Письма в ЭЧАЯ. 2017 Т. 14, № 1(206). С. 125–132

МЕТОДИКА ФИЗИЧЕСКОГО ЭКСПЕРИМЕНТА

МОДЕРНИЗАЦИЯ LGD-КЛАСТЕРА ОИЯИ ДЛЯ ОБЕСПЕЧЕНИЯ ЭКСПЕРИМЕНТОВ ЛЯП

И. В. Бедняков [1], *А. Г. Долбилов, Ю. П. Иванов*

Объединенный институт ядерных исследований, Дубна

С момента создания в 2005 г. вычислительного кластера Лаборатории ядерных проблем (ЛЯП) главной его целью было и остается выполнение вычислительных задач (анализ данных, моделирование и т. п.) для различных научных коллабораций, в работе которых сотрудники лаборатории принимают активное участие. Кластер также служит для обучения специалистов. За прошедшее десятилетие многое изменилось, и для поддержания статуса надежного и современного кластера необходимо было провести модернизацию, нарастить мощность и заменить устаревшее оборудование. В данной статье описан опыт проведения такой модернизации. Этот опыт может быть полезен системным администраторам для быстрого и эффективного ввода в эксплуатацию нового оборудования кластеров данного типа.

Since creation in 2005 of the Computer Cluster at the Laboratory of Nuclear Problems (LNP), its main purpose has been to perform computing tasks (data analysis, modeling, etc.) for a variety of scientific collaborations, where our employees are working too. The cluster also provides for the training of specialists. Over the past decade, much has changed, and it was necessary to modernize, to increase capacity and replace outdated equipment in order to maintain the status of a reliable and modern cluster. This article describes the experience of such modernization. This experience can be useful for system administrators for fast and efficient commissioning of the new equipment of this type of clusters.

PACS: 01.50.Lc; 07.05.Tp; 89.20.Ff

ВВЕДЕНИЕ

В середине 1990-х гг. началась разработка методов распределенных вычислений, получивших впоследствии название GRID [1]. В ОИЯИ работы по созданию GRID-инфраструктуры начались в 2002 г. [2], а в 2004 г. по инициативе дирекции Лаборатории ядерных проблем (ЛЯП) был создан второй в ОИЯИ кластер [3], получивший название LGD (LNP Grid Development). В течение 10 лет кластер ЛЯП успешно и безотказно справлялся со своей задачей: на нем выполнялись различного рода моделирование и отладка программ, предварительный и окончательный анализ данных, проводились разнообразные вспомогательные вычисления и расчеты сотрудников ЛЯП, участвовавших в разных экспериментальных коллаборациях. В значительной мере, особенно на начальном этапе

[1]E-mail: bednyakovi@gmail.com

становления кластера, такого сорта задачи поступали от пользователей ЛЯП, участвовавших в эксперименте ATLAS [4], которые, собственно, и были одними из главных инициаторов создания кластера. Тем не менее сотрудники ОИЯИ, занятые в других экспериментах, также успешно использовали этот кластер для своей работы. За прошедшие годы как в самой лаборатории, так и в сфере компьютерного и сетевого оборудования произошли заметные изменения, что естественным образом привело к необходимости существенной модернизации кластера. Очевидно, что для сохранения статуса надежного и эффективного кластера была необходима замена устаревшего оборудования (в первую очередь всех рабочих узлов кластера), а также существенное наращивание мощности. При этом все мероприятия по модернизации должны были быть проведены в кратчайшие сроки, поскольку работу кластера нельзя останавливать надолго.

В данной статье описан опыт проведения такой модернизации на примере кластера ЛЯП. Эти последовательные действия вполне могут быть полезны любому системному администратору (или просто «продвинутому» пользователю) в случае необходимости быстрой замены оборудования и ввода его в эксплуатацию.

ПРОЦЕДУРА МОДЕРНИЗАЦИИ КЛАСТЕРА

Главным звеном всего процесса модернизации является процедура быстрого переноса общей для всех элементов кластера системной информации с одного (устаревшего) рабочего элемента кластера на другой (новый) рабочий элемент.

Для проведения этой процедуры необходимо наличие версии Linux с возможностью загрузки системы с какого-либо внешнего носителя (CD/DVD, USB флеш-память) или же по сети Ethernet с использованием протокола PXE (Preboot Execution Environment). Дистрибутивы большинства современных версий Linux предоставляют такие возможности. В качестве конкретного варианта такой системы нами был выбран RIP Linux (Recovery is Possible) [5], позволяющий запуск системы со всех типов носителей (CD/DVD/USB), включая удаленную загрузку по сети. Далее в тексте такой способ старта системы будем обозначать как запуск Live CD, без указания конкретного способа загрузки.

Собственно, вся работа заключается в копировании операционной системы с жестких дисков старого рабочего элемента на жесткие диски нового рабочего элемента кластера с последующей адаптацией перенесенной системы к новому оборудованию. Такие действия позволяют избежать длительной и трудоемкой настройки нового рабочего элемента, не-

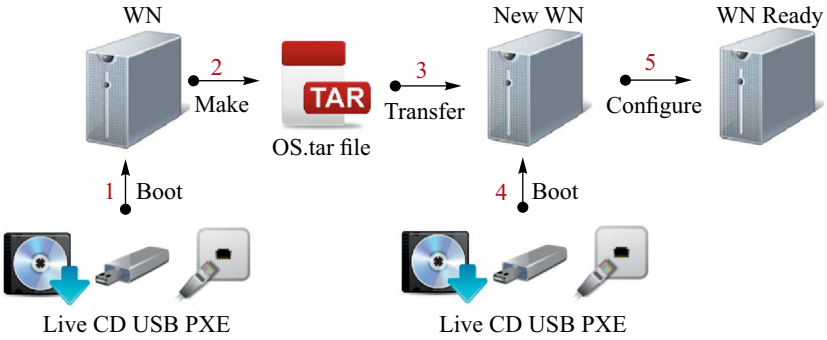

Рис. 1. Схема действий по подготовке к работе нового элемента кластера New WN

избежной при установке системы «с нуля», что позволяет значительно сократить время ввода в эксплуатацию нового оборудования. На рис. 1 схематично показаны все действия, в результате которых новый рабочий элемент кластера, обозначенный как New WN (Worker Node), будет полностью подготовлен к работе.

Всю процедуру можно разделить на три главных этапа (см. рис. 1):

1) создание образа системы старого рабочего элемента (шаги 1 и 2);

2) перенос и развертывание этого образа на новом элементе (шаги 3 и 4);

3) адаптация системы к новому элементу (шаг 5).

Ниже все эти этапы рассмотрены подробно, с соответствующими пояснениями.

Этап 1 — создание образа системы старого рабочего элемента. Этап начинается с запуска системы Live CD на рабочем элементе кластера, с жестких дисков которого необходимо скопировать образ системы. Такая загрузка (т. е. использование сторонней системы) позволяет получить полный доступ к копируемым системным файлам компьютера. В нашем случае вся дисковая система рабочего элемента WN расположена на двух жестких дисках, на одном из которых установлена операционная система, а другой служит для хранения пользовательских данных. По сути дела, именно копирование и последующая конфигурация системных разделов представляет собой основную сложность процесса переноса системы. Копирование же пользовательских данных является элементарной процедурой и может быть сделано после переноса самой системы на новое оборудование. Диск с операционной системой посредством менеджера логических томов LVM (Logical Volume Manager) разбит на разделы, входящие в одну LVM-группу `vg`: root, home и data, на которых расположены разделы `/root`, `/home` и `/data` соответственно. Версия Linux, используемая на кластере, не позволяет загрузку системы с тома LVM, поэтому загрузочный раздел `/boot` размещен на обычном (primary) разделе системного диска `/dev/sda1`. Ключевыми элементами файловой системы являются `/boot` и `/root` разделы.

Сразу после загрузки Live CD дисковые тома LVM неактивны. С помощью команды `lvscan` можно обнаружить существующие разделы LVM. Командой `vgchange -ay` эти тома активируются. Для дальнейшей работы монтируем все дисковые системы рабочего элемента в директорию `/mnt`:

```
cd /mnt
mount /dev/sda1 /mnt/boot
mount /dev/vg/root /mnt/root
mount /dev/vg/home /mnt/home
mount /dev/vg/data /mnt/data
mount /dev/sdb1 /mnt/sdb1
```

Подключение пользовательского раздела `/dev/sdb1` необходимо для дальнейшего сохранения образов boot, root, home и data. Эти образы создаются следующими командами:

```
cd /mnt/root ; tar cf /mnt/sdb1/lgdwn-root.tar . --numeric-owner
cd /mnt/boot ; tar cf /mnt/sdb1/lgdwn-boot.tar . --numeric-owner
cd /mnt/home ; tar cf /mnt/sdb1/lgdwn-home.tar . --numeric-owner
```

Команда `tar` создает файл, содержащий образ логического диска на диске `sdb1`. Используемый параметр `--numeric-owner` важен, поскольку он позволяет избежать возможных несоответствий имен и групп владельцев файлов и директорий в загруженной

(Live CD) и копируемой системе (диски рабочего элемента). По завершении копирования стоит проверить наличие файлов в папке `/mnt/sdb1`. На этом первый этап закончен.

Этап 2 — перенос и развертывание образа системы на новом элементе. В конкретном варианте проведенной модернизации на кластере жесткий диск `sdb` со всеми имеющимися пользовательскими данными просто переставлялся со старой на новую машину (этап 3). Помимо сохранения всех расположенных на нем данных пользователей это позволило также избежать организации промежуточного сервера хранения и передачи образов системы по сети или посредством какого-либо другого более медленного способа.

Итак, после физической установки на новой машине диска `sdb` со всеми необходимыми файлами-образами системы «загружаем» Live CD на новом элементе (этап 4). Для переноса системы необходимо полностью сохранить исходную структуру, т. е. создать точную ее копию с такими же логическими разделами и файловой системой. Сначала с помощью утилиты `fdisk` проводим разметку системного диска `sda` на два раздела: первый `/dev/sda1` предназначен для загрузочных файлов (/boot), а на втором `/dev/sda2` разместим с использованием LVM все остальные системные разделы. Создание и активация логических разделов диска осуществляется последовательностью команд

```
pvcreate /dev/sda2
vgcreate vg /dev/sda2
lvcreate -L 20G -n root vg
lvcreate -L 10G -n home vg
lvcreate -L 16G -n temp vg
lvcreate -L 32G -n swap vg
lvcreate -L 850G -n data vg
vgchange -ay
```

Следующий шаг — форматирование созданных разделов:

```
mkfs.ext4 -L /boot /dev/sda1
mkfs.ext4 -L / /dev/vg/root
mkfs.ext4 -L /home /dev/vg/home
mkfs.ext4 -L /tmp /dev/vg/temp
mkfs.ext4 -L /data /dev/vg/data
mkswap -L swap /dev/vg/swap
```

И наконец, создание директорий согласно «старой» системе WN и монтирование дисков завершается последовательностью команд

```
mkdir -p /mnt/boot
mount /dev/sda1 /mnt/boot
mkdir -p /mnt/root
mount /dev/vg/root /mnt/root
mkdir -p /mnt/home
mount /dev/vg/home /mnt/home
```

Теперь структура нового диска `sda` полностью соответствует структуре исходного рабочего элемента кластера. На нем осталось развернуть саму операционную систему,

которая хранится на втором диске sdb. Для этого монтируем раздел /dev/sdb1 и развертываем систему:

```
mkdir -p /mnt/sdb1
mount /dev/sdb1 /mnt/sdb1
cd /mnt/boot ; tar xf /mnt/sdb1/lgdwn-boot.tar
cd /mnt/root ; tar xf /mnt/sdb1/lgdwn-root.tar
cd /mnt/home ; tar xf /mnt/sdb1/lgdwn-home.tar
umount /mnt/sdb1
umount /mnt/home
umount /mnt/boot
```

Итак, на этом этапе на новом компьютере воспроизведена вся дисковая и файловая структура исходной машины. При этом, естественно, все скопированные файлы, включая «загрузочные» системные файлы, соответствуют старой машине. Осталось провести некоторые операции (включая настройку подсистемы «загрузки» на новом элементе), необходимые для завершения настройки нового рабочего элемента.

Этап 3 — адаптация системы на новом элементе. Копирование системы было проведено в системе Live CD, поскольку подсистема «загрузки» операционной системы на новой машине пока не настроена. Удобнее всего сделать все окончательные настройки, переключив «корневой» раздел на файлы скопированной системы, т. е. для максимально возможного приближения к среде копируемой системы следует с помощью команды chroot перейти в «корень» скопированной системы. При этом система содержит практически все, что есть в нормально стартовавшей системе, кроме ряда динамических разделов Linux, таких как /dev, /proc и /sys. Эти структуры можно взять из Live CD, заранее (т. е. до команды chroot) смонтировав их с параметром --bind в нужные точки файловой системы:

```
mount --bind /dev /mnt/root/dev
mount --bind /sys /mnt/root/sys
mount --bind /proc /mnt/root/
chroot /mnt/root
```

Дальнейшая настройка идет на новом рабочем элементе фактически так же, как если бы уже работала скопированная система, а не операционная система, из-под которой был загружен компьютер (Live CD).

Сначала монтируется «загрузочный» раздел /boot и создается «начальная файловая система» (initfamfs — Initial RAM File System), соответствующая аппаратным компонентам нового элемента и нужной версии «ядра» системы (в примере ниже использована версия «ядра» kernel=2.6.32-573.el6):

```
mount /dev/sda1 /boot
cd /boot
mv initramfs-kernel.x86_64.img initramfs-kernel.x86_64.img-orig
dracut initramfs-kernel.x86_64.img kernel
grub-install /dev/sda
```

На этом этапе оригинальный вариант `intramfs` был на всякий случай сохранен (файл `intramfs...img-orig`). По завершении настройки и при успешной загрузке новой системы этот файл можно удалить.

Загрузочный сектор создан, теперь компьютер сможет уже сам корректно загрузить операционную систему. Но до этого можно провести еще ряд дополнительных настроек как общего плана, так и специфических именно для рабочего элемента кластера LGD. Сначала удаляем ненужные сетевые и дисковые параметры старого компьютера:

- удалить все строчки в файле `/etc/udev/rule.d/70-persisent-net.rules`
- удалить все файлы `blkID.tab*` из директории `/etc/blkid/`

И наконец, монтируем разделы `/data` и `/tmp`, в которых устанавливаем директориям права доступа, необходимые для корректной работы ряда систем (cvmfs, PBS и пр.):

```
mount /dev/vg/temp /tmp
chmod 1777 /tmp
mount /dev/vg/data /data
mkdir /data/cvmfs
chown cvmfs:cvmfs /data/cvmfs
mkdir /data/jobScratch
chmod 1777 /data/jobScratch
```

Настройки новой машины завершены. Можно отключить смонтированные диски, вернуться в основную среду Live CD и перезагрузить компьютер в нормальном режиме:

```
umount /boot
umount /data
umount /tmp
exit
umount /mnt/root/sys
umount /mnt/root/proc
umount /mnt/root/dev
umount /mnt/root
reboot
```

После успешной загрузки системы необходимо проверить работоспособность всех системных компонентов на предмет каких-либо ошибок, а также нормальное функционирование основных служб (пакетной системы PBS и пр.), начав с просмотра файлов регистрации в директории `/var/log` (`boot.log`, `messages` и т. д.). Для проверки системы `cvmfs` следует запустить команду `cvmfs_confige probe`. На этом развертывание нового рабочего элемента закончено.

КЛАСТЕР ЛЯП В 2015 Г.

В начале 2015 г. была успешно и в кратчайшие сроки проведена модернизация кластера с использованием вышеизложенного метода. В таблице приведены сравнительные характеристики кластера (CPU — процессоры, RAM — оперативная память, Disk — общее дисковое пространство и HEP-SPEC06 — совокупное быстродействие в единицах HEP-SPEC06 тестов производительности) до и после модернизации.

Сравнительные характеристики кластера LGD до и после проведенной модернизации

Характеристика	До 2015 г.	После модернизации
CPU, ядра	10×2	10×24
RAM, Гбайт	10×8	10×64
Disk, Тбайт	8	25
HEP-SPEC06 [6]	200	2000

После модернизации кластер ЛЯП (LGD) включает в себя:

— 10 рабочих узлов (Worker Node) $\times$ 24 CPU Xeon 2,1–2,6 ГГц (каждый рабочий узел оснащен двумя дисками, которые, в частности, используются как дисковые элементы распределенных сетевых файловых систем (GlusterFS и XrootD) для хранения пользовательских данных);

— файловую систему AFS + ZFS, где расположены и директории «home».

На рис. 2, *а* можно видеть, что после модернизации кластера только за 2015 г. использовано процессорного времени больше, чем за все предыдущие годы вместе взятые. Также на графике (рис. 2, *б*) приведено количество обработанных задач (по годам с 2008 по 2015).

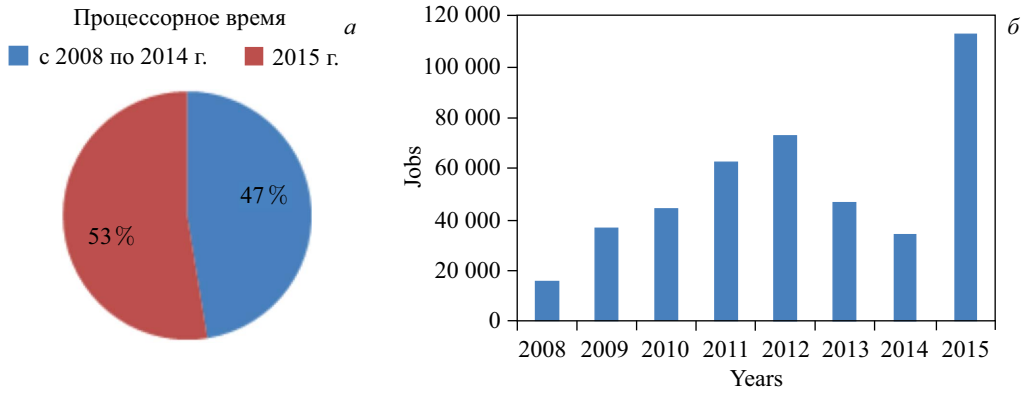

Рис. 2. *а*) Процессорное время; *б*) число задач с 2008 по 2015 г.

Основными потребителями кластерных возможностей LGD являются сегодня следующие эксперименты ЛЯП: эксперимент ATLAS [7] — один из четырех основных экспериментов на Большом адронном коллайдере (LHC) в ЦЕРН [8], эксперимент BES-III [9] на электрон-позитронном коллайдере BEPC-II (Beijing Electron-Positron Collider) в Китае, эксперимент E391a [10] по поиску нарушающих CP-четность распадов нейтральных K-мезонов в Японии (High Energy Accelerator Research Organization, KEK), а также новый базовый эксперимент ОИЯИ — Байкальский нейтринный телескоп Baikal-GVD [11], задачей которого является исследование природных (космических) источников нейтрино и мюонов высоких и сверхвысоких энергий, и другие эксперименты.

ЗАКЛЮЧЕНИЕ

В 2005 г. в ЛЯП был запущен кластер LGD. Более 10 лет он успешно использовался для анализа данных, моделирования различного типа процессов и установок, а также для обучения специалистов. В 2015 г. успешно проведена его существенная модернизация. В данной работе описан опыт проведения этой модернизации, который, являясь достаточно универсальным, может быть полезен системным администраторам для быстрого и эффективного ввода в эксплуатацию нового оборудования кластеров данного типа. Ряд общих вопросов процедуры модернизации был пояснен на конкретных примерах.

В заключение следует отметить, что в связи с постоянным ростом потребностей конечных пользователей, как физиков, так и инженеров ЛЯП, представляется необходимым дальнейшее развитие кластера ЛЯП. Кроме того, кластер всегда являлся стартовой площадкой для запуска и тестирования разнообразных новых задач, связанных с вычислениями и хранением данных. В связи с этим группа администраторов кластера всегда была и остается заинтересованной в сотрудничестве с новыми людьми и в новых проектах.

СПИСОК ЛИТЕРАТУРЫ

1. Грид в ОИЯИ (информационный портал ОИЯИ). http://grid.jinr.ru/?page_id=39.
2. *Долбилов А. Г., Иванов Ю. П.* Элемент GRID-системы LGD-2 в ЛЯП. Сообщ. ОИЯИ Р11-2008-68. Дубна, 2008.
3. *Бедняков И. В., Долбилов А. Г., Иванов Ю. П.* Метод клонирования ГРИД-элемента // Письма в ЭЧАЯ. 2010. Т. 7, № 6(162). С. 699–704.
4. ATLAS Experiment. http://atlas.web.cern.ch/Atlas/Collaboration/.
5. RIP Linux. http://sourceforge.net/projects/riplinuxmeta4s/.
6. HEP-SPEC06. https://w3.hepix.org/benchmarks.
7. ATLAS Experiment. http://atlas.web.cern.ch/Atlas.
8. CERN. http://home.cern/.
9. BES-III. http://bes3.ihep.ac.cn/.
10. KEK-PS E391a. http://www-ps.kek.jp/e391/.
11. Baikal Experiment. http://baikalweb.jinr.ru/.

Получено 23 марта 2016 г.