

Konfiguracja systemu Linux na macierzy dyskowej Compaq MA-8000

LINUX NA MACIERZY DYSKOWEJ

Technologia SAN (Storage Area Network)

oferuje bardzo ciekawe podejście do składowania danych.

W artykule przedstawiłem moje doświadczenia z rozwiązaniem SAN firmy Compaq oraz to, w jaki sposób współpracuje ono z systemami opartymi o jądro Linuksa. Być może informacje zawarte w tym artykule pomogą podjąć decyzję na temat użycia rozwiązania SAN, a także przybliżą tematykę związaną z tą technologią.

PATRYK KOŃCZYK

Wybór rozwiązania SAN w opisywanym przypadku został podyktowany chęcią ujednoczenia sposobu zarządzania podsystemami dyskowymi. Przyczyniło się ono także do lepszego wykorzystania zasobów dyskowych i dało dużą elastyczność w operowaniu pamięcią. Opisywane rozwiązanie wykorzystuje standard Fibre Channel, choć producenci oferują także rozwiązania SAN wykorzystujące standardy Gigabit Ethernet oraz iSCSI.

Standard Fibre Channel

Szczegółowe informacje na temat standardu można znaleźć w specyfikacjach, adresy są podane na końcu artykułu. Poniżej zamieszczam tylko kilka podstawowych informacji na temat standardu.

FC jest standardem definiującym wielowarstwową architekturę służącą przesyłaniu danych przez sieć. Warstw jest pięć od FC-0 do FC-4. Warstwa FC-0 definiuje nośnik fizyczny (w opisywanym przypadku jest to światło), standard definiuje także nośnik międzywarstwowy. FC-1 opisuje standard kodowania i dekodowania danych, jest nim 8b/10b. FC-2 to warstwa sieci zwierająca rdzeń protokołu FC. FC-3 definiuje zewnętrzne funkcje rozciągające się między wiele portów urządzenia FC. FC-4 definiuje interfejs do protokołów wyższych warstw, takich jak protokół IP.

Możliwe są trzy główne topologie point-to-point, arbitrated loop (AL), fabric.

Topologia point-to-point jest to bezpośrednie połączenie dwóch urządzeń, maksymalna liczba urządzeń w tej topologii to dwa.

Arbitrated loop (AL) jest topologią podobną do tych znanych ze standardów token-ring lub FDDI. Urządzenia możemy łączyć w pętle korzystając jedynie z portów w kontrolerach HBA (Host Bus Adapter), jednak wykorzystując huby FC zyskujemy większą niezawodność, gdyż awaria jednego urządzenia nie powoduje przerwania pętli. W topologii AL przepustowość jest podzielona między wszystkich użytkowników pętli.

Topologia fabric opiera się na przełącznikach FC. Jest to topologia najbardziej skalowalna i wydajna. Każdy węzeł ma dostęp do przepustowości 1 Gbps lub 2 Gbps w ramach jednego przełącznika FC. Możliwe jest łączenie przełączników ze sobą.

Każde urządzenie (urządzenie może mieć kilka portów każdy port ma osobny numer), posiada unikalny 64-bitowy numer – WWN (World-Wide Number). Numer ten służy do routingu ramek FC. Jest on wykorzystywany w fazie ustalenia D_ID (destination identifier) i S_ID (source identifier) za pomocą, których ramki są adresowane. D_ID i S_ID są 24-bitowymi adresami. W przypadku topologii arbitrated loop urządzenia troszczą się o unikalność tych identyfikatorów, w topologii fabric zajmują się tym przełączniki FC. Adresowanie za pomocą 24-bitów zamiast 64-bitów zostało wybrane ze względów optymalizacyjnych.

Standard definiuje także sześć klas usług. Każda z klas definiuje inny rodzaj komuni-

kacji między urządzeniami. Trzy główne rodzaje to: klasa pierwsza umożliwiająca dwóm urządzeniom dedykowaną komunikację z maksymalną prędkością, klasa czwarta pozwalająca na przydzielenie określonej przepustowości dla określonych połączeń oraz klasa szósta definiująca komunikację multicast.

Ważną cechą sieci SAN jest zoning. Zoning pozwala na segregowanie urządzeń według ich funkcji, zapobiega sytuacjom, w których kilka systemów ma dostęp do tego samego dysku, oczywiście jeśli taka sytuacja nie jest pożądana. W opisywanym przypadku zoning konfigurowany jest na kontrolerze macierzy.

Poziomy RAID

Macierz obsługuje wszystkie popularne poziomy RAID 0,1,0+1,5,JBOD. Możliwe są w zasadzie dowolne sposoby łączenia poziomów RAID ze sobą. Aby stworzyć mirror na MA-8000 wskazujemy dyski, które mają do niego należeć i dodajemy je do jednostki (unit), po to aby były widoczne z poziomu systemu operacyjnego. Na poziomie jednostki określa się, który host ma do niej dostęp. W ramach jednostki możemy łączyć poziomy RAID. Jednostka może się składać z kilku poziomów podstawowych lub pojedynczych dysków.

Co pod maską?

Opisywany model jest najmniejszym przedstawicielem rodziny. Ma wysokość 22U. Zaglądając do środka widzimy redundantne zasilanie, kontroler macierzy podłączony do switcha FC, pamięć cache oraz podtrzymującą ją baterię, wszystko w oddzielnych modułach hot-swap. Większość komponentów w razie awarii możemy wymienić bez wyłączenia macierzy. Opisywany model znajduje się w podstawowej konfiguracji, gdyż jest wyposażony w jeden kontroler, jedną półkę składającą z 14 dysków SCSI oraz 512MB pamięci cache w kontrolerze. Możliwa jest konfiguracja z dwoma kontrolerami w celu zapewnienia redundancji. Takie rozwiązanie musi być wspierane przez sterownik systemu operacyjnego dla kontrolera HBA. Jednostka jest wówczas widoczna podwójnie. Jedna ze ścieżek jest główna, a w razie awarii jej rolę przejmuje druga ścieżka.

MA-8000 a Linux

Z macierzą został dostarczony kontroler (HBA) Compaq StorageWorks, który instaluje się w slotach PCI. Kontroler został wy-

produkowany przez firmę Emulex i jest to model LP-8000. Firma Compaq oferuje komercyjne sterowniki do swojego produktu, natomiast na stronach firmy Emulex dostępne są sterowniki dla Linuksa na licencji GNU, które są rozwijane jako projekt lpfcxxxx.

Przeszukując sterowniki SCSI dostarczane z jądrem Linuksa można znaleźć sterownik Compaq Fibre Channel HBA, jednak z powodu braku dokumentacji nie został on wykorzystany.

Ze względu na większą homogeniczność niż Linuks na platformie x_86, pierwsze testy mające na celu sprawdzenie poprawności konfiguracji macierzy odbyły się na 64-bitowej platformie SPARC i systemie operacyjnym Solaris 9. Z powodzeniem udało się korzystać z dysków skonfigurowanych na macierzy.

Kolejnym pomysłem było zainstalowanie systemu operacyjnego Linuks na macierzy dyskowej podłączonej do maszyny Intel/Xeon za pomocą kontrolera HBA i spowodowanie, aby system operacyjny uruchamiał się bezpośrednio z macierzy. Z dokumentacji wynikało, że kontroler HBA daje taką możliwość i wystarczy włączyć w nim tak zwany boot code, a w biosie pojawi się nowy dysk SCSI. Za pomocą narzędzi ze strony producenta udało się włączyć boot code w kontrolerze HBA.

Do zainstalowania na macierzy została wybrana dystrybucja Slackware 9.1. Nie było jakiegось szczególnego powodu, dla którego wybór padł na Slackware. Zwyciężyło przyzwyczajenie i fakt, że z nią miałem najdłuższe doświadczenie. Slackware dostarcza standardowe jądro z krenel.org. W tym przypadku było to jądro serii 2.4. Kompilacja modułu odpowiedzialnego za obsługę HBA przebiegała bez problemu. Na stronach firmy EMULEX dostępnych jest kilka wersji sterowników, na pierwszy ogień poszła seria 2.01.

Skompilowany wcześniej moduł został załadowany bez ostrzeżeń, jednak w systemie nie pojawiło się nowe urządzenie. Pomogło przekierowanie do systemu plików proc następującej komendy `echo „scsi add-single-device 0 1 2 3” >/proc/scsi/scsi`, po której pojawiał się dysk `/dev/sda`. Po włączeniu sterownika bezpośrednio do jądra miała miejsce podobna sytuacja i dysk nie był widoczny dopóki do systemu plików proc nie została przekierowana odpowiednia komenda. W takiej sytuacji niemożliwe było zamontowanie partycji z systemem plików root, gdyż dysk, na którym się znajdowała, nie był widoczny podczas startu systemu. Z kolei wersja 4.301 sterownika nie sprawiła podobnego kłopotu. Dysk został wykryty i system udało się pomyślnie uruchomić. Rozwiązanie takie działało stabilnie ponad pół roku, nie było z nim problemu.

Kolejnym krokiem była przesiadka na jądro serii 2.6. Na stronie producenta w chwili pisania artykułu nie było dostępnych sterowników przeznaczonych specjalnie pod jądro 2.6. Z pomocą przyszedł projekt lpfcxxx. Mimo słabej dokumentacji sterownik udało się wkompiłować w jądro. W czasie pobierania sterowników należy zwrócić uwagę na

czajenie i fakt, że z nią miałem najdłuższe doświadczenie. Slackware dostarcza standardowe jądro z krenel.org. W tym przypadku było to jądro serii 2.4. Kompilacja modułu odpowiedzialnego za obsługę HBA przebiegała bez problemu. Na stronach firmy EMULEX dostępnych jest kilka wersji sterowników, na pierwszy ogień poszła seria 2.01.



wersje, gdyż nowsze wydania sterownika dają się jedynie zbudować z jądrami \geq 2.6.10. Konfiguracja oparta o jądro 2.6.10 oraz sterownik z projektu lpfcxxx działa bezawaryjnie do chwili obecnej, czyli około czterech miesięcy.

MA-8000 a Linux na platformie 64-bitowej

Kolejnym urządzeniem, które miało zostać podłączone do macierzy, był 64-bitowy serwer Sun v20z wyposażony w procesor opteron. Maszyna została zakupiona w podstawowej konfiguracji, z przeznaczeniem na serwer bazy danych. Pliki zawierające przesłanie table miały zostać umieszczone na macierzy. Tym razem jako system operacyjny został wybrany Suse Linux Enterprise 9 AMD64. Podczas instalacji można wybrać moduł odpowiedzialny za obsługę kontrolerów Emulex, który nazywa się lpfcdd. Mimo załadowania modułu lpfcdd w systemie nie pojawił się nowy dysk. W tym przypadku nie było wymagane załadowanie systemu operacyjnego bezpośrednio z macierzy i system operacyjny został umieszczony na wewnętrznym dysku serwera. Po instalacji service pack 1 do SLES 9 dyski z macierzy zostały wykryte. Ciężko mi jest cokolwiek powiedzieć na temat stabilności takiego rozwiązania, ponieważ jeszcze nie działa ono w środowisku produkcyjnym.

Niezależnie od zastosowanej architektury i dystrybucji Linuksa kontroler Emulex LP-8000 sprawił pewne kłopoty. Nie udało mi się wykonać zmiany wersji biosu i uruchomienie boot code po Linuksiem, musiałem wykonywać to na Solarisie 9 i architekturze SPARC.

Co dalej?

Możliwość wykorzystania rozwiązań Fibre Channel jest ogromna. W tym artykule zostały przedstawione tylko podstawowe przykłady i informacje na temat rozwiązania firmy Compaq (obecnie HP). Rozwiązania FC otwierają nowe możliwości w zastosowaniach klastrowych i HA (High Availability). Przykładem jest promowany przez Red Hat GFS (Global File System), stanowiący wydajniejszą i stabilniejszą alternatywę dla zdominowanych przez NFS rozwiązań klastrowych. Stworzenie rozwiązania HA przy użyciu technologii FC jest stosunkowo proste i bardzo elastyczne. Sposoby wykorzystania technologii FC można mnożyć, jednak ceny tego typu rozwiązań są duże nawet dla przedsiębiorstw średniej wielkości. ■

Ramka 1: Podstawowe terminy

SAN (Storage Area Network) – jest to sieć, w której zasoby pamięci masowej są transportowane w blokach SCSI (Small Computer System Interface). SCSI (Small Computer System Interface) – standard równoległego przesyłu danych między komputerem a jego urządzeniami poprzez tzw. szynę SCSI. FC (Fibre Channel) – jest to standard definiujący wielowarstwową architekturę służącą przesyłaniu danych przez sieć. Definiuje atrybuty warstwy fizycznej, transportowej, a także wsparcie dla protokołów wyższych warstw, takich jak TCP/IP, SCSI-3 i innych. RAID (Redundant Array of Independent Disks) – to określenie grupy dysków twardych, pracujących wspólnie pod kontrolą jednego oprogramowania zarządzającego. WWN (World-Wide Number) – każde urządzenie w sieci SAN posiada swój unikalny 64-bitowy numer sprzętowy. HBA (Host Bus Adapter) – jest to karta instalowana w komputerze najczęściej w jednej z odmian slotu PCI. Służy do obsługi technologii Fibre Channel.



INFO

[1] Standardy FC i SCSI: <http://www.t10.org> <http://www.t11.org>

[2] Sterowniki: <http://emulex.com/ts/indexemu.html> <http://lpfcxxx.sourceforge.net/>

[3] COMPAQ MA-8000: <http://h18006.www1.hp.com/products/storageworks/ma8kema12k/>