Архітектура GPU вузлів обчислювального кластера

Головинський А.Л., Маленко А.Л., Горенко С.О., Бандура О.Ю.

Інститут кібернетики ім. В.М. Глушкова НАН України, просп. Глушкова, 40, Київ, Україна

golovinsky.andriy@gmail.com, icybcluster@gmail.com

Анотація. *У роботі досліджується ефективність гібридних архітектур обчислювальних вузлів на прикладі тесту Linpack з метою пошуку оптимальної конфігурації в сенсі збалансованості технічних характеристик та загальної вартості.*

Ключові слова

Гібридні обчислення, GPU прискорювачі, обчислювальний кластер, тест Linpack.

1 Вступ

Обчислювальні прискорювачі з'явились у 2006-2007 роках на основі ігрових відеоприскорювачів, і на сьогодні ми спостерігаємо становлення та бурхливий розвиток обчислювачів цього типу. Знаковим кроком у розвитку таких систем стала поява тесту Linpack з підтримкою GPU [1]. Поява Linpack дозволила об'єктивно оцінити ефективність прискорювачів та рейтингувати комп'ютерні системи, побудовані на їх основі.

До останнього часу такі гібридні системи були поза розглядом при побудові систем надвеликого рівня державного значення. На основі гібридних вузлів в основному будувались невеликі тестові сервери та кластери для робочих груп, які не потрапляли до рейтингів Top500 [2] найпотужніших суперкомп'ютерів світу та Top50 суперкомп'ютерів СНД [3].

Природно постала задача вибору оптимальної архітектури на основі гібридних вузлів, що складаються зі звичайних процесорів та GPU, для побудови великих обчислювальних систем.

Для дослідження було обрано два вузли з наступними характеристиками:

- платформа Supermicro SYS-6016GT-TF-FM205;
- два центральних процесора Intel Xeon X5675 3.06 GHz;
- два прискорювачі NVidia Tesla M2050;
- 24 Gb оперативної пам'яті (6xDDR-3 4Gb 1333 MHz ECC Unbuffered);
- інтерконект Infiniband DDR (20 Gbit/sec);
- інтерконект 1G Ethernet (1 Gbit/sec).

Пікова продуктивність обраного процесора дорівнює 73.44 GFlops, прискорювача – 515 GFlops. Мета роботи – дослідження ефективності гібридних архітектур обчислювальних вузлів на прикладі тесту Linpack та пошук оптимальної конфігурації в сенсі збалансованості технічних характеристик та загальної вартості.

2 Методика дослідження

Ми вимірювали залежність продуктивності сервера від швидкості процесора, кількості ядер та обсягу оперативної пам'яті.

Більшість сучасних процесорів підтримує функцію управління тактовою частотою, що дозволяє змінювати енергоспоживання процесора у залежності від поточного навантаження. Цією функцією можна скористатись для запуску тестів з різною тактовою частотою процесора.

В linux-подібних ОС управління частотою можна здійснювати через sysfs. На рис. 1 показано вибір доступних частот процесорів та встановлення частоти 2,8 GHz.

```
# cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_available_frequencies
3067000 2933000 2800000 2667000 2533000 2400000 2267000 2133000
2000000 1867000 1733000 1600000
# for i in /sys/devices/system/cpu/cpu*/cpufreq/scaling_setspeed ;
do echo 2800000 > $i ; done
```

Рис. 1. Отримання доступних частот CPU і встановлення частоти 2,8 GHz

Продуктивність кластера у тесті Linpack збільшується при збільшенні матриці задачі, хоча це зростання не є лінійним [4]. Звідси випливає залежність максимальної продуктивності від обсягу оперативної пам'яті: щоб отримати найбільше значення продуктивності при тесті Linpack, намагаються обирати матриці максимально допустимого розміру, які можна вмістити в оперативну пам'ять.

Під час проведення тестів фактичний обсяг пам'яті не змінювався, змінювались лише налаштування тесту Linpack, у результаті чого тест використовував заданий обсяг пам'яті.

Розглянемо ці налаштування більш детально. Нехай N – розмір матриці. Чим більше N, тим більше арифметичних операцій з плаваючою комою необхідно виконати для її обрахунку. N обмежений обсягом оперативної пам'яті, який операційна система може виділити для тесту.

Один елемент матриці типу double займає 8 байт оперативної пам'яті, тому загальний обсяг пам'яті для матриці розміру N дорівнює $8N^2$. При визначенні максимального розміру матриці ми резервували 1 Gb оперативної пам'яті для операційної системи та службових процесів. Тому розмір матриці N для заданого обсягу пам'яті M Gb обчислювався за формулою:

$$N = \sqrt{2^{30}(M-1)/8}.$$

Linpack, зібраний з підтримкою GPU, працює таким чином, що один процес використовує один GPU. Таким чином, кількість GPU, задіяних у тесті, визначається кількістю запущених процесів. Кількість процесів, що буде запущено під час тесту, задається за допомогою параметра "-n" команди mpirun.

Для того, щоб у одному процесі можна було задіяти декілька процесорних ядер, обчислення в процесі розпаралелюються за допомогою потоків. Оскільки для обчислень на процесорі використовується BLAS із Intel MKL, то кількість потоків на один процес виставляється за допомогою змінної оточення MKL_NUM_THREADS.

Наприклад, якщо потрібно задіяти в тесті 2 GPU і 12 процесорних ядер, то треба запустити 2 процеси по 6 потоків на процес, тобто перед запуском необхідно виконати команду "export MKL_NUM_THREADS=6", а при запуску команді mpirun передати параметр "-n 2".

Параметри продуктивності подібних CPU в тесті Linpack детально досліджені, див. [4, 5]. Тому ми обмежимося базовими налаштуваннями, щоб досягти більш стабільних результатів.

Сучасні процесори фірми Intel підтримують технологію TurboBoost, яка дозволяє тимчасово збільшувати частоту процесорних ядер понад стандартну. Зокрема, ця технологія запускається, коли частина ядер звільняються, а решта починає працювати у форсованому режимі. Для однорідності вимірів при зміні кількості ядер ми відключили цей режим.

3 Аналіз ефективності архітектури гібридного вузла

Гібридний тест Linpack одночасно використовує центральні процесори та прискорювачі. Наближено вважаємо, що загальна продуктивність вузла дорівнює сумі продуктивностей центральних процесорів та прискорювачів окремо. Доданки будемо називати чистою продуктивністю CPU та GPU відповідно.

Бачимо (табл. 1, 2), що обсяг оперативної пам'яті сервера сильно впливає на продуктивність сервера в цілому. Особливо це помітно на комплектації 2 CPU + 2 GPU, що додатково вказує на необхідність використання якомога більшого обсягу оперативної пам'яті з розрахунку на один прискорювач. При збільшенні обсягу оперативної па-

Обсяг оперативної пам'яті, Gb	6	12	24
R_{max}, 1 CPU + 1 GPU	323	356	372
R_{max}, 2 CPU + 1 GPU	363	412	440
R_{max}, 2 CPU + 2 GPU	537	626	690

Табл. 1. Загальна продуктивність сервера в залежності від обсягу оперативної пам'яті, GFlops

Обсяг оперативної пам'яті, Gb	6	12	24
R_{max}, 1 GPU (1 CPU)	254	287	303
R_{max}, 1 GPU (2 CPU)	225	274	302
R_{max}, 2 GPU (2 CPU)	399	488	552

Табл. 2. Чиста продуктивність GPU в залежності від обсягу оперативної пам'яті, GFlops

м'яті з 6 Gb до 12 Gb на GPU продуктивність системи зростала на 10-13%. При збільшенні обсягу до 24 Gb на GPU спостерігалось додаткове збільшення продуктивності на 5%.

Можна очікувати, що при наступному подвоєнні обсягу пам'яті на один GPU, додаткове збільшення продуктивності складе 2-3%. Отже, на нашу думку, мінімальним обягом оперативної пам'яті є 12 Gb, а оптимальним – 24 Gb з розрахунку на один GPU. Тут "оптимальність" розуміється в сенсі співвідношення ціни та продуктивності.

Внаслідок цього існує обмеження кількості прискорювачів на один вузол. На даний час поширеними серверними модулями пам'яті є 4 Gb та 8 Gb. Платформу, яку ми розглядаємо, можна обладнати максимум 48 Gb або 96 Gb оперативної пам'яті, що дасть можливість ефективного функціонування двох або чотирьох прискорювачів відповідно.

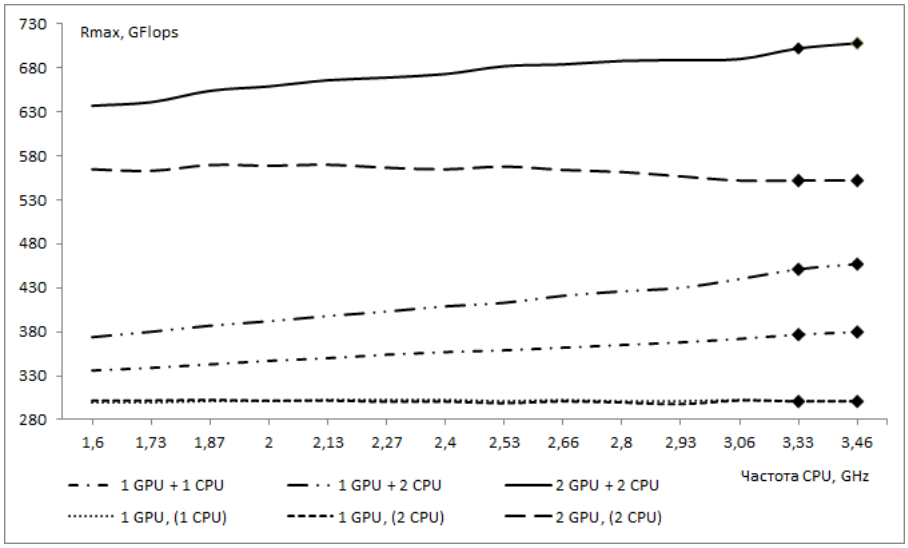

Рис. 2. Продуктивність різних конфігурацій в залежності від частоти CPU

Розглянемо залежність продуктивності від частоти CPU (рис. 2). Помітимо, що у конфігураціях 1 CPU + 1 GPU та 2 CPU + 1 GPU продуктивність самого прискорювача майже незмінна, наближено дорівнює 300 GFlops і не залежить від частоти центрального процесора. Це означає, що прискорювач досягає своєї максимальної продуктивності. Значення, позначені ромбом, є екстрапольованими.

У табл. 3 показано залежність чистої продуктивності прискорювачів від кількості процесорних ядер на один прискорювач. Як бачимо, одного процесорного ядра недостатньо для обслуговування графічного прискорювача. Потрібно не менше двох процесорних ядер на один прискорювач. З рис. 3 бачимо, що спостерігається ефект насичення на двох та чотирьох ядрах. Чиста продуктивність прискорювача на 6-ти ядерній конфігурації спадає.

Кількість CPU ядер на один прискорювач	1	2	4	6
R_{max}, 1 GPU	281	301	305	302
R_{max}, 2 GPU	500	584	588	552

Табл. 3. Продуктивність прискорювачів в залежності від кількості CPU ядер, GFlops

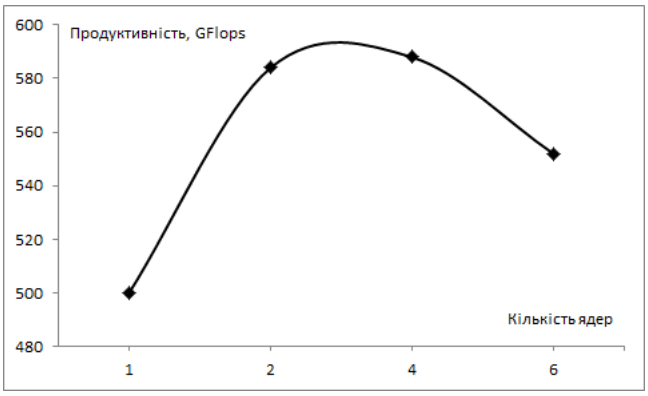

Рис. 3. Продуктивність двох прискорювачів в залежності від кількості процесорних ядер на один прискорювач

Розглядаючи рис. 4, для конфігурації 2 CPU + 2 GPU також спостерігаємо невелике спадання чистої сумарної продуктивності зі зростанням частоти CPU до 2,8–3,06 GHz.

На нашу думку, ці два ефекти пояснюються тим, що при збільшенні частоти та кількості ядер центральний процесор "забирає собі" більше роботи. Додаткова причина подібної поведінки прискорювачів – брак оперативної пам'яті, який відчувається у конфігурації з двома прискорювачами і не настільки помітний для випадку одного GPU (табл. 3). Це стримує повне розкриття усіх можливостей прискорювачів у конфігурації 2 CPU + 2 GPU.

Розглянемо продуктивність роботи кластера з двох вузлів. Ми досліджували їх роботу на основі інтерконекту Infiniband DDR (20 Gbit/sec) та Ethernet (1 Gbit/sec). Значення теоретичної (пікової) та реальної продуктивності показані в табл. 4.

Традиційно для грубої оцінки продуктивності майбутнього кластера на звичайних CPU вузлах спеціалісти Інституту кібернетики імені В.М. Глушкова НАНУ користуються емпіричною формулою, запропонованою С.Г. Рябчуном та А.О. Якубою, яка заснована на спостереженнях поведінки продуктивності кластерного комплекса СКІТ для різної кількості вузлів:

$$\hat{R}_{max} = \alpha n C,$$

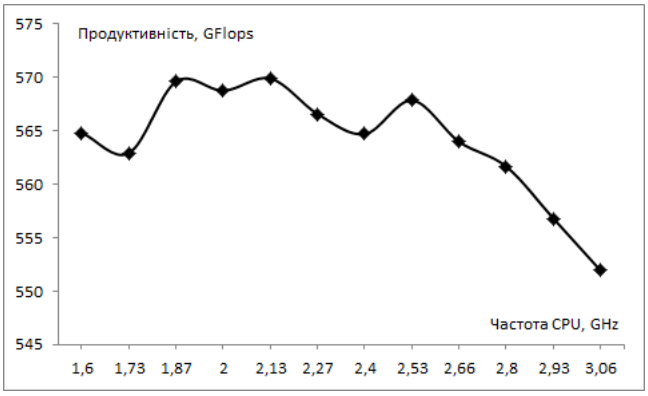

Рис. 4. Продуктивність двох прискорювачів в залежності від частоти CPU

	R_{peak}	R_{max}	α	β
вузол 2 CPU	147	138	0,94	
вузол 2 CPU + 1 GPU	662	440	0,94	0,59
вузол 2 CPU + 2 GPU	1177	690	0,94	0,54
кластер: 2 вузла без GPU, Infiniband	294	257	0,87	
кластер: 2 вузла по 1 GPU, Infiniband	1324	847	0,87	0,57
кластер: 2 вузла по 2 GPU, Infiniband	2354	1178	0,87	0,45
кластер: 2 вузла без GPU, Ethernet	294	219	0,75	
кластер: 2 вузла по 1 GPU, Ethernet	1324	765	0,75	0,53
кластер: 2 вузла по 2 GPU, Ethernet	2354	749	0,75	0,26

Табл. 4. Продуктивність гібридних вузлів у кластері, GFlops

де n – кількість вузлів, C – пікова продуктивність одного CPU вузла, α – коефіцієнт ефективності використання обладнання для такого типу архітектури, $0 < \alpha < 1$.

Для кластера з гібридними вузлами ми пропонуємо користуватись наступною формулою:

$$\hat{R}_{max} = \alpha n C + \beta m G,$$

де m – кількість прискорювачів, G – пікова продуктивність одного прискорювача, β – коефіцієнт ефективності використання GPU прискорювачів.

Отримані коефіцієнти використання обладнання α та β, обчислені для різних конфігурацій, показані в табл. 4. Бачимо, що якість інтерконекту для такої архітектури є критичною. При переході з конфігурації з одним GPU на вузол до конфігурації з двома GPU на вузол спостерігається значне зменшення ефективності використання GPU.

Можна очікувати, що для кластера з більшою кількістю вузлів, а також більшою кількістю GPU на вузол, для ефективного використання потужностей прискорювачів обов'язковим буде використання Infiniband QDR (40 Gbit/sec) або навіть Infiniband FDR (56 Gbit/sec).

4 Оцінка відношення ціна/продуктивність

При проектуванні високопродуктивних обчислювальних систем часто постає питання: яку продуктивність можна отримати, маючи певний обсяг фінансування? Або навпаки: який обсяг фінансування потрібен для отримання певної продуктивності?

Для того, щоб наближено відповісти на таке питання, ми провели аналіз співвідношення вартості одного сервера та його продуктивності в залежності від частоти центрального процесора.

Ми виходили з ціни платформи та цін на процесори за умови придбання їх за безготівковим розрахунком для організацій. Сумарна вартість сервера з процесорами Intel Xeon 5600 series на момент написання статті та продуктивність такої конфігурації наведена в табл. 5. Дані продуктивності для частоти центрального процесора 3,33 GHz та 3,46 GHz є екстрапольованими, виходячи з даних для частот 1,6–3,06 GHz (в таблиці вони позначені зірочкою).

Частота процесора, GHz	2,66	2,8	2,93	3,06	3,33	3,46
Вартість сервера, тис. грн.	75,1	79,44	81,6	84,6	88,1	90,6
R_{max} сервера, TFlops	684	688	689	690	702*	708*
Ціна 1 TFlops, тис. грн.	109,8	115,5	118,4	122,6	125,5	128,0

Табл. 5. Вартість сервера для різних процесорів

Як бачимо (рис. 5), найбільший економічний ефект дає використання молодших моделей з лінійки процесорів (значення, позначені ромбом, є екстрапольованими).

Зауважимо, що в табл. 5 не враховується вартість кластерної інфраструктури, яка також є значною при побудові великої обчислювальної системи. Проектуючи кластер на основі великої кількості менш продуктивних, але дешевших вузлів, архітектор стикається з проблемою збільшення кількості шаф, кондиціонерів, комутаторів інтерконекту тощо, що в свою чергу може нівелювати економію на дорогих процесорах.

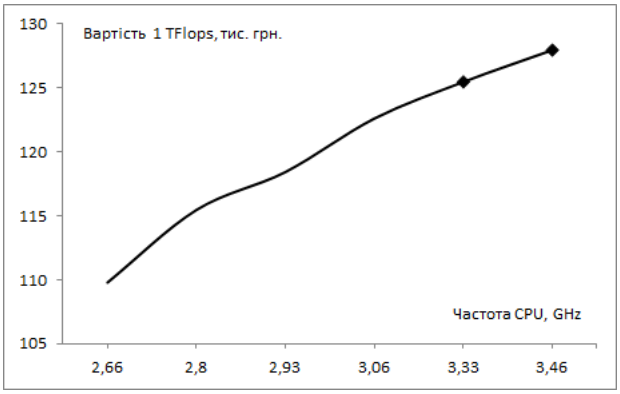

Рис. 5. Вартість одного TFlops в залежності від частоти CPU

5 Висновки

Гібридний вузол є високопродуктивним обчислювальним елементом, придатним для побудови великих кластерних систем.

Продуктивність обчислень гібридного вузла не сильно залежить від швидкості центрального процесора, сильно залежить від обсягу оперативної пам'яті. Для хороших результатів треба мати не менше 12 Gb оперативної пам'яті на один прискорювач. Оптимальним є використання 24 Gb оперативної пам'яті на один прискорювач.

Для ефективної роботи гібридних вузлів у складі обчислювального кластера обов'язковим є використання інтерконекту з високою пропускною здатністю, наприклад, Infiniband QDR або FDR. Крім того, небажаним є використання великої кількості GPU карт в одному вузлі, оскільки інтерконект стане вузьким місцем при обміні даних між вузлами, що не дозволить повністю розкрити можливості прискорювачів.

З точки зору економічної ефективності при побудові кластера, оптимально використовувати молодші моделі з лінійки процесорів, оскільки в такому разі вартість побудови кластера заданої продуктивності є найменшою.

Цікавими питаннями, які залишились поза розглядом, є дослідження інших параметрів системи та тесту Linpack, їх впливу на продуктивність, та отримання рекордних показників продуктивності на вузлах з графічними прискорювачами.

Гібридні вузли значно збільшили загальну продуктивність комплексу СКІТ Інституту кібернетики імені В.М. Глушкова НАНУ.

Література

[1] http://www.nvidia.com/content/PDF/sc_2010/theater/Phillips_SC10.pdf

[2] http://top500.org

[3] http://supercomputers.ru/

[4] http://www.netlib.org/benchmark/hpl/faqs.html

[5] Коваль В.Н., Рябчун С.Г., Сергиенко И.В., Якуба А.А. Суперкомпьютерные кластерные системы - организация вычислительного процесса. Проблеми програмування, НАНУ, Ін-т програмних систем, Київ, т. 2–3, 2006, с. 197–210.