QC/MD GPU-Accelerated Applications

2023/08/24

GPU Spec

	FP64	FP32	Memory	Memory BW
H100 SXM	34 TFLOPS	67 TFLOPS	80GB HBM3	3.35 TB/s
L40		90.5 TFLOPS	48GB GDDR6	864 GB/s
L4		30.3 TFLOPS	24GB GDDR6	300 GB/s
A100 SXM	9.7 TFLOPS	19.5 TFLOPS	80GB HBM2e	2039 GB/s
A100 PCIe	9.7 TFLOPS	19.5 TFLOPS	80GB HBM2e	1935 GB/s
A30	5.2 TFLOPS	10.3 TFLOPS	24GB HBM2	933 GB/s
A40		37.4 TFLOPS	48GB GDDR6	696 GB/s
V100 SXM	7.8 TFLOPS	15.7 TFLOPS	32GB HBM2	900 GB/s
V100S PCIe	8.2 TFLOPS	16.4 TFLOPS	32GB HBM2	1134 GB/s
T4		8.1 TFLOPS	16GB GDDR6	300 GB/s

Benchmark Dataset

Molecular Dynamics

- “JAC” (DHFR; dihydrofolate reductase)
  - 23,588 atoms
- FactorIX
  - 90,906 atoms
- Cellulose
  - 408,609 atoms
- STMV (satellite tobacco mosaic virus)
  - 1,067,095 atoms
- ADH Dodec
  - 95,561 atoms
- ApoA1 (apolipoprotein A1)
  - 92,224 atoms

GPU Benchmarks

NVIDIA HPC Application Performance

NVIDIA HPC Application Performance

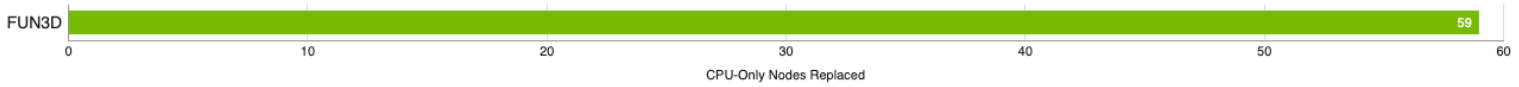
For Deep Learning performance, please go here.

Modern HPC data centers are key to solving some of the world's most important scientific and engineering challenges. The NVIDIA Data Center GPUs fundamentally change the economics of the data center, delivering breakthrough performance with dramatically fewer servers, less power consumption, and reduced networking overhead, resulting in total cost savings of 5X-10X.

The number of CPU-only servers replaced by a single GPU-accelerated server is called the node replacement factor (NRF). To arrive at NRF, we measure application performance with up to 8 CPU-only servers. Then we use linear scaling to scale beyond 8 servers to calculate the NRF. The NRF will vary by application.

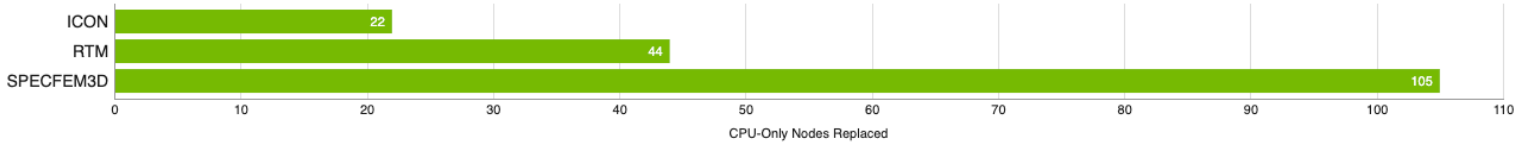
H100 | L40 | L4 | A100 | A30 | A40 | V100 | T4

Engineering



CPU Server: Dual Xeon Gold 6240@2.60GHz | GPU Server: Dual Intel SPR 8480C@2GHz with 4x NVIDIA H100 SXM 80GB | FUN3D Benchmark: dpw_wbt0_crs-3.6Mn_5, CUDA Version: 11.8

Geoscience



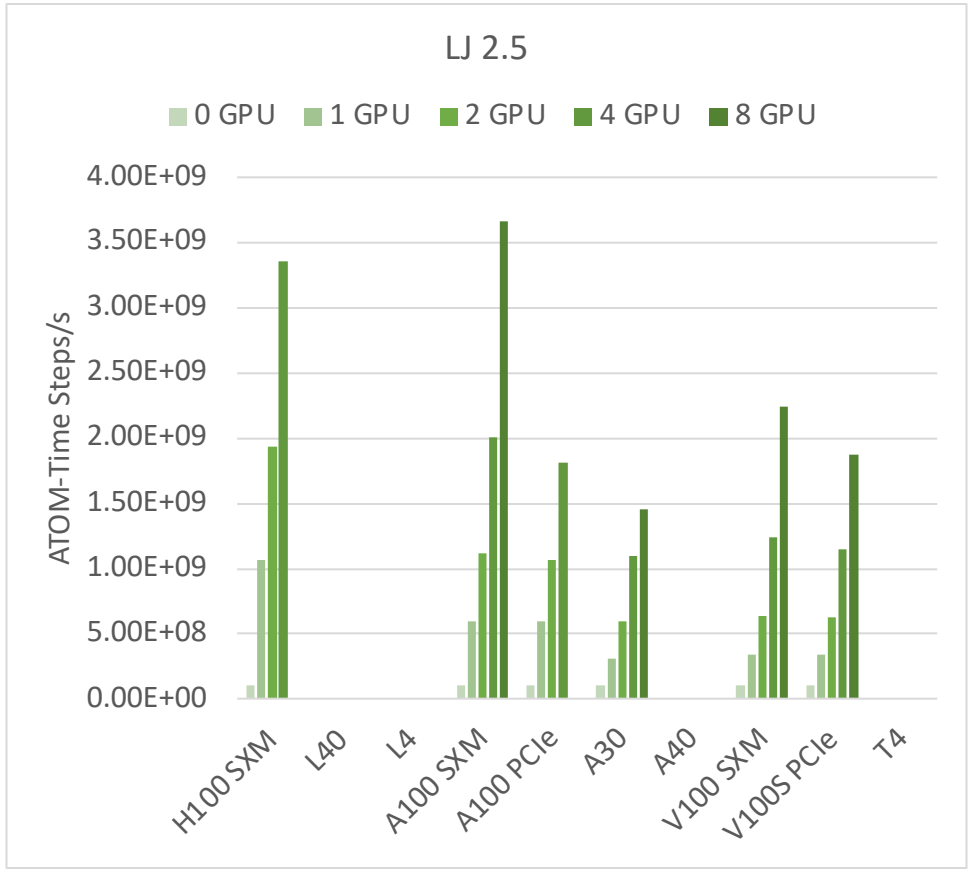
https://developer.nvidia.com/hpc-application-performance

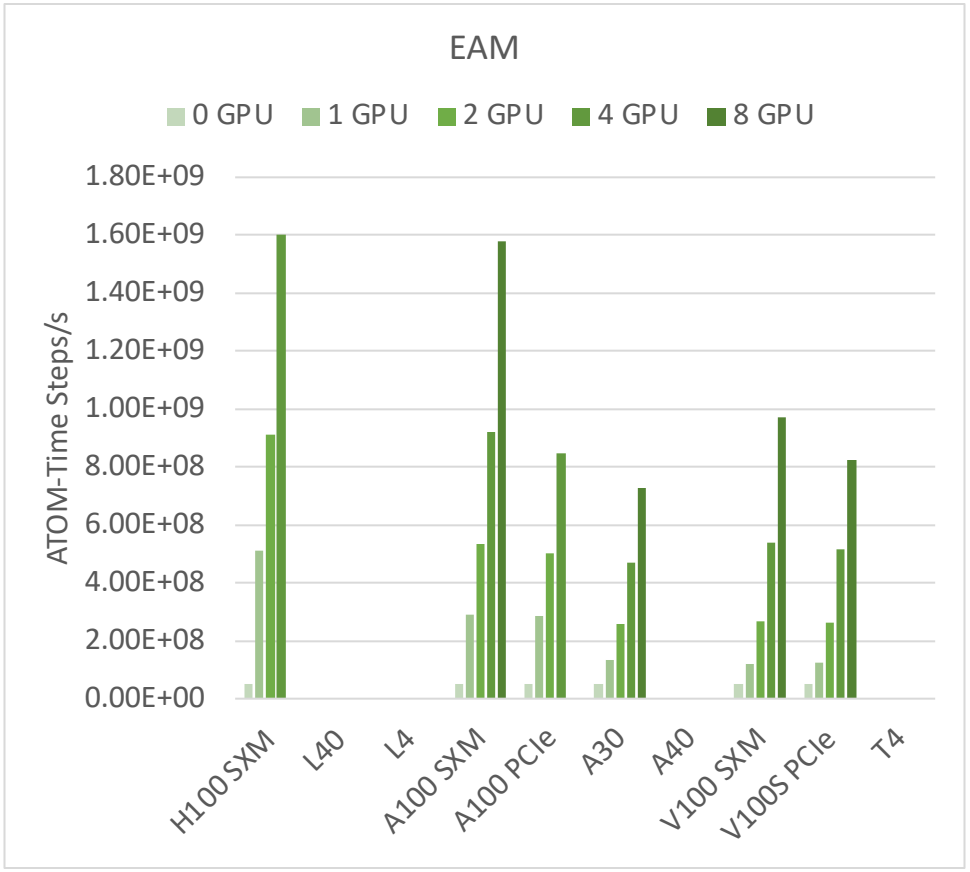
LAMMPS

- Accelerated Features
  - Lennard-Jones, Gay-Berne, Tersoff, many more potentials
- Scalability
  - Multi-GPU, Multi-Node
- Version
  - stable_23Jun2022_update1
- CPU (0 GPU case)
  - Dual Cascade Lake 6240

https://developer.nvidia.com/hpc-application-performance

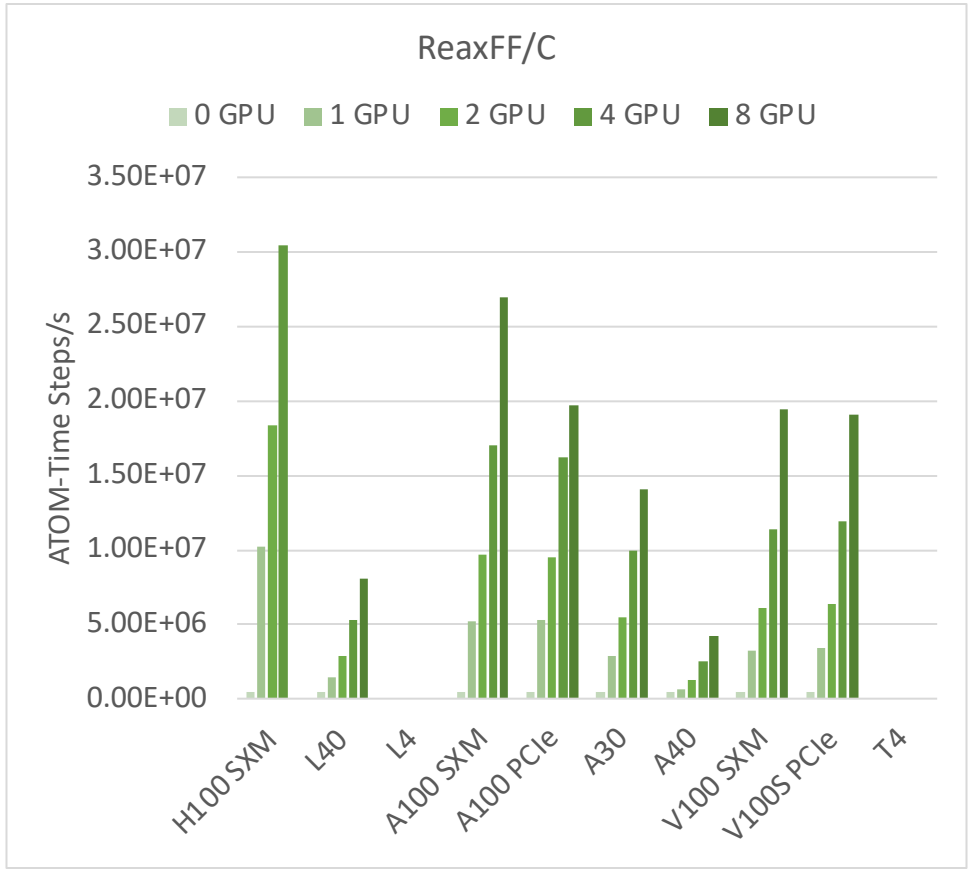
LAMMPS

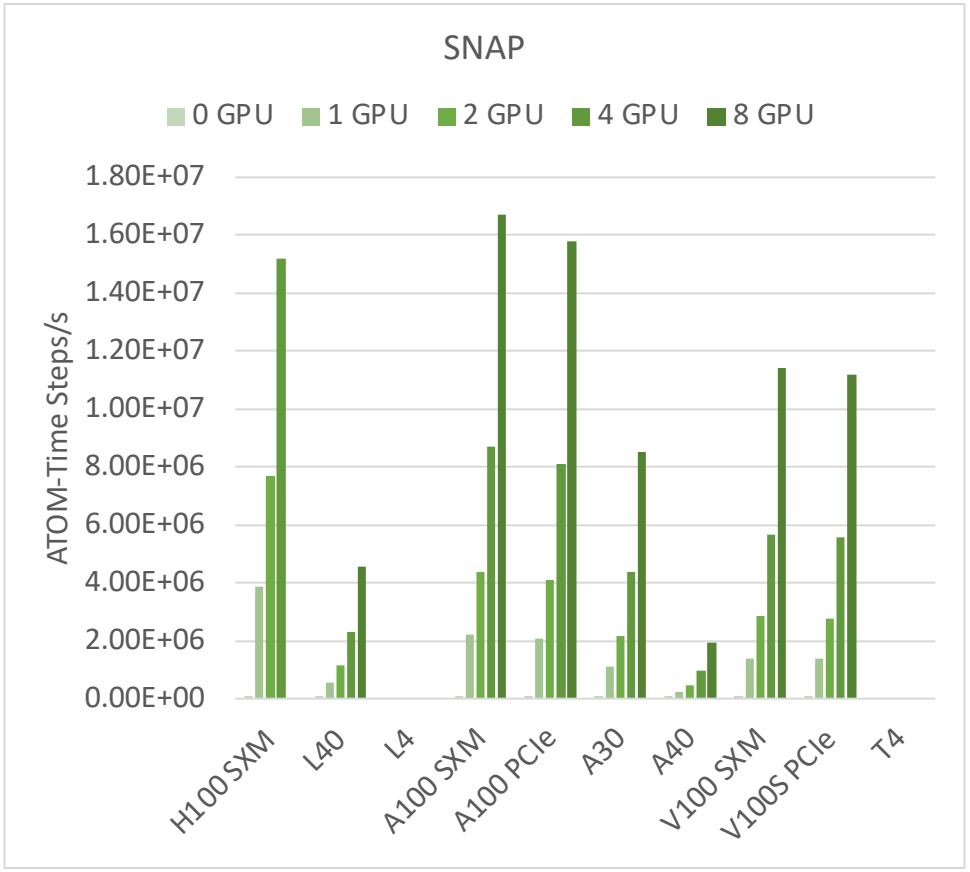




https://developer.nvidia.com/hpc-application-performance

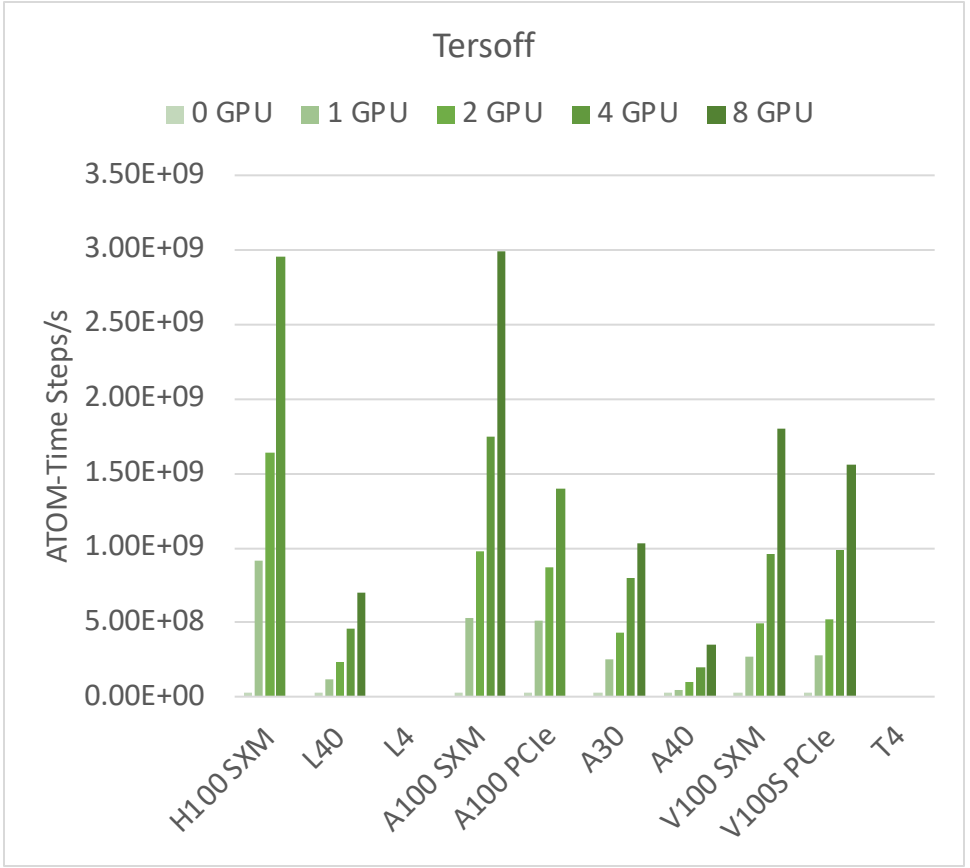
LAMMPS





https://developer.nvidia.com/hpc-application-performance

LAMMPS



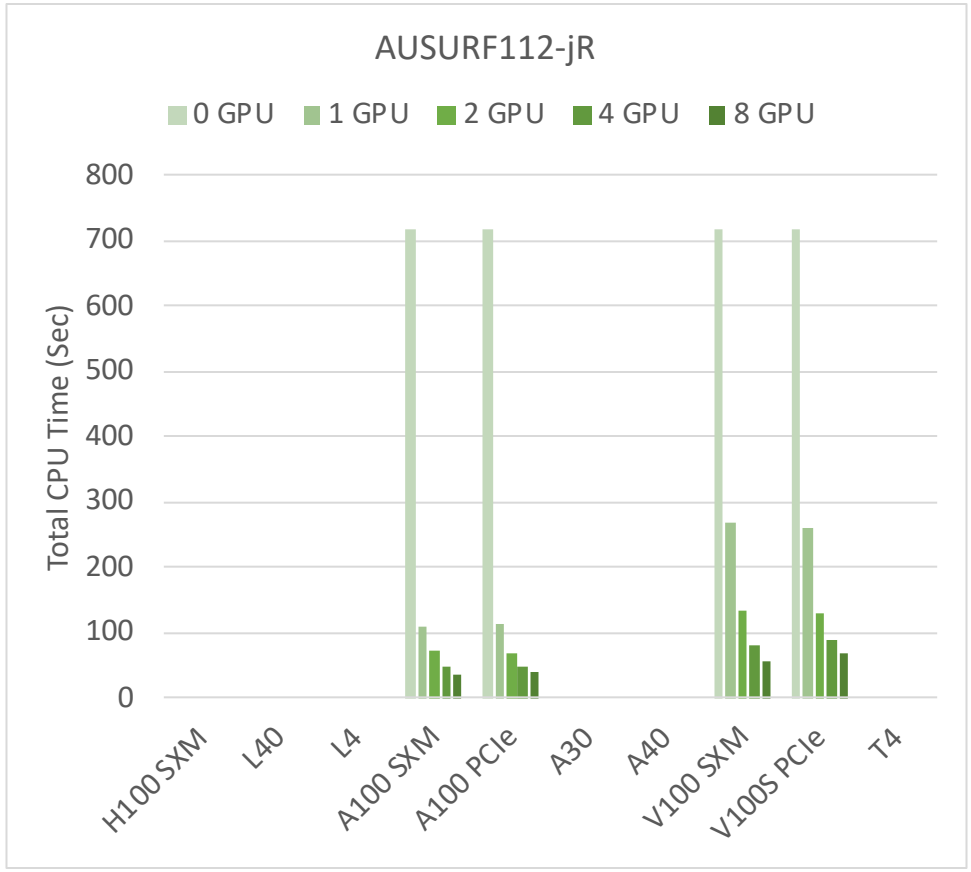
https://developer.nvidia.com/hpc-application-performance

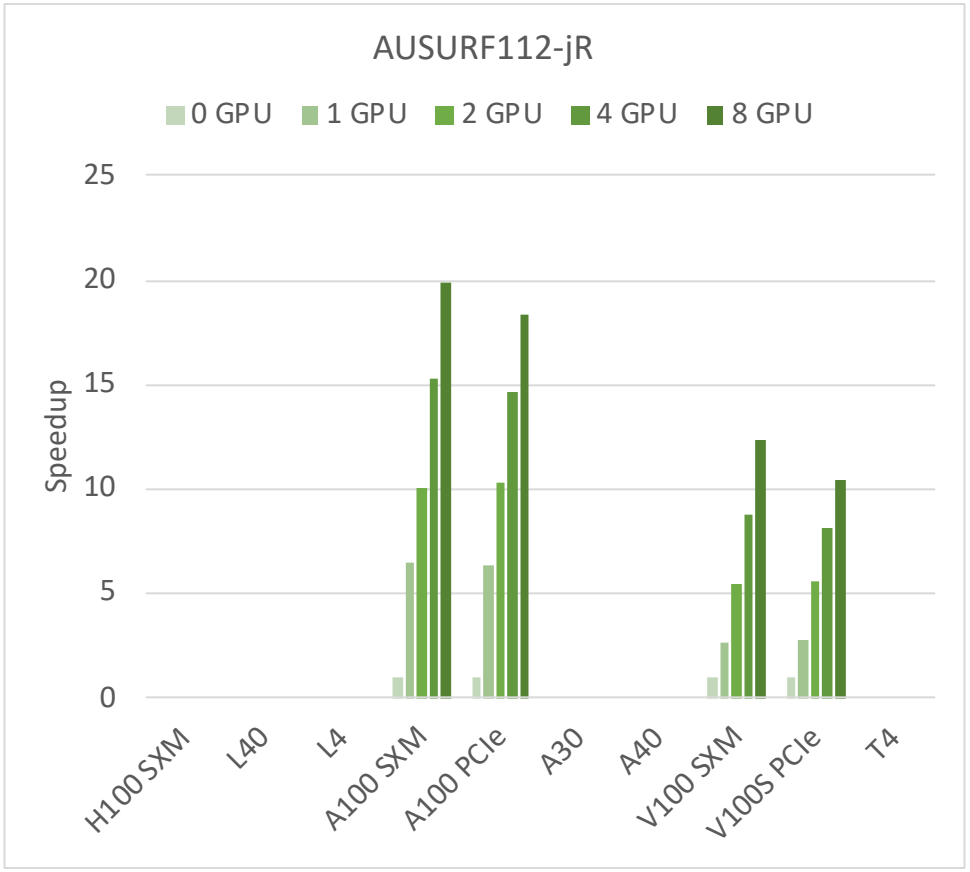
Quantum Espresso

- Accelerated Features
  - linear algebra (matrix multiply)
  - explicit computational kernels
  - 3D FFTs
- Scalability
  - Multi-GPU, Multi-Node
- Version
  - V7.0 CPU; V7.1 GPU
- CPU (0 GPU case)
  - Dual Cascade Lake 6240

https://developer.nvidia.com/hpc-application-performance

Quantum Espresso





https://developer.nvidia.com/hpc-application-performance

GPU Benchmark – VASP

Features Available and Accelerated in VASP 6.3

LEVELS OF THEORY

Standard DFT **(incl. meta-GGA, vdW-DFT)**
Hybrid DFT **(double buffered)**
Cubic-scaling RPA (ACFDT, GW)
Bethe-Salpeter Equations (BSE)

SOLVERS / MAIN ALGORITHM

Davidson **(+Adaptively Compressed Exch.)**
RMM-DIIS
Davidson+RMM-DIIS
Direct optimizers (Damped, All)
Linear response

PROJECTION SCHEME

Real space
Reciprocal space

EXECUTABLE FLAVORS

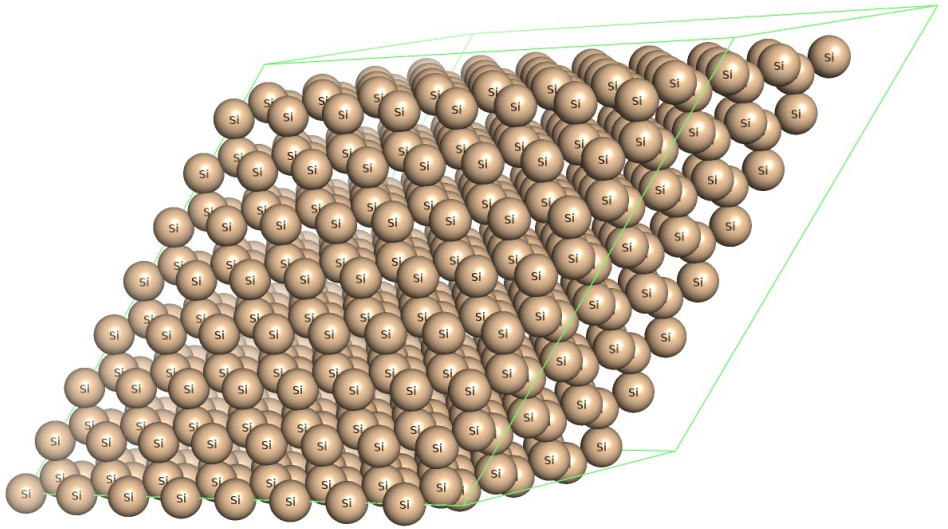
Standard variant
Gamma-point simplification variant
Non-collinear spin variant

- Existing acceleration
- **New acceleration**
- Acceleration work in progress
- On acceleration roadmap

Details on Dataset

Si-Huge

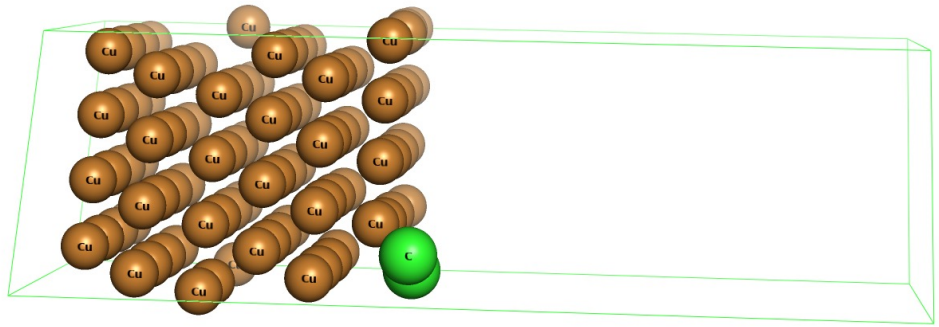
- Cell size: 15.4x30.7x30.7 Å^3
- Atoms: 512 Si
- 14 k-points, 1281 bands, 245.4 eV cutoff Energy, 89 614 PWs
- Standard DFT (GGA: PW91)
- Algo=Normal (Davidson)
- Real-space projection scheme



Details on Dataset

CuC_vdW

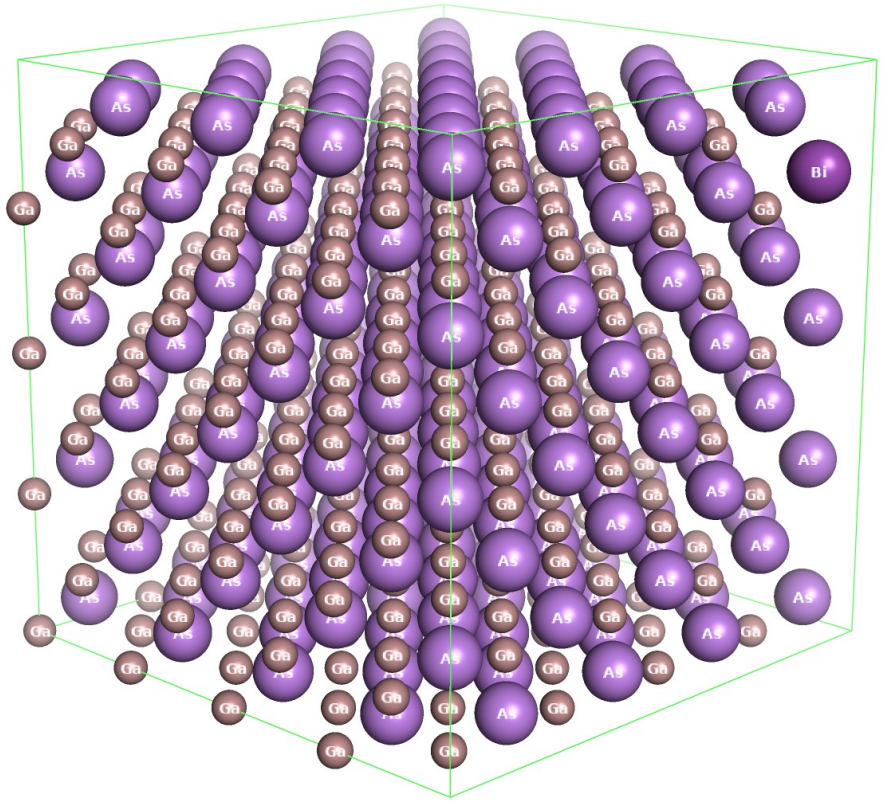
- Cell size: 10.3x10.3x31.5 Å^3
- Atoms: 96 Cu, 2 C (98 total)
- 5 k-points, 638 bands, 400 eV cutoff Energy, 52 405 PWs
- Standard DFT (GGA: PBE)
- Algo=VeryFast (RMM-DIIS)
- Real-space projection scheme



Details on Dataset

GaAsBi_512

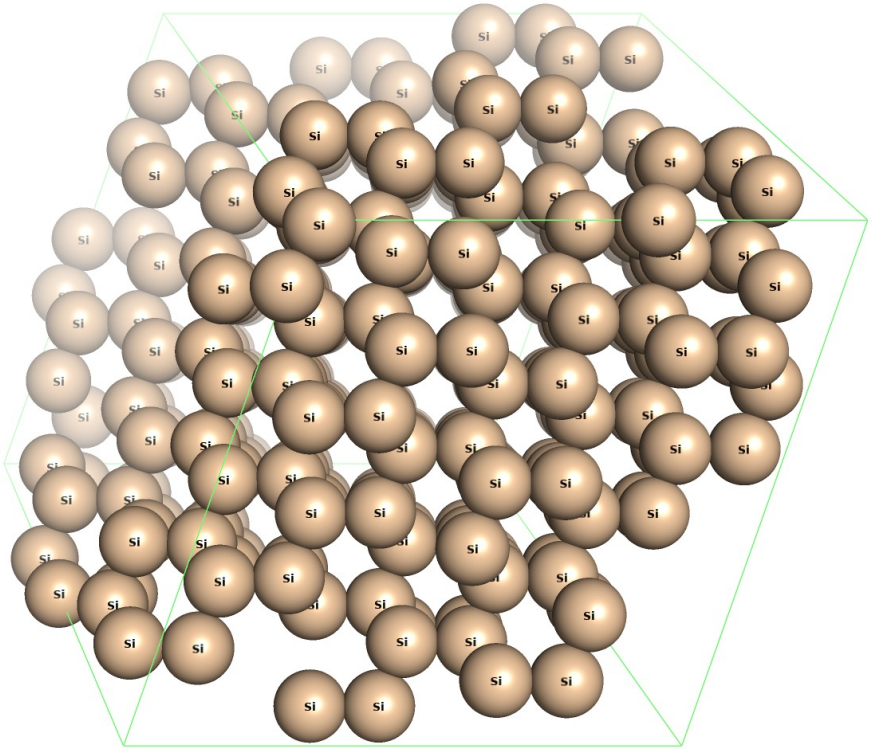
- Cell size: 22.6x22.6x22.6 Å^3
- Atoms: 256 Ga, 255 As, 1 Bi (512 total)
- 4 k-points, 1536 bands, 313 eV cutoff Energy, 145 484 PWs
- Standard DFT (GGA: PBE)
- Algo=Fast (Davidson + RMM-DIIS)
- Real-space projection scheme



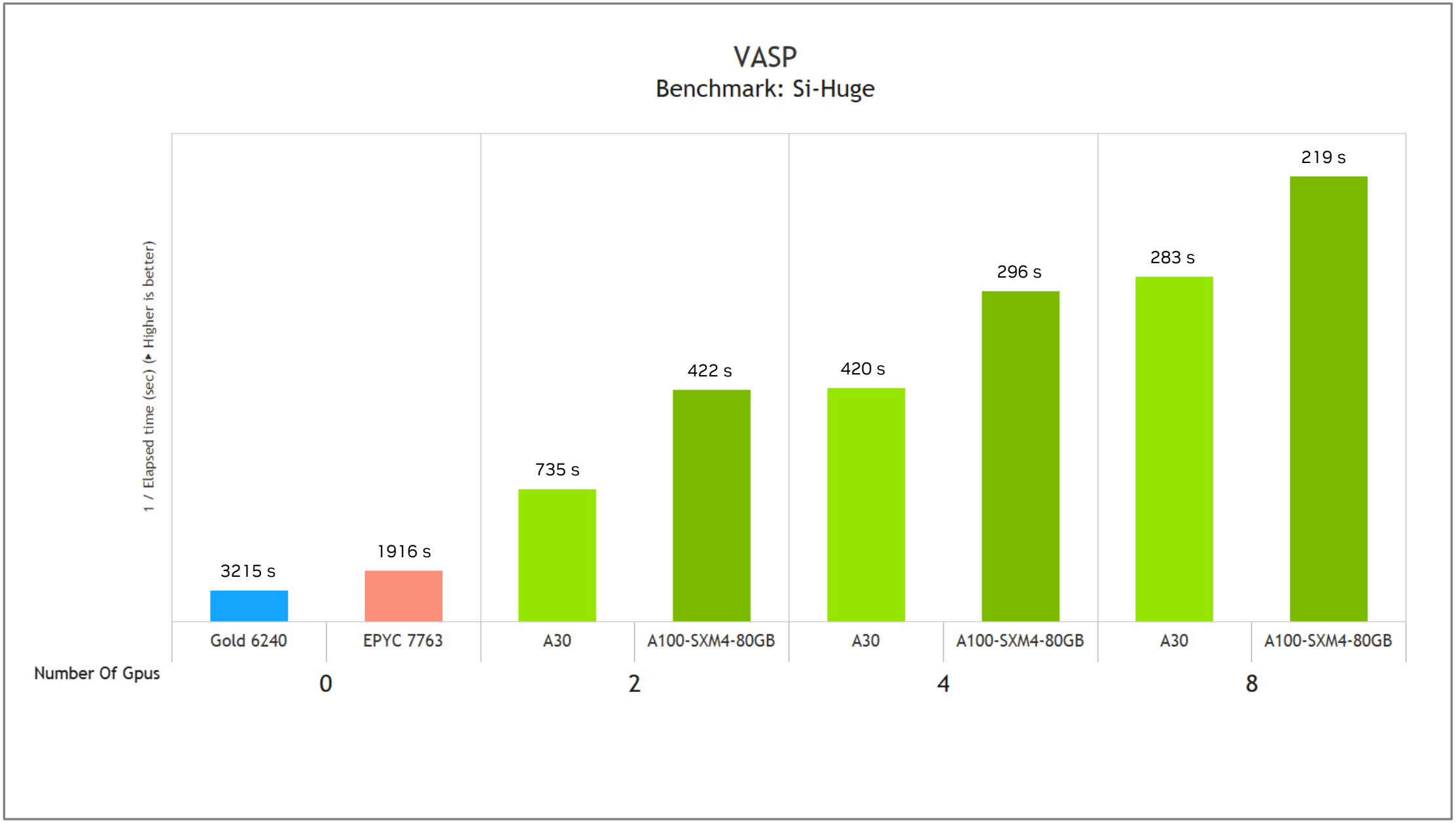
Details on Dataset

Si256_VJT_HSE06

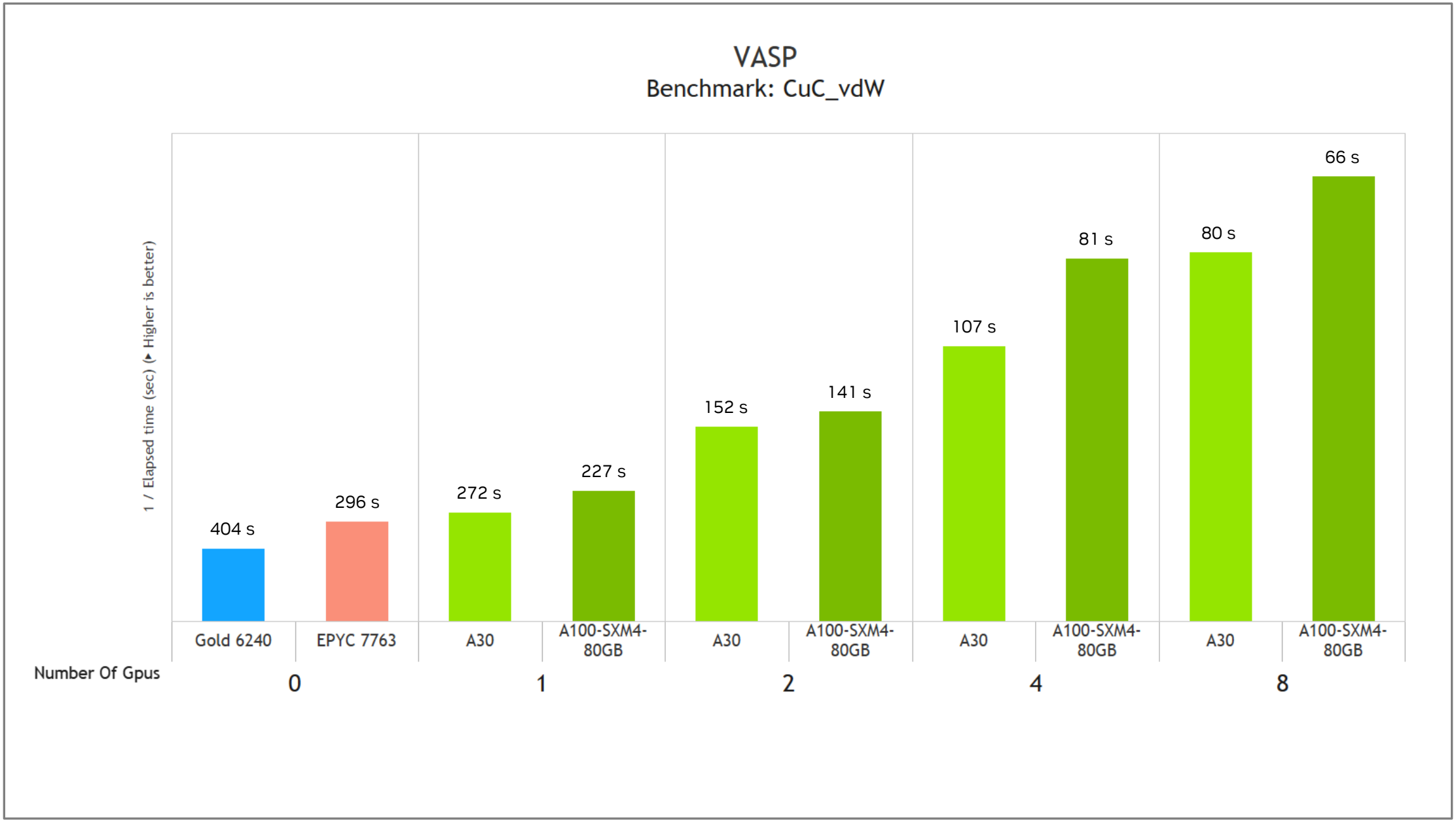
- Cell size: 18.9x18.9x18.9 Å^3
- Atoms: 256 Si
- 1 k-point (Γ), 640 bands, 250 eV cutoff Energy, 23 589 PWs
- Hybrid DFT (PBE0)
- Algo=Damped (direct minimizer)
- Real-space projection scheme



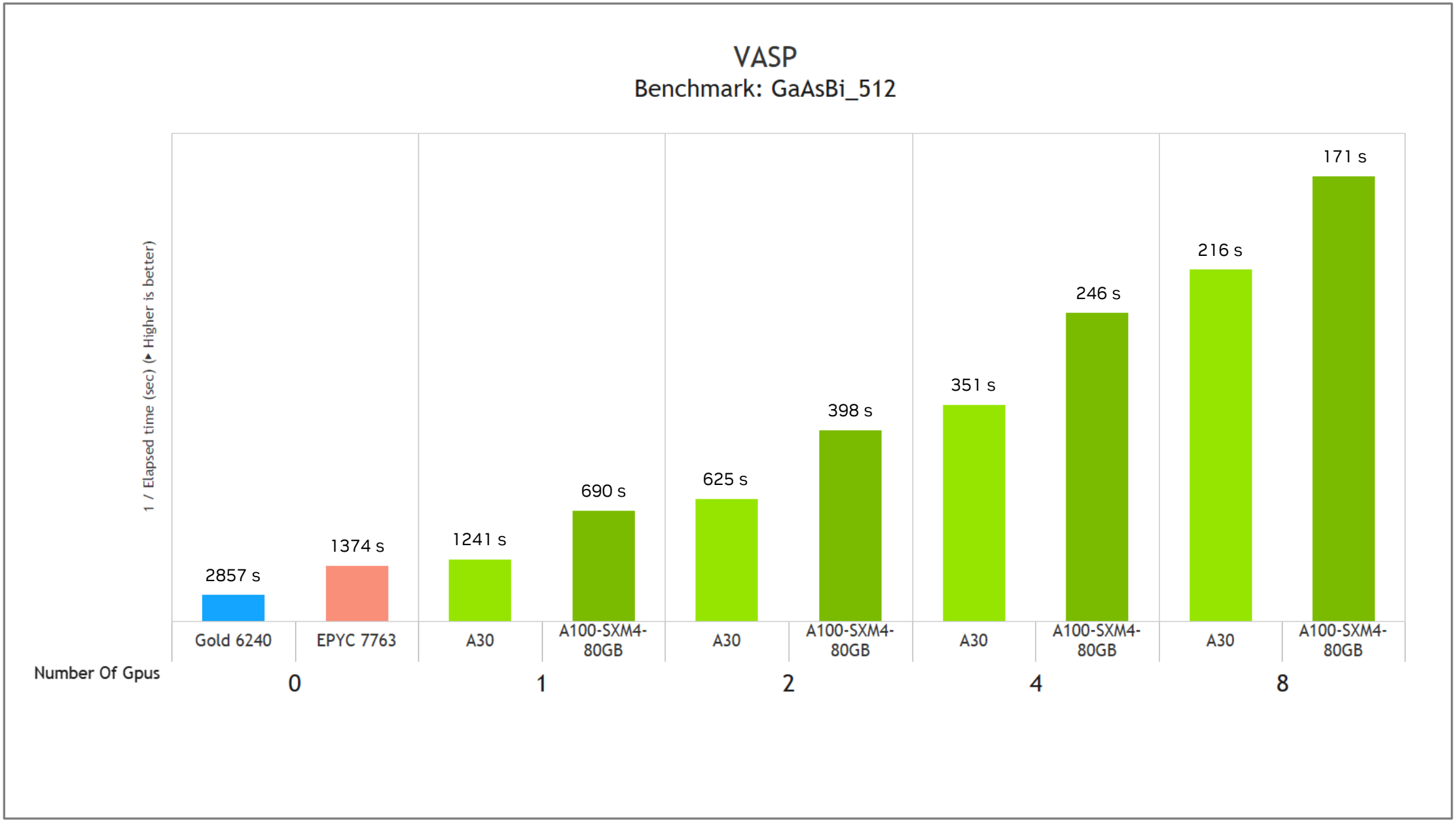
VASP 6.3.0


VASP
Benchmark: Si-Huge
1 / Elapsed time (sec) (▸ Higher is better)
3215 s
1916 s
735 s
422 s
420 s
296 s
283 s
219 s
Gold 6240
EPYC 7763
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
Number Of Gpus
0
2
4
8

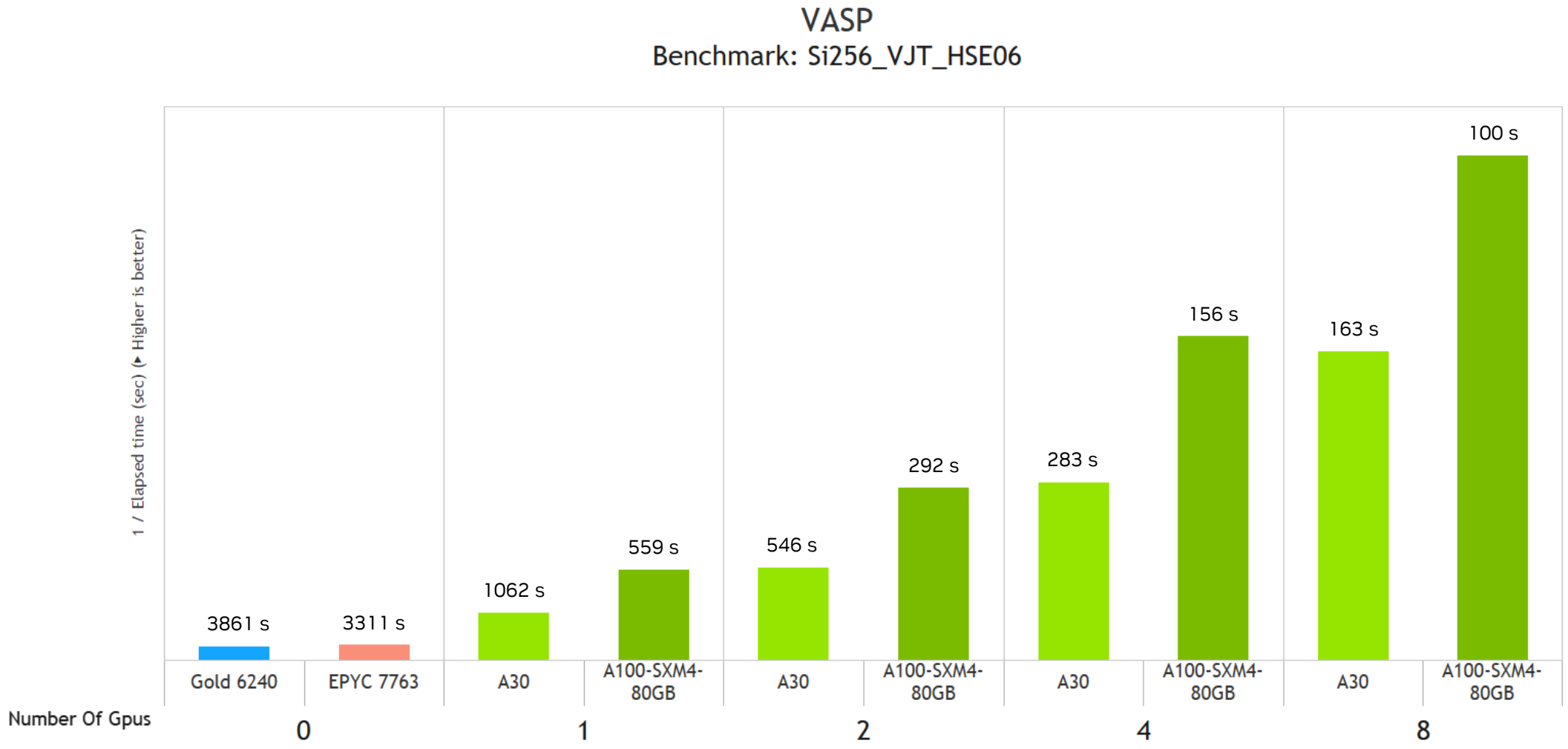
VASP 6.3.0


VASP
Benchmark: CuC_vdW
1 / Elapsed time (sec) (▸ Higher is better)
404 s
296 s
272 s
227 s
152 s
141 s
107 s
81 s
80 s
66 s
Gold 6240
EPYC 7763
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
Number Of Gpus
0
1
2
4
8

VASP 6.3.0


VASP
Benchmark: GaAsBi_512
1 / Elapsed time (sec) (▸ Higher is better)
2857 s
1374 s
1241 s
690 s
625 s
398 s
351 s
246 s
216 s
171 s
Gold 6240
EPYC 7763
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
Number Of Gpus
0
1
2
4
8

VASP 6.3.0


VASP
Benchmark: Si256_VJT_HSE06
1 / Elapsed time (sec) (▸ Higher is better)
3861 s
3311 s
1062 s
559 s
546 s
292 s
283 s
156 s
163 s
100 s
Gold 6240
EPYC 7763
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
A30
A100-SXM4-80GB
Number Of Gpus
0
1
2
4
8

VASP 6.2.0 vs 6.1.2