12/07/2024

PERNON Etienne

M2ID1 – CCI Campus Alsace

Veille Informatique

Groq : La promesse d’une puce

Table des matières

[Introduction 3](#_Toc171444626)

[Objectif 4](#_Toc171444627)

[Abstract 4](#_Toc171444628)

[Méthodologie 5](#_Toc171444629)

[Outils de Collecte 5](#_Toc171444630)

[Passive 5](#_Toc171444631)

[Active 5](#_Toc171444632)

[Mise en relation des sources 5](#_Toc171444633)

[Mindmap 5](#_Toc171444634)

[Archivage 5](#_Toc171444635)

[Analyse 5](#_Toc171444636)

[Contexte 5](#_Toc171444637)

[Historique de la technologie 5](#_Toc171444638)

[Description de la technologie 8](#_Toc171444639)

[Analyse marco-environmental (pestel) 8](#_Toc171444640)

[Perspective 8](#_Toc171444641)

[Concurrence 8](#_Toc171444642)

[Description du marché 8](#_Toc171444643)

[Présentation de la concurrence 8](#_Toc171444644)

[Analyse des options 8](#_Toc171444645)

[Sources 9](#_Toc171444646)

[Groq – officiel 9](#_Toc171444647)

[Benchmark 10](#_Toc171444648)

[Histoire de Groq 10](#_Toc171444649)

[Glossaire 10](#_Toc171444650)

# 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 leurs l’applications.

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[[2]](#footnote-2). 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, c’est 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 de définir si Groq est effectivement une technologie disruptive et qu’elles sont les perspectives que nous donne le marché des processeur dédier à l’IA.

## Abstract

Oui c’est rapide et pas cher

Avec leur image de mark (nom :lpu) et le moment ils s’emparent d’un secteur, il se sont crée une image dans le secteur, une puce pour les

Une techno qui concentre une certaine hype

C’est les premier à aboutir

Mais nvidia le fais déjà et c’est nvidia (h100)

# Analyse

## Contexte

### Description de la technologie

**Qu’est-ce qu’un ASCI**

Les processeurs ASIC sont une famille de processeurs comme les CPU ou les GPU à l’exception qu’il ne pourrait être utilisé en informatique classique. Leur but est d’être particulièrement efficace pour la réalisation de tâches arithmétique bien spécifique. Ces optimisations sont faites en convertissant certains algorithmes et certaines fonctions en composant matériel[[3]](#footnote-3).

**Des gains en vitesse et en énergie**

Ces processeurs présentent des avantages en termes de cout énergétique et de vitesse qui sont incontestable. Les optimisations matérielles apporté viennent soulager les tâches de calcules du processeur.

Toutefois, il faut noter que les gains en énergie de ces puces ne sont pas évidents pour un usage personnel. S’il est vrai qu’une puce ASIC exécutera un calcule spécifique en moins de temps et en consommant moins de Watt qu’un CPU ou un GPU. Il ne faut pas oublier que les puce ASIC à destination du grand public avoisine les 800W, le temps gagner sera utiliser pour exécuter d’autre opération de calcule, dans le cadre d’une utilisation ininterrompue, les factures d’électrice peuvent monter en flèche[[4]](#footnote-4).

**Ils sont fréquemment utilisés dans**

* Les Data center comme ceux de Google [[5]](#footnote-5)
* Pour miner des cryptos money [[6]](#footnote-6)

**Une sous-catégorie des ASIC : les LPU**

Pour Langages Process Unit les LPU sont eux des puces spécialisés dans l’inférence[[7]](#footnote-7). Aujourd’hui la puce LPU est la propriété de la société Groq tout comme la puce TPU de Google. Ainsi, les brevets n’étant pas public, il nous sera impossible d’expliquer le fonctionnement de ces optimisations.

Son concepteur nous révèle tout de même que les 6 premier mois de développement ont été dédié à la conception du compilateur [[8]](#footnote-8)servant à interagir avec la puce. Le CEO de groq assure que cette API est bien plus facile d’accès que la technologie TenserFlow utilisé pour les TPU, un argument de bonne guerre de la part d’un de ces anciens concepteurs.

### Historique de la technologie

Dans les années 2000 quand Google a commencé à s’intéresser à l’IA[[9]](#footnote-9), 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 seraient 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%[[10]](#footnote-10), gagner la compétition d’ImageNet[[11]](#footnote-11) et prouver que les méthodes d’apprentissage profond sont aujourd’hui réalisables. [[12]](#footnote-12)

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 [[13]](#footnote-13) et les besoins en puissance de calcule ne font qu’augmenter.

4 ans plus tard, le modèle AlphaGo[[14]](#footnote-14) défie le monde en devenant le meilleur joueur de Go au monde. Le secret de cette réussit est une nouvelle puce nommée TPU[[15]](#footnote-15) pour Tensor Processing Unit avec laquelle Google entraine sont modèle depuis au moins 1ans.

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[[16]](#footnote-16), tandis que Google continue de commercialiser des variantes de ces processeur comme le Google Tensor G4 [[17]](#footnote-17) présent dans le Pixel 9.

Cependant, un nouvel acteur de la Silicon Valley vient rebattre les cartes d’un marché détenu par ces deux grandes puissances : Groq. Cette entreprise crée par Johnathan Ross en 2016 se fait discrète jusqu’à une levée de fond remarquable[[18]](#footnote-18) en 2021. Son créateur, un des ingénieurs ayant participé à l’élaboration la puce 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 dors est déjà en service pour fournir une interface de chat [[19]](#footnote-19) ainsi qu’une 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[[20]](#footnote-20), une entreprise de « voiture autonome » et une autre de « fintech ».

Aujourd’hui le marché des processeurs dédier à l’IA a explosé, il est nommé « AI Semiconductor market » et pèse cette année USD 30 Billion. Le rapport de [Wood side Cap](https://woodsidecap.com/wp-content/uploads/2023/02/WCP-AI-Semiconductor-2022-Q4.pdf?utm_source=BenchmarkEmail&utm_campaign=WCP_Industry_Report%3a_AI_Semiconductor_Market_Q4_2022&utm_medium=email) pointe les 192 start-up qui ont profités de ces investissement dans le monde entier et donc autant de potentiel concurrent .

## Concurrence

### Description du marché

Le Cloud AI market représente l’ensemble des solutions de data center qui mettent à disposition leur puissance de calcul pour entrainer ou utilisé des modèles d’IA. Ce secteur étant un plein essor depuis les deux dernières années, il entraine avec lui le marché des semiconducteurs dédié à l’IA qui se voit obliger de grandir pour répondre à la demande.

La société McKinsey & Company [[21]](#footnote-21)envisage les principales catégories d’applications à demander de la puissance de calcul seront :

* La génération & l’interprétation de code informatique
* La création de contenu créatif
* L’interaction avec le client
* Des applications innovantes
* Résumer de grand set de data organisé

**Investissements dans ces secteurs**

Pour le secteur du Cloud AI Market la différence entre l’avant et l’après ChatGPT se fait sentir. Si ce marché représentait tout de même USD 46.67 billion en 2023 et arrive jusqu’à USD 60.35 billion en 2023, il devrait atteindre les USD 397.81 billion d’ici 2030[[22]](#footnote-22).

Le secteur des semiconducteurs dédié à l’IA suit une courbe similaire, bien que nivelé vers le bas. L’année 2023 a pesé pour USD 23 billion, la suivante USD 30 billion et les estimations pour 2031 s’élève à USD 198 billion[[23]](#footnote-23).

### Présentation de la concurrence

#### Secteur des semiconducteurs

Comme nous l’avons dit précédemment, 192 start-ups dans ces domaines ont été financé par le fonds d’investissement Wood Side Cap. Certaines comme [Horizon Robotics](https://en.horizon.cc/core-technologies/) travail sur des puces dédié aux voiture autonome, une autre comme [SambaNova](https://sambanova.ai/getstarted) entre en concurrence avec Groq mais uniquement dans le domaine du B2B.

Cet exemple nous montre que d’autre solutions apportant une API, un Data Center et une puce ensemble réunie sont susceptible d’apparaître sur le marché plus vite qu’on ne le pense. Cependant, il n’y a aujourd’hui aucune autre entreprise financée par le Wood Side Cap qui dans les benchmarks porté à notre connaissance.

Du coté des géants comme Nvidia les investissements dans ce secteur ont augmenté depuis 2023 et l’entreprise a sorti en mars de la même année sa tout dernière puce H100. La puce H100 est utilisé dans les Data Center de Microsoft[[24]](#footnote-24), sa puissance équivaut à 567 puces LPU et tous les acteurs se l’arrache. Bien que les processeurs LPU soit pensé pour être utilisé en série et non individuellement, cette puce est de loin la plus aboutie sur le marché et pour cause Nvidia est devenu ces dernières années une des 3 entreprises les plus coté en bourse.

#### Benchmark : cloud computing

Nous avons principalement parler des puces utilisées dans les Data Center, mais nous avons écarté pour l’instant la finalité qui nous intéresse dernière ces avancées : AI API Provider.

Les services dont nous parlons sont ceux qui offre la possibilité de requêter un modèle d’IA entrainé via une API et dont les prix sont calculés en fonction du nombre de Tokens donné / reçu.

Les premiers tests sont frappant [[25]](#footnote-25)pour les fournisseurs les plus populaires qui fleurissent aujourd’hui comme : Amazon Bedrock, Together.ai, Mistral, Replicate, Fireworks. En effet, le service que propose Groq est le meilleur en rapidité. Il renvoie 2x plus de tokens par seconde Groq (553 t/s) que le second Fireworks (251 t/s).

Pour ce qui est du prix, l’offre de Groq s’aligne avec DeepInfra qui propose des résultats médians sauf en vitesse de réponse où ils sont bon dernier. Ce résultat nous montre que l’objectif de cette technologie n’est pas encore atteint, pour rappel : « Make IA free […] as me provide so much compute per dollar that is might as well be free ».

Cependant, ces résultats sont aussi très encourageants. Groq tiens le marché par les deux bouts en se plaçant comme leader des coûts et des performances.

**Projets innovants**

[*https://wow.groq.com/groq-labs/*](https://wow.groq.com/groq-labs/)

## Conclusion

Techno prometeuse qui a besoin de temps

Un premier pas à la hauteur de ses promesses

Pas encore de cassure sur le marché au niveau des prix mais vitesse x2

Nvidia a largement les moyens et le potentiel d’opérer cette fracture lorsque leur prochaine puce dédier à l’IA sera commercialisé. (prix & vitesse)

# Sources

Mon petit doigt

## Groq – site 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/>

## Revue de presse

[La tribune](https://www.latribune.fr/technos-medias/informatique/investir-dans-l-intelligence-artificielle-c-est-investir-dans-la-croissance-anne-bouverot-et-philippe-aghion-992909.html$) : « Investir dans l'intelligence artificielle, c'est investir dans la croissance »

[Community FS](https://community.fs.com/fr/article/a-brief-introduction-to-cpu-gpu-asic-and-fpga.html): Introduction aux CPU, GPU, ASIC et FPGA

## Benchmark

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

## Ressources video

[Cambrian AI](https://cambrian-ai.com/interview-with-ceo-jonathan-ross-of-groq/) : Interview with CEO Jonathan Ross of Groq

## Glossaire

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

1. [Cambrian AI](https://cambrian-ai.com/interview-with-ceo-jonathan-ross-of-groq/) : Interview with CEO Jonathan Ross of Groq [↑](#footnote-ref-1)
2. [La tribune](https://www.latribune.fr/technos-medias/informatique/investir-dans-l-intelligence-artificielle-c-est-investir-dans-la-croissance-anne-bouverot-et-philippe-aghion-992909.html$) : « Investir dans l'intelligence artificielle, c'est investir dans la croissance » [↑](#footnote-ref-2)
3. [Community FS](https://community.fs.com/fr/article/a-brief-introduction-to-cpu-gpu-asic-and-fpga.html): Introduction aux CPU, GPU, ASIC et FPGA [↑](#footnote-ref-3)
4. [Bitproid](https://bitproit.com/asic-vs-gpu-mining/) : ASIC vs GPU Profitability [↑](#footnote-ref-4)
5. [Cloud Google](https://cloud.google.com/tpu?hl=fr): Accélérez le développement de l'IA avec les TPU Google Cloud [↑](#footnote-ref-5)
6. [Youtube](https://www.youtube.com/watch?v=QBJAkz3W6h8&t=1s): CPU vs GPU vs ASIC mining [↑](#footnote-ref-6)
7. [PureStorage](https://www.purestorage.com/fr/knowledge/what-is-lpu.html) : Qu’est-ce qu’une unité de traitement du langage (LPU) ? [↑](#footnote-ref-7)
8. [Youtube](https://www.youtube.com/watch?v=Z0jqIk7MUfE) : "Compute is the New Oil", Leaving Google, Founding Groq, Agents, Bias/Control (Jonathan Ross) 3 : 16 [↑](#footnote-ref-8)
9. [The Chip Letter](https://thechipletter.substack.com/p/googles-first-tensor-processing-unit) : Google's First Tensor Processing Unit: Origins [↑](#footnote-ref-9)
10. [Pine cone](https://www.pinecone.io/learn/series/image-search/imagenet/) : AlexNet and ImageNet: The Birth of Deep Learning [↑](#footnote-ref-10)
11. [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-11)
12. [Arxiv](https://arxiv.org/pdf/1112.6209): Building High-level Features Using Large Scale Unsupervised Learning [↑](#footnote-ref-12)
13. [Youtube](https://www.youtube.com/watch?v=VazSEtXHDcI): Android App Google Speech Recognition Tutorial and Example [↑](#footnote-ref-13)
14. [Discovery](https://discovery.ucl.ac.uk/id/eprint/10045895/1/agz_unformatted_nature.pdf): Mastering the Game of Go without Human Knowledge [↑](#footnote-ref-14)
15. [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-15)
16. [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-16)
17. [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-17)
18. [Wood side Cap](https://woodsidecap.com/wp-content/uploads/2023/02/WCP-AI-Semiconductor-2022-Q4.pdf?utm_source=BenchmarkEmail&utm_campaign=WCP_Industry_Report%3a_AI_Semiconductor_Market_Q4_2022&utm_medium=email) : AI Semiconductor Market Q4 2022 [↑](#footnote-ref-18)
19. [Groq](https://groq.com/): LLM chat [↑](#footnote-ref-19)
20. [Cambrian – AI](https://cambrian-ai.com/interview-with-ceo-jonathan-ross-of-groq/): Interview with CEO Jonathan Ross of Groq 4 : 53 [↑](#footnote-ref-20)
21. [Mc Kinsey](https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry) : The surge of interest in and use of generative AI [↑](#footnote-ref-21)
22. [Fortune Buisness Insigghts](https://www.fortunebusinessinsights.com/cloud-ai-market-108878) : Cloud AI Market Size [↑](#footnote-ref-22)
23. [Edge AI Vision](https://www.edge-ai-vision.com/2024/04/ai-chip-market-to-grow-10x-in-the-next-ten-years-and-become-a-300-billion-industry/) : AI Chip Market to Grow 10x in the Next Ten Years and Become a $300 Billion Industry [↑](#footnote-ref-23)
24. [Learn Microsoft](https://learn.microsoft.com/fr-fr/azure/virtual-machines/nd-h100-v5-series) : Série ND H100 v5 [↑](#footnote-ref-24)
25. [Artificial Analysis](https://artificialanalysis.ai/models/mixtral-8x7b-instruct/providers): Mixtral 8x7B Instruct: API Provider Benchmarking & Analysis [↑](#footnote-ref-25)