УДК 681.322

ВЕБ-ПОРТАЛ СИСТЕМЫ УПРАВЛЕНИЯ СУПЕРКОМПЬЮТЕРОМ

А. Л. Головинский[1], **А. Л. Маленко**[1], **Л. Ф. Белоус**[2]

На сегодняшний день доступ пользователей к вычислительным ресурсам предполагает достаточно высокие требования к знанию дополнительной массы деталей, по сути, чисто технического характера, которые необходимы специалисту в своей области, помимо свойств применяемых пакетов программ и их специфического языка. Не сильно упростило ситуацию появление гриддоступа, скорее наоборот, добавился еще один уровень сложности, обычно предполагающий умение работать в режиме командной строки в операционной среде Unix. С другой стороны, проблема администрирования вычислительного кластера остается непростой и трудоемкой задачей. Рассматриваемая работа является попыткой комплексного решения обеих проблем. Предлагается интегрированное решение в виде веб-портала для системы управления кластером, которое позволяет пользователям управлять прохождением своих задач, не обязуя их изучать многочисленные детали операционного окружения суперкомпьютера. Что касается администрирования кластера, то для этой цели разработан удобный сервис для управления кластером, включающий в себя и режим консоли. Данный проект внедрен и успешно используется в ряде ведущих суперкомпьютерных центров Украины. Статья рекомендована к печати программным комитетом Международной научной конференции “Научный сервис в сети Интернет: суперкомпьютерные центры и задачи” (http://agora.guru.ru/abrau).

Ключевые слова: суперкомпьютер, система управления, веб-портал.

Введение. Опыт использования суперкомпьютеров показывает, что существует ряд проблем, которые препятствуют широкому внедрению практики высокопроизводительных вычислений в различных областях науки и техники. Выделим некоторые из них.

Большинство суперкомпьютеров разрабатывается на основе операционной системы Linux. Это обусловливается большой гибкостью данной системы, ее надежностью, поддержкой высокопроизводительного оборудования и передовых технологий в области компьютерных архитектур. При этом при подготовке научных кадров в основном используется операционная система Windows. Этим обусловлен низкий уровень владения системой Linux среди ученых.

Стандартным интерфейсом работы с суперкомпьютером является консоль (командная строка). Консольный интерфейс позволяет максимально использовать возможности операционной системы, но сложен в изучении, требует глубоких знаний команд. Для неподготовленного пользователя суперкомпьютера консоль оказывается барьером в его освоении.

Для доступа к суперкомпьютеру из среды Windows пользователю требуется устанавливать дополнительные программы, наподобие WinSCP и Putty, для работы с удаленной консолью. В среде Linux эти возможности встроены в систему.

Набирающий популярность академический грид имеет сложный низкоуровневый интерфейс доступа. Средства для его упрощения только начинают разрабатываться.

В академической среде администрированием кластера вынуждены заниматься ученые помимо своей научной деятельности. Администрирование обычно требует ежедневных рутинных операций, которые можно и нужно автоматизировать.

Отсутствие средств самодиагностики суперкомпьютера приводит к тому, что ошибки могут жить неделями. Это сильно сказывается на качестве предоставляемых сервисов.

Часто у суперкомпьютера отсутствует средство сбора статистики использования ресурсов и автоматического построения отчетов. У руководства организации нет общей картины структуры потребления ресурсов.

[1] Институт кибернетики им. В. М. Глушкова НАН Украины, просп. Глушкова, 40, 03187, Украина, г. Киев; А. Л. Головинский, науч. сотр., e-mail: golovinsky.andriy@gmail.com; А. Л. Маленко, вед. математик, e-mail: exipilis@yandex.ru

[2] Физико-технический институт низких температур им. Б. И. Веркина НАН Украины, просп. Ленина, 47, 61103, Украина, г. Харьков; рук. сектора, e-mail: belous@ilt.kharkov.ua

Часто отсутствуют средства оперативного оповещения. Суперкомпьютер не может вовремя сообщить об ошибках, критических ситуациях, перегревах или авариях, что может привести к серьезным последствиям, вплоть до порчи электронных компонентов или пожара.

В настоящей статье рассматривается система, объединяющая интерфейс пользователя и средства администратора. Целью разработки данной системы является решение перечисленных выше проблем. Система SCMS представляет собой веб-портал и программное обеспечение промежуточного уровня, которые взаимодействуют с системным программным обеспечением суперкомпьютера. Общий вид системы показан на рис. 1.

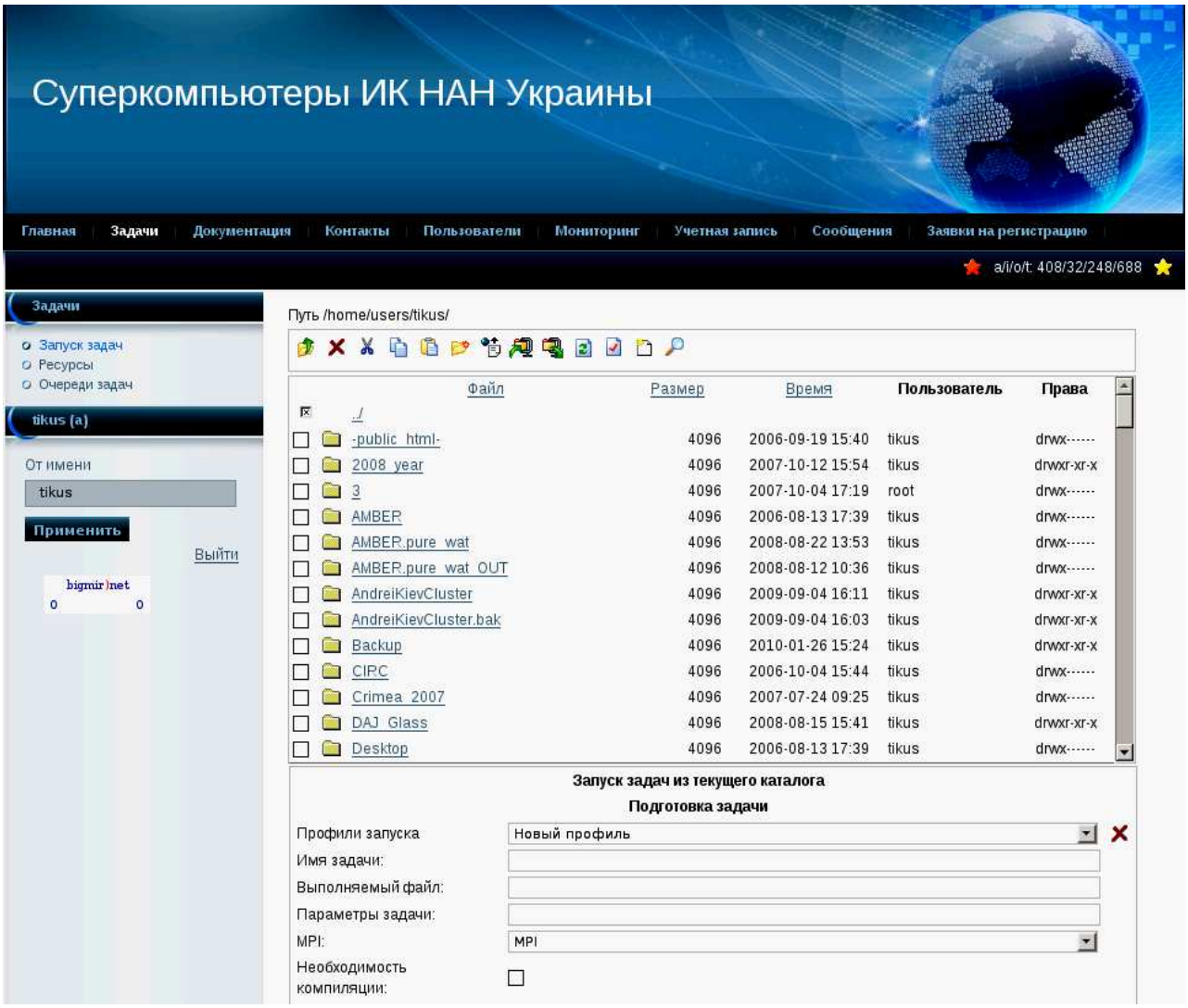

Рис. 1. Файловый менеджер и форма запуска задач веб-интерфейса системы управления суперкомпьютером

1. Основные характеристики системы SCMS. Система имеет характеристики [1], представленные ниже.

1. Система управления поддерживает установку на любой суперкомпьютер с менеджерами ресурсов Slurm, Torque или аналогичными.

2. Система имеет графический интерфейс пользователя, рассчитанный на неподготовленного пользователя. Для использования системы необходимо изучить минимум документации.

3. Рабочим инструментом является веб-браузер, который не зависит от платформы и доступен практически на любой системе. Система стабильно работает во всех современных браузерах: Internet Explorer 6.0+, FireFox 2.0+, Opera 9.0+ и Google Chrome.

4. В системе используются современные технологии Web 2.0, в частности технология асинхронных запросов к серверу Ajax [2]. Передача данных происходит по каналу OpenSSL, обычный пользователь

имеет доступ только к своим файлам и задачам.

5. Интерфейс пользователя поддерживает многоязычность. Это позволяет ученым из разных стран работать на одном суперкомпьютере.

6. Сведения о состоянии системы периодически обновляются в фоновом режиме, пока пользователь работает с другими задачами. Полная информация о состоянии кластера доступна по нескольким нажатиям кнопки мыши. Наличие ошибок проверяется беглым взглядом на экран.

7. Сообщения о критических ошибках, требующих неотложного реагирования, отправляются на e-mail или мобильный телефон. Например, это могут быть: перегрев узлов, выход из строя системы охлаждения, выход из строя жестких дисков хранилища.

8. Система обеспечивает прозрачную работу в гриде, аналогичную работе на локальном суперкомпьютере.

2. Структура и возможности системы SCMS 4.0. Ядро системы составляют скрипты взаимодействия с оборудованием кластера, менеджером ресурсов, ПО грида и т.д. Скрипты выполняют всю работу по обслуживанию запросов от интерфейса пользователя, средств мониторинга и диагностики.

Основой визуальной части системы является веб-портал. В портале размещаются новости, статьи, документация, в него интегрированы функциональные элементы интерфейса пользователя и администратора. Ниже представлена таблица возможностей интерфейсов пользователя и администратора.

Интерфейс пользователя	Интерфейс администратора
компиляция и запуск задач; работа с файлами пользователя; обмен сообщениями с другими пользователями и администраторами	управление пользователями; мониторинг кластера; диагностика; просмотр статистики использования ресурсов; уведомления о неполадках

Работа в гриде происходит точно так же, как и работа с локальным кластером.

При необходимости система может быть дополнена новыми модулями мониторинга и диагностики силами администратора.

2.1. Управление содержимым портала. Система управления порталом позволяет создавать неограниченное количество разделов портала, добавлять статьи, новости, документацию. В составе системы есть встроенный визуальный редактор, аналогичный MS Word, средства работы с изображениями и файлами.

2.2. Графический интерфейс пользователя кластера. Задача интерфейса — обеспечить выполнение всех возможных операций пользователя только средствами интерфейса. Он должен как можно полнее соответствовать пользователю, учитывать его интересы, привычки, специфику его задач.

Большинство ученых работают с готовыми программными пакетами. Им необходима удобная, не перегруженная дополнительными функциями среда для редактирования файлов входных данных, запуска параллельных программ, оперативного просмотра выходных файлов в режиме on-line. Прикладные программисты используют кластер как инструмент настройки параллельных программ, и в дополнение к вышеуказанным средствам им необходима среда для компиляции с поддержкой популярных компиляторов и прикладных библиотек, а также редактор исходных текстов программ.

Основные операции, которые выполняют пользователи комплекса:

— файловые операции;

— выполнение задач на ресурсах кластера;

— слежение за процессом решения и просмотр результатов выполнения задач;

— взаимодействие между пользователями и администраторами;

— работа в гриде.

Рассмотрим операции пользователя более детально.

Файлы пользователей находятся в персональных каталогах, защищенных от доступа посторонних лиц. Интерфейс доступа к файлам обеспечивает все традиционные операции с файлами: создание, редактирование, удаление и т.д. Дополнительно предусмотрены сортировки по полям “имя”, “размер”, “время создания”, что необходимо для удобства работы с каталогами при большом количестве файлов.

Большой объем файловой системы кластера обусловливает необходимость направленного поиска по файловой системе. В графическом интерфейсе предусмотрен поиск по регулярным выражениям и дате создания файла.

Обмен данными между рабочей станцией и кластером осуществляется путем загрузки и отправки файлов. Для эффективной работы предусмотрен также обмен каталогами, который обеспечивается их

предварительной архивацией. Редактирование текстовых файлов осуществляется с поддержкой подсветки синтаксиса популярных языков программирования.

Постановка задачи в очередь осуществляется через соответствующий интерфейс, который позволяет задать все необходимые параметры вычислительной задачи менеджеру ресурсов кластера.

Для исходных текстов программ предусмотрена интеллектуальная система компиляции, которая анализирует текст программы, выбирает соответствующий язык программирования и сценарий компиляции. Поддерживаются сценарии для различных компиляторов, например, Intel и GNU.

В интерфейсе предусмотрен режим работы с общесистемными программными пакетами (Gamess, Gromacs, Abinit и т.д.). В этом режиме некоторые параметры запуска задаются автоматически со значениями по умолчанию, что значительно упрощает работу с ними.

Главным средством контроля процесса выполнения задачи является просмотр журналов задач в реальном времени. В журналах отражается состояние выполнения и обнаруженные ошибки. Для этого поддерживается отдельный режим просмотра файлов — слежение за файлом. Для просмотра журнала выполнения задачи предусмотрена фоновая подкачка хвоста файла, аналогичная tailf. При этом все изменения в файле журнала сразу же отображаются в соответствующем окне. Дополнительными средствами контроля выступают просмотр уровня загрузки процессоров и объема занятой задачей оперативной памяти на узлах.

Для общения между пользователями и администраторами предусмотрена система передачи сообщений и форум.

После добавления грид-сертификата в свою учетную запись пользователь может работать с грид-ресурсами как с локальным суперкомпьютером. Работа с удаленной файловой системой интегрирована в файловый менеджер. Запуск грид-задач ничем принципиально не отличается от запуска задач на локальном суперкомпьютере. После выполнения задачи производится фоновое копирование результатов выполнения на локальный суперкомпьютер.

2.3. Графический интерфейс администратора кластера. Администратор выполняет функции организации вычислительного процесса суперкомпьютера. Каждый администратор имеет учетную запись, как и обычный пользователь, но с расширенными возможностями.

2.3.1. Вход в систему от имени произвольного пользователя. Необходимость в таком средстве появляется при возникновении у пользователя затруднений с работой на кластере. Вход от имени конкретного пользователя позволяет обеспечить повторяемость ошибок и локализовать их в среде, где они возникают.

2.3.2. Администрирование очередей задач. Очереди задач требуют от администратора постоянного надзора. В интерфейсе предусмотрена возможность просмотра параметров задач в очереди, отмены их в случае возникновения ошибки или по каким-либо иным причинам.

2.3.3. Просмотр состояния оборудования кластера. Состояние оборудования требует постоянного внимания со стороны администратора. Своевременное информирование об авариях является одной из основных задач интерфейса.

Подсистема мониторинга взаимодействует с менеджером ресурсов, получает информацию о состоянии узлов. Она имеет модули для контроля состояния аппаратной части и программных компонентов комплекса:

— узлов через IPMI;
— жестких дисков и RAID-массивов серверов;
— файловой системы Lustre [3];
— счетчиков ошибок сетевых коммутаторов Ethernet, Infiniband;
— получает данные о состоянии батарей источников бесперебойного питания;
— контролирует температуру узлов;
— контролирует работоспособность ПО грида.

2.3.4. Управления ресурсами кластера. Основным ресурсом кластера являются вычислительные узлы. Администратор имеет возможность динамически менять общее количество узлов, доступных для назначения задачам. Предусмотрены возможности отключения, включения, блокировки назначения узлов конкретным очередям, возможность блокировки отдельной очереди или всех очередей задач. Предусмотрена также возможность приостановки очередей задач.

2.3.5. Работа с базой учетных записей пользователей. Администратору доступен полный цикл работы с пользователями, начиная от оформления и обработки заявки на регистрацию нового пользователя, редактирования учетной записи пользователя и удаления пользователя. Поддерживаются базы пользователей на основе LDAP [4] и файла PASSWD. Аутентификация пользователей происходит с по-

мощью функций LDAP или PAM.

2.3.6. Запуск диагностических задач. Диагностические задачи — это особенный класс задач. Они позволяют определить характеристики системы, проверить надежность кластера в целом. Запуск таких задач может быть осуществлен как по расписанию, так и по требованию. Предусмотрен интеллектуальный анализ журнала диагностики с определением компонентов с пониженными характеристиками.

Средства диагностики обеспечивают проверку производительности узлов, тестирование работоспособности сети Infiniband и файловой системы Lustre (рис. 2). Специальный модуль защиты от перегрева отключает узлы, температура которых значительно превысила критическую.

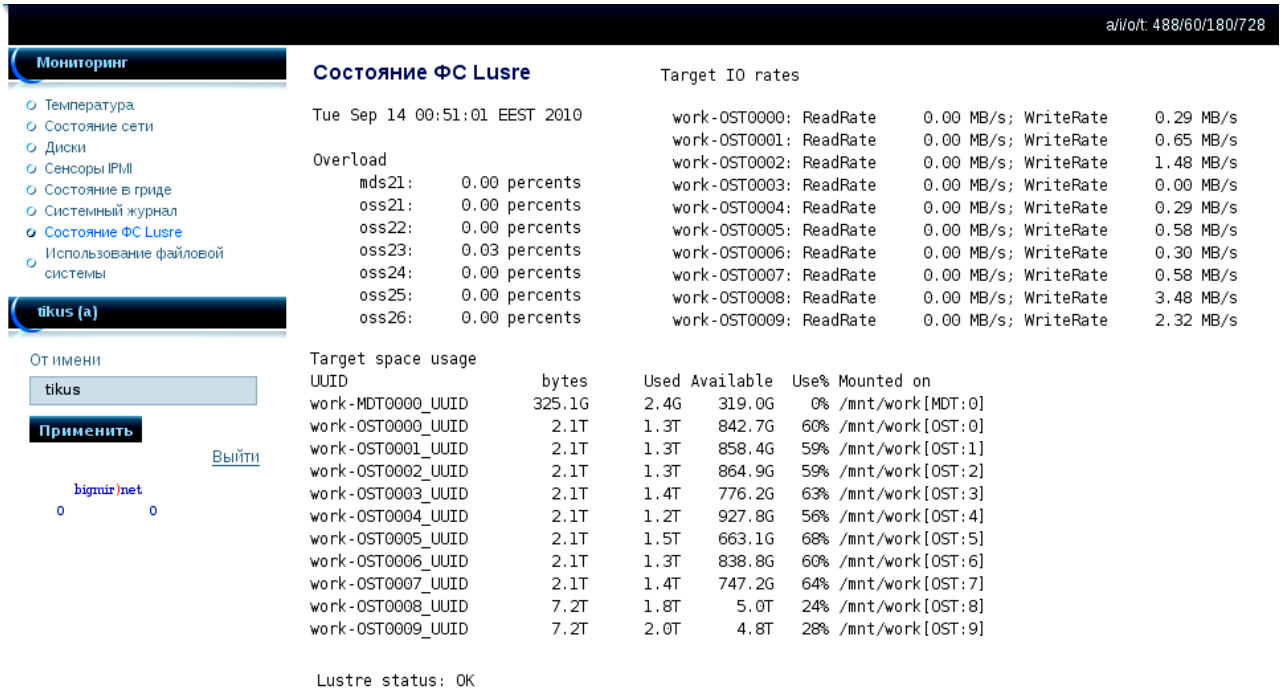

Рис. 2. Контроль состояния ФС Lustre

2.3.7. Просмотр системных журналов. Просмотр журналов компонентов кластера позволяет выявить неизвестные ранее проблемы, детектирование которых не предусмотрено в соответствующих разделах мониторинга.

Предусмотрена возможность фильтрации журналов по определенным ключевым словам, что упрощает анализ больших объемов текста.

2.3.8. Анализ статистики использования ресурсов. Система собирает информацию о выполненных задачах и с датчиков мониторинга. Статистика доступна для просмотра администратором. Есть возможность сгруппировать статистику использования ресурсов по пользователям и организациям. Статистика экспортируется в csv- или Excel-формат.

2.3.9. Уведомление об опасных ситуациях. Система уведомляет администраторов о некритических ошибках по электронной почте, а о критических — высылает SMS сообщение на мобильный телефон. Такими событиями могут быть перегрев узлов, авария системы охлаждения, отказ жестких дисков хранилища.

3. Особенности реализации системы.

3.1. Веб-портал кластерных вычислений. Графический интерфейс должен быть доступным с произвольной рабочей станции пользователя и администратора. Поэтому он должен быть кросс-платформенным и не требовать установки дополнительного программного обеспечения. Именно по этой причине как технологическая основа для реализации данного проекта избраны веб-сервисы. Они предоставляют лучшую кросс-платформенность на сегодняшний день, и их средств достаточно для обеспечения выполнения заданного круга задач.

Общий вид интерфейса показан на рис. 1. Дизайн портала и его функциональность в значительной степени могут быть произвольными и определяется на стадии проектирования под определенный суперкомпьютер.

3.2. Структура ПО интерфейса. Интерфейс имеет модульное строение и состоит из двух слоев программного обеспечения: веб-части, которая отвечает за диалог с пользователем, и ПО промежуточного уровня для взаимодействия с системным программным обеспечением суперкомпьютера (рис. 3).

Среди модулей можно выделить подсистему авторизации и работы с базой пользователей, модуль взаимодействия с менеджером ресурсов кластера и модули диагностики разных подсистем суперкомпьютера. Коротко об их основных функциях.

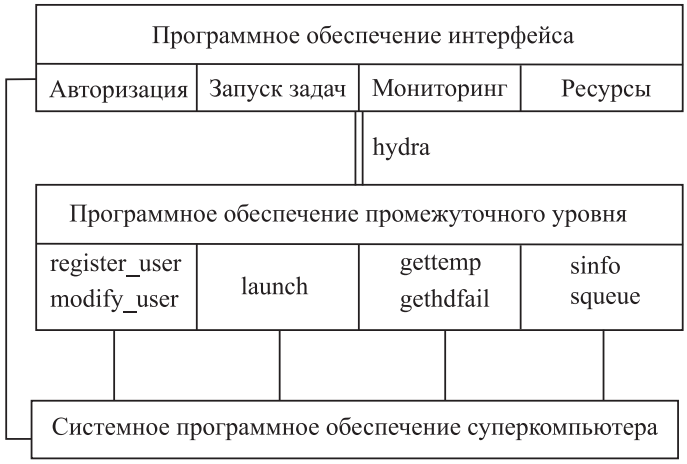

Рис. 3. Структура ПО интерфейса

3.2.1. Модуль авторизации. Поддерживается авторизация с помощью систем PAM и LDAP. Авторизация осуществляется непосредственно, без промежуточного программного обеспечения. Работа с базой пользователей включает в себя операции авторизации пользователя, добавление и удаление пользователей, редактирование их учетных записей. Для работы с LDAP в файле конфигурации указывается адрес ldap-сервиса, суффикс базы, учетные данные администратора базы. Для работы с PAM используется модуль php-auth-pam, который настраивается отдельно.

3.2.2. Модули диагностики. Коллекция тестов анализирует основные параметры работы суперкомпьютера: состояние жестких дисков, сетевых соединений и сети Infiniband, температура узлов, данные сенсоров IPMI. Скрипты диагностики и мониторинга находятся на промежуточном уровне программного обеспечения. Они выполняются от имени суперпользователя на шлюзе кластера. Выполнение осуществляет системный планировщик cron в соответствии с расписанием. Расписание устанавливается в соответствии с потребностью администраторов в предоставляемой информации.

3.2.3. Выполнение административных операций. Часть административных операций выполняется от имени суперпользователя. К таким операциям относятся редактирование базы пользователей и удаление произвольных задач из очереди менеджера ресурсов.

Для выполнения таких задач в суперкомпьютере создается служебный пользователь “суперпользователь”, который через sudo может выполнять заданные команды ПО SCMS от имени root.

3.3. Дистанцирование веб-сервера от суперкомпьютера. Высокопродуктивные файловые системы часто недостаточно надежны. Аварии в файловых системах Lustre, NFS-RDMA [5] чаще всего являются причинами остановок в работе суперкомпьютеров, поскольку ведут к фатальным сбоям в работе компонентов: серверов, узлов, в частности веб-сервера.

Для того чтобы решить проблему надежности работы веб-сервера, разработано специализированное клиент-серверное ПО. Данное ПО позволяет скриптам веб-интерфейса инициировать выполнение команд на шлюзе кластера от имени пользователей и администраторов, а также обмениваться данными. Это ПО работает в следующих режимах:

— выполнение команд пользователя и предоставление результата выполнения;

— чтение данных из файла;

— запись данных, полученных веб-сервером, в файл на файловой системе кластера (реверсный режим передачи данных);

— выполнение команд администратора от имени суперпользователя осуществляется через служебного суперпользователя, который выполняет команду из списка разрешенных.

Таким образом, даже в случае значительной аварии, веб-сервер будет работать и пользователи смогут получить информацию о причине проблемы, сроке ее устранения и т.д.

3.4. Запуск задач с помощью менеджера ресурсов кластера. Запуск задачи на выполнение осуществляется через соответствующую форму интерфейса. На показанном ранее рис. 1 выбрано меню запуска задач с этой формой. Пользователь задает параметры задачи и название файла с программой, параметры командной строки, количество процессоров, время выполнения, выбирает MPI-среду. При необходимости указывается компиляция программы из исходных текстов. В таком случае автоматически создается сценарий компиляции программы из ее исходных текстов, автоматически запускается компиляция, а после ее успешного завершения начинает выполняться программа.

Единожды выполненные действия по заполнению формы запоминаются в виде профиля запуска с

уникальным именем, так что при повторении этого действия достаточно только указать имя профиля (возможно дополнительно отредактировать) и нажать кнопку “запустить”. Через некоторое время создается своего рода библиотека этих профилей, отражающая специфику повторяющихся действий пользователя. С другой стороны, для неподготовленных пользователей такая библиотека может быть заранее составлена системным администратором, что фактически позволит специалистам в своей области вообще не знакомиться с деталями запуска своего пакета, а сосредоточиться только на специфике языка пакета и поведения задачи.

Структура запуска вычислительной задачи показана на рис. 4. Запуск производится через соответствующие модули графического интерфейса. Пользователь задает параметры задачи на странице “Запуск задачи”. Все данные передаются модулю launch, который производит запуск через менеджер ресурсов кластера. Тип менеджера задается в файле конфигурации интерфейса. Дальше задача становится в очередь и после предоставления ей ресурсов начинает выполняться на вычислительных узлах.

3.5. Интерактивное взаимодействие с пользователем. Некоторые компоненты интерфейса отображают информацию, которая часто изменяется. К ним принадлежат модули ресурсов, очереди задач, просмотра файла вывода задачи, которая выполняется. Интерфейс осуществляет обновление информации без необходимости перезагрузки страницы с помощью технологии AJAX. Такой подход позволяет пользователям осуществлять интерактивную работу со своими вычислительными задачами на кластере, что очень важно для многих исследований в областях физики и химии с не полностью формализованными алгоритмами.

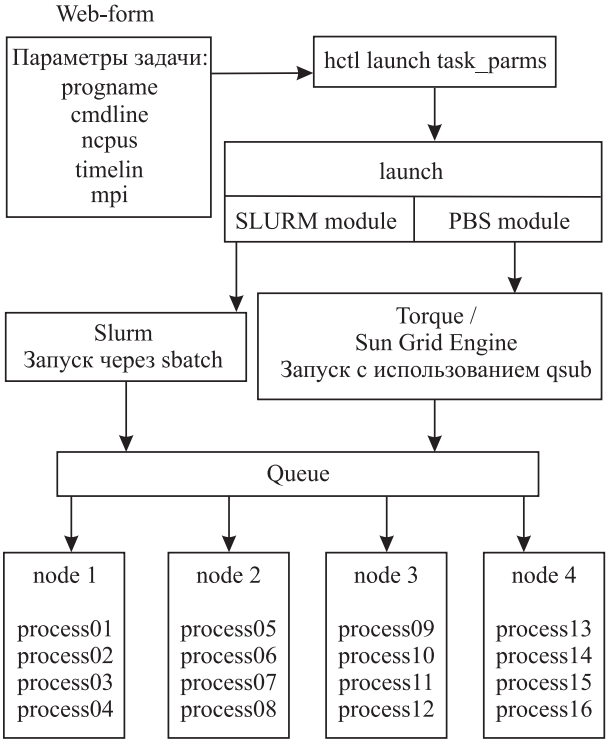

Рис. 4. Структура запуска задачи

4. Заключение. В настоящей статье описана программная система, которая обеспечивает вебинтерфейс системы управления суперкомпьютером для пользователей и администраторов.

С нашей точки зрения, описанная среда способствует более широкому использованию отечественных многопроцессорных вычислительных систем, поскольку сильно упрощает их использование учеными и программистами.

Данный проект внедрен и успешно используется на грид-кластерах Института кибернетики им. В. М. Глушкова НАН Украины, г. Киев [6], в Физико-техническом институте низких температур им. Б. И. Веркина НАН Украины, г. Харьков [7], а также в ряде других академических учреждений, входящих в состав Украинского Академического Грида (УАГ) [8]. Система постоянно улучшается за счет тесного взаимодействия с пользователями суперкомпьютеров.

СПИСОК ЛИТЕРАТУРЫ

1. *Якуба А.А., Головинский А.Л., Бандура А.Ю., Горенко С.А., Ефременюк Д.А.* Портал кластерных вычислений для управления вычислительными процессами на суперкомпьютерном комплексе // Кибернетика и системный анализ. 2009. **6**. 97–105.
2. http://en.wikipedia.org/wiki/Ajax_(programming)/
3. http://www.lustre.org/
4. http://ru.wikipedia.org/wiki/LDAP
5. http://nfs-rdma.sourceforge.net/
6. http://icybcluster.org.ua
7. http://cluster.ilt.kharkov.ua
8. http://lcg.bitp.kiev.ua

Поступила в редакцию
28.10.2010