УДК 681.3

КОНЦЕПЦІЯ СТВОРЕННЯ ГНУЧКИХ ГОМОГЕННИХ АРХІТЕКТУР КЛАСТЕРНИХ СИСТЕМ

С.Д. Погорілий, Ю.В. Бойко, Д.Б. Грязнов, О.Д. Ломакін, В.А. Мар'яновський

Київський національний університет імені Тараса Шевченка,
01033, Київ, вул. Володимирська, 64.
Тел.:(044) 526 0522,
E-mail: sdp@rpd.univ.kiev.ua, lomakin@univ.kiev.ua, vitalik_m@univ.kiev.ua

Запропоновано концепцію побудови гомогенних кластерів. Створено програмний комплекс, який дозволяє здійснювати багатофункціональне застосування комп'ютерного ресурсу організації. Реалізовано метод модифікації MBR-запису та можливість призупинки задач у процесі кластерних розрахунків. Відповідно до запропонованої концепції наведено варіант побудови кластеру, впровадженого в комп'ютерному класі Майкрософт ІТ-академії інформаційно-обчислювального центру Київського національного університету імені Тараса Шевченка.

A concept of construction homogeneous clusters is offered. A software complex, which allows to effect multifunctional use of a computer asset of the organization, is created. A method of change of MBR-record and a possibility to suspend tasks during cluster calculations are realized. According to the introduced conception, a variation of cluster construction, implemented in a computer room in the Microsoft IT Academy in data-computing center of National Taras Shevchenko University of Kiev, is represented.

Вступ

Неухильне просування України до інформаційного суспільства призвело до формування у багатьох державних, комерційних, фінансово-економічних, наукових та інших організаціях і установах значного обчислювального ресурсу у вигляді персональних комп'ютерів. Обчислювальна потужність мікропроцесорів (далі процесорів) цих комп'ютерів практично зрівнялася із потужностями серверів, бо в більшості випадків вони будуються на однакових ядрах та мають однакові робочі частоти. Елементарний підрахунок показує, що в організації із 5-денним робочим тижнем та 8-годинним робочим днем комп'ютерний ресурс використовується менше ніж на 25 %. Під час простою такі комп'ютери можна об'єднати в єдину обчислювальну систему за рахунок наявності комунікаційного середовища між ними.

З іншого боку вищезгадані організації для розв'язання своїх задач (прогнозування, моделювання ризиків, розрахунки у галузі нанотехнологій тощо) дедалі все більше потребують суперкомп'ютерів і в першу чергу кластерних обчислень [1–3].

Для організації локальної мережі між обчислювальними вузлами кластера існує доступна та широко розповсюджена технологія Gigabit Ethernet. Вона повністю задовольняє вимогам до побудови мережі обміну за протоколом Message Passing Interface (MPI) як за пропускною здатністю, так і за параметрами затримок. Прогнозується, що в 2008 році набуде широкого розповсюдження стандарт 10 Гбіт Ethernet. При значеннях затримок 5–10 мкс він може успішно конкурувати зі спеціалізованою технологією Infiniband (високошвидкісна комутована послідовна шина, призначена для внутрішньо- та міжсистемних з'єднань [4]), затримки останніх реалізацій якої лежать у межах 1–5 мкс. Крім того, в 2010 році прогнозується впровадження технології 100 Гбіт Ethernet. На сьогоднішній день більше 40 % кластерів, що входять у Top500, використовують Ethernet [5].

При створенні кластера, в якому необхідно передбачити процес тимчасового призупинення (в робочий час організацій, коли комп'ютери використовуються для інших цілей), виникає проблема призупинення розрахунків, які виконуються на кластері. Відсутність можливості тимчасової зупинки розрахунків є недоліком більшості традиційних кластерних систем.

Для об'єднання персональних комп'ютерів в єдину обчислювальну систему та побудови обчислювального кластеру на основі персональних комп'ютерів з використанням програмного забезпечення Microsoft Windows Compute Cluster Server 2003 (WCCS) [6] в Київському національному університеті імені Тараса Шевченка був розроблений програмний комплекс UACluster.

Стратегічна перспектива проекту базується на широкому розповсюдженні технологій TCP/IP Offload Engine (TOE) та Remote Direct Memory Access (RDMA) (апаратна обробка мережевих протоколів, запис даних, отриманих мережевою картою, безпосередньо в пам'ять без участі процесора), що дозволить звільнити процесор від навантаження по передачі даних. Тенденції розвитку технологій оперативної пам'яті є такими, що її продуктивність підвищують шляхом розширення шини для передачі більшого обсягу даних при майже незмінних затримках. На фоні постійного зменшення затримок у мережі, можна говорити про можливість в майбутньому побудови кластерів з загальною пам'яттю (на основі технології OpenMP). Це приведе до усунення необхідності побудови окремих кластерних обчислювальних систем, достатньо буде об'єднати існуючі ресурси.

Принципи роботи кластеру

Введемо у розгляд сукупність працюючих в єдиному комунікаційному середовищі персональних комп'ютерів організації та визначимо два можливих варіанта їх роботи.

Режим 1. Функціонування у складі кластеру.

Режим 2. Функціонування локальних комп'ютерів, що забезпечують необхідні дії групи різних користувачів в складі мережі.

В неробочий для співробітників час (вночі, вихідні та святкові дні) комп'ютери організації можна об'єднати в кластер (рис. 1), тобто перевести їх у **режим 1** та використовувати для розрахунку різноманітних задач. Для побудови кластера запропоновано використання програмного забезпечення WCCS, що працює під керуванням операційної системи (ОС) Windows 2003 Server (x64) (WS2003).

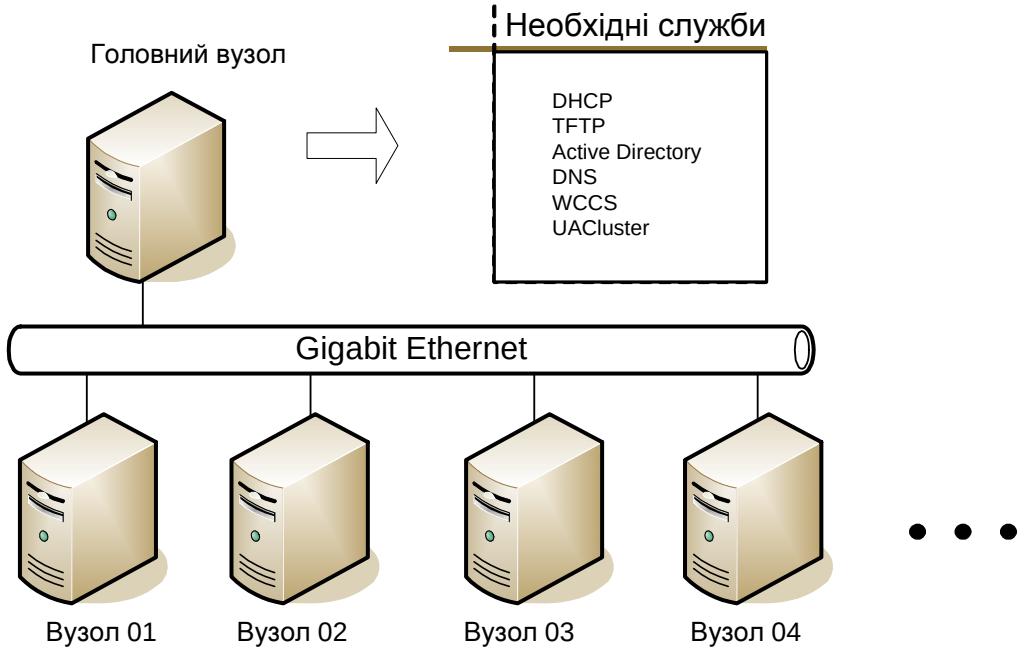

Рис. 1. Організаційна структура кластера

Тоді для функціонування кластеру на базі комп'ютерів організації необхідно забезпечити автоматичне перемикання між різними ОС. Для цього існують такі методи:

1. При інсталяції двох ОС сімейства Windows на персональному комп'ютері в файловій системі однієї з них зберігається файл boot.ini. Який містить інформацію про ОС, яка буде використана при завантаженні. Для зміни ОС необхідно відредагувати файл boot.ini та перезавантажити комп'ютер.

2. Другий метод можливий, коли як клієнтська ОС використовується Linux. У такому випадку, коли комп'ютер працює під керуванням WS2003, за допомогою сценаріїв дистанційно можна зробити активним розділ з Linux. Якщо на комп'ютері працює ОС Linux, також дистанційно за допомогою протоколу SSH можна виконати сценарії, які зроблять Windows-розділ активним, і при перезавантаженні вже буде використана ОС Windows. Даний метод описано в [7].

3. З використанням програмного забезпечення моніторингу за станом комп'ютерів у системі можна змінювати завантажувальну ОС шляхом модифікації запису Master Boot Record (MBR – це спеціальний запис на жорсткому диску, в якому зберігається інформація про розділи). В порівнянні з попередніми методами даний метод більш універсальний та гнучкий. Він може бути використаним для всіх ОС, які для завантаження використовують MBR. У роботі використано саме цей метод для автоматичного керування вибором однієї з ОС на жорсткому диску, яка завантажена.

Процедура інсталяції ОС для реалізації третього методу полягає в наступному.

Крок 1. Виконати інсталяцію клієнтської ОС з обов'язковим резервуванням вільного місця для кластерної ОС. Необхідно зберегти MBR в файл для даної конфігурації – це буде клієнтський MBR для **режиму 2** (рис. 2).

Крок 2. Розділ, в якому встановлено клієнтську ОС, необхідно зробити прихованим (hidden) та неактивним (inactive).

Крок 3. Створити розділ для кластерної ОС. Його треба зробити активним та інсталювати кластерну ОС. Після цього необхідно зняти копію MBR – це кластерний MBR для **режиму 1** (рис. 2).

Крок 4. Після закінчення інсталяції еталонного вузла необхідно зняти образ жорсткого диска цього вузла та розмножити його на всі інші комп'ютери в системі. Гомогенність вузлів є основною вимогою для реалізації методу. За рахунок тиражування образу жорсткого диска копія MBR-запису для завантаження клієнтської та кластерної ОС, яка знята з еталонного вузла, може бути в подальшому використана для всіх інших вузлів.

Рис. 2. Структура розділів жорсткого диска

В результаті виконання цих кроків вигляд MBR-запису жорсткого диска буде визначати, яка саме ОС буде завантажена на певному комп'ютері. Переваги цього методу полягають у тому, що керування процесом завантаження відбувається незалежно від ОС, яка використовується. Лишається тільки створити механізм автоматичної заміни MBR.

Створений програмний комплекс UACluster (деталі створення та роботи програмного комплексу описано на сайті проекту UACluster [8]) складається з двох програм: програма керування завантажувальними записами (ПКЗЗ), пункт керування вузлами (ПКВ).

Заміна MBR відбувається до етапу завантаження ОС. Для цього використовується ПКЗЗ, яка завантажується по мережі та працює в середовищі Preboot eXectuion Environment [9] (PXE – середовище для завантаження комп'ютера по мережі без використання жорсткого диску).

Для керування завантаженням з використанням середовища PXE з боку клієнта необхідна робота двох серверних компонент Dynamic Host Configuration Protocol (DHCP) та Trivial File Transfer Protocol **(**TFTP**)** серверу, на якому зберігається ПКЗЗ та дві копії MBR-розділу жорсткого диска одного з вузлів кластеру. Функції середовища PXE полягають у зверненні до DHCP, отримані параметри налаштування стеку протоколів TCP/IP та розташування завантажувального файлу (ПКЗЗ). Далі, PXE завантажує ПКЗЗ з TFTP-серверу та передає їй керування.

Інша проблема, яку необхідно вирішити для побудови запропонованої концепції створення кластера, полягає в наступному. Перед перемиканням кластера в **режимі 2**, тобто на клієнтську ОС, необхідно тимчасово призупинити всі розрахунки. Для цього, в найпростішому випадку, має вистачити переведення системи до сплячого стану (hibernate). При переведенні системи в такий стан, процеси та задачі, які виконувались, зберігаються на жорсткому диску, а система вимикається. Але на практиці виявилось, що такий механізм зупинки задач спрацьовує не завжди. Тільки після переведення всіх задач одночасно на всіх вузлах до стану паузи (suspend), а ОС до сплячого стану, можна виконати надійну тимчасову зупинку розрахунків. Після перезавантаження кластера (перемикання в **режим 1**) задачі необхідно відновити (resume). Для зупинки задач та відновлення їхньої роботи було використано функції psexec та pssuspend утиліти psTools, яка є частиною пакета Windows Sysinternals [10]. Для перевірки надійності процесу тимчасової зупинки задач був використаний пакет Intel MBI Benchmarks [11].

Процес завантаження кластера керується програмним комплексом UACluster, який встановлено на головному вузлі. В заздалегідь запланований час, коли комп'ютери в **режимі 2** вже не використовуються і вимкнені (кінець робочого дня), головний вузол (він весь час залишається ввімкненим) автоматично переводить всі вузли в **режим 1**. На кожному з вузлів цей процес складається з таких кроків (рис. 3).

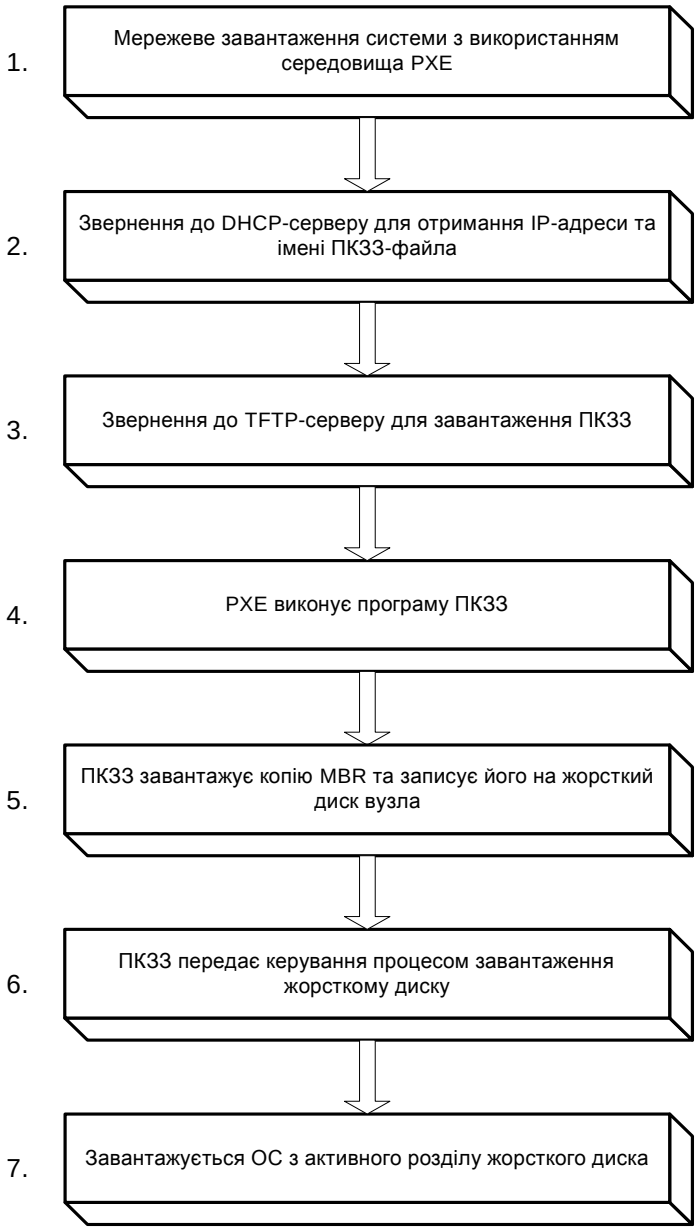

Рис. 3. Завантаження вузла кластера

Крок 1. ПКВ – спеціально створена програма мовою C#, у встановлений час вмикає по мережі вузол, використовуючи технологію Wake-On-Lan (WOL – технологія, яка дозволяє ввімкнути комп'ютер дистанційно, відправивши пакет, сформований спеціальним чином).

Крок 2. Використовуючи мережеве завантаження на стороні клієнта, PXE звертається до DHCP-серверу для отримання параметрів налаштування протоколу TCP/IP: IP-адресу, мережеву маску та IP-адресу шлюзу. Крім налаштувань мережевого інтерфейсу DHCP також повертає параметр boot filename (опція 067), в якому зазначено який саме файл необхідно використовувати для завантаження системи. Далі, PXE намагається прочитати з TFTP-серверу файл, ім'я якого було зазначено в полі boot filename (ПКЗЗ) (перші чотири етапи, рис. 3).

Крок 3. ПКЗЗ змінює MBR-запис жорсткого диска на кластерний (п'ятий етап, рис. 3), внаслідок чого розділ, на якому знаходиться кластерна ОС, стає активним (кластерний MBR).

Крок 4. ПКЗЗ передає керування завантаженням жорсткому диску (шостий етап, рис. 3).

Крок 5. На жорсткому диску завантажується кластерна ОС, вона знаходиться на активному розділі (сьомий етап, рис. 3).

Крок 6. ПКВ відновлює всі задачі, які не встигли виконатися минулого разу. Якщо, таких задач немає, WCCS розпочинає розрахунки нових.

Перед початком нового робочого дня, у встановлений час ПКВ зупиняє всі розрахунки, переводячи їх до стану suspend, а кластер до стану hibernate, та вимикає всі вузли, використовуючи дистанційні команди. Далі

при ввімкнені будь-якого з вузлів завантаження відбувається як і в **режимі 1** (кроки 2-4). З тією відмінністю, що на жорсткий диск записується клієнтська копія MBR (рис. 2). Розділ з клієнтською ОС буде в цьому випадку активним, внаслідок чого вона завантажується.

Метод із заміною MBR забезпечує додатковий рівень надійності, навіть за втратою зв'язку з головним вузлом зберігається можливість локального завантаження ОС.

При необхідності кластерна частина жорсткого диска під час роботи клієнтської ОС може виглядати як вільне місце на диску або специфічна файлова система, тобто вона прихована від користувача і не може бути ушкодженою без прав адміністратора.

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.

Для реалізації проекту було обрано комутатор 3COM SuperStack 3870 48-port.

Шляхи подальшого розвитку

Один із напрямків вдосконалення програмного комплексу UACluster полягає у зберіганні для кожного вузла 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.

Висновки

1. Запропоновано концепцію, відповідно до якої створено програмний комплекс UACluster, який дозволяє здійснювати багатофункціональне застосування комп'ютерного ресурсу організацій. У необхідний момент часу або за розкладом можна активізувати кластер чи локальну мережу. Створено можливість вибору необхідної для завантаження ОС, використовуючи середовище PXE та програму (ПКЗЗ) завантаження необхідного MBR-запису. Програмний комплекс UACluster дозволяє вмикати, вимикати, перезавантажувати **режими 1** та **2** з відновлення розпочатих розрахунків на кластері.
2. В результаті роботи реалізовано методи перезапису MBR та реалізовано можливість зупинки задач, які вже розраховуються на кластері для його тимчасового вимкнення.
3. Відповідно до запропонованої концепції наведено варіант побудови кластера, реалізованого в комп'ютерному класі Майкрософт IT-академії інформаційно-обчислювального центру Київського

національного університету імені Тараса Шевченка. Водночас на ньому виконуються різноманітні розрахунки. Аналогічні гомогенні кластери з гнучкою архітектурою нині впроваджуються в інших навчальних закладах України.

4. Запропонована концепція може бути розповсюдженою на створення гетерогенних кластерних архітектур за рахунок певного ускладнення програмного забезпечення моніторингу стану комп'ютерів вузлів кластеру.

1. Бойко Ю.В., Погорілий С.Д., Шкуліпа І.Ю. Дослідження паралельних схем алгоритму Прима. // Математичні машини і системи, 2007. – № 2. – С. 77 – 89.
2. Судаков О.О., Бойко Ю.В., Третяк О.В., Короткова Т.П. Оптимізація продуктивності обчислювального кластера на базі розподілених слабозв'язаних компонентів. // Математичні машини і системи, 2004. – № 4. – С. 57 – 65.
3. Судаков О.О., Бойко Ю.В. GRID-ресурси інформаційно-обчислювального центру Київського національного університету імені Тараса Шевченка. // Проблеми програмування. – 2006. – № 2/3. – С. 165 – 169.
4. http://ru.wikipedia.org/wiki/InfiniBand
5. http://www.top500.org/stats
6. http://www.microsoft.com/technet/ccs/overview.mspx
7. http://www.microsoft.com/downloads/thankyou.aspx?familyId=1457bc0a-eaff-4303-99ed-b199ab1c0857&displayLang=en
8. http://www.codeplex.com/UACluster
9. http://www.pxe.ca/
10. http://www.microsoft.com/technet/sysinternals/default.mspx
11. www.intel.com