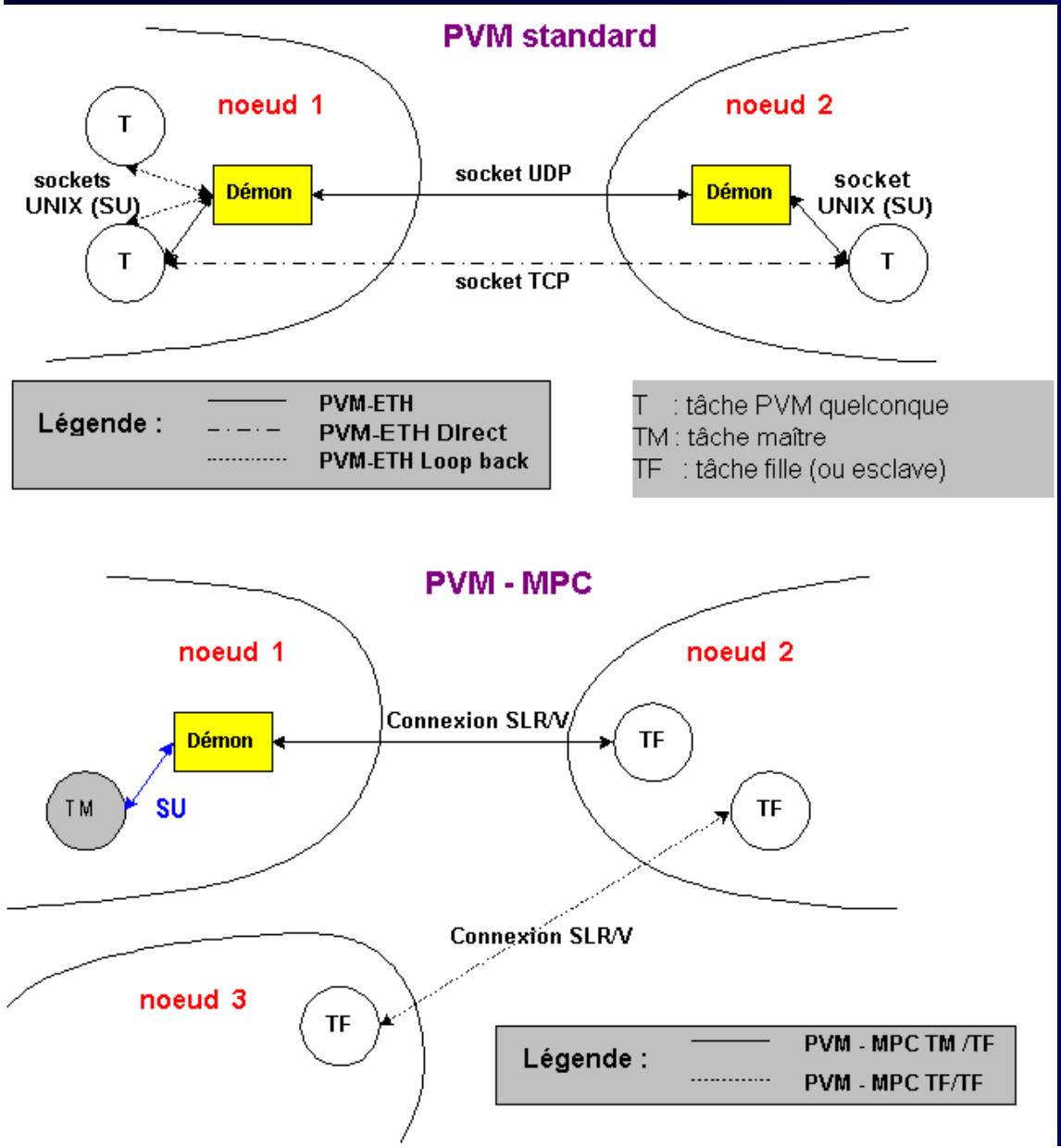
- Modèle à passages de messages - environnements hétérogènes
- Gestion de tampons utilisateur
- Primitives d ’émission et réception





- Portage au niveau de SLR/V
- PVM1 : gestions des tampons, pack
- PVM2 : transfert des fragments sur le réseau natif (appels systèmes - ex. pilote PVM-MPC)

Architecture et communications

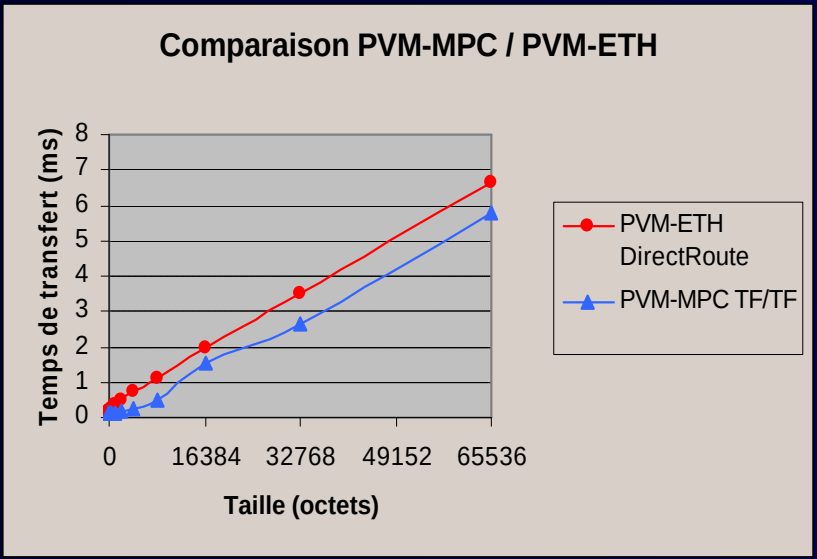


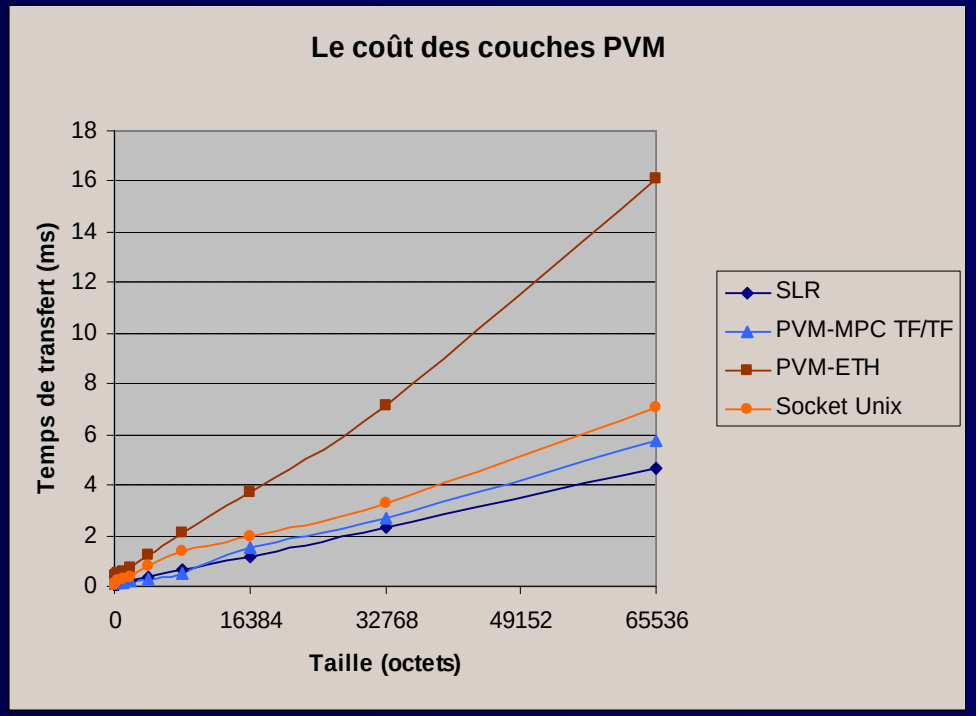
- PVM 3.3 (Ethernet ou MPC)
- Architecture à démon unique (cf. Paragon)

Mesures de performance

- Prog de tests :

  ping-pong (100 aller-retours)





	SLR	SOCK	PVM-ETH LoopB	PVM-ETH	PVM-MPC TM / TF	PVM-ETH Direct	PVM-MPC TF / TF
Latence (µs)	47	151	199	306	1864	203	78
Débit (Mb/s)	113	74	82	33	3	79	91
Pente (REG)	0,07	0,1	0,09	0,23	2,5	0,1	0,087
Corrélation	1,0000	0,9935	0,9955	0,9957	0,9932	0,9991	0,9978

Conclusions sur PVM-MPC

- Couches basses :
  - SLR/V > TCP/IP
  - défauts matériels → débit limité
- Communications PVM TF/TF
  - connexion HSL permanente → bonnes perf.
- Communications PVM TM/TF
  - très mauvais résultats (170 ms pour 64 K)
- Corruption des données sur PVM-MPC
- Coût des couches PVM
  - recopies PVM (> 1 ms pour 64 K) mais SLR zéro copie

Plan de la Présentation

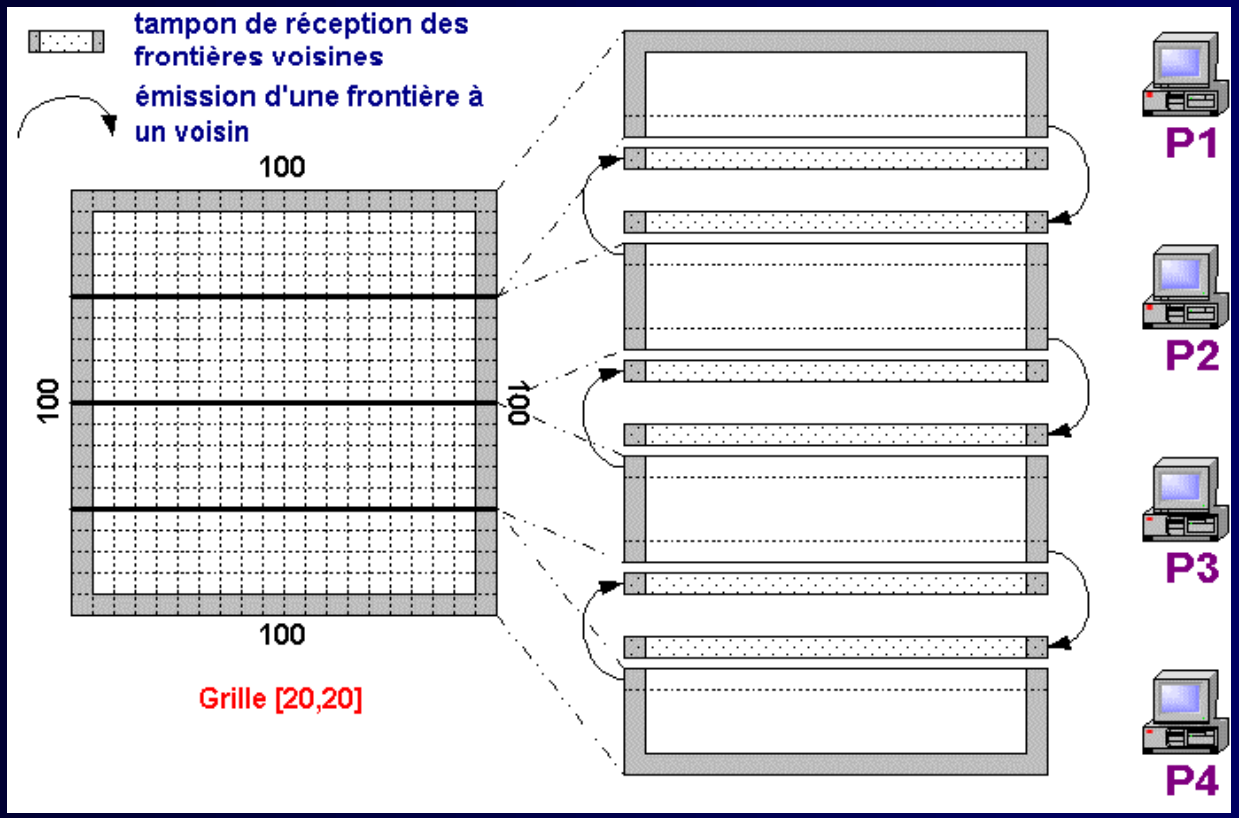
- Introduction / Sujet
- La machine MPC
- L’environnement PVM-MPC
- **L’équation de Laplace**
- Fast-PVM
- Conclusions

L’équation de Laplace

- Principe de parallélisation
  - itération de Jacobi
  - domaine de décomposition
- Méthode R&B
- Mesures sur ETHERNET

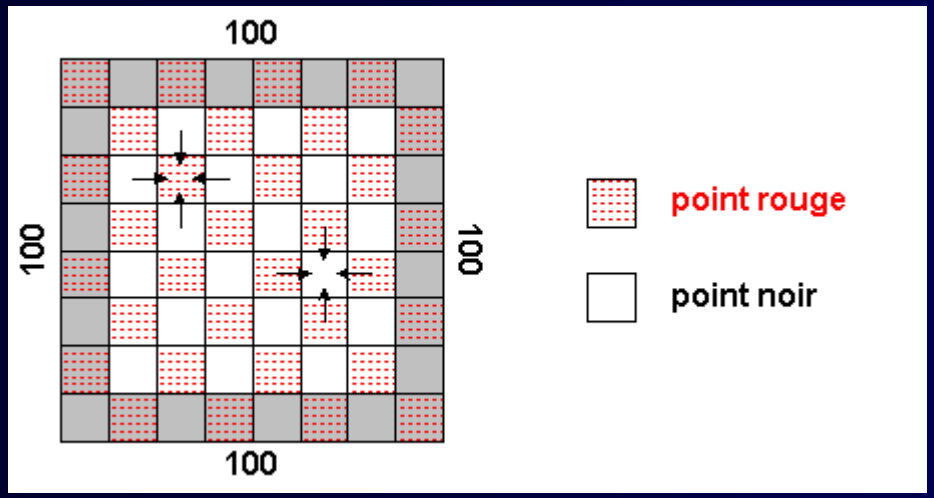
Principe de la parallélisation

- Discrétisation → méthodes itératives
- Itération de Jacobi : $$U_{i,j}^{(k+1)} = \frac{1}{4} * (U_{i+1,j}^{(k)} + U_{i-1,j}^{(k)} + U_{i,j+1}^{(k)} + U_{i,j-1}^{(k)})$$
- Décomposition en bandes



Méthodes R&B

- Décomposition en 2 phases
  - Phase 1 : mise à jour rouge
  - Phase 2 : mise à jour noire
- Convergence plus rapide



Une itération sur un esclave

// Phase 1	// Phase 2
recv(FN)	recv(FR)
calcul(FR)	calcul(FN)
send(FR)	send(FN)
calcul(IR)	calcul(IN)

Mesures sur ETHERNET

- Pas de mesures sur PVM-MPC
- Pas adapté à PVM sur des petites matrices (< 250*250)
- Mesures : 4 processeurs, frontières = 100, précision 0,01

Taille	12*12	120*120	200*200	260*260	524*524
Ratio Tcalcul/Ttotal	0.004	0.046	0.167	0.265	0.663
SPEED-UP	0.004	0.562	0.776	1.504	5.870

SPEED-UP = Temps séquentiel sur Temps //

Plan de la Présentation

- Introduction / Sujet
- La machine MPC
- L’environnement PVM-MPC
- L’équation de Laplace
- **Fast-PVM**
- Conclusions

Fast-PVM

- Objectifs
- Sémantiques
- Contraintes
- Conclusions

Objectifs de Fast-PVM

- Amélioration des performances
- Remplacement des couches actuelles

  → réduire le chemin critique logiciel
- Couches mieux adaptées à l'architecture MPC

  → utilisation directe de PUT
- Ni recopie, ni appel système
- PUT en mode utilisateur
- Sémantique proche de celle de PVM

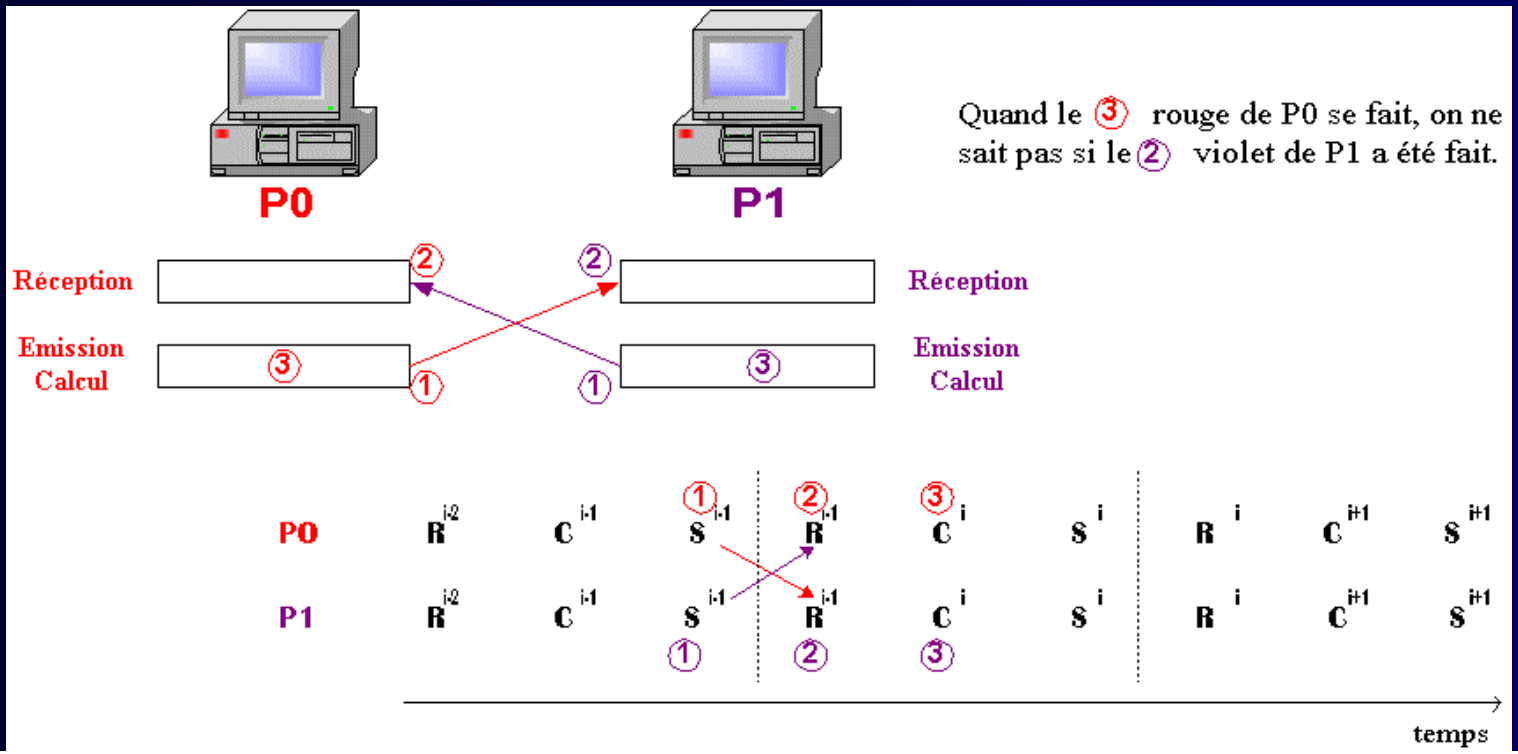
Sémantique de Fast-PVM

- PUT en mode utilisateur

  → 1 seule application à la fois
- 1 zone d ’émission-calcul et 1 tampon de réception par (tid, mstag)
- 3 types de fonctions
  - acquisition / relâchement des ressources réseau
  - gestion des tampons (appel système)
  - émission / réception
- Emission non bloquante (≠ PVM)

Contraintes

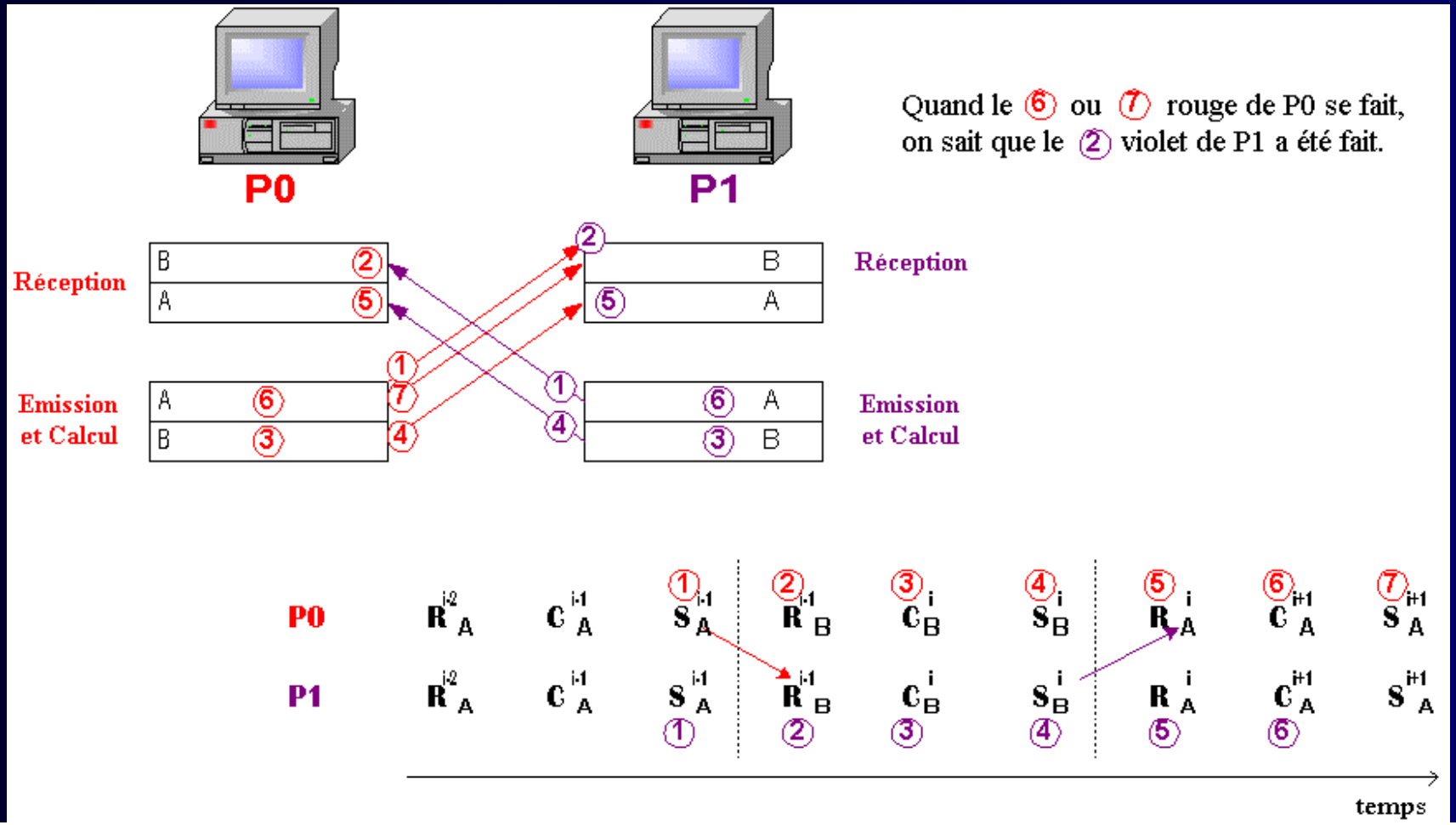
- Réutilisations des mêmes tampons (structure de boucle)
- 2 Problèmes : (exemple d'une séquence Recv/Calcul/Send)
  - modification des données avant l'envoi effectif
  - écrasement des données avant consommation en réception



Contraintes

→ mécanisme de synchronisation (accusé de réception)

Par exemple :



Conclusions sur Fast-PVM

- Adapté à un certain type d ’applications
  - tampons statiques
  - mécanisme de synchronisation
- Application type : méthode R&B
- Problèmes non résolus :
  - partage des ressources réseau
  - PUT en mode utilisateur
  - transformation automatique de portions de code PVM
  - mariage avec PVM

Conclusion

- Bilan du stage
  - JMS pour la machine MPC
  - Etat de fonctionnement de PVM-MPC
  - Benchmark Laplace pour la machine MPC
  - Etudes de nouvelles couches de communication
- Perspectives
  - Exploitation de la machine MPC (JMS/PVM/PM²)
  - Reprendre PVM-MPC
  - Validation de Fast-PVM
  - Projet MPC