Maciej Roszkowski

Wirtualny klaster komputerowy jako narzędzie optymalizacji wydajności infrastruktury technicznej społeczeństwa informacyjnego

Ekonomiczne Problemy Usług nr 87, 479-487

2012

ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 702 EKONOMICZNE PROBLEMY USŁUG NR 87 2012

MACIEJ ROSZKOWSKI
Zachodniopomorski Uniwersytet Technologiczny

WIRTUALNY KLASTER KOMPUTEROWY JAKO NARZĘDZIE OPTYMALIZACJI WYDAJNOŚCI INFRASTRUKTURY TECHNICZNEJ SPOŁECZEŃSTWA INFORMACYJNEGO

Wprowadzenie

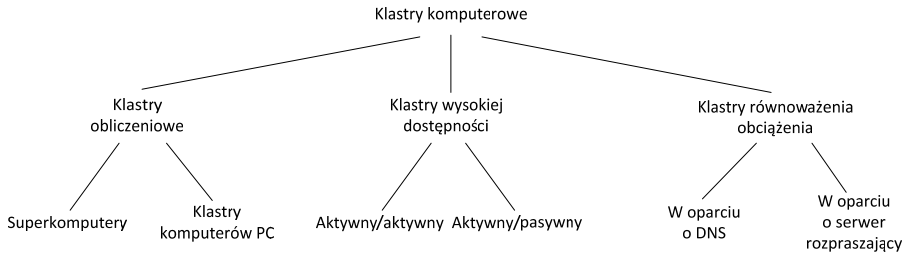
Społeczeństwo informacyjne funkcjonuje w oparciu o technologie informatyczne. Ciągły rozwój społeczeństwa informacyjnego jest możliwy dzięki rozwojowi infrastruktury technicznej. Pomimo ciągłego wzrostu wydajności infrastruktury technicznej zawsze pojawiają się zadania, które przekraczają możliwości używanych urządzeń. Wydajność urządzeń komputerowych zwiększa się poprzez stosowanie szybszych technologii półprzewodnikowych i wzrost częstotliwości taktowania procesorów. Tego typu działania wiążą się ze zwiększeniem emisji energii cieplnej i koniecznością wydajniejszego chłodzenia układów elektronicznych. Miniaturyzacja układów elektronicznych w zasadzie dobiega kresu. Przyszłością są rozwiązania równoległe.

Procesory komputerów są wyposażane w kilka rdzeni (*Multi Core*) oraz technologię hiperwątkowości (Hyper Threading – HT). Wiele rdzeni jest zintegrowanych w obudowie jednego procesora i każdy z nich stanowi procesor fizyczny wykorzystujący ten sam zestaw wyprowadzeń. Przy użyciu technologii HT do każdego rdzenia procesora fizycznego można przypisać dwa procesory wirtualne, przez co podczas obliczeń prowadzonych równolegle dwa niezależne wątki będą mogły korzystać z procesora w tym samym czasie (sprawia to wrażenie wykonania równoległego). Jeżeli aplikacja potrafi pracować wielowątkowo, to może to przyspieszyć wykonywanie programu od kilku do kilkunastu procent. Procesor wykorzystujący HT jest widziany przez system operacyjny jako dwa procesory logiczne. Wzrost wydajności procesorów przy jednoczesnej niskiej cenie umożliwia konstruowanie

coraz bardziej wydajnych komputerów i serwerów, a w konsekwencji całych systemów komputerowych[1].

1. Klaster komputerowy

Klaster komputerowy (*Computer Cluster*) jest to grupa wzajemnie połączonych niezależnych serwerów współdziałających razem, jako pojedynczy zintegrowany system komputerowy. Każdy serwer w klastrze tworzy węzeł klastra. Każdy z serwerów z osobna może pracować niezależnie, bez klastra. Integracja serwerów w klastrze jest możliwa poprzez oprogramowanie zarządzające klastrem. Głównym zadaniem oprogramowania zarządczego jest sterowanie właściwą pracą systemu, dystrybucją zadań, migracją procesów i zarządzanie zasobami systemu.



Rys. 1. Podział klastrów komputerowych

Źródło: opracowanie własne.

Klaster może realizować różne działania. W zależności od przeznaczenia można wyróżnić trzy rodzaje klastrów (rysunek 1), klastry obliczeniowe (ang. Compute Clusters), klastry wysokiej dostępności (ang. High Availability Clusters), klastry równoważące obciążenie (ang. Load Balancing Clusters).

2. Klaster obliczeniowy

Klaster obliczeniowy zapewnia zwiększoną moc obliczeniową. Najczęściej jest wykorzystywany do przeprowadzania obliczeń wysokiej wydajności (*High Performance Computing*), które wymagają wykonania dużej liczby operacji arytme-

[1] K. Lal, T. Rak, *Linux a technologie klastrowe*, Mikom, Warszawa 2005, s. 49–95.

tycznych i logicznych.[2] Obliczenia dotyczące konkretnego zadania są prowadzone w sposób równoległy, za pomocą wielu węzłów działających jednocześnie. Oczywiście zadanie musi zostać wcześniej przekształcone tak, aby każdy węzeł realizował odrębną i niezależną część całego zadania. Obliczenia równoległe znaczenie skracają czas rozwiązania zadań.

Klaster obliczeniowy umożliwia zwiększenie wydajności aplikacji poprzez jej pracę na wielu węzłach klastra równolegle. Do stworzenia aplikacji, która potrafi wykorzystywać równolegle wiele węzłów klastra, niezbędne jest wykorzystanie specjalnej biblioteki programistycznej. Przykładem narzędzi do tworzenia oprogramowania dla obliczeń równoległych jest: Interfejs Transmisji Wiadomości MPI (*Message Passing Interface*), Wirtualna Maszyna Równoległa PVM (*Parallel Virtual Machine*). Klaster obliczeniowy jest często nazywany mianem klastra wysokiej wydajności (*High Performance Cluster*).

Ze względu na architekturę rozwiązania można wyróżnić następujące dwa rodzaje klastrów:

- superkomputery,
- klastry komputerów PC.

Superkomputery to komputery o mocy obliczeniowej znacznie przewyższającej moc obliczeniową komputerów PC. Budowane są na zamówienie najczęściej z seryjnie produkowanych podzespołów komputerowych. Miarą wydajności superkomputerów jest liczba operacji zmiennoprzecinkowych na sekundę (*Floating Point Operations Per Second*). Stale uaktualniana jest lista 500 superkomputerów, które uzyskują najlepszy wynik w teście Lapack (test numerycznego rozwiązywania problemów algebry liniowej).

Klastry komputerów PC wykorzystują do swojego funkcjonowania powszechnie dostępny sprzęt komputerowy. Oferują moc obliczeniową jak najmniejszym kosztem. Ideą funkcjonowania klastrów komputerów PC jest fakt, że większość mocy obliczeniowych komputerów osobistych działających samodzielnie nie jest wykorzystywana. Połączenie komputerów za pomocą sieci lokalnych sprzyja wzajemnej komunikacji i współdziałaniu. Jednym z najpopularniejszych klastrów komputerów PC jest Beowulf, w którym komputery działają w oparciu o system operacyjny Linuks[3].

---

[2] R. Wyrzykowski, *Klastry komputerów PC i architektury wielordzeniowe, budowa i wykorzystanie*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2009, s. 18–19.

[3] W. Stallings, *Systemy operacyjne. Struktura i zasady budowy*, PWN, Warszawa 2006, s. 696–711.

3. Klaster wysokiej dostępności

Klaster wysokiej dostępności (klaster HA) zapewnia dostępność systemu komputerowego w przypadku wystąpienia awarii. Na każdym z węzłów klastra instalowane są usługi, które będą dostępne w trybie wysokiej dostępności. Wszystkie węzły klastra korzystają ze wspólnej macierzy dyskowej (*Shared Storage*). Jeden z węzłów klastra jest węzłem podstawowym (*Primary Node*), a pozostałe węzły są węzłami zapasowymi (*Backup Node*)[4]. Każdy z węzłów klastra wysokiej dostępności może się znajdować w dwóch trybach: aktywnym (*Active*), kiedy posiada uruchomione usługi w trybie wysokiej dostępności, lub w trybie pasywnym (*Passive*), kiedy nie posiada uruchomionych zasobów i znajduje się w stanie gotowości. Ciągły dostęp do zasobów klastra HA jest możliwy poprzez odwołanie się do adresu IP wirtualnego interfejsu, który wskazuje na interfejs sieciowy węzła podstawowego. Awaria węzła podstawowego nie powoduje zmiany adresu IP klastra HA. Oprogramowanie nadzorujące pracę klastra HA wykrywa wystąpienie awarii węzła przy wykorzystaniu usług uczestnictwa (*Membership Services*) i monitorowania zasobów (*Monitoring Services*). Mianem usług uczestnictwa określa się analizę komunikatów dostępności (*Heartbeat*) wysyłanych pomiędzy węzłami klastra. Monitorowanie zasobów to mechanizm okresowego sprawdzania dostępności usług węzła aktywnego. Wystąpienie awarii węzła powoduje rozpoczęcie procesu przejmowania zasobów (*Failover*). Zabezpieczeniem przed przypadkową utrata dostępności powinna być redundancja ścieżek pomiędzy węzłami klastra. Istnieje możliwość przywrócenia pracy macierzystego węzła po jego naprawie (*Failback*). Klaster HA nie zwiększa wydajności systemu komputerowego, jednakże umożliwia wyeliminowanie pojedynczego punktu awarii (*Single Point of Failure*). Podstawowym zadaniem klastra HA jest zwiększenie niezawodności i dostępności systemu komputerowego. Klaster wysokiej dostępności jest często nazywany klastrem pracy awaryjnej lub klastrem niezawodnościowym.

Naturalnym rozwinięciem klastrów HA są systemy stałej dostępności pomimo wystąpienia awarii (*Fault Tolerance* – FT). Główna różnica pomiędzy klastrem HA bez systemu FT a klastrem HA z system FT jest taka, że w przypadku pierwszego z nich zakłada się konkretny czas przestoju na usunięcie ewentualnej awarii (określa się procentowo jego dostępność). Klaster HA z technologią FT jest niezawodny, dopóki posiada redundancję wadliwego komponentu. Wszystkie komponenty w tej technologii mogą być wymieniane w trakcie pracy systemu (*Hot Swapping*). Ze względu na konfigurację węzłów można wyróżnić następujące dwa rodzaje klastrów HA:

– aktywny/aktywny (*Active/Active*),

[4] A. Silberschatz, P.B. Galvin, G. Gagne, *Podstawy systemów operacyjnych*, WNT, Warszawa 2003, s. 20.

– aktywny/pasywny (*Active/Passive*).

W modelu aktywny/pasywny klaster składa się z dwóch węzłów: podstawowego, będącego w trybie aktywnym, i zapasowego, będącego w trybie pasywnym. W przypadku awarii węzła podstawowego jego funkcje przejmuje węzeł zapasowy. W modelu aktywny/aktywny klaster składa się z dwóch węzłów, które są jednocześnie aktywne. Obydwa węzły współdzielą obciążenie (*Load Sharing*) generowane przez klientów korzystających z usług. Ruch sieciowy skierowany do węzła niedostępnego ze względu na awarię zostanie skierowany do węzła aktywnego. Większa liczba węzłów dostępnych w ramach klastra umożliwia dokonywanie modyfikacji konfiguracji. Przy N węzłach aktywnych oraz M węzłach pasywnych (klaster N + M) awaria jednego z węzłów aktywnych powoduje przejęcie jego roli przez jeden węzeł pasywny.

4. Klaster równoważenia obciążenia

Klaster równoważenia obciążenia zapewnia równomierne rozłożenie obciążenia na węzły klastra. Obciążenie jest generowane poprzez strumień zapytań od klientów do serwera usług. Rozłożenie obciążenia (*Load Balancing*) to proces, podczas którego następuje dystrybucja zapytań klientów na węzły klastra za pośrednictwem urządzenia sieciowego. Ideą równoważenia obciążenia jest niedopuszczenie do pełnej zajętości zasobów jednego z węzłów klastra.

Ze względu na konfigurację sieciową można wyróżnić następujące dwa rodzaje klastrów LB:

– klastry równoważenia obciążenia w oparciu o serwer DNS,
– klastry równoważenia obciążenia w oparciu o serwer rozpraszający.

Równoważenie obciążenia bazujące na serwerze DNS (*Domain Name Service*) jest prostym przypisaniem jednej domenie kilku adresów IP (kilku rekordów A) przy wykorzystaniu mechanizmu DNS Round Robin. Algorytm karuzelowy (*Round Robin*) umożliwia szeregowanie zapytań do serwera DNS bez uwzględnienia priorytetów (poszczególne zapytania o rekord A dają w rezultacie naprzemiennie różne adresy IP). Wadą mechanizmu DNS Round Robin jest trudne do przewidzenia równoważenie obciążenia oraz brak mechanizmu wykrywającego awarię węzłów klastra.

Równoważenie obciążenia bazujące na serwerze rozpraszającym zapytania (*Load Balancing Server*) zachodzi w warstwie 4 modelu OSI lub w warstwie 7 modelu OSI[5]. Rozłożenie obciążenia w warstwie 4 modelu OSI (na poziomie IP) polega na dystrybucji żądań od klientów do właściwych serwerów bez potrzeby analizy zawartości pakietu. Rozłożenie obciążenia w warstwie 7 modelu OSI (na

[5] http://www.linuxvirtualserver.org (luty 2012).

poziomie aplikacji) polega na analizie zawartości pakietu i dystrybucji zapytania do określonego węzła klastra. Przy wyborze konkretnego węzła klastra można zastosować różne algorytmy szeregowania zapytań. Zastosowanie serwera rozpraszającego zapytania umożliwia również monitorowanie dostępności węzłów klastra.

5. Propozycja architektury środowiska wirtualnego klastra komputerowego

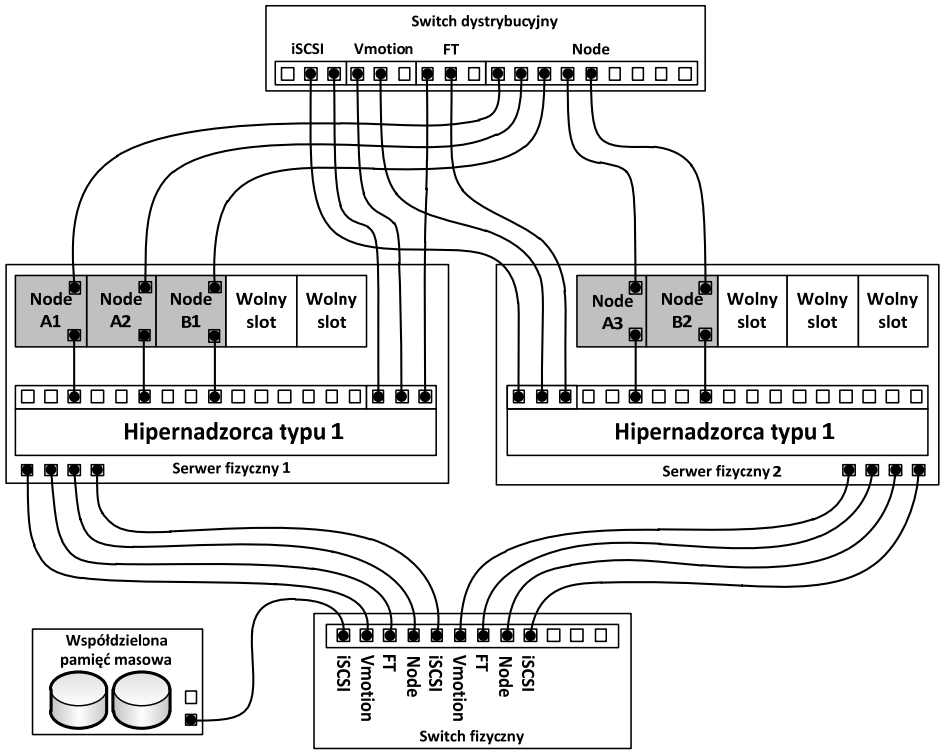
Wirtualny klaster komputerowy jest realizacją trzech głównych rodzajów klastrów: klastra obliczeniowego, klastra wysokiej dostępności, klastra równoważącego obciążenie, w postaci środowiska maszyn wirtualnych (rysunek 2). Celem proponowanej architektury jest możliwość jednoczesnego przeprowadzania obliczeń dla wielu zadań na dowolnej liczbie węzłów przy zachowaniu ciągłej dostępności środowiska obliczeniowego i oszczędności zasobów.

Schemat powstał w oparciu o oprogramowanie wirtualizacyjne VMware vSphere Hypervisor. Do wirtualizacji takiego środowiska można użyć każdego innego oprogramowania wirtualizacyjnego. Jedynym komponentem, który wymaga zastąpienia równoważnym rozwiązaniem, jest switch dystrybucyjny, który jest charakterystyczny dla oprogramowania firmy VMware.

Środowisko składa się z dwóch serwerów fizycznych, switcha fizycznego i współdzielonej pamięci masowej. Serwery fizyczne posiadają cztery karty sieciowe Ethernet, switch fizyczny wykorzystuje dziewięć z dwunastu portów Ethernet, współdzielona pamięć masowa posiada dwa porty Ethernet. Wszystkie komponenty fizyczne są połączone ze sobą według schematu za pomocą kabli typu patchcord. Każda z kart sieciowych serwerów fizycznych jest przeznaczona do przesyłania konkretnego ruchu sieciowego: komunikacja wirtualnych maszyn (Node), komunikacja z macierzą dyskową (iSCSI), ruch sieciowy związany z migracją wirtualnych maszyn pomiędzy serwerami fizycznymi (Vmotion), analiza komunikatów dostępności serwerów fizycznych (FT). Pozostałe komponenty oraz połączenia są wirtualne i powstały w oparciu o oprogramowanie wirtualizacyjne.

Na dwóch serwerach fizycznych został zainstalowany i uruchomiony hipernadzorca typu 1, który działając bezpośrednio na poziomie fizycznego sprzętu, ma pełną kontrolę nad uruchomionymi wirtualnymi maszynami. Każdy hipernadzorca używa wirtualnego przełącznika sieciowego (I/O plane), umożliwiającego komunikację wszystkim wirtualnym maszynom w ramach środowiska wirtualnego danego hipernadzorcy. Każda wirtualna maszyna jest połączona za pośrednictwem wirtualnej karty sieciowej Ethernet z portem sieciowym Ethernet switcha dystrybucyjnego (*control plane*). Na switchu sieciowym (I/O plane) hipernadzorcy są wydzielone porty dla połączeń: iSCSI, Vmotion i FT. Porty sieciowe switcha dystrybucyjnego również podzielone są na grupy, w celu oddzielenia ruchu sieciowego: ISCSI, Node, Vmotion, FT, podobnie jak w switchu fizycznym. Pomiędzy portami przełącz-

nika hipernadzorcy (iSCSI, Vmotion, FT) oraz odpowiadającymi im portami switcha dystrybucyjnego są nawiązane połączenia.



Rys. 2. Proponowana architektura środowiska wirtualnego klastra komputerowego

Źródło: opracowanie własne.

Klaster obliczeniowy jest realizowany na węzłach wirtualnych maszyn. Aktualnie węzły: Node A1, Node A2 i Node A3, realizują zadania obliczeniowe A. Natomiast węzły: Node B1 i Node B2, realizują zadanie obliczeniowe B. Ilość węzłów potrzebnych do obliczeń jest regulowana przez pulę zasobów (ilość pamięci RAM i wolne cykle procesorów serwera fizycznego), wyrażoną w postaci wolnych slotów. Ilość wolnych slotów i ilość wirtualnych maszyn jest uzależniona od parametrów poszczególnych wirtualnych maszyn i puli zasobów serwera fizycznego.

Klaster wysokiej dostępności jest realizowany poprzez technologię stałej dostępności wirtualnej maszyny pomimo wystąpienia awarii serwera fizycznego (*Fault Tolerance*). Jeżeli serwer fizyczny 2 ulegnie awarii, to obliczenia zadań A i B wykonywane przez węzły: Node A3 i Node B2, nie zostaną utracone. Mecha-

nizm FT umożliwi migrację uruchomionych już wirtualnych maszyn (bez restartu) na serwer fizyczny 1. Migracja wirtualnych węzłów jest możliwa dzięki temu, że każdy z węzłów istnieje w postaci plików na współdzielonej pamięci masowej, do której mają dostęp obydwa serwery fizyczne.

Klaster równoważący obciążenie jest realizowany poprzez mechanizm dynamicznego utrzymywania równowagi i alokacji zasobów DRS (*Distributed Resource Scheduler*). Mechanizm ten jest wbudowany w oprogramowanie wirtualizacyjne i pozwala balansować posiadanymi zasobami. Na bieżąco jest monitorowany stopień obciążenia serwerów fizycznych i dokonywana jest migracja uruchomionych wirtualnych maszyn pomiędzy serwerami fizycznymi. Jeżeli wszystkie wirtualne maszyny znajdowałyby się na serwerze fizycznym 1 i mechanizm DRS wykryłby, że wirtualne maszyny wymagają dodatkowej mocy, to serwer fizyczny 2 zostałby włączony i część maszyn zostałaby na niego przemigrowana. Jeżeli sytuacja byłaby odwrotna, serwer fizyczny 2 byłby obciążony w niskim stopniu, to mechanizm DRS migrowałby maszyny na serwer fizyczny 1, a serwer fizyczny 2 zostałby wyłączony.

Podsumowanie

Zaprezentowane rozwiązanie wirtualnego klastra pokazuje, że mechanizm wirtualizacji jest bardzo dobrym narzędziem do optymalizacji wydajności infrastruktury technicznej społeczeństwa informacyjnego. Użycie technologii wirtualizacji do połączenia funkcjonalności klastra obliczeniowego, klastra wysokiej dostępności i klastra równoważącego obciążenie pozwala na uzyskanie wydajnej infrastruktury obliczeniowej odpornej na awarię i przystosowanej do oszczędności zasobów.

Literatura

1. http://www.linuxvirtualserver.org (luty 2012).
2. Lal K., Rak T., *Linux a technologie klastrowe*, Mikom, Warszawa 2005.
3. Wyrzykowski R., *Klastry komputerów PC i architektury wielordzeniowe, budowa i wykorzystanie*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2009.
4. Stallings W., *Systemy operacyjne. Struktura i zasady budowy*, PWN, Warszawa 2006.
5. Silberschatz A., Galvin P.B., Gagne G., *Podstawy systemów operacyjnych*, WNT, Warszawa 2003.

A VIRTUAL COMPUTER CLUSTER AS A TOOL FOR AN EFFICIENCY OPTIMIZATION OF INFORMATION SOCIETY TECHNICAL INFRASTRUCTURE

Summary

The article presents and describes the following computer clusters, Compute Clusters, High Availability Clusters and Load Balancing Clusters. The author designs an architecture of virtual computer cluster environment that enables running calculations for many tasks on many nodes at the same time, maintaining continuous availability of computing environment and the most effective use of resources.

Translated by Maciej Roszkowski