ECE 6930-004
HPC Fault Tolerance
Course Project – Undergraduate Description
Due: 5:00 PM 7 December 2018

Project Overview:
For undergraduates (groups of at most 4) there is a predefined project, but you are free to define your own project. In several papers and lectures, we have studied the reliability of systems, learned about soft errors, and explored hardware and software techniques to detect and correct soft errors. In this project, you will explore the rates of soft errors on Clemson's Palmetto cluster. The Palmetto cluster is a heterogenous system in several meanings of the word. First, the system has nodes that contain both CPUs and GPUs, but also it is built up overtime in **phases**. Thus, the system is heterogenous in terms of the architecture of the network, CPUs, and GPUs some dating back several years.

Over the summer CCIT started collecting daily logs of the rate of bit-flips in the memories of all the GPUs in each phase and node of Palmetto. A subset of one of these logs is shown below where the numbers indicate a count of bit-flip errors of that type in a memory location:

```
Single Bit
                Device Memory       : 271
                Register File       : 0
                L1 Cache            : 0
                L2 Cache            : 0
                Texture Memory      : 0
                Texture Shared      : N/A
                CBU                 : N/A
                Total               : 271
Double Bit
                Device Memory       : 0
                Register File       : 0
                L1 Cache            : 0
                L2 Cache            : 0
                Texture Memory      : 0
                Texture Shared      : N/A
                CBU                 : N/A
                Total               : 0
```

Data like that shown above is logged for each GPU in the system. Note that if a node is offiline due to a fail-stop failure there may not be a report for that node. You will be provided with a

tar.gz or a zip file with the daily log files grouped by phase. This data file will be updated periodically as more data is logged. The exact file format and structure of the logs is defined at the end of this file.

Your tasks for this project are as follows:
Your first job is to write a script that parses the data and stores it in a format suitable for analysis. You are free to create or utilize any data structure/library, but you must document what you use and why. Your code must be runnable on Palmetto. In the future, I plan to use one group's code to study the reliability of the GPUs on Palmetto.

Next, we will compute some statistics/metrics based off this data:

- What are the raw number of single and double bit-flip errors encountered on each phase of palmetto?
- Use the logs to compute the number of bit-flips per GB of memory for each memory type.
- Compute the mean time-between bit-flips for each phase and memory type.
- Compute the mean time-between bit-flips for each phase and memory type scaled per GB.
- Construct a histogram of bit-flip frequency over time for each phase.

Based on the above analysis what GPU is the most reliable? Do you see any resemblance of a bathtub curve with other phases being less reliable than newer phases? Why or why not? Are there any strange events where a phase or a node become very unreliable? If so, which ones? Are some nodes/GPUs very unreliable? If so, which ones? Based on the papers that we have read/discussed in class create 3 research questions on your own and answer them. Clearly state your research questions and explain why you created them.

Each figure/table you add to your final report MUST have accompanying text that explains it. This explanation should go beyond the surface level description and discuss potential impact of the results.

Directory Structure:
The directory structure is as follows:

```
Data
|
|--Directories for each day in the form (YYY-MM-DD)
        |-- Directories for each phase in the form (phase*)
                |-- Files for each node in the phase in the form (node*)
```

File Structure:

Each file contains the report for each GPU on the node. That is, if the node has multiple GPUs, there is more than one GPU report in the file. Here is a complete log file from a practically faulty node with 2 GPUs:

```
==============NVSMI LOG==============

Timestamp                           : Sun Sep 30 06:00:01 2018
Driver Version                      : 396.26

Attached GPUs                       : 2
GPU 00000000:02:00.0
        Replays since reset         : 0
    Ecc Mode
        Current                     : Enabled
        Pending                     : Enabled
    ECC Errors
        Volatile
            Single Bit
                Device Memory       : 28
                Register File       : 0
                L1 Cache            : N/A
                L2 Cache            : 0
                Texture Memory      : 0
                Texture Shared      : 0
                CBU                 : N/A
                Total               : 28
            Double Bit
                Device Memory       : 0
                Register File       : 0
                L1 Cache            : N/A
                L2 Cache            : 0
                Texture Memory      : 0
                Texture Shared      : 0
                CBU                 : N/A
                Total               : 0
        Aggregate
            Single Bit
                Device Memory       : 414
                Register File       : 0
                L1 Cache            : N/A
                L2 Cache            : 8589934590
                Texture Memory      : 0
                Texture Shared      : 0
```

```
                CBU                   : N/A
                Total                 : 8589935004
            Double Bit
                Device Memory         : 0
                Register File         : 0
                L1 Cache              : N/A
                L2 Cache              : 0
                Texture Memory        : 0
                Texture Shared        : 0
                CBU                   : N/A
                Total                 : 0
    Retired Pages
        Single Bit ECC                : 10
        Double Bit ECC                : 0
        Pending                       : Yes

GPU 00000000:82:00.0
        Replays since reset           : 0
    Ecc Mode
        Current                       : Enabled
        Pending                       : Enabled
    ECC Errors
        Volatile
            Single Bit
                Device Memory         : 0
                Register File         : 0
                L1 Cache              : N/A
                L2 Cache              : 0
                Texture Memory        : 0
                Texture Shared        : 0
                CBU                   : N/A
                Total                 : 0
            Double Bit
                Device Memory         : 0
                Register File         : 0
                L1 Cache              : N/A
                L2 Cache              : 0
                Texture Memory        : 0
                Texture Shared        : 0
                CBU                   : N/A
                Total                 : 0
        Aggregate
            Single Bit
                Device Memory         : 0
                Register File         : 0
```

```
                L1 Cache                : N/A
                L2 Cache                : 0
                Texture Memory          : 0
                Texture Shared          : 0
                CBU                     : N/A
                Total                   : 0
            Double Bit
                Device Memory           : 0
                Register File           : 0
                L1 Cache                : N/A
                L2 Cache                : 0
                Texture Memory          : 0
                Texture Shared          : 0
                CBU                     : N/A
                Total                   : 0
    Retired Pages
        Single Bit ECC                  : 0
        Double Bit ECC                  : 0
        Pending                         : No
```

GPUs communicate with the rest of the system over a bus. Typically, this has been the PCI-e bus. Any data sent over this channel has the possibility of suffering a data transfer error and must be resent. The frequency of these errors is recorded on the line **Replays since reset.** Note that this counter is reset each time the node is power cycled, or the diver is reloaded.

For all the GPUs, ECC should be turned on and logs detail the frequency of single-bit and double-bit errors in two modes. The first is **volatile**. These counters are reset anytime the GPU is power cycled or the driver reloaded. Before the counters are reset, they are added to the second mode **aggregate**. These counters persist for the lifetime of the GPU. Therefore, tracking how these counter types change overtime allows us to compute the frequency of bit-flips.

All the GPUs on Palmetto use SEC-DED ECC (Single-bit errors are corrected, and double-bit errors are detected but uncorrectable). Texture memory errors may be correctable via a resend or uncorrectable if the resend fails. Single-bit ECC errors are automatically corrected by the HW and do not result in data corruption. Double-bit errors are detected but not corrected. When double-bit errors occur, most systems kill the application rather than let it compute with corrupted data.

NVIDIA GPUs can retire pages of GPU device memory when they become unreliable. This can happen when multiple single-bit ECC errors occur for the same page, or on a double-bit ECC error. When a page is retired, the NVIDIA driver will hide it such that no driver, or application

memory allocations can access it. **Double Bit ECC** is the number of GPU device memory pages that have been retired due to a double-bit ECC error. **Single Bit ECC** is the number of GPU device memory pages that have been retired due to multiple single-bit ECC errors. **Pending Checks** if any GPU device memory pages are pending retirement on the next reboot. Pages that are pending retirement can still be allocated and may cause further reliability issues.