WHITE PAPER

Enabling Device Intelligence

AI Chip Design from the Data Center to the Edge

Authors

Stelios Diamantidis
Director, AI Products,
Design Group

David Hsu
Product Marketing Director,
Verification Group

Ron Lowman
Product Marketing Manager,
Solutions Group

The Emergence of Artificial Intelligence

The explosive growth in silicon and software for artificial intelligence applications is transforming everything we know about connectivity, energy-efficiency, mobility, and security. Machine learning (ML) techniques are already used in computer vision, object recognition, speech recognition, and big data analytics. Deep learning (DL) algorithms and neural networks are pushing both silicon and software to meet new requirements for processing power, memory latency, and real-time connectivity.

The need for AI acceleration is driven by deep learning tasks: training and inference. Highly specialized processors, or AI accelerators, are emerging to manage the massive and changing compute intensities of these tasks. In data centers, AI accelerators with highly parallel, largely replicated computation topologies are being used to train tens to hundreds of millions of neurons at a fraction of the power requirements of general-purpose CPUs and GPUs. With >8.5 billion smartphone shipments expected by 2021 (ref: Gartner, March 2017), mobile-edge AI accelerators—some as small as 1mm2—will soon become the world's largest computing environment, already able to push trillions of calculations through their trained neurons in mere milliseconds to meet the needs of interactive applications (see Figure 1).

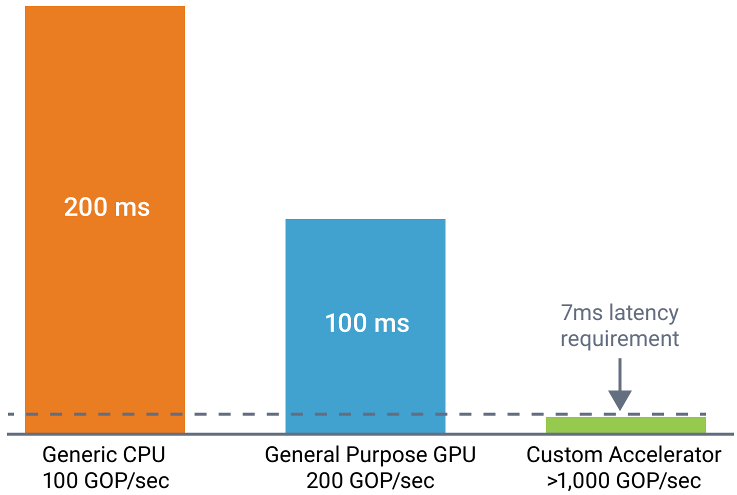

Figure 1: Meeting the inference latency requirements for a 55-layer, 34-million-weight recurring neural network (RNN) used in speech recognition

AI Accelerators: Highly Dynamic and Evolving

A broad range of hardware platforms have emerged to address the computational needs of AI from the data center to the edge. However, creating the chips for AI applications is not a trivial task. Designers must navigate the many technological challenges associated with a very broad set of AI algorithms and corresponding heterogeneity of hardware architectures. Designers also need to overcome the complexity and cost of high-performance, low-power physical design.

Algorithmic Innovation

Research emerging from both university and industry labs is producing new neural network models, enhancing existing ones, and generating massive datasets to train and test them. Once these models are trained, additional innovation is occurring to compress and map those models to different hardware architectures. Innovators need to balance competing requirements such as bandwidth (e.g., number of multiply-accumulate (MAC) operations per second per convolution), impact of quantization or data type selection, area efficiency (frames per second per mm2), memory bandwidth efficiency (MB/frame), and more (see Figure 2).

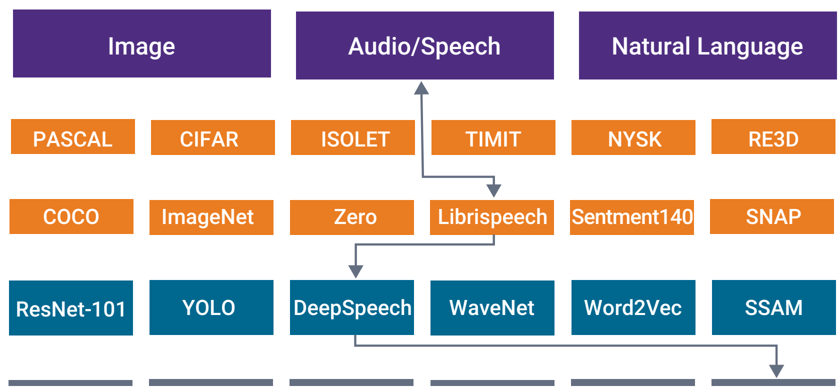

Figure 2: Mapping AI applications to hardware acceleration platform

Diverse Architectures

A broad, heterogeneous range of compute architectures are being proposed to accelerate computations while reducing overall power per operation. Each AI application has specialized compute, memory, and interconnect requirements. Beyond the AI accelerator functionality itself, AI chips include a broad variety of other components. For example, a data center device must have reliable and configurable connectivity to AI data centers, while an edge device will include real-time interfaces to sensors, images, audio, and more (see Figure 3). Memory selection is particularly critical to meeting low-latency access requirements at low power.

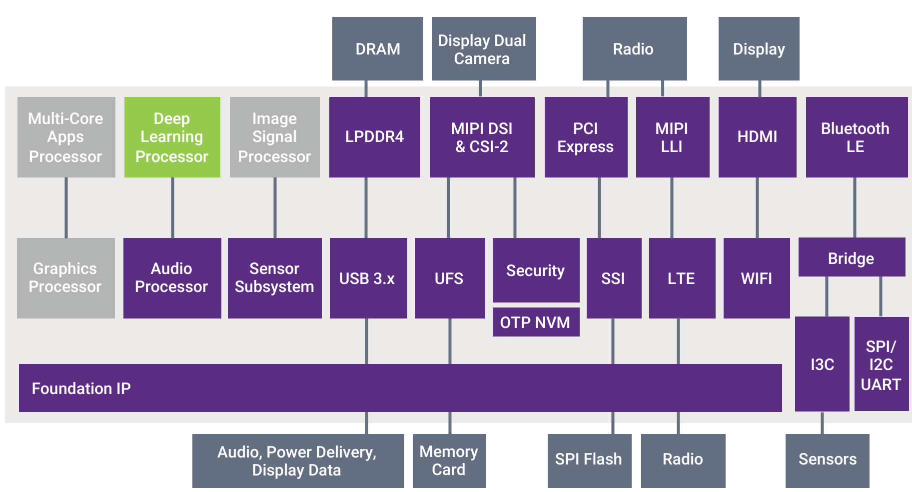

Figure 3: An SoC architecture for augmented/virtual reality at the edge, including a deep learning accelerator

High-performance Design

Recent data center-oriented architectures are packing together more than 20 billion transistors and hundreds or thousands of processing modules at speeds that can exceed 5GHz. New embedded memory topologies to support data locality at low latency are leading to transistor-dominated layouts, with routing congestion and macro timing path closure challenges. At the edge, designers are running 1+ GHz inference engines that need to operate in extreme temperature and voltage corners. Meeting power and thermal budgets that vary per algorithm and dataset is calling for new approaches to power estimation and analysis.

A comprehensive solution for AI chip design must therefore address all these aspects of AI design—from accelerating the algorithmic innovation phase, to reliably piecing together diverse architectures, to finally providing the best possible physical implementation all the way through manufacturing signoff.

Synopsys: Enabling AI Design from the Data Center to the Edge

As the Silicon to Software™ partner for the world's most innovative companies, Synopsys has helped develop many of the electronic products and software applications we rely on every day. Synopsys works with data center and cloud service providers for training and inference; automotive semiconductor leaders for autonomy, vision processing, and decision-making; and mobile/IoT providers for deep learning accelerators to optimize performance and enhance privacy. As a result, practically all AI accelerators in data centers worldwide were designed and verified with Synopsys software. Synopsys is also working closely with many AI startups that benefit from our domain knowledge and solutions optimized for AI.

AI Architectural Exploration

Synopsys' Verification Continuum provides unique solutions for speeding up and optimizing AI architecture exploration. Platform Architect virtual prototyping enables architectural-level performance and power analysis. HAPS® FPGA-based prototyping and ZeBu® emulation make exploration and verification of extraordinarily large and complex RTL implementations practical.

Co-verification of AI Accelerator and Data Framework

RTL simulation of three, sixteen-by-sixteen pixel images on a convolutional neural network (CNN) is beyond the current state-of-the-industry for any software simulator. Synopsys ZeBu is the industry's fastest emulation system and the only proven solution to handle the capacity and speed needs of full AI chip emulation. It offers the highest capacity (19 billion+ gates) and lowest cost of ownership (5X lower power consumption with half the data center footprint) compared to other solutions. ZeBu has been equipped with AI performance visualization capabilities, including graph traceback, tensor graph throughput analysis, memory performance analysis, and more.

Optimization of AI Models for Hardware Acceleration

Synopsys ASIP Designer is the industry-leading tool for designing fully programmable processors and AI accelerators. ASIP Designer automates implementation of highly-parallelized, yet fully software-programmable hardware with custom datapaths, and optimizes AI models for hardware processing and software algorithm iterations (see Figure 4).

Industry's Broadest Range of Silicon-proven AI-ready IP

Synopsys' silicon-proven DesignWare® IP portfolio addresses the diverse processing, memory, and connectivity requirements of AI markets, including mobile, IoT, data center, automotive, and digital home. Processors and convolutional neural network (CNN) engines manage massive and changing compute requirements for machine and deep learning tasks; memory IP solutions support efficient architectures for different memory constraints, including bandwidth, capacity, and cache coherency; interface IP solutions provide reliable connectivity to CMOS image sensors, microphones, and motion sensors for AI applications, including vision, natural language understanding, and context awareness.

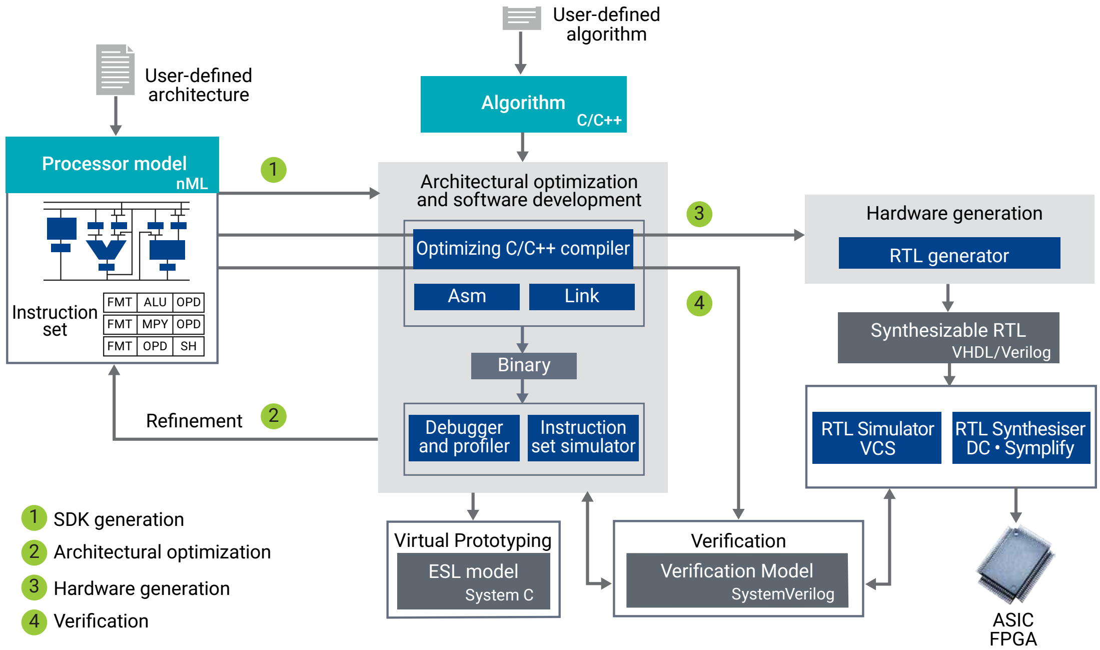

Figure 4: Synopsys ASIP Designer

Pushing the Envelope on Performance, Power, and Area (PPA)

The Synopsys Design Platform with Fusion Technology™ is the preferred choice for developing advanced digital, custom, and analog/mixed-signal designs with the best power, performance, area, and yield. The technology behind 90% of FinFET designs worldwide, Synopsys' digital design implementation solutions have been enhanced with several key AI-focused optimization technologies, including AI chip interconnect planning for hundreds of replicated modules, MAC topology optimization, complete AI IP reference flows, and more. Close collaboration with the world's leading foundries ensures that the Synopsys Design Platform delivers the maximum possible benefit of leading process technologies while meeting the stringent requirements of AI chip schedules.

Innovating with AI-enhanced Design Tools

AI technologies, like machine-learning, can help address high-complexity, high-cost challenges not only for AI designs, but for all kinds of designs. Exemplifying the disruptive power of AI, Synopsys' AI-enhanced Design Platform and VC Formal Regression Mode Accelerator learn continuously and improve performance in customer environments—a marked departure from traditional systems. AI-enhanced tools boost designer productivity by speeding up computationally-intensive analyses, predicting results that drive better decision-making, and leveraging past learning to intelligently guide debug.

Summary

Many innovative applications are driving growth for chips with AI capabilities. Deep neural networks require specialized accelerators that, in turn, introduce new architectural requirements for compute, memory, power, and connectivity. Synopsys offers a comprehensive solution that addresses all aspects of AI design from accelerating the algorithmic innovation phase, to piecing together diverse architectures, to providing the best possible physical implementation while maximizing the benefits of leading foundry process nodes.

Industry leaders, as well as most of the world's most innovative AI startups, rely on Synopsys to implement AI chips from the data center to the edge.

For more information on Synopsys AI design solutions, please visit www.synopsys.com/ai.

09/17/18.CS269318085_AI-Whitepaper.

Pub: Sept. 2018