INFERENCE AT THE EDGE

AI inference processing is still in its infancy, but the market and technologies are growing at a rapid rate. By **Bethan Grylls**

Artificial intelligence (AI) chipsets for edge inference and inference training is set to grow at 65% and 137% respectively between 2018 and 2023, according to a report from market foresight firm ABI Research.

The report anticipates shipments revenues for edge AI processing, which were $1.3 billion last year, to rise, reaching $23bn by 2023.

This exponential growth has resulted in fierce competition between hyperscalers, SMEs and start-ups.

Now, with the launch of several chipsets, each are hoping that their innovations will help them to grab a sizeable share of this fast growing and changing market.

At the edge

"The hot topic is inferencing at the edge," said Andrew Grant, Senior Business Development Director, Vision & AI, Imagination Technologies, "and this is being driven by a couple of major areas, namely automotive and smart surveillance cameras."

Processing at the edge will enable companies to deliver inference without the need to transfer data – not just a costly procedure, but one that can affect latency and accuracy due to a reliance on connectivity and data transfer speeds. These are factors that can diminish user experience, and for some applications could have devastating consequences.

Geoff Tate, CEO of Flex Logix gave autonomous vehicles as one example. "In the future there will be cameras located on the car's exterior and inside too, monitoring, recognising and detecting," he said. "If you want a system that is able to identify other vehicles and pedestrians at 70mph, speed and precision of inference processing is essential."

"The more that can be done in the car, the better," agreed Grant. "If you can reduce the bandwidth around the car by running neural networks at the edge, suddenly new opportunities open up."

And it's not just automotive: running a neural network in surveillance cameras is another new application, according to Grant, that will enable smart cameras to carry out tasks such as gaze tracking and object recognition to enhance safety.

The future car will have a range of cameras situated both inside and around the vehicle, monitoring and detecting

"The hot topic is inferencing at the edge."
Andrew Grant

So why hasn't more AI Inference edge processing been done before? The problem lies with balancing price with power and accuracy.

"Take your smartphone as an example," Tate said, "For an application that does object detection and recognition, to process one picture, you need 227 billion multiples and 227bn additions (operations). A vehicle that can detect objects and surroundings will need to process 30 of those pictures per second, that's now trillions of multiples and accumulates per second.

"The challenge is that customers want an inference chip that can accomplish this – but only costs $20 and burns a few watts."

To reduce some of this computing consumption and cost, chip designers strip some of the training chip's features from the inference chip. The training typically takes place in floating point, such as FP32. Once this is finished, the model is effectively "frozen" and exported into a smaller format such as int8 onto a new chip.

A lot of the original circuitry can be removed, and multiple layers of

the model can be fused into a single computational step to reduce cost and increase speed even more.

To further improve the chip's performance, a new type of processing unit is now being used to run inference models.

CPUs were the original go-to, because of their versatility. However, they're very slow.

As a consequence, companies are increasingly turning to GPU and FPGAs, which have brought huge improvements in terms of performance per watt. Now, Flex Logix says it is taking this one step further.

"Most inference processors use traditional architectures, but these do not efficiently deliver data quick enough to the multiples and accumulator units in time," Tate explained.

"These systems will rely on buses and those have contention as multiple cores fight for access to the same memory."

Flex Logix's InferX X1 leverages the company's interconnect technology from its embedded FPGA and is combined with inference-optimised nnMAX clusters. As a result, the chip is capable of delivering higher throughput in edge applications than existing solutions, and does so with a single DRAM.

According to Flex Logix, the chip has been designed in such a way that even at small batch sizes it is close to data centre inference boards and optimised for large models which need 100s of billions of operations per image. For example, for YOLOv3 real time object recognition, the InferX X1 processes 11.4 frames/second of 2 megapixel images at batch size = 1. Performance is linear with image size: so the frame rate is 22.8 frames/second for 1 megapixel images at batch=1.

Imagination has also taken steps to improve inference chip performance, by developing IP for a neural network accelerator (NNA).

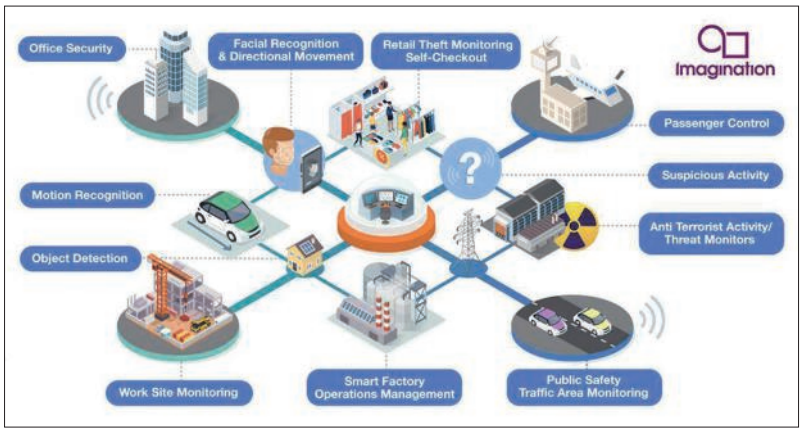
This is a technology which the company regards as a "fundamental class of processor" and "as significant as CPUs and GPUs".

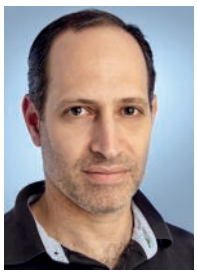
The PowerVR Series 3NX benefits from a 40% performance boost over the previous generation, performing up to 10 tera operations per second (TOPS) from a single core. According to Grant, it currently is able to offer one of the highest performance densities per mm2 in the market.

To further enhance performance Grant explains that it also features, "multicore instances for the 2, 4, 8 and 16 cores. The 16 cores are 16x 10 TOPS, that is roughly 160 TOPS and those are the ones we're finding have the most traction with those using SoC chipsets for automotive."

He continued, "We are working on a GPU where we strip out a lot of the graphics activity and processing and twin this with our neural network accelerator. This means it can run all manners of layers."

Qualcomm is taking a slightly different approach, with its recent announcement of an AI edge cloud inference solution.





"The pace of innovation and change in the inference market is staggering." Yair Siegel

Qualcomm intends to release its Cloud AI 100 - a custom built AI accelerator which it promises will deliver market-leading performance per watt

"From 2018 to 2025, the market for AI inferencing in data centres will grow massively," said Keith Kressin, senior vice president, product management, Qualcomm, "and we see an opportunity to take our low signal processing and communications expertise and create a custom-built AI accelerator that has better performance per watt than what's currently available on the market."

He envisions a scenario where instead of talking to the cloud with 100s of milliseconds of latency, the device will be talking to the cloud edge via 5G.

"This will revolutionise the end-to-end experience," he stated.

The company also intends to support developers with a full stack of tools and frameworks for each of its cloud-to-edge AI solutions, which it believes will extend user experiences including personal assistants for natural language processing and translations, and advanced image search.

"This is a new realm of chip development for us," he continued. "But we think the market is big enough and we have the right pieces in terms of process node leadership, scale and ability to be a major player in the cloud."

"The market for inferencing is tied to the neural networks and the pace of innovation and change is staggering," added Yair Siegel, Director of Segment Marketing, CEVA. "New features and technologies emerge every couple of months. "There is quick growth in computation needs over time, due to more complex networks being developed. For this reason, while very efficient processors are required, they must have enough flexibility and programmability to cope with future and unknown new features in the networks. With frequent updates on the networks onto the device, an automatic migration from training to product is required."