Engineer

レイテンシー

⏱ 要求を送ってから、応答が届き始めるまでの待ち時間

どのようなものか

レイテンシーは、システムが応答を始めるまでの時間です。要求がクライアントを出てから、応答の最初のバイトが戻るまでを指します。技術者は、遅延やラグとも呼び、ネットワークの経路を往復して測る場合は、往復時間（RTT）と呼びます。ウェブ性能では、同じ考えを、最初のバイトまでの時間（TTFB）と呼びます。応答時間は、それより広く、応答の最後のバイトまでを含みます。

スループットや帯域幅と混同しやすいものですが、三つは、それぞれ違う問いに答えます。帯域幅は、一秒に何ビットを回線が運べるかです。スループットは、システムが一秒にどれだけの仕事を終えるかです。レイテンシーは、一つの仕事がどれだけ待つかです。夜間のバッチ処理は、スループットが優れていても、レイテンシーは悪いことがあり、チャットサーバーは、その逆になり得ます。

待ち時間の各ミリ秒は、次のいずれかから生まれます。

- **伝搬**は、信号が距離を移動する時間です。光ファイバー中の光は、一ミリ秒に約200 km進むため、1,000 kmの経路は、ソフトウェアが動く前に、往復で少なくとも10 msかかります。
- **伝送**は、バイト列を回線に送り出す時間です。低速な回線で、大きなペイロードを送る場合に重要です。
- **待ち行列**は、ルーターのバッファー、スレッドプール、接続プール、ロックで、ほかの仕事の後ろに並ぶ時間です。負荷とともに増えるのは、この部分です。
- **処理**は、クエリー、テンプレート、暗号化という、実際の仕事です。
- **プロトコルの往復**は、役に立つ最初のバイトより前の、やり取りです。TCPを使う新しいHTTPS接続には、まずTCPハンドシェイクとTLSハンドシェイクが必要で、TLS 1.3は、完全なハンドシェイクを一往復に減らしました。

IETFのIP性能計測グループは、RFC 2681で、ネットワークの場合を明確に定めました。往復遅延は、送ったパケットの最初のビットから、受け取った応答の最後のビットまでです。この定義の二つの決め方は、どのサービスにも当てはまります。応答が来なければ、標本から黙って外すのでなく、未定義、非公式には無限大として数えます。また、記述する統計は、パーセンタイル、中央値、最小値です。最小値は、伝搬と伝送だけを反映し、それを超える部分は、混雑を示すとも述べます。この区別は、ルーター間の経路だけでなく、APIにも当てはまります。

一つのウェブ要求なら、curlで、待ち時間を段階に分けられます。

```
curl -s -o /dev/null "$URL" -w '
dns        %{time_namelookup}
connect    %{time_connect}
tls        %{time_appconnect}
first byte %{time_starttransfer}
total      %{time_total}
'
```

各値は、開始からの秒数です。そのため、値の差が、名前の検索、接続を開く処理、TLSハンドシェイク、そして最初のバイトまでの、サーバー自身の処理にかかった時間です。

概念を判断に生かす

「レイテンシーとは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 障害が小さいうちに知らせる監視とアラートを整え、安定した状態をどれだけ維持できているかを数値で示します。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 社内初のオブザーバビリティダッシュボードを構築 — 社内システムの状態を初めて可視化

社内初のオブザーバビリティダッシュボードを構築 お問い合わせ →

なぜ重要なのか

利用者が感じるのはレイテンシーで、サーバー料金に表れるのはスループットです。ページが遅ければ、どこに費用をかけるかを決めます。先ほどの五つの部分には、それぞれ違う対策が必要です。そのうち四つを見ると、ハードウェアの購入が万能ではないと分かります。距離には、キャッシュ、CDN、利用者に近いリージョンなど、近くへ移す対策を取ります。往復には、接続の再利用、クエリーの一括処理、並列取得など、回数を減らす対策を取ります。処理には、インデッ

クスや、よりよい実行計画で、仕事を減らす対策を取ります。処理能力を増やすと、待ち行列を短くできる場合があります。プロセッサーの高速化、メモリーの増設、入出力の改善も、それらがボトルネックなら処理時間を短縮できます。規模の大きいサーバーが必ず速いとは限りません。ハードウェアを選ぶ前に、何に時間がかかっているかを測ります。

平均値では、問題が実際より小さく見えます。利用者は、分布の遅い側にも出会うためです。十のバックエンドサービスを並列に呼ぶページは、最も遅いものを待ちます。各呼び出しが互いに独立して1%の確率で遅ければ、$1 - 0.99^{10}$、約9.6%のページ読み込みが、少なくとも一つの遅い呼び出しに当たります。部分ではまれな遅延が、全体ではよくあるものになります。

そのため、厳密な目標は、30日間の要求の99%に300 ms以内で応答するなど、パーセンタイルで記述します。通常、見る価値のある数値も、第99パーセンタイルです。中央値は、よい日の使い心地を示し、高いパーセンタイルは、離れようとしている顧客の使い心地を示します。

実際の活用例

ある地図アプリケーションでは、地図レイヤーの描画と位置データの問い合わせの遅延により、利用者にも、バックエンドの信頼性にも悪影響が出ていました。作業は、従来のMicrosoft SQL Serverデータベースを使うバックエンドで、空間クエリーの処理に焦点を当てました。PostGISを備えたPostgreSQLへ移し、平均応答時間を2.5秒から250ミリ秒未満へ短縮しました。空間クエリーの性能は、10倍に改善しました。空間列にGiSTとSP-GiSTのインデックスを設け、5,000万件を超えるレコードを移行し、同時アクセスに向けて設定を調整しました。バックエンドのCPU負荷は65%低下しました。

一つの行へのアクセスが競合する場合は、レイテンシーが、スループットの上限も決めます。最後のチケットを、厳密に一度だけ売るというノートは、一つの商品について、ロック一回で扱える販売数を、ロックの保持時間で割った値が、おおよその上限だと説明します。ロック中の遅い処理は、どれも上限を下げます。保持したまま決済を呼べば、一秒に約二件の販売まで下がります。ここでは、安全対策にも代償があります。PostgreSQLの同期スタンバイは、すべてのコミットに遅延を加え、上限を求める式の分母である、ロックの保持時間そのものを長くします。

時間を使う箇所を調べるため、ある基盤に組み込んだ独自の監視は、APIのレイテンシーを自身のデータベースに記録し、OpenTelemetryで要求を追跡します。そのため、遅い要求を、推測するのでなく、フロントエンド、API、データベー

スを通して追えます。

問題が起きるところ

平均値が、遅い側の分布を隠します。 平均100 msでも、五十件に一件が三秒かかることは、あり得ます。レイテンシーをヒストグラムに記録し、パーセンタイルを読みます。その後は、集計方法にも注意します。五台のサーバーの第99パーセンタイルを平均しても、全体の第99パーセンタイルにはなりません。Prometheusも、その文書で、サマリー（summary）型で計算した分位数は集計できず、ヒストグラムの各バケットの件数なら、先に合計してから分位数を求められると説明します。

計時を始める場所が違います。 ハンドラーの開始から測るサーバーは、ロードバランサーで過ごした時間、接続の受け入れキューでの待ち時間、プールから接続を得るまでの待ち時間を見落とします。各応答を待ってから次の要求を送る負荷生成器は、システムの処理が滞る、まさにそのときに要求を減らすため、その停滞が結果にほとんど表れません。クライアントからも測り、サーバーには、標準ヘッダーで、自分の受け持った時間を示させます。

```
Server-Timing: db;dur=53, app;dur=47.2
```

やり取りの回数が多すぎます。 一覧の各行に一つのクエリーを実行するページは、一行ごとに一往復分の時間がかかります。データベースが別のリージョンにあり、一往復30 msなら、百行だけで、クエリーが実際の仕事をする前に、三秒かかります。結合を使う一つのクエリーか、一括の呼び出しなら、通信距離による待ち時間は、一回分で済みます。

再試行が、負荷を増やします。 一秒で諦め、三回再試行するクライアントは、一台の遅いサーバーへの通信を四倍にして、さらに遅くします。各要求に全体の期限を与え、試行間の待ち時間は、ランダムな揺らぎを加えて指数的に増やし、再試行回数に上限を設けます。

参考資料

- RFC 2681：IPPMの往復遅延指標 — ビットまで定めた往復遅延です。応答の損失と、パーセンタイルの扱いも明示します。
- RFC 7679：IP性能計測の片方向遅延指標 — 一方向の対応する指標で、Internet Standardとして発行されています。

- Server Timing — サーバーが、どの処理に時間を使ったかをブラウザーに伝える、W3Cのヘッダーです。
- ヒストグラムとサマリー — サマリーの分位数は集計できず、ヒストグラムの各バケットの件数は集計できる理由です。
- RFC 8446：TLS 1.3 — ハンドシェイクの設計と、それが減らした一往復です。

用語集

実際の制約に合わせて選ぶ

デプロイの流れ、運用上のリスク、緊急対応の担当者をお知らせください。検査、復旧手段、明確な責任の所在を伴う変更の範囲を定められます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「レイテンシーとは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く 連絡先カード

心を込めて丁寧に制作： D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/latency/