UNTETHER AI – WHERE PERFORMANCE INTERSECTS WITH EFFICIENCY

SITUATION ANALYSIS

AI is the most disruptive and transformational technology trend we have witnessed. And its unique requirements impact virtually every element of the technology stack – from silicon to compute platforms to software.

While most of the focus has been on the training aspect of AI – the shaping and modeling of datasets – inference, or what to do with trained data, is where the value of this workload is realized. The platforms that perform inference will be broadly deployed because inference happens everywhere and requires real-time responsiveness.

As such, long-term investments in inference will dwarf those in training. Not only do organizations have to acquire more infrastructure, but they need to acquire *different* infrastructure, as the computational needs of inference differ from those of training.

This research brief will explore the AI market, focusing on inference as it emerges as a focus workload for enterprises. Further, this paper will detail the unique needs of this workload and how companies like Untether AI will play a significant role in the market.

BREAKING DOWN AI – TRAINING IS TEMPORARY, INFERENCE LASTS FOREVER

Anybody who has even casually followed the AI market has been enamored by companies such as NVIDIA and AMD, which have held center stage in the hype limelight with GPU technology aimed squarely at ML training. With silicon and instruction sets dedicated to massively parallel processing, matrix multiplications, gradient calculations, and optimization algorithms, these products (and their extreme price tags) seem like science fiction creations.

Undoubtedly, training of large language models has been a topic of virtually every conversation Moor Insights & Strategy (MI&S) analysts have had with enterprise business and technology leaders. Model training and tuning have evolved from science fiction topics to business-critical functions seemingly overnight. But this collective rush to AI has led to an uneven understanding and adoption of actual technology. While

training has been the primary focus of many AI initiatives, inferencing is emerging as a new focal point.

While training will continue to garner the market's attention, MI&S sees inference as entering the AI conversation with considerable emphasis. The nexus between AI and the real world, inference deployments will be initially broad and continue to span the market over time.

Training and inference are two functions of the ML equation that, in many ways, contrast. Whereas training is a workload with a seemingly insatiable thirst for compute at any cost, inference is a workload that is somewhat more lightweight in terms of computational resources but highly dependent on responsiveness. Further, while training is performed in a central location, inference happens anywhere and everywhere – all the time.

Because of the ubiquitous nature of inference, MI&S sees a significant revenue growth opportunity for technology vendors in this market.

THE UNIQUE REQUIREMENTS OF INFERENCE

Because of data regulation and privacy concerns, MI&S believes that most generative AI workload deployments will run on-premises (including the edge) or in a private cloud setting. This translates into an inference solution that is both performant and flexible.

Further, the real-time nature of inference requires the lowest possible latency. Consider the following examples:

- In **advanced driver-assistance systems** (ADAS), inferencing takes data from dozens of sensors and cameras to deliver warnings and assistive measures to drivers in real time. Consider a collision avoidance system. With this system, radar or lidar is used to detect objects in front of a vehicle. Sensors capture object data and in real time track, assess, and decide how to respond. All these steps, powered by AI, need to be accomplished in microseconds.

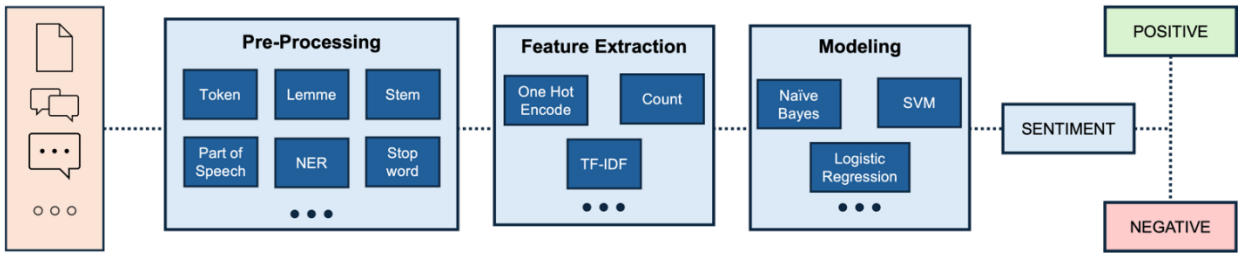
- Similarly, in **agricultural technology**, precision farming in the form of pest and weed abatement can be accomplished using AI-powered object detection, classification, and eradication, in real time.

- Finally, consider real-time **sentiment analysis**. Call centers receive hundreds of calls every hour. As conversations unfold, AI is used to convert voice to text. This

text feeds a natural language processing (NLP) model that is used to discover and gauge customer sentiment to alert service representatives to potential changes in the customers mood. This, in turn, enables the representative to de-escalate mounting tensions and increase customer satisfaction.

These three examples of real-time inference are all unique in how data is collected and where it is processed. However, in each case, the need for real-time inference is critical. Further, the range of deployments outlined above demonstrates the need for inference capabilities that span the traditional datacenter to the edge to even a disconnected state.

FIGURE 1: INFERENCE FOR SENTIMENT ANALYSIS



Real time performance in sentiment analysis

Source: MDPI

These systems must also operate with a power budget that is an order of magnitude less than the power budgets for AI training. Large GPUs that consume upward of a kilowatt of power cannot be deployed in a vehicle. Because most inference occurs on the edge, space, power, and cooling constraints must be considered. As a result, the technology used to drive inference must be capable of deployment in any space with the most minimal power budget. Yet, it must deliver results that exceed those of the most performant GPUs.

Because of these requirements, MI&S sees custom designed, energy-efficient AI inference accelerators as the ideal technology for delivering this capability.

PURPOSE BUILT SHOULD BE PURPOSE DESIGNED

Datacenter infrastructure has largely been unchanged for decades. Servers that run the most critical workloads and applications are designed around the von Neumann architecture – a model or framework conceived in the 1940s that lays out the relationship between a CPU, memory, and I/O devices. A control unit coordinates the

operations, and this nearly 80-year-old architecture also includes a set of instructions that the CPU can execute – an Instruction Set Architecture (ISA).

The von Neumann architecture has held up quite well over time, as evidenced by its utility in modern designs. However, the emergence of AI has led to a silicon ecosystem that must be rethought to handle this workload's computational requirements most efficiently.

As previously mentioned, ML training efficiency has largely been addressed with accelerators designed to handle complex functions efficiently through parallelism and distributed computing. GPUs and ASICs residing in a server (or cluster of servers) enable matrix operations to perform far faster and more efficiently than CPU-based learning.

With inference, the challenge is different. Performance and efficiency require a unique architecture. The effectiveness of an inference solution is related to delivering the fastest (accurate) result while consuming the least amount of power.

Current inference solutions (GPU, CPU) utilizing the von Neumann architecture are insufficient for two reasons. First, as data travels over I/O channels, a significant latency is introduced into the inference equation. In some situations, this latency can be acceptable. However, this latency can literally be the difference between life and death for real-time inference.

The second challenge with utilizing traditional von Neumann-based architectures for inference is power. The most significant consumer of power in inference is rooted in data movement. Estimates from startup Untether AI assign 90% of power consumption to this task of data transfer.

Finally, when considering training versus inference, cost must be accounted for. Whereas training occurs in a central location with extremely powerful and expensive infrastructure, inference is ubiquitous. Because inference happens everywhere at any time, the infrastructure and chips supporting this workload must be economically viable.

Because of these challenges, it should be evident that simply repurposing infrastructure for inference is unwise. High-powered servers and GPUs that require extreme power and don't account for latency are inefficient at best.

MI&S forecasts inference deployments will accelerate to roughly 70% of the AI market. This dynamic means deploying at scale must be a well-thought-out exercise. MI&S

believes the key to fast and efficient inference is in purpose-built accelerators that solve the latency and power challenge. When considering an inference accelerator to deploy widely, organizations should look for a vendor that fully understands these unique requirements and demonstrates this understanding through a portfolio of purpose-designed and purpose-built products.

In a market full of startups, MI&S has found Untether AI to be an inference acceleration vendor worthy of serious consideration.

UNTETHER AI DELIVERS INFERENCE DESIGNED FOR PERFORMANCE, OPENNESS, AND EFFICIENCY

The tech market is full of startups comprised of people with great ideas – great ideas that sometimes turn into products that solve real problems and sometimes fail to translate into anything tangible.

Those startups that deliver real value to the market tend to be organizations that understand how technology translates into a product that solves a business need. This may sound somewhat obvious; however, many startups come and go because of their inability to shape a technology concept into a product or service.

Further, it is best to start with the organization when evaluating startups and their products. Do the people behind the product – the engineers and leadership – have the industry experience and knowledge to inform a fully vetted strategy and solution? While this factor may weigh more or less heavily depending on the service, MI&S believes real-world experience is critical when it comes to silicon design.

Because of these factors, MI&S believes Untether AI stands out in inference acceleration.

Untether AI was founded in 2018, long before the AI craze hit the enterprise IT market. Its leadership, comprised of silicon veterans, saw this need for low-power, low-latency silicon to deliver the real-time inference required across workloads such as FinTech, ADAS, aerospace, and video.

The driving force behind the company's portfolio of technology seems to be the themes discussed previously in this research brief – faster inference at a lower power threshold and a drastically lower cost. It achieves this through the design of inference task-

specific ASICs that can be inserted in industrial systems or packaged for use as a PCIe device in a commercial server or edge device form factor.

The key to Untether AI's performance and efficiency is its implementation of at-memory architecture. Breaking from the legacy of von Neumann removes the bottleneck associated with latency and power. By moving compute to where data resides (at-memory), the company claims incredible reduction (up to 6x) in latency and power consumption.

The company's accelerator cards are equally adept at the edge and the datacenter, with power profiles between 75W and 300W, depending on the configuration.

These cards host the AI inference accelerators (as previously mentioned) and can also be placed directly on a server or system's motherboard. The company's first generation of products, based on the runAI200 devices, proved its at-memory architecture delivers significant performance and energy benefits for AI acceleration. Its second generation speedAI family of devices and PCIe cards, with – astonishingly – greater than 1,400 RISC-V cores at 2 petaFLOPS of performance, will be available this year.

The speedAI family has an affinity for generative AI applications as well as convolutional neural networks (CNNs) and its emerging alternative for image recognition and classification, vision transformers (ViT). This is important to call out as it demonstrates the Untether engineering team's awareness in designing an open architecture that has wide applicability across the variety of use cases.

These accelerators are only useful if there is openness and support for the frameworks deployed by customers. Untether AI's imAIgine SDK enables customers to import models from the most popular frameworks (e.g., TensorFlow, PyTorch, ONNX) to inference on its silicon with great simplicity. The three elements of the imAIgine SDK – the imAIgine Compiler, Toolkit, and Runtime – account for the full development cycle, enabling frictionless AI production.

This combination of vision, experience, and proven delivery is why MI&S is so bullish about the prospects of Untether AI.

SUMMARY

While AI has been used for decades, its applicability in the enterprise market has recently become pervasive. Organizations of all sizes and types want to deploy

generative AI to solve their most pressing needs, develop strategies, and execute financial transactions.

With training dominating the AI discussion and (much-deserved) hype, inference is the long tail that will drive AI in the future. Inference happens everywhere – in the datacenter, on the edge, and in every device. While the required responsiveness of inference may vary, real-time responsiveness measured in microseconds for use cases such as ADAS, financial trading, fraud detection, and defense systems is an absolute requirement.

It is not an exaggeration to say that inference is perhaps a company's most critical investment. And, as such, this investment should be made with much thought and consideration. Latency, power, openness, and cost are critical elements of the inference market as they address performance, deployability, useability, and scale. These capabilities cannot be fully realized unless accounted for at the lowest levels of design.

Inference is a workload that requires purpose-designed and purpose-built silicon that breaks from von Neumann's tradition and addresses the architectural challenges of data movement and data processing.

The hype surrounding AI has ushered in a gold rush of sorts. The technology industry is filled with established vendors reorienting product strategies and messaging and a seemingly infinite number of startups that promise to deliver silicon, systems, and software that will revolutionize the industry (or some similar catchphrase). However, many of these companies are selling a vision and hope rather than a proven portfolio of products and IP that deliver real value.

Untether AI is different. This is a company with a portfolio worthy of serious consideration as organizations begin their inference journey. Its approach to designing accelerators that support the range of uses and deployments that require inference acceleration means real-world relevance today – not sometime in the future. And it can demonstrate this relevance through a list of customers that are using its products in the most demanding environments.

Further, the team that designed this technology portfolio brings considerable industry experience in the silicon, software, and AI markets. This factor instills confidence in the real-world value of the company's technology. Untether AI brings more than just a vision to market: Its technology helps organizations realize the benefits of AI. Because of this,

MI&S sees Untether AI as a company that should be on the short list of consideration for inference – whether in the enterprise or an industrial setting.

For more information on Untether AI and its portfolio, please visit www.untether.ai.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Matt Kimball, Vice President and Principal Analyst, Servers

PUBLISHER
Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts but must be mentioned in the context, displaying the author's name, title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission from Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
Untether AI commissioned this paper. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties regarding such information's accuracy, completeness, or adequacy and shall have no liability for errors, omissions, or inadequacies. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators, not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on the future, they are subject to risks and uncertainties that could materially cause actual results to differ. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of this document's publication date. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.