Deep Learning and the processor chips fueling the AI revolution – a primer

Deep learning processors – converting data into intelligence
We examine digital processing chips that are enabling computing devices to achieve human sight/sound/perception and to continuously learn and make predictions from the explosive amount of unstructured text, images, sound and video. This scale of prediction requires cognitive abilities matching the brain with its 100 billion neurons and 1 trillion bits per second of processing. It requires massive amounts of parallel computation, which is only feasible with graphics processors from Nvidia, programmable chips from Intel and Xilinx, and other ASIC/accelerators from Google, IBM, and others.

10x growth to $10bn potential market by 2020
We estimate the accelerated computing processor chip market across cloud, supercomputers and enterprise applications has the potential to grow ten-fold to over $10bn by 2020E. The ability for machines to see, hear, predict and self-correct has profound implications in healthcare, media, financial, consumer, automotive, defense, gaming, oil/gas, government and education verticals. Applications range from brain cancer detection to weather forecasting to speech recognition to energy market price forecasting to even beating a Korean grandmaster at the 2500 year old game of Go.

Artificial intelligence vs. machine learning vs. deep learning
AI is the overarching concept which refers to a machine exhibiting human intelligence. Machine learning is a subset of AI and consists of taking some data, training a model on that data, and using the trained model to make predictions on new data. During this training phase, the model continuously iterates and gets feedback on its accuracy, with massive amount of computing power required to get the model just right. The "training" phase is followed by the "inference" phase where the model gets put to actual use. Meanwhile deep learning is a way of implementing machine learning, by using multiple hierarchical model layers that mimic the brain’s neural connections.

Processors tradeoffs power, programmability, speed
Nvidia’s graphics processors were instrumental in enabling the deep learning industry, but other options have emerged, including Intel’s Xeon Phi; field programmable gate arrays aka programmable logic from Altera/Intel and Xilinx; Google’s internally designed TensorFlow; IBM’s TrueNorth experimental ASIC; Qualcomm’s Zeroth platform; and new offerings from Nervana (now part of Intel), Knupath, and Wave Computing.

Chart 1: Accelerator TAM $10bn by 2020

Source: BofA Merrill Lynch Global Research

Note: HPC – High Performance computing
Portfolio Manager’s Summary

Accelerated or parallel computing processors and their applications into artificial intelligence and deep learning could be the fastest growing part of technology for the next 10 years. The ability of machines to see, hear, learn, predict and correct with superhuman abilities – aka “artificial intelligence” or AI – has profound implications for a large number of end markets, including VR. We expect a select class of parallel computing processors from Nvidia, Intel and others to be at the forefront of this revolution.

From a processor chip perspective, we estimate that AI, and other parallel computing end-markets could grow ten-fold from $1bn in 2016 to over $10bn by 2020, at a remarkable 75% CAGR, again marking the fastest growing application-market in semiconductors. This $10bn includes $8.4bn in cloud/deep learning with another $1.7bn in high performance computing (HPC) or supercomputing applications. We also expect the 20% of the cloud accelerator TAM in 2020 to be driven by deep learning “inference” market while 80% of the TAM in 2020 will be driven by deep learning “training” market.

**Chart 2: The addressable market for accelerator chips could grow 10x to $1bn by 2020E**

![Chart showing growth of TAM from 2015 to 2020](image)

Source: BofA Merrill Lynch Global Research estimates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud</td>
<td>$125</td>
<td>$404</td>
<td>$841</td>
<td>$1,802</td>
<td>$5,025</td>
<td>$8,404</td>
<td>114%</td>
</tr>
<tr>
<td>HPC</td>
<td>$609</td>
<td>$686</td>
<td>$801</td>
<td>$1,005</td>
<td>$1,283</td>
<td>$1,706</td>
<td>26%</td>
</tr>
<tr>
<td>Total Tam</td>
<td>$734</td>
<td>$1,090</td>
<td>$1,642</td>
<td>$2,806</td>
<td>$6,307</td>
<td>$10,109</td>
<td>75%</td>
</tr>
</tbody>
</table>

Source: BofA Merrill Lynch Global Research estimates

While we currently only look at the HPC and cloud/hyperscale environment, we believe there is an incremental market among private enterprise deploying artificial intelligence and business intelligence to mine data for actionable information and intelligence. Such enterprise customers could choose to deploy their private cloud or on premise equipment, or buy it on on-demand basis from public cloud vendors such as Amazon or Microsoft.

A recent report from BofAML sizes the overall market for artificial intelligence hardware, software and services to be roughly $2.1 billion for 2015, and expects that to grow to $36 billion in 2020 and almost quadruple to $127 billion by 2025. This implies a 76% 5-year compounded annual growth rate (CAGR) and a 51% 10-year CAGR. Tractica estimates the breakdown of AI usage between different end-markets and estimates that 19% of use cases in 2015 were for Ad Service, 16% was for Investments, 12% for
Retail, and another 11% for Media. Some end-markets Legal and Philanthropies are less than 1% and expected to see limited or no growth, while a field such as Medical Diagnostics was only at 4% but was expected to have a lot more AI share in the future.

**Figure 1: Artificial Intelligence revenue by segment 2015-2025**

![Chart showing Artificial Intelligence revenue by segment from 2015 to 2025.](chart1.png)

Source: BofA Merrill Lynch Global Research estimates

**Figure 2: Artificial Intelligence revenue end-market 2015**

![Chart showing Artificial Intelligence revenue end-market in 2015.](chart2.png)

Source: Tractica

**Parallel computing key enabler of accelerated processors**

A serial computer has a central processor (CPU made by Intel or AMD) that can address an array of memory locations where data and instructions are stored. Computations are made by the processor reading an instruction as well as any data the instruction requires from memory addresses. The instruction is then executed and the results are saved in a specified memory location as required. In a serial system, the computational steps are deterministic, sequential and logical, and the state of a given variable can be tracked from one operation to another. In other words, a single problem is broken into a series of discrete instructions that are executed sequentially by a single processor. Only one instruction at a time can be executed, but if the processor has multiple threads (ordered sequence of instructions), it is possible to simulate a parallel computing function through time sharing (alternating between different threads).
In parallel computing, multiple compute resources can be used simultaneously to solve a computing problem (see Exhibit 6). This can be done by breaking a problem into discrete parts that can be solved concurrently, and each part is simultaneously executed on different processors with an overall control mechanism coordinating the compute function. Parallel computing was used by researchers and governments to model difficult problems in many areas of science and engineering such as applied physics, seismology, and weather pattern prediction.

Today, parallel computing increasingly is used in commercial applications such as data mining, web search engines, financial modelling, and virtual reality. Parallel computing could be achieved by combining standalone computers (supercomputers are an example), but for large scale cost effective implementation, it is key to implement parallel computing within a processor itself. GPU has a massively parallel architecture consisting of thousands of smaller, more efficient cores designed for handling multiple tasks simultaneously and efficiently. When it comes to implementing deep learning algorithms, researchers preferred to use GPUs instead of Intel’s CPUs.

An artificial neural network essentially works well with a parallel computing framework. Each core in a processor essentially is tasked with a small function that eventually will be combined to arrive at an outcome. Once the parallel computing of large data sets is completed and the system is trained, a CPU or any other processor could run through new data sets in a serial fashion and make predictions.
Exhibit 3: Parallel computing processors include GPU, Xeon Phi, FPGA, Google TensorFlow, ASICs

Alternative solutions available for inference but programmability is a concern
Facebook uses CPU for running deep learning inference applications but it is well documented that CPU is not the optimum processor for deep learning workloads. There is considerable supporting evidence – Google released Tensor flow ASIC, Microsoft evaluating FPGA (Catapult servers), startups like Nervana came up with custom ASIC solutions and most importantly, Intel promoted one processor (Knights Mill) that can do serial and parallel computing. However, each of these processor options has one key issue that needs to be overcome – programmability. For parallel computing, GPU solution offers the customer better performance and programmability now, largely because Nvidia has been working on a solution for the past 10 years. However, recently Microsoft updated that all the servers in Azure have been equipped with FPGA to run inference workloads and could be a tailwind for the adoption of non-GPU solutions in deep learning workloads.

Can Intel make the transition to parallel computing in time?
In 2016, Intel released a competitive product called Knights Mill within its Xeon Phi family of products, which were designed to address supercomputer market needs (released first in 2013). Knights Mill specifically addresses deep learning. It is designed to execute parallel instructions similar to GPU as well as perform the serial processing functions (both functions embedded within one piece of silicon). However, all the instruction sets are based on Intel homegrown solutions vs more established CUDA framework for GPU and as such, will need additional software optimization.

The key issue is the timely replacement of serial processing Xeon servers with serial/parallel processing Xeon Phi servers. Given that Intel has nearly 100% of the server market now, the transition should not be difficult but in the parallel computing world, Intel products are not the only solution available – GPU, FPGA, ASIC are alternate solutions. Intel acquired FPGA (Altera) and a custom ASIC vendor (Nervana) to diversify its portfolio, but Intel has to solve the ease of programmability concern. In our view, it will take time and will face competition from AMD (CPU/GPU), Xilinx (FPGA), ARM ecosystem (CPU), and most importantly, Nvidia (GPU).

AI vs machine learning vs deep learning
Machine learning is not a new concept but has been in use since the 1980s to solve complex pattern recognition problems. Accelerator computing initially was adopted by university researchers and governments. Super computers were built to accelerate deep learning activities – IBM Watson came out of that initiative. Use cases for supercomputing include weather forecasts which have massive computing needs.
Exhibit 4: Weather simulation can be done at a better resolution and predictions done for longer durations due to GPUs

<table>
<thead>
<tr>
<th>BEFORE GPUs</th>
<th>AFTER GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-Hour Forecasts</td>
<td>24-Hour Forecasts</td>
</tr>
<tr>
<td>2.2km Resolution</td>
<td>1.1km Resolution (2x Higher)</td>
</tr>
<tr>
<td>8 Simulations per Day</td>
<td>8 Simulations per Day</td>
</tr>
<tr>
<td>Medium Range Forecasts</td>
<td>Medium Range Forecasts</td>
</tr>
<tr>
<td>3 Day Forecasts</td>
<td>5 Day Forecasts (2 Days Longer)</td>
</tr>
<tr>
<td>6.6km Resolution</td>
<td>2.2km Resolution (3x Higher)</td>
</tr>
<tr>
<td>3 Simulations per Day</td>
<td>42 Simulations per Day (14x More)</td>
</tr>
</tbody>
</table>

Source: MeteoSwiss Weather Forecasting

More recently, AI has taken the prime stage, and as a result, machine learning has become more important. In our view, deep learning is an extension of machine learning and really enables the implementation of an AI system. Essentially, deep learning or supervised/unsupervised training of deep neural networks (DNN) is necessary in order for AI or systems with predictive capabilities to function efficiently (fewer errors). The accuracy of the AI system depends on the efficiency of architecture used to train the system. At the same time, companies are evaluating various options to minimize the overall cost incurred in implementing AI in a broader scale. In our primer, we conduct a deep dive into the various aspects of deep learning, both hardware and software and present an unbiased view on the potential impact to semiconductor ecosystem.

Exhibit 5: Artificial intelligence has been around since the 1950s but deep learning has just started

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: Nvidia

Deep learning is a subset of machine learning

At a high level, deep learning is a subset of machine learning. Machine learning is a combination of two steps – training and inference. Training involves teaching a neural network to recognize objects, voices etc, just like the neurons in a child’s brain are taught to do so by school teachers. Neural networks are essentially a computer system that is modelled on the human brain and nervous system. In the past, machine learning was used to create models that solve complex pattern recognition problems (such as face recognition or spam filtering.) Recent developments in machine learning algorithms use many optimization layers in order to fine tune the output, and at a high level, this is called deep learning.
Deep learning started to gain prominence in 2011 when IBM Watson computer won against humans in Jeopardy. Following that, innovation in algorithms (convolutional neural nets, etc.) and the availability of higher performance graphics processing units (GPU) allowed for deep learning to progress further (Andrew Ng implemented deep learning at Google – google brain). Google’s Deepmind Alpha Go algorithm beat the former world champion Lee Sedol at Go in early 2016.

Even with the evolution of machine learning algorithms, the core framework remains the same for implementing a deep learning neural network: train a neural network that can run inference computations in the field by using the results of the previous training to classify, recognize and generally process unknown inputs.

**Neural network basics**

Many types of neural networks have been developed, but a few popular ones include convolutional and recurrent neural networks. At a high level, each neural network has three key attributes: architecture, activity rule and learning rule. Architecture specifies the variables involves in the network and the relationship between them. In the neural network, it could be the weight and activities of neurons. Activity rule defines how the activities of the neurons change in response to each other. Third, learning rule defines the way in which the neural network’s weights change with time. In order to fully train the network, the system needs to back propagate (feedback) errors and adjust the weights for neurons accordingly – this is similar to how children learn to identify objects.

In order to increase computational efficiency, the entire operation has to be executed simultaneously. Typically, in order to train a neural network, one million to one billion parameters have to be adjusted, and simultaneous operation is important to minimize time of operation. Throughput, or the number of operations executed per second, is a key attribute for training a neural network. A GPU is the only processor capable of executing billions of operations a second in parallel but new alternatives (Intel Xeon Phi) are emerging.

**Exhibit 6: Training needs more throughput while Inference needs lower latency**

![Diagram](source: Nvidia)

However, the performance goal for inference is slightly different. Typically, the inference batch size is smaller and uses less precise data (8 bits of data at a time) than training (16/32 bits of data at a time), as users don’t want to wait several seconds while the system is accumulating images for a large batch. Moreover, it is important for the results to be communicated to the user faster. Latency, or time taken to deliver an output, is a more important factor for inference. In other words, inference task can use a lower performance, low precision GPU (Nvidia Pascale based P4) than a high performance, high precision GPU (Nvidia Pascal based P100) – we cover this in detail later. Given that throughput is not a concern, a non-GPU solution will also work as long as the processor (ASIC, ASSP, CPU, FPGA) is configured the right way.

**Convolutional vs recurrent neural network (CDNN vs RDNN)**

The key difference between the two networks is the type of input and output – fixed or
variable. In a convolutional neural network, the input and output are of fixed sizes, whereas in a recurrent neural network, the input/output is variable. More importantly, the RNN can leverage prior experience from its memory and fine tune output. For example, if the input is “I bought an apple...I am eating...” an RNN will likely place a higher probability of picking apple. RNNs can be used for natural language translation (like Google Translate) or speech recognition, while CNN is used in image/face recognition. Also, RNNs need a large set of data as well because of complexity. With large sets of data, it is also important to use more parallel computation to save time. In both cases however, GPUs are favored by researchers now, as GPUs deliver higher throughput (operations executed per second) with higher parallel execution of algorithms.

**Exhibit 7: A trained neural network views pictures differently from a human**

![Image](source: Nvidia)

**Deep learning use cases are plentiful**

There are many use cases that have been developed in the past 2-3 years. Facebook uses neural networks for its automatic tagging algorithms, Google for photo search, Amazon for product recommendations, Pinterest for home feed personalization, and Instagram for search infrastructure. In our view, the availability of open source frameworks and higher compute capability will likely lead to a deluge of new use cases. In our view, enterprises can alter the way they do business by leveraging deep learning.

Many startups are creating transferrable neural network algorithms that can be easily applied to a smaller data set and help smaller enterprises to benefit from AI. While enterprises continue to generate data, it is harder to find qualified data scientists who can leverage the data and come up with insights – the startups will likely alleviate this bottleneck. However, larger companies like Amazon, Google, Facebook, Microsoft, and Baidu will continue to drive innovation through organic efforts, and they may acquire smaller, nimbler startups to bolster their internal and external business objectives.

**Table 2: Deep learning use cases**

<table>
<thead>
<tr>
<th>Sound</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice recognition</td>
<td>UX/UI, Automotive, Security, IoT</td>
</tr>
<tr>
<td>Voice search</td>
<td>Handset maker, Telecoms</td>
</tr>
<tr>
<td>Sentiment analysis</td>
<td>CRM</td>
</tr>
<tr>
<td>Flaw detection (engine noise)</td>
<td>Automotive, Aviation</td>
</tr>
<tr>
<td>Fraud detection (latent audio artifacts)</td>
<td>Finance, Credit cards</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment analysis</td>
<td>CRM, Social media, Reputation mgmt</td>
</tr>
<tr>
<td>Augmented search, Theme detection</td>
<td>Finance</td>
</tr>
<tr>
<td>Threat detection</td>
<td>Social media, Government</td>
</tr>
<tr>
<td>Fraud detection</td>
<td>Insurance, Finance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log analysis/Risk detection</td>
</tr>
<tr>
<td>Enterprise resource planning</td>
</tr>
<tr>
<td>Predictive analysis using sensor data</td>
</tr>
<tr>
<td>Business and economic analytics</td>
</tr>
<tr>
<td>Recommendation engine</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video/Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial recognition</td>
</tr>
<tr>
<td>Image search</td>
</tr>
<tr>
<td>Machine vision</td>
</tr>
<tr>
<td>Photo clustering</td>
</tr>
<tr>
<td>Motion detection</td>
</tr>
<tr>
<td>Real time threat detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data centers, security, finance</td>
</tr>
<tr>
<td>Manufacturing, auto, supply chain</td>
</tr>
<tr>
<td>IoT, Smart Home, hardware manufacturing</td>
</tr>
<tr>
<td>Finance, Accounting, Government</td>
</tr>
<tr>
<td>E-commerce, Media, Social Networks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveillance</td>
</tr>
<tr>
<td>Social media</td>
</tr>
<tr>
<td>Automotive, aviation</td>
</tr>
<tr>
<td>Telecom, handset makers</td>
</tr>
<tr>
<td>Gaming, UX, UI</td>
</tr>
<tr>
<td>Security, airports</td>
</tr>
</tbody>
</table>

Source: BofA Merrill Lynch Global Research
Nvidia + TomTom mapping partnership – automotive use case

On Sept 28, Nvidia announced a partnership with TomTom that will allow TomTom to port and run localization and mapping software on Nvidia’s Drive PX 2 AutoCruise. End to end mapping in 3D format allows precise positioning of key landmarks. When a car is in self-drive mode, new images that feed into the AI network (Maps in this case) will be checked against the database and the in-house processor will make decisions on what the car has to do. For example, Baidu is working to create a cloud to autonomous car platform for Chinese and global car makers.

CEVA deep learning tool kit could enable mass market deployment

CEVA, a leading licensor of signal processing IP, recently released a deep learning toolkit that attempts at simplifying the development and deployment of deep learning systems for mass market embedded devices. The tool kit is optimized for CEVA-XM family of imaging and vision DSP (digital signal processors) and enables real time high quality image classification, object recognition and vision analytics. CEVA has been focused on automated driving as a use case with the key value proposition that a DSP can dramatically reduce the power consumption of the overall system while providing complete flexibility. In our view, CEVA’s DSP processor will find use in many low-mid end cars that will eventually have level 5 ADAS features (Autonomous driving) while a combination of CPU/GPU could be used for high end cars with advanced features.

Exhibit 8: CEVA- XM Family of products for AI

Exhibit 9: CEVA vision solution

Source: CEVA, BoFA Merrill Lynch Global Research

Optimization frameworks – depends on use case

The biggest challenge in deep learning is that the training framework must be accurate, efficient, and have the ability to scale to process extremely large amounts of data. From a performance perspective, the time it takes for the machine to be trained to an acceptable level of accuracy. In other words, time-to-train has become critical. If a GPU is used, some training time challenges can be addressed, but it still depends on which neural network framework is used, the quality of data available and the quality of network optimization.

Given that deep learning is a new field, many frameworks (20+) have been developed or are in development. While some frameworks are universal, many have a specific use case in mind. Some are open source (Google’s Tensor flow, Berkeley’s Caffe, Microsoft’s CNTK) and are gaining more popularity.
Use cases dictate which framework will be used and developer’s preference for coding language (C++, Python etc.) will dictate which framework is adopted. In our view, Torch/Theano appear to be good frameworks to conduct research on deep learning itself while Caffe/Tensor flow will be good frameworks for mass adoption of deep learning. In particular, Tensor flow allows for modification of the number of layers during the training, something other framework currently don’t have.

In the end, each potential use case will likely have an optimized neural network, but the timeline for adoption of these deep learning networks by enterprises still depends on the economic value derived. Some use cases like sophisticated artificial intelligence device (a self-driving car) or service offering (Apple Siri/Amazon Echo) appear to have a strong business case (at least as consumers view it) but in most cases, the deployment will likely be transparent to the end user – for example: insurance fraud detection improvement.

Summary of frameworks
Both Nvidia and Intel have built up libraries that can make it easy for developers to optimize neural networks using each company’s unique hardware. OpenCL is another framework that has been created for accelerating algorithms on heterogeneous architectures (FPGA, DSP, and GPU).

CUDA based deep neural networks (CuDNN)
This is a GPU-accelerated library of primitives for deep neural networks. cuDNN provides highly tuned implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers. It allows the developers to focus on training neural networks and developing software applications rather than spending time on low-level GPU performance tuning.

It is important to note that CuDNN is in a different category from the other tools mentioned below. It is a companion library, not a framework itself. Its various function calls were developed in collaboration with the various frameworks.

Intel Math Kernel Library (MKL)
MKL for Deep Neural Networks is an open source performance library for Deep Learning applications that will run on Intel architecture. This contains highly optimized building blocks intended to accelerate compute intensive parts of DL frameworks such as Caffe, Tensorflow, Theano and Torch. The library is implemented in the C++ language but is compatible with Python/Java as well.

DeepBench
DeepBench is an open source benchmarking tool that measures the performance of basic operations involved in training deep neural networks. Baidu Research created this benchmarking tool to help compare the performance of various hardware platforms when operations are executed using neural network libraries. As discussed above, convolutions make up most of the floating point operations in networks that work on images and videos as well as speech/natural modelling and as such, it is only one of the most important layer. DeepBench helps researchers/engineers/hardware vendors to understand how different operations and/or workloads impact the performance of the model.
Open Computing Language (OpenCL)
OpenCL is a framework for writing programs that execute across heterogeneous platforms consisting of central processing units (CPUs), graphics processing units (GPUs), digital signal processors (DSPs), field-programmable gate arrays (FPGAs) and other processors or hardware accelerators. This is created by Khronos group in December 2008 and is open sourced. The frameworks addressed below (Caffe, Torch, Theano) offer partial support to OpenCL framework and as such the level of developer support is still not as high as that for CUDA. DeepCL is another framework that targets only deep learning. In our view, as the industry moves towards implementation of deep learning algorithms across devices (not just cloud/data center), we will see OpenCL gaining more traction.

Torch
Torch is a scientific computing framework that offers wide support for machine learning algorithms. Torch uses programming language called Lua. Currently, Torch is used by Facebook for research but is replaced by Caffe for deployments.

Caffe
Caffe is a deep learning open source framework made with expression, speed, and modularity in mind. It is developed by the Berkeley Vision and Learning Center (BVLC) and by community contributors. Caffe framework uses C++ programming launch along with Mathlab. It has been used in many computer vision applications and is amongst the popular deep learning framework now. Some key advantages include easy model sharing between users, unrestricted use of Alexnet, Googlenet and RNN. As a metric, we note that Caffe based neural networks with AlexNet can process over 60M images per day with a single NVIDIA K40 GPU.

Theano
Theano is a math expression compiler that efficiently defines, optimizes, and evaluates mathematical expressions involving multi-dimensional arrays. Theano is python friendly, with focus on general computation than performance. It is an excellent framework for developing neural network algorithms from scratch. It also has the ability to scale beyond one GPU. New versions of Theano have been optimized to work on GPU and Intel CPU with benchmarks available for drug discovery workloads, convolutional networks and sentiment analysis on movie review.

TensorFlow
TensorFlow is an open sourced software library for numerical computation using data flow graphs, developed by Google’s Machine Intelligence research organization. Google’s TensorFlow is general-purpose in nature and offers a clear, flexible interface to many kind of models and optimizations.
Microsoft CNTK

The Computational Network Toolkit (CNTK) is a unified deep-learning toolkit from Microsoft Research that makes it easy to train and combine popular model types across multiple GPUs and servers. CNTK implements highly efficient CNN and RNN training for speech, image and text data. Recently, Microsoft has shown that this framework scales with more GPUs and has much better performance than comparable frameworks listed above.

Accelerator total addressable market (TAM)

$10bn by 2020, 10x 2016 driven by cloud

We estimate the overall TAM for processors used in accelerated computing (HPC and Cloud) to be around $1bn in 2016 but will likely grow at a 75% CAGR till 2020 to $10bn. We note that HPC (or supercomputers) represents a majority of the TAM (63%) now but cloud will represent a majority by 2020, driven by 1) higher mix of cloud servers (50% of servers in 2020 will be in cloud) and at least one-third of the cloud servers will use accelerators in some form.

We estimate that GPU represents 60% of the overall TAM in 2016. We also estimate that inference TAM will likely be 20% of overall accelerator TAM by 2020 (or $1.5-2.0bn) but could likely be 3-5x training TAM by 2025. We base our calculation based on the following reasons.

- GPU is the main processor used in deep learning training applications and CPU is the main processor for inference applications (in cloud). We estimate the inference market is less than 5% of the overall TAM.
- We expect inference market size could be as high as 3x to 10x that of the training market – in our view, inference processor ASP will be at least 2x lower than training processor ASP, but each server will need more inference processors attached to it in order to run multiple streams of data and accurately make predictions in real time.
- Inference is still a nascent market, but will grow after 2020 and be the largest deep learning market segment by 2025. A large portion of the growth could come from embedded/mobile applications as well (smartphones, self-driving cars etc.)
• Per top 500.org June 2016 update, Nvidia has 72% unit share of the supercomputing market but about 50% of the revenue share.

Table 3: We expect overall accelerator TAM (GPU, Xeon Phi) to grow 10x from 2015 to 2020

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud</td>
<td>$125</td>
<td>$404</td>
<td>$841</td>
<td>$1,802</td>
<td>$5,025</td>
<td>$8,404</td>
<td>114%</td>
</tr>
<tr>
<td>HPC</td>
<td>$609</td>
<td>$686</td>
<td>$801</td>
<td>$1,005</td>
<td>$1,283</td>
<td>$1,706</td>
<td>26%</td>
</tr>
<tr>
<td>Total Tam</td>
<td>$734</td>
<td>$1,090</td>
<td>$1,642</td>
<td>$2,806</td>
<td>$6,307</td>
<td>$8,404</td>
<td>75%</td>
</tr>
</tbody>
</table>

Source: BofA Merrill Lynch Global Research estimates

Cloud TAM of $8.5bn in 2020, up from $0.40bn in 2016

Deep learning is still in its early stages with many use cases still in evaluation mode, but we believe that the TAM is huge. Currently, about 7% of the servers are used in deep learning activities based on Intel and Microsoft. Assuming an annual server unit TAM of 11.7mn in 2016 and about 30% of servers run in cloud, the number of servers running deep learning algorithms now is around 250,000.

One key assumption that drives our model is the GPU attach rate to servers and ASPs. In servers that run deep learning training algorithms, the typical CPU to GPU attach rate is 1 to 8 GPUs, but based on our research, the preferred combination is around 4 GPUs.

From a pricing perspective, we bake in $5,000-6,000 per GPU in 2015-16. With these assumptions, the GPU attach rate was 0.20-0.40% of the overall servers in 2015-16 or 5-6% of the cloud servers in 2015-16.

We assume that ASPs will increase, largely driven by the change in mix toward higher priced Pascal based P100/DGX1 supercomputers as well as a higher mix of Xeon Phi coprocessors.

While Intel expects deep learning to drive a majority of the workloads in 2020, from a server unit perspective, we conservatively assume that one-third of servers shipped in 2020 will be tasked to run deep learning algorithms (training or inference). Assuming Gartner’s server unit forecast, we estimate deep learning will account for about 2.5-3.0 million servers in 2020. We expect that training servers will account for majority of the servers and 80% of the TAM ($6.70bn), while inference servers will account for 20-30% of the servers and 20% of the TAM ($1.70bn).

Making these assumptions, we estimate GPU TAM to be $5bn in 2020 (60% of overall TAM). We estimate that Xeon Phi (Intel), FPGA (Intel, Xilinx) and ASIC solutions will drive the remaining $3-4bn.

Table 4: We expect accelerator TAM to grow to 8.40bn in 2020, with GPU representing 60% of TAM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Server (Gartner)</td>
<td>9.9</td>
<td>10.1</td>
<td>11.1</td>
<td>11.7</td>
<td>12.1</td>
<td>12.6</td>
<td>13.1</td>
<td>13.6</td>
<td>4%</td>
</tr>
<tr>
<td>Cloud %</td>
<td>15%</td>
<td>20%</td>
<td>25%</td>
<td>30%</td>
<td>35%</td>
<td>40%</td>
<td>45%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Cloud servers (mn units)</td>
<td>2.8</td>
<td>3.5</td>
<td>4.2</td>
<td>5.1</td>
<td>5.9</td>
<td>6.8</td>
<td>20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attach rate</td>
<td>3%</td>
<td>7%</td>
<td>10%</td>
<td>15%</td>
<td>30%</td>
<td>36%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>servers using accelerator (mn units)</td>
<td>0.08</td>
<td>0.24</td>
<td>0.42</td>
<td>0.76</td>
<td>1.76</td>
<td>2.46</td>
<td>97%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>accelerator ASP</td>
<td>$1,500</td>
<td>$1,650</td>
<td>$1,980</td>
<td>$2,376</td>
<td>$2,851</td>
<td>$3,421</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>$125</td>
<td>$404</td>
<td>$841</td>
<td>$1,802</td>
<td>$5,025</td>
<td>$8,404</td>
<td>132%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU attach rate</td>
<td>5.0%</td>
<td>5.5%</td>
<td>6.0%</td>
<td>6.5%</td>
<td>7.0%</td>
<td>7.5%</td>
<td>113%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Servers with GPU</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.12</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU ASP</td>
<td>5500</td>
<td>6000</td>
<td>6250</td>
<td>6500</td>
<td>6750</td>
<td>7000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU TAM</td>
<td>91</td>
<td>323</td>
<td>637</td>
<td>1,282</td>
<td>3,331</td>
<td>5,158</td>
<td>124%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: BofA Merrill Lynch Global Research estimates

High performance computing (HPC) TAM of $1.7-1.8bn by 2020

At a high level, HPC computing involves running algorithms that take high levels of computations. The measure is always floating point operations per second or FLOPs. The world’s fastest supercomputer (Sunway Taihulight, China) can perform 93 Peta FLOPs (PFLOPS) or a quadrillion floating point operations per second. This
supercomputer has 40,960 nodes with a combined total of 10.6 million computing cores. In terms of use cases, this supercomputer like many others is used for advanced manufacturing (CAE, CFD), earth system modeling and weather forecasting, life science, and big data analytics.

Top500.org maintains a list of the top 500 most powerful supercomputers ranked by the processing power, but there are many less powerful supercomputers that are not captured. However, the website provides information that help estimate overall TAM.

Based on the website data, the average performance of all the top 500 supercomputers has slowed since 2013 but is still growing at an impressive 55% YoY rate. In 2016, 95 supercomputer systems had more than 1PTLOPs of performance, which is up from 81 systems six months ago (top500.org reports data every six months – June and November). The growth in average performance drives the number of processor cores needed.

While 90% of the supercomputers use Intel Xeon processor currently for normal compute operations, only 20% of these supercomputers use accelerators (GPU, Xeon Phi) to accelerate algorithms. Interestingly, accelerator usage has doubled since 2013 (only 10% of supercomputers used accelerators in 2013) and in our view, will likely double to reach close to 50% of supercomputers by 2020. Our analysis also indicated that Nvidia has commanding market share of the accelerator market (72% market share as of June 2016), but we expect Intel to remain a strong competitor in HPC.

**Chart 5: Nvidia has 72% unit share of the accelerator market in HPC applications**

![Nvidia, 72%
Intel, 28%](chart5.png)

Source: Top500.org June 2016 report, BofA Merrill Lynch Global Research

**HPC Industry growth rates**

**GPU revenue per core grew at 20% CAGR from 2010-15**

We analyzed the HPC market using the number of cores used in order to analyze industry growth rates. The overall unit share (or attach rate to supercomputer servers) for NVIDIA has continued to shrink from 90% in 2011 to 72% in 2016 June, while market share by processing core declined from 81% to 28%. Even with this reduction, Nvidia has been able to grow GPU cores at a 12% CAGR since 2012, slightly faster than the growth in supercomputer attach rate.
Table 5: Nvidia’s unit mkt share is 72% but core mkt share is 25-30%; rev/shr grew at 19% CAGR

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of supercomputers</td>
<td>9</td>
<td>35</td>
<td>50</td>
<td>38</td>
<td>50</td>
<td>66</td>
<td>67</td>
<td>10%</td>
</tr>
<tr>
<td>YoY%</td>
<td>289%</td>
<td>43%</td>
<td>-24%</td>
<td>32%</td>
<td>32%</td>
<td>29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market share</td>
<td>53%</td>
<td>90%</td>
<td>81%</td>
<td>72%</td>
<td>65%</td>
<td>63%</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td># of cores (mn)</td>
<td>0.44</td>
<td>0.69</td>
<td>1.55</td>
<td>1.85</td>
<td>2.27</td>
<td>2.18</td>
<td>2.62</td>
<td>12%</td>
</tr>
<tr>
<td>YoY%</td>
<td>57%</td>
<td>124%</td>
<td>19%</td>
<td>23%</td>
<td>-4%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market share %</td>
<td>68%</td>
<td>81%</td>
<td>31%</td>
<td>32%</td>
<td>23%</td>
<td>28%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nvidia data center revenue</td>
<td>$128</td>
<td>$191</td>
<td>$317</td>
<td>$339</td>
<td>$382</td>
<td>$600</td>
<td>$782</td>
<td>$1,283</td>
</tr>
<tr>
<td>Cloud revenue</td>
<td>6</td>
<td>19</td>
<td>35</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVDA HPC revenue</td>
<td>122</td>
<td>172</td>
<td>282</td>
<td>289</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVDA HPC revenue YoY%</td>
<td>41%</td>
<td>64%</td>
<td>136%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVDA Revenue/core</td>
<td>$0</td>
<td>$0</td>
<td>$78</td>
<td>$93</td>
<td>$124</td>
<td>$133</td>
<td></td>
<td>19%</td>
</tr>
</tbody>
</table>

Source: Top500.org, BofA Merrill Lynch Global Research

We estimate that Nvidia has been able to grow its revenue per core at an impressive 19% CAGR, contributing to the overall HPC revenue growth of 33%. Nvidia launched its Pascal architecture based GPU in April 2016 but it will likely ramp up in late 2016 and drive growth. Given the strong growth in FLOPS in the past (55% CAGR) and the continued demand for higher performing cores, we think the overall industry demand for processors can continue to grow at a 20% CAGR. We note that per our analysis, Intel has a higher revenue per core in HPC ($800) which is about 7x that of Nvidia’s.

High performance computing TAM – $1.7-$1.8bn by 2020

In 2015, Intel’s HPC revenue was 25% of overall DCG revenue (our estimate) or $4bn. Given that 90% of the supercomputing systems use Intel’s Xeon processor for normal compute operations but Nvidia owns a majority share of the accelerated computing market, we assume that 90% of the revenue also comes from Xeon processors with 10% ($350-400mn) from Xeon Phi accelerators.

About 20% of the top 500 supercomputer systems use accelerators (up from only 10% in 2013) and currently, Nvidia has 72% unit share of this accelerated computing systems (Intel has the rest). If we assume that Nvidia’s HPC revenue is around 50% of its reported data center revenue, we estimate $325mn in HPC revenue for Nvidia in 2016. In our analysis, we assume that the revenue per core for GPU and Xeon Phi grows at a 10-15% CAGR, in-line with the historical growth rate. While 50%+ attach rate is possible, we assume that 40% of the supercomputers will use an accelerator by 2020. We also assume that Xeon Phi (co-processor) could reach 50% share of the HPC accelerator market. Net-net, we estimate HPC TAM of $1.7-$1.8bn by 2020.

Table 6: Accelerator TAM can grow at 25-30% till 2020 and reach close to $2bn

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Supercomputer units with acceleration</td>
<td>75</td>
<td>104</td>
<td>93</td>
<td>106</td>
<td>131</td>
<td>161</td>
<td>201</td>
<td>21%</td>
</tr>
<tr>
<td>% of top 500 supercomputers</td>
<td>15%</td>
<td>21%</td>
<td>19%</td>
<td>21%</td>
<td>26%</td>
<td>32%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td># of supercomputers with GPU</td>
<td>61</td>
<td>66</td>
<td>67</td>
<td>71</td>
<td>81</td>
<td>92</td>
<td>104</td>
<td>12%</td>
</tr>
<tr>
<td>GPU % of supercomputers with accelerators</td>
<td>57%</td>
<td>63%</td>
<td>72%</td>
<td>67%</td>
<td>62%</td>
<td>57%</td>
<td>52%</td>
<td></td>
</tr>
<tr>
<td>GPU core (mn)</td>
<td>2.3</td>
<td>2.2</td>
<td>2.5</td>
<td>2.7</td>
<td>3.2</td>
<td>3.7</td>
<td>4.3</td>
<td>14%</td>
</tr>
<tr>
<td>GPU core per system</td>
<td>45355</td>
<td>33030</td>
<td>37612</td>
<td>38740</td>
<td>39515</td>
<td>40305</td>
<td>41112</td>
<td>2%</td>
</tr>
<tr>
<td>Rev/core ($)</td>
<td>$124</td>
<td>$133</td>
<td>$133</td>
<td>$134</td>
<td>$150</td>
<td>$172</td>
<td>$207</td>
<td>12%</td>
</tr>
<tr>
<td>GPU TAM</td>
<td>$282</td>
<td>$289</td>
<td>$334</td>
<td>$367</td>
<td>$480</td>
<td>$636</td>
<td>$888</td>
<td>28%</td>
</tr>
<tr>
<td>GPU market share</td>
<td>47%</td>
<td>49%</td>
<td>46%</td>
<td>48%</td>
<td>50%</td>
<td>50%</td>
<td>52%</td>
<td></td>
</tr>
<tr>
<td>Xeon Phi</td>
<td>25</td>
<td>27</td>
<td>26</td>
<td>35</td>
<td>50</td>
<td>69</td>
<td>96</td>
<td>39%</td>
</tr>
<tr>
<td>Unit market share</td>
<td>33%</td>
<td>26%</td>
<td>28%</td>
<td>33%</td>
<td>38%</td>
<td>43%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>Cores shipped (mn)</td>
<td>4.6</td>
<td>5.0</td>
<td>5.0</td>
<td>5.4</td>
<td>6.0</td>
<td>6.6</td>
<td>7.2</td>
<td>10%</td>
</tr>
<tr>
<td>Revenue/core</td>
<td>$65</td>
<td>$71</td>
<td>$80</td>
<td>$88</td>
<td>$98</td>
<td>$113</td>
<td>$132</td>
<td>12%</td>
</tr>
<tr>
<td>Xeon Phi TAM</td>
<td>$320</td>
<td>$352</td>
<td>$434</td>
<td>$525</td>
<td>$646</td>
<td>$818</td>
<td>$21%</td>
<td></td>
</tr>
</tbody>
</table>

Total accelerator TAM | $609 | $686 | $801 | $1,005 | $1,283 | $1,706 | 26% |

Source: BofA Merrill Lynch Global Research
A deep dive into accelerated computing processors

Deep learning depends on compute capacity, so the semiconductor discussion centers on CPU and GPU. More importantly, the market need is for a processor that can accelerate training activity as well as improve latency during the inference step. Essentially, the industry needs an “accelerator” to improve the performance beyond that offered by general purpose processing. We have discussed the differences between a serial and parallel processor. An accelerator is a parallel processor so has a much better throughput (number of operations per second). For example – in the past, training a neural network took months, but the latest GPU can enable training within a month or days, and in some cases even hours.

Exhibit 11: P40 GPU from Nvidia for Deep learning inference workload

Exhibit 12: DGX1 supercomputer from Nvidia with 8 P100 GPUs

Knights Mill, part of the Xeon Phi product family from Intel, will be launched in 2017 and provide another alternative to GPU but without accurate benchmarks from server OEMs, it is unclear whether Xeon Phi will be a direct replacement of GPU. In the case of inference, there are alternatives either available now or in development. The reason is simple – for inference, the processing needs are much less and an ASSP or ASIC can be designed to do the work by leveraging the trained neural network. Also, power consumption is a key factor for embedded and portable applications. While Nvidia has released Tegra based Jetson TX1 for embedded applications, it will likely face strong competition from just about every other semiconductor processor companies (Intel, Qualcomm, TI, NXP etc.)
Addressing key controversies in deep learning

Given that deep learning is one of the key growth markets, many companies have jumped into the fray to capture a share of the growing pie. Intel and Nvidia are the key competitors when it comes to deep learning accelerators. FPGA has gained popularity as a the mass market processor that can do inference at lower power and has been used for successfully introducing new wired/wireless products (routers, embedded systems) in the past. Microsoft is probably the company who has publicly spoken about FPGA’s benefits but with Intel’s brand and marketing efforts behind FPGA, it is possible that FPGA sees stronger traction outside of Microsoft. We will address some of the controversies below.

GPGPU too clunky for inference, need FPGA or Xeon Phi

GPGPU or General purpose GPU is a perfect tool for running training as discussed above. However, when it comes to inference, throughput is not a concern but latency is. While Nvidia has lower throughput, lower power options for inference applications, in our view, a GPU is not absolutely required for inference. You can use an FPGA, ASIC (Google’s tensor), ASSP (embedded processor from TI, NXP) or just a Core i7 CPU from Intel.

Can Intel’s data center processors satisfy all market demand?

Nvidia was early to the market with products (Kepler and Maxwell generation GPGPU) that addressed the immediate market need (training). In 2016, Nvidia released its Pascal based GPU solutions for gaming and deep learning. The skus specifically for deep learning training include P100 and DGX-1 (supercomputer, which consists of 8 P100). For inference, P40 and P4 are the two skus. All these skus are expected to ramp in Q4 2016 or early 2017. These skus will replace K80/K40 (Kepler architecture, 2012) and M40/M4 (Maxwell architecture, 2014) skus from prior generations over time. We try to address the pushbacks on the long term business case using GPU for deep learning. 1) the products are pricey ($5000 per K80 GPU); 2) Nvidia’s GPGPU products also consume more power (250-300W) vs a CPU (100-150W) and a FPGA (<30W) and 3) GPU performance will degrade due to time lag seen in accessing data from memory. We think otherwise.

Intel solution catching up but Nvidia products still the cost per performance leader

GPGPU’s performance has improved by 60-70x from 2013 levels to around 5.3x TFLOPs (Tera Floating point operations per second, double precision = 64 bits), 21.2 TFLOPs (half precision = 16 bits) vs 3 Teraflops for Xeon Phi Knights Mills – 1.7x Xeon Phi (Exhibit 4). While Nvidia hasn’t announced the price for P100 products, our research suggests pricing could range $5,800-9,500 (source: Microway.com). Higher-end P100 has proprietary interconnects between GPU (NVLink) and higher on die high bandwidth memory (HBM). Taking the mid-range product (16GB memory, PCIe interconnect, Price
estimate $7,374, 18.7 TFLOPs) and calculating the $ per TFLOP, yields a cost of $394. The lowest end P100 with 12GB of memory would cost $313 per TFLOP and still deliver 18.7 TFLOPs of performance.

If we compare this against Intel’s high end Xeon Phi product (7290, Knights Landing) with a performance rating of 13.8 TFLOPs (same memory density as P100) and at a cost of $6300, we get to a cost of $456. We note that Intel has not released specifications for Knights Mill (next gen) and as such we used Knights Landing details. However, we expect Knights Mill to have better performance than 2nd gen Knights landing (and higher price tag) but still have overall performance below Nvidia’s P100. As such, the $ per TFLOP will likely remain $400-500 unless Intel is able to tweak up performance by adding a few more cores and increasing clock speed.

Exhibit 14: Deep learning training performance for various GPU products

Intel’s Xeon Phi solution does cost more on a performance to power basis. Xeon Phi knights landing power consumption ranges from 215 to 245W while P100’s power consumption ranges from 250 to 300W. PCIe interconnects have a max power limitation of 250W but with Nvidia’s NVLink connectivity, power consumption can go up to 300W but performance increases by 13-14% as well. If we compare the $ per TFLOPs per watt between the products compared above, we get $1.60 per TFLOP per W for midrange P100, $1.25 per TFLOP per W for low end and $1.50 per TFLOP per W for high end P100. This compares with Intel high end Knights landing of $1.86 per TFLOP per W.

Historically, GPUs have been disadvantaged due to the time lag created by the back and forth flow of data between main memory, GPU memory and GPU execution cores. Xeon Phi solution takes care of this by using one CPU for performing the serial and parallel processing steps. However, Nvidia’s Pascal based P100 solves it by using NVlink technology, where each link can transfer data at a 20 Giga bytes per sec to a total bandwidth of 160GB/s, 5x higher than the current fastest PCIe bandwidth (Exhibit 6). Also, Nvidia has also further simplified memory access by moving to Unified memory, where the CPU and GPU will share the same memory pool (done through CUDA system software). Once P100 is adopted, we see the advantage tilting towards Nvidia products, although Xeon Phi (Knights Mill) could see higher market penetration given that Intel controls 99.5% of data center market.
GPGPU cannot scale efficiently compared with Xeon Phi
Based on a recent (June) benchmark study conducted by Intel, Intel’s knight landing can scale better than GPUs. Scaling here refers to the number of machines that can run the same neural network code. At about the same time, Baidu research scientist Greg Diamos presented his results on GPU scaling, stating that it is possible to efficiently scale GPUs (8 GPUs to 128 GPUs). According to Diamos, it is possible to improve efficiency by 30x+ by scaling. Benchmark studies generally are biased toward the company presenting the numbers, and we will have to wait to see the results from server OEMs (likely available in Q1 2017).

Google tensor processing unit (TPU) will limit the need for CPU/GPU
Google tensor processing unit or TPU is essentially an ASIC that can be designed to address a certain function in a deep learning system. This product was used in Alpha go, Google search and Google street view. However, Google expects broader adoption of the TPU along with its Tensorflow framework for many different AI applications. In our view, this product targets the inference market, which is still at an early stage in terms of adoption. Like in other applications, custom ASICS are always threat to general purpose processors (ASIC vs FPGA in networking). However, it is still too early to tell if Google will see commercial success of TPU outside of its own AI applications but we think Google still needs high performance GPU or Xeon Phi for training the deep learning algorithms.

FPGA is likely a strong alternative to CPU/GPU
Field programming gate array (FPGA) is a programmable logic device that can be used for a variety of end market applications in industrial, auto, wired/wireless comms, aerospace/defense and others.

Pros and cons of using a FPGA vs ASIC, GPU
An FPGA allows for running a simulation an ASIC in many networking applications. The key advantage of using FPGA in any application vs an ASIC is: 1) faster time to market – no layout, mask steps; 2) no upfront non-recurring expenses (NRE); 3) simpler design.
cycle (software handles routing); 4) more predictable project cycle (due to elimination of potential re-spins, wafer capacities, etc; and 5) field programmability – customers can program a FPGA remotely even on a daily basis. There are many downsides as well. The peak throughput for an FPGA is considerably less than for a GPU. It is also harder to program than a GPU accelerator but this could be because developers haven’t really focused on FPGA as an alternative for GPU.

Exhibit 17: Intel Xeon + FPGA in data center

Nascent data center market opportunity
The new market identified for the FPGA is the data center where it is expected to function as an accelerator, similar to a GPU. Microsoft was the first to consider the use of FPGA for acceleration largely driven by the need to build a scalable deep learning infrastructure but still uses GPU in many servers (about 5-6% of servers run deep learning). Baidu has also evaluated FPGA as a potential method to accelerate SQL (database) at scale. In our view, large companies evaluate multiple options over a period of time and optimize the hardware based on workloads. FPGA could be one of them.

FPGA could exploit potential GPU limitations
We agree that there are some limitations in using GPU across the board in data center. Currently, only a few applications like Facebook’s image recognition can potentially keep GPUs fully utilized and lower the cost of operation. In many applications, GPUs may not be completely utilized and in some cases, the demand needs exceed the limit of GPUs added to a server or a network of servers (limited scaling).

Adding GPUs in every server in the world is also not an option as it will likely increase overall power consumption without necessarily increasing the throughput – we note that CPUs already consume 200W of power and adding 250-300W for every server is not an optimal solution. FPGA might just be turn out to be the solution that fits in between specialized hardware like a GPU accelerator (P100 system from Nvidia) and general purpose hardware (Intel Xeon server) but we note that the power consumption of an FPGA is dependent on the frequency at which it operates and the type of workload.

Research papers (CNNLab: Novel framework for Neural Networks using FPGA/GPU by MaoHua Zhu, UCSB) show that GPU is energy efficient and higher operations per unit of energy metric than FPGA while running certain neural network calculations.

While it is hard to estimate the overall TAM for FPGA, we assume that FPGA could add another $300-400mn TAM (15% attach rate to deep learning servers and $200 ASP). We assume that CPUs can do inference workloads in servers and will be aided by FPGA/ASICs as needed. Intel believes that FPGA will be in one-third of all cloud servers by 2020. Based on Microsoft’s recent announcement that all new Azure servers have a FPGA accelerator card included and given that Azure accounts for 10-15% of total
industry servers, it appears likely that FPGA could be used in one-third of the servers by 2020. However, Microsoft has been working on building out an alternative to GPU for scale out workloads since 2011 (5 years of development) and we haven’t seen any other large companies who have implemented a FPGA based accelerator strategy.

**Exhibit 18: Intel E5, 12 core server with 2 FPGA boards**

Memory bandwidth is key for heterogeneous hardware to outperform GPU
Whichever processor is picked to run the neural network, large memory bandwidth is very important. Essentially, the ideal compute system will be one which can be software programmed to run on different types of networks but with power efficiency close to that of a custom ASIC. The key to high power efficiency is the memory bandwidth. In the case of a typical GPU, the algorithm accesses the GDDR5 memory in the GPU card many times in order to complete a math operation. This tends to clog the memory bandwidth even if the GPU is only utilized 5-10%. Some research has shown it is possible to deepen the pipeline of memory requests and minimize the amount of time spent reaching out to the outside memory –hybrid memory cube or HMC is the proposed solution (Exhibit 8).

**Exhibit 19: FPGA + hybrid memory cube allows for**

OpenCL framework for heterogeneous architecture
OpenCL is an open source, standardized framework for algorithm acceleration on heterogeneous architectures. Programs written in OpenCL can be executed transparently on GPPs, GPUs, DSPs, and FPGAs largely driven by the use of C based
language. Similar to CUDA, OpenCL provides a standard framework for parallel programming, as well as low-level access to hardware. While both CUDA and OpenCL provide similar functionality to programmers. The major difference between OpenCL and CUDA is the ownership of the frameworks – CUDA is a proprietary framework created by Nvidia while OpenCL is open source, royalty-free, and is maintained by the Khronos group. Starting 2013, both Xilinx and Altera (now Intel) adopted OpenCL for their devices and allowed for a much wider software developer ecosystem to develop.

Summary of various AI approaches
AI was born out of the need for improving user experience, and machine learning helped businesses get there. However, there are many ways to develop and implement an AI system. Many startups (Nervana, DeePhi, Wave computing) have cropped up in the past few years and have tried to develop solutions that will simplify the development and implementation of deep learning algorithms. More importantly, companies like IBM (TrueNorth) and Qualcomm (Zeroth) have also come up with alternative approaches to AI. We review a few of the approaches below. In our view, open source frameworks has helped remove the initial obstacle of creating an algorithm and has helped refocus the industry towards creating solutions for real life problems.

IBM TrueNorth
In 2010, IBM started to work with university partners on a quest to build a brain-inspired machine. At a high level, IBM was trying to develop a brain like capability into devices where computation is constrained by power and speed. A brain has around 100 trillion synapses (connection between neurons). In order to replicate and simulate a human brain, we will need 96 Blue Gene/Q (IBM supercomputer) racks. However, even with this firepower, the actual performance was found to be 1500 times slower and more importantly, was found to consume 12GW of power while the actual brain consumes only 20W.

Exhibit 20: IBM TrueNorth technology will help create a holistic computing intelligence system

IBM solved this problem by developing a neurosynaptic device with 4096 cores via an on-chip network to create TrueNorth – this device had 1 million neurons and 256 million synapses. This device consumes only 100mW of power and has a power density 20mW/cm2 vs 317 W/cm2 of a Pascal GP100. Essentially, IBM created a parallel, distributed, scalable and flexible architecture that integrates computation, communication and memory. However, in order to measure the inputs and outputs (which are just spikes of electrical activity), IBM had to create a new algorithm called SyNAPSE. All in, if IBM were to be successful in scaling this device to be closer to a human brain (100 trillion synapses), we can see the architecture solving problems in vision, audition, multi-sensory fusion and could likely be integrated into a smartphone or supercomputer (IBM Watson).

Qualcomm Zeroth – taking deep learning/AI to consumer devices
At a high level, Qualcomm’s zeroth platform mimics the nervous system and brain of a human being and helps push forth embedded cognition through brain inspired computing. Essentially, Qualcomm hopes to convert IBM’s TrueNorth technology into a commercial technology that can be adopted in various consumer applications.

Recently, Qualcomm released a software development kit (SDK) that can allow manufacturers and companies to run limited deep learning programs locally on the
devices. As discussed in this note, most of the deep learning activities currently happen in the cloud now but in the future, broader adoption will likely drive the need for localized computing and real time analytics. The Zeroth platform enables a hardware to anticipate the user needs and share the perception of the world naturally.

**Exhibit 21: Zeroth platform uses perception, reasoning and action reflexes of human beings**

[Diagram showing perception, reasoning, and action]

Qualcomm’s goal is to develop a cognitive computing platform that can adapt and learn from every move made by the user and will ultimately simplify and enrich the user’s daily life. One such use case could be how humans buy products (custom manufactured shows based on a 3D model of the feet) or how to navigate foreign country (translate street signs, speak with locals, etc).

Qualcomm has developed a neural processing unit (NPU) that will reside side by side in future processors for devices (within SoC or system on chip). In order to use this chip, Qualcomm is working with neuroscientists to create mathematical models of the human brain’s biological neuron behavior and perform the processing of real time information in the NPU. The new network called the spiking neural network (SNN) is extremely efficient in how it encodes and transmits information through the brain. Qualcomm hopes to harness this ability for smartphones and other smart devices like robots.

**Exhibit 22: Electrical activity in a real neuron can be replicated to deliver efficient real time analytics**

[Diagram showing real neuron and electrical activity]

Nervana (now Intel) – custom ASIC solution for training/inference

Nervana, which was acquired by Intel, has been working on a custom ASIC alternative that can replace GPU in deep learning. Nervana was an artificial intelligence software company based in Silicon Valley. The company provided a full stack software as a
service platform called Nervana Cloud that enabled businesses to develop customer
deep learning software. Nervana uses an open source deep learning framework called
Neon and is noted to deploy a more efficient algorithm than current frameworks (Caffe,
Theano, Torch and TensorFlow). The company had built its framework to run on Nvidia’s
Titan X GPU but more recently, had developed a custom ASIC solution called a Nervana
engine (using TSMC 28nm technology) which is expected to perform 10x better than
Nvidia’s Maxwell architecture GPUs.

The new solution is expected to have a workaround to a key issue faced with GPU –
communication between GPU cores and memory. The ASIC chip allows for transfer data
on chip through the use of software, does not involve transferring data to memory or
cache and as such improves latency. Nervana believes that the ASIC chip can perform 5-
6x better than Nvidia’s latest Pascal chip. Nvidia has tried to solve the issue by using
NVlink (a series of high speed links between GPUs) but it is yet to be evaluated by
customers.

Nervana’s solution can run both training and inference on the same chip while Nvidia’s
solution has two different solutions (P100 for training and P4/P40 for inference). While
Nvidia’s GPU solution is compelling from a performance point of view, currently, CPU is
the main processor used for inference applications. With Intel’s purchase of Nervana, it
is clear that the future hardware for deep learning inference applications will use a mix
of CPU and ASICs.

**DeePhi – a FPGA solution for deep learning**

DeePhi’s solution hinges on one simple argument – deep learning frameworks are
rapidly evolving and as such, the current hardware solutions cannot keep up with the
framework requirements. As such, DeePhi decided to use a reconfigurable device like an
FPGA to develop a solution (hardware and software) for deep learning. According to
DeePhi, its deep learning processing unit (DPU) can deliver an order of magnitude higher
energy efficiency over GPU on image recognition and speech detection, largely driven by
the co-design approach involving Hardware, Software and Algorithm. However,
programming in FPGA is not as easy – it took one month for the team to optimize the
program and even then, the performance was not as expected. In any case, the company
is in the process of testing its products and benchmarking against key deep learning
workloads. In our view, DeePhi will be one of the many types of solutions developed to
address deep learning use cases but like Nervana, it will likely get acquired (if
technology is sound) prior to mass deployments.

**Wavecomputing (TensorFlow) – parallel computing made easy**

Wave computing has developed an ultralow precision hardware that can outperform
even the most powerful deep learning GPU system (DGX-1 from Nvidia) in training and
inference. The company isn’t trying to replace a CPU or GPU but is positioning itself as
an interesting alternative to companies who will be using TensorFlow deep learning
architecture. The company’s goal is to make TensorFlow models run far faster out of the
box with little help from the user but at a much lower price relative to Nvidia’s DGX-1
appliance. Essentially, Wave computing will be selling a full system for both training and
inference and it will be a plug and play node in a data center network with native
support of TensorFlow. For now, the company plans to commercialize its first products
starting Q2 2017.
Exhibit 23: Wave computing uses 8bit RISC based computing with large memory

Source: Nextplatform.com
Disclosures

Important Disclosures

BoFA Merrill Lynch Research Personnel (including the analyst(s) responsible for this report) receive compensation based upon, among other factors, the overall profitability of Bank of America Corporation, including profits derived from investment banking. The analyst(s) responsible for this report may also receive compensation based upon, among other factors, the overall profitability of the Bank’s sales and trading businesses relating to the class of securities or financial instruments for which such analyst is responsible.

Other Important Disclosures

From time to time research analysts conduct site visits of covered issuers. BoFA Merrill Lynch policies prohibit research analysts from accepting payment or reimbursement for travel expenses from the issuer for such visits.

Prices are indicative and for information purposes only. Except as otherwise stated in the report, for the purpose of any recommendation in relation to: (i) an equity security, the price referenced is the publicly traded price of the security as close of business on the day prior to the date of the report or, if the report is published during intraday trading, the price referenced is indicative of the traded price as of the date and time of the report; or (ii) a debt security (including equity preferred and CDS), prices are indicative as of the date and time of the report and are from various sources including Bank of America Merrill Lynch trading desks.

The date and time of completion of the production of any recommendation in this report shall be the date and time of dissemination of this report as recorded in the report timestamp.

Offices of MLPF&S or one or more of its affiliates (other than research analysts) may have a financial interest in securities of the issuer(s) or in related investments.

BoFA Merrill Lynch Global Research policies relating to conflicts of interest are described at http://go.bofa.com/coi.

"BoFA Merrill Lynch" includes Merrill Lynch, Pierce, Fenner & Smith Incorporated ("MLPF&S") and its affiliates. Investors should contact their BoFA Merrill Lynch representative or BoFA Merrill Lynch Global Wealth Management financial advisor if they have questions concerning this report. "BoFA Merrill Lynch" and "Merrill Lynch" are each global brands for BoFA Merrill Lynch Global Research.

Information relating to Non-US affiliates of BoFA Merrill Lynch and Distribution of Affiliate Research Reports:

MLPF&S distributes, or may in the future distribute, research reports of the following non-US affiliates in the US (short name: legal name, regulator):

- Merrill Lynch (South Africa): Merrill Lynch South Africa (Pty) Ltd, regulated by The Financial Service Board; Merrill Lynch (UK): Merrill Lynch International, regulated by the Financial Conduct Authority (FCA) and the Prudential Regulation Authority (PRA);
- Merrill Lynch Equities (Australia) Limited, regulated by the Australian Securities and Investments Commission (ASIC);
- Merrill Lynch (Asia) Limited, regulated by the Hong Kong Securities and Futures Commission (HKSCF);
- Merrill Lynch (Singapore): Merrill Lynch (Singapore) Pte Ltd, regulated by the Monetary Authority of Singapore (MAS);

This research report has been approved for publication and is distributed in the United Kingdom (UK) to professional clients and eligible counterparties (as each is defined in the rules of the FCA and the PRA) by MLI (UK) and Bank of America Merrill Lynch International Limited, which are authorized by the PRA and regulated by the FCA and the PRA, and is distributed in the UK to retail clients (as defined in the rules of the FCA and the PRA) by Merrill Lynch International Bank Limited, London Branch, which is authorized by the Central Bank of Ireland and subject to limited regulation by the FCA and PRA - details about the extent of our regulation by the FCA and PRA are available from us on request; has been considered and distributed in Japan by Merrill Lynch (Japan), a registered securities dealer under the Financial Instruments and Exchange Act in Japan, is issued and distributed in Hong Kong by Merrill Lynch (Hong Kong) which is regulated by HKSCF (research reports containing any information in relation to, or advice on, futures contracts are not intended for issuance or distribution in Hong Kong and are not directed to, or intended for issuance or distribution to, or use by, any person in Hong Kong), is issued and distributed in Taiwan by Merrill Lynch (Taiwan); is issued and distributed in Singapore to institutional investors and/or accredited investors (each as defined under the Financial Advisers Regulations) by Merrill Lynch International Bank Limited (Merchant Bank) (MLLIBMB) and Merrill Lynch (Singapore) (Company Registration Nos F06872E and 198602883D respectively). MLLIBM and Merrill Lynch (Singapore) are regulated by MAS. Bank of America N.A., Australian Branch (ABN 06 874 531), AFSL License 412901 (BANA Australia) and Merrill Lynch Equities (Australia) Limited (ABN 65 006 276 795), AFSL License 235132 (MLEA) distribute this report in Australia only to Wholesale clients as defined by s.761G of the Corporations Act 2001. With the exception of BANA Australia, neither MLEA nor any of its affiliates involved in preparing this research report is an Authorised Deposit-Taking Institution under the Banking Act 1959 nor regulated by the Australian Prudential Regulation Authority. No approval is required for publication or distribution of this report in Brazil and its local distribution is by Merrill Lynch (Brazil) in accordance with applicable regulations. Merrill Lynch (DIFC) is authorized and regulated by the DFSA. Research reports prepared and issued by Merrill Lynch (DIFC) are done so in accordance with the requirements of the DFSA conduct of business rules. Bank of America Merrill Lynch International Limited, Frankfurt Branch (BAMLFrankfurt) distributes this report in Germany and is regulated by BaFin.

This research report has been prepared and issued by MLPF&S and/or one or more of its non-US affiliates. MLPF&S is the distributor of this research report in the US and accepts full responsibility for research reports of its non-US affiliates distributed to MLPF&S clients in the US. Any US person receiving this research report and wishing to effect any transaction in any security discussed in the report should do so through MLPF&S and not such foreign affiliates. Hong Kong recipients of this research report should contact Merrill Lynch (Asia Pacific) Limited in respect of any matters relating to dealing in securities (and not futures contracts) or provision of specific advice on securities (and not futures contracts). Singapore recipients of this research report should contact Merrill Lynch International Bank Limited (Merchant Bank) and/or Merrill Lynch (Singapore) Pte Ltd in respect of any matters arising from, or in connection with, this research report.

General Investment Related Disclosures:

Taiwan Readers. Neither the information nor any opinion expressed herein constitutes an offer or a solicitation of an offer to transact in any securities or other financial instrument. No part of this report may be reproduced or quoted in any manner whatsoever in Taiwan by the press or any other person without the express written consent of BoFA Merrill Lynch.

This research report provides general information only. Neither the information nor any opinion expressed constitutes an offer or an invitation to make an offer, to buy or sell any securities or other financial instrument or any derivative related to such securities or instruments (e.g., options, futures, warrants, and contracts for differences). This report is not intended to provide personal investment advice and it does not take into account the specific investment objectives, financial situation and the particular needs of any specific person. Investors should seek independent financial advice before making any investment decision, including whether an investment conforms to their investment strategies discussed or recommended in this report and should understand that statements regarding future prospects may not be realized. Any decision to purchase or subscribe for securities in any offering must be based solely on existing public information on such security or the information in the prospectus or other offering document issued with such offering, and not in this research report. Securities and other financial instruments discussed in this report, or recommended, offered or sold by Merrill Lynch, are not insured by the Federal Deposit Insurance Corporation and are not deposits or other obligations of any insured depository institution (including: Bank of America, N.A.). Investments in general and, derivatives, in particular, involve numerous risks, including, among others, market risk, counterparty default risk and liquidity risk. No security, financial instrument or derivative is suitable for all investors. In some cases, securities and other financial instruments may be difficult to value or sell and reliable information about the value or risks related to the security or financial instrument may be difficult to obtain. Investors should note that income from such securities and other financial instruments, if any, may fluctuate and that price or value of such securities and instruments may rise or fall and, in some cases, investors may

Global Semiconductors | 02 October 2016

26
lose their entire principal investment. Past performance is not necessarily a guide to future performance. Levels and basis for taxation may change.

This report may contain a short-term trading idea or recommendation, which highlights a specific near-term catalyst or event impacting the issuer or the market that is anticipated to have a short-term price impact on the equity securities of the issuer. Short-term trading ideas and recommendations are different from and do not affect a stock’s fundamental equity rating, which reflects both a longer term total return expectation and attractiveness for investment relative to other stocks within its Coverage Cluster. Short-term trading ideas and recommendations may be more or less positive than a stock’s fundamental equity rating.

BoFA Merrill Lynch is aware that the implementation of the ideas expressed in this report may depend upon an investor’s ability to “short” securities or other financial instruments and that such action may be limited by regulations prohibiting or restricting “shortselling” in many jurisdictions. Investors are urged to seek advice regarding the applicability of such regulations prior to executing any short idea contained in this report.

Foreign currency rates of exchange may adversely affect the value, price or income of any security or financial instrument mentioned in this report. Investors in such securities and instruments, including ADRs, effectively assume currency risk.

UK Readers: The protections provided by the U.K. regulatory regime, including the Financial Services Scheme, do not apply in general to business coordinated by BoFA Merrill Lynch entities located outside of the United Kingdom. BoFA Merrill Lynch Global Research policies relating to conflicts of interest are described at http://go.bofa.com/coi.

MLPF&S or one of its affiliates is a regular issuer of traded financial instruments linked to securities that may have been recommended in this report. MLPF&S or one of its affiliates may, at any time, hold a trading position (long or short) in the securities and financial instruments discussed in this report.

BoFA Merrill Lynch, through business units other than BoFA Merrill Lynch Global Research, may have issued and may in the future issue trading ideas or recommendations that are inconsistent with, and reach different conclusions from, the information presented in this report. Such ideas or recommendations reflect the different time frames, assumptions, views and analytical methods of the persons who prepared them, and BoFA Merrill Lynch is under no obligation to ensure that such other trading ideas or recommendations are brought to the attention of any recipient of this report.

In the event that the recipient received this report pursuant to a contract between the recipient and MLPF&S for the provision of research services for a separate fee, and in connection therewith MLPF&S may be deemed to be acting as an investment adviser, such status relates, if at all, solely to the person with whom MLPF&S has contracted directly and does not extend beyond the delivery of this report (unless otherwise agreed specifically in writing by MLPF&S). MLPF&S is and continues to act solely as a broker-dealer in connection with the execution of any transactions, including transactions in any securities mentioned in this report.

Copyright and General Information regarding Research Reports:

Copyright 2016 Bank of America Corporation. All rights reserved. iQmethod, iQmethod 2.0, iQprofile, iQtoolkit, iQworks are service marks of Bank of America Corporation. iQanalytics®, iQcst®, iQdatabase® are registered service marks of Bank of America Corporation. This research report is prepared for the use of BoFA Merrill Lynch clients and may not be redistributed, retransmitted or disclosed, in whole or in part, or in any form or manner, without the express written consent of BoFA Merrill Lynch. BoFA Merrill Lynch Global Research reports are distributed simultaneously to internal and client websites and other portals by BoFA Merrill Lynch and are not publicly-available materials. Any unauthorized use or disclosure is prohibited. Receipt and review of this research report constitutes your agreement not to redistribute, retransmit, or disclose to others the contents, opinions, conclusion, or information contained in this report (including any investment recommendations, estimates or price targets) without first obtaining expressed permission from an authorized officer of BoFA Merrill Lynch.

Materials prepared by BoFA Merrill Lynch Global Research personnel are based on public information. Facts and views presented in this material have not been reviewed by, and may not reflect information known to, professionals in other business areas of BoFA Merrill Lynch, including investment banking personnel. BoFA Merrill Lynch has established information barriers between BoFA Merrill Lynch Global Research and certain business groups. As a result, BoFA Merrill Lynch does not disclose certain client relationships with, or compensation received from, such issuers in research reports. To the extent this report discusses any legal proceeding or issues, it has not been prepared as nor is it intended to express any legal conclusion, opinion or advice. Investors should consult their own legal advisers as to issues of law relating to the subject matter of this report. BoFA Merrill Lynch Global Research personnel’s knowledge of legal proceedings in which any BoFA Merrill Lynch entity and/or its directors, officers and employees may be plaintiffs, defendants, co-defendants or co-plaintiffs with or involving issuers mentioned in this report is based on public information. Facts and views presented in this material that relate to any such proceedings have not been reviewed by, discussed with, and may not reflect information known to, professionals in other business areas of BoFA Merrill Lynch in connection with the legal proceedings or matters relevant to such proceedings.

This report has been prepared independently of any issuer of securities mentioned herein and not in connection with any proposed offering of securities or as agent of any issuer of any securities. None of MLPF&S, any of its affiliates or their research analysts has any authority whatsoever to make any representation or warranty on behalf of the issuer(s). BoFA Merrill Lynch Global Research policy prohibits research personnel from disclosing a recommendation, investment rating, or investment thesis for review by an issuer prior to the publication of a research report containing such rating, recommendation or investment thesis.

Any information relating to the tax status of financial instruments discussed herein is not intended to provide tax advice or to be used by anyone to provide tax advice. Investors are urged to seek tax advice based on their particular circumstances from an independent tax professional.

The information herein (other than disclosure information relating to BoFA Merrill Lynch and its affiliates) was obtained from various sources and we do not guarantee its accuracy. This report may contain links to third-party websites. BoFA Merrill Lynch is not responsible for the content of any third-party website or any linked content contained in a third-party website. Content contained on such third-party websites is not part of this report and is not incorporated by reference into this report. The inclusion of a link in this report does not imply any endorsement by or any affiliation with BoFA Merrill Lynch. Access to any third-party website is at your own risk, and you should always review the terms and privacy policies at third-party websites before submitting any personal information to them. BoFA Merrill Lynch is not responsible for such terms and privacy policies and expressly disclaims any liability for them.

Subject to the quiet period applicable under laws of the various jurisdictions in which we distribute research reports and other legal and BoFA Merrill Lynch policy-related restrictions on the publication of research reports, fundamental equity reports are produced on a regular basis as necessary to keep the investment recommendation current.

Certain outstanding reports may contain discussions and/or investment opinions relating to securities, financial instruments and/or issuers that are no longer current. Always refer to the most recent research report relating to an issuer prior to making an investment decision.

In some cases, an issuer may be classified as Restricted or may be Under Review or Extended Review. In each case, investors should consider any investment opinion relating to such issuer (or its securities and/or financial instruments) to be suspended or withdrawn and should not rely on the analyses and investment opinion(s) pertaining to such issuer (or its securities and/or financial instruments) nor should the analyses or opinion(s) be considered a solicitation of any kind. Sales persons and financial advisors affiliated with MLPF&S or any of its affiliates may not solicit purchases of securities or financial instruments that are Restricted or Under Review and may only solicit securities under Extended Review in accordance with firm policies.

Neither BoFA Merrill Lynch nor any officer or employee of BoFA Merrill Lynch accepts any liability whatsoever for any direct, indirect or consequential damages or losses arising from any use of this report or its contents.