The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

ARCHITEK OFFERS VIRTUAL ENGINES

Startup Uses Configurable Architecture for Edge AI

By Linley Gwennap (August 30, 2021)

Among the many edge-AI startups, ArchiTek offers a unique architecture employing fixed-function accelerators that are configurable for different algorithms. Using these "virtual engines," the Aionic architecture can handle many image-processing and AI algorithms. It is quickly reconfigurable for different tasks, enabling it to run multiple algorithms at essentially the same time. This flexibility allows the design to replace GPUs, DSPs, image processors (ISPs), and deep-learning accelerators (DLAs) in camera-based systems.

The Japanese startup has already tested two prototype chips to validate its architecture. It's designing a new chip, called Chichibu, that it intends to sell to customers. The SoC targets peak AI performance of 3.6 trillion operations per second (TOPS) using 8-bit integer (INT8) data with a TDP of just 1.5W. This performance should enable it to process high-definition images at 24 frames per second (fps). ArchiTek expects tapeout in 2Q22 and volume production in 4Q22. It will also license the Aionic architecture for integration into customer chips; the first such design is already under way.

Shuichi Takada is CEO and CTO of the 20-person startup, which he founded in 2011 along with three colleagues from Panasonic. Based in Osaka, the company worked with Toyota to develop a small self-navigating robot using its prototype chip. It also licensed Aionic to Japanese ASIC vendor SocioNext, which is designing a higher-performance chip for an unnamed equipment vendor. ArchiTek has raised ¥1,000 million (about $9 million) and is seeking additional funding to complete the Chichibu design.

Putting AI on IC

The Aionic architecture relies on a set of fixed-function engines that accelerate specific functions, such as fast Fourier transforms (FFTs), computer vision (CV), video/image compression (JPEG), sort, matrix operations, and signal processing. Each engine operates on data in the on-chip cache, as Figure 1 shows. The hardware-based scheduler controls these engines, directing them to perform specific operations on a particular data block. ArchiTek's software takes an algorithm, divides it into basic operations, and downloads a set of instructions to the scheduler. Thus, the architecture can perform a variety of algorithms by simply varying the order in which the scheduler invokes the engines.

Because neural networks are useful for image processing, the architecture includes a larger engine for that function. This DLA implements a scalable array of MAC units that can perform depthwise convolutions for any common kernel size. The units support 2-, 4-, or 8-bit integer weights and feature 32-bit accumulators. In Chichibu, the array is 32x32x6, yielding 3.7 TOPS at 600MHz. The

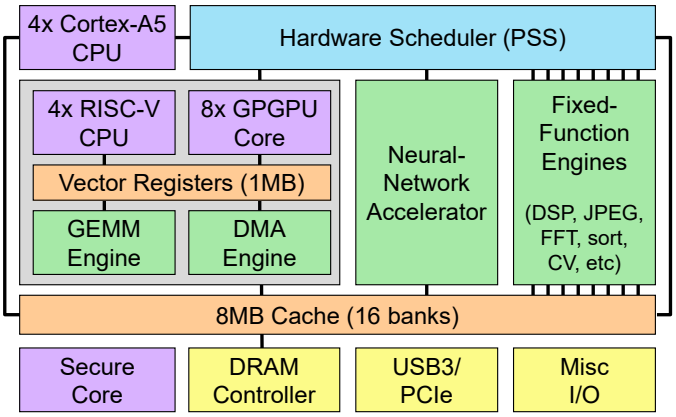

Figure 1. ArchiTek chip design. The company's first product, code-named Chichibu, relies on a data path comprising mainly fixed-function accelerators. Some general-purpose CPUs and GPUs handle functions that the accelerators can't.

neural-network unit also provides acceleration for pooling and activation functions (e.g., ReLU).

The architecture has a cluster of programmable cores for operations unsupported in the fixed-function engines. Chichibu will have eight general-purpose-GPU (GPGPU) cores that can handle floating-point operations the DLA can't. Four small RISC-V CPUs can execute any function the other units can't handle. Although the chip will include four Cortex-A5 CPUs as well, they're mainly for supervisory tasks. By placing the RISC-V CPUs in the data path, ArchiTek reduces the latency for general-purpose operations.

These CPUs implement the standard RISC-V instruction set, but rather than adapting an open-source core, the company chose to create a custom microarchitecture, working in conjunction with a local university. The design implements a superscalar CPU with multiple threads for fast context switching. These CPU and GPU cores share a large (1MB) SRAM-based register file that reduces pipeline stalls due to memory references and that enables higher frequency. The custom GPU design includes a Cordic unit that can compute hyberbolic functions in a single cycle.

Riding the Chichibu

The chip will have 8MB of cache memory that serves the fixed-function engines, the DLA, and the general-purpose cores. The cache also feeds the Arm CPUs and supplies instructions to the scheduler. This memory is divided into 16 banks and, as long as there are no bank conflicts, can serve up to 16 requests per cycle for the numerous engines. For similar reasons, the vector registers are divided into eight banks. The chip can load an image into the cache and then run the desired algorithms, allowing all the cores and engines to access the pixels as necessary.

The cache automatically loads data from the external DRAM as needed. Chichibu will have a single 32-bit channel that supports up to LPDDR4-2400 memory. It also connects directly to external cameras using two MIPI CSI2 ports. A single USB3.0 port connects to an external host processor or other high-speed peripheral; it's configurable as a single PCIe Gen3 lane as well. The design features several low-speed serial ports for connecting other standard peripherals. A secure core holds encryption keys and validates the boot code.

	ArchiTek Chichibu	NXP i.MX8M Plus	Google Edge TPU
Main CPUs	4x Cortex-A5	4x Cortex-A53	None
CPU Speed	1.2GHz	1.8GHz	Not applicable
DLA Cores	Aionic	VIP Nano	TPU
Peak INT8 Perf	3.7 TOPS	2.3 TOPS	4.0 TOPS
On-Chip SRAM	8MB	1MB	8MB*
DRAM Interface	32b LPDDR4	32b LPDDR4	None
PCIe Interface	1x Gen3	1x Gen3	1x Gen2
Camera Interface	2x MIPI	2x MIPI	None
IC Process	12nm	14nm	28nm*
Power (TDP)	1.5W	3.0W	2.0W
Production	4Q22 (est)	3Q20	1Q19
List Price (1ku)	Undisclosed	$10–$20	$15–$20*

Table 1. Edge AI comparison. ArchiTek's first chip is similar to other edge-AI chips in performance despite using less power, but these older products may receive an upgrade before Chichibu reaches production next year. (Source: vendors, except *The Linley Group estimate)

ArchiTek expects the Chichibu chip will require 25mm^2 in TSMC's 12nm technology; the 3.7-TOPS DLA consumes only 4mm^2. At its 1.5W TDP, the chip is rated at a respectable 2.4 TOPS/W. The company estimates it will execute the rigorous Yolo v3 object-recognition model at 24fps at 720p resolution, or 22 million pixels per second (Mpixel/s). This model is far larger than the cache memory, so the chip must load parameters from DRAM as it runs.

As with most AI chips, software is a critical performance factor. ArchiTek has created a proprietary library called aipe_dnn that implements AI operations such as convolutions, normalization, pooling, and activation functions on its DLA. Programmers can call these functions directly or use the startup's tools to convert Keras or ONNX models to aipe_dnn. ArchiTek also offers a computer-vision (CV) library for functions such as filters, remap, and resize. Customers can directly program the CPUs through standard Arm and RISC-V tools, and the startup supplies an LLVM compiler to program the GPGPU cores. Its performance testing employs hand-coded models that use aipe_dnn, however, and thus doesn't indicate the effectiveness of the model-conversion tools.

Chichibu is just one implementation of the Aionic architecture. The company believes it can scale the architecture to 4K resolution by enlarging the DLA by 8x, expanding the cache to 24MB, and doubling the DRAM bandwidth while keeping the rest of the design essentially the same. The resulting chip would deliver 29 TOPS at an estimated 6W, doubling AI performance per watt. Die area would increase less than 3x, tripling performance per yen.

A TOPS/Watt Advantage

ArchiTek's product can serve as the main SoC in some systems or connect to an external host processor in others. To capture these different applications, we compare it with NXP's i.MX8M Plus, a popular SoC with AI acceleration, and Google's Edge TPU, which requires a host processor. A big difference among these chips, naturally, is the main CPUs. The i.MX chip has four 64-bit Cortex-A53 CPUs running at 1.8GHz plus an 800MHz Cortex-M7 MCU, offering considerably more compute power than Chichibu's 32-bit Cortex-A5s. The lower performance limits the applications for which the ArchiTek chip can serve as the host. Since it's designed as a coprocessor, the Edge TPU lacks any Arm cores.

For neural networks, Chichibu offers about the same peak TOPS and SRAM size as the Edge TPU, as Table 1 shows. The TPU chip, however, lacks a DRAM interface, greatly reducing its performance on models that don't fit

into its on-chip memory. Chichibu can run larger models such as Yolo v3 using its DRAM interface. The i.MX8 processor also employs DRAM, but its lower TOPS rate and small on-chip memory restrict performance relative to the ArchiTek design. Whereas the TPU is solely for neural networks, Chichibu can handle image and signal processing as well using its flexible Aionic architecture. The i.MX8 features a HiFi DSP and VeriSilicon GPU to assist with these tasks; the A53 CPUs include Neon SIMD units for software acceleration.

As a coprocessor, the Google chip connects to the host processor via PCIe Gen2 to access cameras and other I/O devices. Both the ArchiTek and NXP chips link directly to two MIPI cameras as well as a variety of other peripherals. The i.MX offers additional high-speed interfaces such as Ethernet, CAN, and HDMI (see *MPR 2/3/20,* "Sharper Vision, Brains in i.MX8M Plus").

ArchiTek expects its first product to require 1.5W TDP, a bit less than the Edge TPU's power. As a result, Chichibu should deliver slightly better TOPS/W than the coprocessor, although the two-year-old TPU is hindered by its trailing-edge manufacturing process. The i.MX processor burns twice as much power as Chichibu, but part of the difference is the greater consumption of the faster CPUs and additional high-speed I/O. ArchiTek withheld the list price, but its chip will have to sell for less than $20 to be competitive with these similar products.

A Flexible Fixed-Function Accelerator

ArchiTek attempts to combine the efficiency of fixed-function accelerators with the flexibility of a programmable processor. Its Aionic architecture offloads compute tasks to the accelerators whenever possible while still providing general-purpose cores to handle overhead functions. Because each accelerator performs a relatively simple task rather than an entire algorithm, the scheduler can mix and match these units to perform a variety of algorithms, including neural networks, image processing, and signal processing. This combination makes the design well suited to camera-based applications that need to clean up an image, perform segmentation and object recognition, and process audio, or drive a motor that moves the camera.

Price and Availability

ArchiTek is developing its first product, code-named Chichibu, which it expects to enter volume production in 4Q22. It withheld pricing. The company also licenses its Aionic architecture for ASIC designs through SocioNext. For more information, access *www.architek.ai.*

Despite its small funding, the startup has produced two prototype chips to validate its hardware design. But it hasn't published data to demonstrate the performance of its architecture. It has made improvements in each generation, and even its Yolo score is an estimate based on the future Chichibu design. Although it touts flexibility, ArchiTek has yet to disclose performance on other neural networks or on any computer-vision or signal-processing tasks.

One problem may be the software stack. The startup currently ports neural networks to its proprietary library by hand, because its conversion tools produce inefficient code. The performance of its computer-vision library is unproven. Other AI startups are spending tens of millions of dollars to develop fully functional high-performance software stacks, even without CV functions. ArchiTek needs its new funding round to advance its software as well as its hardware design.

The Aionic architecture shows promise. The Japanese vendor has convinced one customer to use it in an ASIC, and it hopes to sell its own chip into camera-based monitoring systems beginning late next year. ArchiTek has taken a different approach than other edge vendors, designing a flexible chip instead of a simple AI accelerator. This approach could provide an advantage once the company delivers its initial product. ♦