

Концепция сервиса виртуального скрининга лекарств на базе распределенных вычислений HiTViSc

Е. Е. Ивашко^{1,||}, Н. Н. Никитина¹

¹Институт прикладных математических исследований Карельского научного центра РАН,
Петрозаводск, Россия

^{||}Петрозаводский государственный университет, Петрозаводск, Россия

Аннотация. В работе представлена концепция сервиса High-Throughput Virtual Screening as a Service (HiTViSc), реализующего облачный сервис виртуального скрининга лекарств на базе концепции распределенных вычислений Desktop Grid. Описаны три логических уровня функционирования сервиса, рабочие процессы пользователя и принцип многопользовательского доступа. Концепция описывает сервис, представляющий собой программную систему с необходимой функциональностью для решения задачи виртуального скрининга с использованием практически неограниченно масштабируемых вычислительных ресурсов Desktop Grid.

Ключевые слова: виртуальный скрининг, добровольные вычисления, Desktop Grid, облачные вычисления, Desktop Grid as a Service, Virtual Screening as a Service, High-Throughput Virtual Screening as a Service.

DOI 10.14357/20718632230311

Введение

Высокопроизводительные вычисления играют значительную роль в проведении современных фундаментальных и прикладных исследований, разработках новых материалов, лекарств, новых видов промышленной продукции и в социальной сфере. На сегодняшний день многие задачи невозможно решить без применения высокопроизводительных вычислительных систем. Среди актуальных вычислительноемких задач - моделирование материалов на молекулярном и атомарном уровнях, а также моделирование сложных взаимодействий биологических молекул.

Основным инструментом выполнения высокопроизводительных расчетов на сегодня явля-

ются вычислительные кластеры или суперкомпьютеры; когда требуется задействовать особенно большие объемы ресурсов, географически распределенные суперкомпьютеры объединяют в грид-системы. Однако с развитием каналов связи, соответствующим увеличением скорости и пропускной способности сети Интернет и локальных вычислительных сетей, с ростом производительности персональных компьютеров, все больший потенциал набирают Desktop Grid - грид-сети, объединяющие большое количество географически распределенных, относительно небольших по мощности неспециализированных вычислителей (как правило, персональных компьютеров). В таких системах вычислительные узлы не связаны

между собой и предоставляют вычислительные ресурсы на нерегулярной основе.

Как правило, проекты Desktop Grid строятся на основе добровольно предоставляемых вычислительных ресурсов компьютеров частных лиц и организаций (добровольные вычисления, *volunteer computing*). Однако существуют и примеры построения подобных частных распределенных систем в масштабах организации или группы организаций - Enterprise Desktop Grid. На сегодняшний день одной из наиболее популярных и развивающихся программных платформ для реализации Desktop Grid является BOINC [1].

При этом, как правило, поддержка аппаратной и программной инфраструктуры вычислительных проектов ложится на плечи специалистов, обеспечивающих и содержательную часть проекта. Это приводит к сложностям в запуске, поддержке и развитии проектов Desktop Grid, особенно проектов добровольных вычислений, как следствие, их число в мире явно не соответствует вычислительному потенциалу доступных ресурсов. Одним из решений этой проблемы является привлечение технологий облачных вычислений.

Облачные вычисления хорошо зарекомендовали себя на практике, обеспечивая удобство использования, быстрое развертывание и общедоступность. Исследование различных способов комбинирования Desktop Grid и облачных вычислений для получения преимуществ обеих концепций являлось предметом целого ряда научных работ [2].

Наиболее перспективным подходом, с нашей точки зрения, является концепция Desktop Grid as a Service [3], в рамках которой организуется предоставление высокопроизводительных вычислений на базе Desktop Grid как сервисов SaaS, PaaS и IaaS. В данной работе мы представляем облачный сервис выполнения виртуального скрининга лекарств, реализованный на базе принципов Desktop Grid as a Service.

Создание лекарств представляет собой трудоемкую задачу ввиду значительных временных затрат на поиск веществ, обладающих необходимым терапевтическим действием и соответствующих всем критериям безопасности для человека. С момента начала разработки до вывода

лекарства на рынок проходит в среднем 10-15 лет. Один из начальных этапов процесса разработки лекарства состоит в формировании набора химических соединений-хитов (кандидатов на прохождение на последующие этапы разработки), потенциально имеющих высокую требуемую биохимическую активность по отношению к мишени - макромолекуле, связанной с развитием заболевания. Хиты ищутся среди низкомолекулярных химических соединений, называемых лигандами. Метод роботизированного поиска хитов в лабораторных условиях (*in vitro*) называется высокопроизводительным скринингом (*high-throughput screening*, HTS). В течение продолжительного времени он был основным способом проведения первого этапа разработки лекарства, однако он обладает рядом технических ограничений и требует использования дорогостоящего лабораторного оборудования.

С развитием высокоточных компьютерных моделей белков и процессов связывания лигандов с мишенями, виртуальный скрининг (*virtual screening*, VS) [4] стал *in silico* (компьютерной) альтернативой HTS. В процессе виртуального скрининга проводится компьютерное моделирование взаимодействия лигандов с мишенью и оценивается вероятность образования устойчивых молекулярных комплексов. Лиганды с высокими оценками такой вероятности становятся хитами. Виртуальный скрининг выполняется специальными компьютерными программами молекулярного докинга, такими как AutoDock Vina, CxDock, GOLD, FlexX и др. (современный обзор приведен в [5]).

Проблема проведения виртуального скрининга заключается в потребности в значительных вычислительных ресурсах и времени, необходимых для оценки потенциальных фармакологических свойств множества молекул и для идентификации наиболее подходящих лигандов. Несмотря на то, что оценка взаимодействия отдельного лиганда с наперед заданным белком производится достаточно быстро, объемы библиотек лигандов приводят к необходимости привлечения инструментов высокопроизводительных вычислений для проведения виртуального скрининга. Поэтому востребованы сервисы организации процесса виртуального скрининга с использованием высокопроизводительных

вычислительных ресурсов. Современный обзор таких сервисов на основе облачных вычислений приведен в работах [6,7], на основе Китайской национальной грид CNGrid - в [8], на основе суперкомпьютера Tianhe-2 - в [9]. Виртуальный скрининг также может проводиться на вычислительных ресурсах Desktop Grid в рамках добровольных вычислений [10,11] или Enterprise Desktop Grid [12, 13].

Еще одной особенностью виртуального скрининга зачастую является потребность в специализированных алгоритмах подготовки библиотеки лигандов, организации вычислительного эксперимента и обработки результатов. Так, виртуальный скрининг не ограничен существующими библиотеками синтезированных молекул и может использоваться для оценки соединений, которые еще не были синтезированы, а также фрагментов молекул. Разработка лекарств для отдельных классов заболеваний может подразумевать особые критерии подготовки библиотеки лигандов и выбора хитов (например, [14]). Поэтому проведение виртуального скрининга может требовать существенных технических усилий и времени.

Существует ряд коммерческих решений, предоставляющих полный цикл виртуального скрининга на основе облачных вычислительных ресурсов - Virtual Screening as a Service [15-17]. Такие программные решения являются закрытыми, в их рамках сложно или невозможно реализовать собственные алгоритмы подготовки библиотеки лигандов и вычислительного эксперимента, а объем доступных вычислительных ресурсов ограничен.

При этом, с одной стороны, себестоимость использования облачных (виртуализованных), суперкомпьютерных и грид-ресурсов высока, а политики использования многопользовательских высокопроизводительных вычислительных систем накладывают ограничения как на объемы библиотек химических соединений и генерируемых данных, так и на допустимое число параллельных вычислительных задач; с другой стороны, Desktop Grid хорошо подходит для решения задачи виртуального скрининга, а ресурсы практически неограниченно масштабируются.

Учитывая это, в данной статье мы предлагаем новую концепцию сервиса High-

Throughput Virtual Screening as a Service (HiTViSc), представляющего собой специализированный облачный сервис виртуального скрининга, основанный на высоко масштабируемых ресурсах Desktop Grid. В сравнении с существующими решениями, предложенная концепция объединяет в себе и облачные вычисления с их удобством использования и прозрачным биллингом, и специализированный функционал виртуального скрининга, и вычислительную мощность и масштабируемость Desktop Grid. Данная работа развивает исследования, связанные с развитием концепции Desktop Grid as a Service [3] применительно к задаче виртуального скрининга лекарств, обобщает опыт проекта добровольных вычислений SiDock@home [10] и виртуального скрининга в рамках Enterprise Desktop Grid [13].

Структура работы следующая. В разделе 1 дается краткое описание концепции Desktop Grid и свободно распространяемой программной платформы реализации добровольных вычислений BOINC. В частности, в этом разделе описаны достоинства и недостатки Desktop Grid в целом и BOINC в частности, которые приводят нас к разработке сервиса HiTViSc. В разделе 2 представлено краткое описание задачи виртуального скрининга как одного из этапов разработки лекарств. Раздел 3 посвящен описанию концепции разрабатываемого сервиса виртуального скрининга HiTViSc, использующего распределенные вычисления на базе Desktop Grid. Наконец, в заключении представлены выводы и обсуждение дальнейших направлений работы.

1. Desktop Grid и BOINC, облачные вычисления

Desktop Grid - это система высокопроизводительных распределенных вычислений, использующая простаивающие ресурсы неспециализированных географически распределенных вычислителей, объединенных обычной сетью передачи данных. Как правило, вычислительными узлами Desktop Grid являются настольные компьютеры и ноутбуки добровольцев, подключенные к управляющему узлу через Интернет (эта концепция называется добровольными вычислениями), либо компьютеры организаций,

подключенные через локальную сеть (Enterprise Desktop Grid).

Desktop Grid реализует простой рабочий процесс: исходная вычислительно сложная задача разбивается на большое число относительно небольших подзаданий, каждое из которых может быть в сравнительно небольшое время решено на отдельном вычислительном узле; сервер проекта распределяет по сети подзадания по вычислительным узлам и собирает результаты - такой цикл повторяется до тех пор, пока все подзадания не будут решены; затем сервер объединяет решения отдельных подзаданий в общее решение вычислительно сложной задачи. Такая организация вычислений позволяет объединить сотни тысяч вычислительных узлов, распределенных по всему миру, обеспечивая большую масштабируемость и пиковую производительность, существенно превосходя по этим свойствам традиционные высокопроизводительные вычислительные системы. Однако такая архитектура также существенно ограничивает класс решаемых задач. Desktop Grid хорошо подходит для выполнения вычислительно емких расчетов, в которых исходная задача может быть разделена на множество независимых подзадач. К таким задачам относятся задачи поиска в многомерном пространстве решений (в том числе задача виртуального скрининга лекарств), моделирование методами Монте-Карло и др.

Одной из наиболее популярных и востребованных программных платформ для реализации проектов Desktop Grid является OpenSource-платформа BOINC¹. Эта платформа является стандартом де-факто для организации добровольных вычислений; подробно архитектура и особенности системы описаны в статье [1]. Однако для ученых, имеющих соответствующие вычислительно емкие задачи, BOINC остается технически сложным инструментом, что отражается на числе активных мировых научных проектов добровольных вычислений [18].

На сегодняшний день не существует готового решения для запуска научных расчетов на BOINC без участия технического специалиста. Одним из потенциальных решений является разрабатываемый в настоящее время проект

BOINC Central², разработчики которого планируют реализовать наиболее востребованные научные приложения на инфраструктуре BOINC и предоставлять ученым упрощенный доступ к ним. При этом будет автоматизировано только использование Desktop Grid, но не остальные этапы решения научной вычислительно емкой задачи.

Другим перспективным направлением является концепция Desktop Grid as a Service [3]. Эта концепция нацелена на реализацию преимуществ облачных вычислений в высокопроизводительной вычислительной инфраструктуре типа Desktop Grid. С одной стороны, инфраструктура Desktop Grid as a Service, как и Desktop Grid, задействует простаивающие неспециализированные компьютеры, связанные с управляющим узлом сетью передачи данных, с другой, как и облачные вычисления, она обеспечивает удобный доступ по запросу к общему пулу настраиваемых вычислительных узлов, которые могут предоставляться как частными лицами, так и организациями.

В рамках данной работы мы представляем концепцию сервиса HiTViSc, реализующего подход Desktop Grid as a Service, чтобы предоставить специалистам простое, доступное решение по проведению полного цикла виртуального скрининга на высокопроизводительных вычислительных ресурсах, обладающих всеми преимуществами Desktop Grid.

2. Виртуальный скрининг лекарств

Виртуальный скрининг, как было сказано выше, заключается в компьютерном моделировании взаимодействия белка-мишени с лигандом. Такое моделирование само по себе не требует больших компьютерных ресурсов, однако большое количество лигандов, которые необходимо проверить, делают исходную задачу очень вычислительно емкой (Рис. 1). При этом, так как проверка лигандов выполняется независимо, виртуальный скрининг является одной из прикладных задач, хорошо подходящих для решения с помощью Desktop Grid. Рабочий процесс выглядит следующим образом: сервер выдает

¹ <https://boinc.berkeley.edu>

² <https://github.com/BOINC/boinc/discussions/4965>



Рис. 1. Схематичное изображение процесса виртуального скрининга

Необходимо выбрать множество хитов из химического пространства потенциальных лекарств. Темно-серым цветом обозначены примеры библиотек лигандов. Результат виртуального скрининга - список хитов, упорядоченных по оценкам энергии связывания с мишенью.

каждому вычислительному узлу по одному или несколько лигандов и мишень; узел выполняет молекулярный докинг с помощью специальной программы и возвращает результат - оценку энергии и конфигурацию связывания лиганда с мишенью, после чего получает новое задание. Вычисление на каждом отдельном узле не зависит от других, а множество лигандов задано заранее, поэтому задача эффективно масштабируется в Desktop Grid на доступное количество узлов. При этом решение задачи не требует межпроцессного взаимодействия и/или больших объемов хранения данных, поэтому Desktop Grid является более эффективным инструментом решения задачи виртуального скрининга, чем суперкомпьютер или грид.

Виртуальный скрининг - это вычислительно-емкая задача, в которой объем требуемых вычислений зависит от используемой базы данных (на практике применяются библиотеки от сотен до миллиардов лигандов), используемого приложения молекулярного докинга, точности вычислений, требований ко множеству хитов. Так, например, для проведения виртуального скрининга базы данных GDB-17 потребовалось бы более пяти месяцев непрерывных вычислений на самом мощном суперкомпьютере мира из списка TOP-500 2016 г. [19].

Выполнение виртуального скрининга не всегда связано с полным перебором. Применяются специализированные библиотеки лигандов (например, коммерчески доступных или подготовленных для конкретного вычислительного эксперимента), может быть выбрана различная точность молекулярного докинга, что значительно сказывается на скорости выполнения виртуального скрининга. Поскольку важно не только количество, но и структурное разнообразие хитов, то порядок тестирования лигандов оказывает важное влияние на скорость и качество получения результатов.

Существует и применяется множество различных вариантов при экспертной подготовке библиотек лигандов, выборе и настройке программ молекулярного докинга и протоколов виртуального скрининга. Поэтому большинство ученых проектируют и проводят весь цикл виртуального скрининга самостоятельно, в соответствии со своими потребностями и доступными вычислительными мощностями. Такой процесс является трудоемким и времяемким. Последовательность процедур и генерируемые данные при этом не сохраняются и не доступны другим ученым.

Все перечисленные нюансы должны быть учтены при реализации вычислительной платформы, чтобы обеспечить необходимую

функциональность при проведении виртуального скрининга.

3. Концепция High-Throughput Virtual Screening as a Service

Целью данной исследовательской работы является разработка вычислительного сервиса, который, с одной стороны, предлагает пользователям удобный веб-интерфейс виртуального скрининга, визуализации и первичного анализа результатов, а с другой стороны - обеспечивает выполнение виртуального скрининга с помощью распределенных вычислений на базе Desktop Grid, включая возможности по администрированию доступных вычислительных ресурсов.

Для разработки и реализации специализированного вычислительного сервиса по проведению виртуального скрининга будут применяться, в частности, авторские математические модели и алгоритмы управления заданиями в Desktop Grid, ранее предложенные и апробированные как в рамках Enterprise Desktop Grid [13], так и в действующих проектах добровольных вычислений SiDock@home и RakeSearch³: теоретико-игровая модель на основе игры заполнения для скорейшего нахождения наиболее разнообразных хитов виртуального скрининга [20], алгоритмы эффективной организации виртуального скрининга в Desktop Grid на базе BOINC [21], теоретико-игровая модель формирования пакетов заданий оптимального размера на основе иерархической игры [22], линейно-регрессионная модель прогнозирования времени завершения вычислительного эксперимента [23], алгоритмы оптимальной репликации заданий [24] и выбора кворума [25].

Сервис HiTViSc является реализацией концепции Desktop Grid as a Service [3], по которой высокопроизводительные вычислительные ресурсы предоставляются пользователю в виде специализированного облачного сервиса.

Сервис имеет три логических уровня реализации:

1. Первый уровень - *вычислительный* - обеспечивает выполнение высокопроизводительных

расчетов на базе Desktop Grid, фактически представляет из себя сервер Desktop Grid и реализует функции генерации заданий, коммуникации с вычислительными узлами (назначение заданий, передача исходных данных и расчетных приложений) и учета результатов.

2. Второй уровень - *уровень виртуального скрининга* - реализует специальные функции проведения виртуального скрининга, включая загрузку файла мишени, базы лигандов и расчетного приложения, настройки параметров, протокола молекулярного докинга и обработки результатов, выбор внешних приложений для анализа и визуализации результатов.

3. Третий уровень - *пользовательский* - предоставляет интерфейс пользователя облачного приложения, включая интерфейсы управления и настройки виртуального скрининга, визуализацию прогресса выполнения эксперимента, визуализацию и анализ результатов, а также администрирование вычислительных ресурсов.

В качестве вычислительных узлов выступают клиенты Desktop Grid, обеспечивающие выполнение расчетов.

Основным пользователем системы является ученый (один или в группе), проводящий виртуальный скрининг в рамках трех рабочих процессов: 1) выполнение вычислительного эксперимента, 2) администрирование вычислительных ресурсов, 3) экспертный анализ результатов.

1) Выполнение вычислительного эксперимента происходит следующим образом:

(а) Загрузка мишени.

Для описания мишени при молекулярном докинге широко применяется формат PDB, используемый в ведущей открытой базе трехмерных структурных моделей белков RCSB Protein Data Bank (RCSB PDB) [26]. В сложившейся практике виртуального скрининга пользователь загружает PDB-файл или указывает PDB-идентификатор в базе RCSB PDB. Затем файл подготавливается к молекулярному докингу программными утилитами и/или вручную.

(б) Выбор библиотеки лигандов.

Подготовка библиотеки лигандов является экспертной задачей, результат решения которой влияет на качество виртуального скрининга.

³ <https://rake.boincfast.ru/rakesearch/>

Формат файлов зависит от целевой программы молекулярного докинга. Так, для CmDock используется формат SDF, для AutoDock Vina - формат PDBQT, и т.д. Выбор производится среди библиотек доступных на сервере системы, доступных через онлайн-интерфейсы или загружаемых пользователем. Затем библиотека фильтруется в соответствии с требованиями пользователя. Подготовленная библиотека лигандов хранится на сервере и является важной частью конкретного вычислительного эксперимента виртуального скрининга.

Возможно также задание множества «контрольных» лигандов, которое потребуется пользователю для подготовки мишени и/или при анализе результатов виртуального скрининга.

(в) Выбор программы докинга.

Существует множество программ молекулярного докинга (таких как уже упоминавшиеся CmDock, AutoDock Vina), различающихся как концептуально, так и технически (формат настроек, форматы файлов и т.д.). В сложившейся практике обычно используется программа, доступная на сервере или локально, для которой спроектирован процесс виртуального скрининга.

(г) Настройка протокола докинга.

Вид и формат параметров зависят от конкретной программы молекулярного докинга. Например, для CmDock это 3D-координаты центра сайта связывания и размер области докинга, число повторов молекулярного докинга, различные ограничения и фильтры⁴. Интерфейс системы предоставляет протокол по умолчанию с возможностью правки настроек.

(д) Настройка протокола вычислительного эксперимента.

Протокол вычислительного эксперимента определяет критерий отбора хитов и критерий остановки вычислительного эксперимента.

- Критерий отбора хитов – это значение энергии связывания, при превышении которого лиганды считаются хитами. Это может быть пороговое значение или эффективность лиганда (энергия связывания, нормированная по числу атомов) или более сложный критерий отбора.

- Критерий остановки вычислительного эксперимента - например, заданное число проверенных молекул, заданное число хитов или заданная доля хитов от фиксированного числа проверенных молекул.

(е) Выбор вычислительных ресурсов.

Для выполнения виртуального скрининга пользователю доступны опции выбора вычислительных ресурсов. При этом, в соответствии с концепцией Desktop Grid, основой являются ресурсы неспециализированных вычислителей, в том числе, привлекаемых в систему самим пользователем.

Расчеты в рамках эксперимента могут быть одного из трех типов:

1. тестовые: для проверки корректности настроек эксперимента,
2. публичные: результаты эксперимента, включая мишень, настройки и результаты виртуального скрининга доступны широкой аудитории,
3. частные: результаты эксперимента до
4. из заданной библиотеки,

- количество найденных хитов,
- охват библиотеки (например, число кластеров),
- прогноз времени завершения вычислительного эксперимента.

2) Администрирование вычислительных ресурсов.

Концепция сервиса предполагает три типа вычислительных ресурсов, которые могут быть использованы в виртуальном скрининге как вычислительные узлы Desktop Grid:

- Тестовое окружение. Ограниченный объем высоко доступных и надежных вычислительных ресурсов, используемых для проверки параметров на небольшом объеме вычислений.
- Вычислительные ресурсы системы, предоставляемые в общее пользование добровольцами и другими пользователями HiTViSc и используемые для выполнения экспериментов.
- Собственные и партнерские подключенные к системе вычислительные ресурсы, которые могут использоваться для собственных экспериментов, партнерских проектов или предоставляемые в общее пользование.

⁴ <https://gitlab.com/Jukic/cmdock/-/blob/master/docs/reference-guide/docking-protocol.rst>

Важное значение имеет учет используемых вычислительных ресурсов. Так, одна из возможных политик сервиса может заключаться в том, что приоритет конкретного эксперимента может понижаться или повышаться в зависимости от того, насколько существенные ресурсы пользователь предоставляет другим проектам. При этом, подключение собственных ресурсов осуществляется с помощью специальной ссылки подключения.

3) Анализ результатов виртуального скрининга.

Результаты виртуального скрининга представляют собой список хитов, упорядоченных по оценкам энергии связывания с мишенью. Для каждого хита хранится оценка энергии связывания и конформация - трехмерная модель молекулы, соответствующая данной оценке.

На практике для анализа и визуализации результатов чаще всего применяется метод главных компонент (РСА-анализ) и кластеризация, которые выполняются сторонними программами (такими, как R, Clustvis и др.), реализующими подходящие алгоритмы биоинформатики. Визуализация также является базовым методом для анализа результатов виртуального скрининга. Применяется 2D-визуализация - в виде точек, спроецированных на плоскость (используется для визуальной оценки кластеризации и химического разнообразия) и 3D-визуализация - в виде молекулярных структур, связанных с мишенью (используется для визуальной оценки результатов молекулярного докинга и сравнения с «контрольными лигандами», если они есть).

Наконец, обычная функция выгрузки данных, отфильтрованных по каким-либо критериям, является универсальной возможностью использовать сторонние программы для анализа результатов.

4) Многопользовательский доступ.

Система HiTViSc является многопользовательской. Это означает, что в рамках сервиса большое количество пользователей могут независимо проводить свои вычислительные эксперименты. При этом, в общем случае каждый пользователь имеет доступ ко всему пулу ресурсов, который распределяется между пользователями планировщиком заданий.

Внутри вычислительного проекта также используется многопользовательский подход: владелец (автор) проекта может предоставить другим участникам в той или иной степени ограниченный доступ к настройкам проекта, вычислительным ресурсам и результатам.

Каждому пользователю предоставляются следующие возможности:

1. Доступ к учетной записи для выполнения вычислительных экспериментов, хранения и анализа результатов.
2. Использование вычислительных ресурсов.
3. Использование результатов вычислительных экспериментов по одному из трех вариантов доступа: результаты доступны всем пользователям системы, доступны определенной группе пользователей, или только автору.

Основной планируемый режим работы подразумевает, что все генерируемые данные (подготовленные мишени и библиотеки лигандов, найденные хиты, оценки связывания и др.) могут использоваться широким кругом пользователей для дальнейшего экспертного анализа с сохранением авторства. При этом существование общедоступной библиотеки результатов является, с одной стороны, основанием для привлечения к проекту добровольцев, предоставляющих вычислительные ресурсы (как это широко распространено в добровольных вычислениях), а с другой стороны - станет весомым вкладом в общий научный прогресс в области биоинформатики.

Заключение

Высокопроизводительные вычисления играют значительную роль в проведении современных фундаментальных и прикладных исследований и разработок. На сегодняшний день многие задачи невозможно решить без применения высокопроизводительных вычислительных систем. При этом, для решения конкретной вычислительно сложной задачи должен быть выбран соответствующий инструмент, позволяющий эффективно найти решение.

Системы типа Desktop Grid хорошо себя зарекомендовали при решении задач виртуального скрининга, обеспечивая высокую производительность и масштабируемость. Однако

процесс развертывания Desktop Grid и обслуживание вычислительного процесса, как правило, являются технически сложными для ученых (что, в частности, отражается на числе существующих соответствующих проектов добровольных вычислений). Решением проблемы может стать High-Throughput Virtual Screening as a Service (HiTViSc) - облачный сервис виртуального скрининга на основе Desktop Grid, построенный по принципам Desktop Grid as a Service.

В работе описана концепция сервиса HiTViSc: представлены три логических уровня функционирования (вычислительный, уровень виртуального скрининга и пользовательский), описаны рабочие процессы пользователя, связанные с проведением виртуального скрининга, администрированием ресурсов, визуализацией и анализом результатов, а также принцип многопользовательского доступа. Дальнейшая работа связана с разработкой архитектуры, проектированием и реализацией сервиса HiTViSc.

Литература

- Anderson DP. BOINC: A Platform for Volunteer Computing. *Journal of Grid Computing*. 2020;18:99–122. doi:10.1007/s10723-019-09497-9
- Ivashko E. Desktop Grid and Cloud Computing: Short Survey. In: Voevodin V, Sobolev S. (eds.) *Supercomputing. RuSCDays 2021. Communications in Computer and Information Science*, vol 1510. Springer, Cham; 2021. p. 445–456. doi:10.1007/978-3-030-92864-3_34
- Ivashko E. Desktop Grid as a Service Concept. In: Voevodin V, Sobolev S, Yakobovskiy M, Shagaliev R. (eds.) *Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science* v. 13708. Springer, Cham; 2022. p. 632–643. doi:10.1007/978-3-031-22941-1_46
- Bechelane Maia EH, Assis LC, Alves de Oliveira T, Marques da Silva A, Gutterres Taranto A. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Frontiers in Chemistry*. 2020;8:343. doi:10.3389/fchem.2020.00343
- Torres PHM, Sodero ACR, Jofily P, Silva-Jr. FP. Key Topics in Molecular Docking for Drug Design. *International Journal of Molecular Sciences*. 2019;20(18):4574. doi:10.3390/ijms20184574
- Olğaç A, Türe A, Olğaç S, Möller S. Cloud-Based High Throughput Virtual Screening in Novel Drug Discovery. In: Kolodziej J, González-Vélez H. (eds.) *High-Performance Modelling and Simulation for Big Data Applications. Lecture Notes in Computer Science* v. 11400. Springer, Cham; 2019. p. 250–278. doi:10.1007/978-3-030-16272-6_9
- Singh N, Chaput L, Villoutreix BO. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Briefings in Bioinformatics*. 2021;22(2):1790–1818. doi:10.1093/bib/bbaa034
- Zhang B, Li H, Yu K, Jin Z. Molecular docking-based computational platform for high-throughput virtual screening. *CCF Trans High Perform Comput*. 2022;4(1):63–74. doi:10.1007/s42514-021-00086-5
- Mo Q, Xu Z, Yan H, Chen P, Lu Y. VSTH: a user-friendly web server for structure-based virtual screening on Tianhe-2. *Bioinformatics*. 2023;39(1):btac740. doi:10.1093/bioinformatics/btac740
- Nikitina N, Manzyuk M, Podlipnik Č, Jukic M. Volunteer computing project SiDock@home for virtual drug screening against SARS-CoV-2. In: Byrski A, Czachórski T, Gelenbe E, Grochla K, Murayama Y. (eds.) *Computer Science Protecting Human Society Against Epidemics. ANTICOVID 2021. IFIP Advances in Information and Communication Technology 2021* v. 616. Springer, Cham. p. 23–34. doi:10.1007/978-3-030-86582-5_3
- Mottin M, de Paula Sousa BK, de Moraes Roso Mesquita NC, et al. Discovery of New Zika Protease and Polymerase Inhibitors through the Open Science Collaboration Project OpenZika. *Journal of Chemical Information and Modeling*. 2022;62(24):6825–6843. doi:10.1021/acs.jcim.2c00596
- Zhang B, D'Erasmus MP, Murelli RP, Gallicchio E. Free energy-based virtual screening and optimization of RNase H inhibitors of HIV-1 reverse transcriptase. *ACS omega*. 2016;1(3):435–447. doi:10.1021/acsomega.6b00123
- Ивашко Е.Е., Никитина Н.Н., Меллер С. Высокпроизводительный виртуальный скрининг в Enterprise Desktop Grid на базе BOINC. *Вестник Южно-Уральского государственного университета, серия Вычислительная математика и информатика*. 2015;4(1):57–63. doi:10.14529/cmse150105
- Lloyd DG, Golfis G, Knox AJ, et al. Oncology exploration: charting cancer medicinal chemistry space. *Drug discovery today*. 2006;11(3-4):149–159. doi:10.1016/S1359-6446(05)03688-3
- Krasoulis A, Antonopoulos N, Pitsikalis V, Theodorakis S. DENVIS: Scalable and High-Throughput Virtual Screening Using Graph Neural Networks with Atomic and Surface Protein Pocket Features. *Journal of Chemical Information and Modeling*. 2022;62:4642–4659. doi:10.1021/acs.jcim.2c01057
- Hawkins P. Virtual Screening At Ultra-Large Scale: 1.5 Billion And Counting - Webinars [Электронный ресурс] <http://www.healthtech.com/openeye-scientific-virtual-screening-at-ultra-large-scale/> (Дата обращения: 18.01.2023)
- Blaze Cloud from Cresset. [Электронный ресурс] www.cresset-group.com/products/blaze/#blaze-cloud (Дата обращения: 18.01.2023)
- Ivashko V, Ivashko E. BOINC-Based Volunteer Computing Projects: Dynamics and Statistics. In: Voevodin V, Sobolev S, Yakobovskiy M, Shagaliev R. (eds.) *Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science*, v. 13708. Springer, Cham; 2022. p. 619–631. doi:10.1007/978-3-031-22941-1_45
- Tingting L, Lu D, Zhang H, et al. Applying high-performance computing in drug discovery and molecular simulation. *National Science Review*. 2016;3(1): 49–63. doi:10.1093/nsr/nww003
- Nikitina N, Ivashko E, Tchernykh A. Congestion Game Scheduling for Virtual Drug Screening Optimization.

- Journal of Computer-Aided Molecular Design. 2018;32(2):363-374. doi:10.1007/s10822-017-0093-7
21. Nikitina N, Ivashko E. Optimization of the Workflow in a BOINC-Based Desktop Grid for Virtual Drug Screening. In: Voevodin V., Sobolev S., Yakobovskiy M., Shagaliyev R. (eds.) Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science, v. 13708. Springer, Cham; 2022. p. 686–698. doi:10.1007/978-3-031-22941-1_50
 22. Mazalov VV, Nikitina NN, Ivashko EE. Hierarchical Two-Level Game Model for Tasks Scheduling in a Desktop Grid. In: 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), St. Petersburg, Russia; 2014. p. 541-545 doi:10.1109/ICUMT.2014.7002159
 23. Ivashko E, Litovchenko V. Project Progress Forecasting in a Desktop Grid. In: Voevodin V, Sobolev S. (eds.) Supercomputing. RuSCDays 2021. Communications in Computer and Information Science, v. 1510. Springer, Cham; 2021. p. 487–497. doi:10.1007/978-3-030-92864-3_37
 24. Ivashko E, Nikitina N. Replication of “Tail” Computations in a Desktop Grid Project. In: Voevodin V, Sobolev S. (eds.) Supercomputing. RuSCDays 2020. Communications in Computer and Information Science, v. 1331. Springer, Cham; 2020. doi:10.1007/978-3-030-64616-5_52
 25. Chernov I, Nikitina N. Virtual Screening in a Desktop Grid: Replication and the Optimal Quorum. In: Malyshekin V. (eds.) Parallel Computing Technologies. PaCT 2015. Lecture Notes in Computer Science, v. 9251. Springer, Cham; 2015. p. 258-267. doi:10.1007/978-3-319-21909-7_25
 26. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28:235-242. doi: 10.1093/nar/28.1.235

Ивашко Евгений Евгеньевич. Институт прикладных математических исследований — обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра "Карельский научный центр Российской академии наук"; Федеральное государственное бюджетное образовательное учреждение высшего образования "Петрозаводский государственный университет", Петрозаводск, Россия. Старший научный сотрудник, кандидат физико-математических наук, доцент кафедры прикладной математики и кибернетики. Область научных интересов: высокопроизводительные вычисления, распределенные вычисления, добровольные вычисления, Desktop Grid. E-mail: ivashko@krc.karelia.ru

Никитина Наталья Николаевна. Институт прикладных математических исследований — обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра "Карельский научный центр Российской академии наук", Петрозаводск, Россия. Старший научный сотрудник, кандидат технических наук. Область научных интересов: высокопроизводительные вычисления, распределенные вычисления, добровольные вычисления, Desktop Grid, управление заданиями, теория игр. E-mail: nikitina@krc.karelia.ru

The Concept of a Virtual Drug Screening Service Based on Distributed Computing High-Throughput Virtual Screening as a Service

E. E. Ivashko^{1,†}, N. N. Nikitina¹

¹Institute of Applied Mathematical Research - a separate subdivision of the Federal State Budgetary Institution of Science of the Federal Research Center "Karelian Research Center of the Russian Academy of Sciences", Petrozavodsk, Russia

[†]Federal State Budgetary Educational Institution of Higher Education "Petrozavodsk State University", Petrozavodsk, Russia

Abstract. Drug development is a time-consuming and resource-consuming problem. At the first stages of drug development, virtual screening plays an important role. Virtual screening is *in silico* selection of chemical compounds with potentially high required biochemical activity. To implement it, one typically needs high-performance computing resources, as well as software solutions for organizing the full loop of virtual screening. The paper presents the concept of High-Throughput Virtual Screening as a Service, which is a cloud-based virtual drug screening service based on the concept of Desktop Grid-type distributed computing. We describe three logical levels of service operation, user workflows and the principles of multi-user access. The service will be a software system with the necessary functionality to solve the relevant, computationally intensive problem of virtual screening using the almost unlimitedly scalable Desktop Grid computing resources.

Keywords: Virtual Screening, Volunteer Computing, Desktop Grid, Cloud Computing, Desktop Grid as a Service, Virtual Screening as a Service, High-Throughput Virtual Screening as a Service.

DOI 10.14357/20718632230311

References

1. Anderson DP. BOINC: A Platform for Volunteer Computing. *Journal of Grid Computing*. 2020;18:99–122. doi:10.1007/s10723-019-09497-9.
2. Ivashko E. Desktop Grid and Cloud Computing: Short Survey. In: Voevodin V, Sobolev S. (eds.) *Supercomputing. RuSCDays 2021. Communications in Computer and Information Science*, vol 1510. Springer, Cham; 2021. p. 445–456. doi:10.1007/978-3-030-92864-3_34.
3. Ivashko E. Desktop Grid as a Service Concept. In: Voevodin V, Sobolev S, Yakobovskiy M, Shagaliev R. (eds.) *Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science v. 13708*. Springer, Cham; 2022. p. 632–643. doi:10.1007/978-3-031-22941-1_46.
4. Bechelane Maia EH, Assis LC, Alves de Oliveira T, Marques da Silva A, Gutterres Taranto A. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Frontiers in Chemistry*. 2020;8:343. doi:10.3389/fchem.2020.00343.
5. Torres PHM, Sodero ACR, Jofily P, Silva-Jr. FP. Key Topics in Molecular Docking for Drug Design. *International Journal of Molecular Sciences*. 2019;20(18):4574. doi:10.3390/ijms20184574.
6. Olğaç A, Türe A, Olğaç S, Möller S. Cloud-Based High Throughput Virtual Screening in Novel Drug Discovery. In: Kołodziej J, González-Vélez H. (eds.) *High-Performance Modelling and Simulation for Big Data Applications. Lecture Notes in Computer Science v. 11400*. Springer, Cham; 2019. p. 250–278. doi:10.1007/978-3-030-16272-6_9.
7. Singh N, Chaput L, Villoutreix BO. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Briefings in Bioinformatics*. 2021;22(2):1790–1818. doi:10.1093/bib/bbaa034.
8. Zhang B, Li H, Yu K, Jin Z. Molecular docking-based computational platform for high-throughput virtual screening. *CCF Trans High Perform Comput*. 2022;4(1):63–74. doi:10.1007/s42514-021-00086-5.
9. Mo Q, Xu Z, Yan H, Chen P, Lu Y. VSTH: a user-friendly web server for structure-based virtual screening on Tianhe-2. *Bioinformatics*. 2023;39(1):btac740. doi:10.1093/bioinformatics/btac740.
10. Nikitina N, Manzyuk M, Podlipnik Č, Jukic M. Volunteer computing project SiDock@home for virtual drug screening against SARS-CoV-2. In: Byrski A, Czachórski T, Gelenbe E, Grochla K, Murayama Y. (eds.) *Computer Science Protecting Human Society Against Epidemics. ANTICOVID 2021. IFIP Advances in Information and Communication Technology 2021 v. 616*. Springer, Cham. p. 23–34. doi:10.1007/978-3-030-86582-5_3.
11. Mottin M, de Paula Sousa BK, de Moraes Roso Mesquita NC, et al. Discovery of New Zika Protease and Polymerase Inhibitors through the Open Science Collaboration Project OpenZika. *Journal of Chemical Information and Modeling*. 2022;62(24):6825–6843. doi:10.1021/acs.jcim.2c00596.
12. Zhang B, D’Erasmus MP, Murelli RP, Gallicchio E. Free energy-based virtual screening and optimization of RNase H inhibitors of HIV-1 reverse transcriptase. *ACS omega*. 2016;1(3):435–447. doi:10.1021/acsomega.6b00123.
13. Ivashko E.E., Nikitina N.N., Möller S. High-performance virtual screening in a BOINC-based Enterprise Desktop Grid. *Bulletin of the South Ural State University, Series “Computational Mathematics and Software Engineering”*. 2015;4(1):57–63 (in Russ). doi:10.14529/cmse150105.
14. Lloyd DG, Golfis G, Knox AJ, et al. Oncology exploration: charting cancer medicinal chemistry space. *Drug discovery today*. 2006;11(3-4):149–159. doi:10.1016/S1359-6446(05)03688-3.
15. Krasoulis A, Antonopoulos N, Pitsikalis V, Theodorakis S. DENVIS: Scalable and High-Throughput Virtual Screening Using Graph Neural Networks with Atomic and Surface Protein Pocket Features. *Journal of Chemical Information and Modeling*. 2022;62:4642–4659. doi:10.1021/acs.jcim.2c01057.
16. Hawkins P. Virtual Screening At Ultra-Large Scale: 1.5 Billion And Counting - Webinars [Digital resource] <http://www.healthtech.com/openeye-scientific-virtual-screening-at-ultra-large-scale/> (Accessed by: 18.01.2023)
17. Blaze Cloud from Cresset. [Digital resource] www.cresset-group.com/products/blaze/#blaze-cloud (Accessed by: 18.01.2023).
18. Ivashko V, Ivashko E. BOINC-Based Volunteer Computing Projects: Dynamics and Statistics. In: Voevodin V, Sobolev S, Yakobovskiy M, Shagaliev R. (eds.) *Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science, v. 13708*. Springer, Cham; 2022. p. 619–631 doi:10.1007/978-3-031-22941-1_45.
19. Tingting L, Lu D, Zhang H, et al. Applying high-performance computing in drug discovery and molecular simulation. *National Science Review*. 2016;3(1): 49–63. doi:10.1093/nsr/nww003.
20. Nikitina N, Ivashko E, Tchernykh A. Congestion Game Scheduling for Virtual Drug Screening Optimization. *Journal of Computer-Aided Molecular Design*. 2018;32(2):363–374. doi:10.1007/s10822-017-0093-7.
21. Nikitina N, Ivashko E. Optimization of the Workflow in a BOINC-Based Desktop Grid for Virtual Drug Screening. In: Voevodin V., Sobolev S., Yakobovskiy M., Shagaliev R. (eds.) *Supercomputing. RuSCDays 2022. Lecture Notes in Computer Science, v. 13708*. Springer, Cham; 2022. p. 686–698. doi:10.1007/978-3-031-22941-1_50.
22. Mazalov VV, Nikitina NN, Ivashko EE. Hierarchical Two-Level Game Model for Tasks Scheduling in a Desktop Grid. In: 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), St. Petersburg, Russia; 2014. p. 541–545 doi:10.1109/ICUMT.2014.7002159.

23. Ivashko E, Litovchenko V. Project Progress Forecasting in a Desktop Grid. In: Voevodin V, Sobolev S. (eds.) Supercomputing. RuSCDays 2021. Communications in Computer and Information Science, v. 1510. Springer, Cham; 2021. p. 487–497. doi:10.1007/978-3-030-92864-3_37.
24. Ivashko E, Nikitina N. Replication of “Tail” Computations in a Desktop Grid Project. In: Voevodin V, Sobolev S. (eds.) Supercomputing. RuSCDays 2020. Communications in Computer and Information Science, v.1331. Springer, Cham; 2020. doi:10.1007/978-3-030-64616-5_52.
25. Chernov I, Nikitina N. Virtual Screening in a Desktop Grid: Replication and the Optimal Quorum. In: Malyshkin V. (eds.) Parallel Computing Technologies. PaCT 2015. Lecture Notes in Computer Science, v. 9251. Springer, Cham; 2015. p. 258-267. doi:10.1007/978-3-319-21909-7_25.
26. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Research. 2000;28:235-242. doi: 10.1093/nar/28.1.235.

Ivashko Evgeny E. PhD, Institute of Applied Mathematical Research - separate subdivision of the Federal State Budgetary Institution of Science of the Federal Research Center "Karelian Research Center of the Russian Academy of Sciences", Pushkinskaya 11, 185910, Petrozavodsk, Russia; Federal State Budgetary Educational Institution of Higher Education "Petrozavodsk State University", Lenina 33, 185035, Petrozavodsk, Russia. E-mail: ivashko@krc.karelia.ru

Nikitina Natalia N. PhD, Institute of Applied Mathematical Research - separate subdivision of the Federal State Budgetary Institution of Science of the Federal Research Center "Karelian Research Center of the Russian Academy of Sciences", Pushkinskaya 11, 185910, Petrozavodsk, Russia. E-mail: nikitina@krc.karelia.ru