Getting smarter

FPGAs are set to play a central role in machine learning and AI applications. By Graham Pitcher.

The latest ‘buzz phrases’ to emerge into general use are ‘machine learning’ and artificial intelligence, or AI. Both techniques have wide application, so it’s no surprise to find technology companies of all types pursuing opportunities in this space.

Intel is no exception, addressing the topics through acquisitions and internal development in a move to create ‘end to end’ solutions. Bill Jenkins, senior AI product specialist with Intel’s Programmable Systems Group (PSG, formerly Altera) said: “Intel has formed an AI group, while PSG has a focus on machine learning, so there’s an AI effort across the company; it’s a big topic.”

Jenkins said it’s all about extracting patterns from data and then deciding what to do with the information. “It’s data analytics and perceptual issues; for example, do I understand what something means? How is a word perceived? How is an object used? That’s AI.”

Whilst both terms date back to the 1950s, it is only recently that the level of computing power available has allowed researchers and developers to achieve significant results. “People have been working in this field for some time,” said Jenkins. “But in the 1990s, we didn’t have enough compute power to solve the problems.”

Machine learning (ML) focuses on the ability of a computer to learn without being programmed to do so explicitly, while AI research looks at the development of what might be called ‘human like’ behaviour.

Amongst the targets for ML/AI are: data centres, autonomous vehicles and industrial systems, as well as the analysis of massive data in order to solve unpredictable problems.

Jenkins gave examples of the challenges being faced. “A car is now generating something like 4Tbyte of data a day, while a factory might generate an exabyte of data a day. In the US,” he continued, “94% of data collected is discarded, so the question is how can that data be used?” The answer, Jenkins contends, is to be smarter.

And the hardware and software needed for us to be smarter is available. “For reasonable power and cost,” he said, “we can now solve problems that couldn’t be addressed before. We have the ability to scale up and scale down – it’s a big deal.”

Part of that hardware solution is the FPGA – and one reason why Intel acquired Altera for $16.7billion.

“We’re different,” Jenkins observed in a reference to FPGAs.

“[FPGAs] can be used to create power efficient solutions to complicated analytics problems, with only the hardware needed for the task”

Bill Jenkins

“When you write software, it’s for a fixed architecture. In doing so, you write code in a certain way and people get good at optimising code for a given architecture.

“With FPGAs, you create an architecture for the problem; you control the data path. Rather than having data move through a CPU, then offloaded to memory, it can come right into the FPGA from wherever. It’s then processed inline with the lowest latency and in a deterministic fashion.”

And he contends that FPGAs fit well with AI/ML, offering workload agnostic computing, sensor fusion and unified tools and workflows.

“They can be used to create power efficient solutions to complicated analytics problems,” he said, “with only the hardware needed for the task. That’s the difference; FPGAs are all about system performance and latency.”

While FPGAs have been around for three decades, it’s the latest...
versions of the devices which are picking up the AI/ML load. “Everything we’ve done so far in this field has been done using the Arria 10,” Jenkins said. “That’s because its floating point capability can maintain accuracy. But you can use lower precision because the user doesn’t care whether the device is running at 5bit or 6bit; what they want is accuracy.”

Arria 10 devices offer compute performance of up to 1.5TFLOPs, while Stratix 10 parts handle 10TFLOPs. Both have arbitrary precision data types and what Jenkins says is ‘orders of magnitude’ higher on chip bandwidth.

“FPGAs have a huge amount of internal memory bandwidth,” he said, “something like 8Byte/s, along with Megabytes of on board memory. So you can get data in very quickly, do the analytics and pass the results back with minimal latency. It’s also an order of magnitude better than a CPU; where a CPU needs to run at GHz clock rates, FPGAs can run at 300MHz and sustain throughput. Meanwhile, running the device more slowly and keeping the results on the FPGA saves power.”

Of course, it’s possible to perform ML/AI tasks using an ASIC. “But your problem is that, even if you wanted to build an ASIC, you can’t make changes once it’s done. It’s also expensive to build an ASIC; while the value proposition changes with time, the fundamental need for FPGAs hasn’t.”

Skills development

If you want to get involved with ML/AI, what are the issues? “How do you expand the user base? Part of what we’re doing with our development work,” Jenkins noted, “is not forcing people to become an expert in GPUs or whatever. We want to provide tools that work in the environments in which people are happy, along with libraries and software blocks. The result can then be abstracted

alongside its Xeon server processors. “People are looking at ways to accelerate processing,” Jenkins said, “and so systems are being built which feature accelerators. When you have a CPU, you can run anything, but running it faster means either more processors or accelerators. FPGAs are easier to scale, easy to connect.

“One of the things we’re doing is to offload the machine learning element from Xeon and push it to FPGAs. If there isn’t a primitive in the FPGA, you can use the cache coherent bus to push data back to the CPU, process it and send it back to the FPGA. It’s seamless integration.”

And the concept of FPGA as a Service is emerging. Cloud Service Providers are offering their customers the ability to use FPGAs as accelerators. Two Intel customers are offering this service. In China, Alibaba Cloud offers cloud-based workload acceleration as an alternative to investing in FPGA infrastructure, while in Europe a partnership with Accelize allows enables OVH users to operate pre-built FPGA accelerators, customise them or create new ones.

“I’ve been in FPGAs since the ‘90s,” Jenkins concluded. I’ve always said that if you tell me what an FPGA doesn’t do well, I’ll tell you what’s in the next generation product. And I say do it in an FPGA, unless you can’t.”

Figure 2: FPGAs are fine grained and offer low latency between compute and memory

Figure 3: FPGAs can offer multiple use models for machine learning