INTERVIEW WITH... GEOFF TATE, CEO, FLEXLOGIX

With the Edge AI Summit just under 2 weeks away, we're delighted to bring to you an exclusive interview with Geoff Tate, CEO, FlexLogix. FlexLogix have been highly supportive of the show as a Gold Partner, and we took some time to get Geoff's perspective on how novel compute architectures are enabling neural processing at the edge of the network.

What do you see as the drivers for moving compute close to the source of data and why are innovative new hardware solutions needed for this?

Several things are important to enable edge computing:

- *Neural inferencing throughput must be good for batch=1*; at the edge there is typically one sensor so batching doesn't make sense. Most architectures today do very poorly at batch=1 and instead are optimized for batch sizes from 10 to 100, which may be fine in data centers but not at the edge.

- *The power budget of each edge application varies but puts a ceiling allowable inferencing power*: a backup camera may be able to afford 1W, an autonomous driving ADAS 50W, but every application has a hard ceiling. So lowering TOPS/W at worst-case conditions based on throughput, not peak, expands the throughput possible at a given power budget.

- *The cost budget of each application varies:* lower the cost of the inferencing accelerator and reducing the number of DRAMs required cuts cost and expands the number of edge applications that can use the solution.

- *Scalability*: requirements range from 1 to 100 TOPS, so a point solution will address only a fraction of the market: solutions must scale across at least two orders of magnitude of throughput.

> *"The XFLX interconnect technology invented for our eFPGA would enable very high bandwidth in a reconfigurable way from local SRAM, and not that much of it would be required for existing models like YOLOv3. The advantage this gives us is 10x lower cost and 1/3 the power of existing solutions at similar throughput levels, and we get high speed even at batch=1."*

In which applications do you see the FlexLogix / NMAX solution being most effective?

NMAX is especially effective in edge applications because NMAX' fast-weight-load architecture enables high performance at batch=1: we don't need batching to get high throughput. For ResNet-50 our MAC utilization is 85-90% and for YOLOv3 it is 60-80%, far higher than existing solutions achieve even when they do batching of 10 to 100.

How did you arrive at the position that localized SRAM was the optimal memory solution and what advantage does that give you over other chips?

Cheng Wang, my co-founder, experimented with architectural variations and got customer feedback on priorities. What he heard is that existing solutions need 100's of Gigabytes/second of DRAM bandwidth resulting in high DRAM cost, high DRAM power, and 1000+ pin packages for all the DRAM interface pins. From this he realized that the XFLX interconnect technology he invented for our eFPGA would enable very high bandwidth in a reconfigurable way from local SRAM, and not that much of it would be required for existing models like YOLOv3. The advantage this gives us is 10x lower cost and 1/3 the power of existing solutions at similar throughput levels, and we get high speed even at batch=1.

What developments would you expect to see over the next 12-18 months for AI Hardware at the Edge?

There is actually not much real AI processing done at the edge today: the bulk is key-word recognition and the rest is actually done in data centers. With NMAX and other inferencing chips coming to the market we'll see performance-inferencing take off in edge applications (performance = 1TOPS+).

Hear from the FlexLogix team at the Edge AI Summit in San Francisco on December 11, 2018.

Register online at edgeaisummit.com