Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 449–466.
URL: http://www.matbio.org/2013/Oplachko_8_449.pdf

==========ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
==========ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ==========

УДК: 004.75: 004.738.5: 004.942: 577

Облачные технологии и их применение в задачах вычислительной биологии

Оплачко Е.С.[4,1*], Устинин Д.М.[2,1], Устинин М.Н.[1,3***]**

[1]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия*
[2]*Биологический факультет, Московский государственный университет им. М.В.Ломоносова, Москва, 119234, Россия*
[3]*Пущинский государственный естественно-научный институт, Пущино, Московская область, 142290, Россия*
[4]*ООО "Майкрософт Рус", Москва, 121614, Россия*

Аннотация. Описано понятие облачных вычислений, приведена их классификация, рассмотрены основные архитектурные особенности. Рассмотрены крупнейшие облачные проекты на рынке информационных технологий. Описаны несколько проектов в области вычислительной биологии, использующих облачные вычисления. Подробно рассмотрен проект «Математическая клетка», дана его классификация как облачного ресурса. На примере прямого моделирования внутриклеточных процессов описано использование программного обеспечения и вычислительных ресурсов, предоставляемых пользователям в этом проекте.

Ключевые слова: *технологии облачных вычислений, вычислительная биология, компьютерное моделирование живой клетки.*

ВВЕДЕНИЕ

Облачные вычисления представляют собой сервис, обеспечивающий удаленный доступ пользователя к аппаратным мощностям или программному обеспечению. В качестве прототипа облачных технологий можно рассматривать сервисы электронной почты, такие, как Gmail или Hotmail [1], которые позволяют хранить на своих серверах все письма, персональные данные, файлы и программное обеспечение почтового клиента. Когда пользователю необходимо воспользоваться почтовым сервисом, он просто открывает веб-браузер, переходит на страницу почтового клиента и авторизуется. Возможность использования почтового сервиса обуславливается только наличием доступа в Интернет, то есть пользователь оказывается географически свободен. Кроме того, он не тратит свои аппаратные ресурсы на хранение программного обеспечения и результатов своей работы [2].

Со временем область применения облачных технологий существенно расширилась. Причиной этого стало бурное развитие компьютерных мощностей и линий связи [3, 4]. Создание масштабных вычислительных комплексов и центров хранения данных, а также развитие телекоммуникаций привело к возможности дистанционного предоставления услуг в области информационных технологий. Облачные технологии позволяют пользователю получить доступ к необходимой информации в любое время и из любого места, что избавляет от необходимости заботиться о собственных

*eopl@microsoft.com
**dmitry.ustinin@gmail.com
***ustinin@impb.ru

устройствах хранения информации. При этом организация, предоставляющая облачный сервис, может иметь не только необходимые пользователю аппаратные ресурсы, но и программное обеспечение. Этот факт является привлекательным для тех пользователей, которые не могут приобрести много оборудования и лицензий на программное обеспечение. Университеты, исследовательские организации, малые и средние компании и индивидуальные пользователи могут хранить свои данные на облачных ресурсах. По мере роста или снижения потребностей, пользователь может легко увеличивать или уменьшать необходимый для работы объем ресурсов, тем самым контролируя свои расходы на информационные технологии. Еще одним преимуществом облачных вычислений является доступность информации для всех сотрудников организации из любого места: имея персональный компьютер или мобильное устройство и доступ в Интернет, можно получить доступ к документам и программному обеспечению и вести работу, требующую коллективных усилий.

Сейчас можно увидеть удаленное использование облачных ресурсов как для рутинных операций среднего и малого бизнеса, так и для крупных научных проектов. На рынке информационных технологий можно найти немало предложений в сфере облачных вычислений и хранения данных. Ниже мы рассмотрим некоторые из них, а также дадим обзор применения облачных технологий в вычислительной биологии.

КЛАССИФИКАЦИЯ ОБЛАЧНЫХ РЕСУРСОВ И ИХ ОСНОВНЫЕ ХАРАКТЕРИСТИКИ

Виды облаков

Выделяют несколько видов облачных ресурсов (далее "облаков") (см. рис.1):

1. Публичное облако – доступ к ресурсам осуществляется любым пользователем, имеющим подписку, из любого места, при условии наличия доступа в сеть Интернет.
2. Частное облако – ресурсы доступны только ограниченному числу лиц (например, сотрудникам одной компании).
3. Общественное облако – ресурсы доступны нескольким организациям, имеющим одинаковые потребности с точки зрения информационных ресурсов.
4. Гибридное облако – облако, состоящее из двух и более облаков разных видов, например, публичного и частного.

Конечные домашние пользователи или малый бизнес в основном используют публичные облака.

Рис. 1. Виды облачных ресурсов.

В зависимости от вида облака, им могут владеть и распоряжаться как провайдер, так и пользователь, или и тот и другой. Также могут различаться права доступа к ресурсам (см. таблицу 1).

Таблица 1. Обслуживание и управление различными видами облачных ресурсов

Вид облака	Кем обслуживается инфраструктура	Кто является владельцем инфраструктуры	Где находится инфраструктура	У кого имеется доступ
Публичное	Внешним провайдером	Внешний провайдер	У внешнего провайдера	У любого пользователя
Частное/ общественное	Пользователем или внешним провайдером	Пользователь или внешний провайдер	У внешнего провайдера или у пользователя	У авторизованного пользователя
Гибридное	Пользователем и внешним провайдером	Пользователь и внешний провайдер	У внешнего провайдера и у пользователя	У авторизованных и у любых внешних пользователей

Модели сервиса, предоставляемого в облаке

В зависимости от выбранной подписки, пользователь получает тот или иной набор услуг. Выделяют три основных модели обслуживания:

1. Программное обеспечение как сервис (Software as a Service, SaaS) – пользователю предоставляется доступ как к аппаратным ресурсам, так и к приложению, находящемуся на этих ресурсах. В данном случае пользователь избавляется от необходимости хранить программное обеспечение на своих ресурсах и делать их резервную копию. Более того, появляется возможность доступа к приложению с любого устройства без предустановки. Требуется только доступ в Интернет. При этом пользователь имеет минимум прав для управления приложением.
2. Платформа как сервис (Platform as a Service, PaaS) – данный вид подписки стоит на уровень ниже предыдущего. В данном случае пользователю предоставляются все необходимые компоненты из облака для разработки и эксплуатации программного обеспечения через Интернет.
3. Инфраструктура как сервис (Infrastructure as a Service, IaaS) – при данном виде соглашения пользователь получает набор аппаратных ресурсов, которые он может использовать в соответствии со своими потребностями.

Основные характеристики облачных технологий

Ресурсы, относящиеся к облачным технологиям, должны обладать такими качествами, как высокая доступность, легкая масштабируемость и быть экономически выгодными для подписчиков. С точки зрения инфраструктуры, возникают вопросы, чем отличаются облачные технологии от существовавших ранее подходов к предоставлению аппаратных и программных ресурсов.

Для ответа на эти вопросы Национальный институт стандартов и технологий США (National Institute of Standards and Technology, NIST) [5–7] определил облачные вычисления путем описания пяти основных характеристик:

1. Широкая сетевая доступность (Broad Network Access). Доступ к программному продукту или ресурсам можно осуществить как с традиционных компьютеров или ноутбуков, так и с планшетов и телефонов, воспользовавшись защищенным каналом через сеть Интернет.
2. Легкая масштабируемость (Rapid Elasticity). При необходимости пользователь может быстро подключить к работе дополнительные аппаратные или программные ресурсы.
3. Возможность мониторинга (Measured Service) – облачные системы построены таким образом, что аппаратные ресурсы динамически меняются, а нагрузка

балансируется незаметно для пользователя. При этом облако оснащено системой мониторинга и, как следствие, может быть оценено с точки зрения доступности и стабильности работы.

4. Самообслуживание (On-Demand Self-Service) – при необходимости пользователь может дополнить или изменить набор используемых ресурсов без непосредственного контакта с сервис провайдером.

5. Объединение ресурсов (Resource Pooling) – облачные технологии подразумевают динамическое изменение количества используемых аппаратных ресурсов. При этом сервис-провайдер также может изменять аппаратную часть облака (хранилище данных, оперативная память, процессоры, сетевые компоненты), при этом пользователь не заметит этих изменений.

Помимо пяти основных характеристик, описанных NIST, стоит отметить, что одна из основных технологий, на которых организуются облачные вычисления – это гипервизоры (технологии виртуализации). Данный механизм позволяет запускать несколько операционных систем на одном компьютере, называемом хостом. При этом создаются виртуальные машины на хосте, которые, являясь изолированными друг от друга, делят между собой одни и те же аппаратные ресурсы [8].

Более того, IaaS (как наиболее близкий к аппаратной части системы уровень) предоставляет так называемые Интерфейсы программирования приложений (Application Program Interface, API) которые позволяют пользователю на этом и других уровнях управлять и взаимодействовать с системой.

На рис. 2 показана обобщенная классификация облачных технологий и их основные характеристики.

Рис. 2. Описание облачных технологий по стандартам NIST [5].

Одним из основополагающих качеств облачных технологий является тот факт, что одни и те же ресурсы используются различными пользователями. На рис. 3 иллюстрируется структура облака, в котором одни и те же ресурсы используются в разных моделях сервиса.

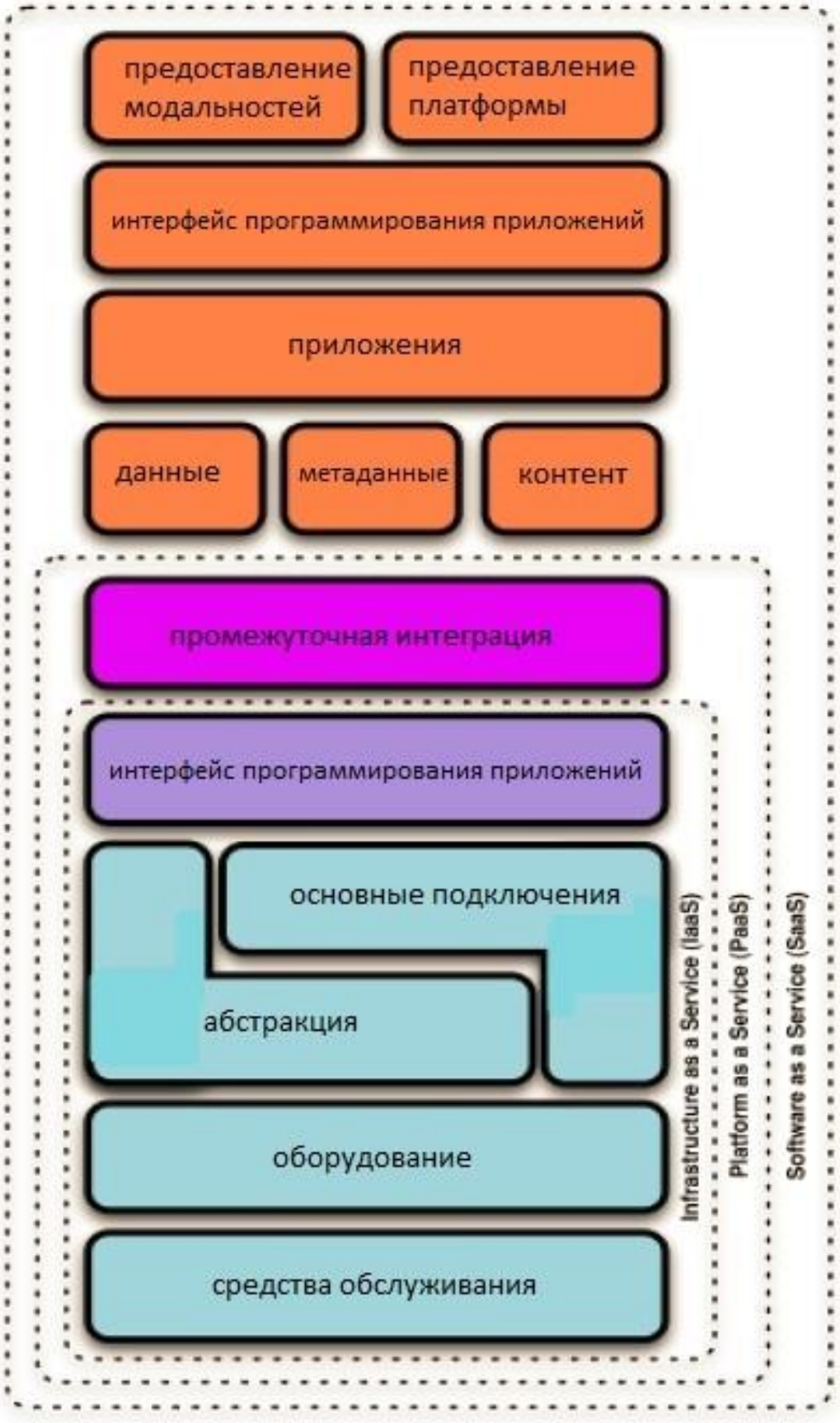

Рис. 3. Структура облака с учетом использования одних и тех же ресурсов для разных целей.

Безопасность информации в облачных ресурсах

Защита данных является важной проблемой в информационных технологиях. Особенно она актуальна для ресурсов облачного типа, предоставляемых дистанционно широкому кругу клиентов. С одной стороны, использование одних и тех же компьютеров и программного обеспечения для разных целей разными пользователями (см. рис. 3) является экономически обоснованным решением. С другой стороны, подобный подход требует повышенного внимания к безопасности, разграничению прав, изолированию данных и программных продуктов, а также к балансировке нагрузки на аппаратную часть.

Рассмотрим проблемы информационной безопасности для разных моделей сервиса в облаке [9]. Прежде всего, следует отметить «вложенность» типов подписки: IaaS-PaaS-SaaS. Соответственно, при уязвимости на самом низком, аппаратном уровне (IaaS), проблемы будут наследоваться и на более высокие слои, вплоть до приложений. "Платформа как сервис" (PaaS) находится фактически поверх "Инфраструктуры как сервис" (IaaS) и сопровождается при этом дополнительными компонентами, включая возможность интеграции с приложениями, антивирусную защиту, базы данных, системы обмена сообщениями и другие. Таким образом, при этой модели сервиса пользователь может разрабатывать, устанавливать и настраивать необходимые ему приложения с помощью языка программирования и инструментов, которые доступны в модели "Платформа как сервис" (PaaS). "Программное обеспечение как сервис" (SaaS) представляет собой замкнутую среду, с помощью которой пользователь получает доступ ко всем необходимым ему приложениям (например, почтовая система, программы для создания презентаций и др.), имея при этом ограниченные возможности управления этими программными продуктами, такие, как настройки интерфейса для удобства использования или задания параметров расчетов, но ни в коем случае не администрирование. Соответственно, уровень безопасности для SaaS выше, чем для PaaS, поскольку при предоставлении провайдером конечного продукта это приложение будет хорошо интегрировано с системой и все потенциальные риски безопасности будут учтены. При работе с PaaS, и тем более с IaaS, пользователь разрабатывает и использует свои приложения, и уровень безопасности зависит от качества используемого программного обеспечения.

Поскольку облачные технологии основаны на обработке и хранении информации вне организации–подписчика, то возникает законный вопрос о защите информации, предоставляемой и/или получаемой пользователем. При этом пользователю трудно определить уровень безопасности предоставляемого сервиса. Разные сервис–провайдеры имеют разные принципы и уровни обеспечения безопасности. С точки зрения защиты данных, поставщик ресурсов должен позаботиться о выборе методов шифрования, о надежности аппаратной части системы, о создании резервных копий данных, об использовании сетевого экрана, и, в случае использования общественного облака, о разграничении прав пользователей.

КРУПНЕЙШИЕ ПРОЕКТЫ НА РЫНКЕ ОБЛАЧНЫХ ТЕХНОЛОГИЙ

В этом разделе дается краткое описание некоторых облачных сервисов, предлагаемых крупнейшими компаниями.

Windows Azure

Сервис *Windows Azure* [10, 11] предлагается компанией Майкрософт (Microsoft). С точки зрения платформы Майкрософт обеспечивает доступ к высоко доступной, масштабируемой системе с высоким уровнем безопасности. При этом Azure обеспечивает управление сервисами в автоматическом режиме. Физически весь сервис расположен в нескольких хранилищах данных (Европе, Юго-Восточной Азии и

Северной Америке). При этом показатели мониторинга и состояние систем можно найти в открытом доступе в реальном времени по адресу: http://www.windowsazure.com/en-us/support/service-dashboard.

Облачная система Azure включает следующие компоненты (см. рис. 4):

1. Windows Azure – непосредственно сама операционная система. На данном уровне осуществляется управление мощностями и осуществляется хранение информации.

2. SQL Azure – база данных, которая может быть предоставлена для пользователя как сервис.

3. Windows Azure AppFabric - программный комплекс, который используются для контроля доступа, интеграции приложений облака и пользовательских данных, обеспечения коммуникации между модулями.

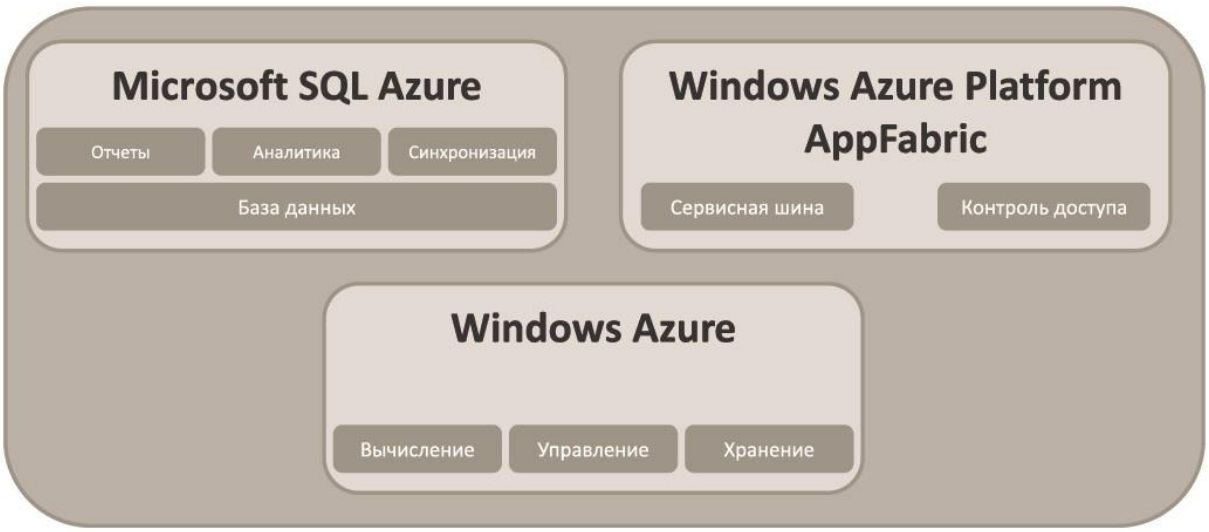

Рис. 4. Компоненты Windows Azure.

С точки зрения технической реализации система Azure основана на технологии виртуализации, схожей с гипервизором Hyper-V. Для организации работы виртуальных серверов используется Windows Azure Fabric Controller. Этот программный слой обеспечивает доступ пользователей к ресурсам, обеспечивает репликацию между географически распределенными серверами и балансировку нагрузки.

Windows Azure предоставляет следующие сервисы:

1. Контейнеры для приложений, которые позволяют помещать в облако приложения, написанные на .NET, Java, Ruby, PHP, Python.

2. Хранение данных. Пользователь может работать в облаке с таблицами, бинарными объектами, другими файлами.

3. Интеграция среды пользователя с сервисами Azure.

4. Сервисы, обеспечивающие безопасность хранения данных. Помимо стандартного разграничения доступа пользователям имеется возможность обеспечить интеграцию с уже существующими средствами аутентификации.

5. Сервисы, позволяющие пользователям разрабатывать собственные приложения, которые можно поместить в облако.

Помимо перечисленных выше возможностей, компания Майкрософт обратила свое внимание на использование облачных технологий для размещения научных разработок. Так возникло понятие "Исследование как сервис" (Research as a Service). Корпорацией был проведен опрос исследователей с целью выяснить, какие достоинства и недостатки видят научные работники в использовании облачных вычислений. 90% исследователей как основное достоинство указали экономию при использовании облаков для создания и размещения их программного обеспечения.

Например, сотрудники Лаборатории географии природных катастроф из Греции (Geography of Natural Disasters Laboratory) разработали приложение, которое может использоваться для моделирования распространения пожаров. Конечными пользователями такого программного обеспечения являются пожарные службы органов гражданской обороны.

Группа Дэвида Хекермана из Майкрософт разработала методику определения причин генетических заболеваний. Для анализа больших объемов данных было использовано 27000 вычислительных ядер на Windows Azure. Результаты этого масштабного исследования в области биоинформатики теперь доступны в качестве облачного ресурса на Windows Azure Marketplace.

Долгосрочная цель Майкрософт заключается в создании самостоятельной научной экосистемы, широко доступной в облаке в виде сервисов и включающей в себя как условия для хранения научных данных, так и их обработки. В дополнение к поддержке хранения данных и организации вычислений, опытные пользователи также будут иметь платформу для создания и продажи необходимых исследовательских услуг.

IBM Smart Business

Компания IBM также предоставляет свои решения для создания как частных, так и публичных облачных систем [12, 13].

Ключевыми спецификациями облаков от IBM являются:

1. Разработка и тестирование решений.
2. Хранение информации.
3. Предоставление результатов работы аналитических систем.
4. Архивирование информации.
5. Совместная работа пользователей с системами семейства LotusLive.

Amazon Web Service

Компания Amazon является одним из первых и ведущих поставщиков облачных технологий. В рамках инфраструктуры виртуальных серверов Amazon Web Service (AWS) [14] предоставляются следующие услуги:

1. Amazon Elastic Compute Cloud (Amazon EC2) – предоставление пользователю виртуальных машин на удаленных аппаратных мощностях с возможностью выбора операционной системы и технических характеристик, а также сетевых параметров (например, получение статического IP адреса).
2. Amazon Simple Storage Service (Amazon S3) – предоставление пользователю ресурсов для хранения информации.
3. Amazon DynamoDB [15] – база данных NoSQL.

Надо отметить, что на основе облачных сервисов Amazon реализованы многие системы, в том числе научный проект Cloud BioLinux.

НАУЧНЫЕ ПРОЕКТЫ В ОБЛАСТИ ВЫЧИСЛИТЕЛЬНОЙ БИОЛОГИИ, ИСПОЛЬЗУЮЩИЕ ОБЛАЧНЫЕ ТЕХНОЛОГИИ

Cloud BioLinux

Ресурс Cloud BioLinux [16] был создан для проведения исследований в области биоинформатики, требующих обработки большого количества данных.

Этот ресурс представляет собой общедоступную виртуальную машину, пользователи которой имеют доступ к ряду заранее сконфигурированных командных строк и графическому приложению, включая полнофункциональный пользовательский интерфейс, документацию и более 135 пакетов для приложений (выравнивание последовательностей, кластеризация, отображение, редактирование). Функциональные

возможности каждой утилиты подробно описаны в документации, доступной через графический интерфейс. В ходе выполнения проекта было создано публичное облако, использующее сервис Amazon EC2, которое является промышленной облачной платформой. Данную платформу разработала компания Amazon.com, для поддержки проекта используются центры обработки данных в США, Европе и Азии. Доступ к утилитам обеспечен через удаленное подключение к экземпляру облака Amazon EC2 с локального компьютера. Был разработан механизм, позволяющий исследователям без доступа к локальным вычислительным кластерам выполнить крупномасштабные вычисления путем аренды виртуальной машины за $0.085/час. Данная виртуальная машина обладает следующими минимальными характеристиками: одноядерный процессор, 1.7 GB оперативной памяти, 160 GB дисковой памяти.

Пользователь может получить доступ к ресурсу Cloud BioLinux через веб-браузер, имея доступ к сети Интернет, выполнив следующие операции:

– зарегистрироваться на сервисе Amazon EC2 и использовать консоль входа;

– в консоли с помощью мастера (используя кнопку "Запуск экземпляра") для запуска Cloud BioLinux и указать образ виртуальной машины «Cloud BioLinux», затем определить количество мощностей, требуемых для выполнения операций с данными;

– скопировать из окна веб-браузера в облако Amazon назначенный Интернет-адрес виртуальной машины ("Public DNS") и вставить его в настройки клиента удаленного рабочего стола. После того, как клиент удаленного рабочего стола установит соединение, пользователь получит доступ к полному рабочему столу с приложениями по биоинформатике.

В качестве альтернативы облаку Amazon, Cloud BioLinux может использовать облако с открытым кодом Eucalyptus или приложение Virtualbox для запуска программного обеспечения непосредственно на компьютере.

Science clouds

Проект «Science clouds» [17] представляет собой неофициальную группу небольших облачных сервисов, созданных в разных институтах на добровольной основе, выполняется с середины 2008 года. Цель проекта: предоставить исследователям облачную инфраструктуру для обработки экспериментальных данных (IaaS), а также обеспечить платформу для разработки приложений, отвечающую требованиям научного сообщества (PaaS).

В группу входят четыре отдельных проекта:

- Nimbus (Университет Чикаго, США);
- Stratus (Университет Флориды, США);
- Wispy (Университет Пердью, США);
- Kupa (Университет имени Масарика, Чехия).

Проект Nimbus представляет собой набор утилит с открытым кодом для предоставления IaaS-сервиса для научного сообщества. Разработчики проекта делали упор на следующие спецификации:

1. Предоставить поставщиков ресурсов для создания частного или публичного облака (IaaS). Служба Nimbus Workspace обеспечивает реализацию вычислительного облака, предоставляя пользователям в аренду вычислительные ресурсы и возможность развертывания виртуальных машин на этих ресурсах. Дополнительно осуществляется квотирование ресурсов хранения информации, что позволяет организовывать несколько реализаций облаков одновременно.

2. Предоставление сервиса IaaS. Для этого используется инструмент Nimbus Context Broker, который создает общий конфигурационный пул и контекст безопасности по ресурсам, предоставленным для нескольких облаков. Также пакет утилит Nimbus содержит инструменты масштабирования, позволяющие пользователям

автоматически работать через несколько распределенных провайдеров, что позволяет переключать задачи между частной и публичной облачными средами.

3. Nimbus позволяет разработчикам расширять, экспериментировать и менять под локальные нужды облачную инфраструктуру. Для достижения этой цели предлагается настраиваемая и масштабируемая реализация облака с открытым кодом. Так, служба Workspace может быть сконфигурирована для поддержки различных реализаций виртуальной среды, также возможно изменять варианты управления ресурсами, интерфейсы (в том числе обеспечивая совместимость с Amazon EC2), и другие параметры.

При реализации Nimbus использовалось программное обеспечение Torque с открытым кодом, которое позволяет управлять распределенными ресурсами для вычислительных кластеров. Оно отвечает за запуск облачных экземпляров, которые были добавлены в кластер, проверяет, выполняются ли назначенные задания и выключает экземпляры, не выполняющие никакой работы.

Однако, подобная организация работы имеет следующие ограничения:

– только определенное количество экземпляров может быть запущено в один момент времени;

– используется только один образ с предустановленным программным обеспечением для всех экземпляров;

– каждый экземпляр может обмениваться информацией с основным узлом, но не с другим экземпляром.

Для устранения этих ограничений были разработан и внедрен (в дополнение к используемым командам Torque) набор скриптов для автоматизации работы узлов кластера. Также был разработан механизм, допускающий обмен информацией между всеми компонентами системы.

Ресурс Wispy относится к компьютерному гриду «TeraGrid». Система представляет собой центр облачных вычислений, позволяющий исследователям создавать программные пакеты команд или виртуальные машины, которые могут быть запущены удаленно в заранее специально сконфигурированной среде. Кластер Wispy содержит 128 виртуальных машин, работающих одновременно.

Magellan

Проект Magellan [18] был разработан с целью создания научной среды, поддерживающей распределенные вычисления и анализ данных. Система содержит два сайта (NERSC и ALCF), которые обрабатывают и хранят большие объемы данных, также работает программное обеспечение, поддерживающее облачные вычисления с настроенным параллелизмом.

В рамках проекта было проведено исследование потенциальной роли облачных вычислений при решении различных задач. В ходе исследования были проработаны несколько вопросов, затрагивающих различные аспекты облачных вычислений, таких как производительность, удобство использования, стоимость. Был разработан стенд для тестирования различных вычислительных моделей с целью определения роли аппаратного обеспечения в работе различных научных приложений.

Основной целью проекта было продвижение открытого интерфейса для облаков, а также получение ответов на следующие вопросы:

- существует ли программное обеспечение с открытым кодом для использования на высокопроизводительных компьютерах;
- отвечает ли оно требованиям безопасности;
- полезен ли модуль для научных целей;
- могут ли уже существующие приложения быть перенесены в облачную среду;
- как облачная среда может быть полезна в науке в целом;

– какова финансовая выгода при использовании высокопроизводительных приложений для научных целей в облаке.

С помощью ресурсов проекта Magellan решается задача анализа геномов микробного сообщества. Выполняется запуск программы BLAST для идентификации сходства парных генов. Данный процесс может занимать до трех недель при работе на 256-ядерном кластере, организованном на ОС Linux. Вопрос производительности обостряется по мере роста базы данных. Программное обеспечение написано таким образом, что основные вычислительные операции распараллелены и не требуют взаимодействия между потоками, что сходно по принципу с работой облачных приложений. Таким образом, необходимость в ресурсах сделала облачные технологии привлекательной платформой для данного вычислительного процесса. Основная часть программы написана на языке программирования Perl с использованием компонентов Java, а так же компонентов, написанных на языках C и C++. Также используются базы данных ДНК (в среднем занимают 16 Гб памяти).

MathCell

Проект «Математическая клетка» [19, 20, 21] является разработкой Института математических проблем биологии РАН. В рамках проекта была создан программный ресурс, который рассматривает различные аспекты моделирования биологической клетки. Доступ к ресурсу осуществляется через веб-браузер (http://www.mathcell.ru). В ходе реализации проекта была создана трехмерная модель эукариотической клетки с возможностью навигации, наглядно представляющая взаимное расположение клеточных органелл.

В рамках проекта предлагается четыре расчетные модели:

- модель переноса заряда в ДНК;
- моделирование связывания белков электрон-транспортной цепи фотосинтеза;
- модель переноса электрона в фотосинтетической мембране;
- расчет энергии растворения биомолекул в воде методом Монте-Карло.

Пользователям также предлагаются обзоры по темам, связанным с моделированием клетки, глоссарий, призванный помочь математикам и программистам понять термины биологии и наоборот, а также список тематических статей со ссылками на оригинальные сайты и ресурсы.

ПРЯМОЕ МОДЕЛИРОВАНИЕ ВНУТРИКЛЕТОЧНЫХ ПРОЦЕССОВ В ОБЛАЧНОМ РЕСУРСЕ MATHCELL

Развитие математического моделирования, компьютерной техники и вычислительных методов сделало возможным детальное описание систем, состоящих из многих миллионов атомов, а успехи молекулярной биологии привели к накоплению большого объёма экспериментальных данных [21]. В результате стало возможно компьютерное моделирование процессов, происходящих в компартментах клетки, прямо учитывающее геометрию системы и свойства каждой из взаимодействующих частиц. В состав Mathcell входят две расчетные программы прямого моделирования, представляющие собой облачные ресурсы, работающие по модели "Программное обеспечение как сервис" (SaaS).

Моделирование связывания белков электрон-транспортной цепи фотосинтеза на примере докинга пластоцианина и цитохрома F

Белок-белковые взаимодействия являются основой большинства биологических процессов. Компьютерное моделирование динамики связывания белков дает важную информацию для понимания механизмов их функционирования. Разработан программный продукт [22], предназначенный для моделирования взаимодействия

макромолекул методом многочастичной броуновской динамики с учетом дальнодействующих электростатических взаимодействий. В модели молекулы белка рассматриваются как броуновские частицы, совершающие поступательное и вращательное движение в вязкой среде. Конформационная подвижность молекул не учитывается (молекулы рассматриваются как твёрдые тела), не рассматриваются также гидрофобные взаимодействия белков. В результате компьютерного моделирования может быть получена статистическая информация об образующихся диффузионно-столкновительных комплексах. Для заданного критерия образования белок-белковых комплексов могут быть получены кинетические параметры реакции – зависимость количества (концентрации) комплексов от времени и, следовательно, константа скорости реакции образования комплексов.

Модельная сцена представляет собой трехмерный реакционный объем, который может быть разделен на несколько компартментов. Каждый компартмент имеет форму прямоугольного параллелепипеда. Трехмерная модель молекул строится по данным из Protein Data Bank (PDB). Перед началом моделирования заданное количество молекул каждого белка распределяется определенным (по умолчанию случайным) образом в пределах отведенного им компартмента. Границы компартментов в зависимости от параметров модели могут быть либо периодическими, либо неупругими (молекула вплотную подходит к границе компартмента, с которой произошло столкновение).

Метод броуновской динамики учитывает взаимодействие частиц с молекулами растворителя за счет введения случайной силы и силы трения в уравнения движения. Взаимодействие между отдельными броуновскими частицами рассматриваются в явном виде. Электростатические взаимодействия являются дальнодействующими и влияют на скорость сближения и взаимную ориентацию макромолекул в процессе диффузии. Контактные взаимодействия проявляются при сближении частиц на расстояние порядка суммы Ван-дер-Ваальсовых радиусов входящих в них атомов и препятствуют взаимному проникновению молекул.

Расчетный модуль программного комплекса написан на C++ с распараллеливанием по реализациям с помощью MPI. Задание параметров в системе MathCell осуществляется через веб-интерфейс, который по данным пользователя формирует конфигурационные файлы расчетного модуля. Расчетный модуль сначала осуществляет предварительное моделирование распределения зарядов и электростатического поля для всех типов взаимодействующих частиц, после чего запускается основной цикл, моделирующий движение и взаимодействие частиц.

При использовании программы в системе MathCell пользователь задает через веб-интерфейс следующие параметры:

- PDB-файлы, описывающие структуру взаимодействующих белков. Если в процессе реакции изменяется окислительно-восстановительное состояние молекул (например, при передаче электрона с одного белка на другой), в предварительном расчете распределения зарядов моделируется конфигурация поля как для окисленного, так и для восстановленного белка каждого типа.
- Размер реакционного объема (длина параллелепипеда, ширина параллелепипеда, высота параллелепипеда)
- Начальное количество и концентрация молекул каждого типа в окисленном и восстановленном состоянии.
- Моделируемое время взаимодействия (время, число шагов)
- Ионная сила раствора

Pdb-файлы для взаимодействующих белков:		
PDB-файл для восстановленного пластоцианина (A.pdb_red):	Browse...	pc_e.pdb
PDB-файл для окисленного пластоцианина (A.pdb_ox):	Browse...	pc.pdb
PDB-файл для восстановленного цитохрома (B.pdb_red):	Browse...	cyt_e.pdb
PDB-файл для окисленного цитохрома (B.pdb_ox):	Browse...	pdb_ox
Размер реакционного объема:		
Длина параллелепипеда (Scene.xmin, Scene.xmax):	-109 109 мм	
Ширина параллелепипеда (Scene.ymin, Scene.ymax):	-109 109 мм	
Высота параллелепипеда (Scene.zmin, Scene.zmax):	-109 109 мм	

Рис. 5. Окно для ввода первичных данных.

После расчета пользователь получает кинетическую кривую, отображающую количество образовавшихся комплексов в зависимости от времени (рис. 6). Пользователь имеет возможность получить результат вычислений в виде текстового файла, а так же просмотреть конфигурационный файл.

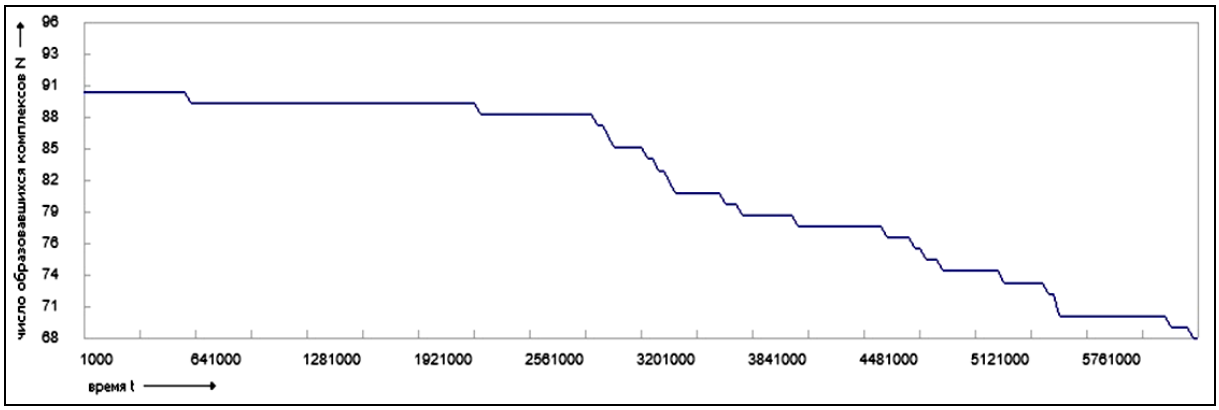

Рис. 6. График кинетической кривой, отображающей количество образовавшихся комплексов в зависимости от времени.

Модель переноса заряда в фотосинтетической мембране

В систему MathCell также включена модель фотосинтетической мембраны, построенная на основе метода многочастичной броуновской динамики [23]. В ней моделируется взаимодействие мобильных белков-переносчиков электрона с встроенными в мембрану пигмент-белковыми комплексами, перенос электрона внутри этих комплексов, перенос протонов через мембрану и диффузия их в межмембранном пространстве (люмене). Модель позволяет имитировать сигналы флуоресценции и электронного парамагнитного резонанса с нескольких реакционных центров системы. Расчетный модуль моделирует движение мобильных переносчиков методом броуновской динамики. Перенос электрона внутри комплексов учитывается с помощью решения системы кинетических уравнений с вероятностями состояний комплексов в качестве переменных [24], диффузия протонов моделируется конечно-разностным методом.

Расчетный модуль реализован на C++ и адаптирован для грид-систем. Выходные данные содержат координаты и состояние всех объектов системы (мобильных переносчиков, пигмент-белковых комплексов, протонов) на каждом шаге моделирования. Модуль трехмерной визуализации, написанный с использованием

OpenGL, позволяет просматривать состояние системы. Можно исследовать траектории отдельных частиц в условиях затрудненной диффузии в межмембранном пространстве, загроможденном белковыми комплексами [25].

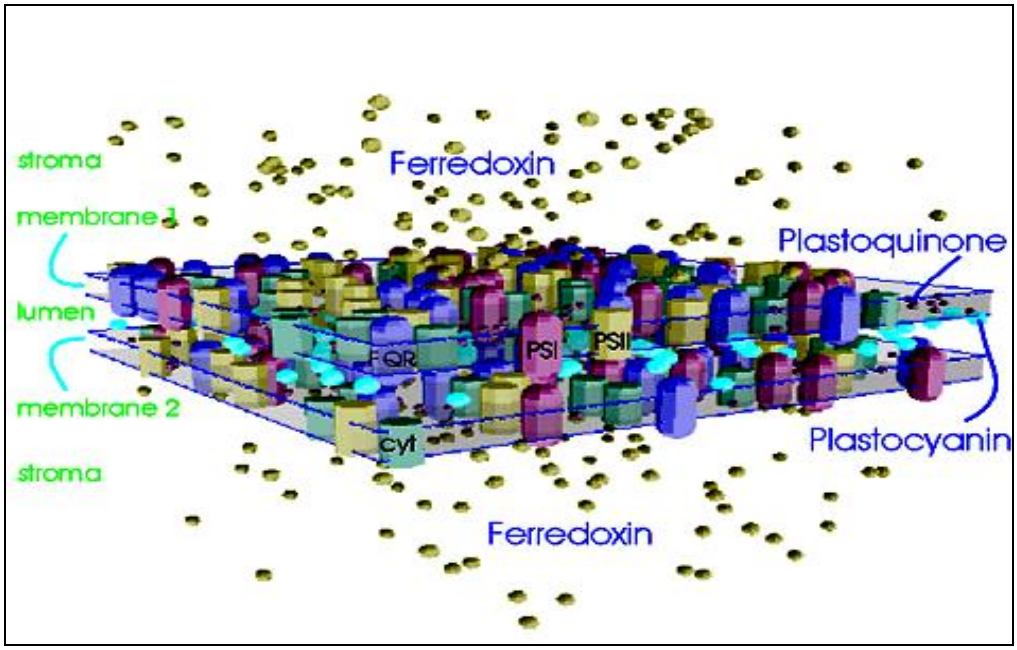

Рис. 7. Вид модельной сцены фотосинтетической мембраны в модуле визуализации.

В системе MathCell пользователь задает конфигурацию моделируемой системы, указывая количество мембранных комплексов и мобильных белков-переносчиков всех типов, параметры взаимодействия между ними (радиус взаимодействия, вероятность образования комплекса при сближении на заданный радиус). Также задается число шагов, коэффициент диффузии протонов, размеры компартментов системы (размер внешней области, расстояние между мембранами, размер участка мембраны).

Выходные данные модели – это кинетические кривые переноса окисления и восстановления всех участников электрон-транспортной цепи, пространственно-временное распределение протонов, траектории мобильных переносчиков. Графики кинетических кривых можно просмотреть через веб-интерфейс, остальные данные выкачиваются в виде файлов и отображаются через модуль трехмерной визуализации (рис. 7), либо внешними программами.

ОБЛАЧНЫЕ ХАРАКТЕРИСТИКИ НАУЧНЫХ ПРОЕКТОВ В ОБЛАСТИ ВЫЧИСЛИТЕЛЬНОЙ БИОЛОГИИ

Оценка облачных характеристик

Опираясь на классификацию NIST (см. рис. 2), мы оценили облачные характеристики нескольких проектов по вычислительной биологии. Результаты представлены в таблице 2. Как следует из таблицы, в мире уже существуют системы, отвечающие понятию «облако для науки», однако проект Cloud BioLinux получил все признаки облака исключительно за счет использования сервиса Amazon, проекты Nimbus и Wispy лишь предоставляют вычислительные ресурсы для проведения научных исследований по модели "Инфраструктура как сервис". Проект Magellan отвечает определению облака и обладает дополнительными функциональными возможностями, которые могут быть использованы биологами для исследований. Проекты, направленные на моделирование биологической клетки, по большинству параметров не представляют из себя облачные технологии, несмотря на использование

вычислительных центров и мощной аппаратуры. Причиной этого является прежде всего то, что эти проекты являются не коммерческими, а исследовательскими.

Таблица 2. Сравнительный анализ проектов в области вычислительной биологии с точки зрения характеристик облачных вычислений

Проект	Широкая сетевая доступность	Легкая масштабируемость	Работу сервиса можно оценить	Наличие пула ресурсов	Использование виртуализации	Балансировка нагрузки
Cloud BioLinux [16]	Да, за счет использования Amazon Web Service (AWS)	Да, AWS	Да, AWS	Да, AWS	Да, AWS	Да, AWS
Nimbus [17]	Да	Да	Да	Да	Да	Да
Wispy [17]	Да	Нет	Да	Частично	Да	Да
Magellan [18]	Да	Да	Да	Да	Да	Да
Virtual Cell [26]	Да	Нет	Нет	Нет	Нет	Нет
E-Cell [27]	Нет	Нет	Нет	Нет	Нет	Нет
MCell [28]	Нет	Да	Нет	Да	Нет	Да
MathCell [19]	Да	Нет	Нет	Частично	Нет	Нет

Проект Mathcell с точки зрения облачных технологий

Рассмотрим более подробно проект «Математическая клетка» для определения текущей позиции системы среди облаков.

Прежде всего следует отметить, что проект предоставляет сервис по типу "Программное обеспечение как сервис" (SaaS), поскольку пользователь выполняет расчеты с помощью программ, написанных и отлаженных провайдером. При этом не предоставляется ни среда разработки, ни платформа для размещения данных, ни инфраструктура в целом.

Далее рассмотрим систему относительно каждой из основополагающих характеристик облачных технологий.

1. Широкая сетевая доступность (Broad Network Access).

Безусловно, проект соответствует данному определению, поскольку все ресурсы доступны через сеть Интернет и имеют так называемый «тонкий клиент», т.е. все операции можно выполнить через браузер, без скачивания и устанавливания каких-либо программ непосредственно на компьютер или мобильные устройства.

2. Легкая масштабируемость (Rapid Elasticity). Проект не соответствует данному критерию, поскольку пользователь не может получить дополнительно доступ к какой-либо части ресурса без взаимодействия с сервис-провайдером.

3. Возможность мониторинга (Measured Service). Проект не соответствует данной характеристике, поскольку не имеет системы мониторинга, по которой можно было бы оценить доступность и работоспособность программного обеспечения.

4. Самообслуживание (On-Demand Self-Service). Проект не соответствует данному критерию, поскольку пользователь не может при необходимости дополнить или изменить набор используемых ресурсов без непосредственного контакта с сервис-провайдером

5. Объединение ресурсов (Resource Pooling). Проект частично соответствует данному критерию. Аппаратные ресурсы не выделяются дополнительно при большой загрузке системы и не перераспределяются при необходимости. Однако, поскольку система работает на Грид-технологиях, балансировка нагрузки в определенном смысле выполняется путем постановки всех заданий в очередь. Таким образом, при нехватке аппаратных ресурсов, запрос выполняется чуть позже, когда появляется такая возможность.

Также стоит отметить, что хотя NIST и не относит виртуализацию и балансировку нагрузки к основным характеристикам облачных систем, они являются неотъемлемой частью качественно разработанных технологических ресурсов. С этой точки зрения проект «Математическая клетка» не соответствует понятию облака, поскольку не использует для своей работы гипервизоры. С точки зрения использования одних и тех же ресурсов разными пользователями для разных целей (понятие Multi-tenancy), проект «Математическая клетка» соответствует понятию облака. В рамках проекта используется Грид-система, которая выделяется под все задачи системы.

Рассмотрим проект «Математическая клетка» с точки зрения обслуживания и управления (таблица 1).

1. Инфраструктура обслуживается полностью сервис-провайдером (в нашем случае коллективом Института математических проблем биологии РАН), который для пользователя выступает внешней организацией.
2. Инфраструктурой владеет также Институт математических проблем биологии РАН, то есть внешняя организация относительно пользователя.
3. Как аппаратная, так и программная части системы находятся на территории сервис-провайдера и являются его собственностью.
4. Проект «Математическая клетка» состоит из нескольких частей, некоторые ресурсы (модель переноса заряда в ДНК, модель переноса электрона в фотосинтетической мембране, расчет энергии растворения биомолекул в воде методом Монте-Карло, моделирование связывания белков электрон-транспортной цепи фотосинтеза) требуют авторизации в системе. Однако другие ресурсы, такие как 3D-модель клетки, тематические обзоры, глоссарий и список тематических статей не требуют авторизационных данных и доступны для всех пользователей из сети Интернет.

Таким образом, проект «Математическая клетка» в основном соответствует понятию публичного облака и предоставляет услуги по модели «Приложение как Сервис» (Software as a Service). После доработки в области мониторинга ресурсов, автоматизации масштабируемости и балансировки нагрузки проект может получить статус полноценного облачного ресурса.

При этом широкий спектр предоставляемой пользователю информации и функциональных возможностей в области вычислительной биологии и биоинформатики выгодно выделяет проект «Математическая клетка» среди других Интернет-ресурсов в данной области науки.

ЗАКЛЮЧЕНИЕ

Развитие облачных технологий в сочетании с прогрессом вычислительной биологии открывают новые возможности научного познания. Мы присутствуем при рождении новой парадигмы науки, в которой анализ крупномасштабных данных выступает не в качестве препятствия для исследователя, а как основа для построения теории, моделирования и эксперимента [29]. Раньше визуализация была единственным способом определить тенденции в огромном количестве научных данных [30]. Теперь же интеллектуальные методы анализа информации могут распознавать закономерности в наборах данных, которые являются слишком большими или сложными для визуализации. Методы прямого моделирования многочастичных систем позволяют детально изучать объекты, ранее недоступные для моделирования. Технологии и научные методы, рассмотренные в данной статье, несомненно, найдут широкое применение в исследовательских и прикладных областях.

Работа выполнена при финансовой поддержке проекта EGEE (European Grid for E-sciencE) и Российского фонда фундаментальных исследований, гранты №№ 11-07-00577, 12-07-00783, 12-07-33036, 13-07-00162, 13-07-12183.

СПИСОК ЛИТЕРАТУРЫ

1. Huth A., Cebula J. The Basics of Cloud Computing. *Carnegie Mellon University*. 2011. URL: http://www.us-cert.gov/sites/default/files/publications/CloudComputingHuthCebula.pdf (дата обращения: 03.09.2013).
2. Soghoian Ch. Caught in the cloud: privacy, encryption, and government back doors in the web 2.0 Era. *J. on Telecomm. and High Tech. L.* 2009. V. 8. P. 359–424.
3. Лахно В.Д., Исаев Е.А., Пугачев В.Д., Зайцев А.Ю., Фиалко Н.С., Рыкунов С.Д., Устинин М.Н. Развитие информационно-коммуникационных технологий в Пущинском научном центре РАН. *Математическая биология и биоинформатика.* 2012. Т. 7. № 2. С. 529–544. URL: http://www.matbio.org/2012/Lakhno_7_529.pdf (дата обращения: 03.09.2013).
4. Исаев Е.А., Корнилов В.В., Тарасов П.А. Научные компьютерные сети – проблемы и успехи в организации обмена большими объемами научных данных. *Математическая биология и биоинформатика*. 2013. Т. 8. № 1. С. 161–181. URL:http://www.matbio.org/2013/Isaev_8_161.pdf (дата обращения: 03.09.2013).
5. Hoff Ch., Simmonds P. Security guidance for critical areas of focus in cloud computing. *Website of Cloud Security Alliance*. 2011. С. 12–20. URL: https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf (дата обращения: 03.09.2013).
6. Khajeh-Hosseini A., Sommerville I., Sriram I. Research Challenges for Enterprise Cloud Computing. *arXiv.org: Cornell University Library*. URL: http://arxiv.org/ftp/arxiv/papers/1001/1001.3257.pdf (дата обращения: 03.09.2013).
7. An overview on cloud computing for research and science published. 2012. URL: http://www.e-irg.eu/news/news/479/an-overview-on-cloud-computing-for-research-and-science-published.html (дата обращения: 03.09.2013).
8. Розенблюм М., Гарфинкель Т. Мониторы виртуальных машин: современность и тенденции. *Открытые системы*. 2005. № 05–06. URL: http://www.osp.ru/os/2005/05-06/185589/ (дата обращения: 03.09.2013).
9. Cloud Computing: 7 проблем безопасности. *Бюро Соломатина*. URL: http://www.bureausolomatina.ru/ru/themes_in_progress/clouds/6 (дата обращения: 03.09.2013).
10. Федоров А.Г., Мартынов Д.Н. *Облачная платформа Microsoft*. 2010. 100 с.
11. Stiefel M. Cloud Computing with Microsoft Azure. *Website of Reliable Software, Inc.* 2009. URL: http://reliablesoftware.com/presentations/Introduction%20to%20Cloud%20Computing%20with%20Microsoft%20Azure.pdf. (дата обращения: 03.09.2013).
12. Облачные вычисления. *Website of IBM*. URL: http://www.ibm.com/ru/cloud/ (дата обращения: 03.09.2013).
13. Что такое облачные вычисления и как их можно использовать. *Website of IBM*. 2008. URL: http://www.ibm.com/ru/cloud/pdf/Understanding_and_Leveraging_Cloud_Computing_RU-1_validated_Feb2_KI_rus_s5_hyperlinks.pdf (дата обращения: 03.09.2013).
14. Varia J. Building GrepTheWeb in the Cloud, Part 1: Cloud Architectures. *Amazon Web Services*. 2008. URL: http://aws.amazon.com/articles/1632?_encoding=UTF8&jiveRedirect=1 (дата обращения: 03.09.2013).
15. Amazon SimpleDB. *Website of Amazon Web Services*. URL: http://aws.amazon.com/simpledb/ (дата обращения: 03.09.2013).
16. Krampis K. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics*. 2012.

17. Clouds. *Website of Science Clouds*. URL: http://scienceclouds.org/infrastructure-clouds/ (дата обращения: 03.09.2013).
18. Ramakrishnan L., Campbell S., Canon R.S., Declerck T., Sakrejda I. Magellan: Experiences from a Science Cloud. In: *ScienceCloud'11* (June 8, 2011, San Jose, California, USA). 2011. URL: http://datasys.cs.iit.edu/events/ScienceCloud2011/p07.pdf (дата обращения: 03.09.2013).
19. Лахно В., Назипова Н., Ким В., Филиппов С., Фиалко Н., Устинин Д., Теплухин А., Тюльбашева Г., Зайцев А., Устинин М. Информационно-вычислительная среда Mathcell для моделирования живой клетки. *Математическая биология и биоинформатика*. 2007. Т. 2. С. 361–376. URL: http://www.matbio.org/downloads/Lakhno2007(2_361).pdf (дата обращения: 03.09.2013).
20. *Математическая клетка*. URL: http://www.mathcell.ru/ (дата обращения: 03.09.2013).
21. Лахно В.Д. Математическая биология и биоинформатика. *Вестник Российской академии наук*. 2011. Т. 81. № 9. С. 812–818.
22. Хрущев С.С., Абатурова А.М., Дьяконова А.Н., Устинин Д.М., Зленко Д.В., Федоров В.А., Коваленко И.Б., Ризниченко Г.Ю., Рубин А.Б. Моделирование белок-белковых взаимодействий с применением программного комплекса многочастичной броуновской динамики. *Компьютерные исследования и моделирование*. 2013. Т. 5. № 1. С. 47–64.
23. Устинин Д.М., Коваленко И.Б., Грачев Е.А., Ризниченко Г.Ю., Рубин А.Б. Метод прямого многочастичного компьютерного моделирования фотосинтетической электронно-транспортной цепи. *В: Динамические модели процессов в клетках и субклеточных наноструктурах.* Москва-Ижевск: *РХД*, 2010. С. 241–262.
24. Устинин Д.М., Коваленко И.Б., Ризниченко Г.Ю., Рубин А.Б. Сопряжение различных методов компьютерного моделирования в комплексной модели фотосинтетической мембраны. *Компьютерные исследования и моделирование*. 2013. Т. 5. № 1. С. 65–81.
25. Kovalenko I.B., Abaturova A.M., Gromov P.A., Ustinin D.M., Grachev E.A., Riznichenko G.Yu., Rubin A.B. Direct simulation of plastocyanin and cytochrome f interactions in solution. *Physical Biology*. 2006. № 3. P. 121–129.
26. Leslie M. Loew, James C. Schaff. The Virtual Cell: a software environment for computational cell biology. *TRENDS in Biotechnology*. 2001. Т. 19. № 10.
27. Tomita M. E-Cell: software environment for whole-cell simulation. *Bioinformatics*. 1999. Т. 15. № 1. С. 72–84.
28. Casanova H. The virtual Instrument: Support for Grid-enabled MCell simulations. *The International Journal of High Performance Computing Applications*. 2004. Т. 18. № 1. С. 3–17.
29. Исаев Е.А., Корнилов В.В. Проблема обработки и хранения больших объемов научных данных и подходы к ее решению. *Математическая биология и биоинформатика*. 2013. Т. 8. № 1. С. 49–65. URL: http://www.matbio.org/2013/Isaev_8_49.pdf (дата обращения: 03.09.2013).
30. Ustinin M.N., Kronberg E., Filippov S.V., Sychev V.V., Sobolev E.V., and Llinás R. Kinematic visualization of human magnetic encephalography. *Mathematical Biology & Bioinformatics.* 2010. V. 5. № 2. P. 176–187. URL: http://www.matbio.org/downloads_en/Ustinin2010(5_176).pdf (дата обращения: 03.09.2013).

Материал поступил в редакцию 19.08.2013, опубликован 04.09.2013.