УДК 681.321

ВИЗНАЧЕННЯ КОНФІГУРАЦІЇ ОБЧИСЛЮВАЛЬНИХ КОМПЛЕКСІВ ВИСОКОПРОДУКТИВНОЇ ОБРОБКИ ВЕЛИКИХ ОБ'ЄМІВ ДАНИХ

Дюжаєв Л.П., к.т.н., доцент; Соколов Д.Ю., магістрант
Національний технічний університет України
«Київський політехнічний інститут», м. Київ, Україна

Вступ

Високопродуктивні обчислювальні комплекси сьогодні активно проникають у всі сфери людської діяльності, дозволяючи успішно вирішувати все більш важкі наукові та виробничі завдання, пов'язані зі складними обчисленнями. Такі завдання відносять до класу високопродуктивних обчислень (high-performance computing, HPC), а для їх вирішення використовуються спеціалізовані суперкомп'ютерні системи [1], побудовані на базі сотень процесорів. З усієї безлічі задач HPC особливо виділяються такі, вирішення яких крім виконання трудомістких обчислень пов'язано з переробкою величезних обсягів даних, що обчислюються терабайтами. Специфіка даного класу задач полягає в тому, що для їх вирішення необхідний не тільки високопродуктивний обчислювальний кластер (ОК), а й високошвидкісна система зберігання даних (СЗД). При виборі конфігурації обчислювальної системи, що складається з обчислювального кластеру і системи збегрігання даних, виникає завдання комплексної оцінки її продуктивності для визначення узгоджених між собою технічних характеристик кластера та системи зберігання. Причому практичне застосування таких систем показало, що збільшення обчислювальної потужності кластера часто не тільки не призводить до зростання загальної продуктивності, але і дає зворотний ефект [2].

Постановка задачі

У даній статті наводиться математична модель обчислювального комплексу (як мережі масового обслуговування), а також описується метод визначення оптимальної (за критерієм продуктивності) конфігурації обчислювального комплексу на основі результатів стандартних промислових тестів продуктивності ОК і СЗД.

Базові засоби побудови математичної моделі

Відзначимо важливі особливості класу задач, що розглядаються.

1. Обсяг оброблюваних даних істотно перевищує обсяг пам'яті обчислювальних вузлів, тому всі дані розбиваються на рівні блоки фіксованого розміру, що розподіляються по томам СЗД, а процес

вирішення завдання поділяється на цикли таким чином, що один вузол ОК обробляє один блок даних за один цикл. Кількість циклів при цьому набагато більше, ніж число вузлів кластера. У кожному циклі інформація проходить в системі по замкнутому контуру: спочатку вихідні дані зчитуються з дисків системи зберігання і завантажуються у вузли кластера, потім вони обробляються обчислювальним вузлом, після чого результати записуються на диск. Процес розв'язання такого завдання кластером показаний на часовій діаграмі (рис. 1).

2. Всі блоки даних ідентичні (будь-який вузол ОК може обробляти будь-який блок даних) і незалежні, що виключає необхідність обміну інформацією між обчислювальними вузлами.

3. Час ініціалізації завдання (розподілу завдань по вузлах ОК і блоків даних по томам СЗД) дуже малий в порівнянні із загальним часом рішення задачі. Затримками, пов'язаними з утворенням черг в комутаторі також можна знехтувати, оскільки в якості комутаційної середовища використовуються високошвидкісні інтерфейси, пропускна спроможність яких на порядок більше ніж пропускна здатність вузлів ОК і контролера СЗД.

Рис.1. Часова діаграма розв'язання завдання кластером

Визначення оптимальної конфігурації обчислювального комплексу

Математична модель комплексу представлена на рис. 2 і являє собою замкнуту стохастичну мережу масового обслуговування (МеМО). Джерелом запитів в мережі є N вузлів (ядер) ОК.

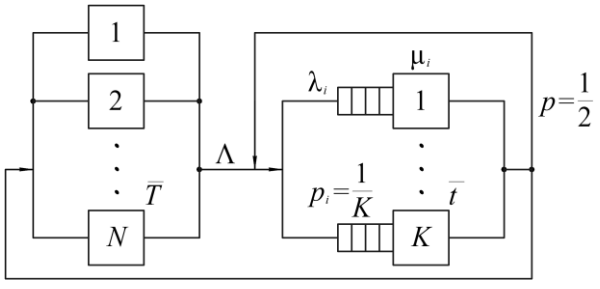

Рис.2. Структурна схема математичної моделі обчислювального комплексу

Всі запити надходять в СЗД, представлену як K незалежних томів, причому кожний запит проходить її двічі (що моделюється петлею зворотного зв'язку з імовірністю 0.5). Розподіл часу обслуговування запиту в системі зберігання є суттєво не експоненціальним. Потік запитів на вході і відповідно розподіл інтервалів часу надходження запитів в систему зберігання також не є експоненціальним. Система зберігання являє собою набір незалежних

томів, кожен з яких має власну необмежену чергу. Використовуючи метод контурів [3] можна скласти нелінійне рівняння балансу запитів в мережі.

Розглянутий в моделі контур замкнутий, а джерелами запитів є тільки обчислювальні вузли (причому новий запит від кожного вузла генерується тільки після обробки попереднього запиту того ж вузла). Рівняння балансу:

$$N \cong \Lambda \overline{T} + 2\Lambda \bar{t}\left[\frac{2\Lambda \bar{t}}{K - 2\Lambda \bar{t}}\left(\frac{C_\lambda^2 + C_\mu^2}{2}\right) + 1\right] \qquad (1)$$

Роз'вязок рівняння (1) щодо Λ дозволить визначити значення інтенсивності надходження заявок в мережу (тобто саме значення Λ) та інші середні характеристики функціонування МеМО ($\overline{T}$ та $\bar{t}$).

Для знаходження оптимальної конфігурації комплексу (кількості ядер ОК і томів СЗД), при якому час виконання завдання буде мінімальним, потрібно сформувати цільову функцію і вирішити задачу оптимізації. Цільовою функцією буде залежність загального часу рішення задачі від параметрів МеМО.

$$T_{FULL} \cong \overline{T}\frac{M}{N} + 2\bar{t}\frac{M}{K}\left[\frac{2\Lambda \bar{t}}{K - 2\Lambda \bar{t}}\left(\frac{C_\lambda^2 + C_\mu^2}{2}\right) + 1\right], \text{ де M - кількість циклів.}$$

В даному випадку мінімізація часу вирішення всього завдання рівносильна мінімізації середнього часу одного циклу, а цільова функція приймає вигляд:

$$T_{CYCLE} \cong \frac{\overline{T}}{N} + \frac{2\bar{t}}{K}\left[\frac{2\Lambda \bar{t}}{K - 2\Lambda \bar{t}}\left(\frac{C_\lambda^2 + C_\mu^2}{2}\right) + 1\right] \qquad (2)$$

Для знаходження екстремуму цільової функції необхідно знайти першу похідну змінної N, а потім прирівняти вираз похідної до нуля і знайти рішення отриманого рівняння. Для обчислення диференціала цільової функції по N потрібно підставити в неї вираз Λ (N), який визначається з нелінійного рівняння балансу запитів в системі (1).

Спочатку розглянемо залежність $\Lambda(N)$. Як видно з рівняння (1) Λ залежить не тільки від N, а також від T, t і K. Побудуємо якісно графіки функції $f = \Lambda(N)$, для різних співвідношень величин T і t (K=const). Вони представлені на рис. 3. Вигляд кожного графіка залежить від співвідношення величин T і t. Чим більше відношення часів T/t, тим крутіший графік, і що найголовніше тим ближче графік наближається до своїх асимптот, однією з яких є пряма $\rho = 1$ відповідна максимальному завантаженні систем масового обслуговування (СМО).

Другою асимптотою є пряма: $\Lambda = \frac{N}{\overline{T} + 2\bar{t}}$. Оскільки при вирішенні завдань НРС час розрахунку істотно перевершує час вводу-виводу, то

завантаження СЗД далеке від одиниці, а для наближених обчислень можна використовувати асимптотичне значення: $\Lambda \approx \frac{N}{\overline{T} + 2\bar{t}}$.

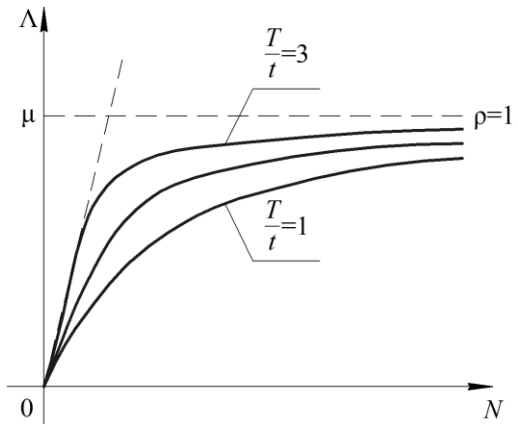

Рис.3. Функція $f=\Lambda(N)$ при різних відношеннях T/t

Крім того, у виразі (2) можна спростити доданок, що містить коефіцієнти варіації. Як показано в роботі [4] коефіцієнт варіації C_μ^2 для дискового масиву дорівнює приблизно 0,15. Важливо зазначити, що час обчислення в кластері є випадковою величиною, з функцією розподілу відмінною від експоненти. Таким чином, сумарний потік являє собою суму випадкових потоків з коефіцієнтами варіації менше одиниці, тобто сходиться до найпростішого потоку знизу, отже, $C_\lambda^2 \to 1$ залишаючись меньше одиниці. Узагальнивши ці твердження, отримаємо: $\frac{C_\lambda^2 + C_\mu^2}{2} \cong \frac{1}{2}$.

З урахуванням вказаних припущень функція (2) набуде вигляду:

$$T_{CYCLE} \cong \frac{\overline{T}}{N} + \frac{2\bar{t}}{K}\left(\frac{K - \bar{t}\frac{N}{\overline{T} - 2\bar{t}}}{K - 2\bar{t}\frac{N}{\overline{T} - 2\bar{t}}}\right) \to \min(N) \qquad (3)$$

Обчисливши похідну функції (3) складемо рівняння:

$-\frac{T}{N^2} + \frac{2\bar{t}}{K}\frac{K\bar{t}(\overline{T} + 2\bar{t})}{\left[K(\overline{T} + 2\bar{t}) - 2N\bar{t}\right]^2} = 0$, розв'язавши яке відносно N отримаємо вираз для оптимального числа обчислювальних вузлів:

$$N_{OPT} \cong K\left(\frac{\overline{T}}{\bar{t}} + 2\right)\frac{1}{2 + \sqrt{2\left(1 + 2\frac{\bar{t}}{\overline{T}}\right)}} \qquad (4)$$

Отримане співвідношення дозволяє легко визначити оптимальне число обчислювальних вузлів кластера при заданих параметрах системи зберігання та відомих тимчасових характеристиках обладнання комплексу. Покажемо тепер, як можна визначити оптимальну конфігурацію комплексу не маючи в наявності даних часових характеристик.

У формулі (4) присутні параметри (середні значення часу обслуговування T та t), для знаходження яких необхідно пов'язати параметри математичної

моделі з характеристиками кластера і дискової системи. Характеристики устаткування можуть бути отримані за результатами стандартних промислових тестів продуктивності. Для обчислювальних кластерів - це тест HPL (Highly Parallel Linpack) з набору тестів HPC Challenge тестів, для систем зберігання даних - тест LFP (Large File Processing) з тестового пакету Storage Performance Council 2 benchmark. Стандартні тести застосовуються для оцінки параметрів продуктивності безлічі промислових систем (їх результати доступні і постійно публікуються); вибір цього способу отримання вихідних даних виправданий і гарний тим, що дозволяє надалі проводити дослідження будь-яких систем.

Покажемо зв'язок параметрів моделі з характеристиками обладнання:

Середній час обслуговування запиту в ОК:

$$\overline{T} = Q\frac{V}{L}, \qquad (5)$$

де L – продуктивність ОК (за тестом HPCC HPL, FLOP/s), V – об’єм оброблюваних даних. Q – питома трудомісткість рішення завдання, що визначає, яка кількість операцій з плаваючою точкою доводиться на один байт оброблюваної інформації.

Саме параметр Q визначає специфіку розв'язуваної задачі (специфіку алгоритму розв'язуваної задачі), тобто кожному класу задач (і кожному методу їх вирішення) буде відповідати своє значення питомої трудомісткості.

Середній час обслуговування запиту в СЗД:

$$\overline{t} = \frac{V}{S}, \qquad (6)$$

де S – пропускна здатність СЗД (за тестом SPC-2 LFP, MB/s), V – об’єм оброблюваних даних.

Якщо підставити в формулу (4) співвідношення (5) і (6), то можна отримати вираз для залежності оптимального числа процесорних ядер кластера від кількості томів системи зберігання та результатів стандартних тестів продуктивності:

$$N_{OPT} \cong K\left(Q\frac{S}{L}+2\right)\frac{1}{2+\sqrt{2\left(1+2\frac{1}{Q}\cdot\frac{L}{S}\right)}} \qquad (7)$$

Висновок

Розглянута мережа масового обслуговування являє собою узагальнену модель обчислювального комплексу, який, крім виконання трудомістких обчислень, також займається переробкою величезних обсягів даних. Отримана формула (7) дозволяє легко визначати оптимальну конфігурацію обчислювального комплексу за характеристиками обладнання, отриманими з результатів стандартних промислових тестів продуктивності

для обчислювальних кластерів і систем зберігання даних.

Література

1. Суперкомпьютерные технологии в науке, образовании и промышленности (второй выпуск) / под редакцией: академика В.А.Садовничего, академика Г.И.Савина, чл.-корр. РАН Вл.В.Воеводина. – М.: Изд. Московского ун-та, 2010. – 208с.

2. Mills R., Sripathi V., Mahinthakumar G., Hammond G., Lichtner P., Smith B. Engineering for Scalable Performance on Cray XT and IBM BlueGene Architectures, 2010.

3. Мясников В.А., Мельников Ю.Н., Абросимов Л.И.. Методы автоматизированного проектирования систем телеобработки данных. – М.: Энергоатомиздат, 1992. – 299с.

4. Lee E.K. Performance Modeling and Analysis of Disk Arrays. University of California at Berkeley, 1993.

Дюжаєв Л.П., Соколов Д.Ю. ***Визначення конфігурації обчислювальних комплексів високопродуктивної обробки великих об'ємів даних****. У даній роботі наводиться математична модель обчислювального комплексу (як мережі масового обслуговування). В якій обсяг оброблюваних даних істотно перевищує обсяг пам'яті обчислювальних вузлів, тому всі дані розбиваються на рівні блоки фіксованого розміру, що розподіляються по томам системи зберігання даних, а процес вирішення завдання поділяється на цикли таким чином, що один вузол обчислювального кластера обробляє один блок даних за один цикл. А також описується метод визначення оптимальної (за критерієм продуктивності) конфігурації комплексу на основі результатів стандартних промислових тестів продуктивності обчислювальних кластерів і систем зберігання даних.*

Ключові слова: *високопродуктивні обчислення, обчислювальний кластер, система зберігання даних, мережа масового обслуговування, математична модель обчислювального комплексу.*

Дюжаев Л.П., Соколов Д.Ю. ***Определение конфигурации вычислительных комплексов высокопроизводительной обработки больших объемов данных****. В данной работе приводится математическая модель вычислительного комплекса (как сети массового обслуживания). В какой объем обрабатываемых данных существенно превышает объем памяти вычислительных узлов, поэтому все данные разбиваются на равные блоки фиксированного размера, которые распределяются по томам системы хранения данных, а процесс решения задачи делится на циклы таким образом, что один узел вычислительного кластера обрабатывает один блок данных за один цикл. А также описывается метод определения оптимальной (по критерию производительности) конфигурации комплекса на основе результатов стандартных промышленных тестов производительности вычислительных кластеров и систем хранения данных.*

Ключові слова: *высокопроизводительные вычисления, вычислительный кластер, система хранения данных, сеть массового обслуживания, математическая модель вычислительного комплекса.*

Dyuzhayev L.P., Sokolov D.Y, ***Determination of the configuration computer systems high-performance handling large volumes of data****. This paper provides a mathematical model of computing complex (a network of queuing). In what volume of processed data outgrows memory compute nodes, so all data are divided into equal blocks of fixed size that are distributed through volumes of data storage and process the task is divided into cycles so that one node computing cluster process one unit data for one cycle. Also describes the method for determining the optimal (by the criterion of productivity) complex configurations based on industry standard benchmarks computing clusters and storage systems.*

Keywords: *high-performance computing, computer cluster, storage system, network queuing, mathematical model of computing complex.*