VeriSilicon Launches VIP9000, New Generation of Neural Processor Unit IP

2019-08-07

Device Will Help Resolve System Level Bottlenecks for More Efficient AI Solutions

Shanghai, China, August 7, 2019 - VeriSilicon, a Silicon-Platform-as-a-Service (SiPaaS®) company, today announces VIP9000, a highly scalable and programmable processor for computer vision and artificial intelligence. The Vivante VIP family's patented Neural Network (NN) engine and Tensor Processing Fabric technology delivers superb neural network inference performance with industry-leading power efficiency (TOPS/W) and area efficiency (mm2/W), with scalable compute capability ranging from 0.5TOPS (Tera-Operations-Per-Second) to 100s of TOPS.

VIP9000 adopts Vivante's latest VIP V8 NPU architecture. According to VeriSilicon's Executive Vice President and GM of Intellectual Property Division Wei-Jin Dai, VIP V8 architecture improves the flexibility of data distribution and processing core configurability to adapt to a wide range of filter shapes and sizes in modern neural networks (e.g. 1x1, Nx1, 1xN, depth wise). VIP9000 enables neural network inference with different data formats based on design choice (INT8, INT16, Float16, Bfloat16). VIP9000 also supports hybrid quantization (mixing data formats between neural network operations) natively.

"Neural Network technology is continuing to grow and evolve and there are so many applications across the board when it comes to computer vision, pixel processing for super resolution, and audio and voice processing," Dai said. "In just a few short years, more than 25 unique licensees have adopted VIP technology across a wide range of applications, from wearable and IoT devices, IP cameras, surveillance cameras, smart home & appliances, mobile phones, laptops, to automotive (ADAS, autonomous driving) and edge servers, which is a true testament to the demand of the products and technology."

Industries with AI Vision, AI Voice, AI Pixel, or AIOT applications will benefit from VIP9000. For smart home and AIOT applications, VIP9000 offers several highly optimized, high precision recognition engines. The release contains the following new features:

- **A more flexible data distributor and processing core configurator**: Brings high MAC utilization to a wide range of filter shapes and sizes in modern neural network models;
- **New data format support for Bfloat16**: On top of existing INT8, INT16, and Float16 support, Bfloat16 delivers better accuracy for AI training;

- **FLEXA API support**: A hardware and software protocol that enables efficient data communication between multiple pixel processing IP blocks. Systems using VeriSilicon's ISP, Video CODEC, NPU, or 3rd party IP compliant with FLEXA API can run AI applications with reduced DDR traffics and low pixel processing latency for applications running thru multiple IPS;
- **Task-specific engines designed for speeding up commonly used AI applications**: This allows for face detection, face recognition, facial landmark detection, object detection and AI Voice. One or more engines can run in parallel inside VIP9000 together with user defined AI programs, due to VIP9000's native multi-task, multi-context support.

VIP9000 supports all popular deep learning frameworks (TensorFlow, Pytorch, TensorFlow Lite, Caffe, Caffe2, DarkNet, ONNX, NNEF, Keras, etc.) as well as programming APIs like OpenCL and OpenVX. Neural network optimization techniques such as quantization, pruning, and model compression are also supported natively with VIP9000 architecture. AI applications can be easily port to VIP9000 platforms through offline conversion by Vivante ACUITY™ SDK, or through run-time interpretation with Android NN, NN API, or ARM NN.
"We are proud to be market leaders as NXP and Broadcom adopt our technology, SiPaaS platforms, OEMs and supplier relationships, to drive further development. Our licensees have integrated Vivante NPUs in more than 50 SoC designs, validating our approach in this market where there are many competing architectures," Dai said.

The biggest thing to happen in the computer industry since the PC is AI and machine learning, according to Dr. Jon Peddie, President, Jon Peddie Research. "It will truly revolutionize, empower, and improve our lives. It can be done in giant machines from IBM and Google, and in tiny chips made with VeriSilicon's neural network processors. By 2020, we will wonder how we ever lived without our AI assistants."

About VeriSilicon

VeriSilicon is a Silicon Platform as a Service (SiPaaS®) company that provides world class SoC and SiP solutions and a leading IP provider with the most comprehensive IP portfolios address markets including mobile internet devices, datacenters, the Internet of Things (IoT), automotive, industrial, and medical electronics. Our turnkey service takes from concept to a completed, tested and packaged chip in record time as performance effective and cost-efficient service for customers including both emerging and established companies, OEMs, ODMs, and large internet/cloud platform companies. VeriSilicon's Vivante® scalable intelligent pixel processing IPs from camera-in to display-out complete solutions include ISP, Neural Network Processor Unit (NPU), GPU and GPGPU, Hantro® video codec, and display controller, which deliver highly differentiated PPA and QOR on the devices, at the edge, and in the cloud. VeriSilicon's scalable ZSP® based solutions widely applied in HD audio/voice and BLE5.0, Wi-Fi,

and NB-IoT. Founded in 2001 and head-quartered in Shanghai, China, VeriSilicon has over 800 employees with 5 R&D centers in US and in China and 10 sales offices worldwide.
For more details,please contact: press@verisilicon.com.