IDC ANALYZE THE FUTURE

IDC ANALYST CONNECTION

Sponsored by: Intel

Shipping systems' total compute capability versus the future data volume to be computed means a compute capability shortage looms. Can artificial intelligence deal with this dilemma, and what must technology providers do to meet its needs?

Artificial Intelligence Requires Tailored Solutions from Technology Providers

August 2019

Analyst: Shane Rau, Research Vice President, Computing Semiconductors

Q. How does IDC define "artificial intelligence" and what does it do?

A. Artificial Intelligence (AI) is the study and research of providing hardware and software that attempt to emulate a human being. Just as human beings collect, process, communicate, store, analyze, and respond to data to perform tasks, so does AI.

Despite the science fiction of AI as general-purpose robots and clones, AI is specialized and takes on many discrete forms. In fact, for every task a human can perform, AI requires a unique combination of hardware and software components. For example, AI semiconductors alone span seven different chip types: Microprocessors, Microcontrollers, Systems-on-chip, Graphics Processors, Field Programmable Gate Arrays, Application Specific Integrated Circuits (custom AI silicon), and Application Specific Standard Products (merchant AI silicon).

The needs of AI thus require one's technology provider to have an extensive portfolio of IP and capabilities, including multiple data processing semiconductor types, integration capabilities, and sophisticated software development environments that help a technology buyer tailor the AI to its specific task.

Q. Why do we need AI and why AI now?

A. The foundation for AI is data. In the past 15 years, the advent of reliable and ubiquitous connectivity—cellular, WiFi, Bluetooth, and many more forms—has enabled the creation and transmission of data from a wide variety of newly connected systems. PCs and servers are no longer the locus of computing. Smartphones, the Internet of Things (IoT), and the Cloud exist because of reliable and ubiquitous connectivity.

The connection of billions of new systems to the Internet has had two major consequences:

- More people connected with more devices that they are using for longer periods of time AND more machines connected to each other has created data of such volume and sophistication that it surpasses the human ability to program with traditional methods of data collection, processing, transmission, storage, and analysis. Data creation

between 2014 and 2023 will grow from 11.5 zettabytes to 103 zettabytes, a compound annual growth rate of 28 percent.

» The ability to connect and compute nearly anywhere faces hardware and software technology developers with the question: What kind of computing should be done where and by what? This is a question which encapsulates the concept known as distributed computing.

The huge amounts of data being generated, processed, transmitted, stored, and analyzed demand that primary clients (PCs, phones, tablets), edge systems (including IoT systems), and datacenter systems (servers, storage systems, networking infrastructure) that are oriented toward the limited bandwidth and capacity of human management and control be reoriented toward the hyper-efficient bandwidth and capacity of machine management and control. This reorientation requires faster processing optimized for AI.

Q. Where is AI having impact now? Five years from now? Ten years from now?

A. The initial wave of AI penetration has been in servers, often built and deployed by cloud service providers using AI to train their AI algorithms to one or more of several leading use cases, including recommendation systems, natural language recognition, supply and logistics functions, and entertainment. As the number of discrete AI accelerators proliferate and microprocessor providers integrate AI acceleration instructions, AI penetration in servers offered by OEMs and cloud-service providers is headed to 90% or more in the next five years.

AI penetration that enables fully autonomous vehicles will take another five to ten years to develop.

The second wave of AI penetration, now and over the next five years, is in consumer smartphones and smart home speaker systems, which need to listen and interpret user requests and respond quickly. This process of inferencing currently relies on the algorithms based in datacenters (today's generation of smart home speakers, like Google Home, don't infer responses locally). However, the need to reduce response times (latency) requires more local intelligence. In 2023, 100% of smartphones shipped worldwide and 100% of smart home speakers shipped will be AI enabled.

Automotive systems, including advanced driver assist systems (ADAS) and automotive control devices, get much attention now for their role in enabling fully autonomous vehicles but, while the need for short latency is compelling, AI penetration that enables fully autonomous vehicles will take another five to ten years to develop. In 2023, 100% of the ADAS (consisting of the radar systems, control systems, and visual systems) that ship in automobiles will be AI enabled.

The field of opportunity for AI is immense. In 2021, 24.5 billion systems across all industries—factory automation, healthcare, communications, retail, transportation, and so on—will ship. If we consider systems that run AI in software with or without a specific accelerator, then the trajectory for AI penetration rises further: 75% of enterprise applications will use AI by 2021. Though AI adoption will vary by industry, AI will become universal over the long term.

Q. What does AI's impact mean for technology providers?

A. Notable among the system categories that will adopt AI is the disparate use cases they will serve. Servers, phones, smart home speakers, and cars vary significantly from each other in the kinds of AI solutions that they will need. To enable these systems, technology buyers will demand disparate solutions of technology providers. No one solution fits all.

Heterogeneous computing is the mixing and matching of the right hardware and software components to serve an intended use case. In the next five years, AI solutions will reflect an increasing diversity to suit a huge variety of use cases based on requirements for:

- Performance: Raw computing power is needed to enable a system to respond quickly to incoming data streams, make decisions, and respond while minimizing latency.
- Cost: For AI to become ubiquitous, it will have to fit within the bill-of-materials requirements of high-volume, low-cost systems, especially in consumer devices.
- Power consumption: Many consumer devices and embedded systems categories like smartphones, portable point-of-sale systems, and remote controls are battery powered. AI, known for compute intensity, thus will have to be designed to be battery sensitive.
- Space: AC-powered and battery-powered embedded systems often share a strong concern for space to physically accommodate components but also to move heat generated from those components. Thus, AI will have to be designed to fit on very small discrete chip dies or be integrated onto the same die with other chips.

Q. What's at stake for AI technology buyers and what should they demand of technology suppliers?

A. Tech buyer spending on discrete AI semiconductors alone (GPUs, FPGAs, AI-ASICs, and ASSPs) will be $8.5 billion by 2023.

To earn that revenue and entrench themselves in customer designs for the larger future opportunity, technology providers will ultimately need to adapt and scale compute capabilities to support the billions of systems across the Internet and the sensors in those systems that are the original source of data capture.

To support heterogeneous computing, tech buyers should ask these key questions of their AI tech supplier:

- How many data processing choices do you have to optimize for my specific AI use case? Do I even need an AI accelerator?
  - Beyond a host microprocessor, if one or more accelerator types will be needed, they would be applied based on the AI use case and data types being processed.
- My use case has significant cost, power, and space restraints. What do you offer for AI integration?

- The ability to integrate AI accelerators into other silicon types or even run AI through software on microprocessors or SoCs will be key to AI's adaptability and ubiquity and its effect on TCO.

- What tools do you offer to program across different discrete and integrated AI chip types?
  - One needs sophisticated software development environments. These environments encompass operating systems, drivers, virtualization and container tools, learning frameworks that can enable systems to deploy AI compute tasks across their available processors and accelerators in real time.

About the Analyst

Shane Rau, *Research Vice President, Computing Semiconductors*

Shane Rau leads IDC's computing semiconductor research within IDC's Enabling Technologies team. Mr. Rau's research covers microprocessors and SoCs, discrete graphics processors (GPUs), FPGAs, and artificial intelligence (AI) accelerators in systems across the Internet, including in the datacenter, in PCs, and at the edge, such as embedded and intelligent systems.

Mr. Rau provides in-depth insight and intelligence on market sizing, forecasting, technology trends, vendors, pricing trends, and market share. Through collaboration with PC, server, and embedded systems analyst colleagues, Mr. Rau spearheads IDC research initiatives into system supply chains, technologies, and interface attach rates, as well as into the changing semiconductor vendor market power dynamics.

IDC Corporate USA

5 Speen Street
Framingham, MA 01701, USA

T 508.872.8200

F 508.935.4015

Twitter @IDC

idc-insights-community.com

www.idc.com