Система збереження даних кластера Інституту фізики конденсованих систем НАН України

О. Я. Фаренюк, Т. М. Пацаган

Інститут фізики конденсованих систем НАН України, вул. Свєнціцького, 1, Львів, Україна

indrekis@icmp.lviv.ua, tarpa@icmp.lviv.ua

Анотація. *Важливою частиною обчислювальних кластерів є система збереження даних. Для кластера Інституту фізики конденсованих система НАН України (ІФКС) в ролі основи системи збереження даних було вибрано розподілену кластерну файлову систему Lustre. Ця система забезпечує високу продуктивність, при тому є гнучкою та зручною в адмініструванні. Ще однією її перевагою є повна прозорість для користувацького програмного забезпечення. Досвід інтенсивної експлуатації протягом семи років підтвердив також надійність Lustre.*

Ключові слова

Системи збереження даних, файлова система, Lustre, кластер

1 Вступ

З кожним роком використання обчислювальних кластерів у фізиці, хімії, біології набуває все більшого поширення в академічних та освітніх закладах України. Завдяки стрімкому росту продуктивності обчислювальної техніки з'являється можливість розв'язувати все більш складні та об'ємні задачі, які, в свою чергу, продукують великі обсяги даних для подальшої їх обробки та аналізу. Тому, при розбудові кластера, збільшенні кількості його вузлів та задач на них, неодмінно виникає необхідність встановлення швидкої та надійної системи збереження даних, яку можна було б з часом легко розширювати. Одним із найуспішніших рішень у цьому напрямку є система на базі розподіленої файлової системи Lustre [1].

Інститут фізики конденсованих систем НАН України (ІФКС), починаючи з 2001 року, веде роботу по розбудові власного обчислювального кластера, на якому працівниками Інституту та ряду інших наукових установ Західного регіону виконуються актуальні фундаментальні та науково-практичні задачі. Більша частина цих задач – це комп'ютерне моделювання конденсованих систем, зокрема, рідких металів, водних розчинів електролітів, іонних рідин, колоїдних суспензій, рідкокристалічних систем, полімерів, поліелектролітів, макромолекул та пористих середовищ. Широкий спектр методів та підходів, які використовуються при цьому, включає в себе першопринципну молекулярну динаміку, квантовохімічні розрахунки, молекулярну динаміку, Монте-Карло, дисипативну динаміку, коміркові автомати тощо. У 2006 році постало питання необхідності системи збереження даних (СЗД). Тому, поряд із значною модернізацією обчислювальних ресурсів, яка тоді відбулася, на кластері ІФКС було встановлено також СЗД на базі дискових RAID-масивів під керуванням розподіленої файлової системи Lustre. Після кількох модернізацій, на сьогодні (початок 2013 року) розмір цієї системи становить 16ТБ, а загальна конфігурація кластера ІФКС має наступний вигляд (див. також рис. 1):

- 22 обчислювальних вузли, 188 ядер, 434Гб RAM, 1.6Тфлопс
- Координуюча машина, 4 ядра, 8Гб RAM
- 1 GPU nVidia Tesla M2050, 448 CUDA cores, 14 мультипроцесорів, Compute Capability 2.0, 3Гб RAM (2.62 Гб з ECC).
- Система збереження даних, 3 вузла, 16Тб доступного користувачам об'єму, RAID6/RAID1, Lustre.

- Infiniband 4x DDR, 8 Гбіт/с в обидва напрямки
- Під'єднання: академічна грід-мережа (1Гбіт/с), Уарнет; локальна мережа ІФКС (100Мбіт/с).
- Енергоспоживання – 10-15 кВт.

Кластер ІФКС є учасником Українського національного гріду (УНГ) [2, 3], і входить в п'ятірку найпотужніших кластерів України [4]. Доступ до кластера здійснюється локально (30 користувачів) та через УНГ в рамках віртуальної організації (ВО) Multiscale [5], яка на даний момент налічує 15 учасників [6]. ВО Multiscale була створена в кінці 2011 року, ставить за мету вирішення науково-прикладних задач в галузі фізики та хімії із використанням методів комп'ютерного моделювання на різних рівнях деталізації досліджуваних систем (від першопринципних розрахунків до мезоскопічного опису). ВО Multiscale надає широкий інструментарій для проведення комп'ютерного моделювання, зокрема для учасників цієї ВО доступні такі програмні пакети: Abinit, CP2k, CPMD, QuantumEspresso, GROMACS, LAMMPS, DL_POLY. Також існує можливість запуску власних програмних розробок. Крім ВО Multiscale, кластер ІФКС надає ресурси ще кільком ВО: medgrid, bitp і bitpedu. Таке навантаження на кластер ІФКС вимагає особливої уваги до обслуговування СЗД та його подальшого розширення.

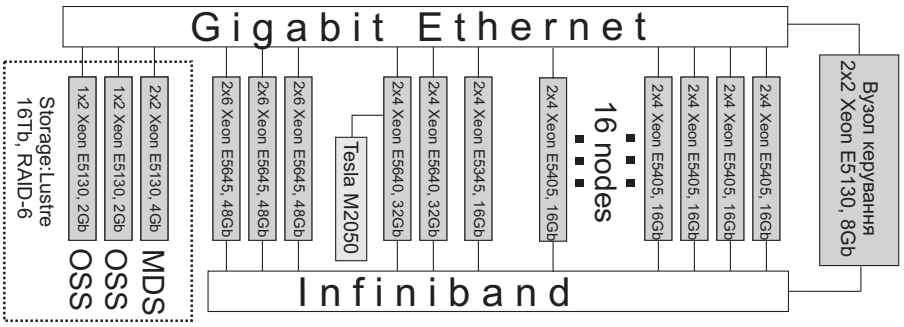

Рис. 1. Будова кластера ІФКС.

2 Розподілена файлова система Lustre

Процес наукових розрахунків методами комп'ютерного моделювання на обчислювальних кластерах зазвичай пов'язаний із потребою зберігати та читати великі об'єми даних. Щоб зробити цей процес ефективним, використовують виділені системи збереження даних (СЗД). Вибір способу організації СЗД – одне із найбільш критичних архітектурних рішень при побудові кластера. Під час вибору архітектури СЗД було сформульовано наступні вимоги:

- Зручність використання на кластері, по можливості — прозорість для користувачів.
  - Надійна робота з OS Linux.
  - Сумісність з POSIX-стандартом розподілених файлових систем.
  - Однаковий доступ з різних вузлів.
- Надійність.
- Висока продуктивність. СЗД повинна працювати за наступних умов:
  - Великі потоки даних.
  - Велика кількість мало пов'язаних (нелокальних) звертань.
  - Доступ одночасно з десятків вузлів.
- Масштабованість:
  - Безболісне збільшення кількості клієнтів.
  - Просте збільшення об'єму – можливість додавати жорсткі диски та вузли СЗД.
- Відкритість, бажано – доступність під Open Source ліцензією.

- Простота та зручність в обслуговуванні.
  - Підтримка квот.

Було проаналізовано багато можливих варіантів. З великою перевагою найбільш оптимальним вибором виявилася кластерна файлова система (ФС) Lustre[1]. Вона, по своїй суті, розподілена, розроблялася якраз для використання в кластерах та датацентрах, з відкритими джерельними текстами, (Open Source) та відтворює POSIX-семантику роботи з файлами, тому прозора для більшості програм. Крім того, вона має дуже хороші рекомендації — використовується в системах від маленьких, з кількома вузлами, до величезних, із десятками тисяч, пропускною здатністю в сотні терабайт на секунду і розміром в десятки пентабайт. Зокрема, станом на листопад 2012 року, 15 з 30 найпотужніших кластерів списку TOP500 використовують саме її [7]. Слід зауважити, що вимога відкритості вибраного рішення себе виправдала з часом. Розробник Lustre, Sun Microsystems, була придбана Oracle, після чого роботи над Lustre було припинено. Однак вона доволі швидко була підхоплена кількома організаціями, зокрема Whamcloud[2], куди перейшли ключові розробники Lustre із Sun [1]. Тому, як покращення, так і виправлення помилок було продовжено.

Структура СЗД на базі Lustre наступна:

- MGS – Management Server, сервер конфігурацій файлових систем, єдиний на кластер. До нього звертаються всі сервери всіх екземплярів ФС Lustre та клієнти. Вимагає для своїх потреб невеликий блочний пристрій – MGT, розміром кілька десятків мегабайт. Необхідний розмір не залежить від розміру файлових систем, що послуговуються MGS. Передбачена можливість суміщати MGS з MDS.
- MDS – Metadata Server, сервер метаданих, тобто імен файлів, директорій, прав доступу, тощо. Він єдиний для конкретної ФС[3]. Фізично метадані зберігаються у відповідному цільовому пристрої, MDT.
- MDT – Metadata Target, цільовий пристрій метаданих. Фактично, це спеціальна файлова система, модифікована ext3/ext4, монтування якої запускає відповідний сервер, а розмонтування – зупиняє.
- OSS – Object Storage Server(s), об'єктні сервери збереження даних, безпосередньо здійснюють ввід-вивід даних та їх передачу мережею клієнтам. Для збереження даних кожен із них використовує одну або більше OST.
- OST – Object Storage Target, об'єктні цільові пристрої збереження даних.
- Клієнти – будь-яка машина з Linux та відповідним клієнтським програмним забезпеченням. Ядро з модифікаціями, що мають відношення до Lustre, не є необхідним! Всі клієнти бачать одну і ту ж, синхронізовану і когерентну ФС, з тими ж правами доступу та іншими атрибутами. Єдина серйозна вимога — UserID та GroupID повинні на всіх вузлах-клієнтах бути тими самими, а їх годинники — синхронізованими (що фактично накладає вимогу використовувати NTP).

Так як MGS/MDS критичні для роботи ФС, передбачено можливість автоматичної заміни при збої, так званий failover.

OSS можна легко додавати до ФС, збільшуючи доступний простір; і навпаки, відключення OSS не впливає на працездатність ФС як цілого. Зрозуміло, що файли, які знаходяться на відключеному OSS стають недоступними. Є можливість керування розташуванням файлів на OSS. Наприклад, перш ніж вимикати конкретний сервер даних, можна вивести всі файли з нього.

Важливе уточнення — клієнтами тут називаються відповідні вузли та драйвери Lustre, що працюють на них. Користувацьке програмне забезпечення в нормі працює з файлами звичайним чином, не знаючи нічого про Lustre. Отож, з точки зору клієнта, робота з ФС виглядає так:

- Клієнт звертається до MDS, отримує метадані файлу та інформацію про його структуру.
- Для читання чи запису інформація про структуру передається LOV (logical object volume), який визначає, які дані з яких саме OST належать цьому файлу.
- Клієнт блокує відповідні ділянки файлу, після чого здійснює серію операцій вводу-виводу, звертаючись безпосередньо до OST.

[1] Слід зауважити, що вибір відбувався в 2005-2006 році. З того часу на ринку з'явилося кілька вартих уваги альтернатив, такі як GlusterFS, Ceph, із Open Source та FraunhoferFS із закритих розробок.
[2] Станом на 2012 рік Whamcloud була куплена Intel, але важливих для користувачів змін це поки не викликало.
[3] В Lustre 2.4 з'явилася можливість створювати багато MDS.

Завдяки такому розподілу функцій, більшість звертань та потік даних не йде через якийсь конкретний вузол, доступ є розподіленим, тому ймовірність виникнення вузьких місць мала. Найважливіше — все це відбувається прозоро для програм, що працюють з файловою системою.

Знаходитися всі описані компоненти, MGS, MDS, OSS та клієнти, можуть як на одному вузлі, так і, практично в будь-яких комбінаціях, на різних. Хоча, для повноцінної роботи ,їм варто все ж знаходитися на різних вузлах. Для зв'язу між вузлами можуть використовуватися звичайний Ethernet, Infiniband, Mellanox, тощо. На даний момент Lustre несумісна з SELinux, і вимагає повного вимкнення цього механізму.

Працюють ці логічні елементи поверх фізичних (backend) пристроїв. "Бекендом" може служити практично будь-який блочний пристрій — розділи дисків, RAID, LVM-томи, loopback-пристрої.

3 СЗД кластера ІФКС

В даний момент на кластері використовується Lustre 2.0, планується оновлення. СЗД включає три вузла. Один із них служить MDS та MGS, однак ці сервери використовують різні блочні пристрої, для спрощення модифікацій та можливості додавання інших Lustre-файлових систем. Як MDT, так і MGT розташовані на дискових масивах RAID1, з двома активними та одним резервним диском. Операційна система теж знаходиться на RAID1 і може завантажуватися із будь-якого диску масиву.

OSS два, кожен з яких використовує одну OST. Фізично кожна з OST знаходиться на дисковому масиві RAID6 із шести 2Тб дисків. Корисна ємність – 8Тб на OSS, повна ємність СЗД – 16Тб. Використання RAID6 виправдане, не зважаючи на зниження корисної ємності, так як в процесі експлуатації вихід з ладу жорстких дисків відбувається доволі часто, в середньому кілька раз на рік, за час експлуатації відбувся принаймні один подвійний збій.

В процесі експлуатації, окрім заміни жорстких дисків, що вийшли з ладу, здійснювалися наступні операції. Проводилося оновлення Lustre, спочатку від версії 1.6 до 1.8, а потім від 1.8 до 2.0. Здійснювалося введення в експлуатацію та виведення з експлуатації OSS. Перш ніж проводити якісь операції на робочій системі, процедура відпрацьовується на модельній Lustre-системі, побудованій на loopback-пристроях.

Із мінусів слід зауважити ряд дрібних збоїв та помилок в коді Lustre. Прикладом може служити помилка LU-129 [8], через яку квоти не працюють, якщо нумерація OSS має розриви.

4 Висновки

Розподілена файлова система Lustre, яка використовується на кластері ІФКС в ролі системи збереження даних показала себе із найкращої сторони, і, не зважаючи на ряд дрібних проблем, повністю виправдала себе. Це ще раз підтверджує світовий досвід її експлуатації та може служити рекомендацією для груп, які займаються вибором майбутньої бази для СЗД.

Література

[1] http://www.whamcloud.com/lustre/

[2] http://grid.nas.gov.ua/

[3] http://infrastructure.kiev.ua/

[4] http://gridmon.bitp.kiev.ua/

[5] http://www.icmp.lviv.ua/multiscale/

[6] http://voms.grid.org.ua/voms/?vo=multiscale

[7] http://www.top500.org/lists/2012/11/

[8] http://jira.whamcloud.com/browse/LU-129