12/07/2024

Veille Informatique

Groq : La promesse d'une puce

PERNON Etienne
M2ID1 – CCI CAMPUS ALSACE

Table des matières

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] CEO de l'entreprise Groq : « The primary limiter for those applications is compute ». La puissance de calcul a souvent été un facteur limitant, que ce soit dans la phase d'apprentissage des modèles, comme dans leurs applications.

Cependant, Google a su montrer qu'il était possible de faire du Deep Learning avec des GPU et aujourd'hui l'intérêt des investisseurs sur le marché de l'IA n'est plus à démontrer[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 ».

[1] Cambrian AI : Interview with CEO Jonathan Ross of Groq
[2] La tribune : « Investir dans l'intelligence artificielle, c'est investir dans la croissance »

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 excèdent mon budget. En effet, de nombreuses entreprises perdent de l'argent sur leur service d'API Mixtral[3].

L'objectif est donc de définir si Groq est effectivement une technologie disruptive et quelles sont les perspectives que nous donne le marché des processeurs dédiés à l'IA.

[3] Semianalysis : Groq Inference Tokenomics: Speed, But At What Cost?

Méthode

Source d'information

Pour recueillir les informations nécessaires à l'élaboration de ce dossier nous avons utilisé les carneaux suivants :

- Forum tel que Reddit pour rester au courant des dernière tendance
- Stack Overflow pour sonder les communauté active à un moment donné
- Les blogs de développeur comme le français Korben
- Les revues de Presse en ligne
- Les vidéos youtube
- Les benchmarks
- Les documentations officielles
- Les publications scientifiques
- Les bilans de fonds d'investissement

Outil récolte

Récolte active

Google fournit un arsenal de paramètres à ses recherches qui permettent une précision extraordinaire. Ces paramètres s'utilisent par l'onglet recherche avancé ou en utilisant les commandes de Google dork[4] qui offre des fonctionnalités plus avancées.

Récolte passive

Pour faire arriver l'information jusqu'à soi, il est possible d'utiliser les Google Alertes. Cette méthode se combine avec les commandes de Google dork puisque les Google Alerte permettent d'enregistrer une google recherche spécifique et d'être notifié

[4] Github : Google Dork cheatsheet

dès que le résultat de la recherche change. L'importance d'une requête Google la plus précise possible est alors évidente, sans quoi vous serez inondé d'alerte.

Il est aussi possible de faire venir l'information à soi en utilisant des flux RSS. Cette technologie a l'avantage d'être très répandue chez revue de presse, ce qui la rend facile à mettre en place tout en respectant une pluralité des sources. De plus, en choisissant bien ces sources, les Flux RSS sont bien plus faciles à utiliser que les Google Alertes. Cette technologie ne nécessite pas réglage complexe et la fréquence des alertes sera toujours en adéquation avec le rythme de parution des sources choisies.

Outil d'analyse

Mind Map

Les cartes mentales sont des outils remarquables. Ils permettent non seulement d'organiser ses idées et de les représenter visuellement, mais aussi de tracer des liens entre certaines idées en apparence éloignées.

Pour la réalisation de ce dossier nous avons renseigné les sources les plus importantes, puis leurs articles sur une carte mentale. Dans un second temps, nous avons sorti des idées et des citations des articles sous forme de : problème, solution, idée. Enfin, nous avons relié les différents acteurs avec leurs idées respectives.

Pour finir, les cartes mentales nous ont aussi été utiles dans un but d'archivage. Tous les documents téléchargés peuvent y être inséré, les liens vers les sources sont conservés, les citations sont liées à leurs sources et leur auteur.

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 arithmétiques bien spécifiques. Ces optimisations sont faites en convertissant certains algorithmes et certaines fonctions en composant matériel[5].

Des gains en vitesse et en énergie

Ces processeurs présentent des avantages en termes de coût énergétique et de vitesse qui sont incontestable. Les optimisations matérielles apportées viennent soulager les tâches de calcul 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 gagné sera utilisé pour exécuter d'autre opération de calcule, dans le cadre d'une utilisation ininterrompue, les factures d'électricité peuvent monter en flèche[6].

Ils sont fréquemment utilisés dans

- Les Data center comme ceux de Google [7]
- Pour miner des cryptos money [8]

[5] Community FS : Introduction aux CPU, GPU, ASIC et FPGA
[6] Bitproid : ASIC vs GPU Profitability
[7] Cloud Google : Accélérez le développement de l'IA avec les TPU Google Cloud
[8] Youtube : CPU vs GPU vs ASIC mining

Une sous-catégorie des ASIC : les LPU

Pour Langages Process Unit les LPU sont eux des puces spécialisés dans l'inférence[9]. 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 dédiés à la conception du compilateur [10]servant à interagir avec la puce. Le CEO de groq assure que cette API est bien plus facile d'accès que la technologie TenserFlow utilisée 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[11], il aurait été impossible à une entreprise d'engager les fonds pour fabriquer un processeur dédié 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 calcul des premiers modèles de reconnaissance d'image, comme AlexNet de Google en 2012. Ce réseau de neurones est le premier modèle à atteindre un taux record d'erreur de 16%[12], gagner la compétition d'ImageNet[13] et prouver que les méthodes d'apprentissage profond sont aujourd'hui réalisables. [14]

[9] PureStorage : Qu'est-ce qu'une unité de traitement du langage (LPU) ?
[10] Youtube : "Compute is the New Oil", Leaving Google, Founding Groq, Agents, Bias/Control (Jonathan Ross) 3 : 16
[11] The Chip Letter : Google's First Tensor Processing Unit: Origins
[12] Pine cone : AlexNet and ImageNet: The Birth of Deep Learning
[13] Data France : « 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 »
[14] Arxiv : Building High-level Features Using Large Scale Unsupervised Learning

Après ces premiers pas éclatant devant le grand public et un apprentissage de 3 jours sur 1.000 machines comptent en tout 16.000 cœurs. Il était devenu nécessaire pour Google de reconsidérer l'option de créer un processeur ASCI spécialisé dans les calculs 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 [15] et les besoins en puissance de calcul ne font qu'augmenter.

4 ans plus tard, le modèle AlphaGo[16] défie le monde en devenant le meilleur joueur de Go au monde. Le secret de cette réussite est une nouvelle puce nommée TPU[17] pour Tensor Processing Unit avec laquelle Google entraîne son modèle depuis au moins 1an.

8 ans se sont écoulés 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[18], tandis que Google continue de commercialiser des variantes de ces processeur comme le Google Tensor G4 [19] 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[20] en 2021. Son créateur, un des ingénieurs ayant participé à l'élaboration de la puce TPU de Google entend défier le leader mondial Nvidia.

L'originalité de Groq réside dans une nouvelle puce ASCI nommée LPU pour Language Process Units. Cette puce spécialement conçue pour des calculs d'inférence est dors est déjà en service pour fournir une interface de chat [21] ainsi qu'une une API

[15] Youtube : Android App Google Speech Recognition Tutorial and Example
[16] Discovery : Mastering the Game of Go without Human Knowledge
[17] Google Cloud Blog : Google supercharges machine learning tasks with TPU custom chip
[18] 01net : Pour ChatGPT, Microsoft a dépensé des millions de dollars… en GPU Nvidia
[19] LeMondeNumérique : Le prochain processeur Google Tensor G4
[20] Wood side Cap : AI Semiconductor Market Q4 2022
[21] Groq : LLM chat

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[22], une entreprise de « voiture autonome » et une autre de « fintech ».

Aujourd'hui le marché des processeurs dédiés à l'IA a explosé, il est nommé « AI Semiconductor market » et pèse cette année USD 30 Billion. Le rapport de Wood side Cap pointe les 192 start-up qui ont profité de ces investissements 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 entraîner ou utiliser des modèles d'IA. Ce secteur étant en plein essor depuis les deux dernières années, il entraîne avec lui le marché des semiconducteurs dédié à l'IA qui se voit obligé de grandir pour répondre à la demande.

La société McKinsey & Company [23]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

[22] Cambrian – AI : Interview with CEO Jonathan Ross of Groq 4 : 53
[23] Mc Kinsey : The surge of interest in and use of generative AI

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[24].

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[25].

Présentation de la concurrence

Secteur des semiconducteurs

Comme nous l'avons dit précédemment, 192 start-ups dans ce domaine ont été financées par le fonds d'investissement Wood Side Cap. Certaines comme Horizon Robotics travail sur des puces dédiées aux voitures autonomes, une autre comme SambaNova entre en concurrence avec Groq mais uniquement dans le domaine du B2B.

Cet exemple nous montre que d'autres solutions apportant une API, un Data Center et une puce ensemble réunie sont susceptibles 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 côté 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 utilisée dans les Data Center de Microsoft[26], sa puissance équivaut à 567 puces LPU et tous les acteurs se l'arrachent. Bien que les processeurs

[24] Fortune Buisness Insigghts : Cloud AI Market Size

[25] Edge AI Vision : AI Chip Market to Grow 10x in the Next Ten Years and Become a $300 Billion Industry
[26] Learn Microsoft : Série ND H100 v5

LPU soient pensés pour être utilisés en série et non individuellement, cette puce est de loin la plus aboutie sur le marché.

En effet, la revue de presse semianalysis spécialisée dans les semiconducteurs compare les deux puces et révèle une meilleure rentabilité à grande échelle pour les puces H100. Pour l'instant il sera plus rentable de monter un Data Center équipé de puces H100, mais faut-il encore se les procurer.

La seule entreprise au monde pouvant graver des puces de silicium en 4nm étant TSMC, la chaîne de production est sous forte tension chez le leader mondial. C'est ici que Groq devient intéressant, la puce LPU est gravé en 14nm ce qui lui permet d'être entièrement construite au USA. A l'inverse, la puce de Nvidia ne peut être construite que par TSMC à Taiwan.

Benchmark : cloud computing

Nous avons principalement parlé 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 offrent la possibilité de requêter un modèle d'IA entraîné via une API et dont les prix sont calculés en fonction du nombre de Tokens donné / reçu.

Les premiers tests sont frappants [27]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

[27] Artificial Analysis : Mixtral 8x7B Instruct: API Provider Benchmarking & Analysis

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 tient le marché par les deux bouts en se plaçant comme leader des coûts et des performances.

Conclusion

En résumé, l'introduction de la technologie LPU de Groq dans le marché des processeurs dédiés à l'IA est un événement majeur qui pourrait révolutionner l'industrie. Avec une puissance de calcul équivalente à 567 puces LPU, la puce H100 d'Nvidia est actuellement la plus aboutie sur le marché, mais la puce LPU de Groq présente des avantages incontestables en termes de coût et de rapidité. Les résultats des benchmarks sont encourageants, avec Groq qui offre la meilleure vitesse de réponse et un coût compétitif.

Cependant, il est important de noter que la concurrence est forte dans ce secteur, avec de nombreuses start-ups et géants technologiques qui investissent dans la recherche et le développement de nouvelles technologies. Il est donc important pour Groq de continuer à innover et à améliorer sa technologie pour rester compétitif.

En fin de compte, l'objectif de Groq de "faire de l'IA gratuite" en fournissant une puissance de calcul élevée à un coût raisonnable est ambitieux, mais il est possible que la technologie LPU soit la clé pour atteindre cet objectif. Avec son potentiel de croissance et ses avantages compétitifs, Groq est bien placé pour devenir un leader dans le marché des processeurs dédiés à l'IA.

Sources

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 : « Investir dans l'intelligence artificielle, c'est investir dans la croissance »

Community FS : Introduction aux CPU, GPU, ASIC et FPGA

Bitproid : ASIC vs GPU Profitability

PureStorage : Qu'est-ce qu'une unité de traitement du langage (LPU) ?

The Chip Letter : Google's First Tensor Processing Unit: Origins

Pine cone : AlexNet and ImageNet: The Birth of Deep Learning

Google Cloud Blog : Google supercharges machine learning tasks with TPU custom chip

LeMondeNumérique : Le prochain processeur Google Tensor G4

Wood side Cap : AI Semiconductor Market Q4 2022

Mc Kinsey : The surge of interest in and use of generative AI

Fortune Buisness Insigghts : Cloud AI Market Size

Edge AI Vision : AI Chip Market to Grow 10x in the Next Ten Years and Become a $300 Billion Industry

Les numériques : Groq, la nouvelle bête noire de Nvidia ?

Le monde informatique : Groq défie Nvidia avec ses accélérateurs LPU

L'Usine nouvelle : Qui est Groq, cette start-up californienne qui veut défier Nvidia dans les puces d'IA ?

L'Usine nouvelle : Les résultats de Nvidia s'envolent sous l'effet de la bulle IA… mais cela peut-il durer ?

01net : Pour ChatGPT, Microsoft a dépensé des millions de dollars… en GPU Nvidia

Semianalysis : Groq Inference Tokenomics: Speed, But At What Cost ?

Semianalysis : Inference Race To The Bottom - Make It Up On Volume ?

Dell : GPU, NPU, FPGA, ASIC, qui est qui et qui fait quoi ?

Mckinsey : Generative AI: The next S-curve for the semiconductor industry ?

Snsinsider : Cloud AI Market Report Scope & Overview

Documentation

Learn Microsoft : Série ND H100 v5

Cloud Google : Accélérez le développement de l'IA avec les TPU Google Cloud

Publications scientifiques

Arxiv : Building High-level Features Using Large Scale Unsupervised Learning

Arxiv : Energy efficiency in Edge TPU vs. embedded GPU for computer-aided medical imaging segmentation and classification

Discovery : Mastering the Game of Go without Human Knowledge

Bilan d’investissement

Wood Side Cap : AI Semiconductor Market Q4 2022

Benchmark

Artificial Analysis : **Mixtral 8x7B Instruct: Mixtral 8x7B Instruct: API Provider Benchmarking & Analysis**

Ressources video

Cambrian AI : Interview with CEO Jonathan Ross of Groq

Youtube : CPU vs GPU vs ASIC mining

Youtube : "Compute is the New Oil", Leaving Google, Founding Groq, Agents, Bias/Control (Jonathan Ross)

Youtube : Android App Google Speech Recognition Tutorial and Example