Jefferson Lab Procurement

USQCD 2021

All Hands' Meeting

Amitoj Singh

Friday, April 30

Jefferson Lab

U.S. DEPARTMENT OF ENERGY | Office of Science

JSA

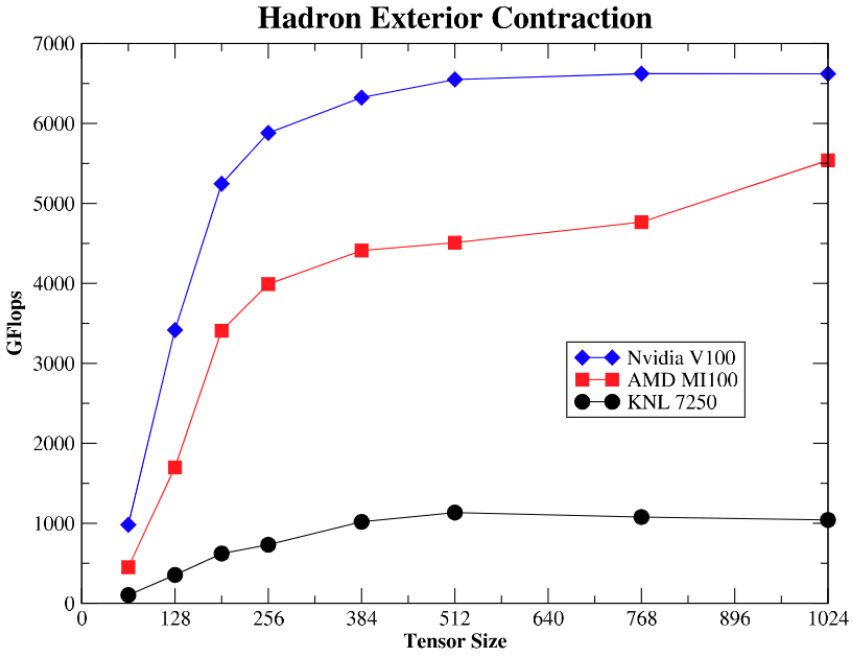
Introduction

- Jefferson Lab currently has two types of clusters, with/without GPU:
  - 440 node Xeon Phi / KNL cluster (16p/18p)
    - Single socket 64 core KNL (with AVX-512 8 double / 16 single precision)
    - 192 (98) GB main memory / node 16p (18p)
    - 16 GB high bandwidth on package memory (6x higher bandwidth)
    - 100 Gbps bi-directional OPA network fabric (total 25 GB/s/node) 32 nodes / switch, 16 up-links to core / switch
    - Total: 3.168 M node-hours = 202.75 M KNL-core-hours
  - 32-node GeForce GPU cluster (19g)
    - Eight-GPU RTX-2080 nodes
    - 8 GB memory per GPU, 192 GB memory per node.
    - Each on 100 Gbps OPA Fabric
    - Total: 230.4 K node-hours = 1.84 M RTX2080-GPU-hours
- Rest of the talk will detail the procurement process for the upcoming Jefferson Lab LQCD cluster.

2020 Hardware Alternatives Analysis

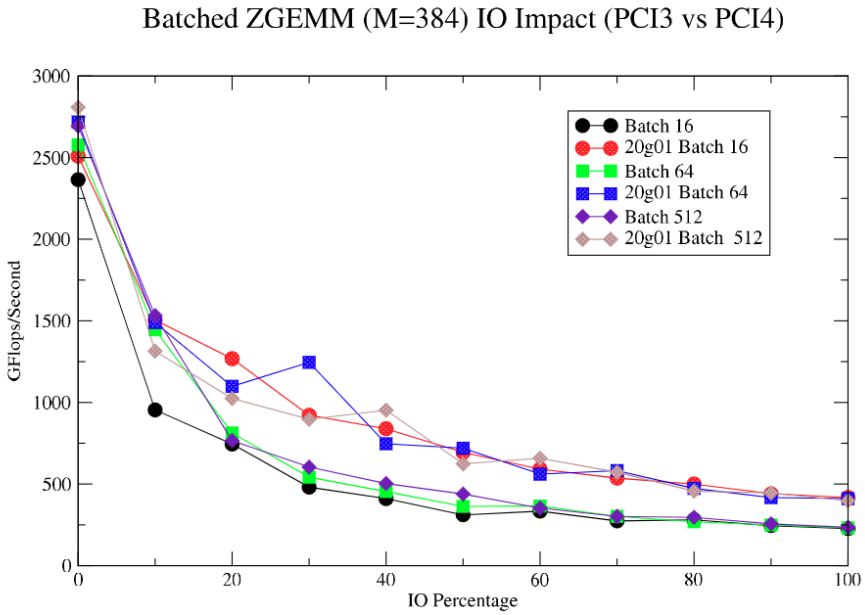
- Budget: $470K in FY20 + $450K in FY21 = $920K.
- Workloads on KNL dominated by contraction-style work so focus of new acquisition should be on propagator and to some extent gauge generation.
- Based on knowledge of existing/emerging projects & input from NP community - rough weighting of overall workloads:
  - 10%-gauge generation
  - 30%-contractions
  - 60%-propagators
- For NP workloads, key performance metrics are:
  - mixed half-double or single-double precision multi-grid clover solvers. *Do not require ECC.*
  - batched ZGEMM complex matrix multiply as well as larger matrix multiplies used in generating propagators and perambulators. *Require ECC.*
- January 2018 FWP had estimated a deployment of 695 TF performance using ECC memory GPUs.

Performance comparison – a few key plots



What the AMD GPU lacks in performance it makes up in price !!

PCI3 vs PCI4 shows a 2x performance bump as expected.



Jefferson Lab

2020 Hardware Alternatives Analysis (cont'd)

- Proposed hardware alternatives using benchmarks Kokkos Dslash & effective ZGEMMs:
  - 8 AMD GPU host (MI50 or MI100) + AMD CPU PCI-gen4.
  - 8 NVIDIA GPU host (V100 or A100). If V100 only PCI-gen3.
  - x86_64, either AMD or Intel based.

- Some observations:
  - Both Frontier & EL Capitan at ORNL will comprise AMD CPUs and GPUs. USQCD codes will have to eventually support this environment.
  - NVIDIA based GPU clusters:
    - Summit with IBM CPUs and V100.
    - Perlmutter@NERSC with A100.


21g

21g Cluster Details

- Early November 2020 a final decision was made on the 21g hardware configuration (2020 HAA), but disbursement of all needed funds was delayed due to a FY21 CR.
- When all needed funding arrived, procurement was able to release the requisition for bid responses on 1/21/21.
- Purchase order for the 21g cluster was awarded to Atipa Technologies on 3/27/21 after a competitive bid process.
- 21g is worth $450K in total. The system consists of:
  - 8 nodes with 8 AMD MI100 GPUs with infinity fabric – 64 GPUs total,
  - AMD Epyc "Rome" CPU, Gigabyte Motherboard, PCIe Gen 4,
  - 1TB Memory and 2TB SSD per node (upgraded, due to favorable lower than predicted final pricing).
  - Mellanox EDR interconnect fabric to a single EDR switch.
  - Anticipate 57.6k node-hours = 0.46M MI100-GPU-hours.
- Delivery is expected by May 19th.
- Open for “friendly user” testing in June.
- Release to production on July 1st to coincide with the new allocation year.

Amitoj Singh

amitoj@jlab.org

Conclusion

- To apply for early access to 21g please contact Robert Edwards edwards@jlab.org. Your feedback will help us.
- Planning to release 21g to the community at large by July 1st.

- Questions ?