AUGUST 2019

National DEFENSE

THE BUSINESS AND TECHNOLOGY MAGAZINE OF NDIA

Viewpoint BY JAMES S.B. CHEW

Realizing the Potential of AI on the Edge

In the late 1970s, the outlook for the internal combustion engine was bleak. Faced with complying to both aggressive fuel economy standards and stringent emission regulations, the American driver seemed destined for a future of small, underpowered cars.

But something funny happened. Engineers and researchers began to study the technological state of those engines. They discovered that while those engines performed adequately, the technology had not progressed much since the development of the Otto cycle. These engines were designed to perform optimally at one set of operating conditions — open road, nonstop driving. Internal combustion research-and-development efforts were executed to develop technologies and components, which would allow the internal combustion engine to operate optimally at all common operating conditions.

These efforts followed a systems approach to re-examine the basic engineering principles to understand and optimize subsystems, then optimize subsystem to subsystem to system interactions. Because of this system-of-systems approach, today we have four-cylinder engines that produce more horsepower and torque than the 1960s-era muscle car engines while achieving average fuel efficiencies of over 30 miles to the gallon. We have V-8 engines that produce horsepower and torque once reserved only for special-purpose racing vehicles — all while achieving fuel efficiencies that are far better than the 1980s-era subcompact cars.

The same system-of-systems optimization approach must be followed to fully realize the potential of artificial intelligence. With the advent of AI at the edge and the associated need to make intelligent decisions in an untethered mobile environment, upwards of three Tera-MACs per watt are needed. To achieve that kind of performance, industry must examine new AI hardware architectures and software solutions that are tailored to the broad range of missions across the military and aerospace landscape. To create these new solutions, new design methodologies, architectures and innovation in the development and verification of AI-related hardware and software is required that go well beyond the frameworks of today.

ISTOCK

Today, continued advancements in microprocessor technology and the availability of big data provide a natural foundation that fuels the interest in AI-enabled devices, as well as machine learning-enhanced design and verification processes. Most of the Defense Department's artificial intelligence efforts have been focused on software algorithm development on existing microprocessors and hardware. Organizations within the department

are leveraging the potential of state-of-the-art microprocessors to develop new AI-enhanced warfighting capability.

As impressive as the results have been, these efforts will not address the Pentagon's need for mass adaptation of AI-enhanced devices. The limitation is that they all rely on being tied to a computer center resource. While there have been significant advances in cloud computing, there is a multitude of defense scenarios in which the resulting data latency would render AI-enhanced warfighting capability useless.

The solution is the development of AI-enabled electronics that can learn and tailor themselves to the goals of the mission in an untethered and mobile environment, thus "AI on the edge." The implementation of such an approach requires the development of optimized AI-specific semiconductors that, just like the engine examples above, can be configured to meet the correct mission parameters in a robust, verified way.

In the commercial sector, application-specific integrated circuits are seeing fast growth for edge applications that span mobile phones as well as medical, drones and industrial applications that include vision and speech. They have a longer-term value proposition for edge applications because of advantages in power and decentralized independence. According to a May 2019 Tractica report on AI, these circuits will represent 52 percent of global deep learning chipset revenue by 2025. The implication is that the train is coming, and technology developers can either get on the train or get in front of it. This is far easier in the commercial sector, but to realize the potential of AI on the edge, the Defense Department and defense industrial base technology must inevitably follow a similar path, only with more stringent requirements around verification, particularly as these edge devices get integrated into existing systems.

Of course, incorporating AI on the edge does not mean the elimination of the cloud, but, better yet, an implementation that yields the most efficient and optimal results for the overall system. This implies a system architecture that leverages the computational resources available in the cloud, with high-performance, low-power system-on-a-chip (SoC) at the edge.

There are two main stages of AI: training/learning and inference. The training/learning stage performs analysis on available data — the more data, the better — develops neural network structures and requires substantial computational resources. The inference stage is where the neural network structures are deployed, and reactions/decisions are made based on the input into the system. To move the inference stage to the edge, certain characteristics are required to drive the feasibility of this efficient architecture such as: low power, high performance/efficiency, optimized processing, etc.

While off-the-shelf devices exist for AI processing at the edge, most of these devices are general-purpose AI engines and none are optimized for their specific tasks at hand. The recent introduction of embedded AI engines, such as the DNA100 DSP from Cadence, allows the development of application-specific system-on-a-chip that delivers the optimal AI-at-the-edge implementation with advanced inference capabilities, so systems can make real-time decisions without the latency —and cost — associated with data transfer and response times with a cloud-based implementation. In addition, these chips also enable a software programmable implementation that allows for the repurposing of SoC-based AI resources, bringing long-term flexibility to the overall system such as adaptability to future algorithm and system requirements.

> "Industry must examine new AI hardware architectures and software solutions ..."

The size, weight and power desire for commercial and defense AI-on-the-edge devices is driving the semiconductor industry to smaller node sizes, stretching the limits of Moore's Law and creating a class of "more than Moore" customers. It has also created a growing number of small, "two-pizza AI-specific hardware" development companies within Silicon Valley.

The reason for the rapid growth of these companies is simple — the state-of-the-art electronic document access tools and processes, as well as the best-in-class emulation devices, can support such development. In fact, the best-in-class emulation devices can be configured to develop an artificial intelligence innovation hub, creating an AI hardware emulation center that can allow for the free flow of hardware design ideas, as well as AI hardware design emulations. By following the commercial electronics industry design best practice of using the best-in-class emulation systems to emulate before fabrication, one can be assured that the final AI on the hard device will achieve first-pass success and be future-proofed.

It wouldn't be prudent for the Defense Department to independently initiate AI-on-the-edge hardware development efforts, or, for that matter, any artificial intelligence effort, without leveraging the billions of dollars of past and current commercial industry investment. The undersecretary of defense for research and engineering, Congress and the office of science and technology policy should require that all current and new defense AI efforts show the commercial industry leverage to ensure the most efficient use of resources. The good news for defense AI science and technology and the acquisition and sustainment communities is that these state-of-the-art electronic document access tools and processes — as well as the best-in-class emulation devices, which can support such development — are now available at several Defense Department facilities.

The foundation for the department to realize the potential of artificial intelligence and machine learning is now in place. It would behoove it to reconfigure and enhance that foundation into a Defense Department AI innovation hub, allowing for the establishment of a leadership position by realizing the potential of this capability. ND

James S.B. Chew is chair of NDIA's Science and Engineering Technology Division and group director, aerospace and defense at Cadence Design Systems. Vic Markarian, Cadence senior group director, Tensilica product line; Steve Carlson, systems solutions architect; and David White, technical fellow and senior group director, research and development, contributed to this article.

This article is reprinted from the August 2019 issue of *National Defense*