MOFFETT AI

Introduction to Sparse Computing Whitepaper

MOFFET ERC-20

MOFFETT AI
MAY 27th
2022

Copyright Notice

Contents

How great things are made

Someone once asked Michelangelo how he carved the statue of David. “Was it difficult?”, they asked. “No,” he replied, “It was easy. I just chipped away the stone that didn’t look like David.”

When we think of building new things, we tend to think about adding materials together to create something larger and more complex. But in the case of David, the creative process was one of gradual subtraction, not addition.

There is a field of science built on applying this principle to how data is processed, called Sparse computing[1]. The concept originated in the US in the 1950s[2] and is now led by large companies like NVIDIA and startups like Moffett AI. Sparse computing has become an important factor in modern chip design.

Sparsity is a smart way to process data. It is vital in applications where there are large amounts of data which need to be processed quickly, accurately and at affordable prices, consuming the least amount of energy.

Sparse computing makes it easier to process large quantities of data by removing the bits of data - and it turns out that there are always many of them - which don’t count. In summary, this vast field of Sparse computing is dedicated simply to how to remove data which is not useful so that the rest can be more easily processed.

The technique is becoming particularly important in the field of artificial intelligence. AI models, which are being used to model climate change, fold proteins, sequence DNA, drive Tesla vehicles, and translate speech and languages. They have grown so large that they are hard to run on normal computers. The largest ones, called “foundation models”, can cost millions of dollars in electricity just to develop, let alone operate. Some of them are so large that they have to be kept on specialized hardware. But by using sparsity to gradually remove unnecessary parameters, researchers are finding ways to make these models hundreds of times smaller – and thus usable at much more cost-effective manners for commercial applications.

[1] We capitalize the initial letter “S” in sparse to define the adjective. We use “Sparcity” as the noun, but “sparse” as the verb.

[2] The paper “Portfolio Selection” by Harry Markowitz of The Rand Corporation, was published in The Journal of Finance in March 1952

The four great computing problems of our age

We believe Sparse computing has an incredibly important role to play outside the laboratory, in the the real world.

Humanity today faces challenges from climate change and global health to wellbeing, safety, education and more. The United Nations Sustainable Development Goals[3] provide a good summary of the main areas of focus and identify 17 major areas where work is required for humanity to address these challenges.

Computing plays a very big role in solving these challenges. At one end, modelling climate change and sequencing DNA gets better the more data there is, and so computers are processing larger and larger quantities of data in order to develop better insights. The more data they can process, the more accurate they get.

This data needs to be processed quickly. If you can process billions of images and data points in situation where time is critical, you can save lives.

At the other end, there are the vast server farms which house these computers, consuming more and more electricity to the extent that they are responsible for at least a few percentage points of the world's total energy consumption and a few percentage points of the world's total greenhouse gas emissions.

Data centers are a vital part of the solution, but also part of the problem: as they process more data, yet they so they consume more power.

[3] The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership. They recognize that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth – all while tackling climate change and working to preserve our oceans and forests. https://sdgs.un.org/goals

That leaves four very clear problems to solve:

1. Increase speed - we need to be able to process more data than ever before in less time and decrease latency.
2. No loss of accuracy - we require computers which can process data with high precision.
3. Decrease cost - we can't simply add more and more processors and electricity, we need to make computing more affordable and reduce the total cost of ownership (TCO) from both capital and operation.
4. Decrease energy consumption - at current levels, our energy consumption is unsustainable and we need to do reduce it as much as possible.

Sparse computing helps all these four things. The more urgency we see in addressing the four compute problems, the more of a role there is for Sparse computing as a part of the solution.

The processor brick wall

The solution to most problems like this used to be to get smarter in how processor chips were designed and developed, specifically how many transistors could be engineered into a single chip.

In 1965, it became possible to create chips with 50 transistors, and it was predicted that this could grow to 1,000 by 1970. Intel made the prediction true with its first 2,300-transistor processor, but the number has since grown to 80 billion.

During this exponential growth, the number doubled approximately every 18 months, and the phenomenon became known as "Moore's Law". For over five decades, Moore's Law reliably provided more and more power but the laws of physics are now preventing further exponential growth and Moore's Law will cease to apply.

This "Moore's Wall" has been anticipated for some time. We saw that the CPU took processing capabilities only so far before the GPU (a graphics processing unit) was developed to handle the higher quantities of data which image processing required.

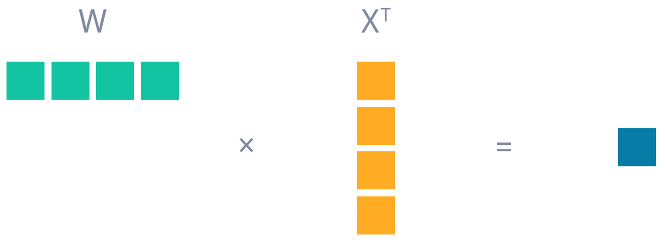
The GPU has triumphed since then. But it, in turn, is now running out of the capacity to keep up with demand.

Scientists and researchers are working with technologies such as Field Programmable Gate Arrays (FPGAs), Quantum Computing, neuromorphic computing, memristors, graphene processors, nanotube transistors, soft machines and system-on-a-chip (SoC) devices to evolve processing performance Sparse computing is one of the possible solutions. Moffett AI, began testing an FPGA Sparse processor in 2019 with public availability on Amazon AWS shortly after. Its Antoum compute platform is a Sparse processor, a system on a chip design which sees public availability in June 2022.

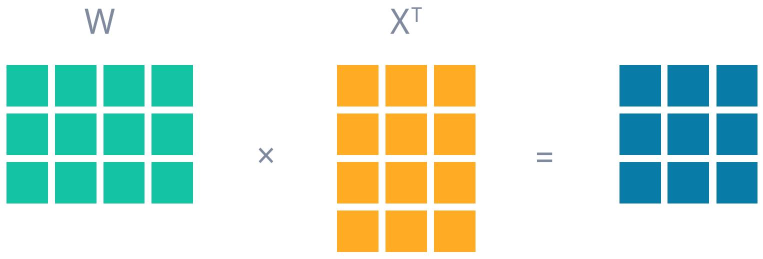
Sparsity in neural networks

Neural networks consist of a set of simulated neurons, analogous to the neurons which make up a biological brain, and a set of weighted connections between them. In these models, the behavior of a single neuron can be written as f(x * w), where w is an ordered list, also called a vector, of N numbers which together represent the synaptic weights of a neuron. They describe the strength of the neuron's connections to N of its neighbors. Meanwhile, x is a vector of equal length which describes the activations of those neighbors.

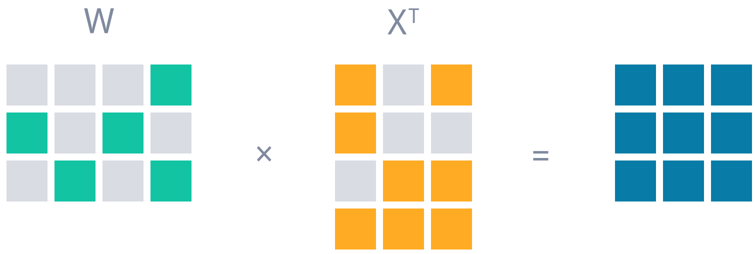
Simulating the behavior of this virtual neuron involves calculating a dot product between the two vectors and then applying the activation function f to the result.



A group of M of these neurons, all with the same number of synaptic connections, could be represented with an M x N matrix where each row is a separate neuron. In the same way, B different activations, each related to a different stimulus, could be stacked atop one another to form a B \times N matrix. Using these groupings, we could compute the responses of a group of M neurons to B different inputs with a single matrix multiplication, W * X^T.



So far, we've assumed that every neuron has a connection to every element in the activation vector. But this is not generally the case, either in biological systems or artificial ones. It's much more common for a neuron to have a few strong synapses and let the rest be zero. Neuron activations tend to be sparse as well. This means that sparse matrices are a natural way to represent both the W and X matrices.

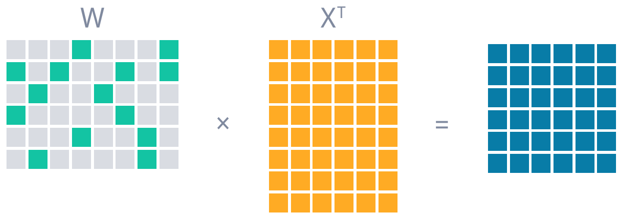


Types of sparsity in neural networks

For early neural network researchers, the importance of sparsity was reinforced when they looked at the human brain. At the time, it was rare for people to stack more than a few layers of neurons atop one another and in these types of "shallow" networks, the total number of synapses would grow quadratically with the number of neurons. But this is not how synapses scale in the brain.

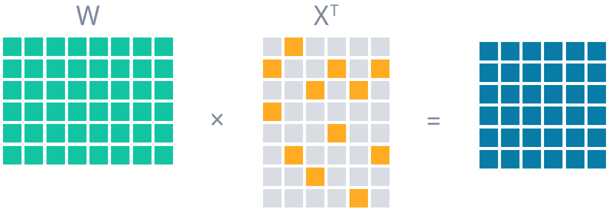
Synaptic sparsity

By some estimates, the human brain has 86 billion neurons and 150 trillion synapses[4]. These numbers imply that only 0.000005%[5] of the possible connections between neurons are actually present. In other words, the connectivity of the brain is 99.999995% sparse. In this regime, the total number of synapses grows linearly with the number of neurons. Each biological neuron gets a fixed number of connections and this number doesn't change even as the total number of neurons increases. Researchers call this property synaptic sparsity.



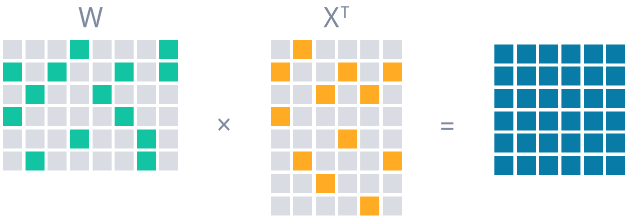
Activation sparsity

The human brain is not only sparse in synapses; it is also sparse in neuron activations. The energy consumed by a biological neuron is roughly proportional to the number of times it fires. So the fewer neurons that fire in the brain, the less energy it consumes. The brain uses this activation sparsity to save energy. By contrast, a simulated neuron as described above consumes the same amount of energy regardless of whether it fires or not. If its output is zero, that zero still gets multiplied with other numbers.



Dual sparsity

These two types of sparsity are complementary to one another. Activation sparsity allows signals to be routed through specific subsets of a network, while synaptic sparsity keeps those subsets small and efficient. Working together, they lead to much greater efficiency gains than would be possible if only one were being used. Researchers suspect that this “dual sparsity” is what permits the brain to be so efficient.



Neural network researchers of the 1990’s were aware of the benefits of sparsity and put a great deal of effort into sparsifying their models, with approaches such as weight magnitude regularization and magnitude pruning to achieve high levels of synaptic sparsity. These works show that sparsity was important even in the early days of AI.

[4] Not all scientists will agree with these numbers but “The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost” authored by Suzana Herculano-Houzel and published in the Proceedings of the National Academy of Sciences of the United States of America, Vol 109 in June 2012, is a good reference.

[5] We obtained this number as follows: $$\quad \textrm{sparsity} = \frac{\textrm{observed number of connections}}{\textrm{possible connections btwn N neurons}} = \frac{150 \times 10^{12}}{N*(N 1)/2} = \frac{150 \cdot 10^{12}}{(86\cdot10^9)(86\cdot10^9-1)/2} = 4\cdot10^{-8} $$

Why aren't neural networks sparse yet?

However, sparse neural networks languished through the 2000's and early 2010's. The early 2000s were known as the AI Winter, a time when funding for neural network research dropped precipitously. Sparsity languished in the early 2010's, even though this was a time when there was a huge interest in and funding for AI. During this "AI Spring", as it came to be known, progress in other areas of AI occurred at a dizzying pace while sparsity stagnated.

It is possible that this was because there were so many other fruitful ways to improve models. First of all, there was better data. By the late 2000's, the internet had exploded in size which made it possible for researchers to construct massive datasets from publicly available data. Second, computing infrastructure grew much better. Not only did computers in general improve, but researchers found that they could massively accelerate their models by putting them on GPUs. A third important event was the rise of automatic differentiation (autodiff) frameworks like Theano, TensorFlow, and PyTorch. These frameworks made it easier to design new models, train them on specialized hardware, and run them on large datasets.

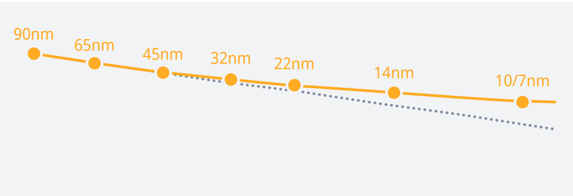
It's worth noting that neither the GPUs nor the autodiff frameworks were built with Sparse computing in mind. And so while they enabled big advances in model size and architecture, they made it very difficult for researchers to reap rewards from sparsity. As long as significant progress was happening in other areas, this was to remain the case. But as the 2010's drew to a close, questions of energy efficiency and the compute-vs-accuracy tradeoff became more pressing and sparsity became much more attractive.

Sparsity becomes more promising

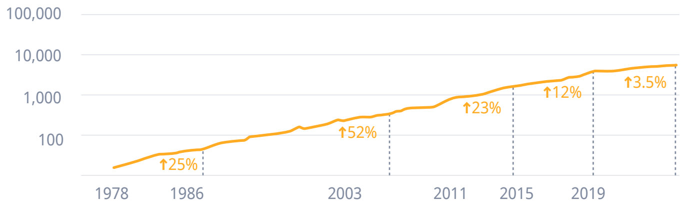
By 2020, AI models and datasets were growing more quickly than the availability of compute power. We can take as an example three famous AI models: ResNet (2014) had 25 million parameters, BERT (2018) had 340 million parameters, and GPT-2 and 3 (2019 and 2020) had 1.5 and 175 billion parameters.

Meanwhile, the rate of improvement in chip technology was beginning to slow. Moore's Law, which had held steady for several decades, was beginning to break down as the sizes of transistors shrank to the limits set by physics. This led to a world in which computing power was increasingly scarce. In this world, the computational benefits of sparsity started to look very attractive[6].

MOORE'S LAW IS NEAR ITS PHYSICSAL LIMITS



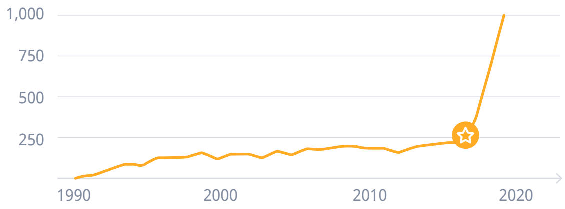
CHIP CAPABILITY INCREASES SLOW DOWN



Steps toward sparsity

One of the early signs that a transition towards sparsity was underway was that academic publications referencing sparsity increased dramatically from 2018 onwards. Then, in 2020, NVIDIA released a chip called the A100 which featured a "sparsity processing unit" (SPU) with a 2x performance boost. In the same year, Google researchers took a first step towards adding sparsity support to Tensor Processing Units[7] ("Sparse-TPU"). Since then, other companies, like Intel and Microsoft, have taken steps in the same direction[8].

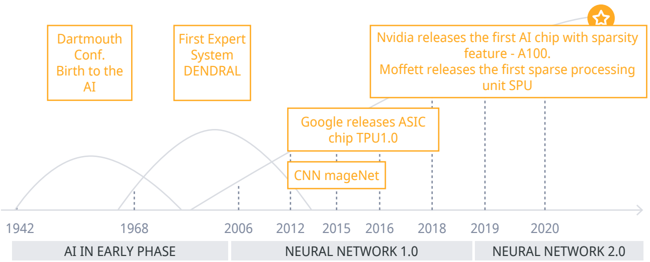
NUMBER OF PUBLICATIONS ON SPARSITY IN ACADEMIA



Sparsity is one of the most popular AI research fields

SPARSITY IN THE INDUSTRY

A milestone in AI 2.0: On May 15, 2020, Nividia launched A100 with 2× Sparsity


Dartmouth Conf. Birth to the AI
First Expert System DENDRAL
Nvidia releases the first AI chip with sparsity feature - A100. Moffett releases the first sparse processing unit SPU
Google releases ASIC chip TPU1.0
CNN mageNet
1942
1968
2006
2012
2015
2016
2018
2019
2020
AI IN EARLY PHASE
NEURAL NETWORK 1.0
NEURAL NETWORK 2.0

The commercialization of sparsity has started, and it will continue to lead the future

[6] "Things that deal with sparse parallelism," said Raja Koduri, Intel's head of chip architecture, "...will give rise to some new architectural ideas that are very different from what we are doing in vector-matrix, which is very mainstream right now." quoted in ZD Net https://www.zdnet.com/article/intel-data-bandwidth-sparsity-are-the-two-biggest-challenges-for-ai-chips in August 2020.

The unrealized potential of sparsity

While these recent developments are moving in the right direction, it is important to put them in context. Researchers have shown that many models can be pruned until they are well over 95% sparse without damaging performance. Naively, this would suggest that such models could be made twenty times smaller and more efficient by adding synaptic sparsity alone. And yet, existing speedups due to model sparsification are only of the order of 2x.

The hardware problem

So far, companies have just made incremental modifications to existing chip architectures and this has not been enough to unlock the order of magnitude gains that are available in theory. One reason that progress has been slow is that sparsity is a difficult hardware problem. Adding full sparsity support means representing matrices and vectors differently on hardware. It means structuring matrix multiplications differently. It means parallelizing computations in different ways. Many experts believe that AI chips need to be rebuilt from the ground up.

Startups

This task, which requires daring and flexibility, is well suited for startups. Indeed, some of the best work being done in this area is happening at small companies. Numenta, a Bay Area startup, recently demonstrated a custom chip with hardware support that runs a popular vision architecture 100 times faster than more traditional chips. Another company, NeuralMagic, offers model sparsification for shrinking foundation models so that they can run on laptop CPUs instead of expensive data center GPUs. But in order to realize the full potential of sparsity, the industry is going to need to design both hardware and software together.
So far, only a few startups have tried to do this. One of the most interesting and ambitious of these companies is Moffett AI.

[7] See “Sparse-TPU: adapting systolic arrays for sparse matrices” authored by Xin He and others in ICS ‘20: Proceedings of the 34th ACM International Conference on Supercomputing, June 2020 https://dl.acm.org/doi/10.1145/3392717.3392751

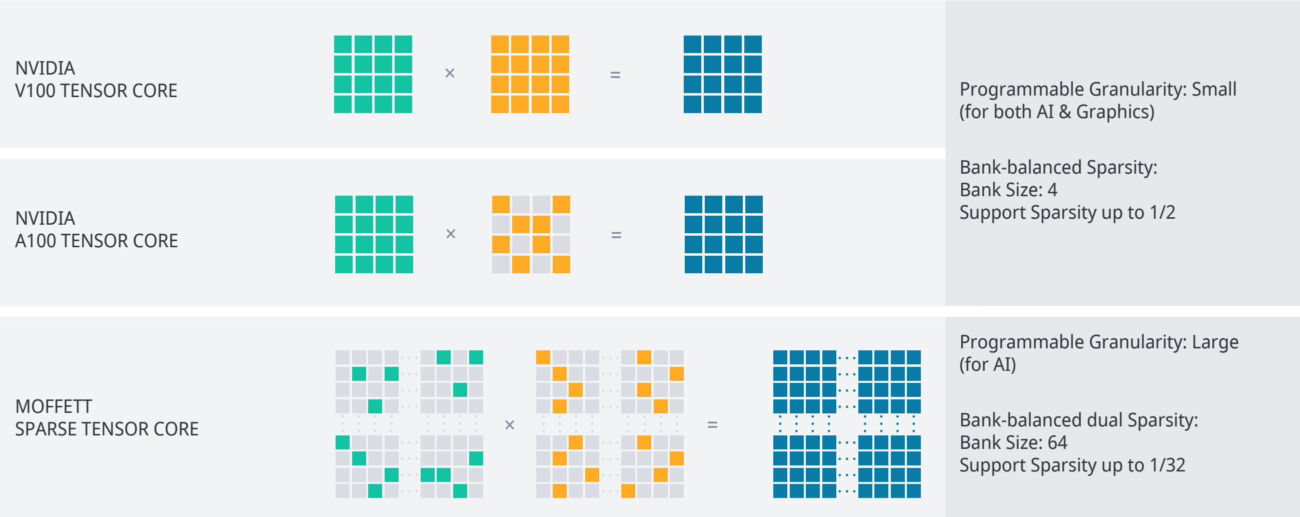
[8] As just one example, in early 2022, Intel advertised an “Intel Neural Compressor” tool aimed at model sparsification https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Quantizing-ONNX-Models-using-Intel-Neural-Compressor/post/1355237

Moffett AI and the path to dual sparsity

Moffett AI was founded in 2018, and is focused specifically on building hardware and software infrastructure to support dual sparsity.

Dual sparsity

As discussed earlier, dual sparsity, refers to sparsity in both the weights and activations of neural networks. The diagram below gives an intuitive comparison of the differences between dense-dense operations, which the majority of AI chips use currently, dense-sparse operations, which some chips like the NVIDIA A100 offer, and “dual sparse” operations which Moffett AI supports. One thing to notice is that using dual sparsity permits researchers to evaluate the products of much larger matrices while using the same amount of memory, compute, and energy. In the upper image, Moffett’s approach leads to a substantial speedup.



Four benefits of dual sparsity

The practical benefits of dual sparsity are fourfold: they include increased speed and decreased latency, no loss of accuracy, less energy consumption and lower cost thereby reducing TCO. To make these benefits more concrete, it is helpful to compare Moffett’s latest dual sparse chip, the Antoum, to its dense-dense and dense-sparse counterparts: the NVIDIA V100 and A100.

objects occur in a scene. Given that objects, even common ones like wheels and eyes, occur infrequently throughout most images, Moffett's researchers realized that it is possible to make the channel dimension very sparse. Starting from this software observation, they adjusted the chip's physical design, allocating less processing power for the channel dimension.

This fertile cross-pollination between hardware and software engineering is rare at larger companies. It's much easier to achieve at small startups like Moffett, where the same researchers tend to be involved in both areas of development. While sometimes this occurs naturally, company structure also plays an important role. In the case of Moffett, for example, its founders chose each other with hardware and software co-design in mind. Two are hardware experts, one is a sparsity software expert, and the fourth has a background in both. Having joined together under a common banner, Moffett's founders have the unique combination of skills needed to make progress in this area.

A vision

Moore's Law in 1965 looked to the heavens to a world where the future would bring more and more transistors. Gordon Moore wrote: "integrated electronics will allow the advantages of electronics to be applied generally throughout society"[9]

and his vision, at that time, was about how economics and manufacturing technology, yields and physics could be mastered by chip makers.

"There remain," he concluded, "many significant problems for the electronics industry to solve in attempting to take advantage of this evolving technology to supply the rapidly increasing electronic requirements of the world."

The vision today in 2022, for the whole Sparse community, and particularly for Moffett AI, no longer lies with the technology, and certainly not with increases in yields.

The vision now is to reduce computation by the use of algorithms, which change the characteristics of the model itself, and combining algorithmic optimization with chip design, as Moffett AI is doing.

The spirit and ambition of Moore is alive and well, but the world of plenty in 1965 has now been replaced by a world of caution and prudence where it is now about doing less, and using sparse techniques to prune and thin out data. Over time, not only should all AI processing become sparse, but all processing could be sparse, and every computer could be a sparse computer.

[9] From "Cramming more components onto integrated circuits" by Dr Gordon E Moore, Fairchild Semiconductor published in Electronics magazine, April 1965.

The mission at Moffett AI is to keep evolving the frontiers of AI processing using Sparse computing, and by extension to see that permeate to all processing, even beyond AI.

Which leads on to a new purpose, to power humanity's most demanding compute requirements with less energy and at lower cost, but faster and more accurately than ever before with Sparse computing.

We think Moore would have approved.[10]

Closing thoughts

Although the industry has started to put more time and energy into sparsity, there are many inefficiencies that have yet to be chiseled away. In coming years, we will need to adapt everything in AI, from chip design to low-level compilers and CUDA kernels to high-level autodiff frameworks, to better accommodate sparsity. Companies like Moffett AI are in a good position to lead this revolution. Perhaps the infrastructure they are building now will, in a few years, be running the most powerful AI models in the world.

[10] Dr Gordon E Moore established the Gordon and Betty Moore Foundation in 2015 to make a "significant and positive impact in the world", tackling large, important issues at scale with the areas of interest including environmental conservation, scientific research, higher education, having observed changes in the natural world and from the dependency of all living species on the planet's health. https://www.moore.org/about/founders-intent

MOFFETT AI