



Photo by Michael Dziedzic on Unsplash

Scoping the AI Chip Landscape - Summarizing Linley Fall 2020 Conference

Published on November 7, 2020



Hamid Reza Zohouri, PhD
Director of Product at Edgecortex

1 article

+ Follow

I had the opportunity of attending the [Linley Fall Processor conference](#) last month where many AI processors were discussed or even introduced for the first time. In this blog, I will give a summary of the presentations in the conference, with specific focus on the ones involving chips for accelerating AI workloads; some presentations not specifically targeted for AI acceleration might be missing from the blog.

Session 1: AI in Edge Devices (Part I)

Flex Logix

Flex Logix introduced their InferX X1 chip for AI inference on the edge. The chip comprises 4 MB of global scratchpad memory, a reconfigurable Tensor Processor (reconfigurability at interconnect level), 1 bank of LPDDR4, and an eFPGA block. The chip die size is 54 mm² and is manufactured at 16 nm technology with a 7 to 13W TDP.

The Reconfigurable Tensor Processor Engine is composed of 16x 1D Tensor Processing Units (DTPU) with each DTPU being capable of performing 64x INT8 or 32x BF16 MAC

operations per cycle for a total of 4096x INT8 or 2048x BF16 operations per cycle. Moreover, this engine has 8 MB of distributed SRAM and its interconnect can be reconfigured in 4 us; the reconfigurability feature is largely used to enable layer fusion and adapting DTPUs to the configuration of each specific layer in a network (i.e. one reconfiguration per layer group).

In terms of performance, the slowest variation of the chip reportedly can run YOLOv3 with input size of 416x416, INT8 precision and a batch size of one at 63 fps, while using 8.5W of power. The chip will be available in multiple frequency variations with ranging from 533 MHz to 933 MHz.

Imagination

Imagination introduced their AI inference IP largely targeted at self-driving cars. They follow a multi-core approach with the IP being configurable up to 8 cores, each core capable of 4096 INT8 MAC/cycle (or 1/4th of that at 16 bits), and 64KB to 256 MB of SRAM. They expect that the IP can run at 1.5 GHz if manufactured using the 7nm process, providing a peak compute performance of 1000 TOP/s.

Their approach to parallelization is "tensor tiling" where each tile in the tensor will be scheduled on a different core, with possibility of splitting the computation over different cores also by splitting the channels for cases where the tensor is small. Their scheduler will handle work splitting and distribution across all the available cores, while each core or a group of cores can be reserved for a particular workload to control latency and throughput trade-offs.

They also briefly touched upon their software stack built on top of the TVM framework which allows accepting Tensorflow, ONNX, etc. as input and covers CNN, LSTM and other types of networks. Quantizer, compiler/scheduler, and performance estimation tools are also provided.

It does not seem like Imagination has any plans of creating any chips based on this IP, and they did not provide any performance or power numbers either.

ThinkSilicon

ThinkSilicon introduced their NEOX SOC with both a RISC-V-based CPU and GPU targeted at AI, GPGPU compute, and graphics rendering workloads. Their design is scalable from 4 to 64 cores to cover all the way from embedded to high-end markets. The existing RISC-V tooling and ecosystem (GCC/LLVM) can be used to program the SOC.

Among interesting features of their SOC, a value cache is employed to skip recomputation if the inputs to a specific function are repeated, with possibility of extending this to cases where the inputs are *approximately* the same as a previous computation, if the application can tolerate approximate computing. Moreover, user-defined instructions can be added to the GPU to enable product differentiation.

The SOC supports FP32|FP16 and INT8|4|2 data types, and ALU utilization in the SOC is maximized by taking full advantage of multi-threading. Moreover, since the CPU and the

GPU use the same ISA, threads can be seamlessly migrated between them. The available GPU configuration choices for the SOC are 4, 16 and 64 cores with each core providing up to 3.2|6.4|12.8|25.6 GOP/s of FP32|FP16|INT8|INT4 compute at 800 MHz, for a maximum of 204.8|409.6|819.2|1638.4 GOP/s for the 64-core configuration. In general, considering the multi-core architecture of the SOC and focus on multi-threading, it is likely the case that this SOC would be more successful in accelerating batched inference rather than batch-1. No performance number from real-world workloads or power efficiency numbers were reported.

Finally, a software SDK is provided to perform conversion from different model formats to TensorFlow Lite format which can be consumed by their graph analyzer, and enable quantization to different formats supported by the SOC.

Global Foundries

Global Foundries talked about their manufacturing innovations that enable next-generation AI processors. No specific chips were introduced here.

BrainChip

BrainChip presented their Akida Neuromorphic processor which uses an event-based approach to process neural networks. Due to the event-based nature of the computation, the chip inherently supports both weight and activation sparsity. On top of that, aggressive quantization of 4 bits and below is employed to minimize the on-chip memory usage of any given network. Activity regularization is also employed during training to increase activation sparsity and enable better compression. The chip also supports on-chip learning to allow incremental learning on the edge.

The obvious drawbacks of the aggressive quantization used in this chip are the need for quantization-aware training (while post-training quantization is typically sufficient for 8-bit quantization), and loss of inference accuracy which can reach 2-3% depending on the model.

The chip is manufactured in the form of an SOC that also includes an ARM M-class CPU, LPDDR4, PCI-E, and converters for converting between network data and events that are fed to and received from the Neuromorphic fabric. The Neuromorphic fabric itself is all-digital logic comprising 80 NPUs and 8 MB of SRAM, manufactured using TSMC 28 nm node for the first implementation. The evaluation is expected to run at 300 MHz.

Multiple application examples for the chip were also presented in the conference, but not much was discussed regarding the software stack/SDK accompanying the chip.

Session 2: Vector-Processing Cores

SiFive

SiFive introduced their RISC-V Vector chip which was claimed to eliminate the need for compute capabilities on GPUs used in embedded SOCs (but GPUs will still be required for graphics). Each vector instruction in this case can operate on 256 bits of data per cycle, composed of 8 to 64-bit data types; floating-point, fixed-point and integer data types are supported. They reported that their IP running on a Xilinx VCU118 FPGA board can beat Cortex-A53 in Mobilenet V1 by 4.3x, and their actual chip is expected to achieve a latency

of 35.21 ms @1.8 GHz for the same network.

Andes

Andes talked about their NX27V vector processor, calling it the first commercial RISC-V processor, with support for 512-bit vector operations on 4 to 64-bit integers, 16 to 64 bit floats, and also bfloat16. On the topic of performance, they reported sustained and peak performance of 96 and 320 GOP/s per core, respectively, for 16-bit data types. For manufacturing, they are targeting the 7 nm process node with a worst-case frequency of 1 GHz and an area of 0.3 mm².

Session 3: Advancing Cloud AI

Groq

Groq mostly talked about the multi-board and multi-chassis scalability of their platform and how, aside from AI workloads which are their main focus, they are now starting to also target linear algebra workloads in HPC/data center applications. Albeit, their performance comparison with the 2-generation-old Nvidia P100 raises the question as to whether they can also take on Nvidia's newer, much more efficient V100 and A100 GPUs or not. One interesting claim in their presentation was that their very wide dot product without any normalization in-between can actually help with network accuracy in AI workloads, even though one might expect the higher stochasticity of having the normalization after every MAC operation to be more beneficial instead. And indeed we are still going to have to wait to see Groq's MLPerf numbers since they missed the deadline for 0.7, too.

Tenstorrent

Tenstorrent talked about their compiler stack which has to enable all forms of parallelism (pipeline, data and model-level) to maximize the utilization of the 120 cores they have on their chip. Data is moved across their chip by splitting them into packets that are routed across the chip based on the header attached to each packet. Their software stack natively supports Pytorch and they are in the process of adding support for other frameworks through ONNX. They believe adding support for conditional processing (i.e. taking advantage of sparsity) is the most important factor in the path to performance for AI workloads; this is something that they are working on right now and matches well with their packet-based data movement. Similar to Groq, they also missed the MLPerf 0.7 submission and they didn't report any new performance results other than their existing NLP results, even though they claimed to have many vision networks running already. They have, however, delivered evaluation boards to customers and opened up their internal development cloud for evaluation by others.

Intel

Intel talked about how hard it is to choose a burger from the restaurant menu, especially during the current lock-downs, and how Intel is helping with that...

Session 4: Automotive Processor Design

Hailo

Hailo spent a lot of time explaining why TOP/s is a bad comparison metric (I hope this becomes common knowledge one day), and then they gave the TOP/s of their own chip... Their Hailo-8 chip is claimed to achieve 26 TOP/s at 3 TOP/s/Watt, and does *not* have any DRAM. In their defense, they also presented two slides at the very end with some batch-1 performance numbers, including ~900 FPS for Resnet-50, ~2600 FPS for Mobilenet V2, and ~1000 FPS for Mobilenet V1 SSD; these numbers reportedly outperform Nvidia's Jetson family offerings. One interesting aspect of their chip is that it is (going to be?) certified for ASIL-B which is crucial in the automotive industry that is their main target market.

ARM

Arm introduced Cortex-A78AE and outlined its functional safety features for ASIL-B and ASIL-D.

Synopsys

Synopsys talked about their automotive chips/SOCs and emphasized that for automotive solutions, **security (protection against malicious attacks) is also required on top of functional safety**, and they have already integrated both at hardware and software level.

Session 5: Security

Cornami

Cornami talked about privacy-preserving Machine Learning and how not only Homomorphic encryption is the solution to this problem, but also claimed that it "is going to make the world a better place". Their software-reconfigurable chip is composed of hundreds of thousands of what they call FracTLCores, and each FracTLCores itself is composed of a RISC-V core and some amount of SRAM, with an NOC enabling communication between the cores. The reconfigurability of their fabric allows it to implement different architectures such as systolic arrays, soft GPUs, RISC-V processors, etc.

Their presentation didn't make it clear what architectural advantage their chip has for Homomorphic encryption, while they made very bold claims such as being able to scale Resnet-50 batch-1 inference performance linearly across any number of cores that are necessary (which was supposedly achieved for up to 131,072 in *simulation* but is simply untrue since it goes against Amdahl's law and AI inference does not have unlimited parallelism regardless of the network, either), to the point that they achieve "the fastest published Resnet-50 numbers of any silicon" despite not having taped out any silicon yet. They also outright claimed that "Every ML Startup out there has built the "wrong" silicon" since they did not consider privacy preservation in their chip design, which is again a claim that would be extremely difficult to prove.

Session 6: The New Infrastructure Edge

Centaur

Centaur talked about their server-class x86 SOC with integrated AI processor and AVX512 support. The AI processor is capable of 20 TOP/s peak at 2.5 GHz operating frequency, supports 4096-byte wide SIMD, has 16MB of SRAM, and supports INT8, UINT8, INT16,

and BFloat16 formats. They have MLPerf results with Resnet-50 single-stream (batch-1) latency of 1.05 ms, which equals ~950 FPS. Their compiler stack is based on MLIR and is capable of optimizing host and kernel code together. They did not mention the power usage of the SOC.

Intel

Intel talked about their new Stratix 10 NX FPGA family. Surprisingly, Stratix 10 NX does not use the same DSP as Agilex and its DSP is designed towards enabling higher math density for AI workloads, at the cost of losing some flexibility for more general (i.e. non-AI) applications. The peak compute performance of the biggest variation of the FPGA is 20 FP32 TFLOP/s, 30 TOP/s INT15/16, 128 TOP/s INT8 or 250 TOP/s INT4. It is unclear what operating frequency has been used to calculate these numbers.

One major issue with the Stratix 10 FPGA family (not limited to NX variation) has been that even though Intel initially touted upwards of 800 MHz as operating frequency, in practice, even with HyperFlex, it is quite difficult to achieve even half of that number. The presenter did report that they have managed to timing close a single-array design with 90% DSP usage at 400 MHz, and a systolic-array design with the same DSP usage at 520 MHz. Power usage in these configurations was claimed to be ~120 watts.

Deep AI

The highlight of this session was probably Deep AI (**This** one and not the 10+ other Deep AIs around). They claimed to have solved the problem of training at *8-bit precision*, which is generally considered not to be enough for most models to converge. Their training solution uses *8-bit fixed-point* and also takes advantage of sparsity during training, rather than performing pruning or drop-out after training. Moreover, they claimed that their solution can dynamically switch between 8-bit training and 8-bit inference without needing any calibration or other operations in-between, making it a perfect solution for training on the edge and also federated/continual learning.

Their solution is FPGA-based (since making ASIC is expensive) and using one Xilinx Alveo U50 (at 8-bit), they claimed to be able to train networks at the same speed as an 8x Nvidia T4 server or a 4x V100 server (32-bit) at much lower power usage and price. Albeit, this doesn't mean their solution can still keep up with the GPU servers if the same 8-bit training is also used on the GPUs, and they made it very clear that they don't want to share any details of the inner workings of their training solution nor do they want to make it available on any other platforms, so we will probably never know.

The Linley report on their solution mentions that their FPGA design is not general and they have a pre-synthesized overlay for each supported network, and apparently they only support Resnet and Yolo right now. This probably means that their solution cannot be readily applied to, or even work for, other models (e.g. Mobilenet-type models which are more sensitive to data precision).

Nvidia

Nvidia first talked about blowing all the competition out of the water in the recent MLPerf v0.7 results using their various solutions targeted for different market segments, and then

introduced their EGX Edge AI platform with a Data Processing Unit (DPU) that has AI capabilities, largely aimed at processing network data on the edge for traffic management and security.

Session 7: AI in Edge Devices (Part II)

Flex Logix

Even though Flex Logix seems to have a working hardware already (discussed in Session 1), they do not seem to have much of a compiler or software stack available yet. They emphasized a lot on the price and performance of their hardware in this session, but apparently they are still trying to get Yolo v3 working at compiler level on their hardware; it seems they followed a hardware-first approach for their solution. Moreover, they mentioned that they consider MLPerf a "politicized" benchmark suite that favors certain manufacturers (Nvidia?) and prefer to work on models their customers want rather than ones used in MLPerf which they claim are of little to no use to their customers. They have a single-chip PCI-E x4 board available right now with a 4-chip PCI-E x8 version and a single-chip M.2 board in the works.

Sima.ai

In complete contrast to Flex Logix, Sima.ai clearly followed a software-first approach and even though they still don't have a working hardware, they have their full software stack working already with the whole compiler flow built on top of TVM, support for every major Deep Learning framework, support for pre-quantized networks along with a proprietary quantization method more suitable for their hardware, complete visualization stack based on Netron, and simulator and [cycle-accurate] emulator fully working.

During their break-out session they demoed their full flow starting from the network down to running on the simulator which can estimate run time and power usage, with what they claimed to be "very good accuracy", in a matter of minutes. They are already shipping their SDK to customers so that they can evaluate their models using the simulator. They claimed their simulator also very accurately simulates DDR memory latency and power. At the same time, they absolutely refused to provide any performance numbers and just shared power efficiency (TOP/s/Watt) which apparently even exceeds 10 TOP/s/Watt in certain cases. I managed to get a glimpse of their *Resnet-50* performance on the simulator during breakout which was *1.08 ms*. Also two quick points about their chip which were revealed during Q&A: it is manufactured using the 16nm node and they use Synopsys Vision IPs on their SOC.

Cadence

Cadence talked about their Tensilica processor/IP suite which is aimed at voice and vision processing for edge devices. They emphasized that these types of processing are not just limited to the CNN processing itself and the pre and post-processing can also take up a lot of cycles and they have an IP/processor for every typical type of processing involved in such applications. Their performance numbers were normalized against an unknown competition (Synopsys?), so it is not possible to extract much useful information out of that. They did report 691 FPS/Watt for MobileNet v1, though.

Session 8: Heterogeneous Computing

Fungible

Fungible talked about their Data Processing solutions for data center workloads such as compression/decompression, encryption, firewall, streaming, and storage acceleration.

Achrnoix

Achrnoix gave an interesting talk about their eFPGA solution. This solution provides customers with the possibility of embedding a custom FPGA (eFPGA) with whatever size and resources that they want into their ASIC. This can provide multiple benefits:

- Since the eFPGA can be configured just like a normal FPGA, it allows offloading certain functionality of the ASIC to the eFPGA to reduce ASIC design time and allow faster time-to-market.
- Having an eFPGA on the chip allows the solution to be valid for a longer time since new functionality/operators not supported by the ASIC can be added in the eFPGA.
- An eFPGA will cost less than a stand-alone FPGA of the same size and use less power since it does not have any external I/O

Achronix's eFPGA IP is apparently compatible with Automotive safety standards already.

Marvell

Marvell also talked about their Data Processing solution composed of a custom 36-core ARM processor, a Data Processing accelerator and lots of I/O, targeted at packet processing, crypto operations, firewall, etc. in data centers.

Session 9: SOC Design

SiFive

SiFive was very proud and enthusiastic to introduce their first PC-grade RISC-V solution that can run a few Linux distributions smoothly and can do basic audio and video playback (they demoed Youtube playback), with support for some limited number of GPUs. Even though this is considered as a major milestone for the RISC-V community, I got the feeling that software-wise, RISC-V is still quite far from becoming a replacement for ARM which is something many have recently started considering more seriously since Nvidia is going to acquire ARM.

ARM

ARM officially announced Cortex X1 and Cortex A78, the former having been designed for absolute performance at the cost of sacrificing power efficiency to some extent, and the latter having been designed to be an incremental upgrade over A77 in terms of both performance and power efficiency. A78 and X1 are ~7% and 30% faster than A77 in the Spec benchmark suite, respectively, while A78 is the most power efficient one among the three.

Arteris

Arteris talked about their NOC IP with support for mixed coherency. They mentioned that many big and small names, including multiple of the companies mentioned in this blog, are using their NOC solution.

Session 10: In-Memory Compute

GSI

GSI talked about their *in-memory* compute chip which packs basic bit operations (AND, OR, XOR, NOT) in-between memory cells for fast "bit processing". Their chip can also perform "word processing" by treating physically consecutive memory cells as vectors. The chip comprises 2 Million bit processors with 96 Mb (= 12 MB) of L1 cache scattered throughout the chip. The chip itself is composed of 4 cores, each of which having its own scheduler, instruction memory, etc. The chip is claimed to support up to 840 trillion bit operations @400 MHz, 25 TOP/s of INT8 ADD performance, and 1 trillion queries per second of TCAM performance in a 60 watt envelope and manufactured as a PCI-E card. The important point to note is that they did not make any claims about the performance of the chip for AI workloads and it doesn't seem to be directly targeted for AI either; their main target market is accelerating applications that rely on search queries and boolean operations (e.g. network firewalls).

Untether AI

Untether AI officially came out of stealth mode during this session and introduced their chip specifically targeted for AI applications (not just CNNs). Their chip performs what they



Home



My Network



Jobs



Messaging



Notifications



Me



Work

version of Cerebras's chip with an island-style design, 200 MB of SRAM organized as 511 banks, 502 TOP/s of INT8 performance, PCI-E Gen 4 x16, two configurable operating frequencies of 720 and 960 MHz, and a maximum power efficiency of 8 TOP/s/Watt which allows the whole board to fit in the 75 Watt power envelope of PCI-E without needing auxiliary power. Their NOC is composed of a set of row-wise and column-wise routes connecting neighboring cores/PEs directly, alongside with a row-based ring to access PCI-E. Each PE has a MAC unit, some registers and also a zero detector that gates the MAC unit to save power. Multiple instances of a custom RISC (not RISC-V) processor are used as controllers to manage PE communication, and also certain operations such as state machines are offloaded to it. *There is no DRAM in their design.*

Their 4-chip PCI-E x16 board provides up to 2,000 TOP/s of peak INT8 compute performance at 250-300 watts (500 TOP/s at less than 75 watts per chip). They claimed 80,000 FPS for ResNet-50 and 12,000 query/s for Bert-base on this chip, which theoretically beats all the existing competition. Accordingly, they claimed 638 FPS/Watt for ResNet-50 and 96 query/s/Watt for Bert. However, none of these numbers are actually measured and the performance numbers seem to be simply the total operations required by the network divided by the peak performance of the device and then multiplied by device utilization which they can get from their scheduler. The power efficiency numbers are also calculated by dividing the aforementioned performance numbers by a fixed 125 watts.


Their software SDK includes automated quantization, post-quantization retraining, visualization that shows chip resource allocation, utilization, etc., compiler that performs optimizations such as layer fusion, and allocator that is capable of partitioning one model across multiple chips or even multiple boards while minimizing chip-to-chip communication, or even sharing one chip between multiple models.

Untether AI in particular was very confident that they can easily scale batch-1 inference performance across multiple chips and multiple boards. I joined their breakout session where they demoed some of their software stack and also their placer/allocator and its visual output. What became very clear there was that their solution, like many others, will have utilization issues when it comes to batch-1 inference. e.g. for Resnet-50, only something like 70-75% of the area of one chip was utilized, while some 10% of the utilized blocks were just used for routing and not actual computation. The Linley report that came out a few days later mentioned that Resnet-50 utilization on their 4-chip board is only 15% (which is roughly equal to 70% of one chip); hence, it is not clear how they are going to scale the performance of this network, or similarly-sized ones, over 4 chips (or multiple boards) to achieve the claimed 80,000 FPS batch-1 inference speed, when the network cannot even fully utilize one chip.

Ambient Scientific

Ambient Scientific introduced their GX-10 AI processor for Micro Edge (i.e. sub-mW power usage). Their SOC consists of an always-on component using 80 μ W for forward

 Like  Comment  Share

   42 · 7 comments

 Messaging

training and inference with the possibility of dynamically switching between them, all power of two bitwidths between 4 and 32 bits and also any mix of them, a peak performance of 512 GOP/s (probably at INT4) and 4.3 TOP/s @ TSMC 40 nm node. Their SOC includes multiple components such as a custom RISC processor (again not RISC-V) for 2D vector instructions, AI cores comprised of analog MAC cores using digital CMOS (quite confusing description) scattered in-between SRAM blocks in a *near-memory* fashion, a 3D SRAM architecture (*not* 3D-stacked), ALU and activation units, and so on.

Their MAC unit gets digital data as input and gives digital data as output, while ADCs and DACs are used to convert between digital and analog signals, and the computation is done using CMOS rather than RLC components. When pressed as to how they made 32-bit ADC/DAC possible, the presenter made it very clear that this is a question everyone asks him and he doesn't want to answer it since it could reveal their chip's secret sauce; he did allude that people should stop thinking about ADCs and DACs and look at their design as a monolithic design (again, confusing description); was he trying to say that there are actually no ADCs and DACs in the design and they just have other components that replicate ADC/DAC functionality?

This presenter in particular was very bold and confident in his company's chip design and its innovative architectural features and even challenged the other presenters in the same session over who has created a better in-memory compute chip. I guess we will have to wait until the chips are out and tested by independent reviews to decide who has done a better job here.

This concludes my blog post on the Linley Fall Processor conference. I am looking forward to the next conference held in spring 2021.

Disclaimer: This article was prepared by Dr. Zohouri in his personal capacity. The opinions expressed in this article are the author's own and do not necessarily reflect the view of EdgeCortex Inc.

Report this

Published by



Hamid Reza Zohouri, PhD

Director of Product at Edgecortex

Published · 6mo

1 article

+ Follow

I had the opportunity of attending the Linley Fall Processor conference last month and I have summarized the sessions involving AI chips in the blog post that follows:

[#artificialintelligence](#)
[#edgecomputing](#)
[#edgeai](#)
[#semiconductor](#)
[#highperformancecomputing](#)
[#fpgas](#)
[#ai](#)
[#datacenter](#)
[#cloudcomputing](#)

Reactions



7 Comments

Most relevant ▾



Add a comment...



Andrei V. · 2nd

6mo (edited) ...

CEO at Scientific Concepts International

Great review of the current state of the art. However, I am puzzled for quite some time. Why smaller companies with smaller budgets can claim better performance, novel compilers and microarchitectures while bigger companies are not?

Like · 🗨️ 1 | Reply · 2 Replies

Hamid Reza Zohouri, PhD · 2nd

6mo ...

Director of Product at Edgecortex

Andrei, the reason is probably something along the lines of the joke I wrote in the article about Intel. Big companies prefer small incremental improvements which have less risk and are expected to generate a steady revenue stream from their existing markets. Start-ups, however, tend to be more open to taking risks since they are not immediately expected to generate revenue. Moreover, start-ups tend to be less concerned with corporate structure and f ...see more

Like | Reply

Andrei V. • 2nd

6mo ...

CEO at Scientific Concepts International

Thank you for teaching me about startups ;)

Like | Reply

Ali Jamali • 3rd+

6mo ...

Electrical/Electronic Manufacturing Professional

congratulations

Like | Reply

[Load more comments](#)



Hamid Reza Zohouri, PhD

Director of Product at Edgecortix

[+ Follow](#)