Artificial Intelligence Hardware (AI HW) Projects

Energy-Efficient and Scalable AI HW Systems through Heterogeneous Integration of Specialized Chiplets (ScaleAI) - $6.7M, Lead: NW AI - Stanford-

This project will drastically improve energy consumption and performance of AI HW systems for DoD applications, reaching the TRL 6 and the MRL 7 upon completion. Existing AI systems use large off-chip memories and spend enormous time and energy shuttling data back and forth between compute and memory chips—the memory wall. This memory wall gets worse as traditional 2D scaling via miniaturization gets increasingly difficult—the miniaturization wall. The deadly combination of the memory wall and the miniaturization wall severely limits the capabilities of existing AI HW for DoD applications such as the Connected Battlespace. This project will use innovations in semiconductor materials, monolithic 3D integration technologies, and AI system architectures to enable new AI capabilities through a dramatic 100-fold improvement in system-level energy efficiency compared to current commercial solutions (GPUs/CPUs). A variety of interconnected, heterogeneous, AI-specialized chiplets, built using leading-edge CMOS and 3D CMOS+X semiconductor technologies such as carbon nanotube transistors, resistive memory, and oxide semiconductors, form the foundation of the overall approach. The transition plans include: 1) AI HW transition via the DIB and 2) establishment of a new 3D CMOS+X technology in a U.S. foundry—a world-first capability—for sustained U.S. leadership in semiconductor innovation. Thus, this project will directly advance the Commons' mission of lab-to-fab translation of critical technologies.

Lab-to-Fab Transfer of complementary metal oxide semiconductor (CMOS) +memristor Chips for Edge Intelligence (Edge-Intel) - $7.9M, Lead: NEMC-University of Massachusetts, Amherst-

This project aims to transfer analog memristor technology from research labs to commercial foundries, creating a thriving AI HW ecosystem for DoD mission-relevant applications. Energy and latency issues are acute at the network's edge and on autonomous and wearable platforms where extreme energy efficiency is mandatory for DoD applications. Commercial AI HW purely based on CMOS does not address these issues and faces challenges in terms of SWaP. Physical computing with analog memristors has been demonstrated as a promising solution as it avoids constant data shuttling between memory and compute chips, and instead performs analog computing where the data is stored using the laws of physics, substantially reducing latency and energy consumption. The greatest obstacle to adapting this technology to real-world applications is the lack of an analog memristor technology in the IC industry. In this project, multi-level memristor devices and processes from universities and startup labs will be transferred to the industrial fabs and integrated with current CMOS technology, offering a commercially viable analog memristor device and system technology on U.S. soil at TRL 7 by the end of the project. Our project will enable compact, fast, intelligent electronic systems for unmanned aerial vehicles, autonomous sensing systems, hypersonic weapons, etc. The AI chips will have substantial SWaP advantages and will address challenges in data-to-decision and autonomy requirements in DoD missions. The proposed project fully engages the NEMC ecosystem and leverages

local and national facilities and infrastructure, including academia, startups, defense contractors, R&D, and commercial foundries. By bringing each team member's expertise together, we will enhance the NEMC capabilities in AI HW development and implement the mission of the Commons in lab-to-fab transfer.

Energy-Efficient, Scalable, and Self-Learning AI HW with 3D Electronic-Photonic-Integrated-Circuits (3DEPIC_AI) - $5.7M, Lead: Northwest-Artificial Intelligence hub (NW AI) -UC Davis-

The project "Energy-Efficient and Scalable, and Self-learning AI HW with 3D Electronic-Photonic-Integrated-Circuits (3D EPIC-AI)" addresses the significant limitations of contemporary AI HW solutions reliant on GPU, CPU, and traditional memory technologies such as HBM, DRAM, and SRAM. By taking the 'best of both worlds' approach, the project innovatively combines photonics and electronics into a compact and advanced 3D integrated circuit module, this initiative aims to create a new class of AI HW that effectively overcomes the innate shortcomings of existing architectures. Today's AI HW systems are hindered by substantial drawbacks, including excessive power consumption, bulkiness, weight, and inadequate processing speed—all resulting from limitations inherent in electronic processing and the von Neumann bottleneck. Furthermore, these systems lack the scalability and adaptability to seamlessly learn new AI models. The proposed innovations in 3D photonic-electronic integrated circuits (3D EPIC) present an unprecedented opportunity to enhance energy efficiency, throughput, and self-learning capabilities, facilitating the development of commercially viable AI HW made in the United States. Key advancements within this project revolve around the design of a novel photonic-electronic AI neural network architecture that incorporates cutting-edge technologies, such as silicon photonics, III-V photonics, Resistive Random Access Memory (RRAM) crossbar arrays, and 3D photonic-electronic integration techniques. The forecasted results promise dramatic improvements in both throughput and power efficiency on a resilient 3D platform, alongside the implementation of self-learning capabilities that significantly enhance learning speeds. The implications of this technology for the DoD are extensive and multifaceted, supporting a wide array of platforms including Unmanned Aerial Vehicles (UAVs), robotics, soldiers in the field, missiles, and naval vessels, among others. With the anticipated performance characteristics of beyond 1 Peta Operations Per Second (POPS), exceeding 10 Tera Operations Per Second per Watt (TOPS/W), and a self-learning capability at the edge, the project targets a solution with more than 10x throughput power density compared to current systems. Achievement of these ambitious goals entails successful transitions from laboratory to fabrication readiness (TRL5/MRL5), demanding engagement with U.S.-based manufacturing foundries. The lab-to-fab transition will leverage collaborative efforts with NW-AI partners, encompassing innovative prototyping in the areas of photonics, electronics, RRAM, and 3D electronic-photonic circuits, ultimately guiding these advancements toward commercial manufacturing at specialized foundries partnered with DoD labs and contractors. Through 3D EPIC-AI, we aim to redefine the landscape of AI HW, delivering systems that are not only high-performing but also significantly more energy-efficient, compact, and capable of evolving in real time to meet the pressing demands of modern defense applications.

CMOS+X: Integrated Ferroelectric Technologies for Ultra Efficient AI HW (FerrAI) - $4M, Lead: NW AI -UC Berkeley-

This project aims to substantially better energy efficiency for future AI HW compared to SOTA. Our team includes UC Berkeley as the lead and Stanford University and Georgia Institute of technology among the academic partners. In addition, the team includes MIT-LL who will be primary fab for lab-to-fab transition, and Intel who will drive system level functions. Further, we have Northrop Grumman Corporation and Raytheon Technologies who will drive defense related applications. The project will exploit unique properties of Ferroelectric materials to lower the power supply voltage of computing hardware as well as achieve non-volatile memory that can be directly integrated with the microprocessor. As the AI workload requires highly memory-centric and parallel operations, low voltage operation together with access to tightly integrated large memory can have a multiplicative effect on the energy efficiency of the AI HW. The impact on future defense systems can be enormous. Essentially, with significantly improved energy efficiency, it will be possible to process data at the edge improving efficiency and speed of decision making, leading to improved safety for defense use cases.

Fab-Accessible BRain-Inspired Chips for Sensors (FABRICS) - $2.1M, Lead: Naval Research Laboratory (NRL)-

FABRICS aims to develop ferroelectric devices made from materials that are compatible with FEOL and BEOL fab processes and that implement synapse- and neuron-like functions for neuromorphic computing applications. To this end, a novel rapid laser annealing process is used to transform hafnium- and zirconium-oxides into crystalline phases that exhibit strong ferroelectric polarization and are stabilized for their final device environment. Simulations of processing conditions and of the device physics, derived from the characterization of prototype devices we are fabricating, will be employed to generate process flows and circuit designs to enable full PDK support for ultimate fab integration of these devices.

CMOS+MRAM Hardware for Energy-EfficienT AI (CHEETA) - $8.7M, Lead: SCMC -Purdue University-

The current landscape of edge AI HW is dominated by mobile SoCs, GPUs and digital accelerators. These platforms achieve processing efficiencies in the range of 1-10 TOPS/W. However, they are all limited by the time and energy required to move data between processing units and memory (the von Neumann bottleneck) and the large leakage of on-chip memories. CHEETA seeks to leverage CMOS+MRAM to address the challenges. To alleviate the von Neumann bottleneck, CHEETA will deploy in-memory computing (IMC) primitives that leverage the non-volatility of MRAM to realize Matrix-Vector Multiplications (MVMs) – the key computational kernel in AI workloads – in situ within MRAM arrays. In addition, the design and fabrication of IMC-Plus primitives will extend the capabilities of the basic IMC primitives in two important ways: 1) ROM-overlays in order to realize operations beyond MVMs such as non-linear activations, Softmax, etc., and 2) stochastic neurons that replace area- and power-demanding ADCs in order to improve the efficiency of IMC. The proposed CMOS+MRAM hardware primitives will enable a new generation of energy-efficient AI HW that push the boundaries of the current state-of-the-art. The end state of CHEETA seeks to develop a TRL-6 technology that drives greater than 100X improvement in energy efficiency compared to current AI HW. Furthermore, the high endurance of MRAM compared to other NVM technologies and the intrinsically radiation-tolerant nature of MRAM will

offer compelling advantages in DoD applications. This project is leveraging the SCMC lab-to-fab transition by the incorporation of the Honeywell radiation-hardened CMOS FEOL, Everspin BEOL, with an end-state transition partner in Northrop Grumman. CHEETA will also be provided with SCMC-supported EDA access during the development process.

Spaceborne Low-Energy AI Computing (SLEAC) - $6M, Lead: SWAP-Arizona State University –

Spaceborne Low Energy AI Computing (SLEAC) aims to extend the power of AI to satellites orbiting our planet. Advancing the performance of satellites through AI has the potential to unlock advances that will yield a major advantage for national defense. Computer hardware designed for spacecraft undergoes a manufacturing process called radiation hardening. It enables the equipment to survive exposure to harsh radiation levels in orbit, but it also compromises performance compared to state-of-the-art computing on Earth. If it were possible to directly integrate a highly efficient, radiation hard AI chip with focal plane array image sensors used in space, satellites could track objects that are too faint or too fast to be detected by current systems. SLEAC project will overcome this challenge by using radiation hard metal oxide-based analog-optimized resistive memory (ReRAM) arrays integrated with foundry-supplied CMOS to create an AIMC prototype processor capable of more than 10 TOPS/W in extreme environments. This would be several times more efficient than modern unhardened systems and over 100 times more efficient than the current state of the art radiation hard systems. ReRAM is a robust form of nonvolatile memory that is being developed at Arizona State University, University of Southern California and Sandia National Laboratories starting from a Sandia baseline process. The project will also leverage resistive switching materials and devices developed for extreme environments at the Center of Neuromorphic Computing under Extreme Environment (CONCRETE) sponsored by the Air Force. It is being optimized and electrically conditioned to have near-ideal properties for AIMC, including stable resistance tunability with a high dynamic range and extremely low noise, making it ideal for scaling to large arrays with high accuracy and efficiency for AI computing under extreme environments. SLEAC plans to integrate ReRAM on the back-end-of-line of commercial radiation tolerant CMOS wafers, using Arizona State University's recently established SWAP capabilities. This will enable demonstration of large-scale ReRAM arrays with high yield, endurance, and low variability, validated across millions of devices. Ultimately this technology will enable demonstration of a radiation hard spaceborne remote sensing systems capable observing phenomena that are currently hidden.

RAVENS 2 chip Demonstration platforms for Deployment (R2D2) - $1.9M, Lead: NRL-

The RAVENS 2 chip Demonstration platforms for Deployment (R2D2) program aims to demonstrate advanced performance in AI tasks through advanced packaging and systems-level integration of the RAVENS2 ASICs developed under the NeuroPipe ARAP into prototypes of applied systems such as a sensory data processors and autonomous vehicle navigation systems. These will be tested at the respective laboratories and warfare centers to demonstrate advances in performance and reductions in SWaP above the established baseline of existing AI HW. The effort will demonstrate the feasibility of memristor-based circuits to dramatically shift computational loads associated with cognitive functions away from digital hardware towards fully analog neuromorphic devices.

Ultra Efficient In-Hardware Prototype Using Hyperdimensional Computing (ENERGY) - $1.6M, Lead: MMEC-Intel Federal-

The Ultra Efficient In-Hardware Learning Prototype Using Hyperdimensional Computing (ENERGY) seeks to provide an AI HW solution for in-hardware learning. AI proliferation faces a number of technical challenges. Current AI HW is large and power hungry which negatively impacts SWaP. In addition, SOTA implementations are sample/training inefficient and susceptible to noise and perturbations. The ENERGY project will develop a hardware prototype that is capable of learning in-hardware directly from the raw image and video data in an incremental fashion that is SWaP efficient in noisy and dynamically changing environments. The effort will build on a breakthrough neural architecture based on Hyper-Dimensional Computing (HDC) which has been developed that enables revolutionary few-shot learning capabilities. HDC, a non-von Neumann computing framework that mimics human memorization and cognition, uses a high-dimensional representation of data called hypervectors (> 1 Kbits), that mimic the representation in the brain which are robust, holistic, and sparse. Recent results have shown that good, few-shot learning performance does NOT require pretraining on a large dataset of labeled examples if the neural representations are defined by tight geometric regions in high dimensions akin to higher-level brain areas. To obtain data hypervectors from complex images for image and video processing, a novel and compact unsupervised/supervised image segmentation module inspired by the early stages of visual processing in the brain will be developed. Hypervectors or symbols for each region of interest will be extracted using streaming encoding techniques. This data-to-symbol converter will be trained offline once with image data in an unsupervised/supervised fashion and then compressed for low SWaP requirements. Initial algorithms will be demonstrated in an FPGA. A longer-term goal is to combine an FPGA with a ferroelectric diode array tailor-made for compute-in-memory to greatly reduce sample, training, and power demand compared to CPU/GPU implementation. The ENERGY project is led by Intel who is responsible for AI and hyperdimensional computing models, FPGA implementation, and program management. AI-Sensation is working with Intel on overall design of algorithm hyperdimensional computing algorithms for joint learning and reasoning. Future activity will include University of Pennsylvania who will work on ferroelectric diode-based compute-in-memory. The AFRL and the ARL will support characterization of the ferroelectric devices. ENERGY is expected to mature Hyperdimensional Computing from a TRL/MRL of 4 to enable breakthrough neural architectures of a TRL/MRL 7.