開
放
的
열린
مفتوح
libre
मुक्त
ముక్త
livre
libero
ಮುಕ್ತ
开放的
açık
open
nyílt
פתוח
オープン
livre
ανοικτό
offen
otevřený
öppen
открытый
வெளிப்படை

open

Utiliser Améliorer Prêcher

Métrologie des IO

guses.org

Sommaire

- Les éléments de la chaîne d'IO
- Les indicateurs, comment les décoder
- Quelques réglages
- Outils de bench

Les systèmes de fichiers

- snapshot
- extend : bloc de taille variable
- journalisation
  - crash recovery
  - écriture séquentielle des méta données
- zfs : oubliez tout ce que vous savez (coming soon)

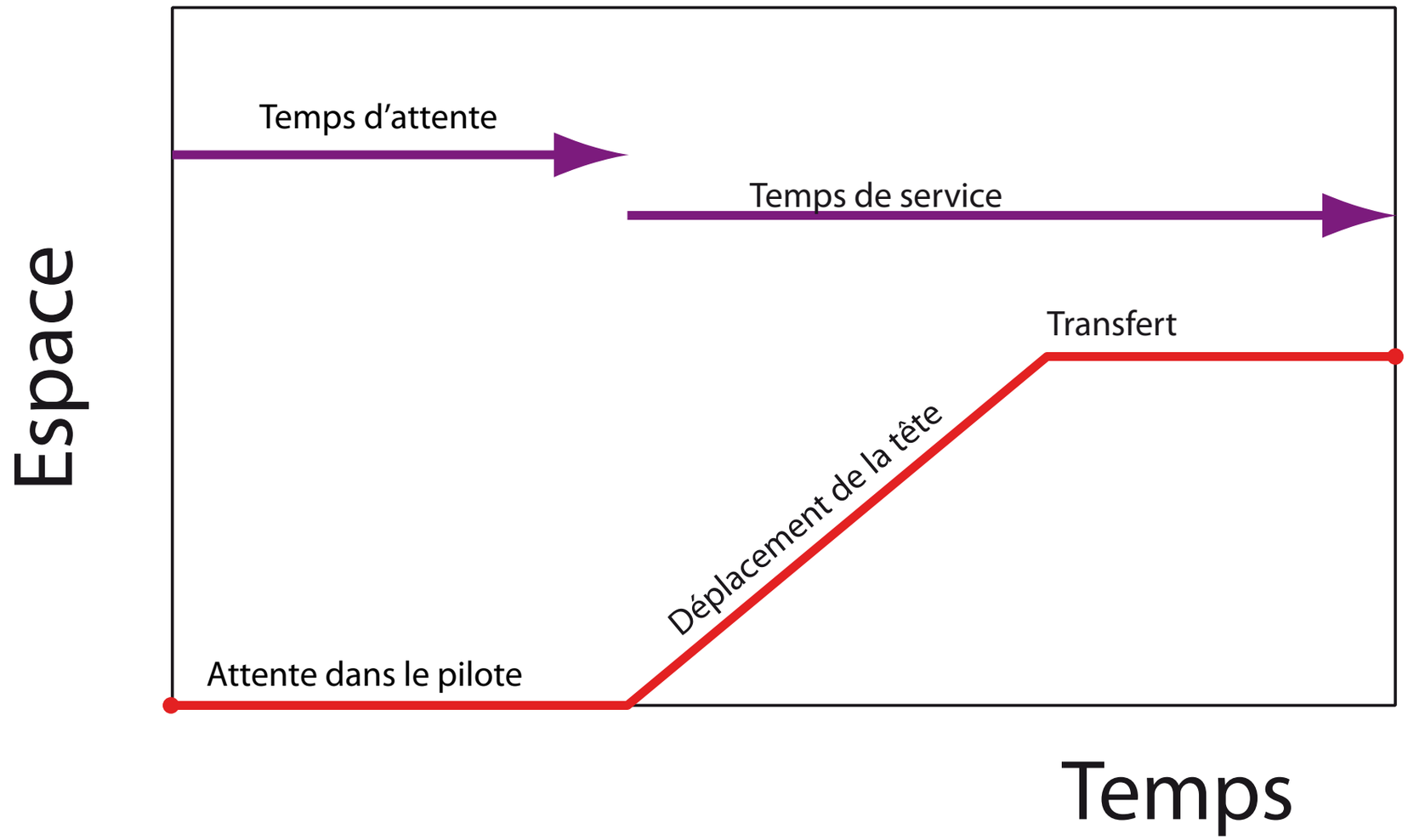
http://www.solarisinternals.com/si/reading/sunworldonline/swol-05-1999/swol-05-filesystem.html

Les caches

- de nom (dnlc)
  - Parcours du VFS (Virtual FileSystem)
- de meta données (buffer cache)
  - par exemple, numéro d'inodes
- de données en lecture
  - permet de lire en avance (prefetch) les lectures, réduire le nombre de petites IO
- de données en écriture
  - permet d'effectuer les écritures en taches de fond, de les regrouper

Éléments de base d'une IO

- Déplacement de la tête
- Vitesse de transfert



L'impact du matériel

- Traitement parallèle : TCQ, NCQ
- Nombre d'axes (RAID)
- Plusieurs niveau de cache en écriture : dans le contrôleur, dans le disque
- Sauvegardé ou non sur batterie, important pour l'acquittement des écritures
- L'algorithme de l'ascenseur

開
放
的
열린
مفتوح
libre
मुक्त
ముక్త
livre
libero
ముక్త
开放的
açık
open
nyílt
פתוח
オープン
livre
ανοικτό
offen
otevřený
öppen
открытый
வெளிப்படை

guses.org

Les indicateurs

Les indicateurs

- Essentiel pour identifier les goulots d'étranglement
- Ne sont pas les mêmes selon le type d'usage :
  - écriture/lecture
  - séquentiel/aléatoire
- Certains sont anciens et n'ont plus de sens, notamment Solaris 10

```
% man mpstat
                    the I/O wait time is no longer calculated as
                    a percentage of CPU time, and this statistic
                    will always return zero
```

IOSTAT

```
% iostat -xnM 10
                    extended device statistics

                    extended device statistics
    r/s    w/s   Mr/s   Mw/s wait actv wsvc_t asvc_t  %w  %b device
  812.4    3.0    1.8    0.6  0.0  0.9    0.0    1.1   0  85 c0t0d0
```

- L'outil de base, le premier à regarder
- usage typique : iostat -xnM 10
- ignorer la première ligne
- r/s et w/s : opération
- Mr/s et Mw/s : débit
- wait : IO en attente dans le pilote
- actv : IO dans le matériel
- wsvc_t : temps moyen d'attente
- asvc_t : temps moyen de service
- %w : % du temps avec des IO en attentes
- %b : % du temps où le disque travail

sar

-a Reports use of file access system routines
-b Reports buffer activity
-d Reports activity for each block device

Dans /var/spool/cron/crontabs/sys
```
* * * * * /usr/lib/sa/sa1
```

Ou « démon »
```
sadc 60 1440 sa16
```

http://ksar.atomique.net/

dtrace et DTraceToolkit

- opensnoop

```
UID    PID COMM          FD PATH
  0  22072 sh            -1 /var/ld/ld.config
  0  22072 sh             3 /lib/libc.so.1
  0    565 sshd           8 /system/contract/process/latest
  0    565 sshd           8 /system/contract/all/16760/ctl
  0  22071 sshd          -1 /etc/hosts.allow
  0  22071 sshd          -1 /etc/hosts.deny
```

- iosnoop

```
UID   PID D    BLOCK   SIZE       COMM PATHNAME
    0 22628 R   779938   3072     bash /usr/xpg4/bin/find
  0 22628 R     1552   8192     find /lost+found
  0 22628 R     1546   1024     find /var/sadm/install
  0 22628 R     1548   1024     find /var/sadm/install/admin
  0 22628 R     1550   1024     find /var/sadm/install/logs
  0 22628 R     1574   1024     find /var/sadm/pkg/SUNWocfd
```

- iotop

```
2009 Jun 16 00:19:31,  load: 0.22,  disk_r:   2411 KB,  disk_w: 240421 KB

  UID    PID   PPID CMD              DEVICE  MAJ MIN D            BYTES
    0    427      1 inetd            sd0      30   0 R             1024
10202  19566      1 java             sd0      30   0 R             2048
    0  22654    427 rpc.rstatd       sd0      30   0 R             3072
    0  22652  22060 bash             sd0      30   0 R            19456
    0  22654    427 inetd            sd0      30   0 R            20480
    0      3      0 fsflush          sd0      30   0 W           180736
    0  22653  22060 cpio             sd0      30   0 W           270336
10202  19566      1 java             sd2      30 128 R           655360
    0  22652  22060 find             sd0      30   0 R           978944
    0      0      0 sched            sd2      30 128 W        297644544
```

http://www.opensolaris.org/os/community/dtrace/dtracetoolkit/

開
放
的
열린
مفتوح
libre
मुक्त
ముక్త
livre
libero
ముక్త
开放的
açık
open
nyílt
פתוח
オープン
livre
ανοικτό
offen
otevřený
öppen
открытый
வெளிப்படை

guses.org

Les réglages

DirectIO

- Pas de segmentation des grandes écritures synchrones
- Suppression du verrou POSIX (single writer lock)
- Pas de double transfert
- Simplification du code
- Utilisation plus efficace de la mémoire

- mount -o forcedirectio
- man directio.3c

http://www.solarisinternals.com/wiki/index.php/Direct_I/O

atime

- mount -o dfratime
- mount -o noatime

/etc/system

- *ufs_HW*, *ufs_LW* volume de dirty
- *maxphys* taille de blocs
- *maxcontig*

http://www.solarisinternals.com/si/reading/fs2/fs2.html
http://docs.sun.com/app/docs/doc/817-0404

開
放
的
열린
مفتوح
libre
मुक्त
ముక్త
livre
libero
ముక్త
开放的
açık
open
nyílt
פתוח
オープン
livre
ανοικτό
offen
otevřený
öppen
открытый
வெளிப்படை

guses.org

Les outils de bench

Ne pas se tromper d'outils

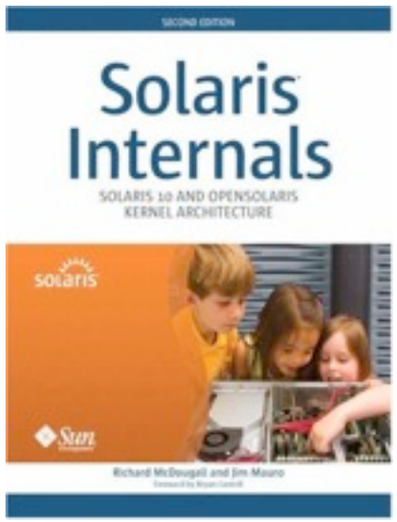
- les IO sont ce qu'ils y a de plus difficile à bencher
- Mesure de latence/débit/nombre d'IO
- Les IO sont très parallèles

Programme

- Micro benchmark
  - dd
  - tar
  - find
- Tests synthétique
  - bonnie++
  - iozone
- Test par simulation
  - filebench
  - fio

En savoir plus

- Solaris Internals second edition
- http://www.c0t0d0s0.org/
- http://www.solarisinternals.com/
- Solaris Tuning Guide



開
放
的
열린
مفتوح
libre
मुक्त
ముక్త
livre
libero
ముక్త
开放的
açık
open
nyílt
פתוח
オープン
livre
ανοικτό
offen
otevřený
öppen
открытый
வெளிப்படை

guses.org

Merci !

“open” artwork and icons by chandan:
http://blogs.sun.com/chandan

Logo GUSES par Philipe Destigny:
mailto:philippe.destigny@free.fr