ПЛЕНАРНЫЕ ДОКЛАДЫ

ВОПРОСЫ СОЗДАНИЯ И ЭКСПЛУАТАЦИИ БОЛЬШИХ КОРПОРАТИВНЫХ НАУЧНО-ОБРАЗОВАТЕЛЬНЫХ ИНФОРМАЦИОННО-ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ (НА ПРИМЕРЕ СЕТИ ПЕРЕДАЧИ ДАННЫХ СО РАН)[1]

Шокин Ю.И.
Институт вычислительных технологий СО РАН, Россия, dir@ict.nsc.ru

В докладе обсуждаются различные аспекты и проблемы развития информационной инфраструктуры Сибирского отделения РАН, наличие которой является необходимым условием обеспечения эффективности научных исследований.

Сибирское отделение РАН является региональным объединением научно-исследовательских, опытно-конструкторских, производственных организаций, а также подразделений, обеспечивающих функционирование инфраструктуры научных центров, расположенных на территории Сибири в семи областях, двух краях и четырех республиках (общая площадь территории около 10 млн. кв. км). Научные центры СО РАН находятся в Новосибирске, Томске, Красноярске, Иркутске, Якутске, Улан-Удэ, Кемерово, Тюмени, Омске, отдельные институты работают в Барнауле, Чите, Кызыле. В состав СО РАН входят более 50 научно исследовательских учреждений, работающих в области физико-математических, технических, химических и биологических наук, наук о Земле, гуманитарных и экономических наук. Примерно половина потенциала Отделения сосредоточена в Новосибирском научном центре.

Интеграция информационных и вычислительных ресурсов в единую среду и организация доступа к ним является одним из важнейших направлений развития современных информационных технологий. Стремительное развитие глобальных компьютерных сетей ведет к изменению фундаментальных парадигм обработки данных вследствие необходимости поддержки и развития распределенных информационно-вычислительных ресурсов. Отметим, что основным направлением работ, определенным правительством РФ, в рамках критических технологий является «создание инфраструктуры, оборудования, алгоритмического и программного обеспечения для инфокоммуникационных систем и создание взаимоувязанной системы стандартов обработки, хранения, передачи и защиты информации».

В настоящее время необходимость разработки механизмов, обеспечивающих функционирование общей информационной инфраструктуры, является приоритетным направлением для задач информационной поддержки научных исследований. Эти вопросы приобретают особую важность для такой организации как Сибирское отделение РАН, в условиях, когда различные группы исследователей, разделенные географически, должны осуществлять совместную работу, обмен данными и знаниями и координировать свои действия с целью оптимизации использования информационно-вычислительных ресурсов, сервисов и приложений.

Для крупного территориально распределенного научного центра, каким является Сибирское отделение, – это один из наиболее действенных способов интеграции научных коллективов и применения результатов их исследований в образовании. Острота вопроса наиболее ощущается в крупных интеграционных проектах и при проведении мультидисциплинарных исследований. Здесь информационные технологии играют

[1] Работа выполнена при поддержке РФФИ (гранты № 08-07-00229 и 09-07-00103), президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН № 4, 116 и 121.

определяющую роль в разработке базовой инфраструктуры исследований, в том числе и при «публикации» результатов исследований.

За годы своего существования информационная инфраструктура СО РАН (сеть передачи данных СПД СО РАН) по числу пользователей и компьютеров, по объемам передаваемых данных, по количеству и качеству накопленных информационных ресурсов, наконец, по разнообразию и качеству предоставляемых услуг превратилась в крупнейшую корпоративную научно образовательную сеть России. В СПД СО РАН зарегистрировано около 150 организаций абонентов. Только в Новосибирске сеть обслуживает более 40 000 пользователей и насчитывает более 12 000 подключенных компьютеров. Кроме того, в региональных научных центрах Отделения находится еще около 30 000 пользователей.

Характерной особенностью информационной инфраструктуры Отделения является наличие огромного количества аппаратно-технических средств. Последние различаются не только по производителю и характеристикам, но и по платформам и технологиям. Объясняется это несколькими причинами: аппаратура приобреталась в разное время; внедрение производилось разными специалистами, которые использовали разные технологии построения информационной инфраструктуры; инфраструктура развивалась и в географическом плане путем присоединения корпоративных сетей региональных научных центров. Такая историческая многоцветность создает серьезные проблемы для информационной совместимости (интероперабельности) ресурсов и для обеспечения системы безопасности СПД.

Современные требования по организации информационной инфраструктуры ориентированы на создание виртуальной среды и системы необходимых сервисов, требующих создания управляющих информационных центров, ответственных за взаимодействие ресурсов, интегрированных в информационную среду. Кроме функций поддержки виртуальной среды, соответствующих сетевых и мультимедийных сервисов, а также управления, синхронизации и диспетчеризации ресурсов СПД, необходимо обеспечивать функции мониторинга за всеми компонентами систем, контроль их параметров, проверку на соответствие, сбор статистики и т.д. Реализация подобных сервисов требует формализации моделей и методов, а также формирования правил (политик) доступа к существующим ресурсам тех или иных проектов.

Для успешного решения большинства задач, связанных с интеграцией информационных ресурсов, необходимы исследования в следующих направлениях [1]:

- разработка стандартов и моделей (профилей) представления информации и метаинформации;
- построение систем доступа к распределенным и разнородным коллекциям (интероперабельность, масштабируемость, обнаружение релевантной информации, интеграция метаинформации);
- разработка интерфейсов пользователей, визуализация и анализ данных;
- анализ и обработка естественного языка, изображений, видео- и аудиоданных;
- поддержка многоязыкового доступа к данным и обслуживание данных на нескольких языках;
- разработка мобильных технологий и «интеллектуальных» агентов;
- разработка алгоритмов автоматической классификации информации, методов и средств поиска, каталогизации, индексирования, а также поддержка целостности и непротиворечивости коллекций, безопасность и защита информации;
- решение вопросов интеллектуальной собственности.

Несмотря на значительные успехи исследований по многим из перечисленных направлений, сдерживающими факторами формирования единого (виртуального) информационного пространства являются:

- иерархичность информационных систем и ресурсов;

- разнородность ресурсов и программно-технических сред, объединяемых в едином сетевом операционном пространстве;
- распределенность элементов информационной инфраструктуры.

Развитие информационных сетей ведет к изменению фундаментальных парадигм работы с информационными ресурсами, в частности, становятся актуальными переход к распределенным ресурсам и создание инфраструктуры для их интеграции в единую информационную систему, обеспечивающую прозрачный доступ к распределенной информации и вычислительным ресурсам.

Любая, в том числе и распределенная, информационная система должна выполнять основные функции, вытекающие из ее основного предназначения, а именно:

- организация хранения информации (организация хранилищ, поддержка систем хранения данных);
- управление информацией (добавление, модернизация, изменение данных);
- управление доступом к информации (контроль исполнения правил регламентации доступа к данным), идентификация данных;
- поиск информации;
- извлечение информации и предоставление ее пользователю в необходимом ему виде;
- визуализация информации в соответствии с требованиями пользователя.

Распределенность и гетерогенность информационных ресурсов налагает следующие дополнительные требования к информационным системам [1]:

- способность систем функционировать в условиях информационной и реализационной неоднородности, распределенности и автономности информационных ресурсов;
- обеспечение интероперабельности, повторного использования неоднородных информационных ресурсов в разнообразных применениях;
- возможность объединения систем в более сложные интегрированные образования, основанные на интероперабельном взаимодействии компонентов;
- осуществление миграции унаследованных систем в новые системы, соответствующие новым требованиям и технологиям при сохранении их интероперабельности;
- обеспечение более длительного жизненного цикла систем.

Информационная инфраструктура любого уровня включает информационные, вычислительные и телекоммуникационные ресурсы. При формировании интегрированной среды основным принципом является функциональная стандартизация или построение функционального стандарта – профиля согласованного набора стандартов и нормативных документов, в котором в формализованном виде зафиксированы потребности в информационных технологиях, обеспечивающих решение какой-либо задачи или класса задач [2]. С учетом этих требований создание развитой инфраструктуры для представления и обмена метаданными является одним из приоритетных направлений формирования единого информационного пространства и совершенствования современной глобальной информационной сети. В настоящее время многие информационные центры, занимающиеся сбором и распространением метаданных, проявляют активную заинтересованность в организации взаимодействия с целью обмена имеющимися у них ресурсами. Как правило, в основе такой интеграции лежит выработка стандарта на форматы представления метаданных с одновременной унификацией массивов нормативно-справочной информации (разработка профиля информационной инфраструктуры [2]).

При разработке профилей необходимо учитывать постоянное появление новых быстро эволюционирующих типов ресурсов (например, мультимедийные ресурсы, интерактивные сервисы сети, электронные модели объектов, электронные карты, телеконференции, электронные коллекции и т.п.), разработка стандартов для которых в силу их динамической

природы и новизны не успевает за темпами развития данных предметных областей. Отметим, что основу интеграции ресурсов составляют технологии работы с метаданными [3], которые:

- обеспечивают механизмы интеграции информационных ресурсов из разных источников сведениями о свойствах этих ресурсов;
- являются источниками сведений о свойствах и содержании информационных ресурсов для механизмов управления данными в информационных системах;
- представляют сведения о системе, ее информационных и других ресурсах для различных приложений и пользователей системы;
- являются источником информации, необходимой для осуществления реинжиниринга информационных систем.

Появление распределенных информационных систем было обусловлено развитием сетей передачи данных, больших корпоративных сетей и глобальной сети Интернет. Задачи распределенных (как и обычных) информационных систем – хранение информации и предоставление ее пользователям в удобном для них виде. Как правило, такие системы могут быть организованы на основе различных технологических решений, направленных на реализацию той или иной парадигмы распределенности. Исходя из основных функций информационных систем, можно рассматривать различные ее аспекты:

- распределенное хранение информации (распределенные хранилища, сетевые системы хранения данных, сетевые файловые системы);
- распределенные СУБД и распределенные высокопроизводительные ресурсы;
- управление доступом к распределенным ресурсам и распределенное управление информационными ресурсами;
- поиск информации и информационных ресурсов;
- извлечение информации;
- визуализация информации из распределенных (разнородных) источников в единых пользовательских интерфейсах.

Переходя к обсуждению конкретных технологических решений, придется вернуться к основным функциям любой информационной системы, сформулированным выше. Несомненно, последняя из них проще всего может быть реализована на основе WEB-технологий. Далее, две предпоследние функции наиболее просто могут быть реализованы в технологиях, связанных с протоколом Z39.50, так как этот стандарт содержит почти все необходимые для это компоненты. Наконец, для реализации функций управления наиболее подходят технологии, основанные на LDAP (упрощенный протокол доступа к каталогам), поскольку именно на его основе сегодня проще всего организовать идентификацию, аутентификацию и авторизацию пользователей в распределенных информационных системах. Немаловажным обстоятельством при этом является тот факт, что LDAP основан на идеологии распределенного хранения информации (деревьев каталогов) на фоне глобальной идентификации всех элементов каталогов, содержит внутри себя определения механизмов и процедур репликаций данных между различными серверами и очень хорошо поддерживается разработчиками прикладного и системного программного обеспечения. Последнее позволяет достаточно просто переходить от локального управления информационными системами и контроля доступа к их ресурсам к распределенному [4].

Современные тенденции развития информационных технологий в организациях, имеющих разветвленную инфраструктуру и многофункциональные информационные системы, требуют создания экономичных и технологически продвинутых решений, позволяющих наиболее эффективно обрабатывать и хранить информацию, обеспечивать ее доступность и защищенность, а также эффективность использования приложений.

Отметим, что большинство «владельцев» информационных и вычислительных ресурсов, к числу которых относятся и учреждения СО РАН, формируют информационные ресурсы исходя из принципа их приватности и ориентированности на внутреннее использование. В результате большие объемы формально публичной информации

труднодоступны или недоступны внешним потребителям, что, в частности, сдерживает мультидисциплинарные исследования и может приводить к курьезным ситуациям. Нередко легче получить научный результат заново, чем узнать о его наличии и получить к нему доступ. Схожие проблемы существуют и в отношении вычислительных ресурсов, которые вследствие отсутствия надлежащих сервисов являются труднодоступными для большинства потенциальных пользователей и поэтому часто остаются недогруженными, в то время как потенциальным потребителям приходится довольствоваться своими локальными ресурсами, плохо приспособленными для проведения вычислений.

Как уже отмечалось, здесь наиболее экономически адекватным и востребованным решением является создание Центров обработки данных (ЦОД), позволяющих аккумулировать мощные вычислительные ресурсы и системы хранения информации, а также резко сократить затраты на обслуживающий персонал и сервисные услуги. Такие Центры могут предоставлять информационно-вычислительные ресурсы как непосредственно через развитую систему сервисов, так и в среде «облачных вычислений». Суть последних состоит в том, чтобы поместить имеющиеся ресурсы в виртуальное «вычислительное облако»' так, что доступ к ним можно было бы осуществлять из любого места по мере необходимости. Например, к услугам «облака» могут обратиться удаленный сервер института в пиковый период или рабочая станция в лаборатории, на которой запустили научное приложение, требующее серьезных вычислительных ресурсов. Инфраструктура вычислительного облака обеспечивает гибкое маневрирование ресурсами и их оптимальной загрузки и детального учета объема потребляемых услуг, предоставляемых ЦОД.

В Институте вычислительных технологий создан прототип Центра обработки данных корпоративной распределенной информационной системы, основанной на стандартных протоколах Z39.50, HTTP, LDAP, проведен предварительный этап его опытно промышленной эксплуатации. На базе ЦОД осуществляется доступ к системе хранения данных объемом 70 Тбайт и использования высокопроизводительного кластера для обработки данных. Создано хранилище, которое регулярно пополняется оперативными данными SPOT 2/4 (по прямому каналу из Зап-СибРЦПОД) и включает архивные данные со спутников серии LandSat на территорию РФ за 1982 – 2002 гг., и создан каталог метаданных, через который осуществляется доступ к информации. Создание прототипа ЦОД еще раз показало необходимость разработки профиля информационной инфраструктуры СО РАН для поддержки фундаментальных исследований.

В заключение остановимся на двух моментах. Во-первых, отметим, что создание информационной инфраструктуры и связанных с ней систем управления ресурсами и информационной безопасности требует значительных людских и материальных затрат. Во-вторых, цель создания информационной инфраструктуры – обеспечение конечного пользователя необходимой информацией и информационно-вычислительными ресурсами.

Литература

1. Жижимов О.Л., Федотов А.М., Чубаров Л.Б., Шокин Ю.И. Технология создания распределенных информационно-вычислительных ресурсов СО РАН // Тр. I Междунар. конф. САИТ-2005. «Системный анализ и информационные технологии». Переславль-Залесский, 2005. Т. 2. 161–165.
2. IEEE Std 1003.23-1998, IEEE Guide for Developing User Organization Open System Environment (OSE) Profiles.
3. Force on Metadata. Summary Report // American Library Association. 1999. Vol. June. 8
4. Барахнин В.Б., Жижимов О.Л., Степанов Ю.Ю., Федотов А.М. LDAP-каталог организации как ядро корпоративной распределенной информационной системы // Инновационные недра Кузбасса. IT-технологии: Сб. науч. тр. Кемерово: ИНТ, 2008. С. 226 –232.