Entrées-Sorties (I/O) & stockage

LIBERATE IT

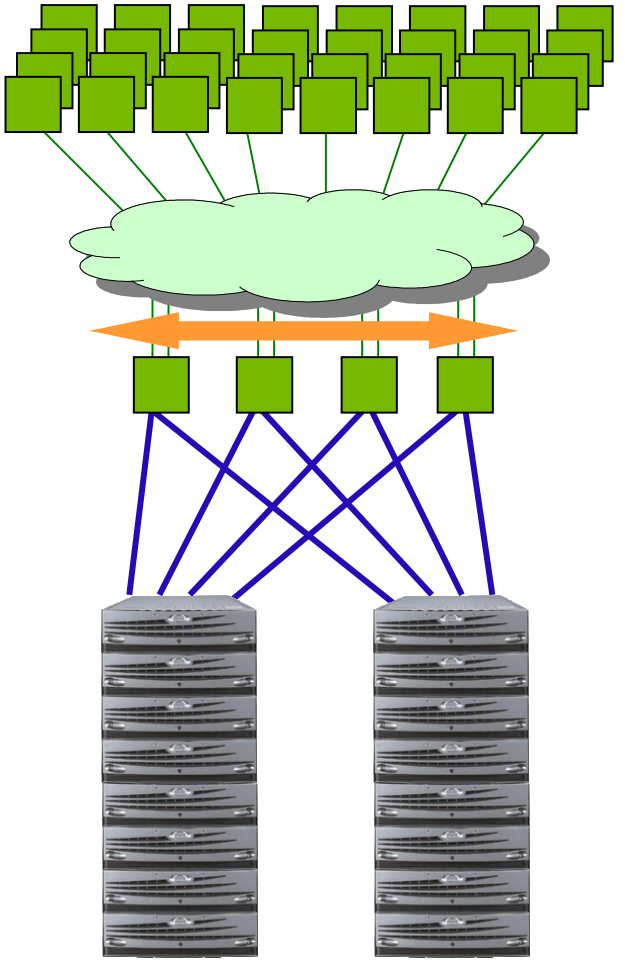
Entrées-sorties - Stockage

- Une problématique complexe
  - « multidisciplinaire »
    - Stockage (hardware)
      - → Technologies disques
        - SATA, SAS, FC, SSD ...
      - → Baies
    - Système de fichiers
      - →
      - → Partages
        - « réels »
        - Performants : Lustre, PanFS ...
        - Simple : NFS
        - Virtuels
    - Réseau d'interconnexion rapide (SAN)

C'est une solution complète !

Définir ces besoins

- Différents aspects pour définir les besoins
  - Volumétrie
    - Baies, nombre de disques
  - Performances
    - Quels débits ? MiB/sec, GiB/sec ?
      - → Nombre d'axes (corrélés à la volumétrie !) → maximiser le ration MB/sec par TB
      - → Interconnexion, serveurs d'I/Os
    - Le nombre d'I/Os pouvant être traités par secondes : IOPS (random, 4 KB).
    - Métada
  - Accès aux données : FS parallèle ?
  - Sécurité
    - Finalité
      - → Stockage long terme
      - → Stockage temporaire (Scratch)
    - Haute-disponibilité ?
  - Densité, consommation électrique ...



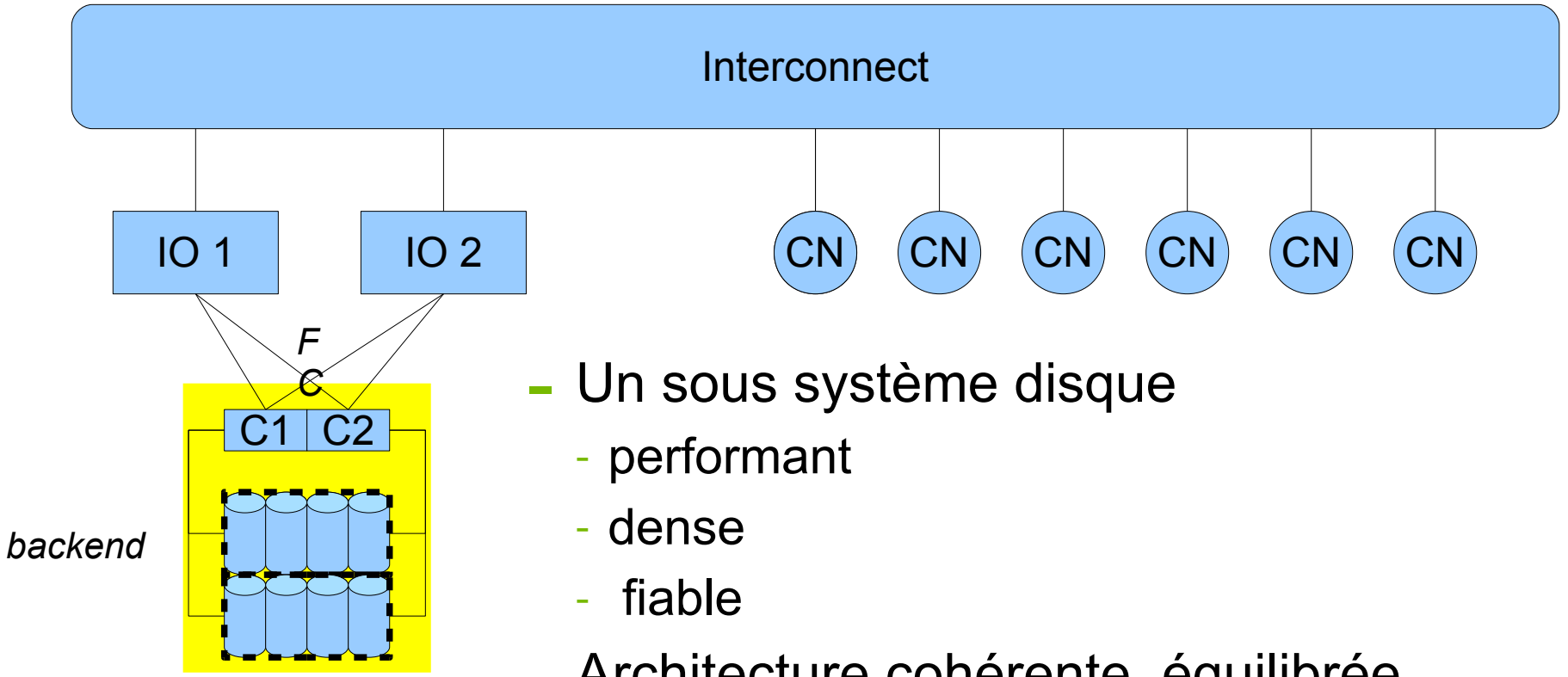
Définir ces besoins

- En matière d'I/Os, tout est possible :
  - De 30 MiB/sec en local, 20 GB/sec partagé, et au-delà ...
- Mais tout à un coût ...
  - De 50 € à plusieurs millions € ...
- Définir clairement ces besoins (production)
  - Habitude des utilisateurs
    - Quel pattern ?
      - → **Séquentiel** (HPC, streaming vidéo...)
      - → Random (base de données)
  - Bonne équilibre avec la taille de la machine
    - Enveloppe budgétaire constante : tout ce qui dans le stockage n'est pas en puissance de calculs ...
  - Pour définir ces besoins, encore faut-il les connaître :
    - Quels sont mes utilisations actuels ?
      - → **Benchmarks** !

Performances

- La performance en débit : multiples de Bytes/sec.
  - Attention aux puissances de 10 et aux puissances de 2 ...
  - Il s'entend par la bande passante global du système (en général différent de la bande passante depuis un seul nœud)
- On donne parfois aussi sa capacité de traitement en IOPS/sec
- Sur les gros systèmes, les métadata ont leur importance
- Le débit global est dépendant de toute la chaîne d'I/Os (baie, serveur, interconnexion (hard, soft), client ...
  - Il faut trouver le « bottleneck »
    - il y en a un c'est certain, et ce n'est pas inquiétant, c'est la vie !
    - Pour une même solution et suivant des conditions d'utilisation que l'on en fait, il n'est pas forcément au même endroit !
- L'ennemi de la mesure : **les effets de caches**
  - Et ils sont nombreux : baies, serveurs, clients ...

Un système équilibré et performant



- Un sous système disque
  - performant
  - dense
  - fiable
- Architecture cohérente, équilibrée
  - Liens FC, interconnect, serveurs d'I/Os (PCI-e) ...
- « Files system » et pile logicielle adaptés.

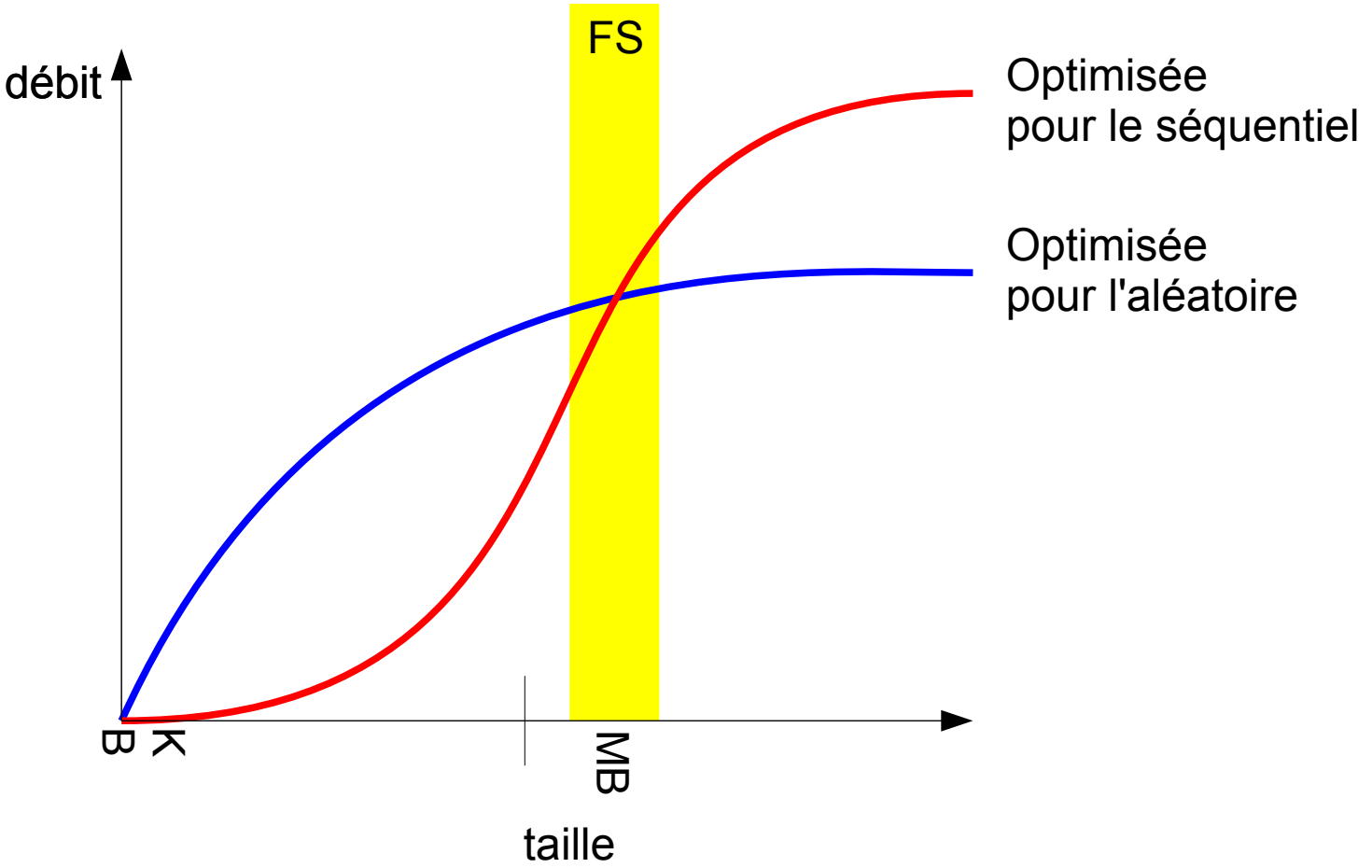
Le stockage : multicouche !

- Il faut être conscient de comment/où je fais mon bench d'I/Os
  - « dd », « iozone » ...
    - Sur le client : au dessus de la couche du FS et de l'interconnect
    - Sur le serveur : directement la couche « bloc » du FS et de la couche bloc de l'OS
  - Outils en mode « raw »
    - Permet de tester les capacités de la baie au plus proche...
    - Désavantage :
      → compliqué
      → destructif (en écriture tout au moins)

Performances

- La performance d'une baie dépends de plusieurs paramètres
  - Sa configuration matérielle (disques, backend, cache ...)
  - Sa configuration : découpage (agrégation, niveau de RAID, découpage en LUN), zoning
  - Le FS avec lequel elle sera utilisée
  - Le profil d'I/O
    - Leur pattern : séquentiel, aléatoire
    - Ratio Écriture/Lecture (1:2)
    - Leur taille
      - → La taille des I/O reçus n'est pas forcément celle que l'on émet : l'I/O traverse le FS !!!
    - Leur nombre (IOPS)

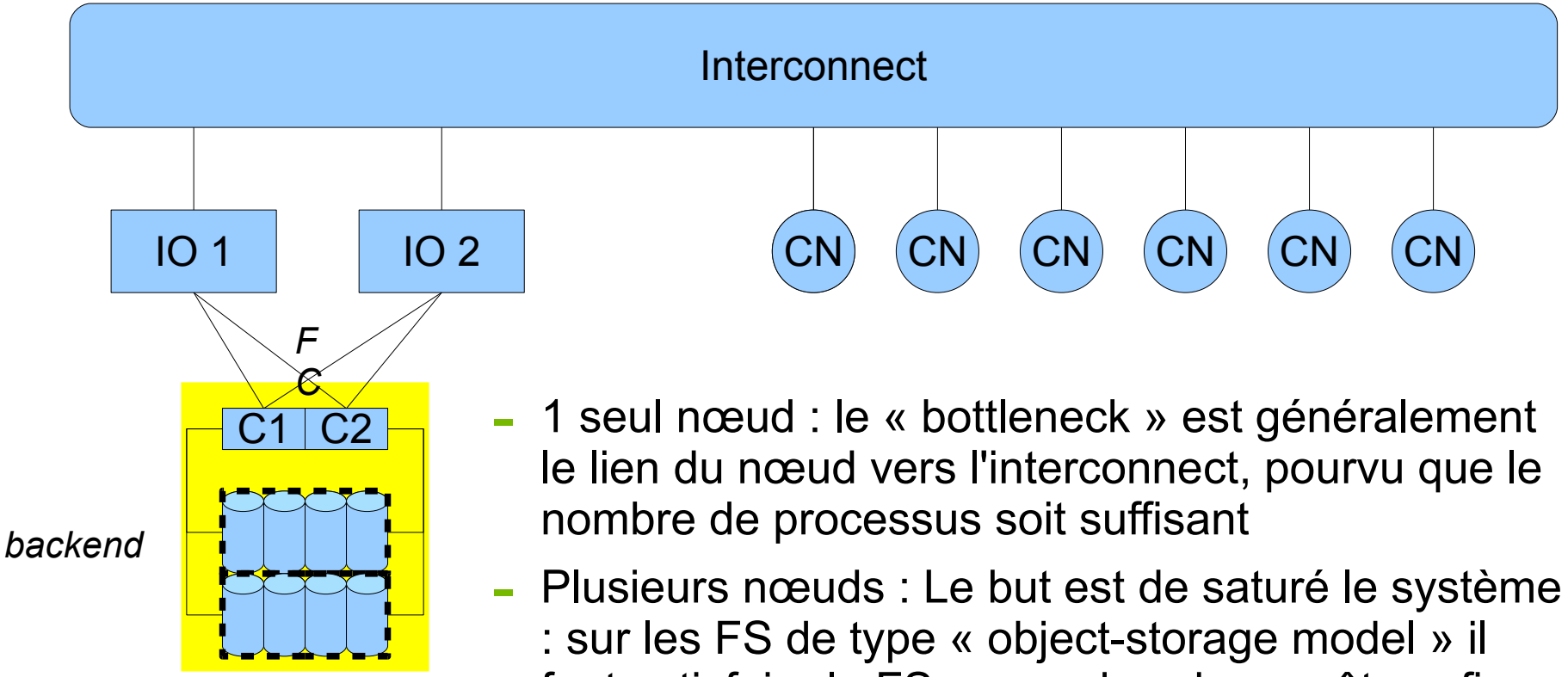
Performances



Benchmark d'I/Os

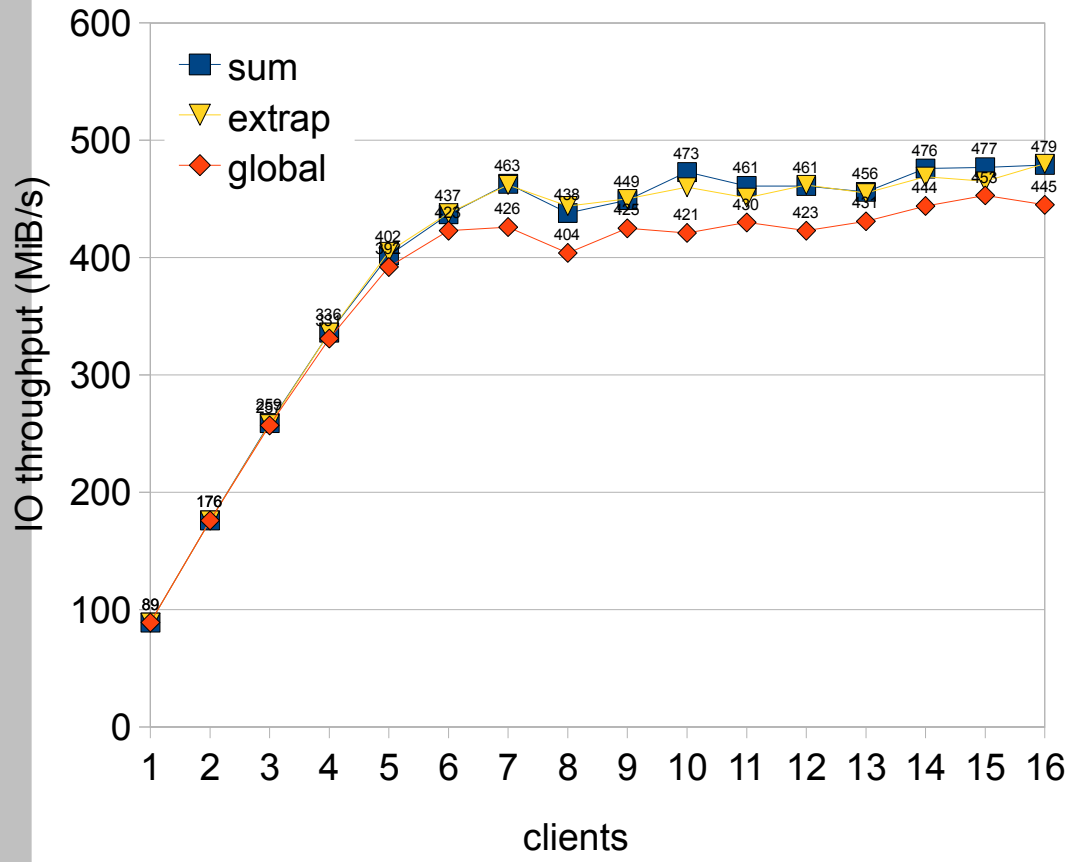
- Plusieurs benchmarks standards
  - Iozone, Bonnie++, IOR, Beff_io
- Benchmarks « maisons » simples, représentatif d'applications utilisateurs :
  - Benchmarks MPI avec
    - 1 noeud, 1 processus
    - N noeuds, 1 processus par noeud
    - N noeuds, M processus par noeud
  - Il donneront à la fois le comportement global de la solution (charge) ainsi le comportement des applications
- Les propriétés du FS sont importantes :
  - Sur un FS de type « object storage model », le « stripping » (nombre d'OST par fichier) est un paramètre primordial
- Il faut se méfier des effets caches : il faut être critiques vis-à-vis des chiffres

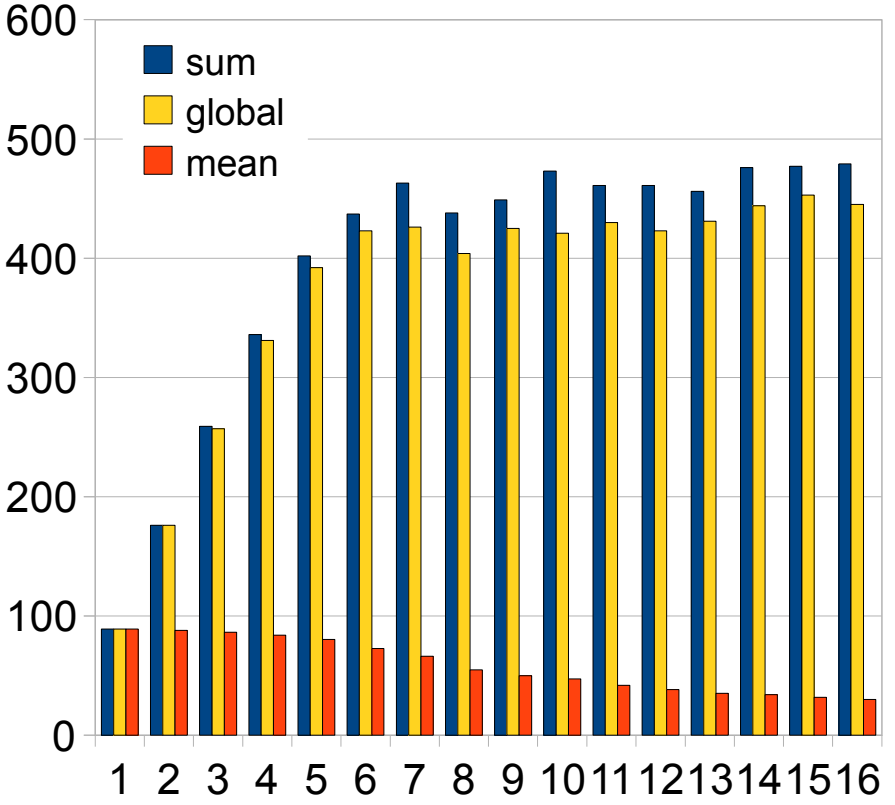
Bottleneck ; j'vous dis qu'y a truc qui bloque ...



- 1 seul nœud : le « bottleneck » est généralement le lien du nœud vers l'interconnect, pourvu que le nombre de processus soit suffisant
- Plusieurs nœuds : Le but est de saturé le système : sur les FS de type « object-storage model » il faut satisfaire le FS en nombre de requêtes afin de saturé le système et se mettre au limite des performances de la baie (si le FS le supporte)
- C'est la chasse au « bottleneck » !

Évolution avec le nombre de clients





- Appliance Panasas (2 tiroirs)
  - Interconnect Gigabits dédié
    - 1 noeud : le lien gigabit sature.
    - Au delà de 6-8 clients : le système d'I/O sature.
  -

Dédé n'est pas votre ami ... (si si...)

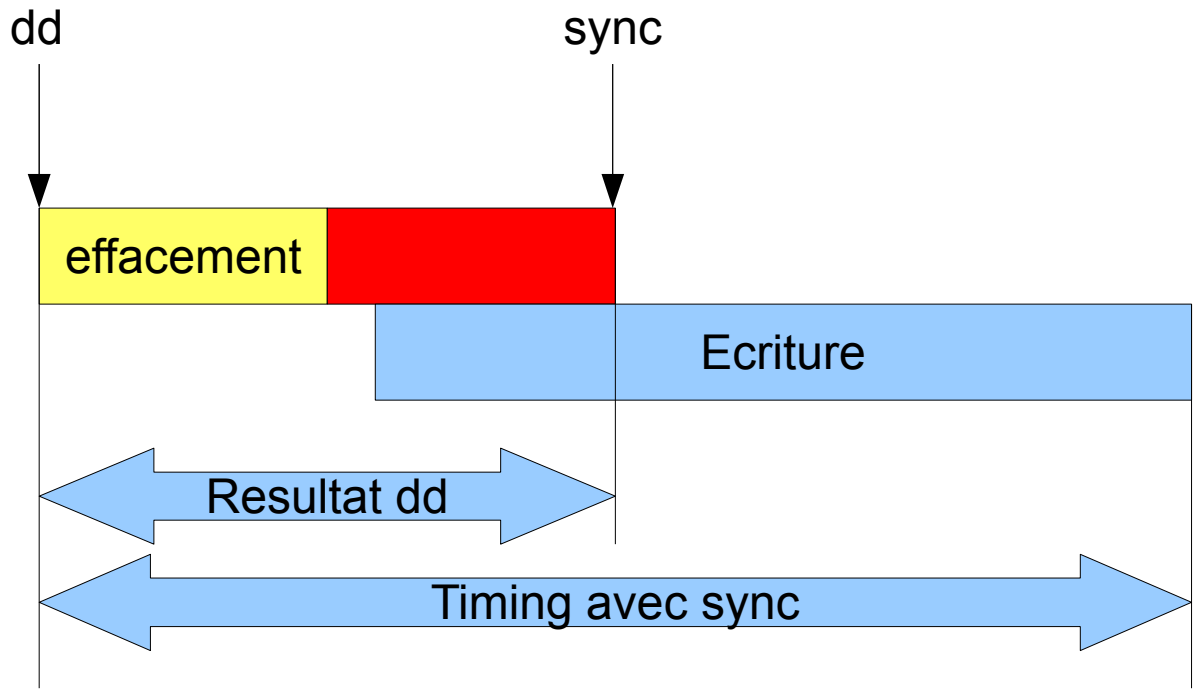
- L'erreur : le 'dd' – sur mon laptop

```
saugel@debian:/tmp$ dd if=/dev/zero of=/tmp/out bs=1M count=10
10+0 records in
10+0 records out
10485760 bytes (10 MB) copied, 0.0324364 s, 323 MB/s
```

Perf de mon disque, 323 MB/sec ... pas mal !!!

- Petite taille : tout est caché par l'OS
  - « dd » ne mesure pas le débit d'un disque mais le temps de remplissage du cache I/O de l'OS ...
  - Il suffit de faire tourner un iostat dans le même temps
    - Attention : ne garantit pas nom plus que les Ios sont sur le disque
  - Faire un sync
  - Faire une écriture directe (flag O_DIRECT)
  - Utilisé des tailles qui dépasse largement la capacité du cache.

Dédé n'est pas votre ami ... (si si...)

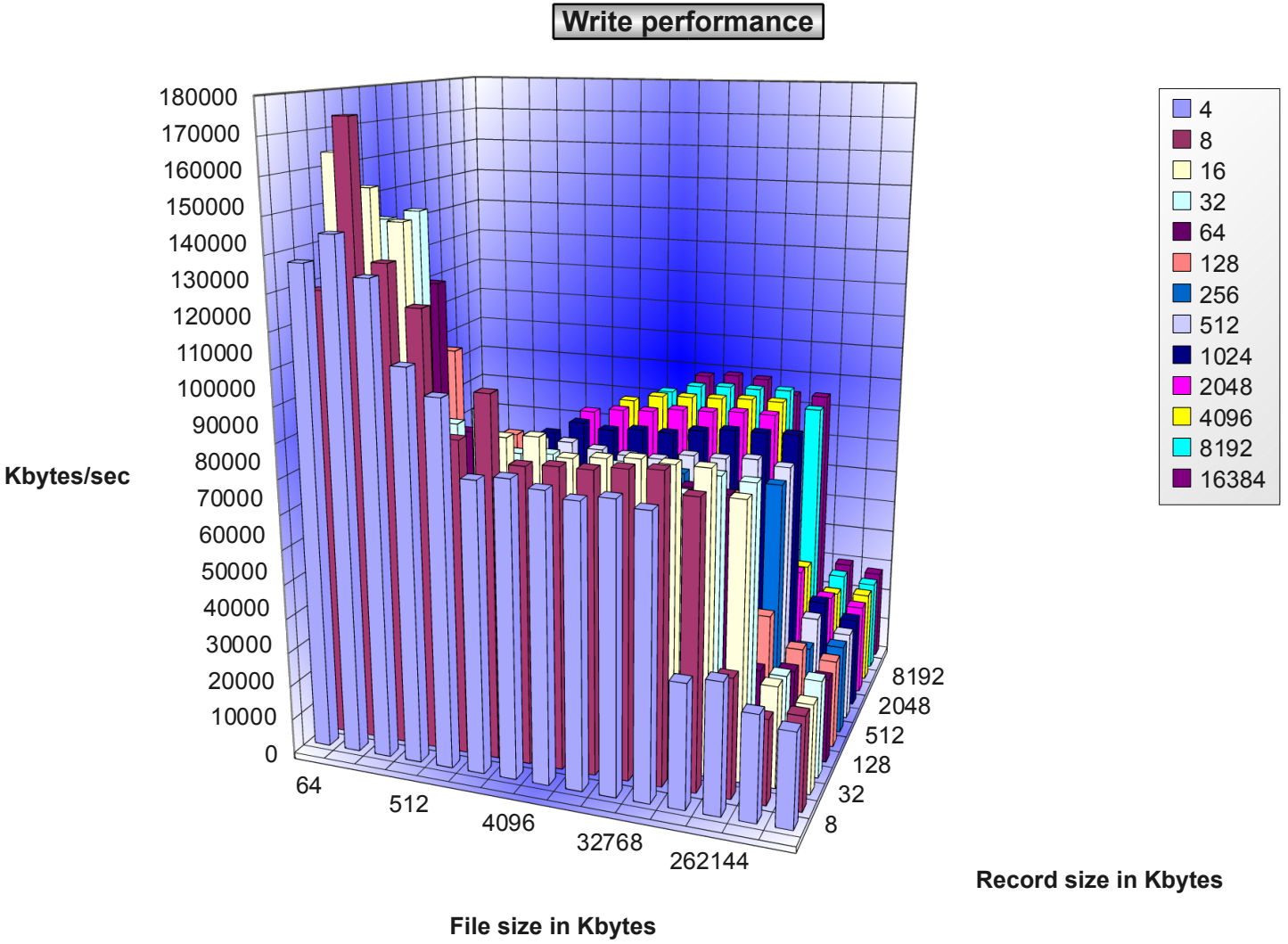


- « dd » n'est clairement pas un outil adapté
  - Difficile à comprendre, sensible à trop d'effets
  - Limité à un client

IOzone

- http://www.iozone.org/
- « Couteau suisse » du bench d'I/Os
  - Multiple pattern d'écriture : *write, read, rewrite,reread, random* ...
  - Multiple taille d'I/Os, multiple volume.
    - Le bench dans sa globalité peut permettre de déceler les effets de caches.
- Portage :
  - C, Posix – multhreading.
  - Simple (Makefile)
- Exécution :
  - De très nombreuse options ....

IOzone



Bonnie++

- http://www.coker.com.au/bonnie++/
- Ecriture séquentiele
  - Charactère (putc), bloc (write), re-écriture
- Lecture séquentielle
  - Charactère (getc), bloc (read)
- Pattern aléatoire : lecture et re-écriture dans 10% des cas.
- Portage:
  - C, Posix
  - Facile
- Execution
  - Facile

Beff_io

- Effective I/O Bandwidth (beff_io) Benchmark
- https://fs.hlrs.de/projects/par/mpi//b_eff_io/index_v1.1.html
- Benchmark orienté MPI-IO
  - Différentes opérations
  - Différents pattern
- Synthétique : nombreux tests, nombreuses données : un score !
- Portage : Facile
  - Nécessite une implémentation MPI-2
- Exécution : Modéré , analyse compliquée ...

IOR

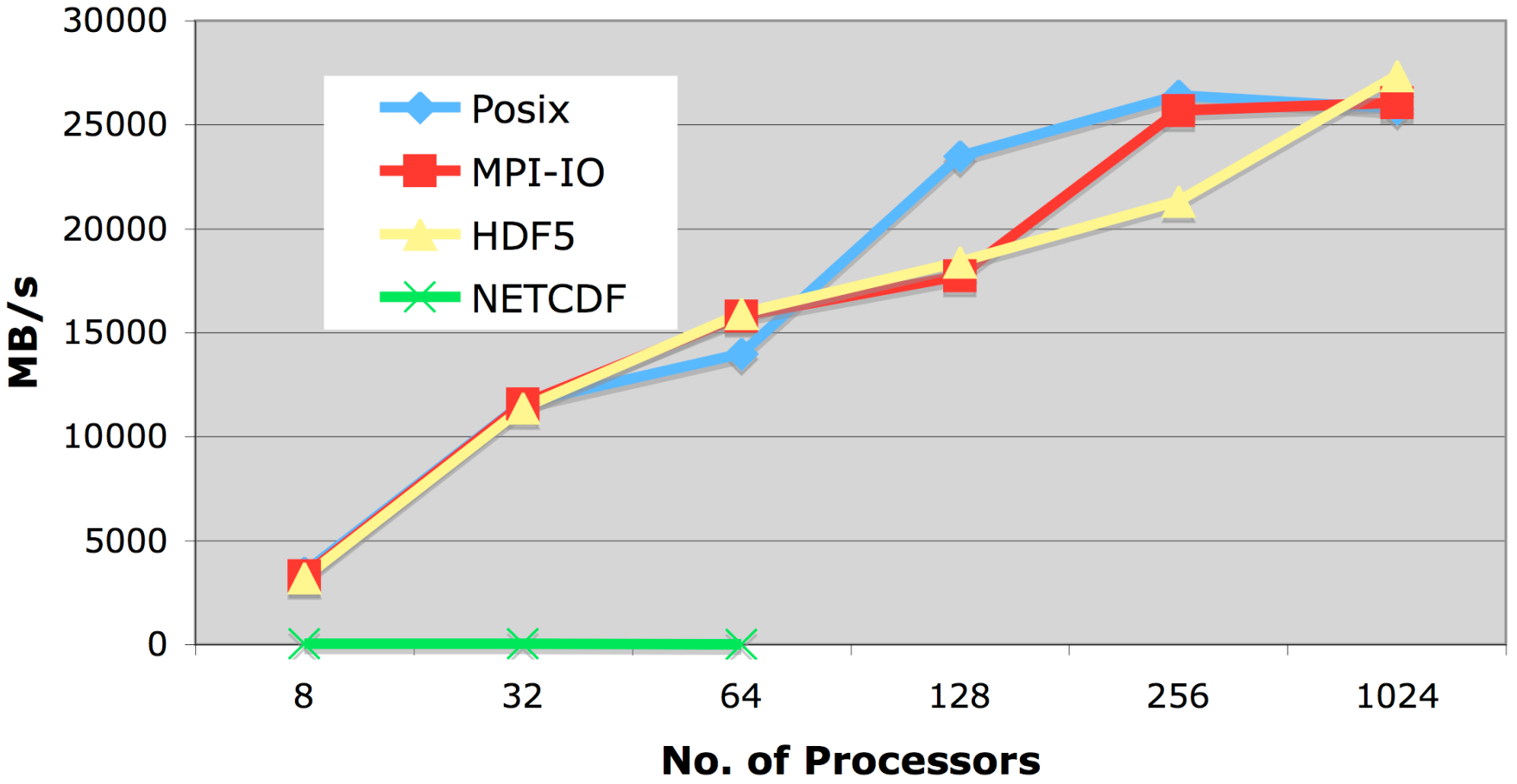
- Interleaved or Random (IOR) benchmarks
- http://sourceforge.net/projects/ior-sio/.
- Ce veut proche des applications
  - Les benchmarks synthétiques précédents ne sont pas représentatifs.
    - Pas parallèles, pas adaptés au HPC (I/O séquentielles)
  - Permet de tester les APIs (parallèles) couramment utilisés par les utilisateurs
    - MPI-IO, parallelNetCDF, pHDF5, posix
    - Par défaut MPI-IO

IOR

- Portage : Modéré
  - C, Posix
  - Requiert une implémentation MPI
  - Dépendances vers vers certaines bibliothèques suivant l'API que l'on souhaite tester
- Exécution : Simple
- Pour ce benchmark, MPI-IO n'est pas adapté à la mesure d'un débit maximale d'une baie. Par contre il parfaitement adapté pour avoir des idées de performances d'applicatifs.

IOR

- Jaguar : 18 DDN 9550 + Lustre (36 GiB/sec) – Noeud : max 1 GB/sec



Conclusion sur les I/O et le stockage

- On ne benchmark pas seulement une baie, mais une solution de stockage pour le HPC
- Il faut une solution équilibrer
- La métrique de base est la bande passante globale (MiB/sec ou GiB/sec)
  - Elle s'entend par la bande passante disponible à partir de plusieurs nœuds
  - Dans le cas contraire, il faut préciser la bande passante désirée par nœud

- Voir aussi la page « Parallel I/O Examples and Benchmark Codes »

http://www.cs.dartmouth.edu/pario/examples.html