Формула подсчёта SLA сервиса

Обозначения и параметры:

- U - **Средний процент доступности** (availability), измеряемый как доля времени, когда сервис отвечает на проверки работоспособности (пробы, *probe_ up*):

$$U = \text{среднее значение } (\text{probe\_up}) \times 100\%$$

  Здесь *probe_ up* - бинарный индикатор, который равен 1, если сервис доступен и отвечает на проверку, и 0, если нет. Среднее берётся за последний период (например, 30 минут).

- S - **Процент успешных запросов** (success ratio):

$$S = \left(1 - \frac{\sum \text{ошибочные запросы}}{\sum \text{всех запросов}}\right) \times 100\%$$

  Этот показатель отражает долю корректно обработанных запросов к сервису за выбранный период (например, последние 5 минут). Ошибочные запросы учитывают любые виды ошибок на уровне обработки.

- avgLatency - **99-процентный перцентиль времени отклика** сервиса, то есть значение времени ответа, которое не превышает 99% всех запросов.

- $\text{SLO}_{\text{latency}}$ - **Целевой порог времени отклика** (Service Level Objective по латентности), установленный для сервиса. Например, 0.5 секунды.

- L - **Показатель соответствия времени отклика SLO** (latency SLO ratio), выраженный в процентах:

$$L = \begin{cases} 100, & \text{если avgLatency} \leq \text{SLO}_{\text{latency}} \\ 100 \times \frac{\text{SLO}_{\text{latency}}}{\text{avgLatency}}, & \text{если avgLatency} > \text{SLO}_{\text{latency}} \end{cases}$$

  То есть, если сервис отвечает быстрее или ровно в целевое время, показатель равен 100%. Если время ответа выше цели, показатель снижается пропорционально.

Итоговая формула SLA:
Общий показатель SLA для сервиса рассчитывается как среднее арифметическое трёх вышеописанных метрик:

$$\boxed{\text{SLA} = \frac{U + S + L}{3}}$$

где

- U - средний процент доступности,
- S - процент успешных запросов,
- L - процент соответствия целевому времени отклика.