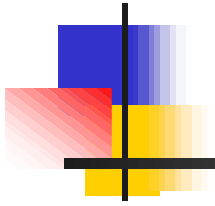
L'interface générique de communication Madeleine II



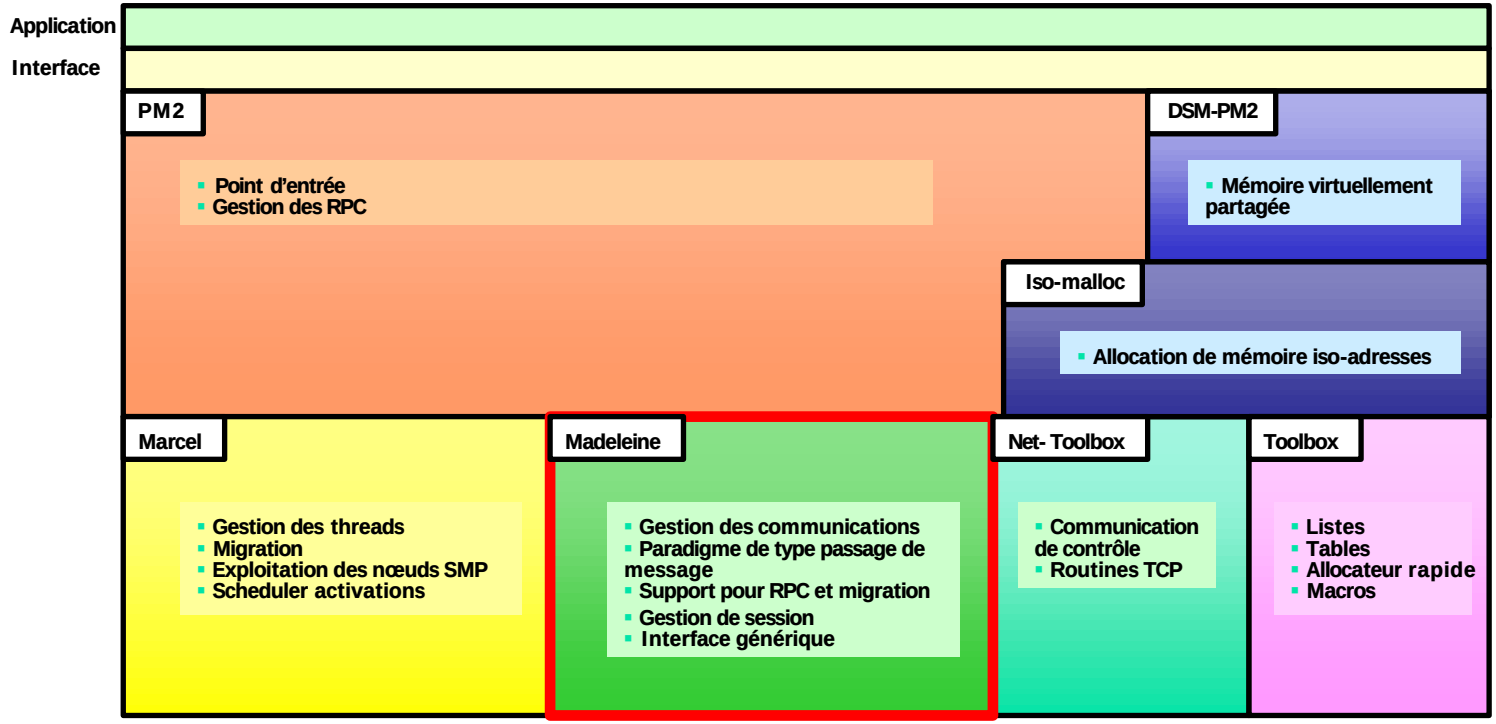
Communications multi-grappes

Olivier Aumage

LIP - ENS Lyon

Olivier.Aumage@ens-lyon.fr

Madeleine, PM2

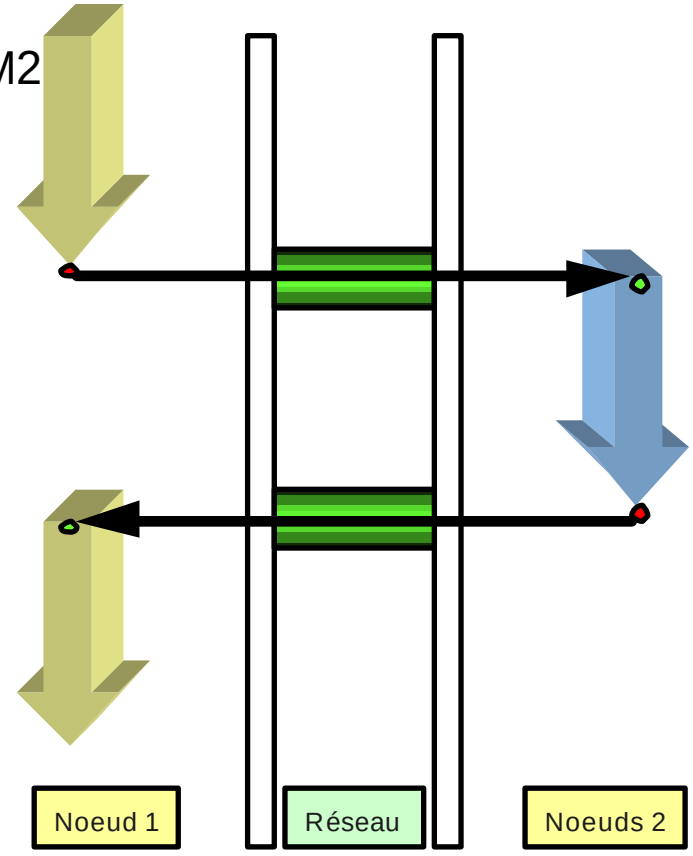


Caractéristiques

- Environnement multi-thread PM2
- Grappes hautes performances
- Communications de type RPC

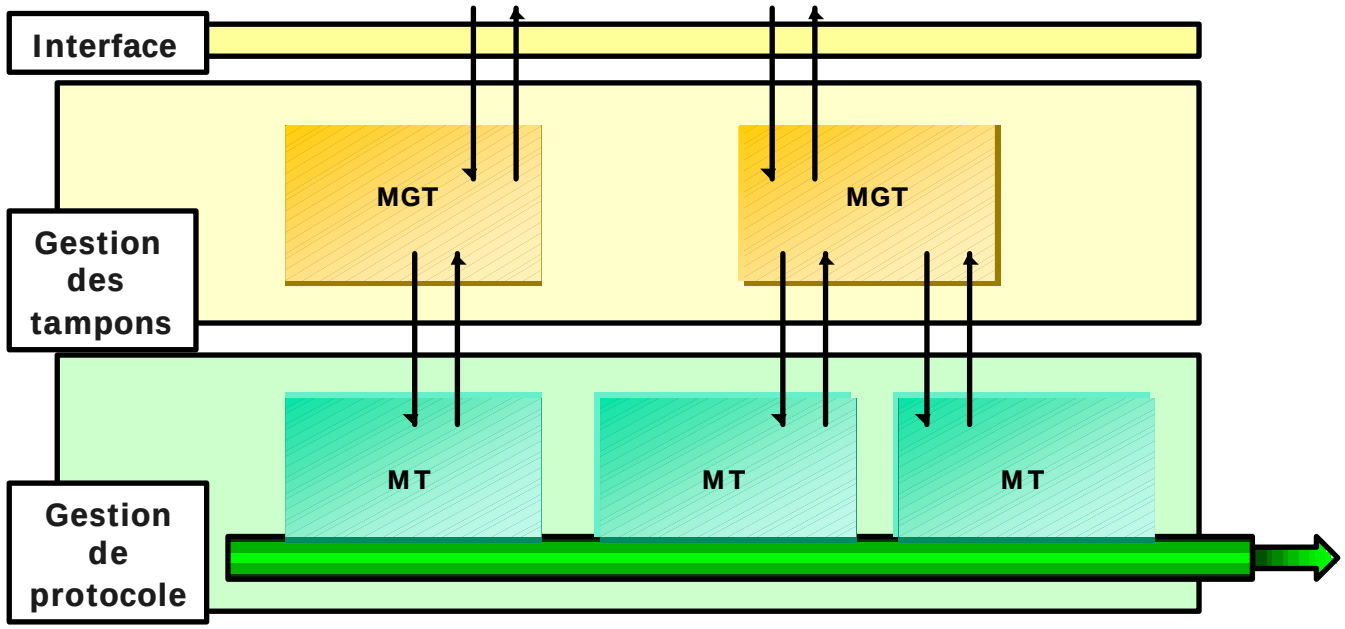
- Efficacité
- Réactivité

- Protocoles multi-paradigmes
- Support multi-protocole
- Support multi-réseau



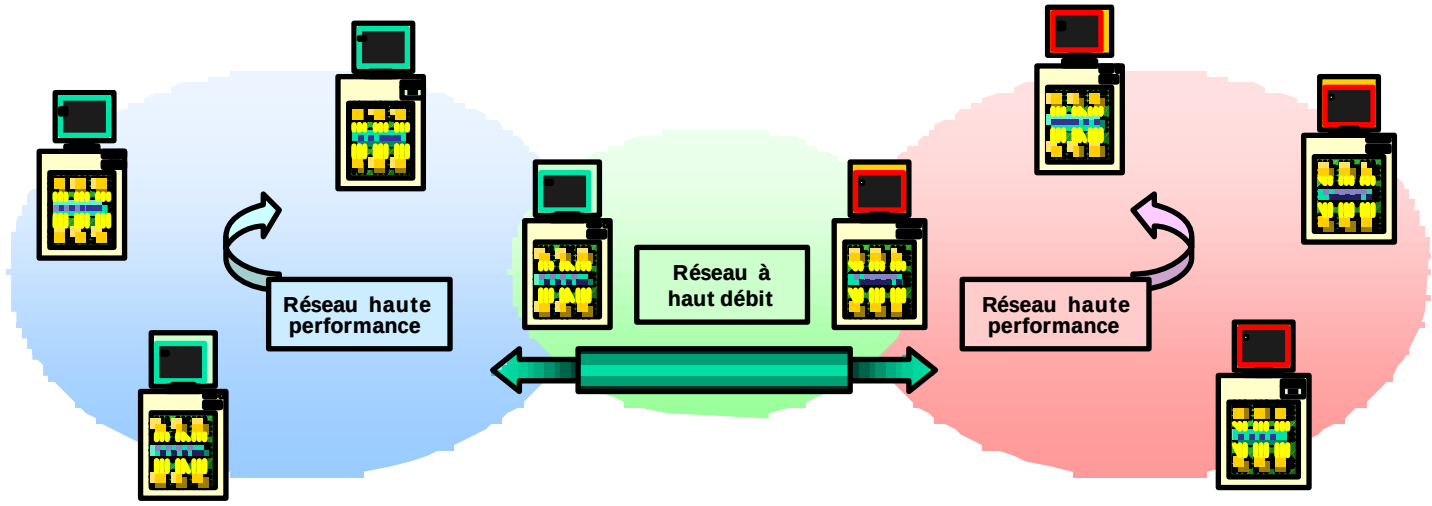
Architecture

- Approche modulaire
  - Module de gestion de tampon (MGT)
  - Module de transmission (MT)



Contexte

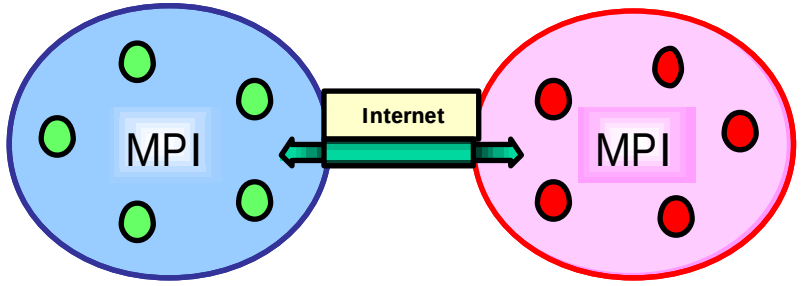
- Grappes de grappes
  - Réseaux intra-grappes rapides
  - Liens inter-grappes rapides
  - Hétérogénéité au niveau réseau

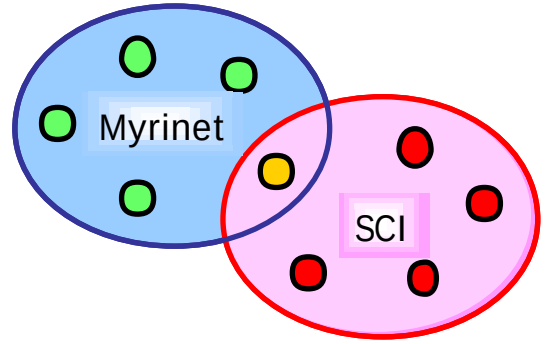


Travaux existants

- Solutions pour grappes hétérogènes
  - PACX-MPI
  - MPICH-G
  - MPIConnect

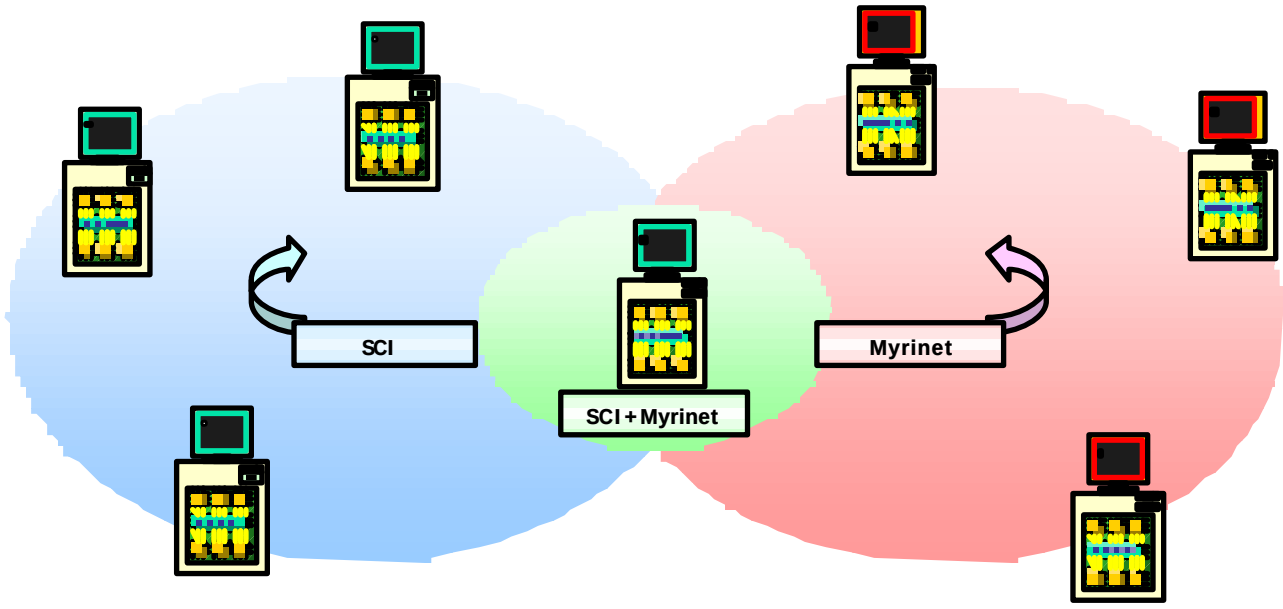
- Caractéristiques communes
  - Utilisation de MPI natifs (communication intra-grappe)
  - Support inter-cluster limité

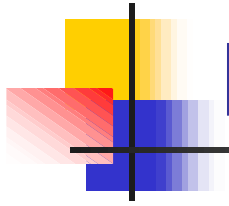




Objectifs

- Exploiter les liens inter-grappes rapides
  - Liens intra-grappes : Gb/s
  - Liens inter-grappes : Gb/s



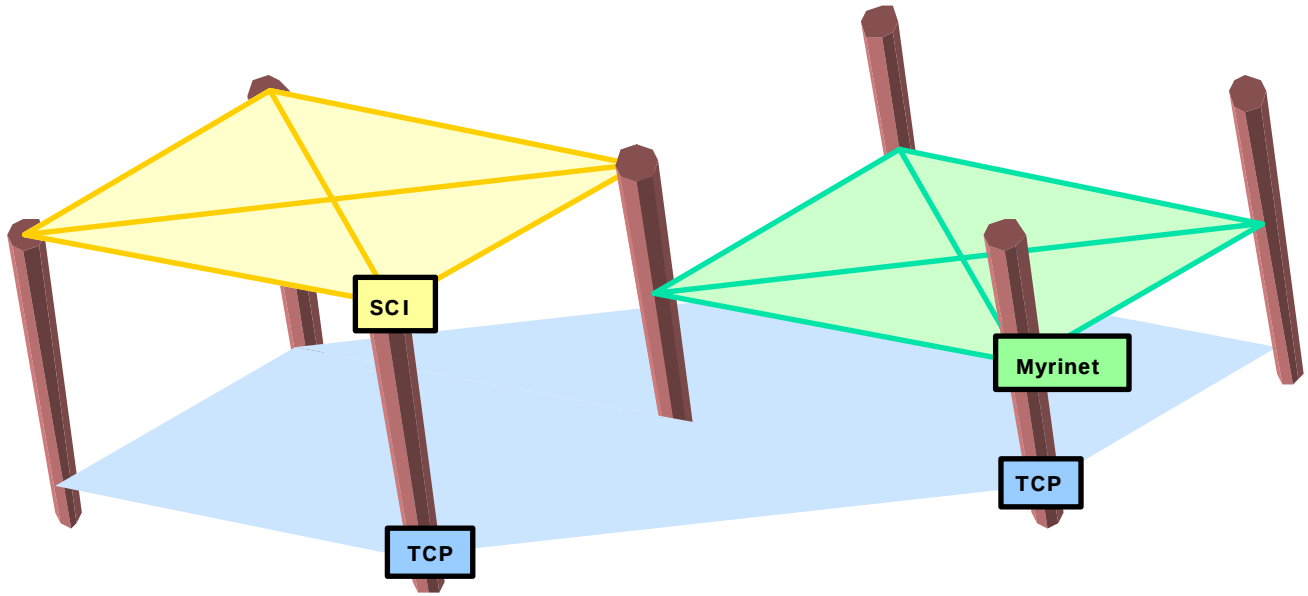


Proposition

- Extension de Madeleine
  - Utilisation du support multi-protocole
- Intégration d'un mécanisme de retransmission
  - Efficace
    - Bande-passante élevée
  - Transparent
    - Invisible pour le programmeur
  - Portable
    - Invisible pour les pilotes de Madeleine

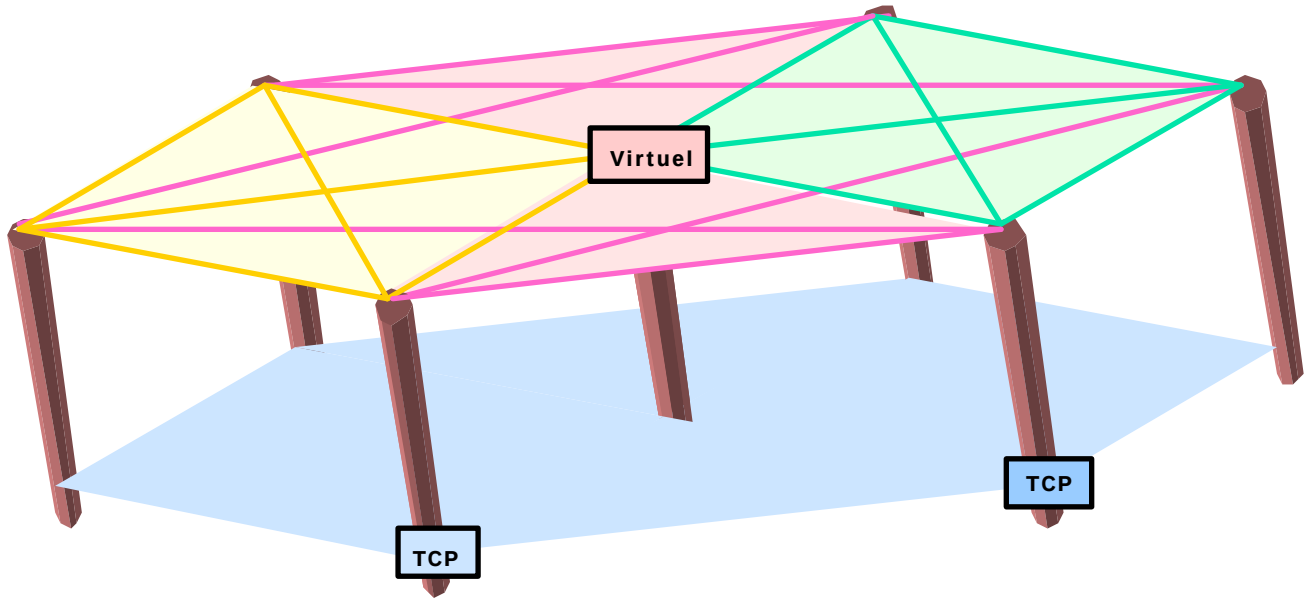
Canaux réels

- Liés à un réseau
- Ne couvrent pas nécessairement tous les noeuds

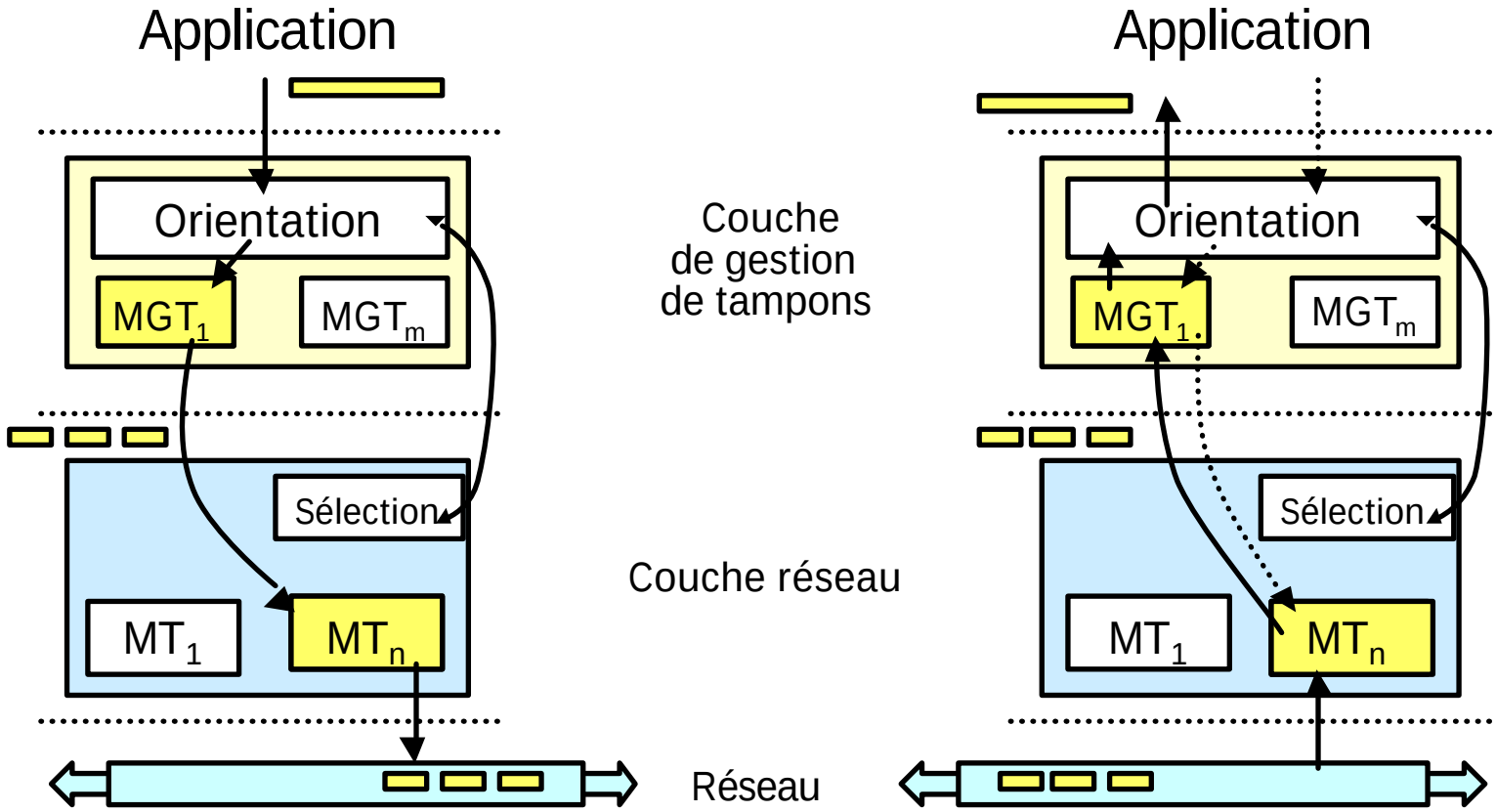


Canaux virtuels

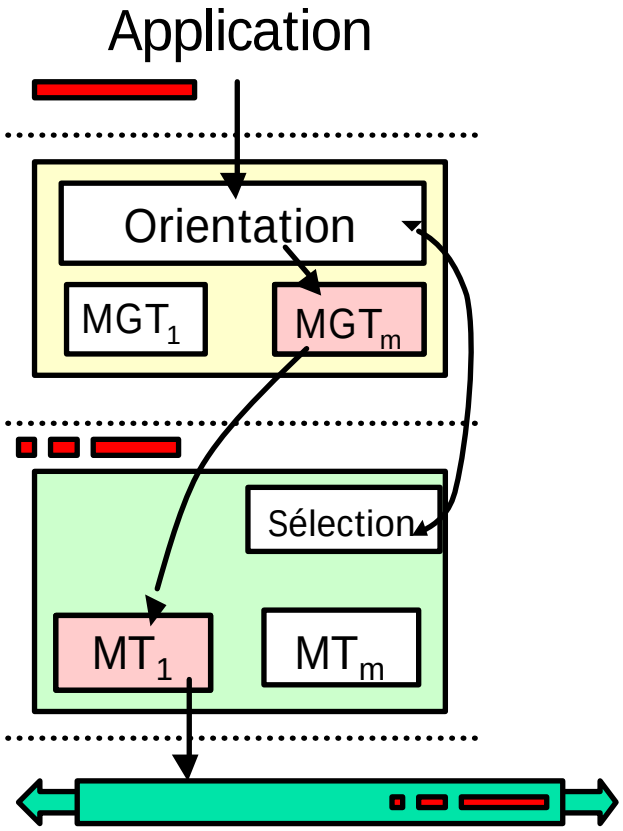
- Couvrent tous les noeuds
- Contiennent plusieurs canaux réels

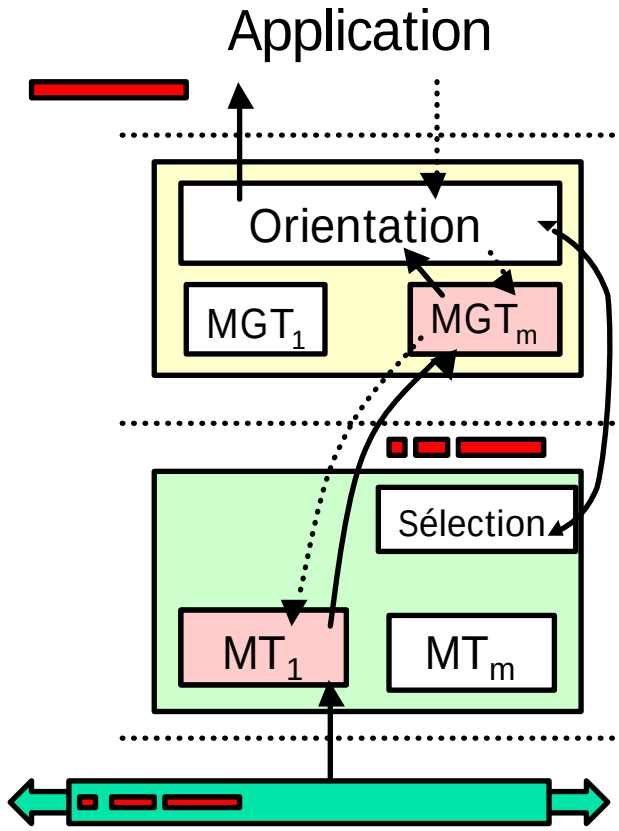


Structure

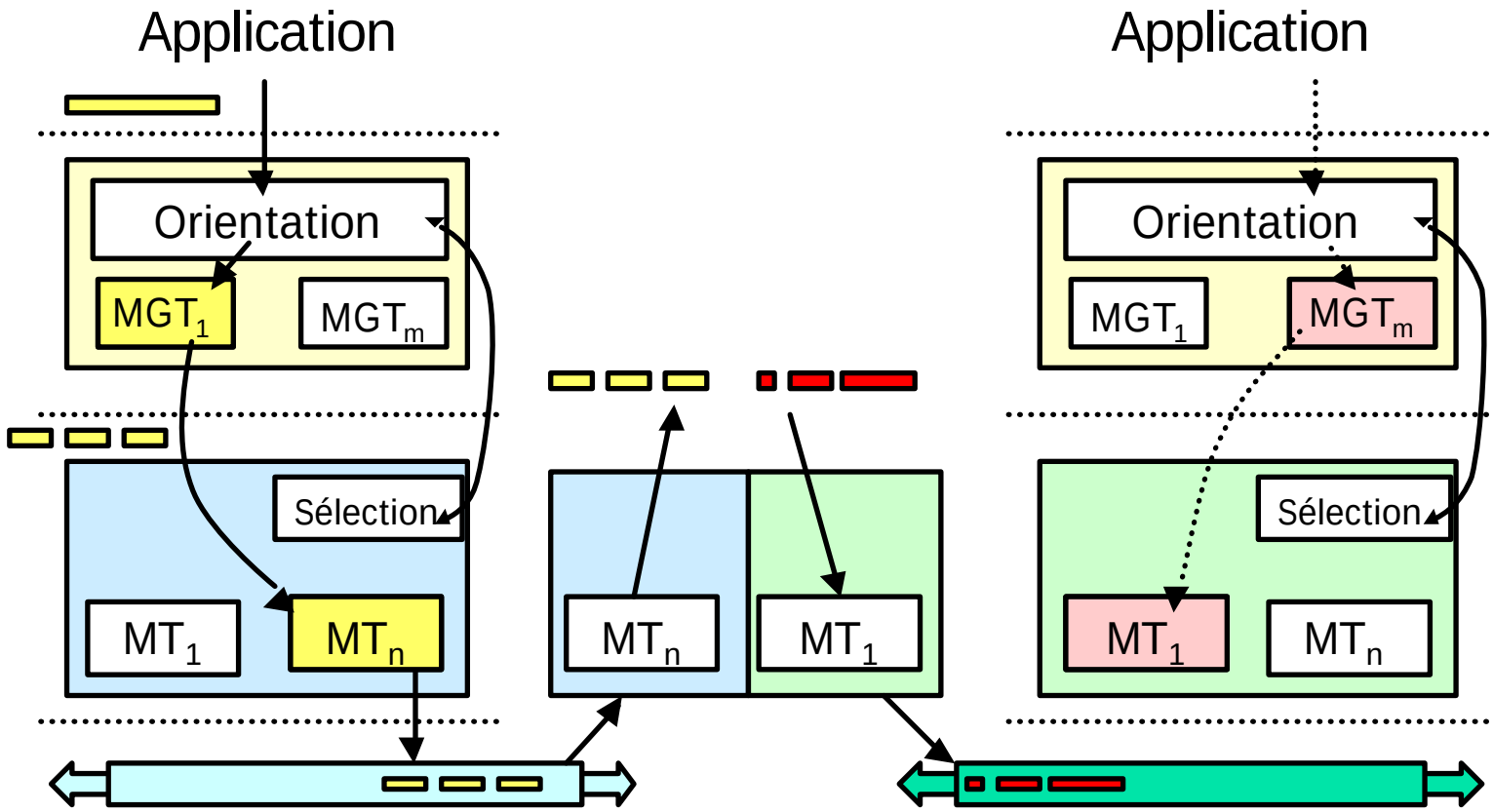

Application
Orientation
MGT1
MGTm
Sélection
MT1
MTn
Couche
de gestion
de tampons
Couche réseau
Réseau
Application
Orientation
MGT1
MGTm
Sélection
MT1
MTn

Un autre réseau



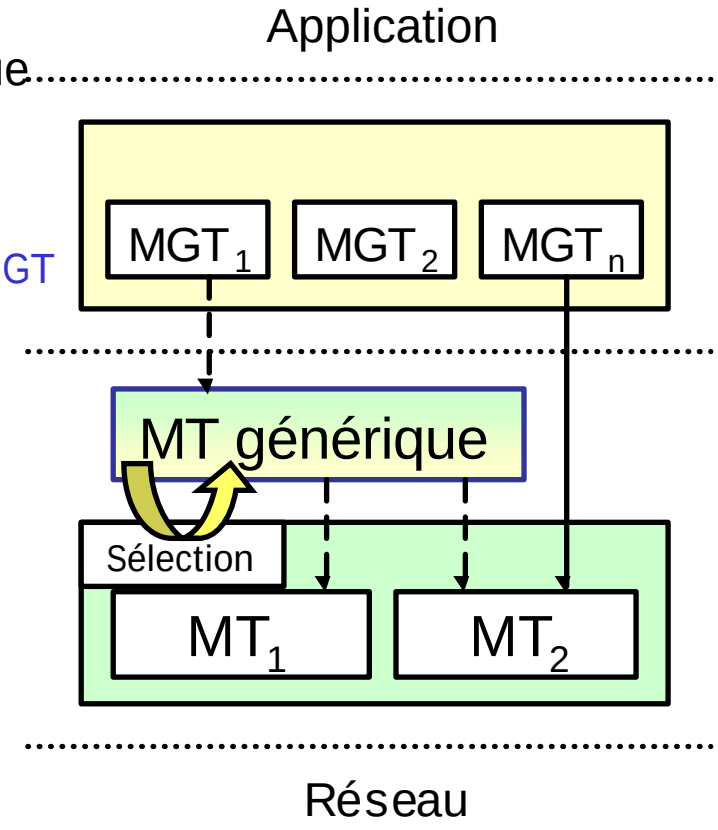


Deux réseaux différents



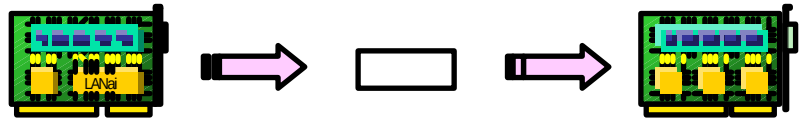
Solution

- Utilisation d'un MT générique
- Propriété
  - Sélection symétrique des MGT
- Caractéristiques
  - Contrôle sur la sélection des MT réseaux
  - Negociation d'une MTU

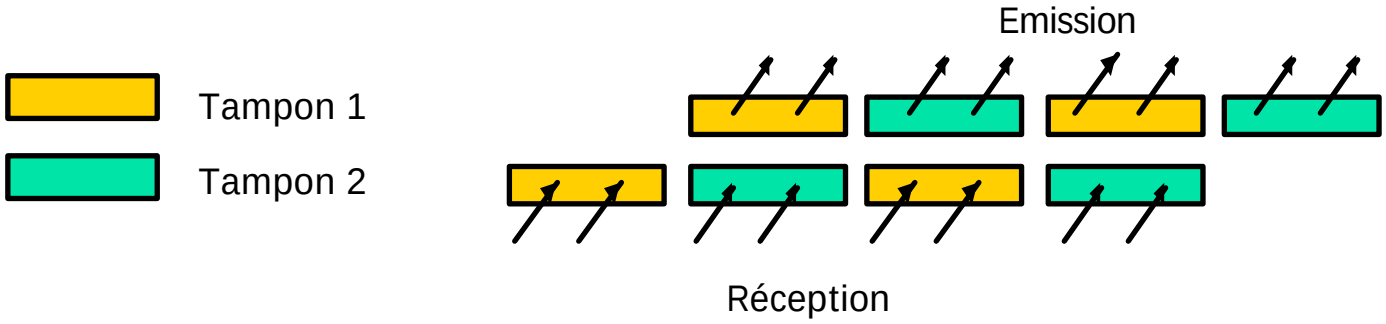


Préservation du débit

- Une copie
  - Même tampon pour la réception et la r é-émission

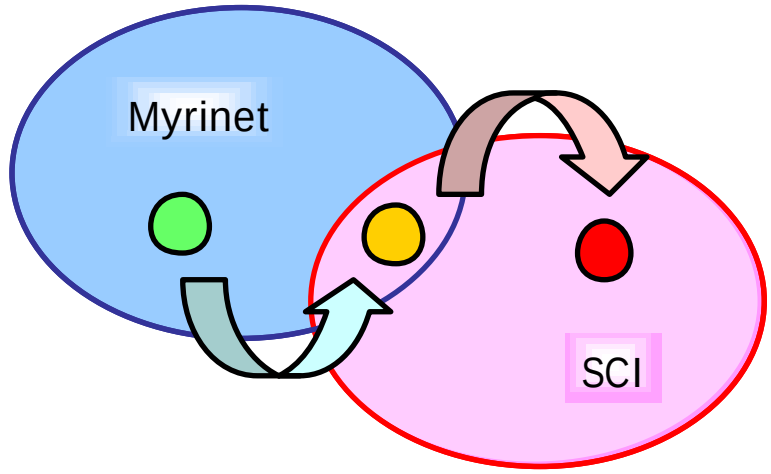


- Pipeline
  - Réception et ré-émission simultanée avec 2 tampons



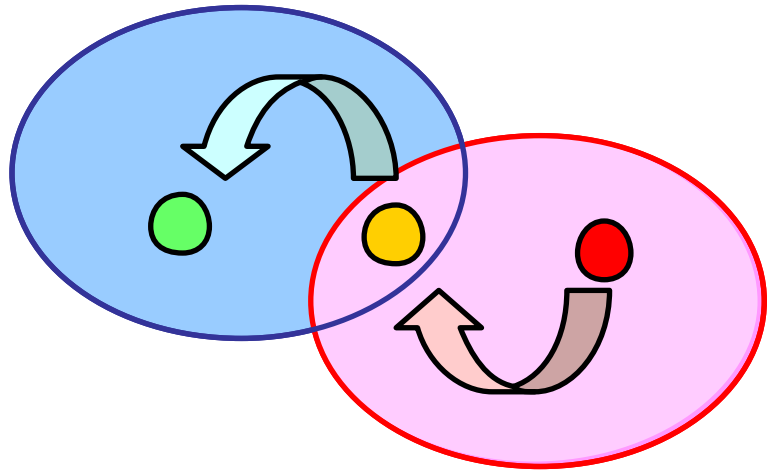
Evaluation bidirectionnelle

- Tests de ping-pong avec trois noeuds
  - Emetteur
  - Récepteur
  - Passerelle

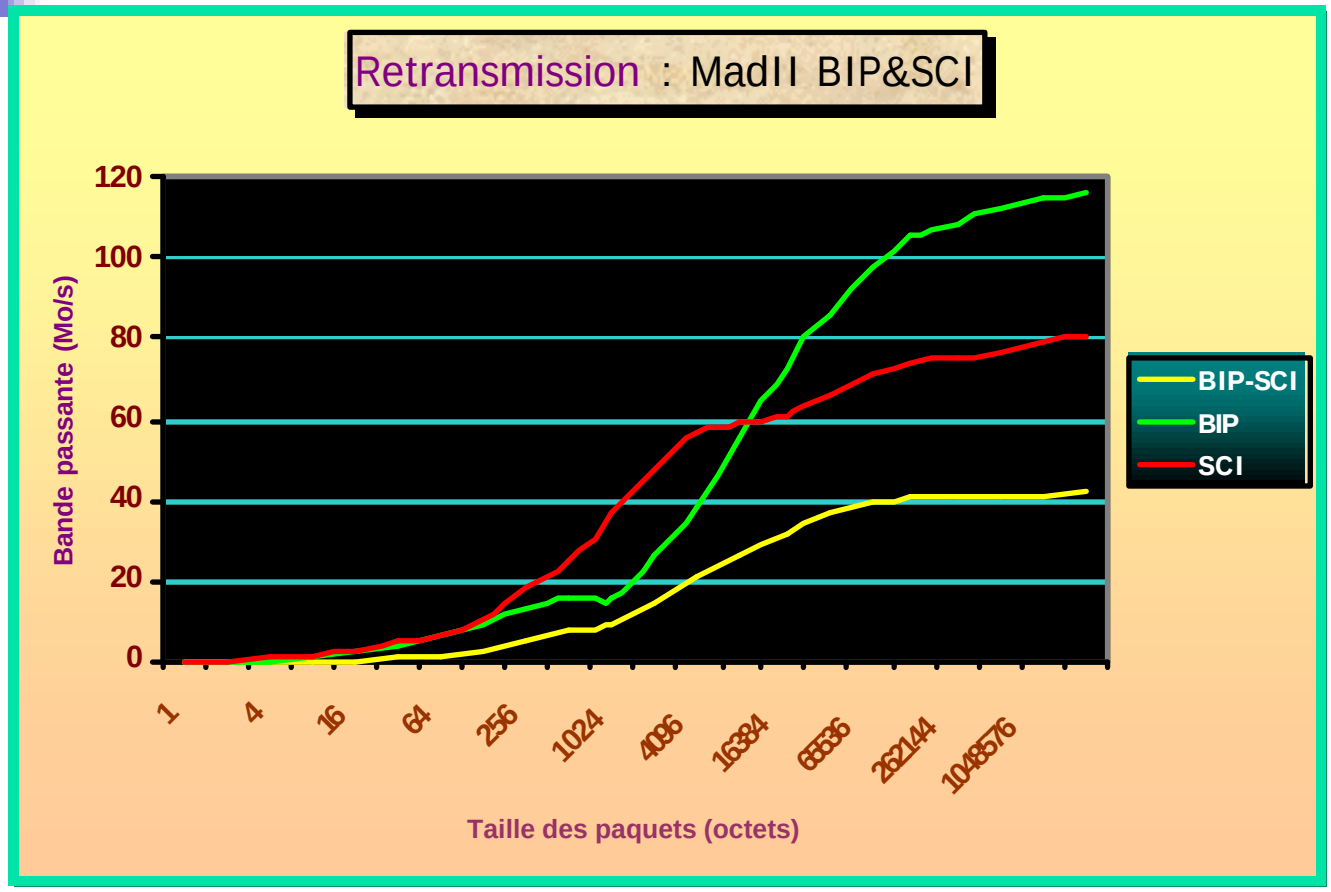


Evaluation bidirectionnelle

- Tests de ping-pong avec trois noeuds
  - Emetteur
  - Récepteur
  - Passerelle

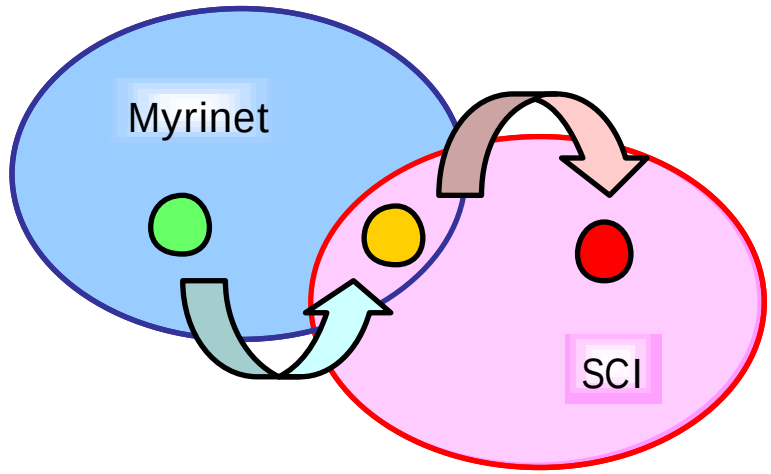


Bande passante



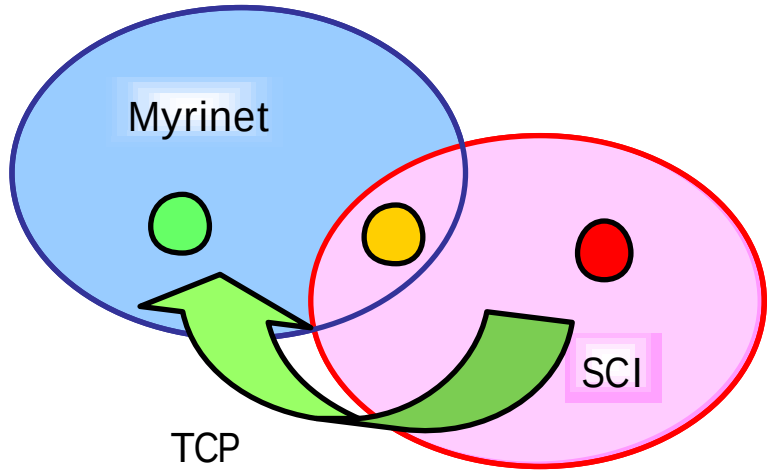
Evaluation unidirectionnelle

- Ping multi-réseau
  - Myrinet vers SCI ou SCI vers Myrinet

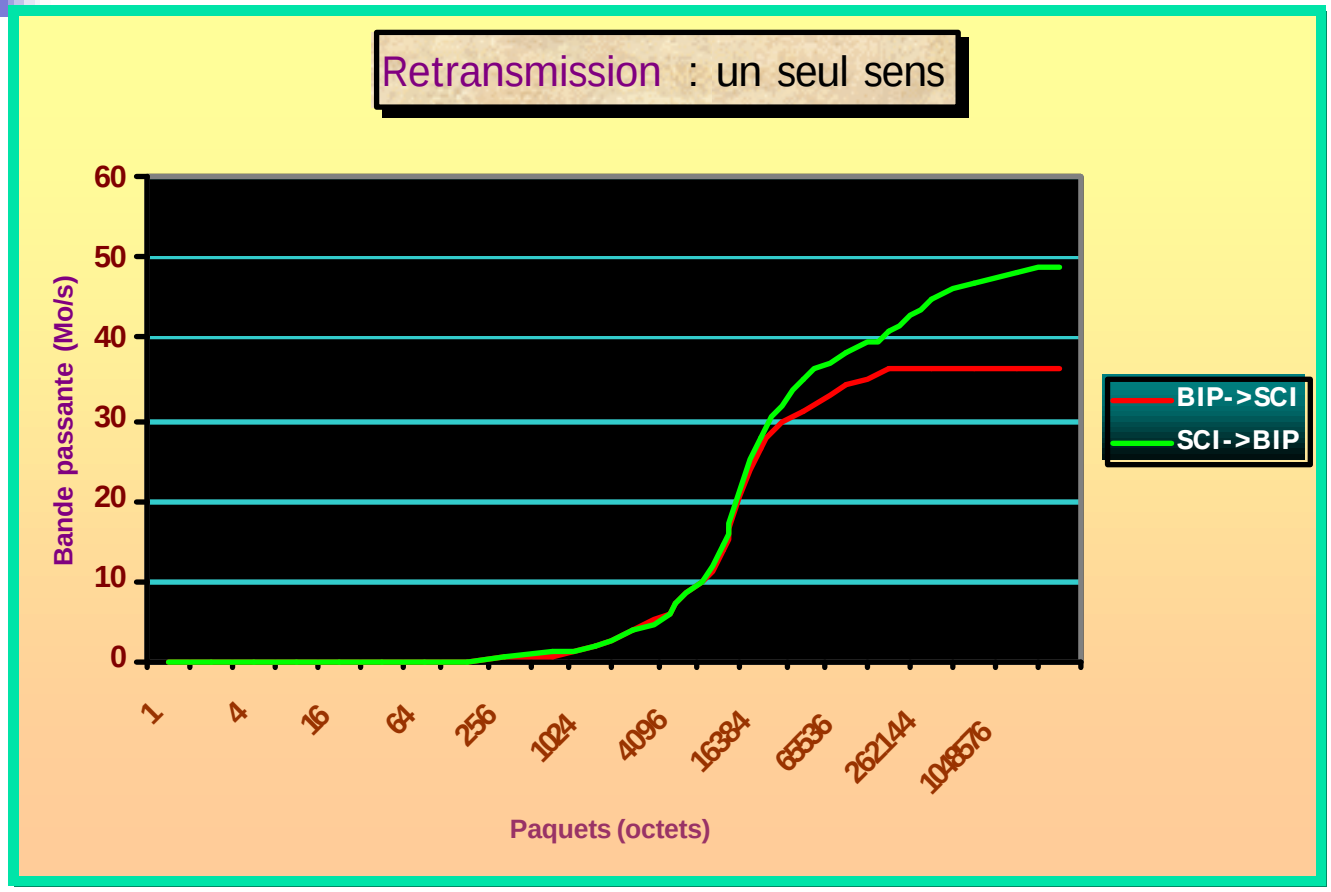


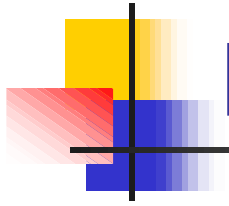
Evaluation unidirectionnelle

- Ping multi-réseau
  - Myrinet vers SCI ou SCI vers Myrinet
- Pong simple
  - TCP avec une latence connue



Résultats

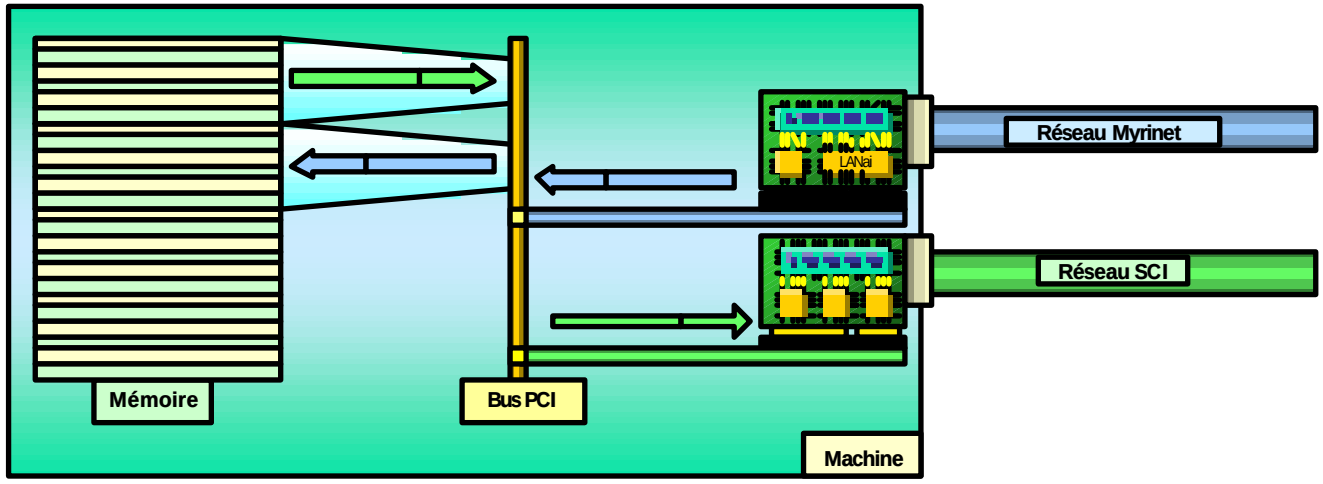

Retransmission : un seul sens
Bande passante (Mo/s)
60
50
40
30
20
10
0
1
4
16
64
256
1024
4096
16384
65536
262144
1048576
Paquets (octets)
BIP->SCI
SCI->BIP

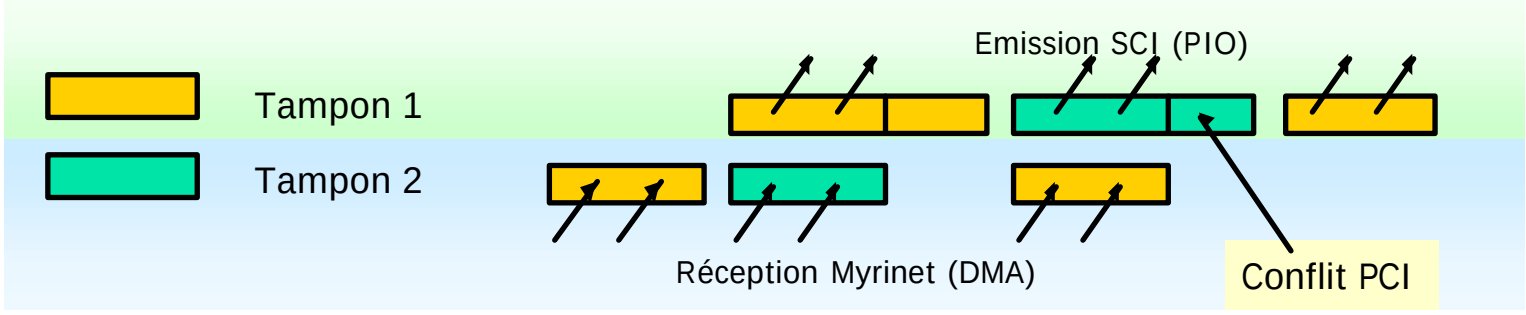


Résultats

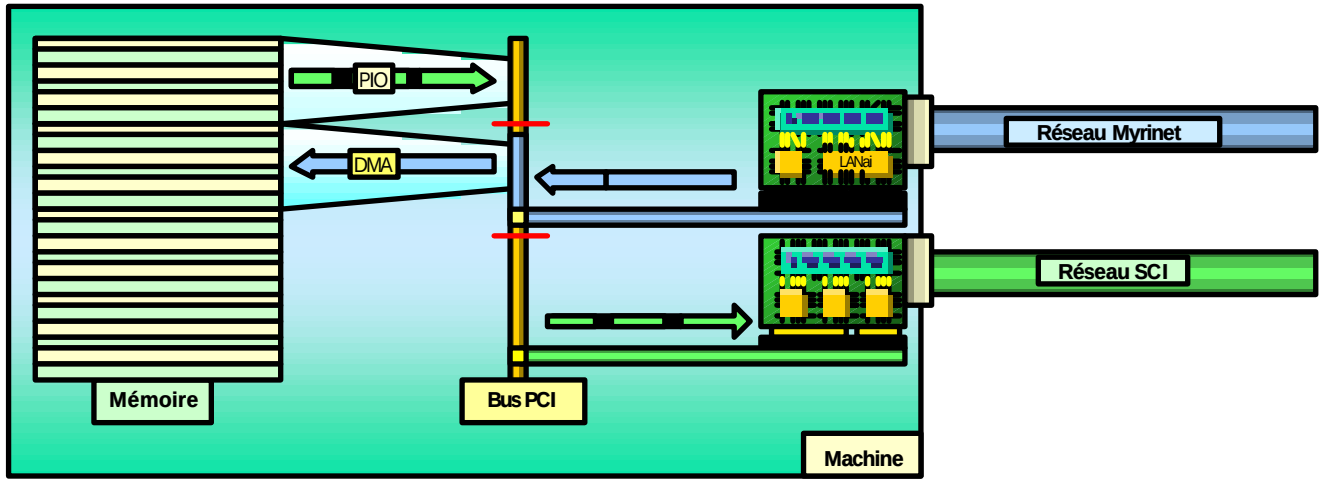
- SCI vers Myrinet
  - Bande passante : 48 Mo/s
  - Limite théorique : 66 Mo/s
    - Bande passante PCI : 132 Mo/s
- Myrinet vers SCI
  - Bande passante : 36 MB/s
  - Perturbation du pipeline
    - Conflits PCI

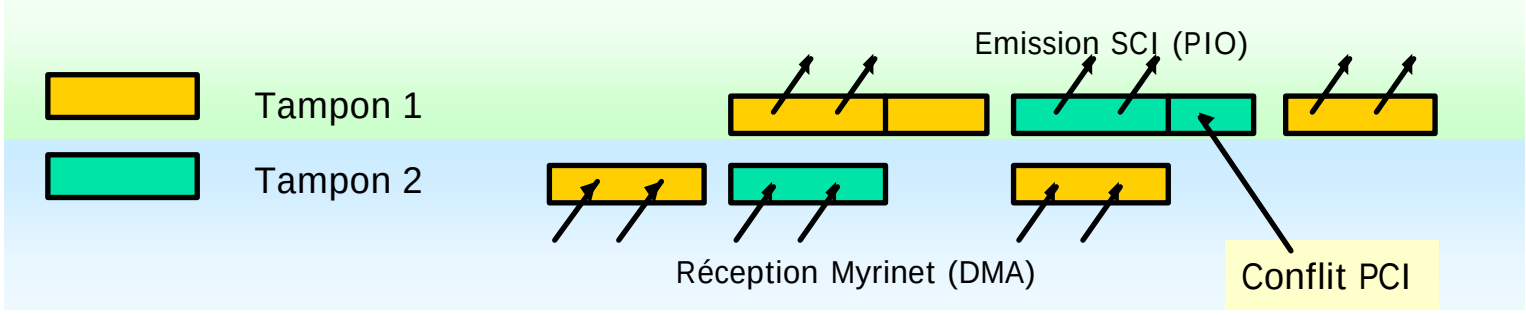
Conflit


Réseau Myrinet
LANai
Réseau SCI
Mémoire
Bus PCI
Machine


Tampon 1
Tampon 2
Emission SCI (PIO)
Réception Myrinet (DMA)
Conflit PCI

Conflit


PIO
DMA
LANai
Réseau Myrinet
Réseau SCI
Mémoire
Bus PCI
Machine

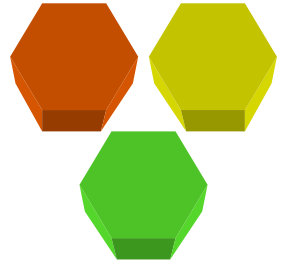

Tampon 1
Tampon 2
Emission SCI (PIO)
Réception Myrinet (DMA)
Conflit PCI

Travaux en cours

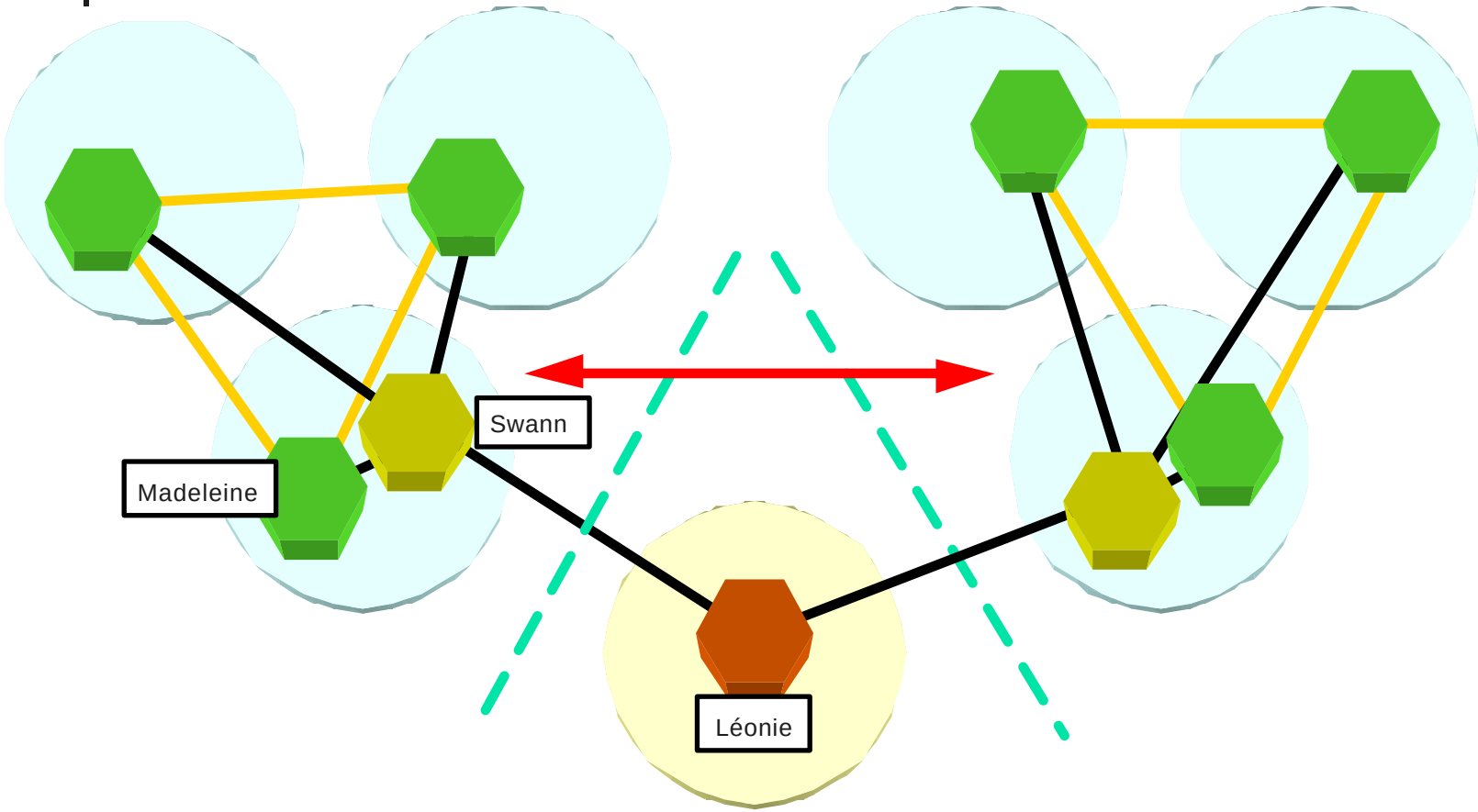
- Démarrage de session

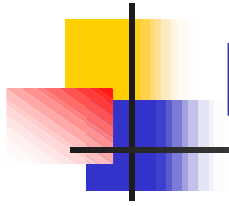
- Une approche modulaire
  - Flexibilité
  - Extensibilité

- Trois modules
  - Léonie
    - Contrôle de session
  - Swann
    - Relais distants
  - Madeleine



Infrastructure

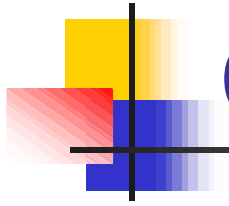




Léonie

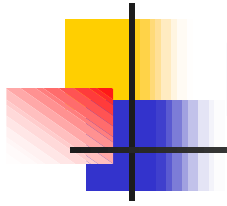
- Sessions
  - Configurations multi-grappes
  - Lancement unifié
    - Déploiement en rafale
    - Support pour lanceurs optimisés

- Réseau
  - Constructions des tables d'information
    - Répertoire des processus
    - Tables de routages des canaux virtuels
  - Ordonnancement
    - Initialisation des cartes, ouverture des canaux



Conclusion

- Mécanisme de retransmission rapide pour passerelle
  - Transparent
    - Canaux virtuels
    - Pas de noeud dédié à la passerelle
  - Portable
    - Module de Transmission générique
  - Efficace
    - SCI vers Myrinet : 48 Mo/s
    - Myrinet vers SCI : 36 Mo/s
- Sessions multi-grappes
  - En cours de finalisation



Perspectives

- Communications multi-réseaux
  - Tests sur cartes SCI récentes (meilleur DMA)
  - PC à double bus PCI
  - Contrôle de congestion spécifique ?
- Sessions
  - Finalisation du support multi-grappes
  - Intégration de lanceurs optimisés
  - Liaison dynamique des pilotes
- Bientôt disponible www.pm2.org