Lattice
Group 11

Team Members
Alexander Kyimpopkin, alxkp@seas.upenn.edu, (EE, ROBO)
Zirun (Danny) Han, dhan1@seas.upenn.edu, (EE, PHYS, SCMP)
Spencer Ware, wares@seas.upenn.edu, (EE, NANO)
Rose Wang, rywang@seas.upenn.edu, (M&T | CIS, Entrepreneurship)

Advisors:
Deep Jariwala dmj@seas.upenn.edu
Troy Olsson rolsson@seas.upenn.edu

Executive Summary

Neural networks are foundational to modern artificial intelligence, yet they are plagued by inefficiencies in their energy consumption and computational speed. The current state of AI relies heavily on digital hardware, where matrix multiplication, a core operation, incurs high costs in time and power. Our project proposes a paradigm shift: developing the first ferroelectric neural network accelerator leveraging compute-in-memory (CIM) technology. By integrating a ferroelectric diode (FeD) crossbar array into a custom-designed printed circuit board (PCB), we aim to reduce the time complexity of matrix-vector multiplication to constant time, giving neural nets a way to scale for free. Further, this significantly decreases inference latency and improves energy consumption by **700x** compared to NVIDIA hardware.

The compute needed for AI can be partitioned into compute for training and compute for inference. Our device is specifically suited for quick matrix-vector multiplication in inference rather than training. With our device, we physically "program" weights of the trained neural network into the device itself by sending voltage shocks to set the conductance state of specific diodes. Our ferroelectric diode-based device saves time moving data from memory, is simpler to manufacture, easier to scale, and significantly more energy efficient than existing solutions.

As the complexity of models only continues to grow exponentially (Fig. 1), we see that improving the performance of these models is increasingly dependent on:

- The size of the network itself (n)
- The amount of time spent in inference (running the same model over the same trained weights and choosing the best outcome)

This is what has led to the performance improvements of recent models like Deepseek R1, GPT4-o1 and -o3, and in Deepseek R1's case, without the use of massive amounts of training compute. Not only does this mean that the decrease in complexity to constant time is more impressive for larger and larger n's, and also means that a lot of that inference time can be significantly decreased through our device.

This innovation addresses critical challenges in the AI ecosystem. Data center power consumption has surged alongside the growth of AI, leading to heightened strain on energy grids and increasing costs for industry stakeholders. With the slowdown of Moore's law, data centers must turn towards specialized hardware accelerators to improve performance per watt. The one-time cost of purchasing these accelerators is outweighed by the energy savings that will decrease the operating expenses of datacenters, both at idle time and during computation.

Our analog neural net accelerator is not just a technological breakthrough but a solution with far-reaching implications for edge AI applications, today's largest hyperscalers, government efforts like Stargate, and environmental sustainability. By making energy-efficient computation accessible, we aim to disrupt the market dominated by traditional GPUs and provide a targeted alternative tailored to fast and efficient inference tasks.

Value Proposition

Our device offers unparalleled speed and efficiency for AI inference tasks by enabling matrix-vector multiplication to be performed directly in memory. The value of making what is currently a digital computation into an analog one are:

- Unlike existing CPUs and GPUs that rely on power-intensive data transfers between memory and computation units, our FeD array integrates these operations, drastically reducing energy usage.
- Since we are computing via hardware, the size of the computation no longer affects the latency and energy demands of the computation. This is shown by the decrease in time complexity from $O(n^2)$ to $O(1)$. This means that the exponential increase in the size of models seen since 2018 (Fig. 1) no longer impacts the time and energy it takes to run these models.

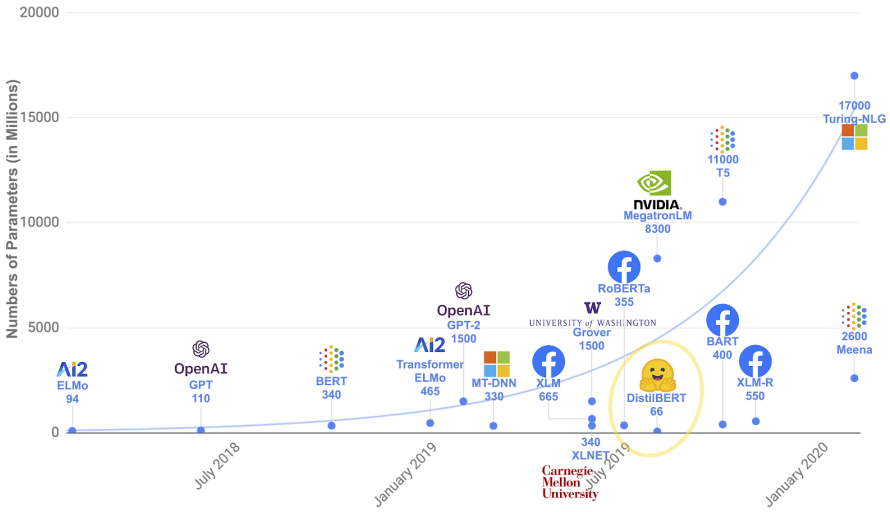


Figure 1. Size of models in millions of parameters year over year.

- With the flattening of fundamental laws like Dennard Scaling, Amdahl's Law, and Moore's Law (Fig. 2), increasing the capabilities of data centers is requiring a greater shift towards parallelism in the form of multi-core CPUs and the movement towards specialized hardware that accelerate specific functions like GPUs and devices like our accelerator.

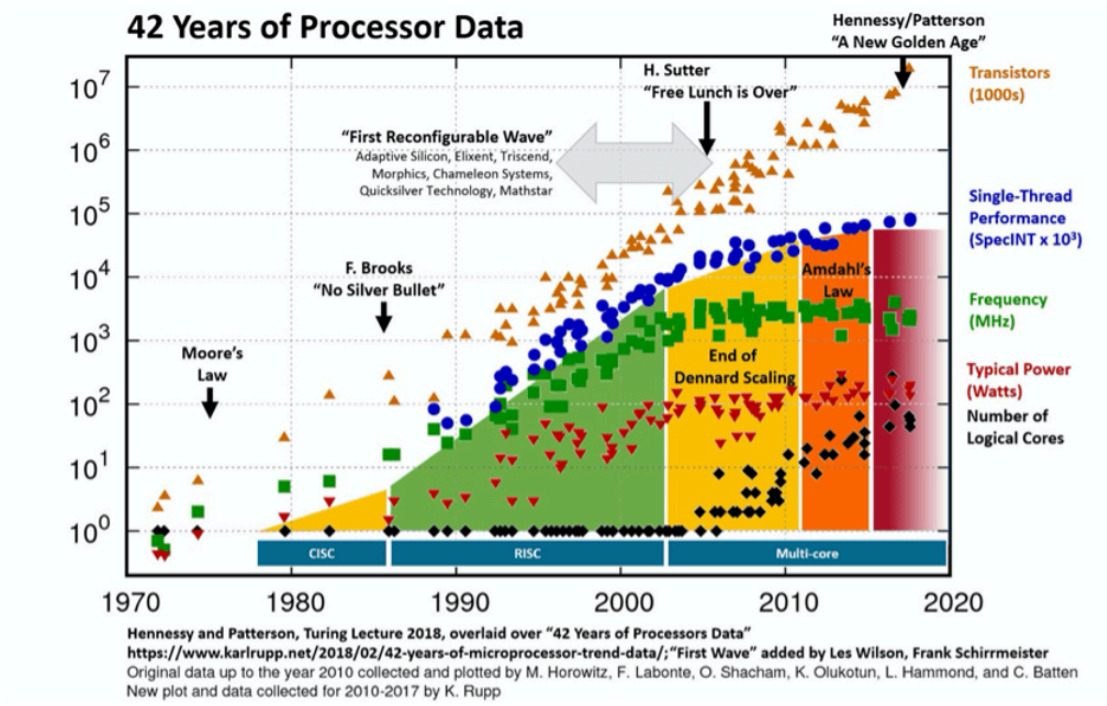


Figure 2. 42 Years in processor data showing the innovations in CPU development in response to the flattening of fundamental laws in semiconductor design.

Preliminary benchmarks suggest that similar compute-in-memory technologies can achieve up to 700x efficiency gains per operation. With our innovation, foundation model companies can cut operational costs, governments can mitigate energy grid stress, and industries focused on edge AI—such as agriculture and space exploration—can access compact, efficient hardware solutions. Furthermore, this solution directly aligns with emerging priorities in onshore chip production, addressing national security concerns and reducing reliance on global supply chains.

By the end of this semester, our goal is to show our device in action, complete with programmable weights, reading to and from the device, ending in a round of inference with the MNIST handwritten digit recognition model.

Stakeholders

Key stakeholders for this project include:

1. **Semiconductor Designers and Manufacturers**: NVIDIA, AMD, and TSMC dominate the semiconductor market but have yet to focus on the specific advantages of compute-in-memory technologies. Our analog compute-in-memory technology provides significant value over other hardware accelerators due to the decrease in energy consumption. GPUs and TPUs have significant problems with both cooling and energy consumption, with high consumption in idle states and even requiring on-device liquid

and plate cooling. By performing computations physically in hardware, our specific technological moat provides us significant improvements over existing solutions.

2. **End-Users**: Edge AI developers, cloud service providers, and foundation model companies will directly benefit from the cost and energy savings offered by our accelerator. Due to the slowdown of the rate of performance improvements in single threaded cores, companies with large data centers like Google, Meta, Amazon, and Microsoft are all investing in hardware accelerators like GPUs and TPUs in order to increase compute efficiency per watt of electricity consumed (Fig. 3). These GPUs and TPUs are either built in house or bought from the wider semiconductor market.

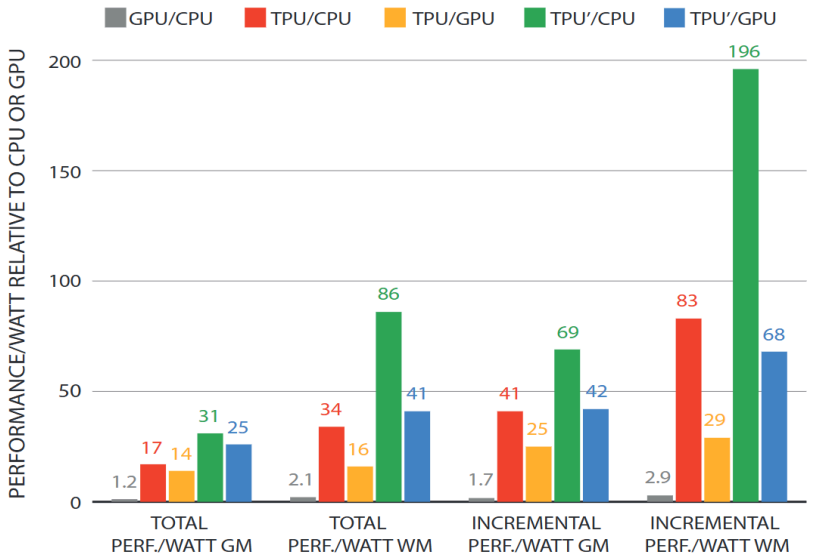


Figure 3. Relative performance/watt (TDP) of GPU server (blue bar) and TPU server (red bar) to CPU server, and TPU server to GPU server (orange bar). TPU' is an improved TPU. The green bar shows its ratio to the CPU server and the blue bar shows its relation to the GPU server. Total includes host server power, but incremental doesn't. GM and WM are the geometric and weighted means.

3. **Government and Public Sector**: Agencies concerned with energy sustainability and national security can benefit from reduced grid stress and localized chip production. The increasing strain of generative AI on the energy grid threatens America's aging infrastructure. Datacenters that power the cloud now have a greater carbon footprint than the airline industry, with one data center consuming the equivalent of 50,000 homes and the total energy consumption at 200 TWh surpassing some nation-states. This is why companies are even looking to their own power sources, with Microsoft looking to restart power generation on Three-Mile Island. This is also why the recent $500 billion Stargate initiative was started, with a focus on investing in American AI infrastructure. In the coming years, we see that demand for power outpaces our ability to generate electricity.

The creation of new technology is necessary to find new ways to significantly reduce energy consumption in the near future.

Market Research

The global semiconductor market is $702.40 billion projected to grow at a compound annual growth rate (CAGR) of 8.7%, reaching $980.80 billion by 2029. Within this landscape, GPUs and accelerators dominate the AI inference market. However, their reliance on digital architectures limits their efficiency for specific tasks. Compute-in-memory technology, by contrast, offers a specialized solution with transformative potential. Existing neuromorphic devices demonstrate power consumption as low as 0.01 pJ per operation, a 700x decrease from existing technologies on the market.

From our revenue calculations, we calculate a total TAM of $79 billion. This TAM is based on our two target customers as revenue streams, and totalling the total revenue of capturing 100% of the market. For more on our TAM calculation, see the Revenue Model section.

Customer feedback and initial advisor consultations indicate strong interest in energy-efficient inference devices, particularly for applications where traditional GPUs are overkill, like inference. Market demand is further corroborated by the increasing focus on sustainability across industries due to increasing operating expenses and the growing shift toward inference optimization over model training.

Customer Segments

Our primary target customers are hyperscalers and edge computing applications.

Foundation Model Companies and Hyperscalers: Businesses running large-scale AI inference operations, where energy costs constitute a significant overhead are looking for ways to increase the overall performance and power efficiency of their data centers proportional with the bursty nature of the demand of computation (Fig. 5). Data centers are often running at 10-50% of capacity, but servers often have poor energy proportionality, requiring large amounts of energy even when idle.

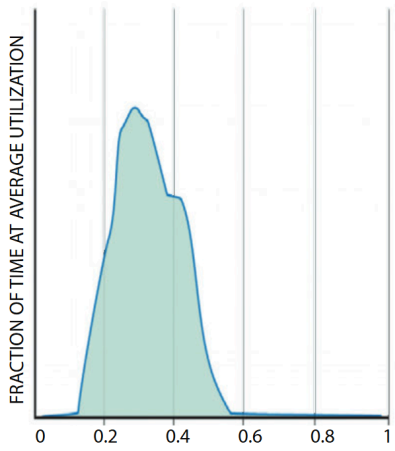


Figure 5. Average activity distribution of a sample of an average shared Google cluster that runs mixed workloads. The cluster contains over 20,000 servers, over a period of 3 months.

In a Google case study modeling the costs per year of a data center at 50% capacity, we see that the overhead cost of running the datacenter (DC Opex), server power, server operating expenses, and the Power Usage Effectiveness (PUE) overhead account for in total 31% of building and running a data center at hyperscale levels (Fig. 6). This implies that improving energy proportionality of hardware and software as well as decreasing the sheer amount of energy consumed for computation, thereby decreasing cooling requirements and cooling overhead, would greatly decrease the day to day costs of running a data center.

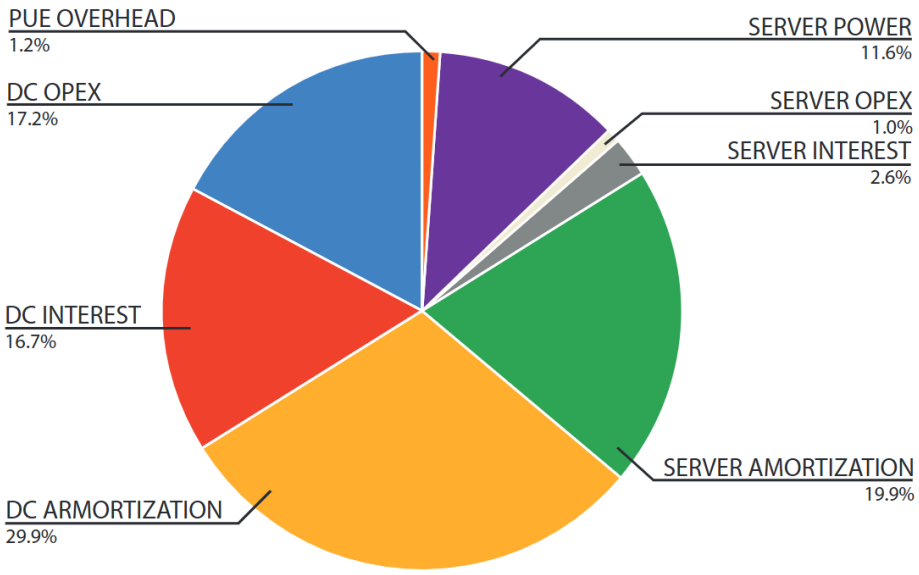


Figure 6. Breakdown of the yearly TCO for a 50% utilized data center among data center and server-related Opex and Capex components.

As our device provides the benefits of in-memory analog computing, we can greatly improve the energy efficiency and proportionality of matrix-vector computation for inference. As foundation model companies and hyperscalers continue to aim to serve not only the developed world but also the rest of the world's population, the sheer scale of inference computations necessary will go from the millions to the billions. The importance of improving inference-time computation efficiency will only become more apparent as the market of AI-based technologies expands.

Edge Computing: Innovators in sectors like agriculture, space exploration, and autonomous systems require compact, efficient hardware for real-time processing. Current edge computing solutions use heavy-handed solutions like NVIDIA Jetson Nano, which have high energy consumption costs, leading to overheating problems. At the edge, speed and energy consumption are evermore important, as batteries are large and heavy and split second decisions must be made constantly. We already see a lot of applications in the autonomous vehicle space. Self driving car companies are looking to use large foundation models to aid autonomous driving decision making, and making these models run quickly and in an efficient manner is already a challenge. Tesla has already been making and using their own silicon to address this challenge, and existing assistive features are all AI based. We believe that light weight solutions that prioritize inference over training, energy utilization, latency, and temperature maintenance would be significantly more attractive for edge computing applications than less powerful energy hungry GPUs.

Competition

The AI hardware landscape is dominated by NVIDIA, AMD, and TSMC, whose GPUs and TPUs set benchmarks for AI computation. However, these devices are designed for general-purpose tasks, leading to inefficiencies for specialized inference operations. Our direct competitors include companies exploring neuromorphic and analog computing technologies, though most focus on niche applications or remain in experimental stages. Some companies have developed related technologies. Micron developed and published a Hafnium Zirconium Oxide (HZO) 32Gb nonvolatile (long memory storage) ferroelectric memory chip in 2023 with high endurance and retention. HZO is a strong candidate for ferroelectric compute in memory due to its highly developed processes in semiconductor manufacturing and its low ferroelectric switching voltage. Micron has not made any announcements on HZO compute-in-memory, but their recent advances point in that direction. Smaller companies like Mythic AI are working on analog compute-in-memory technology, but Mythic uses flash memory and floating gates to encode neural network weights. This technology suffers from charge leakage and endurance, whereas our FeDs are nonvolatile requiring less power to operate over time.

Our differentiation lies in our patented FeD array technology, which integrates computation and memory to deliver unmatched speed and efficiency. By targeting inference-specific workloads, we avoid direct competition with GPUs, positioning ourselves as a complementary technology

rather than a replacement. Furthermore, our modular design and compliance with IEEE standards ensure adaptability and reliability, offering a compelling alternative to traditional solutions.

Cost

Currently, we are fabricating these chips using a 4-inch wafer scale process in the cleanroom facilities at the University of Pennsylvania. Currently, it costs us $7.75 to produce a chip that can encode 16,384 parameters, amounting to $473 per million parameters. The cost breakdown is as follows: The 4-inch sapphire wafer costs $200, and five hours of tool use ($25/hour) and six hours of total cleanroom time ($35/hour accessed), which include the cost of any deposited material, are required to fabricate a full wafer containing 69 arrays.

However, this cost per parameter can be improved dramatically when produced in a commercial fab. First, 90% of the area on the current chip is used by pads for wire bonding and traces that connect these pads to the array - the actual memory itself only occupies 10% of the available area. The area used by pads and traces can be completely **eliminated** when the memory array is fabricated back-end-of-the-line, that is, directly on top of a CMOS silicon processor, as vertical via connections can be used to link the memory array to the processor in a space-efficient manner.

Secondly, the crossbar array design is very amenable to lateral scaling, which has a quadratic effect on device density and per-parameter cost. In our research lab, we have demonstrated the successful scaling of a single device to a diameter of 50 nm while maintaining the same current density, which would increase the device density by 10,000 times. We achieved this feature size by using electron beam lithography, but this is also easily achievable with Extreme Ultraviolet (EUV) photolithography used in commercial foundries.

Our devices can be grown on silicon oxide instead of sapphire, which means commercially, standard 300 mm silicon wafers can be used, further lowering cost. The average cost of production for a fully fabricated 300 mm wafer at TSMC is approximately $6,500, and we can use this as a reliable upper bound to how much our process may cost, as it is significantly simpler than the fabrication process needed for silicon processors. Assuming this scaling, each wafer can support approximately 7 trillion devices, equivalent to 7 trillion parameters, equating to a cost of **0.1 cents per million parameters**.

Revenue Model

There are two potential distribution channels for our device, back-end-of-line and selling full off-the-shelf individual systems. Back-end-of-line is a methodology in semiconductor manufacturing where we would be able to provide our customers with our technology over their existing silicon. We believe that there are applications for this distribution technology in the edge computing customer segment, where devices may be very specific to the use case of the edge

computing application, and space is important. We believe that selling full individual systems will be targeted towards data center use cases, where hyperscalers and small data centers can purchase our hardware accelerators and incorporate them into their data center build to minimize energy consumption and decrease latency.

We projected our revenue using an adjusted version of a ROIC (Return on Investor Capital) tree (Fig. 7).

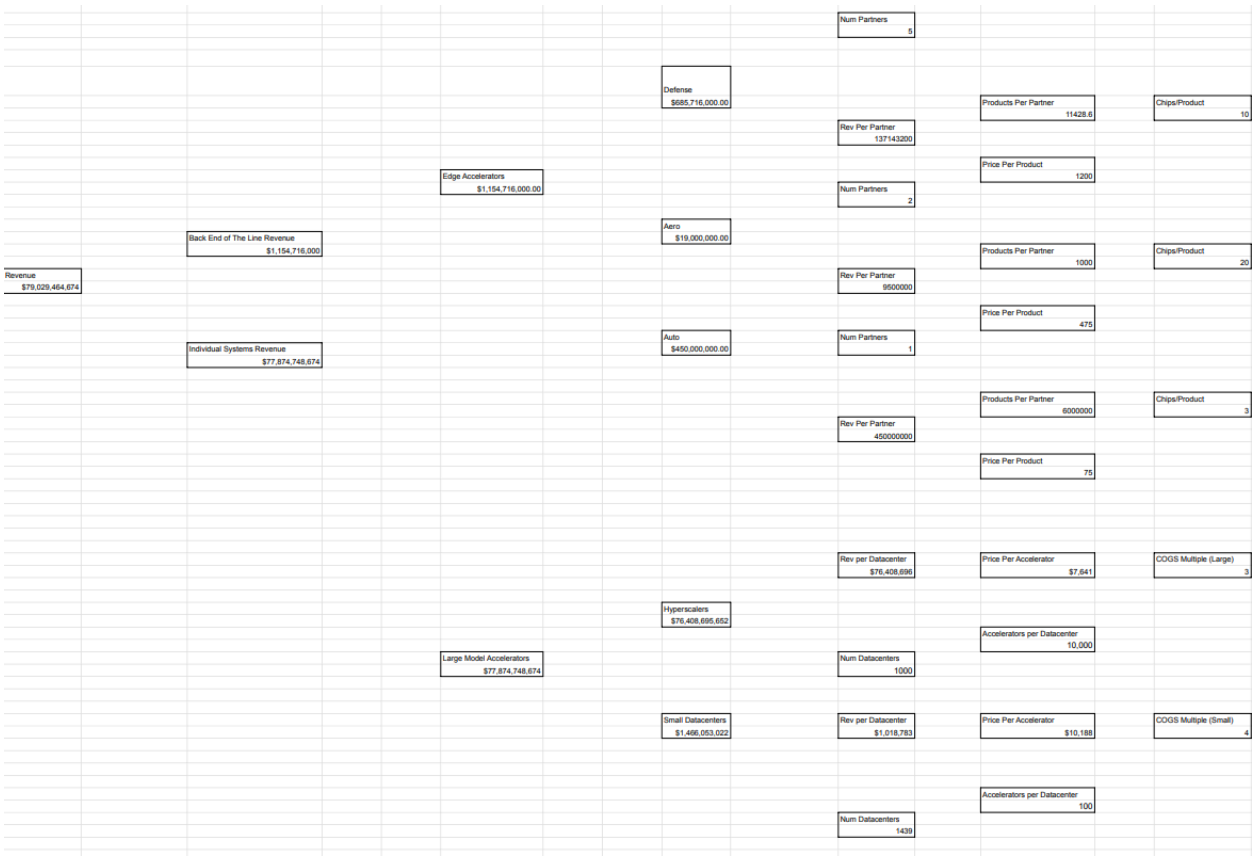


Figure 7. Revenue Model from two revenue streams.

We further segmented the edge computing market into the defense, aerospace, and automotive industries. We projected $685.72 million from defense, $19 million from aerospace, and $450 million from automotive. We derive these values from historical data of defense, aerospace, and automotive company sales numbers.

We also segment the data center market into large (hyperscaler sized) data centers and small data centers. We projected $76 billion in revenue from hyperscaler data centers and $1.5 billion from small data centers. Hyperscalers account for 41% of data centers and recently surpassed over 1000 data centers across the world in early 2024. These data centers are enormous, with one

Amazon data center cluster having over 20,000 GPUs. Smaller data centers are much smaller in terms of the amount of hardware they have, and thus much less lucrative in terms of revenue.

We derive pricing values from pricing of competitive products, like NVIDIA GPUs and NVIDIA Jetson Nanos for edge applications.

Intellectual Property

Our project is grounded in intellectual property developed by Prof. Jariwala's group, specifically Patent Publication Number 20240177759. This patent outlines the use of ferroelectric diodes for compute-in-memory applications, a core component of our accelerator. Additional IP may emerge as we refine our nanofabrication processes and circuit designs, ensuring a competitive edge in the market.

By building on this foundational IP, we not only secure a technological advantage but also create opportunities for licensing and collaboration with academic and industrial partners. Ensuring robust protection for our innovations will be critical as we transition from prototype to production.

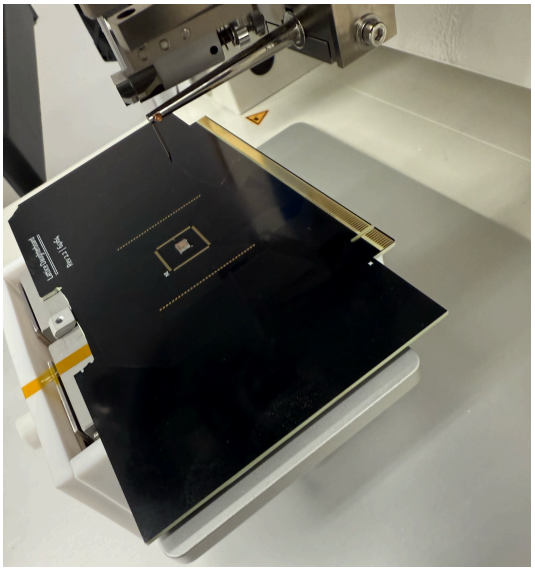
We note that our advisors' labs as well as our group members will have a stake in any IP resulting from this project.

Conclusion

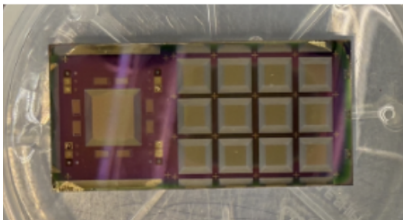
Our ferroelectric neural network accelerator presents a transformative solution to the growing energy and computational inefficiencies in AI inference tasks. By leveraging compute-in-memory technology and utilizing a ferroelectric diode crossbar array, our device reduces the time complexity of matrix-vector multiplication from $O(n^2)$ to $O(1)$, offering up to 700x energy savings compared to traditional digital hardware. This innovation positions our technology as a complementary alternative to existing GPUs and TPUs, targeting energy-intensive inference workloads, which are becoming increasingly crucial as AI models grow in size. With its significant potential to reduce operational costs for hyperscalers and edge AI applications, our accelerator can help address the growing strain on data center power consumption, aligning with both industry needs and sustainability goals. Furthermore, our patented technology, backed by strong intellectual property and scalable manufacturing processes, provides a solid foundation for future commercialization and market disruption.

Demo

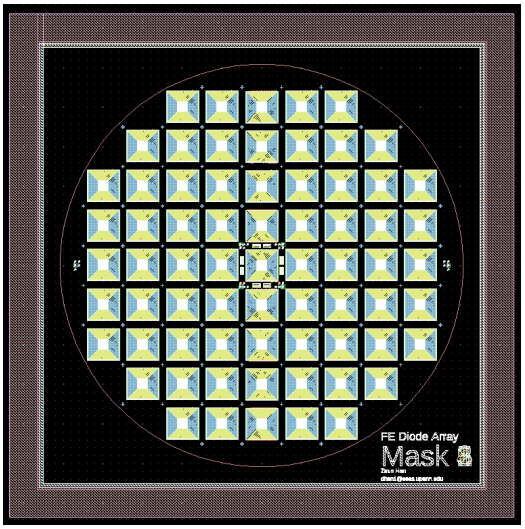
Wirebonding the device to the daughterboard:



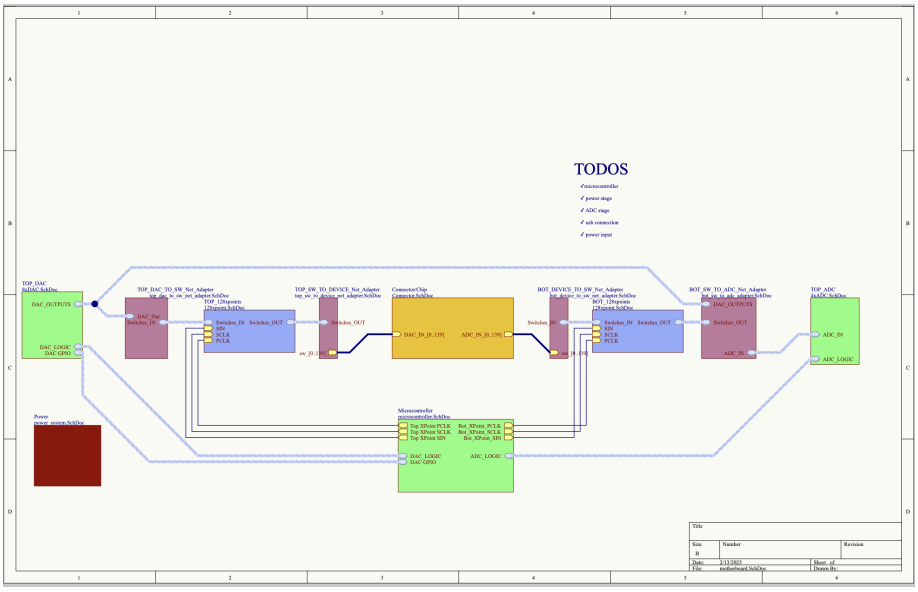
Device fabrication progress:



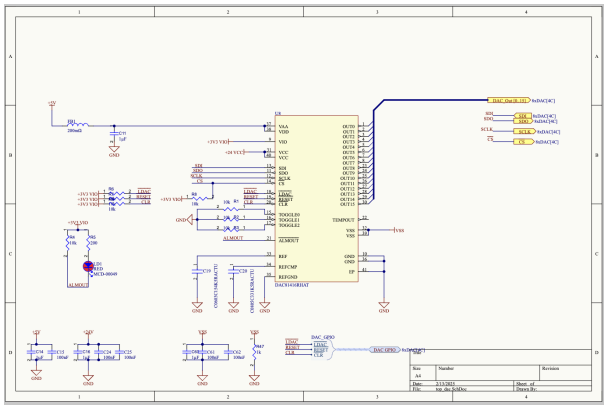
Mask creation to fabricate devices at a larger scale:



Motherboard design schematic:



DAC Design



ADC Design