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

1. 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 отражено в ее автореферате, представленном на 38 стр.

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

Первая глава является обзорной. В ней рассматриваются существующие проблемы планирования облачных вычислений в гетерогенной среде. Выделяются источники неопределенности, возникающей в процессе планирования, и предлагается их классификация. Обсуждаются вопросы, связанные с уровнем обслуживания работ при их выполнении с использованием облачных ресурсов и применением необходимых ограничений, которые должны обеспечивать качество обслуживания.

Во второй главе представлена модель распределенной среды и проведено ее сравнение с моделью многопроцессорной системы с точки зрения планирования работ. Показано, что планирование в рассматриваемой среде является более сложной задачей, чем соответствующее планирование в многопроцессорной системе. Предложены новые алгоритмы планирования работ, как заданных списком, так и поступающих во времени (онлайн). Сформулированы и доказаны оценки границ оптимизации ресурсов, достижимых при использовании данных алгоритмов.

В третьей главе рассматривается онлайн-планирование параллельных работ без прерываний в гетерогенных средах. Предложены алгоритмы планирования, основанные на регулируемой допустимости ресурсов и представлен детальный анализ их работы. Показано, что с точки зрения рассматриваемых в диссертации критериев, стратегии распределения, сужающие допустимое для назначения задания множество ресурсов («допустимые стратегии»), превосходят стратегии, которые используют все доступные ресурсы для распределения работ. Адаптивные допустимые стратегии планирования надежны и стабильны даже в сильно различающихся условиях. Они успешно справляются с различными рабочими нагрузками. Коэффициент допустимости может быть динамически скорректирован для того, чтобы справиться с изменяющейся рабочей нагрузкой.

Четвертая глава посвящена планированию параллельных работ с неизвестным временем их выполнения и регулированием периодов простоя машин. Предложены новые адаптивные схемы динамического изменения алгоритма планирования во время его выполнения для оптимизации поведения системы в целом. Рассмотрены новые алгоритмы, у которых отсутствуют любые знания о состоянии среды и ее ресурсов, кроме числа незавершенных работ и их требований по процессорам. Данные алгоритмы используют пакетное планирование в рамках общего двухфазного подхода. Оно включает две последовательно сменяющиеся фазы выполнения каждого из приложений на одном или на нескольких процессорах. Проведен анализ работы предложенных алгоритмов и обобщены

известные пределы производительности в худшем случае, путем введения двух дополнительных параметров: штрафа за распараллеливание работ и фактора регулирования простоя в дополнение к числу процессоров и максимальным требованиям к процессорам, представленных в публикациях других авторов.

В пятой главе введена простая модель для назначения и планирования работ на основе уровней обслуживания, при построении которой учитывались аспекты расписаний работ реального времени. Рассмотрены сценарии планирования с одной или несколькими одинаковыми машинами. Для этих сценариев получены верхние границы оптимизации распределения ресурсов и оценки их производительности с помощью конкурентного анализа. На основе полученных результатов теоретически подтверждено преимущество использования многопроцессорных систем по сравнению с набором последовательных, что соответствует общему пониманию современного положения дел.

Шестая глава представляет следующие результаты: сформулирована постановка задачи двухкритериальной оптимизации распределения ресурсов с несколькими уровнями обслуживания с учетом дохода провайдера и энергопотребления. Проанализированы сценарии с однородными и разнородными машинами разных конфигураций в условиях разнообразных рабочих нагрузок. Проведено всестороннее экспериментальное исследование жадных алгоритмов приемки работ с известной границей производительности в наихудшем случае и восьми стратегий планирования, учитывающих неоднородность среды. Проведен совместный анализ двух конфликтующих критериев оптимизации. Определена доминирующая стратегия, стабильно работающая в условиях изменения условий функционирования среды, а также обеспечивающая улучшение критериев оптимизации и требуемое качество обслуживания.

Седьмая глава представляет результаты исследования, связанные с планированием работ в облачных системах передачи голосовых сообщений через интернет протоколы (VoIP). Сформулированы и исследованы задачи планирования с учетом динамической нагрузки и прогнозированием начала запуска обеспечивающих сервис виртуальных машин. Для оптимизации используется двухкритериальная модель, учитывающая стоимость услуг в тарификационных часах и качество обслуживания, определяемое скоростью обработки вызовов и временем их ожидания на линиях связи. Экспериментальные исследования в рамках комплексного моделирования облачных систем VoIP, осуществляемого с целью сравнительного анализа разработанных стратегий планирования с используемыми на практике, выполнены на реальных данных. Показано, что предложенные стратегии с прогнозированием рабочей нагрузки существенно превосходят известные стратегии, обеспечивая при этом приемлемое качество обслуживания и более низкую стоимость.

В заключении сформулированы выводы по диссертационной работе и приведены ее основные результаты.

2. Актуальность темы диссертационной работы

Анализ современных тенденций развития моделей и алгоритмов планирования облачных вычислений, базирующийся на теоретических и практических исследованиях ведущих российских и зарубежных ученых, очевидно показывает необходимость применения комплексного подхода к разработке адаптивных планировщиков и математических моделей, учитывающих отсутствие точных знаний при формировании планов работ.

Именно на решение этой актуальной задачи направлено данное диссертационное исследование, целью которого является разработка новых стратегий планирования распределенных вычислений в нестационарных гетерогенных средах с помощью динамических и адаптивных алгоритмов, функционирующих в условиях наличия неполноты сведений о характеристиках работ и ресурсов среды.

3. Научная новизна диссертационной работы

Результаты диссертационного исследования предоставляют возможность оценки различных методов планирования выполнения потоков заданий в облачных гетерогенных средах и предоставляют методы обеспечивающие эффективное управление подобными средами в условиях нестационарности и неопределенности, что определяет их новизну. В диссертационном исследовании получены новые оригинальные теоретические оценки границ оптимизации распределения ресурсов. Кроме того, предложенные приближенные алгоритмы позволяют улучшить оценки, ранее полученные другими авторами. Все результаты, представленные в данной диссертации, являются новыми.

4. Практическая значимость результатов диссертационного исследования

Результаты, представленные в диссертации, нашли свое применение в рамках ряда научно-технических работ, выполненных в российских и зарубежных научно-исследовательских организациях. Их применение обеспечило повышение эффективности планирования параллельных работ в гетерогенных распределенных средах. Апробация путём проведения многочисленных вычислительных экспериментов подтвердила широкие возможности адаптации предложенных алгоритмов к изменению параметров рабочей нагрузки. Кроме того, эти алгоритмы имеют достаточно низкую трудоемкость и обеспечивают возможность получения хороших приближенных решений в реальном времени.

5. Достоверность и обоснованность результатов диссертационного исследования

В рамках диссертации корректно применяются классические методы исследования, приводятся строгие доказательства и проводится анализ эффективности разработанных моделей и алгоритмов на экспериментальных данных. Полученные результаты исследования подтверждаются численными экспериментами.

6. Замечания

К работе имеются следующие замечания:

1) Проблема обеспечения надежности облачных вычислений кратко упоминается в первой главе, хотя, судя по списку публикаций соискателя, он имеет непосредственное отношение к развитию данного направления исследований. Представляется целесообразным более подробное рассмотрение этого аспекта организации распределенных вычислений в гетерогенных средах.

2) Следовало яснее определить круг или классы приложений, подлежащих к планированию для выполнения на распределённых вычислительных ресурсах. Не вполне ясно, имеются ли в виду приложения класса HPC, ориентированных на использование большого числа сильно связанных вычислительных узлов суперкомпьютерных установок, или что-то другое. Если верно первое, то трудно согласится с общим тезисом на странице 133: «Число процессоров, выбираемых пользователем для выполнения работы, как правило, основано на характере задачи или желании пользователя и не учитывает рабочие характеристики многопроцессорной системы». Пользователи соответствующего сегмента хорошо представляют требования к вычислительным узлам и к системе в целом, поскольку без знания таких характеристик обоснованный заказ вычислительного ресурса невозможен. Более того, планирование задач данного класса требует учёта накладных расходов на перенос значительных, в общем случае, объёмов расчетных данных, как стартовых, так и промежуточных, в случае «прерывания» выполнения задания на одной системе и её перезапуска на другой. Соответствующие вопросы не в полной мере представлены в диссертации. Следует отметить, что уровень автоматизации процедуры «прерывания» в случае HPC приложений оставляет желать лучшего, вопреки излишне общему, с моей точки

зрению, тезису на странице 135 «Реальные системы поддерживают прерывания, поэтому предполагаем, что работа может быть прервана и перераспределена при необходимости».

3) В тексте диссертационной работы используется множество терминов и понятий, не все из которых являются общепринятыми и однозначно интерпретируемыми. В совокупности с определённой неаккуратностью оформления текстов диссертации и автореферата, отсутствие глоссария существенно затрудняет их восприятие. В ряде случаев термины, если и объясняются, то значительно позже по тексту, нежели начинают использоваться. Например, два ключевых понятия «конкурентного фактора» и «аппроксимационного фактора» используются постоянно, начиная с одиннадцатой страницы текста, а косвенный комментарий о их смысле присутствует только на странице 50, причём одно определение даётся для двух понятий «$\rho-$конкурентности» и «$\rho-$аппроксимации». На страницах 33, 136 присутствуют множественные опечатки в формулах определения $\bar{\mu}$ и $\underline{\mu}$, там же используется неопределённый в диссертации термин «вычислительная площадь», интерпретировать который можно совершено разными способами. Более чёткое определение основных понятий и обозначений в заметной мере облегчило бы восприятие текста.

5) 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.

7. Общая оценка работы

Отмеченные недостатки не влияют на общее положительное впечатление от рассмотренного диссертационного исследования. Представлена интересная, выполненная на высоком научном уровне работа. В её рамках значительное внимание уделено всестороннему анализу возможных вариантов решения рассматриваемых актуальных задач и, что не менее важно, валидации предлагаемых методов с помощью выполнения убедительных вычислительных экспериментов на основе фактических экспериментальных данных с множества доступных облачных и ГРИД ресурсов, что несомненно является одной из её сильных сторон. Данная работа изложена в хорошем научном стиле, ясно и доказательно. Автореферат верно отражает содержание диссертации.

Основные положения диссертации являются вполне обоснованными. Диссертационное исследование прошло достаточную апробацию на всероссийских и международных симпозиумах и конференциях. Его основные результаты представлены в 63 публикациях, в том числе в 21 статье в журналах из списка ВАК или изданиях, проиндексированных в международные базы данных Scopus и Web of Science. Кроме того, получены 4 свидетельства о государственной регистрации программ для ЭВМ и 1 патент.

Результаты исследований, представленные в диссертационной работе, соответствуют следующим пунктам области исследований в паспорте специальности 2.3.5 (05.13.11) – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей:

- модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем;
- модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования;
- модели, методы, алгоритмы и программная инфраструктура для организации глобально распределенной обработки данных.

8. Заключение

Представленная диссертация является завершенной крупной научно-исследовательской работой по актуальному направлению исследований в области создания методов управления распределенными вычислительными системами, выполненной соискателем на высоком научном уровне. В ней решены важные научные задачи эффективного планирования обслуживания множественных потоков заданий. Практическая и теоретическая значимость полученных результатов, а также весомый вклад диссертанта в развитие соответствующей отрасли знаний не вызывает сомнений. На основании вышесказанного считаю, что работа отвечает всем требованиям ВАК, предъявляемым к диссертационным работам, а ее автор, Черных Андрей Николаевич, заслуживает присуждения степени доктора физико-математических наук по специальности 2.3.5 (05.13.11) – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Официальный оппонент –
доктор физико-математических наук, профессор,
член-корреспондент РАН

Якобовский Михаил Владимирович

23 ноября 2021 г.

Должность: заместитель директора по научной работе
Место работы: Федеральное государственное учреждение "Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук"

Подпись М.В. Якобовского заверяю
учёный секретарь ИПМ
им. М.В. Келдыша РАН
к.ф.-м.н. А.А. Давыдов

« 23 » ноября 2021 г.