УДК 004.002
Мельник В.М.
Луцький національний технічний університет

МОДЕЛЬ СИСТЕМИ УПРАВЛІННЯ ДАНИМИ ЗВ'ЯЗКУ ДЛЯ ПРОЦЕСОРІВ МЕРЕЖЕВИХ ВУЗЛІВ (ПДВЗ) В РОЗПОДІЛЕНОМУ СЕРЕДОВИЩІ ПРОЦЕСІВ.

Процесори мережевих вузлів (ПМВ), які можуть виконувати програми додатків і взаємодіяти один з одним за допомогою високошвидкісних локальних аналітичних вузлів являються важливими компонентами для розподілених сумісно з'єднаних систем. ПДВЗ являється системою управління даними зв'язку для локальних аналітичних вузлів, що надає уніфіковані інтерфейси комунікації і швидкодійний та високонадійний зв'язок. Проаналізовано підходи адресації, які об'єднують програмні інтерфейси зв'язку вузла до терміналу та внутрішні зв'язки між вузлами комунікації. Запропоновано також розширену схему інтерфейсу сокетів, технічну схему для відтворення покращеного зв'язку та технічні вузли для високонадійної групової комунікації між ними на декількох локальних аналітичних вузлах. ПДВЗ може бути реалізована на базі ядра UNIX на відмовостійких системах міні-комп'ютерів, які б працювали фактично у власних он-лайн системах.

Ключові слова: ПРОЦЕСОР ДАНИХ ВУЗЛІВ ЗВ'ЯЗКУ, ПРОЦЕСОР МЕРЕЖЕВИХ ВУЗЛІВ, ЛОКАЛЬНІ АНАЛІТИЧНІ ВУЗЛИ, СИСТЕМА УПРАВЛІННЯ ДАНИМИ ЗВ'ЯЗКУ, ЯДРО UNIX
Рис. 5. Літ. 8.

Мельник В.М. Модель системмы управления данными связи для процессоров сетевых узлов в распределенной среде процессов. *Процессоры сетевых узлов (ПСУ), которые могут исполнять программы приложений и взаимодействовать один с другим через высокоскоростные локальные аналитические узлы являются важными компонентами для распределенных совместно связанных систем. ПДУС является системой управления данными связи для локальных аналитических узлов, что представляет унифицированные интерфейсы коммуникации, быстродействующую и высокосовершенную связь. Проанализировано подходы адресации, которые объединяют программные интерфейсы связи узла к терминалу и внутренние связи между узлами коммуникации. Предложено также расширенную схему интерфейса сокетов, техническую схему для отображения улучшенной связи и технические узлы для высоконадежной групповой коммуникации между ними на нескольких локальных аналитических узлах. ПДУС может быть реализован на основании ядра UNIX на отказоустойчивых системах мини-компьютеров, которые работали бы фактически в собственных он-лайн системах.*

Ключевые слова: ПРОЦЕСОР ДАНИХ УЗЛОВ СВЯЗИ, ПРОЦЕСОР СЕТЕВЫХ УЗЛОВ, ЛОКАЛЬНЫЕ АНАЛИТИЧЕСКИЕ УЗЛЫ, СИСТЕМА УПРАВЛЕНИЯ ДАННЫМИ СВЯЗИ, ЯДРО UNIX

Melnyk V.M. Data communication management system model for network node processors in a distributed processing environment. *Network node processors, which can execute application programs and can cooperate with each other via high-speed local analytic nodes, have become essential components for distributed on-line systems. Data communication node processor is a data communication management system for network node processors. Data communication node processor provides unified communication interfaces and high-performance and highly reliable communication services. We present an addressing technique which unifies program interfaces for node-to-terminal communications and inter-node communications. Socket interface extensions for improving performance and techniques for highly reliable inter-node multicast communications on multiple local analytic nodes are also presented. Data communication node processor has been implemented in UNIX kernel on fault-tolerant mini-computers, and is currently operating in actual on-line systems.*

Key words: DATA COMMUNICATION NODE PROCESSOR, NETWORK NODE PROCESSOR, LOCAL ANALYTIC NODES, DATA COMMUNICATION MANAGEMENT SYSTEM, KRENEL UNIX

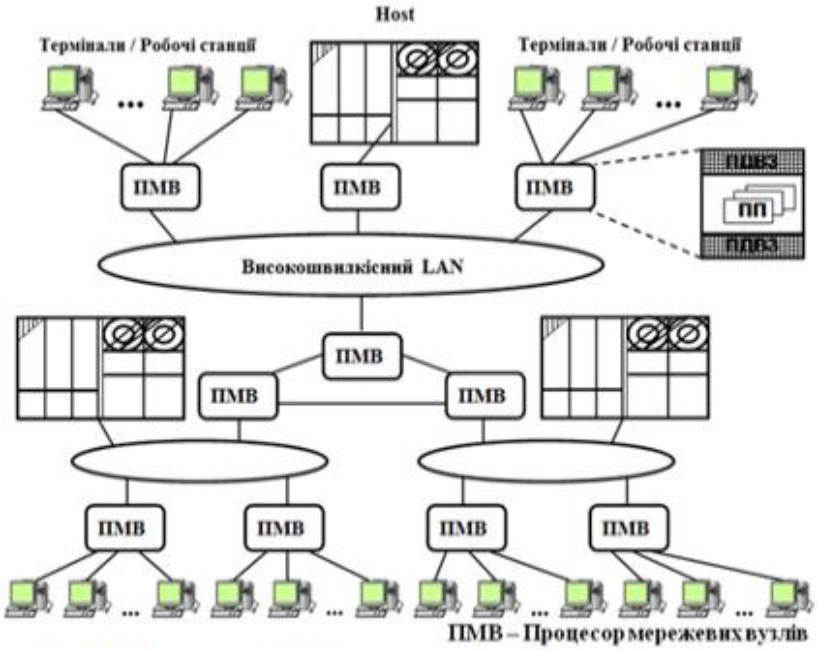


Рис. 1. Процесори вузлів зв'язку і ПДВЗ в розподілених системах он-лайн обробки транзакцій

Постановка проблеми. В архітектурі он-лайн обробки транзакцій система як банківські системи або он-лайн системи інформаційного обслуговування змінювалися від централізованого до розподіленого напрямку. Ці зміни спрямовані на покращення масштабованості системи і обумовлені зменшенням розмірів процесорів та розвитком високошвидкісних локальних аналітичних вузлів. Розподіл функцій он-лайн обробки транзакцій потребує мережевих вузлових процесорів (ПМВ), які можуть керувати декількома терміналами і прикладними програмами host-комп'ютерів, а також можуть співпрацювати один з одним через локальні аналітичні вузли. ПДВЗ являється системою управління даними зв'язку (СУДЗ), що забезпечує програмну платформу для прикладних програм ПМВ в розподіленому середовищі (рис. 1). При побудові розподіленої системи он-лайн обробки транзакцій з ПМВ та локальними аналітичними вузлами накладаються три вимоги на СУДЗ ПМВ. Потрібна висока продуктивність прикладних програм, тому що програми додатків он-лайн обробки транзакцій часто удосконалюються або модифікуються в сторону покращення якості послуг для кінцевого користувача; то ж СУДЗ з ПМВ повинні підтримувати уніфіковані і прості інтерфейси комунікації для спрощення програмування. Висока продуктивність та досить надійні послуги зв'язку між вузлами також необхідні для того, щоб зберегти заслуги високої масштабованості, оскільки розподіл ПМВ створює додаткове навантаження на зв'язок між внутрішніми вузлами, які відсутні в централізованих системах.

Аналіз останніх досліджень і публікацій. Для забезпечення єдиного і простого інтерфейсу зв'язку в розподіленому середовищі, були запропоновані кілька традиційних підходів в розподілених операційних системах та мережевих операційних системах [1], [2]. Однак ці підходи виключають термінали з їх прозорих об'єктів зв'язку, які є важливими в застосуванні прикладних програм для ПМВ.

Для підвищення ефективності взаємодії між внутрішніми вузлами зв'язку, були запропоновані два підходи. Одним з них є використання високошвидкісних і легковагових протоколів [3]. Однак, нестандартні протоколи страждають від обмеженого підключення, що робить його важким для НПП при підключенні до процесорів різних виробників. Інший підхід полягає у створенні зв'язків між внутрішніми вузлами у верхній частині відкритих протоколів та покращенні його роботи. В роботі [4] зв'язок між внутрішніми вузлами побудовано у верхній частині інтерфейсу сокетів [5] і UDP/IP. Проте, як зазначається, інтерфейс сокета

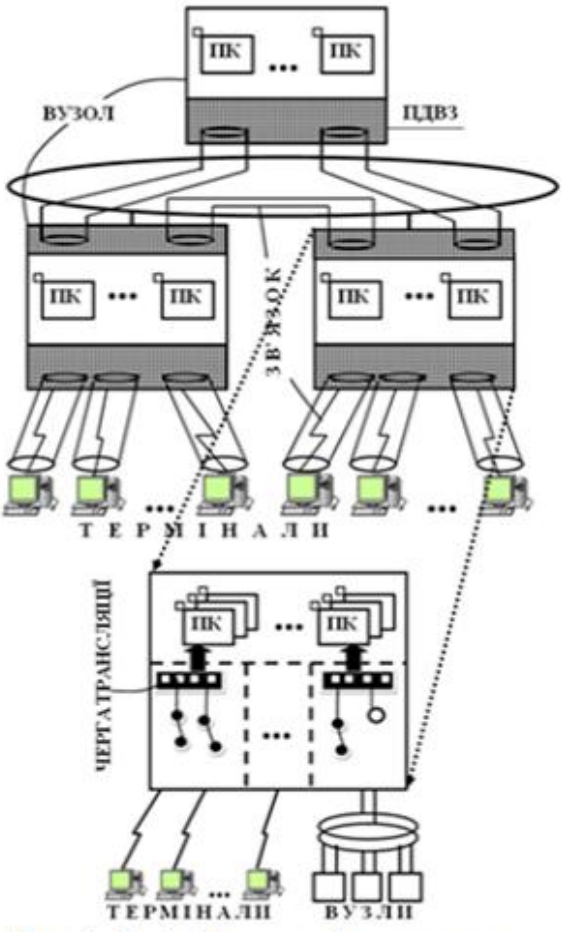


Рис. 2. Логічні мережеві компоненти

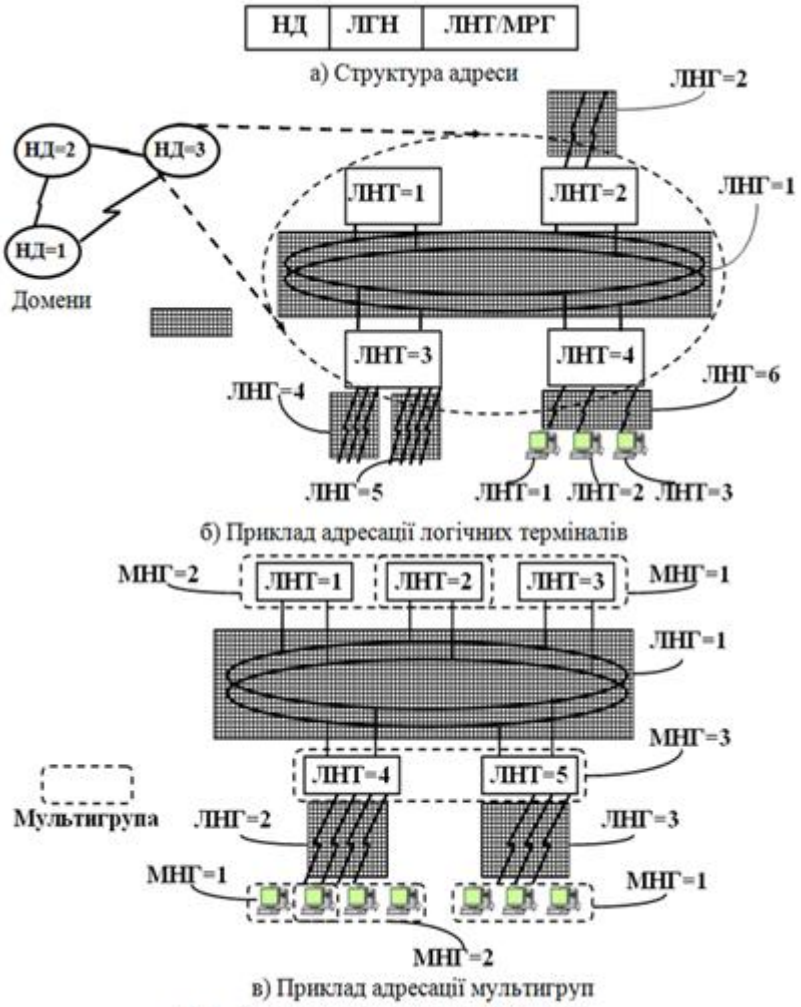


Рис. 3. Система адресації ПДВЗ

має низьку продуктивність. Крім того, управління зв'язку інтерфейсу сокетів не підходить для однакових за рівнем внутрішніх вузлів зв'язку згідно припущень моделі клієнт-сервер.

Для підвищення надійності між внутрішніми вузлами групової комунікації техніка мультиплексних локальних мереж була запропонована в [6], де вузли з'єднані декількома локальними аналітичними вузлами і перекликаються один з одним враховуючи пріоритет кожного локального аналітичного вузла. Вузол-відправник розсилає пакети до кожного локального аналітичного вузла. Вузол-приймач отримує пакети, що передані через найвищий пріоритет локальних аналітичних вузлів. Коли локальний вузол виявляє відсутність найвищого пріоритету, він посилає повідомлення із запитом на зміну пріоритету на інші вузли. Однак, деякі вузли не будуть спроможні отримати це повідомлення-запит, оскільки воно передається з низькопріоритетного локального аналітичного вузла, і вони не можуть отримувати повідомлення після зміни пріоритету цього вузла.

Виділення невирішених раніше частин загальної проблеми. Для вирішення цих проблем спочатку запропонуємо технічне рішення, яке може забезпечити чистоту та уніфіковані інтерфейси терміналу зв'язку і зв'язку між внутрішніми вузлами для однакових за рівнем групових комунікацій. Далі, приймемо застосування буферів пам'яті (систему буферизації пам'яті) не тільки для TCP/IP на базі внутрішнього зв'язку між вузлами, але і для терміналу зв'язку. Для покращення сокетного інтерфейсу запропоновано інтерфейс сокетів реального часу. Сокет реального часу може бути прямо зв'язаний з ПДВЗ менеджером зв'язку між вузлами, який дозволяє асинхронне і однорівневе встановлення TCP-з'єднання між вузлами. Буферизація пам'яті об'єднує буферну систему управління і сокет реального часу може перешкоджати копіювання даних і проміжних процесів з системи, що покращує продуктивність між внутрішніми вузлами зв'язку. Нарешті, можна запропонувати технологію для визначення порядкового номера скидання яка забезпечує високу взаємодію між внутрішніми вузлами групової комунікації і може бути побудована на базі мультиплексних локальних аналітичних вузлів без будь-якого механізму перемикання між ними.

Впровадження ПДВЗ з вищеописаними методами в ядро UNIX може бути здійснене на відмовостійкому міні-комп'ютері, який в даний час підключений і працює в відомих системах інформаційного обслуговування. Опишемо архітектуру системи ПДВЗ і підходи вирішення та методи підвищення продуктивності і надійності між вузлами зв'язку.

Виклад основного матеріалу дослідження.

Архітектура системи ПДВЗ. Мережеві логічні компоненти ПДВЗ показано на рис. 2.

Вузли та термінали: процесор, що працює в ПДВЗ, називається вузлом і термінали кінцевих користувачів та процесори, що працюють від інших СУДЗ, називаються терміналами. Тут немає зв'язків між вузлами типу відношення ведучий-підлеглий чи клієнт-сервер, – всі вузли повинні бути рівні.

Рис. 4. Структура програмного забезпечення ПДВЗ

Черги транзакцій: Чергами транзакції є черги FIFO, де повідомлення від терміналів або вузлів в черзі просуваються відповідно до коду транзакції, який надається для кожного отриманого повідомлення.

Процеси користувачів: Процеси користувачів (ПК), які отримують послуги зв'язку і послуги виділення буфера від ПДВЗ називаються користувацькими програмами і реалізуються UNIX-процесами.

Зв'язок: Зв'язок є логічною лінією передачі між вузлами або між вузлами і терміналами. Зв'язок між вузлами (міжвузловий) реалізується через TCP/IP з'єднання, а зв'язок між вузлом і терміналом здійснюється по термінальній лінії передачі.

В одному напрямку, ПК встановлюють TCP/IP зв'язок самостійно і відправляють повідомлення на інші вузли за допомогою з'єднання. Однак цей підхід вимагає для ПК виконання складної процедури при створенні зв'язку.

У ПДВЗ внутрівузловий менеджер комунікації встановлює TCP/IP-з'єднання з іншими вузлами при запуску системи. ПДВЗ примножує відправку повідомлень в TCP/IP з'єднання через зв'язок, кінець якого знаходиться в вузлі призначення, тому все, що ПК потрібно зробити при відправленні повідомлення, це вказати адресу вузла призначення і дані для посилання, як буде описано нижче.

Вузол або термінал можуть бути ідентифіковані в єдиному варіанті: за багатосимвольним номером домена (НД), лінійним груповим номером (ЛНГ) і логічним термінальним номером (ЛНТ) чи мультиплексним номером групи (МНГ) (див. рис. 3 (а)). Ці числа визначаються таким чином:

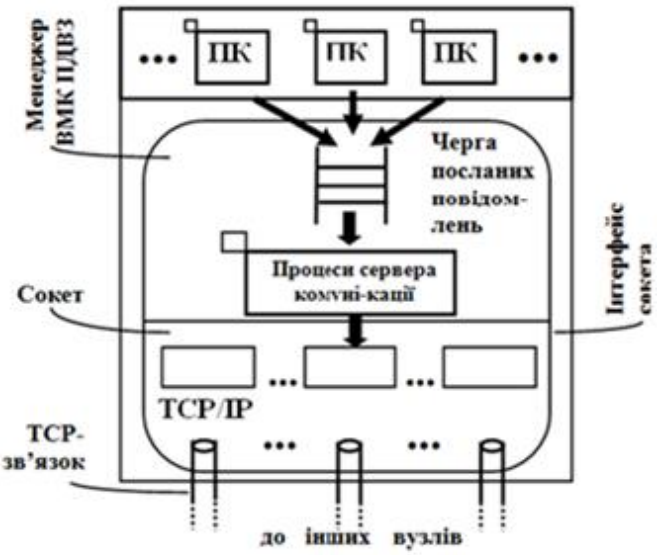
Номер домена (НД): Логічні центри глобальної мережі називаються доменами і НД унікально даються для кожного домена.

Лінійний груповий номер (ЛНГ): група термінальних ліній або локальних мереж, що_підключені до вузлів називається лінійною групою і ЛНГ унікальні для кожної групи. Термінали або вузли, що контактують з цією групою, повинні спілкуватися на одних і тих же протоколах зв'язку. Лінійна група також являється модулем управління ресурсами. Наприклад, черга транзакції генерується для кожної лінійної групи.

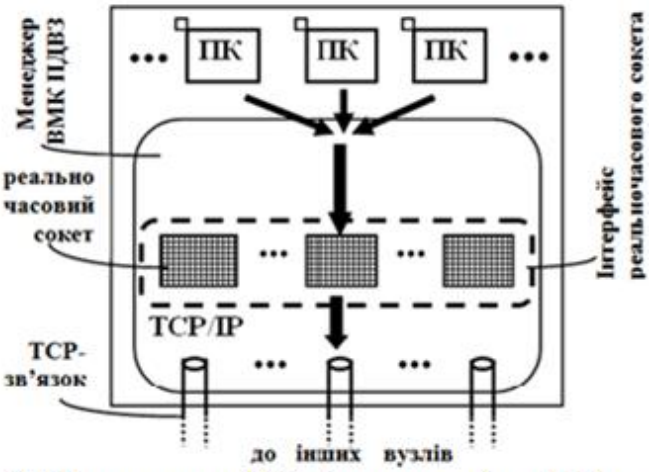
Логічний номер терміналу (ЛНТ): Термінали і вузли, приєднані до лінійної групи називають логічними терміналами, яким ЛНТ призначаються однозначно.

Мультиплексний номер групи (МНГ): Мультиплексна група є підпорядкована логічним терміналам в лінійній групі і являється одним із об'єктів мультиплексного обслуговування. МНГ однозначно призначаються кожній з мультиплексних груп. Багато мультиплексних груп можуть бути визначені в лінійній групі і логічний термінал може належати до декількох мультиплексних груп.

Приклади адресації логічних термінa-



(А) Мультиплексне TCP-з'єднання за допомогою інтерфейсу конвенційних сокетів



Б) Мультиплексне TCP-з'єднання для інтерфейсу сокета реального часу і ПДВЗ

Рис. 5. Мультиплексування TCP-з'єднань

лів і мультиплексних груп представлені на рис. 3 (б) і (в), відповідно. Запропонована система адресації об'єднує зв'язки терміналу і міжвузлові комунікації. Крім того, це звільняє ПК від складних процедур адресації, необхідних в системі IP-адресації [7], коли відправляється повідомлення на вузли з багатьма розгалуженнями, так як лінійна група приховує мультиплексні структури локальних аналітичних вузлів.

Призначені два комунікаційних примітиви, putran і getran, для ПК.

putran: примітив для відправки повідомлень логічному термі налу або мультиплексній групі. Його параметрами є кінцеві адреси, обслуговуючі ID які описують одиничний або мультиплексний, і інформацію для визначення відправки даних: код транзакції (КТ), адресні дані, та дані їх довжини. Якщо ПК вказує ЛНТ, призначений власному вузлу, повідомлення спрямовуються на той же вузол, що призводить до внутрівузлової комунікації.

getran: примітив для отримання транзакції з черги. Його параметрами ЛНГ і (КТ) для ідентифікації черги транзакції, і адреси буфера для збереження повідомлення.

Структура програмного забезпечення вузла працюючої ПДВЗ показана на рис. 4. Коди і таблиці ПДВЗ компактно реалізовані в просторі ядра, і системні процеси в ПДВЗ відсутні.

Методи для підвищення продуктивності

Уніфікована система буферного управління включає в себе метод буферів пам'яті, який може бути використаний для реалізації інтерфейсу сокетів і TCP/IP. У ПДВЗ, метод буферів пам'яті також застосовується до термінальних зв'язків, які не базуються на TCP/IP протоколі. Таке уніфіковане управління буфером дає можливість для ПДВЗ і для сокетів реального часу обмінятися буферами без копіювання даних (сокет реального часу буде описаний нижче). Крім того, як функції вибору відправки (putran) і вибірки (getran), мережа системних буферів, кожний з яких містять різні повідомлення, можуть бути передані до ПК і прийняті від них одночасно без копіювання даних. Такі особливості роблять можливим для ПК обмін повідомленнями між локальними мережами і лінійними терміналами на високій швидкості і пропускній здатності, які, в свою чергу, є важливими функціями ПМВ.

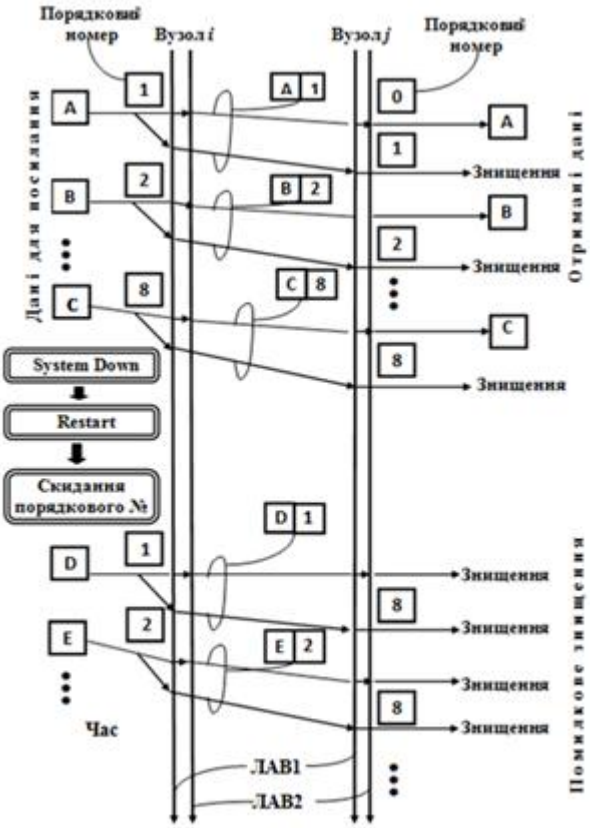


Рис. 6. Помилкове скидання для порядкового номера зв'язку

Буфери пам'яті уніфіковані в систему управління буферами збільшують продуктивність, але це також збільшує і ризик виснаження буфера. Таким чином, в буфер пам'яті додається додатковий механізм регулювання буфером для підтримки надійності системи, як буде описано нижче.

Удосконалення інтерфейсу сокетів полягає в наступному. Зв'язки між вузлами реалізуються через інтерфейс сокета і TCP/IP з'єднання, де ЛНТ вузлів зіставляються з IP-адресами і TCP номерами порта, і повідомлення відправлені від ПК є мультиплексними до TCP/IP зв'язку. Однак, інтерфейси звичайного сокета породжують дві проблеми. По-перше, він допускає відношення клієнт-сервер, а тому серверні вузли повинні починати роботу в першу чергу для того, щоб отримати встановлення з'єднання для запиту від кожного вузла клієнта. Це послідовне створення внутрівузлового зв'язку робить роботу системи більш складною. По-друге, оскільки інші процеси не можуть спільно використовувати сокет, процес комунікації сервера і черга відправки повідомлення потрібні між ПК і сокетом для того, щоб мультиплексувати повідомлення відправки для з'єднання (див. рис. 5 (а)). Важливим при відправленні є те, що процес сервера зв'язку знижує продуктивність ПДВЗ.

Пропонується інтерфейс сокетів в режимі реального часу (ІСРЧ) для подолання цих проблем, які створює звичайний сокет. Коли в режимі реального часу сокет отримує запит встановленого з'єднання від ПДВЗ, він посилає пакет запиту зв'язку (SYN пакет) рівносильному за пріори-

тетом вузлу. Якщо він не може встановити з'єднання, тому що рівноправний вузол недоступний чи відхиляє запит, сокет реального часу утримує сокет в стані з'єднання до тих пір, поки звичайний сокет припинить спробу встановити зв'язок. Ця перевага дозволяє можливість для рівносильного вузла встановити з'єднання як тільки він починає роботу, так що асинхронне створення внутрівузлового зв'язку стає можливим.

Крім того, всі запити безпосередньо зв'язані з сокетом реального часу в неблокуючому режимі, і результати запиту та повідомлення про отримання безпосередньо повідомляють ПДВЗ через переривання програмного забезпечення які супроводжуються параметрами адрес стола сокетів і факторів. ПДВЗ мультиплексує або димультиплексує повідомлення в сторону одного сокета реального часу або від нього. Таким чином, ПК можуть спільно використовувати один і той же сокет і не потрібно ніяких серверних процесів проміжної необхідності, як показано на рис. 5 (б).

Послуги багатоадресного зв'язку також побудовані на сокеті реального часу та UDP/IP.

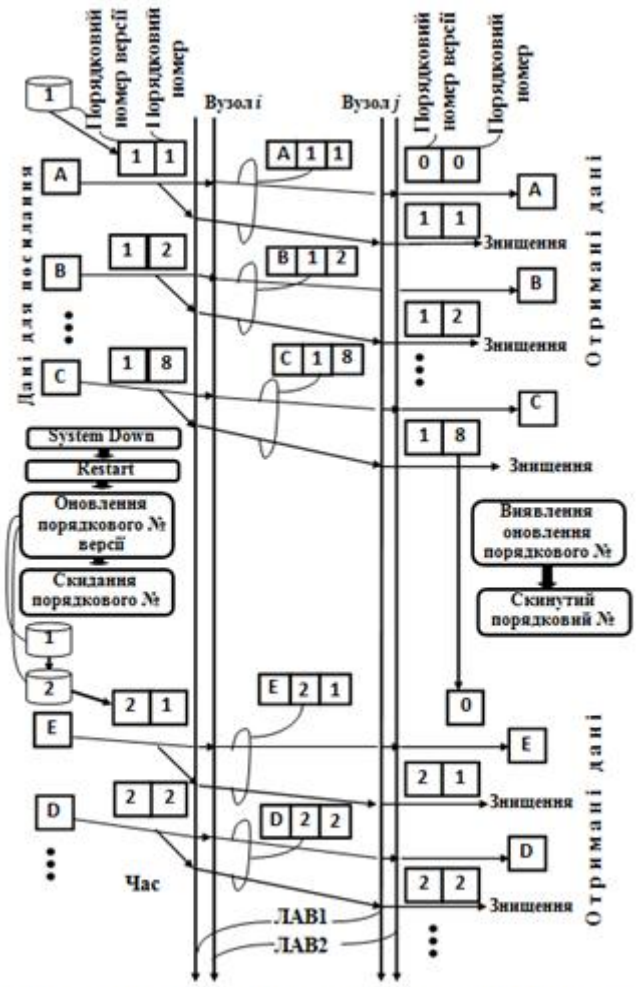


Рис. 7. Виявлення скидання порядкового номера через порядковий номер версії

Методи підвищення надійності

Механізми буферної регуляції включають в себе особливості буферної системи управління ПДВЗ, які вже були описані, і можуть призвести до небажаних побічних ефектів. Наприклад, ненормована ПК або високий трафік лінії терміналу може монополізувати системні буфери, що виснажує їх і спонукає систему до зниження. Щоб уникнути цих проблем, буферний менеджер (mbuf) ПДВЗ регулює буферну систему наступним чином.

В кожен проміжок часу використовується лічильник системних буферів і його верхня межа (MAXBUF) призначена для кожної лінійної групи, де MAXBUF кожної лінії групи визначається як параметр системи конфігурації. Всі запити системних буферів в напрямку до буферного менеджера ПДВЗ повинні вказати ЛНГ і кількість байтів для буферів, де будуть збережені відправлені або отримані повідомлення. Менеджер буферів (МБ) регулює лічильник системних буферів для зазначеної лінійної групи в межах MAXBUF, і якщо він буде перевищувати MAXBUF, то запитувач утримується в стані очікування до тих пір, поки звільниться необхідні буфери або просто відхиляє запит.

Підвищення надійності внутрівузлового зв'язку. Коли вузли з'єднані локальними мережами для підвищення надійності, вони повинні узгодитися про використання в міжвузлових багатоадресних комунікаціях локальної мережі. Приймається механізм паралельного резервування зв'язку, де вузол-відправник мультиадресує пакети на кожен локальний аналітичний вузол з надлишком і вузли-приймачі отримують тільки перший з пакетів який прибув від локальних мультимереж. Надлишкові пакети, що приходять пізніше, відкидаються. Цей механізм не вимагає ні узгодження між локальними мережами, ні використання перемикання між локальними аналітичними вузлами, а тому це усуває комплексний протокол для виявлення відмови таких вузлів або їх збою.

Проблема повинна бути вирішена з виявленням надлишкових пакетів на вузли-приймачі. Механізм традиційних порядкових номерів не достатній, оскільки вузли-приймачі, імовірно, будуть відкидати всі пакети прибуття, коли вузол-відправник скидає порядковий номер після відновлення завершеної його системи. Це помилкове скидання продовжується до тих пір, поки порядковий номер вузла-відправника перевищує останній номер правильно отриманого і збереженого вузлами-приймачами, як показано на рис. 6.

З однієї сторони, вузол-відправник мультиадресує контрольний пакет до кожного члена групи багатоадресної розсилки повідомляючи, що порядковий номер скинуто, і він отримує назад підтвердження. Якщо будь-які вузли не отримають такого підтвердження, вузол-відправник повторно посилає пакет управління для тих вузлів, поки воно не буде правильно прийняте всіма членами групи. Цей підхід вимагає, щоб вузли-відправники знали всіх членів групи багатоадресної розсилки. Однак, не потрібно ніякого протоколу реєстрації приєднання або виходу із групи багатоадресної розсилки між вузлами, щоб дозволити кожному вузлу приєднатися або покинути групу багатоадресної розсилки вільно і швидко. Таким чином, вузли-відправники не мають необхідної інформації про членів багатоадресної групи, і цей підхід не представляється можливим.

Пропонується порядковий номер версії (ПНВ), який дозволяє вузлам-приймачам виявляти скидання порядкового номера у вузлі-відправнику. ПНВ оновлюється при кожному скиданні порядкового номера і розташовується в заголовок пакета в додатку до нього (див. рис. 7). Поточне значення ПНВ зберігається в незалежній пам'яті (наприклад, на диску), а також копія кешується в пам'яті після того, як значення оновлюється при запуску системи. Менеджер комунікації між вузлами (ВМК) використовує кешовані ПНВ. Вузол-приймач захоплює найбільш поточний ПНВ для кожного вузла-відправника і МНГ. Коли вузол отримує пакет з локального аналітичного вузла, він порівнює ПНВ з отриманого пакета з захопленим раніше. Якщо ПНВ з отриманого пакета перевищує захоплений, приймач заміняє значення поточного ПНВ на значення з отриманого пакета, а також заміняє поточний порядковий номер значенням з отриманого пакету. Тому скидання порядкового номера може бути виявлено в вузлі-приймачі, не вимагаючи контрольного пакета для цієї мети і помилкові скидання не відбуваються з тих пір як ПНВ зберігається в незалежній пам'яті, який не може бути втраченим, навіть якщо система зупиняється.

Висновки і перспективи подальших досліджень

Запропонована модель архітектури ПДВЗ-системи і методи для покращення продуктивності та надійності зв'язку. Механізм адресації, заснований на наборі багатосимвольного номера домена, лінійному груповому номері і логічному термінальному номері або багатоадресному груповому номері, надає єдиний простий інтерфейс для ПК. ПДВЗ виключає копіювання буфера і системних процесів за допомогою використання буферів пам'яті, уніфікованого буферного управління та інтерфейсу сокетів реального часу, що підвищує продуктивність процесів зв'язку. Це покращує надійність сервісу комунікації між вузлами через регулювання системних буферів при використанні їх для кожної лінійної групи, і для багатоадресного сервісу локальних мультимереж. Зокрема, послідовний номер версії дозволяє виявляти скидання порядкового номера в кожному вузлі без використання контрольних пакетів, що з легкістю покращує багатоадресну надійність.

Надійний однорівневий сервіс зв'язку також представлений на декількох локальних аналітичних вузлах, де з'єднання встановлюються на кожному з них і кожен вузол вибирає і перемикає зв'язок самостійно. Припускається, що ПДВЗ може бути реалізована в ядрі UNIX на відмовостійких міні-комп'ютерах і в наш час може працювати в реальних онлайн-системах.

Література

1. D. R. Cheriton. The V Distributed System, Communication of ACM. 1988. – vol.31, №.3. – p. 314-333.
2. J. K. Ousterhout. The Sprite Network Operating System, IEEE COMPUTER, 1988. – vol.21, №.2. – p. 23 – 36.
3. D.R. Cheriton. VMTP: A Transport Protocol for the Next Generation of Communication Systems. Proc. of SIGCOM 86, Stowe, Vt. ACM, Нью-Йорк, 1986 рік.
4. E. Mafla., B. Bharat. Communication Facilities for Disaibuted Transaction Processing Systems. IEEE COMPUTER. – 1991. – vo1.24, No.8. – p. 61–66
5. S.J. Leffler. 4.3 BSD UNIX Operating System. Addison Wesley Publishing Company. – 1989. – p. 289.
6. T. Seki, Y. Okataku, **S.** Tamura, A Highly Reliable Broadcast Communication Mechanism on Intellectual Distributed Processing System. Trans. of the IEICE. – 1990. – vol.J73-D-I, №2. – p. 117-125.
7. D. E. Comer. Internetworking With TCP/IP Principles, Protocols, and Architecture. Prentice-Hall, Inc., 1988.