ПОДХОД К ГИБКОМУ УПРАВЛЕНИЮ СУПЕРКОМПЬЮТЕРАМИ

С.А. Жуматий, Д.А. Никитенко

Современный суперкомпьютер, высшего сегмента производительности [1,2] или небольшой, сильно отличается в управлении от обычного персонального. Тут и большое число компонент, и нестандартные процедуры управления, и пользователей, как правило, не меньше десятка, и много других аспектов.

Более того, как раз эти «другие аспекты» будут сильно отличаться для разных организаций и даже разных суперкомпьютеров в одной организации, при этом являясь крайне важными — это могут быть дисковые квоты, лицензии на использование коммерческого ПО, процедуры регистрации пользователей и т.п. Как организовать управление данными, необходимыми для контроля над суперкомпьютером, и как встроить их в процесс поддержки суперкомпьютера? Этим вопросом авторы задались несколько лет назад, но однозначного ответа не нашли, несмотря на большой опыт в эксплуатации суперкомпьютеров [3-5].

Все применяемые ранее средства давали однобокий результат, и приспособить их к решению большинства типичных задач администратора оказывалось крайне трудно или просто невозможно[6,7].

Как это выглядит на практике? Имеется набор скриптов для выполнения стандартных процедур, а необходимая информация о ПО, оборудовании, сбоях, пользователях и квотах хранится отдельно в таблицах, текстовых документах, базах данных и т.п. При этом важно, что зачастую синхронизация и введение новых данных проходит почти полностью вручную. Как показало общение в рамках ежегодных международных суперкомпьютерных конференций серий ISC (Германия) и SC (США), такая ситуация характерна не только для отечественных суперкомпьютерных центров, но и для многих зарубежных. Приведем один из типичных примеров. Информация о пользователях и о текущих проблемных ситуациях с ПО или оборудованием хранится в, вообще говоря, не предназначенной для этого адаптированной help desk системе OTRS [8]. Для ведения статистики по используемым пакетам ПО, пользователям и т.п. — электронные таблицы. Общение с пользователями — электронная почта. Списки рассылки — опять-таки, отдельные таблицы. Дополняет картину набор скриптов у администратора, каждый из которых выполняет какую-то процедуру. В итоге получается достаточно развесистая система, которую сложно синхронизировать, а время администраторов, затрачиваемое на подобные рутины, становится непозволительно большим. Казалось бы, можно один раз связать все эти технологии вместе, автоматизировать процесс, но проблема состоит в том, что появляются новые процедуры, новые типы данных, новые «поля» и т.п. В итоге, регулярная доработка под текущие нужды станет постоянным занятием, а в какой-то момент может потребовать очередной переделки «с нуля». В любом случае, как только система получается достаточно сложной, становится ясно, что требуется или связать воедино используемые отдельные инструменты для их синхронизации, или иметь отдельный инструмент, реализующий весь требуемый функционал.

Типичные подходы предполагают либо полную интеграцию с системным ПО суперкомпьютера, что плохо соотносится с уже работающими системами, либо использование ПО, хранящего информацию независимо, но не способного синхронизировать её с суперкомпьютером. Во обоих подходах подавляющее число рассмотренных авторами вариантов ПО не поддерживает или очень ограниченно поддерживает настройку под специфичные нужды конкретного кластера (точнее, его руководства и администратора).

На основании накопленного авторами опыта по поддержке работы суперкомпьютерного центра Московского университета в НИВЦ МГУ был разработан и предложен подход для создания специализированного ПО, позволяющего решить эти и смежные проблемы. С использованием предложенного подхода силами компании Evrone и НИВЦ МГУ осуществлена программная реализация и успешно внедрён прототип системы для управления суперкомпьютерами, реализованный на базе технологии Ruby [9].

Назначение системы — административная поддержка полного цикла деятельности пользователей и администраторов суперкомпьютерных центров путем предоставления специализированного веб-сервиса.

Основные задачи системы, на решение которых была направлена разработка:

- построение эффективного взаимодействия пользователей и службы поддержки пользователей вычислительных систем;
- обеспечение возможности автоматизированного сбора статистики по текущему составу и динамике выполняемых проектов и проведение анализа этих данных;
- предоставление результатов работы внешних систем сбора статистики и анализа хода выполнения задач на вычислительных системах и интеграция с ними;
- упрощение прохождения административных процедур для пользователей вычислительных центров: регистрация, получение допусков и квот, возможно, еще какие-либо формальности, обусловленные регламентом предоставления услуг по использованию суперкомпьютера;
- снижение нагрузки на администраторов с помощью автоматизации рутинных процессов по управлению учетными записями и связанными задачами;

- минимальное вмешательство в существующее системное ПО. Система не должна заменять ключевые компоненты суперкомпьютера.

Для решения поставленных задач предложено несколько подходов. Первый подход — абстрагирование системы управления от конкретного суперкомпьютера с помощью стандартного набора скриптов. На суперкомпьютере создаётся выделенный пользователь, которому разрешается запуск скриптов из отдельного каталога (недоступного на запись) от имени суперпользователя. Заходя от имени этого пользователя удалённо по протоколу ssh, система управления может выполнять административные задачи — создание и удаление пользователей, изменение квот и т.п. Набор скриптов может быть при необходимости адаптирован для каждого суперкомпьютера отдельно. Например, вместо удаления пользователя может быть выполнено его архивирование и т.п. Такой подход позволяет быстро адаптировать работу системы для любого системного ПО, установленного на суперкомпьютере.

Следующий подход — переход от модели «один человек — один аккаунт на суперкомпьютере» к проектной модели. Вводится понятие проекта, в рамках которого могут работать один или несколько пользователей. В зависимости от потребностей администратора проекту может соответствовать один аккаунт под которым могут работать несколько человек (традиционно используется на большинстве систем), либо проекту соответствует группа, а каждому члену проекта выделяется собственный аккаунт. Последний вариант позволяет гибко управлять квотами и блокировками. Например, если один из участников проекта нарушает правила, то можно ограничить доступ только ему, не затронув остальных участников проекта. Кроме того, упрощается сбор статистических данных — можно точно узнать, сколько ресурсов было потрачено на каждый проект и сколько — каждым пользователем по всем его проектам.

Следующий подход — введение пользовательских типов данных и операций над ними. В качестве пользователя здесь выступает администратор суперкомпьютера. Такой подход позволяет описать объекты учёта, которые могут сильно отличаться в различных системах — программное обеспечение, единицы вычислительного оборудования, запросы в сервис на обслуживание техники и т.п. Это позволит, например, пометить сбойный вычислительный узел, что автоматически выполнит его блокировку в системе управления задачами, отправит запрос в сервис-центр через web-форму и сделает отметку в журнале. В дальнейшем администратор сможет быстро просмотреть список сбойных узлов, или список узлов, которые не вернулись из сервис-центра в течении месяца или более и т.п. При большом числе единиц техники такая задача становится нетривиальной.

На сегодняшний день в работающем прототипе реализованы и успешно работают два первых подхода. Прототип обслуживает суперкомпьютерный комплекс Московского университета. На данный момент это суперкомпьютеры «Чебышёв» (5000 ядер) и «Ломоносов» (50 000 ядер), несколько сотен пользователей и проектов. Пользователи могут самостоятельно создавать проекты, заказывать вычислительные ресурсы, приглашать коллег для работы в своих проектах. После одобрения администратором запросов автоматически создаются аккаунты, добавляются ssh-ключи для доступа. При необходимости пользователь самостоятельно может заменить или добавить ssh-ключ, не обращаясь к администратору.

В рамках системы реализовано общение пользователей с технической поддержкой суперкомпьютера, что позволяет, в отличие от электронной почты, упростить отслеживание истории решения проблем, облегчить совместную работу нескольких сотрудников поддержки, отследить классы проблем и многое другое. Пример рабочего кабинета администратора в данном разделе представлен на Рис.1.

Через интерфейс администратора доступны следующие разделы:

- **Пользователи**
- Вся информация о пользователях: контактные данные, история событий с учетной записью, списки ключей, вхождение в проекты и т.п.
- **Отчеты**
  Описание проектов, представленные в ходе ежегодной перерегистрации
- **Поручительства**
  Документы, подтверждающие ответственность за использование ресурсов
- **Заявки**
  Ходатайства о выделении ресурсов для работ над проектами
- **Проекты**
  Вся информация о проектах: выделенные ресурсы, исполнители..
- **Поддержка**
  Интерфейс общения с пользователями, поддерживающий рубрикацию и шаблоны
- **Консоль**
  Контроль выполнения скриптов системы на стороне кластера
- **Прочее**
  Всевозможные настройки: списки рассылок, опросы и т.д.

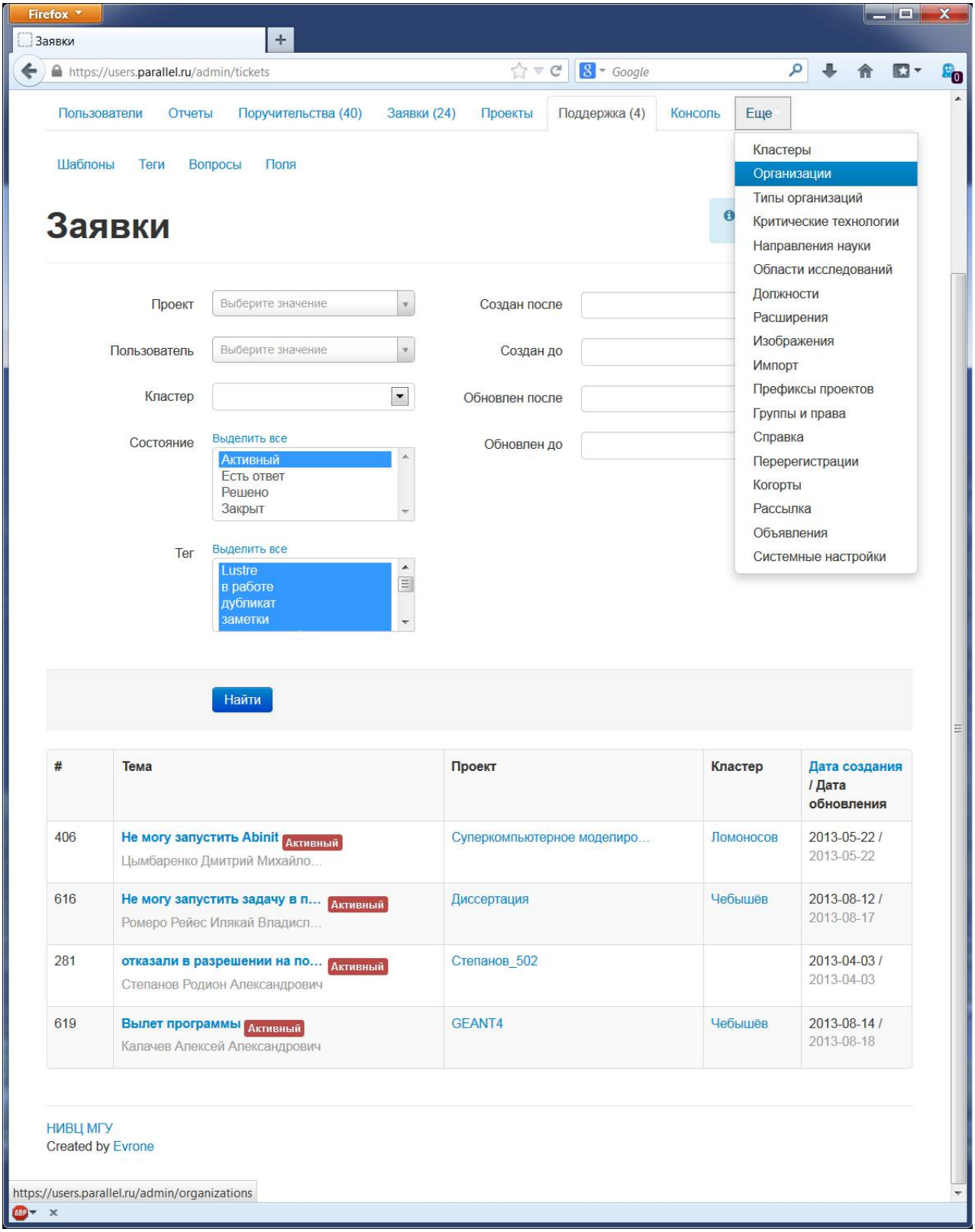

Рис.1. Пример рабочего интерфейса администратора в разделе «Поддержка»

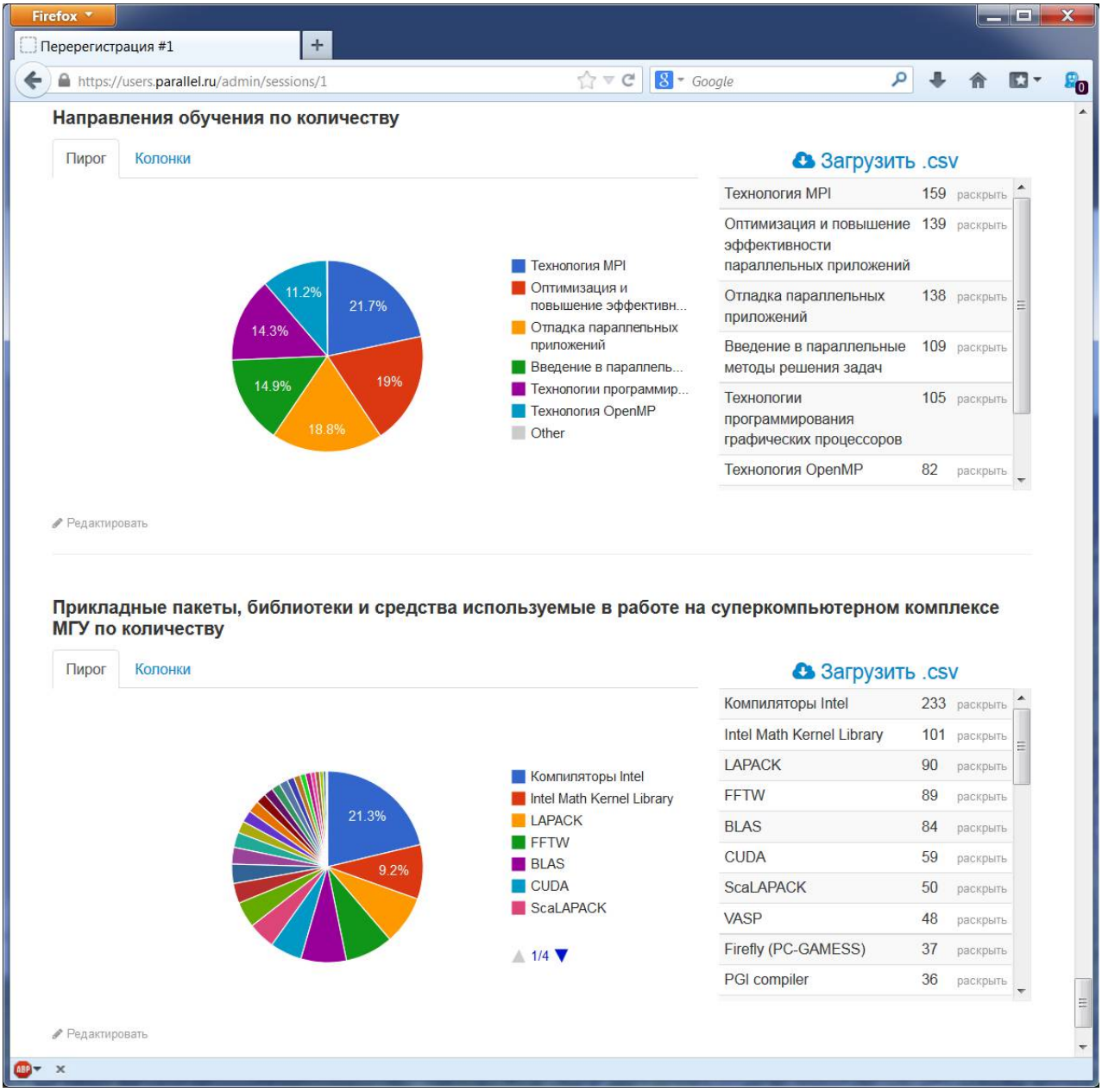

Рис.2. Пример автоматически собираемой статистики на основании результатов перерегистрации (данные получены в ходе тестового режима эксплуатации)

Следует обратить внимание на широкий спектр доступных статистических данных, получаемых в результате проведения ежегодной перерегистрации пользователей. Данная процедура проводится в суперкомпьютерном центре Московского университета в начале каждого календарного года с целью подведения итогов года прошедшего и проведения необходимой корректировки квот, политики доступа и т.п. для поддержания эффективности работы суперкомпьютерного комплекса в целом. На основании результатов проводимых в ходе перерегистрации опросов, руководство может принимать решение о целесообразности и частоте проведения тех или иных мастер-классов (Рис.2), необходимости получения тех или иных лицензий на ПО, да и вообще — лучше понимать весь комплекс потребностей пользователей суперкомпьютерных ресурсов.

Здесь важно отметить изначально заложенную идею широкой интеграции с внешними источниками данных. Таковыми могут быть, например, анализаторы динамических характеристик программ, интеграция с которыми может позволить более эффективно распределять вычислительные ресурсы [10-12]. На текущий момент данный функционал не реализован, но имеющийся у авторов задел позволяет рассчитывать на постоянное расширение функциональных возможностей разработанного комплекса.

Проект развивается и открыт как для установки, так и для совместного развития. Выражаем глубокую благодарность всем пользователям суперкомпьютерного комплекса МГУ и членам экспертной группы, принявшим участие в апробации системы, высказавшим множество ценных замечаний и предложений. Мы

приветствуем любые предложения и отзывы, это крайне важно для развития проекта. Наша цель — помочь всем, кто поддерживает суперкомпьютеры и всем, кто их использует.

Работы проводятся при частичной поддержке РФФИ (Грант №13-07-00750 А).

ЛИТЕРАТУРА:

1. А.С. Антонов, Д.А. Никитенко, С.И. Соболев. 18-я редакция списка Top50 самых мощных компьютеров России: ожидания и перспективы// Параллельные вычислительные технологии (ПаВТ'2013): труды международной научной конференции (г.Челябинск, 1-5 апреля 2013 г.). Челябинск: Издательский центр ЮУрГУ, 2013. С. 258-260.
2. Top500 Supercomputer Sites http://top500.org
3. Воеводин Вл.В., Жуматий С.А., Соболев С.И., Антонов А.С., Брызгалов П.А., Никитенко Д.А., Стефанов К.С., Воеводин Вад.В. Практика суперкомпьютера "Ломоносов" // Открытые системы. - М.: Издательский дом "Открытые системы", 2012. - 7: ISSN 1028-7493.
4. Суперкомпьютерный комплекс МГУ, http://parallel.ru/cluster
5. Moscow University Supercomputing Center, http://hpc.msu.ru
6. Воеводин Вл.В., Жуматий С.А. «Вычислительное дело и кластерные системы».-М.: Изд-во МГУ, 2007. - 150 с.
7. Воеводин Вл.В., Жуматий С.А. Экспонента суперкомпьютерных центров.// Открытые системы. - М.: Издательский дом "Открытые системы", 2008.- 5.- с. 12–15
8. OTRS (Open Ticket Request System), http://otrs.com
9. Ruby Programming Language, http://ruby-lang.org
10. Adinets A.V., Bryzgalov P.A., Voevodin Vad.V., Zhumatii S.A., Nikitenko D.A., Stefanov K.S. Job Digest: an approach to dynamic analysis of job characteristics on supercomputers // Numerical methods and programming: Advanced Computing.- 2012.- V. 13.- Section 2. Pages. 160-166 (http://num-meth.srcc.msu.ru/).
11. Адинец А.В, Брызгалов П.А, Воеводин Вад.В, Жуматий С.А, Никитенко Д.А. Мониторинг, анализ и визуализация потока заданий на кластерной системе// Высокопроизводительные параллельные вычисления на кластерных системах: Материалы XI Всероссийской конференции (Нижний Новгород, 1–3 ноября 2011 г.).- Нижний Новгород: Издательство Нижегородского госуниверситета.- 2011.- с. 10-14
12. Bernd Mohr, Vladimir Voevodin, Judit Giménez, Erik Hagersten, Andreas Knüpfer, Dmitry A. Nikitenko, Mats Nilsson, Harald Servat, Aamer Shah, Felix Wolf, and Ilya Zhujov. The HOPSA Workflow and Tools. Proceedings of the 6th International Parallel Tools Workshop, Stuttgart, September 2012, Springer (to appear).