Downloaded from: justpaste.it/hqs6x

Combien faut-il d'ordinateurs pour traiter les demandes en moteur de recherche et en IA?

Combien faut-il d'ordinateurs à Google pour traiter les demandes en moteurs de recherche ?

Google exploite des millions de serveurs à travers ses centres de données répartis dans le monde entier. Le nombre exact n'est pas officiellement divulgué, mais on estime qu'il dépasse **2 à 5 millions de serveurs**.

Chaque requête envoyée à Google (comme une recherche) est traitée en quelques millisecondes grâce à ces infrastructures massivement parallélisées, optimisées avec des algorithmes avancés et du matériel sur mesure (comme les TPU, des puces conçues pour l'IA).
Pour donner une idée, Google traite environ **100 000 requêtes par seconde**, soit plus de **8 milliards par jour**.

Et l'intelligence artificielle, combien ?

L'intelligence artificielle, en particulier les grands modèles comme ceux de Google (Gemini), OpenAI (ChatGPT), ou Meta (LLaMA), demande encore plus de puissance de calcul que la simple recherche web.

Les chiffres exacts varient, mais voici une estimation :

- **Entraînement d'un modèle** : Un grand modèle comme Gemini ou GPT-4 nécessite des **milliers de GPU ou TPU** fonctionnant en parallèle pendant **des semaines ou mois**. Par exemple, GPT-4 aurait nécessité **des dizaines de milliers de GPU A100 de NVIDIA**.
- **Inférence (réponses aux utilisateurs)** : Chaque requête d'IA peut nécessiter **des dizaines à des centaines de milliards de calculs**, soit bien plus qu'une simple recherche Google.
- **Centres de données** : Google et Microsoft possèdent des fermes de serveurs spécialisés en IA, comptant potentiellement **des centaines de milliers de puces IA** (comme les TPU de Google et les GPU de NVIDIA).

En gros, répondre à une requête d'IA coûte **plusieurs centaines de fois plus** en ressources qu'une recherche classique. C'est pour ça que les entreprises investissent massivement dans des puces optimisées et des techniques pour réduire ces coûts.