AI-Optimized Chipsets
Part III: Key Opportunities & Trends
Aug 2018
An Introduction

Previously in Part I, we reviewed the ADAC loop and key factors driving innovation for AI-optimized chipsets.

In Part II, we review the shift in performance focus computing from general application neural nets and how this is driving demand for high performance computing. To this end, some startups are adopting alternative, novel approaches and this is expected to pave the way for other AI-optimized chipsets.

In this instalment, we review the training and inference chipset markets, assess the dominance of tech giants, as well as the startups adopting cloud-first or edge-first approaches to AI-optimized chipsets.
The training chipset market is dominated by 5 firms, while the inference chipset market is more diverse with >40 players

<table>
<thead>
<tr>
<th></th>
<th>Training Chipsets</th>
<th>Inference Chipsets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No. of Companies</strong></td>
<td>• 5 (i.e. Nvidia, Intel, Xilinx, AMD, Google)</td>
<td>• &gt; 40</td>
</tr>
<tr>
<td><strong>Differentiators</strong></td>
<td>• Computation Performance</td>
<td>• Power Efficiency</td>
</tr>
<tr>
<td></td>
<td>• Usability</td>
<td>• System Latency</td>
</tr>
<tr>
<td></td>
<td>• Innovation Road Map</td>
<td>• Cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Computation Performance</td>
</tr>
<tr>
<td><strong>Barriers to Entry</strong></td>
<td>• R&amp;D Intensity</td>
<td>• Manufacturing Scale Economies</td>
</tr>
<tr>
<td></td>
<td>• Developer Support</td>
<td>• High Switching Costs (End User)</td>
</tr>
<tr>
<td></td>
<td>• Size of End Markets</td>
<td>• Regulatory Requirements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Distribution</td>
</tr>
<tr>
<td><strong>ASP</strong></td>
<td>• USD 2,000 - USD 20,000</td>
<td>• USD 80 - USD 10,000</td>
</tr>
</tbody>
</table>

Source: UBS

• The training chipset market is dominated by 5 firms which have developed massively parallel architectures, well-suited for deep learning algorithms. NVIDIA is the most prominent with its GPU technology stack.

• The early leaders in this market are likely to maintain their lead, but the inference market is large and easily accessible to all, including tech giants and startups. Given the significant market for inference chipsets, many defensible market niches based on high-system speed, low power and/or low Total Cost of Ownership (TCO) products are likely to emerge.

• At least 20 startups and tech giants appear to be developing products for the inference market. According to UBS, within today's datacenter (DC) market, the CPU in particular, appears to be the *de facto* choice for most inference workloads.

Source: UBS
Almost all training happens in the cloud/DC, while inference is conducted in the cloud and at the edge

- The market for AI-optimized chipsets can be broadly segmented into cloud/DC and edge (including automotive, mobile devices, smart cameras and drones).

- The edge computing market is expected to represent more than 75% of the total market opportunity, with the balance being in cloud/DC environments.

- The training chipset market is smaller than that of inference because almost all training is done in DCs. The market for inference chipsets is expected to be significantly larger due to the impact of edge applications for deep learning.

- With the exception of Google that created the Tensor Processing Unit (TPU) specifically for deep learning, other training chipset players entered the deep learning market by repurposing existing architectures that were originally used for other businesses (e.g. graphics rendering for high-end PC video games and supercomputing).

- While a handful of startups plan to commercialize new deep learning training platforms, the vast majority of new players seem focused on inference for large markets (e.g. computer vision). The multitude and diversity of deep learning applications have led many firms to pursue niche strategies in the inference market.

- All training chipset players have developed high performance inference platforms to support sizable end markets in DCs, automotive, and supercomputing. Dozens of startups, many of which are based in China, are developing new designs to capture specific niche markets.
Creating an interesting landscape of tech giants and startups staking their bets across chipset types, in the cloud and at the edge

Key Observations

- At least 45 startups are working on chipsets purpose-built for AI tasks
- At least 5 of them have raised more than USD 100M from investors
- According to CB Insights, VCs invested more than USD 1.5B in chipset startups in 2017, nearly doubling the investments made 2 years ago

Most startups seem to be focusing on ASIC chipsets at the edge and in the cloud/DC

FPGAs and other architectures also appear attractive to chipset startups

Note: Several startups are in stealth mode and company information may not be publicly available.

Source: CrunchBase | IT Juzi | CB Insights | What the drivers of AI industry, China vs. US by ZhenFund | Back to the Edge: AI Will Force Distributed Intelligence Everywhere by Azeem | icons8 | UBS
Most training chipsets are deployed in DCs. The training market is dominated by Nvidia’s GPU

- NVIDIA was first to market, by a wide margin, with semiconductors and products specifically designed for deep learning. It has forward integrated all the way into the developer tools stage of the value chain, with its NVIDIA GPU Cloud offering (which is a container registry, rather than competitor to AWS or Azure), stopping just shy of creating deep learning applications that would compete with its own customers.
- Most training chipsets are deployed in DCs. The training market is dominated by Nvidia’s GPU which has massively parallel architectures, very strong first-to-market positioning, ecosystem and platform integration.
- GPUs have hundreds of specialised “cores”, all working in parallel.
- Nvidia delivers GPU acceleration for both training (Nvidia ® DGX™ SYSTEMS for Data Center, Nvidia GPU Cloud) and inference (Nvidia Titan V for PC, Nvidia Drive™ PX2 for Self-Driving Cars, NVIDIA Jetson™ for intelligent machines).
- In the DC market, training and inference are performed on the same semiconductor device. Most DCs use a combination of GPU and CPU for deep learning.
- The CPU is not well suited for training, but much better suited for inference where code execution is more serial than parallel and low-precision & fixed point logic are still popular.
Microsoft and Amazon both use FPGA instances for large scale inferencing - implying that FPGAs may have a chance in the DC inference market

- Intel bought Altera (FPGA maker) for USD 16.7B in 2015
- Intel sees specialized processors as an opportunity. New computing workloads often start out being handled on specialized processors, only to be “pulled into the CPU” later (e.g. encryption used to happen on separate semiconductors, but is now a simple instruction on Intel CPUs).
- Going forward, Intel is expected to combine its CPU with Altera’s FPGAs
- Intel has secured FPGA design wins from Microsoft for its accelerated deep learning platform and from Alibaba to accelerate its cloud service
- Likely to use FPGAs in its level-3 autonomous solutions for the 2018 Audi A8 and DENSO’s stereo vision system

- Xilinx is one year ahead of Intel in terms of FPGA technology with almost 60% of this market
- It has started developing FPGA chips on a more efficient 7nm node, while Intel is testing the 10nm platform. This could extend Xilinx’s existing 18-month technology lead over Intel in FPGAs
- Baidu has deployed Xilinx FPGAs in new public cloud acceleration services
- Xilinx FPGAs are available in the Amazon Elastic Compute Cloud (Amazon EC2) F1 instances.
- F1 instances are designed to accelerate DC workloads including machine learning inference, data analytics, video processing, and genomics
- Xilinx acquired DeePhi Tech in Jul 2018

By end of 2018, Deloitte estimates that FPGAs and ASICs will account for 25% of all chips used to accelerate machine learning in the DC
At the same time, some tech giants have also started building proprietary chipsets in the cloud/DC

- Alibaba, Amazon, Baidu, Facebook, Google, Microsoft and Tencent (i.e. the “Super 7 Hyperscalers”) exhibit a high degree of vertical integration in the deep learning value chain, capable of building/designing proprietary accelerator(s) by virtue of self-administered hyperscale DC expertise
- If data is the new oil, these firms are emerging as the "vertically integrated oil companies" of tomorrow's data driven global economy. Semiconductor firms must develop strategies to maintain their market power and remain relevant
- The Super 7 hyperscalers, as well as IBM and Apple, have been very active in each stage of the value chain, including algorithmic research, semiconductors, hardware and end-user applications

<table>
<thead>
<tr>
<th>Google</th>
<th>Amazon</th>
<th>Baidu</th>
<th>Tencent</th>
</tr>
</thead>
<tbody>
<tr>
<td>First promoting the idea of a custom chip for AI in the cloud, just launched the third generation TPUs in 2018</td>
<td>Acquired Annapurna Labs in 2012, building AI chip for Echo, deploying FPGAs for all-programmable nature in the AWS cloud</td>
<td>Released XPU demonstrated in FPGAs. Recently unveiled its own AI chip Kunlun for voice recognition, NLP, image recognition, and autonomous driving</td>
<td>Introduced FPGA Cloud Computing with three different specifications based on Xilinx Kintex UltraScale KU115 FPGA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microsoft</th>
<th>Alibaba Group</th>
<th>Facebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uses Altera FPGAs from Intel for Azure via Project Brainwave. Also employs NVIDIA GPUs to train neural networks, designed a custom AI chipset for HoloLens, hiring engineers on AI chip design for cloud</td>
<td>Intel’s FPGAs are powering the Alibaba cloud. It developed Ali-NPU “Neural Processing Unit” to handle AI tasks to strengthen its cloud and acquired a Hangzhou-based CPU designer C-SKY</td>
<td>Big Basin System features 8 Nvidia Tesla P100 GPU accelerators connected by NVIDIA NVLink. It is forming a team to design its own chips and building the “end-to-end SoC/ASIC”</td>
</tr>
</tbody>
</table>

Source: UBS
Google’s ASIC chipset – the cloud TPU is a case in point

Instead of selling the TPU directly, Google offers access via its new cloud service

- Google’s original TPU had helped power DeepMind’s AlphaGo victory over Lee Sedol
- It is making cloud TPUs available for use in Google Cloud Platform
- The TPU is designed to be flexible enough to for many different kinds of neural network models

### TPUv1 (2015)
- 92 teraflops
- Only handle 8-bit integer operations

### Cloud TPU (2017)
- 180 teraflops
- 64 GH HBM
- Added support for single-precision floats

### TPUv2 Pod (2017)
- 11,500 teraflops
- 4 TB HBM
- 2-D toroidal mesh network

### TPUv3 Pod (2018)
- >100,000 teraflops
- More than 8x the performance of a TPUv2 Pod
- New chip architecture + large-scale system

As more compute is pushed to the edge, tech giants are increasingly developing or acquiring solutions in this space

- As these trained models shift to real-world implementation, more compute will be pushed to the edge
- Other edge applications that will contain specialized deep learning silicon include drones, kiosks, smart surveillance cameras, digital signage and PCs

### Chipset Makers

- **Intel**
  - Catching up via acquisitions
  - **Movidius** (ASIC) for computer vision
  - **Mobileye** (Computer Vision for automotive) for computer vision in autonomous driving

- **Qualcomm**
  - Qualcomm’s Snapdragon 845 Neural Processing Engine is used for deep learning inference at the edge

- **NXP**
  - NXP’s S32V234 (vision processor) designed to support automotive applications

- **NVIDIA**
  - Titan V for PC
  - Drive™ Xavier and Pegasus platform for Self-Driving Cars
  - NVIDIA Jetson™ in cameras and appliances at the edge

### Tech Companies

- **Apple**
  - Apple unveils its edge engine called neural engine on iPhone X with approximately 30M units sold in 2017

- **Tesla**
  - Tesla is reportedly developing its own AI chip, intended for use with its self-driving systems, in partnership with AMD

- **Google**
  - Google had built Pixel Visual Core for heavy-lifting image processing while using less power

Source: UBS
Driving a deal-making galore in the chipset industry

Where Chipset Makers are Investing in Private Markets (2015-2017)

• Intel Capital is the most active corporate entity. Close behind is Qualcomm Ventures
• Over 79 deals went to IoT companies — which is to be expected, given their relevance to embedded computing with small chipsets
• 15 of these deals went to drone-specific companies
• AR/VR companies also attracted 25 deals
• 50 deals went to AI companies, many of which were cutting edge computer vision and auto tech plays
• Chipset makers also invested in horizontal AI platforms as well as visual recognition API maker

Source: Where Major Chip Companies Are Investing In AI, AR/VR, And IoT by CB Insights
At the same time, there is a proliferation of new entrants, mostly ASIC-based startups including Graphcore, Wave Computing.

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GRAPHCORE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Located: UK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Founded: 2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage: Series C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Intelligence Processor Unit (IPU) uses graph-based processing</td>
<td>• IPU holds the complete machine learning model inside the processor</td>
<td>• Raised USD 50M Series C led by Sequoia Capital in Nov 2017</td>
</tr>
<tr>
<td>• Uses massively parallel, low-precision floating-point computations</td>
<td>• With computational processes and layers being run repeatedly</td>
<td>• Raised USD 30M Series B led by Atomico in Jul 2017 &lt;br&gt; Investors include DeepMind’s CEO Demis Hassabis and Uber’s Chief Scientist Zoubin Ghahramani</td>
</tr>
<tr>
<td>• Provides much higher compute density than GPUs</td>
<td>• Over 100x more memory bandwidth than other solutions</td>
<td></td>
</tr>
<tr>
<td>• Aims to fit both training and inference in a single processor</td>
<td>• Lower power consumption and higher performance</td>
<td></td>
</tr>
<tr>
<td>• Expected to have more than 1,000 cores in its IPU</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Wave Computing    |                                                                             |                                                                                                               |
| Located: US       |                                                                             |                                                                                                               |
| Founded: 2010     |                                                                             |                                                                                                               |
| Stage: Series D   |                                                                             |                                                                                                               |
|ASIC               |                                                                             |                                                                                                               |
| • Dataflow Processing Unit (DPU) does not need a host CPU | • Characterizes as a coarse-grain reconfigurable array (CGRA) | • Raised an undisclosed amount in Series D round from Samsung Venture investment and other undisclosed investors in Jul 2017 |
| • Consists of thousands of independent processing elements designed for the 8-bit integer operations | • Begins on-premise system testing with early access customers |                                                                             |
| • Each with its own instruction memory, data memory, and registers | • Acquired MIPS Tech to combine technologies to create products that will deliver a single Datacenter-to-Edge platform for AI |                                                                             |
| • Grouped into clusters with a mesh interconnect | |                                                                             |

Source: What Sort of Silicon Brain Do You Need for Artificial Intelligence? by The Register | Venture Beat | Suffering Ceepie-Geepies! Do We Need a New Processor Architecture? By The Register
### KnuEdge, Gyrfalcon Technology...

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Building a “neural chip” to make DCs more efficient in a hyperscale age</td>
<td>- Programmers can program each one of the cores with a different algorithm to run simultaneously for the “ultimate in heterogeneity”</td>
<td>- Raised USD 100M in funding</td>
</tr>
<tr>
<td>- The first chipset features 256 cores</td>
<td>- Multiple input, multiple data</td>
<td>- Planning another round in early 2018</td>
</tr>
<tr>
<td>- Chips are part of a larger platform: [1] KnuVerse, a military-grade voice recognition and authentication technology [2] LambdaFabric technology makes it possible to instantly connect those cores to each other</td>
<td>- Chip is based on “sparse matrix heterogeneous machine learning algorithms”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Preparing to provide cloud-based machine intelligence-as-a-service</td>
<td></td>
</tr>
</tbody>
</table>

### Gyrfalcon Technology

| Located: US  
Founded: 2017  
Stage: N.A.  
ASIC | Located: US  
Founded: 2005  
Stage: Undisclosed  
ASIC |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- The Lightspeeur® series intelligent processors deliver a revolutionary architecture that features low-power, low-cost, and massively parallel compute capabilities to empower AI to consumer electronic devices, mobile edge computing, as well as cloud AI DCs</td>
<td>- The Lightspeeur® 2801S AI Processor features high performance and energy efficiency (9.3 TOPS/Watt)</td>
</tr>
<tr>
<td>- USB 3.0 Stick can be connected to PCs, laptops and mobile phones to enhance image and video-based deep learning capabilities.</td>
<td>- USB 3.0 Stick can be connected to PCs, laptops and mobile phones to enhance image and video-based deep learning capabilities.</td>
</tr>
<tr>
<td>- The multi-chip server board with M.2 and PCIe Interface for cloud applications</td>
<td>- The multi-chip server board with M.2 and PCIe Interface for cloud applications</td>
</tr>
</tbody>
</table>

Source: PitchBook | CB Insights | Forbes
As well as Cerebras, Groq, Tenstorrent - startups with potential offerings in the cloud/DC

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cerebras</strong>&lt;br&gt;Located: US&lt;br&gt;Founded: 2016&lt;br&gt;Stage: Series C Stealth Mode</td>
<td>• New chip to handle sparse matrix math and communications between inputs and outputs of calculations&lt;br&gt;• People familiar with the company claims its hardware will be tailor-made for “training”&lt;br&gt;• Yet to release a product&lt;br&gt;• Ex-AMD team backed by Benchmark Capital</td>
<td>• Raised USD 25M Series B in Jan 2017&lt;br&gt;• Raised USD 27M Series A in May 2016&lt;br&gt;• Plan to raise Series C</td>
</tr>
<tr>
<td><strong>Groq</strong>&lt;br&gt;Located: US&lt;br&gt;Founded: 2016&lt;br&gt;Stage: Undisclosed ASIC</td>
<td>• Ex-Google TPU team backed by Social Capital&lt;br&gt;• Processor is claimed that its first chip will run 400 teraflops which is more than twice Google’s cloud TPU (180 teraflops)&lt;br&gt;• Recently hired Xilinx’s Sales VP Krishna Rangasayee as its Chief Operating Officer</td>
<td>• Funded with USD 10.3M from venture capitalist Chamath Palihapitiya</td>
</tr>
<tr>
<td><strong>Tenstorrent</strong>&lt;br&gt;Located: Canada&lt;br&gt;Founded: 2016&lt;br&gt;Stage: Series A ASIC</td>
<td>• Developing high-performance processor ASICs&lt;br&gt;• Specifically engineered for deep learning and smart hardware&lt;br&gt;• Designed for both learning and inference&lt;br&gt;• Architecture claims to be able to scale from devices to cloud servers&lt;br&gt;• Adopts adaptive computation – does not pre-plan how to split the computation (vs. Graphcore pre-planning)</td>
<td>• USD 12M Series A from Jim Keller, Real Ventures and Clips Inc</td>
</tr>
</tbody>
</table>

Source: PitchBook | CB Insights | Forbes | Cerebras | Groq | Tenstorrent
### At the edge, there are also ASIC offerings by Horizon Robotics, Mythic

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Provides an integrated and embedded AI solutions optimizing both algorithmic innovation and hardware design</td>
<td>Launched embedded AI smart vision processors</td>
<td>• Raised more than USD 100M Series A+ led by Intel Capital in Dec 2017</td>
</tr>
</tbody>
</table>

**Horizon Robotics**
- Located: China
- Founded: 2015
- Stage: Series A+
- ASIC

**Mythic**
- Located: US
- Founded: 2012
- Stage: Series B
- ASIC

- Analog processor-in-memory design
- Server GPU performance at 1/100th of the power draw and cost
- Incubated at the Michigan Integrated Circuits Lab
- Chips interface analog, digital, and flash memory into a monolithic compute and storage block
- Tiled many times over to form a truly unconventional parallel computer.
- Raised USD 40M Series B led by Softbank Ventures with new investors Lockheed Martin Ventures and Andy Bechtolsheim in Mar 2018
- Raised USD 9.5M Series A led by Draper Fisher Jurvetson
- Other investors include Lux Capital, Data Collective and AME Cloud Ventures

Source: 12 AI Hardware Startups Building New AI Chips by Nanalyse | AI Chip Boom: This Stealthy AI Hardware Startup Is Worth Almost A Billion by Forbes | Chip Startups were almost toxic before the AI boom – now investors are plowing money in to them | Pitchbook | Startup Unveils Graph Processor at Hot Chips by EETimes
## Cambricon, Kneron...

### Cambricon

- **Located:** China
- **Founded:** 2016
- **Stage:** Series B
- **ASIC**

- Developing a brain-inspired processor chip to conduct deep learning for both the edge and cloud
- Enable AI for computer vision, autonomous driving and flying, security monitoring, speech recognition, and more
- Also offers an AI software platform for developers

- **Source:** PitchBook | CB Insights

### Kneron

- **Located:** China, US, Taiwan
- **Founded:** 2015
- **Stage:** Series A
- **ASIC**

- Offers edge AI solutions
- Its solutions include Neural Processing Unit (NPU), a dedicated AI processor, and its software solutions
- Kneron’s Reconfigurable Artificial Neural Network (RANN) technology can reconfigure the architecture according to different applications to reduce the complexity

- **Source:** PitchBook | CB Insights

### Solution Highlights

- **Cambricon**
  - Helmed by Chen Yunji and Chen Tianshi, professors at the CAS Institute of Computing Technology
  - Released its AI Chip, called 1A, in 2016
  - Huawei’s Kirin 970 chip, which was developed to power Mate 10, used Cambricon’s IP
  - Unveiled MLU100 for cloud computing and 1M chip for edge computing

- **Kneron**
  - Its NPU can support convolutional neural networks technology for visual recognition and deep learning
  - Its visual recognition software with RANN technology can be used in face recognition, body detection, object recognition, and space recognition
  - Main applications: smart home, smart surveillance, smartphones

### Financing

- **Cambricon**
  - It has raised USD 100M of Series A led by SDIC Chuangye Investment Management in Aug 2017. Alibaba and Lenovo Capital participated
  - It has closed Series B with a reported valuation of USD 2.5B from Capital Venture, SDIC Venture Capital, China Capital Investment Group etc.

- **Kneron**
  - Raised a USD 18M Series A1 round led by Horizon Ventures in May 2018
  - Previously raised a Series A round of more than USD 10M led by Alibaba Entrepreneurs Fund, CDIB Capital Group, and participated by Qualcomm in Nov 2017

Source: PitchBook | CB Insights
<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed a power-efficient edge computing platform based on its Graph Streaming Processor Architecture</td>
<td>Compared to GPU/DSP “node at a time” approach, Thinci is adopting “concurrent nodes” approach</td>
<td>Raised Series B led by Denso International America in Oct 2016</td>
</tr>
<tr>
<td>Provides complete solutions from appliances, modules, add-in accelerators, discrete SOCs, software development kit, libraries of algorithms and applications</td>
<td>This cuts down or eliminates inter-processor communications</td>
<td>Intercept Ventures, Magna and 10 individual investors also participated</td>
</tr>
</tbody>
</table>

Located: US & India
Founded: 2012
Stage: Series B
ASIC

Located: Israel
Founded: 2017
Stage: Series A
ASIC

- A scalable hardware and software processor technology for deep learning, enabling datacenter class performance in an embedded device form factor and power budget
- All-digital, Near Memory Processing (NMP) dataflow approach, focused on edge devices, which eliminates the architectural bottlenecks and reaches near-theoretical efficiency.
- Utilizes properties of NNs, such as quasi-static structure, predefined control plane and connectivity patterns

Raised USD 12.5M Series A
## As well as Syntiant

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Syntiant’s Neural Decision Processors (NDPs) utilize patented analog neural network techniques in flash memory to eliminate data movement penalties by performing neural network computations in flash memory, enabling larger networks and lower power</td>
<td>• Its ultra-low power, high performance NDPS are suited for voice, sensor and video applications from mobile phones and wearable devices to smart sensors and drones</td>
<td>• Closed its Series A round led by Intel Capital in Mar 2018</td>
</tr>
</tbody>
</table>

Source: Syntiant | PitchBook
**Efinix and Achronix - FPGA startups offering cloud/DC and edge solutions**

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Its quantum technology delivers 4X Power-Performance-Area advantage over traditional programmable technologies</td>
<td>• The Trion platform comprises logic, routing, embedded memory and DSP blocks and has been implemented 40nm CMOS manufacturing process by SMIC</td>
<td>• It has raised USD 9.5M in a funding round led by Xilinx and Hong Kong X Technology Fund</td>
</tr>
<tr>
<td>• Ultra-low-power and general edge applications for IoT and mobile applications and advanced applications targeting infrastructure, data centers, and advanced silicon processes</td>
<td>• Quantum-enabled products have a cost and power structure that enables volume production in ways that traditional FPGAs cannot match.</td>
<td>• Other investors include Samsung Ventures Investment, Hong Kong Inno Capital and Brizan Investments</td>
</tr>
</tbody>
</table>

**Efinix**

Located: US  
Founded: 2012  
Stage: Series A  
FPGA

**Achronix**

Located: US  
Founded: 2004  
Stage: Series C  
FPGA

- Offerings include programmable FPGA fabrics, discrete high-performance and high-density FPGAs with hardwired system-level blocks, DC and HPC hardware accelerator boards, and best-in-class EDA software supporting all Achronix products
- Plans to further enhance its high-performance and high-density FPGA technology for hardware acceleration applications in DC compute, networking and storage; 5G wireless infrastructure, network acceleration; advanced driver assistance systems (ADAS) and autonomous vehicles
- Completed a USD 45M Series C in Oct 2011 from Easton Capital, New Science Ventures and Argonaut Partners

Source: PitchBook | CB Insights | CrunchBase | Efinix | Achronix
While ThinkForce, Adapteva offer alternative chipset architectures from CPUs, GPUs, FPGAs and ASICs in their solution(s)

<table>
<thead>
<tr>
<th>Solution</th>
<th>Highlights</th>
<th>Financing</th>
</tr>
</thead>
</table>
| ThinkForce | • Plans to launch its AI chip based on the "Manycore" architecture  
• Helps to complete and optimize the AI cloud virtualization scheduling during the chip-level implementation  
• Efficiency of ThinkForce accelerators is between 90% and 95%  
• Over 5x more efficient in power consumption and cost savings compared with Nvidia | • Partnered with IBM and Cadence  
• Core members were directly responsible for some of the most successful semiconductor chipset launches like the IBM PowerPC, Sony PS3, Microsoft XBOX and the world's fastest 56G SerDes | • Raised RMB 450M Series A led by Sequoia Capital China, Yitu Technology, Hillhouse Capital Group  
• Other participants include Yunfeng Capital |

| Adapteva | • Focuses on Manycore accelerator chips and computer modules for applications with extreme energy efficiency requirements such as real, time image classification, machine learning, autonomous navigation and software defined radio | • Proprietary architecture called Epiphany which is a scalable distributed memory architecture, featuring thousands of processors on a single chip connected through a sophisticated high-bandwidth Network-On-Chip  
• Holds a 10-25x energy efficiency advantage over traditional CPU architectures  
• Shipped to over 10,000 developers | • Raised USD 3.6M Series B from Carmel Ventures and Ericsson in 2014  
• USD 0.85M debt financing in 2012  
• Raised USD 1.5M Series A in 2012 |

Source: PitchBook | CB Insights | CrunchBase | ThinkForce | Adapteva
Notwithstanding this positive outlook, challenges abound for many of these startups

**Execution Risks**
- Might take years to get a chipset to market
- Much of the hardware remains in development for now
- Many startups could fail for being narrowly focused
  - Optimizing for applications that do not become mainstream
  - Similar to startups building processors for 4G wireless applications in the mid-2000s
- However, if startups build chipset(s) that span too wide a set of applications, performance is likely to be sacrificed, leaving them vulnerable to competition

**Competition**
- From 2018, many companies will start releasing their new chipsets, after which the market validation will then begin
- 2019 and 2020 will be the years when a ramp-up will take place and winners will begin to emerge
- Cloud computing giants are already building their own chipsets
- Intel recently announced plans to release a new family of processors designed with Nervana Systems which Intel acquired in 2016. Nvidia is also quickly upgrading its capabilities
- It appears the AI-optimized chipset market is going to be saturated with many players and it is not clear where the exits will be

**Barriers to Entry in DCs**
- The market for chipsets that run massive DCs is a key target for some startups
- Market is currently dominated by tech giants like Amazon, Apple, Facebook, Google, Microsoft, Baidu, Alibaba and Tencent
- Many of these tech giants are already developing proprietary AI-optimized chipsets

**Network Effects & Switching Costs**
- Some industry participants have reservations on market migration to new AI-optimized chipsets given the user familiarity with the tools needed to work with GPUs
- Google will continue to offer access to GPUs via its cloud services as the blossoming market for AI-chipsets spans many different processors in the years to come

“They (GPUs) are going to very hard to unseat ... because you need an entire ecosystem” — Yann LeCun

Source: AI Chip Boom: This Stealthy AI Hardware Startup Is Worth Almost A Billion by Forbes | The Race to Power AI's Silicon Brains by MIT Technology Review | Lux Capital
Conclusion

In **Part I**, we note that deep learning technology has primarily been a software play to date. The rise of new applications (e.g. autonomous driving) is expected to create substantial demand for computing.

Existing processors were not originally designed for these new applications, hence the need to develop AI-optimized chipsets. We review the ADAC loop and key factors driving innovation for AI-optimized chipsets.

In **Part II**, we explore how AI-led computing demands are powering these trends. To this end, some startups are adopting alternative, novel approaches and this is expected to pave the way for other AI-optimized chipsets.

This is the end of Part III, where we took a high-level look at training and inference chipset markets, the activities of tech giants in this space, coupled with disruptive startups adopting cloud-first or edge-first approaches to AI-optimized chips.

Evidently, this is a super-saturated industry with many players. **It is unclear where the exits will be.** Tech giants are either moving forward with their roadmaps (i.e. Nvidia), building their own chipsets (e.g. Alibaba, Amazon, Apple, FaceBook etc) or have already made acquisitions (e.g. Intel).

In Part IV, we look at other emerging technologies including neuromorphic chips and quantum computing systems, to explore their promise as alternative AI-optimized chipsets.

Finally, we are most grateful to Omar Dessouky (Equity Research Analyst, UBS) for his many insights on the semiconductor industry.

Do let us know if you would like to **subscribe** to future Vertex Perspectives.
Thanks for reading!

About Vertex Ventures
Vertex Ventures is a global network of operator-investors who manage portfolios in the U.S., China, Israel, India and Southeast Asia.
Vertex teams combine firsthand experience in transformational technologies; on-the-ground knowledge in the world’s major innovation centers; and global context, connections and customers.

About the Authors

Emanuel TIMOR  
General Partner  
Vertex Ventures Israel  
emanuel@vertexventures.com

Sandeep BHADRA  
Partner  
Vertex Ventures US  
sandeep@vertexventures.com

XIA Zhi Jin  
Partner  
Vertex Ventures China  
xiazj@vertexventures.com

Brian TOH  
Executive Director  
Vertex Holdings  
btoh@vertexholdings.com

Tracy JIN  
Director  
Vertex Holdings  
tjin@vertexholdings.com

Disclaimer
This presentation has been compiled for informational purposes only. It does not constitute a recommendation to any party. The presentation relies on data and insights from a wide range of sources including public and private companies, market research firms, government agencies and industry professionals. We cite specific sources where information is public. The presentation is also informed by non-public information and insights.

Information provided by third parties may not have been independently verified. Vertex Holdings believes such information to be reliable and adequately comprehensive but does not represent that such information is in all respects accurate or complete. Vertex Holdings shall not be held liable for any information provided.

Any information or opinions provided in this report are as of the date of the report and Vertex Holdings is under no obligation to update the information or communicate that any updates have been made.