УДК 004.7

ОПТИМИЗАЦИЯ РАБОТЫ ВЫСОКОНАГРУЖЕННЫХ СЕРВИСОВ С ПОМОЩЬЮ ВНЕДРЕНИЯ ТЕХНОЛОГИИ GRPC

Конончук С.А. (ИТМО), **Самохин Н.Ю.** (ИТМО)
Научный руководитель – преподаватель факультета прикладной информатики Самохин Н.Ю.
(ИТМО)

Введение. В современном мире клиент-серверные приложения получают все большее распространение с каждым годом, что влечет за собой постоянное увеличение нагрузок на серверные системы. В связи с этим перед организациями часто встает проблема масштабирования существующих систем к интенсивно изменяющимся условиям эксплуатации. Справиться с возросшим числом пользователей часто помогает масштабирование, однако этот подход требует высоких финансовых расходов и не позволяет добиться достаточной утилизации ресурсов[1]. Поэтому актуальной задачей остается оптимизация производительности существующей системы и коммуникации внутри нее, где одним из ключевых факторов является сетевое взаимодействие.

Целью работы является оптимизация работы существующей системы высоконагруженных сервисов с помощью внедрения технологии gRPC в качестве протокола межсервисного взаимодействия, а также последующая оценка влияния изменений на производительность системы.

Основная часть. Существенная часть современных сервисов построены на RESTFul архитектуре, которая базируется на прикладном протоколе HTTP/1. Это продиктовано общепринятой практикой и низким порогом входа во внедрение и эксплуатацию этой технологии, однако такой способ становится одним из уязвимых мест системы при значительном увеличении запросов, что затрудняет его применимость в highload-системах. Технология gRPC[2], которая призвана стать альтернативой традиционному подходу, является более современным и эффективным способом межсервисного взаимодействия, повышающим стабильность работы распределенных систем без привлечения существенных вычислительных ресурсов.

В ходе работы была проанализирована микросервисная архитектура highload-системы, а также проведен анализ текущих показателей ее работы и выявлены проблемы в эксплуатации. Согласно полученным измерениям, был сделан вывод о недостаточной стабильности системы, а также о несоблюдении нефункциональных требований к ее работе.

Одна из ключевых проблем системы – это часто возникающие пики времени ответа сервисов при равномерно распределенной нагрузке, которые приводят к нарушению установленного SLO по времени отклика для потребителей. Такое поведение свидетельствует о нестабильности сетевого взаимодействия системы в целом и требует ее оптимизации.

Внедрение технологии gRPC, использующей HTTP/2 вместо текущей HTTP/1, направлено на стабилизацию времени ответа как внутренних сервисов системы, так и сервисов, выходящих на внешнего потребителя[3]. В работе описан процесс внедрения gRPC в систему высоконагруженных сервисов, а также рассмотрены проблемы, возникшие при внедрении, конфигурировании и дальнейшей эксплуатации системы. Также была проведена оценка влияния, оказанного на производительность и стабильность работы системы, внедрением нового сетевого стека.

Выводы. Внедрение gRPC в качестве протокола межсервисного взаимодействия позволяет существенно стабилизировать и увеличить производительность системы, а также ее отдельных компонентов. Предложенный фреймворк обеспечивает эффективное масштабирование системы без привлечения значимых вычислительных ресурсов.

Список использованных источников:

1. Высоконагруженные приложения: программирование, масштабирование, поддержка / Мартин Клеппман; [перевели с англ. И. Пальти, А. Тумаркин]. – Санкт-Петербург: Питер, 2022. – 637 с.: ил. – (Бестселлеры O'Reilly). – ISBN 978-5-4461-0512-0.

2. Официальный сайт фремворка gRPC [Электронный ресурс] – URL: https://grpc.io/docs/what-is-grpc (дата обращения: 12.02.2025).

3. HTTP/2 vs. HTTP/1.1: How do they affect web performance? – [Электронный ресурс] – URL: https://www.cloudflare.com/learning/performance/http2-vs-http1.1/ (дата обращения: 12.02.2025).