# Alibaba crowns its cloud service with powerful AI chip

September 27 2019, by Nancy Cohen


Credit: Alibaba Group

Alibaba's first AI chip is in the news. It's a self developed chip, said Reuters, for cloud computing services.

Reuters said in cloud computing, Alibaba towered over rivals in China. On the global level, Alibaba ranked the third in cloud computing after Amazon and Google, said *Nikkei Asia Review*.

It was not lost on tech observers covering the story that China has been pulling for its own semiconductor technology.

*Abacus* referred to that inclination earlier this month when it remarked that "industry veterans worry that it is a technology gap that may never be closed if China continues down the same path of importing foreign technology instead of developing its own, dooming it to dependence on friends who may become tomorrow's foes."

It has been no secret that Chinese tech companies seek to shift away from reliance on foreign semiconductor manufacturers amid US sanctions, said *DCD*.

Jeff Zhang, chief technology officer at Alibaba, unveiled the chip, called the Hanguang 800, and also described as a high-performance AI inference chip.

Inference chip? Michael Copeland in the Nvidia blog can explain what inference means.

He said, "the trained neural network is put to work out in the digital world using what it has learned—to recognize images, spoken words, a blood disease, or suggest the shoes someone is likely to buy next, you name it—in the streamlined form of an application. This speedier and more efficient version of a neural network infers things about new data it's presented with based on its training. In the AI lexicon this is known as 'inference.'"

As for this new chip from Alibaba, Xinhua said it had computing power

10 times of that of traditional graphic processing units. CNBC said according to claims the chip was able to cut down computing tasks usually taking one hour to five minutes. Alibaba can be pleased about its chip's benefit of enhancing computational efficiency in visual search. Xinhua said the chip could handle "more than 78,500 pictures in a second."

Companies using AI applications require huge amounts of data to train smart algorithms, and that can take several days or weeks.

So, what kinds of tasks could really use the speeding up? Well, Alibaba is using the chip internally, said CNBC. Specific business operations named were product search, automatic translation on e-commerce sites, personalized recommendations, advertising and "intelligent customer services." just the kinds of areas that require extensive computing tasks.

The company is using Hanguang 800 chips with results that show its speed advantage. *EE Times* said that "Using the device, the company's Pailitao service, where users upload pictures of items and search for matching products, had its performance efficiency increased by a factor of 12. This service handles one billion uploaded images each day, requiring an hour to process using the company's GPU-based infrastructure." Hanguang 800 infrastructure processed the same amount of images in far less time.

This is what *SyncedReview* had to say regarding the chip's design and capabilities. "The 12-nm Hanguang 800 contains 17 billion transistors. Given an inference image classification benchmark test on ResNet-50, Hanguang 800's peak performance is 78,563 images per second (IPS). Zhang says the Hanguang 800 is 15 times more powerful than the NVIDIA T4 GPU, and 46 times more powerful than the NVIDIA P4 GPU. The chip's peak efficiency is 500 IPS/W."

Business model? As mentioned, the chip will be offered in its cloud service model. "Alibaba is not directly selling its Hanguang 800 chips to customers at this stage," said *SyncedReview*; instead, developers can rent Hanguang 800 time on the AI cloud service.

*Nikkei Asian Review* heard more about this matter from Sean Yang, an analyst at Shanghai-based CINNO. "The new announcement on AI chips can be viewed not only as an effort to decouple from U.S. chipmakers, but from a business point of view, it's also a way for Alibaba to build more customized and competitive data center cloud service to compete with rivals like Amazon, Google, Tencent and Microsoft," Yang said.



Citation: Alibaba crowns its cloud service with powerful AI chip (2019, September 27) retrieved 16 April 2024 from
https://techxplore.com/news/2019-09-alibaba-crowns-cloud-powerful-ai.html