Conception et réalisation d'un noyau de communication bâti sur la primitive d'écriture distante, pour machines parallèles de type "grappe de PCs"

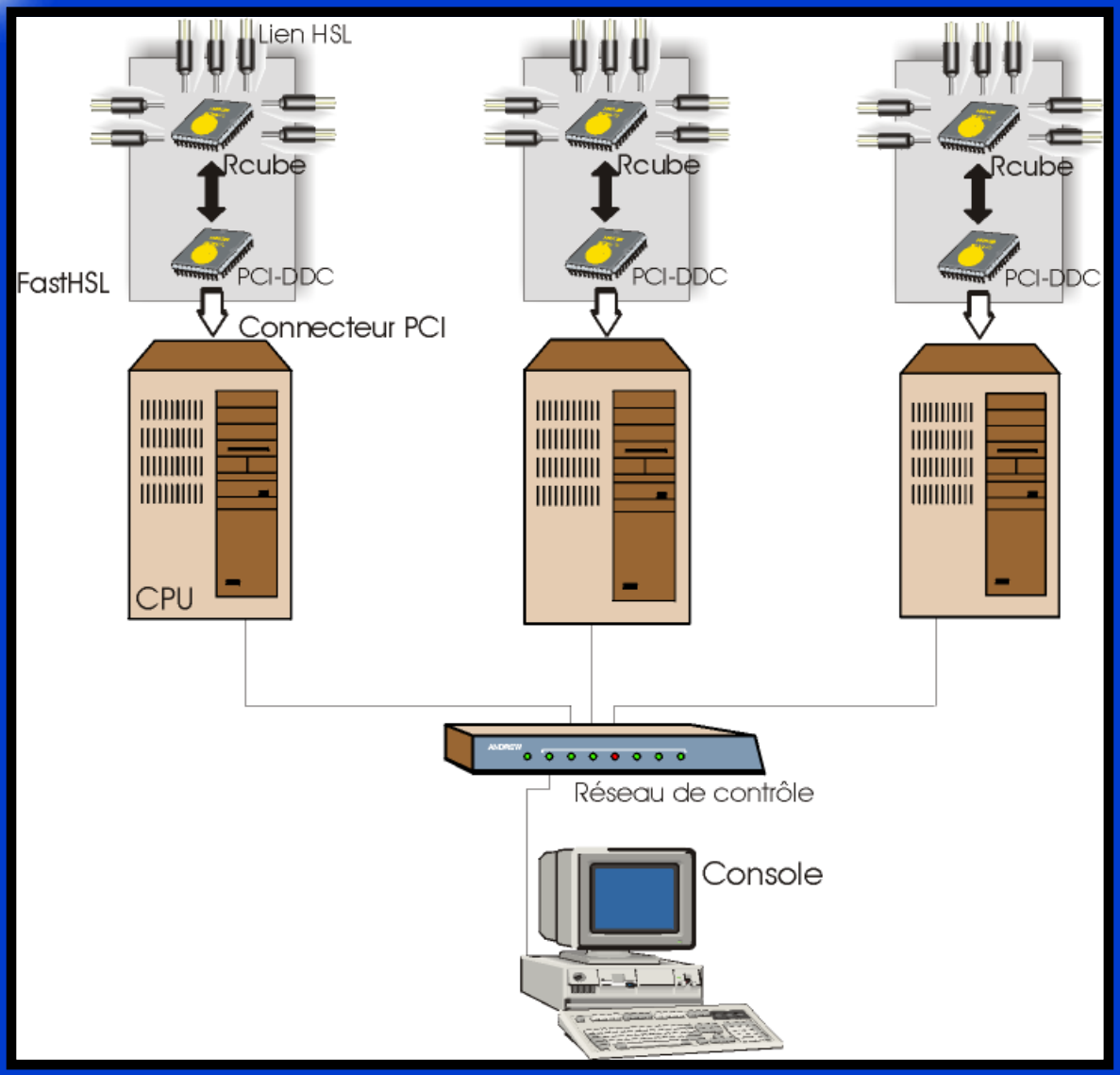
MPC_OS

Alexandre Fenyo
UPMC/LIP6/ASIM

PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

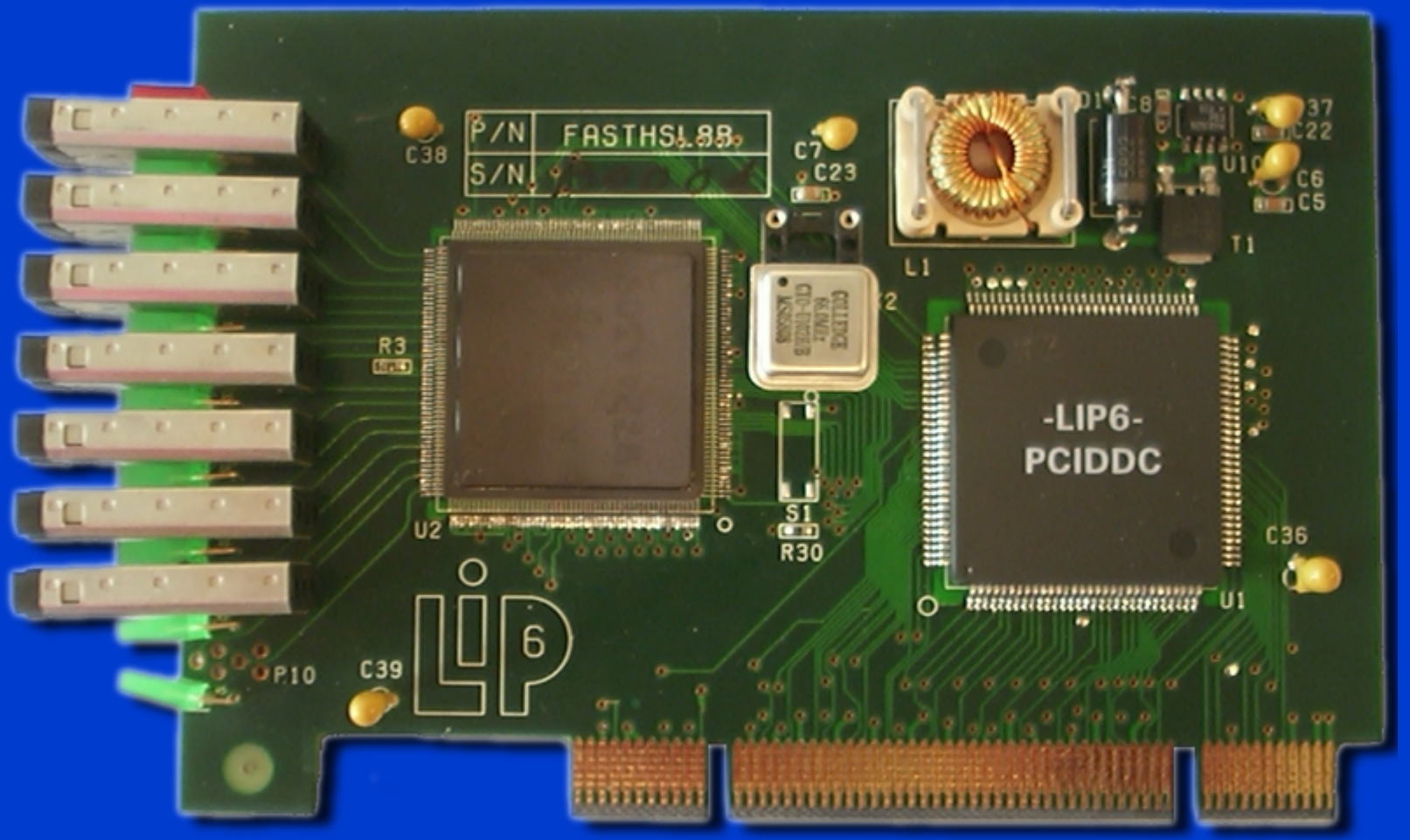
LA MACHINE MPC



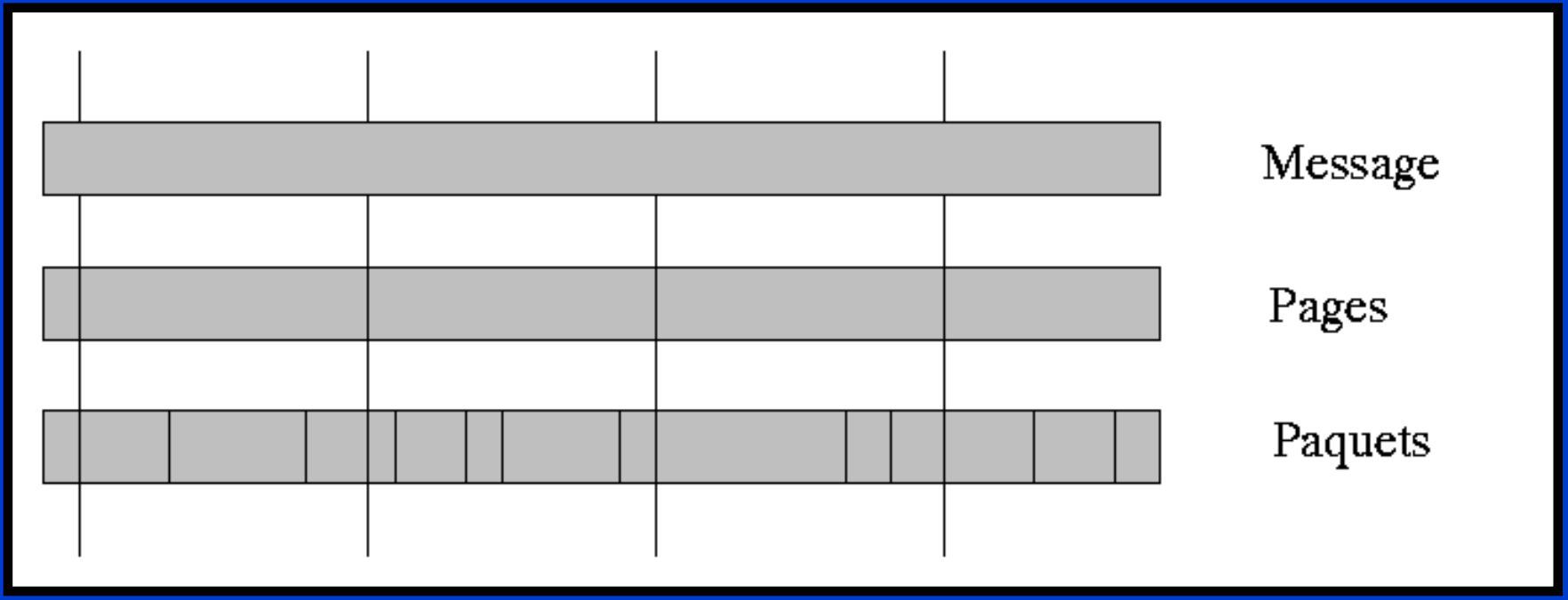
MACHINE MPC (4 nœuds bi-pro)

CONNECTIQUE

LA CARTE FastHSL



DECOUPAGE DES MESSAGES

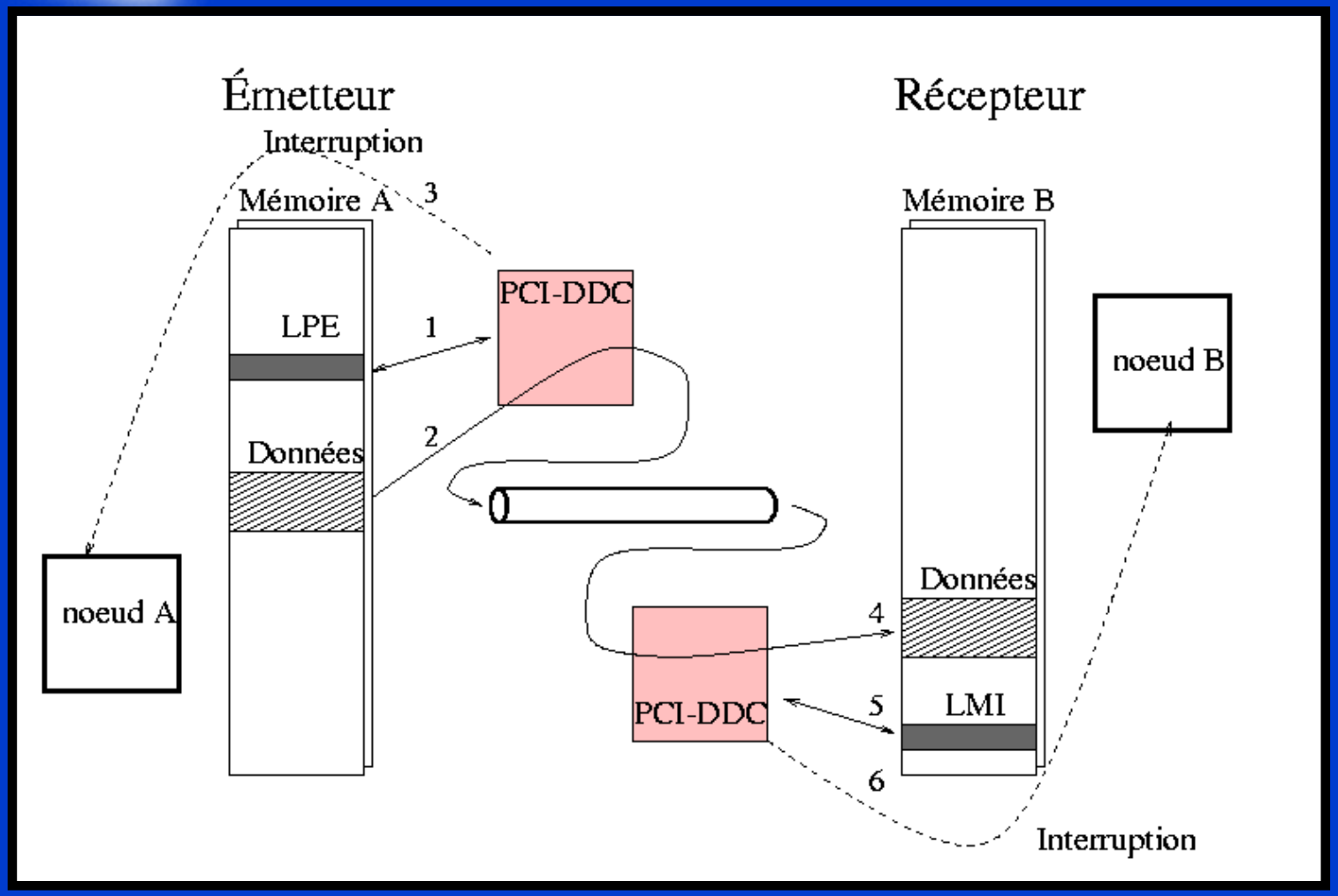


Message : contigu en mémoire virtuelle

Page : contiguë en mémoire physique

Paquet : atomique au niveau réseau

ECRITURE DISTANTE



2 types de messages : normaux et courts

OBJECTIFS

1. Fournir des services à forte valeur ajoutée :
   - canaux de communication entre processus
   - transmissions sécurisées
   - gestion des ressources

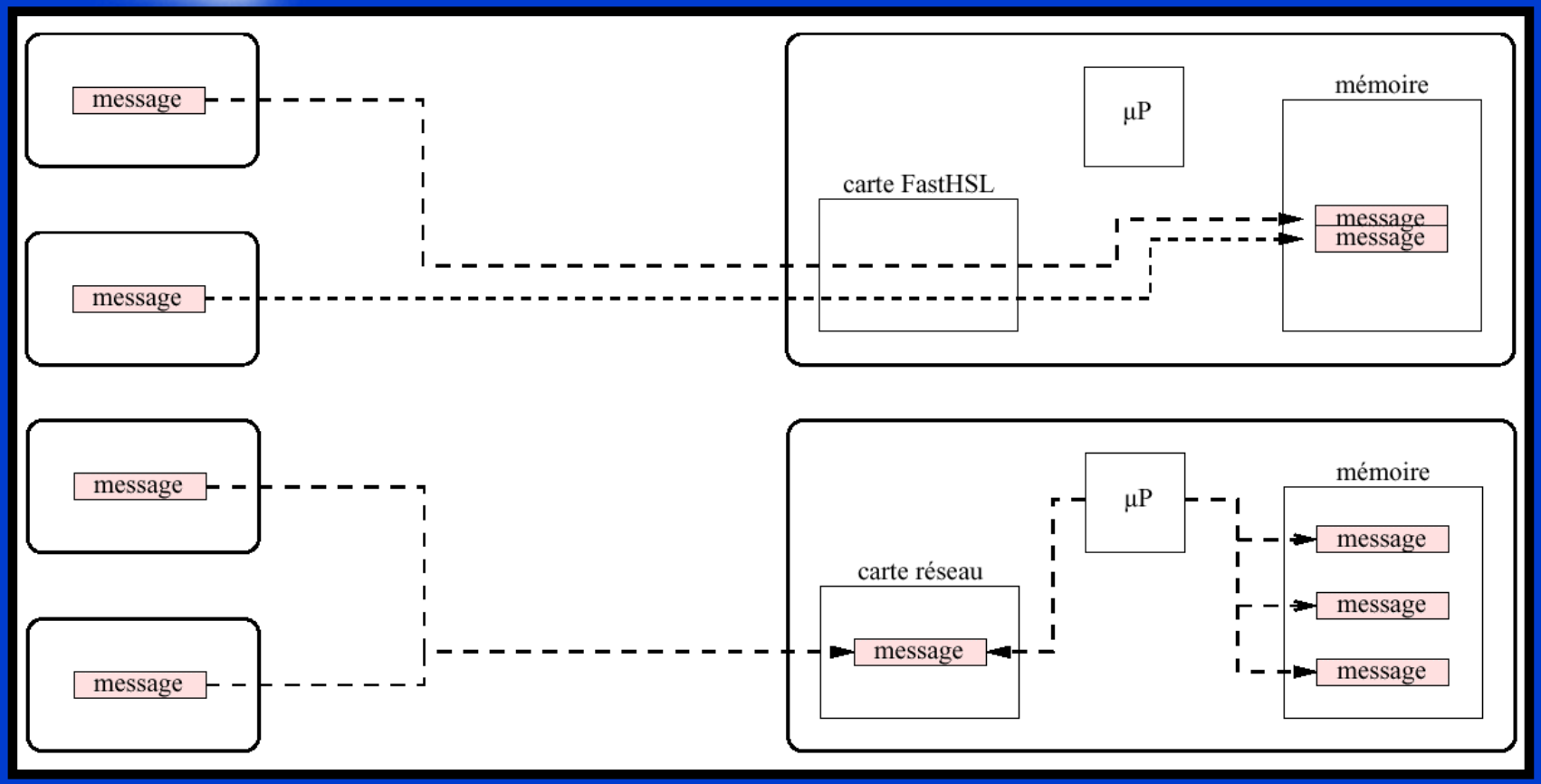
2. Contrainte : écriture distante (Remote DMA)

3. Hautes performances : zéro-copie

PLAN

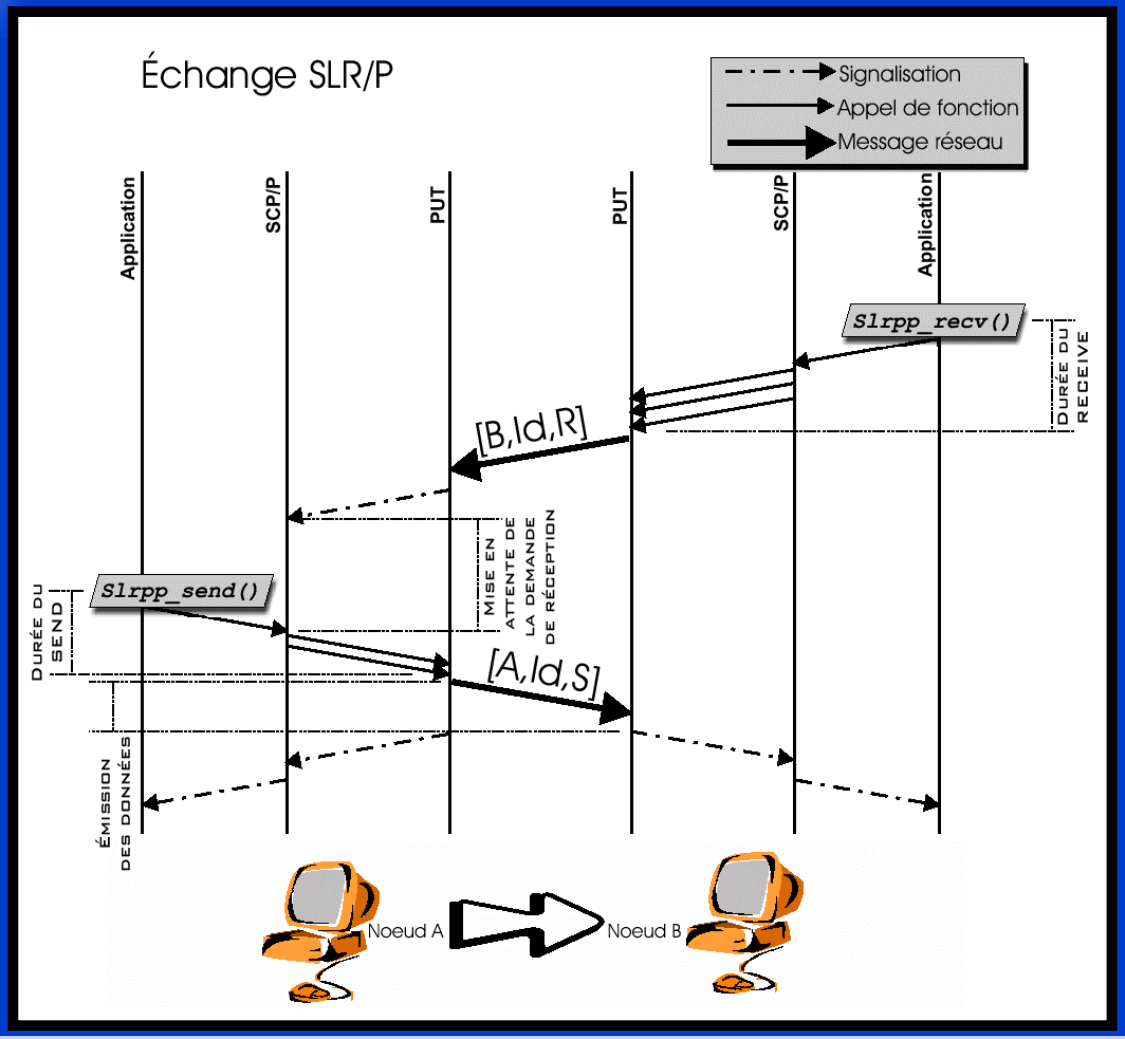
1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

DEPOT DIRECT



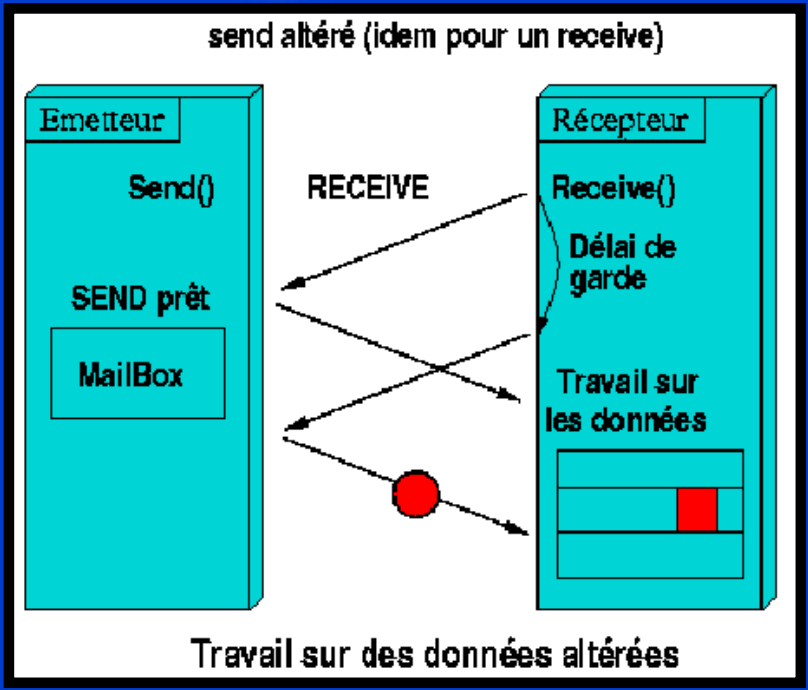
Dépôt direct = Pas de contrôle du côté récepteur
Emplacement du dépôt ? Gestion des fautes ?
Exception : Messages Courts

PROTOCOLE DE RENDEZ-VOUS



Localisation des emplacements physique par Boites-aux-lettres

DIFFICULTES



1- classification des fautes

2- délai de garde

3- problèmes résiduels :

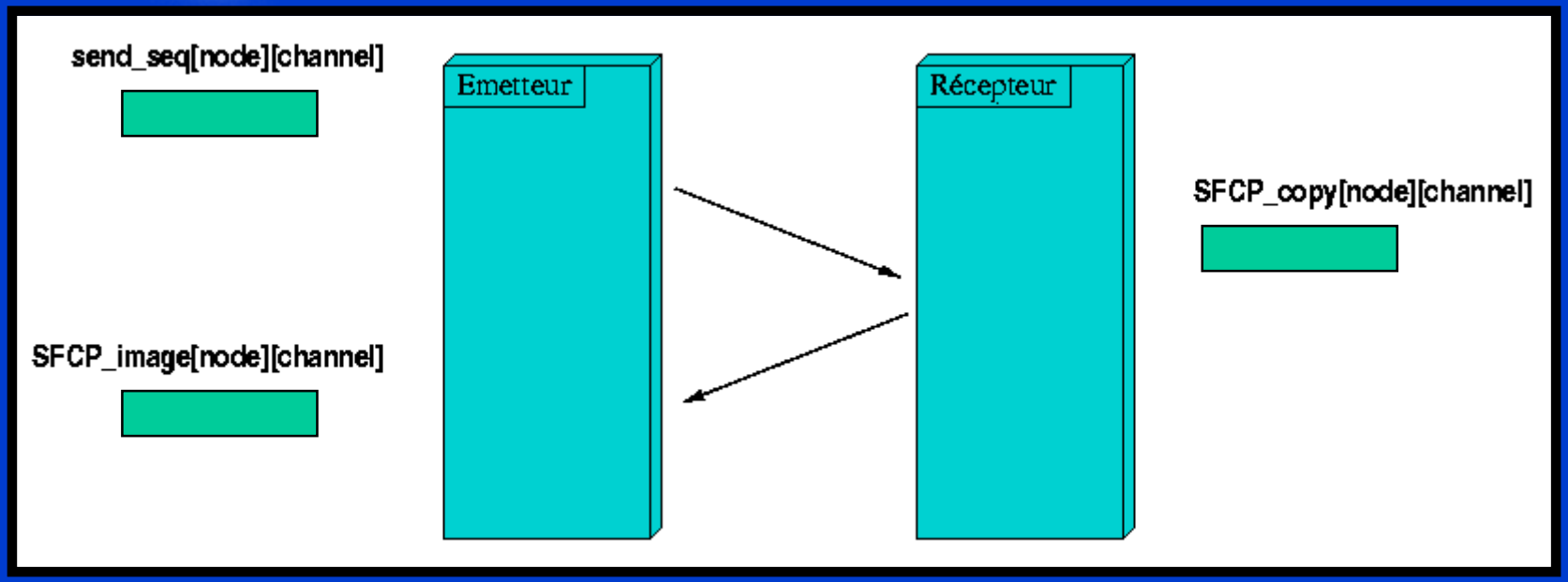
validité des données déposées

libération des tampons et signalisation

4- solution : messages courts

- Perte de SEND
- Perte de RECEIVE
- Paquet corrompu
- Paquet retardé

INFORMATIONS DE CONTROLE



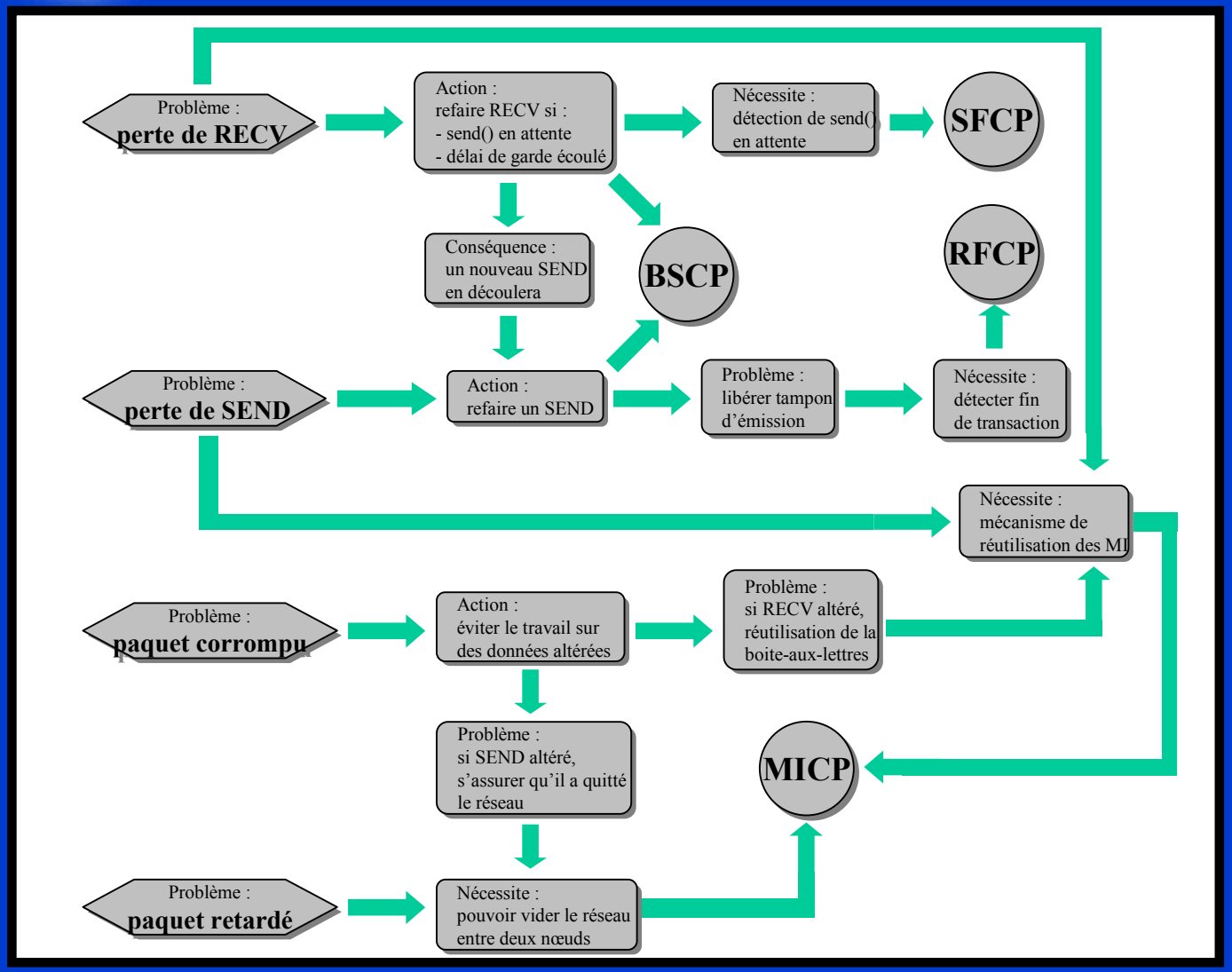
- valeurs croissantes autorisées uniquement
- send_seq[][] ≤ SFCP_copy[][] ≤ SFCP_image[][]
- terminé quand send_seq[][] = SFCP_image[][]

VIDER LE RESEAU

Principe:

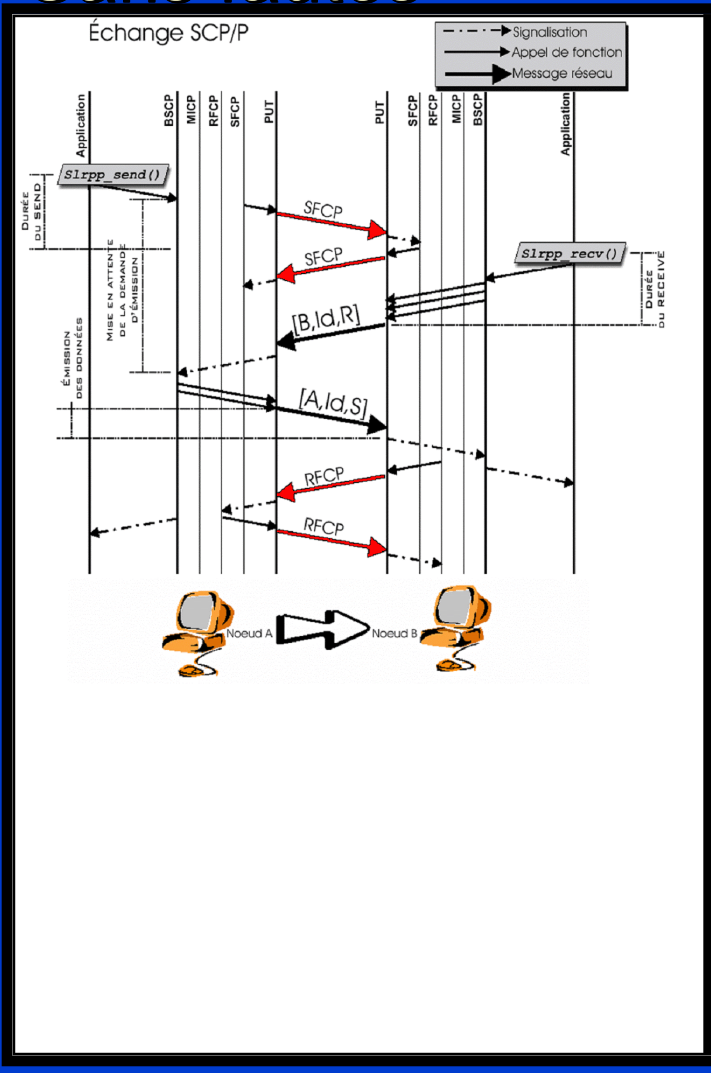
En choisissant des tables de routage non adaptatives, ainsi qu'un protocole de ping/pong analogue au précédent, on peut facilement vider le réseau entre deux nœuds.

DECOMPOSITION DU PROBLEME

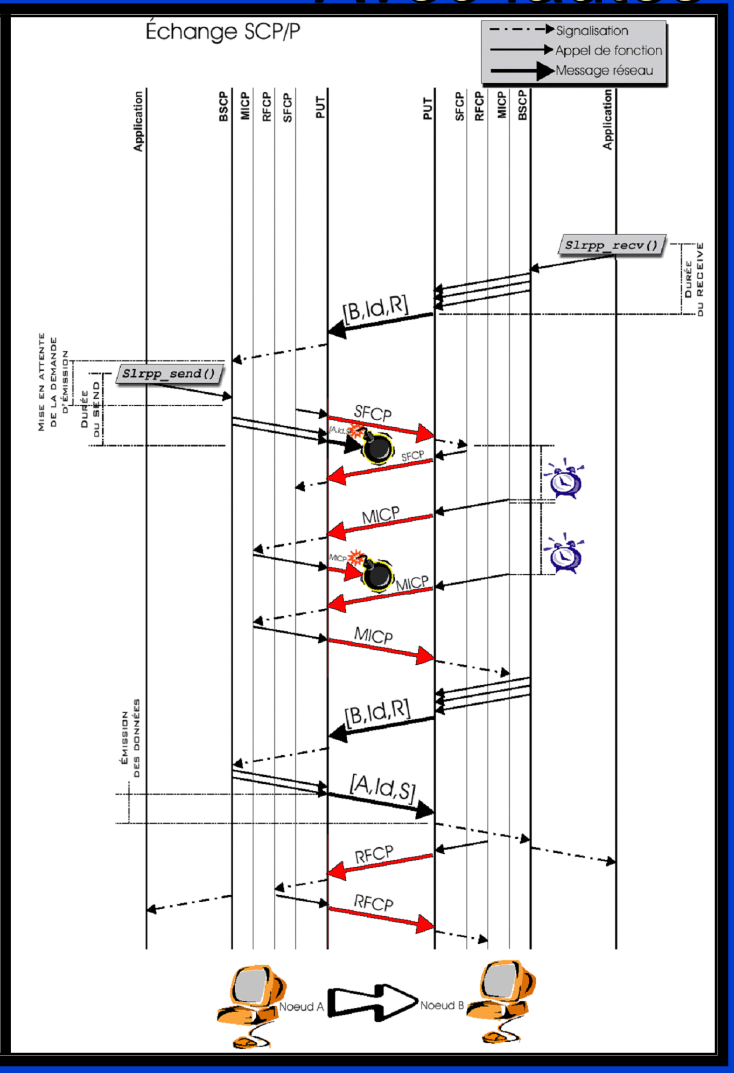


EXEMPLES D'ECHANGES

Sans fautes



Avec fautes

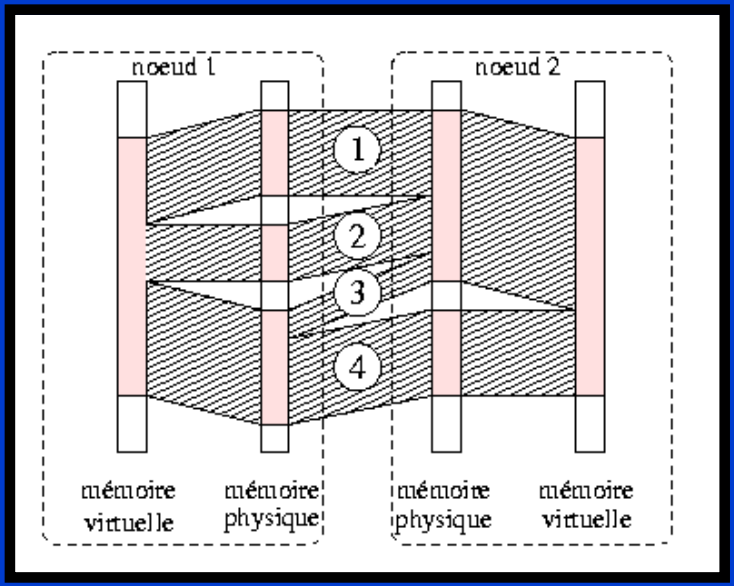


PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

MEMOIRE VIRTUELLE

1- Problématique : dépôt direct & mémoire virtuelle

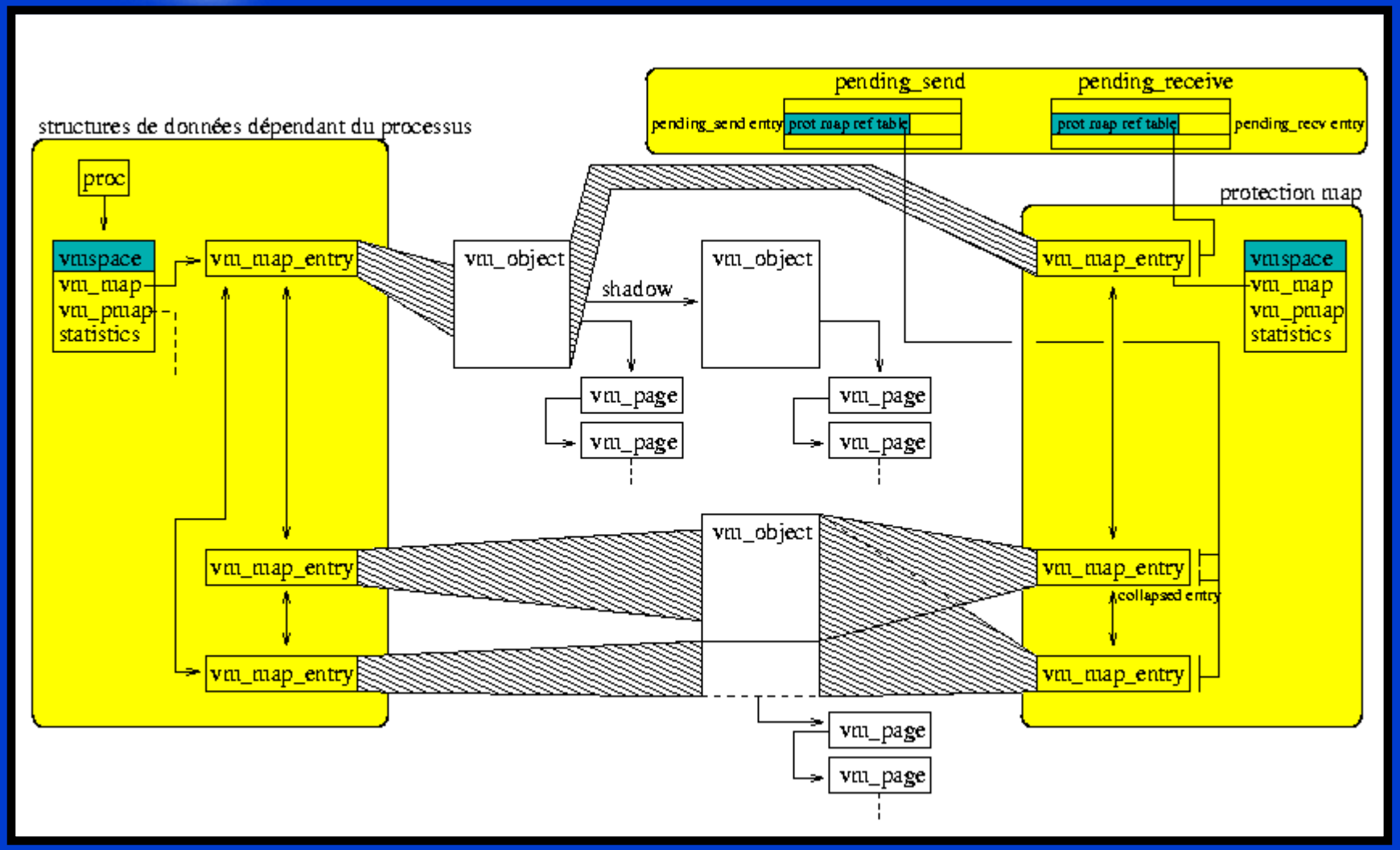


Le récepteur fournit dans les boites-aux-lettres la structure de son espace virtuel

2- Protection contre les fautes des applications

- Décomposition d'un message en pages réseau
- Nécessité de verrouiller les tampons pendant l'échange

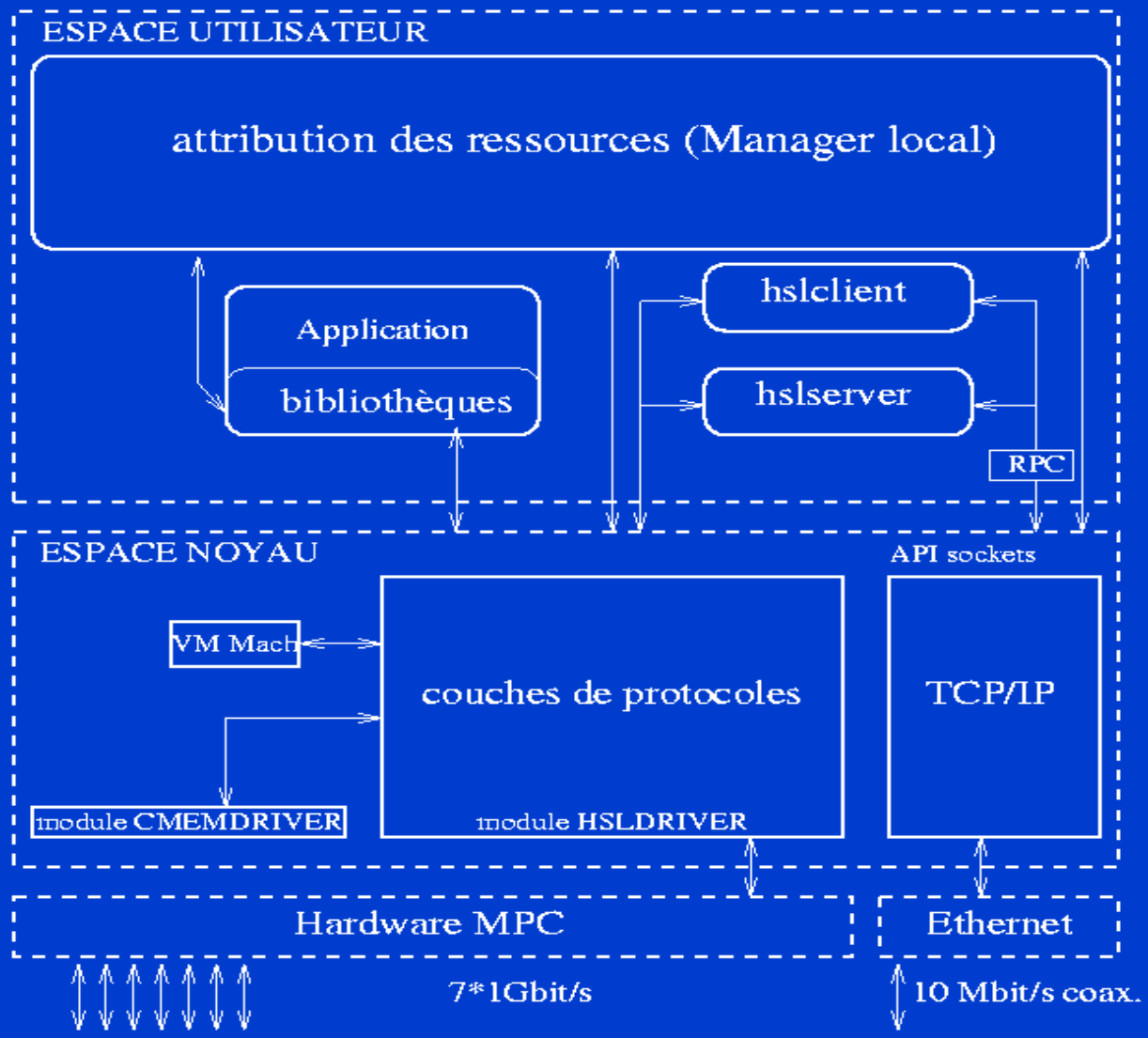
VERROUILLAGE DES OBJETS



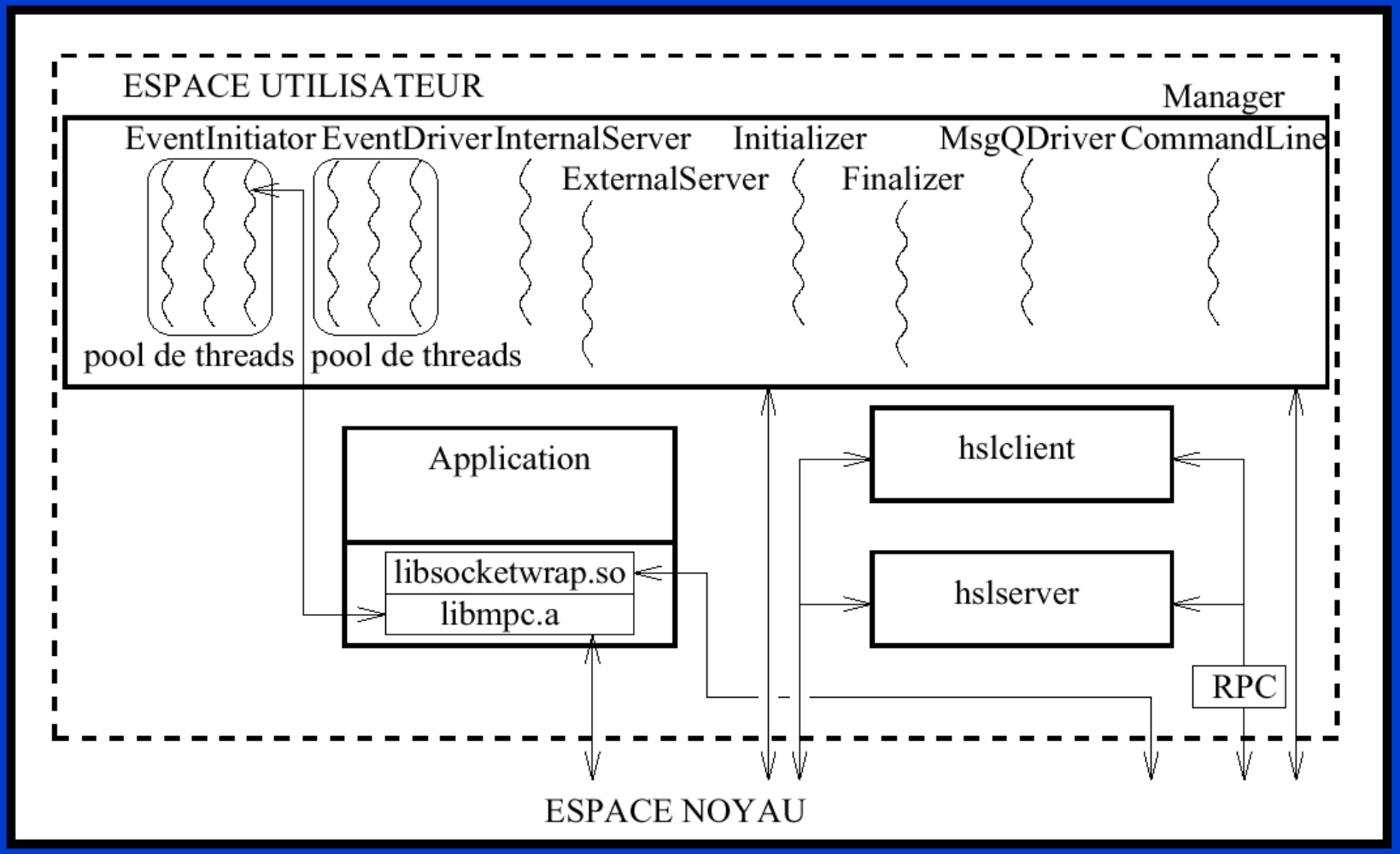
PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

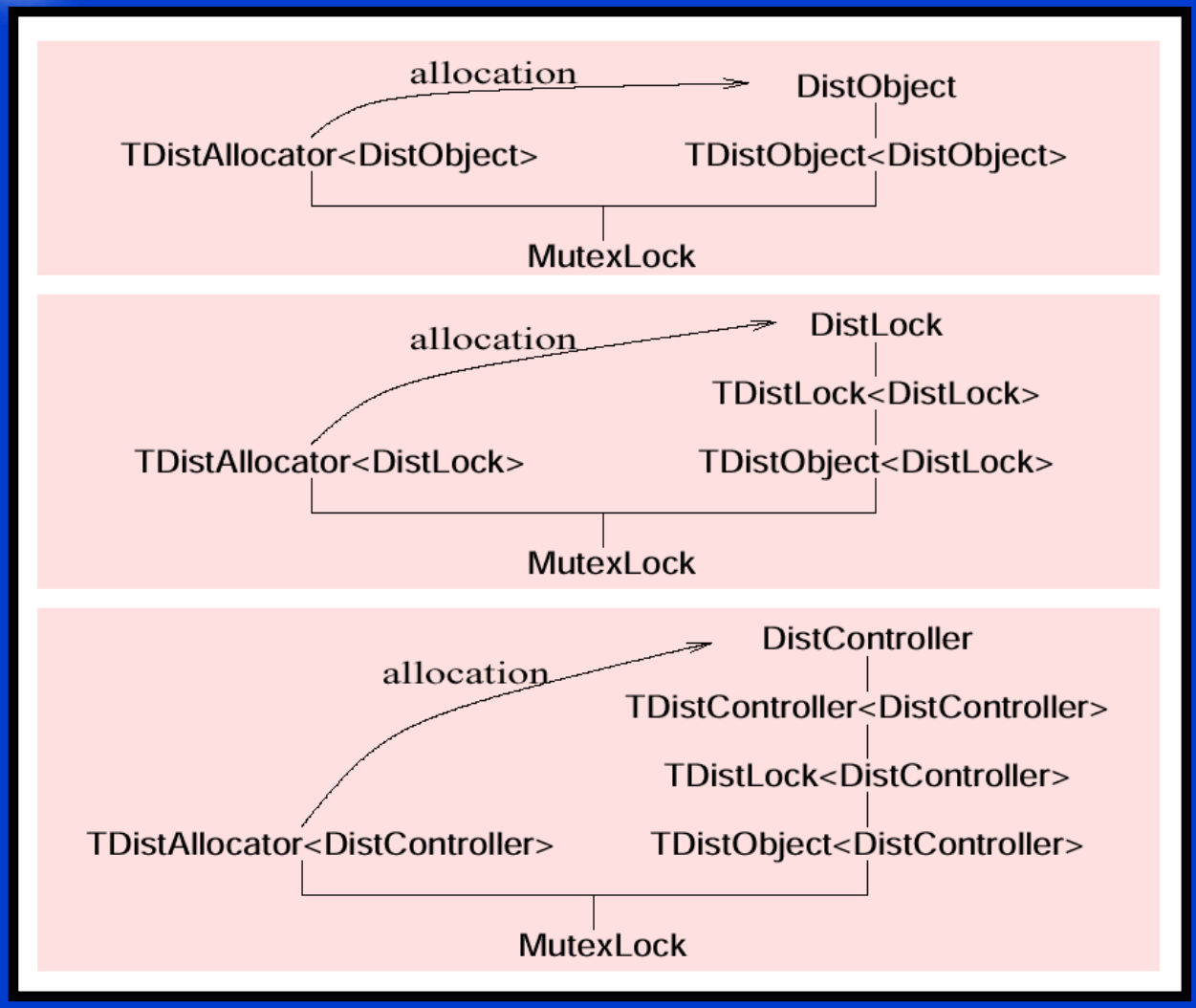
ORGANISATION DE MPC-OS



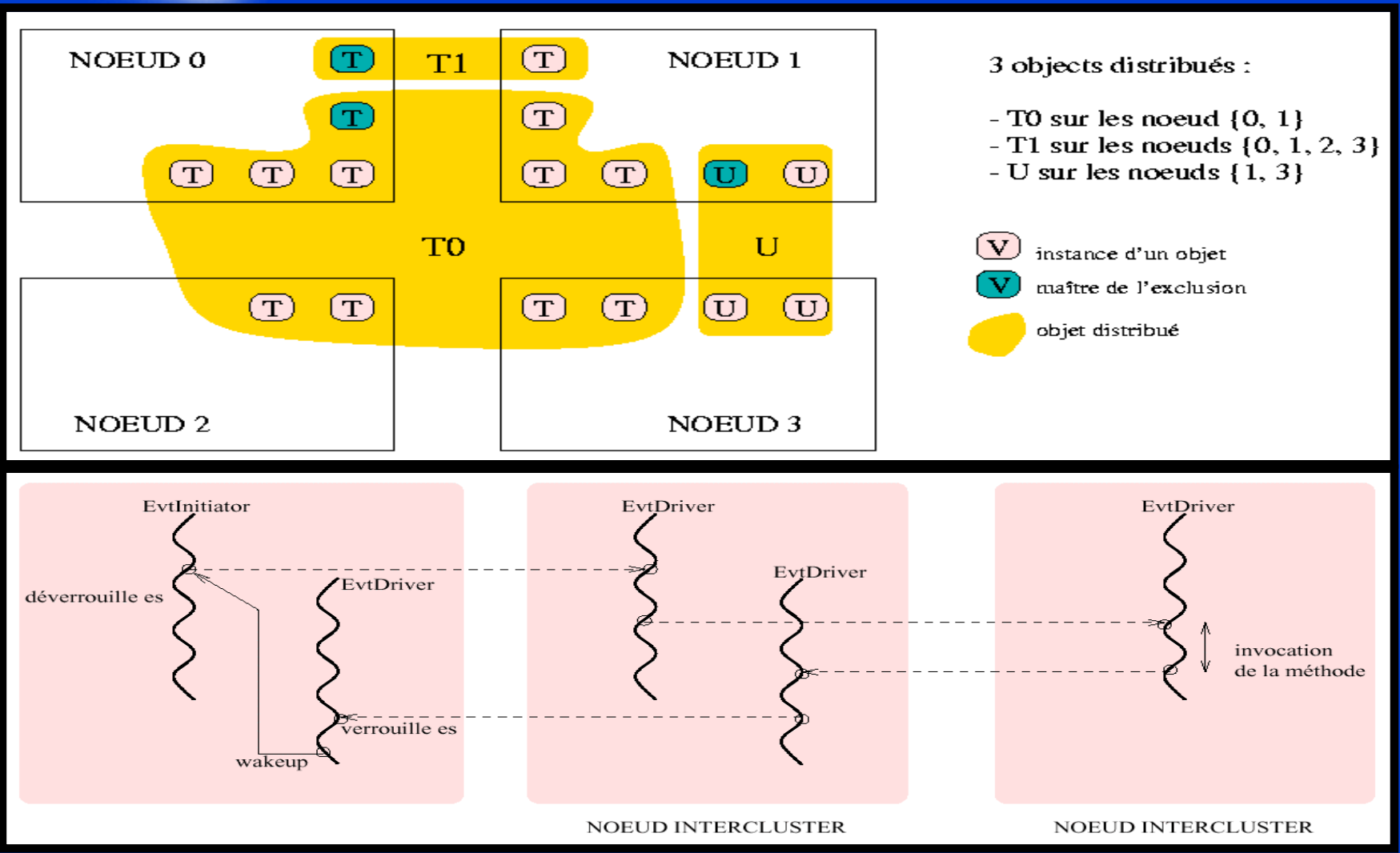
ACTIVITES AU SEIN DU MANAGER



DERIVATION D'OBJETS



OBJETS DISTRIBUES



result = @d_obj,reply,<T>:method(params);

ORGANISATION DU MANAGER

1- Création de tâche

2- Création de canal négocié entre deux tâches :

Nécessite 9 opérations alternativement dans un nœud puis dans l'autre

3- Fermeture de canal pour réallocation future :

. Nécessite 13 opérations alternativement dans un nœud puis dans l'autre, prenant place au sein des noyaux

. Il faut assurer qu'il n'y a plus aucun message en cours de transfert, et ceci en supposant que le réseau est non fiable

PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

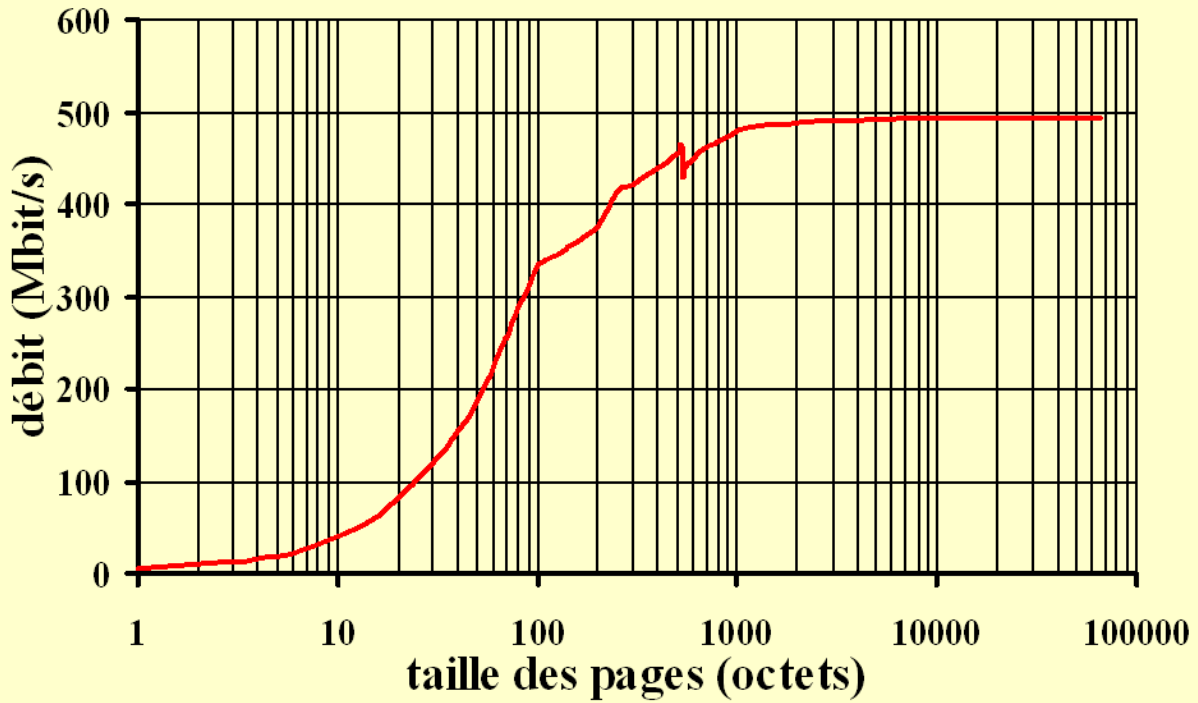
PERFORMANCES BAS NIVEAU

Put()	+	latence matérielle	+	signalisation (polling)	=	latence d 'un transfert
1,9µs	+	>1,7µs <2,1µs	+	>10ns <400ns	=	4,0µs

Surcoût de l'appel système : 1,1µs

Mesures effectuées sur deux PC Pentium II 350MHz, chipset 440BX, carte FastHSL avec oscillateur à 66MHz.

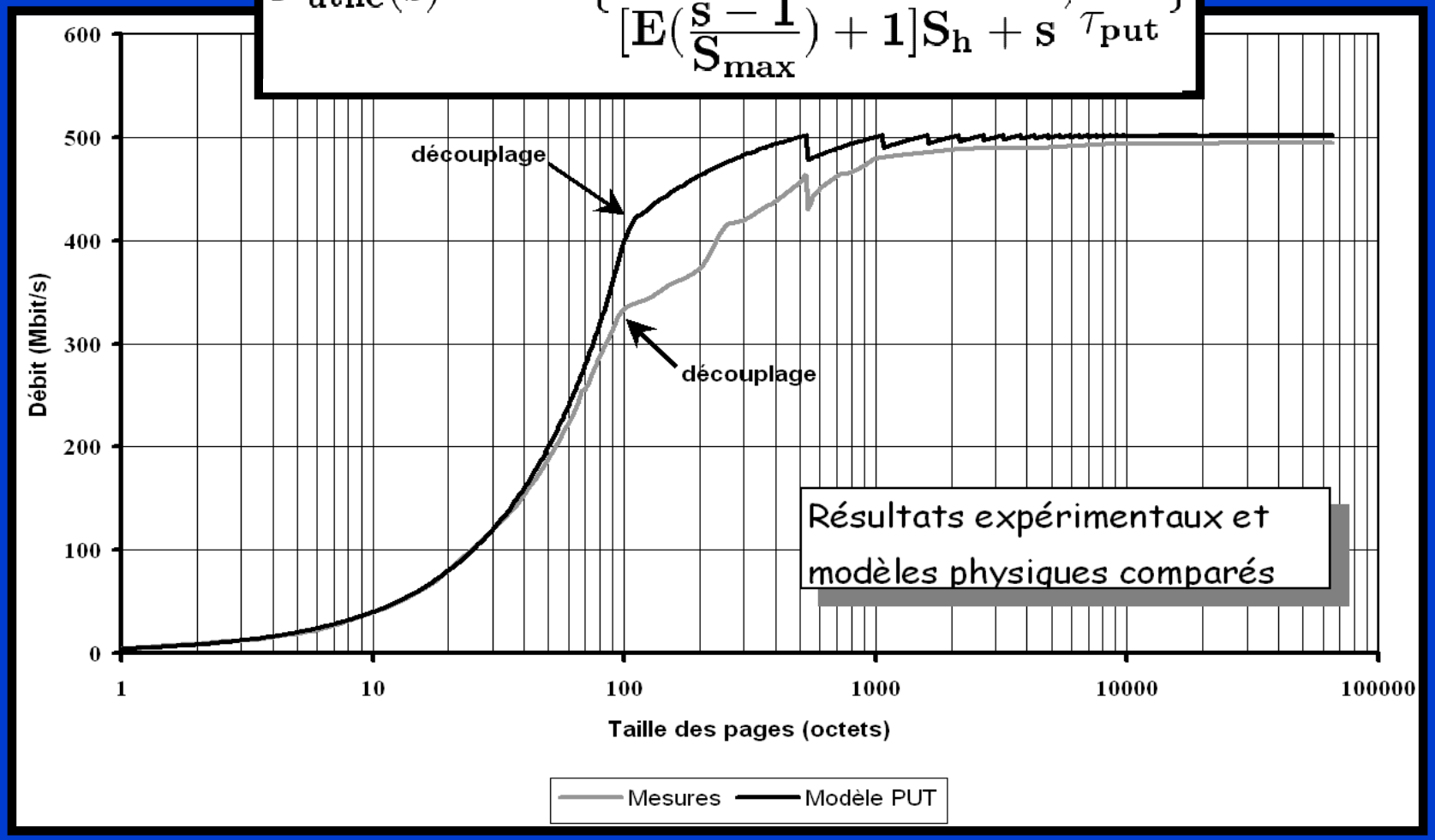
MESURE DU DEBIT



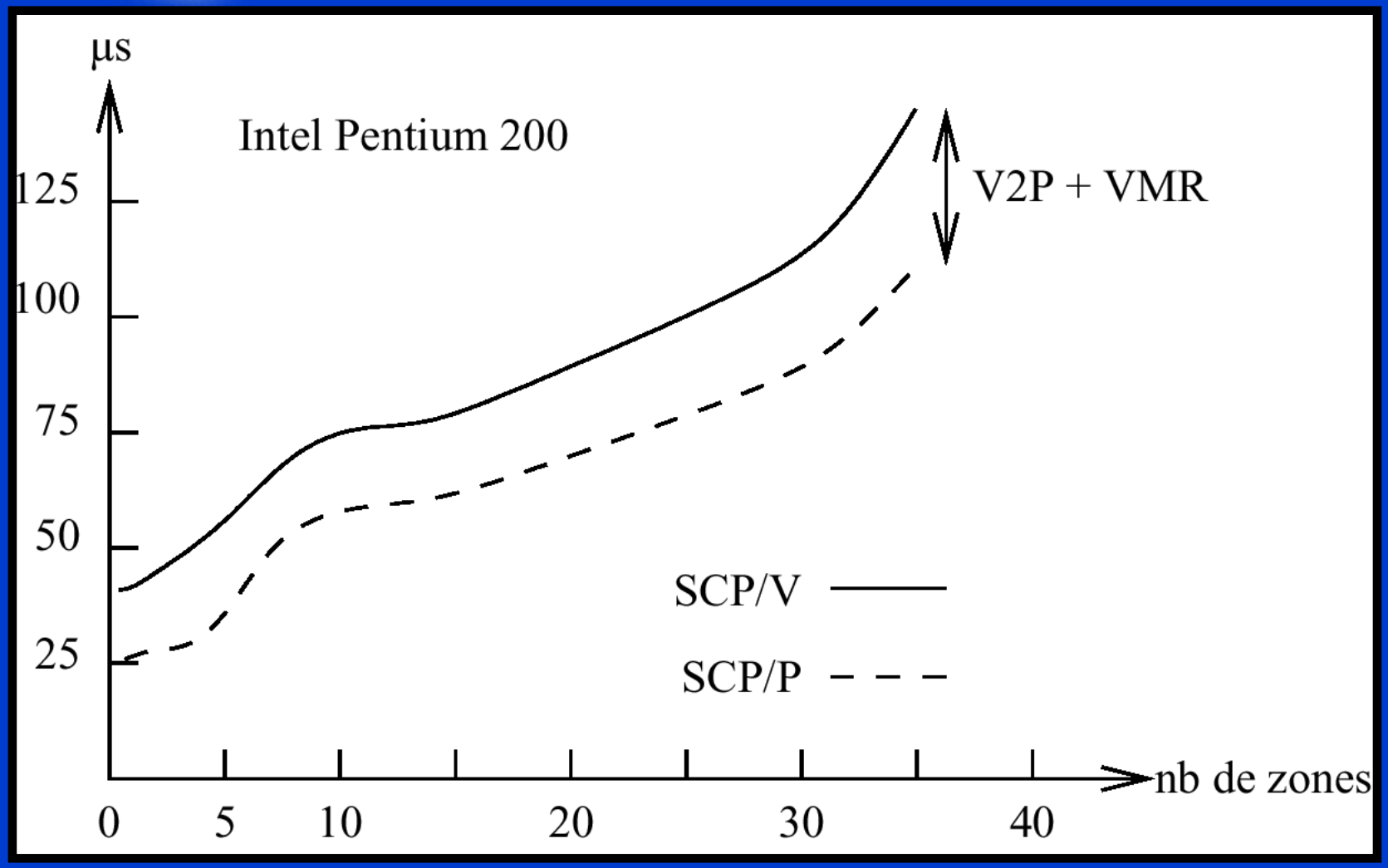
PC Pentium II 350MHz
Chipset 440BX
Débit minimum (1 octet/page) : 3,9 Mbit/s
Débit maximum (65535 octets/page) : 494 Mbit/s
Demi-bande (247Mbit/s) : 66 octets

MODELE DU DEBIT

$$\mathcal{D}_{\text{utile}}(s) = \min\{\frac{s\, D_{DDC/R^3}}{[E(\frac{s-1}{S_{max}})+1]S_h + s}, \frac{s}{\tau_{put}}\}$$



PERFORMANCES HAUT NIVEAU



PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

MODELISATION

X_0, X_1, X_2, etc. : (X_n), inter-arrivées des fautes
Y_1, Y_2, Y_3, etc. : (Y_n), délai de correction de faute

Les X_n suivent la même loi
Les Y_n suivent la même loi

Les X_n et Y_n sont mutuellement indépendants

$\Psi_{X0, Y0}$ représente le délai avant double faute
Ex.: si $X_1(\omega) > Y_1(\omega)$ et $X_2(\omega) \leq Y_2(\omega)$, alors :
$\Psi_{X0, Y0}(\omega) = X_0(\omega) + X_1(\omega) + X_2(\omega)$

MODELISATION

Résultats :

$$\mathbf{E}(\Psi_{\mathbf{X},\mathbf{Y}}) = \mathbf{E}(\mathbf{X_1}) \left[1 + \frac{1}{1 - \mathbf{P}(\mathbf{X_1} > \mathbf{Y_1})}\right]$$

$$\mathrm{V}(\Psi_{X,Y}) = \mathrm{V}(X_1) + \frac{\mathrm{V}(X_1) + E(X_1)^2}{1 - P(\Delta_1)} - \frac{E(X_1)^2}{(1 - P(\Delta_1))^2} + \frac{2E(X_1)}{(1 - P(\Delta_1))^2} \int_{\Delta_1} X_1 \, dP$$

Continuité :

$$X_n \xrightarrow{d} U \text{ et } Y_n \xrightarrow{d} V, \text{ alors } \Psi_{X_n,Y_n} \xrightarrow{d} \Psi_{U,V}$$

MODELE DE LA MACHINE MPC

MODELE

fautes sans mémoire et délai de réparation constant :

$$f_X : \mathbb{R} \longrightarrow \mathbb{R}$$

$$t \mapsto \begin{cases} \lambda\, e^{-\lambda t} & \text{si } t \geqslant 0 \\ 0 & \text{si } t < 0 \end{cases}$$

$$f_Y = \delta_\tau : t \mapsto \delta(t - \tau)$$

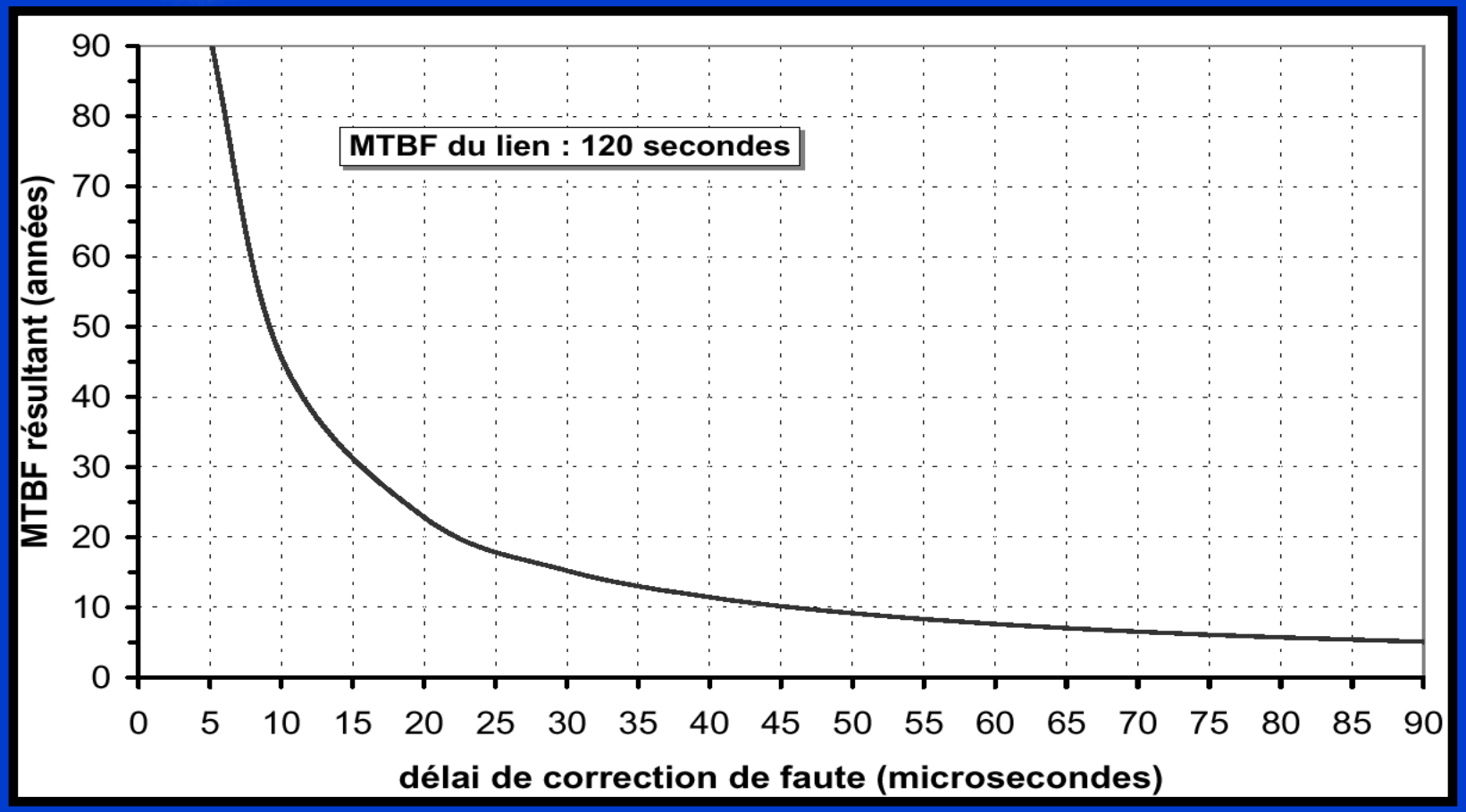
RESULTATS

$$\sigma(\Psi_{X,Y}) \underset{\tau \to 0}{\sim} \frac{1}{\lambda^2\, \tau}$$

$$E(\Psi_{X,Y}) \underset{\tau \to 0}{\sim} \frac{1}{\lambda^2\, \tau}$$

$$MTBF(mpc) = \frac{D_{max}\ MTBF(hsl)}{N\ v\ D_{utile}} \left[1 + \frac{1}{1 - e^{-\frac{\tau}{MTBF(hsl)}}} \right]$$

MTBF DU LIEN



Etude de la répartition : si τ=30µs, seulement 1 millième des doubles-fautes se produisent au cours des 5 premiers jours.

PLAN

1. Introduction : la machine MPC
2. Sécurisation des communications
3. Garantie d'intégrité du système
4. Gestion dynamique des ressources
5. Performances
6. Approche stochastique
7. Conclusion

Conclusion et perspectives

On a développé un noyau de communication fournissant

- des canaux de communication entre processus
- avec transmissions sécurisées
- et gestion dynamique des ressources
- le tout sur la primitive d'écriture distante (Remote DMA)

Conclusion et perspectives

Mais, des limites :

Le caractère zéro-copie conserve le débit, pas la latence

Solution proposée :

un protocole sécurisé stochastique

Conclusion et perspectives

Perspectives : projet ANI

avec un contrôleur réseau reprogrammable (FPGA),

on peut adapter le protocole de bas niveau pour optimiser l'interface logicielle-matérielle afin de construire des protocoles fiables et très efficaces