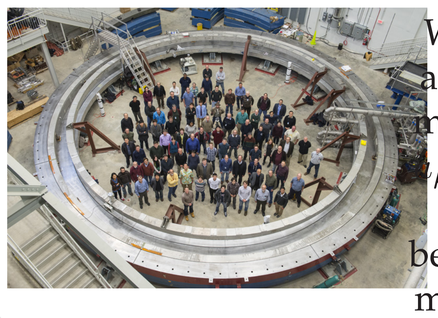
Data Acquisition with GPUs for the Muon g-2 Experiment at Fermilab

Wesley Gohn, University of Kentucky

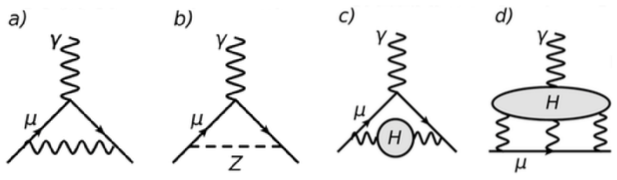
Goals



We will measure the anomalous magnetic moment of the muon, $a_\mu \equiv (g-2)/2$, to 140 ppb, a factor of 4 better than it has been measured previously.

Physics of muon g-2

- In the Dirac model, the muon is a spin 1/2 pointlike particle [1] and has a magnetic dipole moment of $\vec{\mu} = g\frac{q}{2m}\vec{s}$, with $g = 2$.
- Additional effects from QED, electroweak theory, and hadronic factors move the standard model prediction of g away from 2. We 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, extra dimensions, or other new physics.

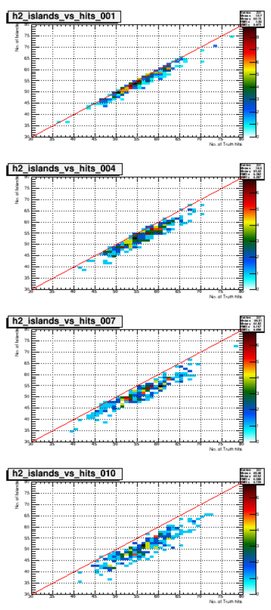


DAQ Requirements

- Accomodate a 12 Hz average rate of muon spills that consists of sequences of eight successive 700 μs spills with 11 ms spill-separations
- Readout, process, monitor and store data obtained from the twenty-four electromagnetic calorimeters, each comprising 9×6 arrays of PbF_2 crystals with SiPMs and read out by 1296 channels of custom 800 MHz, 12-bit, waveform digitizers at a time-averaged rate of 18.6 GB/s.
- Provide both the readout of the raw ADC samples and the derivation of T-method, Q-method, and other calibration, diagnostic and systematic datasets.

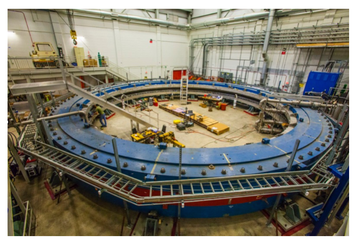
Commissioning

- DAQ is operational at Fermilab and running at it's full capacity, including 28 Tesla K40 GPUs.
- Currently reading data from 7 full μTCA crates each with 60 channels of WFDs and running a simulator mode used to process data equivalent of 24 crates of WFDs.



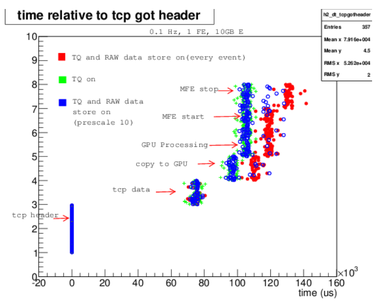
Project Status

- The ring was moved from BNL to FNAL in 2013.
- It was installed in our new MC1 building, cooled to superconducting temperatures, and powered up to full field in October, 2015.
- Shimming of the magnet was completed in August, 2016.
- Installation of detectors will begin imminently.
- Plan for data taking to begin in April 2017, and it will run for 3 years.



Data Processing

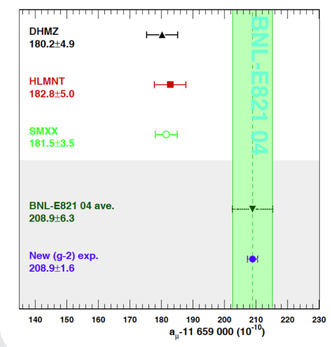
- Most time is spent reading data from TCP socket.
- Significant time copying data into GPU.
- GPU processing of data takes just <4 ms to execute 33M threads.

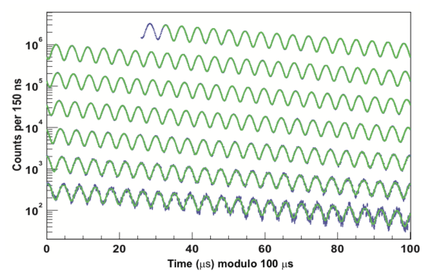


Function	Threads	Time (ms)
Compute pedestals as average of first 100 samples	54	0.1
Determine if threshold is passed for each sample	560k	1.7
Add pre-samples and post-samples to each island	560k	0.1
Check to see if any islands have merged	560k	0.2
Save an array of identified islands	560k	0.2
Sum all waveforms	560k	1.2
Decimate the sum for the Q-method	17.5k	0.3
Make a fill-by-fill sum of waveforms	30M	2.4

Measurements of g-2

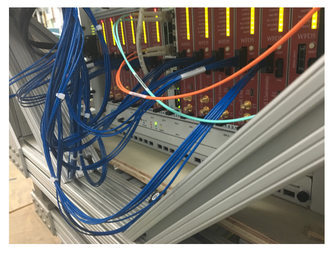
- BNL E821 measured g-2 to have a 3.3σ discrepancy from the standard model (2006) [2].
- Fermilab E989 will measure 21 times the number of muons, reducing the uncertainty on this measurement by a factor of 4.
- Without theory improvements, discrepancy could reach $> 5\sigma$. With expected theoretical improvements [3], it could reach $> 7.5\sigma$.

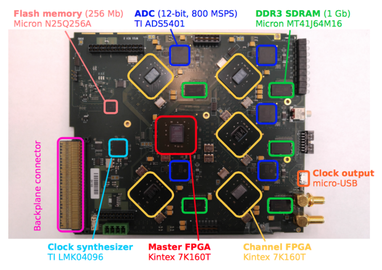




Detectors and Electronics

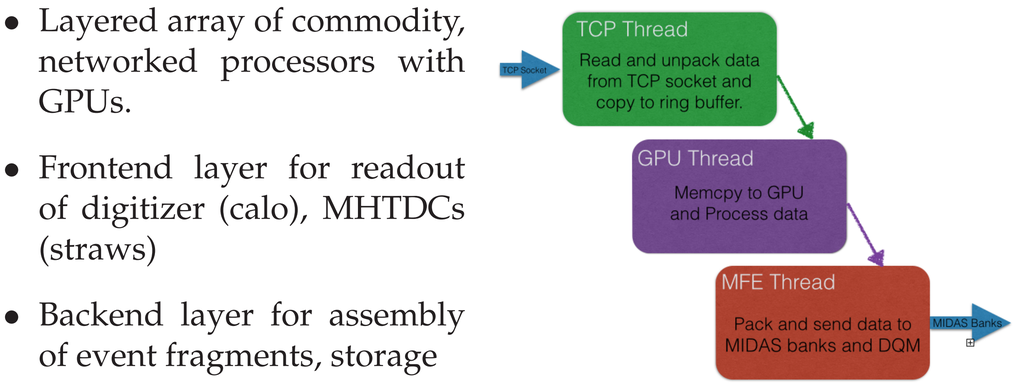
- Measurement will utilize 24 calorimeters (each composed of 54 PbF_2 crystals read out by SiPMs), 3 straw trackers, and several auxiliary detectors.
- Each calorimeter will be readout by a custom WFD in a μTCA crate with an AMC13 control module controlled via IPBus. Data is transferred over 10 Gbps fiberoptic cable.

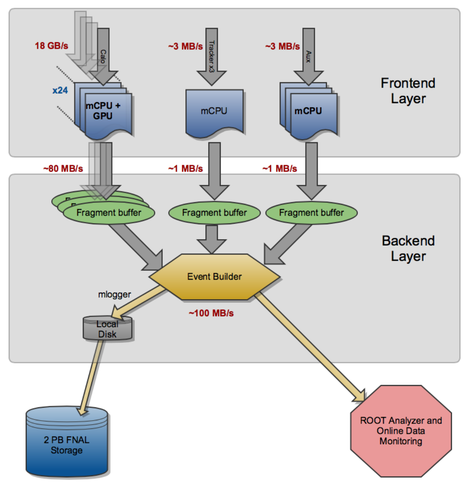




Data Acquisition System

- Layered array of commodity, networked processors with GPUs.
- Frontend layer for readout of digitizer (calo), MHTDCs (straws)
- Backend layer for assembly of event fragments, storage
- Slow control layer for setting, monitoring of HVs, etc.
- Online analysis layer using *art* and node.js to stream data from mserver.





- We are using MIDAS as our data acquisition software [3].
- Much experience with MIDAS in the collaboration from MuLan, MuCap and MuSun.
- Frontend acquisition code written in C/C++ with IPBus libraries for communication with μTCA electronics and CUDA libraries for data processing in GPUs.
- Mutual Exclusion (mutex) locks are used to break CPU processing into three threads.
- Data will be processed in an array of 24 GPUs (One GPU per calorimeter)
- Utilizing NVIDIA TESLA K40 GPUS
  - 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 libraries in C++.

References

1. P.A. Dirac, Proc.Roy.Soc.Lond. A117, 610 (1928).
2. Muon g-2, G. Bennett *et al.*, Phys.Rev. D73, 072003 (2006).
3. Blum *et al*, "The Muon (g-2) Theory Value: Present and Future" (2013), arXiv:1311.2198.
3. http://midas.triumf.ca, TRIUMF MIDAS homepage, 2015.
4. J. Grange *et al.*, Muon (g-2) Technical Design Report (2015, arXiv:1501.06858.

Acknowledge

This research was supported by the National Science Foundation MRI award PHY-1337542. This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Nuclear Physics under Award Number DE-FG02- 97ER41020.