Graphics Processing Unit (GPU) Devices

Edward J. Wyrwas
edward.j.wyrwas@nasa.gov
301-286-5213
Lentech, Inc. in support of NEPP

Acknowledgment:
This work was sponsored by:
NASA Electronic Parts and Packaging (NEPP)

Acronyms

Acronym	Definition
1MB	1 Megabit
3D	Three Dimensional
3DIC	Three Dimensional Integrated Circuits
ACE	Absolute Contacting Encoder
ADC	Analog to Digital Converter
AEC	Automotive Electronics Council
AES	Advanced Encryption Standard
AF	Air Force
AFRL	Air Force Research Laboratory
AFSMC	Air Force Space and Missile Systems Center
AMS	Agile Mixed Signal
ARM	ARM Holdings Public Limited Company
BGA	Ball Grid Array
BOK	Body of Knowledge
CAN	Controller Area Network
CBRAM	Conductive Bridging Random Access Memory
CCI	Correct Coding Initiative
CGA	Column Grid Array
CMOS	Complementary Metal Oxide Semiconductor
CN	Xilinx ceramic flip-chip (CF and CN) packages are ceramic column grid array (CCGA) packages
COTS	Commercial Off The Shelf
CRC	Cyclic Redundancy Check
CRÈME	Cosmic Ray Effects on Micro Electronics
CRÈME MC	Cosmic Ray Effects on Micro Electronics Monte Carlo
CSE	Crypto Security Engin
CU	Control Unit
D-Cache	defered cache
DCU	Distributed Control Unit
DDR	Double Data Rate (DDR3 = Generation 3; DDR4 = Generation 4)
DLA	Defense Logistics Agency
DMA	Direct Memory Access
DMEA	Defense MicroElectronics Activity
DoD	Department of Defense
DOE	Department of Energy
DSP	Digital Signal Processing
dSPI	Dynamic Signal Processing Instrument
Dual Ch.	Dual Channel
ECC	Error-Correcting Code
EEE	Electrical, Electronic, and Electromechanical
EMAC	Equipment Monitor And Control
EMIB	Multi-die Interconnect Bridge
ESA	European Space Agency
eTimers	Event Timers
ETW	Electronics Technology Workshop
FCCU	Fluidized Catalytic Cracking Unit
FeRAM	Ferroelectric Random Access Memory
FinFET	Fin Field Effect Transistor (the conducting channel is wrapped by a thin silicon "fin")
FPGA	Field Programmable Gate Array
FPU	Floating Point Unit
FY	Fiscal Year
GaN	Gallium Nitride
GAN GIT	Panasonic GaN GIT Eng Prototype Sample
GAN SIT	Gallium Nitride GIT Eng Prototype Sample
Gb	Gigabyte
GCR	Galactic Cosmic Ray
GIC	Global Industry Classification

Acronym	Definition
Gov't	Government
GPU	Graphics Processing Unit
GRC	NASA Glenn Research Center
GSFC	Goddard Space Flight Center
GSN	Goal Structured Notation
GTH/GTY	Transceiver Type
HALT	Highly Accelerated Life Test
HAST	Highly Accelerated Stress Test
HBM	High Bandwidth Memory
HDIO	High Density Digital Input/Output
HDR	High-Dynamic-Range
HiREV	High Reliability Virtual Electronics Center
HMC	Hybrid Memory Cube
HP Labs	Hewlett-Packard Laboratories
HPIO	High Performance Input/Output
HPS	High Pressure Sodium
HUPTI	Hampton University Proton Therapy Institute
I/F	interface
I/O	input/output
I2C	Inter-Integrated Circuit
i2MOS	Microsemi second generation of Rad-Hard MOSFET
IC	Integrated Circuit
IC	Integrated Circuit
I-Cache	independent cache
IUCF	Indiana University Cyclotron Facility
JFAC	Joint Federated Assurance Center
JPEG	Joint Photographic Experts Group
JTAG	Joint Test Action Group (FPGAs use JTAG to provide access to their programming debug/emulation functions)
KB	Kilobyte
L2 Cache	independent caches organized as a hierarchy (L1, L2, etc.)
LANL	Los Alamos National Laboratories
LANSCE	Los Alamos Neutron Science Center
LLUMC	Loma Linda University Medical Center
L-mem	Long-Memory
LP	Low Power
LVDS	Low-Voltage Differential Signaling
LW HPS	Lightwatt High Pressure Sodium
M/L BIST	Memory/Logic Built-In Self-Test
MBMA	Model-Based Missions Assurance
MGH	Massachusetts General Hospital
Mil/Aero	Military/Aerospace
MIPI	Mobile Industry Processor Interface
MMC	MultiMediaCard
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MP	Microprocessor
MP	Multiport
MPFE	Multiport Front-End
MPU	Microprocessor Unit
Msg	message
NAND	Negated AND or NOT AND
NASA	National Aeronautics and Space Administration
NASA STMD	NASA's Space Technology Mission Directorate
Navy Crane	Naval Surface Warfare Center, Crane, Indiana
NEPP	NASA Electronic Parts and Packaging
NGSP	Next Generation Space Processor
NOR	Not OR logic gate

Acronym	Definition
NRL	Naval Research Laboratory
NRO	United States Navy National Reconnaissance Office
NSWC Crane	Naval Surface Warfare Center, Crane Division
OCM	On-chip RAM
PBGA	Plastic Ball Grid Array
PC	Personal Computer
PCB	Printed Circuit Board
PCIe	Peripheral Component Interconnect Express
PCIe Gen2	Peripheral Component Interconnect Express Generation 2
PLL	Phase Locked Loop
POL	point of load
PoP	Package on Package
PPAP	Production Part Approval Process
Proc.	Processing
PS-GTR	High Speed Bus Interface
QDR	quad data rate
QFN	Quad Flat Pack No Lead
QSPI	Serial Quad Input/Output
R&D	Research and Development
R&M	Reliability and Maintainability
RAM	Random Access Memory
ReRAM	Resistive Random Access Memory
RGB	Red, Green, and Blue
RH	Radiation Hardened
SATA	Serial Advanced Technology Attachment
SCU	Secondary Control Unit
SD	Secure Digital
SD/eMMC	Secure Digital embedded MultiMediaCard
SD-HC	Secure Digital High Capacity
SDM	Spatial-Division-Multiplexing
SEE	Single Event Effect
SESI	secondary electrospray ionization
Si	Silicon
SiC	Silicon Carbide
SK Hynix	SK Hynix Semiconductor Company
SLU	Saint Louis University
SMDs	Selected Item Descriptions
SMMU	System Memory Management Unit
SNL	Sandia National Laboratories
SOA	Safe Operating Area
SOC	Systems on a Chip
SPI	Serial Peripheral Interface
STT	Spin Transfer Torque
TBD	To Be Determined
Temp	Temperature
THD+N	Total Harmonic Distortion Plus Noise
TRIUMF	Tri-University Meson Facility
T-Sensor	Temperature-Sensor
TSMC	Taiwan Semiconductor Manufacturing Company
U MD	University of Maryland
UART	Universal Asynchronous Receiver/Transmitter
UFHPTI	University of Florida Proton Health Therapy Institute
UltraRAM	Ultra Random Access Memory
USB	Universal Serial Bus
VNAND	Vertical NAND
WDT	Watchdog Timer

Outline

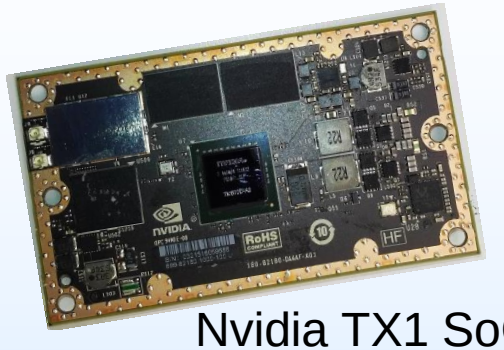
- What the technology is (and isn't)
- Our tasks and their purpose
- Roadmap
- Partners
- Test Readiness
- Comments

Technology

- **Graphics Processing Units (GPU) & General Purpose Graphics Processing Units (GPGPU)**
  - **Are considered a compute device or coprocessor**
  - **Is not a standalone multiprocessor (even when contained in an SoC)**

- **Application workflow:**
  - **Run the sequential part of their workload on the CPU – which is optimized for single-threaded performance**
  - **Accelerate parallel processing using multi-thread performance on the GPU**

Device Packaging



Nvidia TX1 SoC

Qualcomm Adreno

Intel Skylake Processor

Nvidia GTX 1050 GPU

AMD RX460 GPU

Purpose

- **GPUs are best used for single instruction- multiple data (SIMD) parallelism**
  - **Perfect for breaking apart a large data set into smaller pieces and processing those pieces in parallel**

- **Key computation pieces of mission applications can be computed using this technique**
  - **Sensor and science instrument input**
  - **Object tracking and obstacle identification**
  - **Algorithm convergence (neural network)**
  - **Image processing**
  - **Data compression algorithms**

FY18-19: GPU Testing

Description:

- This is a task over all device topologies and process
- The intent is to determine inherent radiation tolerance and sensitivities
- Identify challenges for future radiation hardening efforts
- Investigate new failure modes and effects
- Testing includes total dose, single event (proton) and reliability. Test vehicles will include a GPU devices from nVidia and other vendors as available
  - Compare to previous generations
  - Investigate failure modes/compensation for increased power consumption

FY18-19 Plans:

- Continue development of universal test suite which includes math, output buffer (colors), memory hierarchy and neural networks
- Probable test structures for SEE:
  - Nvidia (16, 14, 10nm)
  - AMD (14, 10nm)
  - Intel (14)
  - Qualcomm (10nm)
- Tests:
  - characterization pre, during and post-rad

Schedule:

Microelectronics T&E	FY18						FY19					
	M	J	J	A	S	O	N	D	J	F	M	A
On-going discussions for test samples												
GPU Test Development		◇						◇				
SEE Testing												
Analysis and Comparison												◇

Lead Center/PI: GSFC/Lentech/Wyrwas
Co-Is: Carl Szabo

Deliverables:

- Test reports and quarterly reports
- Expected submissions for publications

NASA and Non-NASA Organizations/Procurements:

- Source procurements: Proton (MGH), TID (GSFC), Laser (NRL)

GPU Roadmap

Body of Knowledge Document — *BOK* → *eDAA…*

GPUs
- 14nm Nvidia GTX 1050
- 14nm AMD Radeon
- 7nm AMD Vega
- 12nm Nvidia Titan

Radiation Testing (14nm Nvidia GTX 1050)

Collaboration with microprocessors (14nm AMD Radeon, 7nm AMD Vega)

TBD (12nm Nvidia Titan)

System on Chip
- 20nm Nvidia Tegra X1
- 16nm Nvidia Tegra X2
- 14nm Intel+AMD GPU
- 12nm+ Nvidia Xavier
- Qualcomm Snapdragon

Radiation Testing (20nm Nvidia Tegra X1)

TBD (16nm Nvidia Tegra X2, 14nm Intel+AMD GPU, 12nm+ Nvidia Xavier, Qualcomm Snapdragon)

Neural Chips
- KnuEdge Hermosa
- KnuEdge Hydra

TBD (KnuEdge Hermosa, KnuEdge Hydra)

FY17 **FY18** **FY19**

Partners

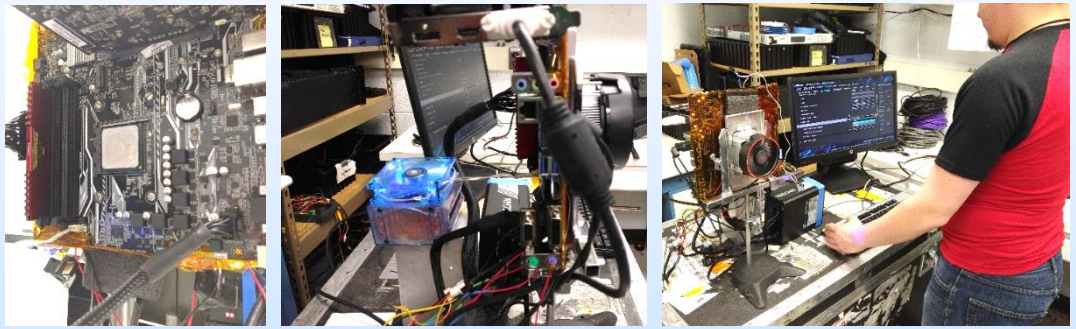
Ongoing and new collaborations:

- JPL (Steve Guertin, Andrew Daniel)
- Navy Crane (Dobrin Bossev, Jonathan Wang)
- NEPP Microprocessors (Carl Szabo)
- Dr. Paolo Rech (UFRGS)
- Cubic Aerospace
- TuSimple
- JSC Human Interface Branch
- GSFC Microwave Branch
- GSFC Photonics Group
- Harris Corporation
- Ball Aerospace
- General Atomics
- LetSAT.org (LeTourneau University)
- Advanced Micro Devices (AMD)

Test Readiness & Results

- A universal test bench is under development to provide a standardized approach to test GPUs with minimal variation between device types. The test bench must perform comparably under Proton, Heavy-Ion, Laser and Total Ionizing Dose tests.

400W Cooling on Bare NVIDIA GTX 1050



180W Cooling on Lidded AMD Ryzen CPU

- A cooling solution created for GPU testing has been refined to also cool socketed CPUs such as an AMD Ryzen microprocessor which contains a GPU. This technique can be applied to System on Module (SOM) devices too.

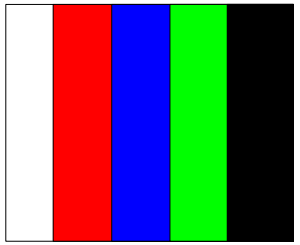
Test Readiness & Results

- Three types of payloads have been created for the GPU test bench: Neural Network, Math-Logic and Colors.
  - The neural network is a convolutional neural network (CNN) which can avoid processor optimizations that recursive neural networks (RNN) primarily benefit from.
  - Math-Logic uses mathematics and conditional logic statements to exercise memory hierarchy.
  - The Colors payload assesses corruption in the output image presented to a display.

Algebra and Matrices

$$A = \pi r^2 \quad OR \quad \begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}$$

Pixel Color Output



Neural Networks

Comments

- The NEPP GPU standardized approach involves:
  - rapid development of cooling system for each DUT form factor and packaging type
  - system implementation using modular COTS' system and network components
  - public domain software that has been excessively tested by the community
  - payloads that can be easily updated to accommodate new DUTs while maintaining the ability to test older DUTs

- References
  - Nvidia Jetson TX1 (SoC) http://hdl.handle.net/2060/20170009004
  - Nvidia GTX 1050 (Discrete GPU) http://hdl.handle.net/2060/20170009005
  - Considerations for testing (overview) http://hdl.handle.net/2060/20170004734
  - NEPP GPU Body of Knowledge document TBD – NEPP Website
  - Standardizing GPU Radiation Test Approaches TBD – SEE Symposium, May 2018

(other documents will be published after review)