NVIDIA Tegra X1 Targets Neural Net Image Processing Performance

Enables Practical Real-Time Neural Network Based Image Recognition

Take-Away

NVIDIA Tegra X1 (TX1) based client-side neural network acceleration is a strong complement to server-side deep learning using NVIDIA Kepler based Tegra server accelerators.

In the software world, “function overloading” refers to using one procedure name to invoke different behavior depending on the context of the procedure call. NVIDIA borrowed that concept for their graphics processing pipelines. They essentially are “overloading” their graphics architecture to be a capable neural network processing architecture.

NVIDIA reuses every portion of their TX1 32-bit floating point pipelines to run two 16-bit pipelines side-by-side, thus doubling numeric throughput at half the precision. Running two 16-bit instructions in the same pipeline as one 32-bit instruction also roughly halves the power consumed per instruction. For applications that do not require 32-bit precision, this arrangement is much more power efficient.

In addition to creating half-width 16-bit floating point graphics pipelines, NVIDIA implemented DSP-like “fused” arithmetic operations in these pipelines. Most notably, the classic FMA instruction (fused multiply-add) enables multiplies and adds to be pipelined into effectively a single-cycle operation when done in bulk. These operations further accelerate neural network recognition functions.

Couple these new advances with the Tegra X1’s I/O offload processors, and you have a system-on-chip (SoC) optimized for recognizing a useful number of features in a real-time 360° high-resolution visual sensory environment.

It would take many years for the industry to pick up an SoC like the TX1 and figure out optimal use cases on its own. NVIDIA looks like it is making substantial software development commitments to neural network based image recognition. These commitments include development environments, application programming interfaces (APIs) and libraries, and hardware reference designs, as well as functional reference applications. This level of commitment is reminiscent of NVIDIA’s initial and very successful investment in CUDA for high performance computing (HPC).

Floating Half Precision

Floating point half precision (16-bits wide, also called FP16) works well for applications that can depend on a wide dynamic range (distance between smallest and largest

number stored) instead of more precise numbers (how many decimal places are stored). Neural network based recognition processing is such an application.

NVIDIA's TX1 pixel shader processes everything in FP32 (“single precision” 32-bit floating point) or FP16. TX1 Graphics RGB is computed in FP16; texture fetch supports FP16 format; and component pixel rendering uses FP16 for format consistency through the pipeline.

TX1’s FP16 format is compliant with IEEE 754-2008’s "binary16" format. ARM’s Midgard GPU (Mali-T6xx, -T7xx and -T8xx), AMD’s Graphics Core Next Gen 3 (GCN 3, in AMD FirePro and Radeon R9 products), and Imagination Technologies’ (ImgTec) PowerVR Series7 also implement IEEE-754 binary16 shader pipelines.

NVIDIA engineered in a little extra design complexity to “overload” their FP32 compute paths, so they can run two FP16 pipelines using the same transistors they use to implement one FP32 pipeline. They can literally run twice the FP16 instructions per clock cycle as FP32 instructions. However, this type of design is not unique.

- **ARM**’s Midgard architecture also “overloads” their FP32 pipeline to run two FP16 instructions in the same clock cycle and on the same transistors. ARM’s Midgard first shipped in Samsung’s Exynos 5250 System-on-Chip (SoC) in November 2012 in the form of a Mali-T604 GPU. What is different is NVIDIA’s scale of implementation in TX1: hundreds of cores versus ARM’s focus on battery powered SoCs and much lower core counts.
- **Imagination Technologies** took a different route and duplicated pipelines in PowerVR Series7 devoting roughly equal area to implement twice the FP16 throughput as FP32. The two pipelines share instruction decode and load/store resources, but otherwise they have completely separate data paths. PowerVR saves power in mobile applications by using FP16 and shutting off the FP32 pipeline. But die area is impacted by having separate FP16 and FP32 resources in the same design. Nevertheless, ImgTec scales their GPU designs to just over a thousand FP16 pipelines.
- **AMD** implemented native FP16 instructions, but they execute on the FP32 pipeline at clock parity. AMD saves power by powering off half of the FP32 transistors while in FP16 mode. FP16 also reduces register file usage and data movement to improve performance; register pressure can be a significant performance constraint for complex shaders. AMD implements close to two thousand pipelines in their designs.

We note that all of the alternatives use FP16 to save power to some extent, while providing close to FP32 graphics quality for various definitions of “mobile” devices.

Fused Operations

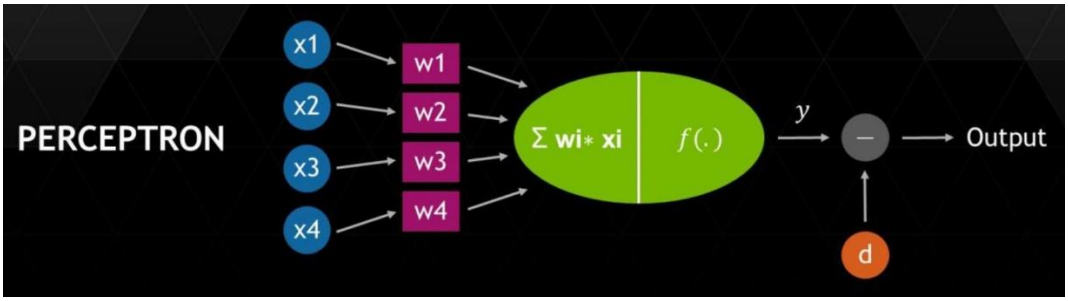
Fused Multiply-Add (FMA) is a single floating-point instruction that implements both a multiply and an add operation in a single instruction pipeline. FMA is typically a three-operand instruction and does not perform rounding operations until the add operation is

complete. Thus FMA saves transistors, improves performance, and reduces power consumption. For the same reasons, DSPs have historically implemented fixed-point Fused Multiply-Accumulate (FMAC), a close cousin of FMA.

FP32 FMA has been part of NVIDIA's graphics architecture since the Fermi generation in 2010. All graphics shaders heavily use FP32 FMA. NVIDIA added FP16 FMA to their new Maxwell architecture to support both graphics and compute. AMD and ImgTec also implement FP16 FMA: AMD because they use a constrained version of their existing FP32 pipeline, and ImgTec because they duplicate exactly their FP32 instruction set in FP16.

It is easy to see that graphics pipelines with this basic DSP-like function should be very good at executing non-branching neural network (perceptron, the digital equivalent to a neuron) input weighting and neural computation functions, as depicted in Figure 1.

Figure 1: A Neural Network "Perceptron"



How is NVIDIA's Approach Different?

NVIDIA is the first of this competitive field to assemble a comprehensive, systems-level approach (hardware and software solution) for real-time neural network based image recognition. AMD, ARM and Imagination Technologies have important pieces of the hardware solution in-hand. But we have not yet seen them invest in stitching those pieces together. Nor have we seen them invest in creating a targeted software development framework for neural network based imaging applications.

NVIDIA's biggest differentiator on the silicon design side is their investment in scaling the number of FP16 pipelines on their TX1 while significantly reducing both power consumption and die area required for each FP16 pipeline.

Usage Model: Internet of Things (IoT) Comes to Car and Driver

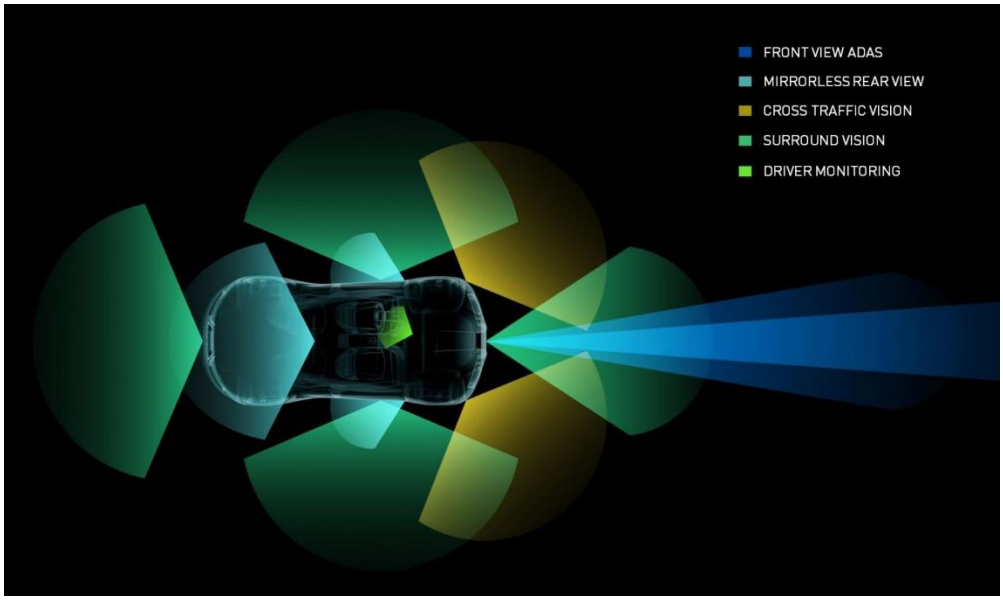
Vehicles are big things, but they are things nonetheless, and they are rapidly being pulled into the Internet of Things (IoT). From driver assistance to autonomous vehicles, creating a 3D surround view of a vehicle and identifying features of the vehicle's surroundings is the core computational function of a smart vehicle.

NVIDIA's DRIVE PX autopilot platform aims to be the video recognition hub for both driver-assisted and driverless cars. NVIDIA will deliver a DRIVE PX developer platform in Q2 2015, and production systems are targeted for 2016.

The DRIVE PX platform uses two TX1 chips running a hardened real-time operating system (RTOS). Built-in functions will range from rear collision monitoring to driver monitoring and self-parking. The ability to add third-party functions also will be supported. These functions will be enabled by a mirrorless 360° surround view of a car's surroundings incorporating up to a dozen video cameras.

In Figure 2, each cone represents the field of view of one camera. The camera is positioned at the sharp point of the fan. The fan shows each camera's field of view. There are 11 outward facing cameras in this depiction (front of car faces to the right) and one camera to monitor the driver.

Figure 2: A Car Full of Video Cameras and Their Overlapping Fields of View

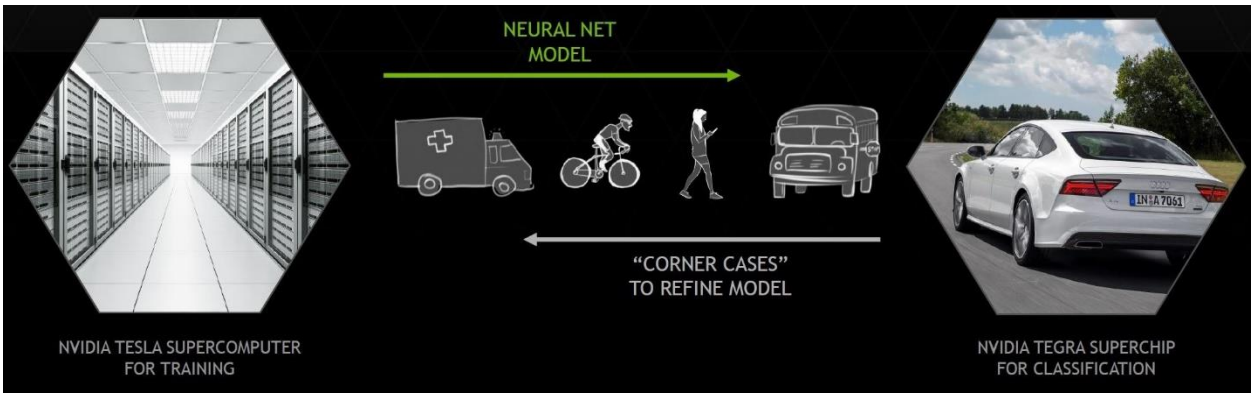


Views from all externally facing cameras are integrated into a surround panorama which is analyzed for hazards and clear paths forward. Figure 3 is a complex test image in which many humans have trouble identifying all of the potential threats.

Figure 3: Not an Average Street Scene with Recognized Objects Labeled

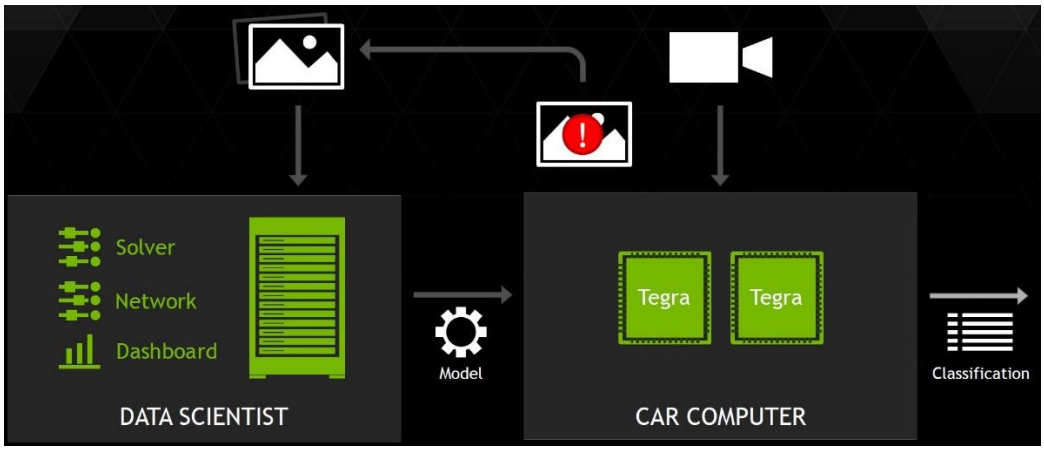
The in-car TX1 based system recognizes what it can with its local neural network configuration. Then, using a cellular network connection, it sends features it cannot recognize back to a datacenter for identification.

Figure 4: High Level View of Car and Datacenter Working Together



Unrecognized features are then classified in the datacenter. After classification, new neural network "rules" are sent back to all cars on the network, so these features can be identified successfully by all vehicles in the future.

Figure 5: Functional View of NVIDIA DRIVE PX Communicating with Datacenter



NVIDIA expects that their eventual full-production, dual Tegra X1 based DRIVE PX system will be capable of identifying about 150 different objects simultaneously at 30fps with a camera configuration similar to that in Figure 2. The DRIVE PX system might be used to alert drivers to potential threats and for emergency maneuvers, or it might be used to gather feature recognition data to improve the performance of autonomous vehicles.

Maybe You Can Drive My Car

Will these systems ever be perfect drivers? And how will we measure their driving acumen? Neural networks of this sophistication cannot be validated completely by deterministic testing. Statistically speaking, they simply must be measurably better than most humans over a wide variety of driving conditions. Consider that many of us humans fail comparatively simple driving tests. Plus we have bad days, injure ourselves, grow old, *etc*—problems that we expect auto-pilots like DRIVE PX not to have.

We give NVIDIA high marks for “completeness of vision” in automotive imaging applications.

Important Information About This Brief

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Licensing

Disclosures
Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.