Слюсар Є. Високонадійна служба центрального каталога даних на базі технології IBM Mainframe в українській національній грід-інфраструктурі / Слюсар Є., Чередарчук А., Бойко Ю. // Вісник ТНТУ. — 2012. — Том 65. — № 1. — С.137-143. — (приладобудування та іформаційно-вимірювальні технології).

УДК 004.75

Є. Слюсар; А. Чередарчук; Ю. Бойко, канд. фіз.-мат. наук

Київський національний університет імені Тараса Шевченка

ВИСОКОНАДІЙНА СЛУЖБА ЦЕНТРАЛЬНОГО КАТАЛОГА ДАНИХ НА БАЗІ ТЕХНОЛОГІЇ IBM MAINFRAME В УКРАЇНСЬКІЙ НАЦІОНАЛЬНІЙ ГРІД-ІНФРАСТРУКТУРІ

Резюме. *Проведено аналіз особливостей застосування служб рівня кооперації української національної грід-інфраструктури, зокрема служби центрального каталогу даних LFC. Сформовано вимоги до реалізації високодоступного центрального каталога даних із використанням високонадійного обчислювального комплексу IBM Mainframe Z800. Показано особливості апаратної архітектури z/Architecture та програмного забезпечення проміжного рівня Nordugrid ARC, що були враховані при розробленні методики формування високодоступної служби LFC. Методику впроваджено для побудови системи з використанням мейнфрейму та обчислювального кластеру Київського національного університету імені Тараса Шевченка. Техніку формування опису грід-завдання для використання конфігурації високонадійної служби запроваджено до реалізації грід-порталу віртуальної лабораторії MolDynGrid в українській національній грід-інфраструктурі.*

Ключові слова*: грід, мейнфрейм, каталог даних, висока доступність, z-архітектура.*

I. Sliusar, A. Cheredarchuk, Yu. Boyko

BUILDING HIGH AVAILABILITY CENTRAL DATA CATALOG SERVICE POWERED BY IBM MAINFRAME TECHNOLOGY IN THE UKRAINIAN NATIONAL GRID-INFRASTRUCTURE

The summary. *Analysis of cooperation level service applications in Ukrainian national grid-infrastructure is conducted focusing on LFC central data catalog services. Methods for building high available central data catalog service employing IBM Z800 Mainframe installation are presented. Specialities of zSeries hardware architecture and Nordugrid ARC middleware software were taken into account for service implementation on top of the mainframe and HPC cluster of Taras Shevchenko National University of Kyiv. Grid job description generation technique for accessing high available LFC service was applied to the implementation of MolDynGrid Virtual Laboratory grid-portal in the Ukrainian national grid-infrastructure.*

Key words: *grid, mainframe, data catalog, high availability, z/Architecture.*

Вступ. Ідея створення грід-систем була запропонована на початку 1990-х років [1]. На той час інфраструктура всесвітньої мережі Інтернет була вже широко розвиненою, крім цього з’явилася велика кількість потужних обчислювальних ресурсів на базі загальнодоступних компонентів. До таких ресурсів насамперед належать обчислювальні кластери та бази даних. Основна ідея полягала в тому, щоб зробити ресурси доступними великій кількості користувачів, а також запобігти простою і перевантаженню наявних ресурсів.

Грід-інфраструктура України була створена впродовж останніх п’яти років. До неї входять обчислювальні ресурси більш ніж двадцяти наукових та освітніх установ України. Центральні грід-служби рівня кооперації, які є критичними для роботи української грід-інфраструктури, працюють у Київському національному університеті імені Тараса Шевченка та Інституті теоретичної фізики ім. М.М. Боголюбова НАН України, оскільки це були перші учасники національного грід-сегменту [2]. Деякі з цих служб, зокрема служба обліку членів віртуальних організацій (Virtual Organization Membership service, VOMS) та центральний каталог даних запущено лише в одному примірнику. Вихід з ладу будь-якої з критичних служб призводить до непрацездатності

всієї грід-інфраструктури. Зокрема, недоступність центральних каталогів даних призводить до неможливості направлення розрахункових завдань, що використовують розподілені сховища для вхідних та вихідних даних, до грід-інфраструктури.

Необхідність інтеграції української грід-інфраструктури у такі європейські структури, як EGI та WLCG накладає підвищені вимоги [3] щодо стабільності роботи критичних грід-служб, які на сьогодні не задоволені повністю.

Мета роботи. У 2011 році в Інформаційно-обчислювальному центрі Київського національного університету імені Тараса Шевченка інстальовано високонадійний обчислювальний комплекс на базі мейнфрейму zSeries Z800 виробництва IBM [4]. Метою роботи є розроблення та впровадження методик забезпечення високої доступності служби каталога даних в українській грід-інфраструктурі із залученням потужностей цього комплексу.

Аналіз останніх досліджень. Інфраструктура керування даними у складі попередніх європейських грід-проектів зазнала значної еволюції. У проекті European Data Grid (EDG, 2002–2004) у якості каталога даних використовувалося програмне забезпечення Replica Location Service (EDG-RLS), реалізоване за допомогою технології Java EE. Каталог даних мав ієрархічну структуру, в якій примірники нижчого рівня публікували свій вміст у каталогу вищого рівня. Засоби інтерфейсу користувача були також реалізовані засобами Java та потребували значних ресурсів при масовому застосуванні, а програмний інтерфейс служби не дозволяв проводити пошук за метаданими. У проекті Enabling Grids for E-sciencE (EGEE, 2004–2010) було розроблено нову реалізацію каталога даних – File and Replica Manager (FiReMan) [5]. Використання протоколу SOAP для доступу до служби та реалізація інтерфейсу користувача мовою C дозволила підвищити ефективність застосувань. Проте реалізація сервера, як і у попередньому проекті, була побудована на основі Java EE та мала обмежену масштабованість. У службі FiReMan вперше було інтегровано підтримку політик доступу на підставі участі користувача у віртуальних організаціях, що засвідчується окремою службою VOMS [6].

У проектах EGI та WLCG застосовується служба LCG File Catalog (LFC) [7]. На відміну від попередніх реалізацій каталога даних EDG-RLS та FiReMan, служба LFC реалізована мовою C, потребує значно менше ресурсів, аніж Java-реалізації, та позбавлена необхідності створення ієрархії примірників служб каталогів. Реалізація LFC використовує єдину зовнішню базу даних для зберігання структури простору імен файлів та місцезнаходження їх ідентичних копій – реплік поміж сховищ грід-інфраструктури.

Постановка задачі. Комплексне вирішення задачі інтеграції примірника служби каталога даних LFC до складу обчислювального комплексу Z800 для створення високонадійного центрального каталога даних для віртуальних організацій української національної грід-інфраструктури передбачає розроблення та реалізацію таких методик:

- створення логічного розділу виконання у процесорному блоці мейнфрейму для виконання окремого примірника операційної системи GNU/Linux;
- перенесення відкритого програмного забезпечення LFC та його залежностей на архітектуру IBM System/390 z/Architecture, на основі якої побудовано мейнфрейм;
- запуску системи керування базами даних у створеному примірнику GNU/Linux для використання програмним забезпеченням LFC;

- запровадження синхронізації вмісту бази даних LFC із базою даних іншого примірника служби LFC, запущеного у складі обчислювального кластера університету;
- забезпечення вчасного архівування вмісту бази даних на енергонезалежні носії та швидкого відновлення із архіву.

Висока доступність LFC. Служба центрального каталога даних необхідна для пошуку місцезнаходження файлів у грід-інфраструктурі. Фактично це – база даних інформації, де вказано місцезнаходження реплік різних файлів та додаткові метадані, автентифікація та авторизація доступу до якої здійснюється за допомогою стандартних механізмів безпеки грід-інфраструктури [8].

При направленні розрахункового завдання на виконання до грід-інфраструктури, користувач або грід-портал віртуальної організації формує опис цього завдання, у якому вказуються вимоги до обчислювальних ресурсів, а також інші параметри, зокрема вхідні та вихідні файли цього завдання. Для забезпечення високої доступності вхідних даних замість посилань на конкретний елемент зберігання даних вказується посилання на відповідну іменовану комірку бази даних у каталозі файлів LFC. Тоді грід-шлюз обчислювального елемента, на який потрапить таке розрахункове завдання, на етапі підготовки вхідних даних звернеться до центрального каталога, використовуючи проксі-сертифікат агента, що направив завдання, та отримає список реплік відповідного файлу. Використовуючи власні алгоритми вибору найближчого елемента зберігання даних із тих, що були вказані як репліки відповідного файлу, обчислювальний елемент здійснить завантаження вхідних даних із обраного сховища.

У реалізації служби обчислювального елемента A-REX, що входить до пакета програмного забезпечення проміжного рівня Nordugrid ARC [9], запроваджено два важливих механізми роботи із вхідними та вихідними даними, що забезпечують надійність доставки вхідних та вихідних даних завдання:

- перемикання на наступну репліку у разі недоступності попередньо обраної репліки при отриманні вхідних файлів;
- відкладена реєстрація файлу у каталозі даних після завантаження вихідних файлів на елемент зберігання даних.

Проте, у випадку недоступності центрального каталога даних LFC, обчислювальний елемент не зможе отримати список реплік вхідних файлів і завдання буде відхилено. Тому без запровадження високодоступної конфігурації центрального каталога, він стає єдиною точкою збою, що відображається на цілісності грід-інфраструктури. Транспортний протокол LFC був розроблений з урахуванням можливості балансування навантаження за допомогою ресурсного запису Round-Robin у системі DNS [10]. Таким чином, для забезпечення високої доступності необхідно відобразити у ресурсному записі DNS IP-адреси усіх синхронізованих примірників центрального каталога LFC.

Відкладена реєстрація файлу у каталозі дозволяє обчислювальному елементу A-REX завантажити вихідні дані грід-завдання на вказане користувачем сховище і виконати кілька спроб реєстрації посилання на завантажений файл у центральному каталозі LFC. Таким чином, якщо центральний каталог не був доступний на запис на момент завершення завантаження вихідних даних, реєстрація у ньому буде відкладена та буде здійснено кілька спроб через вказані у конфігурації проміжки часу.

Отже, критичним з огляду можливості виконання завдань у грід-інфраструктурі є забезпечення високої доступності центрального каталога даних для операцій читання – отримання списку реплік за вказаним логічним ім’ям файлу. Ця обставина накладає свій відбиток на структуру високонадійної служби LFC та реалізацію методики

синхронізації її примірників. Конфігурація, до якої входить один примірник із можливістю доступу як для читання, так і для запису, та один примірник із доступом лише для читання вже забезпечить необхідний рівень доступності для успішного виконання грід-завдань. Таким чином, для синхронізації баз даних вказаних примірників можна застосовувати схему реплікації Master-Slave, яку реалізовано у більшості систем керування базами даних.

Розгортання служби LFC на апаратній платформі System z. Основною перевагою платформи IBM System z є великий час напрацювання на відмову. Час життя обчислювального модуля оцінюється в 20–25 років, а вся система характеризується часом роботи 10–15 років. Підвищена стійкість роботи у випадку проблем з окремими компонентами реалізована через паралельне використання обчислювальних блоків – Z-процесорів та повне дублювання каналів доступу до периферії – мережевих інтерфейсів та сховищ даних. Доступна гаряча заміна всіх компонентів мейнфрейму. Забезпечення цілісності даних реалізовано через використання пам'яті з корекцією помилок та використання додаткових бітів корекції помилок у жорстких дисках системи зберігання даних. Технологія апаратної віртуалізації LPAR [11] дозволяє одночасно використовувати кілька різнотипних операційних систем на одному обладнанні. До складу платформи входить система розподілу та динамічного керування ресурсами для спільного використання апаратних ресурсів примірниками операційних систем, що виконуються у різних розділах LPAR.

Обчислювальний модуль доцільно використовувати для розміщення сервісів, не пов'язаних із виконанням обчислювальних задач. Різниця архітектур не дозволить оптимально збалансувати навантаження між вузлами класичної архітектури Intel x86-64 та z-платформи. З іншого боку, використання програмно-апаратного гіпервізора z/VM дозволить використовувати віртуальні машини з операційною системою GNU/Linux за принципом «один сервіс – один сервер», що забезпечує стабільну роботу та максимальну безпеку всієї системи. Зв'язок віртуальних машин із обчислювальними вузлами та іншими сервісними вузлами здійснюється через чотири вбудовані інтерфейси Gigabit Ethernet.

Для спрощення розгортання грід-служб на мейнфреймі було прийнято рішення сформувати універсальний образ інсталяції операційної системи GNU/Linux, який клонується та відповідна копія вже налаштовується під конкретну службу й слугує основним диском для завантаження відповідного розділу LPAR. За допомогою адміністративних політик у розділі було виділено 1Гб оперативної пам'яті та до двох ядер процесора із динамічним під'єднанням залежно від навантаження. Пристрій зберігання даних, з якого відбувається завантаження LPAR, відображається як пристрій прямого доступу (Direct Access Storage Device, DASD), що може мати певний фіксований розмір, який визначається архітектурою мейнфрейму. Для даної віртуальної машини було обрано диск розміром 50 гігабайт. Для завантаження ядра Linux використовується спеціальна програма zIPL (System z Initial Program Load), що розміщується на перших доріжках диска аналогічно до boot-сектора в архітектурі x86, та здатна зчитати у оперативну пам'ять ядро Linux та передати йому управління. Основні системні утиліти GNU та ядро Linux під z-архітектуру були скомпільовані самостійно із оптимізацією саме під систему Z800, в результаті чого була отримана мінімальна робоча інсталяція операційної системи з усіма необхідними засобами для подальшого розгортання відкритого мобільного програмного забезпечення, зокрема компіляторами GNU GCC. Отриманий таким чином образ DASD-диска було збережено для подальшого розгортання примірників операційної системи GNU/Linux.

Програмне забезпечення LFC підтримує системи керування базами даних MySQL, PostgreSQL та Oracle Database. У якості бази даних для високодоступної конфігурації центрального каталога файлів було обрано систему PostgreSQL, яка є стабільною у роботі та розповсюджується у вигляді вихідних кодів мовою C і без модифікацій успішно компілюється як під z-архітектуру, так і під класичну архітектуру x86-64. Для доступу до даних використовується архітектурно-незалежний протокол низького рівня, що дозволяє створювати високодоступні кластери із вузлів різних архітектур. Система також підтримує схему реплікації Master-Slave без інсталяції додаткових програмних засобів.

Вихідні коди пакета LFC є відкритими, проте служба розроблялася та тестувалася лише на архітектурах персональних комп'ютерів x86 та x86-64. У процесі перенесення цього програмного забезпечення на z-платформу особливу увагу було приділено використанню типів даних, що мають представлення та розмір, сумісний із архітектурою x86. Таким чином, у результуючій адаптованій версії програмного забезпечення LFC під z-архітектуру використовується такий самий протокол обміну, що й у версії під класичну архітектуру. Це дозволяє звертатися до служби за допомогою клієнтських утиліт та програмних засобів, що входять до стандартної поставки пакетів програмного забезпечення проміжного рівня.

Синхронізація та архівування бази даних. При розгортанні високодоступної служби роль головного вузла, що обслуговує операції як читання, так і запису, з міркувань більшої обчислювальної потужності була відведена вузлу у складі обчислювального кластера університету. Таким чином, з точки зору конфігурації системи керування базами даних, цей вузол є керуючим. У розділі LPAR на мейнфреймі система керування базами даних, а також власне служба каталога файлів були налаштовані для обслуговування запитів у режимі тільки для читання. До сценаріїв запуску та керування сервісами на вузлі у складі мейнфрейму було інтегровано засоби, що періодично атомарним чином створюють відбиток умісту бази даних на локальному диску. Для енергонезалежного зберігання архіву відбитків була відповідним чином налаштована система зберігання даних мейнфрейму TotalStorage. У відповідному розділі LPAR процесорного блока мейнфрейму було сконфігуровано доступ через інтерфейс ESCON до стрічкового робота, що входить до складу комплексу. Відбитки бази даних служби LFC записуються на відповідну стрічкову касету за допомогою стандартних утиліт керування стрічковими накопичувачами. Окремий сценарій забезпечує ротацію архівів на стрічці.

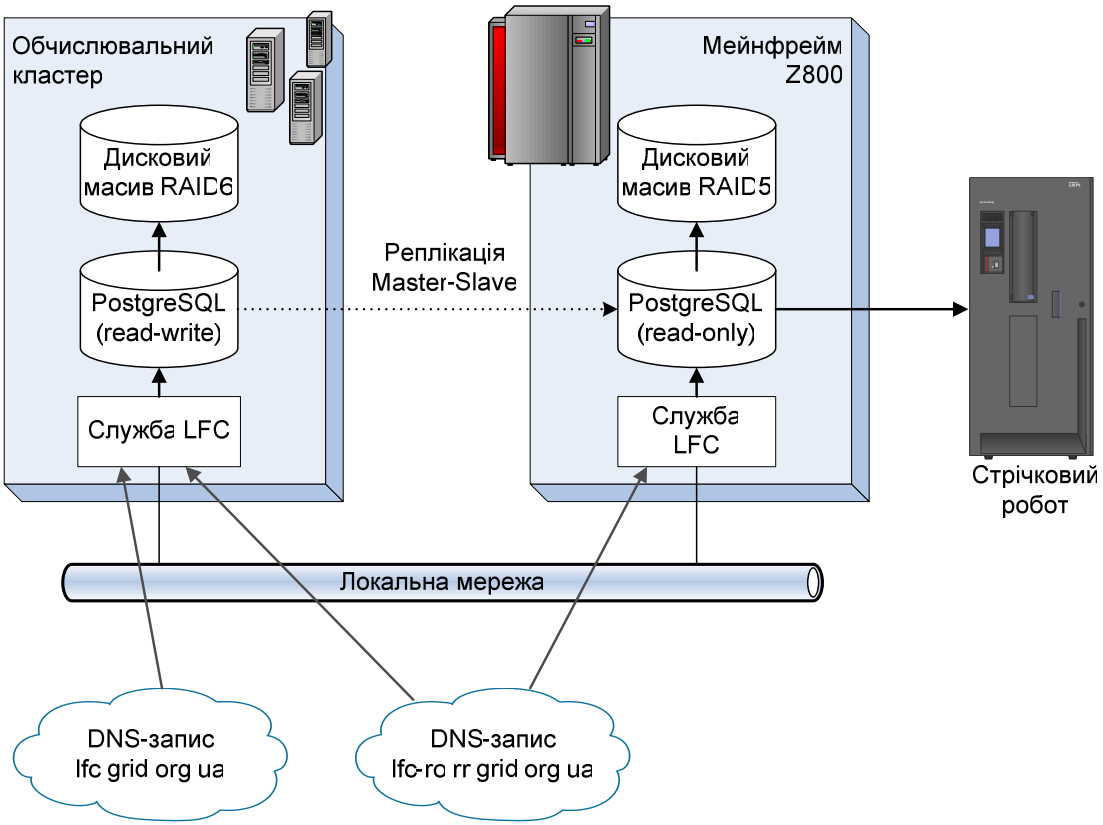

Рисунок 1. Схема високонадійного центрального каталогу даних

Реалізація. Схему високодоступної конфігурації центрального каталога даних української національної грід-інфраструктури зображено на рисунку 1. У системі DNS було створено два ресурсних записи – один для доступу до каталогу як для читання, так і для запису, що посилається на LFC-сервер обчислювального кластера, а другий – в окремій підзоні для Round Robin записів, що посилається на обидва примірники служби LFC для забезпечення високої доступності та балансування навантаження.

Для коректної роботи такої конфігурації, необхідно, щоб в описі грід-завдань у якості адреси вхідних даних вказувалося посилання на логічне ім'я файлу із використанням Round Robin DNS-запису, а для вихідних даних використовувалось посилання на примірник служби LFC обчислювального кластера. Така техніка була реалізована у генераторі опису завдань у складі грід-порталу віртуальної лабораторії MolDynGrid [12].

Висновки. Проведено аналіз особливостей застосування служб рівня кооперації української національної грід-інфраструктури, зокрема служби центрального каталога даних на основі програмного забезпечення LFC. Сформовано вимоги до реалізації високодоступного центрального каталога даних із використанням високонадійного обчислювального комплексу IBM Mainframe. Розроблено методику формування високодоступної конфігурації враховує як особливості апаратної архітектури z/Architecture, так і програмного забезпечення проміжного рівня Nordugrid ARC. Показано, що служба обчислювального елемента A-REX повинна використовувати різні точки входу до служби центрального каталога файлів при роботі з вхідними та вихідними файлами.

Запропоновану методику впроваджено для побудови високонадійного центрального каталога даних LFC на базі мейнфрейму та обчислювального кластеру Київського національного університету імені Тараса Шевченка. Техніку формування

опису грід-завдання, що використовує побудовану конфігурацію запроваджено до реалізації грід-порталу віртуальної лабораторії MolDynGrid, що працює з великими об'ємами даних в українській національній грід-інфраструктурі.

Розроблену методику можна застосовувати для розгортання високодоступних конфігурацій інших центральних служб рівня кооперації грід-інфраструктури та їх адаптації до роботи у складі мейнфрейму. Запропонована техніка формування опису грід-завдання може бути інтегрована до порталів інших віртуальних організацій Українського національного гріду.

Список використаної літератури

1. Foster, I. The Grid, Blueprint for a New computing Infrastructure / I. Foster, C. Kesselman. – Morgan Kaufmann Publishers, Inc., 1998.
2. Український академічний Грід: досвід створення й перші результати експлуатації [Текст] / Ю.В. Бойко, М.Г. Зинов'єв, О.О. Судаков, С.Я. Свістунов // Математичні машини і системи. – 2008. – Випуск 1. – С. 67–84.
3. Ferrari, T. EGI Resource Centre Operational Level Agreement – [Online] v1.1, 12 Jan 2012. https://documents.egi.eu/document/31.
4. z/Architecture and ESA/390 Principles of Operation and Reference Summary // [Online] A2278325 Document 05 SA22-7832-05. http://publibz.boulder.ibm.com/cgi-bin/bookmgr_OS390/XKS/DZ9ZBK07.
5. Kunszt, P. The gLite FiReMan Catalog / Peter Kunszt // LCG Storage Workshop 2005 April 05-07.
6. Stewart, G. LCG Data Management: From EDG to EGEE / Graeme Stewart et al. // In UK eScience All Hands Meeting Proceedings, Nottingham, UK 2005.
7. Calanducci, T. LFC: The LCG File Catalog / Toni Calanducci // gLite Bratislava, 27–30.06.2005.
8. Welch, V. Security for Grid services / Welch, V.; Siebenlist, F.; Foster, I. et al. // in Proc. 12th IEEE International Symposium on High Performance Distributed Computing. – 22–24 June 2003, pages 48–57.
9. Ellert, M. Advanced Resource Connector middleware for lightweight computational Grids / M. Ellert et al. // Future Gener. Comput. Syst. – 2007. – Vol. 23, no. 1. – PP. 219–240.
10. Brisco, T. DNS Support for Load Balancing / T. Brisco // RFC 1794 [Online], April 1995.
11. Singh, K. Security on the Mainframe / Karan Singh // IBM Redbooks [Online] – December 2009 – Chapter 4. Virtualization – pages 24–83. http://www.redbooks.ibm.com/redpieces/pdfs/sg247803.pdf.
12. Salnikov Andrii. Virtual Laboratory MolDynGrid as a Part of Scientific Infrastructure for Biomolecular Simulations / A.O. Salnikov, I.A. Sliusar, O.O. Sudakov et al. // International Journal of "Computing". – 2010. – Vol. 9, no. 4.

Отримано 06.02.2012