Octoshell: система для администрирования больших суперкомпьютерных комплексов

Вл.В. Воеводин, С.А. Жуматий, Д.А. Никитенко

Московский государственный университет имени М.В.Ломоносова

Управление современными суперкомпьютерными центрами и входящими в их состав вычислительными системами представляет собой сложный и комплексный процесс. Традиционное использование многочисленных инструментов для решения отдельных задач по управлению и администрированию суперкомпьютеров становится ограничивающим фактором эффективного использования вычислительных ресурсов при растущих масштабах систем. Разработанная система поддержки работы суперкомпьютерных центров «Octoshell» призвана решить указанную проблему, реализуя в едином интерфейсе основные инструменты администрирования, и позволяет в значительной мере автоматизировать выполнение типовых задач обеспечения эффективного функционирования больших суперкомпьютерных комплексов.

1. Введение

Современный суперкомпьютер представляет собой большой и сложный вычислительный комплекс, включающий в себя десятки тысяч компонент. Их число даже у рядовых систем уже измеряется десятками тысяч [1], при этом у наиболее крупных систем оно насчитывает сотни тысяч, а в некоторых случаях — миллионы [2]. Растет число вычислительных узлов, число вычислительных ядер процессоров и ускорителей и т.п. Большинство специалистов сходятся во мнении, что эта тенденция продолжится и далее.

В рамках суперкомпьютерного центра может функционировать сразу несколько таких систем, и, что очень важно, в работу включается не только управление непосредственно вычислителями, но и множество других сопутствующих процессов. Например, управление лицензионным программным обеспечением, квотами пользователей, техническая поддержка, отслеживание гарантийного и послегарантийного ремонта, и многое другое.

Все эти сущности связаны между собой, и, учитывая количество всех составляющих, очень трудно эффективно организовать рабочие процессы. Помимо значительного числа аппаратных компонент, немаловажный вклад имеют и программные — системное программное обеспечение, оптимизированные библиотеки и программы для высокопроизводительных вычислений, прикладные пакеты. Сами пользователи суперкомпьютера являются неотъемлемой и важнейшей его частью, имея свои особенности, требования и предпочтения, квоты и привилегии, набор используемого ПО, историю использования суперкомпьютера и т. п.

Рассмотрим в качестве примеров некоторые задачи, с которыми приходится сталкиваться каждый день при управлении большой вычислительной системой.

Пример 1. Управление лицензионным ПО. По каждому пакету ПО, а на каждом суперкомпьютере их может быть установлено десятки, необходимо контролировать круг пользователей, которым разрешено его использование, сроки лицензии, обновления, контакты с технической поддержкой производителя. На первый взгляд, всё просто, но только до тех пор, пока вдруг не окажется, что срок действия лицензии истекает, бюджет на закупку не предусмотрен, и т.д., а пользователям надо считать... Или пока не поступит заявка на пакет от пользователя с логином «abc123», и не возникнет вопрос, распространяются ли условия лицензии на этого пользователя? Чем он занимается? Где работает? И вообще, где условия лицензий и контакты? Зачастую — в почте, «в какой-то Excel таблице». Это перестает быть мелочами, если вспомнить о стоимости лицензий, большом числе пользователей и требуемых компонент ПО. Упорядоченность, полноценный учет становятся вдвойне актуальными, когда встают вопросы о необходимости продления лицензии на тот или иной пакет ПО. Для этого необходимо понимать, насколько он был реально востребован, использовался он одной рабочей группой или несколькими, какие проекты от него зависят?

Пример 2. В состав центра входят тысячи единиц техники (вычислительные узлы, коммутаторы, диски, и многое другое), которые не могут не ломаться. Сколько времени уходит на типичный ремонт одного узла? Сколько процессорочасов было потеряно за текущий год по причине ремонта? По какой причине выключен конкретный узел? Просто выключен, отдан в ремонт или уже вернулся из ремонта, но не введен в счетное поле? Эти вопросы незаметны и незначительны для одной лаборатории, офиса, а для большого центра становятся критичными.

Есть и множество иных задач, которые требуют комплексных знаний о составе центра, структуре процессов, происходящих внутри него. От эффективности решения всех этих задач напрямую зависит эффективность работы центра в целом.

Для того чтобы добиться этого, необходимо чётко описать все операции, которые выполняются в центре, формализовать их, описать процессы. Несмотря на актуальность и значительную предысторию, предметная область описана в недостаточной мере. Отсюда вытекала необходимость создания модели работы СКЦ. Составление модели работы суперкомпьютерного центра может помочь изменить и его работу, так как можно найти более эффективные способы организации внутренних процессов. Например, более эффективно построить систему квот и приоритетов. В традиционном подходе квота или допуск к ПО выделяется отдельному человеку, в то время как он может заниматься разными проектами совместно с другими коллективами в рамках нескольких проектов. Поэтому рациональнее выделять квоты и допуски не на людей, а на проект.

2. Основные объекты и решаемые задачи в рамках работы СКЦ

Попробуем рассмотреть эти вопросы подробнее, опираясь на опыт администрирования суперкомпьютерного центра МГУ — одного из крупнейших в России [3]. Центр обслуживает несколько сотен рабочих групп, объединяя более 2500 пользователей. В состав центра входят такие суперкомпьютеры как «Ломоносов», «Ломоносов-2», «Чебышёв». Общее число учетных единиц техники составляет десятки тысяч.

Из чего складывается деятельность суперкомпьютерного центра? Что должно найти отражение в «скелете» модели, какие основные сущности необходимо увязать в единый рабочий процесс?

- Прежде всего, пользователи — специалисты разного уровня подготовки, у которых есть потребность в использовании тех или иных вычислительных ресурсов и ПО.
- Это проекты — реальные прикладные задачи, которые решаются пользователями самостоятельно или в составе рабочих групп.
- Это аппаратура. Те самые разнородные вычислительные ресурсы, необходимые для решения прикладных задач. Вычислительные системы, выделенные их разделы, например, по принципу однородности ресурсов.
- Это системное и прикладное программное обеспечение. Множество прикладных пакетов и библиотек, необходимых для эффективного решения вычислительных задач.
- Механизмы организации работы СКЦ.

Каждая из упомянутых выше компонент обладает массой свойств и особенностей. Причем важно понимать, что в динамике свойства даже уже описанных объектов могут претерпевать изменения. Могут добавляться новые свойства, новые взаимосвязи. Это необходимо учитывать наряду с возможно нетривиальной взаимосвязью этих компонент, отражающих разного рода зависимости [4].

В условиях конкуренции за вычислительные ресурсы, необходимо выстраивать систему приоритетов доступа к ним, а также квотирование. Наиболее рациональным представляется подход, при котором ресурсы выделяются на решение определенной задачи (проект), и не важно, каким числом пользователей она решается. Кроме того, не стоит забывать, что один и тот же пользователь может входить в состав различных рабочих групп и участвовать в решении совершенно не связанных друг с другом задач, играя в их решении принципиально разные роли. В одном случае человек может быть руководителем работ, например, студенческого практикума, когда большинство программ запускается студентами. При этом одновременно он мо-

жет быть единственным исследователем в рамках другого проекта, а в третьем проекте — уже выступать в роли исполнителя наряду с другими коллегами.

Что касается используемого программного обеспечения, то крайне важен уже упомянутый вопрос лицензий. На разных системах может быть установлен различный набор прикладных пакетов или их версий, могут использоваться лицензии с отличающимися условиями. При этом к одному и тому же пользователю могут относиться совершенно разные условия лицензий в зависимости от того, в рамках какого проекта ведется работа.

Помимо учета лицензий необходимо отслеживать статус самого оборудования, чтобы минимизировать время его простоя, повышая потенциальную отдачу от вычислителя в целом.

Отдельно необходимо сказать о процедурах и процессах, непосредственно не связанных с вычислительным этапом работ. Речь идет о комплексе административных процедур и поддержке работы пользователей.

Работа каждого пользователя начинается с заведения учетной записи, при этом учетных записей может быть несколько. Например, отдельная запись пользователя может использоваться для работы в рамках своего проекта. Если в рамках проекта используются ресурсы разных суперкомпьютеров центра, то учетные записи должны создаваться на каждой машине, при этом должна сохраняться информация для идентификации, какие учетные записи используются для работы над какими проектами.

В работе вычислительного центра всегда возникает необходимость связаться с пользователями. Будь то один конкретный человек, участники определенного проекта или, например, сотрудники одной организации. Для адресной связи с пользователем требуются контактные данные, и их требуется поддерживать актуальными. Важно, что все это должно идти в соответствии с действующим законодательством относительно использования персональных данных.

Часто у самих пользователей возникают вопросы, связанные с работой на вычислительной системе. Для корректного описания возникающих проблем администраторам зачастую необходимо «видеть» условия, в которых у пользователя она возникла. Это значит, что от пользователя требуется описание условий возникновения — командная строка, выдачи ошибок, скриншоты, элементы кода и т.п. Представление этих данных через стандартизованный интерфейс, наличие элементов «ситуационного экрана», всех данных о пользователе, его учетных записях и истории обращений в распоряжении администратора кардинально сказываются на эффективности оказания помощи пользователям.

Если для небольшого кластера еще может сработать схема общения по электронной почте и телефону, то для больших систем это становится неприемлемым. Использование же готовой системы поддержки, подразумевает глубокую ее интеграцию с другими инструментами администрирования, что на практике достаточно сложно осуществимо.

Создание классического раздела с частыми вопросами, что вообще говоря, само по себе является шагом необходимым, не всегда решает даже уже хорошо известные ранее встречавшиеся проблемы: к сожалению, многие пользователи предпочитают миновать шаг чтения инструкций или правил и обращаться по каждому поводу непосредственно к администраторам суперкомпьютеров. Это приводит как к увеличению интенсивности самих обращений, так и снижению эффективности самой поддержки: тратится намного больше времени на разбор даже простых ситуаций.

К административным же процедурам может добавиться система проведения экспертизы работы пользователей с целью возможной корректировки их квот и приоритетов.

Изначальное предоставление доступа к вычислительным ресурсам может предполагать создание поручительств и гарантийных писем от организаций, гарантирующих использование вычислительных ресурсов пользователем по назначению.

Все это требуется максимально автоматизировать для ускорения предоставления доступа, минимизации ошибок, повышения эффективной отдачи от суперкомпьютерного центра в целом.

3. Традиционные подходы к управлению СКЦ

Традиционно для организации работы суперкомпьютерных центров используется комбинация систем мониторинга, управления конфигурациями, управления учётными записями, средств организации поддержки пользователей, удалённого запуска программ.

Недостаток такого подхода заключается в том, что решение практически любой частной задачи, например, заведение новой учётной записи или предоставление выделенного времени счёта для некоторой группы пользователей, ведёт к необходимости выполнения множества действий с разрозненными приложениями, потере времени администратора на избыточные действия. Это, в свою очередь, ведёт к ошибкам, потере информации, увеличении времени реакции.

Ситуация усугубляется, когда центр поддерживает несколько суперкомпьютеров, так как вести учёт нескольких систем одновременно не всегда становится возможным и приходится дублировать множество информации.

На данный момент популярными средствами, применяемыми в различных суперкомпьютерных центрах, являются [5]:

а) Средства управления пользователями:
   1) LDAP и оболочки для управления им;
   2) NIS;
   3) batch-системы (torque [6], slurm [7], openpbs [8], maui, lsf, cleo и другие);
   4) ручной учёт в MS Excel, OO calc и аналогичных пакетах.

б) Средства мониторинга оборудования и сервисов: Ganglia [9], Zabbix [10], Nagios [11].

в) Средства удалённого управления: ssh, pdsh; ipmi, ilo; snmp.

г) Средства организации общения с пользователями:
   1) Ticket Management Systems (OTRS [12], Trak, RT и другие);
   2) mailman;
   3) wiki;
   4) форумы.

Использование всех этих средств без интеграции друг с другом приводит к множеству ошибок, несогласованности, дублированию информации. Это связано с тем, что невозможно провести интеграцию подобных средств без чёткого понимания состава и взаимосвязи того, что собственно необходимо учитывать, чем и как управлять.

На основании опыта администрирования и поддержки работы суперкомпьютерного центра МГУ, а также на основании многочисленных обсуждений с административным и техническим составом других суперкомпьютерных центров, включая крупнейшие европейские центры, можно сформулировать следующую проблему.

Существуют инструменты, позволяющие решать отдельные задачи для обеспечения функционирования суперкомпьютерного центра, но при увеличении масштабов вычислительных систем, роста числа обслуживаемых пользователей и т.п. число ошибок несогласования, возникающих при использовании отдельных инструментов, значительно возрастает, что существенно сказывается на отдаче вычислительных комплексов в целом.

4. Octoshell

Ежедневное решение подобных задач в рамках СКЦ МГУ привело к пониманию создания системы, интегрирующей все основные процессы управления и администрирования суперкомпьютерным центром в единый комплекс программных средств. Система получила название Octoshell, т.к. может рассматриваться как оболочка, объединяющая целый ряд отдельных инструментов/модулей. О некоторых деталях реализации будет сказано далее, сейчас же приведем перечень основных сущностей, которыми оперирует разработанная система.

4.1 Основные сущности

4.1.1 Пользователь

Пользователи — те самые реальные исследователи, для которых и существуют высокопроизводительные вычислительные системы. В рамках каждой вычислительной системы СКЦ каждый пользователь может входить в один или несколько проектов. Он может выступать в качестве участника или руководителя проекта. Участие пользователя в проекте должно быть отражено в виде учётной записи на вычислительной системе. Пользователь обязан принадлежать одной или более организаций. Дополнительно пользователь может выступать в роли эксперта и/или администратора.

Каждый пользователь изначально обязан пройти процедуру регистрации для начала работы. В процессе регистрации он обязан указать необходимые персональные данные, включая контактные данные для экстренной связи, и в дальнейшем он должен поддерживать актуальность этих данных.

Пользователь должен иметь доступ к данным статистики, касающимся его проектов или учётных записей, а также к информации о программном обеспечении, текущим или планирующимся процессам обслуживания систем, если они нарушают обычный порядок работы. Важной возможностью, предоставляемой пользователю, является техническая поддержка, с помощью которой он может получить консультации по работе, сообщить о сбое, возникшей проблеме и т.п.

В отличие от обычного пользователя, администратор имеет возможность управления параметрами системы. Возможно существование администраторов с ограниченными правами, например для управления резервным копированием, для создания новых пользователей и т.п. Администраторы имеют доступ к просмотру и изменению данных других объектов в системе. Для администраторов с ограниченными правами могут быть доступны только необходимые для них объекты.

Администратор, в отличие от обычного пользователя, имеет доступ не только к собственным данным определённого типа (например, проекту), а к данным всех пользователей. Важные дополнительные возможности, которые появляются у администратора в дополнении к возможностям пользователя: построение и просмотр отчётов, управление программным обеспечением, квотами и приоритетами, а также управление процессами обслуживания и сопровождения аппаратных компонент.

Наиболее типичные виды администраторов:

- суперадминистратор, обладающий полными правами на управление объектами информационной системы;
- администратор с ограниченными правами — для сопровождения документооборота, решения простых задач пользователей и других типовых работ;
- эксперт, наделенный правами видеть регистрационные данные пользователей, описания проектов и отчеты по проделанной в их рамках работе, включая их историю, вести анонимную переписку с пользователем с целью доработки пользователями отчетов и приведения пользователями указываемой ими информации к надлежащему виду.

4.1.2 Проект

Этот объект определяет деятельность одного или нескольких пользователей в рамках одного направления исследований. Традиционно проекту может соответствовать один или множество учетных записей в зависимости от специфики суперкомпьютерного центра. Соответственно, и выделение ресурсов, и учёт также могут вестись как в терминах учетных записей, так и в терминах проектов.

Проектный подход позволяет описать несколько различных вариантов организации учёта пользователей. Если каждый пользователь имеет одну учетную запись и использует ее во всех работах, то полагаем, что каждый пользователь ведёт один собственный проект. Если на каждый проект выделяется одна учетная запись и ее используют все участники, то считаем, что за него ответственен владелец проекта.

Однако, эти подходы, несмотря на свою распространённость, имеют существенные недостатки, и мы предлагаем использовать подход, в котором каждый участник проекта получает отдельную учетную запись, при этом все участники одного проекта входят в одну группу.

При таком подходе один пользователь должен иметь разные учетные записи для разных проектов, что позволяет проще организовать учёт и квотирование, а также управлять доступом в терминах проектов, а не отдельных учетных записей. Например, если пользователь ведёт свой проект и одновременно участвует в коллективном проекте, то в случае, если коллективный проект будет закрыт или заблокирован, вход для личного проекта останется открытым.

Таким образом, ориентация модели на проекты позволяет как описать традиционные схемы предоставления доступа, так и предложить более эффективную схему. Каждый проект представлен в системе в виде UNIX-группы и одной или более учётной записи. Каждая учётная запись соответствует одному или всем участникам проекта (в зависимости от конечной реализации проектов в системе). Рекомендуется использовать отдельную учётную запись для каждого участника.

Каждый участник проекта должен указать организацию, от имени которой он участвует в проекте. Организация, указанная руководителем проекта, считается ведущей организацией проекта. Такого рода сведения важны, например, для применимости некоторых лицензий на прикладное ПО.

Проекты могут иметь потребность в особом типе ресурсов, ПО, режиме запуска задач и т.п.

По каждому проекту собирается статистика использования, например, число затраченных процессорочасов, число выполненных задач, объём использованного дискового пространства и т.д.

4.1.3 Организация

Организации представляют интересы исследовательских групп и выступают гарантами соблюдения пользователями правил предоставления ресурсов. Это может быть реализовано следующим образом. От имени руководителей организаций или подразделений создаются поручительства для получения доступа, с поручительствами соотносятся пользователи и конкретные проекты. Организации могут быть различных типов. Например, в СКЦ МГУ практикуется следующая градация:

- МГУ и его филиалы;
- российское высшее учебное заведение;
- зарубежное высшее учебное заведение;
- институты РАН;
- российский научная организация (кроме РАН);
- зарубежный исследовательский центр;
- и др.

4.1.4 Вычислительная система

Основная цель при администрировании суперкомпьютерных центров — обеспечить возможность использования вычислительных ресурсов пользователями. Вычислительные системы зачастую являются гетерогенными, кроме того, нередко добавляются узлы нового типа в рамках расширения/обновления вычислительной системы. В рамках одного СКЦ таких систем может быть несколько, и одним из основных объектов модели является вычислительная система. Среди своих атрибутов она имеет набор учётных записей, установленного прикладного программного обеспечения.

Приоритеты

Отдельным проектам и пользователям или учётным записям могут быть предоставлены различные приоритеты на вычислительной системе. Приоритеты могут быть реализованы разными способами в зависимости от особенностей систем. Например, повышенный приоритет может ускорять продвижение задач в очереди или автоматически ставить их выше задач с более низким приоритетом.

Квоты

Квоты отчасти схожи с приоритетами, но имеют другой смысл. С помощью квот можно ограничить объём ресурсов для проекта или учётной записи, например число процессорочасов или объём дискового пространства. Кроме глобального ограничения ресурсов, можно ограничивать «мгновенное» потребление ресурсов, например, число одновременно запущенных задач или число занятых задачами процессоров.

При наличии возможности в планировщике и менеджере ресурсов можно использовать квоты на ресурсы за временные отрезки, например за месяц или неделю. Подобные квоты имеет смысл применять к таким ресурсам, как процессорочасы или число задач. В этом случае пользователь сможет использовать ресурс в нужном ему режиме, не «упираясь» в квоту постоянно: можно запустить несколько задач подряд и выбрать квоту за короткое время, а потом не производить запусков до нового периода квотирования. Это даёт остальным пользователям возможность свободнее работать.

Важно, что на разных вычислительных системах механизмы квотирования и привилегий могут быть существенно отличаться, при этом требуется их согласование при организации доступа к таким разнородным ресурсам, в том числе, и в рамках отдельного проекта.

Непосредственно реализация механизмов квот и приоритетов не является частью Octoshell, но с его помощью администратор может управлять ими из единой точки, получать данные обо всех квотах определённого проекта, пользователя и т.п.

4.2 Администрирование

В работе любого суперкомпьютерного центра есть некоторый общий набор ключевых процессов администрирования. Схемы их реализации могут быть кардинально отличными, в каждом конкретном случае используется своя логика, свои правила. И в рамках отдельно взятого СКЦ изменения в организации работ — не редкость. Именно поэтому необходим «конструктор», позволяющий связать основные объекты и процессы, присущие большинству СКЦ в том логическом виде, который является актуальным. Рассмотрим ключевые процессы в Octoshell.

4.2.1 Регистрация и предоставление доступа

Получение доступа к вычислительным ресурсам начинается с регистрации пользователя в системе Octoshell. Регистрация пользователя включает в себя:

- создание базового объекта пользователя;
- заполнение обязательных данных (контактных и т.п.);
- привязка пользователя к организации и при необходимости, создание новой организации.

Следует помнить, что в соответствии с законодательством РФ не все данные о пользователе могут быть сохранены без применения специальных средств (хотя для работы с суперкомпьютером такие данные, как правило, не требуются) и для сохранения любых персональных данных пользователь должен явно дать разрешение.

После того, как пользователь создан в Octoshell, он получает возможности создания новых проектов, или принятия приглашений на участие в уже созданных.

В зависимости от внутренних регламентов суперкомпьютерного центра статус проекта может быть изменён на активный (т.е. создание учетных записей всем участникам) при соблюдении некоторых правил, таких как:

- получение и одобрение официального поручительства от имени организации;
- определённый статус пользователя;
- прямое указание руководства и др.

Сами правила перевода объектов из одного состояния в другое, условия их создания или удаления, а также обязательность и формат некоторых характеристик определяются исходя из политики работы конкретного центра.

Если все условия для предоставления доступа выполнены, то на вычислительных системах выполняются скрипты и создаются учетные записи для всех пользователей, входящих в проект.

При этом пользователь может использовать один ключ для работы со всеми своими учетными записями, относящимися к различным проектам.

4.2.2 Рассылки и уведомления

В ходе работы суперкомпьютерного центра часто возникают ситуации, затрагивающие определенные группы пользователей или проектов, или даже всех пользователей. Это может быть выведение машины из работы на время профилактики, изменение системного или прикладного ПО и другие изменения.

Для доведения подобной информации до пользователей существуют механизмы рассылок и уведомлений. В рамках Octoshell создан модуль почтовых рассылок, позволяющий создавать и использовать шаблоны рассылок по определенному набору адресатов. Выборка получателей может быть сделана на основании статусов проектов, организаций и т.д.

Рассылки разделены на две группы: информационные и общие. При этом подписка на общую часть является обязательной для пользователей, т.е. если пользователь попадает под целевую группу рассылки, то от получения рассылки нельзя отказаться. Через данный тип рассылок ведется уведомление о важных событиях, связанных с работой СКЦ. Второй тип рассылки используется для уведомления о событиях, не затрагивающих работу пользователей напрямую. Например, объявления о проведении в центре мастер-классов по некоторому прикладному пакету ПО.

Механизм уведомлений является скорее надстройкой над другими модулями. Например, над модулями поддержки и перерегистрации. При возникновении события, требующего реакции от пользователя, соответствующие яркие указатели на вкладках интерфейса системы будут сигнализировать о необходимости реакции от пользователя.

4.2.3 FAQ

Типовые вопросы, такие как правила получения доступа, основные принципы работы на вычислительных системах и их конфигурации и т.п. необходимо должны быть описаны в разделе часто задаваемых вопросов. Чем нагляднее и подробнее будут описания, тем меньше администраторам придется тратить время на такие вопросы, для которых заранее готов ответ.

Конечно, всегда находятся пользователи, которые предпочитают сразу задавать вопрос, не изучая FAQ. Тем не менее, при хорошей структуризации и подробности раздела частых вопросов вероятность того, что пользователь, будучи отправленным к конкретной части раздела, найдет ответы на свои вопросы и в следующий раз все-таки поищет ответ сначала самостоятельно.

4.2.4 Поддержка

Техническая поддержка пользователей суперкомпьютерного центра является одной из ключевых составляющих в администрировании. При обращении в техническую поддержку крайне важной является максимально подробное описание возникшей проблемы. При этом имеют значение все детали, которые имеют хотя бы малейшее отношение к обращению: проект, система, прикладное ПО, текущие квоты, последние данные отчётов и т.п. Наличие этой информации позволяет администратору резко сократить время решения проблемы. Значительная часть этих данных может быть получена автоматически или с помощью пользователя при составлении обращения.

При составлении обращения пользователя в поддержку целесообразно предлагать пользователю ответить на сопутствующие вопросы о типе проблемы и ее обстоятельствах. По ответам можно группировать вопросы по типам обращений, что позволяет, например, объединять однотипные проблемы в одну, исследовать успешные решения подобных проблем и в целом более эффективно решать проблемы.

Все имеющие соответствующий доступ администраторы видят список поступивших обращений и их детали. Если понятно, кто из администраторов ответственный за решение проблемы, его можно назначить. В противном случае ответственным становится первый ответивший

на обращение. При уточнении проблемы ее можно переадресовать, например, более компетентному специалисту.

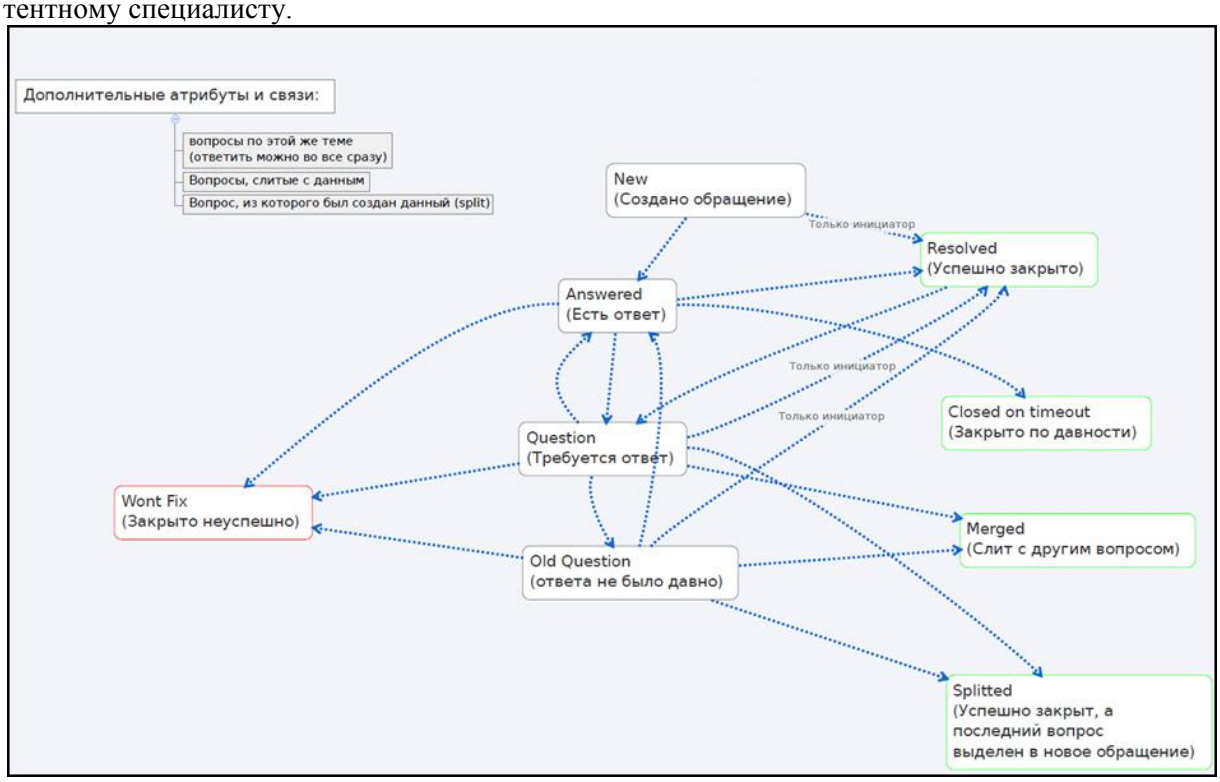

Рис. 1. Диаграмма состояний обращений в пользовательской поддержке.

Все обращения должны сохраняться и дублироваться через оперативные каналы: email, jabber, и т.п. В каждый момент времени каждое обращение должно иметь фиксированный статус: «новое», «есть ответ», «решено», «обновление проблемы» и т.п. (Рис. 1). Как пользователь, так и администратор, должны иметь возможность в любой момент написать сообщение по обращению, которое обновит его статус и будет оперативно отправлено всем участникам переписки.

Важно, что все обращения в поддержку производятся через один интерфейс. Это позволяет сохранять и видеть всю историю обращений и решений проблем со всеми деталями.

Интерфейс позволяет производить выборку и сортировку по новизне обращений, по ответственному администратору, по инициировавшему пользователю, типу обращения, его статусу и другим параметрам.

4.2.5 Процессы обслуживания и сопровождения аппаратных компонент

Так как оборудование и программное обеспечение вычислительных систем центра постоянно нуждается в периодическом и внеплановом обслуживании и обновлении, для обеспечения своевременного принятия каких-либо мер, необходимо отслеживать статус компонент. Прежде всего, это касается больших вычислительных комплексов, где отдельные элементы оборудования даже в надежных системах могут ежедневно выходить из строя просто из-за колоссального числа компонент.

В предлагаемом подходе каждая запись о проведении работ или отправке оборудования в гарантийный ремонт сопровождается следующей информацией:

- состав оборудования или ПО;
- время обнаружения отказа или наступления срока профилактики;
- диагностика сбоя (если таковой был);
- время выведения оборудования или ПО из использования;
- время передачи оборудования в ремонт или обращения в службу поддержки;
- контактные данные по гарантийному ремонту или службы поддержки;

- ожидаемый срок окончания профилактики, ремонта, обновления;
- время возвращения оборудования из ремонта или обновления ПО;
- время передачи ПО или оборудования в использование;
- список изменений, если таковые были (обновление версии, замена блоков и т.п.);
- текущий статус.

Каждая запись дополняется по мере изменения статуса, при этом старые данные остаются. Это дает возможность проследить историю решения проблемы по каждой записи.

Такие данные вносятся вручную техническим персоналом. В некоторых случаях данные о сбое и выводе оборудования из работы могут быть получены автоматически (от системы мониторинга, например), и тогда можно заносить эту информацию соответствующими средствами с обязательным контролем ответственным персоналом.

4.2.6 Компоненты прикладного программного обеспечения

Установка, обновление и поддержка прикладного ПО требуется на каждой суперкомпьютерной системе. Именно поэтому в модели функционирования суперкомпьютерного центра присутствуют эти компоненты. К сожалению, далеко не всегда можно обеспечить корректный автоматический сбор информации о наличии и версиях установленного прикладного программного обеспечения с систем. Поэтому рекомендуется иметь возможность ручного ввода этих данных. Также следует помнить, что некоторые прикладные пакеты могут быть установлены одновременно в разных конфигурациях, например библиотека FFTW может быть одновременно представлена версиями 2 и 3, которые не совместимы друг с другом.

При наличии технической возможности следует обеспечить сбор статистики по использованию установленного прикладного программного обеспечения. Если такой технической возможности нет, стоит получать эти данные от пользователей через опросы наряду с данными о востребованности тех или иных пакетов для оценки целесообразности их установки. Эта информация помогает понять реальную потребность в установленном ПО, а, следовательно, и потребность в его поддержке и обновлении.

Информация об установленном прикладном программном обеспечении должна включать в себя не только название и версию. Целесообразно указать по каждой системе следующие данные по установленному ПО:

- название и версия ПО;
- контакты по вопросам лицензирования;
- контакты для технической поддержки;
- срок лицензии;
- срок поддержки;
- специальные условия лицензии (ограничения);
- стоимость обновления лицензии/поддержки;
- путь в файловой системе, куда установлен пакет;
- описание процедуры установки и настройки пакета с указание конкретных параметров, использованных на системе;
- описание процедуры тестирования пакета;
- краткая инструкция для пользователей по использованию пакета;
- ответственный за установку и обновление пакета;
- контакты локального технического консультанта по пакету.

В зависимости от условий лицензии, доступ к пакету может быть предоставлен не всем, а только части пользователей. Например, академические лицензии позволяют работать только в рамках образовательных проектов.

Некоторые лицензии могут быть ограничены только сотрудниками отдельной организации. Каждый прикладной пакет должен содержать набор отношений к проектам и организациям, которые имеют право на доступ к нему.

4.2.7 Статистика

Возможность сбора и анализа разного рода статистики являются важной составляющей эффективного управления вычислительным центром. Статистические отчёты могут строиться автоматически и в ручном режиме, а также по уже имеющимся данным в системе либо по результатам опросов пользователей.

Автоматически создаваемые статистические отчеты. Это могут быть данные по числу проектов, пользователей, организаций и других объектов в системе управления и их распределениям по различным критериям. При наличии регулярного импорта статистических данных с подключённых систем, в автоматические отчёты могут быть включены такие данные как количество запусков программ, число использованных процессорочасов и т.п., сгруппированных по пользователям, проектам или организациям. Автоматические отчёты целесообразно составлять за регулярные периоды времени, например, сутки или месяц.

Отчёты в ручном режиме могут быть составлены по тем же данным, что и для автоматических, но за произвольный период времени, по специфическим критериям выбора проектов, пользователей или организаций. Это удобно для получения отчётов по отдельным проектами, организациям и их типам, для генерации годовых или промежуточных отчётов и т.п.

Данные для отчётов можно получать, как упоминалось выше, из имеющихся данных и из импортированных данных от систем. Кроме того, можно получать важную информацию от самих пользователей, используя механизм опросов. Такие данные позволяют оценить уровень удовлетворённости пользователей, их потребности, спланировать дальнейшую работу.

4.3 Перерегистрация и экспертиза

4.3.1 Процедура перерегистрации

Возникновение процедуры перерегистрации как таковой обусловлено следующими предпосылками:

- Требуется поддерживать актуальность представленных пользователями персональных данных и данных по проводимым исследованиям
- Требуется контроль использования вычислительных ресурсов. Действительно, при безвозмездном предоставлении доступа к ресурсам целесообразно стимулировать эффективное использование суперкомпьютеров.
- Требуется «держать руку на пульсе» реальных потребностей пользователей, оценивая как динамику объемов запрашиваемых ресурсов отдельными проектами, так и потребность в установке прикладного ПО, закупке лицензий, организации обучения и т.д.
- Вообще говоря, требуется понимание, подкрепленное реальными показателями, подтверждающими, на что именно были потрачены дорогостоящие ресурсы. Вместе с тем, что ресурсы не стоят реальных денег для пользователей сейчас, для держателя системы в лице Университета содержание выливается в очень значимые деньги. Речь идет не только о самих полученных результатах, но и опубликованных книгах, выступления на конференциях, публикациях в высокорейтинговых изданиях, студентах, привлеченных к исследовательскому процессу или прошедших через практикум, и т.д.

В связи с этим ежегодно в СКЦ МГУ проводится перерегистрация. В ее рамках каждый руководитель проекта представляет краткий иллюстрированный отчет о полученных результатах с использованием вычислительных ресурсов суперкомпьютерного комплекса. Вместе с тем заполняются традиционно два опроса.

«Опрос по проектам» адресован руководителям всех активных проектов для оценки потребностей с точки зрения организации проведения исследований и подтверждения актуальности данных по проектам. Опрос также включает форму запроса ресурсов на следующий учетный период.

«Опрос по пользователям» адресован всем активным пользователям. С одной стороны, пользователи подтверждают корректность указанных в профиле данных, с другой — имеют возможность высказать свои пожелания по работе, указав, например, интересующие темы для обучения или потребность в специфическом ПО.

4.3.2 Экспертиза

Направления исследований, проводящихся с использованием СКЦ столь много, что вряд ли найдется отдельный эксперт, способный провести достойную оценку всех работ. В рамках системы создан механизм экспертиз. Каждый эксперт может выделить работы, наиболее близкие по направлению исследований, либо работы могут быть распределены по экспертам принудительно. Если при анализе работы эксперт видит, что целесообразно привлечь эксперта смежной области, есть возможность переназначить эксперта. Эксперты могут видеть представленные отчетные материалы и просить, при необходимости, их доработать. Пользователи, в свою очередь, могут доработать и обновить представленную в системе версию отчета. Эксперты остаются анонимными для пользователей. По результатам экспертизы отчеты оцениваются по ряду критериев.

Неудовлетворительные оценки экспертов дают основание о сокращении объема выделенных ресурсов вплоть до полной блокировки доступа.

Информация о лучших работах представляется на конференциях, авторы приглашаются на специализированные семинары, получают определенные льготы, например, при организации выделенных вычислений.

4.4 Реализация разработанного программного комплекса управления СКЦ

Вся работа пользователей, администраторов и экспертов организована через единую точку входа и осуществляется через web. Система Octoshell создана на основе программной платформы Ruby On Rails. Работа прототипа программного комплекса организована следующим образом:

- система реализована по модульному принципу и может быть доработана до требуемого функционала уже по специфике организации работ в конкретном СКЦ;
- действия на целевой системе выполняются через модуль асинхронных задач, выполняющихся независимо от основного модуля и не блокирующих его работу;
- через аналогичный модуль осуществляется рассылка уведомлений пользователям;
- на целевой системе устанавливается набор скриптов, позволяющих выполнять необходимые действия:
  - создание и удаление пользователей и групп;
  - добавление и удаление ключей доступа;
  - блокирование и разблокирование доступа пользователей;
- создаётся пользователь для выполнения команд на удалённой системе, вход для него разрешается только с помощью ключа, который указывается в модуле для выполнения действий; созданному пользователю на системе предоставляются права выполнения установленных скриптов с привилегиями суперпользователя (через пакет sudo).

Таким образом, достигается максимальный уровень защищённости системы — даже похитив ключ доступа, на системе можно выполнить только заданный набор скриптов.

При необходимости набор действий на целевой системе легко расширяется добавлением новых скриптов. Функциональность комплекса может быть относительно просто расширена.

В качестве хранилища данных используется свободная база данных Postgresql, для реализации модуля асинхронных задач — свободный пакет Sidekiq.

Внедрённый и работающий прототип комплекса расположен в Сети по адресу https://users.parallel.ru. Исходный текст разработанного комплекса доступен по адресу https://github.com/octoshell/octoshell-v2.

В разработанном прототипе комплекса поддерживается гибкое разделение прав доступа — отдельной группе пользователей могут быть предоставлены возможности, например для ответов в технической поддержке, проведении экспертизы отчётов и т. д. При этом каждый пользователь может видеть на странице только те элементы, которые ему доступны.

Подключение новой системы в прототип комплекса производится также через web-интерфейс. Предварительно на системе должен быть заведён пользователь для доступа и настроен набор управляющих скриптов.

Для того чтобы иметь возможность управлять прохождением задач пользователей, например, изменять приоритеты, квоты на процессорочасы и т. п., требуется поддержка со стороны менеджера ресурсов. Не все менеджеры имеют такую поддержку, поэтому для них требуется создание дополнительных модулей или программ. В рамках данного проекта успешно начата работа по созданию такого модуля для популярного менеджера ресурсов SLURM [13].

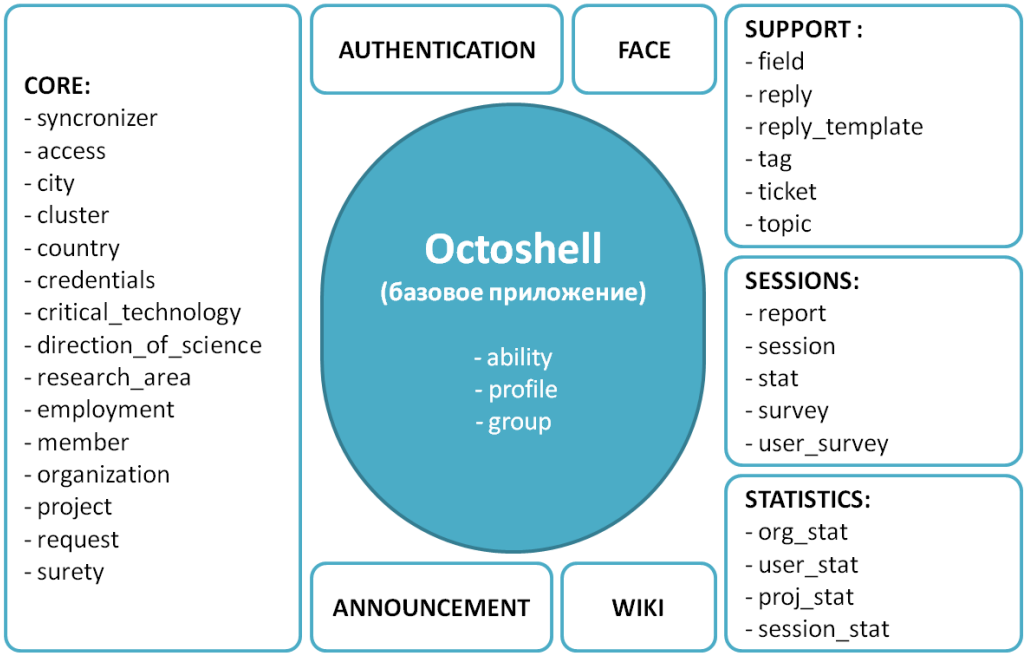

Рис. 2. Модульность системы Octoshell.

Распределение функционала в системе реализовано следующим образом. В базовом приложении реализована минимальная функциональность, большинство действий реализовано на уровне модулей (Рис. 2):

- Базовое приложение:
  - профиль пользователя;
  - формирование группы пользователей по правам доступа;
  - разграничение прав на различные действия.
- CORE — ключевой модуль, отвечающий специфике организации доступа к вычислительным ресурсам, в его реализацию входит:
  - синхронизация данных с вычислительными системами;
  - управление доступом к ресурсам;
  - описание вычислительных систем;
  - запросы пользователей на предоставление ресурсов;
  - механизм поручительств для предоставления доступа;
  - описание проектов;
  - организации;
  - и т.д.
- AUTHENTICATION:аутентификация пользователей системы Octoshell.
- FACE: пользовательский интерфейс.
- ANNOUNCEMENT: механизм рассылок.
- SUPPORT: техническая поддержка.
- SESSIONS: механизм проведения перерегистраций.
- STATISTICS: построение разного рода статистики.

Для самостоятельной установки для скачивания доступна базовая версия, однако, при использовании модели функционирования близкой к СКЦ Московского университета, развертывание системы не потребует серьезных усилий. Дистрибутив устанавливается на отдельную виртуальную машину, представляя собой, де факто, решение «из коробки».

4.5 Апробация

Помимо уже реализованного функционала, в систему добавляются и новые возможности. Приведем некоторые из них.

В рамках разрабатываемого в данный момент модуля реализуется возможность более гибкого управления приоритетами и квотами на процессорное время, что позволит более полно интегрировать менеджер ресурсов SLURM в разработанную систему и дать возможность администратору выделять отдельным проектам больше (или меньше) ресурсов в зависимости от их приоритетности или ранее полученных результатов непосредственно через интерфейс системы.

С каждой системы, как правило, возможно получение статистической информации об активности пользователей и использовании ресурсов, таких как дисковое пространство, число запущенных задач, процессорочасы. Такие данные позволяют получить информацию о фактическом использовании ресурсов. Такие данные могут быть получены различными способами: от систем мониторинга, с помощью системных скриптов для сбора статистики, с помощью специализированных пакетов. Полученные данные могут быть обработаны с целью соотнесения их с другими объектами системы: пользователями, проектами, организациями.

В ближайшем будущем в личных кабинетах пользователей планируется добавить возможность видеть перечень запускавшихся ими задач. Руководители проектов, соответственно, будут видеть данные о запусках задач всеми участниками проекта.

Следующим шагом станет указание более точных данных по задачам: станут доступны средние показатели использования ресурсов CPU, нагрузки на сеть, интенсивности работы с файловой системой, и другие важные интегральные характеристики [14].

Для отдельных задач на следующем этапе развития модуля станут доступны профили использования ресурсов задачей, откроется возможность изучения поведения ключевых динамических характеристик [15]. Для полноценной реализации данного функционала потребуется провести ряд изменений в используемые методы сбора данных системного мониторинга [16].

Доработанная система мониторинга позволит, в том числе, реализовать модуль ситуационного экрана для системных администраторов, отражая степень мгновенной загрузки ресурсов вычислительного комплекса с историей изменений [17].

Важным функциональным дополнением для администраторов станет интеграция с системой обеспечения автономного функционирования суперкомпьютерного комплекса, получив в рамках единого интерфейса возможность видеть аномалии в поведении компонент СКЦ: аппаратуры, ПО, инфраструктуры [18].

В данный момент Octoshell активно используется в суперкомпьютерном центре Московского университета. Вычислительный парк СКЦ включает такие системы как «Чебышев» (60TFlops), «Ломоносов»(1,7PFlops), «Ломоносов-2» (2,5PFlops) — всего около 10000 вычислительных узлов. В завершении, приведем немного цифр, отражающих реальные условия использования системы.

- 616 исследовательских активных проектов;
- более чем 350 организаций;
- в Octoshell зарегистрированы и имеют возможность работать 2677 пользователей;
- службой поддержки за год решается около 1000 обращений пользователей.

Полученные результаты в ходе апробации подтверждают целесообразность использования системы. Перспективы системы Octoshell - эксплуатация и дальнейшее расширение функционала в СКЦ МГУ, с одной стороны, с другой - как разработчики, так и пользователи системы заинтересованы в использовании данной системы и в других суперкомпьютерных центрах с целью выявления новых направлений дальнейшего развития.

Работа ведется при финансовой поддержке гранта РФФИ, грант N13–07–12206–офи_м.

Литература

1. Топ50 Суперкомпьютеры. URL: http://top50.supercomputers.ru (дата обращения: 02.08.2015).
2. Top500 Supercomputer sites. URL: http://top500.org (дата обращения: 02.08.2015).

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, N 7, 2012. С. 36-39.
4. Жуматий С.А., Никитенко Д.А. Подход к гибкому управлению суперкомпьютерами // Труды Международной суперкомпьютерной конференции «Научный сервис в сети Интернет: все грани параллелизма», С. 296-300, Изд-во МГУ Москва г.Москва, 2013.
5. Жуматий С.А., Дацюк О.В. Администрирование суперкомпьютеров и кластерных систем. Издательство Московского университета Москва, 2014. С. 400.
6. Torgue batch system URL:http://www.adaptivecomputing.com/products/open-source/torque/ (дата обращения: 02.08.2015).
7. SLURM workload manager URL:http://slurm.schedmd.com/ (дата обращения: 02.08.2015).
8. OpenPBS URL:http://www.mcs.anl.gov/research/projects/openpbs/ (дата обращения: 02.08.2015).
9. Ganglia Monitoring System URL: http://ganglia.sourceforge.net/ (дата обращения: 02.08.2015).
10. Zabbix monitoring URL:http://www.zabbix.com/ru/ (дата обращения: 02.08.2015).
11. Nagios monitoring URL:https://www.nagios.org/ (дата обращения: 02.08.2015).
12. Open-source Ticket Request System URL:http://www.otrs.org/ (дата обращения: 02.08.2015).
13. Леоненков С.Н. Расширение функциональности менеджера ресурсов суперкомпьютера SLURM // Труды Международной суперкомпьютерной конференции «Научный сервис в сети Интернет: многообразие суперкомпьютерных миров», С. 472-476, Изд-во МГУ Москва г.Москва, 2014.
14. Никитенко Д.А. Комплексный анализ производительности суперкомпьютерных систем, основанный на данных системного мониторинга // Вычислительные методы и программирование: Новые вычислительные технологии (Электронный научный журнал). Том 15, 2014. С. 85–97.
15. Антонов А.С., Жуматий С.А., Никитенко Д.А., Стефанов К.С., Теплов А.М., Швец П.А. Исследование динамических характеристик потока задач суперкомпьютерной системы // Вычислительные методы и программирование: Новые вычислительные технологии (Электронный научный журнал). Том 14, раздел 2, 2013. С. 104-108.
16. Стефанов К.С. Система мониторинга производительности суперкомпьютеров // Вестник Пермского Национального исследовательского политехнического университета. Аэрокосмическая техника. № 39. 2014. С. 17–34.
17. Воеводин Вл.В. Ситуационный экран суперкомпьютера // Открытые системы, N 3, 2014. С. 36-39.
18. Антонов А.С., Воеводин Вад В., Даугель-Дауге А.А., Жуматий С.А., Никитенко Д.А., Соболев С.И., Стефанов К.С., Швец П.А. Обеспечение оперативного контроля и эффективной автономной работы Суперкомпьютерного комплекса МГУ // Вестник Южно-Уральского государственного университета. Серия "Вычислительная математика и информатика", Том 4(2), 2015. С. 33–43.

Octoshell: large supercomputer complex administration system

Sergey Zhumatiy, Dmitry Nikitenko and Vladimir Voevodin

Keywords: computing center, administration, user management, Octoshell

Managing and administering of modern supercomputer centers and HPC systems as a part is a complicated and complex task. The usage of numerous traditional stand-alone tools for administering and management of supercomputers becomes a bottleneck for efficient resource utilization in conditions of growing systems scale. The developed “Octoshell” system for support of running supercomputer centers is aimed at solving this problem. It implements essential tools for administering in a single interface and allows significant automatization of typical management tasks ensuring higher efficiency of large supercomputer complex output as a whole.