Graphics Processor Units (GPUs)

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) Program

Acronyms

Acronym	Definition
BOK	Body of Knowledge (document)
CUDA	Compute Unified Device Architecture
DUT	Device Under Test
GPGPU	General Purpose Graphics Processing Unit
GPU	Graphics Processing Unit
MBU	Multi-Bit Upset
MGH	Massachusetts General Hospital
NEPP	NASA Electronic Parts and Packaging
PTX	Parallel Thread Execution
RTOS	Real Time Operating System
SBU	Single-Bit Upset
SEE	Single Event Effect
SEFI	Single Event Functional Interrupt
SEU	Single Event Upset
SIMD	Single Instruction Multiple Data
SoC	System on Chip
TID	Total Ionizing Dose

Outline

- What the technology is (and isn't)
- Our tasks and their purpose
  - The setup around the test setup
  - Parametric considerations
  - Lessons learned
- Collaborations
  - Roadmap
  - Partners
  - Results to date
  - Plans
- Comments

Technology

- **Graphics Processing Units (GPU) & General Purpose Graphics Processing Units (GPGPU) are considered compute devices that behave like coprocessors**
    - **Take assignments from another device**
    - **Inability to load and execute code on boot by itself**

- **Using high-level languages, GPU-accelerated applications run the sequential part of their workload on the CPU – which is optimized for single-threaded performance – while accelerating parallel processing on the GPU.**

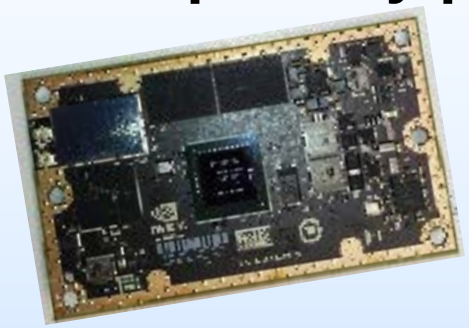
Purpose

- **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**

Device Selection

- **Unfortunately, GPUs come in multiple types, acting as primary processor (SoC) and coprocessor (GPU)**



Nvidia TX1 SoC

Smart Phones

Intel Skylake Processor

Nvidia GTX 1050 GPU

AMD RX460 GPU

Device Software

- **Does it need its own operating system?**
  - **E.g. Linux, Android, RTOS**
- **Can we just push code at it?**
  - **E.g. Assembly, PTX, C**

- **Payload normalization**
  - **Can we run the same code on the previous generation and next generation of the device?**
  - **Cannot with CUDA code; can with OpenCL**

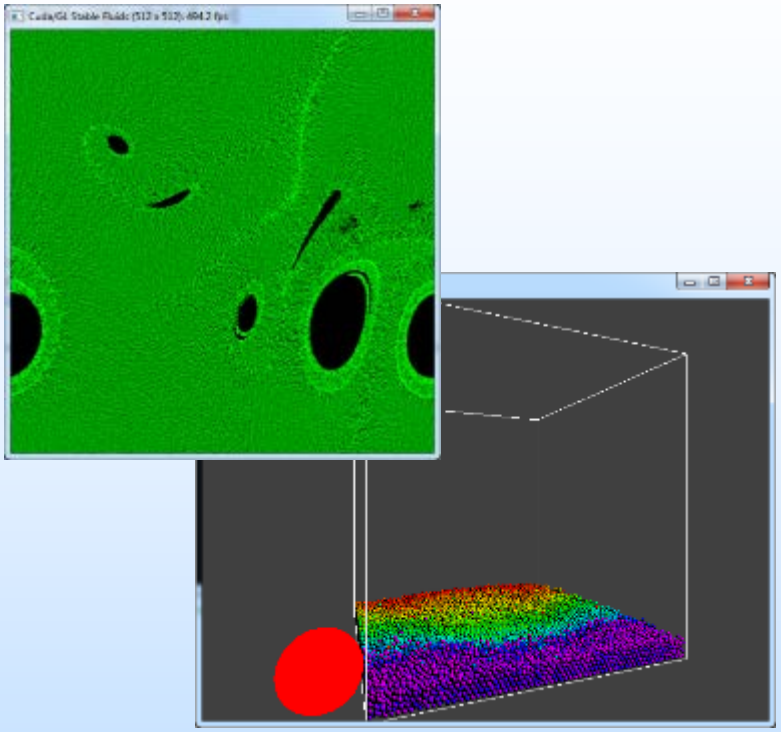
Real-time Operating System (RTOS)

Parallel Thread Execution (PTX)

CUDA is a parallel computing platform and application programming interface (API) model created by Nvidia

Payloads

- **Visual Simulations**
  - **Sample code**
  - **Fuzzy Donut (i.e. Furmark)**
- **Sensor streams**
  - **Camera feed**
  - **Offline video feed**
- **Computational loading**
  - **Scientific computing models**
- **Easy Math**
  - **0 + 0 … wait … should = 0**



Test Setup

- **Things to consider in the test environment**
  - **Operating system daemons**
  - **Location of payload and results**
  - **Data paths upstream/downstream**
  - **Control of electrical sources**
  - **Temperature control (i.e. heaters) in a vacuum**

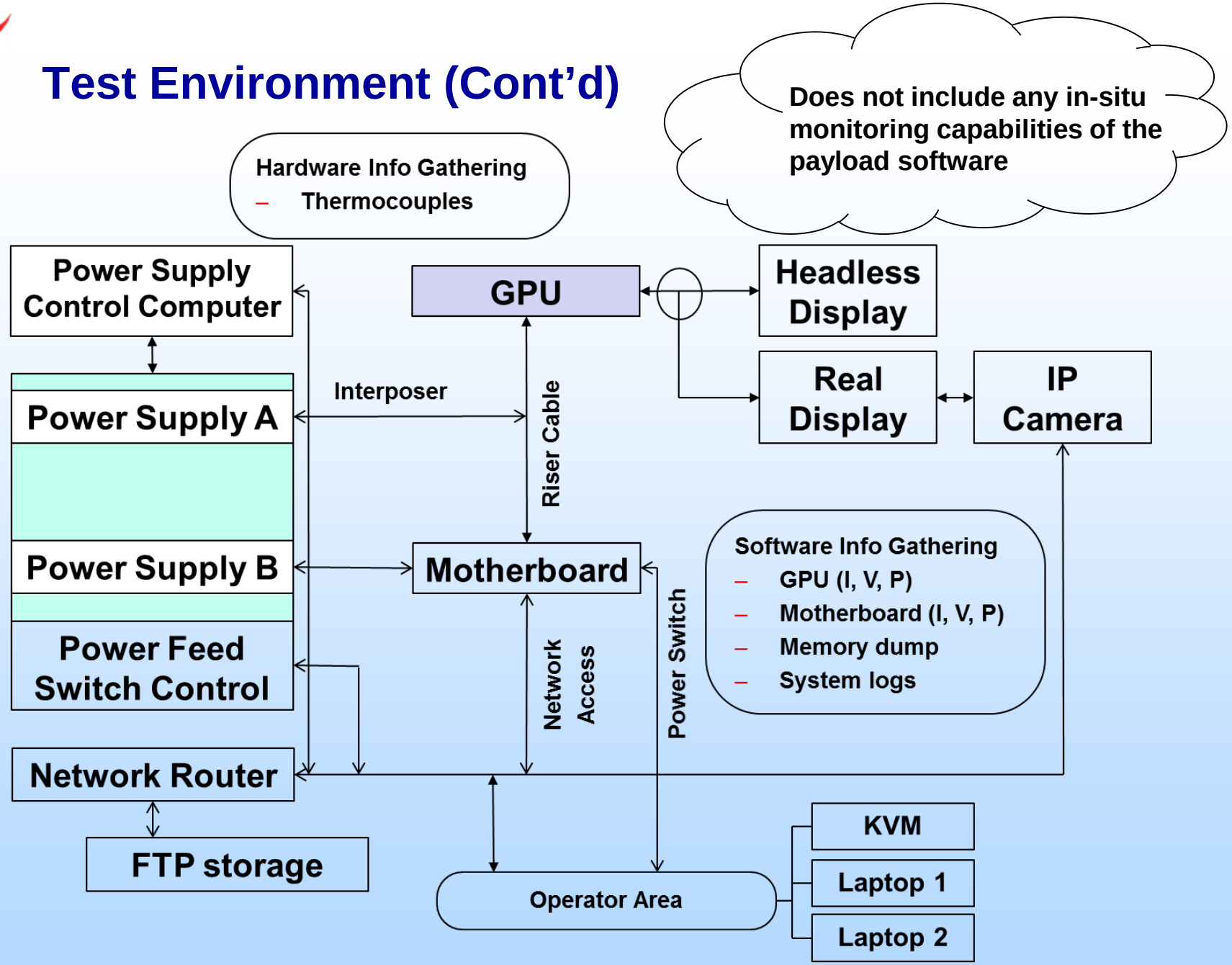
- **Things to consider in the DUT**
  - **Is the die accessible?**
  - **What functional blocks are accessible?**
  - **Which functions are independent of each other?**
  - **Does it have proprietary or open software?**

Test Environment

- **Beam line**
  - **DUT testing zone where collateral damage can happen**
  - **Shielding for everything non-DUT**

- **Operator Area**
  - **Cables, interconnects and extenders**
  - **Signal integrity at a distance**
  - **“Everything that was done in a lab, in front of you on a bench, now must be done from a distance…”**

Test Environment (Cont'd)



Test Environment (Cont'd)

Tripod and mounting

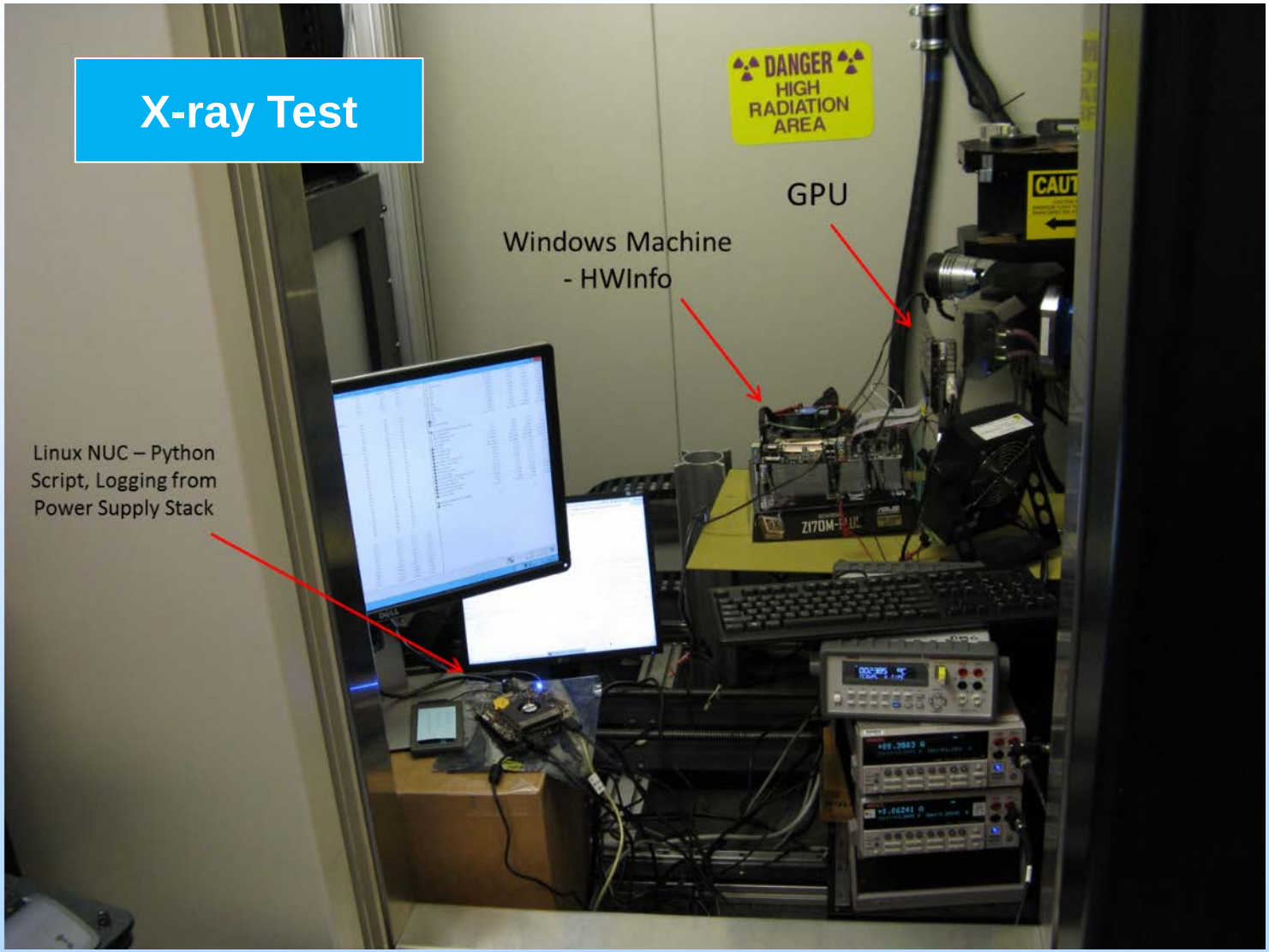
External power

Power injection

Arrows and circle mark locations of the lead and acrylic block fortresses

Pictures are from Massachusetts General Hospital Francis Burr Proton Facility

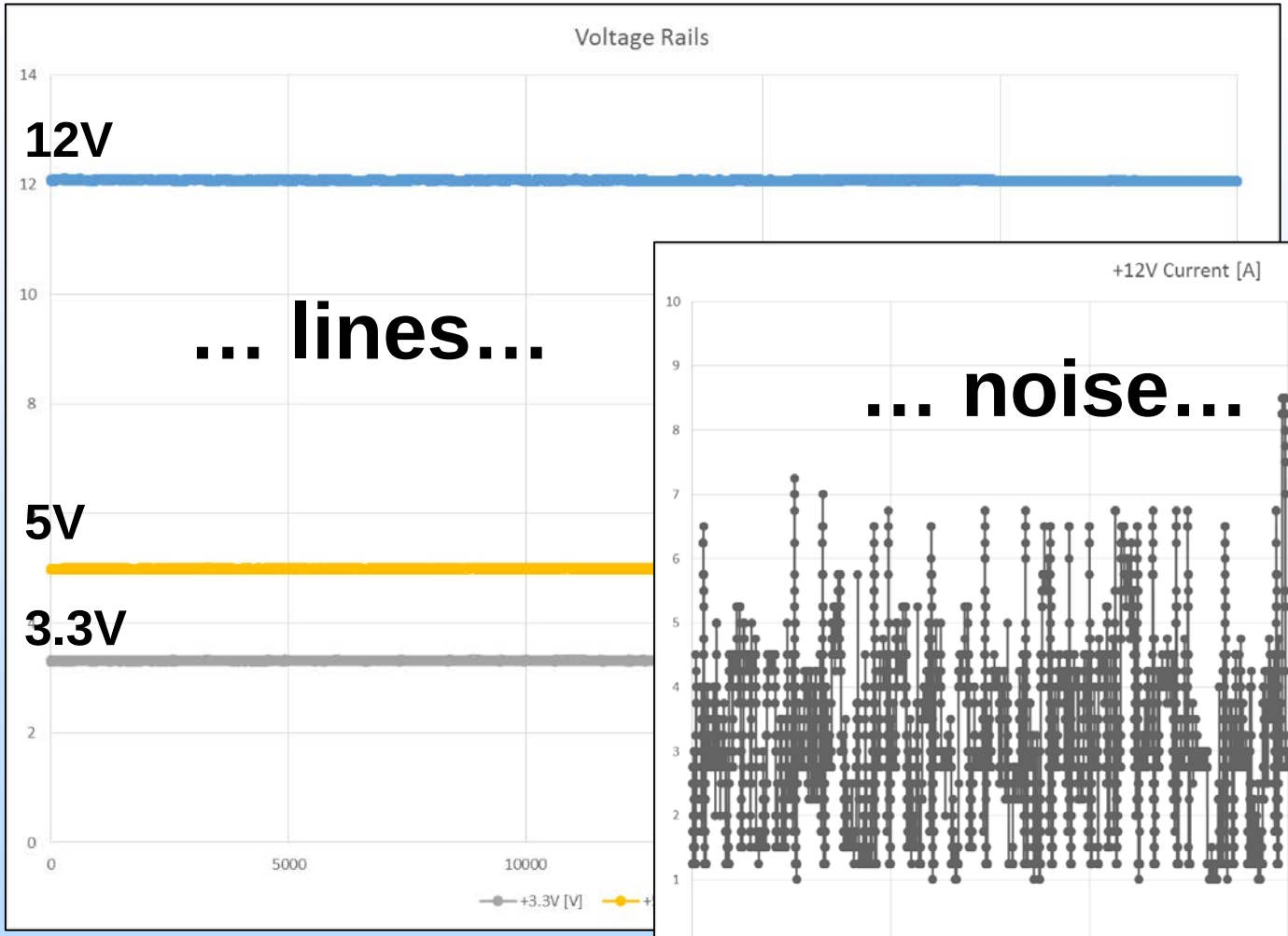
Test Environment (Cont'd)



DUT Health Status

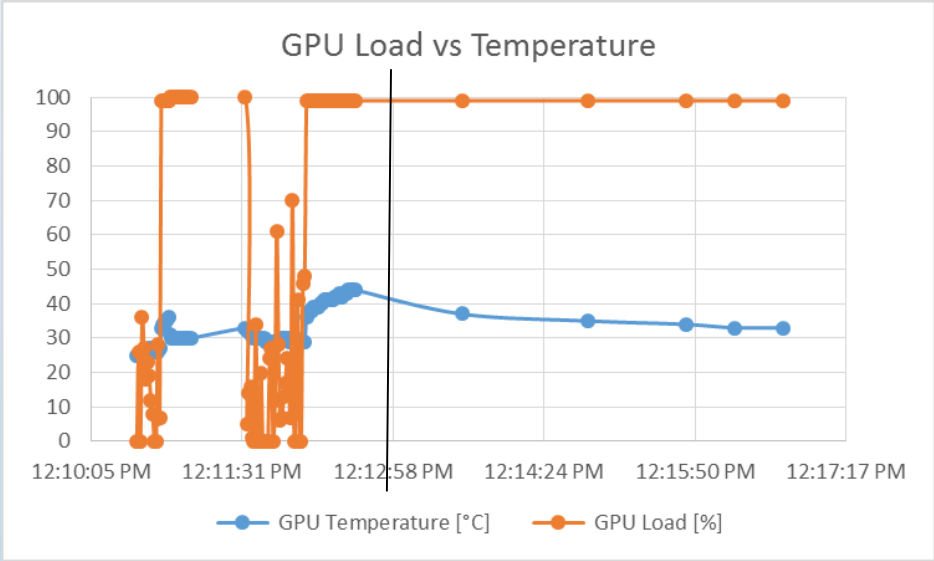
- **Accessible nodes**
  - **Network**
    - **Heart beat by inbound ping**
    - **Heart beat by timestamp upload**
  - **Peripherals response**
    - **“Num lock”**
  - **Visual check**
    - **Remote**
    - **Local**
    - **Local with remote viewing**
  - **Electrical states**
    - **At the system**
    - **At the DUT**

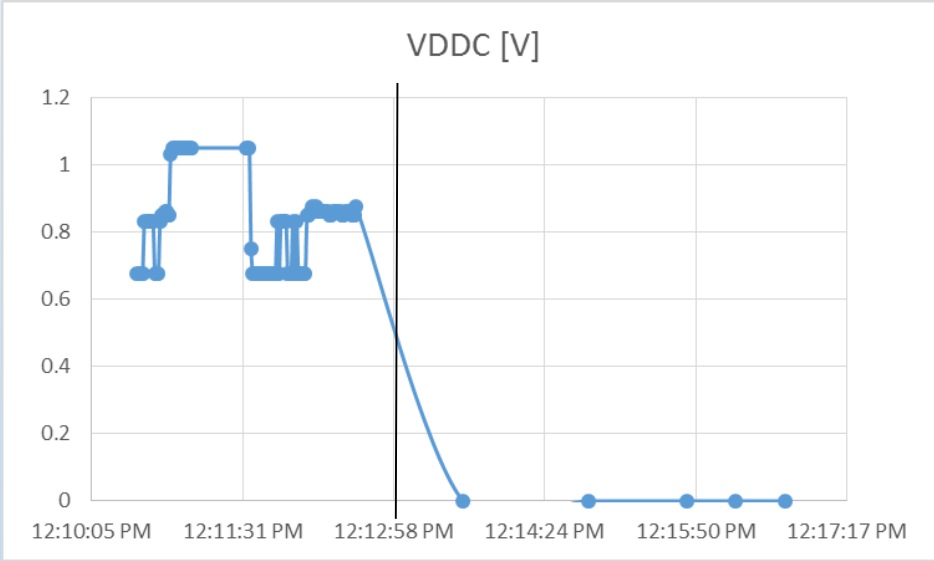
Monitoring Data



Monitoring Data (Cont'd)

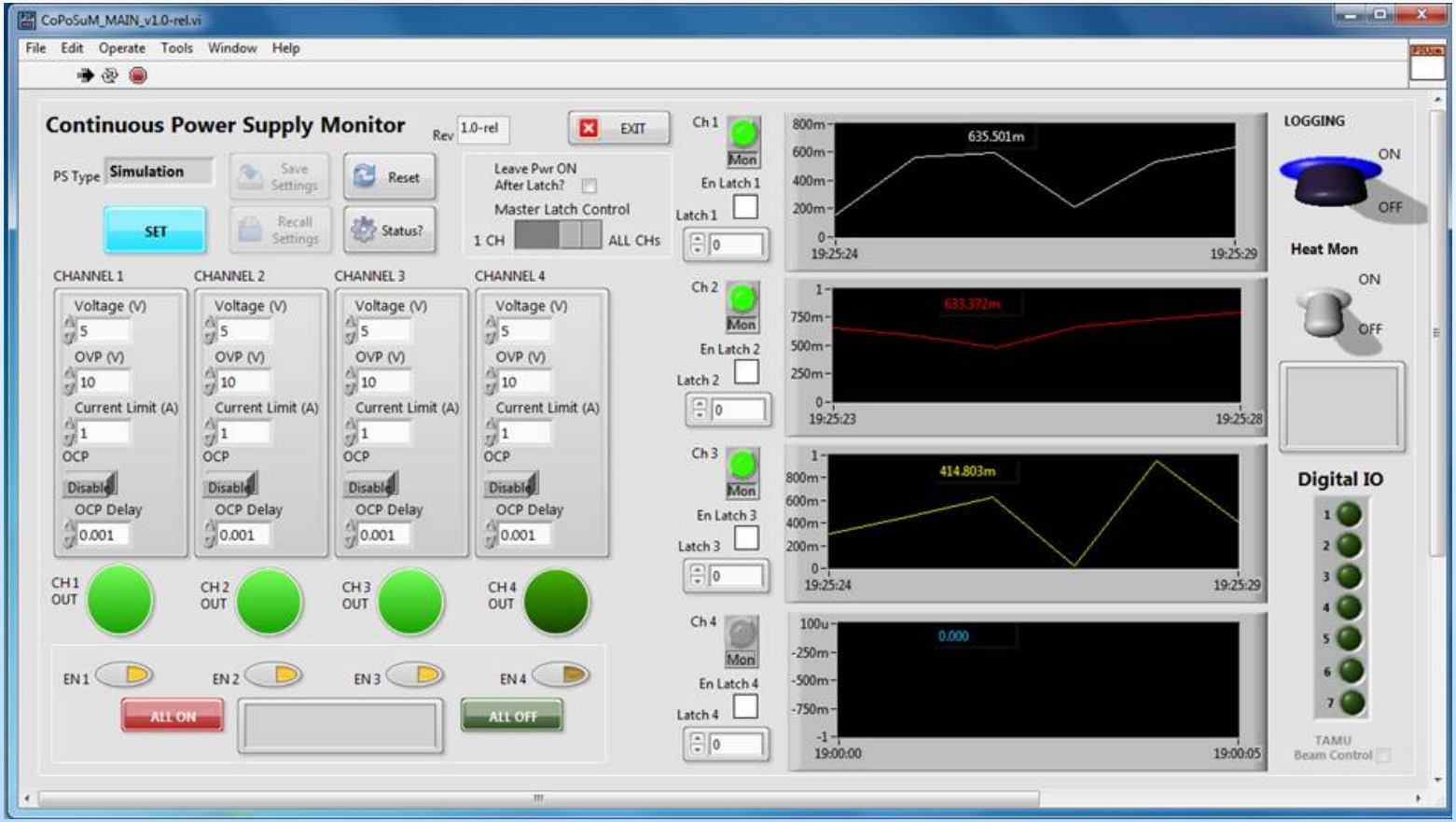
- **Significant digits are important**
- **Resolution is needed for correlation**
  - **Faster sampling speed**
  - **Smaller units (µV or mV, not Volts)**



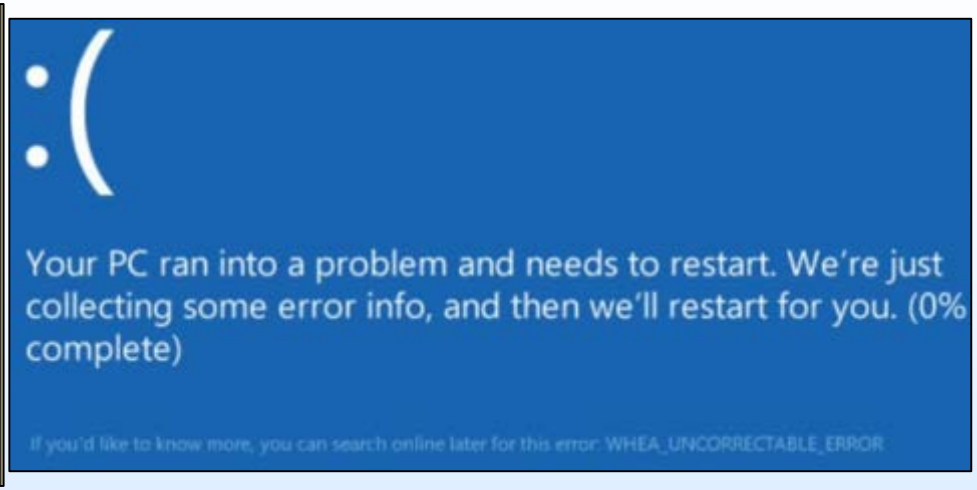


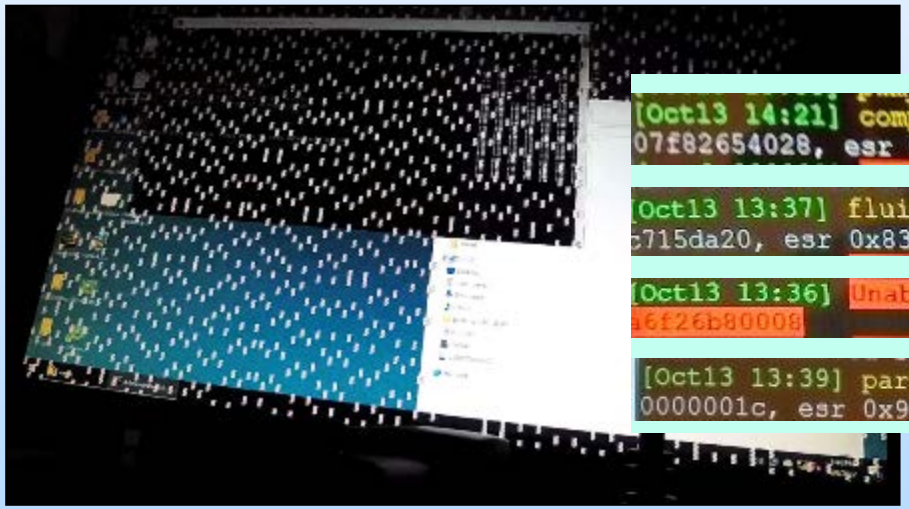
Monitoring Data (Cont'd)

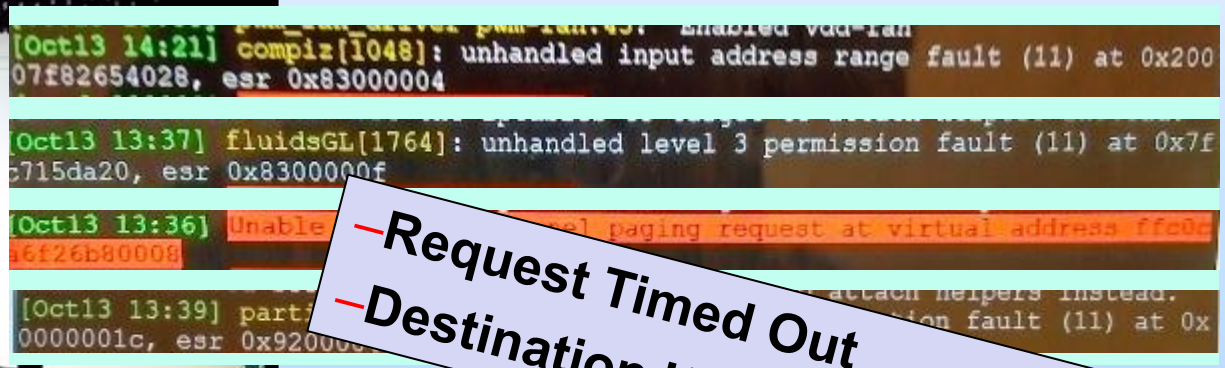
- **Even better (albeit being a mock up):**



What does a failure look like?

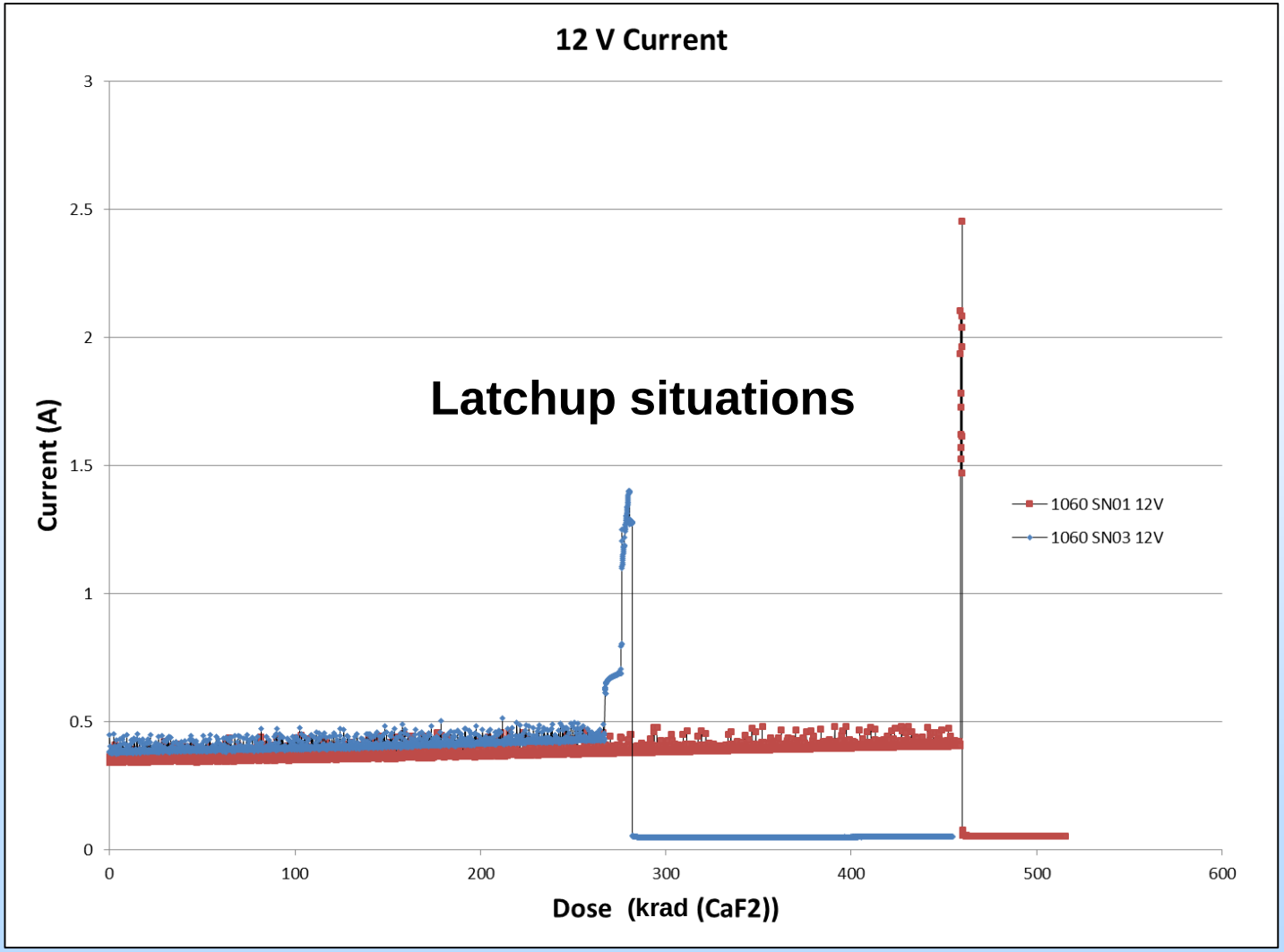






–Request Timed Out
–Destination Host Unreachable

Failures (Cont'd)



Learning Experience

- Every test is another learning experience
  - "Is the laser alignment jig in the beam path…"
  - Nuances with controllable nodes
    - DUT power switch
    - Remote power sources
    - DUT electrical isolation from test platform
    - Thermal paths
  - Improvements are always possible, but preparation time may not be as abundant
  - Prioritization during development is important
    - Software payload
    - Hardware monitoring
    - Remote troubleshooting capabilities

GPU Roadmap

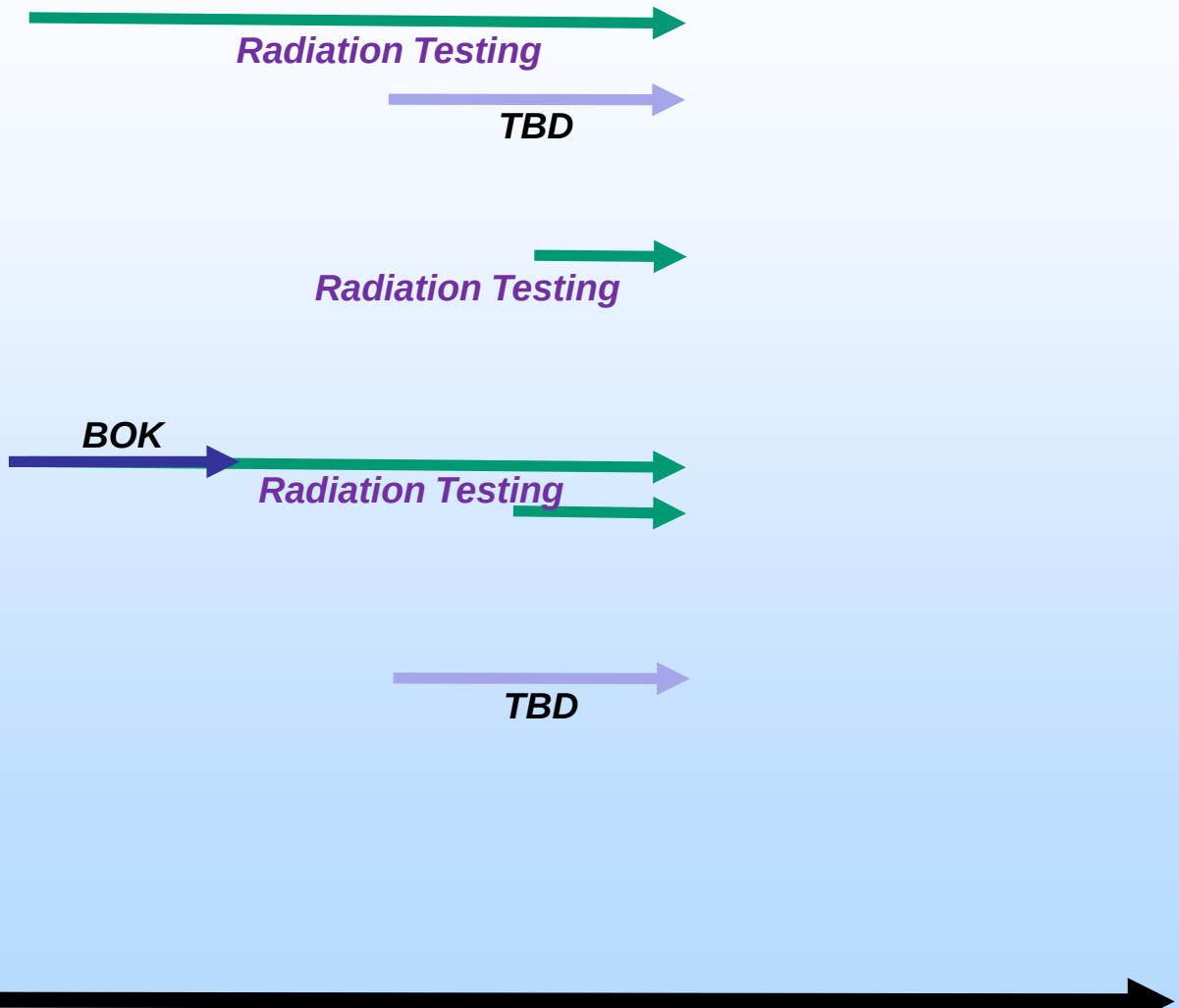
- collaborative with NSWC Crane, others

GPUs
- 14nm Nvidia GTX 1050
- 14nm AMD Radeon

GPGPUs
- 14nm Nvidia Tesla P100

Mobile System on Chip
- 20nm Nvidia Tegra X1
- 16nm Nvidia Tegra X2
- 14nm Intel HD Graphics

Neural Chips
- KnuEdge Hermosa
- KnuEdge Hydra



Partners

- **Navy Crane**
  - **Conducting testing on Nvidia 14nm GPUs**
- **Collaboration with partners is yielding a comprehensive test suite**
  - **L1 and L2 cache**
  - **Registers**
  - **Shared, Internal, Texture and Global memory**
  - **Control logic**

Qualification Guidance

- Creation of GPU Body of Knowledge (BoK) document
  - Technology
    - Silicon
    - Packaging
    - Heterogeneous constituents
  - Reliability
    - Semiconductor mechanisms
    - Package issues
    - Scaling issues
  - Failure categories and trends
  - Software & Hardware sources
- Future guidelines will be developed for this technology to include qualification and test methods

Results to Date

- Developing software for cross platform use
  - Nvidia Tegra X – SoC ARM with embedded Linux
  - Nvidia GPUs – GPU for x86 Windows and Linux
  - Intel Skylake Processor – IP Block for x86 Linux
  - Qualcomm Adreno & Mali GPU – IP Block for ARM Linux

- Proton test result ranges are dependent on physical target within DUT
  - Cross section (σ, cm^2): $1x10^{-7}$ to $9x10^{-9}$
  - Flux ($p/cm^2/sec$): $1x10^6$ to $7x10^6$

Plans (w Schedule)

- More proton testing on 14nm GPUs
  - Test OpenCL payloads
  - Test L1, L2, registers, shared memory & control logic
  - Record die temperature, 12V and 3.3V rail voltages and currents, system events (and observations)

- Two proton test sessions and significant in-lab work has permitted improvements to:
  - Thermal-electrical monitoring of the DUTs – though some more improvements are necessary to achieve the desired resolution
  - Proving out which code libraries won't work for the type of testing we're conducting

FY17-18: 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

FY17-18 Plans:

- Continue development of universal test suite
- Probable test structures for SEE:
  - Nvidia (16, 14, 10nm)
  - AMD (14nm)
  - Intel (14nm)
- Tests:
  - characterization pre, during and post-rad

Schedule:

Microelectronics T&E	FY17							FY18				
	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												◇

Deliverables:

- Test reports and quarterly reports
- Expected submissions for publications

NASA and Non-NASA Organizations/Procurements:

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

PIs: GSFC/Lentech/Wyrwas

Conclusion

- NEPP and its partners have conducted proton, neutron and heavy ion testing on several devices
  - Have captured SEUs (SBU & MBU),
  - Have seen traceable current spikes,
  - But predominately have encountered system-based SEFIs
- GPU testing requires a complex platform to arbitrate the test vectors, monitor the DUT (in multiple ways) and record data
  - None of these should require the DUT itself to reliably perform a task outside of being exercised
- Progress has been made in proving out multiple ways to simulate and enumerate activity on the DUT
  - Narrowing down on a universal test bench
  - End goal is to make test code platform independent

Acknowledgement

- Ken LaBel, NASA GSFC NEPP
- Martha O'Bryan, ASRC Space & Defense
- Carl Szabo, ASRC Space & Defense
- Steve Guertin, NASA JPL
- Adam Duncan, Navy Crane