12/07/2024

PERNON Etienne

M2ID1 – CCI Campus Alsace

Veille Informatique

Groq : La promesse des LPU

Table des matières

[Introduction 2](#_Toc171185491)

[Objectif 2](#_Toc171185492)

[Abstract 2](#_Toc171185493)

[Méthodologie 2](#_Toc171185494)

[Analyse 2](#_Toc171185495)

[Contexte 2](#_Toc171185496)

[Description de la technologie 2](#_Toc171185497)

[Historique de la technologie 3](#_Toc171185498)

[Analyse marco-environmental (pestel) 3](#_Toc171185499)

[Perspective 3](#_Toc171185500)

[Concurrence 3](#_Toc171185501)

[Description du marché 3](#_Toc171185502)

[Présentation de la concurrence 3](#_Toc171185503)

[Analyse des options 3](#_Toc171185504)

[Sources 3](#_Toc171185505)

# Introduction

Depuis DeepMind développé par Google en 2014 jusqu’à GPT 4 par Open AI en 2023, Il est clair que l’IA a déjà pris une place importante dans notre existence. Cependant un frein à son développement n’a pas encore été totalement levé, comme le dit Johnathan Ross[[1]](#footnote-1) CEO de l’entreprise Groq : «  The primay limiter for those applications is compute ». La puissance de calcule a souvent été un facteur limitant, que ce soit dans la phase d’apprentissage des modèles, comme dans l’application de ces modèles.

Cependant, Google a su montrer qu’il été possible de faire du Deep Learing avec des GPU et aujourd’hui l’intérêt des investisseurs sur le marché de l’IA n’est plus à démontrer. De nombreuses entreprise sont prêtes à investir et lever des fonds pour crée la nouvelle génération de processeur ASIC spécialisé dans l’inférence.

L’opportunité que représente ces avancements est énorme, elle représente l’ouverture au plus grand nombre des fonctionnalités de IA, des possibilités d’intégration rentable dans nos applications et des temps de réponse record. Les promesses de Groq sont simples : « Make IA free […] as me provide so much compute per dollar that is might as well be free ».

## Objectif

En tant qu’informaticien dans une PME j’envisage d’intégrer de l’IA dans mes solutions applicative. Cependant, je ne dispose pas d’énorme moyen et pour l’instant les principales API donnant accès à un modèle entrainé excède mon budget. Insert Petit rapport sur les prix

L’objectif est donc d’observer le prix et la latence des solutions *cloud computing[[2]](#footnote-2)* pour les principaux modèles d’IA open source.

~~pour crée un chat bot avec un LLM ou des commandes vocale avec wisper~~. Description des besoin …

Je ne voyais que la perspective d’application qui pouvais se permettre d’attendre.

Je n’avais jamais entendu parler de solution qui permettais de faire d’avoir les réponses d’une IA en temps réel.

Reconnaitre les machines et les pièces en temps réel pour faire remonter les tickets, pré-remplir les inspection des opérateurs.

De ce fait, cette application demande un temps de réponse aussi court que possible.

Ainsi, l’objectif de cette veille sera de déterminer si les solutions de *cloud computing* actuel permettent la mise en œuvre de ma solution et si Groq est réellement le candidat idéal.

## Abstract

Oui c’est rapide et pas cher

# Méthodologie

## Outils de Collecte

### Passive

### Active

## Mise en relation des sources

### Mindmap

### Archivage

# Analyse

## Contexte

### Historique de la technologie

Dans les années 2000 quand Google a commencé à s’intéresser à l’IA[[3]](#footnote-3), il aurait été impossible à une entreprise d’engager les fonds pour fabriquer un processeur dédier à l’IA, alors que personne ne pouvait prédire si les technologies liées seront importantes.

Par chance les GPU ont commencé à être capable de supporter la charge de calcule des premiers modèle de reconnaissance d’image, comme AlexNet de Google en 2012. Ce réseau de neurone est le premier modèle à atteindre un taux record d’erreur de 16%[[4]](#footnote-4), gagner la compétition d’ImageNet[[5]](#footnote-5) et prouver que les méthodes d’apprentissage profond sont aujourd’hui réalisables. [[6]](#footnote-6)

Après ces premiers pas éclatant devant le grand public et un apprentissage de 3 jour sur 1.000 machines comptent en tout 16.000 cœurs. Il était devenu nécessaire pour Google de reconsidérer l’option de crée un processeur ASCI spécialisé dans les calcule d’IA.

De plus, l’entreprise pourrait devenir dépendante de Nvidia à qui elle a déjà commandé 40.000 GPU. Les innovations se font en parelle comme pour les premiers essais de reconnaissance vocale [[7]](#footnote-7) et les besoins en puissance de calcule ne font qu’augmenter.

4 ans plus tard, le modèle AlphaGo[[8]](#footnote-8) défie le monde en devenant le meilleur joueur de Go au monde. Le secret de cette réussit est une nouvelle puce nommée TPU[[9]](#footnote-9) pour Tensor Processing Unit avec laquelle Google entraine sont modèle depuis au moins 1ans.

Depuis 8 ans se sont écoulé et les modèles d’IA ont toutes sortes d’applications. Certains comme le célèbre ChatGPT d’OpenAI en partenariat avec Microsoft ont une architecture essentiellement composée de GPU Nvidia[[10]](#footnote-10), tandis que Google continue de commercialiser des variantes de ces processeur comme le Google Tensor G4 [[11]](#footnote-11) présent dans le Pixel 9.

Récemment un nouvel acteur de la Silicon Valley est venu rebattre les cartes d’un marché détenu par ces deux grandes puissances : Groq. L’entreprise crée par Johnathan Ross, un ingénieur ayant participé à l’élaboration la puce de la TPU de Google entends défier le leader mondial Nvidia.

L’originalité de Groq réside dans une nouvelle puce ASCI nommé LPU pour Language Process Units. Cette puce spécialement conçut pour des calculs d’inférence est en service dans pour fournir une interface de chat [[12]](#footnote-12) et une API pour communiquer avec les principaux modèles de LLM. Mais elle est aussi éprouvée chez deux clients dont l’identité restera à la discrétion de l’entreprise, selon son créateur[[13]](#footnote-13), une entreprise de « voiture autonome » et une autre de « fintech ».

Aujourd’hui le marché des processeurs dédier expose … investissement … secteur prometteur indépendamment de l’entreprise Groq

192 start up développent des puces dédier à l’IA dont Groq qui en 2021a été la 4ème start up la plus financé dans ce secteur.

vise, selon son créateur, à rendre la force de calcule si peux cher qu’elle en serait presque gratuite.

3 :16 les 6 premier mois à travailler sur le compilateur

Groq c’est le seul service à journir l’api > L’archi > la puce

Dans cette équipe de chercheur formé par Google

Google travail en secret sur la nouvelle puce TPU et entraine SERETEMENT deepblue dessus !!!!

https://thechipletter.substack.com/p/googles-first-tensor-processing-unit

This hardware has consistently had a ‘chicken and egg’ problem though. CPUs are not an efficient way of implementing a neural network. But how could firms justify ongoing investment in specialized hardware without proof that neural networks are useful? But for neural networks to work, they needed more powerful hardware. Time and again projects to build dedicated hardware have been abandoned as the hype cycle turned. In the end, it took hardware originally built for another purpose, GPUs, to break out of the impasse.

Pourquoi cette fracture en TPU et LPU >> il faut pouvoir faire l’entrainement et l’inférance pour attirer les clients

### Description de la technologie

ASIC > TPU / google > LPM

Qu’es ce qu’un ASCI à quoi ils sont utilisés

les avantages des processeurs dédier : cout énergétique + vitesse

### Analyse marco-environmental (pestel)

### Perspective

#### Aspect destructif

Analyse des performances

#### Projets innovants

https://wow.groq.com/groq-labs/

## Concurrence

### Description du marché

### Présentation de la concurrence

### Analyse des options

#### Opportunité

#### Force

#### Faiblesse

#### Menace

#### Matrice Swot

# Sources

Mon petit doigt

## Groq – officiel

**Tester Groq :** <https://groq.com/>

**Recherche public 2020 :** <https://wow.groq.com/wp-content/uploads/2020/06/ISCA-TSP.pdf>

**Recherche public 2022 :** <https://wow.groq.com/wp-content/uploads/2024/02/GroqISCAPaper2022_ASoftwareDefinedTensorStreamingMultiprocessorForLargeScaleMachineLearning.pdf>

**Projets associés :** <https://wow.groq.com/groq-labs/>

## Benchmark

**Mixtral 8x7B Instruct: API Provider :** <https://artificialanalysis.ai/models/mixtral-8x7b-instruct/providers>

## Histoire de Groq

## Glossaire

**Cloud computing :** <https://www.cnil.fr/fr/definition/cloud-computing>

1. Interview du Cambrian AI : [lien](https://cambrian-ai.com/interview-with-ceo-jonathan-ross-of-groq/) [↑](#footnote-ref-1)
2. [Selon la Cnil](https://www.cnil.fr/fr/definition/cloud-computing) le cloud computing […] fait référence à l’utilisation de la mémoire et des capacités de calcul des ordinateurs et des serveurs répartis dans le monde entier et liés par un réseau. » [↑](#footnote-ref-2)
3. [The Chip Letter](https://thechipletter.substack.com/p/googles-first-tensor-processing-unit) : Google's First Tensor Processing Unit: Origins [↑](#footnote-ref-3)
4. [Pine cone](https://www.pinecone.io/learn/series/image-search/imagenet/) : AlexNet and ImageNet: The Birth of Deep Learning [↑](#footnote-ref-4)
5. [Data France](https://datafranca.org/wiki/ImageNet): « Base de données d'images annotées par des humains qui est destinée à des travaux de recherche en vision par ordinateur ainsi qu'en apprentissage profond » [↑](#footnote-ref-5)
6. [Arxiv](https://arxiv.org/pdf/1112.6209): Building High-level Features Using Large Scale Unsupervised Learning [↑](#footnote-ref-6)
7. [Youtube](https://www.youtube.com/watch?v=VazSEtXHDcI): Android App Google Speech Recognition Tutorial and Example [↑](#footnote-ref-7)
8. [Discovery](https://discovery.ucl.ac.uk/id/eprint/10045895/1/agz_unformatted_nature.pdf): Mastering the Game of Go without Human Knowledge [↑](#footnote-ref-8)
9. [Google Cloud Blog](https://cloud.google.com/blog/products/ai-machine-learning/google-supercharges-machine-learning-tasks-with-custom-chip?hl=en): Google supercharges machine learning tasks with TPU custom chip [↑](#footnote-ref-9)
10. [01net](https://www.01net.com/actualites/pour-chatgpt-microsoft-a-depense-des-millions-de-dollars-en-gpu-nvidia.html): Pour ChatGPT, Microsoft a dépensé des millions de dollars… en GPU Nvidia [↑](#footnote-ref-10)
11. [LeMondeNumérique](https://lemondenumerique.ouest-france.fr/le-prochain-processeur-google-tensor-g4-prevu-pour-les-pixel-9-pixel-9-pro-et-pixel-fold-2-deja-a-lessai/): Le prochain processeur Google Tensor G4 [↑](#footnote-ref-11)
12. [Groq](https://groq.com/): LLM chat [↑](#footnote-ref-12)
13. [Cambrian – AI](https://cambrian-ai.com/interview-with-ceo-jonathan-ross-of-groq/): Interview with CEO Jonathan Ross of Groq 4 : 53 [↑](#footnote-ref-13)