



УКРАЇНА

(19) UA (11) 57663 (13) U
(51) МПК
G06F 15/16 (2011.01)МІНІСТЕРСТВО ОСВІТИ
І НАУКИ УКРАЇНИДЕРЖАВНИЙ ДЕПАРТАМЕНТ
ІНТЕЛЕКТУАЛЬНОЇ
ВЛАСНОСТІОПИС
ДО ПАТЕНТУ
НА КОРИСНУ МОДЕЛЬвидається під
відповідальність
власника
патенту

(54) МОДУЛЬ ВИСОКОЕФЕКТИВНОЇ БАГАТОПРОЦЕСОРНОЇ СИСТЕМИ ПІДВИЩЕНОЇ ГОТОВНОСТІ

1

2

(21) u201009341

(22) 26.07.2010

(24) 10.03.2011

(46) 10.03.2011, Бюл.№ 5, 2011 р.

(72) ІВАЩЕНКО ВАЛЕРІЙ ПЕТРОВИЧ, БАШКОВ
ЄВГЕН ОЛЕКСАНДРОВИЧ, ШВАЧИЧ ГЕННАДІЙ
ГРИГОРОВИЧ, ТКАЧ МАКСИМ ОЛЕКСАНДРОВИЧ
(73) НАЦІОНАЛЬНА МЕТАЛУРГІЙНА АКАДЕМІЯ
УКРАЇНИ, ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІ-
ЧНИЙ УНІВЕРСИТЕТ(57) Модуль високоефективної багато процесорної системи підвищеної готовності, який містить материнські плати, процесори, оперативну пам'ять, мережеві плати Fast Ethernet та Gigabit Ethernet, мережеві комутатори, призначений для побудови багато процесорних систем, який **відрізняється** тим, що має окрему реконфігуровану мережу й додаткові мережеві інтерфейси з підтримкою режиму channel bonding, які забезпечують високу швидкість обміну даних при зниженні завантаження каналів між кластерами; керовані комутатори, що дозволяють реалізовувати процедуру реконфігурації мережі для заявленого класу задач, підвищити пропускну здатність мережі й високошвидкісний доступ до пам'яті вузлів кластера; проміжні буфери пам'яті комутаторів, які дають змогу "розвантажити" центральний процесор (CPU) в моменти передачі та прийому пакетів між вузлами кластера, при цьому резервування ключових

компонентів і значне зменшення числа критичних вузлів системи підвищило надійність функціонування багато процесорної системи, до того ж інтегрований мережевий інтерфейс майстер-вузла з'єднується за допомогою входу/виходу з портом 00 першого керованого комутатора, інтегровані мережеві інтерфейси першого, другого, третього, четвертого та п'ятого slave-вузлів з'єднуються за допомогою входу/виходу відповідно з портами 01, 02, 03, 04, 05 першого керованого комутатора; крім того, перший режим конфігурації другої мережі реалізовано таким чином, що перший, другий, третій, четвертий та п'ятий slave-вузли з'єднуються за допомогою входу/виходу першого зовнішнього двоспрямованого інтерфейсу відповідно з портами 01, 02, 03, 04, 05 другого керованого комутатора, а також за допомогою другого зовнішнього входу/виходу двоспрямованого інтерфейсу відповідно з портами 01, 02, 03, 04, 05 третього керованого комутатора; крім того, другий режим конфігурації цієї ж мережі реалізовано таким чином, що перший, другий, третій, четвертий та п'ятий slave-вузли з'єднуються за допомогою входу/виходу першого зовнішнього двоспрямованого інтерфейсу відповідно з портами 01, 03, 05, 07, 09 другого керованого комутатора, а також за допомогою другого зовнішнього входу/виходу двоспрямованого інтерфейсу відповідно з портами 04, 06, 08, 10, 02 другого керованого комутатора.

Пропонована корисна модель відноситься до сфери обчислювальної техніки, зокрема, вона стосується архітектури високопродуктивних багато процесорних систем обробки даних і може використовуватися при розв'язуванні задач математичної фізики, екології, при розробці нових технологічних процесів, а також у моделюванні складних технічних систем. Корисна модель може застосовуватись у діяльності науково-дослідних центрів, вузів, конструкторських бюро, а також підприємств, що займаються комп'ютерним моделюванням.

1. У практиці відомий модуль обробки даних для багато процесорної системи, яка містить 5 процесорних модулів, системну шину і 5 зовнішніх

пристроїв. До складу кожного процесорного модуля входять: процесор, локальна пам'ять, комунікаційна пам'ять, регістр команд, регістр даних, регістр адреси і блок керування (Баканов В.М. Персональный вычислительный кластер как недостающее звено в технологии проведения сложных технологических расчетов / В.М. Баканов // Метизы. - 2006. - 2 (12). - С. 33-36.).

Серед недоліків такої багато процесорної системи можна назвати такі:

1. Невисокий ступінь надійності функціонування системи в цілому.

Причина недоліку полягає у застосуванні окремо взятої ПЕОМ для керування кластерною

(13) U

(11) 57663

(19) UA

системою в цілому, а також потреба в налаштуванні й використанні для неї спеціалізованого програмного забезпечення. Таким чином, при виході з ладу керуючої ПЕОМ або збою в роботі програмного забезпечення в цілому перестас функціонувати вся кластерна система.

2. Складність конфігурації і переналаштування обчислювальних вузлів кластера.

Причиною такого недоліку є та обставина, що для розв'язування кожної конкретної задачі необхідно перелаштовувати програмне забезпечення на кожному з обчислювальних вузлів кластерної системи.

3. Функціональна незавершеність кластерної системи.

Недолік пояснюється тим, що, по-перше, кластер не може функціонувати без окремо взятого і спеціально налаштованого комп'ютера, по-друге, окремі вузли кластера (наприклад, маршрутизатор-комунікатор) взагалі винесені за межі кластерної системи.

4. Збільшення енергоспоживання кластерної системи.

Цей недолік є наслідком застосування комплектуючих з малим співвідношенням F_{MFLOPS} / W_{watt} (тобто відношенням продуктивності MFLOPS на одиницю потужності ват), а також результатом присутності індивідуального блока живлення у складі кожного обчислювального вузла. Таким чином, кожен з цих блоків при запуску короткочасно споживає значну потужність (наявність стрибка у напрузі), а також, крім активного навантаження, володіє ємнісно-індуктивним, що спричиняє підвищене енергоспоживання.

2. Існує багато обчислювальних систем з розподіленою пам'яттю, що містять процесори, об'єднані деяким комутаційним середовищем. Серед таких можна назвати Intel Paragon, IBM SP1, Parsytec, Blackford MultiCore та ін. Відмінності між цими системами залежить від типу процесорів та особливостей організації комутаційного середовища. Як аналог можна назвати кластер Blackford MultiCore (Специфікація кластера Blackford MultiCore / Інститут динаміки систем и теории управления СО РАН, г. Иркутск, http://www.mvs.icc.ru/cluster_info.html). Він складається з 20 обчислювальних вузлів. У свою чергу кожен з цих вузлів складається з двох чотириядерних процесорів 2.33 ГГц Intel Xeon 5345 EM64T (Clovertown) та обладнаний кеш-пам'яттю другого рівня 8 Мб, Front Side Bus (FSB) частотою 1333 МГц, Fully-Buffered DIMM, його оперативна пам'ять становить 8x1024 Мб, а також обладнаний двома жорсткими дисками SATA 160 Гб і CD-ROM drive. Керуючий вузол має у своєму складі два чотириядерних процесори 2.33 ГГц Intel Xeon 5345 EM64T (Clovertown), обладнаних кеш-пам'яттю другого рівня обсягом 8 Мб, Front Side Bus (FSB) частотою 1333 МГц, Fully-Buffered DIMM, оперативна пам'ять якого становить 8x1024 Мб, має також Intel SAS RAID-контролер, три жорсткі диски SAS-D 73 Гб, DVD/CD-ROM drive.

Систему змонтовано в 19-дюймову стійку серверної шафи APC AR3100 NetShelter SX 42U, для обчислювальних вузлів використовують корпуси

Intel Server Chassis SR1500 1U, а для керуючого вузла – Intel Server Chassis SR2500 2U. Мережева інфраструктура забезпечена двома комутаторами Gigabit Ethernet AT-GS900/24-XX 1U. Для безперервної роботи застосовується також система відведення тепла APC ACF400 Rack Air Removal Unit SX у вигляді каналів APC ACF127, а моніторинг зовнішнього середовища відбувається за допомогою приладу APC NBWL032 NetBotz 320 Wall з набором датчиків і камер. Модулі розподіленого живлення APC AP7852 і два джерела безперебійного живлення APC Smart SURT8000RMXLI 8 кВА і SURT10000RMXLI 10 кВА забезпечують якісне і безперебійне живлення системи.

Проте належить відзначити такі недоліки багатопроцесорної системи:

1. Низька реальна продуктивність розв'язування сильнозв'язаних.

Цей недолік пояснюється тим, що пікова продуктивність обчислювального вузла дорівнює 2 37.28 ГФл/с, а комунікаційне середовище для всіх вузлів кластерної системи розраховане на використання однієї гігабітної мережі. Таким чином, розв'язуючи задачі з інтенсивним граничним обміном інформації, спостерігаємо перевантаження мережевих ресурсів кластера, тому процесори велику частину часу будуть вимушені простоювати і витрачати додаткові ресурси на процедуру синхронізації, а ця обставина, у свою чергу, і призводить до зниження реальної продуктивності системи в цілому.

2. Висока вартість системи.

Недолік зумовлюється застосуванням спеціалізованих серверних процесорів, серверних корпусів формату 1U/2U, спеціалізованої системи відведення тепла, пристрою для моніторингу зовнішнього середовища, модуля розподілу живлення, дорогих джерел безперебійного живлення (APC Smart 8 кВА, 10 кВА).

3. Високе енергоспоживання і висока вартість експлуатації системи.

Причиною цього недоліку можна назвати необхідність у високому енергоспоживанні для інфраструктури кластерної системи в цілому (8 кВА, 10 кВА), що збільшує витрати на утримання кластера. У разі його модернізації необхідно ставити додаткову серверну шафу, додаткові джерела безперебійного живлення, системи відведення тепла, пристрій моніторингу зовнішнього середовища, модуль розподілу живлення. Отже, виникає необхідність забезпечувати різко зростаючу потужність енергоживлення. Щоб створити умови для надійної роботи кластера, потрібно формувати резерв необхідних комплектуючих кластера, тобто мати запасні блоки живлення, комутатор, процесори, а це, у свою чергу, тягне за собою збільшення вартості експлуатації кластерної системи.

4. Складність експлуатації кластера.

Причини цього недоліку можна пояснити двома обставинами. По-перше, виникає необхідність утримувати штат сертифікованих фахівців для налаштування, експлуатації та ремонту кластерної системи. По-друге, операційна система встановлюється на кожному із обчислювальних вузлів, тому при виникненні збоїв або необхідності вне-

сення змін у системне або програмне забезпечення, потрібно переналаштовувати кожен вузол окремо. Зазначені обставини призводять до збільшення часу простою системи, а також потребують висококваліфікованого обслуговування. Крім того, виникає необхідність в організації спеціального робочого місця (терміналу).

Найбільш близьким до пропонованої корисної моделі за призначенням і архітектурою є УНІФІКОВАНИЙ БАЗОВИЙ МОДУЛЬ БАГАТОПРОЦЕСОРНОЇ СИСТЕМИ З ПРОГРАМОВАНОЮ АРХІТЕКТУРОЮ (патент заявка: 2004136937/09, 16.12.2004, Науково-дослідний інститут багатопроцесорних обчислювальних систем Таганрозького державного радіотехнічного університету (НДІ МВС ТРТУ), <http://fpga.parallel.ru/family.html>), який має у своєму складі групу макропроцесорів, що виконують великі математичні операції; групу мультиконтролерів розподіленої пам'яті, матричний комутатор, який забезпечує прямі просторові з'єднання між усіма компонентами системи. Кожен базовий модуль складається з 16 макропроцесорів, що містять 64 елементарних процесорів, 32 каналів, розрахованих на 16 контролерів розподіленої пам'яті, він споживає 30Вт на тактовій частоті 50МГц із спільною продуктивністю $25 \cdot 10^9$ оп/с (25 Гфл). На основі цього модуля створено кілька типорозмірів обчислювачів з продуктивністю до 100 Гфл, різного призначення, зокрема:

- персональну робочу станцію з продуктивністю $2,5 \cdot 10^{10}$ оп/с;
- модульно-нарощувану МВС ПА «Рись» у складі чотирьох базових модулів з продуктивністю 10^{11} оп/с;
- модульно-нарощувану МВС ПА «Скиф-Т» у складі 8 базових модулів з продуктивністю $2 \cdot 10^{11}$ оп/с;
- модульно-нарощувану МВС ПА «Медведь» у складі чотирьох базових модулів з продуктивністю 200 Гфл.

Програмне забезпечення модуля включає систему складального програмування, компілятор мови програмування високого рівня з неявним описом паралелізму, асемблер, а також пакет прикладних програм. Система складального макропрограмування передбачає такі елементи:

- мова програмування високого рівня з неявним описом паралелізму;
- інтегроване середовище розробника паралельних програм мовою програмування з неявним описом паралелізму;
- база даних (бібліотека) компонентів паралельних програм;
- засоби опису програмних рішень у графічній формі;
- графічний редактор, що дозволяє проектувати як структурно реалізовані макрооперації, так і великі фрагменти завдання - кадри.

Перелічимо недоліки описаного пристрою:

1. Обмежене і спеціально орієнтоване коло розв'язуваних задач, умовна висока ефективність системи.

Цей недолік пояснюється тим, що для кожної конкретної задачі необхідно мати відповідну бібліотеку компонентів паралельних програм. Лише за

таких обставин буде забезпечено високу продуктивність і ефективність обчислювальної системи. У той же час, за відсутності необхідних програмних компонентів бібліотека повинна модифікуватися з урахуванням нових вимог, які до неї пред'являються.

2. Великий обсяг підготовчих робіт до експлуатації системи, що зменшує загальну тривалість виконання завдань і реальну ефективність системи в цілому.

Причина такого недоліку полягає в тому, що висока продуктивність й ефективність системи залежить від наявності бібліотеки компонентів паралельних програм. У той самий час для програмування й налагодження необхідного програмного забезпечення необхідно мати спеціально орієнтовану ПЕОМ. Для створення нового ПЗ і його налагодження потрібно набагато більше часу, ніж на саму експлуатацію модуля багатопроцесорної системи.

3. Відсутність універсальності й здатності програмного забезпечення до перенесення.

Недолік зумовлюється тим, що програмне забезпечення, створене для даної обчислювальної системи, не можна використовувати на інших аналогічних системах, оскільки вони мають різну архітектуру.

4. Використання спеціально орієнтованої елементної бази.

Причина цього недоліку викликана тим, що для функціонування, експлуатації, а в перспективі, модифікації такої обчислювальної системи необхідно використовувати спеціальну елементну базу ПЛІС, що не дозволяє виконувати операції вдосконалення й оновлення системи в умовах швидкого технічного прогресу сучасних інформаційних технологій.

5. Складність експлуатації та супроводження системи.

Причина такого недоліку пояснюється необхідністю утримання штату висококваліфікованих фахівців з великим досвідом роботи у сфері експлуатації ПЛІС-технологій. Крім того, для роботи з такою системою потрібний досвід роботи не лише в межах названих технологій, але й знання спеціалізованих мов асемблера ARGOS, мови програмування високого рівня COLAMO, інтегрованого середовища розробника паралельних програм (трансляторів, бази даних (бібліотеки) паралельних процедур, графічного редактора синтезу компонентів паралельних програм та ін.).

Завдання, на вирішення якого спрямована заявлена корисна модель, полягає у створенні модуля багатопроцесорної обчислювальної системи, реальна ефективність і продуктивність якого була б піковою для заявленого класу задач. До того ж дана система повинна мати підвищену надійність і високу енергоефективність.

Модуль містить один майстер-вузол (PM000) і п'ять обчислювальних slave-вузлів (PN001, PN002, PN003, PN004, PN005), три керовані комутатори (SW1, SW2, SW3), проміжні буфери пам'яті комутаторів, реконфігуровану мережу для обміну даних між обчислювальними вузлами, віртуальні локальні мережі (FS123, VS23, VS012, VS022, VS032,

VS042, VS052, VS013, VS023, VS033, VS043, VS053), механізм резервування ключових компонентів, а також передбачає мережене завантаження вузлів. У майстер-вузлі та slave-вузлах застосовуються одні й ті самі комплектуючі (материнські плати, процесори, інтегровані мережеві плати Fast Ethernet, зовнішні мережеві плати Gigabit Ethernet). Зокрема, майстер-вузли обладнані додатково жорсткими дисками (HDD), CD/DVD, дисковими (FDD).

Технічний результат, що досягається при запровадженні винаходу, полягає в тому, що обмін даними між обчислювальними вузлами винесено в окрему мережу, модель OSI, яка працює на каналному (другому) рівні з використанням механізмів channel bonding і VLAN, що збільшило швидкість обміну даних і знизило завантаження каналу, який з'єднує вузли кластера.

Уведення додаткових керованих комутаторів, які працюють паралельно, дозволило через термінал або WES-інтерфейс змінювати конфігурацію мережі, підвищувати її пропускну здатність, що забезпечило високошвидкісний доступ до пам'яті вузлів кластера й обмін даними між цими вузлами за допомогою комутаційних мереж. Реалізація реконфігурованої мережі дозволяє підвищити ефективність кластерної системи, адаптуючи структуру її мережі для вирішення кожного конкретного типу завдань.

Використання проміжного буфера пам'яті дозволяє «розвантажити» центральний процесор CPU в моменти передачі та прийому пакетів між вузлами кластера, що зумовлює підвищення продуктивності обчислювальної системи в цілому.

Застосовування енергоефективних материнських плат і процесорів, а також мережевого завантаження дозволило завдяки режиму Power on After Power Fail / Former-Sts забезпечити одночасний запуск групи обчислювальних вузлів модуля багатопроекторної обчислювальної системи від одного блока живлення замість кількох (N). Це забезпечує значне підвищення енергоефективності системи і зниження виділення тепла на одиницю її продуктивності, що дозволяє відмовитися від вбудованої системи кондиціонування.

Мережеве завантаження модуля багатопроекторної системи і введення механізму резервування ключових компонентів цього модуля, а також істотне зменшення кількості компонентів системи дозволяє підвищити надійність функціонування вузла багатопроекторної системи.

Для досягнення описаного технічного результату пристрій обладнали додатковими комутаторами, що працюють паралельно з використанням механізмів channel bonding і VLAN, а замість однієї комутаційної мережі, де відбувався процес завантаження, діагностики, керування й обміну даними, введено другу, куди винесено процес обміну даними між вузлами кластера.

З метою реалізації режиму реконфігурування між портами 11, 12, 13, 14, 15 комутатора SW1, портами 00 комутаторів SW2 і SW3 створюється віртуальна локальна мережа (VLAN) VS123, яка регулює трафік у межах окремої VLAN. При цьому обчислювальний вузол PM000 з'єднується за до-

помогою входу/виходу зовнішнього двоспрямованого інтерфейсу (PM000.i2) з портом 15 (VS123.SW1.15) керованого комутатора SW1, двоспрямованого інтерфейсу входу/виходу порту 11 (VS123.SW1.11) керованого комутатора SW1 з 00 портом (VS123.SW2.00) керованого комутатора SW2, двоспрямованого інтерфейсу входу/виходу порту 12 (VS123.SW1.12) керованого комутатора SW1 з 00 портом (VS123.SW3.00) керованого комутатора SW3.

Конфігурація мережі має певні особливості. Зокрема комутатор SW1 утворює мережу керування, завантаження й діагностики кластера. При цьому інтегрований мережевий інтерфейс PM000.i00 вузла PM000, що має функцію мережевого завантаження, з'єднаний за допомогою входу/виходу з портом 00 керованого комутатора SW1, інтерфейс PN001.i01 вузла PN001 з'єднаний за допомогою входу/виходу з портом 01 комутатора SW1, інтерфейс PN002.i01 вузла PN002 так само з'єднаний з портом 02 комутатора SW1, інтерфейс PN003.i01 вузла PN003 - з портом 03 комутатора SW1, інтерфейс PN004.i01 вузла PN004 - з портом 04 комутатора SW1, інтерфейс PN005.i01 вузла PN005 - з портом 05 комутатора SW1, інтерфейс PN011.i01 вузла PN011 - з портом 06 комутатора SW1, інтерфейс PN012.i01 вузла PN012 - з портом 07 комутатора SW1, інтерфейс PN013.i01 вузла PN0X3 - з портом 08 комутатора SW1, інтерфейс PN014.i01 вузла PN014 - з портом 09 комутатора SW1, інтерфейс PN015.i01 вузла PN015 - з портом 10 комутатора SW1.

Щоб налагодити режим агрегації інтерфейсів, застосовують керовані комутатори (SW2 і SW3), які працюють паралельно, підтримуючи функцію агрегації інтерфейсів (channel bonding). Агрегація каналів дозволяє збільшити продуктивність мережі в m разів (m - кількість комутаторів у відповідній локальній мережі), сформувати топологію мережі у вигляді зірки (кожен вузол може передавати/приймати дані будь-якому іншому вузлу в мережі обміну даними).

Вирішення широкого кола завдань за допомогою комутованої обчислювальної мережі відбувається на основі використання двох режимів. Перший режим мережі моделює топологію типу зірки, другий - кільця.

Перший режим. Спочатку формується «розподілена VLAN» VS23 мережа на комутаторах SW2, SW3. При цьому обчислювальний вузол PN001 з'єднується за допомогою входу/виходу зовнішнього двоспрямованого інтерфейсу (PN001.i2) з портом 01 (VS23.SW2.01) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN001.i3) з портом 01 (VS23.SW3.01) керованого комутатора SW3. Обчислювальний вузол PN002 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN002.i2) з портом 02 (VS23.SW2.02) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN002.i3) з портом 02 (VS23.SW3.02) керованого комутатора SW3. Обчислювальний вузол PN003 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього

інтерфейсу (PN003.i2) з портом 03 (VS23.SW2.03) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN003.i3) з портом 03 (VS23.SW3.02) керованого комутатора SW3. Обчислювальний вузол PN004 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN004.i2) з портом 04 (VS23.SW2.04) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN004.i3) з портом 04 (VS23.SW3.04) керованого комутатора SW3. Обчислювальний вузол PN005 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN005.i2) з портом 05 (VS23.SW2.05) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN005.i3) з портом 05 (VS23.SW3.05) керованого комутатора SW3. Обчислювальний вузол PN011 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN011.i2) з портом 06 (VS23.SW2.06) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN011.i3) з портом 06 (VS23.SW3.06) керованого комутатора SW3. Обчислювальний вузол PN012 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN012.i2) з портом 07 (VS23.SW2.07) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN012.i3) з портом 07 (VS23.SW3.07) керованого комутатора SW3. Обчислювальний вузол PN013 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN013.i2) з портом 08 (VS23.SW2.08) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN013.i3) з портом 08 (VS23.SW3.08) керованого комутатора SW3. Обчислювальний вузол PN014 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN014.i2) з портом 09 (VS23.SW2.09) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN014.i3) з портом 09 (VS23.SW3.09) керованого комутатора SW3. Обчислювальний вузол PN015 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN015.i2) з портом 10 (VS23.SW2.10) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN015.i3) з портом 10 (VS23.SW3.10) керованого комутатора SW3.

Другий режим. Частина прикладних завдань передбачає, що граничний обмін даними відбувається між сусідніми обчислювальними вузлами. У такому разі зв'язок між slave-вузлами організовується за топологією кільця. При цьому обчислювальний вузол PN001 з'єднується за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN001.i2) з портом 01 (VS012.SW2.01) керованого комутатора SW2, а також за допомогою входу/виходу двоспрямованого зовнішнього інтерфейсу (PN001.i3) з портом 04 (VS022.SW2.04).

[illegible]

Для уникнення взаємного впливу при граничному обміні (передачі/прийомі даних) між обчислювальними вузлами створюються віртуальні локальні мережі (VLAN), що регулюють трафік в межах окремої мережі (VLAN) комутаторів SW2 і SW3. У кожному комутаторі створюються по 5 віртуальних мереж: VS012 між портами 01 і 02 комутатора SW2, VS022 між портами 03 і 04 комутатора SW2, VS032 між портами 05 і 06 комутатора SW2, VS042 між портами 07 і 08 комутатора SW2, VS052 між портами 09 і 10 комутатора SW2, VS013 між портами 01 і 02 комутатора SW3, VS023 між портами 03 і 04 комутатора SW3, VS033 між портами 05 і 06 комутатора SW3, VS043 між портами 07 і 08 комутатора SW3, VS053 між портами 09 і 10 комутатора SW3.

Режим роботи системи реалізується через використання проміжного буфера пам'яті для зберігання відповідних пакетів. Проміжний буфер пам'яті комутаторів SW1 – SW3 позбавляє від необхідності процедури синхронізації даних при мережевому обміні, коли здійснюється процес відправлення та прийому пакетів під час вирішення сильно зв'язаних завдань, при цьому виникає можливість зменшити завантаження CPU, що підвищує ефективність і продуктивність кластерної системи в цілому.

Реалізована концепція побудови комутаційної мережі модуля багатопроцесорної системи не передбачає застосування маршрутизації/комутації пакетів на мережевому рівні. Запропонована система організації мережі буде доцільною в тих випадках, коли передача інформації здійснюється за допомогою пакетів, що не мають здатності до маршрутизації (наприклад, NETBIOS або протоколи спеціального призначення), але при цьому треба забезпечити високий ступінь захищеності мережі та ізолюваності інформаційних потоків користувача. Разом з тим, таблиця комутації містить записи лише про ті віртуальні канали, що проходять через даний комутатор, а не про всі наявні в мережі вузли (або підмережі, якщо застосовується ієрархічний спосіб адресації).

Режим конфігурування й налаштування програмного забезпечення обчислювальних вузлів спрощується за рахунок мережевого завантаження. При цьому в обчислювальних вузлах відсутні мережеві диски, а завантаження, їх налаштування, діагностика і керування відбувається через мережу комутатора SW1. Такий підхід дозволяє гнучко переналаштовувати конфігурацію ПЗ, оновлювати й адаптувати її під конкретне завдання.

Мережеве завантаження модуля багатопроцесорної системи, резервування ключових компонентів модуля, а також істотне зменшення кількості компонентів системи дає можливість підвищити надійність функціонування вузла.

У винаході також реалізовано режим групової процедури запуску й завантаження обчислювальних вузлів модуля. Енергопостачання обчислювальних вузлів системи здійснюється від одного джерела безперебійного живлення. Стабільне енергоживлення йде через тумблер на блок живлення головного модуля ATX1 і блок живлення обчислювальних вузлів системи ATX2.

Для забезпечення високої надійності системи електроживлення кластера здійснюється через безперебійний блок живлення (UPS), від нього через мережеві інтерфейси (розгалужувачі) струм надходить у блоки живлення головного модуля і резервних модулів. Таким чином, у кожному модулі багатопроцесорної обчислювальної системи присутні два однотипні блоки живлення (ATX): перший живить майстер-вузол (вінчестер, CD/DVD і т. д.), а другий через інтерфейс живлення – п'ять обчислювальних вузлів (материнські плати, пам'ять, процесор). Унаслідок того, що всі обчислювальні вузли кластерної системи однотипні, використання окремого обчислювального вузла як модуля розширення дозволяє підвищити надійність кластерної системи за рахунок процедури резервування. При цьому для кожного обчислювального вузла на панель діагностики і керування П00, П01 і т. д., виводиться інформація про подачу живлення, звернення до жорсткого диска, режим запуску материнської плати, стан кнопок запуску і перезавантаження відповідно.

Причинно-наслідковий зв'язок між сукупністю істотних ознак винаходу і технічним результатом, який досягається, полягає в тому, що введення підмереж завантаження системи, діагностики й обміну даних дозволило розвантажити мережі обчислювальної системи, підвищити її доступність і продуктивність.

Запровадження керованих комутаторів дозволяє підвищити швидкодію обчислень під час вирішення сильнозв'язаних завдань за рахунок режиму використання мереж (VLAN), який дозволяє розподілити частини програм на сегменти, не пов'язані між собою.

Реалізація проміжного буфера пам'яті позбавляє від необхідності виконувати процедуру синхронізації даних при мережевому обміні, коли реалізується процес відправлення та прийому пакетів для вирішення сильнозв'язаних завдань, при цьому виникає можливість зменшити завантаження пристрою CPU, а це підвищує ефективність і продуктивність кластерної системи в цілому. Такий підхід можна пояснити тим, що при передачі пакетів без буферизації передавальний пристрій робить запит приймальному про готовність прийняти пакет (кадр) і чекає підтвердження, потім передає пакет і знову чекає підтвердження про його прийняття. Якщо передача даних передбачає буферизацію, то пакет передавачем передається в буфер і він там зберігається до готовності приймача його прийняти. Крім того, реалізація механізмів віртуальних каналів (virtual circuit та virtual channel) створює в мережі стійкі шляхи прямування трафіка через мережу з комутацією пакетів. Ці механізми враховують існування в мережі потоків даних. Мережі лише забезпечують можливість передачі трафіка вздовж віртуальних каналів, а які саме потоки передаватимуться по цих каналах, визначається самими кінцевими вузлами.

Мережеве завантаження кластерної системи дозволило відмовитися від застосування локальних дисків HDD з метою забезпечення гнучкості й переконфігурації кластера під конкретне завдання,

тим самим забезпечується режим адаптації обчислювальної системи до необхідного типу завдань.

Суть корисної моделі та принцип роботи модуля високоефективної багатопроцесорної системи підвищеної готовності пояснюється рисунками, де зображено:

Фіг.1 – блок-схема будови модуля багатопроцесорної системи підвищеної готовності;

Фіг.2 – схема сполучення інтерфейсів для двох модуля багатопроцесорної системи підвищеної готовності;

Фіг.3 – структура у мережі модуля багатопроцесорної системи підвищеної готовності для реалізації граничного обміну;

Фіг.4 – залежність часу рахунку однієї ітерації від числа вузлів модуля багатопроцесорної системи підвищеної готовності;

Фіг.5 – залежність часу граничного обміну інформації від числа вузлів модуля багатопроцесорної системи підвищеної готовності;

Фіг.6 – залежність часу рахунку однієї ітерації від числа вузлів модуля багатопроцесорної системи підвищеної готовності з урахуванням часу граничного обміну інформацією.

Особливість блок-схеми модуля (Фіг.1) полягає в тому, що всі обчислювальні вузли модуля високоефективної багатопроцесорної системи підвищеної готовності містять процесор (1) C7 CPU приєднаний шиною FSB (Front Side Bus 533/400 МГц) до південного моста CN700 (2) з інтегрованим відеоконтролером VIA UniChrome Pro та відео виходами SVGA (3), TV (4) і інтерфейсом AGP 8X (5), а південний міст підключено до локальної пам'яті (6) стандартів DDR2 533/400 або DDR 400/333/266. Південний і північний мости з'єднані за модульною архітектурою платформ VIA V-MAP (7) (Modular Architecture Platform). Для з'єднання північного моста на чипсеті VT8237A (8) і південного на чипсеті V-MAP передбачено використання шини Ultra V-Link, що працює зі швидкістю 533 МБ/с. До чипсета підключено контролер VIA DriveStation (9), який підтримує інтерфейси SATA, PATA і режим RAID, а також шина PCI Bus з двома рознімними з'єднаннями PCI (10,11), в яких встановлено мережеві інтерфейси з підтримкою режимів channel bonding і Gigabit Ethernet (12, 13). До моста VIA VTS237A підключено інтегрований аудіоконтролер (14) VIA Vinyl™ HD Audio, контролер клавіатури і маніпулятора миші PS/2 (15), а також вісім високошвидкісних портів стандарту USB 2.0 (16), і контролер VT1211 (17), що являє собою повнофункціональний Super I/O-чип з контролером дисководу гнучких дисків, інтерфейсом паралельного порту IEEE-1284, двома послідовними портами 16C550-UART, контролером VFIR (швидкісний інфрачервоний порт), ігровим портом з підтримкою 2-х джойстиків, MIDI-інтерфейсом та інтерфейсом AM FLASH-ROM BIOS (18), інтегрованим мережевим інтерфейсом з підтримкою режимів мережевого завантаження та Fast Ethernet (19).

Для розв'язування деякого класу прикладних задач виникає необхідність розширення обчислювальних потужностей. Закладений принцип модульності дозволяє збільшувати продуктивність обчислювальної системи за рахунок додавання

нових модулів. На Фіг.2. подано схему сполучення інтерфейсів для двох модулів багатопроцесорних систем. На цій схемі зображено головний модуль, що містить один майстер-вузол (PM000) і п'ять обчислювальних вузлів (PN001, PN002, PN003, PN004, PN005), а також модуль як вузол розширення (майстер-вузол PM001 та обчислювальні вузли PN011, PN012, PN013, PN014, PN015). При цьому комутатор SW1 утворює мережу керування, завантаження і діагностики розширеного кластера, всі інтегровані інтерфейси майстер-вузла і slave-вузлів з'єднуються з входами/виходами цього комутатора.

Структура мережі модуля багатопроцесорної системи для реалізації граничного обміну інформацією подається на Фіг.3. Верхня частина схеми моделює топологію типу зірки, нижня - кільця.

Вибраний режим енергоживлення кластерної системи з функцією Power on After Power Fail / Former-Sts забезпечив одночасний запуск групи обчислювальних вузлів модуля багатопроцесорної обчислювальної системи від одного блока живлення замість кількох (N). Реалізований підхід зменшує стрибки напруги при вмиканні блоків живлення, збільшує надійність системи, реалізує режим їхнього оптимального завантаження, дозволяє зменшити споживану електроенергію обчислювальної системи. Описаний інтерфейс запуску обчислювальних вузлів модуля багатопроцесорної системи дозволив спростити структуру цієї операції запуску, а також конфігурації системи в цілому, зрештою істотно знизити вартість обчислювальної системи, використовуючи один UPS на весь модуль та один блок живлення замість кількох серверних спеціалізованих.

Також цей режим енергоживлення системи дозволив збільшити її надійність за рахунок зменшення кількості найбільш критичних компонентів. Тим самим зменшилася вартість системи і витрати на її експлуатацію. Завдяки запровадженню такого режиму енергоспоживання з'явилася можливість відмовитися від спеціалізованих інтегрованих систем кондиціонування, що теж знизило вартість системи в цілому. У той самий час, застосування однотипних компонентів системи, режиму резервування дало змогу підвищити надійність функціонування системи за рахунок оперативної заміни її компонента, що вийшов з ладу, резервним. Такий підхід застосовується і для проведення регламентних та профілактичних робіт без зупинки вузла в цілому, що дозволяє підвищити надійність системи.

Розрахунок ефективності заявленої обчислювальної системи ілюструється поданими нижче аналітичними співвідношеннями.

Стосовно даного класу задач усі обчислення виконуються на різницевій сітці. При цьому для аналізу ефективності багатопроцесорної системи найважливішим параметром буде тривалість однієї ітерації. Тоді в умовах застосування багатопроцесорної системи повний час однієї ітерації визначатиметься на підставі такого співвідношення:

$$T = T_c + T_{ex}, \quad (1)$$

де T_c - час рахунку однієї ітерації відносно об'єкта обчислень, с;

T_{ex} - час граничного обміну між вузлами кластера, с.

При цьому відзначимо, що коли час рахунку ітерації залежить лише від потужності процесора, то час граничного обміну залежить від розміру різницевої сітки, кількості вузлів кластерної системи і пропускної здатності обчислювальної мережі.

Безпосередньо T_{ex} можна визначити на підставі такого співвідношення:

$$T_{ex} = \frac{E}{V}. \quad (2)$$

де E - обсяг даних області граничного обміну, Гбіт;

V - пропускна здатність мережі кластера, Гбіт/с.

В умовах, коли обсяг області обчислень великий, можна вивести формулу для обчислення обсягу даних граничного обміну (для того класу завдань, які вирішуються за допомогою пропонованого кластера). Ця формула матиме такий вигляд:

$$E = 2 \cdot (N-1) \cdot \sqrt{R}, \quad (3)$$

де N - кількість вузлів кластера;

R - обсяг оперативної пам'яті вузла кластера, доступної для завдання, Гбіт.

Зауважимо, що коли $N=1$, то значення величини E перетворюватиметься на нуль, що є повністю очевидним.

За таких обставин можна оцінити час однієї ітерації, який включатиме, власне, час рахунку однієї ітерації при використанні N вузлів кластерної системи і час граничного обміну залежно від кількості вузлів кластера N , тобто

$$T = \frac{T_c}{N} + \frac{2 \cdot (N-1) \cdot \sqrt{R}}{V}. \quad (4)$$

Аналіз співвідношення (4) показує, що при розподілі області обчислень між вузлами кластера навантаження для кожного його леза буде меншим. Реалізуючи режим агрегації каналів можна збільшити швидкість обміну даних у мережі в m разів, при цьому m - кількість каналів, які працюють паралельно, тоді продуктивність мережі буде визначатися співвідношенням:

$$V = m \cdot V_n,$$

де V_n - продуктивність інтерфейсу, Гбіт/с.

Унаслідок того, що вузли багатопроцесорної системи працюють паралельно, то й загальний час ітерації стає меншим. У той самий час, із збільшенням кількості числа вузлів кластера також збільшується обсяг граничних даних і, відповідно, час обміну даними між вузлами.

Згідно з розглянутими співвідношеннями були проведені обчислювальні експерименти із застосуванням багатопроцесорної кластерної системи, блок-схему якої зображено на Фіг.1. Для вихідних даних в експериментах було прийнято такі характеристики класу завдань і самої кластерної системи:

$$T_c = 100 \text{ с}, V = 1 \text{ Гбіт/с}, R = 8 \text{ Гбіт}, m = 2.$$

Результати моделювання показано у вигляді графічних залежностей на Фіг.4-6.

Фіг.4 ілюструє таку ситуацію, коли час рахунку однієї ітерації при збільшенні числа вузлів багатопроцесорної системи зменшується за гіперболіч-

ною залежністю. Разом з цим Фіг.5 показує, що час граничного обміну при збільшенні числа вузлів багатопроцесорної системи збільшується за лінійним законом. Загальну картину зміни часу рахунку однієї ітерації в багатопроцесорній системі ілюструє графічна залежність, яка зображена на Фіг.6.

Аналіз графічної залежності, показаної на Фіг.6 свідчить, що час розрахунків на першому етапі зменшується при збільшенні кількості вузлів кластера. Такий результат можна було б прогнозувати. Проте зменшення такого часу відбувається до певної межі. Якщо кількість вузлів перевищує чотири або п'ять, то час розрахунків починає зростати. Відбувається це на фоні збільшення обсягу даних, які пересилаються між вузлами (Фіг.5). Таким чином, можна відзначити, що при постійному розмірі сітки, в умовах даного завдання збільшувати розмір кластера понад чотири або п'ять лез не має сенсу. За таких обставин якраз і досягається максимальна ефективність кластерної системи.

Особливості функціонування модуля багатопроцесорної обчислювальної системи підвищеної готовності полягають у тому, що після подачі енергоживлення на блок ATX1 і надходження зовнішнього сигналу PUSK з панелі P00 починається запуск та ініціалізація роботи майстер-вузла модуля багатопроцесорної системи. Безпосередньо завантаження ОС може здійснюватися або з жорсткого диска, або з CD/DVD-пристрою. Після завантаження операційної системи запускається конфігураційний скрипт, який налаштовує роботу DHCP-сервера, крім того, тут визначається кількість обчислювальних вузлів системи, у разі потреби налаштовується доступ в середовище Інтернет або зовнішню мережу, при цьому зазначаються основні налаштування й параметри. Оскільки активізація режиму Power on After Power Fail / Former-Sts і подача живлення на відповідний блок (ATX2) запускає всі обчислювальні slave-вузли, то й завантаження операційної системи обчислювальних вузлів відбувається групами, без перевантаження першої мережі. Після завантаження всіх обчислювальних вузлів кластера завершується робота скрипту і кластер готовий до виконання паралельних обчислень.

Майстер-вузол (PM00) через комутатор SW1 забезпечує спрямування потоку даних керування, діагностики та прийому/передачі умов вирішуваних завдань до slave-вузлів. Slave-вузли відповідно до алгоритму вирішуваного завдання і виконуваних процесів реалізують режим необхідних обчислень. Обмін даними між обчислювальними вузлами винесено в окрему мережу, організовану за допомогою керованих комутаторів SW2 – SW3. Для досягнення максимальної ефективності кластерної системи здійснюється процес реконфігурації структури другої мережі залежно від класу вирішуваних завдань. Прийом/передача даних у slave-вузлах відбувається за допомогою проміжних буферів пам'яті. Результати обчислень через першу комунікаційну мережу, за допомогою керованого комутатора SW1, передаються майстер-вузлу (PM00), де здійснюється необхідний режим видачі й обробки даних.

Блоки заявленого пристрою можуть бути реалізовані за допомогою засобів обчислювальної техніки масового виробництва. І тут набула подальшого розвитку технологія FAWN (скорочення від Fast Array of Wimpy Nodes - швидкий масив слабких вузлів). Подібні рішення дозволяють конструювати заявлену багатопроцесорну систему в умовах вузлів, наукових організацій, дослідницьких центрів.

Уведення в пристрій окремої обчислювальної мережі для обміну даних, додаткових керованих комутаторів, що працюють паралельно, проміжного буфера пам'яті комутаторів, режиму мережевого завантаження процесорів, механізму резервування ключових компонентів модуля дозволило:

по-перше, підвищити швидкість обчислень під час вирішення сильно зв'язаних завдань, забезпечити високошвидкісний доступ до пам'яті вузлів кластера й обмін даними між ними, знизити завантаження каналу, що проходить між вузлами кластера, за рахунок формування окремої обчислювальної мережі й реалізації механізмів channel bonding і VLAN;

по-друге, за рахунок проміжного буфера пам'яті «розвантажити» центральний процесор CPU в моменти передачі та прийому пакетів між вузлами кластера, що зумовлює підвищення продуктивності обчислювальної системи в цілому;

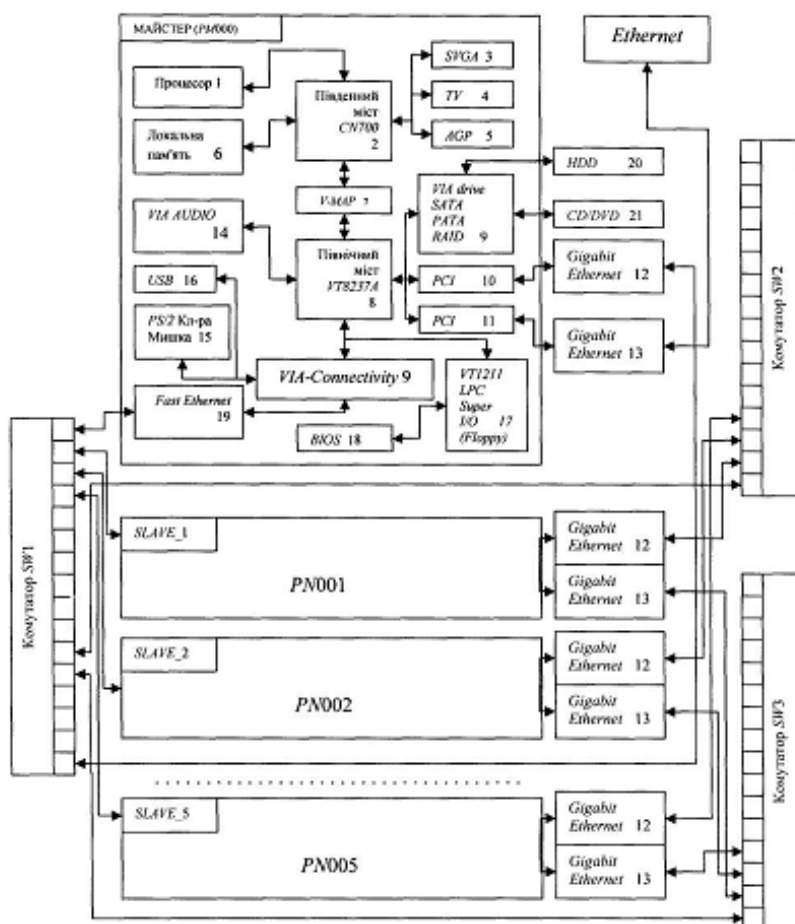
по-третє, підвищити ефективність кластерної системи, адаптуючи структуру її мережі до вирішення кожного конкретного типу завдань;

по-четверте, за рахунок модульного принципу побудови спростити проектування, нарощування або заміну кластерних вузлів, що вийшли з ладу, а також роботу й експлуатацію системи в цілому;

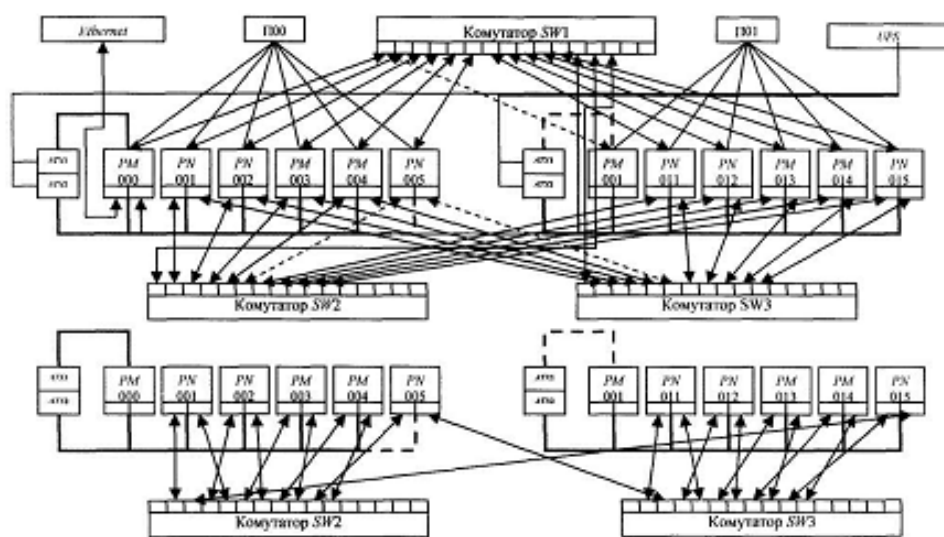
по-п'яте, завдяки мережевому завантаженню й режиму Power on After Power Fail / Former-Sts забезпечити одночасний запуск групами обчислювальних вузлів модуля багатопроцесорної обчислювальної системи підвищеної готовності від одного блока живлення замість кількох, що призводить до значного підвищення енергоефективності системи і зменшення її вартості;

по-шосте, за рахунок мережевого завантаження модуля багатопроцесорної системи підвищеної готовності, а також введення механізму резервування ключових компонентів модуля істотно зменшити кількість її компонентів і тим самим значно підвищити надійність функціонування модуля багатопроцесорної системи підвищеної готовності;

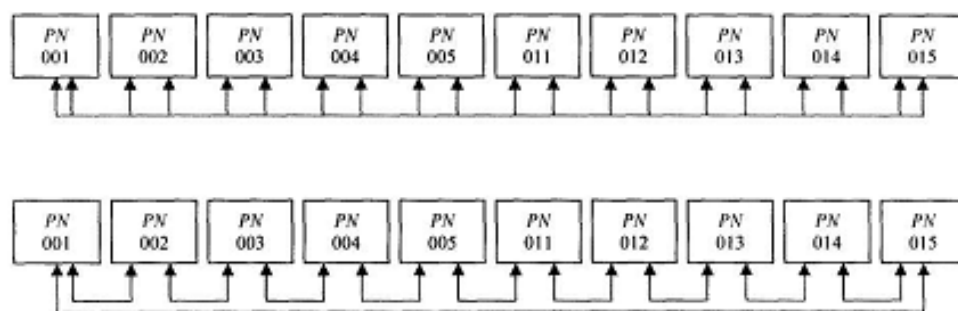
по-сьоме, забезпечити здатність до перенесення програмного забезпечення на інші кластери системи з метою виконання подібних розрахунків.



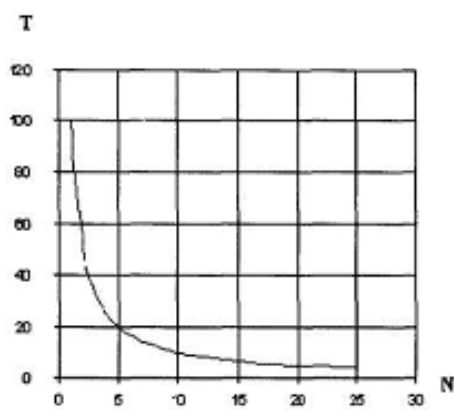
Фиг. 1



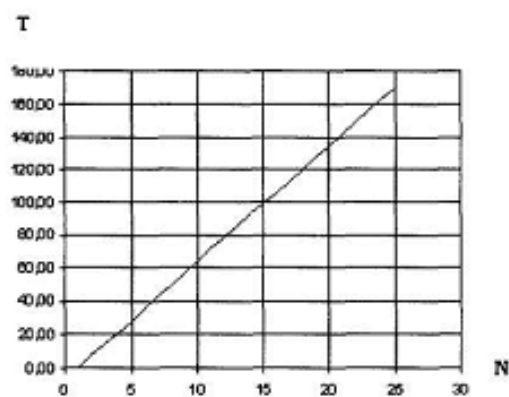
Фиг. 2



Фиг. 3



Фиг. 4



Фиг. 5

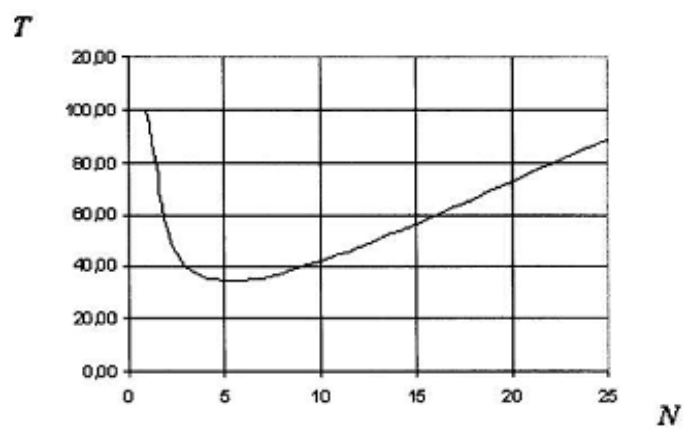


Fig. 6