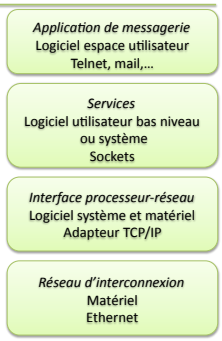
3. Architectures de communication

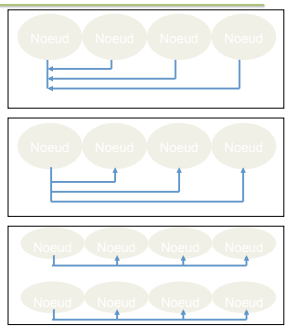
Introduction

- Contexte pour ce chapitre : mémoire distribuée, espaces d'adressages multiples
- Mais la partie réseau d'interconnexion a des applications pour d'autres architectures



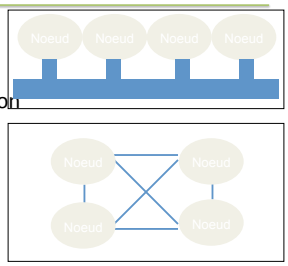
Fonctionnalités

- Obligatoire
  - Connexion point-à-point entre nœuds
- Optionnelle
  - Support communications collectives
  - Réduction
  - Diffusion 1-to-many
  - Multi-diffusion mainy-to-mainy



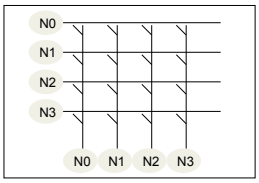
Réseaux élémentaires

- Bus
  - Exclusion mutuelle
  - Point à point et diffusion
  - Extensibilité limitée
- Connexion complète
  - $O(P^2)$ liens
  - O(P) ports/noeud
  - Non extensible



Crossbar

- Relativement extensible
  - P^2 points de croisement
  - 2P liens
  - 1 port/nœud
  - Typiquement –Fin 2009 – 32 ports

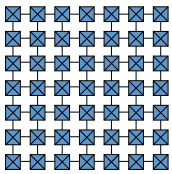


Exemple : Myrinet switch
http://www.myri.com/

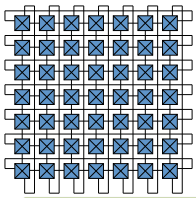
Réseaux incomplets

- Pour le parallélisme massif, réseaux incomplets
- Il n'existe pas de lien physique entre deux nœuds quelconques
- Le matériel doit assurer le routage
  - Transferts point à point entre deux processeurs quelconques
  - Eventuellement collectifs
  - Avec des performances élevées
    - Géométrie régulière
    - Routeurs = switch + contrôle
- Vision duale : la topologie des communication d'un algorithme

Réseaux directs : grilles et tores

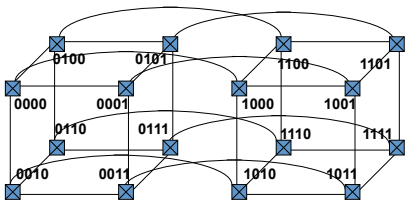


Grille
$N = n^2$
– Extensible
– Non isotrope



Tore $T(n)$
$N = n^2$
– Extensible
– Isotrope

Réseaux directs : hypercube $H(n)$



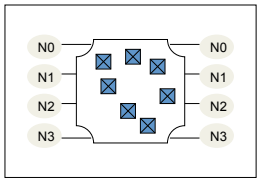
- $N = 2^n$
- Extensible par doublement
- Isotrope

Comparaison des réseaux directs

- Peu significatifs pour les architectures parallèles
  - Diamètre : plus grande distance
    - Hypercube : $\log N$ Tore 2D : $\sqrt{N}$
  - Distance moyenne
    - Hypercube : $\frac{1}{2} \log N$ Tore 2D : $\frac{1}{2} \sqrt{N}$
- Très important pour les performances des algorithmes
  - Bisection géométrique
    - Hypercube : $N/2$ Tore 2D : $2\sqrt{N}$
  - Mais à pondérer par les contraintes technologiques

Réseaux indirects

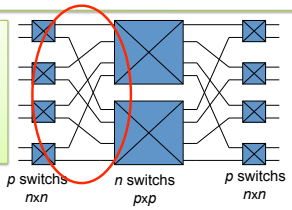
- Les processeurs sont en entrée et sortie du réseau seulement
- *Réarrangeable* si peut réaliser toute permutation des entrées sur les sorties



Reseaux indirects : le réseau de Clos C(n,p)

Connexion shuffle

La sortie i du switch d'entrée j est connectée à l'entrée i du switch j

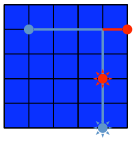


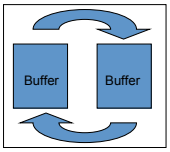
- Le réseau de Clos $C(n,p)$ est réarrangeable
- Réseau de Benes : $n = 2$ et décomposition récursive des switchs intermédiaires

Exemple : Myrinet switch

Routage

- Algorithme de choix des chemins
  - Câblé : routage par dimension
    - Grilles, hypercubes
  - Tabulé
    - Réseaux réarrangeables
  - Déterministe/aléatoire – fixé/adaptatif
    plus court chemin
- Traitement des conflits
  - Tamponnement ou déroutement
    hot-potatoe
  - Résolution d'interblocage par réseau virtuel
- Contrôle de flot
  - Paquet
  - Wormhole





Caractéristiques de la performance

- Des concepts identiques appliqués dans des contextes différents
- Latence isolée : pour un message vide
  - Vue du matériel
  - Vue du programme
- Débit
  - Pour un ping-pong
  - En situation de charge

Composantes de la performance

- Internes au réseau : matériel
  - Barebones latency
- Interface processeur-réseau
  - Matériel : débit critique – problème du bus I/O
  - Logiciel :
    - Couches de protocoles
    - Nombre de copies

Performance of MPICH-MX 1.2.7..1 over MX-10G
Uniprocessor case: one process per node, one NIC per node