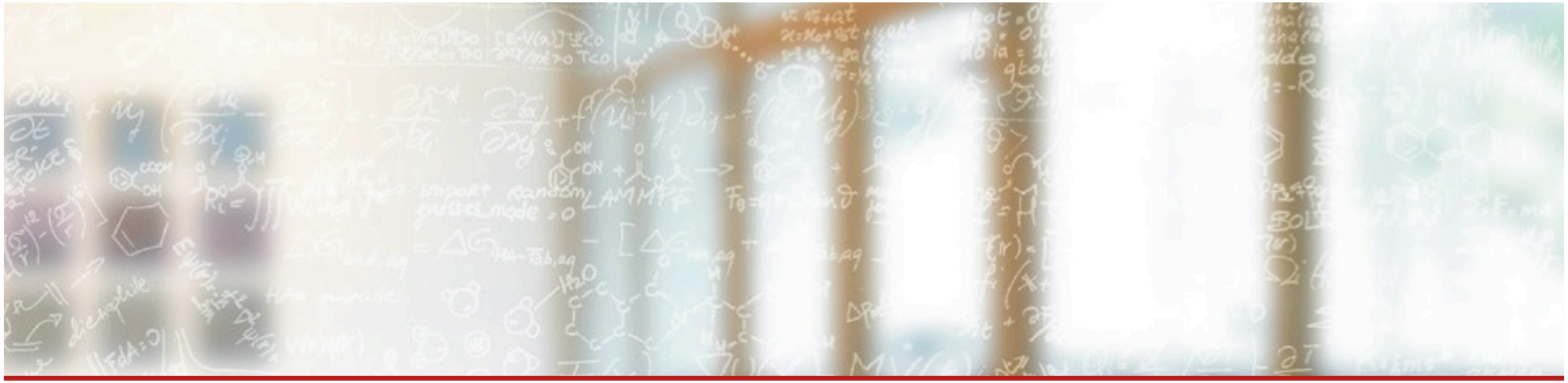
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Detecting and Managing GPU Failures

Cray User Group 2015

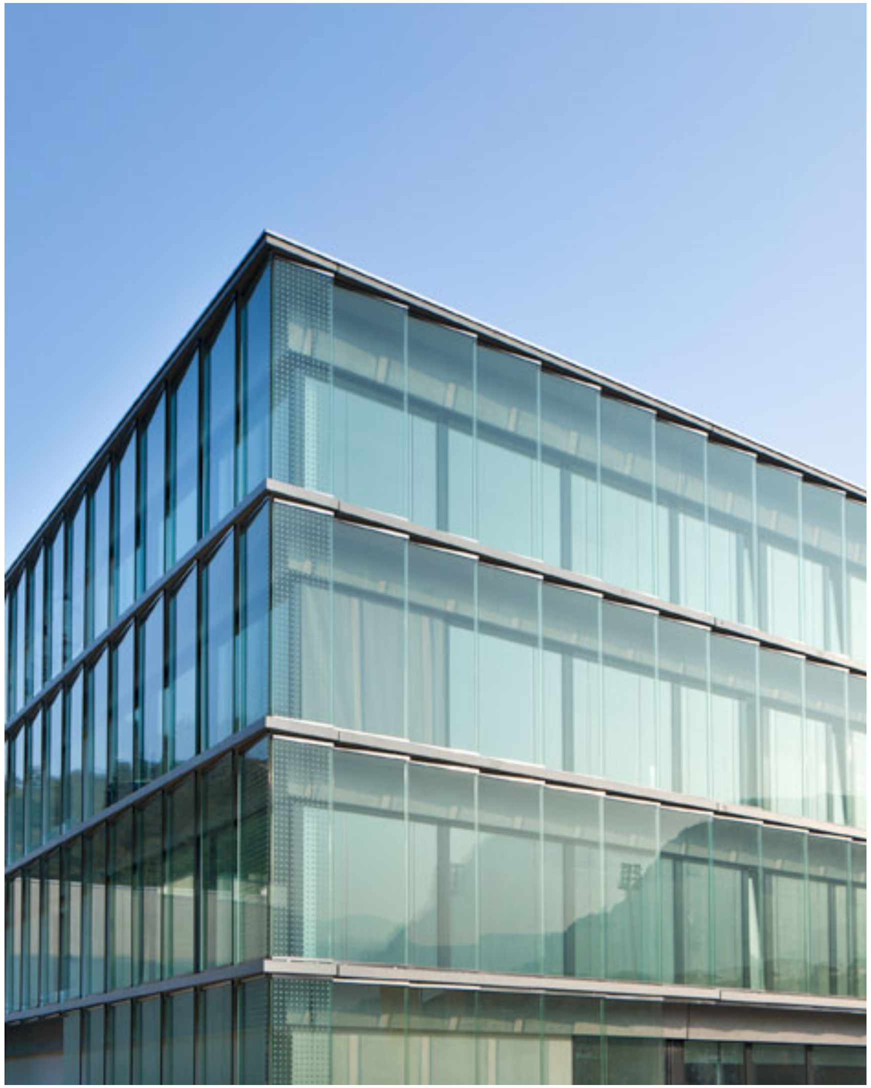
Nicholas P. Cardo, CSCS

April 29, 2015

Theme

> *"Strive for perfection in everything. Take the best that exists and make it better. If it doesn't exist, create it. Accept nothing nearly right or good enough."*
>
> Sir Henry Royce

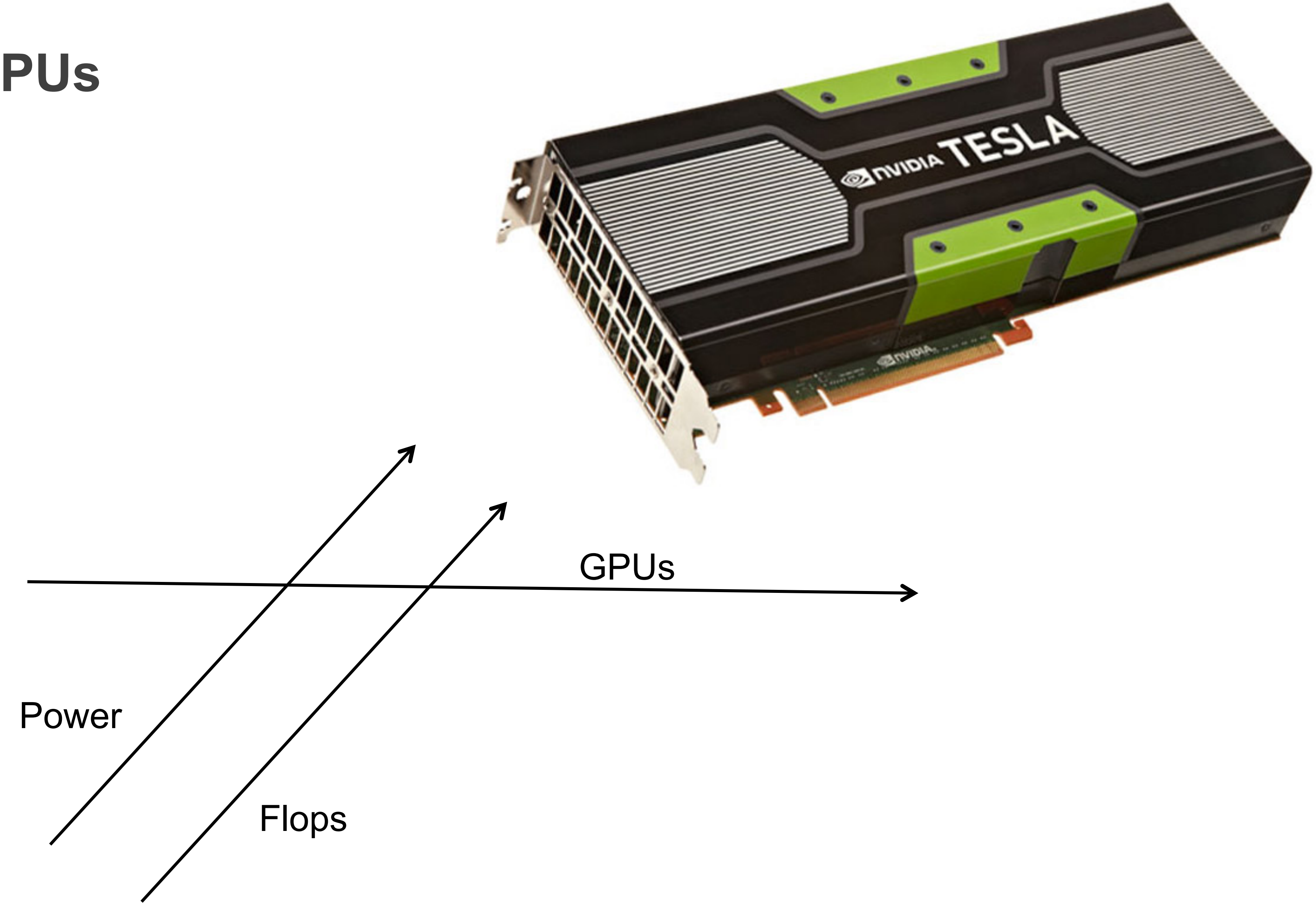
Centro Svizzero di Calcolo Scientifico



The Swiss National Supercomputing Centre, develops and provides the key supercomputing capabilities required to solve important problems to science and/or society. The Centre enables world-class research with a scientific user lab that is available to domestic and international researchers through a transparent, peer-reviewed allocation process.

FRANCE
GERMANY
Rhine
Bodensee
Basel
Winterthur
Zürich
Aare
Biel
LIECHTENSTEIN
AUSTRIA
Lucerne
Neuchâtel
BERN
SWITZERLAND
Davos
Inn
Lake Geneva
Interlaken
Jungfrau
Lausanne
St. Moritz
Montreux
Ticino
Rhône
Geneva
Matterhorn
Lugano
ALPS
ITALY

GPUs



GPUs

Power

Flops

Piz Daint – XC30

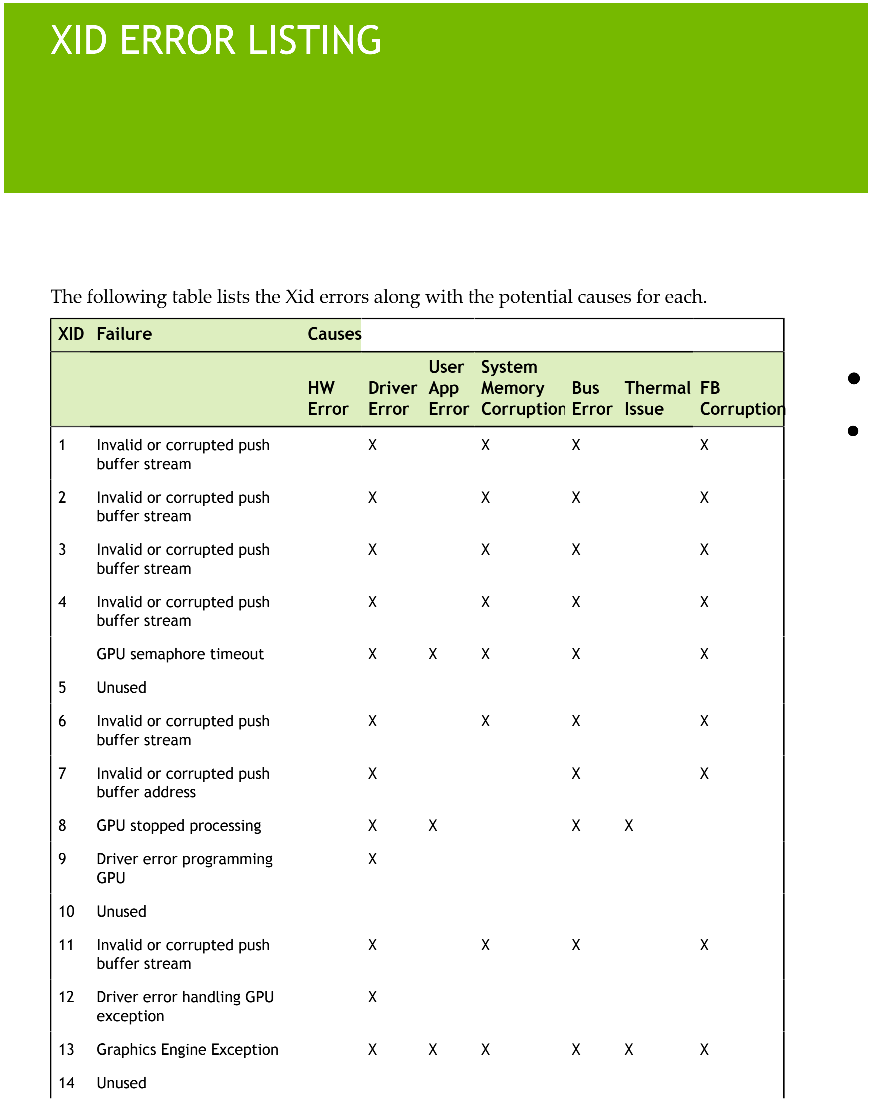
- 5,272 Compute Nodes
  - 8-core E5-2670 2.6 Ghz
  - 32 GB DDR3
  - NVIDIA K20x GPU
    - 6 GB GDDR5
- 52 Service Nodes
- 1.5 TFlops/Node
- 7.8 PFlops/System

Top500: #6 Green500: #9

Failure

- Node Level Diagnostics – *good*
- GPU diagnostics – *lacking*
- Need to detect GPU related issues – *before users*

Technical Error Identification



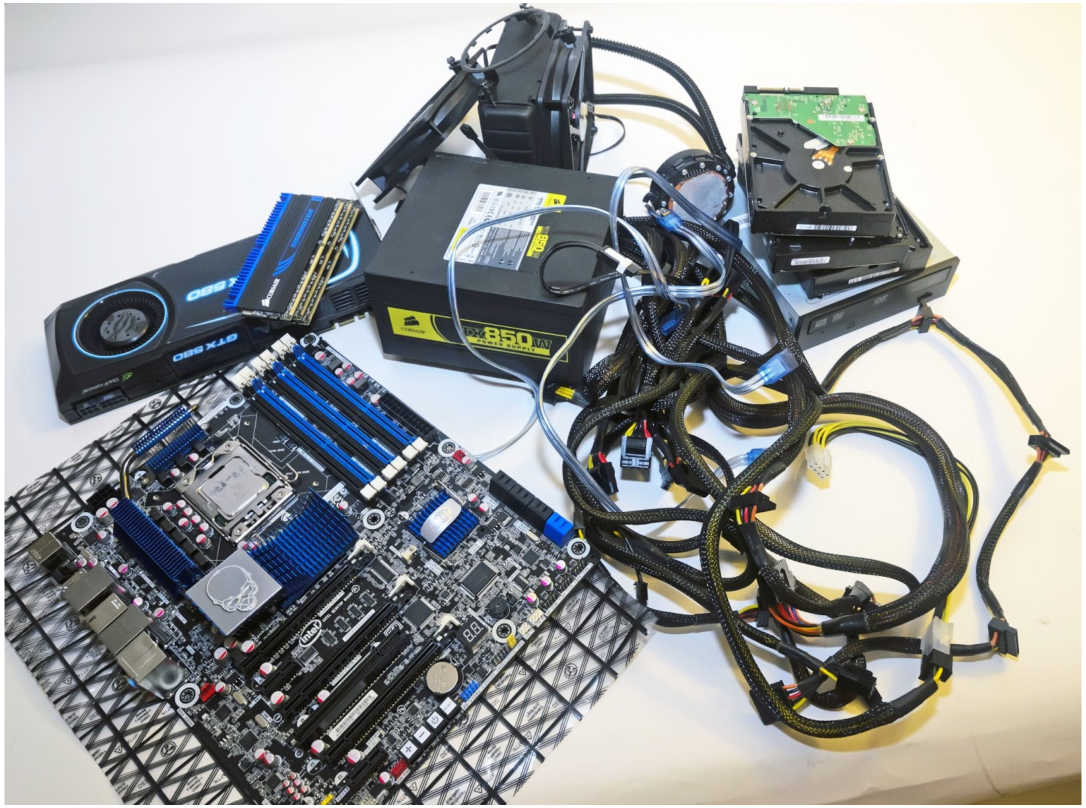
XID ERROR LISTING

The following table lists the Xid errors along with the potential causes for each.

XID	Failure	Causes						
		HW Error	Driver Error	User App Error	System Memory Corruption	Bus Error	Thermal Issue	FB Corruption
1	Invalid or corrupted push buffer stream		X		X	X		X
2	Invalid or corrupted push buffer stream		X		X	X		X
3	Invalid or corrupted push buffer stream		X		X	X		X
4	Invalid or corrupted push buffer stream		X		X	X		X
	GPU semaphore timeout		X	X	X	X		X
5	Unused							
6	Invalid or corrupted push buffer stream		X		X	X		X
7	Invalid or corrupted push buffer address		X			X		X
8	GPU stopped processing		X	X		X	X	
9	Driver error programming GPU		X					
10	Unused							
11	Invalid or corrupted push buffer stream		X		X	X		X
12	Driver error handling GPU exception		X					
13	Graphics Engine Exception		X	X	X	X	X	X
14	Unused							

- Errors are documented
- http://docs.nvidia.com/deploy/xid-errors/

Common Hardware Failures



- Easy to detect
  - IT DON'T WORK

If its broke, fix it…

GPU has Fallen Off the Bus

GPU Has Fallen Off the Bus

- Very annoying
- Could be Hardware or Software
- Bugs Filed
  - 789858, 790527, 804390, 808113, 808114, 814107, 818374
- Corrective Action
  - Remove and clean GPU

GPU Killed by Application

- Symptom
  - User application hangs
  - User application fails to make progress
  - Compute node dies
  - Did I mention Fallen off the Bus?

- Corrective Action
  - Run a GPU optimized HPCG
    - Nodes will crash and die
  - If no hardware failure, clean and reseat GPU

Slow GPUs

- Symptom
  - Users report application performance issues
  - Reports of slow nodes

“one of my jobs running on daint seems to be much slower than the other equivalent jobs. I suspect a broken node.”

- Corrective Action
  - Can reproduce with DGEMM – 40% reduction in performance
- Root Cause 99%
  - Bug 822829
  - Link width reduced from 16x to 8x
  - Detectable with nvidia-smi

```
GPU Link Info
    Link Width
        Max        : 16x
        Current    : 8x
```

System Regression Test Suite

- GPU Related System Checks
  - GOM setting
  - GPU exists
  - Link Width set to 16x
  - ~10 minutes

- GPU Killer
  - ~20 minutes

- Regression Test Suite
  - DGEMM
  - Representative Applications
  - Fully automated
  - ~20 minutes

Metric of Success

- What is the true measure of success?
- Quicker system validation
  - Previously: 3 hours post boot
  - Now: < 1 hour post boot
    - 10,544 node hours returned to service
- Suspect nodes removed from service
  - Detected before users
  - Improved user experience
  - Higher application success rate

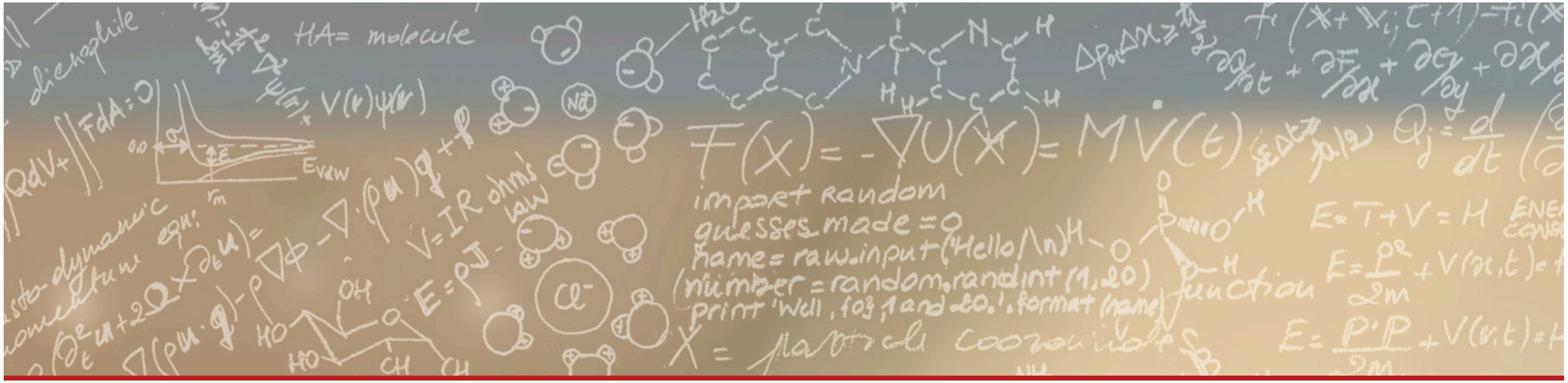
What's Next

- NVIDIA CUDA Toolkit 6.5
  - Improvements in error detection and recovery
  - Could eliminate certain failures
  - Could better identify failure conditions
  - Scheduled for May 11th

- Better Error Tracking
  - Need to track all XID error codes
  - Better error detection, improves reliability

Conclusion

- User Experience Successfully Improved
- Continue to make improvements



Thank you for your attention.