comtegra

Comtegra Sp. z o.o.
ul. Puławska 474, 02-884 Warszawa
tel: +48 22 311 18 00, fax: +48 22 311 18 01
http://www.comtegra.pl

XIII Wydział Gospodarczy KRS, Nr KRS 0000622223
Kapitał zakładowy: 1.000.000,00 w pełni opłacony

icm

comtegra

Budowa superkomputera Okeanos

dla Interdyscyplinarnego Centrum Modelowania
Matematycznego i Komputerowego Uniwersytetu Warszawskiego

Superkomputer

Cechą charakterystyczną dzisiejszej nauki jest jej uzależnienie od danych. Mogą to być na przykład dane eksperymentalne, obserwacyjne lub wyniki symulacji komputerowych.

Projekt w którym firma Comtegra miała okazję wykazać się swoją wiedzą i doświadczeniem obejmował dostawę i integrację elementów systemu teleinformatycznego, w swojej istocie skoncentrowanego wokół całokształtu problematyki dotyczącej Wielkich Danych (Big Data).

Podstawą architektury systemu specjalizującego się w obliczeniach wielkoskalowych, którego pojedynczy przebieg wymaga kilku tysięcy rdzeni obliczeniowych wspieranych kilkudziesięcioma terabajtami pamięci operacyjnej jest Superkomputer. Umożliwia on analizę i przetwarzanie modeli matematycznych dużej ilości danych w bardzo krótkim czasie.

Aby zaprojektować zbudować i wdrożyć taką jednostkę centralną spełniającą jednocześnie wszystkie wymagania ICM, firma Comtegra podjęła współprace z firmą **Cray**.

Michał Odziemczyk
inżynier analizy i wizualizacji danych

- ponad tysiąc węzłów w systemie Cray XC40
- węzły wyposażone w 24 rdzenie procesora Intel Xeon z opcją Hyper Threading(HT) i 128GB pamięci operacyjnej
- procesory Intel Xeon są architekturą typu little-endian – gdzie procesory w obrębie węzła mają spójna pamięć podręczną oraz niejednorodny czas dostępu do pamięci głównej
- całość tworzy 48 logicznych jednostek obliczeniowych
- węzły połączone są wyspecjalizowaną, wysokowydajną siecią typu Cray Aries o prędkości 100Gb/s połączonych w topologii DragonFly

Piotr Siedlak
inżynier systemowy

- 5 x macierzy DDN SFA 12KX
- 25 x półka dyskowa SS8460
- 2100 dysków NL-SAS 6TB 7,2K RPM
- macierz NetApp E2700 wraz z 12 dyskami SAS 900 GB 10K RPM
- 27 serwerów obsługujących file system

Jak zbudować wysokowydajny, pojemny I skalowalny system przechowywania danych?

Zaproponowane przez naszą firmę rozwiązanie dało ICM jeden z najwydajniejszych, najbardziej skalowalnych file systemów w Polsce. System ten umożliwia pracę z ogromną wydajnością sięgającą 150 GB/s, prędkość tę można porównać do nagrania danych na 34 płyty DVD w ciągu sekundy. System plików udostępnia online 10 PB pojemności. Jeśli chcieli byśmy obejrzeć taką ilość filmów Full HD zajęło by to nam około 133 lata, a ułożone jedna na drugiej płyty CD utworzyły by wieżę o wysokości 2 km!

System plików Lustre wykorzystywany jest przez superkomputer **Cray XC40** do składowania danych, które generowane są podczas obliczeń. Całość rozwiązania została oparta o systemy dyskowe firmy **DDN** - światowego lidera w rozwiązaniach dla rynku HPC, serwery **Lenovo** oraz switche infiniband FDR **Mellanox.** Do przechowywania metadanych file systemu wykorzystaliśmy macierz **NetApp** E2700 z dyskami SAS. Rozwiązanie opiera się na systemie operacyjnym **Exascaler** dostarczonym również przez firmę **DDN**. Jest to jeden z najbardziej stabilnych systemów plików Lustre. Przy projektowaniu sieci wymiany danych zdecydowaliśmy się na wykorzystanie technologii infiniband, do której są routowane pakiety z sieci **Aries** będącej układem nerwowym superkomputera **Cray**. Zarówno infiniband jaki Aries to najszybsze technologie przesyłania danych, które dają jednocześnie najmniejsze opóźnienia.

System plików

Big Data

Wdrożyliśmy rozwiązanie w postaci pięciu, niezależnych modułów analitycznych, które choć mają identyczną sprzętową konfigurację, mogą być konfigurowane zupełnie niezależnie. Takie podejście pozwala na wspólną pracę wielu zespołów analitycznych przy jednoczesnym zapewnieniu im odrębnych zasobów.

Niezawodność została uzyskana poprzez wykorzystanie serwerów firmy **Huawei**. Dla zapewnienia wydajności, inaczej niż w przypadku standardowych systemów analizy danych, serwery zostały połączone ze sobą nie tylko standardową siecią Ethernet, ale także superszybką siecią Infiniband zbudowaną w oparciu o rozwiązania firmy **Mellanox**. Takie podejście pozwala na uzyskiwanie wyników analiz szybciej niż przy wykorzystaniu tradycyjnych metod.
Skala projektu stawiała przed firmą Comtegra wiele wyzwań nie pojawiających się w mniejszych wdrożeniach. Samo zainstalowanie tak dużej ilości sprzętu i połączenie go kilkoma kilometrami różnego rodzaju połączeń sieciowych w kilka dni była niezwykłym osiągnięciem. Dodatkowo cały system jest elastyczny i pozwala na szybką rekonfigurację poszczególnych modułów analitycznych – z pomocą przyszły zaawansowane rozwiązania software'owe dostarczone przez firmę **Cloudera** wraz z dystrybucją **Apache Hadoop/Spark**. Systemy operacyjne mogą być automatycznie przeinstalowywane i konfigurowane, wszystkie operacje można wykonywać z centralnego serwera zarządzającego.

- 363 serwery pozwalające na zapisywanie i przetwarzanie danych
- 1440 dysków o pojemności 6TB
- w sumie ponad 8PB przestrzeni
- przesył danych z prędkością do 56Gb/s

Przemysław Dubaniewicz
architekt systemowy

- 4,7 PB RAW w pojedynczej szafie rack (588 x Ultrastar He8 8TB)
- wydajność (aggregate throughput) 3.5GB/s per rack
- możliwość klastrowania dla zwiększenia wydajności, pojemności i wysokiej dostępności
- model Pay-as-you-grow
- obsługa protokołów S3, HTTP REST

Radosław Mirkowski
inżynier systemowy

Coraz większa ilość i wartość informacji stawia organizacjom IT wyzwanie budowania środowisk wysoce skalowalnych i odpornych ale jednocześnie globalnie dostępnych i niedrogich. Aby uzyskać przewagę konkurencyjną, wymaga się aby były one zawsze dostępne niezależnie od tego, gdzie i kiedy są przechowywane.

Zaawansowane chmury storage'owe korzystają z obiektowego systemu przechowywania danych gdzie podstawowym elementem nie jest plik lecz obiekt. Zapewnia to niezbędną redundancję i skalowalność przechowując dane w sposób rozproszony replikując je przynajmniej kilkukrotnie. Jeśli uszkodzeniu ulegnie dysk, po tym jak zostanie on wymieniony, system przepisze aktualną kopię z innych węzłów w sposób dużo szybszy niż przy tradycjonalnym rozwiązaniu RAID.
HGST Active Archive System jest kompletnym rozwiązaniem archiwum obiektowego, w którym dzięki podejściu „out of the box" dostajemy gotowe rozwiązanie (w pełni okablowane, ze wszystkimi niezbędnymi elementami gotowymi do uruchomienia i zintegrowania). Usługa **BitSpread** zabezpiecza dane poprzez bezstratne wydajnościowo kodowanie obiektów w całej hierarchii dysków systemowych i węzłów zapewniając chronione i dostępne dane w przypadku jednoczesnej utraty wielu dysków. Pozwala to zmniejszyć narzut pojemności zarezerwowanej o ponad 60% w stosunku do tradycyjnego storage'u z metodą RAID. Mechanizm **BitDynamics** sprawdza spójność informacji i w przypadku wystąpienia awarii automatycznie przenosi je na inny nie uszkodzony dysk.

Archiwum