Data Acquisition with GPUs
The DAQ for the Muon $g$-2 Experiment at Fermilab

Wesley Gohn

University of Kentucky

August 4, 2016
Introduction

1. Introduction
   - Why GPUs?
   - Choice of GPU Architecture
   - GPU Programming
   - GPUs in HEP

2. Motivation: Muon g-2
   - Project Status

3. The Muon g-2 DAQ
   - System Requirements
   - GPU Processing

4. Conclusion
Why GPUs?

- Potential to dramatically improve performance at a reasonable cost by parallelizing data processing.
- Good for performing a simple algorithm many times, but less good for complex algorithms.
- Technology was developed initially for commercial applications, so it is well supported.
- Reasonably easy to learn programming environments are available.
Introduction

Choice of GPU Architecture

GPU Architectures

Scientific GPU (i.e. Nvidia Tesla K40)
- 2880 CUDA cores at 740 MHz.
- 288 GB/s memory bandwidth
- 12 GB memory
- ECC memory protection.

Gaming GPU (i.e. Black Titan)
- 2880 CUDA cores at 980 MHz.
- 336 GB/s memory bandwidth
- 6 GB memory
- No ECC memory protection.

Xeon Phi
- 68 cores / 272 threads at 1.4 GHz.
- 34 MB L2 Cache.
- PCIe version 2.0 limits data transfer bandwidth.
Other Hardware Considerations

The motherboard matters.

- Want a motherboard that supports PCIe version 3.0.
- Need a motherboard with a PLX chip, which allows it to distribute the load over 40 lanes in order to operate all PCIe slots at 16x.
- We are using the ASUS X99-E WS/USB 3.1.

PCIe version 2.0 vs 3.0.

<table>
<thead>
<tr>
<th>PCIe version</th>
<th>GPU</th>
<th>Host to device, Pageable</th>
<th>Host to device, Pinned</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>K20</td>
<td>3326.6 MB/s</td>
<td>5028.3 MB/s</td>
</tr>
<tr>
<td>3.0</td>
<td>K20</td>
<td>5628.6 MB/s</td>
<td>6003.6 MB/s</td>
</tr>
<tr>
<td>3.0</td>
<td>K40</td>
<td>6647.8 MB/s</td>
<td>10044.3 MB/s</td>
</tr>
</tbody>
</table>
GPU Programming

The choice: CUDA vs OpenCL?

- Proprietary from Nvidia.
- Only works on Nvidia GPUs.
- More efficient code.
- Detailed documentation and references from Nvidia.

- Open Source.
- Works on many architectures, including Nvidia.
- Relies on community for support.

For our g-2 GPU programming, we are using CUDA.
GPUs in Particle Physics trigger and DAQ


- **Mu3e at PSI**: Track and vertex reconstruction, D. vom Bruch (2015)

- **LHCb at CERN**: Track pattern recognition S. Gallorini (2015)

- **Muon g-2 at Fermilab**: Deadtime-free analyzer of 700 $\mu$s waveforms – the rest of this talk.
We moved a 50’ superconducting magnetic ring from BNL to FNAL in 2013 and fully installed it last year (see talk by V. Tischenko on details of muon storage).

Measurement and stabilization of the magnetic field is nearly complete (see poster by B. Kiburg).

Detector systems will be installed this year (see talk by J. Kaspar).

Plan for data taking to begin in early 2017 (see talk by C. Polly tonight for more details).
Motivation: Muon g-2

Experiment details and DAQ requirements

- Muons will be injected into the ring at an average rate of 12 Hz consisting of sequences of eight successive 700 $\mu$s fills with 11 ms fill separation.
- Positrons resulting from muon decays are detected using twenty-four electromagnetic calorimeters, each comprised of $9 \times 6$ arrays of PbF$_2$ crystals read out by SiPMs.
- The full 700 $\mu$s waveforms are read out by 1296 channels of custom $\mu$TCA 800 MHz, 12-bit, waveform digitizer (see poster by D. Sweigart for details).
- For a 12 Hz spill rate the time-averaged rate of raw ADC samples is 18.6 GB/s in total.
Layered array of commodity, networked processors

FE layer for readout of digitizer (calo), MHTDCs (straws)

BE layer for assembly of event fragments, storage

Slow control layer for setting, monitoring of HVs, etc.

Online analysis layer using art+JS for monitoring the integrity of raw data, physics data.
MIDAS

- MIDAS is a data acquisition software developed at PSI and also used extensively at TRIUMF.
- Includes web interface for easy control.
- Frontend acquisition code written in C/C++ with CUDA.
- Javascript based analyzer for online data monitoring via a web gui.
- Data will be written to tape as MIDAS datafiles.
The Muon g-2 DAQ
CPU multithreading with mutex locks

CUDA kernel routines: Data is copied to GPU memory in GPU thread, and then accessed by the following functions to identify and save islands, which are copied back to the computer memory and sent to the event builder.

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of threads</th>
<th>Execution time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute pedestals as average of first 100 samples</td>
<td>54</td>
<td>0.1</td>
</tr>
<tr>
<td>Determine if threshold is passed for each sample</td>
<td>560k</td>
<td>1.7</td>
</tr>
<tr>
<td>Add pre-samples and post-samples to each island</td>
<td>560k</td>
<td>0.1</td>
</tr>
<tr>
<td>Check to see if any islands have merged</td>
<td>560k</td>
<td>0.2</td>
</tr>
<tr>
<td>Save an array of identified islands</td>
<td>560k</td>
<td>0.2</td>
</tr>
<tr>
<td>Sum all waveforms</td>
<td>560k</td>
<td>1.2</td>
</tr>
<tr>
<td>Decimate the sum for the Q-method</td>
<td>17.5k</td>
<td>0.3</td>
</tr>
<tr>
<td>Make a fill-by-fill sum of waveforms</td>
<td>30M</td>
<td>2.4</td>
</tr>
</tbody>
</table>
T and Q Methods

CUDA routines process data with two complimentary methods.

- **T-method**
  - Positron events in the calorimeter are individually identified, sorted and fit to obtain time and energy.
  - All events above an energy threshold are included.
  - $\vec{w}_a$ is determined from a fit to a pileup-subtracted histogram.
  - This was the method used in BNL E821.

- **Q-method**
  - Individual positron events are not identified.
  - Detector current is integrated as a proxy for event energy.
  - No pileup correction is necessary.
GPU cluster containing 28 Nvidia Tesla K40 GPUs is up and running.

Three of the four required backend computers are operational.

We have 40 TB of RAID volume installed for short-term data storage.

Fiber-optic 10 Gbe network has been installed.

A Meinberg GPS unit is used to timestamp fills in the DAQ.
DAQ testing

- DAQ has been tested using a simulator that generates realistic data rates for all 24 calorimeters, which allows us to test our GPU pulse-finding algorithms.
- The full DAQ was used to read data from a full calorimeter at a SLAC test beam in June, 2016.
- Characterization of the 1296 WFDs required for the experiment is underway.
- Full commissioning with all detector systems will take place starting in October, 2016.
The new muon $g-2$ experiment will run at Fermilab beginning in 2017 with the goal of reaching $20 \times$ the BNL statistics.

A new state-of-the-art data acquisition system utilizing parallel data processing in a hybrid system of multi-core CPUs and GPUs is required to achieve the necessary data rates.

The DAQ will acquire data from 1296 channels of custom $\mu$TCA waveform digitizers, as well as straw trackers and auxiliary detectors at a rate of 18 GB/s.

The construction of the DAQ has been completed (with a few small pieces to be added), and commissioning is underway.

First muons expected in March, 2017.
Physics of Muon $g$-2

In the standard model, the muon is a spin 1/2 pointlike particle.

It has a magnetic dipole moment of $\vec{\mu} = g \frac{q}{2m} \vec{s}$, with $g = 2$ for a pointlike particle (Dirac).

Additional effects from QED, electroweak theory, and hadronic factors move the standard model prediction of $g$ away from 2. It has become customary to measure this discrepancy, $g$-2.

If a discrepancy with the standard model value is found, beyond standard model contributions to $g$-2 could come from SUSY, dark photons, or other new physics (NP).

$$a_\mu = a_\mu^{QED} + a_\mu^{EW} + a_\mu^{QCD} + a_\mu^{NP}$$
Measurements of $g$-2

- BNL E821 measured $g$-2 to have a 3.3$\sigma$ discrepancy from the standard model (2006).
- Fermilab E989 will measure 20 times the number of muons, reducing the uncertainty on this measurement by a factor of 4.
- Without theory improvements, discrepancy could reach $>5\sigma$.

\[
a_\mu \equiv \frac{g-2}{2} \\
\vec{\omega}_a = -\frac{Qe}{m}[a_\mu \vec{B} - (a_\mu - \left(\frac{mc}{p}\right)^2) \vec{\beta} \times \vec{E}] 
\]
Measurements of $g$-2

- BNL E821 measured $g$-2 to have a 3.3σ discrepancy from the standard model (2006).
- Fermilab E989 will measure 20 times the number of muons, reducing the uncertainty on this measurement by a factor of 4.
- Without theory improvements, discrepancy could reach > 5σ.

\[ a_\mu \equiv \frac{g-2}{2} \]

\[ \vec{\omega}_a = -\frac{Qe}{m} \left[ a_\mu \vec{B} - \left( a_\mu - \left( \frac{mc}{p} \right)^2 \right) \frac{\vec{\beta} \times \vec{E}}{c} \right] \]
Measurement will utilize 24 calorimeters (each composed of 54 PbF$_2$ crystals read out by SiPMs), 3 straw trackers, and several auxiliary detectors.

Each calorimeter will be readout by a custom WFD in a $\mu$TCA crate with an AMC13 control module controlled via IPBus.
GPU Processing

- Data will be processed in an array of 24 GPUs (One GPU per calorimeter)
- Utilizing NVIDIA TESLA K40 GPUS
  - Peak double precision floating point performance: 1.43 Tflops
  - Peak single precision floating point performance: 4.29 Tflops
  - Memory bandwidth 288 GB/sec
  - Memory size (GDDR5): 12 GB
  - CUDA cores: 2880
- Data processing code is written using CUDA.

Results of bandwidth tests:

<table>
<thead>
<tr>
<th>Frontend Machine</th>
<th>GPU</th>
<th>Host to device, Pageable</th>
<th>Host to device, Pinned</th>
</tr>
</thead>
<tbody>
<tr>
<td>FE01</td>
<td>K20</td>
<td>3326.6 MB/s</td>
<td>5028.3 MB/s</td>
</tr>
<tr>
<td>RAVE01</td>
<td>K20</td>
<td>5628.6 MB/s</td>
<td>6003.6 MB/s</td>
</tr>
<tr>
<td>RAVE01</td>
<td>K40</td>
<td>6647.8 MB/s</td>
<td>10044.3 MB/s</td>
</tr>
</tbody>
</table>
Backup Detectors and Backend Electronics

Event building test

12Hz Event builder data performance

- 100 MB/sec total rate
- >80MB/s
- 50 MB/sec total rate

Data volume per calorimeter frontend per spill (MB)

W. Gohn (U. Kentucky)
Data Acquisition with GPUs
GPU Processing Time

Time is dominated by memcpy to GPU.

- MFE_thread sent data
- MFE_thread got data
- gpu_thread process data
- gpu_thread copy data
- tcp_thread got data
- tcp_thread got header

- 1.25 GB/s physical limit
- 0.75 GB/s required rate
- 1.0-1.25 GB/s measured rate

\[ \text{tcp\_thread} \Delta t = 20-25 \text{ ms} \]
\[ \text{gpu\_thread} \Delta t = 12-15 \text{ ms} \]
\[ \text{mfe\_thread} \Delta t = 2-10 \text{ ms (poll)} \]

meets the FE specs for 60.4 MB 12Hz readout and GPU processing
Test Stands

- DAQ was tested using test stands operating in parallel at Fermilab and U. of Kentucky
- Currently includes backend, frontend, gateway, and $\mu$TCA crate with WFD and AMC13
AMC13 Simulator

- Generates realistic waveforms and packs the data in the AMC13 data format.
- Allows us to exercise the DAQ without dependence on hardware.
- Plan to develop this into a tool that will recreate the full spectrum of DAQ input, which will be used for testing the complete data acquisition system.
Data unpacking in GPU

Unpacking time per fill for CPU, GPU and CPU->GPU algorithms

-unpack from CPU to GPU, via loop over CUDA memcpy's
-unpack in CPU, loop over memcpy's
-copy individual array elements with nsegments*nsamples threads