

Проверено ≠ надежно

Андрей Зарубин





Пока вы не проверили сами,
считайте, что это сломано.

Автор неизвестен

Тестирование надёжности систем



«Одной из основных обязанностей SRE инженеров являются измерение и оценка надёжности обслуживаемых ими систем.

Для этого SRE инженеры применяют классические методы тестирования программ, адаптируя их для своих систем в требуемом масштабе.»

Андрей Зарубин



22 года в IT

Банковское ПО, высоконагруженные
трейдинговые платформы,
геоинформационные системы

Руководжу разработкой с 2009-го,
кросс-функциональными командами с 2017-го

Внедрял Agile, DevOps и другие практики

Сертифицирован по Java, Cloud,
Agile и PRINCE2

План доклада



01 QA & SRE

02 SLI, SLO & SLA

03 Отказоустойчивость

04 Наблюдаемость

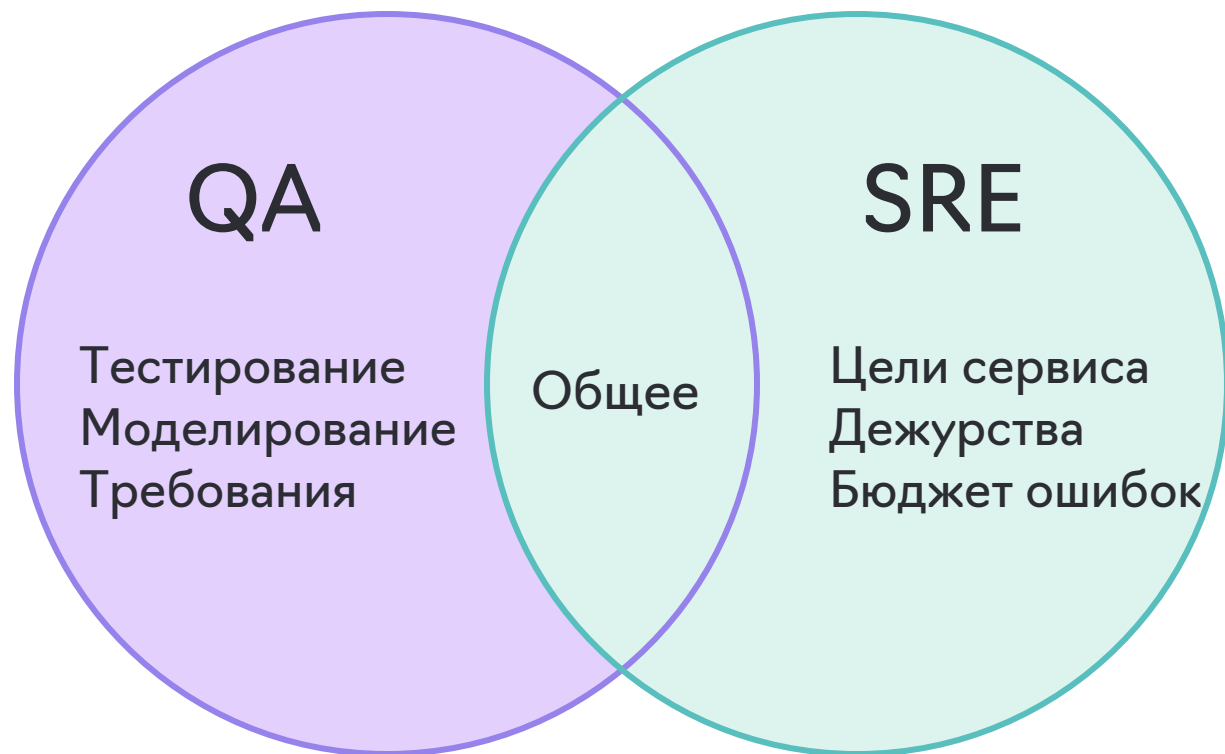
05 Бюджет ошибок

06 QA & Аварии

07 Toil

08 Навыки SRE

QA & SRE



Общая область

Диагностика
Производительность
Эксперименты
Автоматизация
Анализ рисков
Проверка изменений
Последствия отказов

Границы условны



SLI, SLO & SLA

Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Кейс. Описание



Сервис: Небольшой интернет-магазин

Сценарий: Оформить готовый заказ и узнать результат оплаты



За скобками: подтверждение оплаты,
склад, доставка, возвраты и т. п.

Пример внешнего провайдера · Stripe API

Кейс. Требования



≤ 2 с

$\geq 99,9\%$
хороших операций

за 30 дней

Хорошая операция

Подтверждение оплаты
или окончательный бизнес-отказ

Результат соответствует
ответу провайдера

Плохая операция

Технический отказ, нет результата,
неверный ответ
или ответ позднее 2 с

SLI, SLO & SLA



SLI

Показатель уровня сервиса

Доля операций с корректным результатом за ≤ 2 с

SLO

Цель для показателя

$\geq 99,9\%$ хороших операций за 30 дней

SLA

Соглашение об уровне сервиса

Обязательства и последствия их нарушения



SLI, SLO & SLA

Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Механизмы защиты



01 Таймаут

Предел времени одного запроса.

Не отменяет работу на чужой стороне.

Статус операции может остаться неизвестным.

02 Retry

Повтор запроса. Может помочь, если сбой временный. При перегрузке другого сервиса может усугубить проблему.

03 Идемпотентность

Повтор запроса не дублирует эффект.

Например, используем стабильный ключ операции.

Тестирование защитного поведения



- | | | |
|----|---------------------------|--------------------------------|
| 01 | Таймаут | Неопределённый статус операции |
| 02 | Retry | Эффект не дублируется |
| 03 | Отказ вызываемого сервиса | Retry остаются контролируемыми |

После сбоя система возвращается к нормальной работе



SLI, SLO & SLA

Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Четыре золотых сигнала



Latency

Задержка

Сколько ждёт пользователь

Traffic

Объём работы

Новые операции и вызовы,
которые они порождают

Errors

Неуспешные результаты,
включая неправильные ответы

Saturation

Загрузка ресурсов
CPU/RAM, очереди, пулы соединений

Инструменты наблюдаемости



Метрики

Логи

Трейсы

% плохих операций

Статусы заказов

Путь обработки
и длительность
каждого вызова

Без связи между инструментами данные могут противоречить друг другу

Тестирование наблюдаемости



Контролируемый
сбой

Мониторинг

Алёрт



SLI, SLO & SLA

Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Бюджет ошибок



SLO 99,9%

Допустимо 0,1% плохих операций

Ровно 1 000 000 операций за последние 30 дней

$$1\,000\,000 \times 0,1\% = 1\,000$$

плохих операций
в бюджете



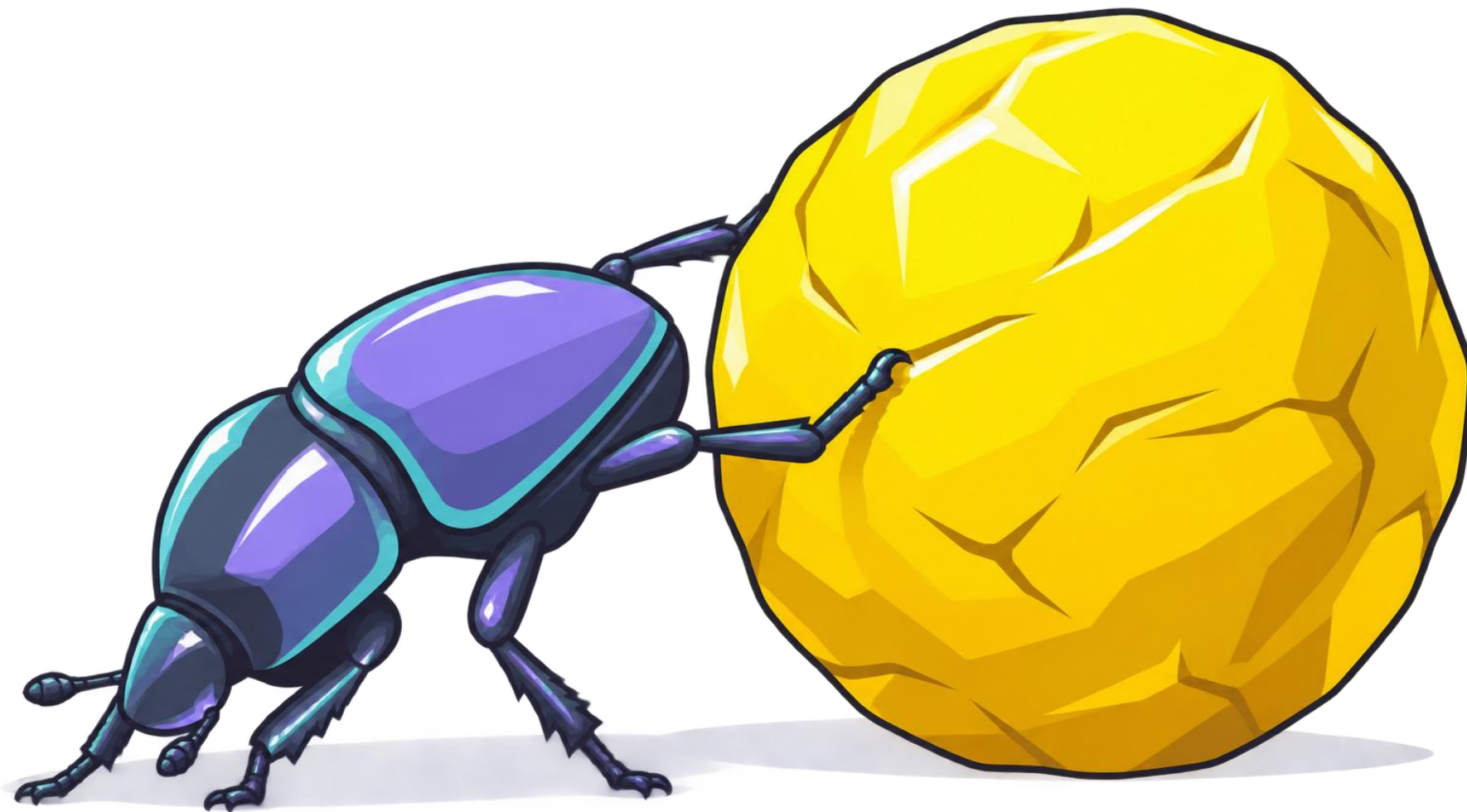
600 · 60% потрачено

400 осталось

400 — в скользящем окне, не до конца месяца.

Этот бюджет нельзя перевести в минуты недоступности.

QA принимают участие в оценке
риска следующего релиза





SLI, SLO & SLA

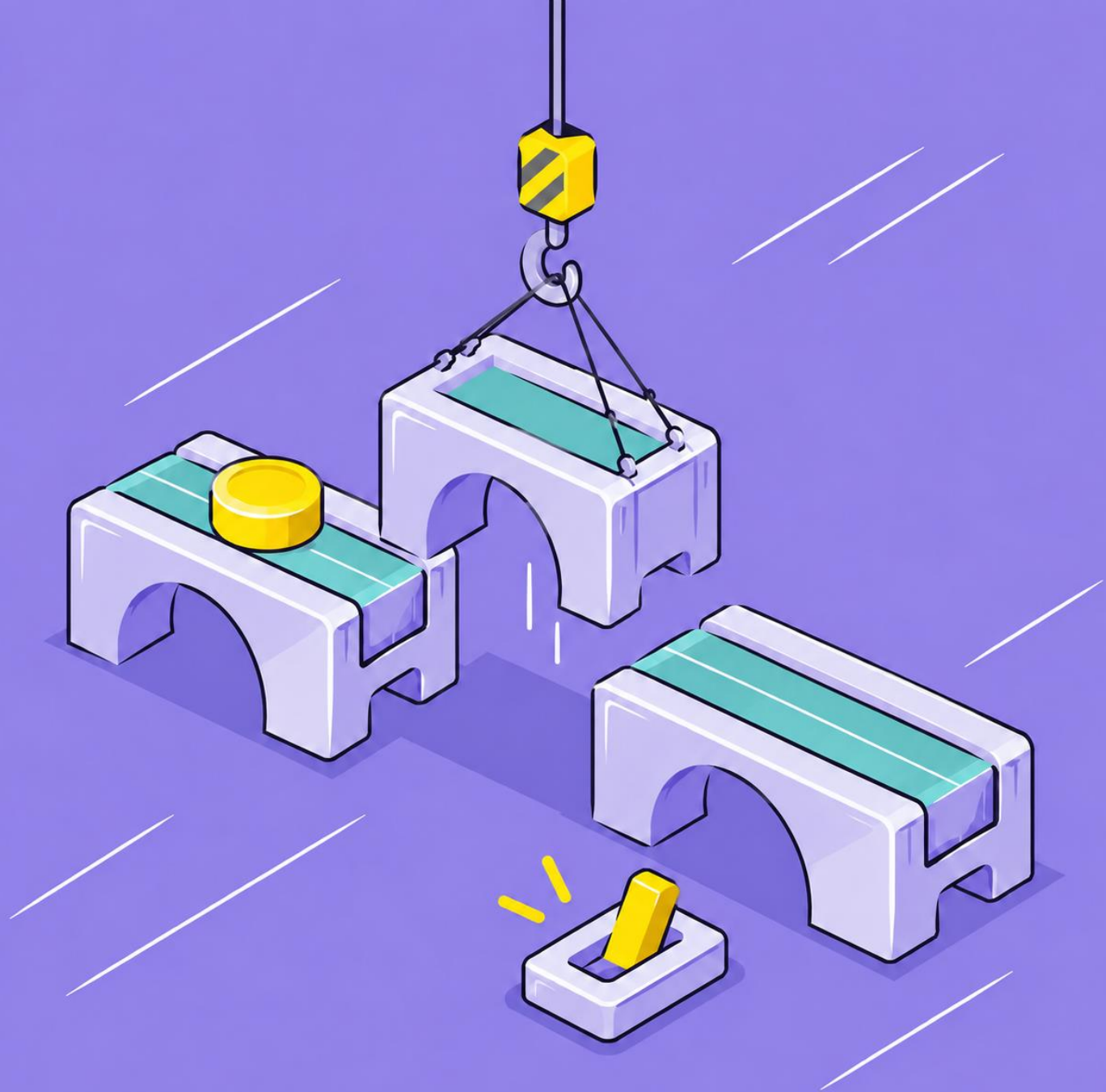
Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Восстановление нужно репетировать



Feature flag

Управляемый переключатель поведения

Rollback

Возврат предыдущей версии или конфигурации

Оба действия проверяем заранее:

- 01 Настройка изменилась на всех экземплярах?
- 02 Сохранённые данные совместимы?
- 03 Что произошло с данными в обработке?

QA готовит сценарий тестовой оплаты и инструкцию

QA во время аварии



01	Проверка	Гипотеза и пользовательский сценарий
02	До исправления	ID проблемных кейсов и временная линия
03	Последствия	Что затронула авария
04	Разбор	Разбор полётов и подготовка к повторному релизу



SLI, SLO & SLA

Отказоустойчивость

Наблюдаемость

Бюджет ошибок

Аварии

Toil



Что считать toil



Повторяющаяся операционная работа



Часто ручная и реактивная

Её можно автоматизировать

Объём растёт вместе с обслуживаемой системой

Не вся ручная или скучная работа — toil

Навыки SRE для QA



Требование

—————→ SLI + SLO

SLI

—————→ Наблюдаемость

Отказоустойчивость

—————→ Восстановление

Считать toil

Спасибо!



Полезные ссылки



SLO и Error Budget

Google SRE Book

[Service Level Objectives](#)

Google SRE Workbook

[Implementing SLOs](#)

Google SRE Workbook

[Example Error Budget Policy](#)

Отказоустойчивость

Google SRE Book

[Testing for Reliability](#)

AWS Builders Library

[Timeouts, retries, backoff with jitter](#)

API

Stripe API

[API reference](#)

Stripe API

[Idempotent requests](#)

Stripe

[Advanced error handling](#)

Toil

Google SRE Book

[Eliminating Toil](#)

Google SRE Workbook

[Eliminating Toil](#)

Наблюдаемость

Google SRE Book

[Monitoring Distributed Systems](#)

OpenTelemetry

[Signals](#)

Google SRE Workbook

[Alerting on SLOs](#)

Аварии

Google SRE Book

[Emergency Response](#)

Google SRE Workbook

[Incident Response](#)