DOI:10.1145/3588591

BY TAO LUO, WENG-FAI WONG, RICK SIOW MONG GOH, ANH TUAN DO, ZHIXIAN CHEN, HAIZHOU LI, WENYU JIANG, AND WEIYUN YAU

Achieving Green AI with Energy-Efficient Deep Learning Using Neuromorphic Computing

DEEP LEARNING (DL) systems have been widely adopted in many industrial and business applications, dramatically improving human productivity, and enabling new industries. However, deep learning has a carbon emission problem.[a] For example, training a single DL model can consume as much as 656,347 kilowatt-hours of energy and generate up to 626,155 pounds of CO_2 emissions, approximately equal to the total lifetime carbon footprint of five cars. Therefore, in pursuit of sustainability, the computational and carbon costs of DL have to be reduced.

Modeled after systems in the human brain and nervous system, neuromorphic computing has the potential to be the implementation of choice for low-power DL systems. Neuromorphic computing features both neuromorphic algorithms, called spiking neural networks (SNNs), and neuromorphic hardware which are dedicated ASICs optimized for SNNs. Spiking neural networks are regarded as the third generation of artificial neural networks (ANNs), in which spikes (represented by "0" and "1" in the computing system, where "0" means the absence of a spike) are used to transmit information between neurons. With such a spiking mechanism, costly multiplications could be replaced by more energy-efficient additions, mitigating the intensity of the computation. Neuromorphic hardware, on the other hand, has a non-von Neumann "processing in memory" architecture, where computations are integrated into or near a distributed memory architecture. Combined with promising emerging memory devices such as non-volatile resistive and magneto-resistive memories (that is, RRAM and MRAM) to store synaptic weights, both static power and power consumed by data movement are significantly reduced.

Though neuromorphic computing is still in its infancy, the market is expected to grow from ~$200M in 2025 to ~$20B in 2035.[b] Significant efforts have been devoted to neuromorphic computing research. As shown in Figure 1(a), players in both academia, including Stanford, and industry, including Intel and IBM, have developed neuromorphic computing systems.

To promote research in neuromorphic computing, Singapore launched an ambitious program in 2017 covering everything from

a http://bit.ly/3YzaDet

b http://bit.ly/3mDqviU

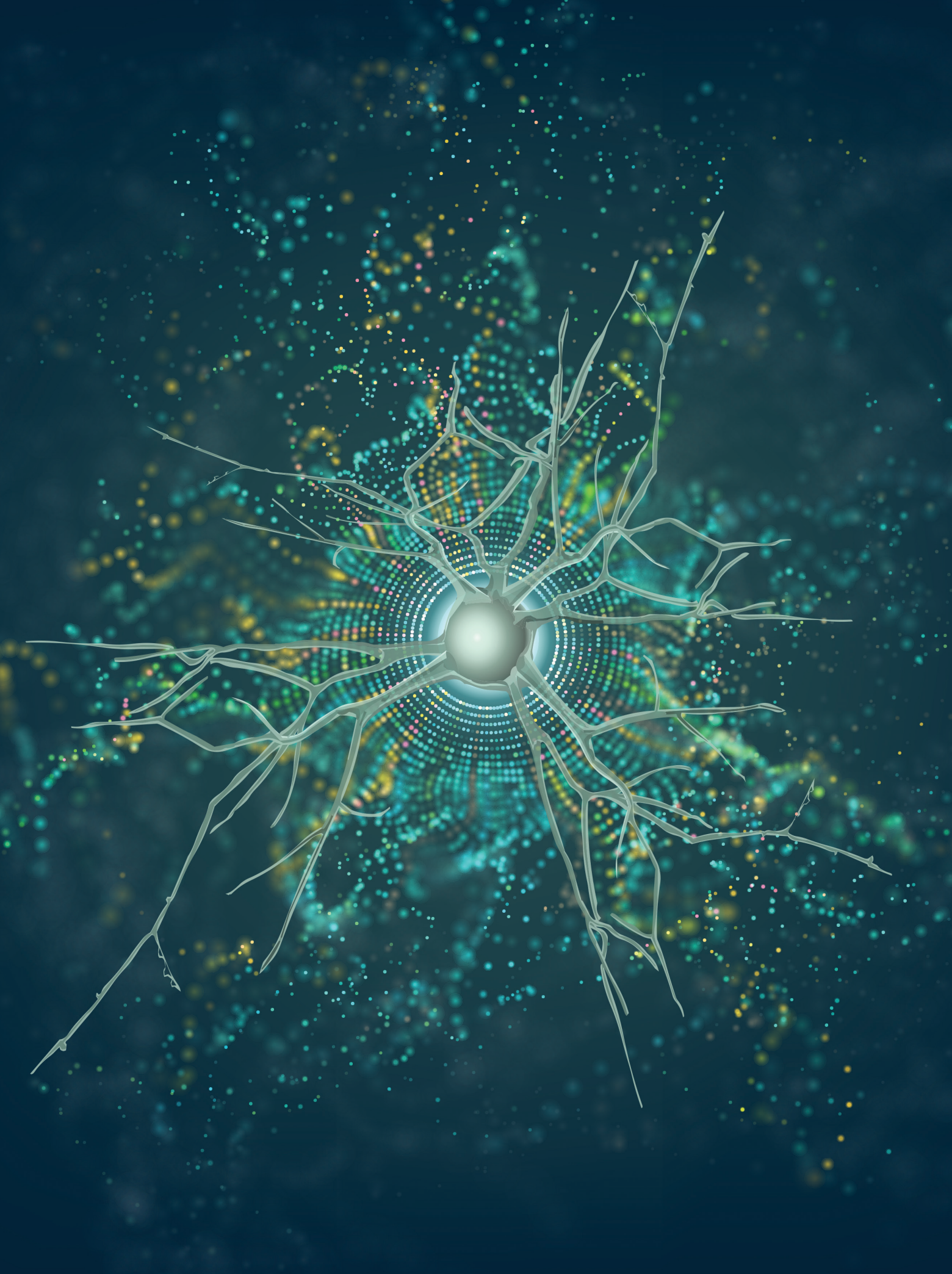
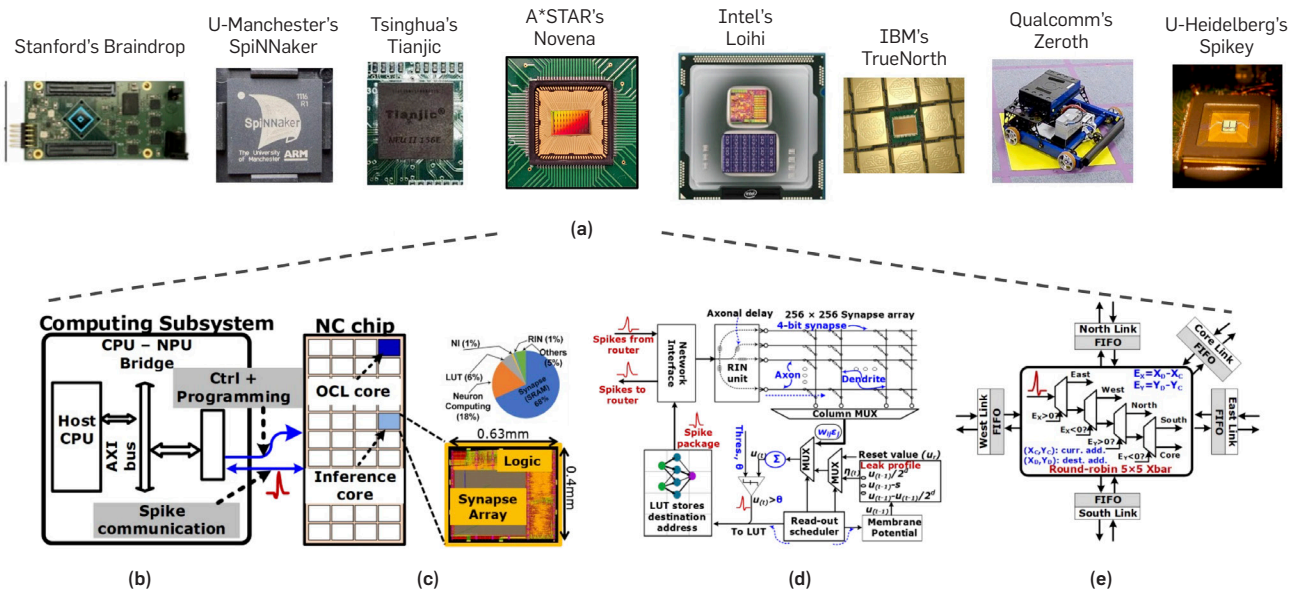


Figure 1. Neuromorphic Computing Projects Zoo.

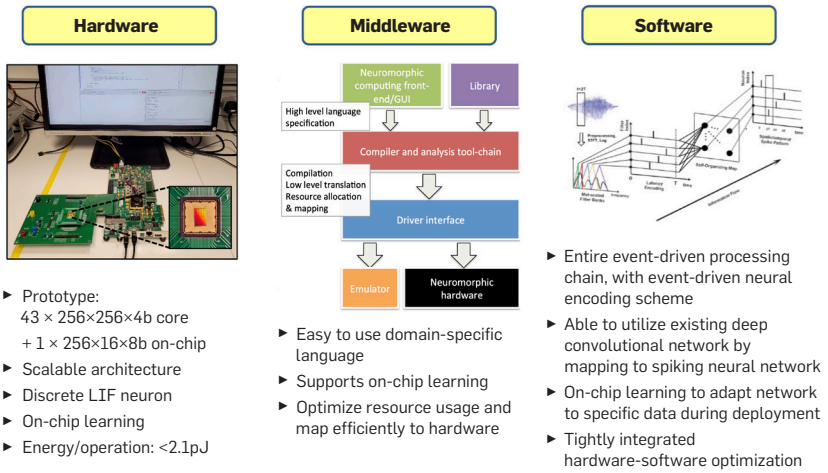
Includes (a) A*STAR's Novena, (b) Overall Novena chips organization and the host CPU Core layout, (c) Area breakdown of each core, (d) Details microarchitecture of each core design, (e) On-chip router supporting 5-input-5-output communication.



hardware to middleware to software, as shown in Figure 2. Through collaborations between the Agency for Science, Technology and Research (A*STAR) and the National University of Singapore (NUS), this program developed an end-to-end neuromorphic computing solution with < 2.1 pJ energy per synaptic operation achieved on our ASIC neuromorphic computing (NC) chip code-named Novena, advancing the current neuromorphic computing research within Singapore.

Designed to develop application-driven solutions for real-world problems, the vision of this program is for every conventional von Neumann computer in the near future to be augmented with a neuromorphic co-processor to handle big data such as text, speech, images, video, and bio-signals that require DL-related solutions. Singapore, as a popular regional hub for datacenters, is working to develop a smart nation while ensuring sustainability as part of its Research, Innovation and Enterprise 2025 Plan (RIE2025). The sheer number of smart devices running DL-related algorithms in datacenters or at edge terminals is expected to have a significant impact on the total carbon footprint of computing. Reducing the energy consumption of such devices is therefore imperative to attain green AI. The project is Singapore's contribution to the advancement of sustainable science and engineering that stands to benefit all of humanity. In this article, we introduce this program and its four aspects, including hardware, middleware, software, and system integration.

Figure 2. Features of the hardware, middleware, and software for the neuromorphic computing program.



- Prototype: 43 × 256×256×4b core + 1 × 256×16×8b on-chip
- Scalable architecture
- Discrete LIF neuron
- On-chip learning
- Energy/operation: <2.1pJ

- Easy to use domain-specific language
- Supports on-chip learning
- Optimize resource usage and map efficiently to hardware

- Entire event-driven processing chain, with event-driven neural encoding scheme
- Able to utilize existing deep convolutional network by mapping to spiking neural network
- On-chip learning to adapt network to specific data during deployment
- Tightly integrated hardware-software optimization

Hardware

The Novena chip was fabricated in a 40-nm CMOS process, occupying a total area of 3.6×5.4 mm^2. Figure 1(b) shows the overall Novena chip[4,5] organization and a host CPU. The CPU configures the NC chip via a separate programming interface. A separate 64-bit bus is used for spike communication. The chip has

44 cores, 43 of which are inference cores while the last is the on-chip learning (OCL) core powered with circuit implementation of an OCL algorithm called Delta Spike Time Dependent Plasticity (Delta STDP). Delta STDP is fundamentally an error-modulated supervised STDP that is customized to perform few-shot learning only on a single layer (the last output layer) of an SNN, where the error is the deviation from target output spike count. Details of the Delta STDP circuits and algorithm can be found in Wong et al.[5] Each inference core has 256 neurons, 256×256 synapses, and 4-bit neurosynaptic weight while the OCL core has 16 neurons, 256×16 synapses, and 8-bits neurosynaptic weight. Details of the area breakdown of each inference core is shown in Figure 1(c).

Figure 1(d) details the neuronal circuit which consists of a 256×256 synaptic crossbar, neuron computation circuit, a look-up table (LUT), and a network interface. Ultra-high-V_{th} is used in SRAM cells, reducing more than 60% of total chip leakage power. To save area, neuron computations such as leak, integrate, and fire are time-multiplexed by a single circuit. To support different applications, neuron leakage profile (that is, fractional or linear) and membrane threshold value θ are configurable. We constrained our fractional leakage to $1-(1/2^d)$ so that a bit-shift operator can be used instead of a full division logic to further reduce power and area. Once the membrane potential u_i of a neuron exceeds its threshold θ, a spike will be generated, and u_i is then reset to a default level which is usually zero. By looking up its address in the LUT, a spike will be sent to the corresponding destination neuron. Output addresses from the lookup table will be queued at the network interface buffer and will only be sent to the corresponding router when it can be assured that no spikes will be dropped.

Figure 1(e) shows the block diagram of the router and neuronal circuit design. The router consists of a round-robin arbiter, interface links, XY routing algorithm logic, and crossbar switch. Each router communicates directly with its own core and sends/receives spikes to/from its four neighbors via its four ports. It is capable of handling spikes as well as debug packets, and can handle indirect addressing for partial summation configurations. The router uses handshaking protocol and thus enables globally asynchronous locally synchronous (GALS) operations for lower power consumption by removing all timing constraints and power overhead on the global clock tree at the top-level integration.

Middleware

Designing a neuromorphic processor with RRAM synaptic memory, on-chip learning circuits, and an architecture that allows system scaling for applications of increasing complexity is not easy. It requires careful software and hardware co-design, with careful considerations to be made on many design choices, with respect to the performance, energy, and area constraints. After the hardware is designed and fabricated, there is also a need for end users to easily program and make use of the chip. To tackle these challenges, in this program, we developed a middleware component to support the development of neuromorphic processors and bridge the gap between applications and the hardware platform.

Figure 2 also shows the architecture of the middleware for the neuromorphic hardware, whose components are specialized to the neuromorphic computing chip. Firstly, a system-level simulator was developed that uses CPUs and GPUs to perform neural core simulation and network-on-chip simulation.[2] It supports simulation of different RRAM material and various RRAM characteristics including stuck-at faults, random telegraph noise, and write variability. The simulator is developed to have high scalability and to support simulation of a neuromorphic chip with up to around 20,000 neural cores that was tested to run on 512 Nvidia A100 GPUs. In addition to the simulator, an FPGA-based hardware emulator for the neuromorphic chip was also developed to accelerate the simulation, which

The sheer number of smart devices running deep learning-related algorithms in datacenters or at edge terminals is expected to have a significant impact on the total carbon footprint of computing.

Surrogate gradient-based learning algorithms have been proposed to resolve the non-differentiable spike function by introducing continuous surrogate derivatives.

achieves ~2,000× speedup compared with multi-threaded execution on the CPU simulator.[3]

Secondly, to allow end users to easily program and use the chip, an end-to-end design framework for neuromorphic computing was developed. It includes a design front-end software compatible with the mainstream design framework including PyTorch and TensorFlow, with extensions to facilitate the design and training of SNNs, and a compilation middleware and analysis tool chain that compiles SNNs from the design front-end to produce configuration data required by our neuromorphic chip. It also optimizes the usage of hardware resources on the neuromorphic chip.[7]

Finally, with the simulator/emulator and end-to-end design framework, software/hardware co-exploration is performed. As the first work of its kind, we successfully tested the ImageNet dataset on a hardware-aware model on our neuromorphic chip architecture using 9,074 neural cores, demonstrating the advantages of our neuromorphic system over the state-of-the-art, achieving promising performance in terms of accuracy, number of neural cores, latency, and energy cost.

Software

Due to the non-differentiable spike function and the complex temporal dependence between spikes, how to efficiently train deep SNNs remains an open question. Various learning algorithms have been proposed. ANN-to-SNN conversion methods transform the knowledge from trained ANNs to SNN counterparts to achieve low-power and low-latency in the inference process.[6] However, ANN-to-SNN methods are unable to yield SNNs that can deal with sequence data processing. Inspired by back propagation through time, surrogate gradient-based learning algorithms have been proposed to resolve the non-differentiable spike function by introducing continuous surrogate derivatives, which however require large computing and memory resources due to frequent update of the synaptic weights at every time step. Spike-driven learning algorithms train the deep SNN in an event-driven manner, in which the spike timing is regarded as a relevant signal for synaptic weights updating.[8]

In this program, we made great progress in training deep SNNs by majorly proposing tandem learning[6] and spike timing dependent back-propagation learning (STDBP).[8] The tandem learning applies rate-based coding and transforms the ANN knowledge to the coupled SNN in a layer-wise manner to reduce conversion errors, achieving competitive performance on both frame-based and event-based datasets. The STDBP places the information in the timing of a single spike, namely temporal coding, and both the inference and learning are in an event-driven manner. STDBP achieves state-of-the-art 99.5% accuracy on the Caltech face/motorbike dataset among spike-driven learning algorithms.

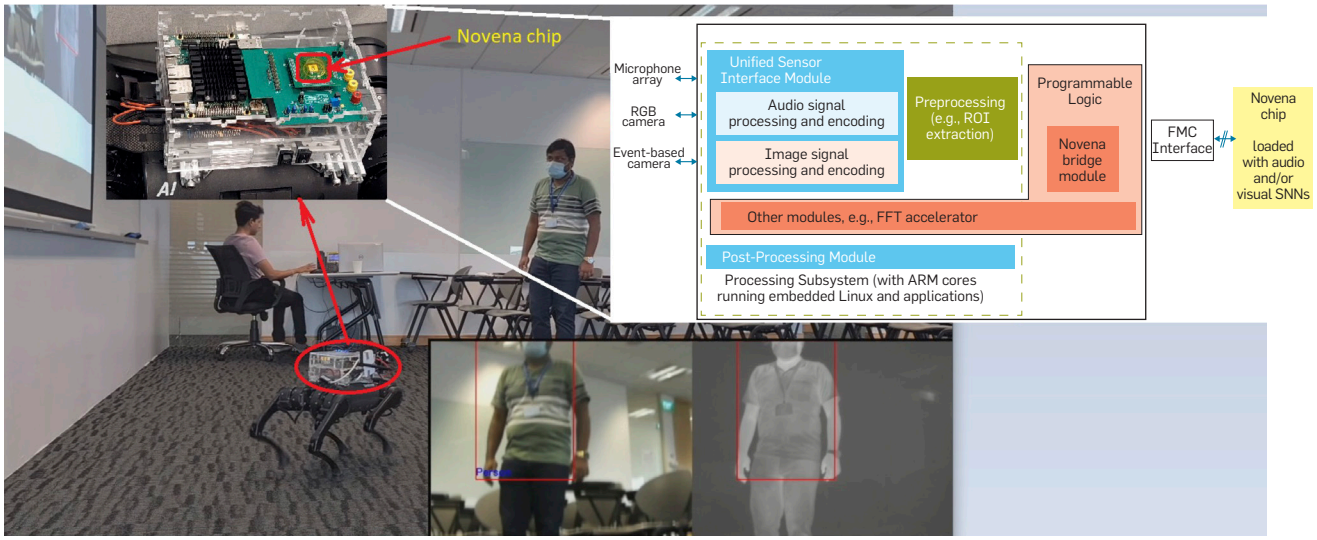
System Integration

To implement a complete neuromorphic computing system, the neuromorphic chip needs to be integrated with a host processor, along with the required sensors. We have chosen to use an FPGA board with ARM cores as the host processor.

First, the FPGA's programmable logic (PL) is programmed with a bridging module to interface with the Novena chip over an FPGA Mezzanine Card low-pin count connector. The PL may also contain other modules such as FFT library implementation to accelerate certain pre-processing steps that would otherwise take considerable cycles in the ARM cores.

Second, the ARM cores, as part of the FPGA processing subsystem, run embedded Linux, which facilitates sensor integration thanks to the availability of device drivers for a wide range of commercial sensors including some event-based sensors. The sensors are then consolidated under a unified sensor interface module, followed by further pre-processing as necessary. One example of pre-processing is to perform region-of-interest (ROI) extraction such as on event-based visual inputs,[1] so as to reduce input dimension and allow

Figure 3. Example neuromorphic computing system using a legged robot as mechanical platform and thermal + RGB camera-based human detection.



resource-efficient implementation on the neuromorphic chip.

Furthermore, a system software programming framework based on a Xilinx SDx environment has been developed to streamline communications between embedded Linux and FPGA PL and to facilitate interfacing with Novena from embedded Linux applications, so that the developer need not be overwhelmed by the low-level details of the Novena bridge interface module in FPGA PL.

Once the Novena chip returns computed outputs, a post-processing module may be used, for example, to further filter outputs over time and improve final accuracy. An illustration of the overall neuromorphic system diagram is shown in the white inset of Figure 3.

Finally, to showcase the capabilities of neuromorphic computing, several demonstrator applications have been built, including live keyword spotting (KWS), live gesture detection at variable distances,[1] and human plus body-part detection for search and rescue applications. Among these, KWS and human detection demos were integrated onto a legged robot, where an operator can use keywords to guide the robot, with the legged robot using a combination of thermal camera and RGB camera where thermal images were used for ROI extraction based on hotspots, followed by human and body-part classification on associated RGB image ROI patch. A snapshot of the human detection demo is shown in Figure 3, along with a zoomed-in view of the FPGA board and Novena chip.

Conclusion

This article gives a description of the six-year effort to develop an end-to-end neuromorphic computing solution in Singapore. From next-generation memory devices, chips, algorithms, middleware, to applications, this ambitious program covered all aspects of neuromorphic computing. Currently, the program is looking into the commercialization of its innovations, especially for energy-efficient edge AI.

Acknowledgment. This work was supported by the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain) under Grant A1687b0033. C

References

1. Fabien, C. et al. Event-based visual sensing for human motion detection and classification at various distances. In *Proceedings of Pacific-Rim Symp. Image and Video Technology*, 2022.
2. Lee, M.K. F. et al. A system-level simulator for RRAM-based neuromorphic computing chips. *ACM Trans. Architecture and Code Optimization 15*, 4 (2019), 1–24.
3. Luo, T. et al. An FPGA-based hardware emulator for neuromorphic chip with RRAM. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems 39*, 22 (2018), 438–450.
4. Nambiar, V.P. et al. Energy efficient 0.5 V 4.8 pJ/SOP 0.93 W leakage/core neuromorphic processor design. *IEEE Trans. Circuits and Systems II: Express Briefs 68*, 9 (2021), 3148–3152.
5. Wong, M.M. et al. A 2.1 pJ/SOP 40nm SNN accelerator featuring on-chip transfer learning using Delta STDP. In *Proceedings of the IEEE 51st European Solid-State Device Research Conf.* 2021.
6. Wu, J. et al. Progressive tandem learning for pattern recognition with deep spiking neural networks. *IEEE Trans. Pattern Analysis and Machine Intelligence, 44*, 11 (2021), 7824–7840.
7. Yang, L. et al. Coreset: Hierarchical neuromorphic computing supporting large-scale neural networks with improved resource efficiency. *Neurocomputing 474* (2022), 128–140.
8. Zhang, M. et al. Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE Trans. Neural Networks and Learning Systems 33*, 5 (2021), 1947–1958.

Tao Luo is a research scientist at the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), in Singapore.

Weng-Fai Wong is an associate professor in the School of Computing (SoC), at National University of Singapore (NUS), Singapore.

Rick Siow Mong Goh is a research scientist at the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), in Singapore.

Anh Tuan Do is a research scientist at the Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), in Singapore.

Zhixian Chen is a research scientist at the Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), in Singapore.

Haizhou Li is a professor in the School of Computing (SoC), at National University of Singapore (NUS), Singapore.

Wenyu Jiang is a research scientist at the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), in Singapore.

Weiyun Yau is a research scientist at the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), in Singapore.