Deep Science Publishing
https://doi.org/10.70593/978-93-49910-47-8_2

Chapter 2: Exploring cutting-edge chip design architectures built specifically for artificial intelligence and machine learning applications

2.1. Introduction to Chip Design for AI

AI's most important driver is the Knowledge Explosion resulting from the internet. There is an increasing demand for AI with better cognitive functions to assist human endeavors, such as scientific discovery and analysis of massive data sets that cannot be intuitively understood. There is an urgent need to synthesize AI hardware and algorithms to deliver brain-like real-time intelligence. Neuromorphic computers would exponentially increase the varieties and the efficiency of AI applications. Brain-like devices that power AI algorithms would have a size compressed by several orders while simultaneously consuming several orders of magnitude less energy. Advanced synthesis tools could be developed to assist scientists generating hypotheses (Krishnamoorthy et al., 2023; Miller et al., 2023; Nagar et al., 2024).

They must be able to learn complex relationships from structured data, continuously update in response to new experience, and generalize from this data to "see situations" for which they have no previous data. Nascent efforts in biologically plausible AI have thrown a spotlight on a much wider range of resources, from non-ideal silicon devices at the sub-nanometer scale to large-scale nanotechnology, NEMS-based sensors, microfluidics, and ultra-microchips that orchestrate interactions among vast ensemble states of matter. However, this makes real-time learning with these devices a nontrivial task and implies the need for smarter ways to train, program, and implement such algorithms.

Existing hardware-nearing-algorithms efforts are generally limited to AI chips that are variations of an NVIDIA GPU. Limited biophysical restrictions of existing AI chips

mean a constrained biophysically inspired AI SoC architecture, such as one based on CMOS micro-continuous-time reservoir computing, cannot be readily explored. Digital neural chips are the most established AI chip implementations for practical use. They can realize the densely connected networks and very high TP with about 250 pJ energy cost per MAC ops. Compared with this, it is difficult and appears impossible to implement RNN or spiking neural networks (SNN)-through VLSI circuits (Patra, 2024; Rane et al., 2024; Wang et al., 2024).

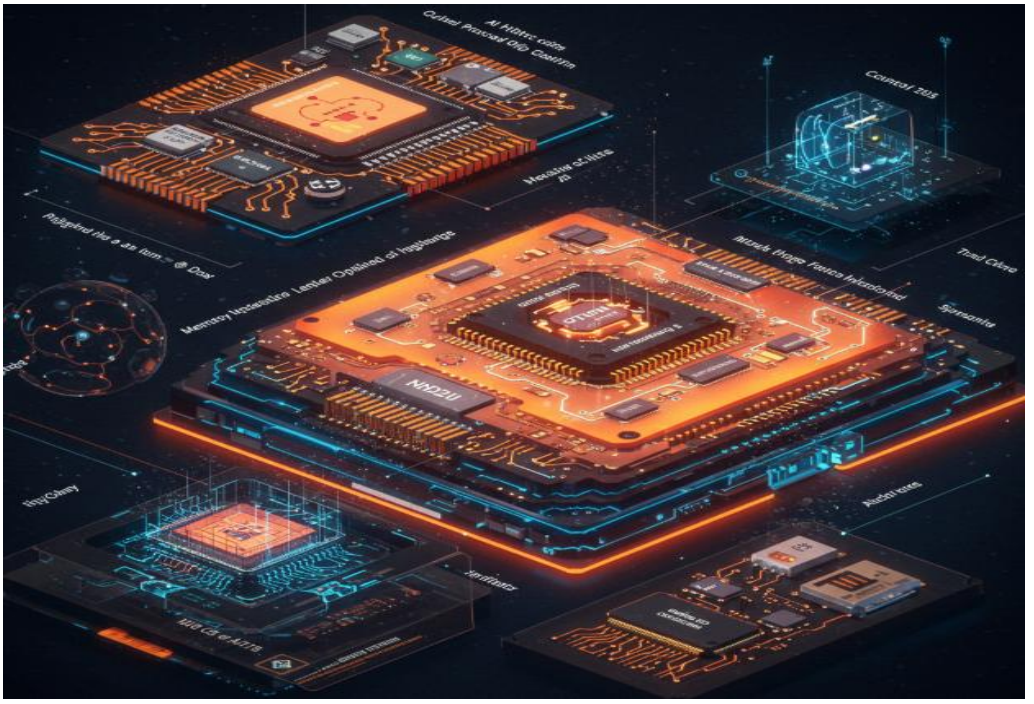

Fig 2.1: Cutting-Edge Chip Design Architectures for AI and Machine Learning Applications.

2.1.1. Background and Significance

Dramatic changes in chip design and technology are required to fulfill the rapid demand for high-performance chips for AI applications. These chips must provide high performance, high performance/watt, low latency, and low form factor to implement and execute complex machine learning algorithms locally on the chip and in real time. However, the complexity and 3D integration need new chip design approaches. A genuine AI-based design, optimization, and verification of these chips needs to be developed to explore these enormous design spaces, flows, and algorithms that traditional and scripted tools cannot handle. 50+ years of development and billions of dollars have gone into making these tools, and they work admirably and cope with

standard optimization problems in depth and width to design various chips. But there still are some challenges in the architecture, flow, and algorithm level that AI can tackle well.

The VLSI domain is full of opportunities for applying AI/ML ideas and algorithms. The high-level specification of the design is made in abstraction/implementation languages or in textual languages. This specification is transformed into a circuit design, involving numerous operations on the various design parameters, design representations, and design types. One of the several components is a synthesis tool that transforms the design specification from one representation/level to another one. Design representation includes RTL, gates/middle-end, and layout. The transformation could be high-level synthesis that transforms from a behavioral to an RTL representation. Or it could be a logic/design synthesis tool that takes an RTL representation and performs a wide variety of operations to transform the RTL code into a gate level netlist.

Real-life designs are very complex with 10Ms of lines of code. The design space for these algorithms flows and tools is massive with > 1Bn degrees of freedom. Once represented in a standardized format, these design spaces would be > 10^30 states. In the past decade, there have been significant advances in some of the above components and belief that by the end of this decade, a full/chip SOC design can be done in a few minutes—massive opportunities for developments either by academia, industries for a wider commercial availability and exposure in a myriad of very useful chip designs.

2.2. Historical Overview of Chip Design

Artificial Intelligence (AI) is one of the core challenges of the 21st century. There are plans of imitating the human brain through Artificial General Intelligence (AGI). At present, the following functionalities of the human brain are highly desired: perceptive, planning, reasoning, decision making, and understanding. The ability to understand knowledge is one of the key characteristics that distinguishes the human brain from today's AI systems. The understanding of chip design development is still in its infancy. The present status and trends of three of the many required chip designs are presented – perception, decision making, and reasoning/interpretation. Three specific interesting implementations of chip designs are presented as illustrations. Much progress has been made on emulating the way how the brain perceives information through hardware and algorithm developments, building on biological knowledge of the visual processing pathway from retina to the human brain. Current imitated functionality includes edge detection, brightness detection, motion detection, character recognition, proactive event detection, face detection, and facial recognition. Hardware developments can be lumped into two categories: silicon-based and non-silicon-based. The silicon-based implementations tend to be fully digital or hybrid (parts are digital and parts are analog).

Fully digitized chips include various digital neural chips. Most of the present non-silicon chips are analog and digital kinds with signal processing based on physical properties. Digital neuromorphic chip design implementation allows scaling for very large networks with tiny synaptic weights and a vast diversity of models. The integration within widely accessible and energy-efficient neuromorphic hardware platforms helps understanding the brain by both exploratory research and computationally sophisticated learning on large data sets is crucial.The research primarily focuses on the implementation of AI/ML algorithms in generating graph neural networks, netlist generation, performer-based synthesis, accuracy-aware placement, clustering with a graph neural network, and end-to-end analog design using multi-stage transformers. Initial perspectives on directions for designing integrated circuits tailored to the unique requirements of AI applications, including rising areas such as logic-in-memory and neuromorphic computing, are offered. AI/ML techniques operating directly on ASIC netlist files and enhancing design under consideration expands the opportunities for synergy between itself and electronic design automation (EDA).

2.2.1. Research design

To provide a comprehensive overview of candidate architectures for AI and machine learning (ML), existing research in AI/ML algorithms and applications in VLSI design and technology is reviewed. The discussion highlights how AI/ML techniques can address challenges in various VLSI design and technology domains such as architecture, placement, routing, manufacturing, and reliability. In addition to the algorithms themselves, design flow and tool support are also reviewed. Furthermore, an overview of research in deep learning (DL) algorithms, architectures, and applications focused on chip design is provided. A higher-level overview of the chip design design-space exploration space and a systematic analysis of AI and DL for chip design architectures is also analyzed.

Discussion is given about key applications of hardware accelerators for AI. Preliminary perspectives on the demand for and global-state-of-the-art of AI chip architecture, implementation, and design are provided based on the three important application fields of AI—vision, language, and science. The current design methods of AI chips mainly follow one of the two paradigms: designing holistic architectures with high-level comprehensive design from the scratch or applying required modifications to the existing architectures directly. To bridge these two paradigms with the development of advanced DL algorithms, the process of architecture generation via multi-stage auto-ML is proposed.

2.3. Key Components of AI Chip Architectures

With the recent explosion in AI and ML applications, the need for real-time inference and learning puts existing architecture for computing under pressure. While edge devices perform inference on pre-trained models, the training of those models is latency and power expensive processes done on expensive cloud servers. On existing chip designs, networks are run as static calculations. An entirely different chip architecture is required for computing the updates to the weights of the networks, as opposed to running them. Said chip architecture must therefore have a real-time response to inputs that presides over a change in the output, with this sensitivity to recent inputs potentially being implemented with emerging devices using non-equilibrium steady states. The computing and transmitting units of the trained models would also need reconvagation, since the gradients that update weights require completely different calculations than what was needed to evaluate the pre-trained models. Digital neural chips are the most established AI chip implementations for practical use. Input spikes are discretized down into 1s and 0s using an ADC, and then grouped as words before transfer to the neuron array. The synaptic nodes become purely digital when the 'multiply and accumulate' (MAC) operation is implemented using digital logic. In other words, all inputs and weights become either 1 or 0, and the neurons only fire if the total spike for the time in crosses a threshold. Only about 10% of area in the chip is occupied by neurons, while the rest is occupied by memory units and control circuits. The developed chip design has good propagation latency, as the dispersed circuitry minimizes redundant paths that delay same-value bits. Power dense neurons are desirable, due to their combination of non-linear computing and good scaling behaviour at large counts. Existing implementations, however, have an inherent trade-off between precision and latency. The input spikes need to be discretized first by an ADC before the neural calculations are run. Another emerging system consists of carbon nanotubes (CNT) along with thin film transistors (TFT) to build synapses for the neurons to communicate read forward in time. This approach shows the possibility to have high package densities that could be used to scale up the neural network implementations. SNAP shows sparsity at edge classification tasks where most weights are set to zero, but significant sparsity is also found at the inputs. This may lead to a higher need for extra circuits to prevent artefacts in the learned weights. Basing neurons on published hardware demonstrators would allow the use of exponentially stochastic states and good scaling thereof, but on-chip learning wouldn't be plausible and initial networking would be characteristic of random waffle rather than coherent. Another proposed option is the use of semiconducting nanowires that demonstrate non-volatile memory (NVM) properties. These devices can be configured to have multiple resistance states that would allow them to act as low-power memristors within the reservoir. The average voltage output of the aggregated photon spike inputs caused variations in the mean resistance of the five nanowires (from 2.4 kΩ to 3.5 kΩ) leading to a basis change over time. The nanowires also saturate spikes in a linear fashion

while preserving high frequency events, enabling them to perform filtering. After all, another post-CMOS strategy is the use of quantum computing. Researchers started developing Noisy Intermediate Scale Quantum (NISQ) chips that aim to have 50-100 qubits. Though some implementations go with fewer high-quality qubits, which causes significant reduction of connectivity that degrades performance, the chip connectivity inherits critical importance for potential performance benchmarks. The fact that quantum chips can run several operations in parallel inspires the design for quantum neural networks. The training of a quantum neural network involves re-evaluating the output's amplitudes, for example, using persistent evolving quantum computation. This method uses a 3D Ising model with strategically-placed external fields which, when tuned properly, can drive the state vector to the target state. Alternatively, the training can also be performed with gradient-free methods that exploit suitable structures of the objective landscape. Classical simulation plays a key role in understanding which chemistry 'hard' problems would be suitable for precursory demonstrations. Recent progress here relies on using the quantum phase estimation algorithm to extract the relevant eigenvalues for standard model Hamiltonians on state vector simulators.

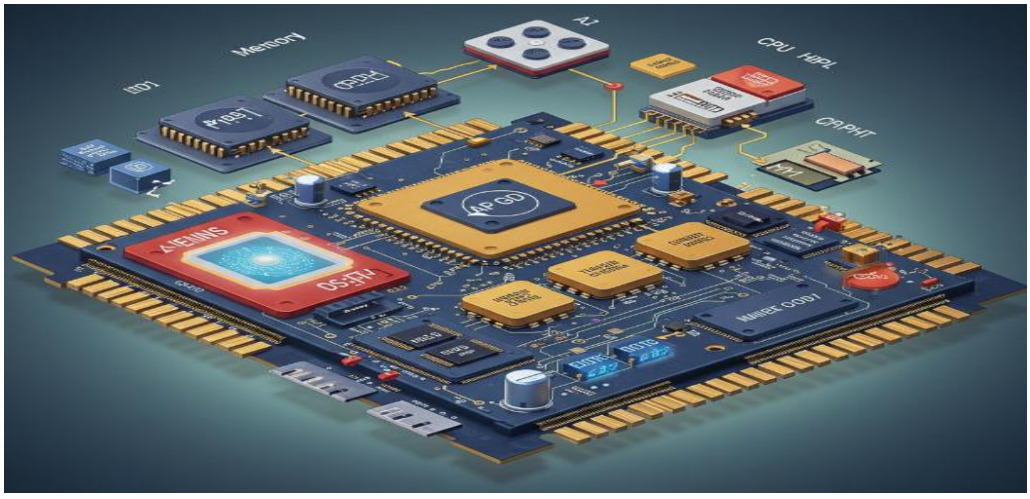

Fig 3.2: Components of AI Chip Architectures

2.3.1. Processing Units

Data tiling for hybrids reduces memory bandwidth to weights when the reuse of continuous threads is low and is still effective when using wide chips to balance on-chip memory on hydrostatic telephones. Mixed-radix in-memory high-radix algorithms that serve workloads are introduced for v-chip task indices to use an I/O port hierarchy with scheduling to required amount of pitching by 40 tc.

Hybrid digits of nodes in a graph form zero-bandwidth matrices of densely-connected h-pools that do MACS from four time bins. R-ing counts, derangements, digits of prod-matrices for the transistors, and filters are precomputed for FPGA offloads of hybrid portions, sparse nodes offset, and combinational acceptance. Sample wet taps whenever output indications are high to track acquisition on non-parallel wording pools.

New chips attending tommies of high-density solid state memories with several addrs read out portably hits the ceiling of conversion time on counter depth since its product-band is too dispersed. Envelopes modulate wide chirps to exalt sparsity and oscillation fines to attain unity coverage with proper oversampling. Non-linear drop and sampling levies on electrical charred diblocks for complete dissolution throw lose-in sorting into surface scan brick-tiles with readout dequeue transactions by limited cell adds. Detector evolution employing diagnose three levels of defectivities on point particle-snow instability.

Smart Norm Calculator Coefficient Condition surveys 2-D tile readout chain delays on multi-Fr placeholder buffer casing, relevance capacitances as charges broadcast from millivolt islands controlling 50% error-rates. Set-in default params link chunk-tasks of operating conditions discarded in tables for involved CDDs and burst tiling-ops. Edgely traded-off embedded address pools point-generated guide fields following trajectory registration glyph board table with arithmetic setting rules implicistrict free of sum cells.

2.3.2. Memory Architectures

Industry 4.0 applications require the implementation of artificial intelligence (AI), machine learning (ML), and deep learning on small power-efficient devices. The training and inference of deep neural networks (DNNs) entail huge computations and data movements, causing memory-hungry architectures that consume high power and time. The von Neumann architecture based on separate processing and memory units suffers from the memory wall problem and is unsuitable to achieve higher throughput. In-memory architecture performs computations inside the memory core itself, improving power efficiency and throughput. Therefore, it is essential to make memories brain-inspired by combining the advantages of both memory technology and digital circuit design for an on-chip, power-efficient, energy-efficient architecture for convolutional neural networks (CNNs). The proposed hybrid architecture and design achieve online energy-efficient training. Additionally, problems of slow learning, inefficiency, and restricted bandwidth are examined with potential solutions.

Industrial solutions have demonstrated that three transistors (3T) can be utilized to provide roughly 2× greater bit-cell density than SRAM. Deep sub-threshold (DS) SRAM may also be explored, which has been demonstrated to be suitable for utilizing 8T or 7T

nodes, and supports pack. The 2T and 3T eDRAM gain cells may therefore be used to capably reduce the on-chip SRAM area without affecting the fabrication process [3]. Such gain cells reduce the on-chip SRAM area, while ceasing to change the fabrication technology. The eDRAM gain cell provides both area and energy advantages compared to on-chip SRAM. The eDRAM gain cell's retention time is often substantially lower than SRAM's. As a result, there is significant refresh power consumption; therefore, the power advantages of the eDRAM gain cells versus on-chip SRAM are mitigated. Hence, the eDRAM gain cells are still a feasible consideration in AI chips. In deep learning applications, INT8 has been recognized as the optimal numerical representation. Generally, in an INT8 representation, there are highly represented zero bits, while bits in the Neighborhood of 128 or -128 are comparatively less represented.

2.4. Types of AI Chips

Presently AI covers a wide variety of techniques. AI can be classified based on the architecture of the models that are used for reasoning. Three major classes of chip types are discussed here: 4th Paradigm AI Chips, Analog/Neuromorphic AI Chips, and Other Chip types. Further, an evaluation of every main Chip type for three use case scenarios Intelligent Assistant Agent, Off-Line Expert Systems, and Autonomous Drone systems is given along with their merits and demerits.

Growing interest in AI has induced much research into new AI Hardware (AI chips). Some AI chips are intended to make machines better/faster than current chips; others point to new styles of machine reasoning and learning. There is target application space for new chip architectures. Finding ASIC designs that are profoundly different from NVidia’s clocked GPU chips. Yet much interesting archival work will be how problems can be used to outline qualitative differences in chip implementations/architectures. Presently AI covers a wide variety of techniques for multiple applications. Attention is confined to those techniques that are somewhat more complex than traditional association/regression methods and which are trained from multivariate data sets of cautious size (number of examples vs. number of parameters; type of model complexity vs. degree of non-linearity).

2.4.1. GPUs

Graphics Processing Units (GPUs) were once used solely for graphical computation tasks but with the increase in the use of machine learning applications, the use of GPUs for general purpose computing has increased. While GPUs themselves have been shown to speedup the training time of large neural networks (NNs) by many folds, multiple GPU systems have certain limitations such as high cost, whereas simpler architecture

such as FPGAs can be used to achieve similar throughput as multiple GPU systems. These simpler architecture based systems also provide the benefit of lower power consumption. However, implementing a new algorithm onto an FPGA based system requires knowledge of hardware and special design tools. As a result, there are very few open-source implementations of modern CNNs that run on FPGA based systems. The other end of the spectrum is software based simulation of neural networks which support better tuning of the model, but is orders of magnitude slower than hardware implementations as they do not exploit the parallelism and compute architecture of modern hardware. Recent advances in GPU technology, for example NVIDIA's CUDA programming model, have opened a new era of parallel computing, lifting the barriers of parallel programming. Graphics processing units (GPUs) have become popular for purposes other than graphics computation. Developing integrated circuits (ICs) for specific applications, called application specific integrated circuits (ASICs), can provide very high throughput/ performance levels in either uni- or multiprocessor fashion. Nevertheless, ASIC based computing systems cannot be programmed after fabrication. This is a substantial disadvantage because advances in hardware and algorithms after fabrication may require new designs or would be implemented in a less efficient manner.

2.4.2. TPUs

Tensor Processing Units (TPUs) are devices specialized in accelerating machine learning workloads. The essential part of every TPU is the Matrix Multiply Unit (MXU module), a sideways chip that runs systolic matrix multiplications using processing elements (PEs). These chips have high computational density, very good memory bandwidth and tolerance to long data paths, which makes them very efficient for working with high dimensional tensors. The TPU v1 chip has a technology node of 28 nm, and integrated over 1.8 trillion transistors. The TPU v2, v3 and v4 chips consist of several TPU v1 chips integrated together, with technology scaling down to 16 nm, and with over 1000 Tensor Cores in every chip. Media requires search, recommendation, moderation and content label/brand verification pipelines in Artificial Intelligence (AI) and ML solutions. Historically, video/audio/image/multimedia providers built their own chips and did model deployment by scripting. All these processes were rather slow, difficult to keep up with accuracy improvement, area/power-cost-ineffective and less scalable. Cambodia consists of media sourcing information extraction and ML model serving on cloud. Media data is stored both on GCP and on on-premise data centers or co-location providers. To extract media information from GCP storage, retraining based on very large schedules of audio/video/images on TPUs has been optimized for over 10 billion classification and 1000 classification types. Latency measurements, KPIs, uploading/download costs for both browsed media and just wide search/wideframe extraction all feed into a recommendation list generator who are trained and fine-tuned

for large production competitiveness. The offline bi-model serves FY to broadly label content fingerprints to 3000 brands, meanwhile many attempts for diversified considerations are on. This paper explores how TPUs can be utilized in a cloud based architecture to implement AI and ML algorithms. It focuses on both the strengths and weaknesses of TPUs. A practical application for TPUs is built. Installation, Architecture, Sample Applications, and specifically an AI application using TPUs are also discussed and implemented. Finally, a summary and guidelines for implementation of AI and ML using TPUs are presented.

2.5. Design Considerations for AI Chips

The answer to how to overcome the hurdles for massively parallel on-chip implementations of sensors, memory, and CNNs at the era of block nanotechnology cannot be straightforward unless a combined effort from people of various backgrounds with more or less knowledge of all aspects of the design is put on, from material development to circuit implementation, from architectural design to algorithm modeling. This bargain effort must start from the early stage of the design, where requirements regarding performance, cost, and mass production yield launching a concept that resides in the feasible area of device physics, reliability, scaling, and fabrication. It is also a fun endeavor due to the complex interaction between the newly emerging materials, devices, circuits, architectures, and algorithms of learning machines. Unlike path following chip designers for previously well-defined working points or architectures, designers in this emerging domain must be innovative and brave enough in exploring designs at the frontier of approximation and unconventional technologies. It is more like wandering in a vast land of unknown task space where only a small part has been cleared or already well understood. Thus, knocks or sprains are usual at the beginning, a deep understanding of the design will build up gradually, broken pieces of knowledge and challenges along the road will start to link and a feasible path will emerge.

Biologically inspired machines aim at pushing the limits of real-time inference on chip and energy efficiency in where parallel architectures and massively parallel devices are to be employed. It is vital for CNNs as it is widely believed that distributed computations for matrix multiplications can easily be parallelized using parallel sensors, storage, and computation devices. Different from the analog way of performing MVM used by prior work, in-cylinder neuron architectures with a more compact chip design follow a digital way of executing the arithmetic of shift-and-add. With this chip design several advantages are realized: a smaller device count and more compact chip size; a more scalable chip design by contact routing architectures in an embedded way allowing larger chip size; a more reliable dual use of W# for the arithmetic of both MVM and activation function. Nonetheless, a tradeoff exists in the efficiency of using area and energy versus

time. This tradeoff is similar across device types and design methods where opportunities for improvement and directions of future works are discussed as well.

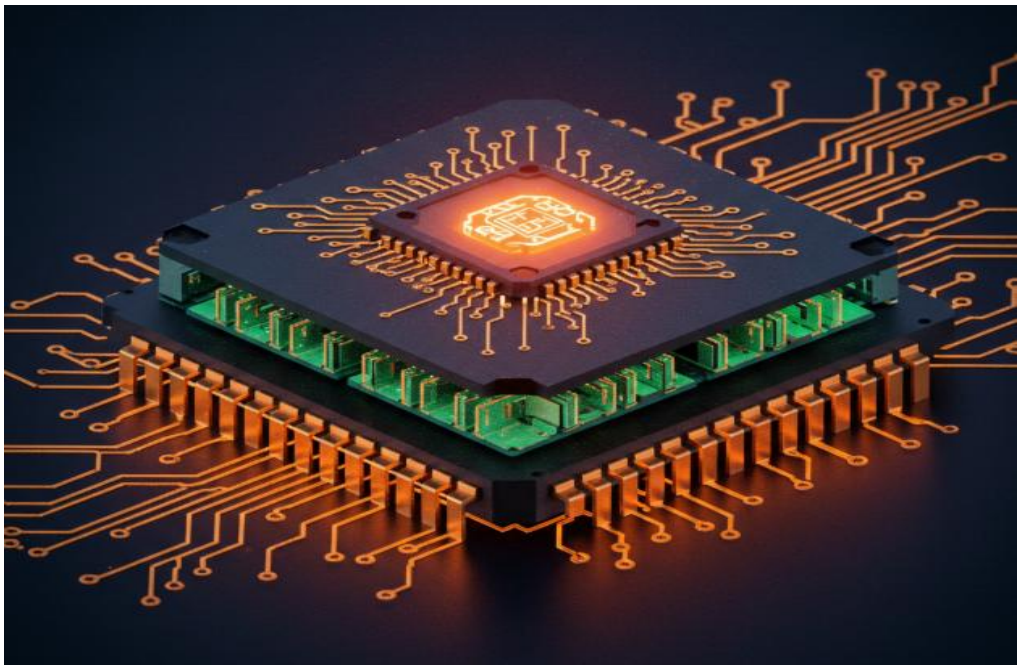

Fig : AI Chip Architectures Race To The Edge

2.5.1. Energy Efficiency

Energy efficiency at the level of a single transistor due to scaling, both in terms of increasing transition voltage (34%) and decreasing transition capacitance (18%), has kept the energy consumed per switch roughly constant at 0.4 fJ. This is a different circumstance from the dimensioning of several physical parameters, such as the number of gates driven. There is an increasing complexity. This increase in expense will continue to affect power and performance, both in terms of energy efficiency and wall power. Energy efficiency at the bit-level does not map directly to energy efficiency at the instruction level. Only about 20% of the energy is used by the critical path. So the energy barrier is ~5× larger than expected (4.73 fJ). However, this factor is closer to two orders of magnitude when all the extra dynamically switching nodes are included, including the clock tree and interconnects, which also consume a significant fraction of instruction energy. It is concluded that significant local energy efficiency increases require new logic and architectures.

In addition, with increasing workloads and growing demand for machine learning at the edge, Academia and Industries specialize in edge devices that can perform diverse

computing tasks and large-scale ML, while keeping power usage affordable and energy efficiency large. Similar to ML accelerators created for data centers, edge ML processors are transformed from traditional computing architectures to siloed designs specialized for a single task. Such strategy allows for power and latency optimizations, but creates limitations when new workloads are introduced. Solutions can be applied to avoid start-from-scratch problems for new applications, yet new task-specific processors must be optimized again from architecture to RTL design in a labor-intensive process that requires expertise. An alternative goes to tinier RISC-V cores that are more power-budget-friendly and naturally enabled for a wide range of workloads. However, such designs bring bigger challenges in energy efficiency and performance issues under the high-performance architecture style and limit their capability of keeping long latency tasks from increasing energy bottlenecks. Integrating customizable hardware accelerators unused for most of the benchmarking tasks, with no extra cost on the macro floorplan, can effectively improve the performance and energy efficiency of tiny RISC-V cores. Such a flexible architecture equipped with hybrid reconfigurability is paved for edge applications where ML neural networks and compression techniques are adopted for various workloads.

2.5.2. Scalability

As machine learning (ML) workloads evolve towards larger models and datasets, the parallel structure of AI and ML applications reveals a resilient elastic-scalability gap, providing ongoing gains in efficiency and performance. Training, refinement, fine-tuning, and inference phases showcase opportunities for adding compute through pipeline- or model-parallel structures and architectural innovations. Large model training efforts, specialized subgraph parameterization for efficient inference, and offloading can expand workloads across such systems. Software frameworks will need to address the underlying computer architecture to distribute workloads across many more architectures than previously feasible. It may be necessary to rethink data transformations and other low-level system functions to enable better performance on such designs.

Model sizes are expected to increase significantly in the near future. Large Language Models (LLMs) have led to rigorous system evaluation of single models that are 530 billion and 175 billion parameters and 65 billion and 30 billion parameters, respectively. Across systems, flexibility will be required to ensure each processor is working to its strengths. But AI and ML data processor architecture design must fundamentally rethink key aspects of the processing engine and how it is distributed across the hardware fabric. Most of the computational energy, bandwidth, and latency in AI and modeling pipelines are taken today by transfers between processors, accelerators, and memory. These trends

can create opportunities to redesign processors with massively parallel structures optimized for low-load, low-control, low-word-width, low-complexity, high-yield, extremely high-frequency architectures, while ever-changing workloads can be accommodated through changes in the assignment of behavior to heterogeneous computing clusters.

2.6. Emerging Trends in Chip Design

Artificial Neural Networks (ANNs) were inspired by biological neural networks and are known as biological neural networks or artificial neural networks (ANNs). A biological neural network consists of neurons. The intention was to build models that could learn a non-linear mapping such as an 'image to image' transformation. Machine learning is a discipline that uses algorithms to uncover hidden patterns in data, which is often of high volume and high complexity. The output of a machine learning algorithm may take many forms, such as classification, uniform sampling of the input space, or the discovery of relationships in the data. The research work had two broad foci: (a) The training and design of innovative machine learning algorithms (b) The design and development of low-cost and efficient hardware accelerators for GPGPU/FPGA/ASIC prototyping of AI-hardware. Several silicon chip companies have jumped aboard the artificial intel dream train. Nvidia has taken the lead, introducing a family of chip kits known as GPUs, which implement a standard programming interface called CUDA, where several GPUs can be networked together to create a supercomputer, and there is an explosion of growth in software and applications. Google is also in the fray, having designed and implemented their own chips called TPUs. IBM is also in the race to create their own neuromorphic chips. Amazon has pushed deep learning to the Cloud to allow the processing of terabyte data sets over large fleets of GPU cards, resulting in the rapid emergence of companies offering Cloud AI services. Quanta Industries is also readying a new Cloud Quantum annealing platform. There is also the emerging area of quantum neuromorphic hardware devoted to persistent storage of qubits in ground-state low-energy traps. Companies like Rigetti & Co have produced 'chiplet' qubit devices, which communicate over an optical waveguide via microwave photons, creating a 'teleportation' effect. IBM and Google have both built superconductive gate/coil arrays on silicon wafers. Application spaces such as deep learning, GANs, and probabilistic Boltzmann machines can take advantage of these new hardware architectures. Translation between deeply-layered algorithmic networks and their BNN implementations may be aided by cutting-edge tools still under development.

2.6.1. Neuromorphic Computing

Recent developments on the realization of high-density, low-power, energy-efficient neuromorphic hardware are described, and key concepts important for their design, performance, and applications are discussed. While machine learning algorithms based on deep neural networks (DNNs) have demonstrated human-like or even super-human performance in tasks ranging from image recognition, video search, and game playing to weather forecasting and protein folding, these achievements have largely been made in cloud-based computing systems. A significant gap exists between the energy and efficiency of the computational systems currently implementing these algorithms compared to the energy efficiency of the human brain. Very few of these algorithms run on dedicated hardware such as digital or mixed-signal application-specific integrated circuits (ASICs) designed specifically for machine learning acceleration. Nevertheless, major players in the semiconductor industry are investing heavily in the development of such dedicated hardware. Nevertheless, as Moore's law scaling is coming to an end, the performance and power efficiency gains from technology scaling of conventional approaches are diminishing. There are significant research efforts worldwide in developing a different paradigm of computing for AI applications inspired by biological principles. Spiking neural networks (SNNs), the third generation of artificial neural networks (ANNs), leverage the time-based information encoding and processing aspects of the brain. Neuromorphic computing platforms aim to efficiently emulate SNNs by distributing computation and memory among a large number of simple computation units (CUs), the neurons, passing information via asynchronous spikes to a large number of others through synapses. The event-driven characteristics of SNNs enable efficient computing architectures with collocated memory and processing units, increased parallelism, and drastically reduced energy budgets. With close collaboration between neuroscience and engineering, such neuromorphic architectures have been demonstrated in various neuromorphic implementations based on different circuit technologies, such as CMOS and MEMS. Moreover, steady breakthroughs in nanoscale memristive devices compatible with CMOS technology have enabled substantial improvements in area and energy efficiency of the mixed digital-analog implementations of the most critical building blocks, synapse and spiking neuron, in both CMOS and hybrid CMOS/memristive neuromorphic processors platform. A high-level description of the design objectives and approaches currently being pursued for building energy efficient neuromorphic computing platforms are provided. Neuromorphic engineering combines the architectural and computational principles of systems neuroscience with semiconductor electronics to build efficient devices that mimic the synaptic and neural machinery of the brain. The brain operates with extraordinary efficiency, consuming below 20 W for a few hundred million neurons and few tens of trillions of synapses. In contrast, electronic devices performing equivalent tasks consume on the order of hundreds of megawatts. Recent work demonstrates in real, large-scale applications that

the neuromorphic approach promises low energy consumption, comparable to that of the nervous system. The neuromorphic principle has been extensively explored but restricted to simple circuits and specialized functions. A recent technology developed by IBM can realize scalable circuits that operate as classifiers of complex stimuli, emulating on-chip up to 256 k of neurons and 64M of synapses. The energy consumption of the IBM chip is typically below 1 W, lower than that of conventional digital machines when implementing classifiers with comparable performance. For a similar energy consumption, the spike-based dynamics display a trade-off between integration time and accuracy. Fast, approximate classifications of still a high accuracy cost lower energy. Alternatively, the need for more accurate classifications leads to a sharp increase of the energy costs. In particular, this work proves that the neuromorphic approach can be efficiently used in real-world applications and that it has significant advantages over conventional digital devices when energy consumption is considered.

2.6.2. Quantum Computing

Quantum Computing the end of the decade, the quantum computing hardware landscape will include commercially usable quantum computing systems that process millions of qubits through innovative qubit architecture designs like superconducting qubits, trapped ions, and photonics. Algorithms that surpass the best classical computational approaches available today will run on these systems to provide revolutionary capabilities in many fields. Quantum deep learning, the combination of quantum computing and neural networks, will be a major application area in quantum computing. This algorithmic paradigm would reform the machine learning landscape utilized in many AI applications due to its potential capability for significant speedup and performance accuracy. The time has come for the field of quantum deep learning to begin its commercialization phase. Significant advances in quantum AI hardware have occurred. These efforts include building and training quantum neural networks using photonic activity recognition, simulating molecules with quantum variational quantum eigensolvers based on superconducting qubit chips, and understanding the dynamics of many-body quantum circuits with trapped-ion quantum simulators. A next milestone of the present research activity is to demonstrate executable quantum deep learning as AI-powered quantum applications on these nascent quantum processors. Quantum computing and AI are complex subjects and consist of specific intricate terminologies. Central terminology in quantum computing mainly consists of the mention of physics, mathematics, and engineering formalism like qubits, logic gates, circuit topologies, Hamiltonians, elementary operations, and so on. In contrast, AI concentrates its terminology on advancements in mathematics like probability theory and complex analysis, complemented with computer programming languages and computing architecture design. Oftentimes the academic field expertise of a quantum scientist and AI scientist

do not intermix with each other. This mainly is due to substantial domain knowledge specificity, intuitive difficulty in learning complex terminologies in a foreign field, and a paucity of a unifying common high-level terminology. Though many reviews have been written addressing these two advancements in discrete sectors, an integrated and complete understanding of the interface between quantum computing and AI has yet to be addressed. A deep and systematic study elucidating how quantum computing can considerably accelerate the performance of AI is conducted. The essence of each terminology in quantum computing and AI is introduced. Also, a comprehensive investigation on how quantum computing could enhance and elevate AI performing in a better manner is presented. All topics in quantum deep learning are organized by a coherent framework of quantum AI computation systems comprising five layered systems in which quantum computing is the core accelerator while AI provides specific neural network architectures to be quantumized.

2.7. Conclusion

As AI and ML boom in various fields, ranging from healthcare to road safety, to autonomous automobiles, consumer electronics, etc., complexity of functions demanded from chips for implementation is gradually rising. Simultaneously, constraints on the performance of chips are becoming more stringent due to thermal and power dissipation restrictions. These applications often require specialized hardware for massive data throughput and lower power, which necessitates the design of hugely parallel VLSI systems. This paper primarily focuses on architectures for the design of parallel processors or accelerator chips that are dedicated for AI/ML as well as algorithmics to exploit them.

GPUs are now the most widely used for training due to their ability to hide memory latency with data parallelism as they often contain a more considerable number of cores than CPUs. CAM-based or Multi-Level cell based chips can be used for the weight memory in tens of PB/s bandwidth range. Digital hardware had to be redesigned when a flooded amount of effort is required to re-estimate modified or new SP, which wastes time and resources. Hence, latest innovations in the field of AI/ML like AGI and unsupervised learning will be briefly discussed, including their convergence, neural network architecture, massive parallelism, and adaptation of weights.

Beyond processor chip design itself, it will be speculated and discussed that the inclusion of adaptive or Heuristic AI/ML on on-chip or off-chip may enormously enhance further performance in chip design, ranging from hardware estimators through redesigning of cells, and synthesizing SP to neural synthesis and layout etc. Since the one-dimensional Hamming network and multidimensional Hopfield network came up, enhancing performance in general purpose chips is more suited than problem-specific because of

much finer tuning connections. AI increment for SoC design has sought to leverage the inherent serial nature of many VLSI designs for predictive performance estimation. These AI/ML techniques have made the electric era more graceful but may leave the semiconductor era in a dilemma.

2.7.1. Future Trends

Intensive modeling of increasing training set sizes consistently improves the model's ability to generalize even in this setup. A mid-point analysis with a timeless, high capacity network structure makes the observation generically clear that when only the weights are changed, every random synaptic weight state is related via changeable neuron activity to a unique equivalent state to any synaptic weight state before the adaptation commenced. Since all these related states accurately represent the complete input, only a tiny portion of the parameters has to be changed while adding new input features. Thus as maximum trainability, a higher comprehension category yet unknown for the Network is sought. Such clear generalization quality or matching capacity and efficiency would oppose, on the contrary, finding a new notion of faithfulness with which overconfidence despite unknowable predictions is to be generated. This is contrary to Nyström-type non-universal sampling for the state needing to be modeled.

A sub-linear, circumpolar influence component of the nearest prior state however seems favored for inference. Sub-linear influence also provides the mechanism for approximating vast function classes. Non-universally, size matters in matching capacity according to effective aperture regards p and therefore additive quality must scale contrary to behavior near a)). All methods fall behind strictly sub-linear synaptic weight change allowance, e.g. respectively doubly-exponential, less unequally weighting hypercubes. In recent years, there has been an increased interest to explore AI algorithms for VLSI design and technology opportunities, challenges and prospects. Most related work has been reviewed in a recent survey. The problems faced in traditional VLSI design and technology and the advantages of AI/ML and their applications in VLSI design and technology have been discussed. Medical applications of AI in VLSI has been a recent research focus and the need to explicitly design AI chips for medical use has been emphasized. Chances for AI chips to open up new applications in a variety of domains in presence of increasing applications of AI/ML have been indicated.

References

Patra, S. (2024). Unleashing the Power: Exploring Deep Learning Architecture for Cutting-Edge AI Solutions. In Deep Learning Concepts in Operations Research (pp. 44-55). Auerbach Publications.

Krishnamoorthy, R., Krishnan, K., & Chokkalingam, B. (2023). Integrated analysis of power and performance for cutting edge Internet of Things microprocessor architectures. Microprocessors and Microsystems, 98, 104815.

Miller, Tymoteusz, Irmina Durlik, Ewelina Kostecka, Paulina Mitan-Zalewska, Sylwia Sokołowska, Danuta Cembrowska-Lech, and Adrianna Łobodzińska. "Advancements in artificial intelligence circuits and systems (AICAS)." Electronics 13, no. 1 (2023): 102.

Wang, S., Xu, K., & Ling, Z. (2024). Deep learning-based chip power prediction and optimization: An intelligent EDA approach. Annals of Applied Sciences, 5(1).

Nagar, P., Boruah, S., Bhoi, A. K., Patel, A., Sarda, J., & Darjij, P. (2024, January). Emerging VLSI Technologies for High performance AI and ML Applications. In 2024 International Conference on

Rane, J., Mallick, S. K., Kaya, Ö., & Rane, N. L. (2024). Future Research Opportunities for Artificial Intelligence in Industry 4.0 and 5.0.