Тема №8 Програмування для багатоядерних систем

Питання:
1. **Вступ.**
2. **Масивно-паралельні системи.**
3. **Симетричні мультипроцесорні системи.**
4. **Системи з неоднорідним доступом до пам'яті.**
5. **Паралельні векторні процесори.**
6. **Технологія Fork-Join для програмування багатоядерних систем**

Вправи і завдання до теми №8

1. Вступ. Основним параметром класифікації паралельних комп'ютерів є наявність спільної (SMP) чи розподіленої пам'яті (MPP). Щось середнє між SMP і MPP являють собою NUMA - архітектури, де пам'ять фізично розподілена, але на логічному рівні загальнодоступна. Кластерні системи є дешевшим варіантом MPP. За підтримкою команд обробки векторних даних говорять про векторно - конвеєрні процесори, які, у свою чергу можуть об'єднуватися в PVP - системи з використанням загальної чи розподіленої пам'яті. Усе більшу популярність дістають ідеї комбінування різних архітектур в одній системі і побудови неоднорідних систем.

При організації розподілених обчислень у глобальних мережах, наприклад, Інтернет говорять про мета - комп'ютери, які, строго кажучи, не є паралельними архітектурами.

Розглянемо основні типи комп'ютерів за такими ознаками: *особливості архітектури, приклади конкретних комп'ютерів, перспективи масштабованості, типові особливості побудови операційних систем, найхарактерніша модель програмування.*

2. Масивно - паралельні системи (MPP)

Блок-схема MPP наведена на рис.6.1, характеристики – в табл.6.1.

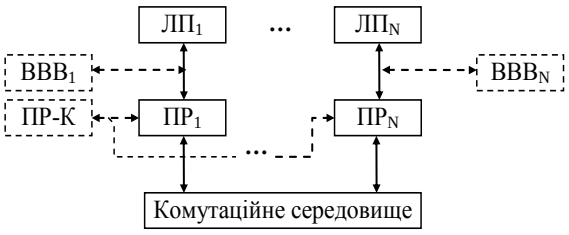

Рис.6.1. Блок-схема MPP

На рис.6.1 позначено: ЛП – локальна пам'ять, ПР – процесор, ПР-К – процесор комутаційний, ВВВ – вузол вводу-виводу. Пунктирними лініями позначені необов'язкові вузли і лінії зв'язку.

Перевагами MPP є хороша масштабованість і висока швидкодія (серед розглянутих схем – найшвидша); недоліками – великий час міжпроцесорного обміну, використання кожним ПР обмеженого об'єму локального банку даних, висока ціна програмного забезпечення.

1. Симетричні мультипроцесорні системи (SMP)

Блок-схема SMP наведена на рис.6.2, характеристики – в табл.6.2.

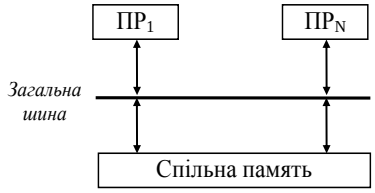

Рис.6.2. Блок-схема SMP

Перевагами SMP є: простота організації, універсальність при програмуванні, простота експлуатації, невисока ціна; недоліком – погана масштабованість.

Таблиця 6.1. Характеристики MPP

Архітектура	Система складається з однорідних *обчислювальних вузлів*, що включають: - один чи кілька центральних процесорів (звичайно RISC); - локальну пам'ять (прямий доступ до пам'яті інших вузлів неможливий); - комунікаційний процесор чи мережний адаптер; - іноді - тверді диски (як у SP) і/чи інші пристрої введення/виведення. До системи можуть бути додані спеціальні вузли введення/виведення і вузли керування. Вузли зв'язані через деяке комунікаційне середовище (високошвидкісна мережа, комутатор і т.п.)
Приклади	IBM RS/6000 SP2, Intel Paragon/Asci Red, SGI/CRAY T3E, Hitachi SR8000, трансп'ютерні системи Parsytec.
Масштабованість	Загальне число процесорів у реальних системах досягає декількох тисяч (Asci Red, Blue Mountain).
Операційна система	Існують два основних варіанти: 1. Повноцінна ОС працює тільки на керуючій машині (front - end), на кожнім вузлі працює сильно урізаний варіант ОС, що забезпечують тільки роботу розташованих в них паралельних задач. Приклад: Cray T3E. 2. На кожнім вузлі працює повноцінна UNIX - подібна ОС (варіант , близький до кластерного підходу). Приклад: IBM RS/6000 SP + ОС AIX, що встановлюються окремо на кожнім вузлі.
Модель програмування	Програмування в рамках моделі передачі повідомлень (MPI, PVM, BSPlib)

Таблиця 6.2. Характеристики SMP

Архітектура	Система складається з декількох однорідних процесорів і масиву загальної пам'яті (звичайно з декількох незалежних блоків). Усі процесори мають доступ до будь-якої комірки пам'яті з однаковою швидкістю. Процесори підключені до пам'яті або за допомогою загальної шини (базові 2 - 4 процесорні SMP - сервери), або за допомогою crossbar - комутатора (HP 9000). Апаратно підтримується когерентність кешів
Приклади	HP 9000 V - class, N - class; SMP - сервери і робочі станції на базі процесорів Intel (IBM, HP, Compaq, Dell, ALR, Unisys, DG, Fujitsu і ін.)
Масштабованість	Наявність загальної пам'яті спрощує взаємодія процесорів між собою, прооте накладає сильні обмеження на їхнє число - не більш 32 у реальних системах. Для побудови масштабованих систем на базі SMP використовуються кластерні чи NUMA - архітектури .
Операційна система	Уся система працює під керуванням єдиної ОС (звичайно UNIX - подібної, але для Intel - платформ підтримується Windows NT). ОС автоматично (у процесі роботи) розподіляє процеси/нитки по процесорах (scheduling), але іноді можлива і явна прив'язка.
Модель програмування	Програмування в моделі загальної пам'яті. (POSIX threads, OpenMP). Для SMP – систем існують порівняно ефективні засоби автоматичного розпаралелення.

2. Системи з неоднорідним доступом до пам'яті (NUMA)

Блок -схема NUMA наведена на рис.6.3, характеристики – в табл.6.3.

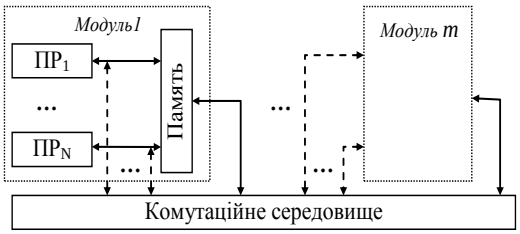

Рис.6.3. Блок-схема NUMA

Переваги системи NUMA забезпечуються зручністю систем з спільною пам'яттю і дешевизною систем з розподіленою пам'яттю.

Таблиця 6.3. Характеристики NUMA

Архітектура	Склад: однорідні базові модулі (комірки) з невеликого числа процесорів і блоку пам'яті. Модулі об'єднані за допомогою високошвидкісного комутатора. Підтримується єдиний адресний простір, апаратно підтримується доступ до віддаленої пам'яті. Доступ до локальної пам'яті в кілька разів швидший, ніж до віддаленої. У випадку апаратного підтримання когерентності кешів у всій системі говорять про архітектуру cc - NUMA (cache - coherent NUMA)
Приклади	HP HP 9000 V - class у SCA - конфігураціях, SGI Origin2000, Sun HPC 10000, IBM/Sequent NUMA - Q 2000, SNI RM600.
Масштабованість	Обмежується об'ємом адресного простору, можливостями апаратури підтримки когерентності кешів і можливостями ОС по керуванню великим числом процесорів.
Операційна система	Система працює під керуванням єдиної ОС, як у SMP. Можливі варіанти коли окремі "частини" системи працюють під керуванням різних ОС (наприклад, Windows NT і UNIX у NUMA - Q 2000).
Модель програмування	Аналогічно SMP.

3. Паралельні векторні системи (PVP)

Блок - схема PVP наведена на рис.6.4, характеристики – в табл.6.4.

Таблиця 6.4. Характеристики PVP

Архітектура	Основна ознака - наявність спеціальних векторно – конвеєрних процесорів, у яких передбачені команди однотипної обробки векторів незалежних даних, що ефективно виконуються на конвеєрних функціональних пристроях. Як правило, кілька таких процесорів (1 - 16) працюють одночасно над спільною пам'яттю (аналогічно SMP) у рамках багатопроцесорних конфігурацій. Кілька таких вузлів можуть бути об'єднані за допомогою комутатора (аналогічно MPP).
Приклади	NEC SX - 4/SX - 5, лінія векторно - конвеєрних комп'ютерів CRAY: CRAY - 1, CRAY J90/T90, CRAY SV1, серія Fujitsu VPP.
Модель програмування	Ефективне програмування має на увазі векторизацію циклів (для досягнення розумної продуктивності одного процесора) і їх розпаралелення (для одночасного завантаження декількох процесорів однією задачею).

Перевагами PVP є висока швидкодія і практично відсутність проблеми взаємодії між процесорами; недоліком – висока вартість.

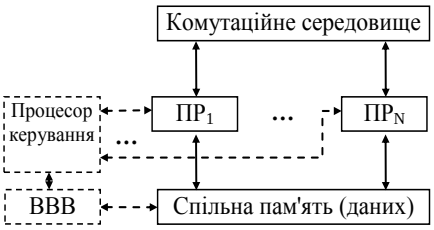

Рис.6.4. Блок-схема PVP

Апаратні та програмні платформи

Вимоги до апаратного забезпечення головного вузла базуються на апаратних вимогах для встановлення

ОС WS2003. Тому обладнання повинно мати підтримку платформи x64.

Мінімальні вимоги апаратного забезпечення головного вузла:

процесор з підтримкою архітектури x64:

- AMD Opteron;
- AMD Athlon 64;
- Intel Xeon з підтримкою Intel EM64T або Intel 64;
- Intel Pentium з підтримкою Intel EM64T або Intel 64;

512 Мб оперативної пам’яті;

4 Гб дискового простору для встановлення системи.

Оптимальна конфігурація апаратного забезпечення така:

процесор з підтримкою архітектури x64 з частотою:

- не менше 2 ГГц для AMD Athlon 64, AMD Opteron, Core 2 Duo, Intel Xeon (Core);
- не менше 3 ГГц для Intel Pentium4, Intel Xeon (NetBurst);

2 Гб оперативної пам’яті.

Для реалізованого проекту використано таку конфігурацію: процесор 1x Pentium D 935 (dual core, 3.2 ГГц, 4 Мб cache); системна плата Asus P5L-VM 1394;

оперативна пам'ять 2x 1 Гб Corsair DDR2-667 (Value RAM);

жорсткий диск складається з трьох розділів: 16 Гб системний, 60 Гб дані користувачів та 8 Гб розділ кластерної ОС.

Вимоги до апаратного забезпечення вузла також базуються на апаратних вимогах для встановлення ОС

WS2003. Мінімальні вимоги для вузла збігаються з мінімальними вимогами для головного вузла.

Оскільки всі операції з обробки паралельних програм користувачів відбуваються на вузлах, від їх конфігурації залежить загальна продуктивність кластера. Збільшення продуктивності можливо або шляхом збільшення кількості вузлів, або модернізацією апаратного забезпечення окремих вузлів. Щодо останнього варіанта, можна дати загальні рекомендації з розширення мінімальної конфігурації та підбору обладнання, щоб отримати оптимальну швидкодію та забезпечити подальшу модернізацію.

Системна плата має задовольняти таким вимогам:

підтримка чотирьохядерних процесорів;

підтримка не менше 4 Гб оперативної пам’яті та двоканального режиму.

Програмне забезпечення для функціонування системи складається із стандартних мережевих служб ОС

WS2003 та програмного комплексу UACluster. Необхідні такі мережеві служби:

Active Directory (AD) – реалізація розподіленої служби каталогів, сумісної з Lightweight Directory Access

Protocol. Призначена для централізованого керування доступом до мережевих ресурсів.

Compute Cluster Pack (CCP) – пакет, що забезпечує функціонування обчислювального кластера під керуванням ОС WS2003. Містить реалізацію Microsoft MPI для обміну повідомленнями між вузлами у процесі паралельних обчислень та набір сервісів і прикладних програм для керування завданнями та адміністрування кластера. Разом з AD є обов'язковими службою для функціонування програмного комплексу UACluster.

Domain Name System – служба доменних імен, що є стандартною службою мережі ОС WS2003 та є частиною стеку протоколів TCP/IP. Призначена для трансформації символьних імен мережевих вузлів та ресурсів у IP-адресу та навпаки. Є обов'язковою службою для функціонування служби каталогів AD.

DHCP – служба динамічного конфігурування вузла. Використовується для автоматичної видачі мережевих налаштувань.

TFTP – служба простої передачі файлів між вузлами без аутентифікації. Використовує передачу фіксованими блоками по 512 байт, як транспортний протокол виступає UDP. Служба є частиною стандартної служби Remote Installation Services (RIS) ОС WS2003. Разом з DHCP є обов'язковою службою для функціонування середовища PXE.

RIS – служба дистанційної інсталяції ОС. Використовує підготовлені образи ОС для подальшого їх

розгортання по мережі на велику кількість вузлів. Може бути використана для розгортання ОС на вузлах.

Windows Deployment Services – наступник RIS, який підтримується в Service Pack 2 для сімейства WS2003,

є стандартним для сімейств Vista та Longhorn. В пакеті CCP підтримується, починаючи з Service Pack 1 (SP1).

Як основну мережеву технологію обрано Gigabit Ethernet. Вона має на сьогодні найкраще співвідношення пропускної здатності, затримок та ціни.

Мережеві адаптери мають відповідати таким вимогам: підтримувати передачу даних за стандартом 1000Base-T; підтримувати функцію ввімкнення по мережі WOL; підтримання мережевого завантаження з використанням PXE; мають бути доступні драйвери WS2003;

підключення по PCI-E (рекомендація для зовнішніх адаптерів для досягнення оптимальної швидкодії).

Оскільки реалізації PXE відрізняються в залежності від виробника, тому мережеві адаптери мають проходити перевірку на сумісність з програмними компонентами системи. На сьогоднішній день пройшов випробування адаптер Intel 1000 GT.

Комутатори для побудови обчислювального кластера на основі обраної технології (Gigabit Ethernet) мають обиратися за такими критеріями:

- підтримувати передачу даних за стандартом 1000Base-T;
- кількість портів мають забезпечувати підключення всіх вузлів (включаючи головний вузол) та серверів додаткових служб, плюс 1 для додаткового обладнання, та плюс 10 % резервних;
- при плануванні розширення кластера слід мати окремий порт або модуль для підключення інших комутаторів. Рекомендується обирати комутатори з підтримкою порту розширення на швидкості 10 Гб/с;
- мати якомога кращі значення швидкодії переключення, пропускної здатності за обсягом даних та кількістю пакетів, що передаються за секунду, і характеристики за затримками.

Один із напрямків вдосконалення програмного комплексу UACluster полягає у зберіганні для кожного вуз- ла окремої копії клієнтського та кластерного MBR. У такому випадку, кожний з вузлів зможе мати своє власне розбиття жорсткого диска та використовувати різноманітні клієнтські ОС на різних вузлах для режиму 2.

При виконанні критичних розрахунків, для яких важливу роль відіграє проблема безпеки, доцільною є реалізація, в якій кластерна ОС із результатами всіх розрахунків, буде зберігатися на окремому сервері. В цьому випадку завантаження кластерної ОС необхідно виконувати по мережі.

У програмному пакеті UACluster передбачається, що керування вибором необхідного MBR відбувається за рахунок маніпуляцій опцією 067 boot filename DHCP-серверу. Тому для роботи UACluster на комп'ютерах, де використовується сторонній DHCP-сервер, необхідно передбачити можливість вибору необхідного MBR- запису на основі конфігураційного файла. Конфігураційний файл, як і копії MBR, можна зберігати на TFTP- сервері.

Технологія Fork-Join для програмування багатоядерних систем

Fork-Join - метод, застосовуваний у комунікаційних і комп'ютерних системах і служить для прогнозування продуктивності виконання великої кількості робочих задач.

Метод полягає в тому, що кожна задача розбивається на безліч синхронізованих задач, які обробляються паралельно на різних серверах.

Суть методу проста: велика задача розбивається на задачі поменше, ті, у свою чергу, на ще більш дрібні задачі, і так доти, поки це має сенс.

У самому кінці отримана тривіальна задача виконується послідовно. Даний етап називається Fork

Результат виконання послідовних задач об'єднується вгору по ланцюжку, поки не вийде рішення самої верхньої задачі.

Даний етап називається Join. Виконання всіх задач відбувається паралельно.

Вправи і завдання до теми №8

1. Є дві системи. Одна має швидкі процесори і повільні канали зв'язку, а інша – повільні процесори і швидкі канали зв'язку. Які переваги і недоліки кожної системи? На якій системі програми будуть мати кращу масштабованість?

2. Наведіть приклад реальної обчислювальної системи з розподіленою пам'яттю і комутаційною мережею, що має топологію двовимірного тора.

3. Допустимо, що перемножуються дві квадратні матриці. Які появляться особливості в організації обчислювальних процесів, якщо взяти матриці максимального розміру і старатися розв'язати задачу максимально швидко? Розгляньте варіанти обчислювальних процесів:

а) з спільною пам'яттю і універсальними процесорами;

б) з спільною пам'яттю і конвеєрними суматорами, перемножувачами і пристроями ділення;

в) з розподіленою пам'яттю і універсальними процесорами;

г) з розподіленою пам'яттю і конвеєрними суматорами, перемножувачами і пристроями ділення.