Arrêts systèmes et Dumps

Gérard MADIOT / Christophe ROMAC

Introduction

Dans un certain nombre de cas, le système n'est plus apte à assurer un fonctionnement normal. Cela peut provenir de problèmes matériels, logiciels liés au noyau, de blocages à cause de contention de ressources, ou de boucle infinie dans un code privilégié.

Ces cas sont, entre autres :

- Une interruption inattendue.
- Le noyau qui reçoit une instruction lui demandant de s'arrêter.

Lorsque cela se produit, le noyau appelle des routines, dites de "*dump*", qui vont copier un certain nombre d'éléments de mémoire (mémoire réelle uniquement) dans un fichier, le "*dump logical volume*", qui seront nécessaires à la détermination du problème.

Différents cas d'arrêts du système

Il existe différents cas qui provoquent l'arrêt du système.

Systèmes avec afficheur

Sur les systèmes pourvus d'un afficheur, le code d'arrêt apparaît sous la forme de 888 clignotant.
Le fait d'appuyer plusieurs fois sur le bouton "*reset*", jusqu'au retour du 888 clignotant, permet d'afficher les codes complémentaires qui indiquent la raison de l'arrêt .

888 102 300 0cx

Indique un "*Data Storage Interrupt*" (DSI).
Le système essaie de charger une donnée mais n'y parvient pas pour diverses raisons :

- Faute de page pendant le traitement d'une interruption.
- Lecture ou écriture à une adresse mémoire qui n'est pas présente en mémoire réelle (*mapped*).
  Exemple : mauvais registre ou segment.
- Erreur d'entrée-sortie lors d'une écriture directe de données.

888 102 400 0cx

Indique un "*Instruction Storage Interrupt*" (ISI).
Cette erreur est similaire au DSI, à ceci près que le système essaie de charger une instruction et non une donnée.

888 102 700 0cx

Program Interrupt.
Le système exécute une instruction de type "*trap interrupt*", c'est-à-dire programmée dans le code par une macro-instruction "***assert***" ou un "appel système" (*System Call*) "***panic()***".

888 102 800 (*)

Erreur pendant l'exécution d'une instruction en virgule flottante.

888 102 5xx (*)

Erreur due à un "*external interrupt*".

(*) Note : Ces deux derniers cas d'erreur sont extrêmement rares.

888 102 207

Ce type d'erreur est typiquement accompagné d'un "*machine_check*" dans la *log* d'erreurs.
Ce code indique un problème matériel (*hardware*).

888 103 XXX YYY

Erreur de type matériel (*hardware*).
XXX YYY est un code SRN (*Service Request Number*), c'est-à-dire la référence de la pièce détectée comme étant défectueuse.

Les codes d'état du dump 0cx

- **0c0** Indique que l'écriture du *dump* s'est correctement terminée.
- **0c1** Erreur d'E/S durant l'écriture du *dump*. Non exploitable, cela relève du matériel.
- **0c2** Ecriture du *dump* en cours (forcée par l'utilisateur par "***sysdumpstart***" ou bouton "***reset***").
- **0c4** Indique que l'unité de *dump* est trop petite. Le *dump* est néanmoins EXPLOITABLE.
- **0c5** Erreur des routines de *dump*.
- **0c8** Indique que l'unité de *dump* n'est pas définie.
- **0c9** Ecriture du *dump* en cours (déclenchée par le système).

Hang ou loop

Dans ce cas, le système est inopérant sans code d'erreur affiché.
Plus aucune opération n'est possible sur la machine, aussi bien en "*local*" qu'en "*remote*".
Cela est dû, généralement :

- à un verrouillage (*lock*) fait par un processus et qui n'est jamais relâché,
- ou à une boucle dans une routine interne du noyau.

Sysdumpstart

La commande "***sysdumpstart***" permet de générer un "*system dump*".

Cette commande, peu usitée, permet de vérifier que la prise de *dump* se fait correctement et que tout est bien configuré. Elle permet également, lancée à la demande des laboratoires,

d'obtenir l'image mémoire du système dans des cas très particuliers.

Systèmes sans afficheur

Sur les machines n'ayant pas d'afficheur, on peut néanmoins connaître le code d'arrêt en consultant la *log* d'erreurs pour y trouver une entrée : "SYSDUMP_SYMP". Pour cela, il faut utiliser la commande d'affichage détaillé de la *log* : ***errpt -a***

Configuration d'une unité de dump

Unités de dump principale et secondaire

Lors d'un arrêt du système, l'image de la mémoire (*dump*) est écrite sur l'unité définie comme unité principale de *dump*.
En cas de problème d'écriture, les routines de *dump* essaient d'écrire sur l'unité secondaire.

L'unité principale de *dump* est, par défaut, le "*paging space*" (hd6), sauf sur les noeuds SP2 (voir, plus loin, le paragraphe "Spécificités SP2").
En effet, suite à un blocage, le volume logique dédié à la pagination ne sert plus en tant que tel.

L'unité de *dump* secondaire peut être utilisée lorsque l'utilisateur le spécifie dans la commande "***sysdumpstart***".

Par défaut

Après installation de l'AIX, l'unité de *dump* par défaut est le *paging space*.
Un *dump* copié sur le *paging space* doit être recopié ailleurs lors du *reboot*.

Par défaut, le système essaye de copier le *dump* sous "***/var***" dans le répertoire "***/var/adm/ras***" pendant le processus de *boot*.
Dans ce cas, le fichier de l'image mémoire s'appellera "***vmcore.0***" pour le premier, "***vmcore.1***" pour le second, etc.
Si "***/var***" n'est pas assez grand pour acceuillir le *dump*, le système demande à l'utilisateur d'insérer un média pour y copier le *dump*. Vu la taille d'un *dump* il est inutile d'essayer de le mettre sur disquette. Seule une bande s'avère être la bonne solution.

La taille approximative d'un *dump* peut être obtenue en cours de fonctionnement normal par la commande : ***sysdumpdev -e***

```
$sysdumpdev -e
0453-041 Estimation de la taille du cliché (en octets) : 39845888
```

Pour toutes les versions antérieures à l'AIX V4.3.3, les *dumps* copiés sur le *paging space* ne peuvent pas être analysés si celui-ci est en mode *mirroring*.
Dans le cas où l'unité de pagination est en mode *mirroring* il faut dédier une unité de *dump* spécifique.

Création d'une unité de dump spécifique

- Il faut créer un volume logique de type SYSDUMP dont la taille sera définie par la formule :
  taille LV = sysdumpdev -e + 10%

- Il faut ensuite définir ce "*dump logical volume*" comme unité de *dump* principale par la commande :

```
smitty dump
        ==> changement d'unité de cliché principale
```

Paramètres modifiables

- Pour vérifier la configuration actuelle de l'unité de *dump*, utiliser la commande : sysdumpdev -l

```
$sysdumpdev -l
        principal               /dev/hd7
        secondaire              /dev/sysdumpnull
        Répertoire de copie     /var/adm/ras
        option de copie imposée TRUE
        cliché toujours autorisé TRUE
```

- On peut alors apporter les modifications suivantes :

  - Répertoire de copie :
    On peut affecter un autre répertoire où le *dump* sera copié si "hd6" est l'unité de *dump*.
    Le répertoire spécifié n'est pas utilisé si un "*dump logical volume*" est dédié.

  - Cliché toujours autorisé :
    Cette option permet de définir si un *dump* peut être forcé, ou non, par l'utilisateur en appuyant sur le bouton "*reset*" de la machine.
    ==> Ce paramètre doit être forcé sur "**VRAI**" pour pouvoir obtenir un *dump* lorsque le système est bloqué ou en boucle.

Récupération des éléments

Une fois le *dump* copié, que ce soit via le *paging space*, par un "*dump logical volume*" dédié, ou que ce soit sur bande lors du *reboot*, la récupération de tous les éléments indispensables est la même.

Procédure

Passer les commandes :

- snap -a
  Cette commande va créer, sous "*/tmp*", un répertoire "*/tmp/ibmsupt*" où tous les éléments seront copiés.
- snap -o /dev/rmtx
  Cette commande copie ensuite les éléments sur la bande.

Notes :

- Si le *dump* a été copié directement sur bande lors du *reboot*, il va de soi qu'il ne faut pas passer la commande "*snap -o*" avec la même bande sous peine d'écraser le *dump*.
- S'il n'y a pas assez de place sous "*/tmp*", la commande "*snap -a*" ne se terminera pas et le signalera. Dans ce cas, il y a deux possibilités :
  - Soit agrandir "*/tmp*"
  - Soit créer un *File System* temporaire avec un point de montage en définissant IMPERATIVEMENT : "***/tmp/ibmsupt***".

Spécificités SP2

- Pour obtenir les codes d'arrêt, utiliser la commande "***reset***" de SPMON ou perspectives.
- Les nodes d'un SP2 ne comportent pas de lecteur de bande.
  C'est pourquoi, l'unité principale de *dump* est définie par défaut sur un volume logique dédié, hd7 ou lv00, suivant la version du PSSP.
  Il n'y a donc pas à se préoccupper de vérifier si la taille de "*/var*" est suffisante pour contenir un fichier "*vmcore.x*".
- La commande "***snap***" doit toujours être utilisée pour collecter les éléments nécessaires à l'étude du *dump*.
  - La méthode recommandée consiste à passer les commandes :
    - snap -a
    - cd /tmp/ibmsupt
    - tar -cvf PMR#.tar *
      (où PMR# est le numéro d'incident ouvert auprès du Point Service AIX)
    - ftp de "PMR#.tar *" sur la CWS ou une machine disposant d'un lecteur de bande
    - tar -cvf /dev/rmtx PMR#.tar

  - Une autre méthode consiste à faire un montage NFS de "*/tmp/ibmsupt*" sur la CWS puis de passer la commande "*snap*".

Erreur due à un problème matériel

Code 888 102 207

- Il faut analyser la log d'erreurs pour des entrées de type: "*machine_check*", "*bus_error*", "*scan_err_chrp*".

Code 888 102 300 0c4

- La taille du *dump* est de 0 *byte* (on peut le vérifier par la commande : "*sysdumpdev -L*").
  Dans la *log* d'erreurs, rechercher la valeur d'**EXVAL** dans l'entrée DSI_PROC.
  Si cette valeur est "**0000 0005**", cela signifie qu'il y a eu une erreur d'E/S en accédant à le *paging space*. Ceci démontre que c'est un problème matériel, lié au disque contenant le *paging space* ou à sa carte de contrôle, au microcode, etc.
  ==> Faire appel au service de support matériel.

Code 888 102 700

- Ce code, indiquant généralement une erreur logicielle, peut indiquer une erreur matérielle dans le cas où le processeur a reçu une instruction à exécuter qu'il estime invalide.
  Pour le savoir, analyser la *log* d'erreurs à l'entrée PROGRAM_INT, et noter le contenu du *Machine State Save Register 1* (MSSR1) :
  - Si sa valeur est de "0008 0000", il s'agit d'un problème matériel.
  - Si sa valeur est de "0002 0000", il s'agit d'un problème logiciel correspondant à une instruction de type "*trap interrupt*".

Code 888 103 XXX YYY

- Les codes XXX YYY indiquent le SRN (*Service Request Number*) dont la signification est donnée dans la documentation "*Diagnostic Information*" livrée avec le système MCA ou PCI. Ce type de code indique systématiquement un problème matériel.

MODS et bosdebug

MODS est l'acronyme de *Memory Overlay Detection System*.

Un certain nombre d'erreurs ou de blocages ont pour origine la gestion de la mémoire : son allocation, son accès ou sa libération. Par exemple, dans le cas de la mémoire *kernel* :

- Fragmentation de la mémoire sur des systèmes dont les accès réseaux sont importants.
- Pointeurs situés encore dans le cache déréférencé explicitement par une extension du *kernel*.
- Appel de routines de gestion de la mémoire avec de mauvais paramètres.
- Ecriture au-delà de la zone de mémoire allouée.
- Tentative d'utilisation d'une zone de mémoire non allouée.

Jusqu'à la version 4.1 d'AIX, il était nécessaire de demander aux laboratoires un noyau spécial, dit de "*debugging*", incluant des routines de vérification d'allocation mémoire.
Depuis la V4.2, ces routines sont intégrées au *kernel* et peuvent être activées ou désactivées par la commande "***bosdebug***" :

- Activation

```
bosdebug -M
bosboot -a
shutdown -Fr
```

- Désactivation

```
bosdebug -o
bosboot -a
shutdown -Fr
```

La commande "***bosdebug -L***" permet de connaître le statut de la fonction MODS :

```
$bosdebug -L

Débogueur de mémoire        on
Tailles de mémoire          0
Tailles de mémoire réseau   0
Débogueur de noyau          off
```

Rappels

- On peut faire un *mirroring* du *LV dump* à partir de la version 4.3.3 d'AIX.
  Dans ce cas, le *dump* sera écrit sur la première copie du LV.

- Le fichier "***/unix***" (*kernel*) doit être celui qui avait le contrôle lors du problème.
  Ceci signifie qu'on ne pourra pas analyser un *dump* si le noyau du système a été mis à jour par un correctif avant de passer la commande "***snap***".

  Exemple :

  Les éléments **obtenus seront corrects** pour la séquence :

```
crash
snap -a
update kernel
```

  Les éléments obtenus seront incorrects pour la séquence :

```
crash
update kernel
snap
```

Envoi des éléments

Nous vous rappelons, ci-dessous, l'adresse à laquelle envoyer vos éléments à analyser, **uniquement** à la demande du technicien du Point Service, **sans oublier** d'indiquer votre **numéro d'incident** (référence de dossier) :

Compagnie IBM France
M./Mme _____
Incident n° _____
Point Service AIX et SP
Service 1061
1, Place Jean-Baptiste Clément
93881 NOISY-LE-GRAND Cedex

■