ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Черных Андрея Николаевича

«Методы и алгоритмы решения задач оптимизации ресурсов в нестационарных распределенных гетерогенных вычислительных средах»,

представленную на соискание ученой степени доктора физико-математических наук по специальности 2.3.5 (05.13.11) – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Актуальность темы диссертационной работы

Высокие темпы развития вычислительной техники, увеличение роли сетевых информационных ресурсов, повсеместная интеграция высокотехнологичных мобильных цифровых устройств и виртуальных сервисов в быт современного человека в последние десятилетия послужили причиной резкого увеличения объема цифровых данных. Поиски методов эффективной обработки больших массивов данных привели исследователей к идее объединения ресурсов различных вычислительных устройств в облачные структуры. Развитием данной идеи стала концепция распределенной гетерогенной вычислительной среды (ГРИД), объединяющей вычислительные ресурсы территориально удаленных друг от друга разнотипных групп вычислительных устройств. Данный подход открыл принципиально новые возможности для эффективного решения фундаментальных и прикладных задач, обеспечивающих конкурентное преимущество в научной, производственной, экономической и других сферах человеческой деятельности. Однако данный подход также обнаружил новые проблемы, не характерные для многопроцессорных систем, имеющих схожий функционал и назначение. В отличие от многопроцессорных систем, гетерогенная вычислительная среда характеризуется нестационарностью, что сильно усложняет процесс планирования. Она в большей степени подвержена модификации, увеличению разнородности, масштабности и т. д. Влияние указанных факторов на эффективность вычислений трудно прогнозируемо, в связи с чем появляется фактор неопределенности при их планировании.

Именно решению этой актуальной задачи посвящено данное диссертационное исследование, целью которого является разработка новых стратегий планирования в нестационарных гетерогенных средах с помощью динамических и адаптивных алгоритмов.

Содержание диссертационной работы

Представленная диссертационная работа выполнена в Федеральном государственном бюджетном учреждении науки Институт системного программирования имени В. П. Иванникова Российской академии наук. Она включает введение, семь глав, заключение, список литературы из 223 наименований и два приложения. Общий объем работы 325 стр., в том числе 294 стр. основного текста, включающего 72 рисунка и 49 таблиц. Содержание диссертации адекватно отражено в автореферате общим объемом 38 стр.

Во введении сформулированы цель и задачи диссертационной работы, обоснована актуальность, подчеркнуты научная новизна и практическая значимость полученных результатов, а также приведены основные положения, выносимые на защиту.

В первой главе описаны основные источники неопределенности, приведена их классификация, сформулирована проблема планирования работ в нестационарных распределенных гетерогенных вычислительных системах, которая является актуальной нерешенной ранее научной задачей. Приведен краткий обзор основных современных моделей и методов, используемых при решении этой задачи с помощью теории расписаний. Показано, что нестационарные распределенные гетерогенные системы – это сравнительно новый динамично развивающийся подход к организации распределенных вычислений особенно актуальный при создании облачных сервисов. Использование известных частных решений, разработанных как для стационарных, так и для динамических систем, приводит к необходимости существенной доработки алгоритмов под конкретные условия использования. Для решения этой актуальной проблемы предложен комплексный подход к построению адаптивных планировщиков и математических моделей, учитывающих отсутствие точных знаний при формировании плана работ чтобы уменьшить влияние нестационарного поведения системы на ее производительность.

Во второй главе представлен подход к моделированию ГРИД с учетом неопределенности характеристик, динамики распределенных вычислений, масштабирования, отсутствия точной информации о времени вычисления работ и т. д. для предоставления надежных решений, где основной целью является не поиск абсолютно оптимальных решений, а нахождение приближенных решений нечувствительных к изменению параметров среды. Показано, что производительность классического списочного алгоритма планирования Гарея и Грэхэма хуже в распределенных системах, чем в мультипроцессорах. Поскольку обычное списочное расписание не подходит для ГРИД, представлен планировщик, использующий несколько списков работ для каждой машины. Каждый из этих списков не требует какого-либо конкретного порядка. Представлен алгоритм планирования, гарантирующий конкурентный фактор 5 и аппроксимационный фактор 3. Этот алгоритм может быть реализован с использованием подхода «кражи работ» и может хорошо подходить для использования для алгоритмов планирования в реальных системах.

В третьей главе анализируются алгоритмы планирования работ, интегрирующие распределение работ по машинам и их локальное расписание. Предлагается и анализируется адаптивная схема распределения работ с использованием концепции допустимого распределения (англ., admissible allocation). Основная идея этой схемы заключается в том, чтобы установить ограничения по распределению работ на машины и динамически адаптировать их к различным рабочим нагрузкам и свойствам системы снижая возможный дисбаланс нагрузки. Представлены 3-х-и 9-аппроксимационные и 5-ти и 11-конкурентные алгоритмы. Проведена экспериментальная оценка двухэтапных стратегий планирования с допустимым коэффициентом, включенным в политику планирования. Представлено подробное исследование его влияния, на общую производительность ГРИД. Показано, что с точки зрения рассматриваемых критериев, стратегии распределения с допустимым коэффициентом превосходят алгоритмы, которые используют все доступные машины для распределения работ. Такие адаптивные стратегии планирования надежны и стабильны даже в сильно различающихся условиях и способны успешно справляться с различными рабочими нагрузками.

Четвертая глава посвящена задаче планирования, оптимизация которой не всегда эффективна классическими методами из-за разнообразия реальных параллельных и распределенных платформ и/или сред. Показано, что адаптивные алгоритмы, способные динамически изменять распределение работ во время исполнения для оптимизации поведения глобальной системы, являются лучшей альтернативой для решения этой задачи. В главе рассматривается планирование параллельных работ с известными требованиями к ресурсам, но с неизвестным временем выполнения, фокусируясь на регулирование периодов простоя машин при выполнении. Рассматривается новое семейство стратегий планирования, основанных на двух фазах, которые сочетают последовательное и параллельное выполнение работ. Обобщаются известные предельные границы производительности в наихудшем случае (аппроксимационный фактор), учитывая, помимо числа процессоров и максимальных требований к процессору, рассматриваемых в литературе, два дополнительных параметра, а именно: штраф за распараллеливание работ и коэффициент регулирования простоя. Кроме того, доказывается, что регулирование простоя может улучшить аппроксимационный фактор планирования параллельных работ в режиме разделения пространства. Эта схема балансирует потребности пользователя (работ) с потребностями компьютерной системы.

В пятой главе рассматриваются проблемы планирования виртуализированных вычислительных ресурсов, предоставляемых пользователям через Интернет в форме облачного сервиса. В типичном сценарии Infrastructure as a Service – IaaS поставщик инфраструктуры предлагает свои ресурсы по требованию и с различными уровнями сервиса своим клиентам. Эти уровни сервиса в основном отличаются количеством вычислительной мощности, которое заказчик гарантированно получит в течение определенного периода времени. Для формализации уровня сервиса вводится слак-фактор и цена за единицу времени обработки. Предлагаются различные алгоритмы и приводится конкурентный анализ для обсуждения различных сценариев для данной модели. Эти сценарии объединяют фиксированные уровни обслуживания с одной или несколькими машинами. Демонстрируется, как можно достигнуть лучшего конкурентного фактора.

Шестая глава рассматривает модификацию модели IaaS, приведенную в главе 5 анализируя более реалистичные сценарии, где для клиентов предоставляются несколько уровней обслуживания, а также рассматривается двухкритериальная оптимизация: увеличение дохода провайдера и снижение энергопотребления. Алгоритмы используют минимальную информацию, что важно в нестационарной среде, и характеризуется небольшой вычислительной сложностью. Тем не менее, они позволяют достичь хорошего улучшения критериев и обеспечивают гарантированное качество обслуживания.

В седьмой главе исследуются проблемы планирования, связанные с облачными системами телефонии VoIP. Они учитывают динамическую нагрузку, вариативность времени запуска VM, свойства работ, инфраструктуры, наличия других пользователей, которые совместно используют сервис, и т. д. Сформулированная задача рассматривается как частный случай динамической упаковки в контейнеры. Контейнеры представляют собой виртуальные машины, а высота элементов определяет вклад вызова в загрузку виртуальной машины. Планировщик знает только вклад вызова в загрузку VM. Все решения принимаются без информации о продолжительности вызова, скорости поступления вызовов и т. д. Принципиальной новизной данной проблемы является временное существование элементов (вызовов) в упаковке контейнера. В отличие от стандартной формулировки, контейнеры всегда открыты и динамичны, даже если они полностью заполнены. Элементы в контейнерах могут быть удалены (завершение вызова), а утилизация VM может быть уменьшена в любой момент времени, тогда VM могут использовать свободное пространство для обработки новых вызовов. Параметры

алгоритмов могут быть динамически адаптированы к различным предпочтениям, рабочим нагрузкам и свойствам облака. Результаты экспериментов на реальных данных компании MiXvoip показывают, что предложенные алгоритмы с методами прогнозирования нагрузки превосходят известные стратегии, обеспечивая высокое качество обслуживания и более низкую стоимость и могут быть эффективно использованы в облачной среде VoIP.

Научная новизна диссертационной работы

В работе сформулирована проблема планирования в ГРИД в условиях ее нестационарности, предложены подходы к ее решению путем динамической адаптации к изменению системных параметров, получены оригинальные теоретические оценки границ оптимизации распределения ресурсов, улучшающие известные оценки.

Все результаты, полученные в рамках диссертационной работы, являются новыми, достоверными и соответствуют основным положениям ВАК РФ.

Теоретическая и практическая значимость результатов диссертационного исследования

Работа носит теоретический характер, в ней предложены новые приближенные алгоритмы с лучшими по сравнению с известными в литературе аппроксимационными и конкурентными факторами. Применение результатов диссертационного исследования обеспечивает повышение эффективности планирования параллельных работ в современных вычислительных системах путем его адаптации к изменению параметров рабочей нагрузки и системы. Также следует отметить, что предложенные алгоритмы имеют низкую вычислительную сложность и при этом позволяют получать хорошие приближенные решения. Значимость полученных результатов и вклад диссертанта в развитие соответствующей отрасли знаний подтверждается более чем внушительным числом цитирований результатов в международных изданиях: 2512 ссылок в Google Scholar (h-index = 25) и 1412 ссылок в Scopus (h-index = 20). Основные результаты диссертационного исследования были использованы в рамках научно-технических работ, среди которых проекты: Министерства науки и образования Российской Федерации, РФФИ – Российского фонда фундаментальных исследований, CICESE Research Center (Мексика), ANII – Национального агентства по изучению и инновациям (Уругвай), FNR – Национального фонда научных исследований (Люксембург), BSC – Барселонского суперкомпьютерного центра (Испания) и других научно-исследовательских организаций.

Достоверность и обоснованность результатов диссертационного исследования подтверждена корректным применением классических методов исследования, строгими доказательствами и анализом эффективности разработанных моделей и алгоритмов. Результаты проведенных численных экспериментов корректны. Они коррелируют с теоретическими утверждениями диссертации и текущим уровнем научного знания, а также доказывают целесообразность и актуальность применения предложенных автором стратегий, моделей, методов и алгоритмов планирования вычислений.

Основные результаты работы докладывались и обсуждались на 88 международных и 14 всероссийских научных конференциях и семинарах. Всего по теме диссертационного исследования автором было опубликовано 64 статьи, в том числе 21 статья в журналах из списка ВАК или в приравниваемых к ним журналах, включенных в международные базы данных Scopus и Web of Science, 42 работы опубликованы в трудах российских и международных конференций, получен 1 патент. Также получены 4 свидетельства о регистрации программ для ЭВМ. Высокий индекс цитируемости журналов и уровень конференций, в которых опубликованы и на которых апробировались научные статьи

соискателя, говорит о высоком уровне значимости и качестве результатов исследований, представленных в данных работах. Диссертация не содержит заимствованных материалов или отдельных результатов без ссылок на авторов и источники заимствования.

Рекомендации по использованию результатов и выводов диссертации. Разработанные стратегии, модели, методы и алгоритмы могут быть в явном виде использованы при организации работы планировщиков распределенных облачных и ГРИД ресурсов для повышения их эффективности. Рекомендуется внедрение учебных курсов, основанных на материалах диссертационной работы соискателя.

Тематика работы и основные результаты диссертации соответствуют следующим областям исследований паспорта специальности ВАК 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»: п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем», п. 8 «Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования», п. 9 «Модели, методы, алгоритмы и программная инфраструктура для организации глобально распределенной обработки данных».

Замечания.

В разделе 7.7.2.2 диссертации рассматриваются вопросы, связанные с настройкой нейронной сети для прогнозирования нагрузки. Однако накладные расходы на обучение сети не приводятся. Их учет мог быть полезен при определении рационального объема обучающей выборки в случае чрезмерного возрастания временных затрат, требуемых на обучение.

В распределенных системах появляются новые угрозы информационной безопасности, было интересно исследовать вопрос влияния механизмов информационной безопасности на показатели оптимальности использования ресурсов распределенной системы.

Некоторые термины, которые являются «калькой» англоязычных слов затрудняют чтение, например, «слак-фактор» и «стрейтч-фактор», хотя в целом работа читается хорошо.

Заключительная оценка работы.

Указанные замечания не являются критическими и не снижают научную и практическую ценность работы и проведенных исследований.

Таким образом, диссертация Андрея Николаевича Черных является законченным трудом, в котором на основании выполненных автором исследований и разработок осуществлено решение крупной научной проблемы разработки фундаментальных основ и новых адаптивных алгоритмов планирования нестационарных ресурсов, а также проведения их анализа для различных сценариев функционирования ГРИД. Совокупность разработанных теоретических положений можно квалифицировать как новое достижение в развитии перспективного направления специальности «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертационная работа отвечает требованиям ВАК, предъявляемым к докторским диссертациям по специальности 2.3.5 (05.13.11) – "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей", а ее автор,

Черных Андрей Николаевич, заслуживает присуждения ему ученой степени доктора физико-математических наук по специальности 2.3.5.

Официальный оппонент –
Доктор технических наук (специальность 05.13.15 «Вычислительные машины, комплексы и компьютерные сети»), доцент
заместитель директора по научной работе – директор Межведомственного суперкомпьютерного центра академии наук – филиала Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук» (МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН)

Шабанов Борис Михайлович

«22» ноября 2021 г.

Подпись директора, д.т.н. Б.М. Шабанов заверяю

Заместитель директора по научной работе
МСЦ РАН – филиала ФГУ ФНЦ НИИСИ РАН П.Н. Телегин