



ДЕРЖАВНА СЛУЖБА
ІНТЕЛЕКТУАЛЬНОЇ
ВЛАСНОСТІ
УКРАЇНИ

УКРАЇНА

(19) UA

(11) 105438

(13) U

(51) МПК

G06F 15/16 (2006.01)

(12) ОПИС ДО ПАТЕНТУ НА КОРИСНУ МОДЕЛЬ

(21) Номер заявки: **u 2015 07019**
(22) Дата подання заявки: **14.07.2015**
(24) Дата, з якої є чинними права на корисну модель: **25.03.2016**
(46) Публікація відомостей про видачу патенту: **25.03.2016, Бюл.№ 6**

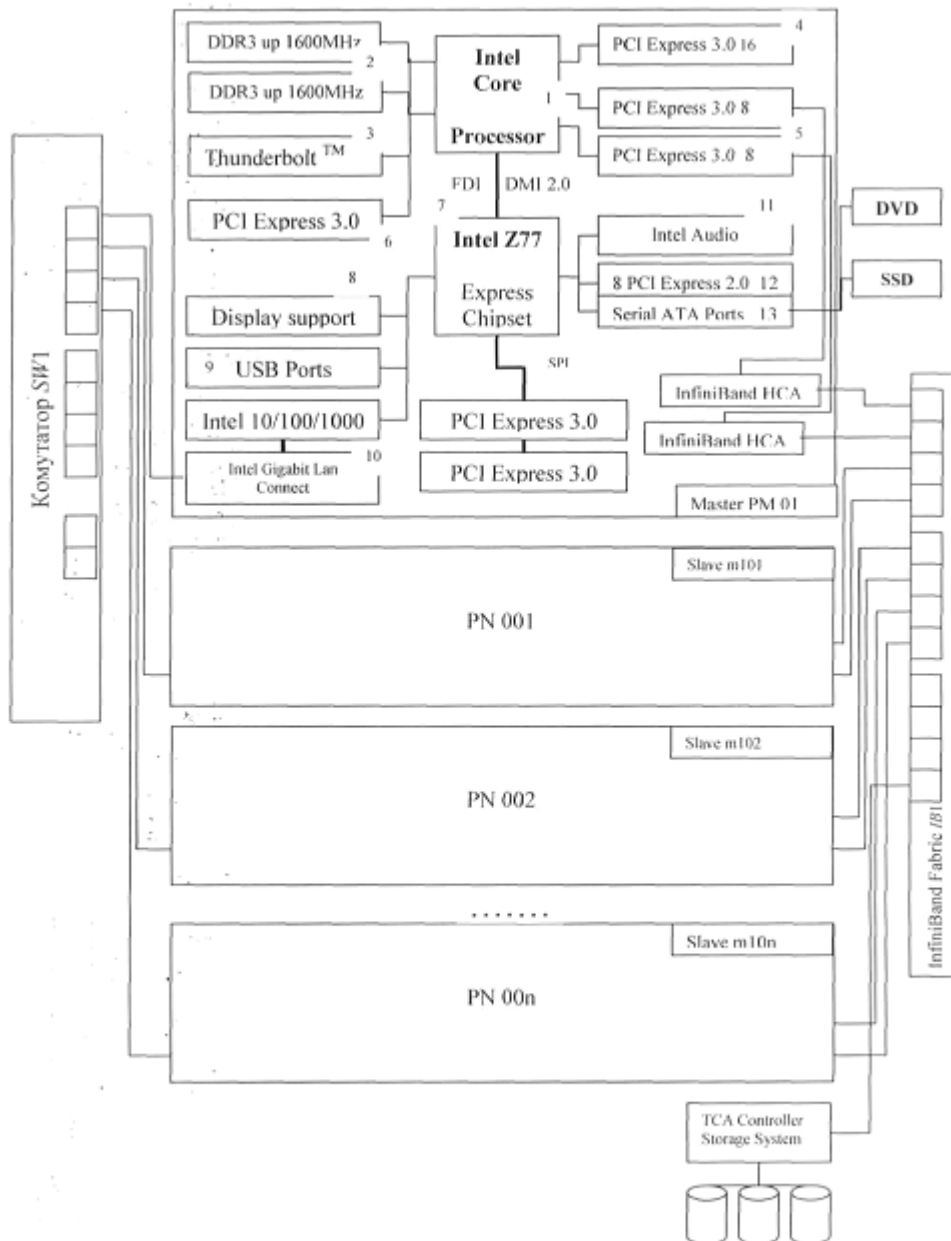
(72) Винахідник(и):
**Іващенко Валерій Петрович (UA),
Башков Євген Олександрович (UA),
Швачич Геннадій Григорович (UA),
Ткач Максим Олександрович (UA),
Щербина Павло Олександрович (UA),
Волнянський Владислав Васильович (UA)**
(73) Власник(и):
**НАЦІОНАЛЬНА МЕТАЛУРГІЙНА
АКАДЕМІЯ УКРАЇНИ,
пр. Гагаріна, 4, м. Дніпропетровськ, 49600 (UA),
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ
ТЕХНІЧНИЙ УНІВЕРСИТЕТ,
пл. Шибанкова, 2, м. Красноармійськ,
Донецька обл., 85300 (UA)**

(54) МОДУЛЬ ВИСОКОЕФЕКТИВНОЇ БАГАТОПРОЦЕСОРНОЇ СИСТЕМИ З РОЗШИРЮВАНОЮ ОБЛАСТЮ ОБЧИСЛЕНЬ

(57) Реферат:

Модуль вискоефективної багатопроцесорної системи з розширюваною областю обчислень містить майстер-вузол і обчислювальні slave-вузли, призначений для побудови багатопроцесорних систем. Модуль містить два керовані комутатори SW1 (GigabitEthernet) та JB1(InfiniBand), систему локального збереження результатів та проміжних обчислень, механізм резервування ключових компонентів, а також передбачає мережеве завантаження вузлів по мережі GI через комутатор SW1. Кожен вузол модуля має персональний блок живлення стандарту ATX, slave-вузли додатково обладнані двома двопортовими зовнішніми адаптерами HCA InfiniBand. Між портами 0, 1, 2, 3, 4, 5, 6, 7, 8 комутатора SW1, створюється віртуальна локальна мережа (VLAN). Майстер-вузол модуля в мережі діагностики, завантаження та управління з'єднується за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу мережевої плати Gigabit Ethernet з портом 0 керованого комутатора SW1. В мережі обміну даних майстер-вузол з'єднується зі slave-вузлами двома двопортовими мережевими адаптерами InfiniBand з портами керованого комутатора IB1 обчислювальні вузли в мережі діагностики. Завантаження та управління з'єднується за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу з портами керованого комутатора SW1. В мережі обміну даних з'єднується за допомогою двох двопортових мережесхем адаптерів InfiniBand з портами керованого комутатора IB1, систему локального збереження результатів та проміжних обчислень під'єднано до порту керованого комутатора IB1, інтерфейс налаштування керованого комутатора IB1 з'єднується з портом керованого комутатора SW1.

UA 105438 U



Фиг. 1

Корисна модель належить до сфери обчислювальної техніки, зокрема вона стосується архітектури високопродуктивних багатопроцесорних 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.

I. 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 (Спецификация кластера Blackford MultiCore /Институт динамики 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 8×1024 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 8×1024 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 розраховане на використання одногігітної 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 внесення 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).

II. Відомий аналогом є модуль обробки даних 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 30 Вт на тактовій частоті 50 МГц 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 Argus, мови програмування високого рівня COLAMO,

інтегрованого середовища розробника паралельних програм (трансляторів, бази даних (бібліотеки) паралельних процедур, графічного редактора синтезу компонентів паралельних програм та ін.).

III. Найближчим аналогом до корисної моделі є модуль високоефективної багатопроцесорної системи підвищеної готовності (Пат. 61944 Україна, МПК C21D1/26, G06F15/16 (2011.01). Модуль високоефективної багатопроцесорної системи підвищеної готовності /Іващенко В.П., Башков Є.О., Швачич Г.Г., Ткач М.О.; патентовласники Національна металургійна академія України, Донецький національний технічний університет. - № u201009341; заявл. 26.07.2010; опубл. 10.03.2011, Бюл. № 5), який містить один майстер-вузол (MNode001) і п'ять обчислювальних slave-вузлів (NNode001, NNode002, NNode003, NNode004, NNode005), три керовані комутатори (SW1, SW2, SW3), проміжні буфери пам'яті комутаторів, реконфігуровану мережу для обміну даними між обчислювальними вузлами, віртуальні локальні мережі, механізм резервування ключових компонентів, а також передбачає мережеве завантаження вузлів. Комутована мережа багатопроцесорної обчислювальної системи працює в двох режимах. Перший моделює топологію типу зірки, другий - кільця. Така кластерна система побудована на базі лезових технологій. Із цих причин вона являє собою щільно упакований модуль з процесорами лезового типу, встановленими в стійці. Усередині стійки містяться вузли, апаратура для ефективного з'єднання компонентів, апаратура керування внутрішньою мережею системи і т. д. Кожне лезо кластера працює під керуванням своєї копії стандартної операційної системи. Склад і потужність вузлів можуть бути різними в рамках одного модуля, проте в даному випадку було розглянуто однорідний модуль. Взаємодія між вузлами кластерної системи встановлюється за допомогою інтерфейсу програмування, тобто спеціалізованих бібліотек функцій.

Серед недоліків найближчого аналога можна назвати такі:

1. Відсутня можливість використання такої системи для розв'язування задач з розширюваною областю обчислень.

Недолік зумовлюється тим, що комунікаційне середовище для всіх вузлів кластерної системи розраховане на використання одногігітної мережі. При розв'язуванні задач з розширюваною областю обчислень, буде спостерігатися перевантаження мережевих ресурсів системи, тому процесори будуть вимушені простоювати і система буде працювати тільки на організацію обміну даними між її вузлами. Тим самим непродуктивно використовуються обчислювальні вузли кластера, а час розв'язування задач значно збільшується і визначатиметься спроможністю комунікаційної мережі передавати дані обчислень.

2. Невисока реальна продуктивність системи для сильнозв'язаних задач.

Причина недоліку полягає у використанні одногігітної мережі, особливість якої полягає в тому, що при розв'язуванні задач з інтенсивним граничним обміном інформації, спостерігається перевантаження мережевих ресурсів кластера, тому процесори велику частину часу будуть вимушені простоювати і витратити додаткові ресурси на процедуру синхронізації, а ця обставина, у свою чергу, і призводить до зниження реальної продуктивності системи в цілому.

3. Обмежене та спеціально орієнтоване коло задач, які розв'язуються за допомогою такої системи.

Цей недолік пояснюється тим, що вирішення задач за допомогою комутованої обчислювальної мережі системи відбувається тільки на основі використання двох режимів. Перший режим мережі моделює топологію типу зірки, другий - кільця. Такі режими роботи орієнтовані для реалізації граничного обміну даних в залежності від того обмеженого класу задач, який розв'язується за допомогою запропонованого кластера.

4. Обмежена розширюваність багатопроцесорної системи. Причина цього недоліку обумовлена використанням одногігітної мережі, тому при розширенні кластерної системи кількість її лез буде обмежена через перевантаження мережевих ресурсів.

5. Висока латентність обчислень для сильнозв'язаних задач. Недолік зумовлюється використанням в мережі обміну даних одногігітної технології, яка має високу латентність при розв'язуванні сильнозв'язаних задач (до 80 мкс), а це означає, що при передачі коротких пакетів даних основний, час буде витрачатися на ініціалізацію та синхронізацію повідомлень, тому процесори велику частину часу будуть вимушені простоювати.

В основу корисної моделі поставлена задача створити модуль багатопроцесорної обчислювальної системи, реальна ефективність і продуктивність якого була б піковою при розв'язуванні сильнозв'язаних задач та задач розширюваною областю обчислень. До того ж дана система повинна мати підвищену надійність і високу енергоефективність. Блоки заявленого пристрою повинні комплектуватися за допомогою засобів обчислювальної техніки масового виробництва.

Поставлена задача вирішується тим, що модуль містить майстер-вузол і обчислювальні slave-вузли, згідно з корисною моделлю, два керовані комутатори (SW1, IB1), систему локального збереження результатів та проміжних обчислень, механізм резервування ключових компонентів, а також передбачає мережеве завантаження вузлів по мережі GI через комутатор SW1, при цьому між портами 0, 1, 2, 3, 4, 5, 6, 7, 8 комутатора SW1, створюється віртуальна локальна мережа (VLAN), кожен вузол модуля має персональний блок живлення стандарту ATX, slave-вузли додатково обладнані двома двопортовими зовнішніми адаптерами HCA InfiniBand, майстер-вузол модуля в мережі діагностики, завантаження та управління з'єднується за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу Мережевої плати Gigabit Ethernet з портом 0 керованого комутатора SW1, в мережі обміну даних майстер-вузол з'єднується зі slave-вузлами двома двопортовими мережевими адаптерами InfiniBand з портами керованого комутатора IB1, обчислювальні вузли в мережі діагностики, завантаження та управління з'єднуються за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу з портами керованого комутатора SW1, в мережі обміну даних з'єднуються за допомогою двох двопортових мережеских адаптерів InfiniBand з портами керованого комутатора IB1; систему локального збереження результатів та проміжних обчислень під'єднано до порту керованого комутатора IB1, інтерфейс налаштування керованого комутатора IB1 з'єднується з портом керованого комутатора SW1. Зокрема, майстер-вузол додатково обладнано накопичувачами жорстких дисків (SSD) та DVD. Комутована мережа багатопроцесорної обчислювальної системи працює в чотирьох режимах: зірка, кільце, лінійка, решітка. Такі режими роботи були орієнтовані на реалізацію граничного обміну даними, що відображають особливості задач, які розв'язуються за допомогою пропонованої багатопроцесорної системи.

Технічний результат, полягає в тому, що обмін даними між обчислювальними вузлами винесено в окрему мережу з використання технології InfiniBand, що збільшило швидкість обміну даних і суттєво знизило латентність (завантаження каналів) мережі, яка з'єднує вузли кластера.

Застосування комутованого середовища в мережі обміну даних зі сполуками "точка-точка", на відміну від ранніх варіантів мереж, які використовували шинне з'єднання дозволило суттєво збільшити швидкість передачі даних між вузлами багатопроцесорної системи та зменшити латентність у середовищі передачі пакетів даних. Це пояснюється тим, що всі передачі починаються та закінчуються на HCA-адаптері каналу (host channel adapter).

Уведення режиму QDR (Quad Data Rate) у двопортових мережеских плат MCX353A-FCBT (максимальна швидкість передачі даних режимі 2*56 Гбіт/с, відповідає режиму FDR (Fourteen Data Rate)) дозволило узгодити обчислювальні можливості процесора та мережі передачі даних по інтерфейсу PCI Express (2*32=64 Gb/s).

Використання принципу RDMA (Remote Direct Memory Access – віддалений прямий доступ до пам'яті) дозволяє передавати дані без додаткової буферизації й не вимагає активної роботи ОС, а також бібліотек або додатків. Це дозволило суттєво зменшити навантаження на процесори системи під час передачі даних, тим самим звільняються обчислювальні ресурси процесорів і зменшується латентність у середовищі передачі даних.

Застосування двопортових HCA-адаптерів за рахунок режиму 4x агрегації каналів мереженого інтерфейсу дозволило змінювати конфігурацію мережі, підвищивши її пропускну здатність. Реалізація реконфігурованої мережі Дозволяє підвищити ефективність кластерної системи, адаптуючи структуру її мережі для вирішення кожного конкретного типу завдань.

Як обчислювальні платформи було вибрано процесори Intel Core I7 та материнські плати на базі чипсетів серії Intel Z77, яка містить два вільних слоти PCI Express 3.0 8 bit, це дозволило суттєво збільшити швидкість обчислень та передачі даних (максимальна швидкість передачі даних в дуплексному режимі відповідає 64 Гбіт/с).

Мережеве завантаження системи і введення механізму резервування ключових компонентів, а також використання блоків живлення для кожного леза багатопроцесорної системи дозволяє підвищити надійність функціонування модуля системи.

Для досягнення описаного технічного результату пристрій обладнано комутатором з підтримкою технології Infiniband та двопортовими мережевими HCA-адаптерами 4x з роз'ємними з'єднаннями стандарту QSFP+ (Quad Small Form-factor Pluggable), які оснащені пасивними мідними кабелями. Крім того, модуль використовує механізми віддаленого прямого доступу до пам'яті RDMA, технологію channel bonding 4x, віртуальні обчислювальні мережі VLAN, систему Локального збереження результатів та проміжних обчислень TCA Controller Storage System.

Кожен вузол модуля має персональний блок живлення стандарту ATX, а також передбачає мережеве завантаження вузлів по мережі GI через комутатор SW1, slave-вузли додатково обладнані двома двопортовими зовнішніми адаптерами HCA InfiniBand, при цьому між портами

0, 1, 2, 3, 4, 5, 6, 7, 8 комутатора SW1 створюється віртуальна локальна мережа (VLAN); майстер-вузол модуля в мережі діагностики, завантаження та управління з'єднується за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу мережевої плати Gigabit Ethernet з портом 0 керованого комутатора SW1, в мережі обміну даних майстер-вузол з'єднується зі slave-вузлами двома двопортовими мережевими адаптерами InfiniBand з портами керованого комутатора /51; обчислювальні вузли в мережі діагностики, завантаження та управління з'єднуються за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу з портами керованого комутатора SW1, в мережі обміну даних з'єднуються за допомогою двох двопортових мережеских адаптерів InfiniBand з портами керованого комутатора IB1, систему локального збереження результатів та проміжних обчислень під'єднано до порту керованого комутатора IB1, інтерфейс налаштування керованого комутатора IB1 з'єднується з портом керованого комутатора SW1.

Вирішення широкого кола задач за допомогою комутованої обчислювальної мережі відбувається на основі використання чотирьох режимів. Перший режим мережі моделює топологію типу зірки, другий - кільце, третій - лінійки, четвертий - решітки.

Перший режим. Основна особливість такої топології полягає в тому, що всі процесори системи мають зв'язок з управляючим процесором. У такому разі зв'язок між обчислювальними slave-вузлами організовується за топологією типу зірка.

Спочатку між портами керованого комутатора IB1 формуються чотири "розподілені VLAN" мережі: VS11 між портами 01, 05, 09, 13, KS12 між портами 02, 06, 10, 14, VS13 між портами 03, 07, 11, 15, VS14 між портами 04, 08, 12, 16.

Налаштування комутатора IB1 та його конфігурування виконується майстер-вузлом за допомогою двох портів стандарту Gigabit Ethernet (IB1 Gl.i1 - керування, /IB1 Gl.i2 - масштабування). Систему локального збереження результатів та проміжних обчислень TCA Controller Storage System під'єднано до 16 порту керованого комутатора IB1.

Другий режим. Частина прикладних задач передбачає, що граничний обмін даними відбувається між сусідніми обчислювальними вузлами. У такому разі зв'язок між обчислювальними slave-вузлами організовується за топологією типу кільце.

Спочатку між портами керованого комутатора IB1 формуються вісім "розподілених VLAN" мереж: KS01a між портами 03 і 05, KS01b між портами 04 і 06, мережі VS12a між портами 07 і 09, VS12b між портами 08 і 10, мережі VS23a між портами 11 і 13, VS23b між портами 12 і 14, мережі KS03a між портами 15 і 01, VS30b між портами 16 і 02.

Налаштування комутатора IB1 та його конфігурування виконується майстер-вузлом за допомогою двох портів стандарту Gigabit Ethernet (IB1 Gl.i1 - керування, /IB1 Gl.i2 - масштабування). Систему локального збереження результатів та проміжних обчислень TCA Controller Storage System під'єднано до 16 порту керованого комутатора IB1.

Третій 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 типу лінійки, це частинний приклад топології типу кільце.

Спочатку між портами керованого комутатора IB1 формуються сім "розподілених VLAN" мереж: VS01a між портами 03 і 05, VS01b між портами 04 і 06, VS12a між портами 07 і 09, VS12b між портами 08 і 10, VS23a між портами 11 і 13, VS23 між портами 12 і 14, мережі VS03aб між портами 01, 02, 15, 16 і 17.

Налаштування комутатора IB1 та його конфігурування виконується майстер-вузлом за допомогою двох портів стандарту Gigabit Ethernet (IB1 Gl.i1 - керування, IB1 Gl.i2 - масштабування). Систему локального збереження результатів та проміжних обчислень TCA Controller Storage System під'єднано до 16 порту керованого комутатора IB1.

Четвертий режим. Частина прикладних завдань передбачає, що граничний обмін даними відбувається між сусідніми обчислювальними вузлами за топологією типу замкнута сітка.

Спочатку між портами керованого комутатора IB1 формуються шість "розподілених VLAN" мереж: VS011-013 між портами 05 і 07, VS012-024 між портами 06 і 12, мережі VS021-023 між портами 09 і 11, VS022-034 між портами 10 і 15, мережі VS031-033 між портами 13 і 15, VS032-014 між портами 14 і 08.

Налаштування комутатора IB1 та його конфігурування виконується майстер-вузлом за допомогою двох портів стандарту Gigabit Ethernet (IB1 Gl.i1 - керування, IB1 Gl.i2 - масштабування). Систему локального збереження результатів та проміжних обчислень TCA Controller Storage System під'єднано до 16 порту керованого комутатора IB1.

Як конструктив було вибрано єдиний корпус, що являє собою осередок обчислювальної шафи. Це пов'язане з тим, що, з одного боку, при необхідності можна декілька модулів розміщати в єдиному корпусі, а з іншого боку - при такому підході забезпечується компактність,

успішне охолодження й легкий доступ до гнізд і елементів плат, які налагоджуються. Обчислювальна система включає вертикальне, паралельне стосовно одне одного, розташування системних плат, що відповідає ідеї "Blade" - серверів.

Після подачі живлення на блок (ATX) майстер-вузла та зовнішнього сигналу PUSK з панелі модуля керування П01 розпочинається запуск та ініціалізація роботи майстер-вузла багатопроцесорної системи.

Безпосередньо завантаження ОС може здійснюватися з жорсткого диска, або з CD/DVD-пристрою. Після завантаження операційної системи запускається спеціально орієнтований конфігураційний скрипт, який налаштовує роботу DHCP-сервера. Крім того, тут визначається кількість обчислювальних вузлів системи, у разі потреби налаштовується доступ до середовища Інтернет, або до зовнішньої мережі. При цьому задаються основні налаштування й параметри. Послідовна подача живлення на блоки живлення (ATX) та ініціалізація slave-вузлів зменшує необхідну потужність блока UPS, запускає всі обчислювальні slave-вузли та завантажує на них операційні системи. Завантаженням та налаштуванням усіх обчислювальних вузлів кластера завершується робота відповідного скрипту і система готова до виконання паралельних обчислень.

Майстер-вузол (PM001) через комутатор KG1 SW1 забезпечує спрямування потоку даних, пов'язаних із керуванням, діагностикою. В свою чергу slave-вузли відповідно до алгоритму розв'язування задач і перебігу процесів реалізують режим необхідних обчислень. Обмін даними між обчислювальними вузлами та завантаженням умов задач винесено в окрему мережу, яка організована за допомогою керованого комутатора KIB IB1. Для досягнення максимальної ефективності роботи кластерної системи використовуються одно чи двоportові адаптери Infiniband та здійснюється процес реконфігурації структури другої мережі відповідно до специфіки розв'язуваних задач. Результати проміжних та остаточних обчислень передаються в майстер вузол через керований комутатор Infiniband KIB. При цьому управління та передача відповідних даних із slave-вузлів відбувається за допомогою мережевих адаптерів HCA (Host Channel Adapters). Безпосередньо зберігання даних обчислень з метою їх подальшої обробки виконується за допомогою мережевого адаптера TCA (Target Channel Adapters).'

Причинно-наслідковий зв'язок між сукупністю істотних ознак корисної моделі 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.

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 першу 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.

Для забезпечення високої надійності системи електроживлення багатопроцесорної системи напруга подається через безперебійний блок живлення (UPS), який під'єднаний до модуля керування П01, від нього через силові мережеві інтерфейси (розгалужувачі) струм надходить у блоки живлення головного модуля (ATXm) і slave-вузлів системи (ATX). Таким чином, у кожному лезі модулі багатопроцесорної обчислювальної системи присутні однотипні блоки живлення. Даний підхід зменшує стрибки напруги при вмиканні блоків живлення, збільшує надійність системи, реалізує режим їхнього оптимального завантаження та дозволяє зменшити споживану електроенергію обчислювальної системи в цілому. Описаний інтерфейс електроживлення обчислювальних вузлів модуля багатопроцесорної системи дозволив спростити структуру цієї операції, зрештою істотно знизити вартість обчислювальної системи, використовуючи один UPS на весь модуль.

Завдяки запровадженню зазначеного 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 енергоживлення та режиму її резервування дало змогу підвищити надійність функціонування такої системи.

Обмін даними між обчислювальними вузлами винесено в окрему мережу з використання технології InfiniBand, що збільшило швидкість системи в цілому і суттєво знизило латентність (завантаження каналів) мережі, яка з'єднує вузли кластера.

Використання принципу RDMA (Remote Direct Memory Access – віддалений прямий доступ до пам'яті) дозволяє передавати дані без додаткової буферизації й не вимагає активної роботи ОС, а також бібліотек або додатків. Це дозволило суттєво зменшити навантаження на процесори системи під час передачі даних, тим самим звільняються обчислювальні ресурси процесорів і зменшується латентність у середовищі передачі даних.

Застосування двопортових HCA - адаптерів за рахунок режиму 4 x агрегації каналів мереженого інтерфейсу дозволило змінювати конфігурацію мережі, підвищивши її пропускну здатність. Реалізація реконфігурованої мережі дозволяє підвищити ефективність кластерної системи, адаптуючи структуру її мережі для вирішення кожного конкретного типу завдань.

Як обчислювальні платформи було вибрано процесори Intel Core i7 та материнські плати на базі чипсетів серії Intel Z77, яка містить два вільних слоти PCI Express 3.0 8 line, це дозволило суттєво збільшити швидкість обчислень та передачі даних (максимальна швидкість передачі даних в дуплексному режимі відповідає 64 Гбіт/с).

Мережеве завантаження системи і введення механізму резервування ключових компонентів, а також використання блоків живлення для кожного леза багатопроцесорної системи дозволяє підвищити надійність функціонування модуля системи.

Корисна модель та принцип роботи високоефективної багатопроцесорної системи з розширюваною областю обчислень пояснюється кресленнями, де зображено;

фіг. 1 - блок-схема будови модуля багатопроцесорної системи;

фіг. 2 - схема сполучення інтерфейсів для двох модулів багатопроцесорних систем;

фіг. 3 - структурна схема модуля багатопроцесорної системи;

фіг. 4 - структурна схема мереженого інтерфейсу типу зірка;

фіг. 5 - структурна схема мереженого інтерфейсу типу кільце;

фіг. 6 - структурна схема мереженого інтерфейсу типу лінійка;

фіг. 7 - структурна схема мереженого інтерфейсу типу сітка;

фіг. 8 - криві залежності часу обчислення однієї ітерації від розміру області обчислень багатопроцесорної системи.

Особливість блок-схеми модуля (фіг. 1) полягає в тому, що всі обчислювальні вузли модуля високоефективної багатопроцесорної системи підвищеної готовності містять процесор Intel Core LGA1155f(1) з інтегрованим двоканальним контролером пам'яті DDR3 up 1600MHz (2), а також інтегрованим контролером Thunderbolt(3), з підтримкою інтерфейса PCI Express 3.0 16 line (4), двох каналів PCI Express 3.0 8 line(5), а також підтримкою PCI Express 3.0 8 line (6), процесор приєднано шиною DMI2.0 до північного моста Intel Z77 (7), чипсет має підтримку інтерфейса Display support (&), в нього інтегровано контролер стандарту (9) USB 2.0 на 10 портів та USB 3.0 на 4 порти, міст містить інтегрований контролер (10) стандарту Gigabit Ethernet зі спеціальним інтерфейсом (LCI/GLCI), до моста підключено інтегрований аудіоконтролер High Definition Audio 7.1(11), чипсет підтримує до 8 портів стандарту PCIe 2.0 × 1 (12), в міст інтегровано 2 порти стандарту (Serial ATA) SATA600 та 4 порти стандарту SATA300 з можливістю організації RAID-масиву (рівня 0, 1, 0+1 (10), 5) та підтримує функцію Matrix (13).

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 (PM001) і три обчислювальні вузли (PN001, PN002, PN003), а також модуль як вузол розширення (майстер-вузол PM011 та обчислювальні вузли PN011, PN012, PN013). 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 даних подається на фіг. 4-7: фіг. 4 моделює топологію типу лінійки, фіг. 5 - кільця, фіг. 6 - зірки, фіг. 7 - сітки.

Розрахунок ефективності заявленої обчислювальної системи ілюструється поданими нижче аналітичними співвідношеннями.

Отже, розглядається задача розширення області обчислень шляхом збільшення числа вузлів кластерної системи. При цьому вважатимемо, що область обчислень рівномірно розподіляється між вузлами кластерної системи. При таких умовах визначено T_{ex} - час граничного обміну даними між вузлами кластера, с Відзначимо, що якщо час обчислення ітерації залежить лише від потужності процесора, то час граничного обміну даними визначається розміром різницевої сітки, кількістю вузлів кластерної системи і пропускну

спроможністю обчислювальної мережі. Отже, величину T_{ex} можна визначити за таким співвідношенням:

$$T_{ex} = \frac{m \cdot N \cdot \sqrt{\frac{S}{\pi}}}{k \cdot d \cdot V_p} \quad (1)$$

Значення m може дорівнювати одиниці для одностороннього режиму граничного обміну

даними або двом для двостороннього, V_p - пропускна спроможність порту мережевого інтерфейсу (Гбіт/с), N - число вузлів багатопроцесорної системи, S - загальний обсяг області обчислень багатопроцесорної системи, k - кількість каналів зв'язки обчислювальної мережі, які працюють одночасно (кількість обчислювальних мереж), d - напівдуплексний ($d=1$) або дуплексний ($d=2$) режим роботи обчислювальної мережі кластерної системи.

В даному класі задач усі обчислення виконуються на базі різницевої сітки. До того ж, для аналізу ефективності багатопроцесорної системи найважливішим параметром буде час обчислення однієї ітерації (T_{it}) відносно області обчислень. Тоді в умовах застосування багатопроцесорної системи загальний час однієї ітерації визначатиметься на підставі такого співвідношення:

$$T_{it} = T_c^N + T_{ex} \quad (2)$$

Тут T_c^N є часом обчислення однієї ітерації при використанні N обчислювальних вузлів, c . Очевидно, для випадку, коли $N=1$, одержують, що

$$T_{it} = T_c^1 \quad (3)$$

Тут T_c^1 - час обчислення однієї ітерації для одно процесорної обчислювальної системи.

Аналіз співвідношення (1, 2) показує, що при збільшенні області обчислень в N разів час обчислення задачі збільшується як $N^{3/2}c$ деяким коефіцієнтом, що залежить від обсягу оперативної пам'яті вузла, пропускної спроможності мережі кластера і характеру обміну даними між обчислювальними вузлами, тобто:

$$T_{it} = T_c^N + N^{3/2} \cdot f(m, R, V) \quad (4)$$

Аналіз співвідношення (4) показує перспективність застосування сучасних комунікаційних технологій, таких як InfiniBand, а також багатоядерних обчислювальних платформ.

На фоні проведених досліджень розглянемо випадок гіпотетичного комп'ютера з необмеженим обсягом пам'яті. Так, з урахуванням співвідношення (3), одержуємо:

$$T_c^1(S) = \frac{S_i}{V_c} \quad (5)$$

У виразі (5) загальний обсяг області обчислень гіпотетичного комп'ютера можна подати так:

$$S_i = i \cdot R \quad (6)$$

тут i - коефіцієнт, що визначає зміну області обчислень гіпотетичного комп'ютера.

Аналіз співвідношень (5), (6) показує, що при збільшенні загального обсягу обчислень в N разів, час обчислення задач росте лінійно з деяким коефіцієнтом, що залежить від обчислювальних можливостей тих процесорів, які використовуються в системі.

Відповідно до виведених співвідношень були проведені обчислювальні експерименти, для комп'ютерної платформи, оснащеної процесором Intel Core i7. Результати моделювання представлені у вигляді графічних залежностей на фіг. 8. Як видно з фіг. 8, час обчислення однієї ітерації при збільшенні області обчислень багатопроцесорної системи змінюється за

нелінійною залежністю (крива 1, T_{it}). Така залежність показує, що при збільшенні області обчислень в N разів час обчислення задачі зростає як $N^{3/2}$ з деякими коефіцієнтами, що залежать від обсягу оперативної пам'яті вузла кластера, пропускної спроможності мережевого інтерфейсу та характеру обміну даними між обчислювальними вузлами. Разом з цим, час обчислення однієї ітерації для гіпотетичного комп'ютера з необмеженою пам'яттю, як і

очікувалося, збільшується за лінійним законом (лінія 2, T_{id}). При цьому кут нахилу такої лінії

визначається характеристиками обчислювальної платформи, яка використовується в системі. Результати моделювання показали таку загальну тенденцію.

Визначення 1. Точку перетину ліній часу обчислення однієї ітерації ідеального комп'ютера та реальної багатопроцесорної системи називається точкою ідеальної рівноваги.

5 Визначення. 2. Точка ідеальної рівноваги, що відповідає деякому значенню області обчислень S_n називається ідеальним значенням області обчислень S_{id} .

При цьому очевидно, що, у випадку, коли $S_n < S_{id}$, час обчислення багатопроцесорної системи стає менше часу обчислення ідеального комп'ютера". Це пояснюється збільшенням

числа процесорів багатопроцесорної системи. З іншого боку, коли $S_n > S_{id}$, то в силу істотного впливу часу, граничного обміну даними на загальний час розв'язування задачі, на фоні розширення області обчислення, час розв'язування задачі для реальної багатопроцесорної системи істотно збільшуватиметься в порівнянні з ідеальним комп'ютером. При цьому стає очевидним, що перспективним режимом використання багатопроцесорної системи є випадок, коли $S_n < S_{id}$.

15 Особливості функціонування модуля багатопроцесорної обчислювальної системи підвищеної готовності полягають у тому, що після подачі живлення на блок живлення майстер-вузла (ATX) та зовнішнього сигналу PUSK з панелі модуля керування П01 розпочинається 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 спеціально орієнтований конфігураційний скрипт, який налаштовує роботу DHCP-сервера. Крім того, тут визначається кількість обчислювальних вузлів системи, у разі потреби налаштовується доступ до середовища Інтернет, або до зовнішньої мережі. При цьому задаються основні налаштування й параметри. Послідовна подача живлення на блоки живлення. (ATX) та ініціалізація slave-вузлів зменшує необхідну потужність блока UPS, запускає всі обчислювальні slave-вузли та завантажує на них операційні системи. Після завантаження та налаштування всіх обчислювальних вузлів кластера завершується робота відповідного скрипту і система готова до виконання паралельних обчислень.

30 Майстер-вузол (PM001) через комутатор KGI SW1 забезпечує спрямування потоку даних, пов'язаних із керуванням, діагностикою. В свою чергу slave-вузли відповідно до алгоритму розв'язування задач і перебігу процесів реалізують режим необхідних обчислень. Обмін даними між обчислювальними вузлами та завантаженням умов задач винесено в окрему мережу, яка організована за допомогою керованого комутатора KIB IB1. Для досягнення максимальної ефективності роботи кластерної системи здійснюється процес реконфігурації структури другої мережі відповідно до специфіки розв'язуваних задач. Прийом/передача даних у slave-вузлах відбувається без буферизації за допомогою керованого комутатора IB. Результати проміжних та остаточних обчислень передаються в майстер вузол через керований комутатор Infiniband KIB. При цьому управління та передача відповідних даних із slave-вузлів відбувається за допомогою мережесовмісних адаптерів HCA (Host Channel Adapters). Безпосередньо зберігання даних з метою їх подальшої обробки виконується за допомогою мережевого адаптера TCA (Target Channel Adapters).

45 Уведення в багатопроцесорну систему окремої обчислювальної мережі обміну даними стандарту InfiniBand, реалізації механізмів агрегації каналів мережевого інтерфейсу та підтримка VLAN, спеціально організованого режиму обміну даними в мережі керованого комутатора KIB, а також шляхом розробки режиму мережевого завантаження процесорів та механізму резервування ключових компонентів модуля дозволило:

50 - по-перше, завдяки застосування технології InfiniBand було закладено наступні пріоритети: низька латентність, масштабованість, можливість резервування, можливість підбору необхідних швидкостей із заданого діапазону швидкостей, що, в свою чергу, дозволило застосовувати розроблену систему для розв'язування сильнозв'язаних задач та задач з розширюваною областю обчислень;

- по-друге, через термінал, або WEB-інтерфейс змінювати конфігурацію обчислювальної мережі, адаптуючи її структуру для вирішення кожного конкретного типу задач;

55 - по-третє, завдяки застосуванню засобів RDMA технології InfiniBand, формуванню окремої обчислювальної мережі, агрегації каналів та реалізації механізмів VLAN з'явилася можливість прямого обміну даними між оперативною пам'яттю вузлів багатопроцесорної системи, що дозволило підвищити швидкість обчислень під час розв'язування сильнозв'язаних задач,

забезпечити високошвидкісний доступ до пам'яті вузлів кластера, розвантажити CPU при обміні даними та знизити завантаження мережі;

- по-четверте, за рахунок використання адаптерів ConnectX забезпечуються нові можливості "коннективності" з різними обчислювальними середовищами, що зумовлює підвищення продуктивності всієї обчислювальної системи та дозволяє розвантажити центральний процесор від навантаження по обслуговуванню трафіку InfiniBand;

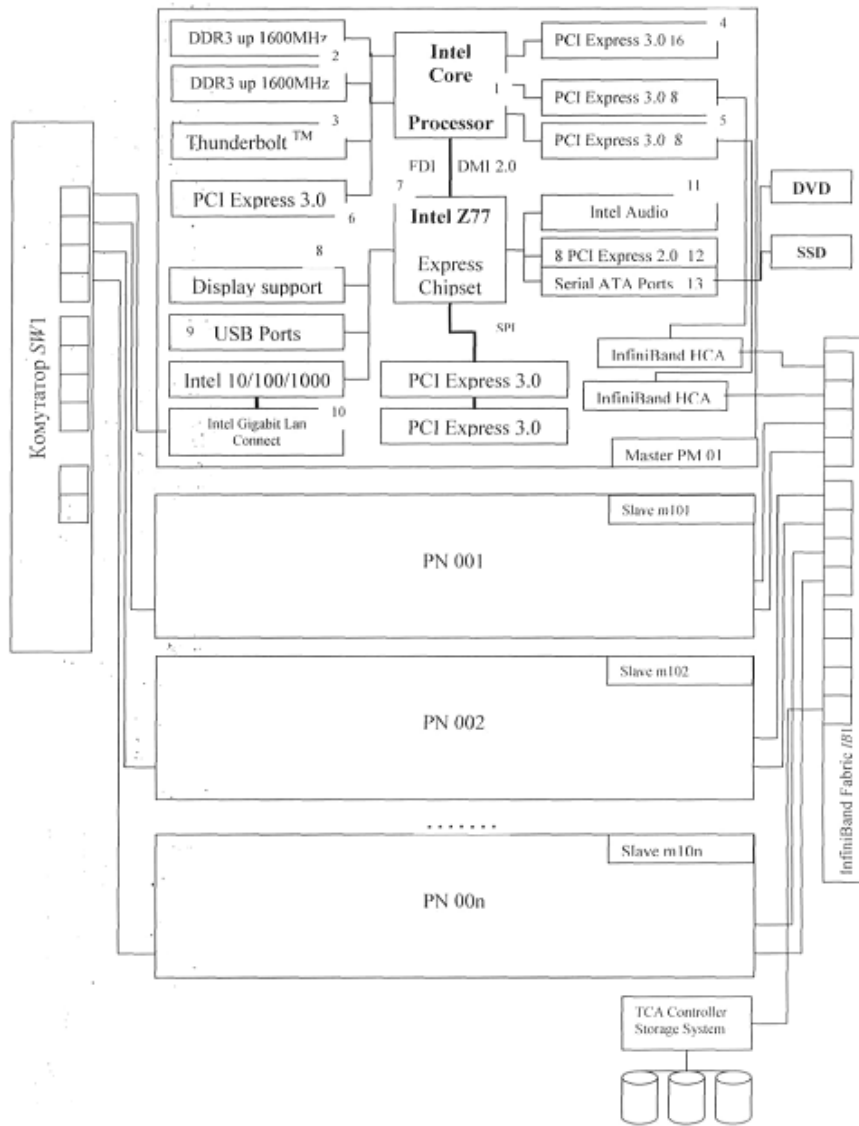
- по-п'яте, підвищити ефективність кластерної системи, адаптуючи структуру її мережі до розв'язування задач кожного конкретного типу;

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

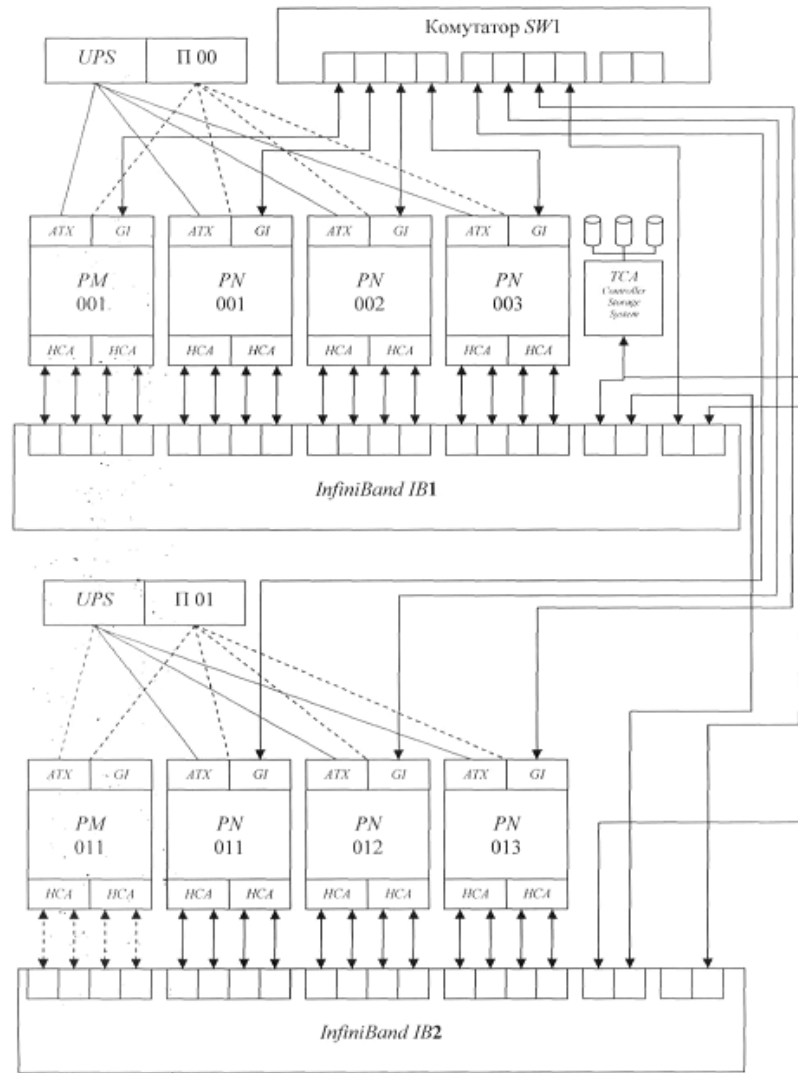
- по-сьоме, суттєво збільшити швидкість передачі даних між вузлами багатопроцесорної системи та зменшити латентність у середовищі передачі пакетів даних завдяки застосуванню комутованого середовища в мережі обміну даних зі сполуками "точка-точка".

ФОРМУЛА КОРИСНОЇ МОДЕЛІ

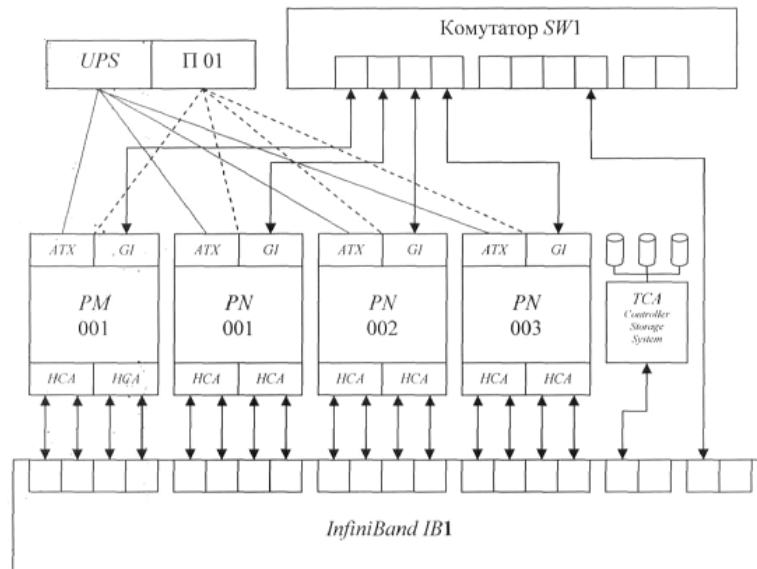
Модуль високоефективної багатопроцесорної системи з розширюваною областю обчислень, що містить майстер-вузол і обчислювальні slave-вузли, призначений для побудови багатопроцесорних систем, який **відрізняється** тим, що містить два керовані комутатори SW1 (GigabitEthernet) та JB1(InfiniBand), систему локального збереження результатів та проміжних обчислень, механізм резервування ключових компонентів, а також передбачає мережеве завантаження вузлів по мережі GI через комутатор SW1, кожен вузол модуля має персональний блок живлення стандарту ATX, slave-вузли додатково обладнані двома двопортовими зовнішніми адаптерами HCA InfiniBand, при цьому між портами 0, 1, 2, 3, 4, 5, 6, 7, 8 комутатора SW1 створюється віртуальна локальна мережа (VLAN); майстер-вузол модуля в мережі діагностики, завантаження та управління з'єднується за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу мережевої плати Gigabit Ethernet з портом 0 керованого комутатора SW1, в мережі обміну даних майстер-вузол з'єднується зі slave-вузлами двома двопортовими мережевими адаптерами InfiniBand з портами керованого комутатора IB1, обчислювальні вузли в мережі діагностики, завантаження та управління з'єднуються за допомогою входу/виходу внутрішнього двоспрямованого інтерфейсу з портами керованого комутатора SW1, в мережі обміну даних з'єднуються за допомогою двох двопортових мережеских адаптерів InfiniBand з портами керованого комутатора IB1, систему локального збереження результатів та проміжних обчислень під'єднано до порту керованого комутатора IB1, інтерфейс налаштування керованого комутатора IB1 з'єднується з портом керованого комутатора SW1.



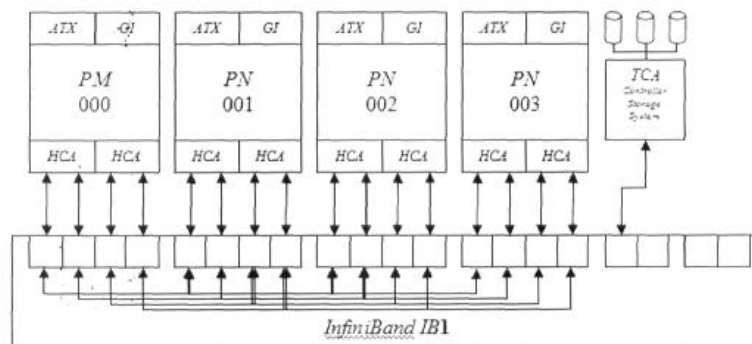
Фиг. 1



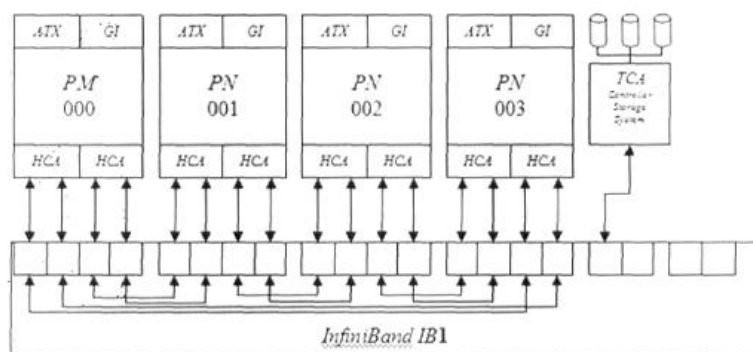
Фиг. 2



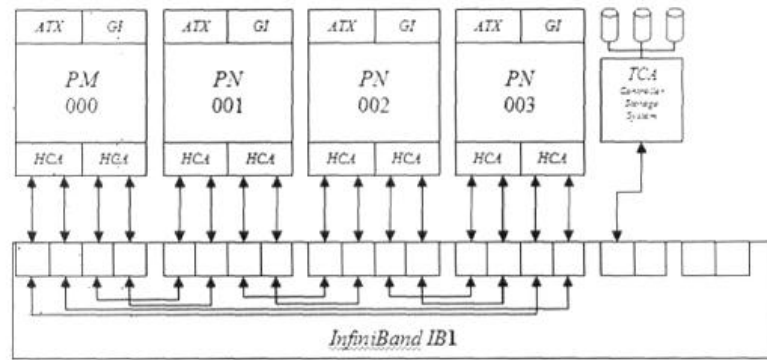
Фиг. 3



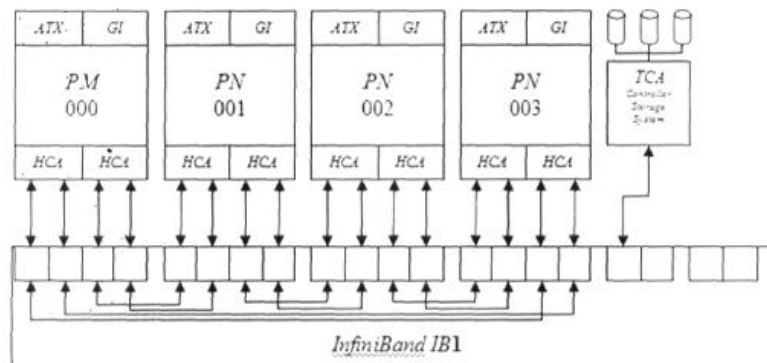
Фиг. 4



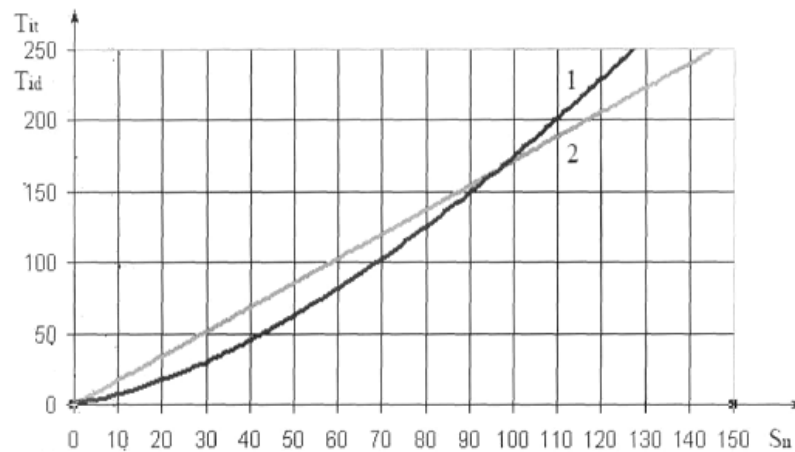
Фиг. 5



Фиг. 6



Фиг. 7



Фиг. 8

Комп'ютерна верстка В. Мацело

Державна служба інтелектуальної власності України, вул. Василя Липківського, 45, м. Київ, МСП, 03680, Україна

ДП "Український інститут інтелектуальної власності", вул. Глазунова, 1, м. Київ – 42, 01601