

Klaster UKŁAD i serwer GROM w WCSS

– pomiary wydajności w teście High Performance Linpack

mgr inż. Paweł Dziekoński
mgr inż. Maciej Dyczkowski

16 lipca 2003 roku

1 Wprowadzenie

W kwietniu 2003 roku klaster UKŁAD [1], zainstalowany we Wrocławskim Centrum Sieciowo – Superkomputerowym [2], powiększył się o 20 nowych węzłów, osiągając wielkość 30 komputerów, w tym 1 serwer komunikacyjny oraz 29 węzłów obliczeniowych. Komunikację w klastrze zapewnia przełącznik sieciowy 100Mbps, który spięty jest z serwerem linkiem gigabitowym. Klaster pracuje pod kontrolą systemu operacyjnego Linux 2.4.

W czerwcu 2003 roku rozbudowano natomiast serwer obliczeniowy GROM (SGI Origin 2400), pracujący we Wrocławskim Centrum Sieciowo – Superkomputerowym od 1997 roku. Wymianie podlegały wszystkie procesory, zwiększono także ilość zainstalowanej pamięci operacyjnej. Zainstalowano także najnowszą dostępną wersję systemu IRIX 6.5.20.

Rozbudowa zasobów KDM spowodowała konieczność przeprowadzenia nowych testów wydajności wszystkich nowych elementów, całego superkomputera, a także klastra jako całości. W tym celu użyty został popularny test NETLIB High Performance Linpack [3]. Zastosowanie tego samego narzędzia do wszystkich testów dało również możliwość porównania zupełnie odmiennych architektur sprzętowych do jakich należą Klaster UKŁAD i Serwer GROM.

Po przeprowadzonej modernizacji, konfiguracja sprzętowa testowanych komputerów przedstawia się następująco:

Konfiguracja serwera dostępowego w klastrze:

- Serwer: obudowa Intel SC5200, rack 5U,
- 2 procesory Intel Xeon 1.8GHz, Hyper-Threading, 512KB cache L2,

- płyta główna Intel SHG2,
- 3 GB RAM (DDR/266 ECC Registered),
- kontroler RAID Intel SRCMR i macierz SCSI-160 w konfiguracji RAID-10/1 – 360 GB,
- interfejsy sieciowe: 2x 1000 + 1x 100Mbps Ethernet.

Konfiguracja 9 starszych węzłów klastra:

- obudowa "super tower" SuperMicro, dedykowana do zastosowanych płyt głównych,
- 2 procesory Intel Xeon 1.7GHz, 256KB cache L2,
- płyta główna SuperMicro P4DCE+,
- 1 GB RAM (RDRAM/400),
- 40 GB przestrzeni na pliki tymczasowe (LVM 2x20 GB IDE/ATA100),
- interfejsy sieciowe: 1x 100Mbps Ethernet,
- grafika: NVidia GeForce2 GTS,
- monitory: iiyama Pro451.

Konfiguracja 20 nowych węzłów klastra:

- obudowa Intel SR1300, rack 1U,
- 2 procesory Intel Xeon 2.8GHz, Hyper-Threading, 512KB cache L2,
- płyta główna Intel SE7501WV2,
- 2 GB RAM DDR/266,
- 40 GB przestrzeni na pliki tymczasowe (LVM, IDE/ATA100),
- interfejsy sieciowe: 2x 1000Mbps Ethernet.

Konfiguracja serwera GROM - SGI Origin 2400:

- 32 procesory MIPS R14000 500MHz,
- 32 GB pamięci RAM,
- 420 GB przestrzeni dyskowej,
- 4x 100Mbps Ethernet,
- 1x 155Mbps ATM.

2 Test High Performance Linpack

Test High Performance Linpack (HPL, pełna nazwa to *A Portable Implementation of the High-Performance Linpack Benchmark for Distributed Memory Computers*) stał się w ostatnich latach referencyjną metodą pomiaru wydajności komputerów dużej mocy (KDM). Test ten jest także oficjalnym testem organizacji TOP500 [4] publikującej listę najszybszych KDM na świecie (TOP500 nie narzuca implementacji NETLIB HPL a jedynie określa procedurę testową w sensie matematycznym; w praktyce dominuje wykorzystanie implementacji NETLIB HPL).

Test HPL polega na rozwiązywaniu złożonego układu równań z danymi podwójnej precyzji (64 bity). Dane do obliczeń generowane są losowo. Praca maszynowo równoległa możliwa jest dzięki implementacji algorytmu komunikacji w standardzie MPI-1.1 [5]. Tym samym test pozwala szacować wydajność nie tylko procesorów ale także zastosowanych do połączenia procesorów rozwiązań sieciowych. Istotne znaczenie ma także przepustowość magistrali systemowej (komunikacja z interfejsami sieciowymi) oraz magistrali pamięci.

Najbardziej czasochłonnym krokiem obliczeń jest szereg operacji macierzowych. Z tego względu zastosowanie wydajnej implementacji biblioteki BLAS (*Basic Linear Algebra Subprograms*) [6] ma kluczowe znaczenie. Najlepiej jeśli zastosowana biblioteka została napisana specjalnie dla danej platformy sprzętowej. Biblioteki takie udostępniane są zwykle przez producentów KDM lub można pobrać ich darmowe implementacje z Internetu.

Wynikiem testu jest wydajność badanego KDM mierzona w miliardach operacji zmiennoprzecinkowych na sekundę (Gflops). Test pozwala na kontrolę dokładności uzyskiwanych wyników.

Test HPL posiada wiele parametrów konfiguracyjnych. Najważniejsze z nich to:

- PxQ – wymiary tablicy która definiuje ilość używanych procesorów,
- N – wielkość problemu obliczeniowego,
- NB – wielkość bloku wymiany danych,
- threshold – próg badania dokładności wyników,
- PFACTs – *panel factorization*,
- NBMINs – kryterium zatrzymania rekursji,
- NDIVs – ilość paneli w rekursji,
- RFACTs – *recursive panel factorization*,

- BCASTs – rodzaj topologii sieci,
- DEPTHS – *lookahead depth*,
- *swapping threshold*.

3 Wyniki testów

3.1 Klaster UKŁAD

Obecnie klaster posiada 29 węzłów obliczeniowych i 58 procesorów. Teoretycznie możliwe jest uruchomienie przez użytkownika jednego zadania obliczeniowego na wszystkich procesorach. Nie ma to jednak sensu z punktu widzenia wydajności takich obliczeń. Wynika to z dwóch przyczyn. Przyczyna pierwsza to stosunkowo niska wydajności zastosowanej sieci. W teście HPL wszystkie procesy testu komunikują się pomiędzy sobą. Wymaga to zastosowania bardzo wydajnej sieci (najlepiej specjalizowanych sieci typu Myrinet [7] lub Quadrics [8]). Zastosowana przy budowie klastra UKŁAD sieć typu Fast Ethernet nie spełnia takich wymagań. Drugą przyczyną to różnice w konfiguracji starszych i nowszych węzłów. Główny proces testu rozdziela zadania procesorom potomnym w równych porcjach i oczekuje na wszystkie wyniki zanim rozdzieli kolejne dane. Algorytm taki jest często realizowany w wielu programach masywnie zrównoleglonych. Oznacza to, że proces uruchomiony na szybszej maszynie skończy swoje obliczenia szybciej i będzie czekał na procesory wolniejsze.

Przesłanki opisane powyżej spowodowały, że przyjęto następującą procedurę testową:

1. wyznaczono wydajność sub-klastra starszych węzłów,
2. niezależnie wyznaczono wydajność sub-klastra nowszych węzłów,
3. wyznaczono wydajność serwera klastra jako pojedynczej maszyny.
4. Końcowa wydajność klastra to suma tych trzech wyników.

3.1.1 Porównanie wybranych bibliotek BLAS

W chwili obecnej najpopularniejszą implementacją BLAS dla platformy IA-32 jest ATLAS (*Automatically Tuned Linear Algebra Software*) [9]. Pakiet ten pozwala na etapie kompilacji uzyskać biblioteki BLAS automatycznie dostosowane do konkretnego komputera. W teście wykorzystany został ATLAS w wersji 3.2.1, zainstalowany z pakietów dystrybucji Debiana Woody (pakiety atlas2-sse2 i -base 3.2.1ln-7). Biblioteka ta dawała lepsze wyniki niż kompilowana lokalnie wersja 3.4.x.

Parametr	Wartość
N	4000
NB	80
P	1
Q	2
PFACT	Left Crout Right
NBMIN	4 8
NDIV	2 3 4
RFACT	Left Crout Right
BCAST	1ring 1ringM 2ring 2ringM
DEPTH	1
SWAP	Mix (threshold = 60)
L1	transposed form
U	transposed form
EQUIL	no
ALIGN	8 double precision words

Tabela 1: Konfiguracja testu porównującego biblioteki BLAS.

Biblioteka	Parametry	Czas [s]	Wynik [Gflops]
Goto	W10C2C8	6.94	6.155
ATLAS	W10L2C4	8.25	5.177

Tabela 2: Wyniki testu porównującego biblioteki BLAS.

Drugą znaną implementacją jest biblioteka GOTO [10] napisana przez Kazu-shige Goto. Wykorzystana wersja to 0.6, prekompilowana dla procesorów Intel Pentium4 z cache'em 256 lub 512 kB.

Dostępna jest także biblioteka Math Kernel Library Intela. Użycie jej do tych testów jednak było nie możliwe ze względu na błędy linkowania.

Porównanie wykonano przez uruchomienie testu na nowym węźle dwuprocesorowo. Ponieważ jest to test porównawczy i interesujące są wielkości względne uzyskanych wyników, więc wielkość problemu (N) ustalono na niewielką wartość (4000) co pozwoliło na wykonanie pełnego skanu pozostałych parametrów testu w rozsądnym czasie. Pełna konfiguracja testu przedstawiona jest w Tabeli 1.

Uzyskane wyniki zestawiono w Tabeli 2. Biblioteka Goto daje prawie 19% przyspieszenia w stosunku do ATLAS-a i będzie używana jako biblioteka podstawowa w dalszych pełnych testach.

3.1.2 Porównanie wybranych implementacji standardu MPI

W teście porównano wyniki uzyskane dla implementacji MPICH [11] w wersji 1.2.5 ("release) of : 2003/01/13 16:21:5") oraz LAM/MPI [12] w wersji LAM 6.5.9/MPI 2 C++/ROMIO.

Implementacja	Sieć [Mbps]	Parametry	Czas [s]	Wynik [Gflops]
LAM	100	W00L2L4	618.16	11.48
CH	100	W12R2C4	609.09	11.66
LAM	1000	W00L2L4	470.85	15.08
CH	1000	W00L2L4	471.87	15.05

Tabela 3: Wyniki testu porównującego implementacje standardu MPI.

Test porównawczy przeprowadzono na 4 procesorach na 2 węzłach z użyciem przełączanej sieci Ethernet 100 i 1000 Mbps. Wielkość problemu ustalono na $N=22000$ a blok NB na 80.

Wyniki zestawione w tabeli 3 wskazują na wyższą wydajność implementacji MPICH w sieciach o gorszych parametrach wydajności. W przypadku sieci gigabitowej osiągnane wyniki są praktycznie takie same. Ponieważ klastr UKŁAD oparty jest na sieci Ethernet 100 Mbps, do dalszych testów użyto MPICH.

3.1.3 Wydajność starych węzłów klastra UKŁAD

W roku 2002 wyznaczono wydajność klastra składającego się z 9 starych węzłów na 17.87 Gflops. Wydajność jednego węzła wyznaczono na 4.302 Gflops. Testy te przeprowadzono z użyciem biblioteki ATLAS. Zastosowanie biblioteki GOTO pozwoliło uzyskać wynik 4.935 Gflops dla pojedynczego węzła. Cały subklastr nie był testowany ze względu na uruchomione zadania użytkowników. Z tego względu całkowita wydajność starszego sub-klastra została przeskalowana o zysk wynikający z użycia biblioteki GOTO wg. wzoru: $(17.87/4.302) * 4.935$, co daje wynik 20.50 Gflops.

3.1.4 Wydajność nowych węzłów klastra UKŁAD

Test wykonano przy użyciu trzech konfiguracji sieci: bezpośredniego połączenia maszyn interfejsami sieciowymi (maks. 3 węzły) oraz przy użyciu przełączników sieciowych - gigabitowego i 100 Mbps.

Tabela 4 zestawia wyniki testów przy użyciu przełączników sieciowych.

Nowo zainstalowane węzły posiadają po 2 porty gigabitowe. Bezpośrednie połączenie dwóch komputerów port w port pozwala wykonać transfer FTP z prędkością prawie 900 Mbps. Tabela 5 zestawia wyniki testów 2 i 3 węzłów połączonych bezpośrednio ze sobą. Konfiguracja taka pozwala wyeliminować czynnik opóźnienia generowany przez aktywne urządzenia sieciowe takie jak przełączniki. Test przeprowadzono dla wielkości problemu 22000 (2 węzły) lub 27000 (3 węzły), oraz bloku NB=80.

Wykonano również test z użyciem 2 maszyn i przy wykorzystaniu tylko 1 procesora na każdej z nich (konfiguracja 2x1). Wielkość problemu ustalono na 15500,

Ilość węzłów	Sieć [Mbps]	N	NB	Parametry	Topologia	Czas [s]	Wynik [Gflops]
1		15500	80	W10L3R2	1x2	312.37	7.949
2	1000	22000	80	W11R2R8	2x2	491.55	14.44
3	1000	27000	80	W11C2C4	2x3	648.93	20.22
4	1000	31000	80	W11C2L4	2x4	787.39	25.23
5	1000	35000	80	W12C2C4	2x5	930.64	30.72
6	1000	38500	80	W10L2R4	3x4	1093.15	34.80
2	100	22000	80	W12R2C4	2x2	609.09	11.66
3	100	27000	80	W11C2R4	2x3	907.91	14.45
4	100	31000	80	W11C2L4	2x4	1223.82	16.23
5	100	35000	80	W12R2C4	2x5	1447.26	19.75
6	100	38000	80	W14L2L4	3x4	1855.55	19.72
20	100	68000	160	W04L3L8	5x8	3899.30	53.76

Tabela 4: Wyniki testu dla całej instalacji nowych węzłów.

Ilość węzłów	Parametry	Czas [s]	Wynik [Gflops]
2	W00L2L4	471.87	15.08
3	W10L2C4	605.92	21.66

Tabela 5: Wyniki testu z użyciem bezpośredniego połączenia interfejsów sieciowych węzłów.

NB=80, pozostałe parametry: W00L2L4, czas: 302.33, wynik: 8.213 Gflops. Ponieważ wydajność 1 węzła w teście 2-procesorowym (konfiguracja 1x2) wyniosła mniej, tj. 7.949 Gflops, można podejrzewać, że konfiguracja 2x1 ujawnia wąskie gardło wydajności węzła, prawdopodobnie w przepustowości magistrali pamięci.

3.1.5 Test węzła serwerowego klastra UKŁAD

Test węzła serwerowego wykonano dla wielkości problemu 19000, NB=80, pozostałe parametry testu to W00L2L4. Uzyskana wydajność: 5.110 Gflops.

3.2 Serwer GROM

3.2.1 Architektura komputerów klasy SGI Origin [13]

Dla skutecznego przeprowadzenia testów serwera GROM istotne było uwzględnienie architektury testowanej maszyny. Jest to klasa komputerów S2MP (*Scalable Shared Memory Multiprocessor*), a więc wyposażonych we wspólną pamięć dzieloną dostępną dla każdego z wielu procesorów. Dla zapewnienia odpowiedniej skalowalności procesory są połączone ze sobą w topologii hiperkostki, przy czym cała dostępna pamięć jest rozproszona pomiędzy tzw. moduły. Moduł jest elementem konstrukcyjnym, na którym oprócz pamięci za-

instalowane są także dwa procesory. Moduły połączone są zgodnie z topologią hiperkostki za pomocą ruterów i tzw. craylinków.

Należy pamiętać, że dostęp do pamięci nie jest jednolity. Pamięć umieszczona w tym samym module jest dostępna niemal natychmiast, za to informacje odczytywane z najodleglejszych komórek pamięci muszą zostać przetransmitowane przez minimum trzy rutery. To powoduje, że wynik testu mocno zależy od zastosowanego do obliczeń algorytmu.

Testom poddano dwie konfiguracje - całą maszynę złożoną z 32 procesorów i 32 GB pamięci RAM oraz jej wycinek złożony z dwóch procesorów umieszczonych we wspólnym module z 2 GB pamięci RAM. Pierwszy pomiar wyznacza całkowitą moc obliczeniową komputera w teście HPL, będącą pochodną sumarycznej prędkości dostępnej liczby procesorów oraz ilości dostępnej pamięci i jakości mechanizmów komunikacji międzyprocesorowej. Drugi pomiar ma na celu ułatwienie porównania serwera GROM z klastrem UKŁAD oraz umożliwienie oceny skalowalności maszyny w teście HPL.

3.2.2 Przebieg testów serwera GROM i wyniki

Test HPL jest testem wieloprocessorowym, w którym znaczącą rolę odgrywa komunikacja. Do wyboru jest pięć modeli komunikacji, przy czym wszystkie są modyfikacją topologii pierścienia (ring). Fizycznie komunikacja odbywa się jednak w sposób na jaki pozwala budowa maszyny, dlatego w celu skrócenia czasu komunikacji skorzystano z narzędzia dostępnego dla systemu operacyjnego IRIX – dplace.

Narzędzie dplace pozwala kontrolować zachowanie się procesów, ich rozmieszczenie na poszczególnych procesorach oraz umiejscowienie wykorzystywanych przez każdy proces stron pamięci. Osiągnięto w ten sposób następujące cele:

- wymuszono "przywiązanie" procesu do procesora, na którym został on uruchomiony,
- kolejne procesy rozmieszczane były na procesorach w taki sposób, aby sąsiadowały ze sobą w topologii hiperkostki,
- pamięć dla każdego procesu przydzielana była możliwie lokalnie - z tego samego modułu.

Pozostałe parametry testów wyznaczono empirycznie i przyjęto wartości, przy których wyniki testów próbnych dawały najlepsze rezultaty. Parametry testu, dla którego uzyskano najlepszy wynik dla całej maszyny, zostały zamieszczone w tabeli 6.

Parametr	Wartość
N	40000
NB	100
P	4
Q	8
PFACT	Right
NBMIN	2
NDIV	2
RFACT	Right
BCAST	2ringM
DEPTH	1
SWAP	Binary-exchange
L1	no-transposed form
U	no-transposed form
EQUIL	no
ALIGN	8 double precision words

Tabela 6: Konfiguracja testu całej maszyny Origin 2400 .

Parametry testu, dla którego uzyskano najlepszy wynik dla dwóch procesorów, zostały zamieszczone w tabeli 7.

Parametr	Wartość
N	16000
NB	192
P	2
Q	1
PFACT	Right
NBMIN	2
NDIV	2
RFACT	Crout
BCAST	1ring
DEPTH	0
SWAP	Binary-exchange
L1	no-transposed form
U	no-transposed form
EQUIL	no
ALIGN	8 double precision words

Tabela 7: Konfiguracja testu dla dwóch procesorów.

3.2.3 Wyniki testów serwera GROM

Czas trwania obliczeń testowych dla pełnej konfiguracji maszyny wyniósł 1847 sekund, a moc maszyny obliczono na 23.10 Gflops. W przypadku dwóch procesorów z wydzieloną dostępną pamięcią 2GB RAM czas trwania testu wy-

niósł 1750 sekund, a moc wyznaczona w ten sposób wyniosła 1.56 Gflops. Sumując wyznaczoną wydajność wszystkich modułów otrzymujemy wartość 24.96 Gflops. Uzyskany wynik stanowi więc prawie 93% maksymalnej wartości.

4 Podsumowanie

4.1 Uzyskane wyniki końcowe:

Klaster UKŁAD:

- wydajność sub-klastra starszych węzłów = 20.50 GFlops,
- wydajność sub-klastra nowszych węzłów = 53.76 GFlops,
- wydajność serwera klastra = 5.11 GFlops,
- wydajność całej instalacji = 79.37 GFlops.

Serwer GROM:

- wydajność dwuprocessorowego modułu = 1.56 GFlops,
- wydajność całej maszyny 23.10 GFlops.

5 Wnioski

1. Uzyskany wynik plasuje klaster UKŁAD na drugim miejscu w rankingu KDM prowadzonym przez KBN [14]. Biorąc to pod uwagę pozycja serwera GROM w tym samym rankingu nie ulegnie zmianie.
2. Możliwe jest zwiększenie wydajności klastra – wymaga to zainstalowania infrastruktury sieciowej o lepszych parametrach przepustowości i opóźnienia. Jedną z dróg osiągnięcia tego celu jest zakup przełącznika gigabitowego. Specjalizowane sieci typu Myrinet są nieprzydatne ze względu na niewielką ilość realizowanych przez użytkowników zadań masywnie równoległych. Możliwa jest także wymiana starszych procesorów 1.7 GHz szybsze, oraz zwiększenie pamięci do 2 GB na wszystkich węzłach.
3. Wydajność serwera GROM jest zbliżona do sumy wydajności poszczególnych jego elementów. Wskazuje to na wysoką skalowalność maszyny. Obecna konfiguracja zapewnia więc jej optymalne wykorzystanie.
4. Wejście klastra na listę TOP500 (wg. stanu z czerwca 2003) wymagałoby zwiększenia jego wielkości około czterokrotnie - przy użyciu podobnych

komponentów, lub około dwukrotnie w oparciu o procesory Itanium2 (przy założeniu, że próg wejścia nie zwiększy się znacząco w następnych edycjach listy).

Literatura

- [1] Klaster UKŁAD, <http://uklad.wcss.wroc.pl>
- [2] Wrocławskie Centrum Sieciowo – Superkomputerowe, <http://www.wcss.wroc.pl>
- [3] High Performance Linpack, <http://www.netlib.org/benchmark/hpl/>
- [4] Projekt TOP500, <http://www.top500.org/>
- [5] Message Passing Interface, <http://www.mpi-forum.org/>
- [6] BLAS – Basic Linear Algebra Subprograms, <http://www.netlib.org/blas/>
- [7] Myricom, Inc., <http://www.myri.com/>
- [8] Quadrics Ltd., <http://www.quadrics.com/>
- [9] ATLAS – Automatically Tuned Linear Algebra Software, <http://math-atlas.sourceforge.net/>
- [10] Biblioteka GOTO: High-Performance BLAS by Kazushige Goto, <http://www.cs.utexas.edu/users/flame/goto/>
- [11] MPICHi, <http://www-unix.mcs.anl.gov/mpi/mpich/indexold.html>
- [12] LAM/MPI, <http://www.lam-mpi.org/>
- [13] Performance Tuning Optimization for Origin2000, <http://www.cineca.it/manuali/sgi/origin/O2000Tuning.0.html>
- [14] Lista komputerów dużej mocy obliczeniowej KBN, <http://www.kbn.gov.pl/pub/info/dep/di/top-kdm.html>