**Пользовательские Внутренние Показатели**

Механизм в библиотеке gRPC, который позволяет пользователям внедрять пользовательские метрики на сервере gRPC и использовать их на клиентах gRPC для создания собственных алгоритмов балансировки нагрузки.

Обзор

Простые решения по балансировке нагрузки могут быть приняты с учётом локальных или глобальных данных о нагрузке на серверную часть, например, о загрузке процессора. Более сложные решения по балансировке нагрузки могут быть приняты с учётом данных о конкретном приложении, например, о глубине очереди, или путём объединения нескольких показателей.

Функция пользовательских показателей производительности предоставляет API, которые позволяют пользователям внедрять обратную связь по показателям в свои политики LB.

Варианты использования

Эта функция предназначена в основном для расширенных сценариев использования, в которых пользовательская политика балансировки нагрузки применяется для более интеллектуальной маршрутизации трафика на список внутренних серверов для повышения производительности маршрутизации, например, взвешенная циклическая политика балансировки нагрузки.

gRPC традиционно позволяет пользователям подключать собственные политики балансировки нагрузки, см. руководство. Для пользователей xDS можно настроить пользовательский балансировщик нагрузки для выбора пользовательской политики LB.

Отчетность по метрикам

Open Request Cost Aggregation (ORCA) — это открытый стандарт для передачи информации о внутренних показателях. gRPC использует стандарты сервисов и показателей ORCA и поддерживает два механизма отчётности о показателях:

Отчёт о показателях для каждого запроса: серверная часть присоединяет пользовательские показатели к завершающим метаданным, когда соответствующий RPC-запрос завершается. Обычно это полезно для коротких RPC-запросов, таких как унарные вызовы.

Внеполосная отправка метрик: серверная часть периодически отправляет клиенту данные о метриках, например об использовании процессора и памяти. Это полезно во всех ситуациях: при однократных вызовах, длительных RPC-вызовах в потоковых вызовах или при отсутствии RPC-вызовов. Однако при внеполосной отправке метрик не отправляются данные о стоимости запросов. Частота отправки метрик настраивается пользователем и указывается в пользовательской политике балансировки нагрузки.

На схеме показана архитектура, в которой пользователь создаёт собственную политику балансировки нагрузки, реализующую обратную связь по внутренним показателям.

Реализация

Для получения более подробной информации, пожалуйста, ознакомьтесь с предложением gRPC A51.

Языковая Поддержка

ЯзыкПримерJavaПример JavaВпередПерейти к примеруC++Пример предстоящего

Последнее изменение 29 февраля 2024 г.: используйте абсолютные пути вместо абсолютных URL (#1268) (4f733b4)