Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.051

Старостин
Илья Дмитриевич

Методы внедрения эффективного кода для сбора информации о производительности распределенных систем

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Неборский С.Н.
к.т.н.

Минск 2017

КРАТКОЕ ВВЕДЕНИЕ

Программное обеспечение является неотъемлемой частью современного мира. Оно развивается и совершенствуется из года в год, создаются новые технологи и паттерны разработки ПО. Усложняется всё тем, что требования к производительности ужесточаются, а нагрузка увеличивается. Особенно хорошо такая динамика прослеживается на сфере веб-технологий. Связано это с двумя основными факторами: количество пользователей сети интернет увеличивается из года в год, средний возраст пользователей растёт из года в год.

Не смотря на постоянно совершенствующееся аппаратное обеспечение рабочих станций и серверов, часто обработки только одной задачи в одно и тоже время может быть недостаточно. По этой причине приложения создаются как распределённые. В простейшем случае это может быть распараллеливание обработки на несколько логических потоков. В более сложных случаях приложение создаётся с возможностью выполнения на нескольких физических серверах, связанных между собой сетью. И именно с такими приложениями возникает проблема сбора данных о производительности и стабильности программного комплекса, о производительности различных компонентов системы и др., так-как нету чётко структурированной последовательности потоков данных.

Основой данной работы послужил реальный случай из практики, когда после интеграции сторонней системы сбора данных о производительности время отклика основного программного комплекса значительно увеличилось. Целью данного исследования является анализ существующих средств сбора данных о производительности системы, частичный реверс-инженеринг протокола сетевого взаимодействия, а также проектирование метода внедрения кода для эффективного сбора данных о производительности данных систем.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Цель магистерской диссертации – провести исследование существующих методов сбора данных о производительности распределённой системы, и предложить улучшения.

Для достижения поставленной цели необходимо решить следующие *задачи*:

1. Провести исследование существующих средств сбора данных о производительности распределённых приложений.
2. Выделить основные методы, используемые в существующих ССД, а так же сетевые протоколы и форматы данных, используемые для передачи данных о производительности.
3. Выделить критерии сравнения методов, протоколов, форматов данных, и провести сравнение.
4. Предложить улучшения и оптимизации к существующим решениям.

Объектом магистерской диссертации являются системы и методы сбора данных о производительности распределённых систем.

Предметом является оптимизация существующего метода для уменьшения влияния ССД на производительность анализируемого ПО. Основная *гипотеза*, положенная в основу: существующие методы сбора данных о производительности не являются оптимальными и могут сильно влиять на производительность анализируемого ПО.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «Методы внедрения эффективного кода для сбора информации о производительности распределенных систем» (ГБ № 16-2004, № ГР 20163588, научный руководитель НИР – Н. В. Лапицкая).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя С. Н. Неборского, заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

По теме диссертации опубликована 1 печатная работа в сборниках трудов и материалов международных конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, библиографического списка и одного приложения. В главе 1 приводится исследование существующих программных комплексов для сбора и анализа данных о производительности. Глава 2 посвящена исследованию методов сбора данных о производительности. В главе 3 исследуется механизм организации передачи данных о производительности по сети. В главе 4 подводятся итоги, выделяются критерии сравнения и проводится сравнительный анализ методов.

Общий объем работы составляет 51 страница, из которых основного текста – 51 страница, 18 рисунков на 16 страницах, 14 таблиц на 12 страницах, список использованных источников из 31 наименований на 2 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** произведён обзор и базовый анализ существующих коммерческих решений по сбору информации о производительности системы. Было выделено 5 методов сбора данных о загруженности аппаратного обеспечения и о производительности анализируемого приложения, такие как:

1. Подключаемая библиотека.
2. Расширение для платформы.
3. Приложение мониторинга.
4. Сбор данных об ОС через системные вызовы.
5. Сбор данных о запущенных виртуальных машинах.

Так же были выделены протоколы передачи (http, https) и форматы данных (xml, json), используемые для передачи собранной информации на центральный сервер для обработки.

Вторая глава посвящена анализу методов методы, выделенных в первой главе. Были определены основные достоинства и недостатки каждого метода. Было определено, что некоторые методы действительно могут сильно влиять на производительность всей системы. Это может завесить от используемой связки технологии и метода, так и просто от самого метода.

Для оптимизации был предложен новый, гибридный метод, который является квинтэссенцией идей методов «подключаемой библиотеки» и «сбора данных через системные вызовы» с определёнными улучшениями и доработками. Данный метод подразумевает использование того-же способа подключения и работы с анализируемым ПО как и в метода «подключаемой библиотеки», однако для непосредственно сбора данных о загруженности аппаратного обеспечения использовать драйвера или низкоуровневые системные вызовы (в зависимости от технологии и ОС).

В **третьей главе** происходит анализ сетевого взаимодействия средств сбора данных с центральными серверами, а именно анализ протоколов и форматов данных. Было установлено, что используемые протоколы являются не оптимальными при использовании в средствах сбора данных, так как требуют отправки большого количества технических данных, что загружает сетевой интерфейс и понижает эффективность средств сбора данных. Было предложено 2 протокола, призванных оптимизировать процесс передачи данных, такие как UDP и http2.

Используемые форматы данных так же несут большое количество мета информации. В качестве решения описано применение специальной версии формата json с компрессией данных, не используемые в современных средствах сбора.

Четвёртая глава посвящена сравнительному анализу рассмотренных и предложенных методов, протоколов и форматов. Первая часть главы является теоретическим анализом. Проведя сравнительный анализ на основе предложенных критериев, было подтверждено что предложенные методы\протоколы\формат данных являются более эффективными чем используемые в современных средствах сбора данных. Предложенный гибридный метод сбора данных хоть и не является полностью универсальным, однако должен уменьшить потребление процессорного времени средством сбора данных. Так же был рассчитан объём траффика, который можно будет сэкономить на сетевом интерфейсе анализируемого ПО – он составляет примерно 15 гигабайт в год.

Во второй части главы описаны результаты практического эксперимента, в котором проводились замеры падения производительности при использовании:

1. самого распространённого метода сбора данных с распространёнными протоколом и форматом передачи данных.

2. предложенного гибридного метода, с протоколом передачи данных UDP и кодированием данных в модифицированной форме формата JSON.

Эксперимент показал, что в использование модуля сбора данных реализованного в соответствии с пунктом 2 увеличивает производительность анализируемого программного средства на 9%.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведено исследование существующих решения для сбора данных о производительности распределённых систем. Выделены основные методы интеграции модулей сбора данных с анализируемым ПО, а так же их сетевой стек – протоколы передачи и форматы данных.

2. Рассмотрены существующие методы сбора данных о производительности. Проведён сравнительный анализ методов.

3. Предложен эффективный «гибридный» метод сбора данных о производительности.

4. Рассмотрены протоколы передачи данных, используемые в существующих средствах сбора данных. Произведён сравнительный анализ протоколов передачи данных в контексте ССД.

5. Предложены 2 протокола для повышения эффективности сбора и отправки данных о производительности.

6. Рассмотрены форматы передачи данных, используемые в существующих средствах сбора данных. Произведён сравнительный анализ форматов данных в контексте ССД.

7. Предложен формат передачи данных снижающий нагрузку на сетевой интерфейс.

8. Реализован тестовый модуль сбора данных, основанный на предложенном методе, протоколе и формате данных.

9. Произведено тестирование влияния внедряемых модулей сбора данных на производительность анализируемого приложения. В протестированном случае удалось добиться увеличения производительности анализируемого программного средства на ~9%, а так же снизить накладные расходы на сеть (~5.68гб в год с одного экземпляра приложения).

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для реализации и внедрения эффективного кода сбора данных о производительности. Они могут быть использованы при создании новых ССД, а также для внедрения в существующие системы сбора.

2. Предложенные методы могут служить основой для долгосрочного процесса стандартизации методов сбора данных о производительности, а также быстрого внедрения для решения исследовательских и коммерческих задач.

3. Результаты могут использоваться для реализации эффективных модулей сбора данных каких-либо аналитических систем, получения поведенческих метрик и отслеживания событий, происходящих в системе.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Старостин, И. Д. Оптимизация передачи данных о производительности для распределённых систем / И. Д. Старостин // Наука и образование в XXI веке. Электронный научный журнал по материалам междунар. науч.–практ. конф., Москва, 30 ноября 2016 г. – Москва, 2016. – с. 104–107. – ISSN 2414-5041