МГУ имени М.В. Ломоносова

4.08.Лаборатория параллельных информационных технологий

№ госрегистрации
АААА-А21-121011690003-6

УТВЕРЖДАЮ
Директор/декан
_____
_____
«__»_____ _____ г.

УДК

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по теме:
Разработка программных средств поддержки жизненного цикла и обеспечения эффективности суперкомпьютерных приложений, систем и центров
(промежуточный)

Зам. директора/декана
по научной работе
_____
«__»_____ _____ г.

Руководитель темы
Воеводин В.В.
_____
«__»_____ _____ г.

Москва 2024

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель темы: ведущий научный сотрудник, доктор физико-математических наук, член-корреспондент, профессор по специальности	(Воеводин В.В.)
Исполнители темы: ведущий научный сотрудник, кандидат физико-математических наук	(Антонов А.С.)
специалист	(Антонова А.П.)
доктор исторических наук, профессор по кафедре	(Винокуров В.И.)
заведующий лабораторией, кандидат физико-математических наук	(Воеводин В.В.)
специалист	(Гамаюнова Т.С.)
	(Горейнов С.А.)
	(Дергунов Н.В.)
	(Зубков Д.А.)
	(Игнатенко А.П.)
	(Колганов А.С.)
младший научный сотрудник	(Копылов К.Е.)
техник	(Крымский С.А.)
ассистент	(Кулагин А.В.)
техник	(Личманов Д.И.)
программист 1 категории, кандидат педагогических наук	(Майер Р.В.)
	(Матвеев В.А.)
ведущий научный сотрудник, кандидат физико-математических наук	(Никитенко Д.А.)
младший научный сотрудник, кандидат химических наук	(Нилов Д.К.)
старший научный сотрудник, кандидат химических наук	(Панин Н.В.)
техник	(Паокин А.В.)
	(Пионткевич А.Г.)
аспирант	(Подшивалов Д.Д.)
	(Попов А.С.)
	(Попов С.М.)
программист 2 категории	(Сетяев А.В.)

специалист	_____	(Сидоров И.Ю.)
	_____	(Скрябин Г.Д.)
	_____	(Струков П.В.)
другие должности	_____	(Фатеев И.Д.)
	_____	(Фатеева А.А.)
техник	_____	(Шайхисламов Д.И.)
заведующий лабораторией, доктор химических наук, профессор по специальности	_____	(Швядас В.К.)
ведущий специалист	_____	(Шоков В.Н.)
программист 2 категории	_____	(Шубин М.В.)
научный сотрудник, кандидат химических наук	_____	(Щербакова Т.А.)

РЕФЕРАТ

Ключевые слова:
эффективность, суперкомпьютерный центр, суперкомпьютер, суперкомпьютерные приложения
Ключевые слова по-английски:
supercomputing center, supercomputer, supercomputing applications, efficiency

В рамках проведения детального анализа всех важных аспектов работы суперкомпьютерного центра, связанных с эффективностью его функционирования, выполнена апробация системы для оценки качества использования суперкомпьютерных ресурсов на реальных данных.

В 2024 году были начаты работы по облегчению возможности использования системы поддержки суперкомпьютерного центра Octoshell в других суперкомпьютерных центрах, добавлено несколько новых инструментов и было продолжено устранение случаев некорректной работы системы.

Развитие подходов к построению систем мониторинга для суперкомпьютеров заключалось в расширении функционала и повышении работоспособности системы мониторинга DiMMon.

В рамках работ по разработке и развитию методов сохранения данных о состоянии суперкомпьютера и построения "срезов" на любой заданный момент времени начаты работы по реорганизации раздела статистики в системе поддержки функционирования суперкомпьютерного центра.

В ходе работ по развитию Открытой энциклопедии свойств алгоритмов выполнена разработка сервиса загрузки больших файлов и технологий визуализации архитектур суперкомпьютеров.

ВВЕДЕНИЕ

В ходе этапа "Тестовая эксплуатация разработанного программного обеспечения на вычислительных установках с нетрадиционными архитектурами, установленных в МГУ" была продолжена апробация программных компонентов и средств, созданных в рамках следующих направлений исследований:

- Проведение детального анализа всех важных аспектов работы суперкомпьютерного центра, связанных с эффективностью его функционирования;
- Разработка модульного программного комплекса поддержки суперкомпьютерного центра;
- Развитие подходов к построению систем мониторинга для суперкомпьютеров;
- Разработка и развитие методов сохранения данных о состоянии суперкомпьютера и построения "срезов" на любой заданный момент времени;
- Развитие Открытой энциклопедии свойств алгоритмов, направленное на замыкание цепочки от особенностей решения конкретных вычислительных задач до их эффективной реализации на вычислительных системах.

ОСНОВНАЯ ЧАСТЬ

В 2024 году была проведена масштабная апробация разработанной ранее системы оценок (интегрирована в ПО TASC), которая позволяет анализировать эффективность использования различных типов ресурсов в пользовательских приложениях. В частности, был осуществлен перенос данной системы совместно с TASC на два больших суперкомпьютера, входящих в списки наиболее мощных систем России.

Была осуществлена подстройка и адаптация решения для более точного соответствия новой архитектуре суперкомпьютеров. Это касалось, в частности, компоненты TASC, связанной с первичным анализом проблем с производительностью в пользовательских приложениях с помощью набора правил. Так, было необходимо модифицировать значения констант, которые задают пороги для различных характеристик производительности и строения вычислительных узлов в соответствии с новой архитектурой. Помимо этого, было необходимо адаптировать набор самих правил, поскольку некоторые из них стало невозможно вычислять на новых системах из-за различий в доступных входных данных (это касалось данных о работе с коммуникационной сетью и файловой системой), зато появились другие входные данные, которые позволяют обнаруживать новые типы проблем.

Также в процессе данного переноса была выполнена модификация разработанного решения, позволяющая сделать его более универсальным, что заметно облегчит последующие переносы в будущем. Ранее в некоторых компонентах TASC была явная привязка к стилю именования вычислительных узлов, было необходимо явно в нескольких местах указывать список разделов, а также предполагалось, что в рамках одного раздела все вычислительные узлы одинаковы. Чтобы устранить эти недостатки, был предложен новый универсальный способ, в рамках которого при переносе требуется лишь создать один небольшой файл с описанием общей архитектуры суперкомпьютера, при этом никаких дополнительных правок в исходном коде TASC не требуется. Такой подход позволил решить все три проблемы: теперь именование узлов может быть любым, список и названия разделов автоматически подхватываются из указанного файла, и разделы могут быть разбиты на группы, в рамках каждой из которых вычислительных узлы одинаковые (что позволяет иметь любое количество разнородных узлов в рамках одного раздела). Помимо этого, были выполнены и другие работы по созданию более универсального решения. Так, для компоненты визуализации на основе Redash был подготовлен переносимый набор запросов к БД, графиков и дашбордов, который содержит только требуемую функциональность и является универсальным. Теперь для разворачивания этой части компоненты визуализации требуется лишь установить сам Redash, импортировать в ее БД созданный бекап (содержащий упомянутый набор) и указать небольшой набор данных конфигурации (логины, пароли и имена БД с входным данными).

Помимо этого, был проведен детальный анализ статистики, собранной за год с помощью системы оценок на суперкомпьютере Ломоносов-2. Были изучены общие показатели эффективности использования суперкомпьютерных ресурсов, проанализированы распределения оценок в целом и отдельные оценки для пользователей, проектов и прикладных пакетов в частности. Например, было обнаружено, что пользователи нередко собирают

собственные версии прикладных пакетов, однако при это эффективность таких сборок зачастую ниже, чем эффективность централизованно установленных версий этих пакетов. Помимо этого, были обнаружены пользователи, задания которых показывают очень низкую эффективность использования предоставленных им ресурсов. Так, у приложений одного из пользователей среднее число активных процессов было более 100, что в несколько раз превышает оптимальное значение и создает заметные накладные расходы, которые мешают эффективной работе таких приложений.

В 2024 году были начаты работы по облегчению возможности использования Octoshell в других суперкомпьютерных центрах, добавлено несколько новых инструментов и было продолжено устранение случаев некорректной работы системы.

Был разработан инструмент для подсчета узлочасов, количества запусков проектов как на всей вычислительной системе, так и по её отдельным разделам с возможностью предварительного отбора интересующих проектов с помощью фильтров. Добавлены инструменты для мониторинга активности экспертов, проверяющих ежегодные отчеты пользователей, и инструменты, позволяющие сравнивать похожие проекты и пользователей, что облегчает процедуру ежегодной отчётности.

Для повышения отчуждаемости Octoshell были выделены настройки, индивидуальные для каждого суперкомпьютерного центра: политика конфиденциальности, используемый SMTP-сервер, эксплуатирующая суперкомпьютерный центр организация, адреса серверов, ответственных за информирование пользователей через телеграмм и др. Эти параметры предлагается редактировать в отдельном файле или на специальной странице в веб-интерфейсе в зависимости от необходимости наличия определенных параметров на стадии запуска приложения.

Для повышения удобства процесса учёта потраченных ресурсов удалённые руководителем проекта участники теперь хранятся не только в логах работы пользователей системы, но и в отдельном отношении, что ускоряет получение результатов запросов в СУБД. Самое главное, что теперь аккаунты таких удалённых участников участвуют в синхронизации для того, чтобы заблокировать их доступ на вычислительные системы.

Остальные исправленные в 2024 году ошибки почти не ухудшали опыт использования пользователей и администраторов системой Octoshell. Среди них отметим следующие. В ознакомительном туре раздела “Эффективность вычислительных задач” не отображались демонстрационные задачи. При синхронизации доступа пользователей на вычислительные системы Octoshell иногда неправильно интерпретировал состояние доступа пользователя из-за появления символа перевода строки в конце состояния, что на практике влияло только на выполнение лишних шагов при синхронизации.

В 2024 году был осуществлен перенос системы мониторинга DiMMon на два больших суперкомпьютера, входящих в списки наиболее мощных систем России. В ходе данного переноса была проверена применимость DiMMon на другой целевой аппаратуре. Было определено, что система мониторинга в целом легко переносится на другие системы. Исключение составляет модуль сбора показателей аппаратных датчиков процессора, в силу существенного различия в списках доступных процессорных датчиков и их именовании. Были изучены и реализованы подходы для модификации данного модуля с учетом особенностей новой целевой аппаратуры, что потребо-

вало изучения документации по строению процессоров различной микроархитектуры, небольшой модификации исходного кода модуля и проверки корректности работы измененного модуля на практике. Также был выполнен ряд модификаций, позволяющих упростить и автоматизировать процесс сборки системы мониторинга. В первую очередь это касается автоматизации работы с внешними библиотеками: теперь они поставляются вместе с самим DiMMon, и процесс их компиляции и сборки почти полностью автоматизирован. На данный момент сборка DiMMon на новой системе осуществляется одной командой, для запуска нужно выполнить две команды.

Помимо этого, было выполнено расширение функционала системы мониторинга DiMMon за счет реализации нового модуля, позволяющего собирать информацию о работе сетевой файловой системы NFS. Данный модуль собирает данные о частоте выполнения операций чтения и записи в файлы, а также о частоте открытия и закрытия файлов. Этот модуль применялся на одном из двух суперкомпьютеров, куда был выполнен перенос общего решения, поскольку на нем не использовалась файловая система Lustre (применяемая на суперкомпьютере Ломоносов-2), а вся работа с файлами в пользовательских приложениях выполняется посредством NFS. Стоит отметить, что данный модуль позволил собирать данные, которые раньше не были доступны (о частоте выполнения операций чтения и записи), что позволило реализовать новые правила для определения проблем с производительностью при работе с файловой систмеой. Полученный модуль был интегрирован в DiMMon и может применяться в дальнейшем в составе общего решения.

Также было проведено исследование причин периодических падений системы мониторинга DiMMon на суперкомпьютере Ломоносов-2. Было определено, в каком месте программы происходит падение, а также выявлены ситуации, приводящие к падениям (они происходят во время выполнения скриптов пролога SLURM). На данный момент выполняется анализ первопричин и поиск методов устранения этих падений.

Начаты работы по реорганизации раздела статистики в системе поддержки функционирования суперкомпьютерного центра. Проведена систематизация имеющихся источников, определены требования к формированию срезов на основе временных меток и этапов в жизненном цикле прикладных проектов.

Основными объектами для анализа являются: - Проекты - основная единица для выделения ресурсов на прикладные исследования.

- Пользователи - непосредственные исследователи, выполняющие прикладные проекты. Могут имеет разные роли (руководитель проекта, эксперт, пользователь, администратор и т.д.)

- Учетные записи - представление пользователя на вычислительной системе в рамках конкретного проекта.

- Организации - у любого проекта должна быть организация-поручитель или заказчик работ.

- Обращения в поддержку - индикатор числа и темы актуальных проблем.

- Отчеты и опросы - единица контроля целевого использования ресурсов.

- Задача - отдельный запуск приложения на вычислительной системе, в совокупности позволяют определить и объем, и хронологию, и эффективного использования ресурсов.

Разработаны функциональные требования к разделу, в т.ч. к экспорту

данных, визуальным средствам анализа, включая сравнительный, хронологический и другие. Полученный задел позволит на следующем этапе реализовать качественно новый функционал как для администраторов, так и для рядовых пользователей.

Задача загрузки больших объёмов пользовательских данных решается в рамках разработки компоненты PerfData проекта Algo500. Основой PerfData является репозиторий данных, в котором хранятся конфигурационные файлы для конкретной вычислительной системы, сведения о выбранной для запуска программы совокупности вычислительных узлов суперкомпьютера, а также данные о входных аргументах, параметрах и результатах прогона программы. Кроме того, необходимо хранить программные реализации алгоритма, программы, генерирующие различные начальные данные, скрипты, предназначенные для компиляции программ и сборки исполняемых файлов, необходимые сторонние библиотеки программ и т. д. Объем таких программных компонент может быть весьма существенным. Возникает потребность в приложении, позволяющем передавать на сервер данные больших объёмов, а также создавать и редактировать необходимую структуру директорий.

В 2024 году реализован микросервис для передачи больших файлов, в основе которого лежит библиотека JavaScript Resumable.js. Микросервис позволяет загружать на сервер файлы размером нескольких десятков гигабайт. При разрыве соединения загрузка файла продолжается с того же места, где была остановлена, благодаря возможности возобновляемой загрузки. Протестирована передача файлов большого объёма. Корректность передачи данных подтверждается сравнением контрольных сумм, что свидетельствует о стабильности соединения и успешной передаче файлов без ошибок. Проведены оценки времени передачи больших файлов на сервер, составлено описание необходимых функций пользователя, являющихся основой сервиса создания и редактирования директорий.

Структура платформы Algo500, основанная на подсистемах хранения данных, изначально состояла из трех взаимосвязанных компонент: AlgoWiki, CompZoo и PerfData, позднее была также выделена компонента RatingLists. В рамках данной работы реализуется решение, позволяющее получать данные из базы данных компоненты CompZoo проекта Algo500, представляющее описания архитектур суперкомпьютеров, и затем отображать эти данные в удобном виде, предоставляющем пользователям возможности по изучению и анализу.

Исследованы возможные подходы для отображения данных об архитектуре суперкомпьютеров компоненты CompZoo. Выбран подход, наиболее удовлетворяющий необходимым критериям. Спроектирована объектно-ориентированная структура решения по визуализации данных, выстроена соответствующая диаграмма классов. Произведено встраивание разработанной структуры решения в проект Algo500. Разработано решение по визуализации данных описаний архитектур суперкомпьютеров компоненты CompZoo на основе применения вспомогательных javascript-библиотек – полностью в виде дерева, ограниченно – в виде таблицы. Разработано решение по визуализации данных описаний архитектур суперкомпьютеров компоненты CompZoo с использованием чистого JavaScript в табличном и графовом видах – в полном объеме.

ЗАКЛЮЧЕНИЕ

В ходе данного этапа были выполнены следующие работы:

1. Выполнена апробация системы для оценки качества использования суперкомпьютерных ресурсов на реальных данных.

2. Начаты работы по облегчению возможности использования системы поддержки суперкомпьютерного центра Octoshell в других суперкомпьютерных центрах, добавлено несколько новых инструментов и было продолжено устранение случаев некорректной работы системы.

3. Выполнено расширение функционала и повышение работоспособности системы мониторинга DiMMon.

4. Начаты работы по реорганизации раздела статистики в системе поддержки функционирования суперкомпьютерного центра.

5. Выполнена разработка сервиса загрузки больших файлов и технологий визуализации архитектур суперкомпьютеров.

Запланированные работы выполнены в полном объеме.

ПРИЛОЖЕНИЕ А
Объем финансирования темы в 2024 году
Таблица А.1

Источник финансирования	Объем (руб.)	
	Получено	Освоено собственными силами