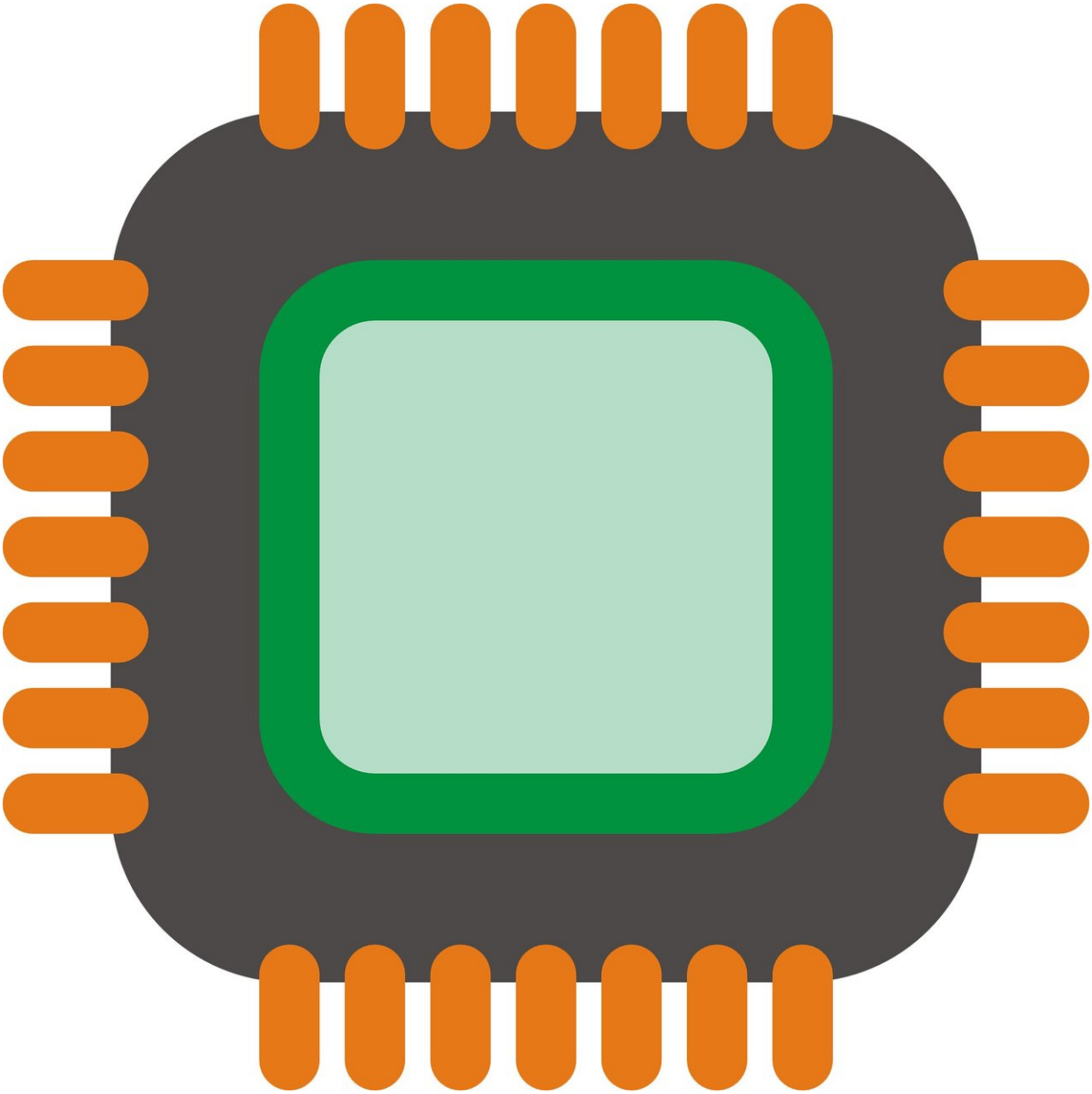
Intel: Hot Chips event details AI-strength processors

August 22 2019, by Nancy Cohen



Credit: CC0 Public Domain

Tech watchers this week got an earful of impressive AI accelerator work at Intel, namely revelations at the Hot Chips 2019 event, where Intel presented details of its Nervana neural network processors, (1) NNP-T for training and (2) NNP-I for inference.

Let's first revisit the inference (Spring Hill) work, and NNP-1, having been developed at its facility in Haifa, Israel. NNP-1 stands for Neural Network Processor for Inference.

Its construct is going to allow it "to cope with high workloads using minimal amounts of energy," said Steven Scheer, Reuters.

Describing its function, Ravie Lakshmananin, *TNW,* said it "specifically targets the inference aspect of AI to deduce new insights. By making use of a purpose-built AI inference compute engine, NNP-I delivers greater performance with lower power."

Where does the name Nervana come from? Nervana Systems is the company it acquired in 2016. At the time, analyst Karl Freund told *EE Times* that it made a lot of sense for Intel to do so. Acquiring Nervana was a way of getting into the deep-learning market.

This week, the question of why was turned into a why-not by an Intel spokesperson.

"In order to reach a future situation of 'AI everywhere', we have to deal with huge amounts of data generated and make sure organizations are equipped with what they need to make effective use of the data and process them where they are collected," said Naveen Rao, founder of

Nervana and now general manager of Intel's artificial intelligence products group, in a report from Reuters. "These computers need acceleration for complex AI applications."

Tech watchers including *SiliconANGLE* were saying that the NNP-1 was good for large data centers running AI workloads. *Fossbytes* said a comprehensive set of RAS features was to ensure that it can be easily deployed in existing data centers.

Joel Hruska in *ExtremeTech*: "Intel claims the NNP-I can deliver ResNet50 performance of 3,600 inferences per second when running at a 10W TDP. That works out to 4.8 TOPS/watt, which meets Intel's overall efficiency goals (the company claims that NNP-I is most efficient at lower wattages)."

The other item of interest at the 2019 Hot Chips conference was the NNP-T, which stands for Intel Nervana Neural Network Processor for Training. Intel described the NNP-T (code named Spring Crest) as purpose-built (1) to train complex deep learning models at massive scale, and (2) simplify distributed training with out-of-the-box scale-out support.

Paul Alcorn, *Tom's Hardware*, wrote about how "the NNP-T is designed to scale gluelessly from chassis-to-chassis, and even rack-to-rack, without a switch." He said the network was specially designed with high bandwidth and low latency in mind; in turn, the architecture is to handle "massive models that scale to 5 or 8 billion parameters, or beyond."

Naveen Rao commented:"Intel Nervana NNP-T pushes the boundaries of deep learning training. It's built to prioritize two key real-world considerations: how to train a network as fast as possible and how to do it within a given power budget." The architecture was built from the ground up, with no legacy workloads to support.

In the bigger picture, *The Times of Israel* said "Companies such as Intel, Nvidia, Qualcomm and Google and startups globally are all on the hunt for new technologies in this field, which involves among other things creating the hardware to enable the processing of huge amounts of information."

The processing hardware has two purposes, wrote Shoshanna Solomon: (1) training the computers to do new tasks and (2) teaching them to infer and thereby reach insights.

All in all, Intel is doing its bit to enable data scientists to do both as they work on unstructured and complex data.

Citation: Intel: Hot Chips event details AI-strength processors (2019, August 22) retrieved 2 May 2024 from https://techxplore.com/news/2019-08-intel-hot-chips-event-ai-strength.html