Superkomputery

Paweł ARTYMOWICZ
University of Toronto

Superkomputery pomagają w badaniach przyrody, projektowaniu urządzeń i leków. Czym są, jak działają, jakich używają procesorów, jak szybko liczą? Odpowiedzi na te pytania zilustrujemy przykładami kilku superkomputerów, w tym czterech najszybszych na świecie oraz największego w Polsce.

Zdefiniujmy superkomputer

W latach 60. (to starożytna epoka technik obliczeniowych; geometryczny postęp technologii komputerowej opisaliśmy w *Delcie* 5/2017) uznawano za superkomputery maszyny firm Cray i CDC. Były 3 do 10 razy szybsze niż inne komputery. Później nazywano tak wszelkie systemy zdolne do obliczeń o rząd wielkości szybszych niż pojedynczy, średni komputer. Pod koniec XX wieku międzynarodowa grupa informatyków podjęła się uaktualniania dwa razy na rok szczegółowej listy 500 najszybszych systemów obliczeniowych świata. Jeśli przyjąć treść listy za ich definicję, to wiadomo ściśle, ile jest superkomputerów (500). Listę TOP500 łatwo jest odszukać w sieci.

Od chaosu do Linuxa

Do pierwszych lat obecnego wieku panowała trudna dziś do wyobrażenia różnorodność i konkurencja kilkunastu rodzajów procesorów. Mikroprocesory o architekturze x86 korporacji Intel zdobyły w ostatniej dekadzie całkowitą przewagę w dziedzinie wysokowydajnych obliczeń HPC (*High Performance Computing*), pozostawiając konkurencyjnej (także amerykańskiej) architekturze Power firmy IBM mniej niż 10% rynku. To samo nastąpiło wśród maszyn domowych. Np. komputery Apple miały kiedyś procesory Motoroli, potem PowerPC, ale od 2005 r. przeszły na procesory z rodziny x86. Tylko w telefonach przewagę zyskały procesory ARM (o zredukowanym słowniku komend procesora). Azja i Europa (mimo że ta druga nie produkuje wiodących procesorów) myślą o zawojowaniu w następnej dekadzie z pomocą oszczędnej architektury ARM rynku superkomputerów.

Podobna unifikacja zaszła w oprogramowaniu podstawowym superkomputerów. Z chaosu dawnych zmagań wyłonił się jako zwycięzca system operacyjny Linux (potomek Unixa). System Windows nie jest spotykany w świecie HPC. Linux jest nie tylko bardziej logiczny i niezawodny, ale jest też systemem otwartym i darmowym. Mikroprocesory Intela są natomiast niezawodne, szybkie i popularne, ale pilnie strzeżone patentami i drogie. Produkcja odbywa się w 75% w USA, a reszta w Irlandii, Izraelu i Chinach, po czym wysyłane są zwykle do końcowej integracji i testowania w Malezji. Zamiana piasku (krzemu) na procesory to działalność opłacalna, lecz kapitałochłonna. Skoro mowa o pieniądzach, to Intel oferuje zwykle studentom wszystkich uczelni świata za darmo swoje drogie oprogramowanie. Warto skorzystać!

Anatomia superkomputera: rdzenie, pamięci i łącza

W dawnych czasach superkomputer miał 4 do 8 niezależnych procesorów (rdzeni) i unikatowy system łącz danych. Mieścił się w jednym dużym monolicie. Obecne superkomputery są zupełnie inne: to klastry, czyli bardzo liczne ($\sim 10^4$) węzły obliczeniowe umieszczone w standardowego wymiaru szafach komputerowych, połączone szybkimi łączami. Węzeł ma kilka procesorów obliczeniowych (1–8), a z fizycznej konieczności uzasadnionej w wyżej cytowanym artykule *Delty* każdy procesor ma wiele rdzeni liczących równocześnie (od 4 do 240). Duży superkomputer ma dziś zatem wiele milionów rdzeni obliczeniowych, z których każdy może wykonywać kilka równoległych wątków obliczeń. Zadanie musi zostać podzielone na mnóstwo współbieżnych części, dopasowanych jak najlepiej do wielopoziomowej hierarchii zarówno kalkulatorów, jak i pamięci: od pamięci podręcznej (*cache*, do kilkudziesięciu MB) i pamięci operacyjnej (10 do 100 GB, tj. gigabajtów), aż po najwolniejsze i najbardziej pojemne pamięci stałe lub dyskowe (czasami w sumie kilka petabajtów, PB = 10^{15}B). Szybkość dostępu do pamięci z dowolnego procesora jest zasadnicza, dlatego łącza komunikacyjne są niesłychanie ważne, będąc potencjalnie wąskim gardłem obliczeń. W najpopularniejszych łączach Infiniband wiązką drutów lub światłowodem płynie strumień danych 50–100 GB/s. Komputer ma więcej łączy niż węzłów, np. w topologii wielowymiarowego torusa. Dane rozpoczynają płynąć szerokim strumieniem już po mikrosekundzie od wysłania komendy. W największych instalacjach w czasie sekundy transmitowany jest siecią łącz petabajt, tj. zawartość trzeciej co do wielkości na świecie biblioteki na Uniwersytecie Toronto (54 miliony dokumentów). Gdy porównamy ten strumień liczb ($\sim 10^{14}$ liczb/s) z sumarycznym tempem działań arytmetycznych w superkomputerze (prawie 100 PFLOP/s = 10^{17}/s), to zrozumiemy, że rdzenie obliczeniowe dużo szybciej produkują wyniki działań, niż pozyskują dane. Sztuka HPC sprowadza się do tego, by liczyć jak najbardziej lokalnie, nie dając się rdzeniom nudzić w oczekiwaniu na dane z odległej pamięci. Pomagają w tym nieco kompilatory, tłumaczące program w jednym z języków komputerowych (C++, Fortran i in.) na binarne instrukcje w kodzie x86. Optymalizacja nie odbywa się jeszcze w pełni automatycznie. Nadal ważną rolę odgrywają umiejętności programisty decydującego o strukturach danych i programu.

Platformy obliczeniowe i systemy

Superkomputer umieszczony jest w odpowiednio zaprojektowanym budynku, w sali o powierzchni 100–600 m^2. Zużywa do kilkunastu MW mocy, a jego potężny układ chłodzenia wodnego lub powietrznego znajduje się w innej części budynku. Pracuje w komputerze tyle wentylatorów i pomp, że na uszach trzeba mieć ochraniacze, jeśli przebywa się w pobliżu ponad 15 minut dziennie.

W czołówce najpotężniejszych maszyn (pierwsze 3 wiersze tabeli na następnej stronie) większość obliczeń robią nie procesory główne (CPU), lecz dodatkowe koprocesory, każdy

o dużej liczbie rdzeni, choć o wolniejszych od CPU zegarach taktowych, ze względu na limit energii. Omawiając tabelę, poznamy dwa główne rodzaje platform obliczeniowych nazywanych MIC i GPU. W tabeli figurują: rok budowy systemu i ówczesne miejsce w rankingu, liczba rdzeni obliczeniowych, moc zasilania, rodzaj głównej platformy obliczeniowej i dwie miary prędkości: P, największa prędkość w testach, i $P_{\max}$, prędkość osiągalna teoretycznie (w jednostkach petaflop/s, PF $= 10^{15}$ operacji podwójnej precyzji na sekundę). Podajemy też koszt systemu i koszt względny (stosunek koszt/prędkość w tysiącach dolarów za 1 teraflop/s). Największe systemy nie są, jak widać, tanie, ale dla porównania – nie są droższe niż samolot pasażerski Airbus A380.

Niektóre superkomputery (styczeń 2017 roku). Pierwsze cztery to najszybsze obecnie systemy.

kraj	nazwa	ranking, w roku	rdzeni	P [PF]	($P_{\max}$) [PF]	zasil. [MW]	rodzaj proc.	koszt [mln \$]	koszt/P
Chiny	Sunway TaihuLight	1, 2016	0,6 mln	93	(125)	15	MIC	237	2,5
Chiny	Tianhe-2	1, 2013	3,1 mln	34	(55)	18	MIC	390	11
USA	Tytan	1, 2012	0,6 mln	18	(27)	8	GPU	97	5
USA	Sekwoja	1, 2012	1,6 mln	17	(20)	7	CPU	250	13
Polska	Prometeusz	38, 2015	41 tys.	1,7	(2,3)	0,9	CPU	12	7
Kanada	SciPhi	>500, 2016	3,4 tys.	0,07	(0,08)	0,02	MIC+GPU	0,06	0,9

Systemy tradycyjne oparte na CPU: Prometeusz i Sekwoja

Najszybszym komputerem w historii Polski, przez pewien czas 38. na świecie (dziś 59.), jest Prometeusz, zainstalowany w 2015 r. w krakowskiej Akademii Górniczo-Hutniczej przez firmę Hewlett-Packard. Jest oparty na CPU serii Haswell Intela i osiąga 1,7 PF. Na powierzchni zaledwie 13 m^2 mieści 15 szaf sprzętowych z 30 t sprzętu zużywającego prawie megawat mocy elektrycznej. Podobnie jak poprzedni superkomputer AGH, Zeus (ciągle na liście TOP500), pomaga polskim naukowcom w badaniach podstawowych i stosowanych, zwłaszcza w zakresie chemii fizycznej, medycyny i bioinformatyki, a także fizyki i inżynierii. Tym zajmują się też wszystkie inne superkomputery na świecie. Większym odpowiednikiem Prometeusza jest amerykańska Sekwoja, przez pół 2012 roku najszybszy superkomputer świata, który zawiera 16-rdzeniowe procesory centralne IBM PowerPC i mieści się w 98 szafach w laboratorium Lawrence Livermore koło San Francisco. To był zapewne ostatni najszybszy system zbudowany wyłącznie z CPU o niewielkiej liczbie rdzeni. Jego najbardziej znanymi osiągnięciami były symulacja kosmologiczna z $3{,}6{\cdot}10^{12}$ cząstkami oraz symulacja elektrofizjologii serca. Sekwoja ma intrygująco podwójną osobowość: projektuje broń jądrową, a jednocześnie stara się uchronić ludzkość od skutków ocieplenia klimatu.

Obliczenia nietradycyjne: Tianhe-1A i Tytan

Pierwszy najszybszy komputer oparty nie na CPU, lecz głównie na GPU (*Graphics Processing Unit*, czyli procesor graficzny) powstał w 2010 roku w ChRL; był to niewymieniony w naszej tabeli Tianhe-1A (Droga Mleczna 1A) w mieście Tiancin. Dwa lata później podobną maszynę numer 1 zbudowała firma Cray w ośrodku jądrowym w Oak Ridge, Tennessee. Tianhe-1A miał 7 tys. kart graficznych Nvidia Tesla generacji Fermi, zaś Tytan 19 tys. kart Nvidia Tesla generacji Kepler. Nieprzypadkowo te właśnie procesory znajdują się w użytkowych kartach graficznych do gier komputerowych. GPU rozwinęły się, kiedy CPU przestał być wystarczająco wydajny, aby tworzyć kadry animacji kilkadziesiąt razy na sekundę. GPU stały się wielordzeniowe i bardzo wielowątkowe. Paradoksalnie na początku obecnego wieku CPU w komputerze naukowca nie nadążał w arytmetyce za GPU służącym do zabawy. Brakowało jednak możliwości łatwego programowania GPU. Problem ten rozwiązali inżynierowie Nvidii 10 lat temu. Do obliczeń naukowo-technicznych na swych kartach graficznych firma udostępniła darmowo rozszerzenie języka C o nazwie CUDA, dodając stopniowo całą gamę pomocy dla programistów, takich jak biblioteki współbieżnego oprogramowania matematycznego. Później powstał OpenCL (*Open Compute Language*), w zamierzeniu język programowania wszystkich urządzeń obliczeniowych. Naukowcy przy użyciu masowej paralelizacji swych symulacji, gdzie liczba równoległych wątków sięga wielu tysięcy, dokonywali niemal cudów. GPU nie są co prawda tak inteligentne, jak CPU, lecz masowo powtarzając proste, identyczne sekwencje instrukcji na zmieniających się danych, są najefektywniejsze. Jednak programowanie GPU jest bardziej złożone niż programowanie CPU, a pewne zadania wręcz nie znoszą rozdrobnienia. Obiad dla jednej rodziny kilku kucharzy zrobi szybciej niż jeden, ale dwustu powolnych i niezbyt inteligentnych kucharzy tylko spowolni tę pracę. Tytan ma setki tysięcy „kucharzy”, dlatego receptury (programy) dla niego są wybierane bardzo uważnie. Liczy naraz tylko około 5 zoptymalizowanych zadań; w sumie tylko około 30 rocznie. Symulowano m.in. fizykę spalania w projektowanych silnikach diesla, elektrownie jądrowe, nowe polimery, przewidywano zmiany klimatu. Symulowano wybuchy supernowych oraz dynamikę naszej Galaktyki tak dokładnie, że każda z 200 mld gwiazd miała odpowiedniczkę w komputerze!

Droga Mleczna-2 i procesory MIC

Tianhe-2, obecnie numer 2 w rankingu, oparto w 2013 roku na nowej platformie obliczeniowej MIC (*Many Integrated Cores*, liczne zintegrowane rdzenie). To klasa procesorów firmy Intel, znana też jako Xeon Phi (koprocesor KNC, czyli Knights Corner z 2013 r., oraz procesor KNL, czyli Knights Landing z 2016 r., który ma wszystkie zasadnicze funkcje CPU i może zastępować CPU). MIC to doprowadzony

do granic efektywności schemat CPU, o standardowej architekturze 64-bitowej x86, z liczbą rdzeni pomiędzy 57 a 72. Rdzenie są bardziej złożone niż w GPU, ale prostsze niż w CPU. Nie mają, na przykład, zdolności przestawiania kolejności instrukcji w programie dla jego przyspieszenia ani przewidywania, jak potoczy się dalsze wykonanie programu, aby zawczasu zażądać danych i instrukcji z pamięci. Dlatego MIC „lubi" pracochłonne, ale nieskomplikowane rodzaje obliczeń, np. niektóre macierzowe, i jest w nich około dwukrotnie szybszy niż CPU zbliżonej generacji. Warto porównać tu procesory MIC i GPU. W biologii zachodzi zjawisko ewolucji konwergentnej, kiedy zupełnie różne gatunki przybierają zbliżony wygląd i zachowanie zgodne z wymaganiami środowiska. Podobna ewolucja MIC i GPU spowodowała, że mimo niekompatybilnych schematów budowy wewnętrznej ich karty obliczeniowe trudno odróżnić, zawierają równie wielką liczbę tranzystorów (5–10 mld), zbliżone liczby fizycznych rdzeni (nazywanych w GPU SMP, multiprocesorami symetrycznymi), mają to samo ograniczenie na zużywaną moc i wymuszoną przez to częstotliwość zegara $f = 1$–$1{,}5$ GHz, i w końcu – różniącą się zazwyczaj nie więcej niż dwukrotnie moc obliczeniową. To świadczy dobrze o inżynierach, którzy i w MIC, i w GPU optymalnie wykorzystują tranzystory dostępne w danej technologii procesorowej. Są jednak istotne dla użytkownika różnice. CPU i MIC liczą zarówno w pojedynczej, jak i podwójnej precyzji (7 i 15 dziesiętnych miejsc znaczących). Projektanci GPU poświęcali natomiast podwójnej precyzji niewiele obwodów (są niepotrzebne do gier ani do sztucznej inteligencji) i to spowalnia karty graficzne w zastosowaniach naukowych. W odróżnieniu od kart graficznych karty Xeon Phi uruchamiają odchudzony system operacyjny Linux, stając się komputerami wewnątrz komputera, z własnym adresem sieciowym. Z czasem różnice między CPU i MIC znikną, nie zajdzie to zaś w przypadku GPU (innych niż firmy Intel).

Wracając do Tianhe-2, warto zauważyć, że użyto w nim koprocesorów MIC z 57 rdzeniami w tak zmasowanej liczbie (3 mln rdzeni), że detronizacja tego systemu z pierwszego miejsca rankingu zajęła nietypowo długo, bo aż 3 lata. Historia Tianhe-2 jest frapująca. Intel planował rozszerzyć swe wpływy w Azji, oferując procesory KNC instytutowi Ludowej Armii Chin. Tianhe-2 powstał tam na bazie chińskich łączy i płyt głównych oraz najnowszych produktów Doliny Krzemowej. Na rok 2015 zaplanowano podwojenie mocy obliczeniowej. Intel zaczął produkcję dodatkowych procesorów. Wtedy władze USA, nie tłumacząc, czy chodzi o ekonomię, czy bezpieczeństwo, zakazały eksportu technologii MIC Intela do „czarnej listy" odbiorców w ChRL. Kiedyś taka reakcja miała szansę powodzenia. Obecnie, jak zobaczymy, spaliła na panewce. Intel pozbywając się po niskiej cenie nadmiarowych procesorów do Tianhe-2, pozwolił autorowi niniejszego tekstu zbudować eksperymentalny klaster wyszczególniony w ostatniej linii tabeli jako SciPhi, oparty na MIC, GPU i CPU. Ma moc obliczeniową równą 1/25 Prometeusza i nie potrafiłby realizować wszystkich zadań przez niego wykonywanych, za to jest aż 200 razy tańszy.

Sunway TaihuLight, obecny lider rankingu

Chiny rozwinęły produkcję własnych procesorów o parametrach konkurencyjnych w stosunku do objętych embargiem mikroprocesorów Xeon Phi. Jesienią 2016 roku dwie godziny drogi na zachód od Szanghaju, w mieście Wuxi, powstał największy superkomputer świata Sunway TaihuLight, oparty na 41 tys. chińskich procesorów SW26010. Jest szybszy od Tianhe-2 aż o czynnik 3. Po pół roku działania umożliwił zespołowi fizyków, meteorologów i programistów obliczenie na 10 mln rdzeni dynamiki atmosfery, przy użyciu nowego algorytmu wyróżnionego nagrodą Gordona Bella, przyznawaną za najlepsze obliczenia równoległe. Komputer pomoże inżynierom w projektowaniu budowli i symulacji nowych technologii, chemikom w badaniu trójwymiarowej struktury wielkich molekuł ważnych dla życia, a ekonomistom w prognozowaniu zmian gospodarki. TaihuLight jest znakiem nowych czasów: po raz pierwszy ChRL posiadała w 2016 r. nie tylko dwie największe maszyny, ale też więcej komputerów na liście TOP500, o większej sumarycznej mocy obliczeniowej niż USA.

W niedalekiej przyszłości

Przewaga Chin chwilowo może się dodatkowo zwiększyć, gdy obecny rząd USA zrealizuje zapowiedź głębokich cięć budżetowych. Jednak zmiany są bardziej długofalowe i nieuniknione: udział USA w obliczeniach superkomputerowych zmniejszył się w ciągu ostatniej dekady o połowę, do 31%, podczas gdy Chin – wzrósł z 3% do 37%. Ameryka będzie walczyła o odzyskanie pierszeństwa, ale jak to powiedział kiedyś bejsbolista amerykański L. Berra, „przyszłość nie jest już taka, jak kiedyś". A Polska? Porównanie polskich zasobów superkomputerowych do światowych nie wypadło w 2016 roku najgorzej. Udział Polski w mocy obliczeniowej 500 superkomputerów wyniósł około 1%. To 3 razy mniej niż wkład Wielkiej Brytanii albo Francji, i 5 razy mniej niż Niemiec. Jednak Polska wyprzedziła m.in. Kanadę, Hiszpanię, Szwecję, Indie i Rosję, zajmując 11. miejsce w rankingu krajów. To powód do zadowolenia, choć to na pewno nie nasz kraj zbuduje pierwszy system o wydajności eksaflopowej (eksa = 10^{18}, chodzi więc o miliard miliardów działań na sekundę – dla porównania, eksa to więcej niż liczba sekund, które upłynęły od Wielkiego Wybuchu). Kilka krajów: Chiny, Japonia i USA, są na drodze do realizacji tego symbolicznego celu w roku 2020. W następnym odcinku z tej serii przedstawimy superkomputer w zupełnie innej skali, do zrealizowania u siebie w domu i zastosujemy go do słynnego problemu grawitacyjnego N ciał.