NVIDIA's CORNERSTONE for UNMATCHED DOMINANCE in AI (CUDA)

*"**We Have No Moat, And Neither Does OpenAI**" proclaimed a leaked paper circulated internally at Google*[1]. We believe that this is true for most AI businesses. However, one company does have a moat, a quite large moat, and that company is Nvidia.

Nvidia has been the ultimate "pick and shovel" investment - playing a key role, dating back to 2012, in enabling the application of Artificial Intelligence (AI) technologies. This strategy has catapulted it from a $30bn gaming hardware company to a big tech juggernaut now valued at more than $1T. This article will not only show how Nvidia turned CUDA into their AI moat, but also show why PyTorch and Triton are not competitors. It will also detail CUDA's true competitor (AMD's ROCm) and show why it is vastly inferior.

CUDA

The enabling force behind Nvidia's rise is a software library named Compute Unified Device Architecture, or CUDA. CUDA was originally released in 2007 as a software development kit allowing for general-purpose computing on Graphical Processing Units (GPUs), or a General-Purpose Graphical Processing Unit (GPGPU). Nvidia made this software free to use - enabling any engineer who wished to take advantage of the speed with which a GPU could perform parallelized floating point

[1] Patel, "Google 'We Have No Moat, And Neither Does OpenAI.'"

math operations to do so. Early uses of GPGPU processing were for cryptography, sorting algorithms, and building physics simulators similar to those used in virtual gaming worlds. Founder Jensen Huang and Nvidia hoped that by enabling the use of GPUs for accelerated computing tasks, they would sell more GPU cards. Today, we are witnessing the manifestation of this vision.

The catalyzing event that ushered in this new age of AI and cemented Nvidia as one of its dominant players occurred in 2012. In that year a computer vision model named AlexNet won the annual ImageNet competition. This competition consisted of using computers to identify objects in images from a set of 12 million pictures spanning 22,000 different objects.

At the time, the best algorithms programmed on CPUs could do no better than 74% accuracy, with most doing far worse. AlexNet, an algorithm trained with CUDA on GPUs, achieved 85% accuracy, trouncing the competition in 2012[2]. AlexNet used a Convolutional Neural Network (CNN) architecture, which at the time was thought to be impractical due to the amount of computation needed to train it. **AlexNet was Nvidia's coming out party**.

The following year every competitor was using a CNN. And every CNN was using CUDA on a Nvidia GPU[3]. The software development world took notice, building nearly all of the subsequent machine learning libraries upon CUDA.

[2] "AlexNet and ImageNet."

[3] "ImageNet Large Scale Visual Recognition Competition 2013 (ILSVRC2013)."

PyTorch & Triton: Challengers to the CUDA Moat?

Two machine learning libraries, PyTorch and Triton, that have achieved broad industry adoption, have been touted by popular media outlets as proof that Nvidia's CUDA moat is eroding[4]. However, due to the technical nature and some easily glossed over details, we believe these outlets are over emphasizing the threat to Nvidia.

PyTorch is a machine learning software framework originally created by Meta (Facebook) in 2016. Software engineers write code in either Python or C++ (traditional programming languages), which both use PyTorch to leverage the built-in machine learning tools. Meta released PyTorch as open-source, meaning it is freely available for anyone to use, contribute to, or modify for a specific use. However, as alluded to above, **the numeric operations that PyTorch performs are accelerated by CUDA, on Nvidia GPUs**.

PyTorch has since seen wide adoption by the industry - from small-scale hobbyist projects to large-scale projects like the Tesla Autopilot system. Members of the PyTorch community had the reasonable desire to remove vendor lock-in and provide the option to utilize PyTorch on AMD GPUs as well as CPUs. Adding this capability required a major revision to the framework, which resulted in PyTorch 2.0[5] (released March 15th, 2023) ultimately enabling AI applications to be compiled for use on either AMD or Nvidia hardware. However, it is not so simple as to 'switch' from Nvidia to AMD GPUs with pre-existing code. This is because CUDA and ROCm do not have the same functionality. Therefore a transition from Nvidia to AMD may require re-architecting

[4] "Josh Wolfe on Where Investors Will Make Money in AI."

[5] "PyTorch 2.0."

the codebase and potentially eliminating functionality. The process by which PyTorch can execute code on an AMD GPU will be discussed in more detail below.

Similarly, Triton is a programming language developed by OpenAI that obfuscates CUDA so that an engineer does not have to program with CUDA directly. A common misconception is to call Triton a competitor to CUDA. This misunderstanding stems from the conflation of the fact that just because a Triton programmer does not write CUDA code, it does not also mean that Triton does not use CUDA. In fact, Triton, in its current form, only executes on Nvidia GPUs (via CUDA). That said, the OpenAI team has left the door open to support other hardware configurations, not just AMD GPUs but also CPUs (listed as "under development")[6].

> *We're releasing Triton 1.0, an open-source Python-like programming language which enables researchers with no CUDA experience to write highly efficient GPU code—most of the time on par with what an expert would be able to produce.*
>
> *- OpenAI Website*

In the two years subsequent to Triton's release, OpenAI has not implemented the support for AMD GPUs or CPUs. The general belief, potentially arising from the idea that Triton is a CUDA competitor, is that OpenAI's intention was to remove CUDA from the software stack. **We find this view naive, and lacking technical understanding**. A more likely explanation is that finding a workforce of engineers experienced programming to the CUDA interface was more difficult than developing a software abstraction, removing the need for engineers to directly use, or learn CUDA.

[6] "Triton."

AMD's ROCm

Meta and OpenAI are not natural competitors with Nvidia; rather, they are Nvidia's customers. As large customers, they reasonably don't like having a sole supplier of a critical component of their businesses. They would prefer to have AMD as well as Nvidia GPUs available for AI workloads. The Radeon Open Compute platform (ROCm), first released in 2016 (nearly a decade after CUDA), is AMD's attempt to make that possible and ultimately compete with Nvidia in this rapidly growing market.

ROCm is directly analogous to CUDA, it is a software stack for programming AMD GPGPU processors. As with CUDA, PyTorch 2.0 uses ROCm when an engineer wishes to develop their software for an AMD processor. While there is much to be said for learning from a prior example, ROCm has been playing catchup ever since it was launched. Nvidia recognizes that CUDA is their most important moat and they spend lavishly to maintain that advantage. This is supported by the fact that Nvidia now employs more software engineers than hardware engineers[7]. ROCm has always been a step or two behind as far as introducing new capabilities, supporting its entire line of GPUs[8], or even running on Windows[9]. Additionally, the supporting infrastructure desired for a software library just isn't there for ROCm. For example, the software documentation, tutorials, and support is nowhere near what is offered by the CUDA community. Even some of the best engineers have had a difficult time finding a list of which AMD GPUs even properly run ROCm[10].

---

[7] "NVIDIA Corporation Presents at Citi 2019 Global Technology Conference."

[8] Mujtaba, "AMD CEO Teases ROCm Support Coming To Radeon Consumer GPUs Soon."

[9] Ansari, "Why NVIDIA Keeps Winning.... And AMD Does Not."

[10] *This AMD Documentation Is Basically Unusable*.

In Summary...

For the non-programmers reading this, the following analogy is more relatable (at least to the traditional finance crowd). **Just as you would build a financial model in Excel, you would build your new, clever AI chatbot using PyTorch (or Triton)**. **PyTorch relies on either CUDA or ROCm and similarly, Excel relies on an operating system, either Windows or MacOS. The catch, as there always is, is that PyTorch on ROCm is limited in just the same way that Excel on MacOS is annoyingly inferior to Excel on Windows.**

With how fast artificial intelligence applications are being developed, and certain aspects of it have been categorized as a "race", no one other than those running the largest stable AI workloads have felt enough pain to go through the effort of migrating away from an Nvidia GPU architecture. It's probable that one day, these two GPGPU software interfaces will reach parity. An investor in any of these companies will need to pay attention to this topic specifically, looking for signs that Nvidia's moat is starting to erode. However, as of today, the CUDA moat shows no signs of shrinking and its biggest threat is the lack of availability of Nvidia H100s.

IMPORTANT DISCLOSURES. The information herein is provided by Top Mark Capital Management LLC ("Top Mark Capital") and: (a) is for general, informational purposes only; (b) is not tailored to the specific investment needs of any specific person or entity; and (c) should not be construed as investment advice. Top Mark Capital makes no representation with respect to the accuracy, completeness or timeliness of the information herein. Top Mark Capital assumes no obligation to update or revise such information. In addition, certain information herein has been provided by and/or is based on third party sources, and, although Top Mark Capital believes this information to be reliable, Top Mark Capital has not independently verified such information and is not responsible for third-party errors. You should not assume that any investment discussed herein will be profitable or that any investment decisions in the future will be profitable. Investing in securities involves risk, including the possible loss of principal.

"AlexNet and ImageNet: The Birth of Deep Learning | Pinecone." Accessed July 31, 2023. https://www.pinecone.io/learn/series/[object%20Object]/imagenet/.

Ansari, Tasmia. "Why NVIDIA Keeps Winning.... And AMD Does Not." Analytics India Magazine, October 12, 2022. https://analyticsindiamag.com/why-nvidia-keeps-winning-and-amd-does-not/.

Bloomberg.com. "Josh Wolfe on Where Investors Will Make Money in AI." July 17, 2023. https://www.bloomberg.com/news/articles/2023-07-17/josh-wolfe-of-lux-capital-on-investing-on-ai-and-computing.

"ImageNet Large Scale Visual Recognition Competition 2013 (ILSVRC2013)." Accessed July 31, 2023. https://image-net.org/challenges/LSVRC/2013/results.php.

Mujtaba, Hassan. "AMD CEO Teases ROCm Support Coming To Radeon Consumer GPUs Soon." Wccftech, June 19, 2023. https://wccftech.com/amd-ceo-teases-rocm-support-coming-to-radeon-consumer-gpus-soon/.

"NVIDIA Corporation Presents at Citi 2019 Global Technology Conference," September 5, 2019.

Patel, Dylan. "Google 'We Have No Moat, And Neither Does OpenAI,'" January 16, 2023. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

"PyTorch 2.0." Accessed July 31, 2023. https://pytorch.org/get-started/pytorch-2.0/.

This AMD Documentation Is Basically Unusable, 2023. https://www.youtube.com/watch?v=Zsh6lPqvAcw.

"Triton." C++. 2014. Reprint, OpenAI, July 31, 2023. https://github.com/openai/triton.