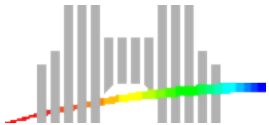
Réseau longue distance et application distribuée dans les grilles de calcul : étude et propositions pour une interaction efficace

Ludovic Hablot

17 décembre 2009

Thèse effectuée au Laboratoire de l'Informatique du Parallèlisme (LIP) de l'ENS Lyon, dirigée par Olivier Glück et Pascale Vicat-Blanc Primet.



Plan

Plan

Contexte

Les applications parallèles

- Besoins en puissance de calcul grandissant pour différents domaines, tels que physique, astronomie, biologie, prévisions météorologiques
- Division des calculs pour gagner en temps d'exécution

Le standard MPI

MPI (Message Passing Interface) est un standard pour programmer une application parallèle :

- il fonctionne par passage de messages
- il en existe plusieurs implémentations
- il s'appuie sur les protocoles de transport existants
- il propose des fonctions point à point et collectives

Contexte

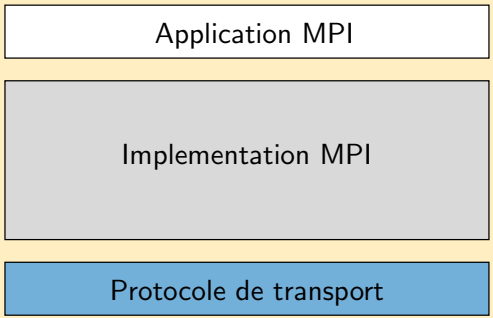
Les applications parallèles

- Besoins en puissance de calcul grandissant pour différents domaines, tels que physique, astronomie, biologie, prévisions météorologiques
- Division des calculs pour gagner en temps d'exécution

Le standard MPI

MPI (Message Passing Interface) est un standard pour programmer une application parallèle :

- il fonctionne par passage de messages
- il en existe plusieurs implémentations
- il s'appuie sur les protocoles de transport existants
- il propose des fonctions point à point et collectives



Contexte

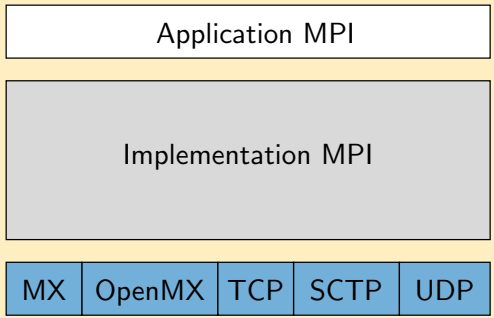
Les applications parallèles

- Besoins en puissance de calcul grandissant pour différents domaines, tels que physique, astronomie, biologie, prévisions météorologiques
- Division des calculs pour gagner en temps d'exécution

Le standard MPI

MPI (Message Passing Interface) est un standard pour programmer une application parallèle :

- il fonctionne par passage de messages
- il en existe plusieurs implémentations
- il s'appuie sur les protocoles de transport existants
- il propose des fonctions point à point et collectives



Contexte

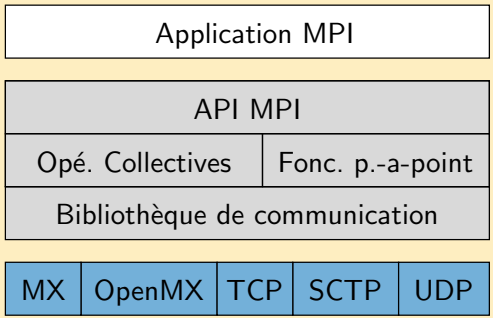
Les applications parallèles

- Besoins en puissance de calcul grandissant pour différents domaines, tels que physique, astronomie, biologie, prévisions météorologiques
- Division des calculs pour gagner en temps d'exécution

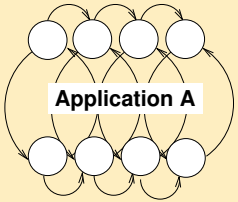
Le standard MPI

MPI (Message Passing Interface) est un standard pour programmer une application parallèle :

- il fonctionne par passage de messages
- il en existe plusieurs implémentations
- il s'appuie sur les protocoles de transport existants
- il propose des fonctions point à point et collectives

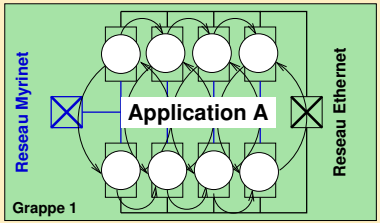


Les grilles



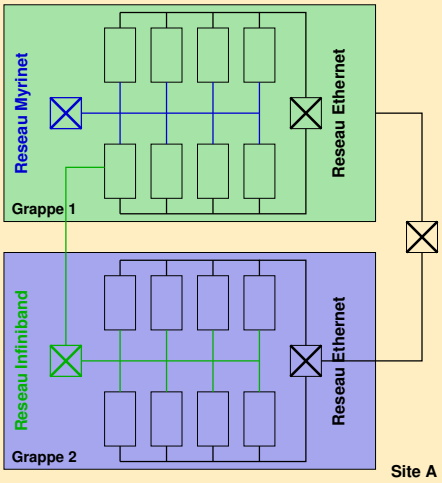
Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Les grilles



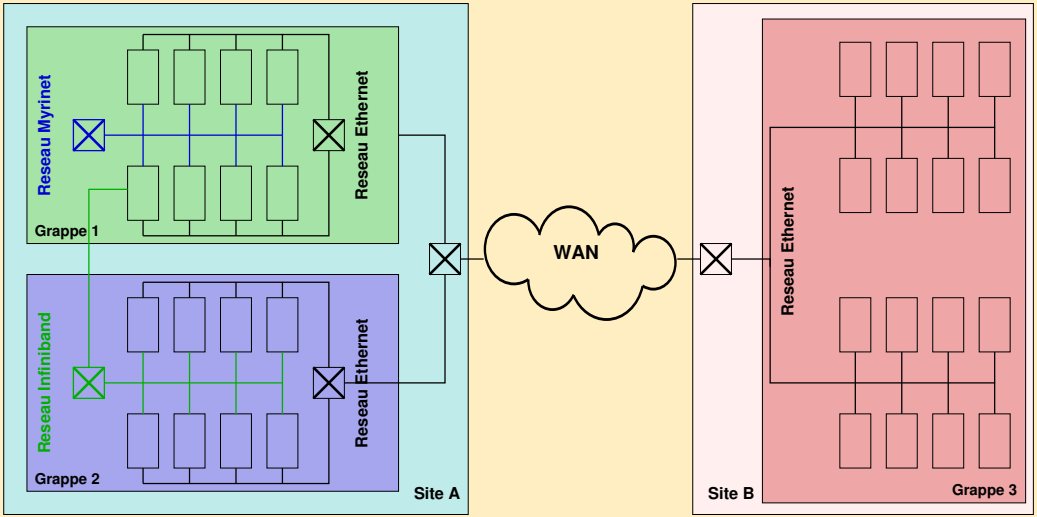
Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Les grilles



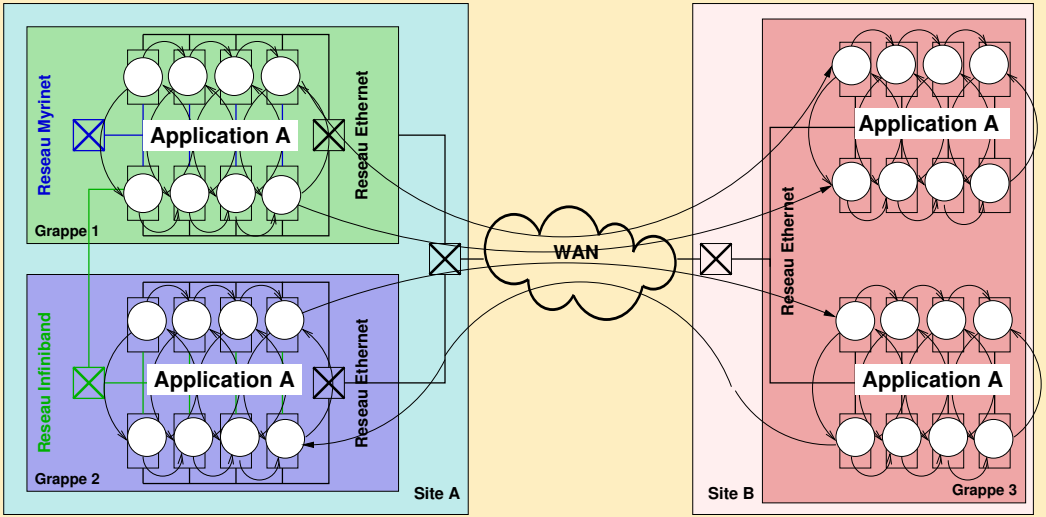
Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Les grilles



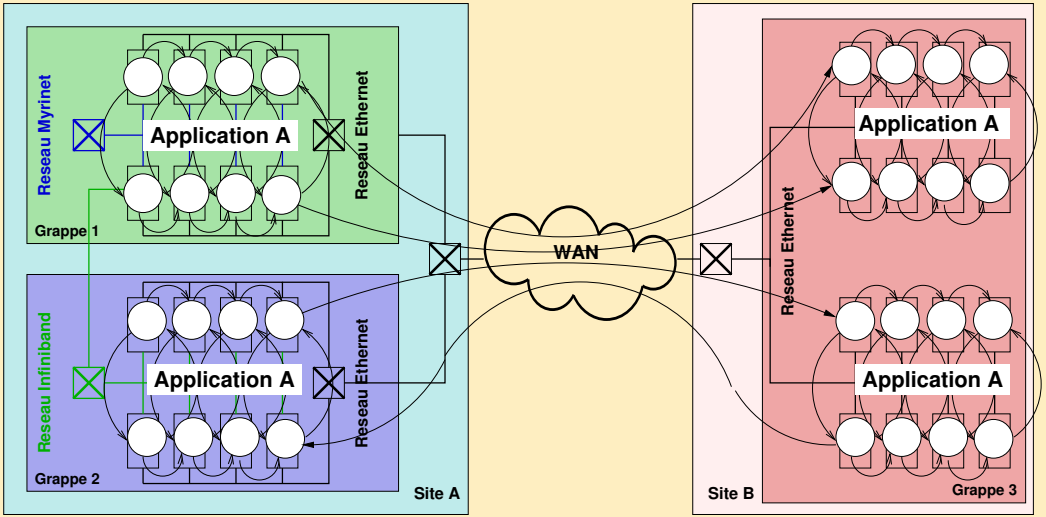
Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Les grilles



Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Les grilles



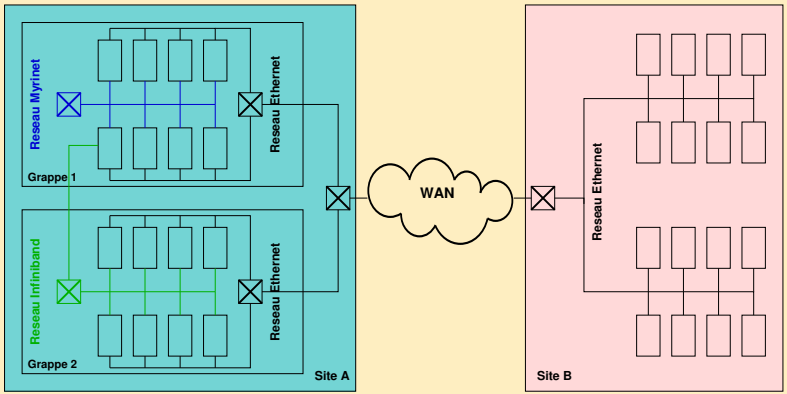
Notre définition : Les grilles sont une aggrégation de grappes ou de grappes de grappes, géographiquement éloignées et interconnectées par un réseau longue distance. Ce dernier est constitué d'un WAN (Wide Area Network) par opposition au LAN (Local Area Network).

Plan

Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

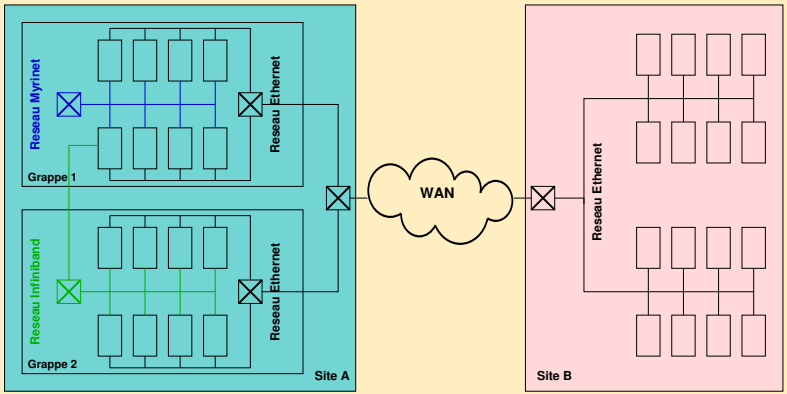
- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

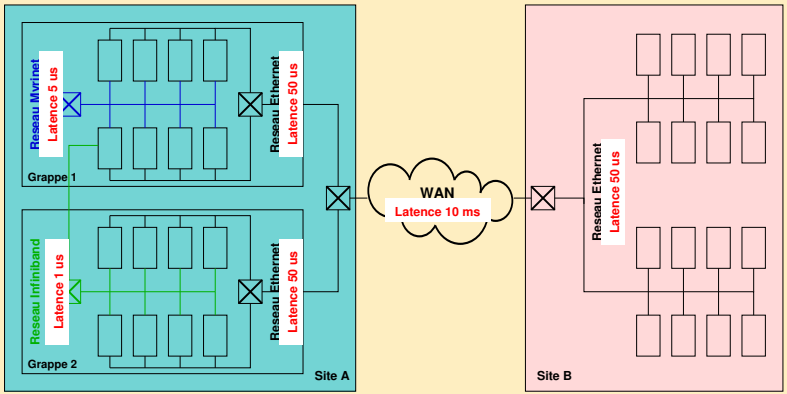
- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

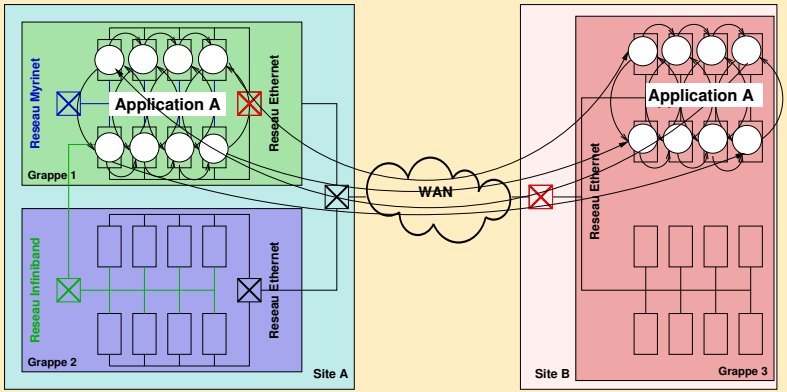
- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

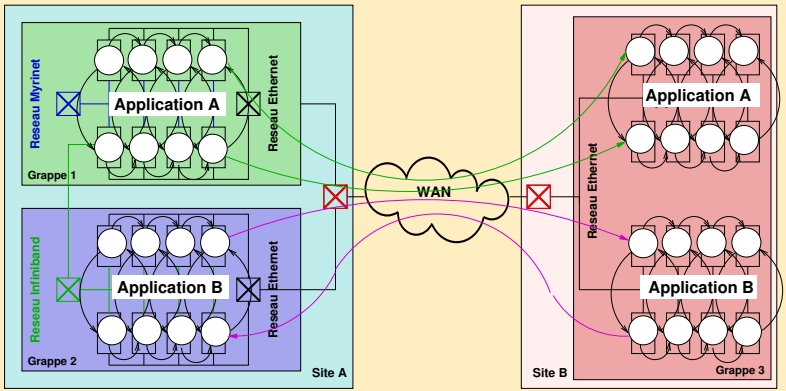
- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

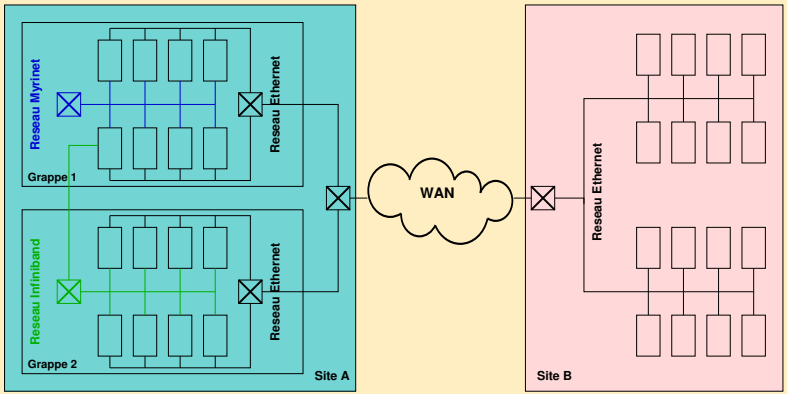
- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Spécificités de la grille

La grille soulève de nouveaux problèmes liés à ses spécificités :

- hétérogénéité des machines : problème résolu par un placement approprié des processus MPI
- hétérogénéité des réseaux rapides des clusters : problème résolu en utilisant une implémentation capable communiquer sur des réseaux différents
- latence plus grande sur le WAN que sur le LAN
- goulot d'étranglement du WAN : bande passante du WAN inférieure à la la somme des noeuds qui peuvent communiquer dessus
- partage des ressources, notamment des ressources réseau



Problématiques

Contraintes

- Transparence vis à vis de l'utilisateur : on garde intacte l'application MPI
- Transparence par rapport à l'implémentation MPI
- TCP est le protocole utilisé pour les communications sur le réseau longue distance des grilles

Comment exécuter au mieux des applications MPI sur une grille de calcul dont le protocole de transport sur le réseau longue distance est TCP, en optimisant l'interaction entre ces deux couches ?

Sous-questions

- Comment se comportent les applications MPI sur un réseau longue distance ?
  - Etude des caractéristiques des communications : taille, fréquence, synchronisme
  - Etude des points problèmatiques de la grille mentionnés précédemment
- Quels paramètres de TCP limitent les communications des applications MPI dans un réseau longue distance ?
  - Etude du contrôle de congestion et du contrôle de fiabilité
- Comment réduire l'impact de TCP sur les communications MPI longue distance ?
  - Différentiation des communications (locales ou longue-distance)
  - Adaptation des communications pour rendre le protocole de transport plus réactif

Problématiques

Contraintes

- Transparence vis à vis de l'utilisateur : on garde intacte l'application MPI
- Transparence par rapport à l'implémentation MPI
- TCP est le protocole utilisé pour les communications sur le réseau longue distance des grilles

Comment exécuter au mieux des applications MPI sur une grille de calcul dont le protocole de transport sur le réseau longue distance est TCP, en optimisant l'interaction entre ces deux couches ?

Sous-questions

- Comment se comportent les applications MPI sur un réseau longue distance ?
  - Etude des caractéristiques des communications : taille, fréquence, synchronisme
  - Etude des points problèmatiques de la grille mentionnés précédemment
- Quels paramètres de TCP limitent les communications des applications MPI dans un réseau longue distance ?
  - Etude du contrôle de congestion et du contrôle de fiabilité
- Comment réduire l'impact de TCP sur les communications MPI longue distance ?
  - Différentiation des communications (locales ou longue-distance)
  - Adaptation des communications pour rendre le protocole de transport plus réactif

Problématiques

Contraintes

- Transparence vis à vis de l'utilisateur : on garde intacte l'application MPI
- Transparence par rapport à l'implémentation MPI
- TCP est le protocole utilisé pour les communications sur le réseau longue distance des grilles

Comment exécuter au mieux des applications MPI sur une grille de calcul dont le protocole de transport sur le réseau longue distance est TCP, en optimisant l'interaction entre ces deux couches ?

Sous-questions

- Comment se comportent les applications MPI sur un réseau longue distance ?
  - Etude des caractéristiques des communications : taille, fréquence, synchronisme
  - Etude des points problèmatiques de la grille mentionnés précédemment
- Quels paramètres de TCP limitent les communications des applications MPI dans un réseau longue distance ?
  - Etude du contrôle de congestion et du contrôle de fiabilité
- Comment réduire l'impact de TCP sur les communications MPI longue distance ?
  - Différentiation des communications (locales ou longue-distance)
  - Adaptation des communications pour rendre le protocole de transport plus réactif

Etat de l'art : implémentations existantes

	Gestion de l'hétérogénéité	Optimisation des comm. longue distance	
		Opérations coll.	Optimisation TCP
PACX-MPI	X	X	
MagPIe		X	
MPICH-GQ			Limitation de débit
MPICH2			
MPICH-VMI	X	X	
MetaMPICH	X		
MPICH-G2	X	X	Flux parallèles pour les gros messages sur le WAN
MPICH-Madeleine	X		
GridMPI	X	X	Diminution du RTO, limitation de débit, pacing au démarrage, chgt. fenêtre cong.,
OpenMPI	X	?	

Etat de l'art : implémentations existantes

	Gestion de l'hétérogénéité	Optimisation des comm. longue distance	
		Opérations coll.	Optimisation TCP
PACX-MPI	X	X	
MagPIe		X	
MPICH-GQ			Limitation de débit
MPICH2			
MPICH-VMI	X	X	
MetaMPICH	X		
MPICH-G2	X	X	Flux parallèles pour les gros messages sur le WAN
MPICH-Madeleine	X		
GridMPI	X	X	Diminution du RTO, limitation de débit, pacing au démarrage, chgt. fenêtre cong.,
OpenMPI	X	?	

Etat de l'art : implémentations existantes

	Gestion de l'hétérogénéité	Optimisation des comm. longue distance	
		Opérations coll.	Optimisation TCP
PACX-MPI	X	X	
MagPIe		X	
MPICH-GQ			Limitation de débit
MPICH2			
MPICH-VMI	X	X	
MetaMPICH	X		
MPICH-G2	X	X	Flux parallèles pour les gros messages sur le WAN
MPICH-Madeleine	X		
GridMPI	X	X	Diminution du RTO, limitation de débit, pacing au démarrage, chgt. fenêtre cong.,
OpenMPI	X	?	

Etat de l'art : implémentations existantes

	Gestion de l'hétérogénéité	Optimisation des comm. longue distance	
		Opérations coll.	Optimisation TCP
PACX-MPI	X	X	
MagPIe		X	
MPICH-GQ			Limitation de débit
MPICH2			
MPICH-VMI	X	X	
MetaMPICH	X		
MPICH-G2	X	X	Flux parallèles pour les gros messages sur le WAN
MPICH-Madeleine	X		
GridMPI	X	X	Diminution du RTO, limitation de débit pacing au démarrage, chgt. fenêtre cong.
OpenMPI	X	?	

Plan

Instrumentation des applications MPI et de TCP

Pourquoi instrumenter ?

- Analyse des pertes de performances des applications lors du passage sur la grille
- Deux couches accessibles TCP et MPI : analyse des communications longue distance au niveau de ces deux couches

- Nombre, taille et fréquence des communications
- Schéma de communication

→

InstrAppli

- Date système des événements
- Surcharge des fonctions de l'API socket
- Source et destination des données
- Fonction appelée et paramètres de celle-ci

- Evolution de la fenêtre de congestion de TCP
- Instant des retransmissions

→

tcp_probe modifié

- Date système des événements
- Espace libre dans les tampons d'émission de TCP

Instrumentation des applications MPI et de TCP

Pourquoi instrumenter ?

- Analyse des pertes de performances des applications lors du passage sur la grille
- Deux couches accessibles TCP et MPI : analyse des communications longue distance au niveau de ces deux couches

- Nombre, taille et fréquence des communications
- Schéma de communication

→

InstrAppli

- Date système des événements
- Surcharge des fonctions de l'API socket
- Source et destination des données
- Fonction appelée et paramètres de celle-ci

- Evolution de la fenêtre de congestion de TCP
- Instant des retransmissions

→

tcp_probe modifié

- Date système des événements
- Espace libre dans les tampons d'émission de TCP

Instrumentation des applications MPI et de TCP

Pourquoi instrumenter ?

- Analyse des pertes de performances des applications lors du passage sur la grille
- Deux couches accessibles TCP et MPI : analyse des communications longue distance au niveau de ces deux couches

- Nombre, taille et fréquence des communications
- Schéma de communication

→

InstrAppli

- Date système des événements
- Surcharge des fonctions de l'API socket
- Source et destination des données
- Fonction appelée et paramètres de celle-ci

- Evolution de la fenêtre de congestion de TCP
- Instant des retransmissions

→

tcp_probe modifié

- Date système des événements
- Espace libre dans les tampons d'émission de TCP

Instrumentation des applications MPI et de TCP

Pourquoi instrumenter ?

- Analyse des pertes de performances des applications lors du passage sur la grille
- Deux couches accessibles TCP et MPI : analyse des communications longue distance au niveau de ces deux couches

- Nombre, taille et fréquence des communications
- Schéma de communication

→

InstrAppli

- Date système des événements
- Surcharge des fonctions de l'API socket
- Source et destination des données
- Fonction appelée et paramètres de celle-ci

- Evolution de la fenêtre de congestion de TCP
- Instant des retransmissions

→

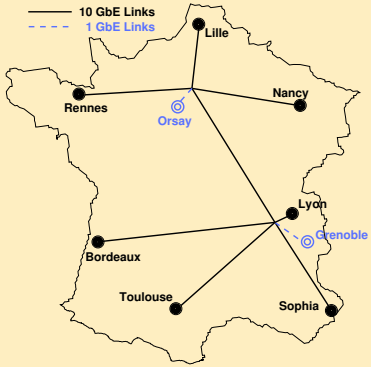
tcp_probe modifié

- Date système des événements
- Espace libre dans les tampons d'émission de TCP

Plateforme de tests : Grid'5000

Grid5000

- Grille de recherche française qui regroupe 9 sites,
- interconnectés à 1 ou 10 Gb/s.

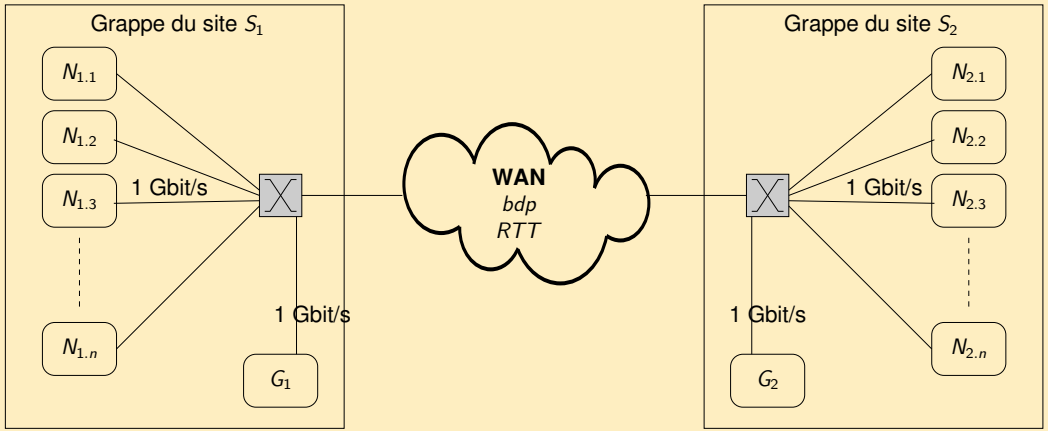


Expérience type

- Réservation de noeuds
- Déploiement d'un environnement
- Paramétrage des noeuds
- Lancement de l'expérience
- Récupération et regroupement des résultats

Banc d'essai

- 2 sites connectés au WAN à 1 Gb/s
- 1, 2 ou 8 noeuds par site selon les expériences



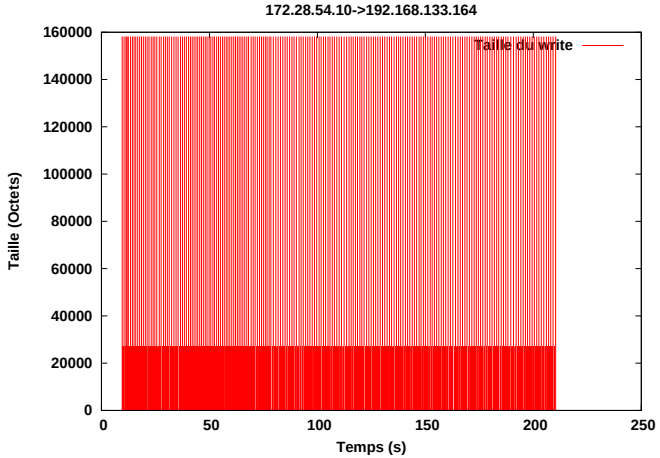
Analyse des NAS Parallel Benchmark (NPB)

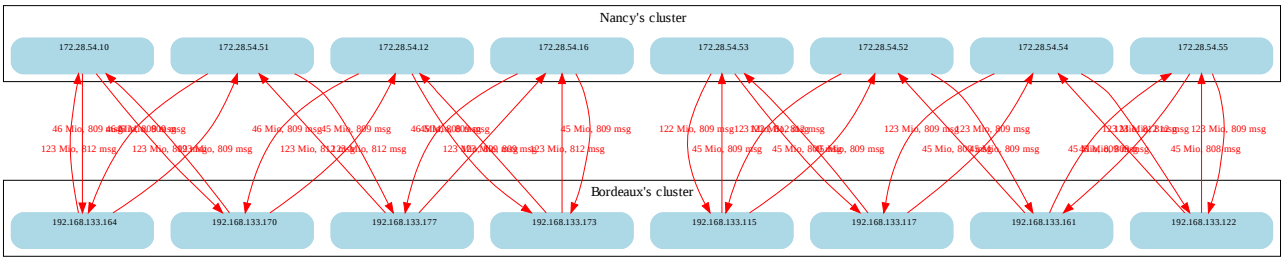
Les NAS Parallel Benchmark [Bailey et al. 1994]

- Les NPB sont des applications représentatives des applications MPI :
  - BT (Block Tridiagonal)
  - CG (Conjugate Gradient)
  - FT (Fast Fourier Transform)
  - IS (Integer sort)
  - LU (Lower-Upper symmetric Gauss-Seidel)
  - MG (MultiGrid)
  - SP (Scalar Pentadiagonal)
- différentes tailles de problème

Analyse de BT à l'aide de InstrAppli

Grâce à InstrAppli et tcp_probe, analyse des communications longue distance des NPB.





Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

Classification des NPB

Classification des NPB

		Qualité		
		Faible	**Moyenne**	**Grande**
Métrique	**Fréquence des comm.**	**x > 1s** FT, IS	**0.1s < x < 1s** BT, SP	**x < 0.1s** CG, MG, LU
	Taille des comm.	**x < 1ko** LU	**1ko < x < 200ko** BT, SP, CG, MG	**200ko < x** FT, IS
	Synchronisme des comm.	BT, SP, LU	CG, MG	FT, IS

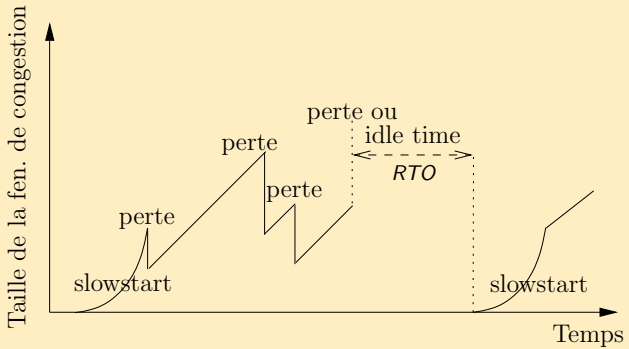
Plan

TCP

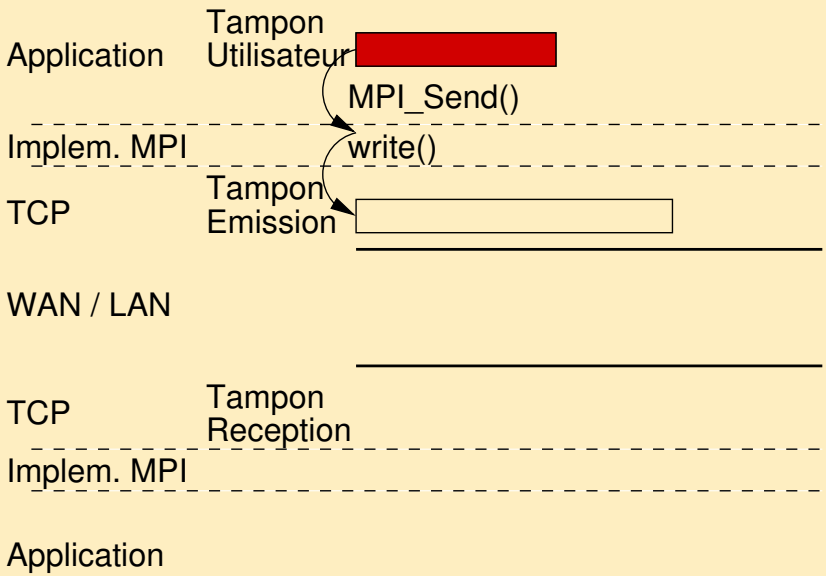
- TCP (Transport Control Protocol) a été créé pour proposer un transfert *fiable* et *ordonné* de données aux applications Internet.
- 3 mécanismes principaux :
  - Contrôle de fiabilité : retransmission des données en cas de perte ou d'erreur
  - Contrôle de flux : prévention de la perte de données si un récepteur est trop lent
  - Contrôle de congestion : partage équitable de la bande passante et utilisation maximale des liens.

On distingue deux phases :

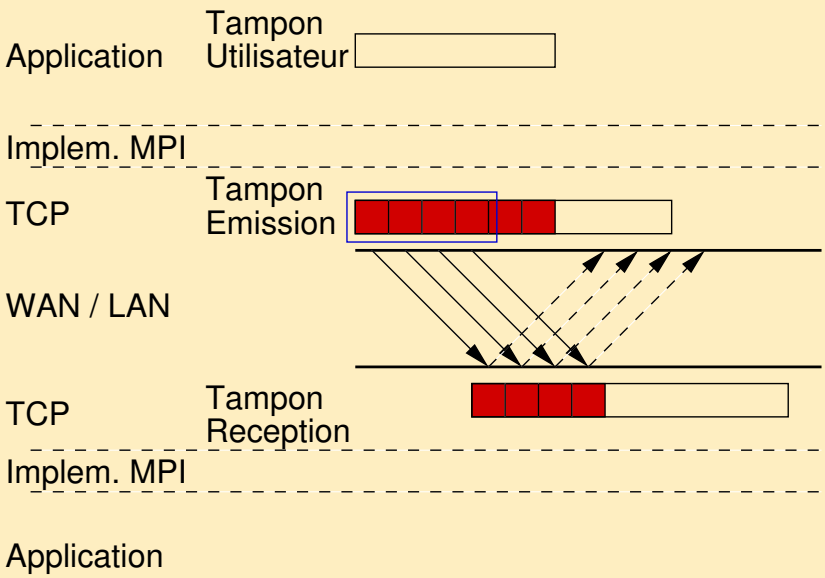
- le démarrage lent (slowtstart)
- la phase d'évitement de congestion



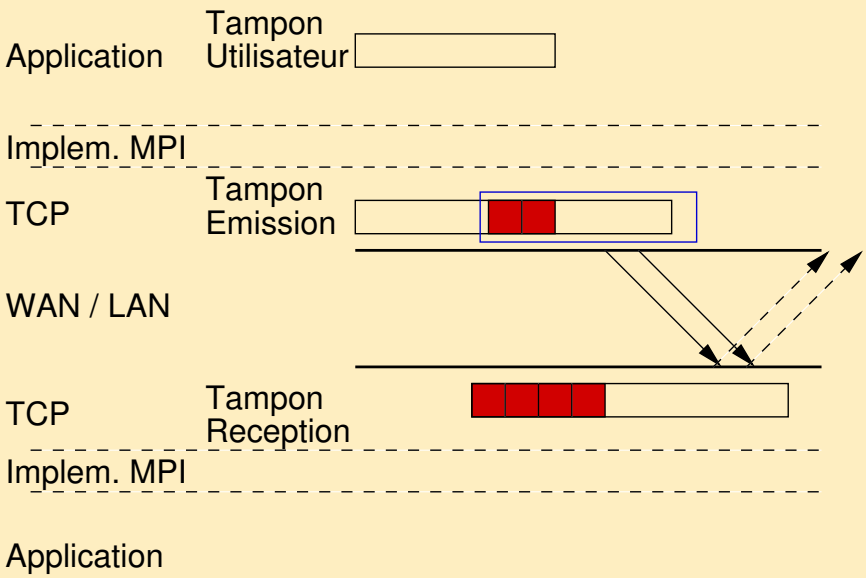
Interaction entre MPI et TCP



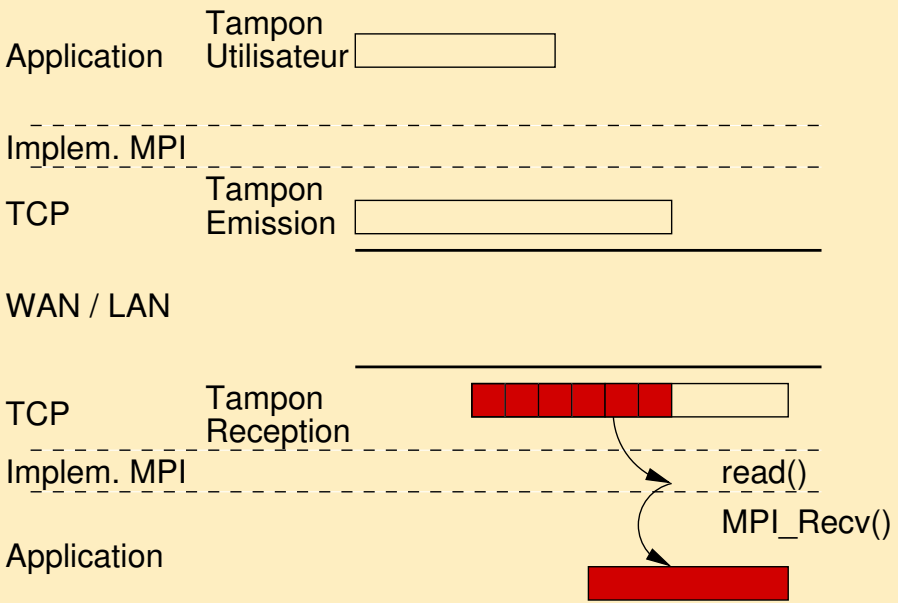
Interaction entre MPI et TCP



Interaction entre MPI et TCP



Interaction entre MPI et TCP



Suppression du démarrage lent pour les applications MPI

Le démarrage lent permet de déterminer le débit d'émission approprié sur un lien. Il intervient à 3 moments :

- au démarrage d'une connexion
- après une rafale de perte
- après un temps d'inactivité : particulièrement coûteux pour les applications MPI qui communiquent peu souvent.

Peut-on supprimer le démarrage lent après une période d'inactivité pour les applications MPI ?

Suppression du démarrage lent pour les applications MPI

Le démarrage lent permet de déterminer le débit d'émission approprié sur un lien. Il intervient à 3 moments :

- au démarrage d'une connexion
- après une rafale de perte
- après un temps d'inactivité : particulièrement coûteux pour les applications MPI qui communiquent peu souvent.

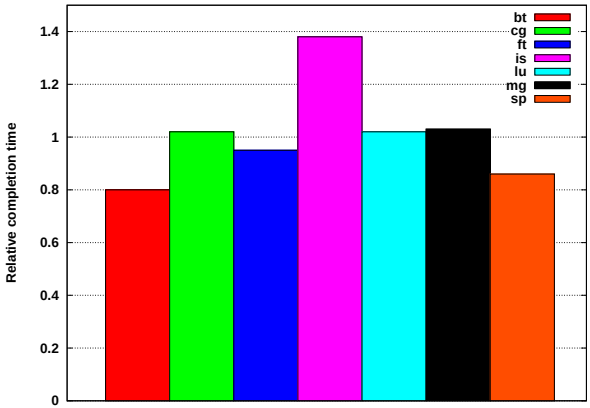
Peut-on supprimer le démarrage lent après une période d'inactivité pour les applications MPI ?

Suppression du démarrage lent pour les applications MPI

Le démarrage lent permet de déterminer le débit d'émission approprié sur un lien. Il intervient à 3 moments :

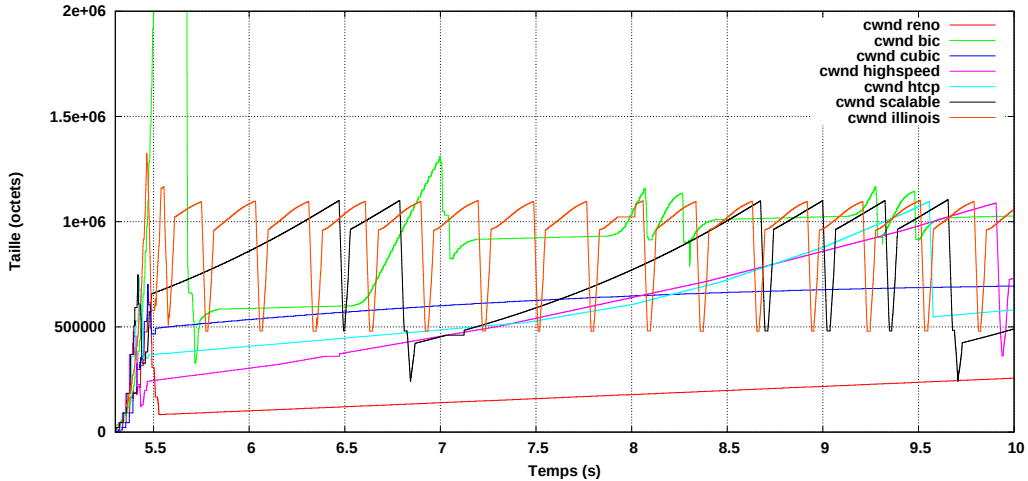
- au démarrage d'une connexion
- après une rafale de perte
- après un temps d'inactivité : particulièrement coûteux pour les applications MPI qui communiquent peu souvent.

Peut-on supprimer le démarrage lent après une période d'inactivité pour les applications MPI ?



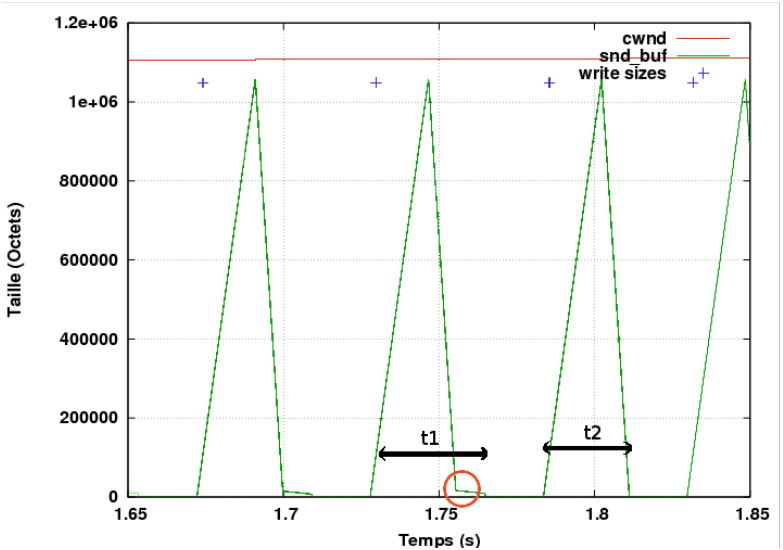
Les différentes variantes de TCP

- TCP New Reno pose problème sur les liens avec un grand débit et/ou une forte latence.
- Différentes variantes de TCP améliorent l'agressivité de l'algorithme : changement des facteurs d'augmentation et de diminution de la fenêtre de congestion.
- On peut citer : BIC, CUBIC, Highspeed, Hamilton TCP (H-TCP), Scalable, Illinois.



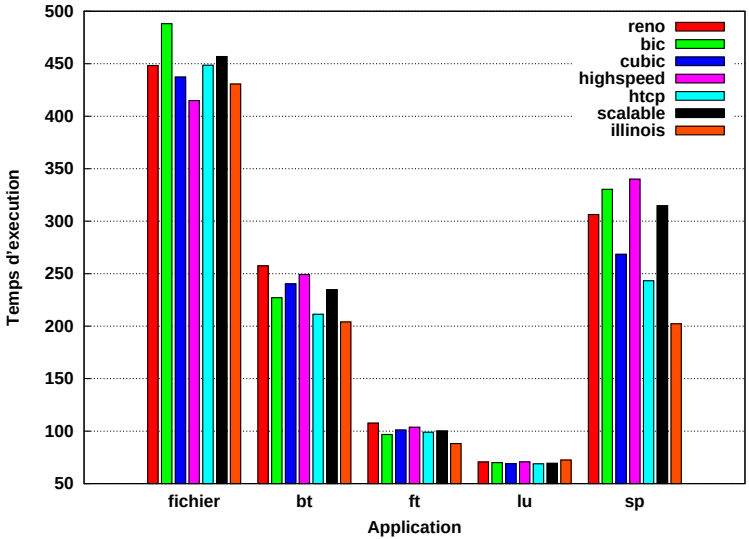
Impact du contrôle de congestion sur les applications MPI

Le contrôle de congestion limite l'émission des données MPI (et ralentit l'exécution d'une application)



Changement de variante de TCP pour les applications MPI

Est-ce qu'une variante de TCP plus agressive permet de limiter le phénomène de rétention des messages MPI en garantissant une plus grande fenêtre de congestion ?

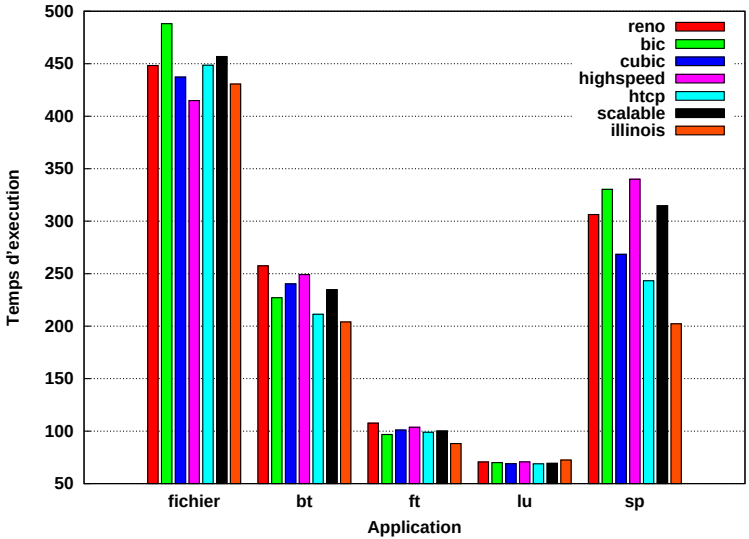


Quelle variante choisir ?

- pour les applications MPI, Illinois semble la plus appropriée dans nos tests
- nombreux paramètres : latence, bande passante, taux de congestion, taux de multiplexage ...

Changement de variante de TCP pour les applications MPI

Est-ce qu'une variante de TCP plus agressive permet de limiter le phénomène de rétention des messages MPI en garantissant une plus grande fenêtre de congestion ?



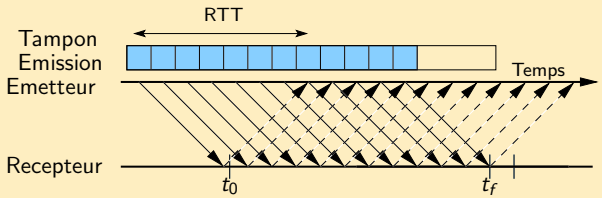
Quelle variante choisir ?

- pour les applications MPI, Illinois semble la plus appropriée dans nos tests
- nombreux paramètres : latence, bande passante, taux de congestion, taux de multiplexage ...

Impact du contrôle de fiabilité

- Effectue les retransmissions des paquets perdus ou erronés.
- La détection d'une perte se fait : par la réception de ACK dupliqués ou l'expiration du délai de retransmission.

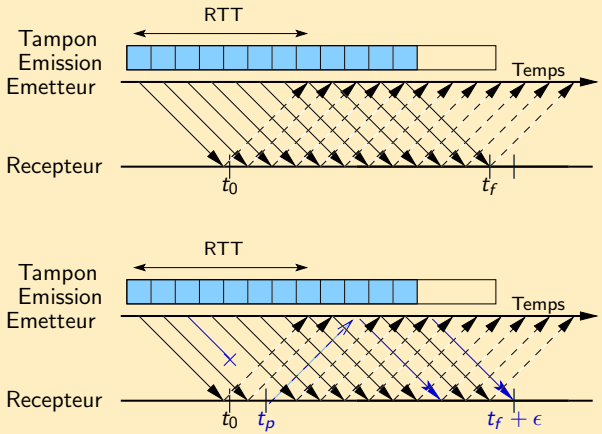
Impact d'une perte sur le transfert d'un fichier



Impact du contrôle de fiabilité

- Effectue les retransmissions des paquets perdus ou erronés.
- La détection d'une perte se fait : par la réception de ACK dupliqués ou l'expiration du délai de retransmission.

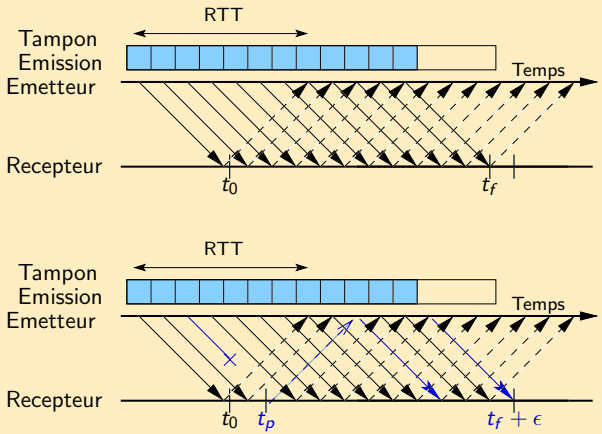
Impact d'une perte sur le transfert d'un fichier



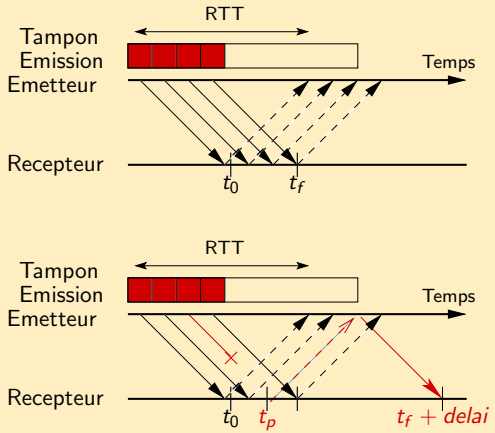
Impact du contrôle de fiabilité

- Effectue les retransmissions des paquets perdus ou erronés.
- La détection d'une perte se fait : par la réception de ACK dupliqués ou l'expiration du délai de retransmission.

Impact d'une perte sur le transfert d'un fichier



Impact d'une perte sur le transfert d'un message MPI



Synthèse

- Le contrôle de congestion et le contrôle de fiabilité ralentissent les applications MPI
- Ces deux mécanismes sont basés sur le RTT qui est très grand comparé au temps d'émission d'un message MPI
- Certaines variantes de TCP permettent une amélioration sensible des performances.
- La granularité de TCP n'est pas assez fine pour les communications MPI

Comment réduire l'impact de TCP sur les applications MPI ?

Synthèse

- Le contrôle de congestion et le contrôle de fiabilité ralentissent les applications MPI
- Ces deux mécanismes sont basés sur le RTT qui est très grand comparé au temps d'émission d'un message MPI
- Certaines variantes de TCP permettent une amélioration sensible des performances.
- La granularité de TCP n'est pas assez fine pour les communications MPI

Comment réduire l'impact de TCP sur les applications MPI ?

Synthèse

- Le contrôle de congestion et le contrôle de fiabilité ralentissent les applications MPI
- Ces deux mécanismes sont basés sur le RTT qui est très grand comparé au temps d'émission d'un message MPI
- Certaines variantes de TCP permettent une amélioration sensible des performances.
- La granularité de TCP n'est pas assez fine pour les communications MPI

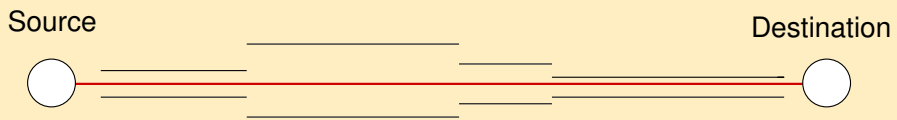
Comment réduire l'impact de TCP sur les applications MPI ?

Plan

Eclatement des connexions TCP

SplitTCP [Kopparty et al. , 02]

- créé dans le contexte des réseaux sans fil
- a pour but de différencier les liens traversés



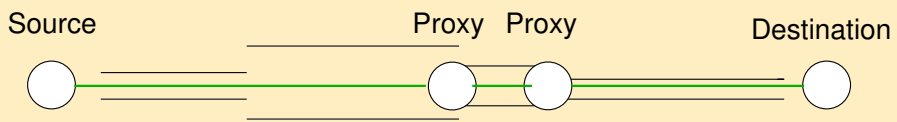
Pour MPI, l'éclatement des connexions permet :

- de rendre visible le réseau longue-distance
- de proposer des optimisations au niveau des passerelles

Eclatement des connexions TCP

SplitTCP [Kopparty et al. , 02]

- créé dans le contexte des réseaux sans fil
- a pour but de différencier les liens traversés



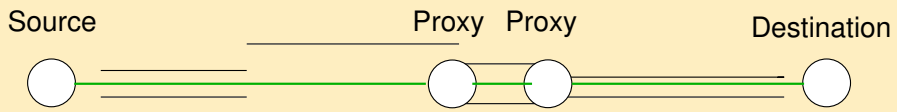
Pour MPI, l'éclatement des connexions permet :

- de rendre visible le réseau longue-distance
- de proposer des optimisations au niveau des passerelles

Eclatement des connexions TCP

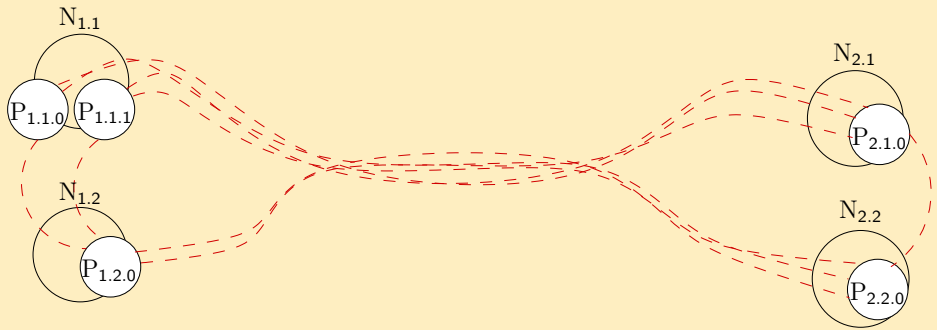
SplitTCP [Kopparty et al. , 02]

- créé dans le contexte des réseaux sans fil
- a pour but de différencier les liens traversés



Pour MPI, l'éclatement des connexions permet :

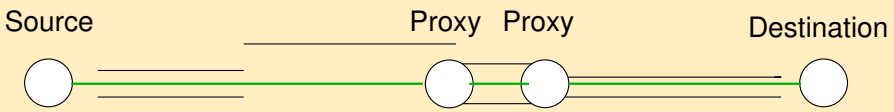
- de rendre visible le réseau longue-distance
- de proposer des optimisations au niveau des passerelles



Eclatement des connexions TCP

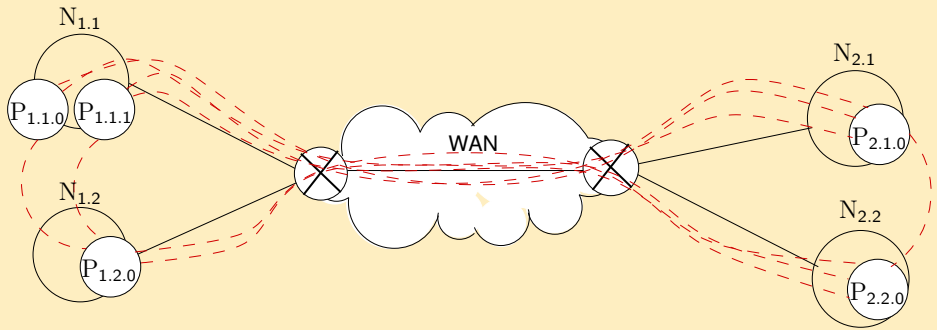
SplitTCP [Kopparty et al. , 02]

- créé dans le contexte des réseaux sans fil
- a pour but de différencier les liens traversés



Pour MPI, l'éclatement des connexions permet :

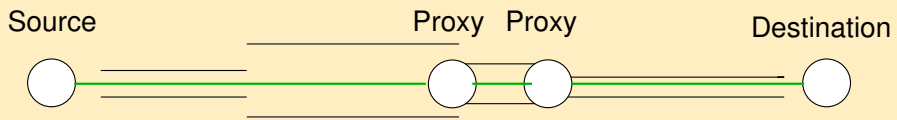
- de rendre visible le réseau longue-distance
- de proposer des optimisations au niveau des passerelles



Eclatement des connexions TCP

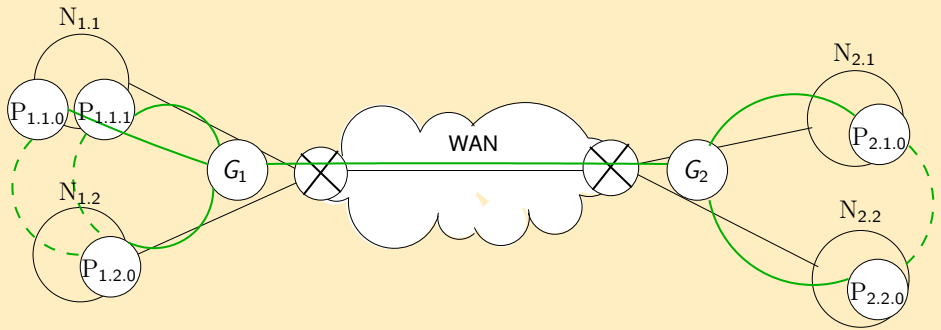
SplitTCP [Kopparty et al. , 02]

- créé dans le contexte des réseaux sans fil
- a pour but de différencier les liens traversés



Pour MPI, l'éclatement des connexions permet :

- de rendre visible le réseau longue-distance
- de proposer des optimisations au niveau des passerelles



Eclatement des connexions TCP : avantages et inconvénients

Avantages :

- Diminution du nombre de connexions et donc de la quantité de mémoire utilisée
- Diminution des pertes longue distance
- Fenêtre de congestion plus proche de la capacité réelle du lien longue distance
- Détection de pertes plus rapide

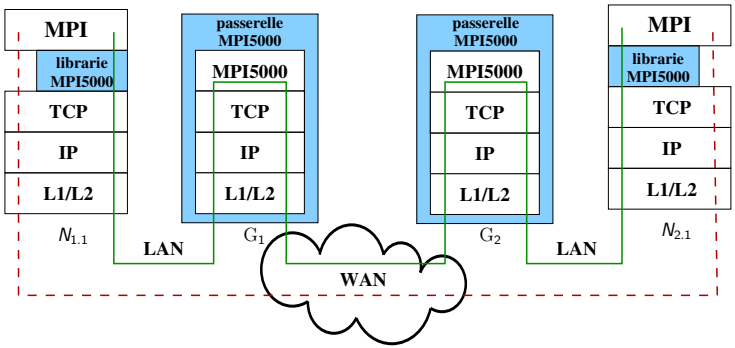
Inconvénient :

- Coût de recopie au niveau des passerelles

Optimisation possibles grâce à l'utilisation de passerelles

- Utilisation d'une variante de TCP différente sur le WAN et sur le LAN (par exemple Reno sur le LAN and HighSpeed TCP sur le WAN)
- Réservation de bande passante entre les passerelles pour limiter la congestion
- Utilisation de différentes stratégies en fonction de la taille des messages

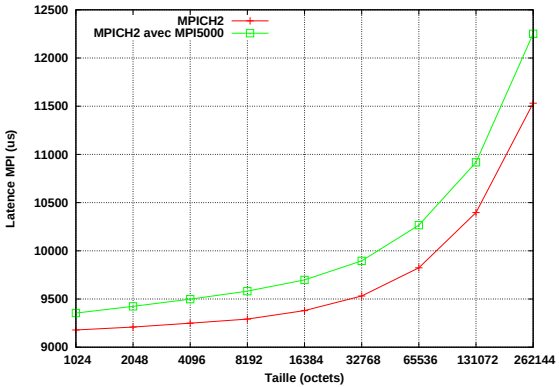
MPI5000 : mise en oeuvre de l'éclatement des connexions à base de passerelles

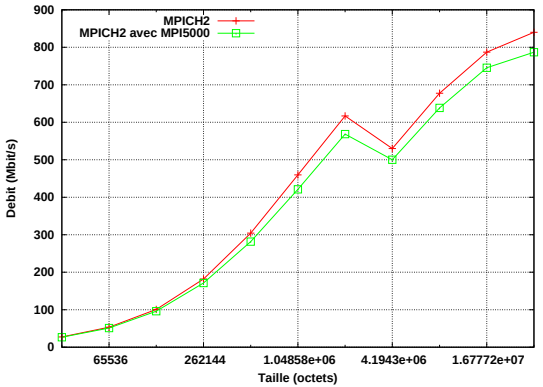


Trois éléments dans MPI5000 :

- Librairie : permet l'interception des appels aux fonctions de l'API socket pour rediriger les connections vers la passerelle du site. Cette librairie est lancée de manière transparente.
- Passerelles : retransmettent les données vers une autre passerelle ou vers les noeuds locaux.

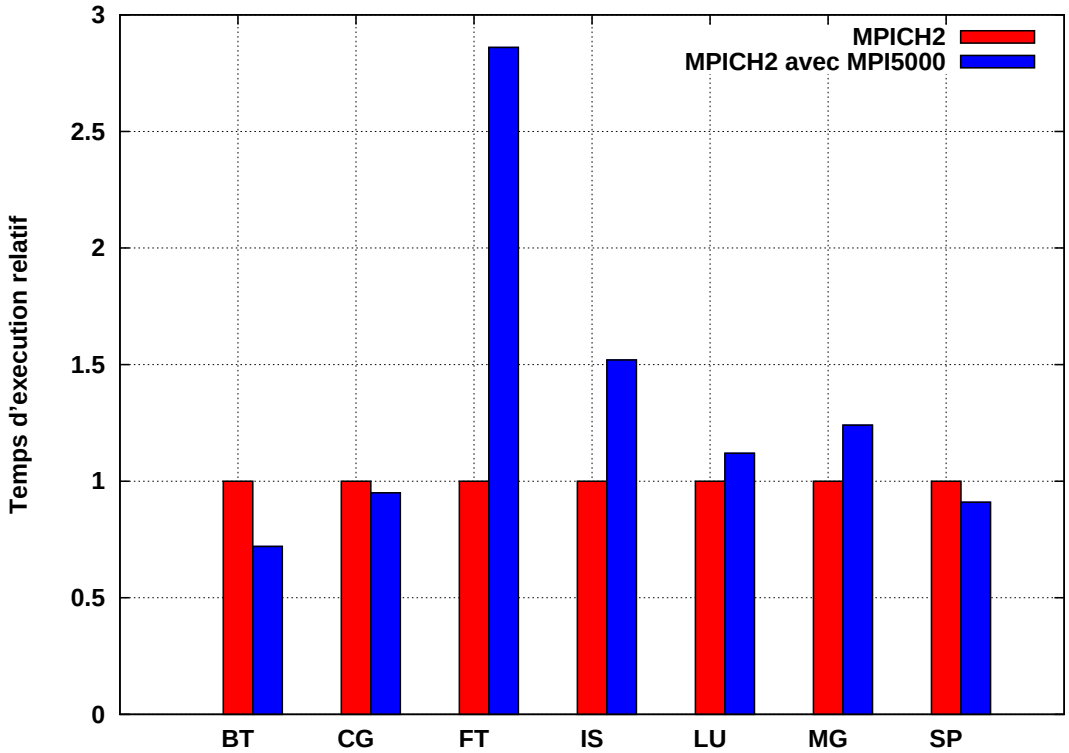
Evaluation de MPI5000 : surcoût





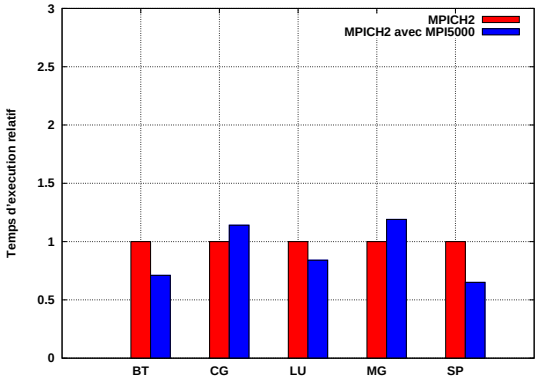
- Ajout de 141 μs en latence dû :
  - aux recopies dans les passerelles
  - au temps aller-retour entre les passerelles et le noeud
- La bande passante diminue de 7% (de 840 à 785 Mb/s)

Evaluation de MPI5000 : Exécution des NPB



Evaluation de MPI5000 : Réduction du nombre de pertes sur le WAN

- Est-ce que MPI5000 permet de diminuer le nombre de pertes sur le réseau longue distance ?
- Peu de pertes dans le cas précédent : ajout de trafic concurrent.

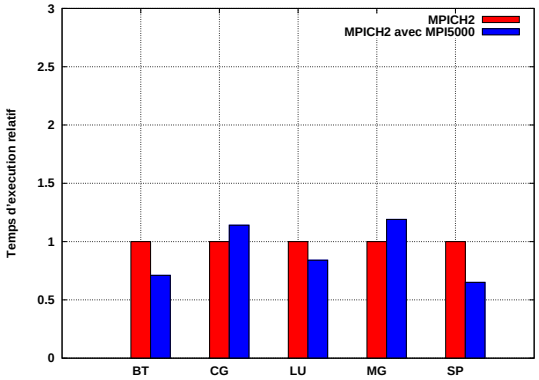


	MPICH2 sans MPI5000		MPICH2 avec MPI5000			
	Distant		Local		Distant	
NPB	DupAck	RTOs	DupAck	RTOs	DupAck	RTOs
BT	757	56	4	1	320	1
CG	78	25	0	0	54	19
LU	327	232	0	0	174	41
MG	94	53	7	0	48	4
SP	1409	778	8	0	667	131

- Diminution du nombre des pertes sur le longue distance pour tous les NPB
- Diminution faible pour CG et MG : MPI5000 n'améliore pas le temps d'execution
- Diminution significative pour BT, LU et SP : MPI5000 améliore le temps d'exécution

Evaluation de MPI5000 : Réduction du nombre de pertes sur le WAN

- Est-ce que MPI5000 permet de diminuer le nombre de pertes sur le réseau longue distance ?
- Peu de pertes dans le cas précédent : ajout de trafic concurrent.

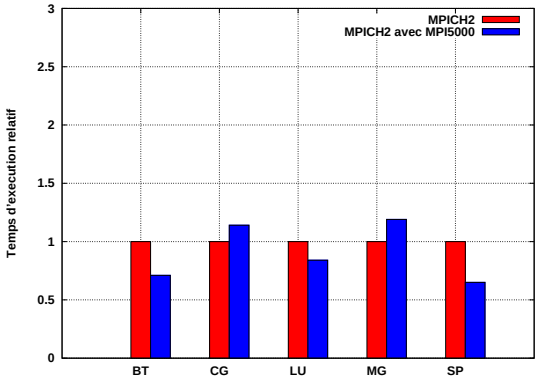


	MPICH2 sans MPI5000		MPICH2 avec MPI5000			
	Distant		Local		Distant	
NPB	DupAck	RTOs	DupAck	RTOs	DupAck	RTOs
BT	757	56	4	1	320	1
CG	78	25	0	0	54	19
LU	327	232	0	0	174	41
MG	94	53	7	0	48	4
SP	1409	778	8	0	667	131

- Diminution du nombre des pertes sur le longue distance pour tous les NPB
- Diminution faible pour CG et MG : MPI5000 n'améliore pas le temps d'execution
- Diminution significative pour BT, LU et SP : MPI5000 améliore le temps d'exécution

Evaluation de MPI5000 : Réduction du nombre de pertes sur le WAN

- Est-ce que MPI5000 permet de diminuer le nombre de pertes sur le réseau longue distance ?
- Peu de pertes dans le cas précédent : ajout de trafic concurrent.



	MPICH2 sans MPI5000		MPICH2 avec MPI5000			
	Distant		Local		Distant	
NPB	DupAck	RTOs	DupAck	RTOs	DupAck	RTOs
BT	757	56	4	1	320	1
CG	78	25	0	0	54	19
LU	327	232	0	0	174	41
MG	94	53	7	0	48	4
SP	1409	778	8	0	667	131

- Diminution du nombre des pertes sur le longue distance pour tous les NPB
- Diminution faible pour CG et MG : MPI5000 n'améliore pas le temps d'execution
- Diminution significative pour BT, LU et SP : MPI5000 améliore le temps d'exécution

Plan

Conclusion

- Analyse des communications MPI sur le réseau longue distance :
  - implémentation de deux outils (InstrAppli et tcp_probe)
  - application aux NPB
- Etude de l'interaction entre MPI et TCP :
  - Impact de la fenêtre de congestion si elle est trop petite
  - Impact disproportionné du contrôle de fiabilité sur les messages MPI

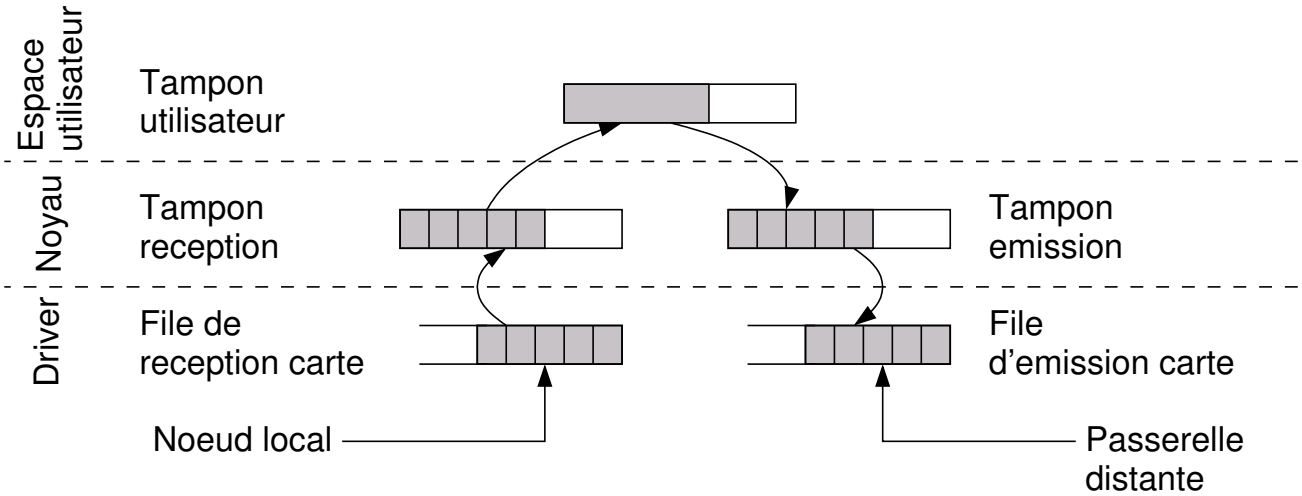
  $\rightarrow$ le RTT n'est pas une granularité assez fine par rapport au temps d'émission d'un message MPI
- Proposition d'éclater les connexions TCP pour les applications MPI : implémentation et évaluation d'une architecture à base de passerelles.
  - Passerelles coûteuses sur les gros messages
  - Diminution du temps d'exécution de BT et SP de l'ordre de 30%
  - Validation de l'approche : réduction des pertes longue-distance

Perspectives

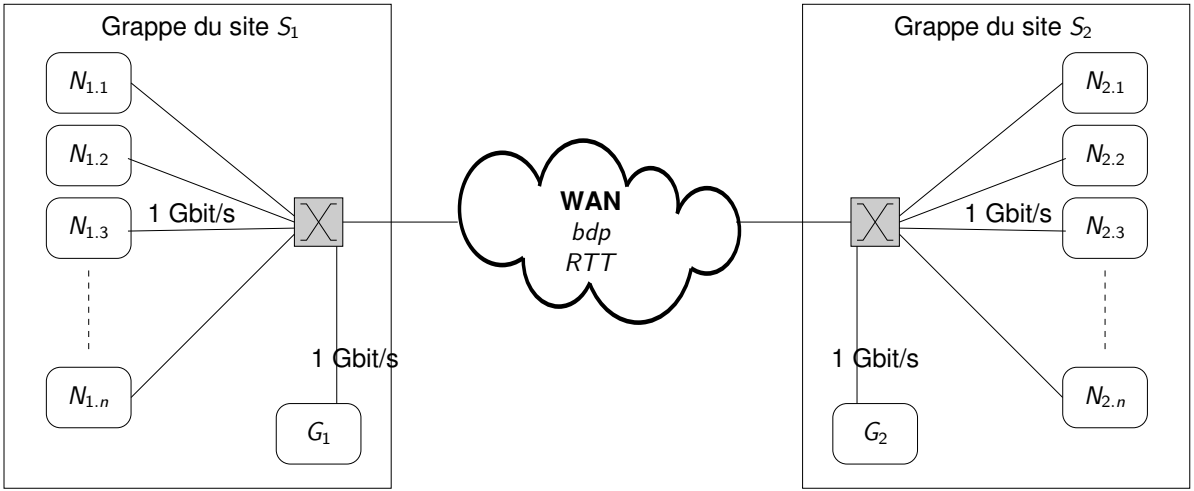
- Optimisation des passerelles
- Limitation de débit sur le réseau longue distance
- Modèlisation des communications MPI sur TCP : modélisation de la fenêtre de congestion
- Adaptation du protocole de transport : trouver une taille de fenêtre de congestion plus appropriée pour des messages de type MPI

Questions

Cout des recopies



Banc d'essai


Grappe du site S_1
$N_{1.1}$
$N_{1.2}$
$N_{1.3}$
$N_{1.n}$
1 Gbit/s
1 Gbit/s
G_1
WAN
bdp
RTT
Grappe du site S_2
$N_{2.1}$
$N_{2.2}$
$N_{2.3}$
$N_{2.n}$
1 Gbit/s
1 Gbit/s
G_2