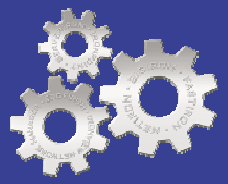
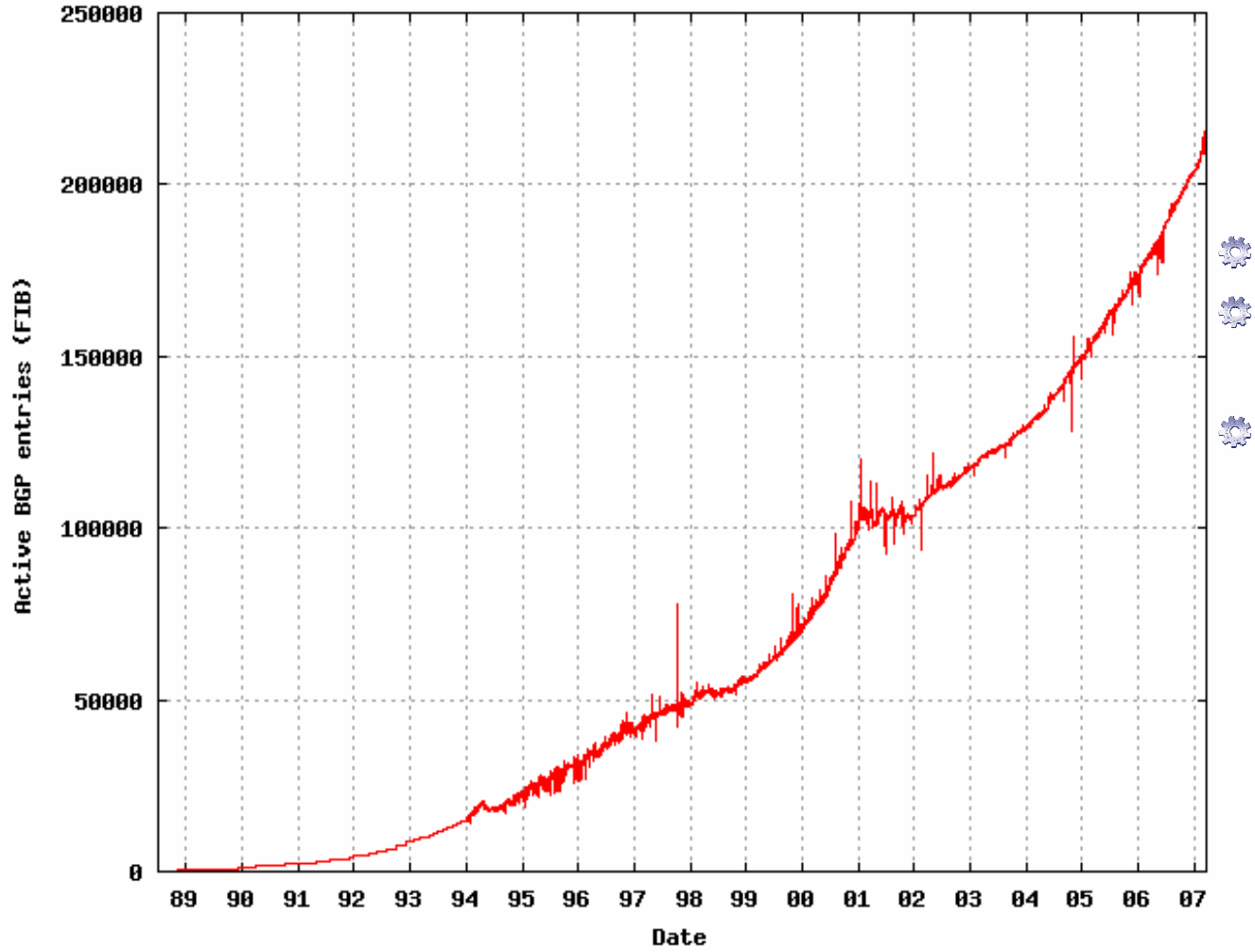
Table de routage IPv4 : Ses implications sur les routeurs



Agenda

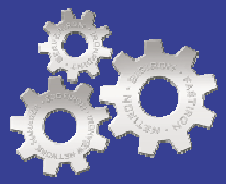
- Croissance continue de la FIB : La base du problème
- Au coeur d'un routeur Multi Gigabit
  - Technologies disponibles et à venir
  - Gestion des routes de la FIB
- Optimisations possibles
- "Un coût toujours plus élévé" … Mythe où réalité ?
- Conclusion

Croissance continue de la FIB : La base du problème



- Environ 235K routes
- Croissance : ~17% par an / 4 dernières années
- ~400-500 routes ajoutées par semaine

Source: CIDR Report, March 23 2007

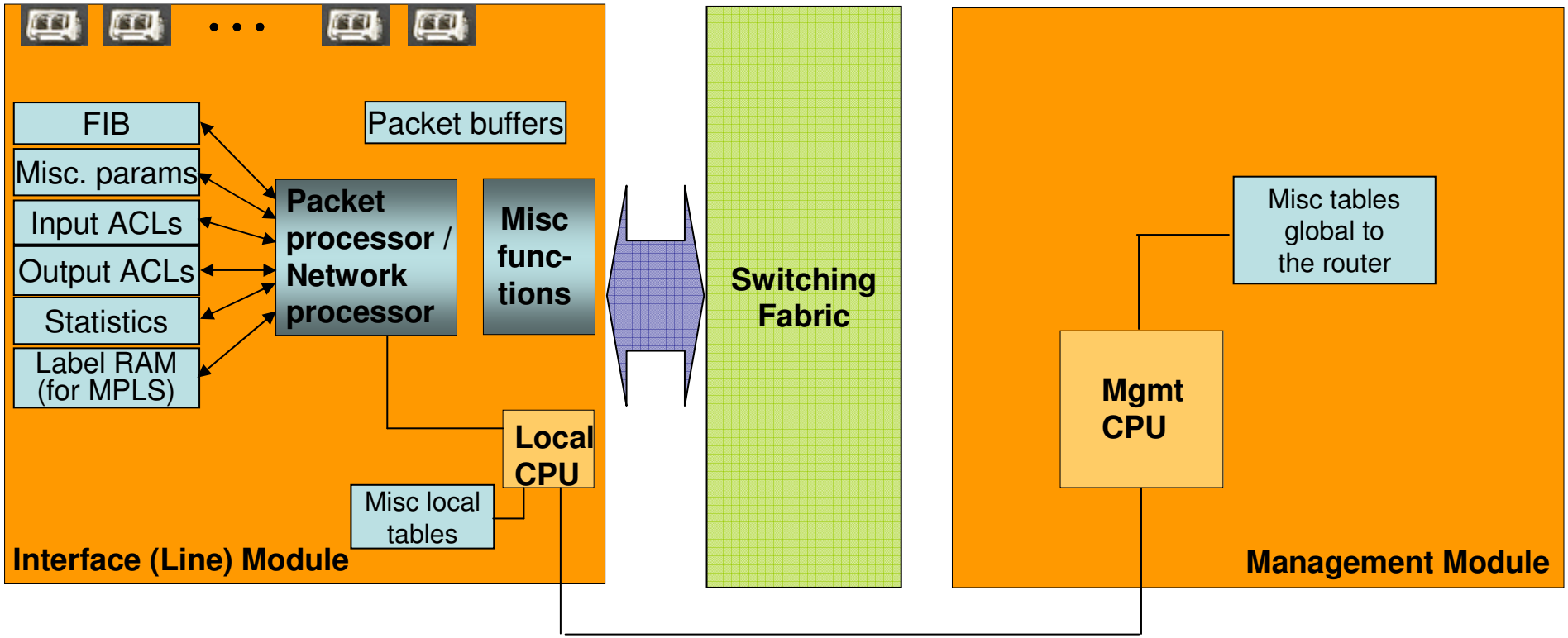


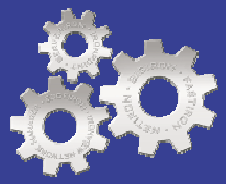
Quelle est la gravité du problème ?

- Cela dépend :
  - Si il s'agit d'un routeur dont les limites de la FIB sont atteintes
    - Alors il s'agit d'un problème à gérer dés aujourd'hui
  - Si il s'agit d'un routeur ayant encore des ressources disponibles. En s'appuyant sur les prévisions de croissance, soit :
    - Les ressources permettent de voir les années à venir de manière sereine,
    - Les ressources ne sont pas suffisantes et dans ce cas, il faut planifier :
      - Le remplacement de l'équipement,
      - La mise à jour ce celui-ci

- Facteurs influençant cette croissance :
  - Multi-homing
  - Fragmentation excessive des blocs attribués
  - Règles d'ingénieurie
  - Fusion et acquisition de companies
  - …

Au coeur d'un routeur Multi Gigabit

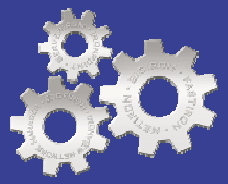




Quels composants pour stocker la FIB (1)?

- Plusieurs options disponibles : CAM, SRAM, DRAM

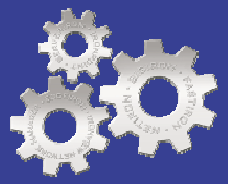
- CAM :
  - Le moyen le plus rapide/déterministe de rechercher un “longest prefix”
  - La plus coûteuse des 3 options
  - Les composants les plus denses en grande quantité supportent 18 Megabits :
    - 1 entrée IPv4 utilisant 32 bits, 512K routes IPv4 peuvent donc être aisément stockées dans les CAM les plus denses disponibles aujourd’hui.
  - De multiples CAM peuvent être utilisées :
    - En cascade pour accroître la capacité de stockage si une CAM ne suffit pas
    - Les recherches peuvent alors être faites en parallèle
  - Des CAM de 36 Megabit devraient être disponibles en production de masse d’ici peu.
    - Possibilité de mettre 1M de routes IPv4 dans un seul composant.
  - Conclusion :
    - La technologie CAM existe depuis longtemps et est éprouvée
    - Elle permet de passer le 1M de routes IPv4 si le marché le demande.



Quels composants pour stocker la FIB (2)?

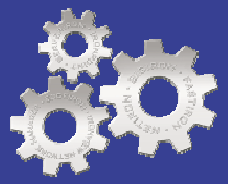
- SRAM :
  - Temps de “lookup” plus long car il faut réaliser de multiples opérations
    - Moins déterministe que les CAM, mais beaucoup plus que la DRAM
    - Un composant de stockage dont les mécanismes de “lookup” sont à implémenter
    - Des améliorations continuent à être faites pour rendre la SRAM plus déterministe
  - Moins coûteuse que les CAM, mais plus chère que la DRAM
  - Capacité maximale disponible aujourd’hui en masse : 72 Megabits
  - Conclusion:
    - La SRAM permet de passer les 1M de routes IPv4 si le marché le demande
    - Optimisations à réaliser sur les recherches

- DRAM :
  - Le coût peut varier en fonction du type de DRAM choisie (SDRAM, DR, DDR, DDR2)
  - Le temps de “lookup” peut ne pas être déterministe du tout
  - Le D de DRAM signifie “Dynamic” :
    - Un rafraîchissement périodique est nécessaire
  - Conclusion :
    - Même si il est possible d’utiliser la DRAM, son coté non déterministe la rend plus difficilement utilisable dans des environnement avec énormément de routes.



Ou sont utilisés ces composants ?

- La croissance du nombre de route affecte les éléments suivant :
  - La FIB utilisée dans les packet processor, network processor et asics
    - Typiquement le domaine des technologies CAM ou SRAM
  - La copie de la FIB dans le module de supervision/management
    - Technologie de type DRAM
  - La copie de la FIB dans les modules d'interfaces
    - Technologie de type DRAM
  - La table RIB maintenue au niveau du module qui gère les protocoles de routages
    - Généralement le module de supervision/management
    - Technologie de type DRAM
    - De la mémoire tampon est également nécessaire pour gérer les messages d'update venant des peers



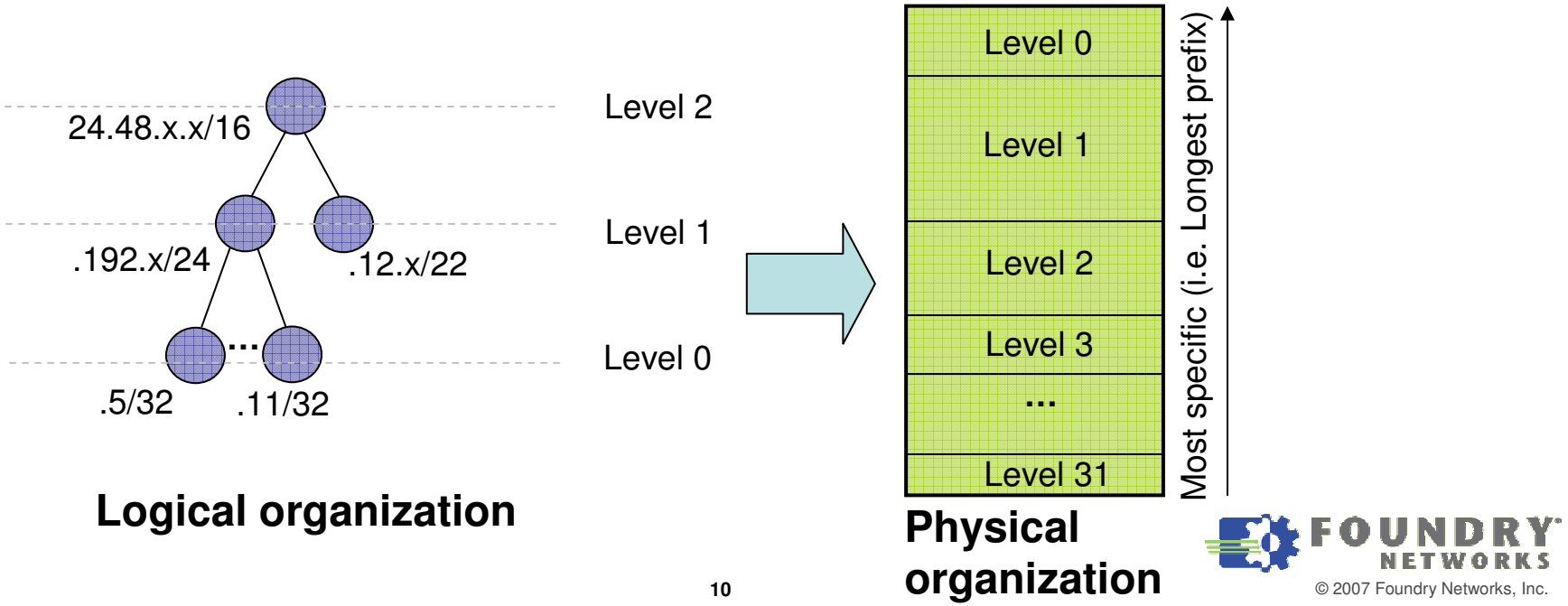
Organisation des routes dans la FIB (1)

- En plus du dimensionnement de la mémoire, une structure de données appropriée doit être définie pour organiser les routes
  - Pour les FIB basées sur la SRAM, c'est un pré-requis comme la SRAM est une mémoire de stockage.
  - Pour les FIB basées sur la CAM, l'ordre des entrées est important
    - Pour rechercher le préfixe le plus long, ceux-ci doivent se trouver dans les premiers niveaux de la CAM.

- Actuellement, la structure de données la plus optimale pour mémoriser la FIB est une structure en arbre appelée "trie" (pour re**trie**val)
  - Détermination aisée du temps de résolution le plus long.
    - Dépend de la longueur de la clé
  - Chaque noeud d'une structure de type "trie" réprésente une portion du préfixe.
  - 2 éléments clés :
    - Pas (Stride) : Nombre de bits inspectés simultanément
    - Profondeur de l'arbre : Chemin entre la racine et le noeud le plus éloigné de la racine

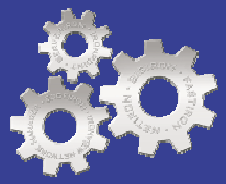
Organisation des routes dans la FIB (2)

- Quelques type de “tries” :
  - M-level trie
    - Exemple : un “trie” à 3 niveaux avec des pas de 16, 8, 8 signifie :
      - Au premier niveau, un /16 est maintenu (Soit un total de 64K entrées)
      - Au second niveau, un /8 additionnel est maintenu (Soit un total de 256 entrées par parent)
      - Au troisième niveau, un /8 additionnel est maintenu (Soit un total de 256 entrées par parent)
  - Dynamic prefix trie (DP-trie)
    - Le nombre de niveaux peut varier, mais permet une bonne optimisation de la mémoire utilisée



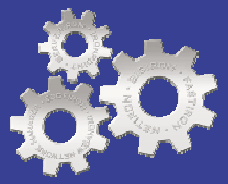
Logical organization

Physical organization



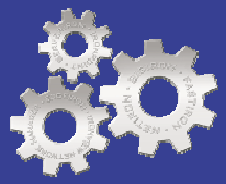
Organisation des routes dans la FIB (3)

- Points for de l'approche en "trie" :
  - Typiquement les 4 premiers niveaux prennent en compte une très grande majorité (>99%) des routes
  - Lors d'une mise à jour de routes, les "updates" au sein d'un même niveau sont triviaux :
    - Revient à considérer que la longueur du préfixe le plus long pour la nouvelle route est la même que pour une route existante
    - Puis ajouter la nouvelle route dans le même niveau.
  - L'opération d'ajout d'un niveau lors d'une mise à jour des routes, même si elle est plus longue, est pratiquement imperceptible
  - L'allocation dynamique de chaque niveau permet une grande flexibilité
- De nouvelles formes d'optimisations pour organiser la recherche des "longest prefixes" (nouvelles structures d'arbres par exemple) affectent uniquement le logiciel.
- Ces optimisations peuvent potentiellement permettre des mises à jour plus rapides vers les mémoires (CAM, SRAM, …)



Optimisations possibles

- Pour chaque problème résolu, il faut s'assurer que les contraintes induites soient bien connues
  - Temps de traîtement par rapport à l'économie de place
- La technologie n'est pas une contrainte :
  - Les capacités des CAM et SRAM sont bien au-delà de ce que demande une "full view".
- Les possibilités d'optimisations incluent :
  - Mises à jour de la RIB
  - Mises à jour et distribution de la FIB
  - Autres optimisations possibles : Net Aggregation, filtrage des routes, …



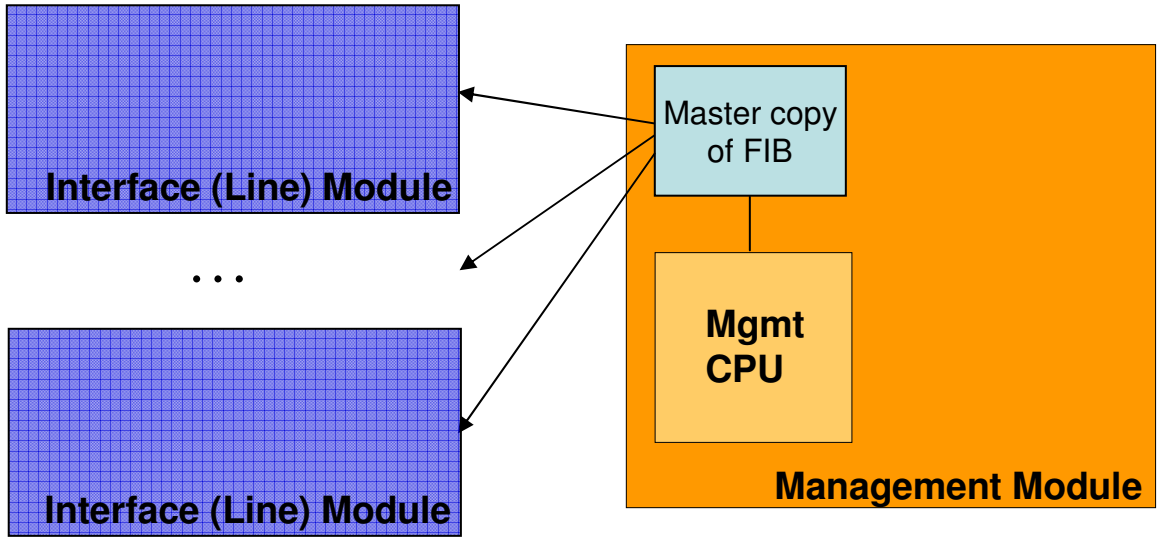
La RIB et ses mises à jour

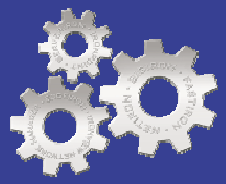
- RIB
  - Contient l'ensemble des routes apprises par les différents protocoles de routage
  - Plusieurs fois la taille de la FIB
  - Nécessite une bonne optimisation dans sa gestion : ressources CPU et mémoire
- Optimisation au niveau des routes-map pour mieux contrôler les routes reçues/apprises
- Le temps de mise à jour de la RIB dépend en partie des ressources CPU disponibles :
  - Activité des autres processus
  - Filtrage des routes
  - L'operating system est-il distribué ?
    - Les CPU des cartes d'interfaces peuvent-elles assister le processeur du module de management ?
- Du point de vue de la CPU
  - Processeurs multi coeurs disponibles
  - Fréquences élevées même pour les processeurs embarqués

Mise à jour et Distribution de la FIB

- Sur des systèmes multi processeurs :
  - Le module de supervision maintient une copie centralisée de la FIB
  - Une copie locale est conservée sur chacun des modules d'interface

- Il faut donc que la distribution de la FIB soit optimale, avec par exemple :
  - Utilisation d'un réseau interne, dédié, en haut débit pour la distribution des copies
  - Compression des mises à jour transmises par le module de supervision vers les modules d'interfaces





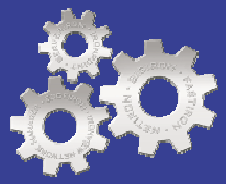
Autres Optimisations possibles

- Net aggregation
  - Agréger les prefixes contiguës qui ont le même next-hop dans un préfixe unique plus grand
  - Objectif : renverser la fragmentation parfois excessive de certains blocs (si le next-hop le permet)
  - Local au routeur

  - Avantages :
    - Permet d'accroître la durée de vie de l'équipement (Si le produit supporte cette fonction).
    - Peut devenir plus importante dans le futur si la fragmentation des blocs continue.

  - Inconvénients :
    - A cause du flapping et de la nature dynamique de l'Internet, ce qui était agrégé hier ou pouvait l'être, peut ne plus l'être demain.
    - L'espace occupé par la FIB peut être préservé, mais cela nécessite une activité plus importante de la CPU pour réaliser ces agrégations.

- Filtrage des routes apprises
  - Vues partielles,
  - Routes par défaut.



“Un coût toujours plus élevé” … Mythe où Réalité?

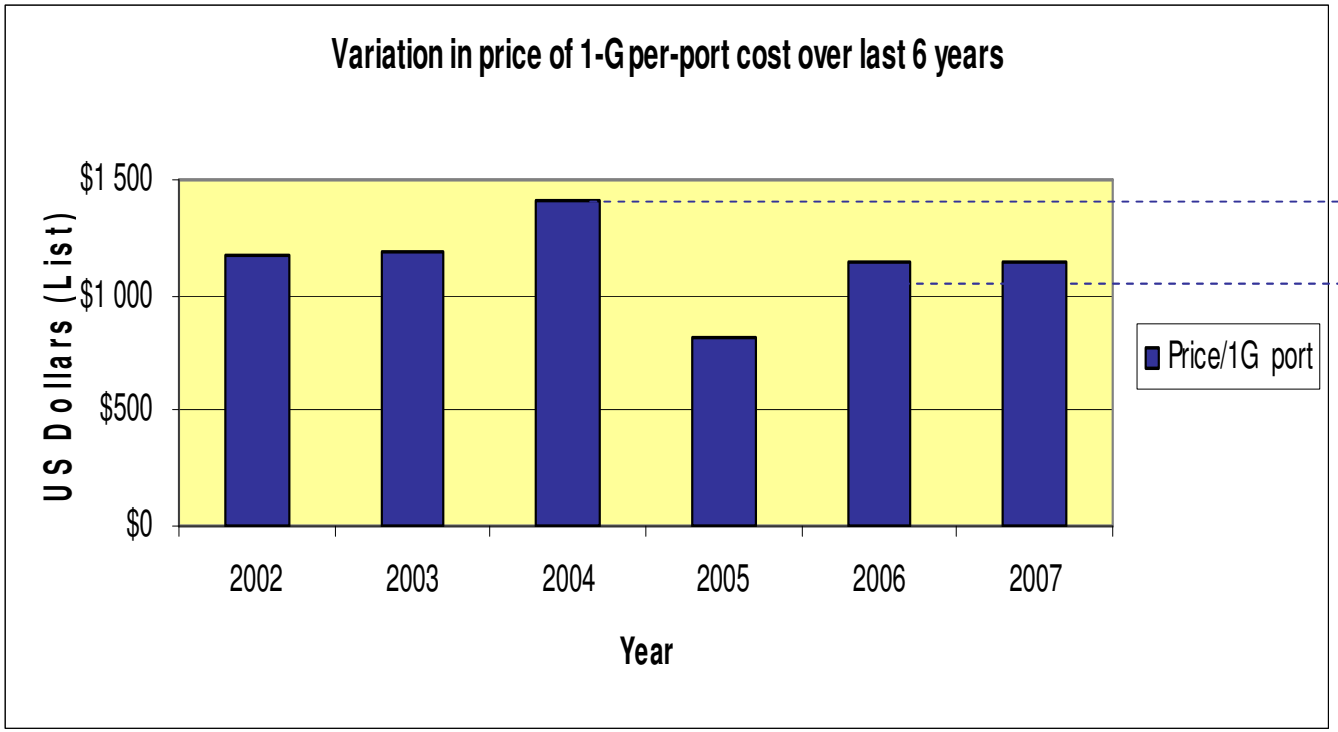
- Pour déterminer cela, 6 années de données collectées :
  - Prix par Gigabit en commutation IP
  - Prix par 10 Gigabit en commutation IP
  - Prix par module de supervision (Intéressant car il conserve la RIB)

- Même si les informations passées ne préjugent pas du futur, elles permettent tout de même de se faire une idée

- En partant des conditions suivantes :
  - Pour chaque année, l’équipement le plus dense disponible au premier trimestre est choisi (Même vendeur)
  - Equipement chargé au maximum et redondé (optiques exclues)

Quelle est la variation de prix ?

Prix par Gigabit en commutation IP / 6 ans

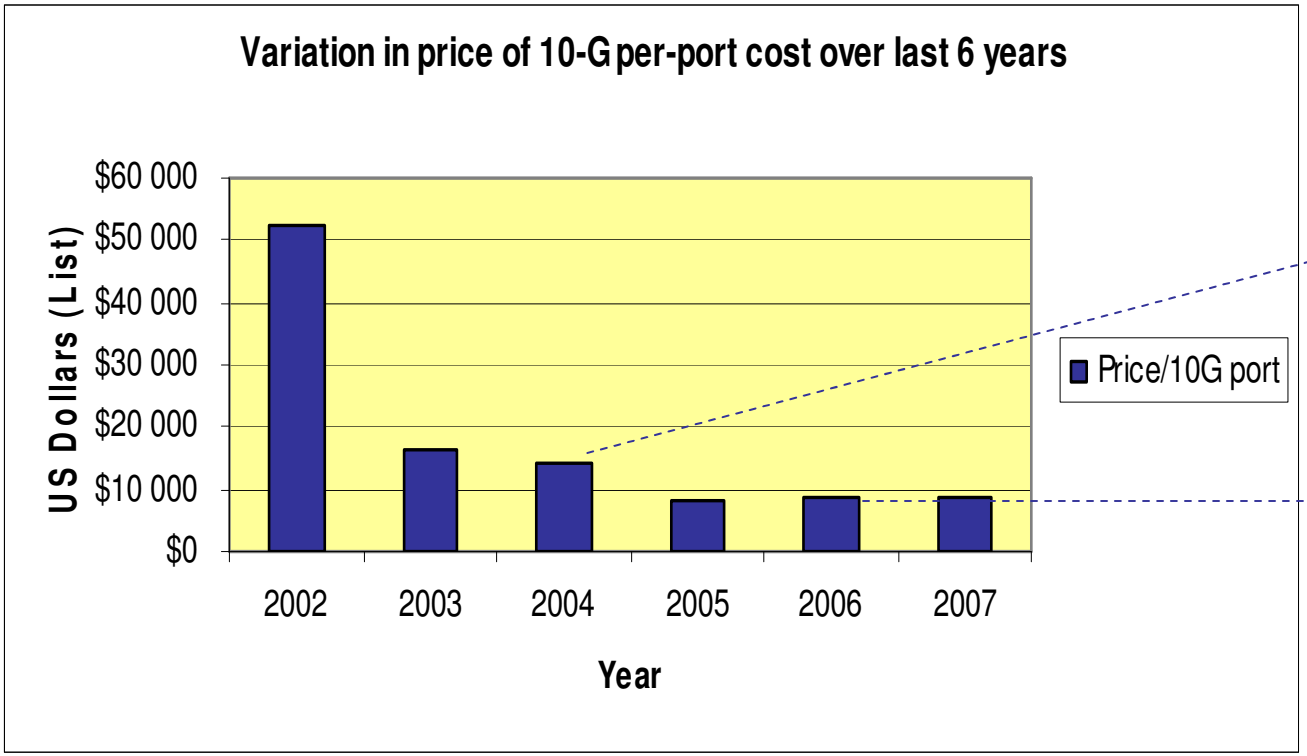


IPv6 en hardware

IPv4/v6/MPLS en hardware; 1M de routes IPv4 FIB; Quantité d'ACLs accrue

- Produit le plus dense disponible en Q1 de chaque année pour le même vendeur
- Châssis plein, redondant
- Optiques non incluses
- Données issues de Foundry Networks

Prix par 10 Gigabit en commutation IP / 6 ans

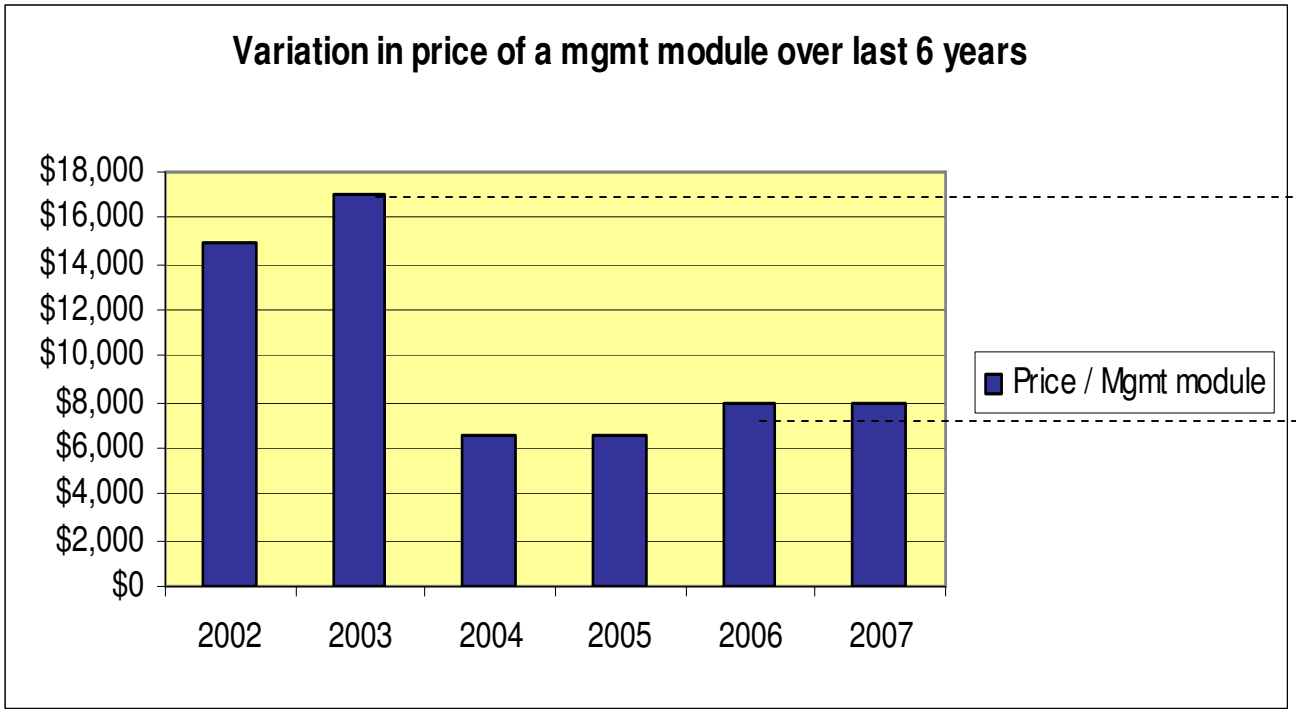


IPv6 en hardware

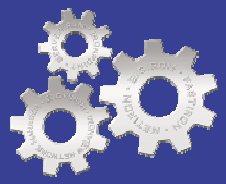
IPv4/v6/MPLS en hardware; 1M de routes IPv4 FIB; Quantité d'ACLs accrue

- Produit le plus dense disponible en Q1 de chaque année pour le même vendeur
- Châssis plein, redondant
- Optiques non incluses
- Données issues de Foundry Networks

Prix par module de supervision / 6ans



•Données issues de Foundry Networks



Conclusion

- La croissance du nombre de routes IPv4 est un problème qu'il faut :
  - Prendre en compte dès maintenant
  - Fixer au plus vite si nécessaire
    - Soit par l'investissement de nouveaux équipements,
    - Soit par une optimisation de la RIB et de la FIB.
- Mais, les technologies disponibles aujourd'hui permettent de s'affranchir de ce problème.
- Choisir un équipement supportant moins de 512K routes dans la FIB aujourd'hui serait un risque non négligeable
  - A moins d'avoir la volonté de travailler uniquement sur des vue partielles et/où des routes par défaut
- L'idée du coût des équipements de routage suivant la croissance de la table des routes IPv4 n'est pas aussi triviale que l'on pouvait le penser
  - Une économie substancielle sur les routeurs de Edge peut également réalisée

FOUNDRY®
NETWORKS

Questions ?

Sales : Antoine Gayon : +33 139 304 159/+33 662 399 066
agayon@foundrynet.com

SE : Laurent Gallampois : +33 139 304 154/+33 607 476 066
laurent@foundrynet.com

THE POWER OF PERFORMANCE™

FOUNDRY®
NETWORKS

Merci !

THE POWER OF PERFORMANCE™