ПРОГРАМНО-ТЕХНІЧНІ КОМПЛЕКСИ

А.Л. ГОЛОВИНСЬКИЙ, І.В. СЕРГІЄНКО, В.Г. ТУЛЬЧИНСЬКИЙ, А.Л. МАЛЕНКО, О.Ю. БАНДУРА, С.О. ГОРЕНКО, О.Ю. РОГАНОВА, О.І. ЛАВРІКОВА

УДК 004.386

РОЗВИТОК СУПЕРКОМП'ЮТЕРІВ СЕРІЇ СКІТ, РОЗРОБЛЕНИХ В ІНСТИТУТІ КІБЕРНЕТИКИ ІМ. В.М. ГЛУШКОВА НАН УКРАЇНИ У ПЕРІОД З 2002 ПО 2017 РОКИ

Анотація. У 2017 р. виповнилося п'ятнадцять років проекту створення вітчизняної обчислювальної системи з кластерною архітектурою СКІТ. У роботі вперше наведено дані про еволюцію архітектури комплексу СКІТ та статистику функціювання суперкомп'ютерного центру в період з 2002 по 2017 рр. Ці фактичні дані будуть корисні розробникам обчислювальних кластерів та дослідникам, які проектують нові алгоритми керування ними.

Ключові слова: суперкомп'ютер, кластер, паралельне програмування.

ВСТУП

Обчислювальні кластери серії СКІТ (суперкомп'ютер для інтелектуальних технологій) є несерійними дослідними системами, які будувалися на найкращій технологічній основі свого часу та втілювали оригінальні вітчизняні розробки в напрямку системного програмного забезпечення та бібліотек обчислювальної математики.

У статті наведено ретроспективний аналіз розвитку архітектури СКІТ та впорядковано статистичні дані користування кластерами.

Статистичні дані роботи користувачів аналізують у багатьох великих суперкомп'ютерних центрах. Так, у роботі [1] наведено такий аналіз для суперкомп'ютера Кракен Оакріджської лабораторії (США). Попри різницю у кількості ядер і географічне розташування ця система та СКІТ мають схожі структури обчислювальних задач за галузями науки та подібні параметри їхнього розподілу.

ЕВОЛЮЦІЯ АРХІТЕКТУРИ ПРОЕКТУ СКІТ

Розвиток архітектури кластерів серії СКІТ складається з п'яти основних етапів, показаних в табл. 1.

Перший дослідний зразок обчислювального кластера СКІТ-0 було створено у період з 2002 по 2004 рр. Основні ідеї проекту описано в [2, 3]. Кластер мав чотири вузли на AMD Duron 800, інтерконект 100 Мб Fast Ethernet, сховище даних обсягом 160 Гб на основі масиву RAID-1.

Оригінальною особливістю цього кластера, яка відрізняла його від класичної архітектури Beowulf, був бездисковий принцип, коли всі вузли є клонами, завантаженими з одного образу операційної системи. Це дозволило полегшити адміністрування та швидко і синхронно вносити зміни у системне та прикладне програмне забезпечення. На цьому кластері було відпрацьовано архітектурне рішення — спільне дискове сховище для домашніх тек користувачів, що одночас-

но монтувалось на всі вузли кластера. Ці рішення у наступних проектах майже не змінювалися. Кластер СКІТ-0 не мав спеціалізованої системи керування та системи черг, MPI-задачі запускалися з візуальним контролем зайнятості вузлів. На цьому кластері було розв'язано низку важливих задач моделювання зміни клімату, фільтрації ґрунтів та розрахунку міцності конструкцій.

Таблиця 1. Продуктивність кластерів серії СКІТ

Рік	Назва	Кількість процесорних ядер	Продуктивність у Linpack, Rpeak, Тфлопс
2002	СКІТ-0	4	0.02
2004	СКІТ-1	48	0.19
2005	СКІТ-2	64	0.28
2007	СКІТ-3	725	5.3
2012	СКІТ-4	448	18

У 2004–2006 рр. майже одночасно було створено два принципово різних кластери: СКІТ-1 та СКІТ-2, основні ідеї їхнього розроблення описано в [3, 4]. У 2006 р. кластери було запущено в експлуатацію.

Кластер СКІТ-1 (посідає 13-у позицію у першій редакції рейтингу «TOP 50 суперкомп'ютерів СНД» [5] від 07.12.2004 р.) мав 24 вузли на процесорах Intel Xeon з 32-бітною архітектурою i386, інтерконект Infiniband SDR.

Кластер СКІТ-2 (посідає 8-у позицію у другій редакції рейтингу «TOP 50 суперкомп'ютерів СНД» від 05.04.2005 р.) мав передову технологію IPMI-1.1, яка дозволяла керувати живленням вузлів та контролювати стан обладнання. Ця система стала базовою на платформах Intel ia64. Кластер СКІТ-2 мав 32 вузли з інноваційними процесорами Intel Itanium-2 з 64-бітною архітектурою ia64. Встановлений інтерконект Dophinics SCI забезпечував нижчу латентність і вищу швидкодію у порівнянні з Infiniband, але меншу надійність.

Кластери СКІТ-1 та СКІТ-2 були обладнані власними керувальними серверами і спільною системою збереження даних обсягом 480 Гб на основі контролерів MegaRAID типу RAID6. Для керування ресурсами кластерів було розроблено оригінальну систему керування MVS, яка формувала чергу обчислювальних задач, надавала зручний діалоговий інтерфейс користувача та засоби задання пріоритетів задач, забезпечувала денний і нічний режими роботи. У денному режимі кластер ставив на обрахунок тільки короткі задачі, а великі багатодобові задачі обчислювалися вночі. Таким чином звільнялись ресурси для швидкого оброблення коротких задач у робочий час.

У 2007, 2008 рр. було створено кластер СКІТ-3 (посідає 5-у позицію у шостій редакції рейтингу «TOP 50 суперкомп'ютерів СНД» від 11.04.2007 р.), концепцію якого описано в [6]. Кластер мав 125 вузлів та 725 процесорних ядер. В основу кластера лягли новітні на той час 64-бітні процесори: 2-ядерний Intel Xeon 5160 3.0 ГГц та 4-ядерний Xeon E5345 2.33 ГГц з архітектурою x64. Як мережа передавання даних використовувся Infiniband DDR зі швидкодією 20 Гб/с.

Важливим кроком у побудові системи збереження даних був перехід на паралельну файлову систему Lustre. Ця файлова система мала у складі набір вузлів збереження, які підтримували одночасну паралельну роботу з вузлами кластера, що дозволило досягнути швидкодії дискової системи понад 1 Гб/с у паралельних обчислювальних задачах, а її обсяг збільшити до 33 Тб. Так, у період з 2002 по 2005 рр. обсяг сховища на основі файлової системи NFS становив 4 Тб, у 2006 р. обсяг сховища на основі паралельної файлової системи Lustre дорівнював 160 Гб, у 2007 р. — 600 Гб, у 2008 р. — 5.3 Тб, у 2009 р. — 17 Тб, у 2010, 2011рр. — 33 Тб, у 2012–2014 рр. — 80 Тб, у 2015 р. — 120 Тб, у 2016 р. — 170 Тб.

Як систему керування обчислювальними ресурсами було обрано менеджер ресурсів SLURM, який надавав гнучкість у контролі за потоком задач та обчислювальними вузлами і забезпечував просту конфігурацію.

У 2013 р. було створено кластер СКІТ-4 [7, 8], який мав 28 вузлів та 448 процесорних ядер. Важливою модифікацією архітектури стало використання гібридних обчислювальних вузлів, у яких окрім процесорів Intel Xeon E5-2600 2.6 ГГц з архітектурою x64 використовувалися графічні прискорювачі NVidia Tesla M2075.

Графічні прискорювачі дозволяють частину обчислювального алгоритму виконувати на власних спеціалізованих процесорах, значна кількість яких розташована на одній платі (448 для M2075, 1344 на одному вузлі). У реальній роботі прискорювачі показали себе неоднозначно. З одного боку, в спеціально перероблених для цього алгоритмах досягалася висока продуктивність, а з іншого — більшість програмних пакетів, які застосовували користувачі СКІТ, не мали GPU версій, а ті, що підтримували GPU, мали реалізацію лише деякої підмножини алгоритмів пакету.

Передбачається, що кластер СКІТ-5 стане наступною системою в серії суперкомп'ютерів, його архітектура зараз перебуває у стані розроблення і дослідження різних перспективних рішень: перехід на SSD-накопичувачі у системі збереження даних, впровадження постійних запам'ятовувальних пристроїв, базованих на технології NVMe, дослідження прискорювачів Intel XeonPhi, перехід від RAID масивів до файлових систем нового покоління, таких як ZFS, тощо.

СТАТИСТИКА НАДІЙНОСТІ ОБЛАДНАННЯ СКІТ

Апаратну архітектуру суперкомп'ютерів зазвичай добре висвітлено в анонсах чергових модернізацій. Однак майже не публікують даних щодо надійності обладнання протягом тривалого часу. Насамперед це пов'язано з високим темпом модернізації американських та європейських обчислювальних систем.

За показниками журналу ремонтів СКІТ, за десять років накопичилися статистичні дані щодо обладнання. Звичайно, воно застаріло і не може бути орієнтиром при створенні нового кластера, але дає деяке уявлення про роботу кластера впродовж певного періоду.

Проаналізуємо вузли кластерів СКІТ, побудованих на таких типах платформ:

- Tyan S5370;
- Supermicro TWIN X8DTT-F;
- Supermicro GPU Nvidia Tesla M2050;
- HP Proliant SL250 Gen 8 з GPU Nvidia Tesla M2075;
- HP Proliant SL230 Gen 8.

За спостереженнями користувачів, надійність платформ залежить від класу виробника. Найгіршими виявилися платформи Tyan, на яких одразу виникли проблеми і за дев'ять років вийшло з ладу 10 %. До платформи Supermicro суттєвих зауважень немає — втрати за вісім років склали 5 %. Найкраще себе зарекомендував виробник класу A Hewlett Pakkard, у якого за чотири роки не було втрат.

В обладнанні вузлів основні проблеми виникають з платами IPMI, материнськими платами та пам'яттю. Найбільш надійними виявились процесори Intel Xeon різних серій.

У 2012, 2013 рр. проводилися дослідження можливості використання десктопних компонентів у вузлах та серверах, оскільки вони набагато дешевші за серверні. Було створено кілька вузлів з материнськими платами на Intel Core i5, прискорювачами Radeon 9660 та блоками живлення на 700 Вт. Їхня робота показала, що така концепція невиправдана, оскільки більшість обладнання не витримала роботи в режимі $24/7*365$ і вийшла з ладу впродовж року.

Жорсткі диски у сховищі даних на базі ОС Lustre здебільшого мають обмежений термін експлуатації. Використовувалися серії:

- Hitachi Deskstar 500 Гб;
- WD 2 Тб;
- Seagate 3Тб;
- Hitachi Deskstar NAS 6Тб.

Диски Hitachi 500 виявилися найкращими, за вісім років експлуатації вийшло з ладу лише приблизно 5 %, інші морально застаріли і були виведені з основної конфігурації, але продовжують працювати на менш відповідальних ділянках.

Диски WD обиралися з енергоощадливої десктопної лінійки, Seagate — з недорогої десктопної серії, а термін їхньої експлуатації склав три роки, як зазначено в гарантії, після чого 80 % вийшло з ладу.

Щодо використання Hitachi Hitach Deskstar 6 Тб маємо лише півроку спостережень, за цей час жодних проблем не виникало.

Виходячи з наведеного, необхідним є постійне підтримання резерву у два–п'ять дисків кожного типу, що використовують у системі збереження.

Надійність комутаторів мережі Ethernet та контролерів мережі Infiniband в цілому відповідає сучасним передовим технологіям. За 15 років вийшло з ладу два із 17 комутаторів і приблизно 3 % мережевих контролерів у вузлах.

Як Ethernet-обладнання використовувались керовані комутатори HP ProCurve та DLink різних серій. Комутатори високого класу HP ProCurve виявилися найкращими, а всі інші виходили з ладу чи мали дефекти в роботі після трьох–п'яти років експлуатації.

Використовувались APC Smart UPS різних серій. Ці блоки безперебійного живлення виявились в цілому надійними — за 15 років з 20 блоків вийшов з ладу лише один, якщо не враховувати батареї, які працювали в середньому лише п'ять років, після чого потребували заміни.

СТАТИСТИКА РОБОТИ КОМПЛЕКСУ СКІТ

Для забезпечення ефективного функціювання кластерного комплексу СКІТ та його пристосування до потреб користувачів надзвичайно важливо збирати та аналізувати статистичні дані щодо роботи кластера. Завдяки цій інформації можна визначити структуру обчислювальних задач, встановити, якому прикладному програмному забезпеченню необхідно найбільше приділяти увагу, коригувати різноманітні параметри черги задач, створювати нові глобальні профілі запуску задач у порталі кластерних обчислень. Також актуальним є контроль за використанням кластерних ресурсів окремими користувачами або організаціями.

У 2006 р. на кластерах СКІТ було встановлено систему накопичення статистики роботи користувачів.

При обробленні даних від менеджера ресурсів кластер може отримати такі статистичні показники за певний період:

- кількість задач, які були запущені на виконання, — показник активності застосування кластера користувачами;
- структура задач за статусом завершення як в абсолютних, так і відносних величинах дозволяє визначити, яка частка задач серед поставлених у чергу виконана успішно;
- середня тривалість однієї задачі визначає спосіб використання кластера — екстенсивний чи інтенсивний;
- сумарне завантаження кластера в процесоро-хвилинах визначається як сума часу виконання кожної задачі, помноженої на кількість зайнятих нею процесорів.

Окремо необхідно розраховувати статистичні дані як щодо кластера в цілому, так і відносно користувачів та організацій, при цьому додатково обчислювати такі параметри:

- частка задач окремого користувача чи організації у загальній кількості задач за період дозволяє визначити його активність стосовно запуску задач у порівнянні з іншими користувачами;
- частка завантаження кластера користувачем у його сумарному завантаженні дозволяє відстежити основних споживачів кластерних ресурсів СКІТ.

Загалом за період 2002–2016 рр. кластери СКІТ надали 11.8 млн процесоро-годин обчислювальних ресурсів, виконавши більше 70 тис. задач користувачів. Зведені статистичні дані за роками наведено у табл. 2.

Найбільше обчислювальних ресурсів кластери надавали у докризові 2009, 2010 рр., надалі кількість процесоро-годин зменшувалась, що спричинене економічними факторами, насамперед обмеженим фінансуванням витрат на електроенергію.

У період з 2013 по 2016 рр. на кластерах Інституту кібернетики ім. В.М. Глушкова НАН України працювало 70 наукових груп користувачів з 25 установ НАНУ і МОН. Серед них виділимо такі: Інститут молекулярної

Таблиця 2. Характеристики роботи кластерів серії СКІТ у період з 2006 по 2016 рр.

Рік	Кількість задач	Процесоро-години, тис.	Середній час виконання, хв
2006	1050	2.3	34
2007	12017	207.8	27
2008	11043	107.2	142
2009	11475	2440.7	479
2010	17251	2870.1	310
2011	16334	2340.2	300
2012	12784	1845.3	320
2013	9678	877.8	373
2014	7107	1039.2	1049
2015	1548	47.7	425
2016	3450	1200.5	500
Всього	103737	12978.8	—

біології і генетики; Інститут геохімії, мінералогії та рудоутворення ім. М.П. Семененка; Інститут математики НАН України; Інститут проблем матеріалознавства ім. І.М. Францевича НАН України; Інститут геофізики ім. С.І. Субботіна НАН України; Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»; Ужгородський національний університет; Інститут органічної хімії НАН України; Чернівецький національний університет імені Юрія Федьковича; Фізико-технічний інститут низьких температур ім. Б.І. Вєркіна НАН України; Національний науковий центр «Харківський фізико-технічний інститут»; Запорізький національний університет.

Середня річна кількість активних користувачів складає 50 чоловік. Розглянемо зведені статистичні дані щодо типів обчислювальних задач:

- квантово-хімічні розрахунки (приблизно 100 тис. процесоро-годин на рік) — дослідження активних центрів біологічних молекул, пошук нових ліків;
- молекулярна динаміка (приблизно 100 тис. процесоро-годин на рік) — дослідження органічних молекул із складною тривимірною структурою (білків, ДНК тощо);
- алгебра і теорія чисел (приблизно 10 тис. процесоро-годин на рік) — чисельні експерименти для доведення теорем, підтвердження гіпотез;
- оброблення результатів експерименту ALICE на LHC (10 тис. процесоро-годин на рік);
- обчислювальні задачі гідродинаміки (25 тис. процесоро-годин на рік).

ВИСНОВКИ

За п'ятнадцять років у такій динамічній галузі, як обчислювальні системи змінилось багато трендів: з'явились і занепали грід-технології, з'явились і потужно заявили про себе хмарні системи. Однак суперкомп'ютери залишаються основним інструментом розв'язання складних науково-технічних задач.

Суперкомп'ютерний комплекс СКІТ є найбільшим обчислювальним центром України, орієнтованим на розв'язання широкого кола наукових задач. Доступ до нього здійснюється як через веб-інтерфейс на офіційному сайті [9], так і за традиційним SSH-протоколом. Спеціалісти Інституту кібернетики ім. В.М. Глушкова НАН України проводять дослідження в галузі архітектури високопродуктивних обчислювальних систем, розробляють засоби керування ними та діагностики.

СПИСОК ЛІТЕРАТУРИ

1. Haihang You, Hao Zhang. Comprehensive workload analysis and modeling of a petascale supercomputer. 16th International Workshop, JSSPP 2012, Shanghai, China, May 25, 2012. Revised Selected Papers. P. 253–271.
2. Коваль В.Н., Савьяк В.В., Сергиенко И.В. Тенденции развития современных высокопроизводительных систем. *УсиМ*. 2004 № 6. С. 31–43.
3. Коваль В.Н., Савьяк В.В. Мультипроцессорные кластерные системы: планирование и реализация. *Искусственный интеллект*. 2004. № 3. С. 3–43.
4. Сергієнко І., Коваль В. СКІТ — український суперкомп'ютерний проект. *Вісн. НАН України*. 2005. № 8. С. 3–13. URL: http://nbuv.gov.ua/UJRN/vnanu_2005_8_1.

5. TOP 50 суперкомп'ютерів СНД. URL: http://top50.supercomputers.ru/.
6. Головинский А.Л., Рябчун С.Г., Якуба А.А. Гетерогенный кластерный комплекс Института кибернетики НАН Украины: средства построения. *Искусственный интеллект.* 2006. № 16. С. 107–112.
7. Головинський А.Л., Маленко А.Л., Сергієнко І.В. Енергоефективний суперкомп'ютер СКІТ-4. *Вісн. НАН України.* 2013. № 2. С. 50–59.
8. Головинський А.Л., Маленко А.Л., Бандура О.Ю, Ємченко І.О. СКІТ-4 — суперкомп'ютер Інституту кібернетики ім. В.М. Глушкова НАН України. *High Performance Computing*. 2012. С. 149–151.
9. Сайт кластерного комплексу СКІТ. URL: http://icybcluster.org.ua/.

Надійшла до редакції 19.12.2016

А.Л. Головинский, И.В. Сергиенко, В.Г. Тульчинский, А.Л. Маленко, А.Ю. Бандура, С.А. Горенко, Е.Ю. Роганова, Е.И. Лаврикова

РАЗВИТИЕ СУПЕРКОМПЬЮТЕРОВ СЕРИИ СКИТ, СОЗДАННЫХ В ИНСТИТУТЕ КИБЕРНЕТИКИ ИМ. В.М. ГЛУШКОВА НАН УКРАИНЫ В ПЕРИОД С 2002 ПО 2017 ГОДЫ

Аннотация. В 2017 г. исполнилось пятнадцать лет проекту создания отечественной вычислительной системы с кластерной архитектурой СКИТ. В работе впервые рассмотрена эволюция архитектуры комплекса СКИТ и статистика функционирования суперкомпьютерного центра в период с 2002 по 2017 гг. Эти фактические данные будут полезны разработчикам вычислительных кластеров и исследователям, которые проектируют новые алгоритмы управления ими.

Ключевые слова: суперкомпьютер, кластер, параллельное программирование.

A. Golovynskyi, I. Sergienko, V. Tulchinkyi, A. Malenko, O. Bandura, S. Gorenko, O. Roganova, O. Lavrikova

DEVELOPMENT OF SCIT SUPERCOMPUTERS FAMILY AT THE INSTITUTE OF CYBERNETICS OF NAS OF UKRAINE IN 2002–2017

Abstract. 2017 marks the 15th anniversary of the supercomputer project SCIT, which allows us to summarize the results and draw conclusions. In this paper, we discuss evolution of SCIT architecture and supercomputing center statistics for years 2002–2017. These data will be useful to computer cluster developers and researchers who design resource management algorithms for computing clusters.

Keywords: super computer, cluster, parallel programming.

Головинський Андрій Леонідович,
кандидат техн. наук, старший науковий співробітник Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: icybcluster@gmail.com.

Сергієнко Іван Васильович,
академік НАН України, директор Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: incyb@incyb.kiev.ua.

Тульчинський Вадим Григорович,
доктор фіз.-мат. наук, завідувач лабораторії Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: dep145@gmail.com.

Маленко Андрій Леонідович,
кандидат фіз.-мат. наук, старший науковий співробітник Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: icybcluster@gmail.com.

Бандура Олександр Юрійович,
науковий співробітник Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: icybcluster@gmail.com.

Горенко Сергій Олександрович,
науковий співробітник Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: icybcluster@gmail.com.

Роганова Олена Юріївна,
інженер-програміст І категорії Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: olena.rogan@gmail.com.

Лаврікова Олена Іванівна,
науковий співробітник Інституту кібернетики ім. В.М. Глушкова НАН України, Київ, e-mail: icybcluster@gmail.com.