Ponte Vecchio Across the Atlantic

Single-Node Benchmarking of Two Intel GPU Systems

Thomas Applencourt[1], Aditya Sadawarte[2], Servesh Muralidharan[1], Colleen Bertoni[1], JaeHyuk Kwack[1], Ye Luo[1], Esteban Rangel[1], John Tramm[1], Yasaman Ghadar [1], Arjen Tamerus [3], Chris Edsall [3], Tom Deakin[2]

November 18, 2024

[1]Argonne National Laboratory

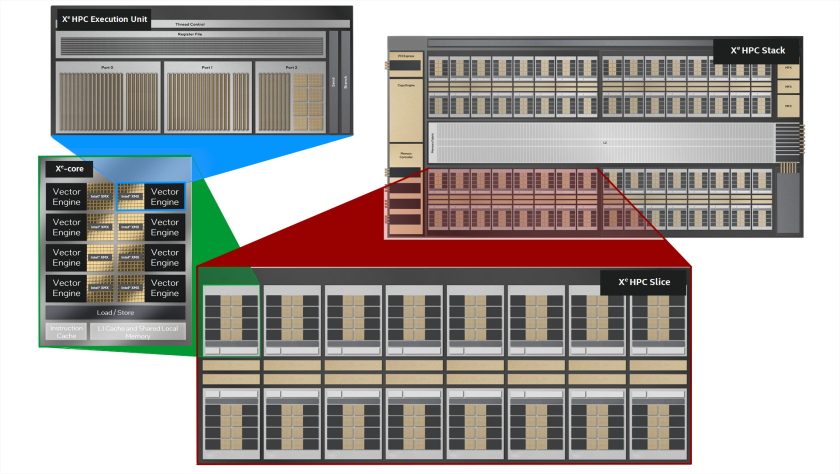
[2]University of Bristol

[3]University of Cambridge

Table of Contents

Hardware

PVC Architecture



Goal (and Non-goal) of This Talk

- Demonstrate PVC performance (and implicitly the quality of the software stack)[1]
- "A little bit of everything" (micro benchmark, comparison to other hardware, analysis of different configuration)
- Not a deep-dive of anything[2]

[1] *The Aurora work was done on a pre-production supercomputer with early versions of the Aurora software development kit.*

[2] Future Papers, collaboration are welcome

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Aurora versus Dawn: Node difference

- Number of GPUs per Node: Aurora 6 PVC (12 Xe-Slice) / Dawn 4 PVC (8 Xe-Slice)
- Number of Xe-Cores per Xe-Slice: Aurora 56 / Dawn 64

Now some pictures![3]

[3] Last picture before tables of numbers! You've been warned ...

Dawn

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Aurora

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmarks

Micro-benchmark Details

	Programming Model	Description
Peak Compute	OpenMP	Chain of FMA to measure FLOPS
Device Memory Bandwidth	OpenMP	Triad used for HBM bandwidth
Host to Device Transfer Bandwidth	SYCL	Compute the Bandwidth of the PCIe datatransfer
Device to Device Transfer Bandwidth	SYCL	Measure the Bandwidth between 2 Ranks (Stacks on the GPU & between GPUs)
General Matrix Multiplication (GEMM)	SYCL	DGEMM, SGEMM, ...
Fast Fourier Transform (FFT)	SYCL	Backward and forward
Lats	SYCL, CUDA, HIP	Measure the access latency of different levels of the memory hierarchy

Micro-benchmark Results: Comparison

	H100*	MI250*	1x GCD MI250x	1x PVC Stack (512EU)
Double Precision Peak Flops	34.0 TFlop/s	45.3 TFlop/s	-	20 TFlop/s
Single Precision Peak Flops	67.0 TFlop/s	45.3 TFlop/s	-	26 TFlop/s
DGEMM	-	-	24.1 TFlop/s	17 TFlop/s
SGEMM	-	-	33.8 TFlop/s	25 TFlop/s
Memory Bandwidth	3.4 GB/s	3.2 TB/s	1.3 TB/s	1.0 TB/s
PCIe Unidirectional Bandwidth	128.0 GB/s	64.0 GB/s	25.0 GB/s	54.0 GB/s
GPU to GPU	-	-	37.0 GB/s	15uni, 23bi GB/s

* Indicates theoretical values

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results

	Aurora (PVC)			Dawn (PVC)		
	One Stack	One PVC	Six PVC	One Stack	One PVC	Four PVC
Double Precision Peak Flops	17 TFlop/s	33 TFLop/s	195 TFlop/s	20 TFlop/s	37 TFlop/s	140 TFlop/s
Single Precision Peak Flops	23 TFlop/s	45 TFlop/s	268 TFlop/s	26 TFlop/s	52 TFlop/s	207 TFlop/s
Memory Bandwidth (triad)	1 TB/s	2 TB/s	12 TB/s	1 TB/s	2 TB/s	8 TB/s
PCIe Unidirectional Bandwidth (H2D)	54 GB/s	55 GB/s	329 GB/s	53 GB/s	54 GB/s	218 GB/s
PCIe Unidirectional Bandwidth (D2H)	53 GB/s	56 GB/s	264 GB/s	51 GB/s	53 GB/s	212 GB/s
PCIe Bidirectional Bandwidth	76 GB/s	77 GB/s	350 GB/s	72 GB/s	72 GB/s	285 GB/s
DGEMM	15 TFlop/s	26 TFlop/s	151 TFlop/s	17 TFlop/s	30 TFlop/s	120 TFlop/s
SGEMM	21 TFlop/s	42 TFlop/s	242 TFlop/s	25 TFlop/s	48 TFlop/s	188 TFlop/s
HGEMM	207 TFlop/s	411 TFlop/s	2.3 PFlop/s	246 TFlop/s	509 TFlop/s	1.9 PFlop/s
BF16GEMM	216 TFlop/s	434 TFlop/s	2.4 PFlop/s	254 TFlop/s	501 TFlop/s	2.0 PFlop/s
TF32GEMM	107 TFlop/s	208 TFlop/s	1.2 PFlop/s	118 TFlop/s	200 TFlop/s	850 TFlop/s
I8GEMM	448 TIop/s	864 TIop/s	5.0 PIop/s	525 TIop/s	1.1 PIop/s	4.1 PIop/s
Single-precision FFT C2C 1D	3.1 TFlop/s	5.9 TFlop/s	33 Tflop/s	3.6 TFlop/s	6.6 TFlop/s	26 TFlop/s
Single-precision FFT C2C 2D	3.4 TFlop/s	6.0 TFlop/s	34 Tflop/s	3.6 TFlop/s	6.5 TFlop/s	25 TFlop/s

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results: Floating-point Ratio

	Aurora (PVC)			Dawn (PVC)		
	One Stack	One PVC	Six PVC	One Stack	One PVC	Four PVC
Double Precision Peak Flops	17 TFlop/s	33 TFLop/s	195 TFlop/s	20 TFlop/s	37 TFlop/s	140 TFlop/s
Single Precision Peak Flops	23 TFlop/s	45 TFlop/s	268 TFlop/s	26 TFlop/s	52 TFlop/s	207 TFlop/s
Memory Bandwidth (triad)	1 TB/s	2 TB/s	12 TB/s	1 TB/s	2 TB/s	8 TB/s
PCIe Unidirectional Bandwidth (H2D)	54 GB/s	55 GB/s	329 GB/s	53 GB/s	54 GB/s	218 GB/s
PCIe Unidirectional Bandwidth (D2H)	53 GB/s	56 GB/s	264 GB/s	51 GB/s	53 GB/s	212 GB/s
PCIe Bidirectional Bandwidth	76 GB/s	77 GB/s	350 GB/s	72 GB/s	72 GB/s	285 GB/s
DGEMM	15 TFlop/s	26 TFlop/s	151 TFlop/s	17 TFlop/s	30 TFlop/s	120 TFlop/s
SGEMM	21 TFlop/s	42 TFlop/s	242 TFlop/s	25 TFlop/s	48 TFlop/s	188 TFlop/s
HGEMM	207 TFlop/s	411 TFlop/s	2.3 PFlop/s	246 TFlop/s	509 TFlop/s	1.9 PFlop/s
BF16GEMM	216 TFlop/s	434 TFlop/s	2.4 PFlop/s	254 TFlop/s	501 TFlop/s	2.0 PFlop/s
TF32GEMM	107 TFlop/s	208 TFlop/s	1.2 PFlop/s	118 TFlop/s	200 TFlop/s	850 TFlop/s
I8GEMM	448 TIop/s	864 TIop/s	5.0 PIop/s	525 TIop/s	1.1 PIop/s	4.1 PIop/s
Single-precision FFT C2C 1D	3.1 TFlop/s	5.9 TFlop/s	33 Tflop/s	3.6 TFlop/s	6.6 TFlop/s	26 TFlop/s
Single-precision FFT C2C 2D	3.4 TFlop/s	6.0 TFlop/s	34 Tflop/s	3.6 TFlop/s	6.5 TFlop/s	25 TFlop/s

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results: Floating-point Ratio

- FP64: 8 (EU per Xe-Core) * 8 (SIMD width) * 2 (FMA) *2 (double pipeline) * 56 Xe-core per stack * 1.4 Ghz = 20 Tflops per stack (17 measured)
- FP32: 8 (EU per Xe-Core) * 16 (SIMD width) * 2 (FMA)* 56 Xe-core per stack * 1.6 Ghz = 23 Tflops per stack (23 measured)

Key points:

- Frequencies are different between FP64 and FP32
- High percentage of theoretical peak achieved

Micro-benchmark Results: GEMM Peak

	Aurora (PVC)			Dawn (PVC)		
	One Stack	One PVC	Six PVC	One Stack	One PVC	Four PVC
Double Precision Peak Flops	17 TFlop/s	33 TFLop/s	195 TFlop/s	20 TFlop/s	37 TFlop/s	140 TFlop/s
Single Precision Peak Flops	23 TFlop/s	45 TFlop/s	268 TFlop/s	26 TFlop/s	52 TFlop/s	207 TFlop/s
Memory Bandwidth (triad)	1 TB/s	2 TB/s	12 TB/s	1 TB/s	2 TB/s	8 TB/s
PCIe Unidirectional Bandwidth (H2D)	54 GB/s	55 GB/s	329 GB/s	53 GB/s	54 GB/s	218 GB/s
PCIe Unidirectional Bandwidth (D2H)	53 GB/s	56 GB/s	264 GB/s	51 GB/s	53 GB/s	212 GB/s
PCIe Bidirectional Bandwidth	76 GB/s	77 GB/s	350 GB/s	72 GB/s	72 GB/s	285 GB/s
DGEMM	15 TFlop/s	26 TFlop/s	151 TFlop/s	17 TFlop/s	30 TFlop/s	120 TFlop/s
SGEMM	21 TFlop/s	42 TFlop/s	242 TFlop/s	25 TFlop/s	48 TFlop/s	188 TFlop/s
HGEMM	207 TFlop/s	411 TFlop/s	2.3 PFlop/s	246 TFlop/s	509 TFlop/s	1.9 PFlop/s
BF16GEMM	216 TFlop/s	434 TFlop/s	2.4 PFlop/s	254 TFlop/s	501 TFlop/s	2.0 PFlop/s
TF32GEMM	107 TFlop/s	208 TFlop/s	1.2 PFlop/s	118 TFlop/s	200 TFlop/s	850 TFlop/s
I8GEMM	448 TIop/s	864 TIop/s	5.0 PIop/s	525 TIop/s	1.1 PIop/s	4.1 PIop/s
Single-precision FFT C2C 1D	3.1 TFlop/s	5.9 TFlop/s	33 Tflop/s	3.6 TFlop/s	6.6 TFlop/s	26 TFlop/s
Single-precision FFT C2C 2D	3.4 TFlop/s	6.0 TFlop/s	34 Tflop/s	3.6 TFlop/s	6.5 TFlop/s	25 TFlop/s

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results: GEMM

- $15/17 \approx 88\%$
- $21/23 \approx 90\%$

Note: Aurora DGEMM Number are not the number of the paper. We where able to squeeze more performance since the paper was finalized.

Micro-benchmark Results: Stream Peak

	Aurora (PVC)			Dawn (PVC)		
	One Stack	One PVC	Six PVC	One Stack	One PVC	Four PVC
Double Precision Peak Flops	17 TFlop/s	33 TFLop/s	195 TFlop/s	20 TFlop/s	37 TFlop/s	140 TFlop/s
Single Precision Peak Flops	23 TFlop/s	45 TFlop/s	268 TFlop/s	26 TFlop/s	52 TFlop/s	207 TFlop/s
Memory Bandwidth (triad)	1 TB/s	2 TB/s	12 TB/s	1 TB/s	2 TB/s	8 TB/s
PCIe Unidirectional Bandwidth (H2D)	54 GB/s	55 GB/s	329 GB/s	53 GB/s	54 GB/s	218 GB/s
PCIe Unidirectional Bandwidth (D2H)	53 GB/s	56 GB/s	264 GB/s	51 GB/s	53 GB/s	212 GB/s
PCIe Bidirectional Bandwidth	76 GB/s	77 GB/s	350 GB/s	72 GB/s	72 GB/s	285 GB/s
DGEMM	15 TFlop/s	26 TFlop/s	151 TFlop/s	17 TFlop/s	30 TFlop/s	120 TFlop/s
SGEMM	21 TFlop/s	42 TFlop/s	242 TFlop/s	25 TFlop/s	48 TFlop/s	188 TFlop/s
HGEMM	207 TFlop/s	411 TFlop/s	2.3 PFlop/s	246 TFlop/s	509 TFlop/s	1.9 PFlop/s
BF16GEMM	216 TFlop/s	434 TFlop/s	2.4 PFlop/s	254 TFlop/s	501 TFlop/s	2.0 PFlop/s
TF32GEMM	107 TFlop/s	208 TFlop/s	1.2 PFlop/s	118 TFlop/s	200 TFlop/s	850 TFlop/s
I8GEMM	448 TIop/s	864 TIop/s	5.0 PIop/s	525 TIop/s	1.1 PIop/s	4.1 PIop/s
Single-precision FFT C2C 1D	3.1 TFlop/s	5.9 TFlop/s	33 Tflop/s	3.6 TFlop/s	6.6 TFlop/s	26 TFlop/s
Single-precision FFT C2C 2D	3.4 TFlop/s	6.0 TFlop/s	34 Tflop/s	3.6 TFlop/s	6.5 TFlop/s	25 TFlop/s

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results: PCIe Concurrency

	Aurora (PVC)			Dawn (PVC)		
	One Stack	One PVC	Six PVC	One Stack	One PVC	Four PVC
Double Precision Peak Flops	17 TFlop/s	33 TFLop/s	195 TFlop/s	20 TFlop/s	37 TFlop/s	140 TFlop/s
Single Precision Peak Flops	23 TFlop/s	45 TFlop/s	268 TFlop/s	26 TFlop/s	52 TFlop/s	207 TFlop/s
Memory Bandwidth (triad)	1 TB/s	2 TB/s	12 TB/s	1 TB/s	2 TB/s	8 TB/s
PCIe Unidirectional Bandwidth (H2D)	54 GB/s	55 GB/s	329 GB/s	53 GB/s	54 GB/s	218 GB/s
PCIe Unidirectional Bandwidth (D2H)	53 GB/s	56 GB/s	264 GB/s	51 GB/s	53 GB/s	212 GB/s
PCIe Bidirectional Bandwidth	76 GB/s	77 GB/s	350 GB/s	72 GB/s	72 GB/s	285 GB/s
DGEMM	15 TFlop/s	26 TFlop/s	151 TFlop/s	17 TFlop/s	30 TFlop/s	120 TFlop/s
SGEMM	21 TFlop/s	42 TFlop/s	242 TFlop/s	25 TFlop/s	48 TFlop/s	188 TFlop/s
HGEMM	207 TFlop/s	411 TFlop/s	2.3 PFlop/s	246 TFlop/s	509 TFlop/s	1.9 PFlop/s
BF16GEMM	216 TFlop/s	434 TFlop/s	2.4 PFlop/s	254 TFlop/s	501 TFlop/s	2.0 PFlop/s
TF32GEMM	107 TFlop/s	208 TFlop/s	1.2 PFlop/s	118 TFlop/s	200 TFlop/s	850 TFlop/s
I8GEMM	448 TIop/s	864 TIop/s	5.0 PIop/s	525 TIop/s	1.1 PIop/s	4.1 PIop/s
Single-precision FFT C2C 1D	3.1 TFlop/s	5.9 TFlop/s	33 Tflop/s	3.6 TFlop/s	6.6 TFlop/s	26 TFlop/s
Single-precision FFT C2C 2D	3.4 TFlop/s	6.0 TFlop/s	34 Tflop/s	3.6 TFlop/s	6.5 TFlop/s	25 TFlop/s

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Micro-benchmark Results

	Aurora (PVC)		Dawn (PVC)	
	One Stack-Pair	Six Stack-Pairs	One Stack-Pair	Four Stack-Pairs
Local Stack Unidirectional Bandwidth	197 GB/s	1129 GB/s	196 GB/s	786 GB/s
Local Stack Bidirectional Bandwidth	284 GB/s	1661 GB/s	287 GB/s	1145 GB/s
Remote Stack Unidirectional Bandwidth	15 GB/s	95 GB/s	-	-
Remote Stack Bidirectional Bandwidth	23 GB/s	142 GB/s	-	-

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Mini-apps

Mini-apps: Summary

	Science Domain	Language	Programming model	Characteristic	Scaling	Figure-of-Merit
miniBUDE	BioChemistry	C++	SYCL, HIP, CUDA	FP32 flop-rate bound	N/A	$\frac{Billion\ Interactions}{time(s)}$
CloverLeaf	Computational Fluid Dynamics	C++	SYCL, HIP, CUDA	Memory bandwidth bound	Weak	$\frac{N_{cells}}{time(s)}$
miniQMC	Material Science	C++	OpenMP	Compute/Memory BW bound CPU congestion bound	Weak	$\frac{N_w N_e 10^{-11}}{diffusion\ time(s)}$
GAMESS RI-MP2 mini-app	Quantum Chemistry	Fortran	OpenMP	DGEMM bound	Strong	$\frac{1}{time(h)}$

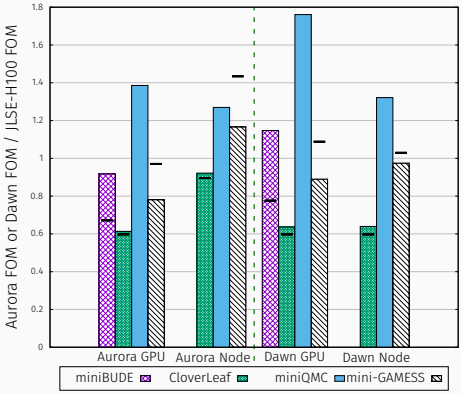
Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Mini-apps: Result

	Aurora (PVC)			Dawn (PVC)			JLSE (H100)		JLSE (MI250)	
	One Stack	One GPU	Six GPU	One Stack	One GPU	Four GPU	One GPU	Four GPU	One GCD	Four GPU
miniBUDE	293.02	-	-	366.17	-	-	638.40	-	193.66	-
CloverLeaf	20.82	40.41	240.89	22.46	41.92	167.15	65.87	261.37	25.71	192.68
miniQMC	3.16	5.39	15.64	3.72	6.85	16.28	3.89	12.32	0.50	0.90
mini-GAMESS	19.44	38.50	197.08	24.57	43.88	164.71	49.30	168.97	-	-

Figure of Merits on Aurora (6) and Dawn (4) Relative to JLSE-H100 (4)

The black bars are the expected relative performance based on known bounds.



Apps

Preliminary Conclusion

- They run! [4]
- Performance are good

[4] Impressive as first generation hardware and full new software stacks (compiler, UMD, KMD, ...)

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Apps: Summary

	Science Domain	Language	Programming model	Characteristic	Scaling	Figure-of-Merit
OpenMC	Particle Transport	C++	OpenMP	Memory latency/bandwidth bound	Weak	$\frac{\textit{Thousand particles}}{\textit{time(s)}}$
HACC	Cosmology	C++	SYCL, HIP, CUDA	CPU memory BW bound, GPU FP32 flop-rate bound	Weak	$\frac{N_p N_{steps}}{\textit{time(s)}}$

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Apps: Result

	Aurora (PVC)			Dawn (PVC)			JLSE (H100)		JLSE (MI250)	
	One Stack	One GPU	Six GPU	One Stack	One GPU	Four GPU	One GPU	Four GPU	One GCD	Four GPU
OpenMC	-	-	2039	-	-	-	-	1191	-	720
HACC	-	-	13.81	-	-	12.26	-	12.46	-	10.70

HACC

To compare across the different architectures, we calculate the scaled performance[5] of HACC on the GPUs.

- Relative to JLSE-H100, the Dawn and Aurora single-GPU performance is 0.954 and 0.947
- Relative to JLSE-MI250, the performance is 0.987 and 0.981

Long story short: HACC performs as well on PVC as on other architectures (a little better raw performance thanks to a better CPU / GPU ratio on Dawn / Aurora versus JLSE H100/MI250).

[5] The scaled performance is the single precision theoretical peak multiplied by the aggregate GPU time.

OpenMC

The Aurora 6× PVC node design results in 1.7× the performance of the JLSE 4× H100 node.

Excellent performance on PVC! (More information on the OpenMC port in "Performance Portable Monte Carlo Particle Transport on Intel, NVIDIA, and AMD GPUs"[6])

[6] https://doi.org/10.1051/epjconf/202430204010

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE

Conclusion

Take-aways

- Development of micro-benchmarks
  - Show pros and cons of PVC Hardware (”low” memory BW, good scalability, concurrent PCI transfers, High peak GEMM performances)
- PVC performance is on-par with other recent HPC hardware (compared using micro-benchmarks and characterized mini-apps)
  - Absolute: We show that the figure-of-merit of the mini-apps on a single PVC ranges from 0.6–1.8X the performance of an H100, and 0.8–7.5X of a MI250.
  - Relative: Matches expectations (proving good software capabilities!)
- Competitive performance of two science applications (comparing PVC based node to H100, MI250 node)

Argonne NATIONAL LABORATORY University of BRISTOL UNIVERSITY OF CAMBRIDGE