

XIII Konferencja PLOUG
Kościelisko
Październik 2007

Przygotowanie systemu IBM AIX do instalacji Oracle RAC

Maciej Przepiórka
IBM Polska

maciej.przepiorka@pl.ibm.com

Abstrakt. Przygotowanie platformy systemowo-sprzętowej dla instalacji Oracle Real Application Cluster w wersji 10g jest bardzo mocno zależne od specyfiki tejże platformy. Referat ma za zadanie przedstawić aspekty, na które trzeba zwrócić szczególną uwagę przygotowując system IBM AIX 5L w trzech różnych architekturach:

- Oracle na wolumenach surowych z wykorzystaniem Automatic Storage Devices
- Oracle na IBM General Parallel File System
- Oracle na wolumenach surowych utworzonych jako grupa zasobów IBM HACMP

Opisano także wady i zalety każdej architektury, co pozwoli na wybór konkretnego rozwiązania zanim nastąpi proces implementacji.

1. Wprowadzenie

Oracle Real Application Clusters jest rozwinięciem Oracle Parallel Server znanego z wersji 8i i wcześniejszych. Każdy z węzłów (a więc każda instancja) ma dostęp do tych samych danych, a więc tych samych dysków. Idea, która przyświeca Oracle RAC, to zapewnienie wysokiej dostępności oraz wydajności poprzez skalowalność w poziomie. Dokładając kolejne węzły do RACA zwiększamy dostępność bazy danych (baza danych teoretycznie jest dostępna nawet wtedy, gdy tylko jeden węzeł pozostaje dostępny) oraz wydajność (zwiększa się całkowita ilość procesorów).

Z punktu widzenia administratora systemu operacyjnego instalacja Oracle RAC wymaga:

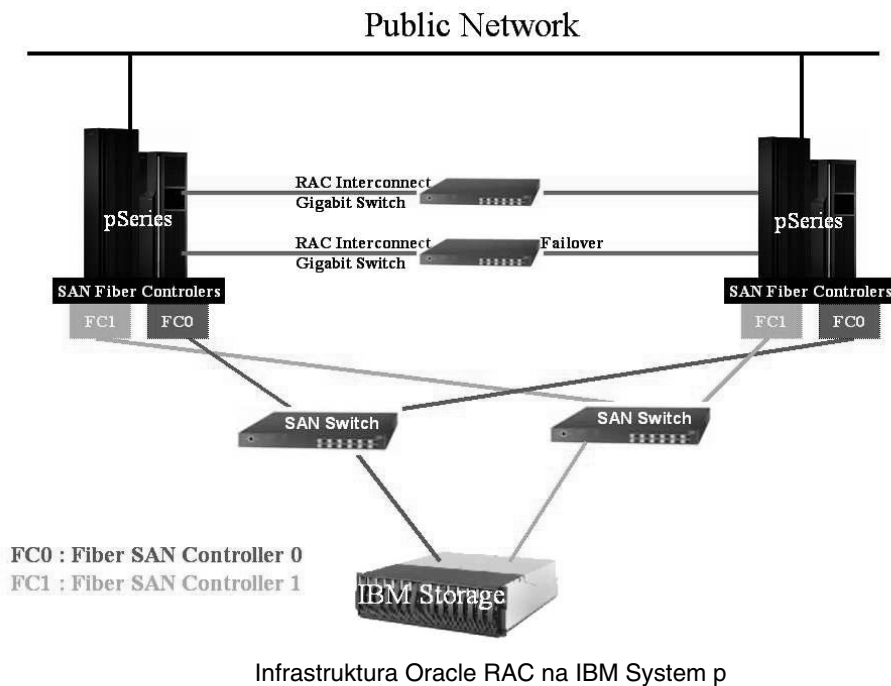
- utworzenia wspólnych zasobów dyskowych,
- zapewnienia połączenia sieciowego między wszystkimi węzłami (Oracle RAC interconnect) w warstwie IP.

O ile drugi warunek jest w miarę prosty do spełnienia, bo wystarczy zainstalować (najlepiej gigabitowy) switch ethernet i po jednym dodatkowym adapterze ethernetowym w każdym z węzłów, to pierwszy wymaga oprócz zainwestowania w sprzęt (macierz, przełączniki i adaptery SAN), zastosowania specjalnego oprogramowania, aby równoległy dostęp do dysków nie spowodował utraty spójności danych.

We wszystkich wersjach Oracle do 9i RAC włącznie administrator miał możliwość wyboru pomiędzy ulokowaniem plików bazy danych na wolumenach surowych zarządzanych przez AIX CLVM (Concurrent Logical Volume Manager) a równoległym systemem plików GPFS (General Parallel File System). Każde rozwiązanie wymagało zakupu dodatkowego oprogramowania – HACMP lub HACMP i GPFS. Od wersji Oracle 10g dostępna jest trzecia opcja, za darmo dostarczana przez producenta bazy danych – Oracle ASM (Automatic Storage Management) zarządzający fizycznymi urządzeniami dyskowymi. Poza tym, implementacja na systemie plików GPFS nie wymaga oprogramowania HACMP, gdyż jego funkcjonalność została zastąpiona przez Oracle Clusterware (CRS).

2. Elementy wspólne

Niezależnie od sposobu umiejscowienia plików bazy danych na dyskach, infrastruktura sprzętowa Oracle Real Application Clusters na IBM AIX wymaga w wersji minimalnej dwóch serwerów IBM System p, macierzy dyskowej SAN (w opisywanym przypadku, gdyż możliwa jest także praca na macierzy SCSI podłączonej do dwóch serwerów lub w sieci NAS), przełącznika LAN i SAN oraz odpowiedniej ilości adapterów zainstalowanych w obu serwerach. W większych instalacjach, gdzie krytyczne staje się zapewnienie wysokiej dostępności środowiska, stosuje się przynajmniej dwa przełączniki SAN i dwa przełączniki LAN obsługujące ruch Cache Fusion.



Do celów testowych można oczywiście użyć prostszej konfiguracji. Wystarczy jeden serwer IBM System p z wydzielonymi trzema logicznymi partycjami (LPAR) – jedną ze środowiskiem Virtual IO Server i dwiema z zainstalowanym systemem operacyjnym AIX. W takiej konfiguracji można zwirtualizować wewnętrzne zasoby dyskowe serwera i utworzyć wirtualne połączenia LAN między systemami operacyjnymi, jednak należy pamiętać, że nie jest to konfiguracja certyfikowana.

Należy zaznaczyć, że przełącznik LAN dla ruchu Cache Fusion jest konieczny nawet w dwuwęzłowej wersji RAC. Połączenie bezpośrednie (na krótko karta-karta) może być zastosowane jedynie w celach testowych i nie jest wspierane przez pomoc techniczną Oracle. Połączenie typu EtherChannel (agregacja adapterów LAN w celu zapewnienia wyższej dostępności i większej przepustowości) jest jak najbardziej wspierane i w wielu przypadkach zalecane.

System operacyjny AIX, grupy i użytkownicy

Na każdym węźle (na każdym serwerze bądź partycji logicznej) Oracle RAC powinna być zainstalowana dokładnie ta sama wersja systemu operacyjnego. Aktualnie zalecana wersja systemu AIX to 5.3. AIX 5.2 jest również certyfikowany i wspierany. AIX 5.1 nie jest wspierany dla Oracle w wersji 10g.

Konkretne wersje poprawek dla systemów operacyjnych to:

- dla AIX 5.2: Maintenance Level 04 lub nowszy,
- dla AIX 5.3: Maintenance Level 02 lub nowszy.

Po zainstalowaniu systemu operacyjnego należy się upewnić, czy zawarte są w nim wymagane przez Oracle RAC pakiety (*filesets*):

- bos.adt.base
- bos.adt.lib
- bos.adt.libm
- bos.perf.perfstat

- bos.perf.libperfstat
- bos.perf.proctools
- rsct.basic.rte
- rsct.compat.clients.rte
- xlc.aix50.rte.7.0.0.4
- xlc.rte.7.0.0.1

Kolejnym etapem przygotowania systemu operacyjnego jest utworzenie użytkownika oracle oraz zniesienie niektórych limitów nałożonych na środowisko systemu operacyjnego. W przypadku Oracle RAC, grupa i użytkownik, który będzie właścicielem instancji, muszą mieć ten sam GID i UID na wszystkich węzłach. Edytując plik `/etc/security/limits` zdejmujemy wszelkie limity (można także dla wszystkich użytkowników edytując sekcję `default`) wpisując jako wartość parametr `-1`.

Nazwa limitu	Wartość zalecana dla użytkownika oracle	Wartość zalecana dla użytkownika root
Soft file size	-1 (Unlimited)	-1 (Unlimited)
Soft CPU time	-1 (Unlimited)	-1 (Unlimited)
Soft data segment	-1 (Unlimited)	-1 (Unlimited)
Soft stack size	-1 (Unlimited)	-1 (Unlimited)

AIX jest o tyle przyjemny w porównaniu do innych systemów UNIX, że jądro systemu operacyjnego dynamicznie przydziela obszar pamięci dzielonej i nie trzeba modyfikować (bo ich nie ma) parametrów takich jak `shmmmin` czy `shmmmax`.

Użytkownik oracle powinien mieć odpowiednio skonfigurowane środowisko. W systemie AIX konfiguracja polega na edycji pliku `.profile` znajdującego się w katalogu domowym użytkownika. Poniżej przykładowe wpisy w profilu użytkownika oracle:

```
export ORACLE_SID=RAC1
export ORACLE_SCOPE=/oracle
export ORACLE_HOME=/oracle/crs
#export ORACLE_HOME=/oracle/ora10g
export ORACLE_CRS=/oracle/crs
export ORACLE_CRS_HOME=/oracle/crs
export ORA_CRS_HOME=/oracle/crs
export LD_LIBRARY_PATH=/oracle/crs/lib:/oracle/crs/lib32
export PATH=$ORACLE_HOME/bin:$PATH
export AIXTHREAD_SCOPE=S
export NLS_LANG=american_america.ee8iso8859p2
export NLS_DATE_FORMAT='YYYY-MM-DD HH24:MI:SS'
export TEMP=/tmp
export TMP=/tmp
export TMPDIR=/tmp
umask 022
```

Zmienna `ORACLE_SID` powinna być różna dla każdego z węzłów RAC (np. `RAC1`, `RAC2`, `RAC3`), a `ORACLE_HOME` w tym przypadku określa katalog instalacji plików binarnych Oracle Clusterware.

Sieć IP – interconnect oraz adresy publiczne i wirtualne

Konfiguracja IP w przypadku Oracle RAC jest niezmiernie istotna. Wiele problemów podczas instalacji wynika z nieprawidłowości, błędów i przeoczonych lub czynności pominiętych na tym etapie. Poprawne skonfigurowanie tego elementu zapewni swobodny przepływ danych przez interconnect, który jest jednym z najważniejszych elementów infrastruktury.

AIX widzi adaptory (porty) ethernet jako urządzenia enX, gdzie X jest liczbą naturalną lub zerem. Przykładowo, w systemie z dwoma kartami ethernet urządzenia te mogą mieć nazwy en0 i en1. Podczas instalacji Oracle Clusterware administrator definiuje, która z kart będzie kartą publiczną (przez nią będzie obsługiwany ruch z zewnątrz do bazy danych), a która kartą prywatną (będzie ona obsługiwała ruch interconnect CacheFusion). Ponieważ definiowanie odbywa się po nazwie urządzenia, wskazane jest aby karty LAN używane jako adaptory publiczne miały identyczne nazwy na wszystkich węzłach klastra. Innymi słowy, jeżeli na pierwszym węźle dla ruchu publicznego przeznaczymy kartę en1, to na innych węzłach też należy użyć adaptera en1. To samo dotyczy się kart typu private.

Z punktu widzenia wydajności warto pamiętać, że Oracle interconnect przesyła między węzłami klastra duże ilości bloków danych. Zapewnienie dużej przepustowości i minimalnych opóźnień na tym połączeniu jest krytyczne, jeśli klastrer ma się skalować w poziomie. Połączenie gigabit ethernet jest wysoce zalecane do tego celu, choć w konfiguracjach testowych można użyć sieci 100Mbps. Do zastosowań transakcyjnych o bardzo dużych wymaganiach wydajnościowych można zastosować adaptory typu InfiniBand, w których opóźnienia są nawet kilkukrotnie mniejsze niż w przypadku kart 1Gbps. Spowoduje to zmniejszenie oczekiwań typu global cache w instancjach bazy danych.

Parametry sieciowe, które należy ustawić na poziomie systemu operacyjnego, mają na celu umożliwienie przesyłania ramek sieciowych o dużej wielkości. Ustawiane na poziomie warstw TCP, UDP oraz IP na wszystkich węzłach klastra pozwalają na zapewnienie wysokiej wydajności komunikacji Oracle Cache Fusion. Nazwy parametrów wraz z sugerowaną wartością przedstawia tabela:

Nazwa parametru	Sugerowana wartość
ipqmaxlen	512
rfc1323	1
sb_max	1310720
tcp_recvspace	65536
tcp_sendspace	65536
udp_recvspace	655360 Zalecana wartość tego parametru to 10 razy wartość parametru udp_sendspace, ale mniej niż wartość sb_max.
udp_sendspace	65536 Zalecana wartość to 4kB + wartość <i>db_block_size</i> pomnożona przez <i>db_file_multiblock_read_count</i> .

Zmiany poszczególnych parametrów dokonuje się używając polecenia:

```
/usr/sbin/no -o parametr=nowa_wartość
```

Nazewnictwo hostów i adresacja IP są kolejnym ważnym elementem. Oracle RAC jest czuły na błędy w tym obszarze. Należy przede wszystkim zadbać, aby zawartość pliku `/etc/hosts` była zgodna na wszystkich węzłach. Podczas instalacji Oracle RAC używane są adresy publiczne, adresy prywatne (RAC interconnect) oraz adresy wirtualne (VIP). Adres wirtualny jest aliasem IP na karcie publicznej i powinien być wyznaczony w tym samym segmencie podsieci IP. Przykładową adresację IP w klastrze czterowęzłowym przedstawia tabela:

Adresy publiczne (karta en0)		Adresy wirtualne (karta en0)		Adresy prywatne (karta en1)	
Nazwa	Adres IP – maska 255.255.255.0	Nazwa	Adres IP – maska 255.255.255.0	Nazwa	Adres IP – maska 255.255.255.0
aix1	10.10.10.101	aix1-vip	10.10.10.111	aix1-int	10.0.0.101
aix2	10.10.10.102	aix2-vip	10.10.10.112	aix2-int	10.0.0.102
aix3	10.10.10.103	aix3-vip	10.10.10.113	aix3-int	10.0.0.103
aix4	10.10.10.104	aix4-vip	10.10.10.114	aix4-int	10.0.0.104

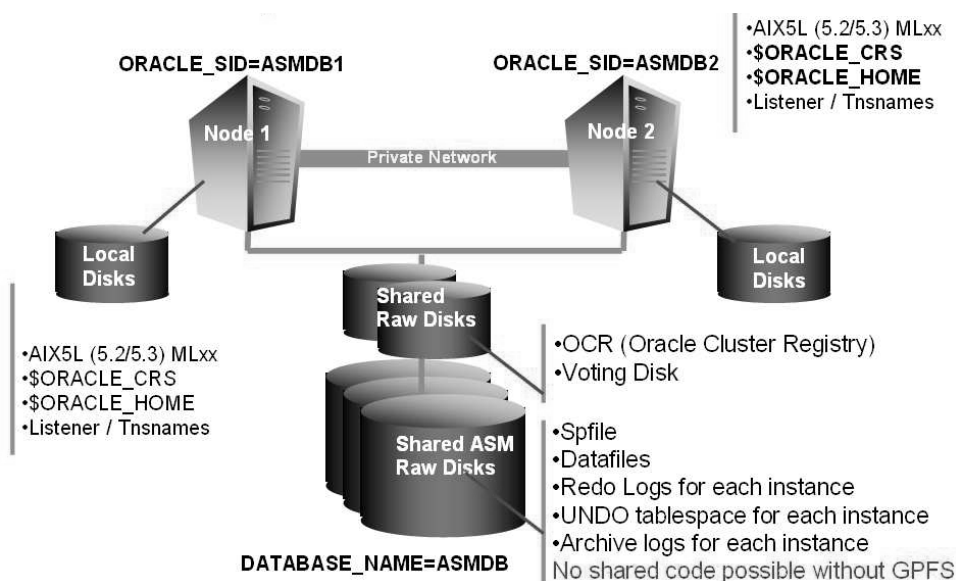
Ostatnim elementem wspólnym dla przygotowań instalacji Oracle RAC jest konfiguracja protokołu ssh na wszystkich węzłach w taki sposób, aby użytkownicy *oracle* oraz *root* mogli uruchamiać procesy i mieć dostęp do plików na pozostałych węzłach klastra bez użycia hasła. Ten etap konfiguracyjny jest niezbędny, aby pomyślnie przejść przez proces instalacji oprogramowania Oracle. W późniejszym etapie funkcjonalność ta może być bezpiecznie wyłączona. Po pomyślnym skonfigurowaniu "ekwiwalencji" ssh, użytkownicy *oracle* oraz *root* na każdym z węzłów powinni móc logować się na pozostałe węzły bez podawania hasła za pomocą polecenia *ssh aixN*, gdzie *N* będzie numerem dowolnego węzła tego klastra.

3. Wariant I – pliki bazy danych na Oracle ASM

ASM (Automatic Storage Management) jest instancją Oracle, która zajmuje się wyłącznie zarządzaniem dyskami na potrzeby bazy danych. Można używać tej funkcjonalności dla baz w klasycznej konfiguracji nieklastrowej (single instance). ASM łączy w sobie dwa niegdyś odrębne podejścia do konfiguracji bazy danych Oracle – system plików (prosty w obsłudze) oraz urządzenia surowe (skomplikowane w administracji, ale za to wydajniejsze od systemu plików).

ASM zarządza surowymi urządzeniami dyskowymi, na poziomie poniżej warstwy AIX LVM, a więc dyski, które mamy zamiar przeznaczyć dla bazy danych nie powinny należeć do żadnej grupy wolumenów ani posiadać sygnatury PVID. Jednak zanim zostaje uruchomiony ASM, startują usługi klastrowe Oracle. One także wymagają równoległego dostępu do (przynajmniej dwóch w minimalnej konfiguracji) dysków, a więc tam ASM niczego nie ułatwi. Binaria bazy danych również są skazane na rezydowanie na rozłącznych systemach plików. Podsumowując, na zasobach ASM mogą być umieszczone:

- plik spfile
- pliki danych (datafiles)
- redo logi
- przestrzeń undo
- archiwalne redo logi
- przestrzeń flashback



Urządzenia dyskowe w konfiguracji z wykorzystaniem ASM

W minimalnej konfiguracji musimy więc utworzyć 3 zasoby dyskowe (LUNy) i udostępnić je wszystkim węzłom klastra. Będą to:

- *CRS disk* (około 200MB) przechowujący konfigurację (topologię) naszego klastra,
- *VOTE disk* (około 200MB), który jest również używany przez CRS i zabezpiecza przed zjawiskiem split- brain (bardzo nieprzyjemne),
- *ASM disk* (przynajmniej tak duży, jak baza danych, która ma się na nim znajdować), który byłby w tym przypadku zasobem przeznaczonym dla plików bazy danych.

W konfiguracjach bardziej rozbudowanych mamy możliwość:

- utworzenia i równoległego używania dwóch zasobów typu *CRS disk*, aby w razie awarii jednego z nich konfiguracja klastra nie była utracona,
- utworzenia do trzech urządzeń dyskowych *VOTE disk*, co jest zalecane w przypadku, gdy używamy więcej niż jednej macierzy dyskowej,
- utworzenia wielu urządzeń dyskowych dla ASM. Jest to powszechnie stosowane, gdyż tylko wtedy ASM potrafi zapewnić redundancję danych oraz większą wydajność (dane są rozrzucone po dostępnych dyskach).

W systemie operacyjnym AIX napędy dyskowe są widziane jako `hdiskX` i reprezentowane przez dwa urządzenia: `/dev/hdiskX` oraz `/dev/rhdiskX`. Różnica między nimi jest zasadnicza: dostęp do `/dev/hdiskX` odbywa się w trybie blokowym, a do `/dev/rhdiskX` w trybie znakowym. Ten, kto próbował podłączać do Oracle urządzenia w trybie blokowym wie, że kończy się to falą niepowodzeń. Oracle CRS jak i ASM oraz sama baza danych, nie ważne czy pracuje w klastrze czy nie, wymaga dostępu do urządzeń w trybie znakowym, takich jak `/dev/rhdisk2`.

Mając zdefiniowane LUNy na macierzy dyskowej i widząc je w systemie operacyjnym, musimy zadbać o możliwość równoległego dostępu do nich. W tym celu na urządzeniach dyskowych administrator ustawia wartość parametru `reserve_policy` na `no_reserve`. Na przykład dla dysku `hdisk5` musimy wydać komendę:

```
chdev -l hdisk5 -a reserve_policy=no_reserve
```

W przypadku niektórych modeli macierzy dyskowych dostępny jest inny parametr konfiguracyjny tryb współdzielenia dysku, `reserve_lock`. Wtedy administrator wyda polecenie:

```
chdev -l hdisk5 -a reserve_lock=no
```

Ponieważ nasze urządzenia dyskowe mogą mieć różne oznaczenia (liczbowe) na różnych węzłach klastra (przykładowo *hdisk2* na węźle *aix1* może być widziany na węźle *aix2* jako *hdisk3*), należy wyeliminować możliwość błędnego przypisania dysków do poszczególnych funkcji. Identyfikując wspólne zasoby dyskowe tworzymy urządzenia o innych nazwach w taki sposób, aby na każdym węźle były reprezentowane identycznie. W pierwszej kolejności odnajdujemy numery major i minor danego urządzenia dyskowego:

```
aix1:root > ls -l /dev/hdisk3
brw----- 1 root system 20, 3 Aug 11 11:12 /dev/hdisk3
```

```
aix2:root > ls -l /dev/hdisk3
brw----- 1 root system 21, 3 Aug 11 11:12 /dev/hdisk3
```

Następnie tworzymy urządzenie o dostępie znakowym z takimi samymi numerami major i minor jak widziany na danym węźle dysk:

```
aix1:root > mknod /dev/vote_disk c 20,3
aix2:root > ls -l /dev/vote_disk
crw----- 1 root system 21, 3 Aug 11 11:12 /dev/vote_disk
```

```
aix2:root > mknod /dev/vote_disk c 21,3
aix2:root > ls -l /dev/vote_disk
crw----- 1 root system 21, 3 Aug 11 11:12 /dev/vote_disk
```

Urządzenie */dev/vote_disk* ma na węźle *aix1* takie same numery major i minor jak *hdisk2*, podobnie na węźle *aix2*. Podczas instalacji Oracle CRS wskażemy urządzenie */dev/vote_disk* jako dysk służący do odkładania głosów i unikniemy błędów wynikających z różnej numeracji dysków.

Podobną czynność należy wykonać dla pozostałych dysków, które mają być dostępne przez CRS bądź ASM.

Ostatnią niezbędną przed wystartowaniem instalatora Oracle CRS czynnością jest zmiana właściciela urządzeń dyskowych (*chown*) i nadanie odpowiednich uprawnień (*chmod*), aby użytkownik oracle mógł z nich korzystać:

```
chown oracle:dba /dev/vote_disk
chown oracle:dba /dev/ocr_disk
chown oracle:dba /dev/rhdisk5
chmod 660 /dev/vote_disk
chmod 660 /dev/ocr_disk
chmod 660 /dev/rhdisk5
```

Dalsza część instalacji powinna przebiegać bezproblemowo i nie różni się zasadniczo od instalacji Oracle RAC 10g na innych platformach sprzętowych.

4. Wariant II – pliki bazy danych na IBM GPFS

GPFS (General Parallel File System), jak sama nazwa wskazuje, jest klastrowym systemem plików. Jest produktem dojrzałym, rozwijanym przez firmę IBM od wielu lat. GPFS ma wiele możliwości konfiguracji i umożliwia nawet rozproszenie systemu plików na wielu lokalnych dyskach serwerów w sieci. Dla potrzeb Oracle RAC można go wykorzystać do współdzielenia między wieloma serwerami (węzłami klastra) zasobów plikowych znajdujących się na macierzy.

GPFS wymaga osobnej sieci IP służącej do synchronizacji między węzłami. W ostateczności można użyć tej samej sieci, która świadczy usługi na rzecz CacheFusion, ale nie jest to zalecana konfiguracja.

Konfiguracja General Parallel File System składa się z trzech etapów:

- instalacji binariów GPFS na wszystkich węzłach
- utworzenia klastra GPFS
- utworzenia systemu plików GPFS

Pierwszy etap polega na standardowej operacji zainstalowania plików binarnych (np. za pomocą narzędzia `smrit`). W drugiej części administrator tworzy plik `/var/mmfs/etc/node.list`, w którym wpisuje nazwy węzłów (muszą być uprzednio wpisane w `/etc/hosts`) oraz ich przeznaczenie:

```
aix1:root > cat /var/mmfs/etc/node.list
aix1-gpfs:1:quorum
aix2-gpfs:2:quorum
```

Następnie utworzenie klastra poprzez wywołanie komendy:

```
aix1:root > mmcrcluster -t lc -p aix1-gpfs -s aix2-gpfs -n ./node.list
```

W tym przypadku węzeł `aix1-gpfs` jest węzłem podstawowym GPFS, a węzeł `aix2-gpfs` węzłem zapasowym.

Utworzenie systemu plików, podobnie jak poprzednia czynność, wymaga w pierwszym etapie edycji pliku, który definiuje przeznaczenie dysków dla GPFS. W przykładowie dyski `hdisk2`, `hdisk3`, `hdisk4` i `hdisk5` będą współdzielonymi zasobami klastra składającego się z węzłów `aix1-gpfs` i `aix2-gpfs`:

```
aix1:root > cat /var/mmfs/etc/nsddisk.list
hdisk2:aix1-gpfs:aix2-gpfs:dataAndMetadata
hdisk3:aix1-gpfs:aix2-gpfs:dataAndMetadata
hdisk4:aix1-gpfs:aix2-gpfs:dataAndMetadata
hdisk5:aix1-gpfs:aix2-gpfs:dataAndMetadata
```

Później następuje czynność utworzenia wspólnego zasobu dyskowego:

```
aix1:root > mmcrnsd -F /var/mmfs/etc/nsddisk.list
aix1:root > mmlsnsd
File system  Disk name  Primary node      Backup node
-----
(free disk)  gpfs1nsd   aix1-gpfs         aix2-gpfs
(free disk)  gpfs2nsd   aix1-gpfs         aix2-gpfs
(free disk)  gpfs3nsd   aix1-gpfs         aix2-gpfs
(free disk)  gpfs4nsd   aix1-gpfs         aix2-gpfs
```

Ostatnią czynnością jest utworzenie właściwego systemu plików:

```
aix1:root > mmcrfs /oradata /dev/oradata -F /var/mmfs/etc/nsddisk.list \
-A yes -B 512k -n 2 -N 8000
```

oraz jego zamontowanie na każdym węzle:

```
aix1:root > mount /oradata
```

Parametr `-B` określa wielkość bloku systemu plików. Jest to również wielkość `stripe size` używanego przez GPFS do rozrzucania danych między dyskami.

Stworzenie systemu plików GPFS jest relatywnie prostym zadaniem. Do głównych zalet tego podejścia w przypadku Oracle RAC zalicza się możliwość instalacji binariów na wspólnym zasobie dyskowym oraz późniejszą łatwość w administracji bazą danych. Wszystkie pliki związane z bazą danych i oprogramowaniem CRS mogą być umieszczone na takim systemie plików, a dodatkowo można tam przetrzymywać dowolne inne pliki, które chcielibyśmy współdzielić między

serwerami. GPFS jest rozwiązaniem wysoce skalowalnym – pozwala na współdzielenie pojedynczego zasobu między 1530 węzłami (w przypadku AIXa) i rozproszenie systemu plików nawet na 1500 urządzeniach dyskowych. Limit wielkości systemu plików (przetestowany w rzeczywistości) to 2PB.

Na koniec warto pamiętać, że najnowszą certyfikowaną wersją tego systemu plików dla Oracle jest wersja 2.3.

5. Wariant III – pliki bazy danych na wolumenach IBM HACMP

HACMP (*High Availability Cluster Management Protocol*) jest komercyjnym produktem IBM, który umożliwia budowę klastrów niezawodnościowych w topologiach od 2 do 32 węzłów. Jedną z jego funkcjonalności jest CLVM – Concurrent Logical Volume Manager. Dzięki niemu możliwy jest jednoczesny dostęp do zasobów dyskowych z wielu węzłów klastra. Zasoby współdzielone są surowymi wolumenami logicznymi (*raw devices*), a więc mogą być użyte przez bazę danych Oracle jako przestrzeń dla plików danych.

W przypadku Oracle RAC 10g, certyfikowanymi wersjami HACMP są 5.1 i 5.2, jesienią 2007 roku powinna być certyfikowana także wersja 5.3, a na początku 2008 roku wersja 5.4. Instalacja HACMP w AIX jest prosta i polega na wgraniu jego pakietów. Administracja klastrem oraz zasobami HACMP może odbywać się całkowicie za pomocą narzędzia Smitty (`smit hacmp`) z jednego z jego węzłów. W przypadku Oracle 10g RAC konfiguracja jest uproszczona, bo nie wymaga definiowania adresów sieci serwisowej ani bootującej, ale sieciowe połączenie między węzłami HACMP jest jak najbardziej konieczne i używana jest sieć prywatna.

W ramach klastra HACMP administrator definiuje zasoby. W tym przypadku musimy zdefiniować przynajmniej jedną równoległą (*concurrent*) grupę wolumenów na wybranych wcześniej urządzeniach dyskowych wystawionych z macierzy do wszystkich węzłów, na których ma działać Oracle RAC. Ponieważ w tym wariantcie nie używamy systemu plików (i tak nie dałoby się tego zrobić), ale wolumenów surowych, na takiej grupie kreujemy wiele równoległych wolumenów logicznych (*concurrent logical volumes*), po jednym dla każdego pliku danych, wolumeny VOTE disk oraz CRS disk dla CRS, pliku `spfile` itd. Dostęp do nich będzie równoległy ze wszystkich węzłów klastra.

Ważnym elementem konfiguracji jest dodanie do grupy `hags` użytkownika `oracle`. Po tej czynności można startować klaster HACMP, co powinno poskutkować pojawieniem się nowej grupy wolumenów na wszystkich węzłach klastra. Będzie ona podłączona w trybie dostępu (*VG Mode*) `concurrent`:

```

aix1:root > lsvg oradatavg
VOLUME GROUP:   oradatavg          VG IDENTIFIER:   00003100000000ee20634c17
VG STATE:       active              PP SIZE:         32 megabyte(s)
VG PERMISSION:  read/write                     TOTAL PPs:      1084 (34688 megabytes)
MAX LVs:        256                  FREE PPs:        657 (21024 megabytes)
LVs:            2                     USED PPs:        427 (13664 megabytes)
OPEN LVs:       0                     QUORUM:          2
TOTAL PVs:      1                     VG DESCRIPTORS:  2
STALE PVs:      0                     STALE PPs:       0
ACTIVE PVs:     1                     AUTO ON:         no
Concurrent:     Enhanced-Capable      Auto-Concurrent: Disabled
VG Mode:        Concurrent
Node ID:        1
MAX PPs per PV: 1016
LTG size:       128 kilobyte(s)
HOT SPARE:      no                    BB POLICY:       relocatable

```

Po zdefiniowaniu logicznych wolumenów w ramach tak utworzonej grupy i przyporządkowaniu im odpowiednich nazw można przystąpić do instalacji Oracle CRS i tworzenia bazy danych.

Wariant instalacji Oracle RAC z użyciem HACMP jest dość popularny ze względu na jego szerokie wykorzystanie w środowiskach Oracle 9i RAC. Ponadto samo HACMP daje o wiele więcej funkcjonalności niż sama warstwa CLVM – potrafi przełączać różnego rodzaju zasoby (dyski, sieci, aplikacje) w ramach rozbudowanych topologii klastrowych, zarządzać sprzętowo replikacją macierzy dyskowych (w wersji XD) i stanowić podstawę dla wysoce dostępnych rozwiązań biznesowych. Jednak do zastosowań wyłącznie pod Oracle 10g RAC wydaje się być powoli wypierany przez mechanizm Oracle ASM, bo zarządzanie samą bazą danych jest w tym drugim przypadku prostsze – nie wymaga działań ze strony administratora systemu operacyjnego.

Podstawową wadą tego rozwiązania jest brak możliwości tworzenia wspólnych zasobów dla jakichkolwiek innych plików niż pliki danych i wolumeny CRS. W konfiguracji Oracle RAC każdy węzeł będzie musiał oferować odrębny system plików dla archiwalnych dzienników powtórzeń, co w razie potrzeby odtwarzania jest utrudnieniem.

6. Podsumowanie

Instalacja Oracle 10g RAC na systemie AIX umożliwia wybór metody zarządzania dyskami. Która z nich jest najlepsza? Niezaprzeczalną zaletą HACMP jest jego stabilność i dojrzałość. ASM, jak się go już skonfiguruje, jest prosty w użyciu. GPFS jest sprawdzony, równie łatwy i przyjemny jak ASM i można na nim również umieszczać wolumeny CRS, ale w świadomości niektórych osób będzie mniej wydajny niż wolumeny surowe.

Podsumowując i wymieniając zalety i wady każdego z rozwiązań:

ASM:

- łatwe zarządzanie plikami bazy danych za pomocą mechanizmu Oracle Managed Files, można swobodnie używać funkcjonalności AUTOEXTEND itp.,
- możliwość przechowywania w jednym wspólnym miejscu archiwalnych plików dziennika powtórzeń ze wszystkich węzłów,
- bardzo dobra wydajność,
- jest dostępne w cenie bazy danych bez dodatkowych opłat,
 - wolumeny Oracle CRS muszą być przechowywane poza ASM,
 - skomplikowana konfiguracja, niestety nie opiera się na mechanizmach zarządzania dyskami wbudowanych w system operacyjny,
 - przeznaczony jak na razie (w wersji 10g) tylko dla plików bazy danych,
 - rozwiązanie ciągle młode i niedojrzałe, ale agresywnie wypychane na rynek, przez co szybko ewoluuje.

GPFS:

- względnie prosta instalacja i konfiguracja,
- produkt dostępny od wielu lat, sprawdzony w trudnych warunkach u wielu klientów,
- umożliwia przechowywanie i współdzielenie wszystkich rodzajów plików, nie tylko bazy danych,
- prosta i przyjemna administracja bazą danych, jak w przypadku klasycznego systemu plików,
- dobra wydajność i bardzo dobra skalowalność,
 - rozwiązanie komercyjne, dodatkowo płatne.

HACMP:

- sprawdzony i dojrzały produkt, daje o wiele więcej możliwości niż potrzebuje Oracle RAC,
- dobra wydajność,
- duża kontrola nad środowiskiem, bardzo dużo szczegółowych opcji konfiguracji,

- rozwiązanie komercyjne, należy dokupić osobną licencję,
- utrudniona administracja ze względu na ciągłą konieczność współpracy między administratorem bazy danych a administratorem AIXa,
- nie jest systemem plików i nie umożliwia przechowywanie innych plików niż wolumeny CRS i pliki danych.

Wybór rozwiązania na pewno nie jest prosty, ale nikt nie zabronił rozważania kilku podejść jednocześnie. Można umieścić bazę danych na surowych wolumenach zarządzanych przez HACMP, a pozostałe elementy (binaria, archiwalne logi) przechowywać na systemie plików GPFS. AIX jest jednym z wiodących systemów operacyjnych na rynku, dlatego w przypadku jego wyboru pod Oracle RAC mamy szeroką możliwość wyboru oferowanych rozwiązań. W każdym przypadku zalecane jest dobre zaplanowanie instalacji, bo bez tego żaden system produkcyjny nie będzie działał długo.

Bibliografia – źródła

Dokumentacja Oracle Database 10g Release 2

Cookbook: Quick Installation Guide, Oracle 10g RAC R2 on IBM pSeries running AIX5L SAN Storage – European Oracle/IBM Joint Solutions Center Montpellier, January 2006

Dokumentacja GPFS:

<http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/topic/com.ibm.cluster.gpfs.doc/gpfsbooks.html>

Disaster recovery with general Parallel File System, IBM Corp, August 2004

Dokumentacja HACMP:

<http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/topic/com.ibm.cluster.hacmp.doc/hacmpbooks.html>