УДК 658.012

ЭВОЛЮЦИЯ АРХИТЕКТУРЫ ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ БОЛЬШОГО НАУЧНО-ТЕХНИЧЕСКОГО ПРОЕКТА

А. В. Жучков[(1)], Н. В. Твердохлебов[(2)]

[1] *Институт химической физики им. Н. Н. Семенова, г. Москва;*
[2] *Институт проблем управления им. В. А. Трапезникова, г. Москва*

Рассмотрены проблемы организации корпоративных информационно-вычислительных ресурсов и управления ими при проведении научных исследований, связанных с интенсивным обменом данными. Описаны возможные решения указанных проблем на основе технологии понятийных сетей. Указаны перспективы применения Grid-технологий для реализации крупномасштабных понятийных сетей (на примере биомедицинских исследований).

ВВЕДЕНИЕ

Создание информационно-вычислительных ресурсов (ИВР) крупномасштабного научно-исследовательского проекта и управление этими ресурсами в ходе реализации проекта является нетривиальной задачей. Причины трудностей заключаются в слабой формализации задач информационно-вычислительной поддержки исследовательских работ, значительной разнородности вовлекаемых ресурсов при необходимости их интеграции и существенной эволюции требований к ИВР со стороны участников работ в ходе выполнения проекта. В рамках межведомственного проекта, осуществляемого десятками научных центров в течение ряда лет, одновременно существуют несколько точек зрения на эффективность использования ИВР и, как правило, некоторые из них являются конкурирующими. Более того, эти точки зрения изменяются, так как в ходе реализации проекта возникает более глубокое понимание задач и проблем исследований, и потребности в ИВР меняются и перераспределяются в рамках сообщества, которое в соответствии с современной терминологией мы будем далее называть «виртуальной организацией» [1]. Управление распределенными ИВР виртуальных организаций (ВО) осуществляется как применением определенной политики предоставления ресурсов, так и путем периодической реконфигурации архитектуры всего информационного пространства ВО.

За последние два десятка лет произошло несколько радикальных изменений архитектуры ИВР. Мэйнфреймы сменились на мини-ЭВМ, их сменили персональные компьютеры, которые вскоре объединились в одноранговые сети. Следующим этапом стал переход к архитектуре «клиент—сервер» и, наконец, сегодня активно развивается архитектура Grid [2]. Технологии высокопроизводительных вычислений, массового хранения информации (баз данных), создания и использования интеллектуальных систем первоначально развивались в значительной степени независимо. К настоящему времени в таких областях науки, как биотехнология, геофизика, химия горения и высокомолекулярных соединений и др., поставлены на повестку дня задачи, требующие для их решения совместного и согласованного применения перечисленных технологий. Реализация этих требований в корпоративных информационно-вычислительных системах, масштаб которых в ряде случаев стал достигать глобальных размеров (например, корпоративная сеть научных центров концерна «Novartis» [3], инфраструктура LCG для обработки данных Большого адронного коллайдера в CERN [4]), поставила ряд новых проблем. В первую очередь, это обеспечение интероперабельности при реализации метакомпьютинга на разнородных распределенных вычислительных ресурсах и семантическая интеграция разнородных распределенных федеративно-администрируемых информационных ресурсов [5].

Перечисленные проблемы и возможные пути их решения рассмотрены на примере эволюции ИВР Межведомственной научно-технической программы «Вакцины нового поколения и медицинские диагностические системы будущего» (далее по тексту МНТП). Меди-

ко-биологические исследования по проектам МНТП проводятся с 1999 г. в более чем 30-ти ведущих научных центрах России. Современные технологии стремительно превращают эту область знаний в высококомпьютеризированную отрасль научной индустрии, в которой важную роль играет возможность осуществления интенсивных операций с огромными массивами данных. В связи с этим, с самого начала реализации МНТП в составе ее проектов было выделено направление, связанное с информационно-вычислительной поддержкой проводимых исследований.

1. ЭВОЛЮЦИЯ ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ КОРПОРАТИВНОЙ СЕТИ МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Создание единого информационного пространства МНТП началось с построения коммуникационных ресурсов, способных обеспечить интенсивные потоки данных, возникающих при биомедицинских исследованиях. Например, объем каждой генетической базы данных сегодня составляет около 300—500 Гбайт и растет экспоненциально. Опуская «эмбриональные» этапы развития того, что выросло в корпоративную сеть МНТП, можно сказать, что при рождении это была быстро расширявшаяся сеть peer-to-peer. Магистральные коммуникационные ресурсы были созданы на основе Южной Московской опорной сети (ЮМОС), но включали в себя также и элементы других крупных сетей, в том числе в Санкт-Петербурге и Новосибирске. Ядро коммуникаций составило FDDI-кольцо ЮМОС с пропускной способностью 100 Мбит/с, связавшее наиболее крупные научные центры, участвующие в работах по Программе [6]. К концу 2003 г. все организации, участвующие в проектах МНТП, получили разветвленные внутрикорпоративные сети и высокоскоростной Интернет для доступа к библиографическим ресурсам (Medline, EBSCO, Springer Verlag, Elsevier и др.). Быстрое развитие технологий позволило с 2004 г. начать переход на гигабитные коммуникации не только на магистральных, но и на внутрикорпоративных линиях. Технология Интранет стала основой внутрикорпоративных информационных обменов. В результате поток пересылаемых и накапливаемых данных возрос на несколько порядков и стала очевидной необходимость организации управления потоками данных и ресурсами их хранения как в рамках Интранет-сетей отдельных организаций, так и в рамках корпоративной сети МНТП в целом.

Начальный этап развития вычислительных ресурсов в рамках МНТП проходил по типовому сценарию. Отсутствие общей стратегии развития и раздробленность источников финансирования привели к накоплению большого числа разнородных персональных компьютеров (ПК). Об эффективности их использования и тем более о совместном использовании не могло быть и речи. Помимо ПК в некоторых организациях накапливались и серьезные вычислительные ресурсы, предназначенные для исследований, связанных с геномикой, протеомикой и компьютерным конструированием лекарств. Число таких организаций ограничено — Институт молекулярной биологии (ИМБ) РАН, Институт биоорганической химии РАН, Институт биомедицинской химии (ИБМХ) РАМН, к тому же технологическое развитие стремительно понижает статус ранее созданных вычислительных ресурсов. Тем не менее, к настоящему времени в рамках МНТП накоплен значительный потенциал, совместное использование которого в корпоративной сети позволило бы решать задачи, практически недоступные отдельным институтам. С 2002 г. на основе существующих коммуникационных и вычислительных ресурсов МНТП создан пул распределенных вычислительных ресурсов для задач биомедицинских исследований. В его состав вошли Linux-кластер ИХФ (Институт химической физики им. Н. Н. Семенова РАН), два Linux-кластера Телекоммуникационного центра «Наука и общество» и гетерогенный Windows-кластер ИМБ. В 2004—2005 г г. к имеющимся ресурсам были добавлены Linux-кластер с оперативной памятью 16 Гбайт для задач химической кинетики (ИХФ РАН) и высокопроизводительный Linux-кластер на базе процессоров Opteron для задач молекулярной динамики при исследованиях протеинов и ферментов в Институте биохимической физики (ИБХФ) РАН.

Первоначально доступ к распределенным вычислительным ресурсам осуществлялся традиционным путем создания учетных записей для удаленных пользователей. Однако, неравномерность загрузки ресурсов, их разнородность и, главное, необходимость взаимного обмена свободными ресурсами в рамках ВО привели к постановке задачи виртуализации вычислительных ресурсов корпоративной сети МНТП, а также задачи управления и взаимного учета использования ИВР.

2. ИНФОРМАЦИОННАЯ СИСТЕМА КОРПОРАТИВНОЙ СЕТИ МНТП

Активное применение компьютеров привело к резкому ускорению накопления данных в биомедицинских исследованиях. Было очевидно, что группы исследователей, совместно работающих в рамках отдельных проектов МНТП, потребуют создания в рамках корпоративной сети выделенных серверов данных для хранения информационных массивов, нарабатываемых в рамках этих проектов, а также организации совместного доступа к этим ресурсам. Такие работы были выполнены, и в рамках одноранговой корпоративной сети были созданы базы данных общего пользования на основе сетевых версий СУБД MS Access, MS FoxPro и Paradox. Эти ресурсы позволили осуществить начальное накопление данных и обучение исследователей методологии работы с применением информационных технологий.

На начальном этапе эти формально сетевые информационные ресурсы на самом деле являлись вполне локальными. Сетевыми они были только для маленьких групп исследователей, работавших вместе над конкретной проблемой. В то же время, исследования в рамках МНТП были спланированы как комплексные работы, и ценность их в значительной мере определялась не только масштабностью, но и продуманным сочетанием различных направлений, логически связанных между собой. В связи с этим, к 2002 г. была сформулирована задача организации единого информационного пространства МНТП с целью обеспечить согласованную работу исследователей из разных проектов над общим информационным полем.

Для решения этой задачи в 2002—2003 г г. была разработана клиент-серверная архитектура с целью создания в рамках корпоративной сети МНТП набора феде-

ративно-администрируемых информационных ресурсов, доступных всем участникам МНТП в соответствии с определенными для них полномочиями. Эти ресурсы представляют собой совокупность разнородных предметно-ориентированных баз данных и инструментальных средств поиска, обработки и представления информации, унифицированных в рамках МНТП. При разработке информационных ресурсов МНТП пришлось преодолевать ряд проблем, которые не возникают в других приложениях или, по крайней мере, не имеют столь выраженного характера. Общепринятая схема разработки баз данных состоит в выявлении и формализации системы запросов, которые будут осуществляться к создаваемой базе. В нашем случае не только не удавалось сформулировать совокупность фиксированных запросов, но даже и сама формализация информационных объектов, присутствующих в процессе медико-биологических исследований, оказалась весьма сложной задачей. В ходе исследований выяснилось, что большинству участников МНТП необходимо создание достаточно простых нереляционных баз данных для хранения и анализа библиографической и фактографической информации. Такая схема хранения данных не позволяет конструировать сложные запросы, однако информационные потребности исследователей в процессе работы столь изменчивы, что для большинства из них формирование формализованных запросов к базе данных представляет собой неприемлемо сложную задачу. Вместо этого был разработан механизм индексирования баз данных и поиска по ключевым словам. В тоже время в МНТП имеются исследовательские группы, достаточно подготовленные, чтобы создавать собственные базы данных с реляционными СУБД (например, база эпитопов вируса гепатита С в ИБМХ на основе СУБД Interbase). Эта информация представляла интерес для других групп исследователей, например, занятых разработкой ДНК-вакцин, в связи с чем потребовалось обеспечить взаимодействие с различными СУБД. Опуская технические детали, отметим, что для организации работы различных групп исследователей в общем информационном поле МНТП потребовалось разработать унифицированный интерфейс приложения для поиска и представления информации из разнородных баз данных. Это потребовало создания программного обеспечения промежуточного уровня (middleware), обеспечивающего согласованное взаимодействие этого приложения с различными СУБД, имеющимися в составе ИВР, включая нереляционные, а также взаимодействие с внешними ресурсами, в том числе Medline (PubMed) и некоторыми другими [7].

Созданные средства и технология работы в едином информационном пространстве МНТП обеспечили возможность быстрого доступа ко всей совокупности информации по проблемам разработки вакцин и диагностических систем. Однако открывшийся доступ к столь значительным информационным ресурсам породил новую проблему — пользователи оказались не в состоянии эффективно работать со столь большим объемом разнородной информации. Для решения этой проблемы необходимо радикальное изменение технологии работы с информационными ресурсами.

3. ОТ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПРОЕКТА К ПОНЯТИЙНОЙ СЕТИ ВИРТУАЛЬНОЙ ОРГАНИЗАЦИИ

Стремительное развитие глобальных информационно-вычислительных сетей ведет к изменению фундаментальных парадигм обработки данных. Их можно охарактеризовать, с одной стороны, как переход к исключительно распределенной схеме создания, поддержки и хранения ресурсов, а с другой — как стремление к виртуальному единству посредством предоставления свободного доступа к любым ресурсам сети через ограниченное число «точек доступа» (например, порталов). Мы полагаем, что для решения возникающих при этом проблем разного уровня будут создаваться так называемые «понятийные сети» (ПС) — комплексы предметно-ориентированных программно-аппаратных средств и организационных мероприятий для создания, поддержки и коллективного использования разнородных распределенных ИВР в рамках виртуальных организаций пользователей, являющихся коллаборациями экспертов по конкретной прикладной области и экспертов по технологиям знаний (knowledge management). Отметим, что важная особенность ПС заключается в участии в ВО экспертов по управлению знаниями, задача которых состоит в формировании модели предметной области на основе имеющихся данных и предметно-ориентированных онтологий, обеспечивающих связь между разнородными данными и необходимую формализацию понятий. Физическую основу ПС составляют разнородные распределенные ресурсы — базы данных различной архитектуры, библиотеки прикладных решений, специализированные программные пакеты, потоковые данные, базы знаний, библиотеки онтологий и метаданных, а также аппаратная часть — вычислительные узлы и узлы хранения. Функциональное наполнение ПС составляет широкий набор сервисов, предназначенных для поддержки ВО и виртуальных лабораторий, проблемно-ориентированных вычислений, автоматизированного обучения и консультирования, поиска данных и знаний, визуализации различного рода, поддержки Интернет-порталов как точек доступа в ПС, публикации научных работ [8].

Номенклатура функциональностей сервисов конкретной ПС зависит от предметной ориентации ВО — ее пользователей. Однако независимо от прикладной направленности функционирование всех элементов ПС должно опираться на сервисы и ресурсы единой операционной среды. Базовыми функциями этой среды являются аутентификация и авторизация пользователей, администрирование среды в целом и взаимодействия ресурсов, сервисов и пользователей, информационное обслуживание, управление заданиями в среде распределенных ресурсов, управление информационными потоками и другие функции такого рода. В рамках развития информационной среды МНТП на основе технологии «клиент—сервер» была реализована значительная часть этих функций. Имеется множество разнородных ИВР, некоторые элементы управления виртуальной организацией и небольшой набор сервисов для биомедицинских вычислительных задач и работы с разнородными информационными ресурсами. Однако в рамках сильно гетерогенной среды сложно обеспечить безопасность (важный аспект биомедицинских исследований),

взаимный учет использования ресурсов и оперативное предоставление свободных ресурсов ВО «по требованию». В то же время сегодня уже имеются и активно развиваются комплексная технология и средства, позволяющие реализовать необходимые для ПС функциональности операциональной среды и при этом обеспечить интероперабельность разнородных элементов ПС в масштабах ресурсов трансевропейских ВО и в перспективе — в глобальном масштабе. Это технология, именуемая термином Grid.

4. GRID — ПЕРСПЕКТИВНАЯ ОПЕРАЦИОНАЛЬНАЯ СРЕДА ПОНЯТИЙНЫХ СЕТЕЙ

Создание и исследование крупномасштабных сетей распределенных ИВР, построенных с помощью технологии Grid, является сегодня наиболее динамично развивающейся областью Computer Sciences. От пионерских работ, основанных на использовании множества независимых подключенных к Интернету ПК (SETI@Home, FightAids@Home), исследования перешли в стадию реализации крупномасштабных международных проектов и создания программного обеспечения промежуточного уровня для интеграции вычислительных фабрик и распределенных хранилищ данных. Проекты построения сетей Grid осуществляются и в США (например, GridAlliance), и в Японии (BioGrid). В Европе сегодня выполняется 81 международный Grid-проект, финансируемый центральными органами ЕС в рамках Шестой рамочной программы (FP6).

Полный доступ ко всем ресурсам удаленного компьютера — одна из главных особенностей, отличающих Grid от Интернета, и эта особенность требует организации более совершенных систем управления распределенными ресурсами, особенно аутентификации, авторизации и обеспечения безопасности как сети Grid в целом, так и ее отдельных узлов. Однако полный доступ к ресурсам — это только одна из особенностей. Grid — это нечто большее, чем удаленный доступ. По нашему мнению, главное отличие Grid от существующих сетей — это виртуализация информационно-вычислительных ресурсов. Основываясь на нашем опыте построения и исследования сегмента Grid-сети в рамках международных (EU DataGrid) и российских (RGrid) проектов, мы полагаем, что наиболее адекватно следует определить «компьютерную сеть по технологии Grid», как совокупность аппаратно-программной и организационной инфраструктур, обеспечивающих виртуализацию информационно-вычислительных ресурсов, совместно используемых в рамках ВО. Отметим важный факт, что локальное управление (администрирование) этими ресурсами осуществляется их владельцами на федеративной основе и, таким образом, Grid-сеть принципиально децентрализована, несмотря на возможность наличия нескольких уровней ее организации.

Основу реализации Grid-сети составляет организующее программное обеспечение промежуточного уровня, стандартом де-факто которого сегодня является Globus Toolkit. За пять лет развития и тестирования достаточно хорошо отработаны такие функции, как аутентификация и авторизация пользователей (и их задач, выполняемых в Grid-сети) на основе системы сертификационных центров, информационные службы учета доступных ресурсов и службы управления заданиями. Благодаря этому Globus Toolkit стал серьезной основой для построения крупномасштабных ПС.

В рамках участия в европейском проекте EU DataGrid в 2000—2001 гг. в Москве на базе оптоволоконных линий ЮМОС была создана опорная магистраль российского сегмента RGrid с пропускной способность 1,0 Гбит/с. Сотрудничество с рядом европейских Grid-проектов позволило авторам изучить и применить для организации вычислительных ресурсов МНТП программное обеспечение промежуточного слоя (middleware) Globus Toolkit, что обеспечило решение проблемы аутентификации и авторизации пользователей, а также запуска задач при реализации метакомпьютинга на имеющихся вычислительных ресурсах МНТП. В 2002 г. на основе Globus Toolkit версии 2.0 эти ресурсы были организованы в сегмент RGrid и разработана технология распределенного решения таких задач большой размерности, как моделирование конформаций ДНК (совместно с ИБХФ) и сравнительный анализ генетических последовательностей (совместно с ИМБ) [9].

В третьей версии Globus Toolkit был радикально переработан в соответствии с концепцией Open Grid Service Architecture (OGSA), основанной на взаимодействии служб (сервисов) Grid-сети, доверяющих друг другу на основе сертификатов. В настоящее время продолжается сближение Web и Grid на основе концепции Web Services Resource Framework, которая, как предполагается, даст возможность полнее реализовать преимущества сервис-ориентированной архитектуры Grid-систем и в еще большей степени удовлетворить потребности ПС в функциональностях операциональной среды. Разработка Grid-сервисов для организации распределенных вычислений на виртуализированных вычислительных ресурсах и для взаимодействия пользователя ПС с разнородными, распределенными и виртуализированными хранилищами данных является самой актуальной задачей развития Grid-технологии. Наиболее перспективной представляется технология OGSA-DAI (www.ogsadai.org), обеспечивающая создание и функционирование совокупности Grid-сервисов для доступа к разнородным виртуализированным информационным ресурсам. В соединении с тематическими онтологиями, представляющими собой модели различных разделов предметной области и позволяющими связывать разнородные данные и ключевые понятия (концепты), сервисы OGSA-DAI образуют каркас информационных связей, обеспечивающий семантическую целостность информационного пространства ПС. Отметим, что для эффективного использования семантических связей при поиске информации в ПС необходима разработка и стандартизация методов и Grid-сервисов формирования и использования метаданных.

Концепция OGSA, Globus Toolkit, технология OGSA-DAI, а также разработанные в ходе исследований сегмента RGrid тематические онтологии и структуры метаданных послужили основой для реализации операциональной среды понятийной сети МНТП.

5. ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОГО ПРОСТРАНСТВА МНТП НА ОСНОВЕ АРХИТЕКТУРЫ OGSA

В 2003 г. информационно-вычислительные ресурсы МНТП были реконфигурированы на основе новой версии middleware Globus Toolkit 3.0 в соответствии с кон-

цепцией OGSA, а ранее разработанные приложения реализованы в виде выполняемых по запросу Grid-сервисов. Для новой конфигурации вычислительных ресурсов сегмента RGrid были проведены исследования по организации вычислений и управлению заданиями. Благодаря сервис-ориентированной архитектуре стала возможной виртуализация вычислительных ресурсов ПС в МНТП в рамках сегмента RGrid и даже были сделаны попытки организовать доступ к этим ресурсам «по требованию» (on-demand). Однако развитие этой архитектуры и эффективность ее применения серьезно сдерживаются недоработанностью middleware Globus Toolkit, которое, несмотря на его признанность в качестве международного стандарта де-факто, не является законченным программным продуктом и требует при его использовании не только высокой квалификации специалистов, но и некоторой доработки применительно к конкретной конфигурации программно-аппаратных ресурсов, а иногда и к решаемым задачам. Реализация Grid-технологий для создания операционной среды ПС в МНТП обнаружила целый пласт проблем, которые необходимо решить. Наряду с известными проблемами, связанными с информационной безопасностью распределенных ресурсов, их федеративным администрированием и другими, выявилась проблема несоответствия структуры «тяжелой» вычислительной задачи предлагаемым в сети Grid разнородным вычислительным ресурсам. Это несоответствие может возникать как вследствие сложности распараллеливания вычислительных алгоритмов, так и в связи с распределением огромных объемов, необходимых для вычислений данных по элементам вычислительного ресурса.

Одна из целей исследований при создании ПС в МНТП заключалась в разработке архитектуры ИВР и проверке адекватности сервис-ориентированной архитектуры Grid для реализации высокопроизводительных вычислений при решении задач молекулярной динамики, докинга и сравнительного генетического анализа. В качестве тестовой задачи биомедицинской направленности был выбран сравнительный анализ генетических последовательностей. Пул вычислительных ресурсов ПС в МНТП при проведении исследований включал в себя Windows-кластер на основе локальной сети ИМБ, имеющий выход на гигабитную магистраль сегмента RGrid, а также два различных Linux-кластера регулярной структуры в ИХФ и Телекоммуникационном центре «Наука и общество». Программное обеспечение, реализующее алгоритм сравнительного генетического анализа BLAST, является международным стандартом де-факто в геномике, распространяется бесплатно Национальным центром биотехнологической информации (США) и доступно для всех основных вычислительных платформ. Особенность данной задачи состоит в использовании в процессе анализа гигантской генетической базы данных, например, SWISS-PROT. Эффективность использования вычислительных ресурсов в данной задаче зависит, главным образом, от соответствия распределения частей этой базы и поступающих счетных задач по узлам вычислительного ресурса. Эксперименты с тестовыми генетическими последовательностями показали, что для эффективного использования вычислительных элементов необходим предварительный анализ поступающих заданий и частей генетической базы данных, имеющихся в данный момент на вычислительных элементах, с целью максимально возможного согласования заданий и данных. В противном случае накладные расходы на пересылку данных делают неэффективным использование вычислительных кластеров. В качестве механизма адаптации узлов междисциплинарного сегмента Grid-сети к конкретным вычислительным задачам для повышения эффективности их решения предложены и исследованы концепция и алгоритмы локальных библиотек стратегий [10]. Эти стратегии представляют собой набор алгоритмов для локальной системы управления заданиями (например, PBS), заранее разработанных для оптимизации определенного класса задач на конкретном вычислительном ресурсе и учитывающих его архитектурные особенности, производительность отдельных элементов и другие факторы. С помощью заранее разработанных предметно-ориентированных алгоритмов управления распределением задач и данных по вычислительным элементам Linux-кластера под управлением PBS суммарное время выполнения задачи генетического анализа удалось сократить на 30—60 % в зависимости от объема требуемой части базы SWISS-PROT.

Библиотеки стратегий управления распределением потоков задач и данных могут быть применены аналогичным образом и на уровне Grid-сегмента. Условием их применения является обеспечение системы управления ресурсами корректной информацией о возможностях узлов данного сегмента. В технологии Grid для сбора и предоставления такой информации имеются специальные сервисы (GIIS, GRIS). Однако, разработка методов управления Grid-сегментами оказалась белым пятном при создании концепции Grid, и исследования в этой области находятся в начальной стадии.

Наиболее важная задача информационной поддержки исследований в рамках МНТП заключается в организации совместной работы групп исследователей в общем информационном пространстве. Эта задача достаточно общая для такого рода проектов. Например, схожая задача решается в рамках проекта LHC-Grid под руководством CERN (lcg.web.cern.ch/lcg). Для большинства проектов, в том числе и LHC-Grid, характерно существенное единообразие элементов информационного пространства. Несмотря на разбиение на некоторое число уровней иерархии, в целом в рамках проекта, как правило, соблюдается единообразие применяемых методов организации хранения и поиска данных. В рамках МНТП такое единообразие невозможно вследствие большого числа исследовательских коллективов, члены которых принадлежат к различным организациям, а решаемые задачи радикально разнятся как по методам использования внешней и собственной информации, так и по степени подготовленности исследователей к использованию новейших достижений информационных технологий. В связи с этим, сервис-ориентированная операционная среда на основе технологий Grid представляется наилучшим решением для обеспечения связности информационного пространства при биомедицинских исследованиях. Для объединения информационных ресурсов МНТП, представляющих собой разнородные базы данных и информационные массивы, на соответствующих локальных компьютерных системах развернуты комплексы программного обеспечения промежуточного уровня на основе пакета Globus Toolkit, интегрирующие эти ресурсы в сегмент RGrid. Для по-

иска и представления данных из разнородных источников разрабатываются соответствующие Grid-сервисы. Интеграция всего информационного пространства сегмента RGrid осуществляется с помощью набора Grid-сервисов, реализующих связывание элементов распределенных информационных ресурсов на основе предметно-ориентированных онтологий. Такая концепция, активно разрабатываемая в Grid-сообществе, позволяет бесконфликтно накладывать на множество разнородных информационных ресурсов неограниченное число связей в соответствии с задачами и предпочтениями каждой предметной области и даже отдельной группы исследователей (ВО). Grid-сервисы для создания, хранения и представления онтологий реализуют семантическое связывание данных и дают исследователям возможность динамично формировать и развивать многофакторные модели изучаемых объектов и процессов, а также своей предметной области в целом [11].

ЗАКЛЮЧЕНИЕ

Опыт проводимых с 2000 г. работ по созданию российского сегмента RGrid как элемента трансевропейской сети DataGrid и работ по формированию виртуальных организаций и понятийной сети МНТП «Вакцины нового поколения и медицинские диагностические системы будущего» в операциональной среде сегмента RGrid показал перспективность технологии Grid и концепции OGSA при решении вычислительных задач большой размерности и интеграции разнородных распределенных баз данных в единое информационное пространство [12]. Одновременно был выявлен ряд нерешенных в рамках Grid-технологий проблем, ключевые из которых, на наш взгляд, следующие:

— управление потоками задач и данных в среде разнородных распределенных федеративно-администрируемых ресурсов;

— создание согласованных систем метаданных для описания содержания и структуры хранилищ информации.

Существующие средства управления заданиями (например, в middleware Globus Toolkit) при выборе вычислительного ресурса в качестве критерия используют соответствие параметров, указанных в задании, и статических характеристик ресурса. В такой ситуации невозможно говорить о каких-либо оценках эффективности использования ресурсов сети в целом. Создание методов и средств управления Grid-сетями требует разработки моделей функционирования Grid-сети в целом, а также методов и моделей для количественного оценивания параметров (особенно динамических) ресурсов, потоков задач и данных в Grid-сетях, например, с привлечением методов теории массового обслуживания.

Разработка методов и средств создания, хранения и анализа метаданных для распределенных хранилищ информации является сегодня областью интенсивных исследований. Однако, сравнение различных систем метаданных и средств их реализации затруднено вследствие размытости критериев оценки и эмпирической природы этих оценок. Для исправления ситуации необходима, по-видимому, разработка принципов и методов формализованного оценивания не только скорости доступа к данным, но также и ценности данных, извлекаемых из хранилища информации, в том числе с использованием при этом метаданных.

Создание моделей ресурсов Grid-сетей и методов формализованного оценивания их статических и динамических параметров позволит целенаправленно разрабатывать адекватные задачам методы и средства управления ресурсами Grid-сетей и потоками задач и данных.

ЛИТЕРАТУРА

1. *Foster I., Kesselman C.* The Grid: Blueprint for a New Computing Infrastructure. — San Francisco: Morgan Kaufmann Pub., 1999.
2. *Жучков А. В., Ильин В. А., Кореньков В. В.* Некоторые аспекты создания глобальной системы распределенных вычислений в России // Тр. Всерос. науч. конф. «Высокопроизводительные вычисления и их приложения» / Черноголовка, 2000. — С. 227—231.
3. *Peitsch M.* Knowledge management and informatics in drug discovery // Drug Discovery Today: BIOSILICO. — May 2004. — Vol. 2, iss. 3. — P. 94—96.
4. *Butler D.* The Grid: tomorrow's computing today // Nature. — 2003. — P. 799—800.
5. *Арнаутов С. А., Жучков А. В.* Цифровые библиотеки в распределенной среде // Открытые системы. — 2001. — № 2. — С. 46—48.
6. *Жучков А. В.* Организация информационной поддержки МНТП «Вакцины нового поколения и медицинские диагностические системы будущего» через глобальные информационные сети // Аллергия, астма и клиническая иммунология. — 1999. — № 9. — С. 136—138.
7. *Создание* и развитие информационных ресурсов корпоративной сети МНТП «Вакцины нового поколения и медицинские диагностические системы будущего» / А. В. Жучков, С. В. Голицын, Н. В. Твердохлебов, А. К. Яновский // Аллергия, астма и клиническая иммунология. — 2003. — № 9. — С. 216—218.
8. *От информационной* системы проекта (учреждения) к электронной библиотеке в понятийной сети / А. В. Жучков, Н. В. Твердохлебов, С. А. Арнаутов, С. В. Голицын // Тр. V Всерос. конф. «Технологии информационного общества — Интернет и современное общество» IST/IMS-2002 / СПб., 2002. — С. 46—49.
9. *Система* распределенного хранения и анализа геномной информации / А. А. Черный, К. А. Трушкин, В. А. Боковой и др. // Молекулярная биология. — 2004. — Т. 38, № 1. — С. 104—109.
10. *Жучков А., Твердохлебов Н., Яновский А.* Локальные библиотеки стратегий как элемент управления ресурсами Grid // Тр. II Междунар. конф. «Параллельные вычисления и задачи управления» PACO'2004 / Ин-т пробл. упр. — М., 2004.
11. *Grid-Based* Onto-Technologies Provide an Effective Instrument for Biomedical Research / A. Joutchkov, N. Tverdokhlebov, I. Strizh, et al. // From Grid to HealthGrid. Studies in Health Technology and Informatics. Ed. by T. Solomonides, R. McClatchey. — London: IOS Press, 2005. — P. 37—46.
12. *Libraries* of Strategies and Ontology-driven Subject Area Models as «Corner Stones» in Grid Development / A. Joutchkov, N. Tverdokhlebov, A. Yanovsky, et al. // Methods of Information in Medicine. — 2005. — Vol. 44, N 2. — P. 249—252.

☎ *(495) 135-78-46*

E-mail: nickhard@chph.ras.ru ❑