Brookhaven
National Laboratory

New hardware options for USQCD

Peter Boyle
High Energy Theory
Physics @ BNL

With thanks to USQCD Requirements Committee

Zihua Dong, Chulwoo Jung, Imran Latif, Meifeng Lin (BNL)

Amity Singh (JLAB)

Jo Fazio, Jim Simone, Alan Prosser (FNAL)

Outline

Aim to summarise work, results and considerations of the BNL procurement requirements committee.

Final report delivered to procurement committee yesterday

BNL Lattice QCD Cluster Requirements Committee Report

Peter Boyle (chair)[1], Zihua Dong[1], Josephine Fazio[2], Chulwoo Jung[1], Imran Latif[1], Meifeng Lin[1], Alan Prosser[2], James Simone[2], and Amitoj Singh[3]

[1]BNL
[2]FNAL
[3]JLAB

April 2024

1 Background

The DOE has made provision for dedicated Lattice QCD computing serving the particular needs of the US-wide lattice gauge theory community at the BNL, JLAB and FNAL national laboratories on a continuing basis of computing installations 2003.

BNL has hosted for USQCD:

- 2004: Columbia, BNL, RIKEN and UKQCD designed QCDOC computers, jointly developed with IBM Research, based on a custom SoC design using PowerPC processors.
- 2011: IBM BlueGene/Q hardware designed in part by Columbia and Edinburgh with IBM Research.
- 2016: Intel Knight's Landing many-core processor cluster
- 2018: Intel Skylake cluster as part of the SDCC institutional cluster.

This document is the report of a committee formed to identify technical requirements for a 2024 hardware installation at Brookhaven National Laboratory.

In the project execution plan, the programme description and functional requirements are as follows. These describe the basic phases of the Lattice QCD workflow: ensemble generation, valence quark propagator calculation and the contraction of quark propagators into correlation functions.

1.1 Program Description

The purpose of the LQCD computing program is to provide the USQCD user community with the mid-scale computing resources required to meet the computational needs of the lattice quantum chromodynamics (QCD) research program.

1.2 Functional Requirements

Three classes of computing are done on lattice QCD machines. In the first class, a simulation of the QCD vacuum is carried out, and a time series of configurations, which are representative samples of the vacuum, are generated and archived. The second class, the quark-propagator phase, computes the propagation of quarks in these snapshots. These jobs obtain solutions to a large, sparse linear system of equations, so they also require large floating-point capabilities The third class, the analysis phase, uses hundreds of thousands of files of hadron correlation functions, which are obtained by sewing together various quark propagators on each configuration. The ensemble averages of these files yield physically meaningful information, such as masses, matrix elements, or cumulants. Heavy-duty reprocessing of these files is needed to estimate statistical and systematic uncertainties.

2 Technical interpretation of the workload

It is useful to reduce the workload to the primitive computing operations that dominate the behaviour. In both dynamical gauge configuration generation with dynamical fermions, and the calculation of valence quark propagators, the bulk of the floating point operations and run time are spent in running numerical inversion of the QCD Dirac matrix. This in practice involves repeatedly applying the Dirac matrix to a current residual vector, and the performance of the Dirac matrix subroutines are the single most critical representative proxy for overall code performance.

Translating workload into technical requirements

Workload requirements (as per project execution plan):

- Gauge configuration generation
  - Benchmark proxy: Multinode Dirac operator
- Quark propagator inversion
  - Benchmark proxy: single node Dirac operator
  - Benchmark proxy: coarse grid preconditioning via batched C/ZGEMM
- Observable contraction
  - Benchmark proxy: Memory bandwidth

Benchmarking

- Aim to fairly compare real world obtainable performance across platforms.
- Used Grid because
  - Supports HIP, SYCL, CUDA and CPU vectorization
  - Used in multiple procurements, including FNAL last year.
  - Wilson, Domain wall, Staggered operators
- Introduced Benchmark_usqcd & benchmarked:
  - GPUs: Nvidia (A100/Perlmutter, H100/SDCC) ; AMD (MI250X, Frontier); Intel (PVC, Aurora)
  - CPUs: Intel SPR/HBM ; Intel SPR/DDR ; AMD Genoa
- Produces a CSV spreadsheet of results for each run.
  - Also developed a new machine burn-in test with bit-level reproduce testing.

8U Universal GPU Systems (4)

	X13 NEW	H13 NEW	X13
			
	SYS-821GE-TNHR DP Intel 8U System with NVIDIA HGX H100 8-GPU and Rear I/O	**AS -8125GS-TNMR2** DP AMD 8U System with AMD MI300X	**SYS-821GV-TNR** DP Intel 8U System with Intel Data Center GPU Max 1550
GPU	8	8	8
GPU-GPU	NVIDIA® NVLink® with NVSwitch™	AMD Infinity Fabric™ Link	Intel® Xe Link Bridges
CPU	2	2	2
CPU Type	5th Gen Intel® Xeon®/4th Gen Intel® Xeon® Scalable processors	AMD EPYC™ 9004 Series Processors	5th Gen Intel® Xeon®/4th Gen Intel® Xeon® Scalable processors
DIMM Slots	32	24	32
Drive Size	2.5"	2.5"	2.5"
Drives	19	18	19
Networking	4x 10G, 2x 25G		2x 10G

GPU A+ Server AS -4145GH-TNMR (Complete System Only ⓘ)

4U AMD quad APU system with 4 AMD Instinct™ MI300A accelerators

Artificial intelligence market has boomed but has high margins, increasing prices

MI300X - AMD next gen GPU **256MB cache, up from 8MB!**

MI300A - AMD hybrid CPU-GPU with 128GB of HBM & no DDR

H100 - Nvidia next gen GPU (yeah, B100 announced)

Intel SPR - Xeon CPU with HBM (option)

AMD Genoa - x86 CPU with excellent performance

Nvidia Hopper 8x H100 @ SDCC

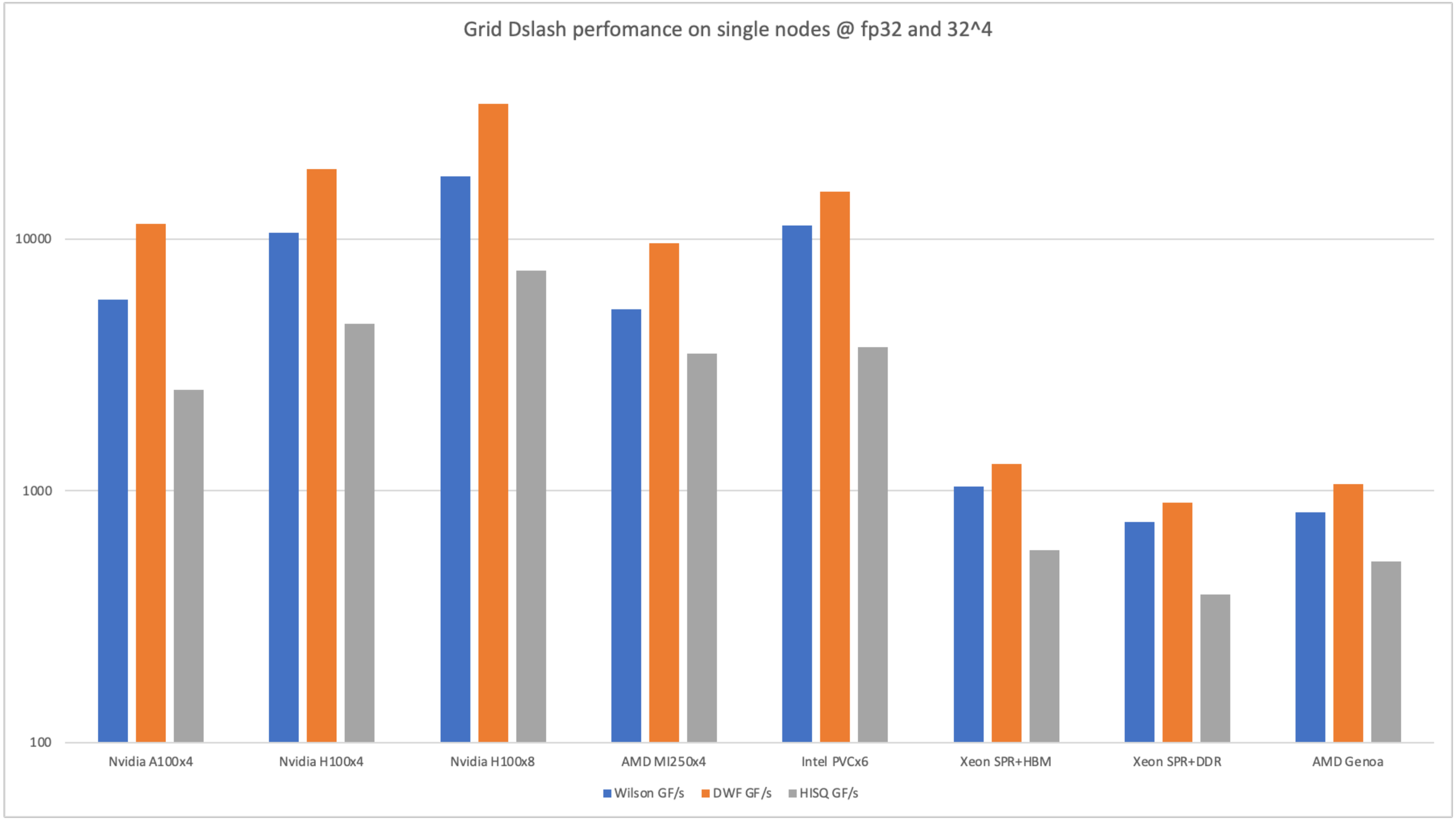
A	B
Memory Bandwidth	
Bytes	GB/s per node
6291456	254.227647
100663296	3500.053754
509607936	10351.75926
1610612736	14353.04629

Communications		
Packet bytes	direction	GB/s per node
4718592	2	207
4718592	3	208
4718592	6	190
4718592	7	204
15925248	2	308
15925248	3	288
15925248	6	296
15925248	7	301
37748736	2	332
37748736	3	339
37748736	6	333
37748736	7	336

ZGEMM				
M	N	K	BATCH	GF/s per rank
16	8	16	256	5.787642
16	16	16	256	275.941053
16	32	16	256	256.925207
32	8	32	256	652.809961
32	16	32	256	180.69161
32	32	32	256	2346.463776
64	8	64	256	2338.287944
64	16	64	256	4660.337778
64	32	64	256	6563.214083
16	8	256	256	1950.83907
16	16	256	256	3647.22087
16	32	256	256	5604.080501
32	8	256	256	2874.041285
32	16	256	256	5651.272758
32	32	256	256	9905.36738
64	8	256	256	4444.295629
64	16	256	256	8589.934592
64	32	256	256	14851.20089
8	256	16	256	1928.415632
16	256	16	256	3717.942604
32	256	16	256	4652.260936
8	256	32	256	2995.931429
16	256	32	256	5122.814046
32	256	32	256	8310.695232
8	256	64	256	4350.655689
16	256	64	256	8146.75132
32	256	64	256	12874.60221

Per node summary table				
L	Wilson	DWF4	Staggered	GF/s per node
8	192	2295	64	
12	974	8668	323	
16	2959	17458	910	
24	9732	28906	3605	
32	17661	34400	7477	

Frontier MI250X4 | SDCC-A100x4 | SDCC-H100x4 | **SDCC-H100x8** | Aurora PVCx6 | SPR Xeon+HBM | SPR+DDR | AMD CPU Genoa - 32x2

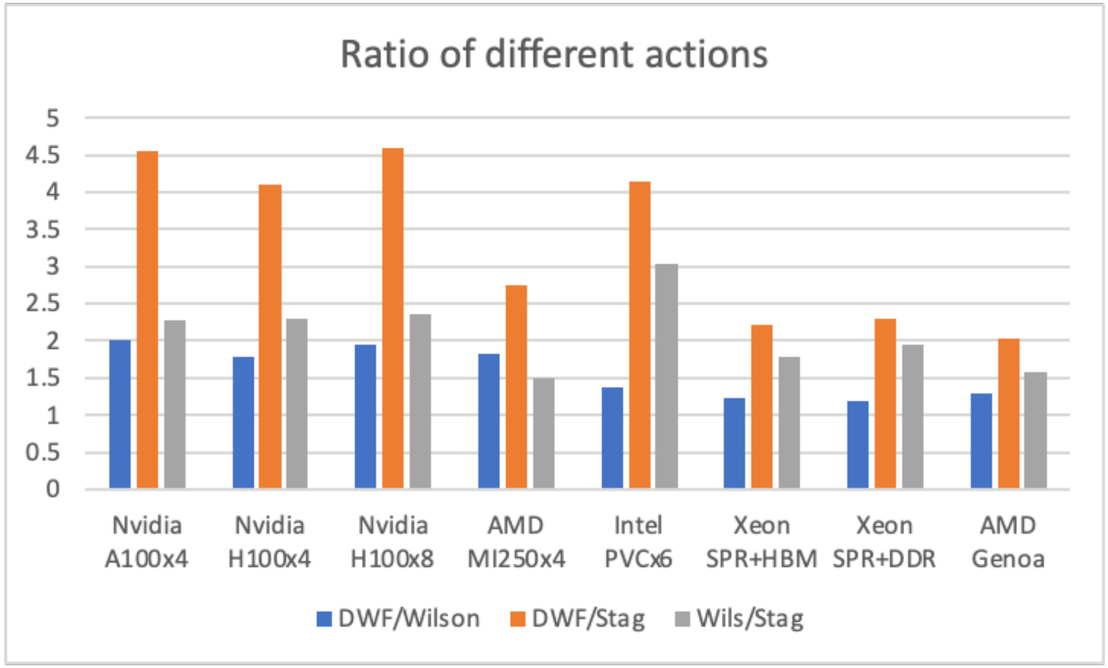

Grid Dslash perfomance on single nodes @ fp32 and 32^4
10000
1000
100
Nvidia A100x4
Nvidia H100x4
Nvidia H100x8
AMD MI250x4
Intel PVCx6
Xeon SPR+HBM
Xeon SPR+DDR
AMD Genoa
Wilson GF/s
DWF GF/s
HISQ GF/s

Metric-1 : DWF performance per node

- Wilson and Staggered less representative
  - Less optimized
  - MultiRHS versions better for valence analysis
  - Results correlate across architectures fairly well: DWF favors GPUs a bit more

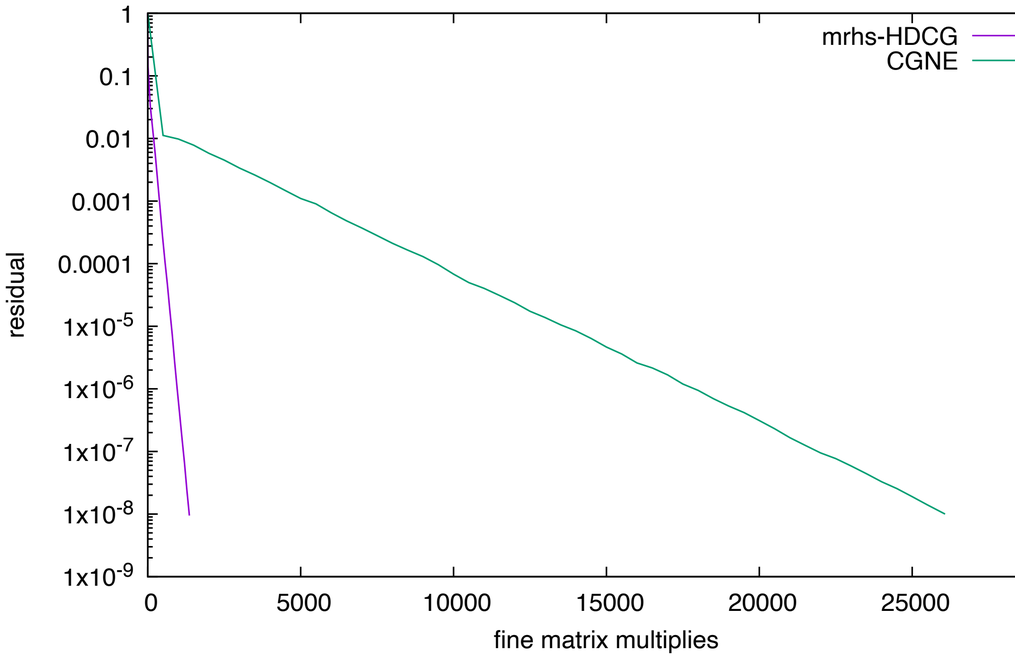
AMD MI250X has a tiny 8MB L2 cache

This shows up as hobbling gauge link reuse in DWF

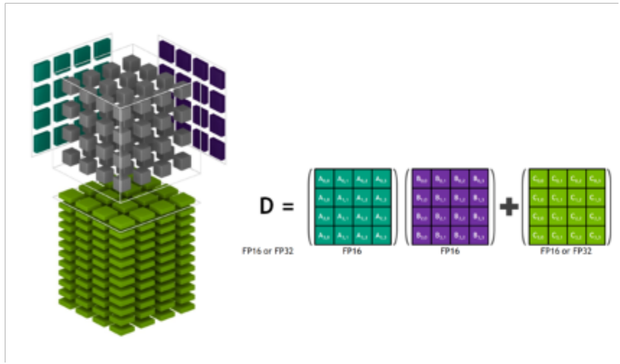


Why batched GEMM ? Multiple RHS multigrid

- mrhs-HDCG solves twelve RHS in 725s on 18 nodes of Frontier
  - CGNE 770s for 1 RHS
- 13x speed up wall clock and 17x reduction in fine matrix multiplies (26000 vs. 1500)
- batched BLAS ZGEMM on GPU on red named routines: 30x speedup!

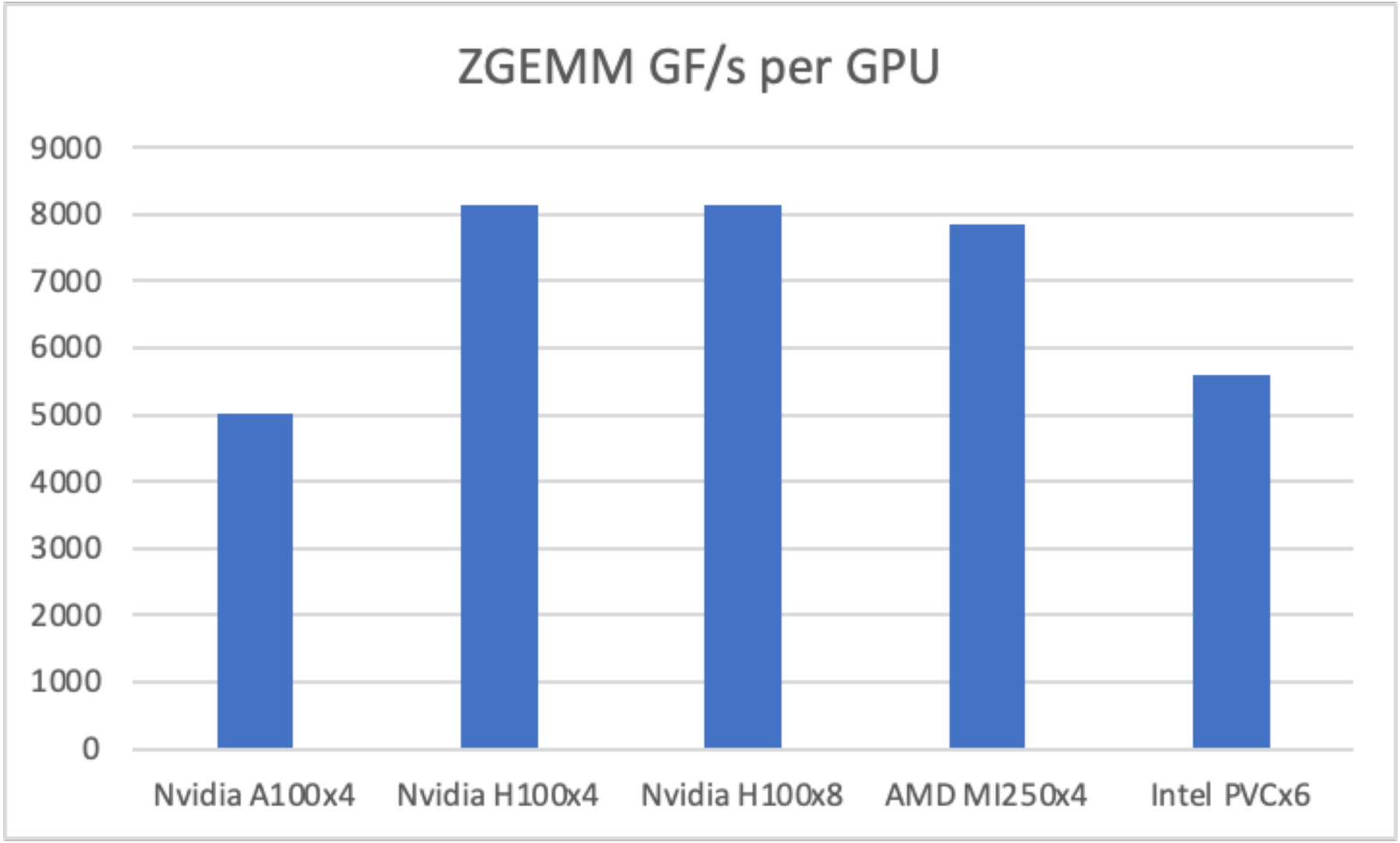


Total	725s
FineSmoother	430s
CoarseSolver	159s
FineResidual	100s
FineLinalg	25s
FineToCoarse	6s
CoarseToFine	5s
Deflate	0.3s



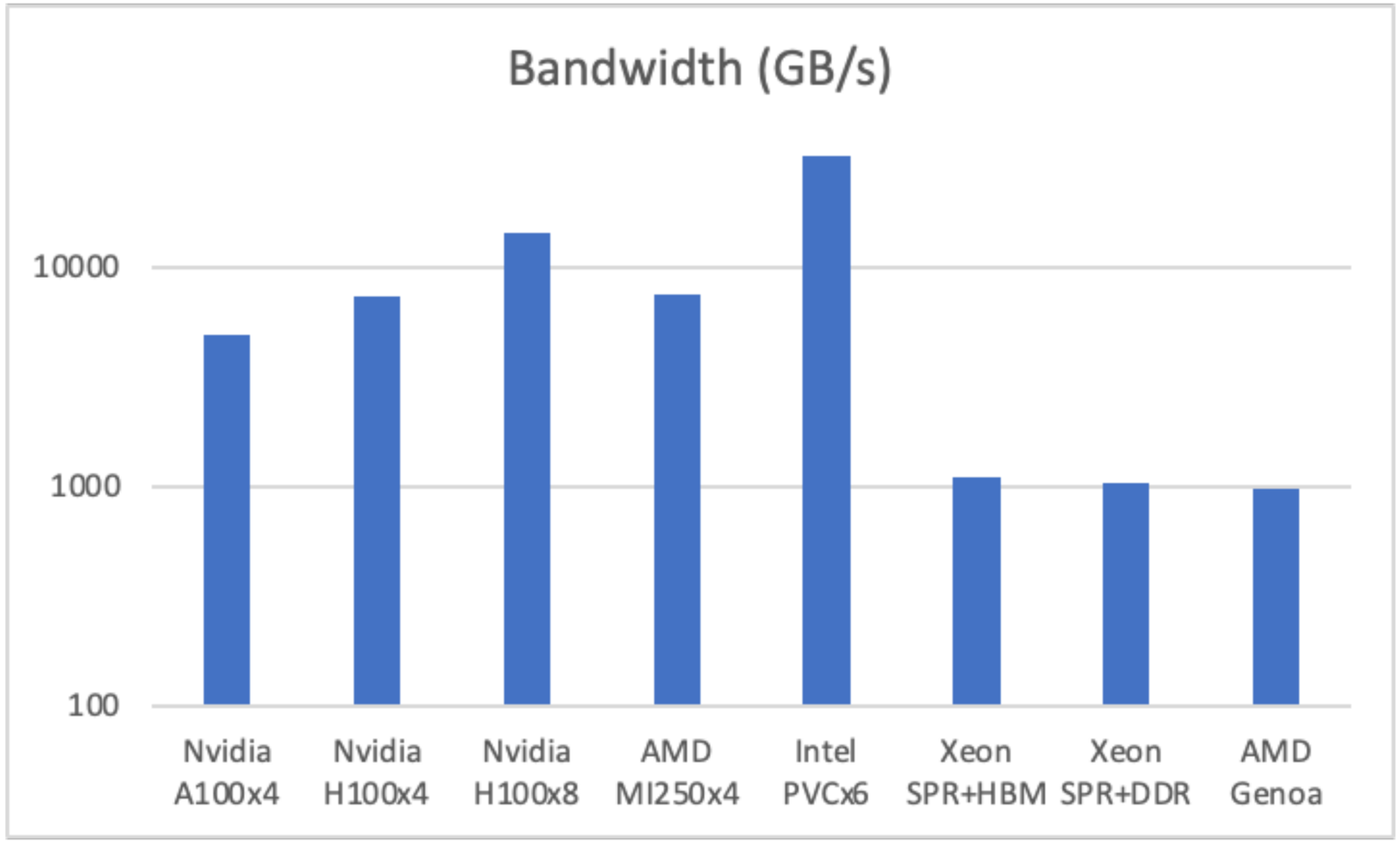
Uses GPU tensor cores to make coarse grid efficient

Batched ZGEMM for multiRHS-multigrid



General computing (i.e. contraction) best measured by memory bandwidth

PVC is flattered here by large cache (vectors did not spill to HBM)



Benchmarking summary:

- (to a good approximation) multi-GPU nodes are 10x faster than CPU nodes.
- Price ratio is around 6x to 8x or more in current market
- Significant lead time on some GPU parts
- Competition between an FNAL style 18 x 4 GPU system vs JLAB still 100x CPU nodes
  - 100 TF/s CPU and 100TB RAM … OR 180TF/s GPU and 18 - 36TB RAM
  - Can use SSD as a memory expander, but programmer overhead
  - Which is better? Depends on what problem you are solving ! If it doesn’t fit the performance is zero!

Type	Vendor	Platform	Performance	GPU count	HDR-200 NICs	Memory
CPU	Intel	Xeon+DDR	0.9		1	≥1TB
CPU	Intel	Xeon+HBM	1.3		1	≥1TB
CPU	AMD	EPYC Genoa	1.1		1	≥1TB
CPU	Nvidia	Grace/ARM	-		1	≥1TB
GPU	Nvidia	H100 x4	19	4	2	≥2TB
GPU	Nvidia	H100 x8	36	8	4	≥2TB
GPU	AMD	MI300X x4	19	4	2	≥2TB
GPU	AMD	MI300X x8	36	8	4	≥2TB

Table 3: The types of nodes that are considered acceptable to satisfy USQCD requirements, and the corresponding score value associated per node. The performance score for MI300X is estimated and must be confirmed via vendor providing us with benchmarking access if bid.

User survey

- Thanks to all responses
- High level of GPU readiness.
    - 14/15 Nvidia ; 8/15 AMD ; 5/15 Intel GPU
- Memory footprint is an issue. 18TB is less than many require.
- Ideally schedule multiple jobs at once rather than timeshare whole cluster
- Local scratch SSD may help mitigate but requires software work

Is your software able to make use of GPU's ?	Does your software plan enable GPUs in the near future?	Which GPUs can your software use?	Assuming greater execution throughput is available from GPU's fixed price, how much faster need
Yes	Yes	Nvidia, AMD, Intel	>4x
Yes	Yes	Nvidia, AMD, Intel	>2x
Yes	Yes	Nvidia, AMD	Always preferable
Yes	Yes	Nvidia, AMD	Always preferable
No	Maybe	None	
Yes		Nvidia	>4x
Yes	Yes	Nvidia, AMD	Always preferable
Yes	Yes	Nvidia, AMD, Intel	>2x
Yes	Yes	Nvidia	Always preferable
Yes	Yes	Nvidia, AMD, Intel	>2x
Yes	Yes		>2x
Yes	Yes	Nvidia	>2x
Yes	Yes	Nvidia	Always preferable
Yes	Yes	Nvidia	Always preferable
Yes	Yes	Nvidia, AMD, Intel	Always preferable

Recommendations

Recommended the following technologies be quoted and competitively assessed:

- AMD MI300X - requires benchmark access.
    - Substantially better memory system than MI250
- Nvidia H100
- AMD Genoa
- Intel SPR + HBM
- Lower priority: Intel SPR + DDR, Nvidia Grace ARM

Considerations

- Software readiness on PVC makes it a risk.
  - Aurora not in production for QCD yet, unlike Frontier
  - AMD and Nvidia GPU’s run QCD on Frontier daily with Grid, QUDA and Chroma ported
- AMD MI300X and Nvidia H100 should be solid platforms
- MI300A integrating CPU and GPU with HBM is a beautiful idea, but we can’t afford enough of them to have adequate memory
- A100 and MI250X are close to end of life cycle products. 5 year spares lifecycle may be required.
  - Lead times and current AI/ML market conditions make procurement challenging. Vendor competition needed.
- CPU’s are substantially slower, but also substantially cheaper.
- Propose score be geometric mean of cluster TF/s and cluster TB memory
  - Rule of thumb: double the memory can be traded for half the throughput due to flexibility around our expected budget

Duration:

- The cluster nodes should be expected to operate for at least five years.

Node technology:

- Technology selection should remain open and made in response to bids
- All compute nodes should be identical to each other.
- CPU or GPU nodes could satisfy requirements.
  - AMD MI300X and Nvidia H100 GPU nodes are acceptable, subject to full speed NVlink/Infinity link and minimum 2TB host memory and 8TB SSD per node.
  - AMD Genoa or Intel Sapphire Rapids CPU nodes are acceptable, with lower core counts likely more price/performance efficient. They should have a minimum of 1TB memory per node. The greater score given to SPR with HBM should be noted.
  - Nvidia Grace CPU nodes are acceptable with a minimum of 1TB per node, but require the vendor to run and submit our benchmark results.

Interconnect technology:

- One infiniband HDR-200 interface should be included per CPU node
- One HDR-200 interface per two GPUs if GPU nodes are selected.
- Breakout cables to NDR ports can be used if necessary.

Memory and storage

- The global cluster host memory must equal or exceed 20TB. (mandatory)
- It is desirable that the global cluster host memory equal or exceed 40TB. (desirable)
- 1000TB parallel file system disk must be allocatable by USQCD over the machines duration of service, including additional disk in the procurement if required. This should include continuity for the current allocation of 800TB with 450TB used. If the growth can be managed within operations budgets, then the capital spend on computing nodes should be maximised and growth managed as needed.
- An additional 3PB of tape must become available to USQCD over 3 years in addition to the 4.6PB used (long and short term) and 2.4 unused capacity bringing the total to 10PB. Continuity of access to data on long and short term tape for USQCD is required, with users being asked to migrate from short term to long term if required. If the growth can be managed within operations budgets, then the capital spend on computing nodes should be maximised.

Procurement recommendations

- The rule for geometric mean of memory capacity (TB) and scored performance throughput (TF) should be the overall scoring criterion.
- The procurement should request bids for one or more of several options for the entire cluster,
  - Option: Intel Sapphire Rapids nodes with $\geq$ 1TB memory and 1 HDR200 interface, with and without HBM.
  - Option: AMD Genoa nodes with $\geq$ 1TB memory and 1 HDR200 interface.
  - Option: Nvidia H100 GPU nodes with $\geq$ 2TB memory, $\geq$2 HDR200 interfaces and 8TB local SSD.
  - Option: AMD MI300X GPU nodes with $\geq$ 2TB memory and $\geq$2 HDR200 interfaces and 8TB local SSD.
  - Any vendors proposing Nvidia Grace CPU nodes should be equally considered providin benchmarking access is available to assess a score.
- The procurement should not up-front commit to buying any specific one of these technologies. Judgement should be made on the basis of bid scoring.
- Performance scores for each category of node listed in this document can be used by the vendors to avoid redundant benchmarking, save for the MI300X and Nvidia Grace platforms for which we have had no benchmark access and would require access to benchmark.
- Maintenance period should ideally be 5 years on all components.

Feedback welcome !