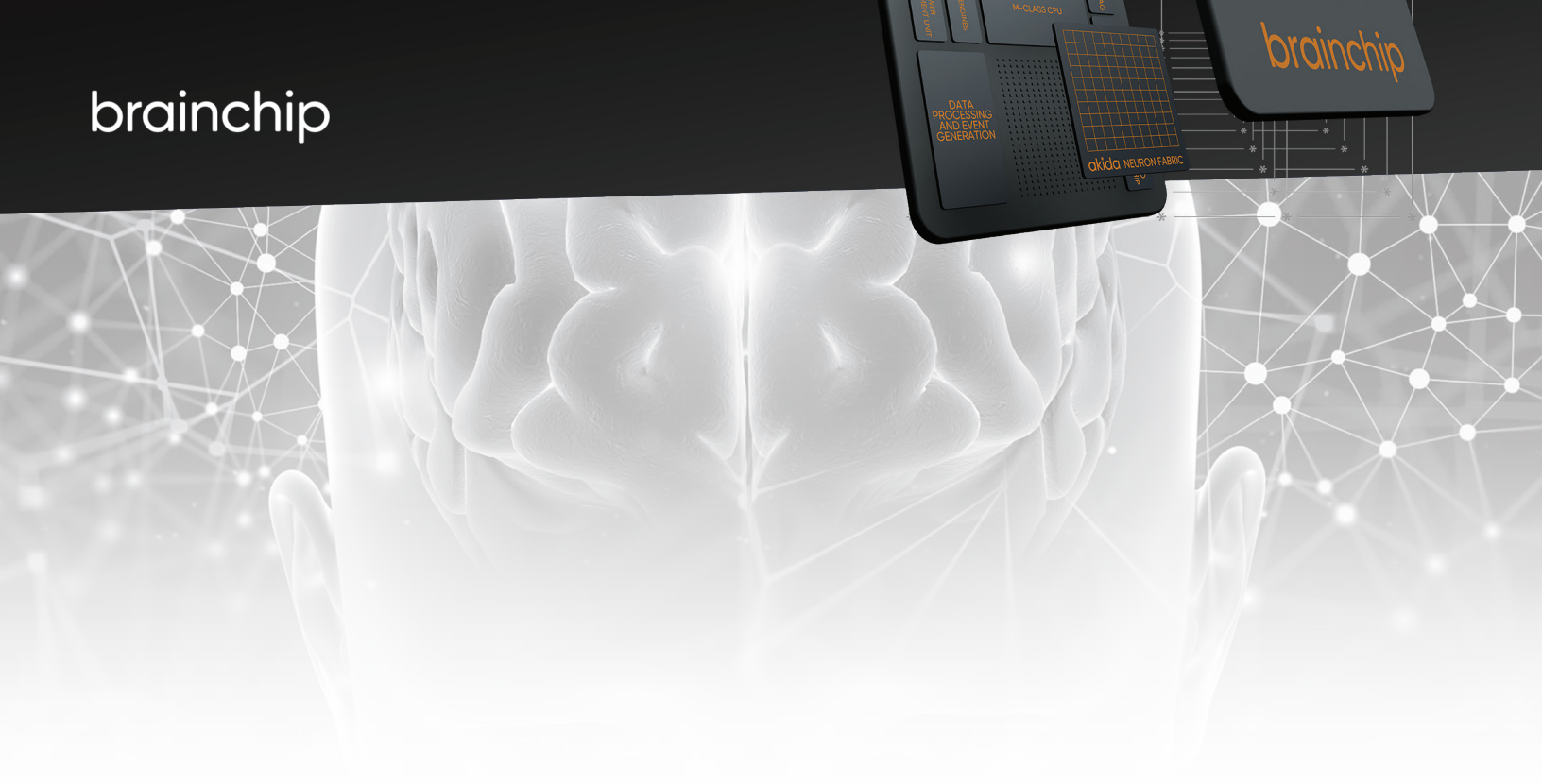
brainchip

The advantages of Unified Deep Learning technology

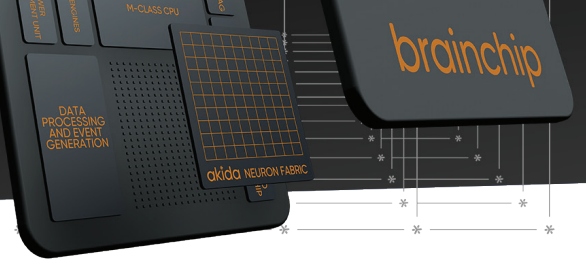


Executive summary

The growth of Artificial Intelligence (AI) has been rapid, with the market heavily focused on training large generative models, such as Open AI's chatGPT. Deep learning models currently rely heavily on neural networks (NNs), which require high-power, massively parallel computing to calculate the millions of values needed for each inference instance. This results in the cost of training a model of such a large caliber requiring millions of dollars in computational power.

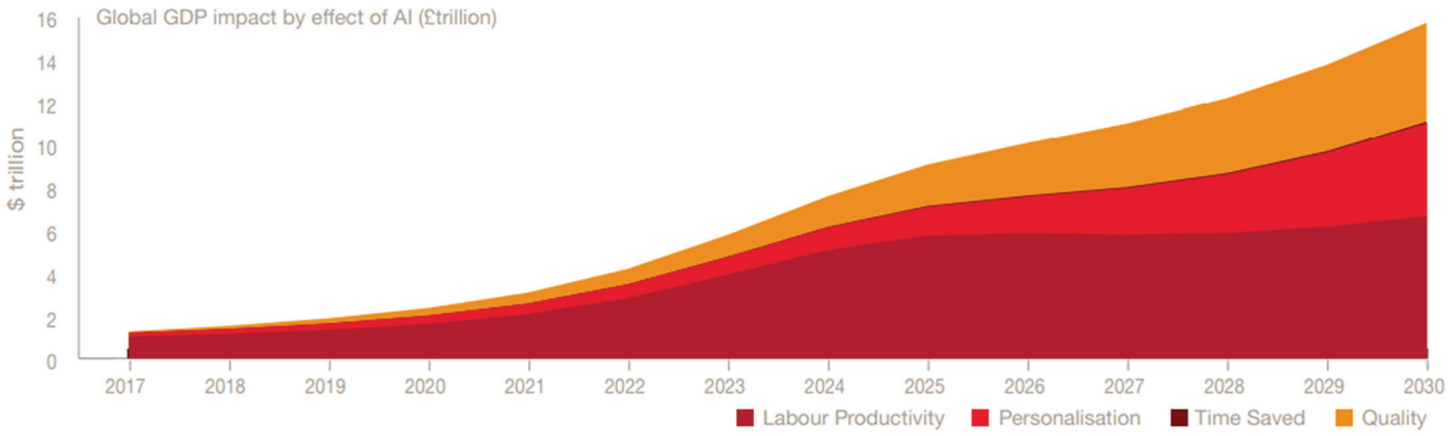
This paper explores an alternative, neuromorphic computing, which can be up to 1,000 times better performance and 10,000 times better efficiency compared to traditional high-performance computing hardware, such as CPUs and GPUs. Neuromorphic computing also reduces the need for high-power cloud inference, raw data traffic, and congestion in networks. However, there are very few Spiking Neural Network (SNN) models that are ready for use on the market.

BrainChip has developed the Akida technology, which combines convolution functions efficiently with a fully digital, neuromorphic computing core. The Akida technology is capable of executing most deep learning networks and performing inference at an energy cost that is a fraction of conventional solutions, such as convolutional neural networks (CNNs) and deep neural networks (DNNs). The radical energy efficiency of neuromorphic computing opens the capability of learning in real-time and in the field, enabling immediate customization of AI-enhanced products.



Introduction

AI is a technology paradigm that continues to rapidly penetrate everyday products, making them smarter, safer, easier to use, and more reliable. AI will continue to enhance the function of medical diagnostic devices, transport systems, security systems, home appliances, industrial robots, industrial safety, and quality assurance systems. Studies predict the annual (positive) impact of AI on world GDP will top $15T in 2030.

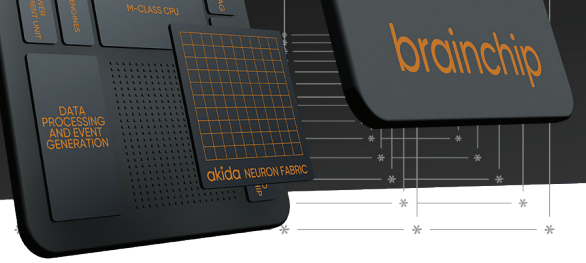


Labour productivity improvements are expected to account for **over 55%** of all GDP gains from AI over the period 2017 - 2030.

As new technologies are gradually adopted and consumers respond to improved products with increased demand, the share of impact from product innovation increases over time.

58% of all GDP gains in 2030 will come from consumption side impacts.

Fig.1 Global GDP impact due to AI (source: PWC analysis)
Source: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence.html



Where we are today

The most prevalent branch of artificial intelligence is machine learning. Deep learning, a subset of machine learning, gives machines an amplified ability to find and enhance even the smallest patterns within data. At the heart of deep learning algorithms are neural networks, which need to be trained with millions of examples and infer patterns from that data. For example, a model could be shown many images of a cat. Once well-trained, the model would be able to identify an image of a cat on its own. The most common type of neural network used to process visual data (i.e. videos, pictures, images, etc.) is a convolutional neural network (CNN).

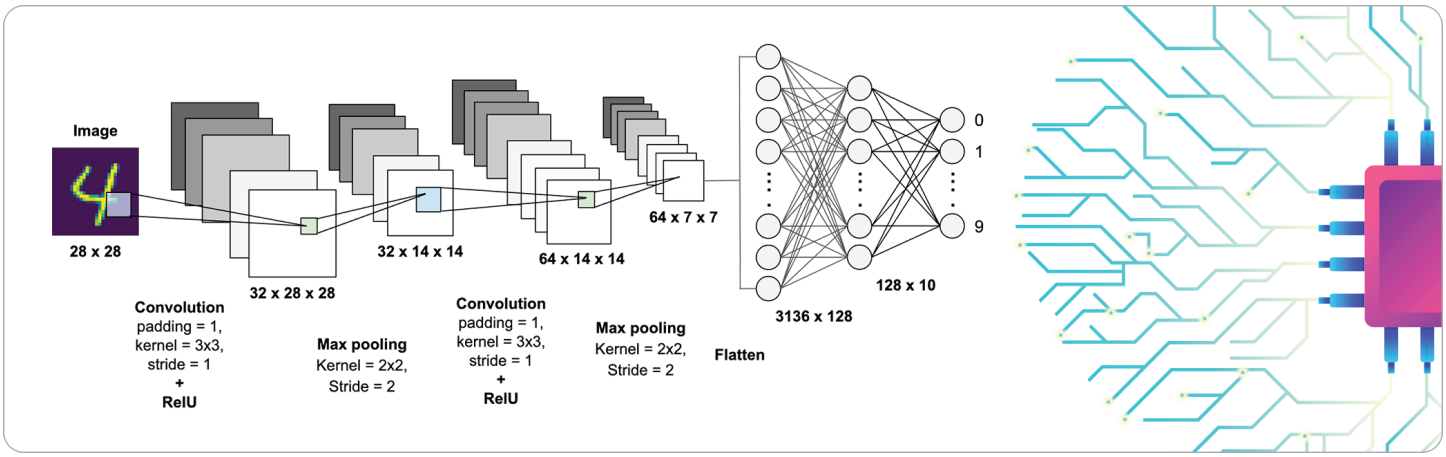
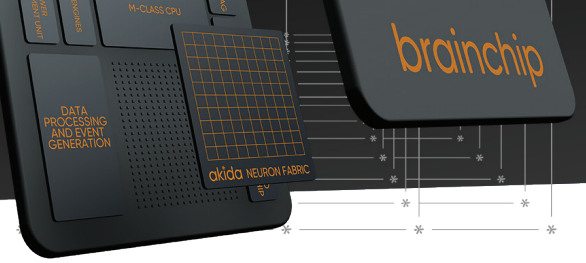


Fig.2 CNN training

Today's neural nets are able to take advantage of massively parallel processors due to the nature of these mathematical calculations. Each inference requires doing intensive computations and neural nets have reached high accuracy levels, but at the cost of high energy consumption, excessive heat generation, and a large environmental impact.

At the same time, there have been several significant commercial advancements in generative AI. Generative AI is a type of artificial intelligence capable of creating new and original data (i.e. images, music, text, videos, voice, etc.) that is extremely similar to data generated by humans. Generative AI models rely on massive amounts of training data, such as all of Wikipedia or all of Reddit. Training one generative AI model is done in the cloud and can cost millions of dollars. Hence, the model is not often completely retrained.



While the total cost of training the model costs tens of millions, running the model is the true expense. Although the computational power required for answering and responding to queries (i.e. each inference) is much less, it is necessary each and every time a user needs the service. Hence, inferences are used billions of times more. Running OpenAI's chatGPT reportedly costs the company $100,000 USD per day. "Deploying current ChatGPT into every search done by Google would require 512,820 A100 HGX servers with a total of 4,102,568 A100 GPUs," writes Dylan Patel and Afzal Ahmad in SemiAnalysis, "The total cost of these servers and networking exceeds $100 billion of Capex alone, of which NVIDIA would receive a large portion."

Given that AI processing is only going to grow, the environmental and commercial costs can become prohibitive, creating a serious challenge to the broader proliferation of AI.

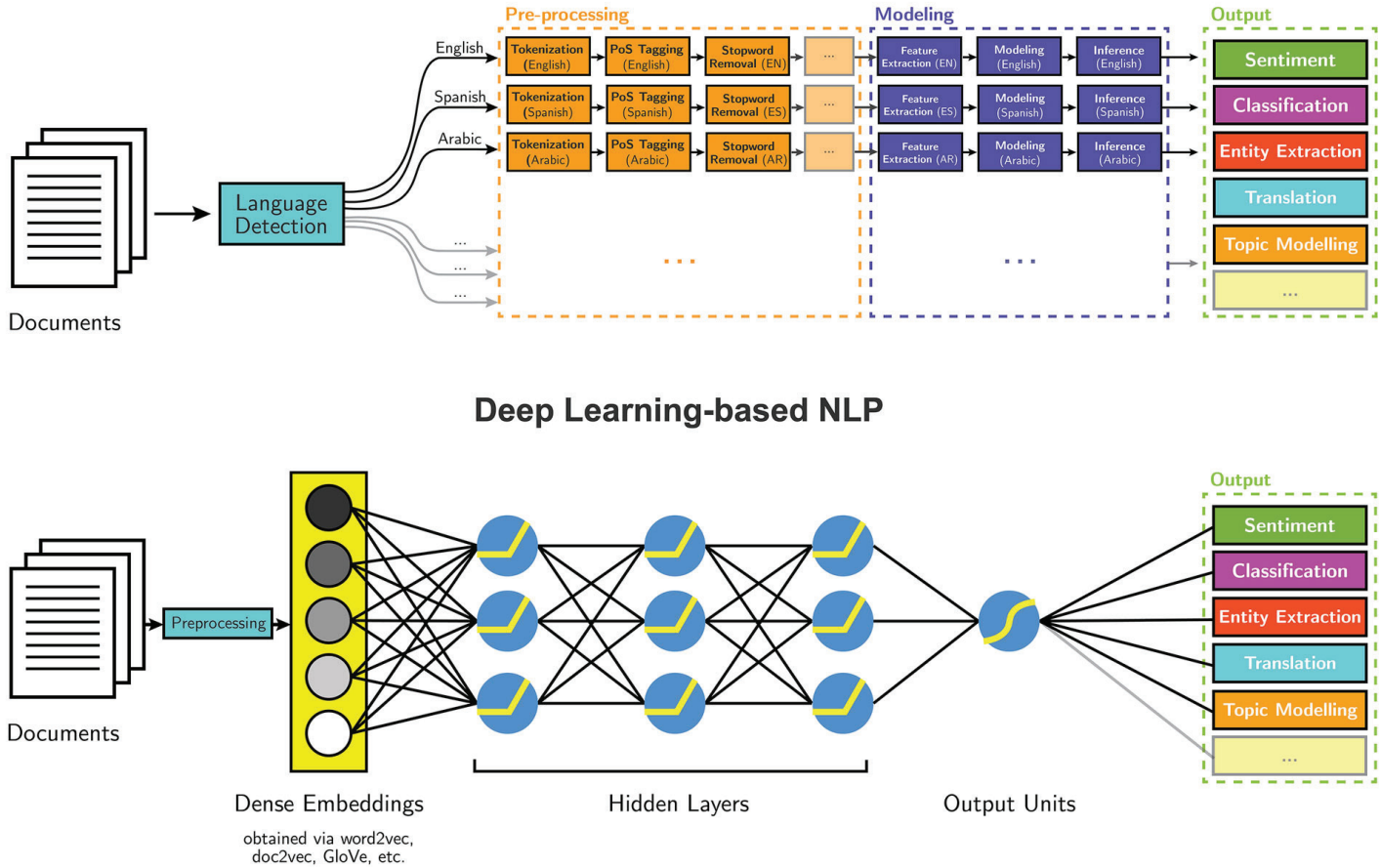
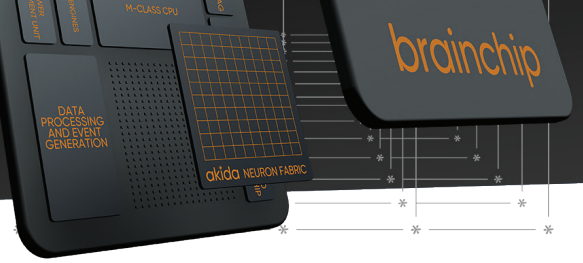


Fig.3 NLP models growing



Addressing the challenge

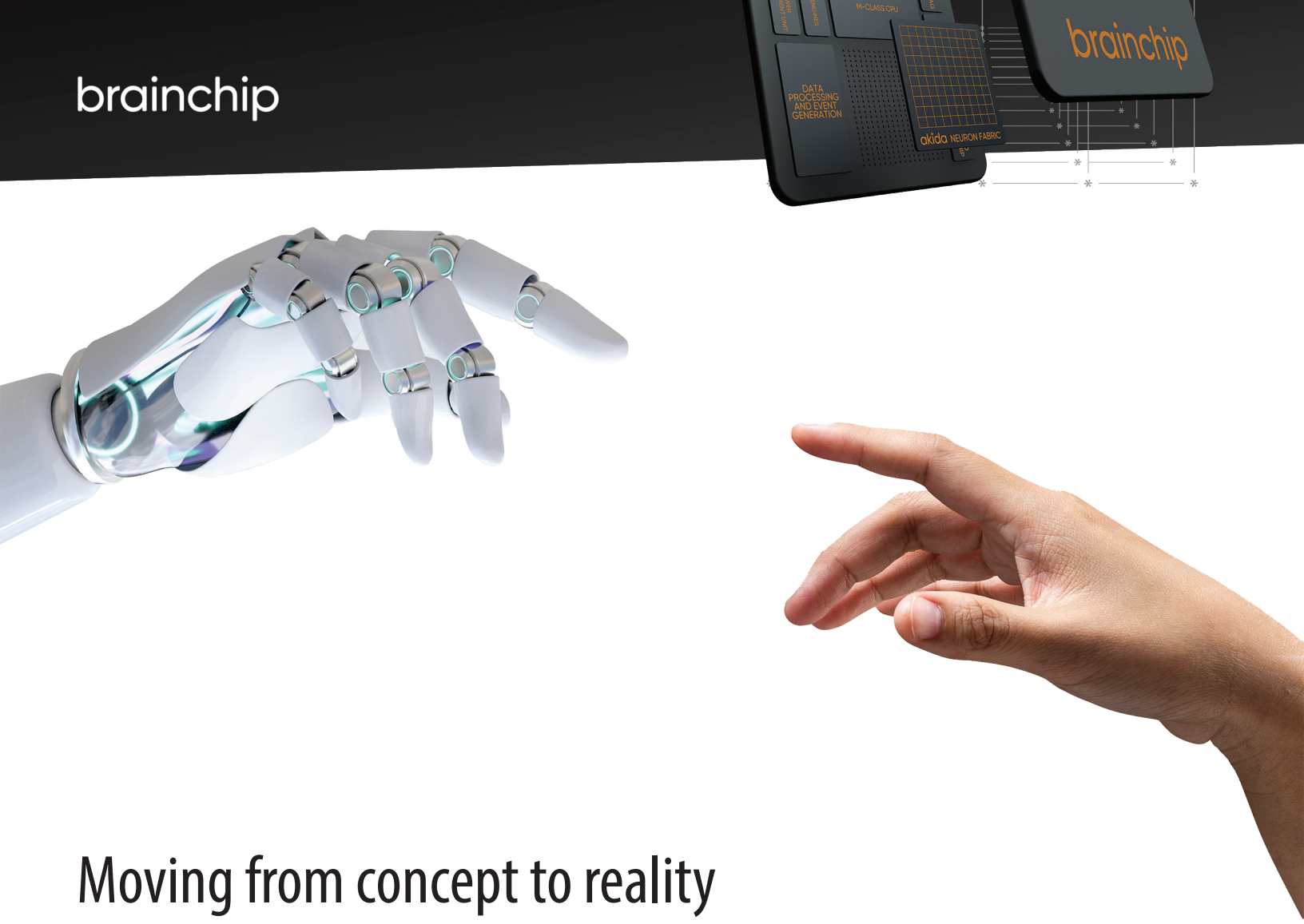
One alternative approach is to accurately mimic the brain, the most efficient "computer" known to humans, using a method called neuromorphic computing. Neuromorphic computing gains its performance power from its unique organization. According to an article in Nature Computational Science[1], it offers up to 1,000 times better performance and 10,000 times better efficiency compared to high-performance computing. The concept is to implement the ACH brain cell (i.e. a neuron) in a hardware circuit that is connected to all other cells but operates in parallel. The software models for this type of computation are called Spiking Neural Networks (SNNs), which are a type of artificial neural network that is inspired by the all-or-nothing way neurons communicate in the brain. In comparison to traditional neural nets, which use continuous valued signals, SNNs use discrete, binary signals.

This means that SNNs use energy only when they're active, which results in much higher performance at a low energy cost, in the range of 1/1000th of a watt.

Decades and hundreds of billions of dollars have been invested in CNN and DNNs, but very few ready-to use SNN models exist in the industry. The advantage of neuromorphic computing is that it can be energy efficient enough to do much more accurate inference in devices at the Edge, such as smart sensors, smart cameras, and smart meters. SNNs reduce the need for high-power cloud inference. They also decrease raw data traffic and congestion in networks. This also reduces security risks from exposing sensitive data as well. Neuromorphic computing overall leads us toward more distributed computing, where localized AI contributes to and enables a leaner, more efficient global AI.

Many industry leaders realize that neuromorphic computing is the future. Companies such as IBM™ and Intel® have already invested in neuromorphic technology and demonstrated working chips. However, these investments still focus on accelerating SNN models which are, at the moment, academic. The actual tools and deployment ecosystems are in their infancy. This makes using the prototypes difficult, if not impossible in real-world deep learning CNN environments. Developing real-world applications is complex and has plagued the adoption of neuromorphic technologies.

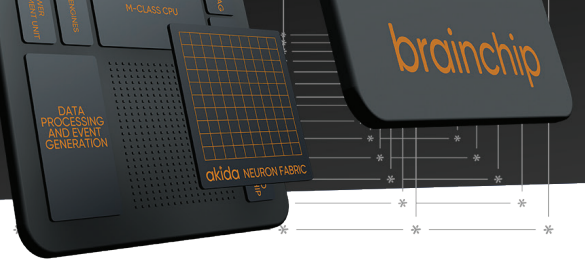
1 Nature computational science. Published: 31 January 2022: "Opportunities for neuromorphic computing algorithms and applications" by Catherine D. Schuman, Shruti R. Kulkarni, Maryam Parsa, J. Parker Mitchell, Prasanna Date & Bill Kay https://www.nature.com/articles/s43588-021-00184-y



Moving from concept to reality

BrainChip identified this challenge early on while developing its neuromorphic offering. The team devised a solution that combined convolution functions efficiently with an already hyper-efficient, fully digital, neuromorphic computing core. This unified product delivers the best of both worlds: offering the accuracy and capabilities of the vast, existing deep learning ecosystem, and the radical energy efficiency of neuromorphic computing. The BrainChip Akida technology can execute most deep learning networks and perform inference at an energy cost that is a fraction of conventional solutions.

A unique aspect of unified Akida technology is that the neuromorphic level can learn in real time and in the field (i.e. without having to do expensive retraining in the cloud), enabling the fast customization of AI-enhanced products. The convolutional layers are used as feature extractors, while the last layer is learning within milliseconds to combine these features into new objects. This customization also prevents the exposure of sensitive information, providing the benefit of enhanced privacy and security.



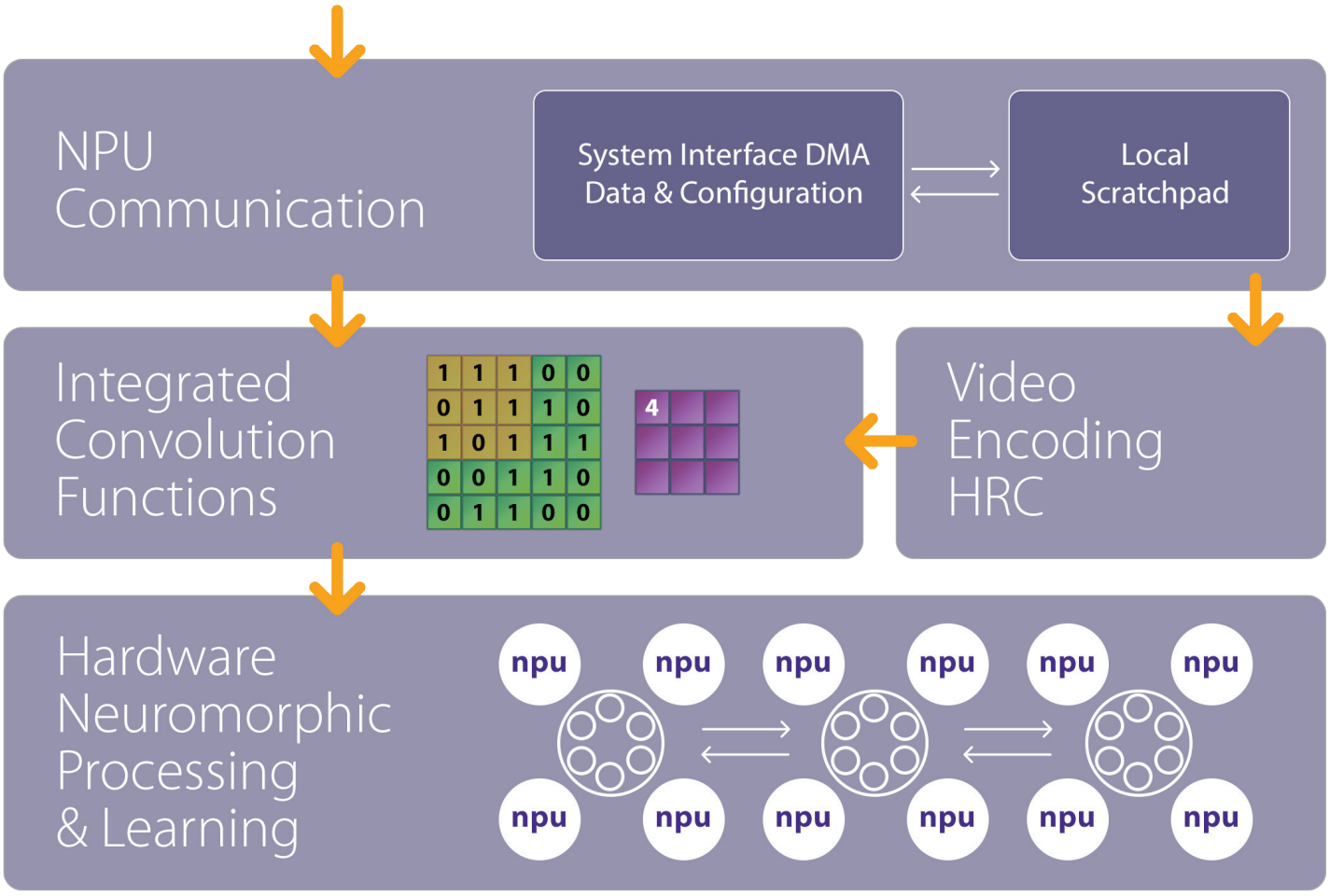


Fig.4 The all-digital Akida IP and chip combine the best features of both CNN and neuromorphic processing, incorporating in hardware several different convolution sizes, configurations, and methods, including pooling and ReLU. The network is executed in an array of digital neuron cells, thereby preserving MCU cycles.

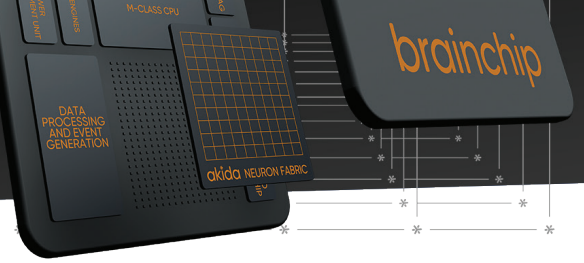
The beauty of the solution is in its implementation. BrainChip has built multiple silicon platforms to enable the development of unified deep learning solutions. More importantly, the solution is focused on simplifying the model and software development with MetaTF™ for developers in their preferred environments, along with many common models already available in the Akida Models Zoo. The Brainchip Akida architecture is mature and proven in silicon, with major corporations already having licensed the IP.

Fig.5 The BrainChip AKD1000 evaluation board contains the Unified Convolutional Neuromorphic chip. BrainChip also provides a free tool called MetaTF™ that integrates seamlessly with popular frameworks to allow developers to build and deploy AI smoothly. More information is available on BrainChip's Website. https://brainchip.com/akida-enablement-platforms/

Looking ahead

Neurons in the brain operate in the analog domain, integrating thousands of input signals over time and outputting a single bit "spike" when the input signals match stored parameters. The brain contains billions of such cells organized in a functional structure that can do amazing things that are still beyond today's Artificial Intelligence capabilities[2]. The brain functions on the equivalent operating power of approximately 230 picowatts (2.3 * 10-10) per cell, or around 20 watts for the entire brain. Any analog function can be simulated in a digital circuit. Imagine, for instance, the evolution of music recording from analog tapes and records to digital MP3 and downloadable music files. The Brainchip Akida technology is accomplishing the same feat by bringing the elegance of biological neurons into an entirely digital world.

2 Watson, D. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. Minds & Machines 29, 417–440 (2019). https://doi.org/10.1007/s11023-019-09506-6



The neuromorphic approach has many advantages beyond the extremely low power consumption in processing. There is no need for a fan to cool the device, which is typically necessary for Edge AI. The design is very stable due to its digital nature, enabling the reuse of training data between devices. All devices are identical without discrepancies in process properties and drift seen in analog technologies. The on-chip learning capabilities available through the neuromorphic core can be combined with deep learning derived features to provide in-the-field reconfigurability.

BrainChip continues to look ahead, and the fundamentals of neuromorphic technology have strong benefits in more advanced learning with lesser data, like the human brain. This can significantly scale future training and inference.

The unified approach is a launchpad from perception to cognition, which AI may use to accelerate from machine learning to machine intelligence in the march toward Artificial General Intelligence (AGI).

Contact for more information:

23041 Avenida De La Carlota, Suite 250
Laguna Hills CA 92653

+1 949 784 0040
(United States)

sales@brainchip.com