Deep Science Publishing
https://doi.org/10.70593/978-93-49910-47-8_6

DeepScience
Open Access Books

Chapter 6: Developing low-power, high-throughput artificial intelligence chips for edge devices and real-time inference systems

6.1. Introduction

Artificial Intelligence (AI) chips are a class of hardware accelerators specifically designed for real-time inference of AI algorithms. Digital AI chip(s) are custom CPUs designed for efficient implementations of DNN (Deep Neural Networks) workloads. AI chips differ from general-purpose CPUs/GPUs in optimizing the compute engine architecture such as memory access, and processing elements organization to minimize power and latency. Their architecture is driven by their workload(s). Developers are continually looking for smarter architectures that can do more with less, hence optimizing area, power, and latency. Development of new architectures and circuits for the evolving demands of AI workflows is a big challenge. This is an actively explored area and is core to the functioning of AI chips. Possible architectures include: (1) conventional architectures used in CPU/GPU with optimizations in the access mechanisms, (2) architectures similar to FMCW radar that can provide on-chip memory and has high fan-in which helps accelerate specific workloads such as DNN, (3) processing-in-memory architectures that improve energy efficiency by moving compute closer to memory (Lin et al., 2021; Cheng et al., 2024; Omidsajedi et al., 2024).

AI chips target an increasingly growing number of applications, each with its own unique workloads which require the ability to execute diverse algorithms of different complexities. Real-time inference systems that need continuous inference are critical in this epoch of automated processes and performance chasing machines. However, state-of-the-art AI chips, architectures, and tools struggle to meet the conflicting demands of the number crunching characteristics of inference workloads on latency/throughput and

flexibility requirements. Demarketing continuous inference with a microarchitecture analysis is performed to identify workloads that stress various microarchitecture components of RAIC including conditions that render it inefficient. This identifies opportunities to improve area-efficient AI chips. Various techniques based on hardware-centric software changes and accelerator-specific coils are presented to optimize performance in terms of throughput and latency and demonstrate their efficacy on RAIC. With continued miniaturization devices in edge devices will be incapable of extracting sufficient electrical power from conventional battery chemistry to support growing storage, communication and processing demands. Design choices must minimize energy consumption per operation, communication costs and fabrication time which will likely lead to proposals based on new material systems (Venkataramani et al., 2021; Shuvo et al., 2022; Santoso & Surya, 2024).

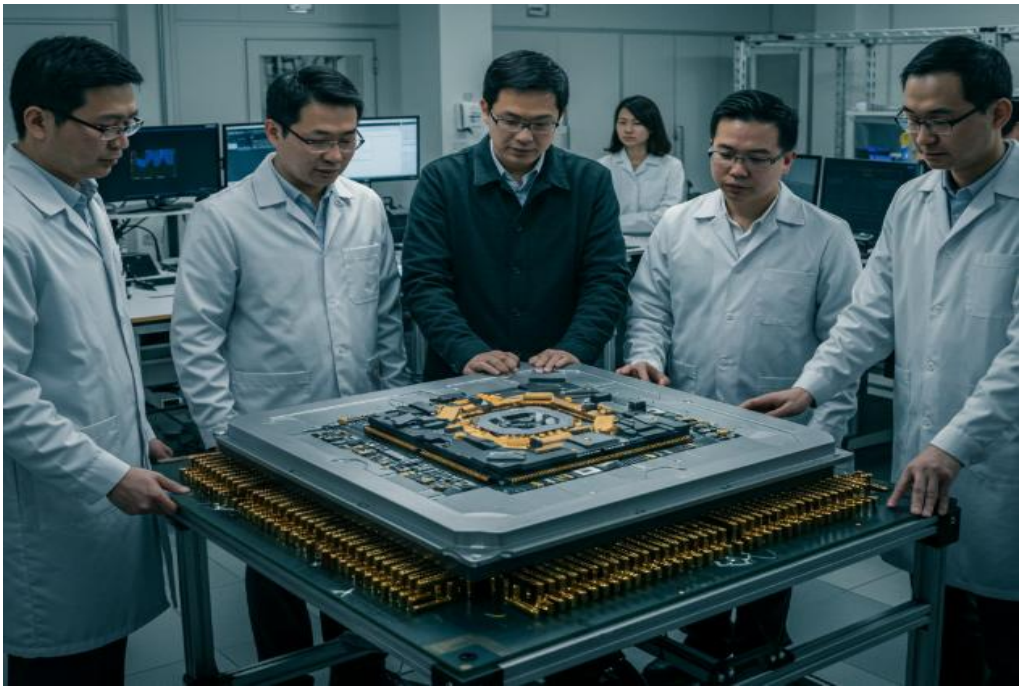

Fig 6.1: Developing Low-Power, High

6.1.1. Background and significance

The surging interest in converging artificial intelligence (AI) applications with ultra-low-power "edge" devices, such as personal health monitoring in smartwatches, autonomous vision systems in surveillance and robots, real-time video and audio analytics in home and wearables, has stimulated the massive effort in developing low-power, high-throughput AI chips. Such chips, by incorporating in-memory computing and parallel

architecture, could eliminate the bottleneck of the off-chip data movement that generally exists in the von Neumann architecture and perform on-chip rapid computation and inference with minimum energy consumption. But with such chips, there are still challenges to smoothly integrate them into the edge inference systems for practical use. Existing software stacks and neural network frameworks for data center and cloud use are mostly developed based on the AMD or NVIDIA platforms using GPUs. Country-specific model compression accelerators have been recently proposed to reduce the computational complexity. They could enable pre-trained AI models from open-source model libraries faster inference with performance unreduced on edge devices. However, they still rely heavily on state-of-the-art AI chips and frameworks originating from the cloud infrastructure.

Real-time learning and adaptation of models with knowledge auto-transferring and sharing are necessary to accommodate edge applications' dynamic environment. The flexibility and adaptability of neural structures utilizing some evolving mechanisms like neuron/connection growth and death could make replicated networks for the same tasks even more efficient. The emerging memory technologies and the memristor synapses' potential in realizing massive parallel MAC operations make real-time learning with emerging devices to be a nontrivial task. This also implies the need to have smarter ways to train, program, and implement algorithms. Performance scalability becomes a hot research point given the ever-growing model complexity or input size. Digital neural chips are the most established AI chip implementations for practical use. The synaptic nodes become purely digital when the 'multiply and accumulate' (MAC) operation is implemented using digital logic. Other emerging systems include carbon nanotube (CNT) and thin film transistor (TFT) to build synapsis, which could scale up neural network implementations. Another upcoming strategy is the use of quantum computing, with researchers developing Noisy Intermediate Scale Quantum (NISQ) chips aiming to run several operations in parallel, inspiring designs for quantum neural networks.

6.2. Background and Motivation

Artificial Intelligence (AI) has gained significant attention in the past decades due to the emergence of machine learning (ML) algorithms and big data processing devices. These led to the development of the first high level algorithm such as deep neural network (DNN). Beyond simple pattern recognition tasks such as image and voice recognition, the goal is to build intelligent machines that can learn the task on their own. Memory-centric real-time learning AI chips such as Resistor Based Artificial Neural Network (RAN) Chip are needed to facilitate this. The objective of this project is to develop RAN chips by using the low-power highly scalable Analog Crossbar architecture, along with Emerging synaptic devices as memory resources. As an Exploring Project, two

prototypes with technologies including 40nm CMOS back-end of line and 160 nm CMOS rail-to-rail digital are proposed for rapid prototyping and high-speed signal verification respectively. RAN chips are aimed to be used in the next generation real-time running devices such as self-driving vehicles and edge devices in IoT.

Recently, developing low-power, high-throughput AI chips for edge and devices have gained significant attention. Towards this goal, a novel memory-centric architecture, i.e., Resistor-Based AI Chip (RAN Chip), is proposed and implemented by using resistive crossbar circuits, which achieve large-scale, low-power, and high-throughput computing. A crossbar access scheme of the RAN Chip is also proposed, enabling programmability and introducing In-Stage-Processing architecture to eliminate redundant read-modulation delay, while RAN chip architecture relies on externally programmable analogue circuit. Emerging memristor devices are also gaining momentum for AI chips, with unique advantages of high speeds in device and programming, improved endurance and variability. Conventional memristors use crystallinity manipulation to switch the resistance, and are hence incompatible with scaled nodes primarily fabricated with amorphous. In addition, RRAMs rely on ion diffusion with asymmetric switching speed, leading to poor SRAM compatibility. Earlier, an emerging device such as Memristor was integrated with CMOS technology to achieve the highest density of an 8Gb RAM chip with cell retention.

6.2.1. Research design

The objective of this research is to explore new design techniques that can realize low-power, high-throughput AI chips for edge computing devices and real-time inference systems. Research needs to fulfill both architecture designs and circuit designs in the following aspects: (1) A new AI accelerator architecture based on parallel model compression and model quantization that seeks a trade-off between performance and model size, (2) A comprehensive design flow that includes synthesis flow, quantized HDL simulation platform, and CAD tools for performance validation and power error analysis focused on model quantization, (3) Circuit designs of mixed-signal digital-to-analog converters based on regenerative voltage-mode current steering architectures, which achieve low area and energy cost on-chip while enabling ultra-low-power, high-throughput chip designs. This research employs a simulation system-level design for architecture design exploration. Model compression and quantization techniques at the algorithm level will be comprehensively studied and validated. Additionally, some critical techniques, e.g., error-cancellation techniques, matched-designed amplifiers, and operation-dependent biasing, will be considered for circuit designs.

The research will incorporate the following four techniques for architecture exploration, performance validation, and power error analysis: (1) Model compression techniques

focusing on weight pruning and low-rank approximation, (2) Model quantization techniques focusing on weight quantization and activation quantization, (3) A multi-port-adaptive architecture along with a distributed-interconnect topology for mixed-signal multi-chip designs, and (4) Simulation tools for language design and circuit partition that also include power error analysis on various types of quantized distributions like linear, non-linear, or trainable. Researchers will develop tools to automatically check the fidelity of weight pruning and low-rank approximation on the effect of classification accuracy for deep neural networks. They will also study and implement metaheuristic algorithms for network pruning and block-wise low-rank decomposition.

6.3. Overview of AI Chip Architectures

The AI Chip architectures can be divided into the following five categories: 1) Custom chip with only 2D MAP devices, 2) Digital SOC with custom chip for only S/N and 2D MAP devices, 3) Mixed-signal SOC for only S/N and 2D MAP devices, 4) Mixed-signal SOC for only S/N and R/2D MAP devices, and 5) Digital SOC for only S/N and 2D MAP devices.

1) Custom chip with only 2D MAP devices: The simplest architecture consists of a few MAP devices. The proposed method can route weights only using the blank cell and no add circuits for non-zero weights, as the original devices have non-zero weight values. This can implement different weights and long variable storage using only 2D MAP devices. However, due to the low yield factor of the 2D MAP devices, the number of devices may be smaller than expected, and the performance of the remaining devices could vary severely. In this way, the AI chip could fall short of performance. This could be a good choice for chip validation.

2) Digital SOC with custom chip for only S/N and 2D MAP devices: To increase the yield rate and avoid performance variation issues, the overall design could be implemented using a digital chip. The co-simulator validates the accuracy of the proposed algorithm and the hardware implementation using only digitized S/N and 2D MAP devices. The MAP devices could be the same as in the previous one. The digital implementation of switching circuits could be tested on FPGA to verify the functionality of MAP-based networks.

3) Mixed-signal SOC for only S/N and 2D MAP devices: To increase the yield rate and reduce total power consumption, the overall design could be implemented using a mixed-signal chip. The S/N input signals could be digitized and encoded using the ADCs in the digital design chip. The MAP-based networks could comprise SOC design with a

chip composed of digital blocks such as a shift circuit, timing circuit, decoder, and custom analog circuit.

4) Mixed-signal SOC for only S/N and R/2D MAP devices: Due to the difficulty in fabrication, the structure of 2D MAP devices could be changed to R, and the map-based networks could be implemented using the newly redesigned SOC structure. S/N input signals could be digitized and encoded using the ADCs in the digital design chip. Depending on the availability of 2D or R devices, either design could be chosen.

5) Digital SOC for only S/N and 2D MAP devices: If the performance of S/N devices is higher than that of R devices, then the final architecture could be as simple as the previous mixed-signal design. The custom ISA would support MAP functions, and the resource utilization and routing could be performed using routing assignment algorithms.

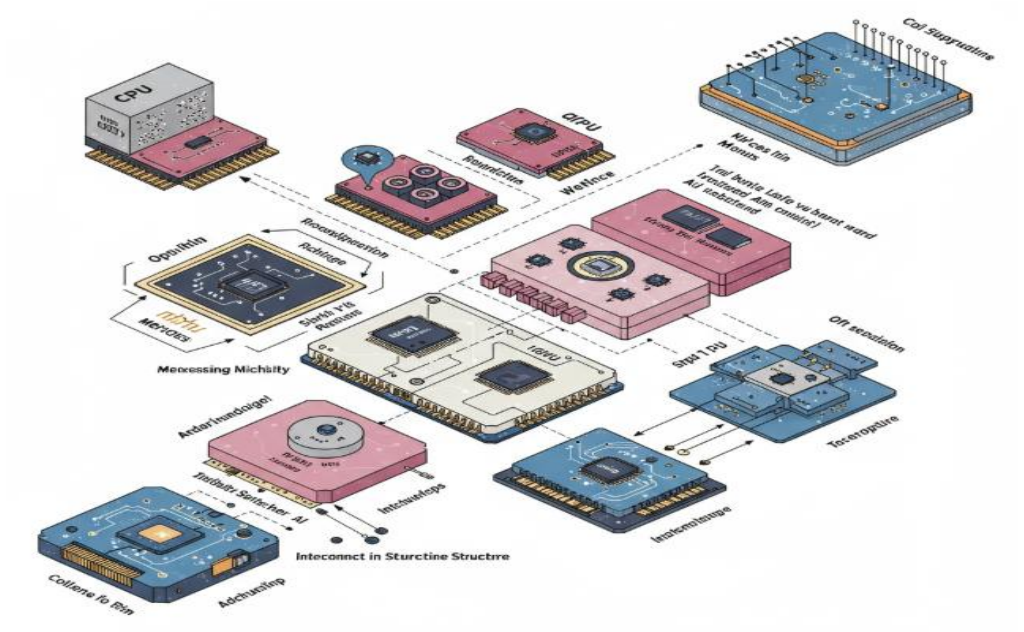

Fig 6.2: Overview of AI Chip Architectures

6.3.1. Traditional Architectures

Conventional chip architecture is characterized by fixed function processing elements mapped to fixed function on-chip interconnects to achieve a fixed execution schedule for max performance while consuming fixed off-chip bandwidth. The assumed hardware for computer architecture has been transitioned from room filling multi-board

architecture with huge power budget to monolithic implementation of a large scale multi-core processor with multi-level cache hierarchy. Deep learning algorithms have either been implemented using parallel processing on highly scalable chip multiprocessors, or, rely on FPGA-like reconfigurable architectures for inference efficiency. All conventional architecture assumes a large budget for chip area, power, and bandwidth, thus are not suitable for energy efficient processing at the edge due to stringent resource constraints imposed by mobile, battery powered, and low-cost deployment.

High-throughput chip architecture is organized hierarchically, processing elements (PE) are grouped into clusters that are connected by a higher level crossbar switch network to form a multi-chip system on chip. Each chip integrates multiple clusters connected through moderate off-chip bandwidth, and each cluster, in turn, consists of a fixed number of PE on a shared local high-bandwidth on-chip interconnect. Such hierarchy allows low power and high throughput while a simpler vertex-centric design remains applicable. High-throughput architecture organizes a unified cluster of massively parallel low-power processing cores (i.e, accelerator) that only support fixed function computation for streaming inference. It overcomes the von Neumann bottleneck by designing energy-efficient interconnects and routing, thus achieving high throughput and is widely adopted in ML DSP. In high density ML with highly burst switched execution, it is proven that an inter-core communication graph with a topology close to the 2D Hamiltonian cycle achieves the maximum throughput by requiring minimum on-chip routing.

6.3.2. Emerging Architectures

Edge AI is a collection of systems and frameworks that may be intelligent and able to make predictions or decisions about a given environment based only on data acquired by local sensors. These systems must be ultra-low-power and inexpensive compared to the tremendous computing power available at inexpensive data centers across the globe. Key methods to accomplish this goal include specialized low-power IC design circuits, novel low-power architectures, improved algorithmic efficiency, and aggressive internal representations to minimize energy costs. In addition, in on-chip data storage, efficient non-volatile components or resistive architecture cells are emerging but will require new low-power reading circuits.

Neuromorphic architectures pushed Moore’s law to its limits and new parameters have to be modeled to fill the gaps between bio-plausibility and functionality compatibility with AI. Silicon photonics for optical incoherent SPIKEs or DOPED for mathematically coherent likening of crossbar architecture and other emerging technologies are currently researched for future neuromorphic implementations. E-in-memory neuromorphic systems are on-chip and run energy-saving embedded AI operations while DA is a

compute-in-memory strategy limited to production testing of larger chips. These chips have to be tailored to existing deep learning network design and thus benefit less from the neuro-biological research ideas. Continuing model-hardware co-design on sparsity, compression, quantization, or binary/multiplicative networks should retain its utility in innovative A1 chip design and runtime acceleration.

A compact size low-power RFM-based edge AI accelerator chip for CNNs was presented for real-time AI object detection and tracking applications with on chip data management for best bandwidth utilization. The Montgomery Clarity 64-nm TSMC technology can provide more parallel resources retaining the compact size with reduced RFM array, and fast analog Varga approximated CNN inference operation with improved hidden states utilization. The future work is planned to further enhance the tracking throughput and edge-supporting intelligent applications without offline training, data downloading, schedule optimization, or high-frequency bandwidth telemetry.

6.4. Low-Power Design Techniques

Low-power design is fundamentally important for the deployment of AI chips on edge devices. Power consumption must scale down as computational workloads become intense, as data locality is not fully possible. There are many chip design techniques designed for low-power AI chips published in the past decades. Interconnect design techniques were firstly discussed in the low-power DSP project and can be applied in AI chips as well since AI chips typically have dedicated on-chip buses, interconnects, and multi-core fabrics similar to DSP chips. AI chips also study better logic optimization techniques. Logic can be simplified through re-formulating the hardware-accelerated algorithms, and resultant simpler logic can save dynamic and leakage power. There are many power-scaled processing elements so that they remain proportionally commensurate with the change in clock frequency through power-scaled voltage reduces such means. Overall, recent low-power boosting techniques offer a diverse range of solutions that can be implemented in AI silicon chips in different domains and layers of abstraction.

6.4.1. Dynamic Voltage and Frequency Scaling

Voltage and frequency scaling of the processor impacts the performance of the system. The higher clock frequency requires higher supply voltage, leading to high power consumption and heat generation. Moreover, supply voltage must be reduced to meaningful levels to save power. The processing speed of the processor depends on supply voltage. A conventional voltage and frequency scaling technique would scale the

supply voltage and frequency based on workload requirements. However, this conventional DDVFS approach has a fixed mapping of clock frequency and supply voltage pairs. A dynamic voltage and frequency scaling (DVFS) estimator was developed to track the workload requirements and provide the output for portable microprocessor systems at the lowest power without missing deadlines. This estimator predicts the worst case execution time of the portable systems and provides the corresponding supply voltage and frequency pairs. Development of a microprocessor that monitors the workload of the system and thereby saves power. A powerful embedded DSP processor-ring architecture was also suggested to address mobile system requirements for on-chip communication, power efficiency, and area constraints.

Dynamic voltage scaling (DVS) has been developed as an effective technique for conserving energy in portable computing. A facility was provided for continuous switching between different voltage frequency pairs at run time to speed up the computation-intensive tasks and slow down the simpler ones. An evaluation of the tradeoffs in value over cost by incorporating dynamic voltage scaling into wireless base stations confirmed a power saving of 15%. Clock frequency (f) and supply voltage (Vdd) pairs, where the maximum throughput of the coarse-grained reconfigurable architecture lies in a variable voltage domain. The supply voltage ranges from 0.85 V to 1.25 V. A control strategy is suggested to optimize the frequency and voltage consistently as per the throughput requirement. Enhanced voltage/frequency pair mapping is provided to examine the anticipated reconfiguration architectures within the same voltage domain. A forward voltage/recombinable design of VCOR is implemented in a custom-designed dual-path GaAs general-purpose analog floating-point multiplier.

6.4.2. Power Gating Strategies

The traditional approach to power gating focuses on using large transistors to connect or disconnect power to relevant circuit sections. In a low-threshold voltage circuit, the deep sleep state has generally been implemented by shutting down device bias. This takes the circuit to a very low current state, where dynamic power is reduced effectively. However, waking up such designs takes a longer time, so the circuit may be in this mode only if the internal states are not required. A no-change state in ramping a design to its operating state has been a big requirement in designs using state-retaining latches or similar flops. Here, the state is retained but power to internal sections is isolated using large transistors and appropriate boundary conditions need to be provided to avoid contention. If adequately handled, power dissipation improves drastically. Many architectures use division of circuits into a sequence of states, each of which helps to reduce active count, viz., memory blocks. A reduced transition circuit is used to drive it out of its no-change state. Feedback from existing designs involving this partitioning is used to validate and improve newer designs in either blocks or complete chip layouts.

This process is very relevant to designs for mobile applications. In standalone designs, bottleneck identification is of crucial importance. Understanding of possible loads and their placement provides a basis for power gating strategy selection. Long-term investigations regarding partitioning of designs into hot and cold components have given aid in mitigating bottlenecks to performance. This is of special interest to purchase designs. Here, the design may be fixed but systems level modifications involving analysis of packages can aid in tuning implementations to improve performance and reduce power. This type of enhancement is most effective when builds of original designs have experience from which to draw verification, but can also be applied to unintended designs.

6.5. High-Throughput Design Strategies

The chip-to-system message-passing buses and register file management, which on-chip memory access protocols would allow faster than off-chip access to DRAM-like technologies, would allow on-chip level chips similar to FPGAs, TPUs, and Graph-Core chips. Image and Scene Processing are fast-growing AI accessory processing in a sensing application that benefit greatly from low-power AI acceleration. At the same time, most of the existing designs are newly developed architectures based on existing processor-conceptions without any consideration and integration at chip-level interconnect management. AI Chips at the chip-level local-to-global interconnects should be analyzed through proposals, which would provide chances to breakthrough retroactively. Implementation case studies would be presented to illustrate the designs, assumptions, and trade-offs of designs from the industry professionals.

Low-precision is the first order knob for achieving higher Artificial Intelligence Operations. However, the algorithmic space for sub-8-bit precision compute is diverse, making FPGAs a natural choice for Deep Neural Network inference. A programmable architecture needs to be proposed with the following goals: Flexibility/Versatility and Performance. A set of performance benchmarks and design considerations are the inputs to the architecture. Finally, how to design an interface, a programming model, and a compiler are to be addressed.

Low-level sensory data processing in many Internet-of-Things devices pursue energy efficiency by utilizing sleep modes or slowing the clocking to the minimum. To curb the share of stand-by power dissipation in those designs, near-threshold/subthreshold operational points or ultra-low-leakage processes in fabrication are employed. The upcoming sub-1V solid-state electronics, operating in near-threshold voltage, provides unprecedented energy efficiency at peak performance. With assumptions in operand distribution, data aggregator architecture to convert the address/interface toggles to logic-state toggles, and an SRAM architecture in low-leakage technology targeting up-

to-20 nm, benefits of near-threshold operational point and massive parallelism are optimum energy consumption per instruction operation and minimized memory roundtrips. Examples of rapidly growing applications include Advanced Driver Assistance Systems, data gathering using drones, surveillance systems and service robotics.

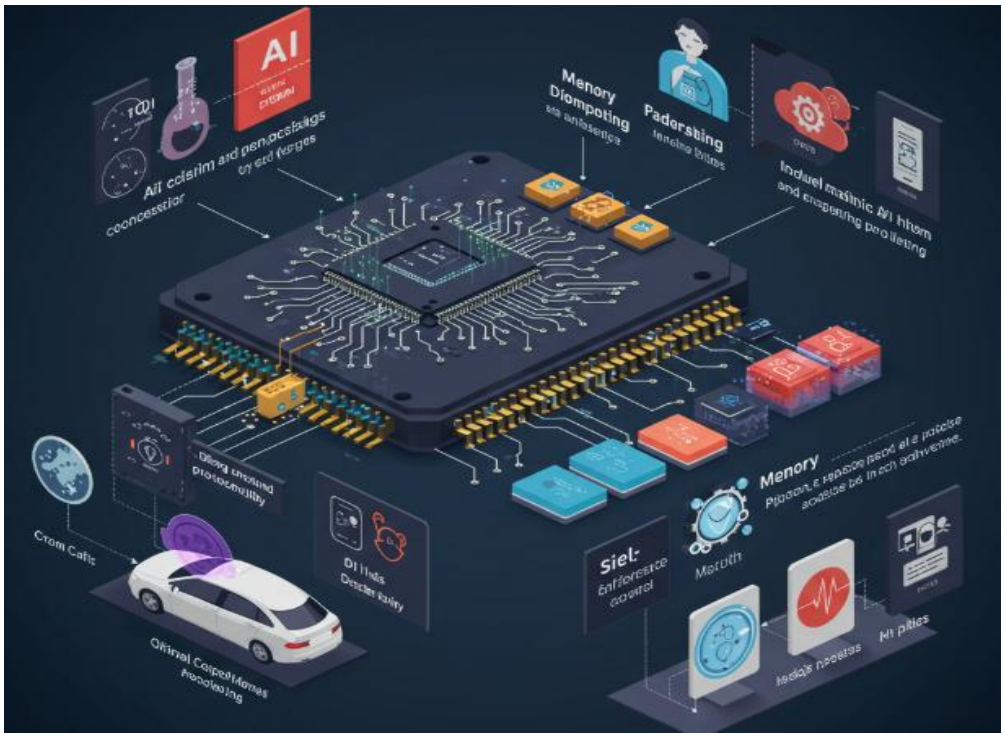

Fig : AI Chips Explained How AI Chips Work, Industry Trends, Applications

6.5.1. Parallel Processing Techniques

In a parallel computing architecture, multiple processors perform calculations and processes simultaneously, thus achieving high throughput while consuming low power. First, processing flow, algorithm structure, and processing unit structure are studied. Both output-oriented and data-flow-oriented parallelisms are adaptive in processing flows. Early polling mechanisms reduce the delay time for a manageable data population. Data flow structures rely on data to decide processing engines and processing routes. Fixed-rate parallelism achieves real-time inference of slow-moving objects while processing units are also adopted for lower power consumption. Localized parallelisms are applied to algorithms with a neighborhood or hierarchical settings and include maximum area trees in object detection and adaptive grid-searching in key-points

matching. Within processing units, layer-wise parallelisms are adapted or newly proposed to enable parallel computing of various layers in CNNs. Weight reuse can save parameter storage by changing interconnects in convolution operations while directly storing parameters into routers can save frequency cycles for fast weight-fetching. Post-firing data representation can adaptively avoid dropping throughput at low neuron firing rates. Pixel-architectured MASH can perform high-speed and low-power operations of upsampling, convolving, and deleting hotspots pixel-wise. Letting devices perform initial NMS but still leveraging GPGPU for final processing can reduce process board area and complexity and save power while achieving high throughput.

Second, architecture-level pruned algorithms are designed as prototypical architectures. Dividing maps to multi-pipelines and processing separately can increase throughput. Reduced networks decimate computation while retaining accuracy rates. Modified region linking rapidly connects regional pools with exhaustive searches for specific pairs. Cross-scanning designs maximize coverage for sharp foveation and uneven center-point distribution. Design-level re-routed MacSL can enable parallel computing of not only weight but also data. Multi-rolling-tiles designs can achieve more than completed sign-adding of weights while consuming less than. Searching for coarse aspect ratios greatly shrinks the searching space while saving reselection reiterations for regular grids. Latency-insensitive designs can make the frame process shareable. Rescaling and division in histogram calculation can enable parallel computation while preserving quantization storage for higher precision. Cap-based ports avoid blocking by adjusting access forces while still keeping simplicity.

6.5.2. Pipeline Architectures

The design of real-time inference systems for edge devices requires the careful co-design of hardware and algorithms. On one side, large Deep Neural Networks (DNNs) must be compressed to a size suitable for on-device inference while guaranteeing a maximum achievable accuracy. On the other side, these compact models must be mapped onto specialized hardware, such as low-power AI accelerators. This latter step implies the adaptation of algorithmic design, data representation, and computation in order to maximize the effective usage of both parallelism and sparsity in modern DNNs.

High-throughput devices normally rely on SRAM-type Memory Hierarchies (MHs) and attentively designed dataflow to offer a high bandwidth to the computing units. However, such a design can be overly expensive in terms of area and power/energy. In contrast, light-weight and restricted edge devices target cheaper and lower-range technologies. They typically rely on a low power DRAM-like architecture for weight storage, while the slow burden is alleviated through concurrent data transfer and

computation at multiple time scales. As a result, these devices may not need to be fixed with a large-integer data representation as in high-throughput devices.

Such a distinct design space also brings many changes in the algorithm side of DNNs for edge devices. For example, popular model compression techniques allow for weight-copy minimizing retraining. By virtue of this design goal, local sensitivity-aware truncated gradient updates play a key role to prevent performance plunge in data-format adaptations. Partnerships are also necessary for the co-design of DNNs and cross-layer architectures given their high intertwining. DNNs for edge inference devices require the exploration of extra robustness characteristics that are not desirable in high throughput edge devices.

6.6. Conclusion

Real-time AI analysis, processing, and decision-making are integral and rapidly growing capabilities of edge devices and subsequent autonomous systems. This will further catalyze the adoption of various edge AI applications, such as smart homes, smart surveillance cameras, smart factories, autonomous vehicles, and drones. Significant progress has been made in the AI algorithms and methods for deep learning and data analytics. However, the advances in enabling circuits and systems have not kept pace with the exponential increase in AI model size, complexity, and compute requirements. New architecture, circuit, and system strategies must be developed to address the concerns of high power/energy consumption, performance, cost, and die area. Given the fast-growing edge AI market, these issues are even more pressing for real-time system developments. Notably, deep learning and segmentation model size and edge device power/energy budgets must be considered in architecture spatiotemporal design. Infrastructure innovation, including scalable modeling/simulation and pre-silicon deployment software, is also essential for rapid commercialization and fast design and implementation iterations. Notably, circuit and system technologies for wide-bandwidth I/O and A/D with high energy efficiency, linearity, noise performance, and flexibility are critical for next-generation dynamic vision sensor designs. Multi-modal sensing-input fusion for vision-edge is a promising direction for achieving noise robustness, low power consumption, and compactness. As AI makes increasingly autonomous decisions and controls robotic platforms, the traditional separation of perception and control tasks remains a significant challenge. Continuing efforts in novel chip architectures, near-sensor computations, algorithm-computable mappings, and co-design between the algorithm and system level will offer edge-based, energy-efficient, and low-latency solutions. Further funding is needed by both government and industry in collaboration with academic institutions to develop foundational chip and system technologies for autonomous perception, processing, and robotics. This is essential for fostering a self-

sustaining cycle between autonomous systems, infrastructures, applications/speculations, and landscape transformations across industries and societies and nations.

6.6.1. Emerging Trends

The introduction of Artificial Intelligence (AI) systems and algorithms across a range of applications has prompted researchers to develop hardware architectures to accelerate a key category of algorithmic approaches: neural networks. With rapid advancements in computation power, neural networks have been adopted for many commercial applications. These high-performance-and-low-power chips enable edge AI, which refers to performing AI inference on the edge and is key to next-generation smart devices. The design of these chips typically tackles two interrelated criteria: throughput and power efficiency, which in turn depends on the architecture and engine. Emerging devices such as nonvolatile memory can effectively reduce chip power and area usage, performing computation and storage in the same place, which is key to enabling more powerful and compact architectures. Even as new devices continue to emerge, dedicated Digital Signal Processing (DSP) and Tensor Processing Unit (TPU)-type chips are required to process widely used conventional deep learning algorithms, and few alternative solutions have proven attractive at scale.

The interest in Edge AI has prompted a flood of research in accelerators and new architectures for neural networks run on digital systems. Digital neural chips have continued demonstrating accelerated performance, leading to deployment in major applications such as target tracking, computer vision, and natural language processing models. As methods develop outside of the typical Deep Learning (DL) architecture, the diversity of general-purpose structures will become increasingly important. Post-CMOS approaches have again flourished alongside digital systems. New physics have sparked novel devices that enable computation to occur with orders of magnitude improvement in efficiency, with properties that could either be exploited or replicate effective behavior in more robust architectures. Emerging edge devices, typically battery-powered portable devices, generate data locally. On-device AI is important to ensure user privacy and reduce network bandwidth. However, such devices continue to be limited by computing and battery life constraints. At present, mobile or edge devices for efficient AI computations typically utilize GPUs to compute fast, with the expected trade-off being poor power performance compared to alternatives for small networks. Proposed new hardware architectures include orthogonal/core devices and in-memory computing chips. The task of the model post-training quantization is to concentrate activity, make minimal modifications to the weight vector, and in-rode data into weight buffers.

References

Shuvo, M. M. H., Islam, S. K., Cheng, J., & Morshed, B. I. (2022). Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. Proceedings of the IEEE, 111(1), 42-91.

Lin, W., Adetomi, A., & Arslan, T. (2021). Low-power ultra-small edge AI accelerators for image recognition with convolution neural networks: Analysis and future directions. Electronics, 10(17), 2048.

Venkataramani, S., Srinivasan, V., Wang, W., Sen, S., Zhang, J., Agrawal, A., ... & Gopalakrishnan, K. (2021, June). RaPiD: AI accelerator for ultra-low precision training and inference. In 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA) (pp. 153-166). IEEE.

Omidsajedi, S. N., Reddy, R., Yi, J., Herbst, J., Lipps, C., & Schotten, H. D. (2024). Latency optimized Deep Neural Networks (DNNs): An Artificial Intelligence approach at the Edge using Multiprocessor System on Chip (MPSoC). arXiv preprint arXiv:2407.18264.

Santoso, A., & Surya, Y. (2024). Maximizing Decision Efficiency with Edge-Based AI Systems: Advanced Strategies for Real-Time Processing, Scalability, and Autonomous Intelligence in Distributed Environments. Quarterly Journal of Emerging Technologies and Innovations, 9(2), 104-132.

Cheng, L., Gu, Y., Liu, Q., Yang, L., Liu, C., & Wang, Y. (2024). Advancements in accelerating deep neural network inference on AIoT devices: A survey. IEEE Transactions on Sustainable Computing, 9(6), 830-847.