Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 3): Edge and Automotive

Publication Date: 21 July 2020. Product code KR303.6

Michael Azoff

Kisaco Research View

Motivation

Today Artificial intelligence (AI) is out of the research laboratory and in the realm of practical engineering applications. AI engineering today is largely about running machine learning (ML) models on digital computers, and these models are typically simulations of brain-inspired models such as neural networks, with deep learning (DL) being the most successful example today. With the plateauing out of CPU performance improvements and the end of Moore's law, even with multi-core CPU machines, the community has turned to hardware accelerators to run their AI models.

While the cloud has become the marker for our current age of computing, the edge is set to take over the limelight and the most straight forward reason is that it is where most of the data is generated and we are moving to technology that can process it at source rather than create lag and throughput bottlenecks in shifting it first to the cloud.

The AI hardware accelerators needed for edge computing and for the automotive market are in stark contrast to those needed in the data center (DC) and for high performance computing (HPC). Whereas in the DC AI models are typically trained and inferenced, on the edge AI models are typically just inferenced (training can be done at the edge but the chips we review are designed for the most common use case of edge inferencing). Size and power constraints are also significant factors in the edge which have less effect in DC/HPC choices.

The edge represents a spectrum of use cases and so we focus on small sized chips for small edge scenarios such as embedded AI in consumer products, security systems, sensors, and a host of smart devices. The automotive market, which has distinct requirements from other edge computing, also covers a spectrum of use cases in the vehicle: smart controllers, ADAS, and autonomous vehicles (AV). We focus on AI chips suitable for the AV market. Thus, this report contains a KLC for each of these market segments: small edge and automotive-AV.

Definitions are important

AI

Not everyone talking about AI has the same idea in their head, so clarity of definitions is essential in any conversation on AI, especially to avoid fueling hype. In our companion report, "*Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 1): Technology and Market Landscapes*, June 2020", we level set by defining what we mean by AI.

Briefly: we use the term AI broadly, as a label for the space, without prejudice to its state of progress in achieving its goals. We define the AI community's current state of progress as being in the age of "machine intelligence", short of narrow AI (by maybe a decade or two) and some long way from general AI. There are no standard definitions of narrow AI and general AI, but for us narrow AI's most defining characteristic is rapid learning from a few examples, and general AI is achieving parity with human intelligence. For more in-depth coverage of these terms see the companion report.

The small edge

What constitutes edge computing and internet of things (IOT) is an open question as the spectrum of possibilities is open to human ingenuity in bringing compute to some end/local node and connecting it online. For cloud hyperscalars the edge could be a data center in some far-flung remote location or an on-premises extension of their data center, or it could be a smart consumer wearable device. The scale between these examples is vast. The mobile phone is clearly an edge device, but the mobile industry is a mature one that has evolved a distinct technology stack with a well-regulated regime and set of standards. For this reason, most discussion of edge computing excludes the mobile phone. Another criterion used within the industry, for example Linux Foundation Edge (LF Edge), is to define it in terms of local compute and storage with a latency less than 20ms at the point where the compute is required.

So, for our purposes in this report, we start with the above LF Edge definition, exclude mobile phones, and further confine ourselves to what we term the “small” edge: local environments that may have constraints of chip size, power availability, mobility, and connectivity. Some writers refer to “end nodes” or “end devices” when referring to the limit of the edge (or even extreme edge). Given that edge computing represents a spectrum with local edge servers and edge gateways being part of the edge fabric, referring to end devices may provide further clarity. We prefer in this report small edge as the chips we review are designed for small products.

We include a KLC for in-vehicle applications separately from the small edge KLC for reasons we go into detail below, being a quite distinct industry with its unique set of requirements and only a subset of AI chip vendors play in this market.

A final word on definitions, AI for the edge is sometimes called artificial intelligence of things (AIOT), which is a nice play on IOT, and also the intelligent edge, or edge intelligence.

Key findings

- The edge computing market is set to grow on the back of convergence of adjoining technologies: AI, 5G, cloud native technology, and expansion of IOT.
- We find it useful to focus on the small edge here (i.e. small devices at the end of the edge spectrum operating within 20ms latency) as the edge spans large variations in power availability, size of components, and latency.
- Automotive industry uses for AI span from AV to ADAS and smart controllers. We focus here on AV as this represents an extreme set of requirements for AI chips. AV has gone through a hype cycle and currently there is more focus on progressing the capability than manufacturers boasting when their level 5 vehicles will be on the road.
- We find the market for AI accelerator chips splits into three extremes, with the DC/HPC, small edge, and automotive at each node of a triangle having distinct features, while the rest of the market in the triangle middle varies the most in AI chip characteristics depending on application.
- The market for edge AI inferencing is the most active segment of the AI chip market. We found more players competing in this space and the intense competition drives the unit chip costs down – the winner offers the highest performance for the lowest power consumption and at the lowest chip cost. While the margins are low the volumes can be huge.
- We identified three leading players in the automotive-AV AI chip market participating in our report. We believe this market is particularly tough for new entrants given the greater hurdles required to satisfy the auto manufacturers.

- Selecting an AI accelerator chip simply because it has the highest operations per sec processing capability may not be relevant if a model can be found that requires less processing to be performed for the same or better accuracy.
- Neural processing units designed for different application use cases will need different characteristics – there is no one accelerator to fit all jobs.
- AI training can be done at the edge but in most commercial applications, the approach is to train systems in the DC and then embed the AI applications in inference mode into products.

Companion reports

This report is part three of a three-part series of reports, the companion reports are:

- **KLC on AI Hardware Accelerators 2020-21 (part 1): Technology and Market Landscapes**: Provides a broad view of the technology and market landscape with analysis. It includes a list of all the players in the AI accelerator space, from established to startups, with a snapshot of key products.
- **KLC on AI Hardware Accelerators 2020-21 (part 2): Data centers and HPC**: Assessment of key vendors with KLC charts, one for DC/HPC AI training and one for AI inferencing. Deep profiles of participating vendors with strengths and weaknesses analysis are included.

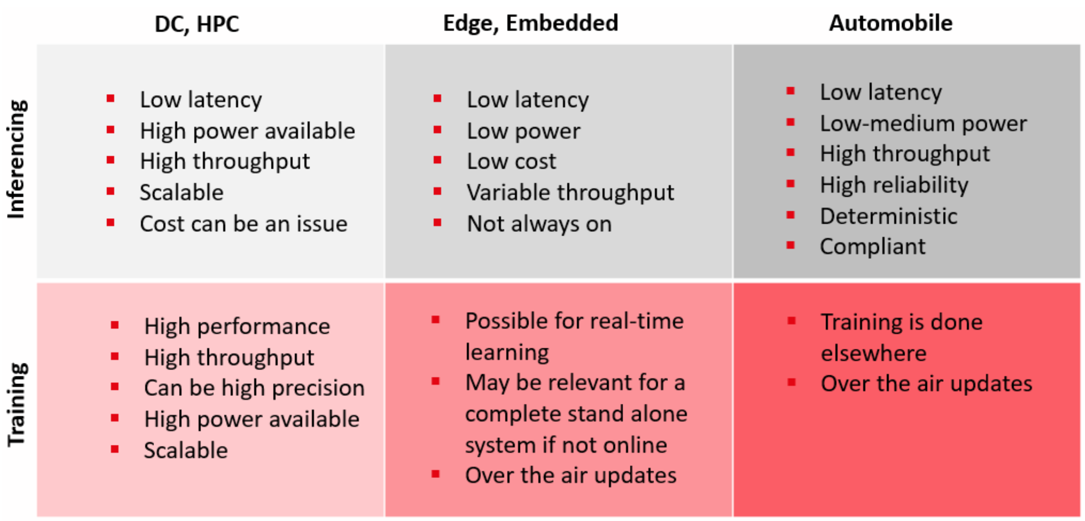
Solution analysis: AI inferencing on the edge

Technology and market trends

Market segments

Our market segments table, Figure 1, shows the characteristics for AI training and inference by the three main market segments covered in our series of reports. In this report we look at the middle and right columns and at inferencing in particular.

Figure 1: Market segments AI training and inferencing characteristics



	DC, HPC	Edge, Embedded	Automobile
Inferencing	Low latency High power available High throughput Scalable Cost can be an issue	Low latency Low power Low cost Variable throughput Not always on	Low latency Low-medium power High throughput High reliability Deterministic Compliant
Training	High performance High throughput Can be high precision High power available Scalable	Possible for real-time learning May be relevant for a complete stand alone system if not online Over the air updates	Training is done elsewhere Over the air updates

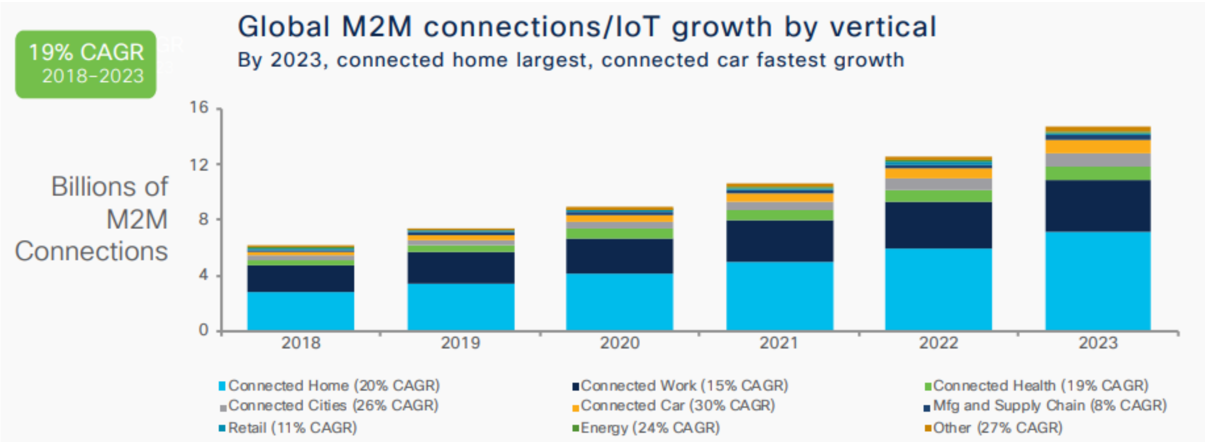
Source: Kisaco Research

AI training can be done at the edge but in most commercial applications, the approach is to train systems in the DC and then embed the AI applications in inference mode into products. There are notable differences between edge and automotive: while both edge and automotive generate big data, edge systems vary in throughput, for example static surveillance cameras may see little change over a period, or a device only one has one input sensor. In-vehicle AI systems typically process big data streaming from multiple sensors while in motion and throughput will be high. Low latency is always an advantage for inferencing, for edge and automobile AI applications that operate in real-time it can be a must have for safety-critical reasons. We discuss these constraints in more detail below.

Small edge

According to CISCO, the amount of global machine-to-machine (M2M) traffic is forecast to grow strongest for connected home applications (see Figure 2), which includes home automation, home security and video surveillance, connected white goods, and tracking applications, representing 48% of total M2M connections by 2023. Connected car applications such as fleet management, in-vehicle entertainment systems, emergency calling, Internet, vehicle diagnostics and navigation, and more, will be the fastest growing category, at a 30 percent CAGR.

Figure 2: Convergence of technologies and impact on edge applications



Source: Cisco Annual Internet Report (2018-2023) White Paper.

Edge computing (in the broadest sense, including autonomous driving) is a market predicted to outgrow cloud computing in the decade ahead, and the provision of AI chips to the edge is set to explode. The small edge represents the largest volume of objects connected at the edge and spans multiple types of devices: home appliances, sensors, cameras, and a host of smart products such as smart speakers and smart earbuds.

There are five crucial trends taking place right now that inter-connect each other and is leading to this edge computing revolution that is part of the broader fourth industrial revolution and digital transformation taking place. These trends are:

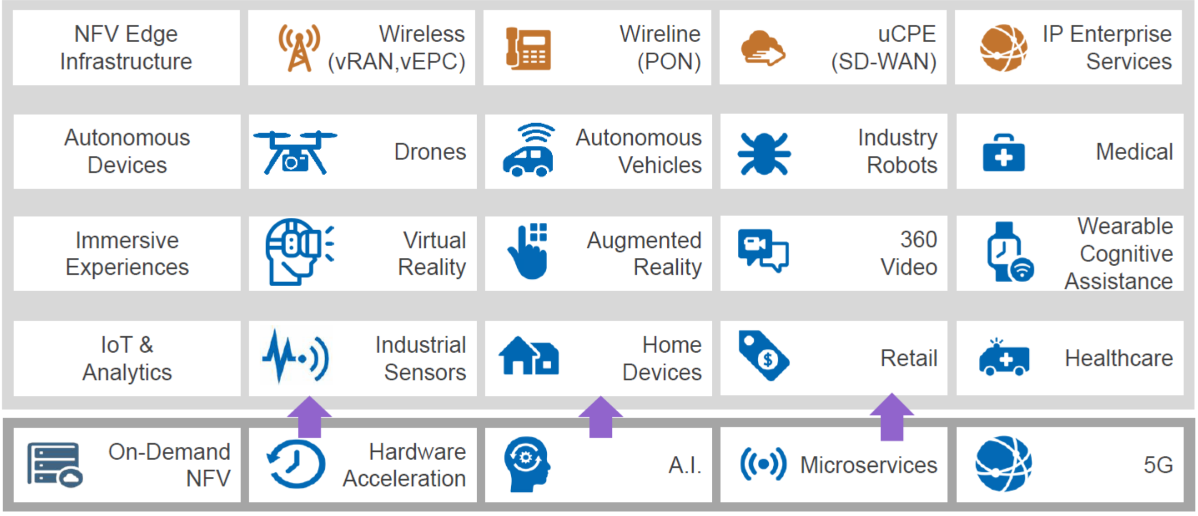
- 5G telecommunications: how next generation communications makes connectivity continuous and ubiquitous.
- AI: how intelligent processing can be performed.
- IOT: how all objects are connected by the internet.
- Edge computing: how processing takes place locally and can generate big data.

- Cloud native technology (DevOps, microservices, containers, Kubernetes, and serverless): how application delivery can be automated, and made rapid and reliable, decoupling applications from infrastructure.

To give an example of the impact 5G will have once roll-out is completed: Cisco predicts average mobile network speeds will increase from current 4G of 10 Mbps, by 2023 5G will offer 575 Mbps.

These technologies and applications that are affected are captured in a LF Edge graphic, see Figure 3.

Figure 3: Convergence of technologies and impact on edge applications



NFV Edge Infrastructure | Wireless (vRAN,vEPC) | Wireline (PON) | uCPE (SD-WAN) | IP Enterprise Services

Autonomous Devices | Drones | Autonomous Vehicles | Industry Robots | Medical

Immersive Experiences | Virtual Reality | Augmented Reality | 360 Video | Wearable Cognitive Assistance

IoT & Analytics | Industrial Sensors | Home Devices | Retail | Healthcare

On-Demand NFV | Hardware Acceleration | A.I. | Microservices | 5G

Source: LF Edge

The edge can generate huge amounts of data (big data) but much of this may not be useful information, so processing at the edge becomes an important filter to extract important data for further system use or for sending out to edge gateways or to the cloud. Local aggregation of data may also be performed. AI plays a large role in extracting meaning from local data, in multiple ways: pattern recognition, vision, audio, and more.

Autonomous driving

Autonomous driving research is being conducted by every major car manufacturer and a multitude of startups. Success in this market will lead to a massive demand for AI chips to provide the intelligence in the automation.

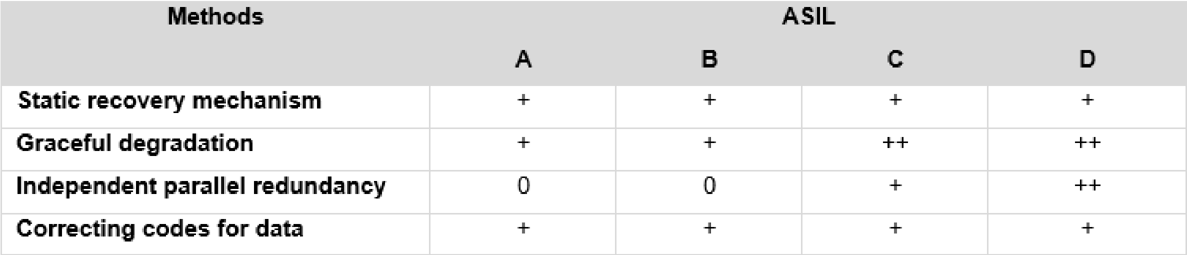
An important criterion for AV inference applications is real-time response, vehicles must have a response fast enough to react to hazards when the car is travelling at high speed. Low latency is therefore an important characteristic of chips in this space. And needless to say, the computing must be performed locally in the vehicle to achieve fastest response, it would only make sense to send data to the cloud for analysis that is less time constrained and no doubt filtered for only vital data.

Another criterion is deterministic behavior. Governing control systems must know that an answer is to be provided from an AI sub-system within a specified time limit, otherwise they need to assume a failure and take corrective action. An AI system response that is non-deterministic may pose challenges in this respect.

The certification relevant for software application, which includes DL, is ISO 26262 (“Road vehicles – Functional safety”) part 6. It specifies requirements for product development at the software level for automotive applications such as software safety requirements, software architectural design, software unit design and implementation, software unit verification, software integration and verification, and testing of the embedded software.

An important part of ISO 26262 part 6 is the Automotive Safety Integrity Level (ASIL) risk classification system for an automotive software application. The classification runs from ASIL-A the lowest risk to ASIL-D the highest risk level. For example, Figure 4 show the error handling methods recommended depending on the ASIL rating.

Figure 4: ISO 26262 part 6: error handling methods by ASIL rating: ++ = highly recommended, + = recommended, 0 = no recommendation



Methods	ASIL			
	A	B	C	D
Static recovery mechanism	+	+	+	+
Graceful degradation	+	+	++	++
Independent parallel redundancy	0	0	+	++
Correcting codes for data	+	+	+	+

Source: R Salay et al, *An Analysis of ISO 26262*, 2017, arXiv:1709.02435v1.

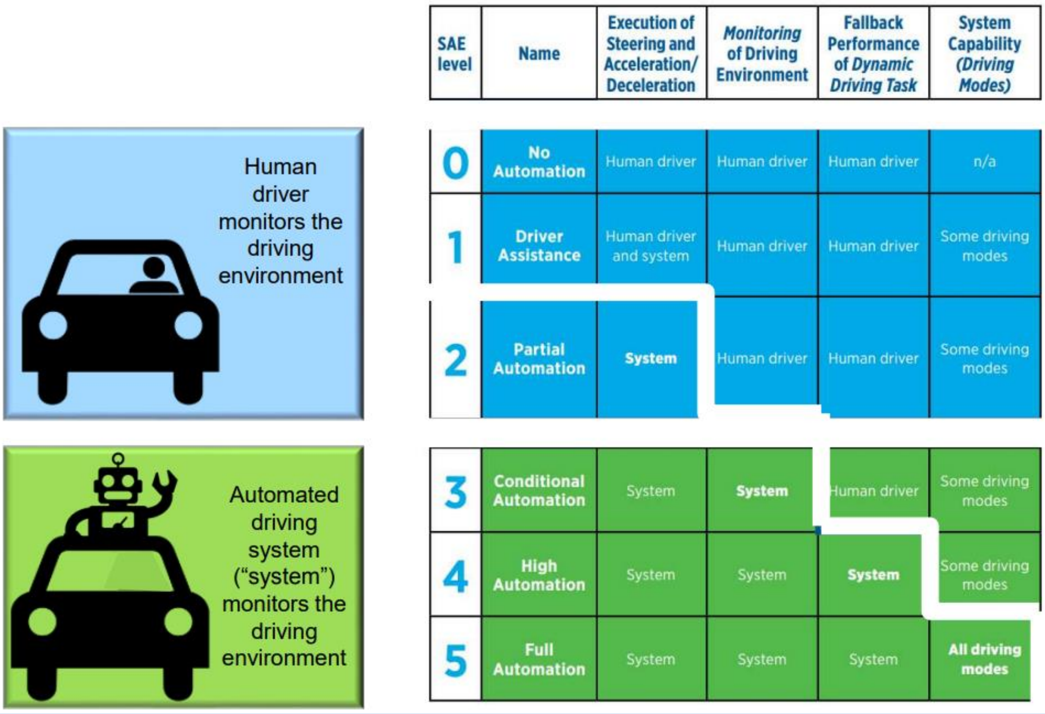
The industry is undergoing change as it accommodates the way ML applications work in contrast with traditional software. The challenges for AI chip suppliers to the industry is not just the certification processes needing to catch up with ML based software, but the challenges of an industry that is in the middle of converting itself into a software-centric one, as are many other industries. The top-end vehicles today carry more than 100m lines of software source code, this is a massive degree of complexity, to which AI only increases the issues the industry must grapple with.

To mention one major challenge of supplying this industry, the technology change cycle in software methods and in AI technology are rapid, six months for example can witness major shifts. To this the car industry builds a new model over, say, two to three years and will maintain the vehicle for a number of decades. So, an order for AI chips today going into a new model to be launched in three years' time means the technology in the car on launch will be three years old, and parts must be available for a long period thereafter. Given this long supply time the car manufacturers want evidence that the chip supplier has a long-term track record, rather difficult for a startup. So not straight forward but also not insurmountable given the need the industry has for new AI technology. For those suppliers that step up the rewards can be huge as cars sell in millions.

We focus our report on AI chips for AV so we should address the question of when it is likely that AV will be on the road. In recent times AV has gone through a hype cycle and currently there is more focus on progressing AV capability than for manufacturers to boast when their level 5 vehicles will be on the road. We believe that the intense research being conducted in AV will yield vehicles operating at level 4 in the next decade. Some limited environment taxi and bus operations are already running at level 3.

The Society of Automotive Engineers (SAE) International defines five levels of autonomous vehicles – see Figure 5. In the first wave of AV it is possible that commercial fleet operations will have a degree of remote monitoring taking place, with the capability for humans to take over if necessary, so where Figure 5 states human driver, it could be a remote driver.

Figure 5: SAE driving automation levels defined


Human driver monitors the driving environment
Automated driving system ("system") monitors the driving environment

SAE level	Name	Execution of Steering and Acceleration/ Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
0	No Automation	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	System	Human driver	Human driver	Some driving modes
3	Conditional Automation	System	System	Human driver	Some driving modes
4	High Automation	System	System	System	Some driving modes
5	Full Automation	System	System	System	All driving modes

Source: Ali Maleki, Ricardo

AI accelerator power, size, and cost constraints

The availability of power at the edge is one of the significant constraints that affects local computing. Large or networked edge installations may have a mains electricity supply or their own generators, but the options for small edge devices is often between battery (long life or rechargeable) and harvested, such as solar power.

What drives the trends in the DC, for example the Cerebras wafer-scale chip which makes the chip as large as possible because on-chip traffic is the most efficient, are the opposite in the small edge: the smallest chip will consume the least power. The small edge concerns small products, so the chip has to fit within these constraints power and size. Cost forms a third constraint when these small devices are inexpensive, so the AI chip component has to fit within a budget.

Vendors in the market make distinctions between the edge and ultra-edge or extreme edge, and this refers to the above constraints meeting certain criteria. Figure 6 shows some example numbers.

Figure 6: Example constraints at the edge with some typical values

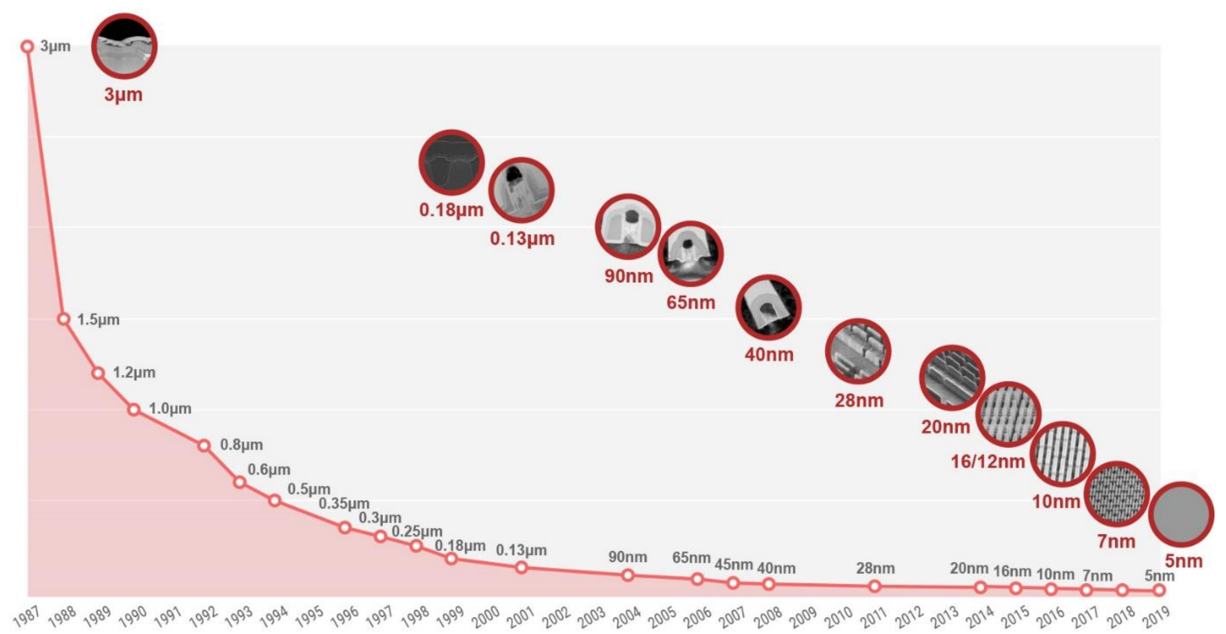
AI processor constraints	Small edge	Extreme edge
Power	~0.1W – 100W	~1mW – 0.1W
Size of chip	~25 mm^2	~2.5 mm^2
Cost (mass volume)	$10	$1
Compute	1 GOPS – 1 TOPS	1 MOPS – 1 GOPS

Source: Kisaco Research

The choice of chip fabrication process plays a role in managing costs for the chip maker - Figure 7 shows how the fabrication processes have reduced over the years. Given the budget allocation for memory size, necessary TOPS, and size of chip, the fabrication process provides another parameter. Generally, the cost per chip reduces for smaller fabrication processes, as each chip area reduces, and cost is based on the die area. However, the more readily available larger fabrication processes provide an economy of availability and also designing chips for the older generation fabrications are cheaper than designing for the latest smallest process.

Balancing and trading off the different factors, a chip maker may decide to use an older process to gain cost economy, and have the advantage at a later stage of moving to a smaller process and gaining advantages such as faster operation, lower power consumption, and a smaller form factor.

Figure 7: The reduction in fabrication process size by year



Source: TSMC

Example small edge application: keyword spotting

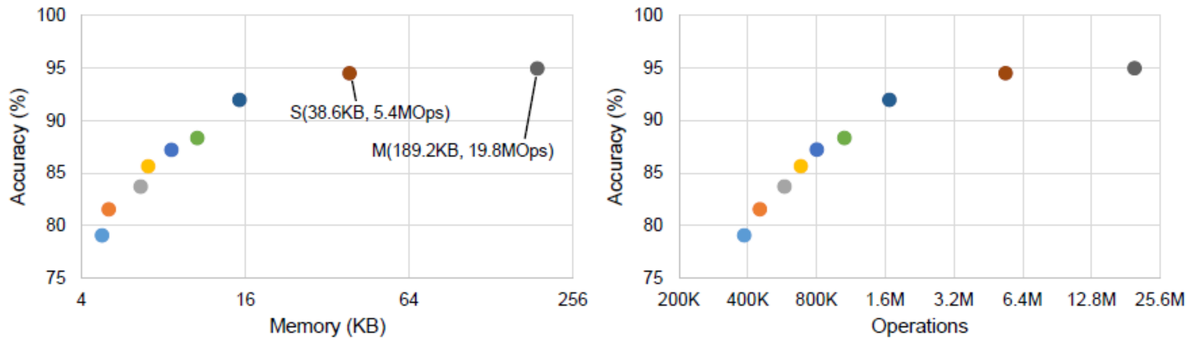
The task of keyword spotting (KWS) is how voice command always-on devices, such as smart home speakers, are activated. The challenges are a typical noisy household environment, instant recognition of spoken commands without prior training on an individual, fast real-time response, and high accuracy.

Research conducted by Yundong Zhang, a Stanford University intern at ARM, with support from Google, found that optimally balancing memory and compute on hardware optimized neural network architecture is the key to success. Using the open Google speech commands dataset, Zhang and team trained different neural net models comparing accuracy and memory requirements versus operations per inference, from the viewpoint of fulfilling requirements in embedding the application in microcontroller systems. Such systems typically contain a processor core, on-chip SRAM, and on-chip embedded Flash, and the defining constraints are low cost and high energy efficiency, meaning as low a memory footprint as possible and as few operations per second as necessary to achieve the desired KWS accuracy.

In Zhang's experiments the processor cores were various ARM Cortex processors. While some ARM Cortex CPUs have integrated SIMD and MAC acceleration features, the accelerator chips reviewed in our report (which some have ARM Cortex integrated in the chip) have dedicated cores to super accelerate DL computations. The 32-bit floating point KWS models were quantized by Zhang into 8-bit fixed-point versions (quantizing weights and activations) and found to have no loss in accuracy.

Zhang's results for various Depthwise Separable Convolutional Neural Network (DS-CNN) models are shown in Figure 8. The ideal model is high accuracy, small memory footprint and the lowest number of computations: this represents the lowest power consumption and the fastest response time for the given accuracy. Figure 8 shows how the different sized DS-CNN models scale with memory footprint on the left graphic, with accuracy dropping for memory below around 38 KB. The right graphic shows how accuracy drops once the number of operations drops below 3M.

Figure 8: Accuracy vs. memory and operations of different ML models



Source: Y Zhang et al, *Hello Edge: Keyword Spotting on Microcontrollers*, 2018, arXiv:1711.07128v3. S and M on the left graph refer to small and medium sized neural networks. The colored points represent sized variations of DS-CNN models.

The addition of an AI accelerator to the system can boost the performance (inferences per sec) and the ideal is a low cost accelerator that can allow high accuracy models to be implemented (which typically requires a minimum number of operations per sec) within the budget, size, and power constraints of the small edge product. Zhang's experiments also highlight how inferencing with integer fixed-point arithmetic is sufficient without accuracy degradation of DL models trained with floating point arithmetic - this approach is standard practice in the AI industry.

Solution analysis: vendor comparisons

Kisaco Leadership Chart on AI hardware accelerators 2020-21: edge and automotive

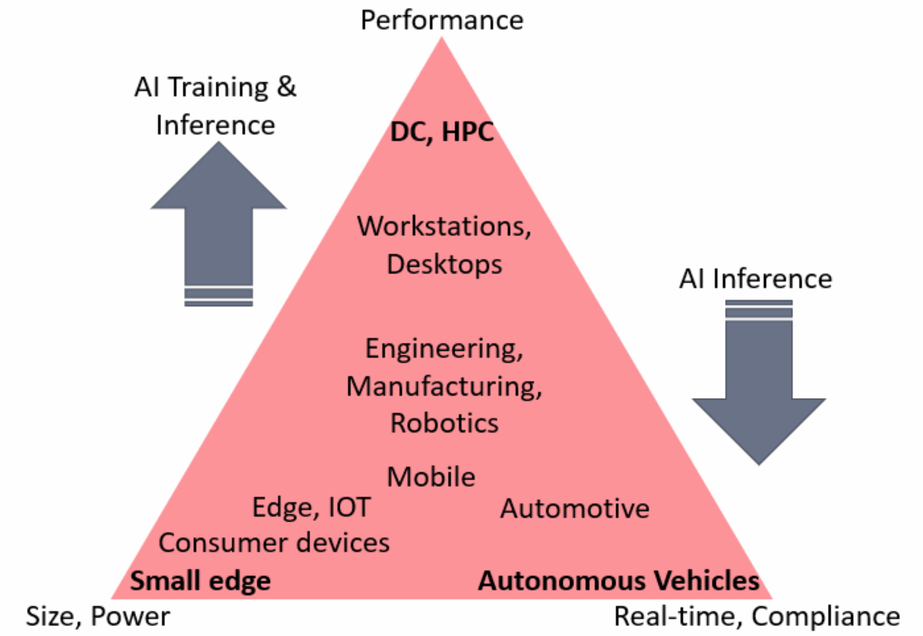
Introduction

In our companion landscape report (part 1 of this series) we identified 80 startups and 34 established players in the AI chip market (each one has a snapshot profile in that report). Therefore, our participating vendors in this report, 10 for the small edge and six for automotive-AV, are a part of the market – but note that we have on board the most significant players.

In a market as diverse as that for AI hardware accelerators, the challenge is how to ensure that our comparisons are performed with like for like products. We first split the AI hardware accelerator market into environment segments where the applications will run as follows:

- Autonomous driving
- Advanced driver-assistance systems
- Consumer electronic devices
- Data center
- Desktop/workstation
- Engineering, manufacturing, and robotics
- HPC
- IOT edge, other embedded
- Mobile and handheld devices

Figure 9: Market segment applications for AI hardware accelerators



Source: Kisaco Research

Having analyzed the responses from our participating vendors we identified three distinct categories plotted in Figure 9, falling into a triangle with vertices that have the most defining characteristics (and note these are not exclusive of each other, just the most essential): performance; chip size and power consumption; and real-time response and highly regulated. The figure also shows where AI training and inference are mostly likely to take place.

Based on this view we selected three application areas, DC/HPC (AI train and inference) – covered in the companion report; and in this report AI inference for small edge, and automotive – autonomous vehicles (AV).

Mobile devices are a distinct and separate market with many players that mostly do not crosscut the vendors we worked with in the other categories, so we have not assessed the mobile market. Consumer electronic devices largely overlaps with edge/IOT so is subsumed in that. That leaves a middle ground of desktop/workstations and engineering/manufacturing/robotics which has such diverse needs that performing like-with-like comparisons is a challenge and beyond the scope of this series of reports.

The KLC charts for AI hardware accelerators: AI inference for small edge

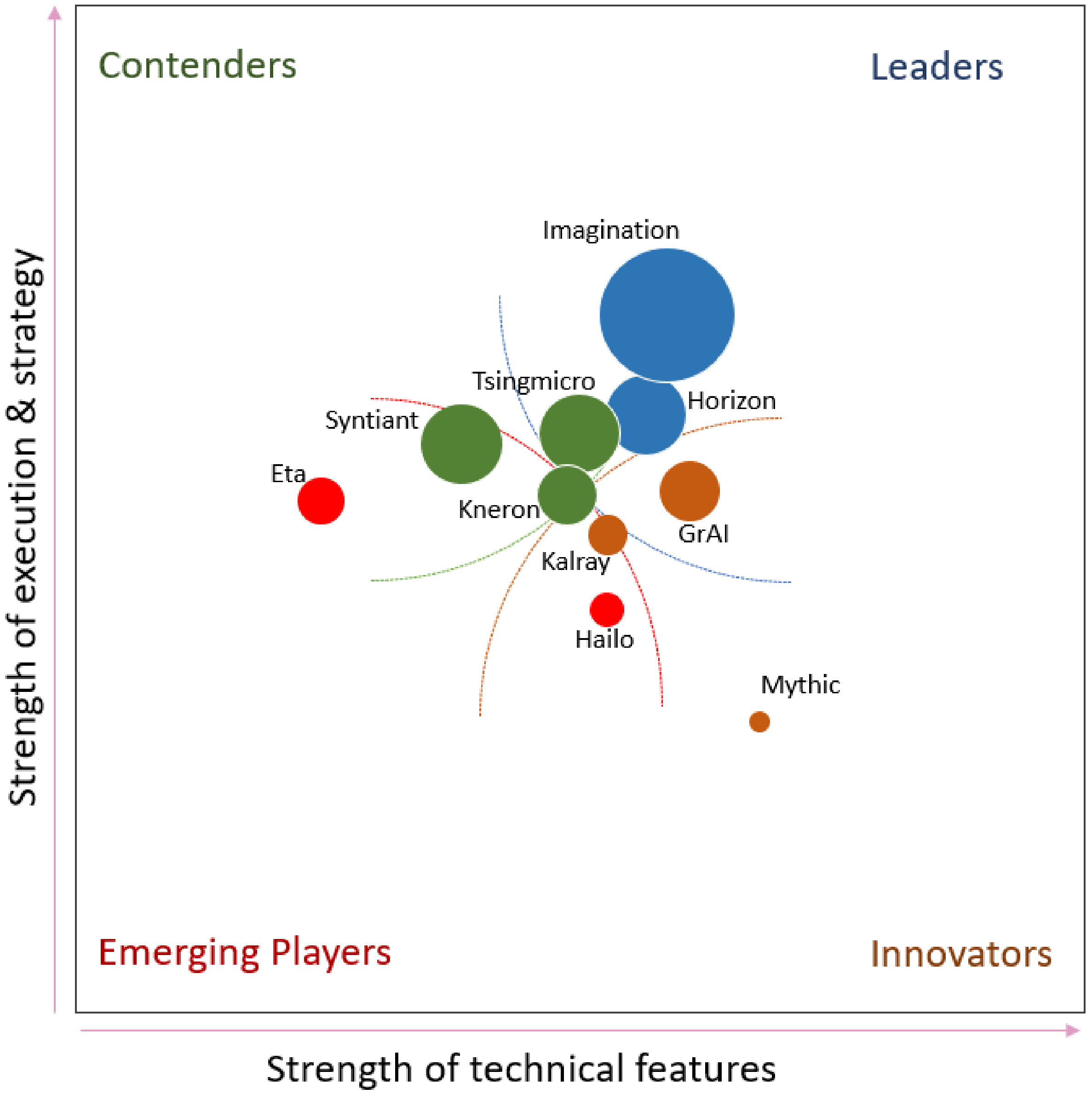
The small edge has constraints of size and power and AI chips are typically used for AI inferencing in embedded applications used in products such as for audio wake word detection, image recognition, and more. In the small edge inference mode KLC we covered 10 vendors as follows:

- **Eta Compute**: ECM3532.
- **GrAI Matter Labs**: GrAI One and anticipated next generation.
- **Hailo**: Hailo-8.
- **Horizon Robotics**: Journey 5.
- **Imagination Technologies**: NNA Series3NX.
- **Kalray**: Coolidge MPPA3-80. Contains 80 Kalray 64-bit cores
- **Kneron**: Second generation AI SOC processor to the KL520.
- **Mythic**: IPU processor. No product yet launched. We have based our analysis on the expected performance of the chip.
- **Syntiant**: NDP10x.
- **Tsingmicro**: TX510.

All the above processors require a CPU, but many contain an embedded ARM Cortex CPU, and these can in some products allow the processor to perform in stand-alone mode without needing an external CPU.

The KLC has three dimensions. The x-axis represents strength of technical features, across a range of features essential for a successful AI accelerator, from hardware performance figures to completeness of the software stack. The y-axis represents strength of market execution and its strategy, from go-to-market strategy to depth of customer support and partner network. The third dimension is the circle size and is representative of market share, normalized to the market leader revenue. The results of our analysis are shown in Figures 10 and 12. And the tables in Figures 11 and 13 tabulate the ranking of the vendors.

Figure 10: Kisaco Leadership Chart on AI Hardware Accelerators 2020-21: small edge – AI inference



Source: Kisaco Research. Circle size is representative of market share.

Figure 11: Kisaco Leadership Chart on AI Hardware Accelerators 2020-21: small edge – AI inference: ranking of vendors

Leader	Innovator	Contender	Emerging Player
Horizon Robotics	GrAI Matter Labs	Kneron	Eta Compute
Imagination Tech.	Kalray	Syntiant	Hailo
	Mythic	Tsingmicro	

Source: Kisaco Research

The two leaders in the small edge KLC both have significant revenue in the market and released new generation chips to address AI inference on the edge. The contenders and innovators are earlier in their maturity cycle and have potential to rise and challenge the leaders. Mythic is an outlier, it has truly ground-breaking in-memory technology but has yet to release its processor into the market. The two emerging players Eta Compute and Hailo are each intriguing: Eta Compute has clever technology built on off-the-shelf components, it should be evaluated on power consumption and cost basis and is an ideal choice for the right workload. Hailo is a new player with strong technology and yet to establish itself in the market.

The KLC charts for AI hardware accelerators: AI inference for automotive-AV

In the automotive-AV inference mode KLC we covered six vendors as follows:

- **Hailo**: Hailo-8.
- **Horizon Robotics**: Journey 5.
- **Imagination Technologies**: NNA Series3NX.
- **Kalray**: Coolidge MPPA3-80. Contains 80 Kalray 64-bit cores.
- **Nvidia**: New processors are expected based on the new Ampere architecture, such as Nvidia EGX A100 and Drive AGX.
- **Tsingmicro**: TX510.

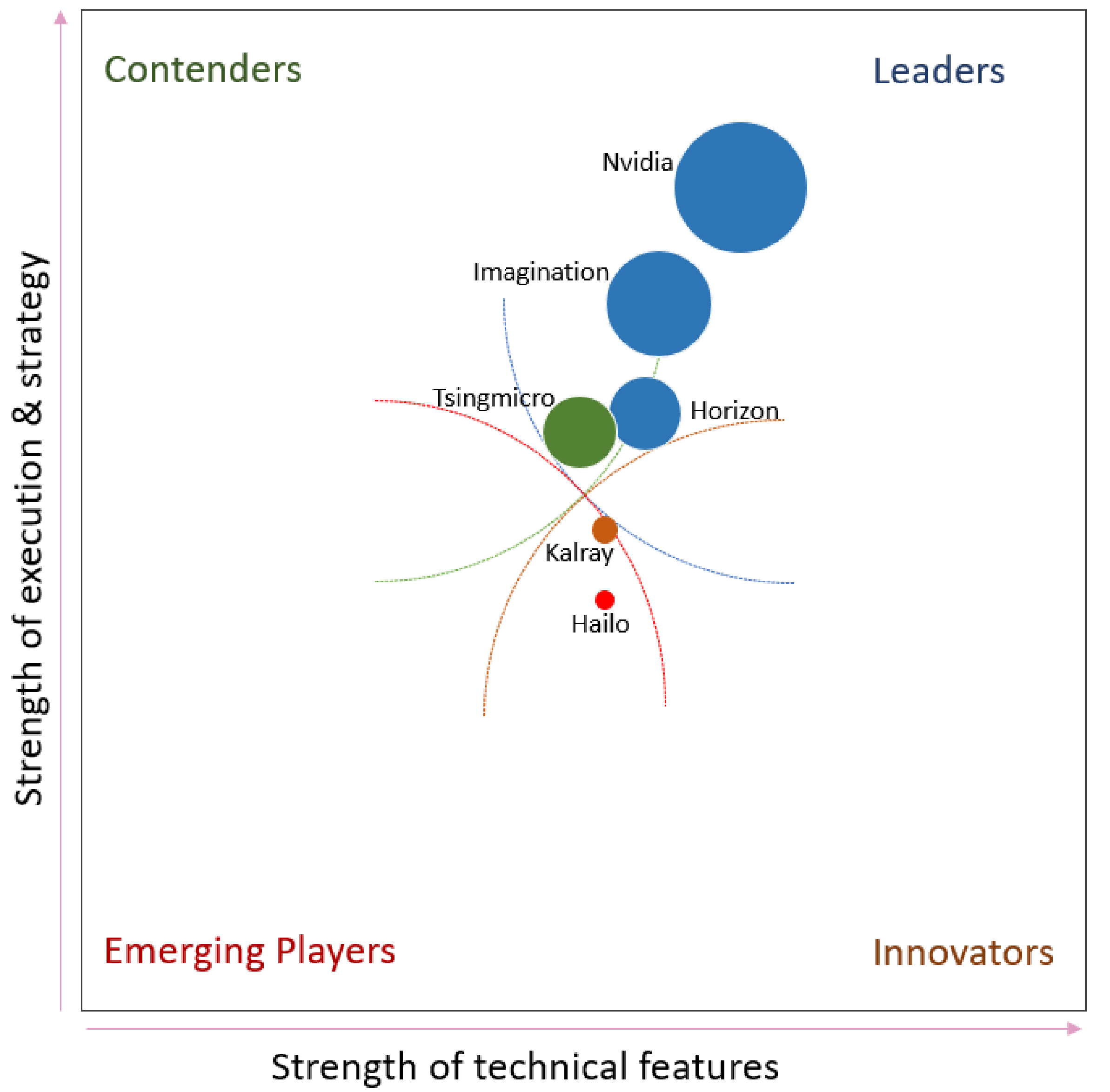
All the above processors require a CPU, but many contain an embedded ARM Cortex CPU, and these can sometimes allow the processor to perform in stand-alone mode without needing an external CPU.

The automobile industry has strict requirements on suppliers, with compliance certification, company viability as a long-term supplier, and long lead times, as well as long term support for vehicle that will be on the market in customer hands for decades. Only a sub-set of the AI chip makers are prepared to enter this market and we review six players in our KLC.

We identified three leading players in the automotive-AV AI chip market participating. We believe this market is particularly tough for new entrants given the greater hurdles required to satisfy the auto manufacturers.

Nvidia earned £700m in the automotive market segment according to its 2020 annual report, and with its new Ampere GPU technology launched recently we make it a clear leader. Imagination Technologies and Horizon Robotics are also important players in this market, and we rate them leaders. Tsingmicro is a close contender to the leaders, and Kalray and Hailo we expect will improve their standing as they mature and grow in the market.

Figure 12: Kisaco Leadership Chart on AI Hardware Accelerators 2020-21: automotive-AV – AI inference



Source: Kisaco Research. Circle size is representative of market share.

Figure 13: Kisaco Leadership Chart on AI Hardware Accelerators 2020-21: automotive-AV – AI inference: ranking of vendors

Leader	Innovator	Contender	Emerging Player
Horizon Robotics	Kalray	Tsingmicro	Hailo
Imagination Tech.			
Nvidia			

Source: Kisaco Research

Data centers and HPC (a companion report)

The data center and HPC environments are where AI systems are typically trained as well as inferenced. Adding accelerators to the CPU has proved essential for AI workloads based on deep learning as well as other AI models. This report reviews the leading AI accelerators for these environments and has two KLCs, one for training mode and one for inference mode (which includes accelerators dedicated to inferencing and not training). The vendors covered are:

- Cerebras
- Graphcore
- Imagination Technologies
- Habana
- Kalray
- LightOn
- Mythic
- Nvidia
- Xilinx

The analysis will be found in the companion report.

Vendor analysis

Eta Compute, Kisaco evaluation: Emerging Player

Product: Eta Compute ECM3532, neural sensor processor with CVFS.

Eta Compute was co-founded in 2015 by Chief Technologist Gopal Raghavan and is led by President and CEO Ted Tewksbury. The company is headquartered in Greater LA, CA, and has 30 people across four sites: its HQ, Westlake Village, CA, San Jose, CA, Austin TX, and Bangalore India for its software division.

The company has 15 patents (some currently in progress) bringing innovation to its mission, which is to build efficient AI inference processors at low and ultra-low power for the edge. It started with work on spiking neural networks but pivoted to traditional neural networks and deep learning, exploiting its patented continuous voltage and frequency scaling (CVFS) technology. The approach taken by Eta Compute in the AI accelerator space is to use off-the-shelf components inside its chip combined with its innovative CVFS hardware-based technology to yield outstanding performance for a range of applications, including AI.

Eta Compute’s first product, the ECM3531, was launched in 2018, and won in that year awards at ARM TechCon for the Best Use of Advanced Technologies and Design Innovation of The Year in the Speech, Image, and Video category. In Feb 2020, the company announced its production grade product the ECM3532, going into production in Summer 2020. The fabrication process is standard 55nm which is highly economical with proven technology.

Early customers are using Eta Compute for voice activation/wake keywords being used in edge device consumer space such as ear buds. The second type of customer is in the IOT space for applications in hotel room temperature monitoring, or video/camera people detection, for example in home security, and sound classification. The company is looking at automotive applications such as tire and passenger monitoring. Another application area being explored is in intelligent computer interfaces, such as voice commands and voice ID.

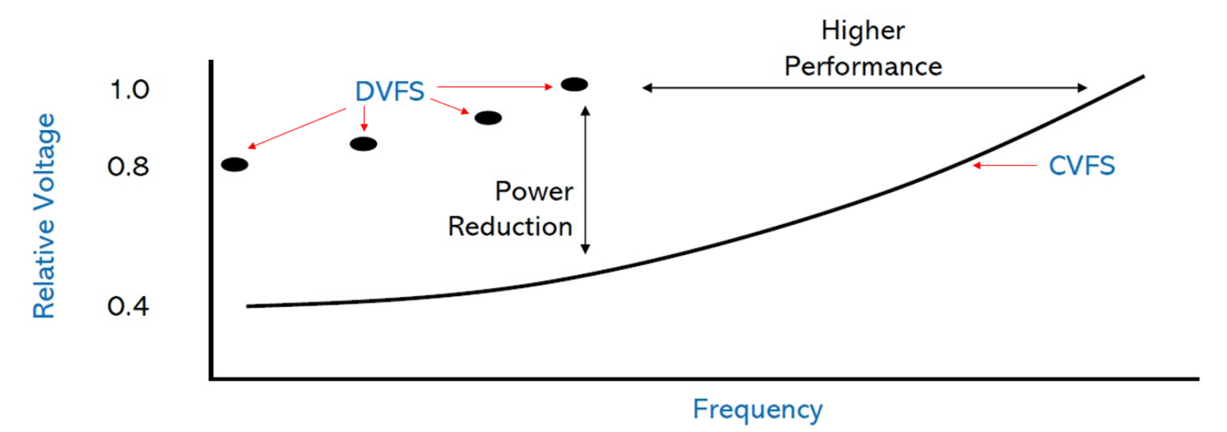
Eta Compute divides the edge between the 0.1–100W processor power range, characteristic devices are cameras, cars and mobile phones, and extreme edge in the 1mW–10mW processor power range, where the characteristic devices are wearables and low power monitoring sensors (e.g. fire detectors). The promise of IOT is met by the following challenges and needs AI to overcome them:

- Data storage, data maintenance and communication cost.
- Battery killing RF transmissions.
- Security and privacy.
- Poor real time response (for example intelligent home speakers can take about 2 seconds to respond).

However, the AI component must meet certain constraints:

- **A power budget of a few mW**: due to reliance on battery or harvesting energy. Hearing devices must operate for 5 hours with a 50 milliampere hour battery. AI can add hearing improvement, event detection, etc. A building's power consumption is about 19% of its operating cost, smart sensors can help reduce that.
- **Less than $10 system cost**: extreme edge devices are generally already low-cost commodity products (microcontrollers sell for < $10) and so added smartness must keep the cost low. On the plus side the microcontroller is typically already a powerful SOC device.
- **As high performance as possible within above constraints**: trading off the constraints, as clearly performance will be limited by the power and cost, the performance must be achievable within a microcontroller SOC operating at a few 100MHz.

Figure 14: Eta Compute's patented technology CVFS improves over traditional DVFS



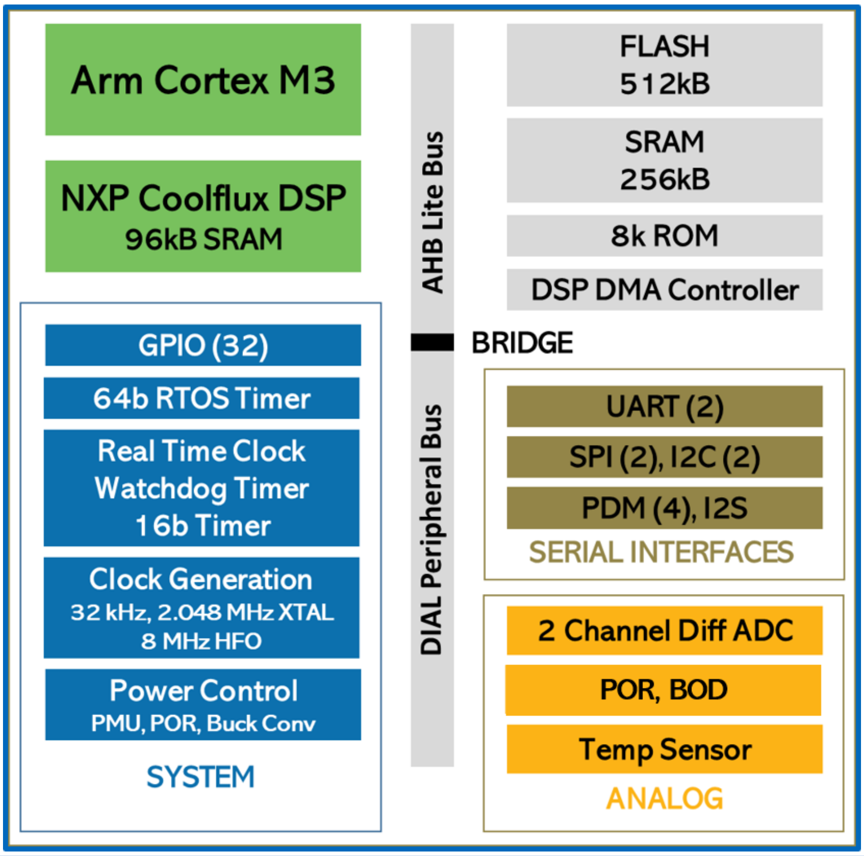
Source: Eta Compute

The novel approach taken by Eta Compute in the ECM series of chips is self-timed AI cores using CVFS – see Figure 14. In micro-controller units (MCUs) the logic works in synchrony against clock cycles that operate within a voltage range, usually 0.9-1.2V. To reduce power one option is to reduce voltage, which has a quadratic relationship to power and is therefore a highly effective strategy. However, reducing voltage reduces the maximum operating frequency (and hence speed of operation). Dynamic voltage frequency scaling (DVFS) is the traditional approach to control voltage and frequency together and hence reduce device power consumption. But there is a limit to how DVFS can be used in fully synchronous designs, as varying voltage across components in devices causes inconsistent logic timing, i.e. the voltage variation behavior is device dependent. This has meant DVFS is limited to a small voltage range, a few 100 mV, and to a few pre-defined discrete voltage levels – these discrete points are indicated in Figure 14.

Self-timed CVFS allows each device to adapt to these device logic-voltage variations and adjust voltage and frequency automatically to prevent timing violations. CVFS allows devices to have voltage and frequency modulated continuously from low to high voltage and run most efficiently for any given workload. In Figure 14 the performance curve for CVFS is shown as a continuous curve, and able to operate at lower voltages than DVFS and with higher device performance.

Eta Compute has taken a radically different approach in the AI hardware accelerator market as its ECM3532 device takes off-the-shelf components, the ARM Cortex M3 and dual NXP 16-bit MAC DSP with 96KB SRAM and adds its CVFS to ramp up performance. The device architecture is shown in Figure 15. The device implements a complete AI process from signal ingestion, feature extraction, inference, analysis, and connection with the application, with network connection to the cloud where necessary; all the steps required for a smart IOT node. The ECM3532 is called a neural sensor processor and offers analog power management to improve the neural network performance. This approach is highly versatile as the device can implement any DSP targeted calculations, such as FFT, and machine learning algorithms: deep learning with CNN, RNN, search trees, linear regression etc.

Figure 15: Eta Compute ECM3532, neural sensor processor with CVFS



Source: Eta Compute

Compared with running a neural network on a traditional MCU, the Eta Compute advantage with ECM3532 is an overall 1000-fold improvement in efficiency. It performs in the field with 3mW or less power consumption.

The advantage of using existing chip components is that the software comprises proven software from providers like ARM. However, Eta Compute is working on a software solution (for release in July 2020) in partnership with an embedded edge software specialist, Edge Impulse. The solution is called TENSAI and it takes TensorFlow and ONNX compatible models and runs the TENSAI compiler that also performs an optimization for implementing an inference deep learning model on the ECM3532. The automated neural network optimization includes standard techniques such as pruning and merging, as well as optimization that is designed for the components in the chip. This can reduce the number of weights and operations required to achieve a given accuracy and uses less memory consequently.

Kisaco Assessment

Strengths

- Eta Compute's CVFS is a key technology that allows the company to produce exceptional performance out of standard industry building blocks, pushing the boundaries of what can be achieved through the algorithms that control the power and frequency in the chip's operation.
- The approach of implementing hybrid multi-core engines on a single chip allows a variety of edge AI inferencing applications. The use of standard off-the-shelf components means they are readily available and reduce the cost of producing the chip.
- Eta Compute with its partner has developed AI software optimization to further improve the performance of DL algorithms using the TENSAI algorithm. This reduces the size of the neural network leading to improved performance.

Weaknesses

- The performance of the Eta Compute chip ECM3532 in terms of TOPS per Watt is less than that of custom-built AI chips on the market for edge inferencing. However, within the constraints of power consumption and component cost, the ECM3532 will meet the needs of multiple edge inference use cases and should not be judged purely on such performance figures.
- We believe Eta Compute should work with independent benchmark organizations to create an energy efficient benchmark suitable for tiny ML chips used in edge inferencing. We understand the company is working with EEMBC, another option is to submit benchmarks in the open section of MLPerf.
- We believe the merits of Eta Compute are under appreciated and would be suitable for many edge AI inferencing use cases where low power and low cost are more important than sheer compute. Eta Compute should better highlight these use cases in the market.

GrAI Matter Labs, Kisaco evaluation: Innovator

Product: GrAI Matter Labs, GrAI One.

GrAI Matter Labs (GML) is led by CEO Ingolf Held and was founded in 2016, based in Paris, France. The company has received some $15m in Series A venture funding, holds over 15 patents, and has around 45 employees. GML's mission is to bring AI inference to the edge, and the challenge that it addresses foremost is that by its estimate, some 95% of power is wasted at the edge by traditional AI accelerators because they ignore sparsity in the sensor data. While sparsity in structure and computation can be exploited in model development, GML is also able to leverage real-time data sparsity for significant gains. GML brings light AI to the edge, with low power consumption and compressed computation that has ultra-low latency than competing approaches.

Edge computing applications are often running in real-time, such as smart devices responding to inputs, or human-machine interfaces (voice, visual), tightly coupled feedback systems, autonomous systems (driving, piloting) etc. The edge input data streams are also typically continuous in real-time. The data rate is also much higher than the useful information content within the data. GML provides two examples to elucidate this:

- **Home intelligent speaker**: two microphones picking up voice control will sample at 16k samples/s at 16 bits, which equates to 512k bits/s. However, most of the time a person is not talking to the system, when she/he does speak it is on average at a rate of 39 bits/s.
- **Always on UXGA access control video system**: generates 79MB/s, whereas when no caller is present there is no information at all. The lossless compression can be higher than 95%.

These examples illustrate the time sparsity occurring in time for typical edge use cases. To delve deeper into the video case, with continuous video surveillance in a static location most of the data streaming is unchanged, for example in a residential road and a car driving through the only data to change is the location of the car as it moves across the field of view, the change is sparse, less than 5% of the total data content. Furthermore, the changes are highly localized and are correlated in space and time. Correlation is at the root of the sparsity benefit: an AI competition dataset like ImageNet that shows random images to an AI system has no correlation, whereas the series of images input into a security camera will be highly correlated, showing the same environment with a few changes due to moving objects.

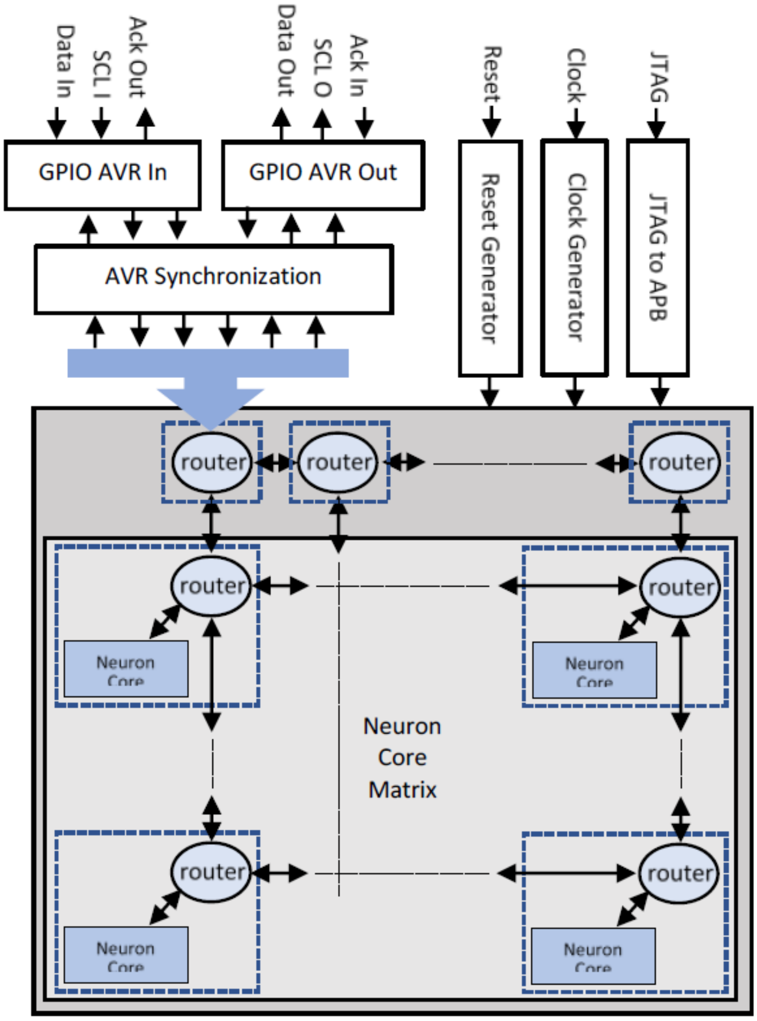
Occurrence of sparsity can be summarized in four ways:

- **Input data in space**: As data dimension grows (e.g. 1K pixels to 4k pixels), the proportion of null data points (taking differences in data) grow exponential, but the amount of useful information stays low.
- **Input data in time**: Change happens slowly.
- **Neural network model connectivity**: only significant weights carry useful information.
- **Neural network model activation**: Less than 40% of neurons may be activated by an upstream change.

Traditional approaches will ignore sparsity related redundancy summarized above, instead processing all input data and all model weights, wasting much energy and increasing latency as a result. GML has therefore addressed both aspects of sparsity in what it calls event-based computation of networks, where only data changes are processed. The first frames in any application will be processed in a traditional manner but subsequent passes through GML networks exploit redundancy, so only changes get propagated through the network. The model is reset after a period with all the data once many changes have accumulated. GML's approach combines software and hardware to exploit sparsity. The company quotes an autonomous driving data-set workload example (inspired by Nvidia, a small 6-layer tiny PilotNet model built with TensorFlow) implemented on GrAI One, in which the number of million operations (MOps) reduced from 5.14 to 0.26 per frame for the same model accuracy, resulting in a 20 fold improvement.

There are two ways to exploit sparsity: one can process very high frame rates, and as only about 1/20th of operations are needed to be processed this can be performed more efficiently than with other accelerators, or one can run the accelerator at a low clock rate and save on power. This also impacts utilization: in the former low latency scenario you can run the chip at maximum utility, in a low power scenario you can shutdown large areas of the chip.

Figure 16: GML GrAI One architecture



Source: GrAI Matter Labs

GML's first proof of concept chip GrAI One was released in Dec 2019, it is a 28nm process silicon prototype, is fully programmable comprising 14x14 array of Neuron Cores, each core features 1024 neurons, supporting a total of 200,704 neurons. Cores are connected by a proprietary network-on-chip design. The chip has a dataflow architecture, shown in Figure 16. The device is less than 4.5x4.5mm and is fully self-contained with no external DRAM necessary. A large part of each Neuron

Core comprises local SRAM. Building this device also allowed the end-to-end software stack to be developed and prove the concepts with software and hardware.

GML has demonstrated GrAI One using an external host system with three example edge cases: the Nvidia six-layer PilotNet model, a keyword detection application, and an event-based camera that responds to gestures. The event generation for standard cameras (calculating frame differences or 'deltas') is done in software but in the upcoming product chip will be done in hardware. GML calls its implemented sparsity exploitation technology NeuronFlow. The upcoming product chip is planned for 2021, and will be a smaller process than the current 28m fabrication SOC. It will also feature a number of enhancements including a larger core, and industry standard interfaces.

The SDK to support GrAI One is named GrAIFlow and allows general software programming in C++ and Python to work against the Neuron Core API, as well as read TensorFlow models. This SDK has direct network support, integrated simulator and debugger, and graphical code editor. This stack has a Mapper that optimally maps a model to the hardware. An important aspect is setting in the software the various localized sparsity thresholds.

Kisaco Assessment

Strengths

- GML has developed several innovations that combined have produced an outstanding AI accelerator. Foremost is the exploitation of sparsity at three levels that allows considerable reduction in computation for a given task, compared with other approaches. While exploiting sparsity in avoiding computing with low valued neural network weights is now more commonly practiced.
- It is exploiting sparsity (spatial and temporal) in the input data that is the stand-out approach taken by GML. This again reduces the amount of computation required, for example in vision tasks compared with standard approaches.
- The neuron cores are designed with near-memory SRAM that reduces data traffic bottlenecks and as a result the GML approach does not require slower, external DRAM. This also reduces the overall power consumption.

Weaknesses

- This review is partly based on GrAI One, which is a prototype, and on the general release DSA/ASIC chip expected in 2021. We do expect GML to deliver.
- We would like all the vendors participating in this report to take part in independent benchmarks and urge GML (and all the others not yet releasing benchmark data) to join MLPerf and others.
- As a startup it is perhaps not surprising to see a lack of maturity in customer support beyond direct support. The AI market is new and awareness of how AI chips can make a difference is a learning experience for all parties: manufacturers, suppliers, and software developers. We hope to see better support for accelerating that learning curve.

Hailo, Kisaco evaluation: Emerging Player Profile

Product: Hailo-8 Deep Learning Processor for edge devices.

Hailo is based out of Tel-Aviv and was founded in 2017 by founders CEO Orr Danon, CTO Avi Baum, and CBO Hadar Zeitlin. The company is currently 90+ people with strong technical background, and recently completed $60m in B-round funding with new strategic investors NEC and ABB, bringing its financing to a total of $88m. The new funding will help Hailo to expedite the deployment of new levels of edge computing capabilities to smart devices and intelligent industries around the world, including mobility, smart cities, industrial automation, smart retail, and beyond.

The mission from the get-go was to develop the most efficient deep learning inference chip for use in edge scenarios, including vision applications in videos and cameras, automotive ADAS and autonomous vehicles, intelligent consumer devices, smart cities, smart retails, IOT machine learning in assembly line quality checking, and robotics. The company holds 10+ patents on its distinctive structure-defined dataflow architecture and announced its first chip Hailo-8 in mid-2019, with full scale production targeted for second half of 2020.

The market for edge devices is predicted to grow and with it the need for intelligence on the edge. The sweet spot for Hailo is low power consumption and high performance and to achieve this it has designed a domain specific architecture processor named Hailo-8, a square device about the size of a thumb nail 17x17 mm, with the following characteristics and features:

- High performance: 26 TOPS.
- High efficiency: 3 TOPS/W.
- Low latency: no external memory required.
- Can run self-contained or as a co-processor.
- Has a fully programmable and comprehensive SDK.
- Will be certified for automotive industry applications: ASIL-B (D), AEC-Q100 Grade 2, and A-SPICE certified toolchain.

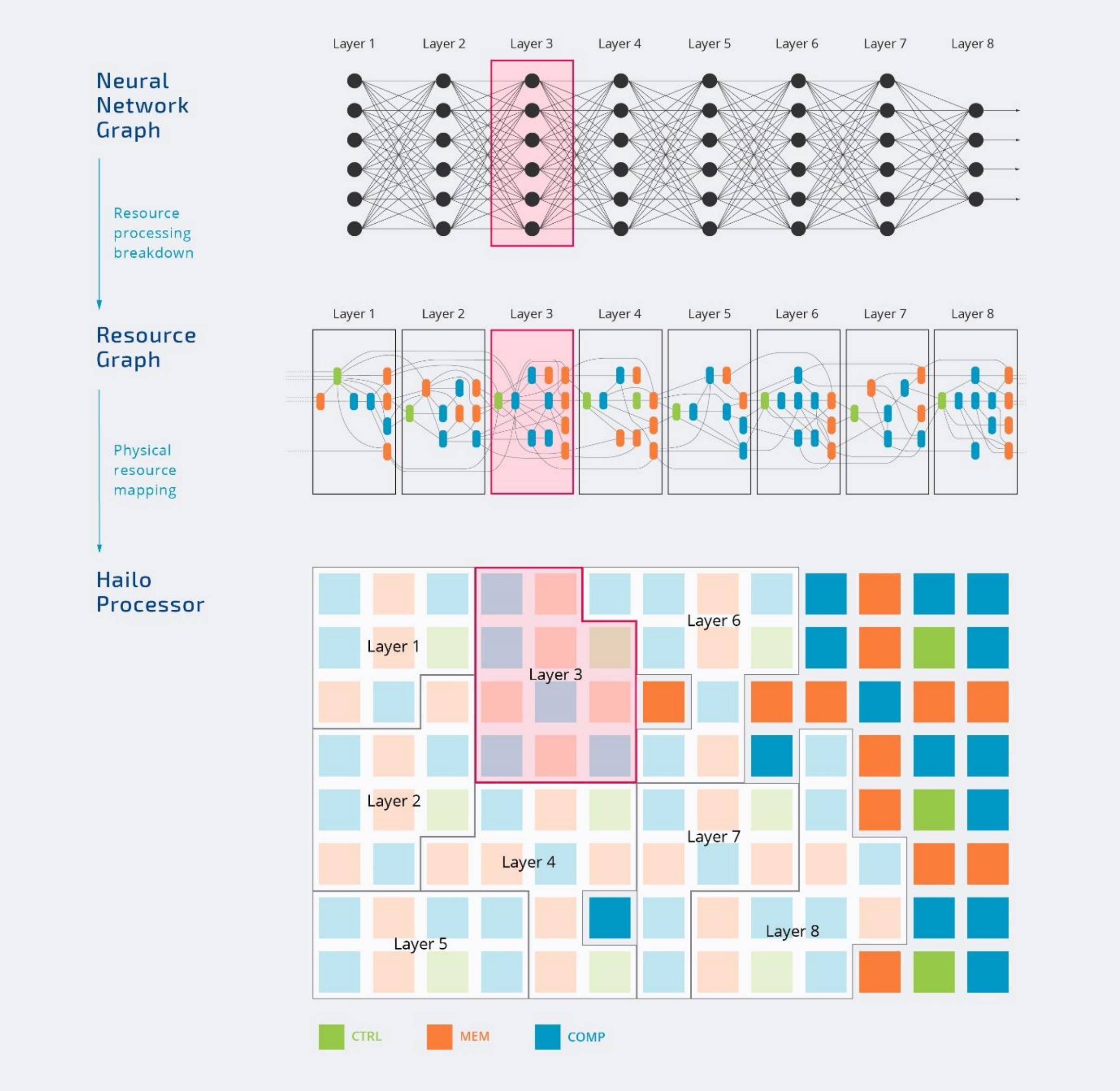
The ResNet-50 classification benchmark (224x224 image resolution fed at 780 FPS in a batch size of 1 and with the processor running in 8-bit precision), has total power consumption of 1.7W and the power efficiency is 2.8 TOPS/W. Hailo is specific to point out that this is the original ResNet-50 model and not a pruned and reduced size version. Further benchmark figures include MobileNet-Single Shot Detector at 720p reaches 1.9 TOPS/W, and Fully Convolutional Network 16x up sampled semantic segmentation at 1080p which achieves 2.6 TOPS/W.

The Hailo-8 advanced structure-defined dataflow architecture translates into higher performance, lower power, and minimal latency, enabling more privacy and better reliability for smart devices operating at the edge. With reference to Figure 17, the Hailo-8 neural network core architecture comprises multiple patterns of three hardwired functions:

- **Control** (colored green in diagram): compile time flexibility (allocation), and fully deterministic in runtime.
- **Memory** (orange): near memory model, parameters and partial sums are localized, and layer outputs move only nearby.
- **Compute** (blue): Designed for recurring MAC operations at mostly low precision.

As shown in Figure 17, while the pattern of these three fundamental elements are fixed, the compiler (offline) allocates the units to each layer in the neural network, so that each layer has one control unit and multiple compute and memory units as required. The performance can be scaled out by cascading several chips via the chip-to-chip interface.

Figure 17: Hailo-8: structure defined dataflow architecture.



Source: Hailo

The Hailo-8 processor contains one neural network core (as described above), an ARM Cortex-M4 embedded CPU, an H.264 video compression encoder circuit, and an image signal processing (ISP) unit. The device has two primary use cases:

- **Hailo Centric System**: A sensor connects to Hailo-8 via a MIPI or parallel interfaces, and output connectivity is via ethernet or Secure Digital Input Output interface. In this configuration the Hailo-8 acts independently of an external CPU and is suitable for low end use cases that have budgetary constraints.

- **Hailo as co-processor**: In this configuration the Hailo-8 is fed tasks by an external CPU application processor. Sensors can be connected to the application CPU or directly to the Hailo and output connectivity is also to this external CPU. Hailo-8 and the external CPU communicate via MIPI, ethernet or PCIe.

The Hailo Dataflow Compiler (an offline SDK) can have as an input a TensorFlow or ONNX standard file and converts it into a binary file to run on Hailo-8 using its compiler, and includes a model translator, numeric translator, and resource allocator. In addition, the Dataflow Compiler has an off-line emulator and profiler, which allows the user to estimate the frames per second and power efficiency in offline mode without the need for the actual Hailo-8 device. The HailoRT orchestrates the computation on the core processor.

To help speed up adoption, Hailo offers a Hailo-8 Fast Track Program developer suite comprising:

- Development kit.
- SDK license.
- Industry standard neural network examples.
- Training.
- Technical support.

Kisaco Strengths and Weaknesses Assessment

Strengths

- The structure-defined dataflow architecture is Hailo’s stand-out, patented innovation: this allows the Dataflow compiler flexibility in how it allocates compute regions. The low latency design also does not require external DRAM, which helps conserve power. The low power of the chip helps to yield the high TOPS/W rating. Since the launch of Hailo-8 in May 2019, Hailo managed in a short time to engage with large scale companies such as NEC, ABB, and Foxconn.
- Hailo is targeting the inference edge and this includes autonomous driving, where it is going through a number of certifications to compete in a market that is expected to grow significantly.
- Hailo offers a comprehensive SDK encompassing translation to internal numeric representation including neural network weight quantization. This includes an off-line emulator.
- We applaud Hailo for being one of a small number of participating vendors in this report to have submitted a benchmark to MLPerf.

Weaknesses

- The Hailo-8 chip is currently sampling and is available for selected customers. The chip is expected to be in production in the second half of 2020 so there is a degree of anticipation in how we have rated Hailo. We expect to increase its market execution position in our next report iteration.
- As with many startups we are covering, the challenge will be to have design wins once the chip goes to production. We expect this to happen as the company grows.

- There are certain types of neural networks that cannot be modelled in Hailo, such as Bayesian. This should not be an issue, as most DL neural models for the inference edge use more common types of neural networks which Hailo supports.

Horizon Robotics, Kisaco evaluation: Leader

Profile

Product: Horizon Robotics, Journey 5 AI processor and its embedded brain processing unit (BPU).

Horizon Robotics was created in June 2015 by CEO Kai Yu, VP Algorithms Chang Huang, and COO Annie Tao and others. The company has raised $600m in Series B funding in February 2019, bringing its valuation to several billion dollars, with major investors including SK Hynix. Investors, including from previous rounds, include Intel, CMBC Capital, Morningside Venture Capital and Hillhouse Capital, amongst others. The company has around 900 employees and holds approximately 600 patents. The company has offices in China with HQ in Beijing, and an office in Silicon Valley staffed with AI researchers, deep learning algorithm and application developers, as well as automotive sales and marketing associates. The company focus is on AI inference for low-power edge applications, including smart mobility that encompasses ADAS, autonomous driving and intelligent cockpit for diverse types of vehicles such as personal cars, robotaxi and roboshuttle, delivery robots, trucks, and more.

The company is building an open edge AI platform on four pillars:

- **BPU based AI processors**, exploiting the BPU’s deep learning capabilities and innovative MIMD architecture, such as the Journey 2 auto grade, edge AI processor, with dual BPU cores.
- **Matrix perception compute system**: Vehicle-ready AI computer technology and associated frame grabber and HIL toolkit, to support software testing and validation, that can also be used as a hardware reference design for customer’s system deployment and mass production. A Matrix 2.0 device contains four Journey 2 chips.
- **OpenExplorer**: Horizon has developed an open AI toolchain for model training, quantization, optimization, and deployment onto the BPU based processors. OpenExplorer can take in models trained on popular AI frameworks: TensorFlow, MXNet and PyTorch. The stack middleware can implement a message subscription and distribution mechanism or a data flow-based programming framework. Customers can choose whether to use Horizon algorithms or their own on the Horizon processors and ensure that the deep neural networks are automatically optimized to run on the BPU, with a high utilization rate of the engine.
- **AI algorithms** for perception and mapping localization, designed and optimized for BPU and exploiting camera and LiDAR data.

Horizon’s autonomous driving technology is being used in robotaxi and delivery vehicles world-wide (USA, Europe, and China), and for ADAS in personal vehicles, currently in test and validation stages of development. Prior to creating the ASIC BPU technology, Horizon developed its algorithms and BPU architecture on several generations of FPGAs and these are in production today. The BPU implemented on Journey 2 processor adds a further level of performance capability and cost effectiveness and is already in production and used in a series vehicle by a major automotive brand in China (Changan UNI-T model).

On top of the Horizon platform's processing optimization (low latency, high accuracy) and software flexibility (open choice of development tools), it offers cost advantages for large scale mass production of processor units, and energy efficiencies.

Horizon is marketing its processors into the competitive automotive sector. Journey 2 is qualified under the AEC-Q100 grade2 automotive quality standard (-40/+105 deg. C). Journey 2 is a system-on-chip fabricated in the 28nm CMOS process. It incorporates a dual-core CPU, delivers 4 deep learning TOPS in INT8, enables less than 60ms latency from camera exposure to final perception result output, and consumes less than 3W when running a complex suite of detection, classification and pixel level segmentation algorithms, at 30fps. The next generation, Journey 3, is a 16nm fabrication device achieving 5 INT8 TOPs, with a Quad-core CPU, improved ISP and Security features and customer samples are available now.

On June 4th, Journey 2 was awarded the Vision Product of the Year award as the Best Automotive Solution at Embedded Vision Summit 2020. Previously the system compute platform Matrix 2.0 won industry awards for its use in autonomous driving, such as the CES 2019 Innovation award and the Automotive Vision Product of the Year award in 2019.

Example customer deployments include Audi for platooning on highways in China, Continental for road-side solutions, SK Telecom for crowd sourced map updates, and Faurecia-Clarion for intelligent cockpit, ADAS and AV development.

The AI algorithms developed by Horizon for Journey 2 and Matrix 2.0 include 2D and 3D detection, with recognition of six vehicle sub-types within 150m, pedestrians, cyclist, traffic sign recognition of 376 types in US and 179 types in China, and more. Semantic segmentation of up to 23 classes, including free space, lanes and boundaries, poles, construction zones, road markings, and vegetation. The device robustness has features such as weather classification, freeze, blockage, and glare recognition, and calibration loss detection. The Matrix system competes with Tesla's full self-driving (FSD) computer, which also processes neural networks for the Tesla Autopilot system, and the appearance of FSD helped Horizon educate the market as to what its processor was achieving. The next generation of Journey will compete even more effectively with Tesla FSD and offer the auto industry access to this technology.

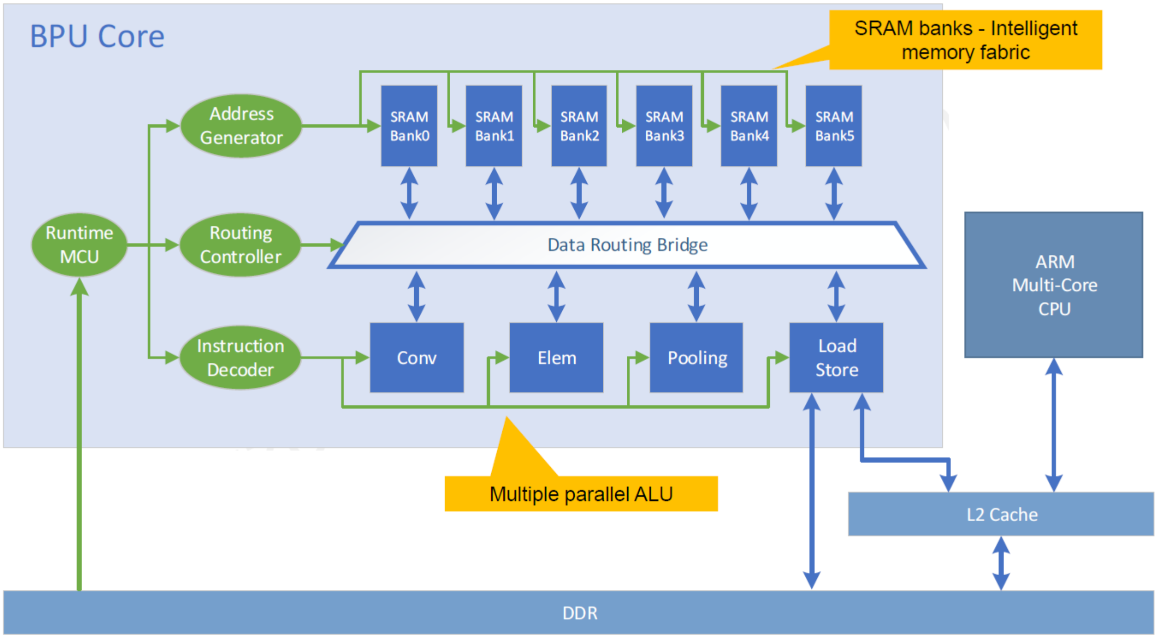
The Matrix 2.0 system is designed to work with 12 cameras (eight narrow field of view, and four fisheye) – currently a vehicle carries three Matrix 2.0 devices to achieve 360 surround vision perception. Its latency in such a use case is less than 100ms and power requirement is less than 25W per Matrix 2.0, which can be passively cooled. This system is designed to support SAE Level 4 autonomous driving. As Horizon releases its next plus one, new generation Journey chip, only one Matrix system will be needed per vehicle.

Horizon is also working with LiDAR technology, creating perception algorithms exploiting the Horizon original variable group convolutional networks (VargNet) backbone, running optimally on the BPU of Journey 2. The solution optimizes workloads across the dual ARM Cortex A53 CPUs and the MAC-rich dual core BPUs of Journey 2. The fusion of the LiDAR and camera information produces improved perception robustness. In a typical test, the LiDAR takes up around 90% utilization of a single BPU core and around 40% utilization of the CPUs to deliver up to 40fps at only 2W power dissipation. For SAE level 3 and above, Horizon sees its customers putting both camera and LiDAR on their vehicles. LiDAR manufacturers are moving the technology into solid state devices so the cost will reduce over time.

The BPU is a domain specific architecture comprising a collection of tensor (MAC) engines, optimized for running convolutional deep learning neural networks. The software and hardware teams worked closely together to design the BPU. An SRAM memory fabric is set close to the MAC cores to reduce latency. The BPU is an MIMD system (see Figure 18) whereby separate ALUs run in parallel on the chip, performing convolution (Conv in diagram), element-wise operations (Elem), pooling and memory load and store. The SRAM memory can be paired with each ALU as required.

Horizon makes the point that its processor technology is targeted at the automotive grade, which is highly exacting, and has dictated the choices of proven technology in the architectural design, usually a generation behind other application sectors because the test and validation cycles take longer. Compliance with ISO 26262 for security and safety is highly demanding and for a SOC all the components need to be compliant including the CPU and other components, making the task complex and the reason only a small number of vendors are prepared to go through this process. The new generation Journey processor will be the one being designed for compliance with ASIL B. The current Journey device is targeted at mainly ADAS and in-cabin AI applications.

Figure 18: Horizon Robotics, BPU core based on heterogeneous MIMD architecture



Source: Horizon Robotics

Kisaco Assessment

Strengths

- Horizon Robotics has advanced beyond beginning startup, despite its young age, and has established a significant set of eco-system partners and collaborators. Its current AI chip, Journey 2, is already in production for automotive applications, its next immediate processor is being sampled now, and the company is focused on developing its next plus one processor, with significant performance improvements, aimed for 2021.
- The Journey chips are qualified for the automotive industry, which demands stringent certification. This opens up a wide range of opportunities for the business across AI inferencing in the edge to applications in the automotive industry (autonomous driving, ADAS, robotaxis etc.) and this is reflected in its diverse partner eco-system.

- Horizon has created a full software stack to support the Journey chips. This includes support for popular ML frameworks. It also supports most high-level development languages and also the interface standard ONNX.

Weaknesses

- As with most edge inference AI accelerators, Horizon's technology is designed to maximize TOPS performance in Integer precision, which is optimum for DL models but not beyond where floating point processing is desirable.
- We encourage Horizon Robotics to participate in an independent benchmark standard such as MLPerf and believe the company is planning to do so.
- While Horizon Robotics has grown fast, we perceive its customer engagement is limited based on the web site which has a startup look and feel, with limited online engagement. However, we understand Horizon is engaged strategically with top Tier1s and OEMs around the world, which are its target markets, and these collaborations will yield results for the company.

Imagination Technologies, Kisaco evaluation: Leader

Product: Imagination Technologies, PowerVR Series3NX neural network accelerator (NNA).

Imagination Technologies has a long history: first launched in 1985 as VideoLogic, it was first listed on the London Stock Exchange in 1994, and in 1999 it refocused on IP licensing and changed its name to the current one. In 2017 it went back to being private under a new owner, Canyon Bridge, a private equity fund based in California, USA. It is currently led by interim CEO Ray Bingham, Founder of Canyon Bridge.

The company has presence in three markets: embedded GPUs (spanning mobile, automotive, IOT, gaming, and consumer markets), AI accelerators (and 'AI Synergy' combining GPUs and accelerators), and wireless connectivity and ethernet packet processing. The company holds fundamental patents in GPU technology (Imagination is in the top 10 UK patent holdings companies). To date it has shipped over 11 billion chips with Imagination IP. The company has over 900 employees, of which 80% are engineers, and has R&D centers in UK and China.

In the AI field Imagination launched the Series2NX NNA in Dec 2017, which has since been licensed into multiple application areas. This NNA device series is built from the ground-up to perform massively parallel MAC operations with (dynamic) fixed integer precision logic, and hence suitable for AI inference.

The company has found that the market has moved beyond the need to be educated about the potential of AI technology, to now coming forward to Imagination with AI RFPs. The company has been involved in AI research for the past seven years and has proven silicon IP, with over 80 AI related patents filed/granted on top of over $75m in AI R&D. Imagination's focus is on low power AI inference applications.

Key segments for Imagination NNA accelerators, are automotive, AIoT/security, consumer (mobile/home), and datacenter. The NNA performs quantization, allowing lower-precision numerical formats to be used, leading to advantages for edge computing: lower power and faster computation, while retaining accuracy.

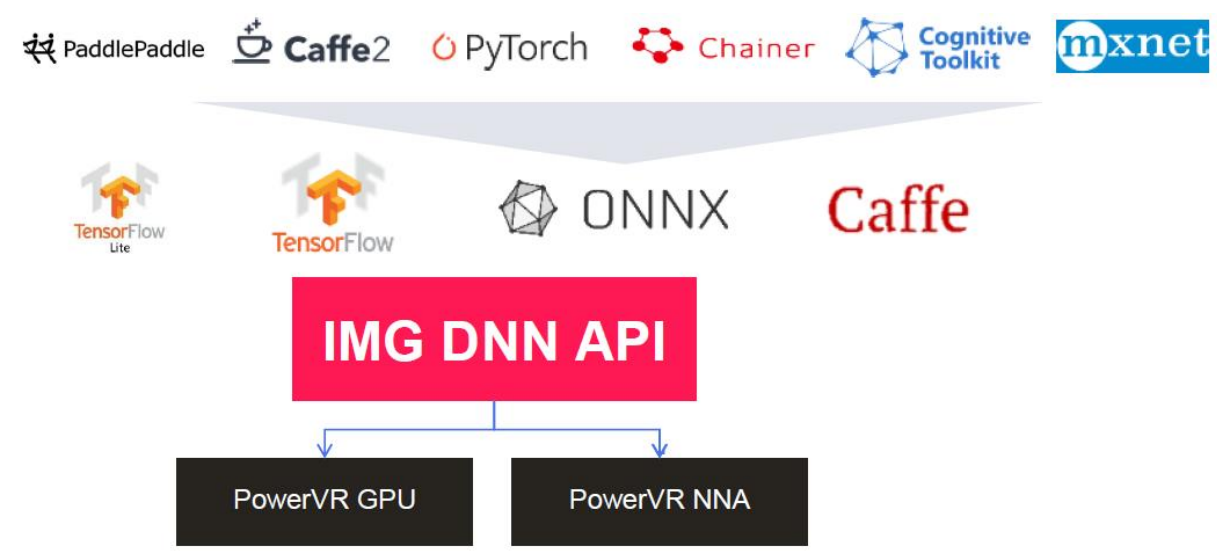
Imagination launched its Series3NX in Dec 2018, which is just now feeding through into silicon (reflecting the time lag to get a new product into the market). It has fixed-point logic and lossless weight compression which allow smaller models to be run on the edge and providing 70% improvement over the previous model in terms of inferences per sec (inf/s). Imagination refers to the power, performance, and area (PPA) features of its technology, and Series3NX has 40% improved inf/s/mm^2 and up to 35% lower bandwidth than Series2NX. Series3NX also has variable bit-depth support, so it is possible to fine tune per neural network layer the ideal bit precision (any of 4, 5, 6, 7, 8, 10, 12, and 16-bit precision). Imagination's off-line software tools support all these features.

Series3NX also offers improved security and support for the latest Android 10 OS. For IOT edge applications the Imagination AX3596 offers its current highest performance at 10 TOPS, and for the automotive and video market the UH16X40 offers 160 TOPS. Target nodes are for 16nm TSMC processes, running at 1.2GHz. Imagination makes the point that how accelerators behave varies with the application, so there is no ideal simple performance metric. The company has scaled down versions of its chips to fit the needs of different sized devices on the market and the associated level of AI computation required, so wake words require less computation than full-scale speech recognition. Conversely Imagination has seen (for example in automotive) its customers' demand for performance double with every meeting, showing the pace at which innovation is accelerating. The higher TOPS devices contain multiple units of the NNA devices.

The NNA A-Series3NX can utilize Imagination's AI Synergy approach to bring its GPUs and NNAs together to produce an optimum accelerator for applications such as in the automotive sector. Based on technology originally developed as a virtualization layer to run different OSs, these accelerators have what Imagination calls Hyperlane technology, which reserves processing capabilities in a GPU for certain tasks, so that graphics and AI tasks can run in parallel and have reserved spaces. The reservations are also variable depending on needs.

The Imagination software stack PowerVR AI is a platform with a single API for both GPUs and NNAs – see Figure 19. Popular frameworks are supported such as TensorFlow and Caffe and support for the open standard ONNX, making it easy for software developers not familiar with the embedded space to write AI algorithms to run on the Imagination chips.

Figure 19: Imagination Technologies PowerVR AI software stack



Source: Imagination Technologies

Kisaco Assessment

Strengths

- Imagination Technologies has a relatively long history in the chip business and although it launched its first AI chip in 2017, it was built on years of expertise in advanced chip architecture and development. The NNA series chips offer high-performance for AI inferencing, targeting markets the company understands well: mobile and automotive, as well as other markets.
- The company works closely with clients to develop the right scale of device for their needs. With so many variables and constraints needed to be traded against each other in embedded and edge scenarios there is often no one-size-fits-all solution and being able to scale to customer needs is a strength of Imagination.
- The capability to fine tune bit precision at the neural network layer level gives algorithm developers fine control on the performance of their models.
- Imagination understands well the need for a strong software stack, and it has built one to support NNA, with support for the most popular DL frameworks.

Weaknesses

- NNA does not perform floating point calculations and this will restrict some application areas. It does read and convert floatingpoint-32 however.
- Imagination appears in the AI-Benchmark standard but specific to mobile phones. It would be good if the company participated in the MLPerf benchmark, which is designed to characterize an AI chip across either standard AI models or open for the participant to run their own model.
- We believe Imagination has scope to improve its market execution as its play in AI becomes better known. The Jan 2020 announcement of a new deal with Apple will help.

Kalray, Kisaco recommendation: Innovator

Product: Kalray, massively parallel processor array (MPPA).

Kalray was founded in 2008 and has offices in Grenoble (HQ, France), Los Altos (USA) and Tokyo (Japan), and is led by CEO Eric Baissus. The company is a spin-off from one of the largest European research labs ("CEA") and became listed on EuroNext Growth Paris in June 2018. Kalray is a processor company, outsourcing manufacturing of its chips. It has created a manycore architecture chip, the MPPA for AI in the edge, data center and for HPC applications. While the edge today is a small part of the overall AI semiconductor spend, dominated by cloud/data centers, the market for edge AI is expected to grow significantly. The need will grow to have high compute capabilities without going all the way to the cloud.

The main emphasis for Kalray is HPC and AI applications in inference mode. Examples areas that Kalray is exploring opportunities include HPC with the Open Compute Project (OCP) Foundation for scaling NVMe Flash SSD storage, and for computational storage applications (which includes running AI). The company has invested over 80m euro in R&D and owns 30 patent families. It has 81 employees, most of which are engineers and PhDs. A key differentiation of the MPPA processor is the capability to accelerate a wide number of data crunching functions in real time, and pipe them very efficiently, with applications ranging from AI, to path planning, signal processing, network stacks or storage services.

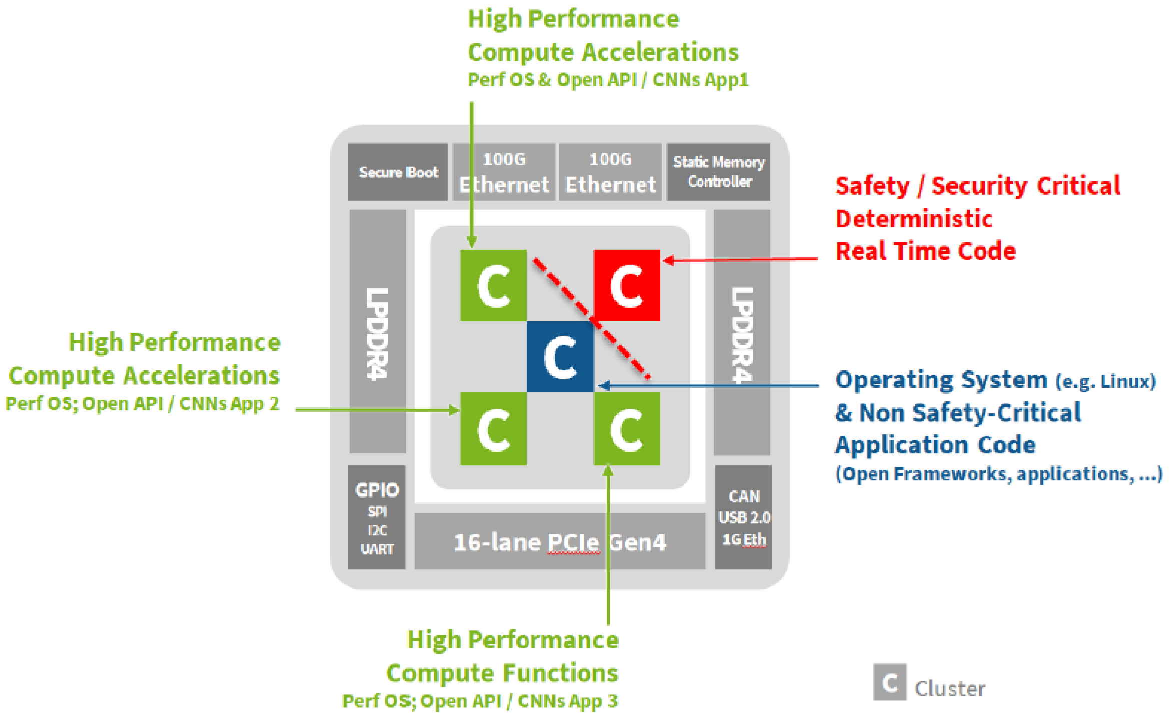
There are two target markets today: first is data center acceleration, offering a complete and standards based PCIe board and associated software solution for appliance makers and data centers; second is the automotive market for ADAS and autonomous driving, addressing the needs for delivering SAE levels 2 to 5, with testing currently being conducted in concept cars. For automotive Kalray has a strategic shareholder, car makers Alliance Renault-Nissan-Mitsubishi, and a strategic partnership with NXP.

The MPPA design allows the number of cores to scale exceptionally well. The first and second generation MPPA samples went to early customers for evaluation, and the latest MPPA third generation, named Coolidge, has been available in samples since early 2020 and will go into mass production at the end of 2020. It increases the performance of the previous chip by factor of 25, it has dropped the price by half, and it supports high-speed interfaces (PCIe 4, 200 Gigabit Ethernet). The Coolidge MPPA is available on a Kalray open and intelligent card (KONIC) which is a programmable and reconfigurable PCIe card. Users have access to a comprehensive software stack named AccessCore.

For the automotive market Kalray has formed a strategic alliance with NXP, the world’s premier automotive semiconductor manufacturer. For the Data Center market, an important client is Wistron, an original design manufacturer (and the manufacturing arm of Acer), which is integrating KONIC in its next storage server and supporting NVMe-TCP. Another client is 2CRSI, a French supplier of appliances to data centers, enabling rapidly composable modular solutions for a variety of tasks, including AI.

Coolidge has been designed to process big data on the fly, for data centers applications such as storage and smart NIC. Kalray has reworked its network-on-chip technology to improve the memory-compute communications in its MPPA design. The ease of programmability has also been improved, so software developers can work with familiar tools (TensorFlow and Caffe, and on the roadmap is ONNX support) and not be concerned with the MPPA itself, which is optimized by the Kalray MPPA compiler. Legacy software can be run on the MPPA, for example Linux, allowing a host of Linux tools to be run. The standard convolutional neural networks (CNNs) can be run, so models developed for the GPU or FPGA can be ported to the MPPA.

Figure 20: Kalray MPPA architecture



Source: Kalray

For edge scenarios the MPPA is a replacement for GPUs and FPGAs where power consumption or lock-in to proprietary software are issues. In the context of automotive applications, Coolidge becomes a safe alternative for running multiple types of environments on one device, so certification and security only needs to be assessed once. Automotive certification such as AEC-Q100/QM for Coolidge1 and for the next generation Coolidge2 ASIL B / ISO 26262, is a long and complex process making the barrier to entry high, but for vendors that succeed the rewards are high.

When working with alternative technologies manufacturers must deal with multiple chips: CPU, GPU, FPGA, etc, and they need to isolate the compute capabilities to ensure freedom from interference, i.e. if one component fails it needs to be ensured that the impact is contained and not affecting other parts of the system. Today this is commonly achieved with software virtualization (hypervisors) but this is complex to program and complex to maintain. In contrast the MPPA has by hardware design the capability to isolate computing. For the auto-industry it means users do not need to run a software virtualization layer which makes it less complex and easier to certify the system.

In addition, MPPA is a platform that can support heterogenous computing acceleration, either involving multiple stages of an end-to-end use case (i.e. pre-processing, DL-based processing, post-processing), or in the context of multiple use cases support on the same platform.

The Coolidge architecture shown in Figure 20, has five compute units, called clusters, and in turn each of these contains 16 very long instruction word (VLIW) cores and 16 co-processors designed for rapid CNN processing. Each of these clusters has 4MB of SRAM, so in total an MPPA has 20MB of SRAM. And there are two DDR external memories for relaying data into MPPA.

VLIW processors allow program instructions to execute in parallel, and in total there are 80 VLIW cores and 80 co-processor accelerators. The compute clusters are isolated from each other and connected by the network-on-chip, allowing different applications to be off-loaded from a host and run on different clusters. So different types of CNN models can be run in parallel. There is parallel programing support for OpenCV and OpenCL and it is possible to chain applications, so that computer vision with CNN is chained to a post-processing application.

Another option is not to use an external host to off-load applications to the MPPA but to use it as a stand-alone solution. So, one of the five cluster can be dedicated to run Linux and then other cores can be used for acceleration or to run other applications altogether. As a CPU, a MPPA core is comparable with an ARM Cortex A53.

The Coolidge device consumes power in the 5-30 Watt range, and this translates to 200 inferences per Watt in an AI application. The version Coolidge2 is being certified for the auto-industry and the one to be deployed there. Coolidge2 adds improvements in the compiler optimization leading to greater performance for the given power consumption.

Kisaco Assessment

Strengths

- Kalray has built an exceptional many core processor designed for real-time AI inference and heterogenous workloads which make it suitable for automotive and Edge Computing applications. The Coolidge architecture is in third generation, with low latencies achieved with technologies such as on-chip networking and memories, making it suitable for real-time applications and low power scenarios.
- To prevent runaway failure mode, for example important in safety critical systems, it is desirable to isolate workloads so that failure in one does permeate other functionality. Kalray enables freedom of interference through hardware-based isolation. This isolation also allows multiple operating systems to run on the same MPPA chip, without software hypervisor, and also allows the chip to run without needing an external CPU in a stand-alone mode.
- The MPPA has deterministic computation, allowing processing time to be predictable, an important requirement for the automotive industry. The alliance with NXP Semiconductors and Renault-Nissan-Mitsubishi is a strong validation of Kalray's suitability for the automobile industry.

Weaknesses

- Kalray works with partners in its customer relationships but we believe direct customer support in terms of use cases, and good practices materials could help open up potential new applications and greater awareness of the potential for Kalray MPPA.
- We encourage Kalray to participate in MLPerf benchmarks. We believe this helps the AI community by having an independent assessment of AI chips.
- Kalray has strong features for small edge and automotive AI inferencing and its position has improved with a move to the right compared to its position in the DC/HPC KLC. However, the inferencing market is also more competitive, and Kalray is still on the brink of improving its execution in the market.

Kneron, Kisaco evaluation: Contender

Product: Kneron, neural processing unit (NPU).

Kneron was founded in 2015 by CEO Albert Liu and Frank Chang, and is based in the Greater San Diego area, CA, with offices in Shenzhen, China and Taipei, Taiwan. The company mission is to build ultra-efficient NPU hardware and lightweight CNN algorithms to empower AI capability for edge devices. To date the company has raised a total $73m, the latest series A2 round for $40m was in Jan 2020, and investors include Alibaba, Qualcomm, and Sequoia. The company is targeting the AI inference market, which is expected to grow at a faster rate than the AI training chipset market, with demand for AI on the edge expected to grow the fastest.

The challenge for Kneron is that the edge AI inference market is the most contested in the overall AI hardware market. The edge AI is also quite distinct from cloud AI characteristics:

- There is a need for greater privacy protection.
- Real-time (RT) or near RT response requires low latency computation.
- Bandwidth constraints limit connectivity with the cloud.
- There is a need for low power consumption.

Furthermore, billions of edge sensors create a vast mountain of data and moving this data from the edge to the cloud for processing by intelligent systems is inefficient, as well as limited by bandwidth and being unsuitable for real-time responses.

Kneron’s answer to these challenges is to build an edge AI network or mesh, where computing elements on the edge can process the large amount of data generated by sharing computing loads across the mesh. The mesh can also act with proactive intelligence, where elements in one locality can forewarn other elements moving into a zone, for example where an accident or hazard has occurred.

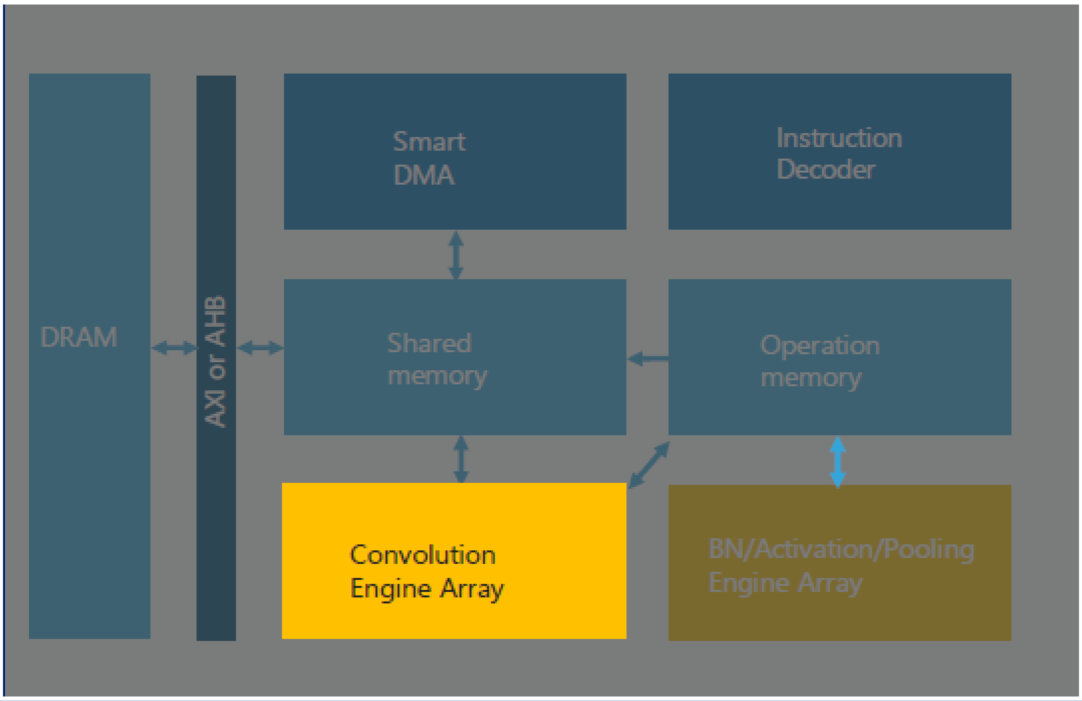
In practice, both types of intelligent systems are needed, centralized on the cloud, and localized at the edge. By suitable balancing of system control between cloud and edge, only necessary data needs to be moved. For example, to protect privacy, the edge mesh can remove details before sending information to the cloud.

The Kneron NPU ASIC, shown in Figure 21, has a reconfigurable architecture, with a basic building block of 8-bit MACs that can be combined into 16-bit MACs as required. There is SRAM near to the computing cores and DRAM connected via AXI or AHB interface. Kneron has built turn-key solutions for the most popular AI edge use cases:

- Facial recognition.
- Object recognition.
- Body/gesture recognition.
- 3D sensing.
- Defect detection.
- Prediction maintenance.

Kneron provides easy to use software to develop on the NPU for the CPU or DSP toolchain. The NPU supports popular AI software frameworks TensorFlow, Keras, Caffe, PyTorch, and Darknet, and the ONNX standard. Example CNN models run on the NPU include Vgg16, Resnet, GoogleNet, YOLOv2v3, Tiny YOLOv2v3, Lenet, MobileNet, Unet, and Densenet. Kneron optimizes the neural

Figure 21: Kneron NPU architecture



Source: Kneron

networks using its compression techniques, which maintains accuracy while reducing model size (and hence storage) and computation cost. Its patented compression technology uses dynamic fixed-point quantization to convert floating point to fixed point logic with less than 0.5% accuracy loss.

Kneron has built to date two large scale commercialization AI SOCs, the KL520 and its second-generation version:

- **SOC KL520**: This SOC contains the KDP520 NPU, with an ARM Cortex M4 for system control and a second M4 acting as AI co-processor. The SOC can contain 32MB or 64MB of LPDDR2, is rated at 576MACS/cycle, supports a host on interconnects, consumes around 0.5W and comes into package sizes: BGA 8x8mm and LQFP-128 14x14mm.
- **Second generation version of KL520**: This device is similar to the KL520 in configuration, is rated at 1024MACS/cycle, swaps in a DSP for the AI co-processor, has 128MB of LPDDR3 memory, consumes around 1.5W, and is packaged in Companion mode BGA 9x9mm, and Host mode 11x11mm. This SOC is projected to be sampled in Q3 2020.

Customers include smart air-conditioning from Gree, smart lock for Datang Semiconductor, the Bank of Taiwan, and industrial applications at AAEON and Acer. Kneron demonstrated how its technology for face recognition, which includes liveliness recognition (facial recognition security algorithm), that will not be tricked a high quality 3D static mask, whereas in officially approved tests, wearing the same mask gave access at a WeChat payment terminal and at China’s high-speed railway station gates.

Kisaco Assessment

Strengths

- The Kneron NPU offers high compute performance at low power making it suitable for a range of edge inference applications. Edge inferencing is a highly competitive market and Kneron has sought to differentiate itself by allowing its AI inference chips to combine in a mesh network (called vision-edge AI net) and leverage the network to make better informed decisions, for example in smart city applications.
- Kneron has developed algorithms to run on its NPU that excel in tasks such as face recognition. For example, where other systems are foiled by masks, Kneron is able to identify a mask due to its static facial impression that lacks giveaway muscle movements.
- The NPU has a full stack of software to support software development using popular DL frameworks. As is common with many AI hardware accelerator startups, the origins of the company are in the embedded hardware world and engaging with more traditional software development is a cultural barrier to overcome. Beyond the full software stack Kneron offers a developer section on its web site and a forum, which is more than we have seen across other startups.

Weaknesses

- Kneron has not participated in independent benchmarks such as MLPerf and we encourage it to publish these benchmarks.
- The NPU is not programmable but rather offers fixed functions with some configurability for typical DL workloads. However, the reconfigurable architecture allows the compiler to program together basic building blocks (like Lego), which are the optimally NPU programmed blocks.
- Kneron NPU, as is common with many edge inference AI accelerators, is a chip designed for maximizing TOPS and not FLOPS (i.e. there is no floating-point support). Kneron does point out that its fixed-point design is best for edge AI and AIoT for low power and power efficiency. Its INT8 solution has almost no performance loss compared with floating point solutions. Furthermore, Kneron says its second-generation chip has a 16-bit fixed point design which is close to floating point.

Mythic, Kisaco evaluation: Innovator

Product: Mythic Intelligent Processing Unit (IPU).

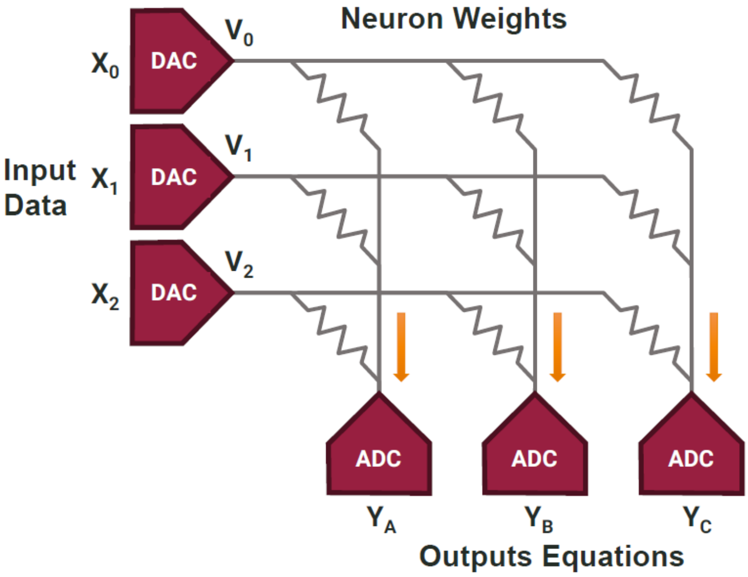
Mythic was founded by CEO Mike Henry and CTO Dave Fick in 2012 with current offices in Redwood City CA, and Austin TX. The company has 100+ employees, and a lineup of 11 investors, including Threshold Ventures, Lux Capital, DCVC (Data Collective), SoftBank Ventures Asia, and Valor Equity Partners. Mythic is currently in a pre-commercial phase,

While AI data center training and inference is set to grow over the next decade, proportionately edge computing inference is expected to grow faster, especially with 5G rollout and the continued pace of digital transformation. The edge inference market is therefore a key focus for many players in the AI hardware accelerator space.

Mythic's technology is based on three advanced areas of computing: compute in-memory, dataflow architecture, and analog computing. As an in-memory processor, the processing is also deterministic, with the advantages of being able to ascertain how long a computation will take.

Mythic stands apart from the bulk of competitors in being an in-memory processing AI solution provider, where most others offer near memory SRAM and it does this using innovative analog technology based on a CMOS process using non-volatile Flash memory, where the Flash transistors act as variable resistors and equivalent to computer memory. These embedded Flash elements store weights in non-standard ways, as analog values. Looking at Figure 22, a voltage input is applied to these elements and summation is performed across the columns providing the outputs, making the matrix MAC operations possible, as required in deep learning.

Figure 22: Mythic exploits embedded Flash transistors as variable resistors to hold weights



Source: Mythic

The Mythic device is used for AI inference, so the weights are now locked in value and do not change once set. This technology is ideal for inference, whereas in training the weights need to be updated often and Flash technology is not ideal for fast weight changes, so Mythic has not explored this AI mode.

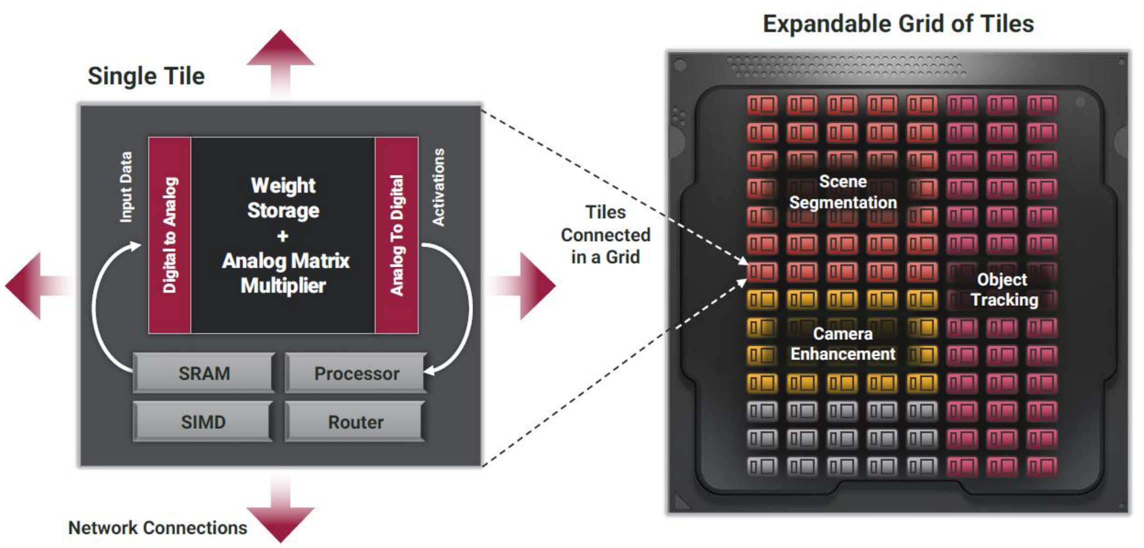
The first device is built with a 1k x 1k array of analog memory elements. A key IP technology are the DAC and ADC convertors, which Mythic has patented. The company can mass produce these in the thousands and operate these with high efficiency, high accuracy, and low power, making the whole concept of analog computation feasible.

The DAC runs voltages across the inputs and by the laws of electric currents in resistor networks, linear superposition takes place, performing MAC operations. The voltages appearing at the outputs are then digitized by the ADC units. When voltages appear at the inputs, the voltage change at the outputs occur at the speed of electromagnetic waves in the chip wire channels, which varies from 50% to 90% the speed of light – for all practical purposes it is instantaneous. The Flash components are standard off-the-shelf, the innovation required to make this work was in mass producing efficient DACs/ADCs; latest Mythic chips will typically carry some 27k of these. This is in stark contrast with typical SOCs that will carry about 6 of these convertors.

The matrix math engine is implemented in tiles, see Figure 23, with each tile also possessing a RISC processor, SRAM memory, a digital SIMD vector processor, and a router, and then multiple tiles fill the chip space; depending on application a chip could be built with 100s of tiles. The RISC-V nano-processor is royalty free and is 1000 times smaller than a micro-sized processor, performing limited control tasks.

The current chip, Mythic IPU, has 108 tiles built in. Fabrication is in a standard 40nm process, with consequent improvements in specifications as 28nm and 14nm chips are built on the roadmap. The architecture is dataflow-based so tiles run applications independently, and this allows multiple models to run on a single chip, for example performing tasks such as scene segmentation, object tracking, and camera enhancement as shown in Figure 23. Another use case scenario is to run the same job in parallel, with different regions allocated to each model instance. The allocation of work to the chip tiles is performed at compile time.

Figure 23: Mythic DNN chip with deep learning neural network tiled architecture



Source: Mythic

The advantage of the analog approach is that the weights are highly densely packed in the tiles. Mythic talks about weight capacity rather than GB of memory allocated to weights, so one eight-bit weight is equivalent to one Flash transistor, and with 1k x 1k Flash array per tile, holding one mega weight per tile, means 108 tiles, one current Mythic chip, can hold 113M+ million weights. Unlike other AI chip accelerators, the near-memory SRAM is not used for holding weights but for holding intermediate computations, data that is held before being presented at the next tile input. Because the Mythic weights never move there is greater efficiency in this design compared with weights that are stored in near on-chip SRAM or worse, in off-chip DRAM.

Applications require the calculations to be quantized in INT4, INT8, or INT16 with no support for floating point calculations. However, off-chip layers are supported, and they could run in floating-point logic.

The Mythic IPU is available on a PCIe accelerator card and works with a host CPU or SOC. There is no DRAM required. For extra-large tasks, multiple Mythic chips can be built onto a single PCIe card. The PCIe card is targeted at high-value real-time edge inference computing requiring high performance, low latency, and low power consumption, such as industrial machine vision, smart city, video surveillance, drone and other aerospace applications, automotive applications, and installation in hyperscaler data centers on the cloud edge where a card with 16 Mythic chips could process very large models with 1.8 billion weights. A four chip PCIe card is designed for on-premises servers.

The power range for its device is in the <5Watt range, with efficiency of 8+ TOPS/W, although Mythic says theoretically this could be 10 times higher but would cost more to build. Even with the 1.8 billion weights card the power consumption would be in the 100Watt range.

Mythic provides a full stack software suite, supporting TensorFlow, Keras, PyTorch, Caffe, and ONNX support. The SDK includes profiling and logging, and the graph compiler. The optimization suite has post-training quantization and retraining libraries. Although the compute technology is analog, it provides an interface similar to digital AI chips and therefore does not require any changes to standard AI software.

Kisaco Assessment

Strengths

- The Mythic accelerator architecture has three significant innovations: analog computing, in-memory computation, and dataflow process. Each of these adds considerable performance boosts making this processor a rare product in the market. The use of analog computing is novel in the market, with a couple of potential near-competitors currently in stealth and optics-based analog computation also in early stage, Mythic has a lead in this type of general purpose analog computation engine for AI inferencing.
- The Mythic tiled architecture allows the compiler to allocate different areas to be dedicated to specific workloads, making the chip highly efficient. A PCIe switch can be used to cascade work across multiple IPUs.
- The innovation required to make the analog, in-memory computation possible is surprisingly as much to do with the conversion of digital signals to analog (and the reverse), and being able to fabricate these efficiently and in large numbers (over 20k) on the chip, while staying within a low power budget, gives Mythic a lead on potential imitators.

Weaknesses

- The company is due to launch its products in 2020 but possibly due to the pandemic we have had no further news about a release date. Our assessment is based on early samples and anticipation of market launch, so we expect to raise Mythic’s market execution position.
- Without a product yet on the market there is clearly no opportunity yet to provide an independent benchmark, we encourage Mythic (as we do all vendors) to publish on MLPerf or similar standard.
- The company has little material for direct customer education or support such as case studies or business starter materials. We expect this will improve as the company grows.

Nvidia, Kisaco evaluation: Leader

Note: Below we repeat a profile on Nvidia GA 100, a product built on the Ampere A100 GPU which we published in the DC/HPC part 2 report. For edge inferencing we evaluated the anticipated features of Nvidia EGX A100, an edge inferencing GPU also based on A100 and expected at the end of 2020. At the end of the profile we have added material relevant for the new Nvidia EGX A100, Nvidia EGX Jetson Xavier NX, and Nvidia Drive AGX Orin.

Product: Nvidia GA100, Ampere architecture.

Nvidia, founded in 1993 and based in Silicon Valley, is today the most prominent supplier of compute acceleration for AI and more: HPC (five of the top 10 supercomputers as listed on Top500 use Nvidia technology), ray tracing, and computer gaming graphics. The company under leadership of CEO Jensen Huang has taken its GPUs into the AI hardware acceleration space since deep learning algorithms were first ported to its general-purpose computing GPUs (GPGPUs) back in 2011. The introduction of GPGPUs was in turn a response to the saturation of Moore's Law for CPUs and the timing was perfect. The company rapidly became the first stop for AI acceleration, and its top end GPU architectures have continually evolved, with new generations appearing at a steady twice-yearly pace, raising the performance bar each time.

The success of Nvidia (revenue is in the $10-11b range) has also ushered in a new era of startups aiming to compete with GPUs for AI hardware acceleration, and against the greater competition Nvidia continues to innovate: the latest architecture, Ampere, was announced in May 2020, together with details of the GA100 GPU. The first implementation device, A100 Tensor Core GPU, based on GA100, is immediately available. It has 108 streaming multi-processors (SMs) to the 128 in GA100.

Nvidia has been active in acquisitions in the last two years, prior to that was PortalPlayer in 2006. The recent acquisitions were Cumulus Networks, a networking OS player, acquired in May 2020, the Mellanox (high speed data center networking) acquisition for $7b completed in April 2020, SwiftStack for data storage and management in Mar 2020, and (an outlier to the pattern) Parabricks for genomic analysis in Dec 2019. The main pattern in these recent acquisitions is the data center and recognition that east-west traffic dominates over north-south traffic. Mellanox technology is evident in the networking available in the Nvidia DGX A100 appliance, built for heavy AI workloads.

The Nvidia GA100 GPU is designed for data center AI and HPC applications. The device is built on the latest 7nm TSMC fabrication process, a graphic of the architecture is shown in Figure 24 with parts of the structure removed to provide legibility. The NVLink for in-chip and chip-to-CPU networking is in third generation.

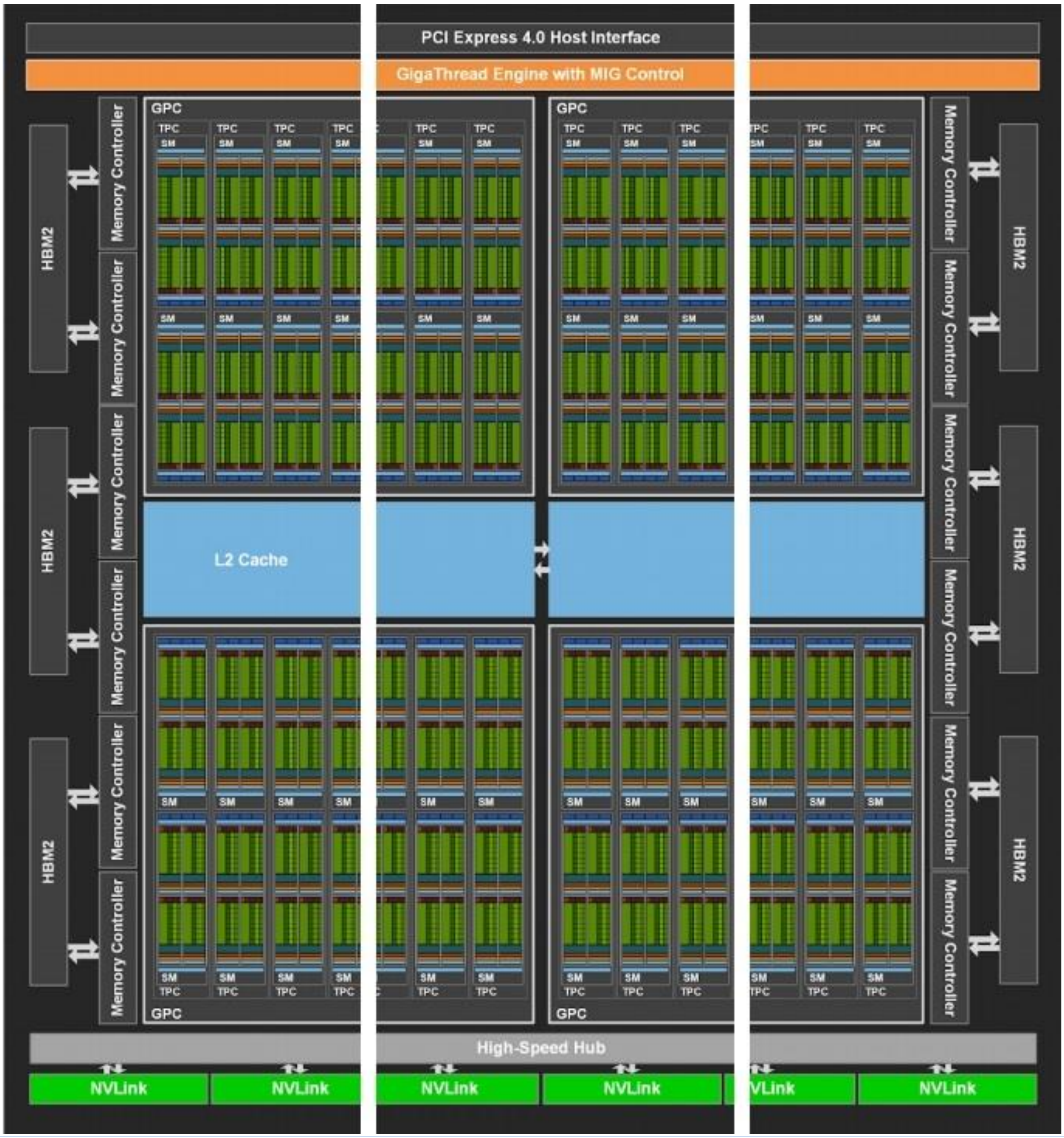
An important innovation in Ampere is the multi-instance GPU (MIG) which manages the partitioning of the GPU in up to seven independent GPU instances, making it possible to run one large application across the whole GPU or run seven independent applications, and variations in-between.

The SM is shown in detail in Figure 25, it is where CUDA kernels are run. The SM contains multiple registers that execute application threads, there is shared SRAM memory for fast data exchange, various precision cores, and the tensor core is designed for accelerating mixed-precision matrix multiply and accumulate (MAC) operations.

Another innovation in Ampere is the introduction of the new math precision TensorFloat-32 (TF32), which speeds up single-precision work used in AI, while maintaining accuracy and using no new code. TF32 runs in the GA100 SM tensor cores speeding up operations by up to 10x compared with single precision FP32. TF32 sits between FP16 and FP32, so it uses the 10-bit mantissa of FP16 and the 8-

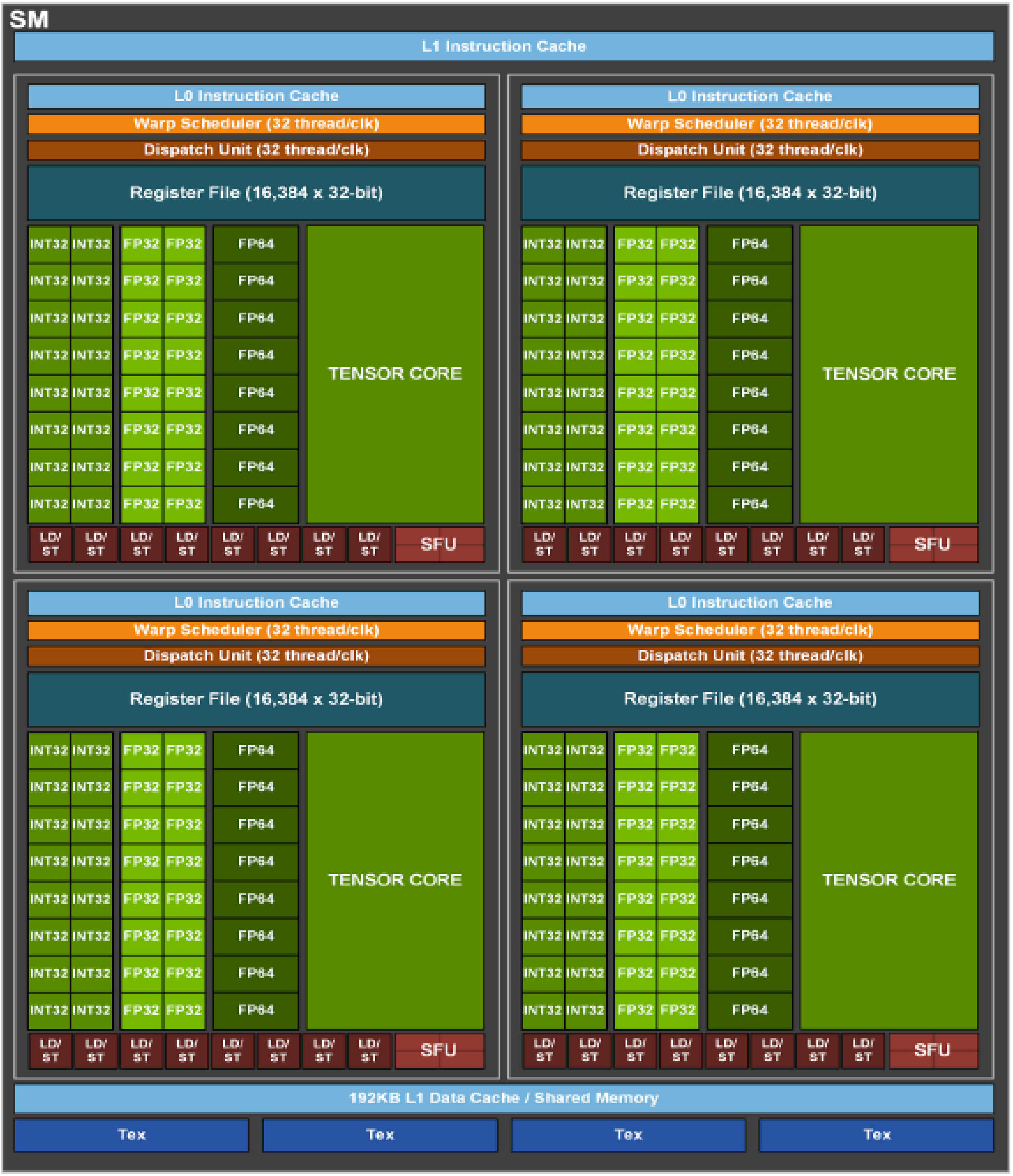
bit exponent of FP32, making a total of 18 bits plus one more bit for sign). TF32 is a name not a description of the number of bits used, the point is that it offers a balance of FP32 performance and the economy of FP16.

Figure 24: GA100 GPU with 128 SMs – with strips removed to provide legibility.


PCI Express 4.0 Host Interface
GigaThread Engine with MIG Control
HBM2
Memory Controller
GPC
TPC
SM
L2 Cache
High-Speed Hub
NVLink

Source: Nvidia

Figure 25: Internals of a GA100 SM.

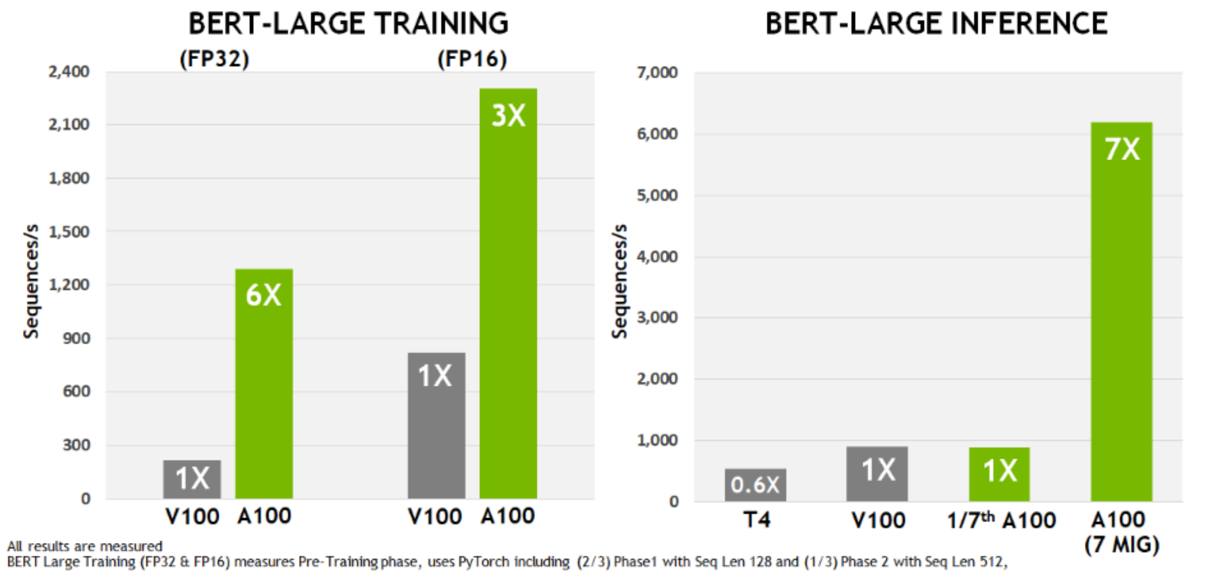

SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
TENSOR CORE
LD/ ST
SFU
192KB L1 Data Cache / Shared Memory
Tex

Source: Nvidia

A final Ampere architecture innovation covered here is the introduction of sparsity and pruning techniques for neural networks in inference mode. The new algorithm prunes two out of four non-zero weights. Weights are further compressed for a reduction in data footprint allowing the tensor cores to effectively double the processing. Zero weights are skipped by the sparsity reduction. The performance can be seen in the NLP BERT benchmarks that Nvidia released with the Ampere announcements in Figure 26. The tensor cores in training use TF32. In inference mode the A100 using all seven MIG partitions provides a x7 performance improvement over previous Nvidia generation Volta V100.

Performance figures for the A100 Tensor Core GPU with sparsity/pruning on are 312 TFLOPS in TF32 precision and 1248 TOPS in INT8. There are numerous other improvements, combined they have raised the capability of Ampere to new levels above the previous Volta GPU.

Figure 26: A100 GPU performance in BERT deep learning training and inference modes comparing Tesla V100 and Tesla T4.



All results are measured
BERT Large Training (FP32 & FP16) measures Pre-Training phase, uses PyTorch including (2/3) Phase1 with Seq Len 128 and (1/3) Phase 2 with Seq Len 512, V100 is DGX1 Server with 8xV100, A100 is DGX A100 Server with 8xA100, A100 uses TF32 Tensor Core for FP32 training
BERT Large Inference uses TRT 7.1 for T4/V100, with INT8/FP16 at batch size 256. Pre-production TRT for A100, uses batch size 94 and INT8 with sparsity

Source: Nvidia

Nvidia has also enhanced its Rapids suite of open source software libraries with an accelerator for Apache Spark 3.0 for GPU-accelerator data analytics. Apache Spark 3.0 is the first fully integrated release with GPU acceleration, for on-premises or on the cloud applications.

Products: Nvidia EGX A100, Nvidia EGX Jetson Xavier NX, and Nvidia Drive AGX Orin, all based on the new Ampere architecture.

NVIDIA EGX A100 is a single board GPU bult with the latest Ampere A100 architecture and NVIDIA Mellanox ConnectX-6 Dx SmartNIC. Nvidia has enabled its GPUs on Kubernetes which allows Kubernetes to manage EGX clusters. This means GPU accelerator nodes can be managed just like CPU nodes. EGX can deployed on Kubernetes (x86 and ARM) or with Red Hat OpenShift (x86). This device has enhanced security features with a new security engine and secure, authenticated boot. In addition, the ConnectX-6 has L4 firewall, TLS crypto engine, IPsec crypto engine, and hardware root of trust, all performed on the GPU board and reducing latencies.

The Jetson Xavier NX is a small (credit card sized) edge inferencing GPU, offering 21 TOPS at INT8 and 15W, and 14 TOPS at 10W. It can support data streaming from up to 32 1080p IP cameras and supports cloud native technology such as Kubernetes. Nvidia's eco-system partners already have 20 different products on the market implementing this GPU, for AIOT, retail, industrial computers, and IOT gateways.

Nvidia Drive AGX Orin provides an embedded supercomputing platform processing data from camera, radar, and lidar sensors for autonomous driving. Nvidia says most car manufacturers are testing AV systems using Nvidia Drive technology and choose either to use all the capabilities including Nvidia supplied DL algorithms or may mix and match with their in-house algorithms. Nvidia has built Orin to satisfy ISO 26262 ASIL-D.

Kisaco Assessment

Strengths

- The latest Nvidia Ampere architecture series of GPUs place Nvidia at the top of our ranking in the KLC AI hardware accelerators for DC/HPC AI training. The computer box DGX A100 offers excellent value for AI training mode and forms a formidable challenge to the startups competing in this space. Nvidia's presence in the market is also a formidable barrier to entry: the reality at time of writing is that the first choice for training acceleration is a Nvidia GPU.
- Multi-GPU instances (one to seven) in a single A100 GPU, provide greater versatility for varying workloads across training and inference scenarios. Mixed precision and Nvidia's own TensorFloat-32 precision offer a good balance for improving processing speed and precision.
- The inclusion of features to address and exploit sparsity in DL models puts the Ampere architecture at the forefront of DL techniques.
- The Rapids software suite has been updated to support Spark 3.0 directly on the A100 GPU without the code needing to go to the CPU first, improving the processing speed.

Weaknesses

- The Nvidia vision is being an AI computing company. To date this means running AI applications on GPUs. As the startup field moves beyond first chips, we believe Nvidia will acquire a startup to consolidate its position. While GPUs have proved fortuitously useful for DL training, we expect startup architectures, designed from the ground up to accelerate DL, will pose a threat to the GPU.
- Nvidia is the first choice for training DL applications in the DC and for HPC. The company has said it does not intend to compete in the edge/inference mass market but has given its Nvidia DL accelerator (NVDLA) IP to the open source hardware community. To date there is no evidence of any take up and Nvidia remains weak for a whole class of small edge inference applications, as this market is the most competitive across startups.
- While leading players in the high-tech space with a deep interest in AI are researching (and publishing papers) on a range of alternate hardware architectures, from spiking neural networks to analog technologies, we do not hear of any such research within Nvidia. At this point in the company's maturity we believe it should reveal something about its AI hardware research if it exists or create such a research arm.

Syntiant, Kisaco evaluation: Contender

Product: Syntiant NDP100, Audio Neural Decision Processor.

Syntiant was founded in 2017 by four experienced professionals in the high-tech industry: CEO Kurt Busch, COO Pieter Vorenkemp, CTO Dr Stephen Bailey, and Chief Scientist Prof Jeremy Holleman. The company mission is to create a new type of processor for machine learning, targeting deep learning inference mode for the edge, what it calls “intelligence of things”, as an IOT version 2. Syntiant identifies four advantages of bringing processing to the edge:

- **Privacy**: Processing at the edge means less data is sent to the cloud.
- **Reliability**: With less reliance on a cloud connection, the edge device will be more reliable.
- **Responsiveness**: With processing performed locally the latency is reduced.
- **Battery life**: Low powered local processing will prolong battery life.

The company currently has 60 employees, about a third are hardware specialists and the rest are ML experts and has raised $30m to date in funding (Intel led A-round and Microsoft led B-round). It produced its first processor sample in July 2018 and first production orders went out in Sep 2019. The first product line is aimed at intelligent voice. Current targeted applications are for sensors in event detection, wake words for intelligent command driven devices, and on the roadmap are speaker identification and conversational speech.

Syntiant’s team of software and hardware developers sit together to learn from each other. The Syntiant design of high efficiency overcomes limitations of existing deep learning inference processors:

- **Memory and compute**: Traditional Deep Learning (DL) processors are limited by memory bandwidth and large amounts of data movement which increase costs. Syntiant’s near memory architecture reduces these bottlenecks.
- **Deterministic compute**: While parallel processing is now the essential mush have for DL, the use of caches and far memory leads to non-deterministic processing times, whereas Syntiant’s design allows for deterministic execution, important where the compute latency needs to be known in advance.
- **Reduced precision**: With DL it is possible to achieve the same level of accuracy while reducing precision below 8-bit, with benefits in efficiency.

While the I/O of the NDP100 is designed for voice applications, the core DL engine is application agnostic and other potential use cases are possible on the roadmap.

The Syntiant NDP100 chip has performance characteristics of 3.4 micro Joules per inference, with a maximum of 100 frames per sec. In terms of energy per decision the Syntiant device is highly favorable even against high powered and high-performance chips on the market, but the workloads are radically different comparing data center high power scenarios and inference on the edge.

Syntiant tackles the challenge of bringing DL projects into production, and the barrier is often the lack of a process to deploy applications, which is why Syntiant places great value in the software stack to support the hardware. Syntiant talks of the three pillars of deep learning products:

- **Silicon**: The advantage of NDP100 in terms of power, performance, and cost metrics.
- **Pipeline**: Full production DL pipeline. The Syntiant software stack (in Python) takes in raw data, models it, validates it and deploys the solution onto the device. This lifecycle approach reduces the time from development to production.
- **Data advantage**: Syntiant has collected millions of utterances in an anonymized database it provides its clients for helping to train new models. Furthermore,

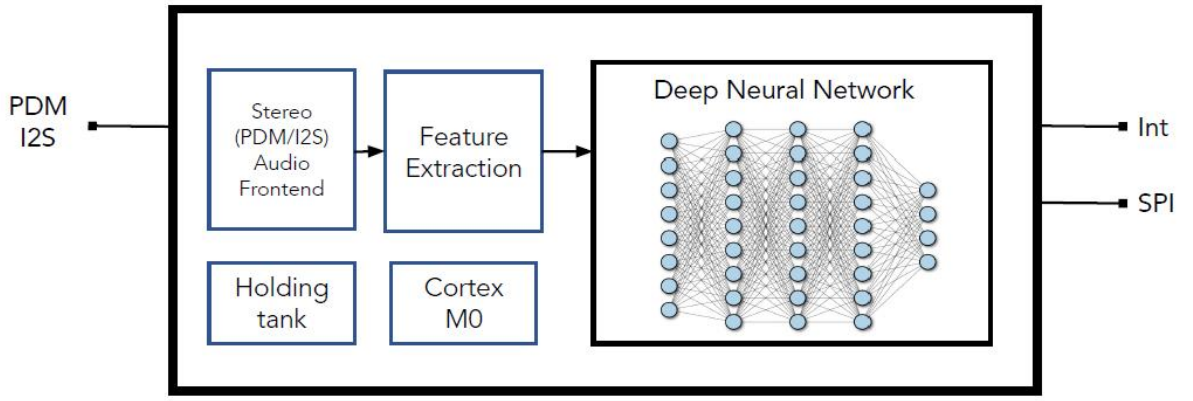
deployed devices capture data to add to the client's version of the database, further enhancing the data collection.

Syntiant can read TensorFlow models directly without needing an intermediate middle layer such as ONNX. Syntiant has customers using other modeling tools such as Caffe and Keras (which can sit on top of TensorFlow).This approach simplifies how the TensorFlow dataflow model is mapped to the silicon, which is often done by a compiler, whereas Syntiant loads the weights directly onto the device. The pipeline approach cuts the time spent developing a new model; one Syntiant customer replaced years of development with three months' work using Syntiant.

Performance is essentially a trade-off between false accept and false reject (false reject = 1 – correct accept), which can be plotted in a ROC graph, and the more data available the sharper the curve, hence the advantage of the Syntiant utterances database.

The Syntiant NDP100 contains a core processing device that is only 1.4mm x 1.8mm (WLBGA12 chip) and adds a stereo front-end (PDM/I2S), an ARM-Cortex M0, a feature extractor, and a holding data (external memory). The device is designed for low-speed tasks such as wake word detection. It contains a fully connected deep neural network with 5 layers, comprising 1600, 256, 256, and 64 neurons – see Figure 27. The always on power consumption is 140 micro Watts.

Figure 27: Syntiant NDP100 architecture



Source: Syntiant

The device uses SRAMs to store the weights, which is data that does not move. The MACs are tightly coupled to the SRAMs and uses a dataflow architecture. The device is built on ultra-low power 14 LPU (14 nm) fabrication process, which is highly available and relatively cheap for mass market manufacturing.

Syntiant goes to market in two ways, a device with pre-configured DL net, i.e. trained with the keywords the customers requested. The second way is to provide a TensorFlow model of the device and have the customer perform the training. Most customers don't have the expertise to perform their own training, so the pre-configured approach is more popular. The barrier to entry in the edge voice market is having the data to perform the training which is Syntiant's strong point. Syntiant can also provide partially trained DL on devices so that the customer can train re-train the last layer.

Syntiant has commercial clients in products such as mobile phones, earbuds, smart speakers, laptops, remote controls, and hearing aids. The company is clearly in its early days, the current device manages dozens of words, and increasing the capability all the way to full conversational speech is on the roadmap for future Syntiant processors.

Kisaco Assessment

Strengths

- Syntiant has set out to make a small sized AI inference accelerator for the mass market with ultra-low power (less than 150 microWatts) usage and at low cost: in this mission it has eminently succeeded. The chip is available at an exceptionally low cost for million-unit volumes. Syntiant has NDP chips in production and client contracts in place; the company expects revenue of $m's in 2020.
- The initial focus is for audio applications, such as voice applications. Syntiant has collected a large audio database (millions of utterances for wake words and commands) for developing DL voice applications and this is an important asset for engaging with clients.
- There is a full software stack available for developers to build and deploy applications. The hardware and software engineers who developed the chip worked together and the attention to providing a full software stack was part of the company mission. Software developers can use popular DL frameworks such as TensorFlow, PyTorch and more.

Weaknesses

- In terms of AI hardware acceleration performance irrespective of power and size then Syntiant NDP falls short compared with other products on the market, however as an ultra-low-power device it excels and within this category it is more than a Contender: this needs to be understood, that our positioning of Syntiant is within the broader market for AI accelerators.
- The type of neural networks supported currently on the NDP are geared towards audio, so convolutional neurons, which are used in vision DL applications, are not supported.
- We encourage Syntiant to submit its devices for independent benchmarking such as MLPerf, and we also appeal to MLPerf to provide metrics that are relevant for ultra-low-power devices, such as TOPS/Watt, which is more relevant for this category AI chip.

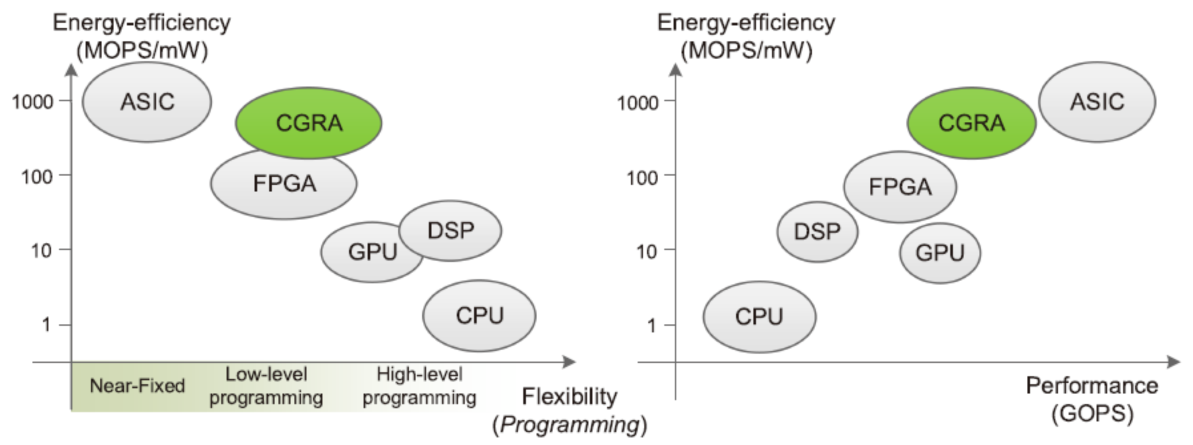
Tsingmicro, Kisaco evaluation: Contender

Product: Tsingmicro, TX210 and TX510 high energy efficient reconfigurable processors.

Tsingmicro is a Beijing-based startup spin-off from the Institute of Microelectronics (IME), Tsinghua University. The team includes experts from IME and senior management from major companies including from the chip industry. The company's origins are in reconfigurable architecture technology that won National Technology Invention in 2015, which led to the development at IME of Thinker-I in 2016 (TSMC 65nm fabrication, with 5 TOPS/W), Thinker-II in 2017 (TSMC 28 nm fabrication with 90 TOPS/W) and Thinker-S also in 2017 (consuming 1mW). Tsingmicro was then founded in 2018 and in mid-2019 the TX210 MP processor was launched. The team continued to win awards for their processors throughout this history.

The technology implemented can be described as coarse-grained reconfigurable architecture (CGRA). Figure 28 shows where CGRA sits within the spectrum of architectural types. The ultimate aim in architecture design is to achieve ASIC levels of performance with what is normally a trade-off: programmability, or flexibility. The CGRA provides an alternate option with attractive attributes of high performance, programmability, and low power consumption. In the Tsingmicro implementation of CGRA the compiler dynamically configures the processing elements (PEs) that carry out a set of functions and orchestrates the data flows between PEs. Data resides in near memory SRAM, located around the block of PEs – see Figure 29.

Figure 28: CGRA in the context of other processor architecture types.



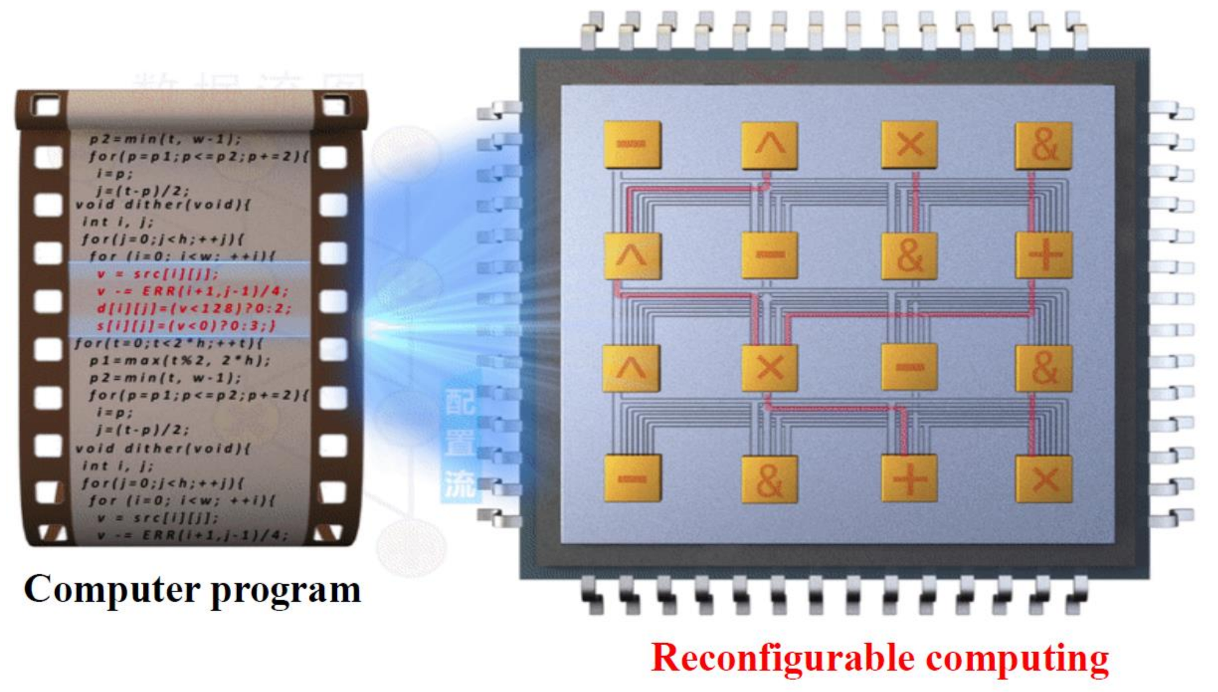
Source: A survey of course-grained reconfigurable architecture and design. L Liu et al. doi.org/10.1145/3357375

Thinker-I was the first chip produced for general-purpose neural networks that could implement multiple architectural network types and consumed power in the range 4mW to 447mW with a maximum of 5.09 TOPS/W. The current generation of processors are the TX210 for voice AI applications such as wake up trigger and voice command recognition, and the TX510 multi-modality based intelligent computing chip that supports speech and vision intelligent processing.

The TX210 is small and highly energy efficient hence suitable for products such as wireless earbuds, smart headphones, smart watches and more. Applications include voice AI for wake words and commands. It contains 128 PEs. Tsingmicro will ship over 10m units in 2020.

The TX510 is designed for more complex applications ranging from piloting drones, smart payment, intelligent security, smart factory and more, and has also been mass-produced. It contains 1000 PEs. Chip shipments are expected to be more than 1m in 2020. These processors can be combined in arrays to tackle more challenging applications, and these multi-array chips can be combined with external chips to create a larger processing system.

Figure 29: Tsingmicro: the compiler configures the PE functions (orange squares) and data flow in real-time.



```
  p2=min(t, w-1);
  for(p=p1;p<=p2;p+=2){
   i=p;
   j=(t-p)/2;
void dither(void){
 int i, j;
 for(j=0;j<h;++j){
  for (i=0; i<w; ++i){
   v = src[i][j];
   v -= ERR(i+1,j-1)/4;
   d[i][j]=(v<128)?0:2;
   s[i][j]=(v<0)?0:3;}
for(t=0;t<2*h;++t){
  p1=max(t%2, 2*h);
  p2=min(t, w-1);
  for(p=p1;p<=p2;p+=2){
   i=p;
   j=(t-p)/2;
void dither(void){
 int i, j;
 for(j=0;j<h;++j){
  for (i=0; i<w; ++i){
   v = src[i][j];
   v -= ERR(i+1,j-1)/4;
```

Source: Tsingmicro

Tsingmicro has developed a software stack to support development on its chips. Developers can use standard frameworks like TensorFlow, Caffe and PyTorch to build machine learning models which can be loaded onto the chips.

Kisaco Assessment

Strengths

- Tsingmicro's chip architecture is the only CGRA in the evaluations we have conducted for this report. It offers high energy efficiency to near ASIC standards with the advantage of programmability (to ASIC's none). The software stack to support this type of chip is essential for AI software engineers to write DL algorithms and Tsingmicro has developed such a full stack.
- The CGRA approach offers dynamic configuration, so that hardware resources are interconnected to form a computing path to perform data-driven calculations. This dataflow architecture approach overcomes the von Neumann bottleneck of fetching and decoding operations in the traditional instruction-driven computing architecture.
- Tsingmicro has already put its innovative processors into production in 2019 and is a proven product on the market. The company is shipping chips and expecting to reach 10m.

Weaknesses

- Tsingmicro has achieved significant success in the market with targeted industrial buyers. We perceive its wider customer engagement is still at a startup level of maturity and we see scope for improving its support for developers building applications on Tsingmicro chips.

- Like many of the vendors we have covered, Tsingmicro has not published independent benchmarks such as MLPerf and we encourage it to do so.
- Like many AI inference chips for the edge market, Tsingmicro chips do not support floating point processing but rather integer precision. This is not an issue for the target market but is a general limitation of these chips for wider AI workloads.

Appendix

Vendor solution selection

Inclusion criteria

In general, the KLC is not designed to exhaustively cover all the players in a market but a representative set of the leading players. Kisaco also invites smaller, possibly niche vendors that have innovative solutions and are on a fast growth path. With this flexibility we consider each participant on its merits as a good fit to the KLC topic.

The criteria for inclusion of a vendor product in this report are as follows:

- Vendor has an offering fitting the topic of AI hardware acceleration processor.
- There are two categories of vendor that are considered for inclusion in this evaluation:
    - Vendor has significant market share relative to peers and is either a recognized leader in the market or has the potential to become one.
    - The vendor is a niche player or an emerging player with outstanding market leading technology.

Exclusion criteria

We exclude vendors that are too early stage with no product that has at least fabricated samples.

Methodology

- Vendors complete a comprehensive capability questionnaire in a spreadsheet, covering the three dimensions of the KLC. The resultant matrix of responses is appropriately scored, and these scores are plotted to produce the KLC.
- We hold comprehensive briefings with all participating vendors, including a demonstration where possible.
- Supplemental information is obtained from vendor literature and publicly available information.

Definition of the KLC

The KLC spans three assessment dimensions.

Technical Features

Kisaco Research has developed a series of features and functionality that provide technology differentiation between the leading solutions in the marketplace, covering hardware and the supporting software stack.

Market execution and strategy

Kisaco Research reviews the capability of the solution and the vendor’s performance in executing its strategy around key areas such as vision of the business, go-to-market strategy, customer engagement, and market execution.

Market share

Market share is a metric normalized to the market leader and is based on the solution's global revenue. Where revenue data is unavailable, Kisaco provides a representative estimate.

Kisaco Research ratings

- **Leader:** This vendor appears in the top right of the KLC chart and has established a significant market position with a product that is technologically advanced compared with peers and its market execution is strong.
- **Innovator:** This vendor appears in the bottom right of the KLC chart and has established a significant technological lead compared with peers but may be still early in its market execution.
- **Contender:** This vendor appears in the top left of the KLC chart and has established an excellent record executing on its market vision. The product is technically strong compared with peers but may be still early in its development.
- **Emerging player**: This vendor appears in the bottom left of the KLC chart and has a strong enough product to have participated in the KLC. The vendor may be still in early stages of establishing itself in the market, or it may be a niche player with a product aimed at a narrower range of customers.

Further reading

Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 1): Technology and Market Landscapes, KR301, July 2020.

Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 2): Data Centers and HPC, KR302, July 2020.

Kisaco Leadership Chart on ML Lifecycle Management Platforms 2020-21, July 2020.

Acknowledgements

I would like to thank all participating vendors for their time to provide briefings and answer many questions, as well as fill out our comprehensive questionnaire.

Author

Michael Azoff, Chief Analyst

michael.azoff@kisacoresearch.com

Kisaco Research Analysis Network

We are running a network for AI chip users, buyers, and people in AI related decision-making roles for their business. We will run surveys, members will receive free reports on the results, and we will also run unique events of interest to the network. To register interest please email: analysis@kisacoresearch.com with your contact details and "Kisaco Research Analysis Network" in the subject line.

Copyright notice and disclaimer

The contents of this product are protected by international copyright laws, database rights and other intellectual property rights. The owner of these rights is Kisaco Research Ltd. our affiliates or other third-party licensors. All product and company names and logos contained within or appearing on this product are the trademarks, service marks or trading names of their respective owners, including Kisaco Research Ltd. This product may not be copied, reproduced, distributed or transmitted in any form or by any means without the prior permission of Kisaco Research Ltd.

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Kisaco Research Ltd. nor any person engaged or employed by Kisaco Research Ltd. accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard - readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Kisaco Research Ltd.

Kisaco Research

CONTACT US

www.kisacoresearch.com

michael.azoff@kisacoresearch.com