COMMUNAUTÉ ÉCONOMIQUE ET MONÉTAIRE DE L'AFRIQUE CENTRALE (CEMAC)

Institut Sous-régional de Statistique et d'Économie Appliquée (ISSEA)

Organisation Internationale

B.P : 294 Yaoundé

www.issea-cemac.org

Projet de Big Data

ANALYSE DES DONNEES LOG DU SERVEUR DNS PUBLIC DE GOOGLE VIA PYSPARK.

Rédigé par :

BESSALA Junior Serges Edouard

DAH Fongnéma

KALEFACK NGUEPI Sergeo

KEOUL Maab Mara

MAHAMAT NGAGUEDI Eric

MOUSSAVOU MOUSSAVOU Lloyd A.

NDOKO SOUAMOUNOU Geddy S.

TAKOUGOUM Steeve Rodrigue

TAMIBE EZECHIEL Pallaye

Élèves Ingénieurs Statisticiens Économistes, ISE 2

https://github.com/Githubsssssssssssssssssssssss/Big_Data

Sous l'encadrement de :

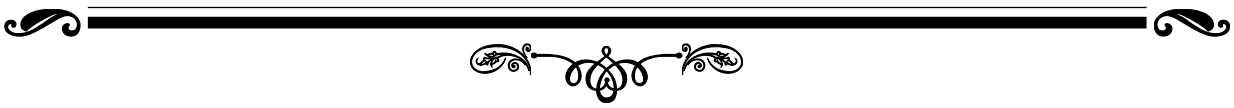
M. Serge NDOUMIN

EnSeignant à l'ISSEA-CEMAC

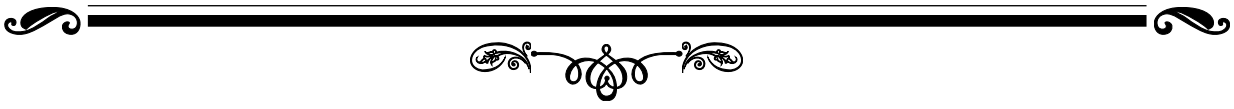
Année académique 2024-2025

SOMMAIRE

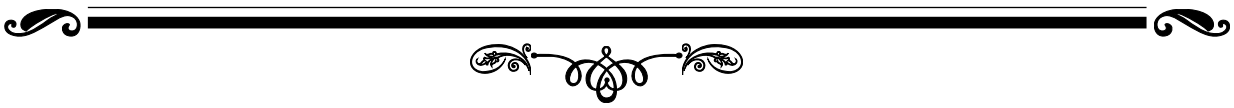
LISTE DES TABLEAUX



LISTE DES FIGURES



INTRODUCTION GENERALE



Contexte et justifcation

À l'ère du Big Data, l'analyse des masses de données constitue un enjeu majeur pour comprendre et optimiser les infrastructures numériques. Le Domain Name System (DNS), pierre angulaire d'Internet, représente un cas d'étude particulièrement pertinent dans ce contexte. Ce service fondamental, qui traduit les noms de domaine en adresses IP, est essentiel à la navigation fluide des utilisateurs sur Internet. Sans cette infrastructure, les utilisateurs seraient contraints de mémoriser des séries de chiffres complexes plutôt que des noms de domaine intuitifs, rendant l'accès aux ressources en ligne considérablement plus difficile.

L'infrastructure DNS mondiale traite quotidiennement des milliards de requêtes, générant un volume considérable de données exploitables. Les serveurs DNS publics de Google (8.8.8.8), largement utilisés à travers le monde, offrent un excellent terrain d'investigation pour analyser les performances d'un service critique. Malgré l'implémentation de techniques d'optimisation comme la mise en cache DNS, qui stocke temporairement les résultats des requêtes fréquentes, les performances des serveurs DNS continuent de montrer des variations significatives qui méritent une analyse approfondie, particulièrement dans un contexte où la dépendance aux services en ligne ne cesse de croître.

Problématique

Dans un environnement où la rapidité et la fiabilité des infrastructures réseau sont essentielles, le serveur DNS public de Google (8.8.8.8) joue un rôle clé dans l'acheminement des requêtes internet. Cependant, des fluctuations de performance,

notamment en termes de latence et de perte de paquets, peuvent affecter l'expérience utilisateur et la stabilité des connexions. Dès lors, quelles sont ces variations de performance et comment les expliquer ? Quels sont les facteurs influençant la stabilité du serveur et les périodes de dégradation des performances ?

Objectifs

L'objectif principal de cette recherche est d'évaluer la performance du serveur DNS de Google, avec une attention particulière portée à la latence. De façon spécifique, nous visons à :

- Mesurer et analyser les fluctuations des temps de réponse du serveur DNS ;
- Évaluer les taux de perte de paquets et leur impact sur la connexion
- Evaluer la stabilité du routage des requêtes envoyées au serveur ;
- Analyser la stabilité et la variabilité des performances du serveur ;
- Identifier les périodes de latence élevée et leurs causes potentielles. ;

Méthodologie et outils

Notre démarche analytique s'articule autour d'une étude exploratoire initiale, utilisant des techniques de statistique descriptive pour caractériser les distributions des temps de réponse et autres métriques pertinentes. Nous approfondirons ensuite notre analyse en exploitant des outils adaptés au traitement de grands volumes de données comme PySpark, permettant une exploration plus fine des tendances et anomalies présentes dans notre jeu de données.

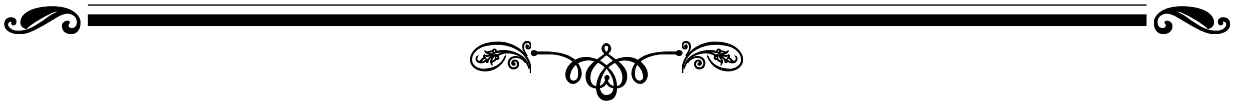
Plan du travail

La première partie de cette étude concerne la présentation des données et l'analyse exploratoire. Dans cette section, nous allons décrire les données utilisées pour ces analyses et nous ferons une breve analyses exploratoire. La deuxième partie de l'étude concerne

une analyse approfondie de la base avec PySpark. Elle inclut des statistiques descriptives, une analyse des variations du Time To Live (TTL) des requêtes DNS et la détection de tendances et d'anomalies, en utilisant des outils analytiques avancés pour obtenir des insights approfondis. La troisième partie est réservée à l'interprétation et à la discussion des résultats obtenus. Enfin, la dernière partie concerne la conclusion.

CHAPITRE :

1. PRESENTATION DES DONNEES ET ANALYSES EXPLORATOIRES



1.1 Présentation des données

1.1.1 Source et nature des données

Les données utilisées dans cette étude proviennent d'une commande `ping` exécutée pour mesurer la latence d'accès au serveur DNS public de Google (`8.8.8.8`). À chaque exécution de la commande, des paquets ICMP (Internet Control Message Protocol) sont envoyés à l'adresse cible à un intervalle d'une seconde. Chaque ligne de réponse représente l'état de la requête envoyée et contient des informations sur la latence ainsi que d'autres paramètres relatifs à la requête.

Informations clés des données

Chaque ligne contient plusieurs paramètres essentiels, qui fournissent des informations sur l'état de la requête ICMP envoyée ainsi que la latence observée. Les lignes peuvent être classées en plusieurs catégories selon l'état de la requête :

- **Requête aboutie** : Représentée par une ligne contenant le temps de réponse en millisecondes (latence), comme par exemple `time=XXX ms`.
- **Requête non aboutie** : Certaines requêtes échouent, et la ligne contient un message indiquant l'échec de l'envoi, comme par exemple `ping: sendto: No route to host`.
- **Requête aboutie sans retour** : Dans certains cas, une requête peut atteindre la destination sans retour de données, résultant en un message de `Request timeout for icmp_seq=XXX`, signifiant que la requête a été envoyée mais aucune réponse n'a été reçue dans le délai imparti.

Structure d'une requête reussie (réponse)

Une ligne représentant une requête réussie contient plusieurs informations structurées, qui permettent de mesurer et d'analyser la latence ainsi que d'autres caractéristiques du réseau. La structure d'une ligne typique de réponse réussie est la suivante :

- **Taille du paquet reçu** : `64 bytes from 8.8.8.8` (indique que le paquet contient **64 octets** et provient de l'adresse IP cible).
- **Numéro de séquence ICMP** : `icmp_seq=XX` (**XX** représente un identifiant unique incrémenté à chaque envoi).
- **TTL (Time-To-Live)** : `ttl=105` (indique le nombre maximal de sauts restants avant que le paquet soit rejeté).
- **Temps de réponse (latence)** : `time=XXX ms` (le temps **aller-retour** entre l'ordinateur source et l'adresse cible, exprimé en millisecondes).

Ces quatre éléments permettent de définir et d'analyser les caractéristiques d'une requête réussie, comme le temps que le paquet met à voyager entre l'ordinateur source et le serveur cible, ainsi que la fiabilité du réseau (indiquée par le TTL).

La date et l'heure de départ de la commande `ping` sont enregistrées dans la première ligne de la capture. Dans ce cas précis, la commande a été exécutée le :

- **Date et heure de départ** : Vendredi 31 janvier à 09:23:12.
- **Adresse cible** : 8.8.8.8 (serveur DNS public de Google).
- **Intervalle de collecte** : Chaque seconde entre chaque requête envoyée.

1.2 Analyses exploratoires

Dans le cadre de l'analyse de la latence du serveur DNS public de Google (IP 8.8.8.8), un total de 80 222 requêtes ping ont été émises afin d'évaluer la stabilité et la performance de la connexion. Une brève exploration a permis d'analyser le nombre de requêtes reçues avec succès, les erreurs rencontrées et les variations des TTL observées.

Sur les 80 222 tentatives, 76 529 requêtes ont bien été envoyées, tandis que 3 693 n'ont pas pu être émises, affichant une erreur "No Route to Host" ce qui signifie une absence de route valide vers la destination. Parmi les requêtes envoyées, 65 722 ont été traitées avec succès et ont reçu une réponse du serveur. En revanche, 10 807 requêtes ont atteint Google sans recevoir de réponse, générant une erreur "Request Time Out".

2. ANALYSES DE LA PERFORMANCE DU RESEAU AVEC PYSPARK

L'analyse de la performance réseau repose sur l'examen de plusieurs indicateurs clés permettant d'évaluer la stabilité et l'efficacité des connexions. Dans ce chapitre, nous exploitons PySpark pour traiter et analyser dix indicateurs de performance, offrant ainsi une vision détaillée des variations et des tendances observées.

2.1 Indicateurs de tendance centrale

L'analyse du temps moyen et de la médiane nous offre une vue d'ensemble sur la distribution des latences.

Tableau 1: Mesures de tendance centrale

Mesure	Valeur
Médiane	133.824 ms
Temps moyen	634.931 ms

Source : Données collectées

Le temps moyen de 634.931 ms dépasse largement la médiane de 133.82 ms, ce qui indique une distribution asymétrique avec une queue droite étendue. Cela suggère la présence de latences extrêmes, c'est-à-dire des délais de réponse particulièrement longs, qui ont un impact significatif sur la moyenne. En revanche, la médiane, étant moins sensible aux valeurs aberrantes, fournit une meilleure indication de la latence centrale typique, représentant de manière plus fiable l'expérience de la majorité des requêtes.

2.2 Indicateurs de dispersion

Tableau 2: Mesures de dispersion

Mesure	Valeur
Premier quartile	131.07 ms
Troisième quartile	140.17875 ms
Temps maximal	927,118.712 ms
Temps minimal	35.597 ms
Ecart Type	15,968.798 ms

Source : Données collectées

Le tableau présente des statistiques descriptives des temps de latence des requêtes ping vers le serveur DNS public de Google. On note :

- Le premier quartile (131,07 ms) montre que 25 % des requêtes sont rapides, sans congestion notable.
- La médiane (133,824 ms) indique une latence stable pour la moitié des requêtes.
- Le troisième quartile (140,17875 ms) suggère que certaines requêtes rencontrent des chemins plus lents ou des congestions.

Le temps minimal (35,6 ms) reflète une faible congestion, tandis que le temps maximal (927 118,712 ms) signale des paquets fortement retardés ou bloqués par des congestions sévères. L'écart type de 15 968,798 ms révèle une grande variabilité, indiquant des latences globalement rapides, mais avec des exceptions notables dues à des congestions temporaires ou des réacheminements.

2.3 Fréquences d'apparition des ttl

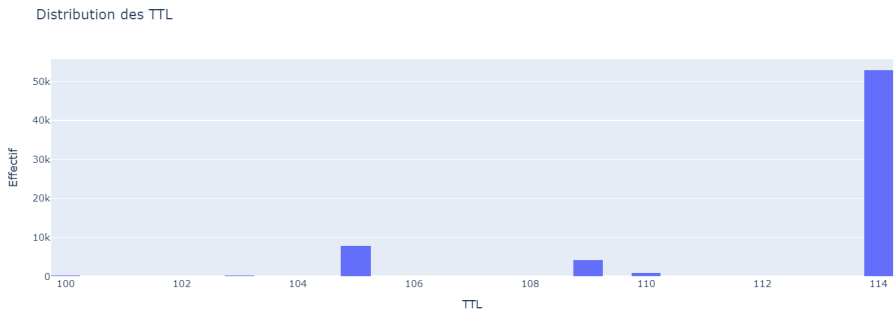
Tableau 3: Nombre d'occurrences pour chaque TTL

TTL	Nombre d'occurrences
105	7 810
100	61
103	190
114	52 883
110	730
109	4 048

Source : Données collectées

Le tableau ci-dessus montre que dans la majeure partie des cas il y a environ 14 routeurs entre le serveur DNS de Google et l'émetteur des requetes. En effet, cette fréquence élevée du nombre d'occurrence du 'TTL=114' peut signifier qu'elle représente une route principale emprunter par les requêtes ou alors que la connexion utilise par l'émetteur est optimisée.

Graphique 1: Repartition du nombre d'occurence par TTL

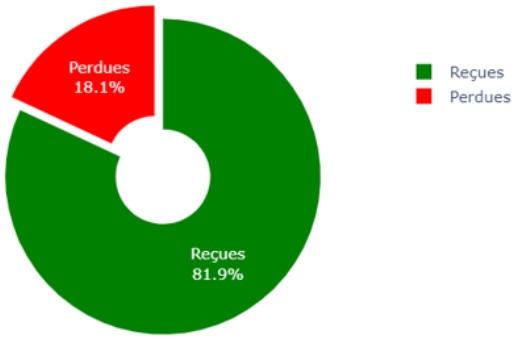


Source : *Auteur a partir de python*

2.4 Taux de perte de paquets

Le taux de perte de paquets est un indicateur essentiel pour évaluer la qualité d'une connexion réseau. Il représente le pourcentage de paquets envoyés qui ne parviennent pas à destination.

Graphique 2: Repartition des requêtes envoyées



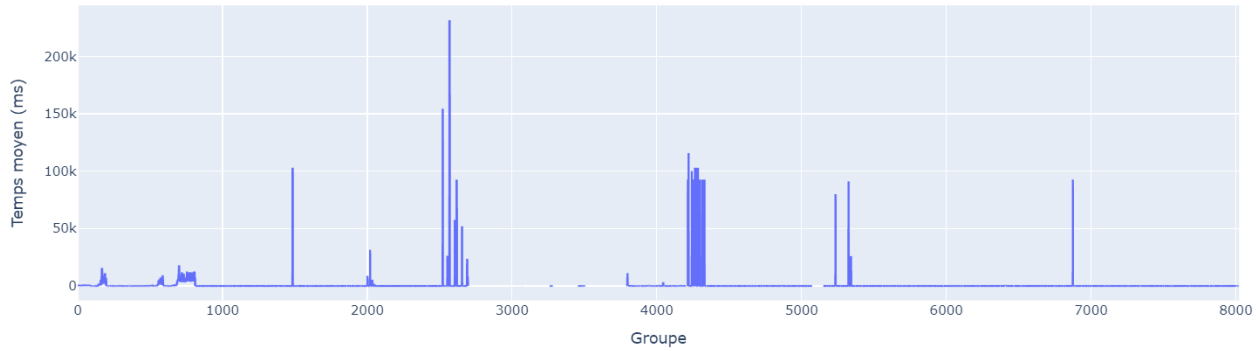
Source : *Auteur a partir de python*

Un taux de perte de 18.1 % sur les requêtes envoyées est préoccupant et suggère l'existence de problèmes notables, potentiellement liés à la congestion du réseau, à des erreurs de configuration ou à des défaillances matérielles. Une investigation approfondie est nécessaire pour identifier précisément l'origine de ces pertes et mettre en place des mesures correctives adaptées.

2.5 Moyenne par groupe de secondes

2.5.1 Par groupe de 10 secondes

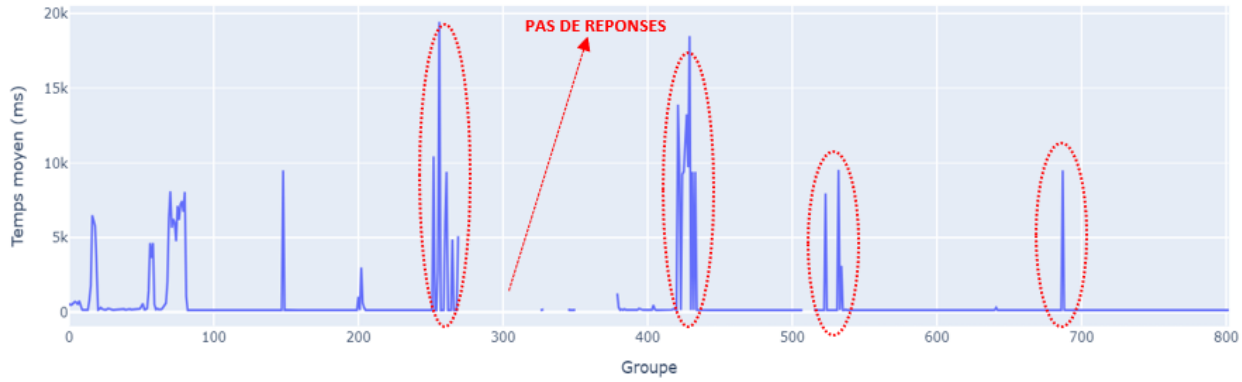
Graphique 3: Temps moyen par groupe de 10 secondes



Source : *Auteur a partir de python*

2.5.2 Par groupe de 100 secondes

Graphique 4: Temps moyen par groupe de 100 secondes



Source : *Auteur a partir de python*

Les graphiques ci-dessus fait état des niveaux moyens des temps de latence des requêtes par intervalle de 10 et 1000 secondes. On peut remarquer que une grande partie du temps, la connexion semble avoir un comportement plus ou moins stable. Cependant, on observe plusieurs zones globalement repartie sur la période d'étude durant lesquelles les temps moyens des trajets des requêtes sont extrêmement au-dessus de la normale. Ces anomalies pourraient s'expliquer par une congestion temporaire du réseau, des politiques de filtrage ICMP appliquées par Google.

L'abscense de temps de latences sur le graphique correspond aux periodes pour lesquelles aucune réponse n'a été reçue. Cela signifie qu'aucune latence n'a pu être observée, ce qui peut être dû à plusieurs facteurs :

- **Perte de paquets** : Les requêtes envoyées n'ont pas abouti à une réponse en raison de pertes dans le réseau, causées par une congestion, des erreurs de transmission ou des problèmes de routage.
- **Timeouts** : Les délais de réponse peuvent avoir dépassé un seuil défini, entraînant une absence d'enregistrement des latences.
- **Blocage ou filtrage des requêtes** : Certains pare-feu ou politiques réseau peuvent empêcher les réponses d'être envoyées, créant ainsi des périodes sans données.
- **Défaillance du serveur cible** : Si le serveur destinataire est temporairement indisponible ou en surcharge, il peut ne pas répondre aux requêtes.

2.6 le nombre de requetes où la latence est anormalement élevée

L'analyse des latences montre que 9 685 requetes présentent des valeurs anormalement élevées, identifiées comme valeurs aberrantes selon la règle de l'intervalle interquartile (IQR).Le temps de latence correspondant est supérieur 153.84 secondes

2.7 le nombre de requetes où la latence est anormalement faible, relativement à l'ensemble

L'analyse des latences montre que 9 685 occurrences présentent des valeurs anormalement élevées, identifiées comme valeurs aberrantes selon la règle de l'intervalle interquartile (IQR).Le temps de latence correspondant est inférieur à 117.4 secondes L'observation de **71 occurrences** avec une latence anormalement basse, identifiées comme **valeurs aberrantes inférieures**, suggère plusieurs scénarios possibles :

- **Optimisation ponctuelle du réseau** : Certaines requêtes ont pu bénéficier d'une transmission ultra-rapide en raison d'un faible trafic ou d'une proximité avec le serveur.
- **Mécanismes de mise en cache** : Si certaines requêtes sont servies via un cache, elles peuvent afficher des temps de réponse bien inférieurs à la normale.

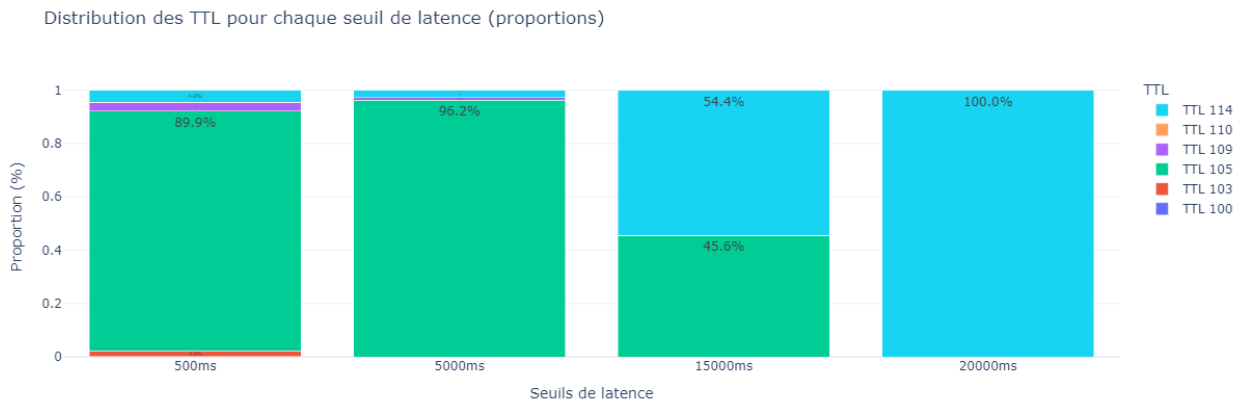
2.8 Anomalies où le temps de latence dépasse un certain seuil

Ces données révèlent une forte variabilité des temps de réponse du réseau, avec un nombre significatif de requêtes présentant des latences élevées :

- **3 257 requêtes** ont dépassé **500 ms**, signalant déjà une dégradation notable des performances.
- **1 144 requêtes** ont dépassé **5 000 ms**, indiquant des délais critiques affectant gravement l'expérience utilisateur.
- **57 requêtes** ont excédé **15 000 ms**, traduisant des temps d'attente extrêmement longs.
- **30 requêtes** ont franchi **20 000 ms**, ce qui suggère des quasi-échecs de transmission ou des congestions réseau sévères.

2.9 Repartition des TTL en fonction des seuil de latence anormales

Graphique 5: Repation des TTL en fonction du temps de latence



Source : *Auteur a partir de python*

Le diagramme présente la répartition des occurrences de TTL en fonction de niveaux de latence précis. Il apparaît clairement qu'à mesure que le temps de latence augmente, la

proportion de requêtes avec un TTL de 114 devient de plus en plus prépondérante. Cette tendance, d'une part, confirme que le TTL 114 est l'occurrence la plus fréquente parmi l'ensemble des requêtes, et d'autre part, qu'il est fortement associé à des latences élevées.

Cette prépondérance suggère que la voie empruntée par les requêtes affichant un TTL de 114 est probablement saturée. En effet, la concentration de requêtes à haute latence sur ce TTL laisse penser que le réseau rencontre une congestion sur cette route, ce qui ralentit le traitement des requêtes et accroît leur temps de latence.

2.10 Variation moyenne du délai de transmission des paquets ou gigue

La gigue, ou *jitter*, mesure la variation du délai de transmission des paquets dans un réseau. Elle est particulièrement importante pour les applications en temps réel, telles que le streaming vidéo, où une forte variabilité des délais peut entraîner une dégradation de la qualité de service. La variation moyenne entre les valeurs de RTT successives est donnée par :

$$Jitter = \frac{\sum |RTT_i - RTT_{i-1}|}{N-1} \tag{2.10.1}$$

où RTT_i (Round-Trip Time) est le temps qu'un paquet met pour aller d'un émetteur à un récepteur et revenir lors de la i-ème requête envoyée, et N est le nombre total de requêtes.

D'après https://www.iptis.fr/blog/comprendre-la-qualite-de-sa-connexion-internet, une gigue inférieure à 20 ms est considérée comme optimale pour garantir une connexion stable et fluide. Or, dans notre cas, elle atteint 500.122 ms, une valeur extrêmement élevée qui indique une forte variabilité des délais entre les paquets de données. Une telle gigue peut entraîner des temps de réponse imprévisibles et dégrader significativement l'expérience utilisateur, rendant la navigation sur Internet lente et irrégulière. Cette instabilité résulte d'une fluctuation importante du temps que met chaque paquet pour traverser le réseau, avec des écarts pouvant aller de 0 à 500 ms, compromettant ainsi la qualité des services nécessitant un délai minimal, comme le streaming en temps réel.

3. INTERPRETATION ET DISCUSSION DES RESULTATS

Dans le cadre de l'évaluation de la latence du serveur DNS public de Google (IP 8.8.8.8), 80 222 requêtes ping ont été émises afin d'analyser la stabilité et la performance de la connexion. L'étude a permis d'examiner les requêtes réussies, les erreurs rencontrées et les variations des TTL observées.

3.1 Statistiques générales sur l'émission et la réception des requêtes

Sur les 80 222 requêtes envoyées, 76 529 ont été transmises avec succès, tandis que 3 693 ont échoué en raison de l'erreur *"No Route to Host"*, indiquant une absence de route valide vers la destination. Parmi celles qui ont été envoyées, 65 722 ont reçu une réponse du serveur, tandis que 10 807 ont atteint le serveur sans réponse, générant l'erreur *"Request Time Out"*. Les erreurs *"Request Time Out"* peuvent s'expliquer par plusieurs facteurs :

- Filtrage ICMP appliqué par Google, qui pourrait ignorer certaines requêtes.
- Congestion réseau, entraînant une perte de paquets.
- Mécanismes de limitation du trafic ICMP mis en place par Google pour éviter les abus.

Les erreurs *"No Route to Host"* sont plus probablement liées à des interruptions réseau côté client, une perte de connectivité temporaire ou un problème de routage externe.

3.2 Analyse des TTL et de la stabilité du routage

L'étude des TTL montre que 80% des requêtes ont un même TTL, ce qui indique une route réseau stable et prédominante pour la majorité du trafic. Les 20% restants affichent des TTL différents, suggérant :

- L'existence de routes alternatives occasionnelles utilisées en cas de congestion.
- Des ajustements dynamiques du routage pouvant impacter certaines requêtes.
- Une instabilité intermittente sur certains segments du réseau.

Toutefois, la grande majorité des requêtes suivant une route fixe suggère que le routage dynamique n'est pas un facteur clé dans les variations de latence. Les fluctuations observées sont donc plus probablement liées à des problèmes de congestion réseau, de gestion des priorités ICMP ou de filtrage par Google.

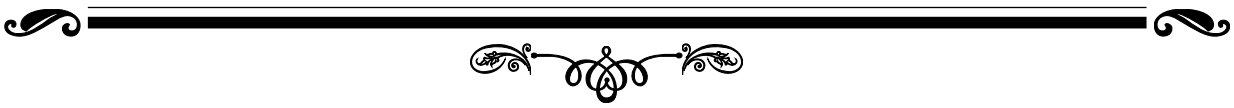
3.3 Analyse des temps de latence

Les statistiques descriptives des latences révèlent une grande variabilité des temps de réponse. Bien que le temps médian soit de 133.8 ms, la moyenne est nettement plus élevée (634.9 ms) en raison de valeurs extrêmes pouvant atteindre 927 118 ms. Cette dispersion des latences peut être expliquée par :

- Des pics de congestion réseau à certaines périodes, augmentant le délai de transmission.
- Des routes alternatives plus longues, utilisées temporairement pour certaines requêtes.
- Un traitement différentiel des paquets ICMP par Google, pouvant pénaliser certaines requêtes.

En observant l'évolution des latences par intervalle de temps, plus de 75% des requêtes présentent une stabilité relative, mais certaines périodes montrent des augmentations brutales des délais, probablement dues à des interférences ou une congestion temporaire du réseau.

CONCLUSION



En conclusion, cette étude montre que la connexion au serveur DNS public de Google est généralement stable, comme le témoigne la constance du TTL pour près de 80 % des requêtes. Cette stabilité suggère l'utilisation prépondérante d'une route principale, garantissant une connectivité fiable. Toutefois, les anomalies détectées — illustrées par des erreurs "Request Time Out" et "No Route to Host", ainsi que des pics de latence notables — indiquent que des fluctuations temporaires se produisent. Ces dernières sont probablement dues à des congestions réseau, à des mécanismes de filtrage ICMP ou à des réacheminements dynamiques. Bien que la majorité des requêtes aient donné de bons résultats, ces observations soulignent la nécessité d'une surveillance continue et d'analyses approfondies pour identifier et résoudre les défaillances potentielles du réseau. En définitive, cette investigation constitue une base solide pour comprendre les facteurs influençant la latence et offre des pistes concrètes pour optimiser la qualité du service DNS, notamment en traitant les interruptions intermittentes et en assurant l'efficacité du réseau.