Университетский Кластер НИЯУ МИФИ

Руководство пользователя

Версия 1.0

Москва, 2012

Все замечания и предложения по этой документации с благодарностью принимаются по адресу unicluster@mephi.ru.

Создано при помощи XeLaTeX.

Последнее изменение: 28 июля 2012 г.

Содержание

1. Общие положения

Университетский кластер НИЯУ МИФИ предназначен для выполнения ресурсоёмких и/или распределённых вычислений сотрудниками и учащимися Университета при выполнении научных, исследовательских и образовательных задач, в частности, для обучения использованию современных HPC-технологий.

В рамках имеющихся аппаратных ресурсов пользователям предоставляется возможность выполнять задачи с использованием следующих технологий на основе ОС Linux:

- PBS (менеджер ресурсов Torque [1], шедулер Maui [2]);
- MPI (реализация OpenMPI [3]);
- OrangeFS [4] (распределённая параллельная виртуальная файловая система, поддерживающая ROMIO [5]).

2. Как стать пользователем

Перед тем как подавать заявку на использование ресурсов кластера, пожалуйста, ознакомьтесь с правилами использования. Фактом подачи заявки Вы подтверждаете своё согласие с данными правилами и обязательность их исполнения Вами.

2.1. Правила использования

1. Ресурсы университетского кластера НИЯУ МИФИ предназначены для поддержки фундаментальных и прикладных научных исследований, исследовательских и образовательных задач, требующих привлечения HPC.
2. Пользователями могут быть только сотрудники и учащиеся НИЯУ МИФИ.
3. Пользователям категорически запрещается передавать свою учётную запись, пароль к ней или секретный ssh-ключ иным лицам.
4. Пользователь обязуется не использовать кластер для задач не указанных в п.1, в т.ч. для какой-либо деятельности, противоречащей законодательству РФ.

5. Установка и использование нелегального программного обеспечения категорически запрещена, как и любое нарушение лицензий на используемое ПО (например, попытка использования на кластере бесплатного только для частного использования ПО).

6. Пользователям запрещается пытаться обойти систему защиты, квот или административных ограничений кластера, в частности, эксплуатировать уязвимости.

7. В случае обнаружения уязвимости системы, пользователь обязан незамедлительно сообщить об этом администраторам по e-mail unicluster@mephi.ru.

8. Администрация кластера выполняет тщательный аудит действий пользователей.

9. В случае обнаружения нелегитимной активности или нарушения данных правил со стороны пользователя, администрация оставляет за собой право блокирования соответствующей учётной записи с возможным удалением нелегальных данных.

10. Пользователи в публикациях работ, выполненных при помощи данного кластера, обязуются ссылаться на использование ресурсов Университетского кластера НИЯУ МИФИ. Для русскоязычных работ следует использовать формулировки вида «при проведении работ был использован Университетский кластер НИЯУ МИФИ», для англоязычных — "our work was performed using NRNU MEPhI University cluster".

11. Администрация прилагает все возможные усилия для безотказной работы кластера и сохранности пользовательских данных. Однако, в силу объективных обстоятельств, невозможно гарантировать абсолютную стабильность и сохранность информации, поэтому пользователи должны регулярно сохранять полученные результаты и хранить копии особо важных данных вне кластера.

2.2. Получение учётной записи

Для получения учётной записи Вам необходимо заполнить заявку http://report.ut.mephi.ru/unicluster_request/. Затем с Вами будет согласовано время визита в В-123 для идентификации пользователя (возьмите с собой пропуск в МИФИ или удостоверение) и Вам будет выдан логин/пароль.

При желании Вы сможете изменить пароль с помощью команды passwd, но при этом он должен будет соответствовать строгим требованиям безопасности как по длине, так и по сложности. Система не позволит Вам установить слабый пароль.

2.2.1. Получение учётных данных по e-mail

Так же можно получить учётные данные по электронной почте в зашифрованном виде. Для этого необходимо создать PGP-ключ для указанного e-mail и разместить его открытый подключ на любом из публичных GPG-серверов. Период синхронизации всех серверов составляет около суток, поэтому создавайте ключ заблаговременно. Инструкцию по созданию ключей и работе с GnuPG можно найти в работе [6].

Обратите внимание, что данный механизм получения пароля не отменяет необходимость персональной идентификации пользователя.

2.2.2. Регистрация группы пользователей

При необходимости получить учётные записи для большой группы пользователей, можно подать заявку сразу на всю группу в виде служебной записки от руководителя группы или подразделения на имя начальника управления информатизации Романова Николая Николаевича. В служебной записке должны быть перечислены ФИО пользователей, их e-mail и цель работ. Ответственность за достоверность предоставленных данных находится на руководителе, подавшем служебку.

Данный механизм авторизации требует использования цифровых ключей для почты пользователей, описанных в разделе 2.2.1. Пароли для всей группы на руки не выдаются, участникам группы приходить в В-123 не нужно.

Для групповой заявки также рекомендуется заполнить он-лайн заявку `http://report.ut.mephi.ru/unicluster_request/`.

2.2.3. Использование ssh-ключей

При желании, пользователь после получения пароля может использовать ssh-ключи для доступа к Университетскому кластеру. Для этого необходимо *на клиентской машине* создать ключ с помощью:

```
ssh-keygen -b 521 -t ecdsa
```

Не забудьте указать сложный пароль для защиты ключа! Затем поместите ключ в файл `~/.ssh/authorized_keys` на кластере. Обратите внимание, что передача ключа иным лицам категорически запрещена.

3. Структура кластера

3.1. Аппаратная часть

Вычислительные ресурсы кластера составляют:

- 128 ядер;
- 512 GB RAM;
- 1.5 TB полезного дискового пространства;
- сеть 1 Gbit/s;
- пиковая производительность $\sim$ 1.5 TFlops.

Кластер состоит из 16 узлов. Каждый узел состоит из:

- 2 x E5450 Intel Xeon CPU;
- 4 физических ядра на CPU.
- 32 GB RAM;
- 120 GB HDD;
- BCM5715S Gigabit Ethernet.

3.2. Программная инфраструктура

Кластер работает на базе операционной системы Linux, дистрибутив Gentoo [7], используемое ядро 3.2.18 (LTS).

Пользоваетли непосредственно работают по ssh только с мастер-нодой unicluster.campus.mephi.ru, она же предназначена для компиляции приложений. Работа с вычислительными узлами осуществляется последством инструментов PBS без прямого доступа пользователя.

На кластере предоставляются следующие инструменты:

PBS На кластере используется система управления распределёнными вычислениями (PBS, Portable Batch System) на основе менеджера ресурсов Torque [1] версии 3.0.5 и диспетчер задач Maui [2] версии 3.3.1. В рамках PBS реализована поддержка MPI задач.

MPI Поддержка MPI (Message Passing Interface, инструмент для обмена данными между параллельно работающими задачами на разных узлах) реализована с помощью пакета OpenMPI [3] версии 1.5.5. С точки зрения пользователя, MPI-приложение работает внутри PBS-задачи. Поддерживается ROMIO [5] I/O API.

Работа MPI-приложений ускорена с использованием технологии KNEM [8], особо эффективной для передачи больших объёмов данных, асинхронного и векторного обмена данными.

PVFS2 На Университетском кластере применяется распределённая параллельная виртуальная файловая система OrangeFS [4] версии 2.8.5, являющаяся ветвью PVFS2. Данное решение позволяет максимально полно использовать имеющиеся ресурсы дискового пространства, а также предоставить пользователям все достоинства параллельного ввода-вывода данных, что позволяет на грамотно спроектированных приложении получать скорости доступа к файлам, ограниченные лишь пропускной способностью сети.

Distcc Предусмотрена возможность использования кластера для помощи в распределённой компиляции во внутренних сетях НИЯУ МИФИ с использованием технологи distcc [9]. На данный момент находится в стадии тестирования.

GCC Для компиляции пользовательских приложений предоставляется стандартная для Linux коллекция компиляторов GCC версии 4.5.3, поддерживающая следующие языки: C, C++, Assembler, Fortran, Objective C, Objective C++. Поддерживается технология OpenMP [10]. Для компиляции MPI приложений следует использовать соответствующие команды с префиксом «mpi»: mpicc, mpic++, mpif77, mpi90.

Текстовые редакторы Представлены текстовые редакторы Vim, Emacs и, для новичков, mcedit и nano.

Взаимодействие основной инфраструктуры ПО кластера с пользовательскими задачами отображено на рис. 1.

Обратите внимание, что X-сервер на нашем кластере не поддерживается. Аппаратные акселераторы визуализации физически отсутствуют. Кластер предназначен для вычислительных задач, а не для визуализации полученных результатов, которую необходимо выполнять на клиентских системах.

На кластере имеется предустановленное ПО для задач области физики частиц: ROOT, Geant, Pythia. При поступлении заявок, иное научное программное обеспечение может быть установлено общесистемно

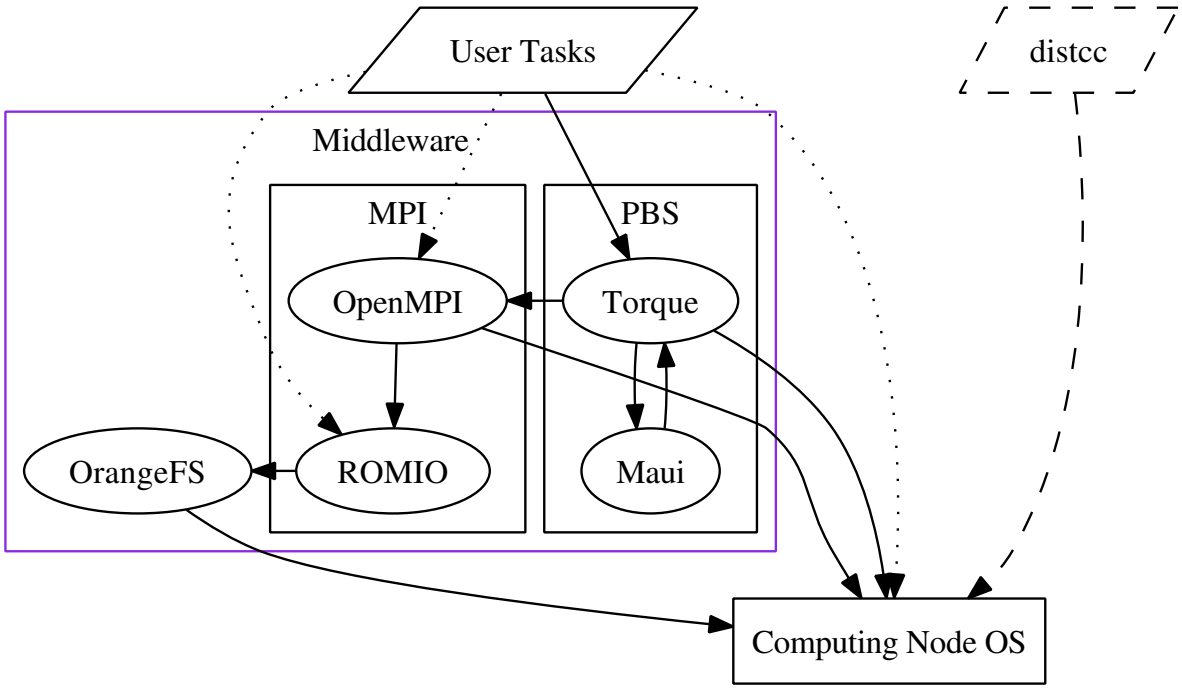

Рис. 1. Основная инфраструктура ПО кластера

и поддерживаться системными администраторами, при условии что оно доступно в стандартных репозиториях Gentoo (portage, science overlay).

3.3. Политика обновлений

С целью поддержания актуальности и безопасности установленного общесистемного ПО, будут проводится регулярные обновления системы. Плановое обновление будет проходить раз в полгода (или раз в год, в зависимости от обстановки), между учебными семестрами. Администраторы оставляют за собой право проводить экстренное обновление отдельных компонент при выявлении серьёзных проблем безопасности.

4. Использование кластера

4.1. Предварительные требования

Для работы с кластером пользователю необходимо обладать базовыми навыками работы в ОС Linux; в частности, необходимо владеть `bash`, `ssh`, `gcc`, одним из установленных текстовых редакторов (`vim`, `emacs`, `nano`, `mcedit`) и обладать навыками программирования, достаточными для компиляции пользовательских приложений.

Большинство этих аспектов выходит за рамки данного руководства, которое посвящено описанию предоставляемой рабочей среды (см. раздел 3.2) и особенностям, специфичным для Университетского кластера. Для каждого приложения на кластере установлена документация, доступная посредством `man`, `info` и в `/usr/share/doc`. Однако, начинающим в Linux можно порекомендовать следующие материалы для ознакомления: Linux in a Nutshell [11], Advanced Bash-Scripting Guide [12].

4.2. Работа с файловой системой

Пользователям предоставляется три вида дисковых хранилищ:

- /home на NFS [13];
- /tmp;
- ~/pool на OrangeFS [4].

4.2.1. /home

Домашняя директория пользователей расположена на NFS-4 разделе и обладает полной POSIX-совместимостью. В силу ограниченности ресурсов, дисковое пространство для каждого пользователя в `$HOME` ограничено размером 2 GB и количеством файлов 100 000. Ограничения можно кратковременно превышать в определённых пределах, однако, при длительном превышении (7 дней и более) пользователь будет автоматически заблокирован. Узнать о Вашей текущей дисковой квоте можно с помощью команды `quota`.

Данный раздел предназначен для компиляции пользователями своих приложений, а также для использования специальных файлов (сокеты, fifo и т.п.) приложениями, для которых нужно, чтоб данные файлы были доступны для всех запущенных процессов на разных узлах одновременно.

4.2.2. /tmp

Для хранения временных данных рабочего процесса на конкретном вычислительном узле предназначен раздел /tmp, так же обладающий полной POSIX-совместимостью. На дисковое пространство действуют те же ограничения, что и на раздел /home (см. 4.2.1), но срок временного превышения квоты сокращён до трёх дней.

Обратите внимание, что размер квоты относится ко всем пользовательским процессам, работающим на данном узле, суммарно. Файлы на /tmp, не используемые в течении 7 дней, будут автоматически удалены.

4.2.3. ~/pool

Основное хранилище данных предоставляется пользователям в виде директории на кластерной распределённой параллельной виртуальной файловой системе OrangeFS [4], доступной как ~/pool.

Хранилище предназначено для исходных данных, результатов обработки и прочих пользовательских данных, а также для установки пользовательских приложений. Система квот не применяется, но пользователям не рекомендуется занимать более необходимого и более 1/3 от общего объёма хранилища. На данный момент на всей файловой системе доступно 1.4 TB для всех пользователей суммарно (это предел технических возможностей оборудования).

Как и при работе с любой распределённый параллельной файловой системой, при работе с OrangeFS следует учитывать её характерные особенности. Данная файловая система является ветвью PVFS, предназначенной для HPC вычислений и оптимизированной для работы MPI приложений. Она не является полностью POSIX-совместимой, в частности, отсутствует механизм блокировок (POSIX file locks) — целостность данных гарантируется за счёт атомарности операций. Так же на ней нельзя создавать специальные файлы и жёсткие ссылки (для этих задач используйте `$HOME` или /tmp), но можно использовать символические ссылки.

Файловая система хорошо оптимизирована для параллельного доступа большого числа процессов с разных узлов, т.о. Вы можете использовать эффективный ввод-вывод данных при работе распределённых MPI-приложений. Подробнее о возможностях файловой системы можно узнать в вики [14] проекта.

При работе с OrangeFS настоятельно рекомендуется максимизировать размер данных в операциях чтения/записи (вплоть до 1MB). Если Ваше приложение будет производить работу с файлами блоками данных по несколько десятков байт, Вы получите резкое падение скорости чтения или записи. При невозможности исправить приложение для

корректной работы с распределёнными параллельными файловыми системами, рекомендуется использовать /tmp для локального кеширования данных. Обратите внимание, что данное требование распространяется так же на любой сетевой обмен данными, будь то NFS или MPI.

В процессе штатной работы доступ к обычным файлам на OrangeFS с точки зрения пользовательского процесса ничем не отличается от работы с локальной файловой системой. На случай возникновения аварийных ситуаций есть специальные утилиты доступа к данным, начинающиеся с префикса `pvfs-`, подробно описанные в соответствующих `man` руководствах. При возникновении проблем нужно сообщить администраторам по форме http://report.ut.mephi.ru/unicluster/.

4.3. Подготовка пользовательских приложений

В общем случае, пользовательское приложение должно быть откомпилировано на мастер-ноде кластера. Для этого используется стандартный набор компиляторов `gcc`, описанный в разделе 3.2. Также доступны отладчики `gdb` и `valgrind`.

На кластере предоставляется широкий набор системных и научных библиотек. Для того, чтоб узнать, есть ли библиотека (или любой пакет) в подключенных репозиториях и установлена ли она в системе, необходимо использовать команду eix, например (для библиотеки быстрых Фурье-преобразований fftw):

```
$ eix fftw
[U] sci-libs/fftw
     Available versions:
        (2.1)   2.1.5-r8
        (3.0)   3.2.2 (~)3.2.2-r2 (~)3.3.2
     {{altivec avx doc float fortran mpi neon openmp paired-single
quad sse sse2 static-libs threads zbus}}
   Installed versions:  3.3.1(3.0){tbz2}(03:00:31 PM 04/10/2012)
(fortran mpi openmp sse sse2 threads -altivec -avx -doc -neon -quad
-static-libs -zbus)
     Homepage:            http://www.fftw.org/
    Description:        Fast C library for the Discrete Fourier
Transform
```

Подробно использование данной команды описано в `man eix`.

Если Вам необходима библиотека, имеющаяся в репозиториях, но не установленная в системе, сообщите об этом администраторам, используя форму http://report.ut.mephi.ru/unicluster/. Если Вам нужна версия библиотеки, отличная от установленной, то в некоторых слу-

чаях возможна установка доплнительной версии, если не возникает конфликта с основной системной.

Если Вам необходимо приложение, имеющееся в официальных репозиториях, но не установленное на кластере, также обратитесь к администраторам. Но обратите внимание, что X сервер и сопутствующие приложения не поддерживаются. Вычислительные задачи можно откомпилировать без поддержки X.

4.4. Запуск задач

Запуск и удаление задач выполняются с помощью инструментов менеджера ресурсов Torque [1], исчерпывающе описанных в работе [15] и в соответствующих страницах `man`.

В простейшем случае для запуска задачи достаточно выполнить команду:

```
qsub myjob.sh
```

В результате скрипт myjob.sh будет поставлен в очередь long и впоследствии запущен на одном из вычислительных узлов, с возможностью использовать одно ядро и 4GB оперативной памяти, с ограничением на время исполнения в 168 часов (1 неделя). Ограничение по времени исполнения астрономическое (walltime) и не зависит от степени загрузки CPU пользовательским процессом.

Независимо от директории, из которой была выполнена команда, приложение будет запущено в `$HOME` пользователя. После завершения работы программы, stdin и stdout будут размещены в файлах вида `~/${jobname}.o${job_id}` и `~/${jobname}.e${job_id}` соответственно.

Если Вашему приложению нужно передать аргументы или запустить его в директории, отличной от `$HOME`, необходимо использовать скрипт для выполнения соответствующих действий и с помощью qsub запускать этот скрипт, а не само приложение.

4.4.1. Запрос ресурсов

Вы можете явным образом указать необходимые для работы ресурсы, в частности, если необходимо запросить несколько ядер или изменить время исполнения. Например, следующая команда:

```
qsub -q medium -l nodes=2:ppn=8,walltime=10:00:00 job.sh
```

поставит задачу `job.sh` в очередь `medium`, запросив 2 узла с 8 ядрами на каждом и ограничив время исполнения до 10 часов.

Аналогичный результат можно получить, используя специально сформированный заголовок задачи, содержащий директивы `#PBS`:

```
#!/bin/bash
#
#PBS -q medium
#PBS -l nodes=2:ppn=8,walltime=10:00:00
```

В этом случае для запуска задачи достаточно выполнить:

```
qsub job2.sh
```

При наличии параметров как в заголовке задачи, так и в опциях командной строки, учитываются и те и другие с приоритетом за опциями командной строки.

Пожалуйста, объективно оценивайте необходимые ресурсы и максимально точно их указывайте — это позволяет повысить эффективность работы диспетчера задач и уменьшить задержки в очередях. Полное описание доступных ресурсов можно найти в [16].

Обратите внимание, что задачи, превысившие отведённое им время walltime, уничтожаются.

4.4.2. Очереди задач

Задачи распределяются по очередям в зависимости от запрошенного времени исполнения. Чем меньшее время исполнения разрешено в очереди, тем выше её приоритет и тем быстрее начнут исполняться задачи. Это сделано для того, чтоб можно было быстро просчитать небольшие задачи без помех со стороны долгих задач. Кроме того, каждая очередь имеет свои ограничения по числу процессов, которые могут быть в ней одновременно запущены.

Можно явно запросить очередь с помощью:

```
qsub -q $queue_name
```

Естественно, если при этом указан walltime, он не должен противоречить параметрам очереди, если же он не указан, то применяются параметры по-умолчанию, установленные для данной очереди.

Просмотреть список доступных очередей можно с помощью:

```
qsub -Q
```

и детальную информацию по каждой очереди:

```
qsub -Q -f
```

Доступные очереди задач приведены в табл. 1. Для каждой очереди приведено минимальное время исполнения задачи t_{min}, максимально время t_{max} и присваиваемое время по-умолчанию t_{def}.

Если не задана ни очередь задачи, ни время исполнения, задача помещается в очередь long (168 часов) и для интерактивных задач — short (6 часов).

Очередь	t_{min}	t_{max}	t_{def}
short	0:00:00	6:00:00	6:00:00
medium	6:00:01	24:00:00	24:00:00
long	24:00:01	168:00:00	168:00:00
xxl	168:00:01	4320:00:00	2160:00:00
auto	Автоопределение по walltime		

Таблица 1. Очереди задач

4.4.3. Запуск MPI-приложений

Для запуска MPI-приложений необходимо запросить нужное число ядер и/или узлов (см. раздел 4.4.1) и использовать mpirun внутри скрипта запуска задачи для выполнения нужного приложения. Указывать число процессов для mpirun не нужно: оно будет определено автоматически исходя из суммарного запрошенного числа ядер. В остальном работа с MPI-задачами не отличается от обычных задач.

Пример файла описания задачи:

```
#!/bin/bash
#
#PBS -l nodes=2:ppn=8,walltime=05:00:00

cd ~/workdir/
mpirun ./my_mpi_program
```

4.4.4. Интерактивные задачи

При возникновении проблем, в отладочных целях удобно использовать интерактивные задачи, которые предоставляют возможность отладки приложения непосредственно на вычислительном узле.

Для запуска интерактивной задачи следует использовать:

```
qsub -I
```

Интерактивные задачи не могут быть поставлены в очереди long или xxl.

4.5. Мониторинг и управление задачами

Пользователь может просмотреть статус *собственных* задач с помощью команды `qstat`. С параметрами загрузки очередей диспетчера задач Maui [2] можно ознакомиться используя `showq`.

Для удаления задач следует использовать: `qdel $job_id`.

5. Контакты

Сообщить о проблеме можно используя форму http://report.ut.mephi.ru/unicluster/.

Подать заявку на использование ресурсов кластера можно по адресу http://report.ut.mephi.ru/unicluster_request/, но перед этим Вы *должны* тщательно ознакомиться с разделом 2.

В остальных случаях связаться с нами можно по e-mail unicluster@mephi.ru.

Список литературы

[1] http://www.adaptivecomputing.com/products/open-source/torque/ 3, 6, 12

[2] http://www.clusterresources.com/pages/products/maui-cluster-scheduler.php 3, 6, 14

[3] http://www.open-mpi.org/ 3, 7

[4] http://www.orangefs.org/ 3, 7, 9, 10

[5] http://www.mcs.anl.gov/romio/ 3, 7

[6] https://www.pgpru.com/chernowiki/rukovodstva/bezopasnostj/upravleniekljuchami/podkljuchiopenpgp 5

[7] http://www.gentoo.org/ 6

[8] http://runtime.bordeaux.inria.fr/knem/ 7

[9] http://distcc.org/ 7

[10] http://openmp.org/wp/ 7

[11] http://books.google.ru/books/about/Linux_in_a_Nutshell.html?id=wXIvheS3r_gC 9

[12] http://tldp.org/LDP/abs/abs-guide.pdf 9

[13] http://linux-nfs.org/wiki/index.php/Main_Page 9

[14] http://www.orangefs.org/trac/orangefs/wiki/WikiStart 10

[15] http://www.adaptivecomputing.com/resources/docs/torque/3-0-3/ 12

[16] http://www.adaptivecomputing.com/resources/docs/torque/3-0-3/2.1jobsubmission.php 13