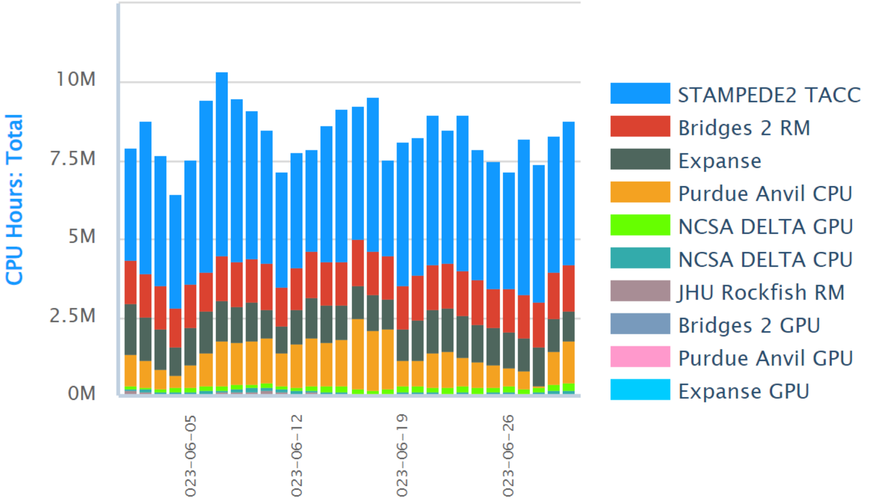
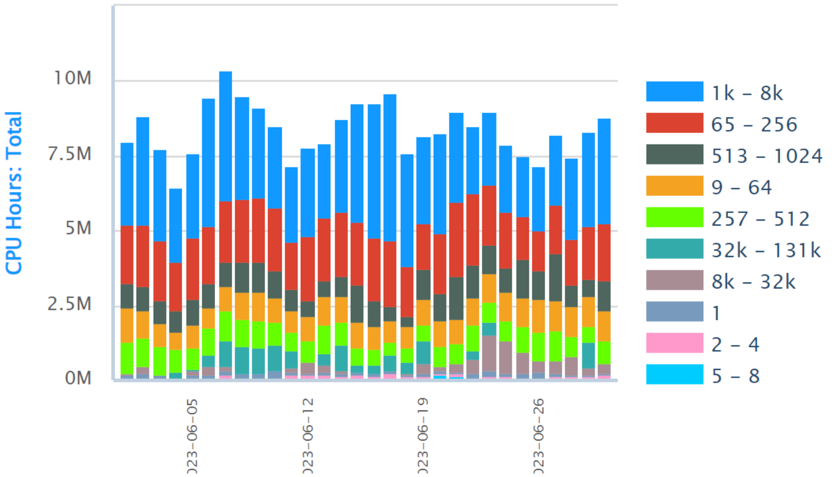
ACCESS Advancing Innovation

XDMoD METRICS ON DEMAND





Feasibility of Application-Agnostic Speed and Cost Performance Metrics on an Example of Gromacs, a Molecular Dynamics Application

Nikolay A. Simakov, Robert L. DeLeon

NSF Award: 2137603

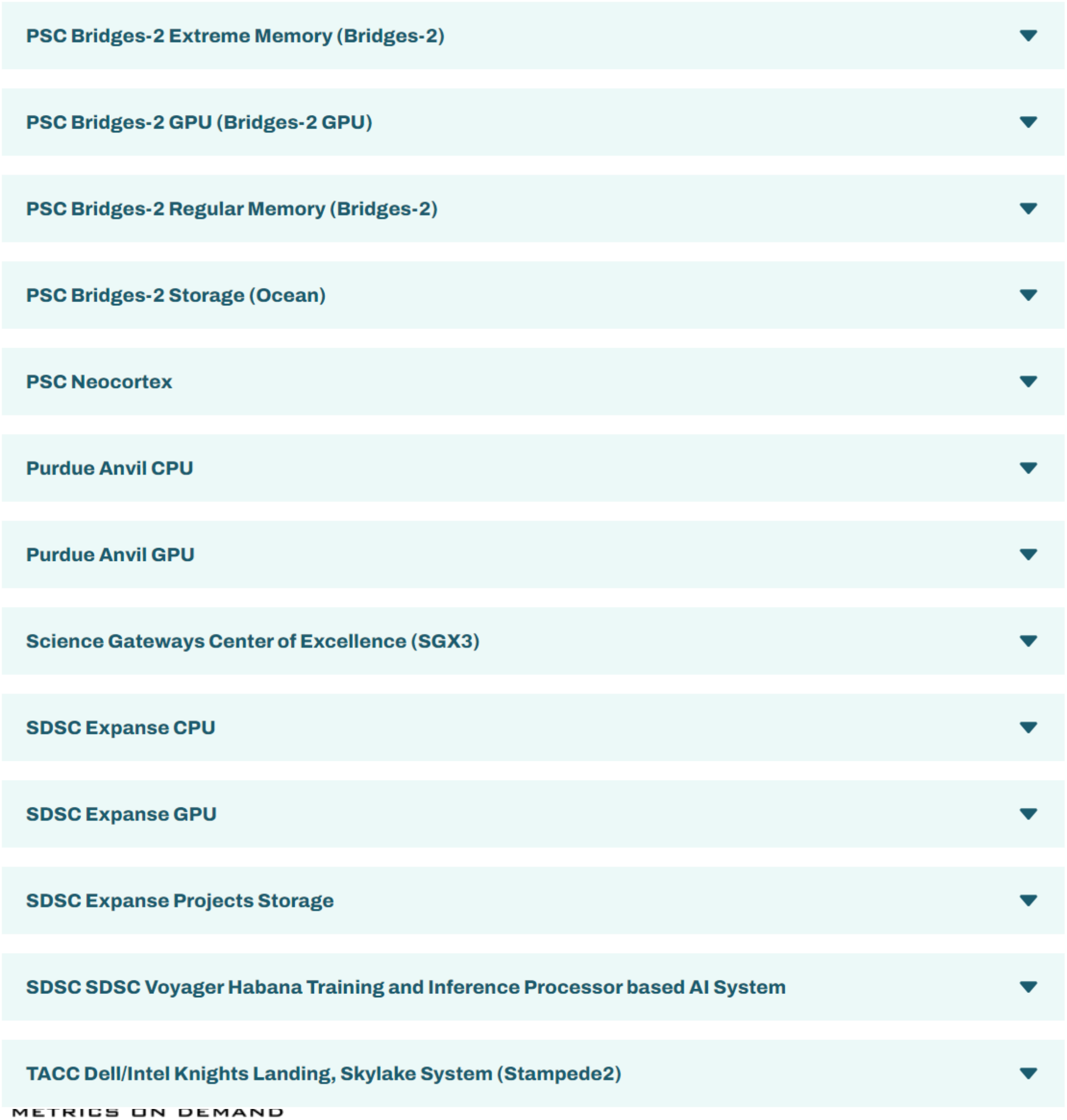
Metric23, Denver, Colorado, November 11 – 12, 2023

Abstract

The majority of modern HPC resources are significantly heterogeneous. Even same-generation compute nodes can be sufficiently diverged in the configuration as well as price. The nodes can include regular compute nodes, large memory nodes, fast memory (HBM) nodes, GPU-HBM nodes, and GPU-DDR nodes. Applications can exhibit different performances over all these nodes in the sense of performance per time and performance per currency. From an economic point of view, it makes sense to utilize the nodes with the best performance per currency. Unfortunately, such a metric is often unavailable for HPC users because the price of a machine is not readily available, and performance measurements itself require additional effort. Can we devise calculation speed and economic metrics relying on the actual users' application execution? In this work, we examine two such metrics, namely floating point operations (FLOP) per second and per dollar for giving application, and correlate them to well-established Gromacs metrics like simulated nanoseconds per day and per dollar. The grouping by used memory is used to approximate the problem sizes. The comparison is done for three test systems executed on multiple platforms and compared to FLOP per second and per dollar from actual HPC system users.

Motivation

ACCESS Allocations



PSC Bridges-2 Extreme Memory (Bridges-2)

PSC Bridges-2 GPU (Bridges-2 GPU)

PSC Bridges-2 Regular Memory (Bridges-2)

PSC Bridges-2 Storage (Ocean)

PSC Neocortex

Purdue Anvil CPU

Purdue Anvil GPU

Science Gateways Center of Excellence (SGX3)

SDSC Expanse CPU

SDSC Expanse GPU

SDSC Expanse Projects Storage

SDSC SDSC Voyager Habana Training and Inference Processor based AI System

TACC Dell/Intel Knights Landing, Skylake System (Stampede2)

METRICS ON DEMAND

Modern HPC resources are heterogeneous

- price differ greatly
- performance vary as well

Where to run?

- The fastest?
- The most money/energy efficient?

When running on cloud and/or paying actual money many users would benchmark first and choose most economical, fastest or balanced.

With hour-allocated or fair-share HPC resources many users don’t run benchmarks and run where they can or where the lines are shorter

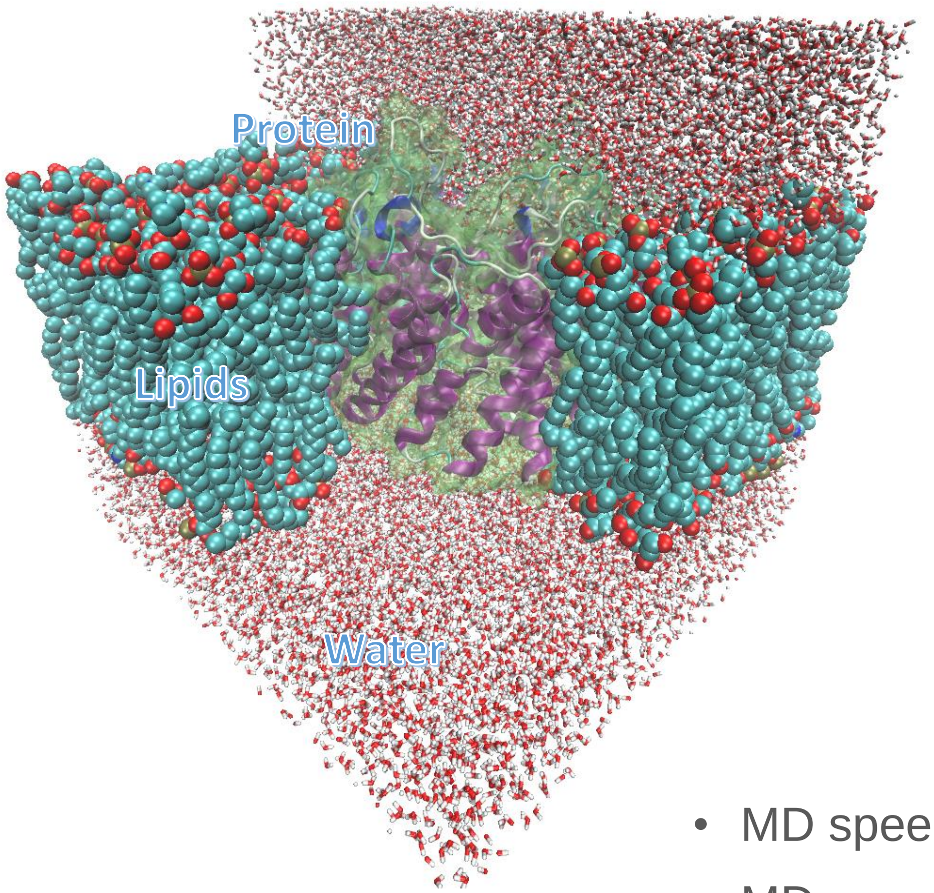
ACCESS credits is attempted to address some of this issues but users still are not much incentivized to run benchmarks themselves

Users might use performance data for resource and sub-resource selection.

ACCESS Metrics

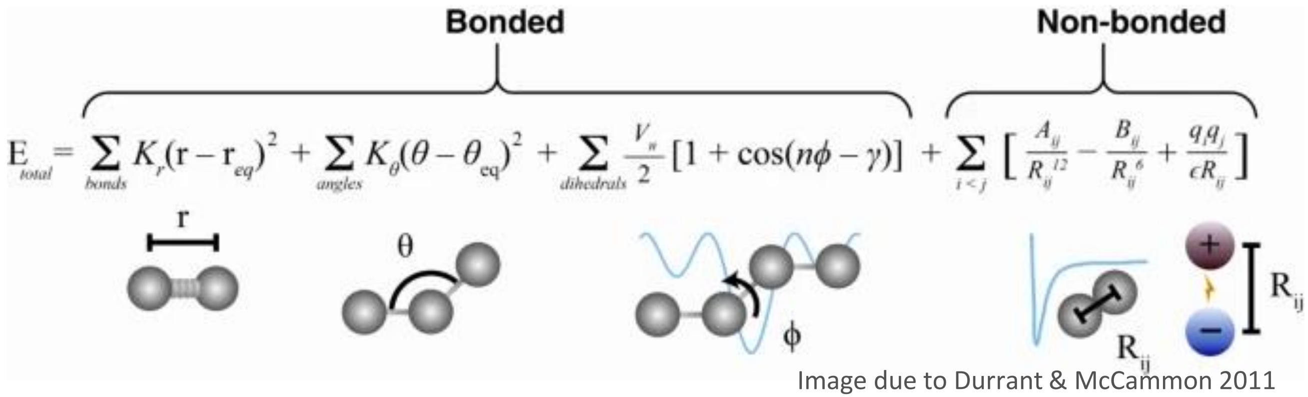
NSF

All Atomic Molecular Dynamics (MD) Simulation of Biomolecular Systems



- Solve second-Newton equation for classical-physics representation of biomolecular system

$$m_i \ddot{\boldsymbol{r}}_i = -\frac{\partial E_{total}}{\partial \boldsymbol{r}_i}$$



$$E_{total} = \overbrace{\sum_{bonds} K_r(\mathrm{r}-\mathrm{r}_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1+\cos(n\phi-\gamma)]}^{\text{Bonded}} + \overbrace{\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]}^{\text{Non-bonded}}$$

Image due to Durrant & McCammon 2011

- MD speed performance metric is **simulated ns / real days (ns/day).** Higer better.
- MD economy performance metric is $ / **simulated ns ($/ns) .** Smaller better.

10 times more money for 2 times slower performance !

- Price and performance comparison for Gromacs for two systems

Multi-node run can be faster than single GPU run, but for 6-67 times more money

Provider	GPU(NVIDIA)	CPU	Physicsl Cores	Price per hour	PEP 1,200,000 Atoms		MEM 82,000 Ator	
					Speed, ns/day	Price, $/ns	Speed, ns/day	Price, $/ns
vast.ai	RTX 4090 Consumer grade GPU			0.51	3.82	3.20	284.82	0.04
runpod	RTX 6000 Ada			1.14	2.37	11.57	245.40	0.11
runpod	L40			1.14	2.53	10.82	160.23	0.17
runpod	H100-PCE			4	2.88	33.37	183.92	0.52
google	L4			1.16	0.98	28.41	142.04	0.20
amazon	a10g			1.64	1.76	22.31	160.53	0.25
Stony Brooks University	A100	ICL	64	4.10*	2.42	40.69	242.62	0.41
		Intel SKX	40	3.84	0.42	218.70	51.40	1.79
		Intel SKX	40x8	30.72	2.94	251.13	204.00	3.61
		AMD Milan	96	8.29*	0.92	216.33	95.31	2.09
		AMD Milan	96x4	33.17*			298.04	2.67
		Intel SPR	96	9.67*	1.18	196.89	203.64	1.14
		Intel SPR	96x2	19.35*	2.33	198.97		
		Intel SPR-HBM	96	13.54**	1.20	270.05	206.10	1.58

- Gromacs can efficiently use a single GPU

  - The price is based on comparable AWS system
  - Scaled by 30% to account for increase in HBM additional cost

Two times slower for ten times more money

The Goal

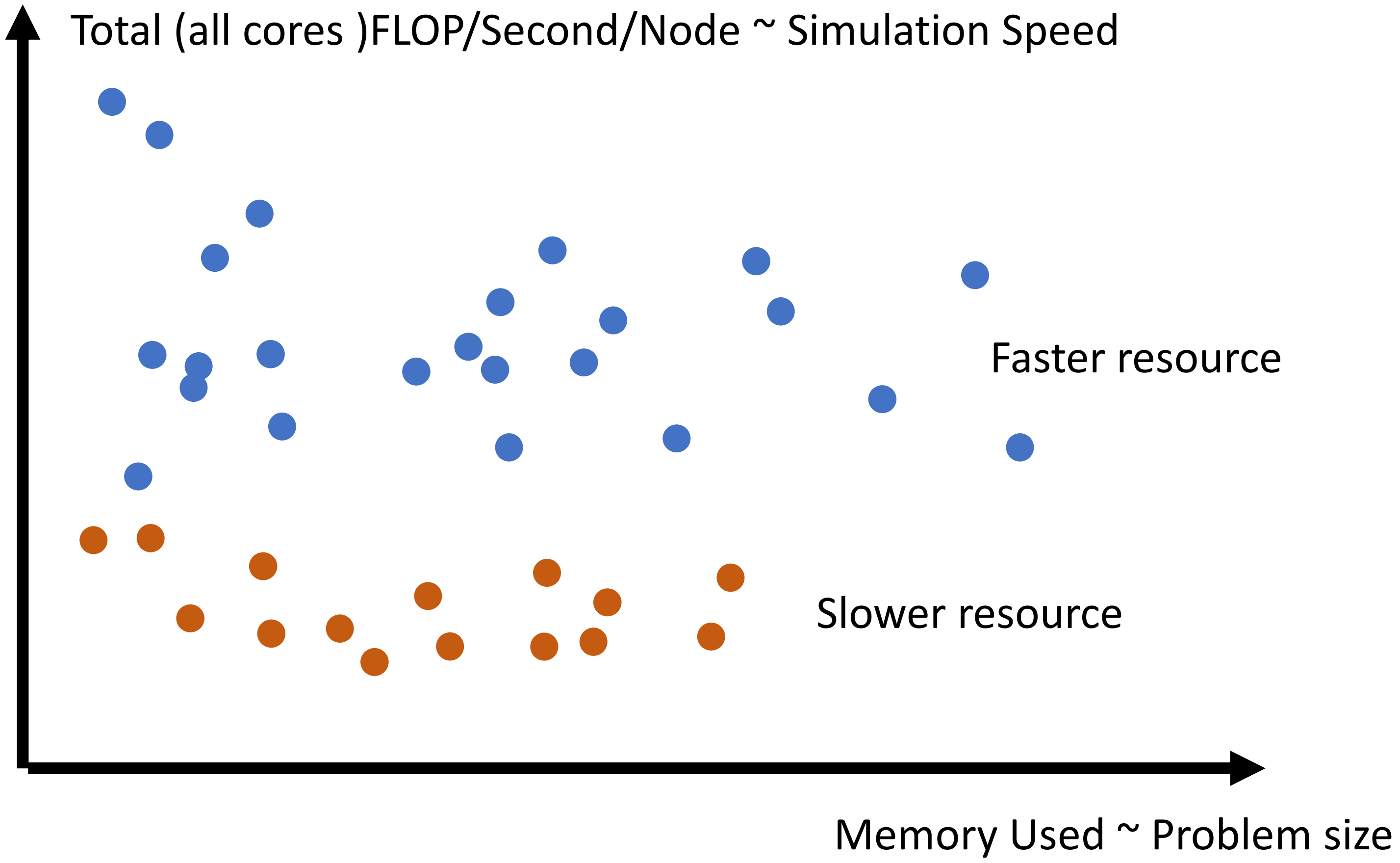
The goal is to provide an estimate on an application performance on a HPC resource without running benchmarks and only relying on the users' actual calculations.

We are interested in performance in both speed and economic aspects

Importantly we want to compare different resource

The Idea

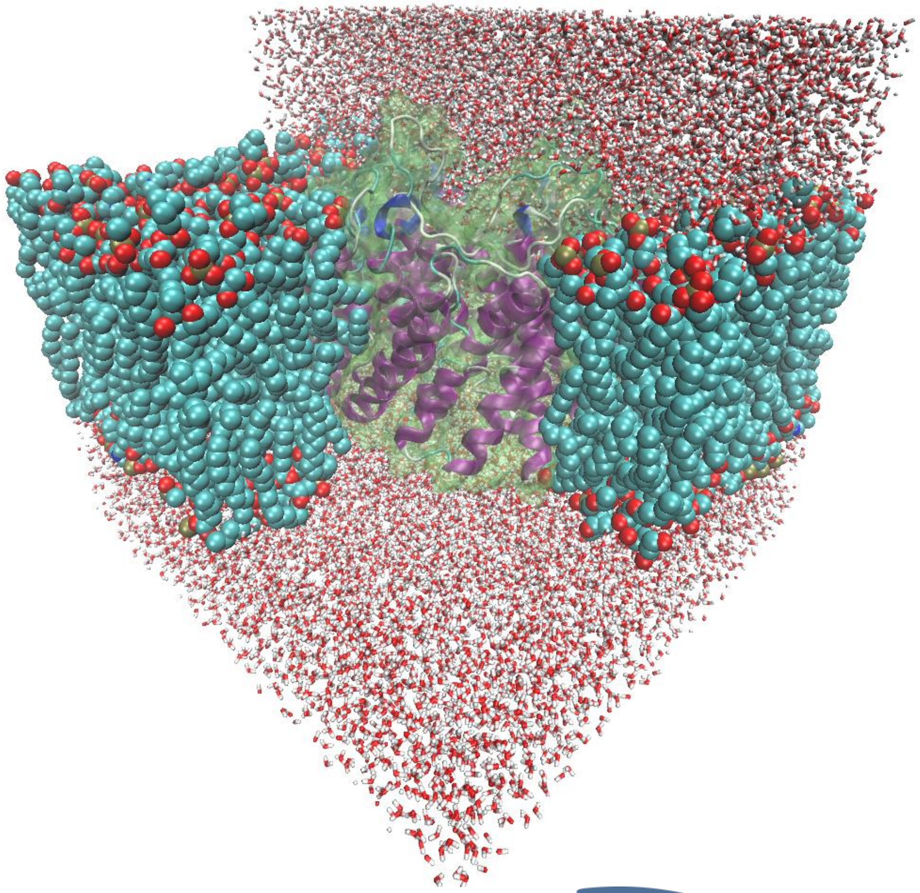
- **FLOPs/second can be used as such metric**
- FLOPs/$ (or FLOPs/ACCESS credit) can be used as economic metric



Challenges:

- **Different methods within application can exhibit different behavior**
  - Possible remedies:
    - Assume similar jobs executed across all platforms (Probably not true)
    - Provide data for Cluster
    - Show raw results and let user decides (Something is better than nothing).

Controlled Gromacs Benchmarks



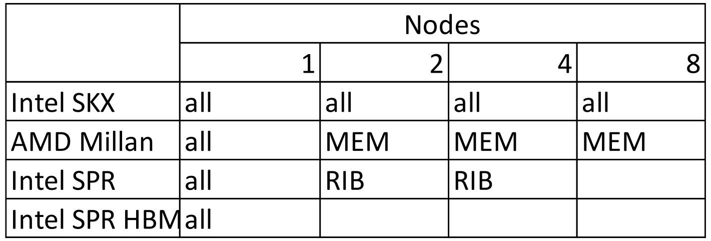
FAST. FLEXIBLE. FREE.
GROMACS

XDMoD
METRICS ON DEMAND

Want to compare the Gromacs intrinsic performance metric to general FLOPs/second and FLOPs/$

- **Three test systems:**
  - MEM (small membrane protein), 82,000 atoms, 2 fs time step
  - RIB (ribosome) 2,000,000 atoms, 4 fs time step
  - Pep (peptides solution) 12,000,000 atoms, 2 fs time step
- **Several systems at SBU**
  - Intel SKX
  - AMD Millan
  - Intel SPR
  - Intel SPR HBM

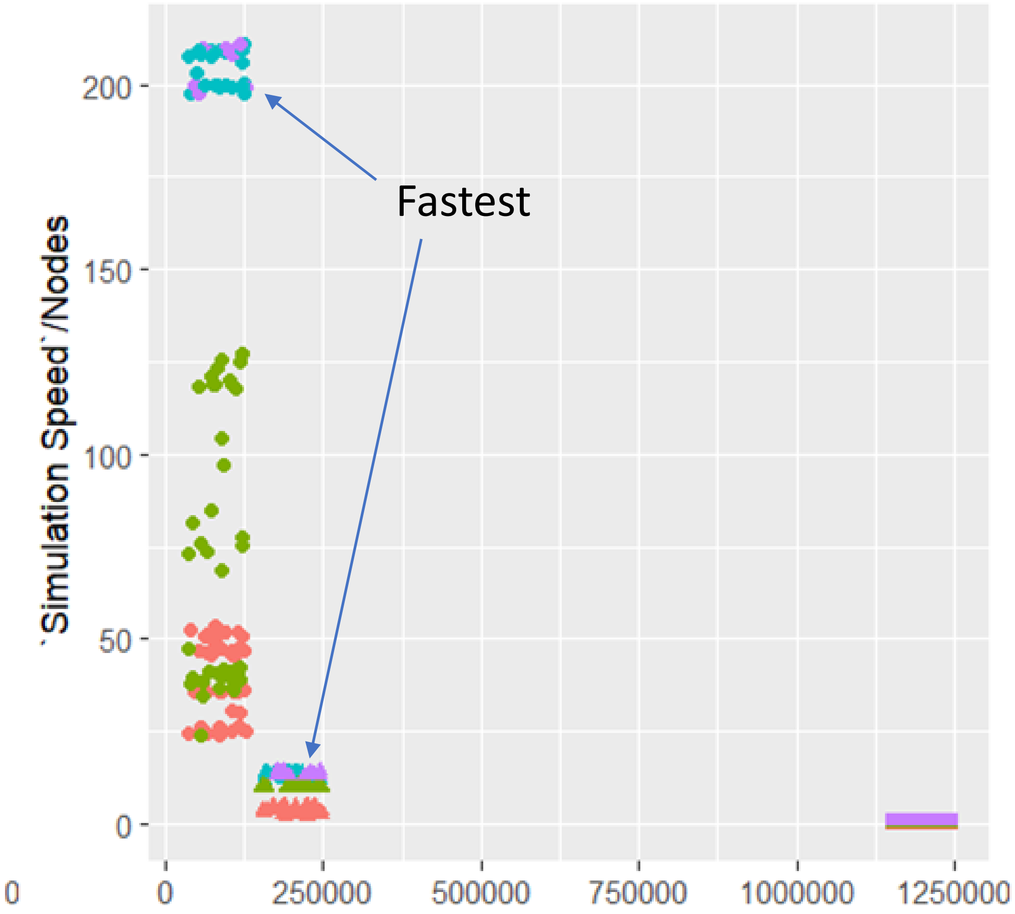
- Calculations most often done in **single precision**
- All test system run on single node, on some platforms we run multi-nodal as well:

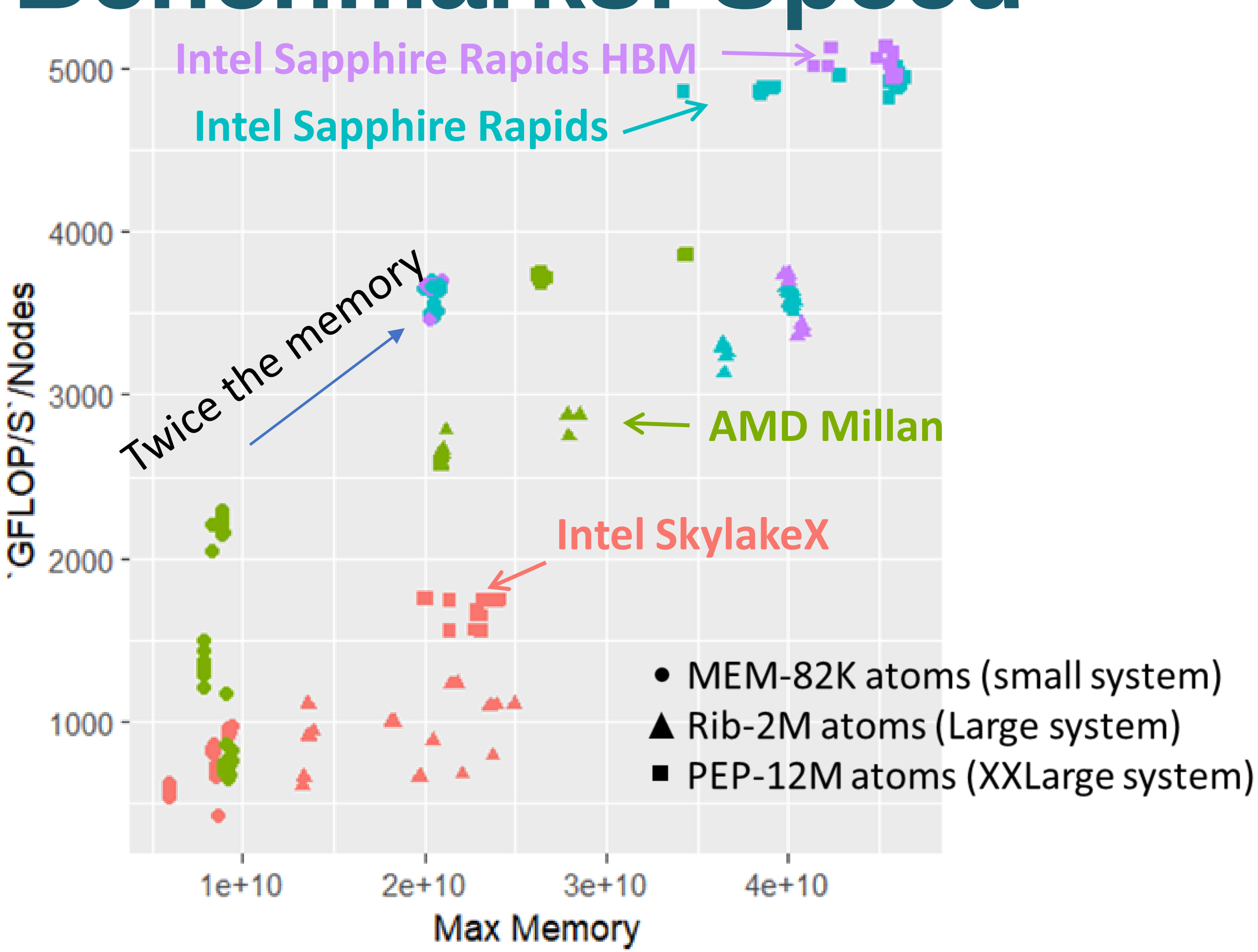


	Nodes			
	1	2	4	8
Intel SKX	all	all	all	all
AMD Millan	all	MEM	MEM	MEM
Intel SPR	all	RIB	RIB	
Intel SPR HBM	all			

ACCESS
Metrics

Controlled Gromacs Benchmarks: Speed



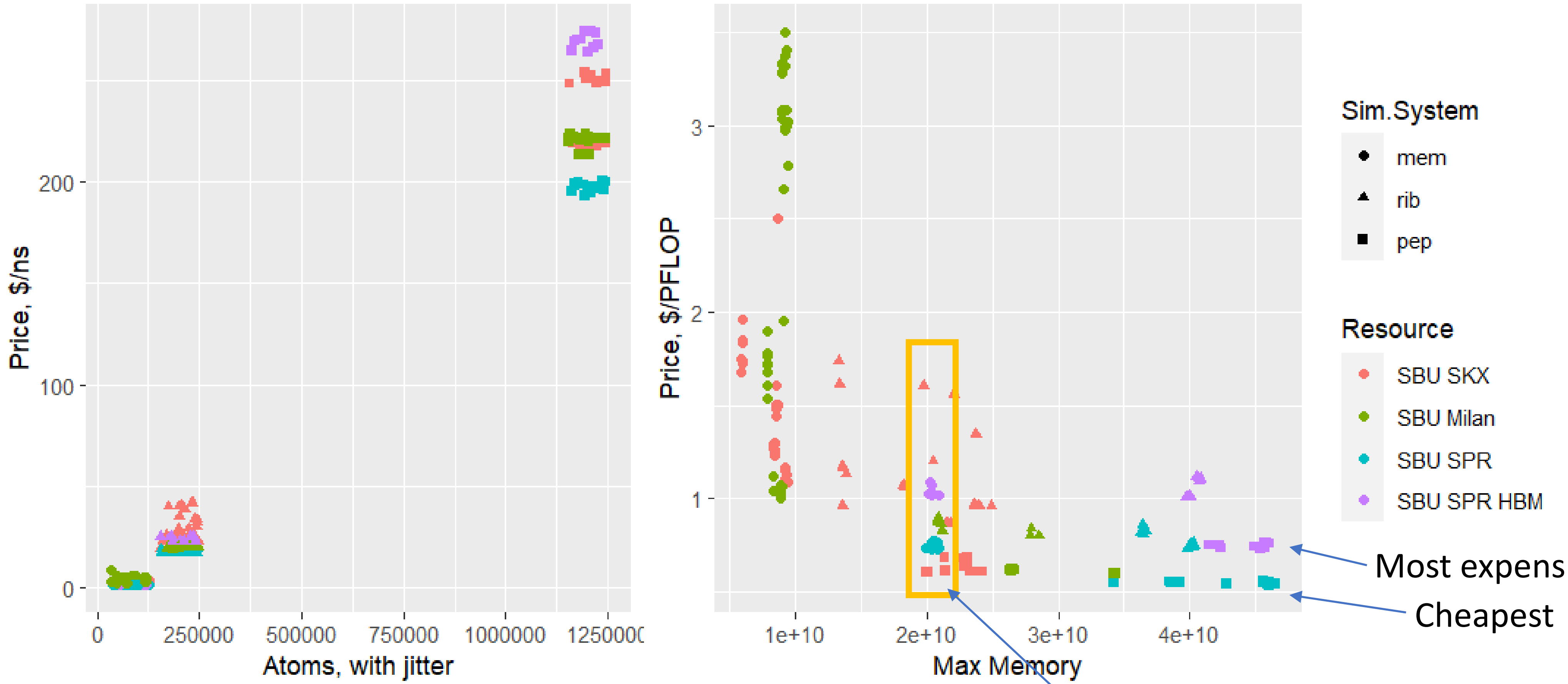


CPU	Physicsl Cores	Price per hour	PEP 1,200,000 Atoms	
			Speed, ns/day	Price, $/ns
Intel SKX	40	3.84	0.42	218.70
Intel SKX	40x8	30.72	2.94	251.13
AMD Milan	96	8.29*	0.92	216.33
Intel SPR	96	9.67*	1.18	196.89
Intel SPR	96x2	19.35*	2.33	198.97
Intel SPR-HBM	96	13.54**	1.20	270.05

- **Faster resource is on top, but memory footprint on different systems is different**

ACCESS Metrics

Controlled Gromacs Benchmarks: Cost



Most expensive

Cheapest

Ranking: SKX<SPR<Millan<SPR HBM<SKX?

CPU	Physicsl Cores	Price per hour	PEP 1,200,000 Atoms	
			Speed, ns/day	Price, $/ns
Intel SKX	40	3.84	0.42	218.70
Intel SKX	40x8	30.72	2.94	251.13
AMD Milan	96	8.29*	0.92	216.33
Intel SPR	96	9.67*	1.18	196.89
Intel SPR	96x2	19.35*	2.33	198.97
Intel SPR-HBM	96	13.54**	1.20	270.05

- **Hard to see the difference!**
- **Most of them are similarly priced**
- **What about ACCESS credits?**

Comparing the Historic HPC Users’ Jobs

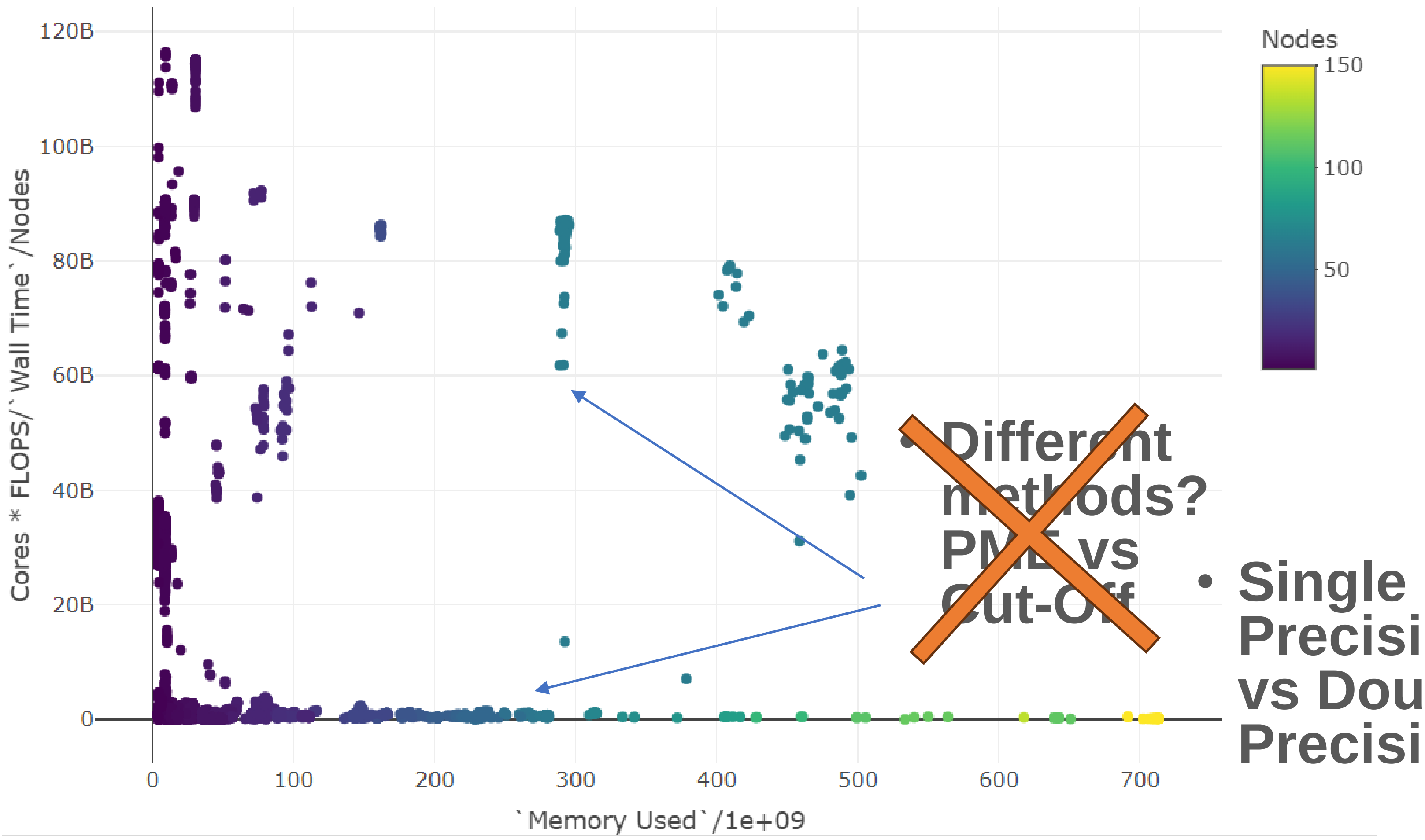
- Right now, FLOPS data is available only for
  - TACC-Stampede (Intel Xeon E5-2680, Sandy Bridge, Q1 2012)
    - Only double precession (not enough counters for both SP and DP)

- Collected but not yet properly converted
  - TACC-Stampede 2 SKX (Intel Xeon Platinum 8160, Skylake, Q3 2017)
    - Only double precession (not enough counters for both SP and DP)

- Not collected but existing tools can be relatively easy extended
  - TACC-Stampede 2 ICX (Intel Xeon Platinum 8380, Ice Lake, Q2 2021)
    - Enough counters for both single and double precession

Data obtained from TACC-Stat and processed by XDMoD

Gromacs: Users jobs at Stampede 1



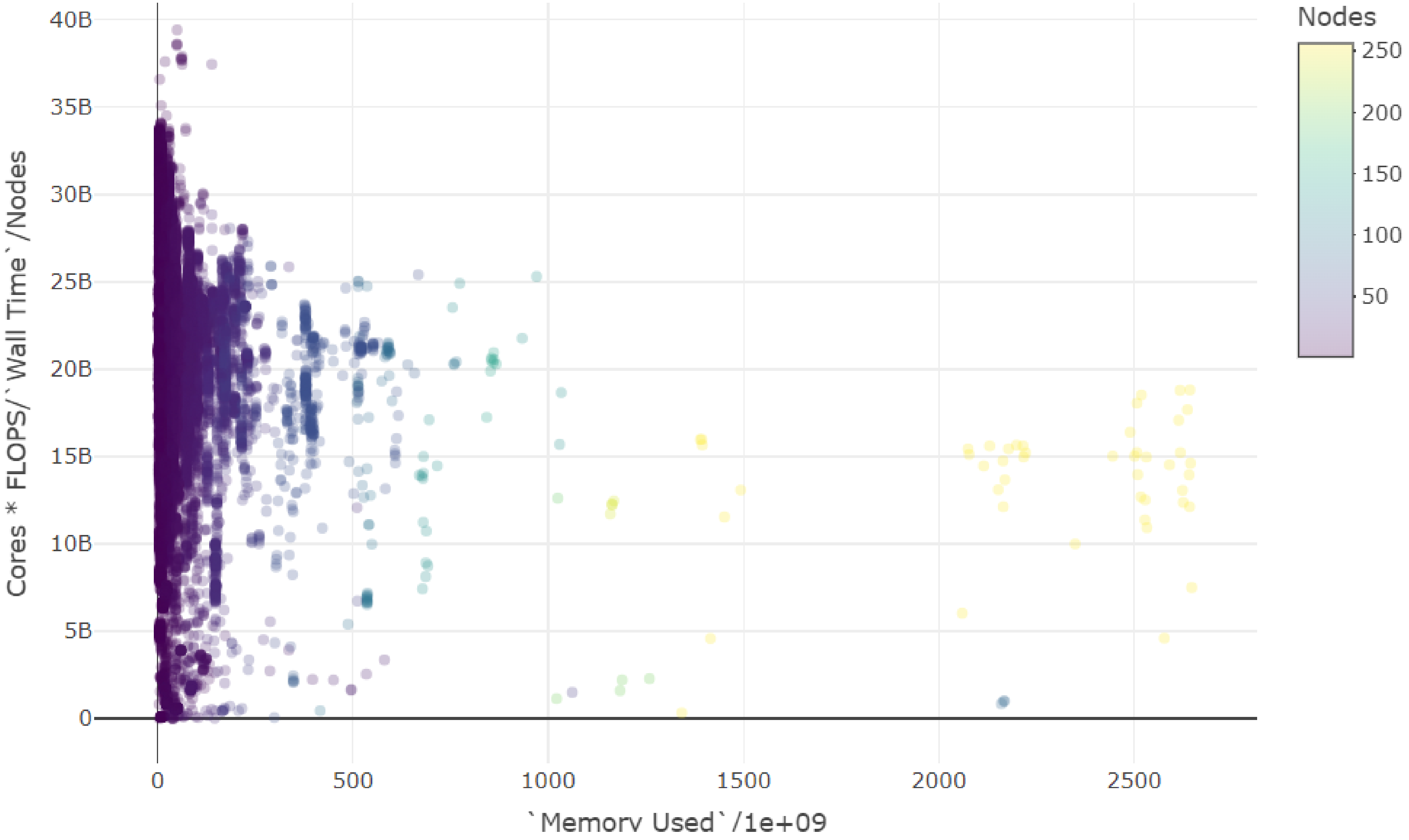
- ~~Different methods? PME vs Cut-Off~~

- Single Precision vs Double Precision

Gromacs most often uses single precision FP, some calculations would require double precision

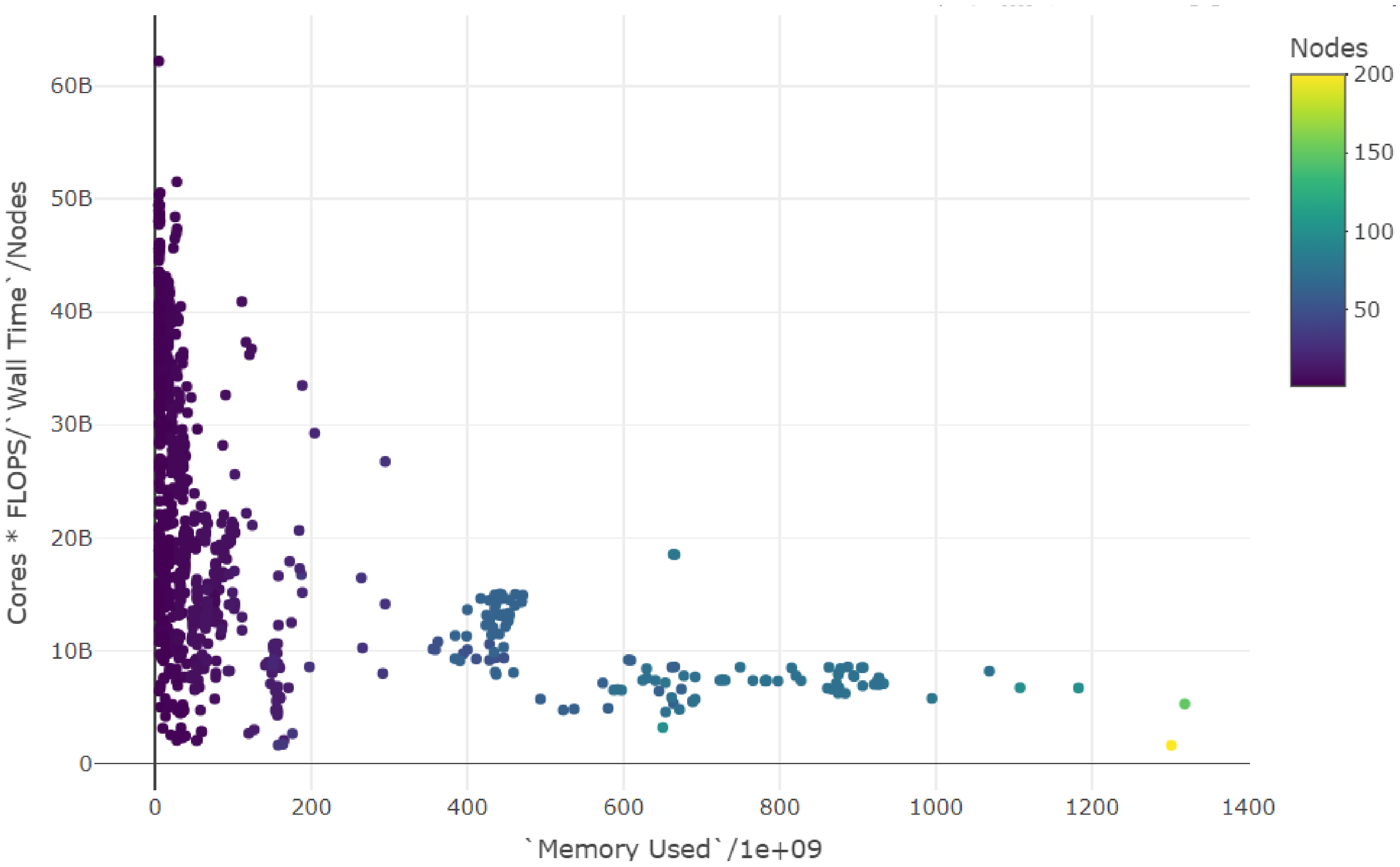
Data obtained from TACC-Stat through XDMoD

NAMD: Users jobs at Stampede 1



- **More uniform than Gromacs**

NWChem: Users jobs at Stampede 1



Conclusion and Future Plans

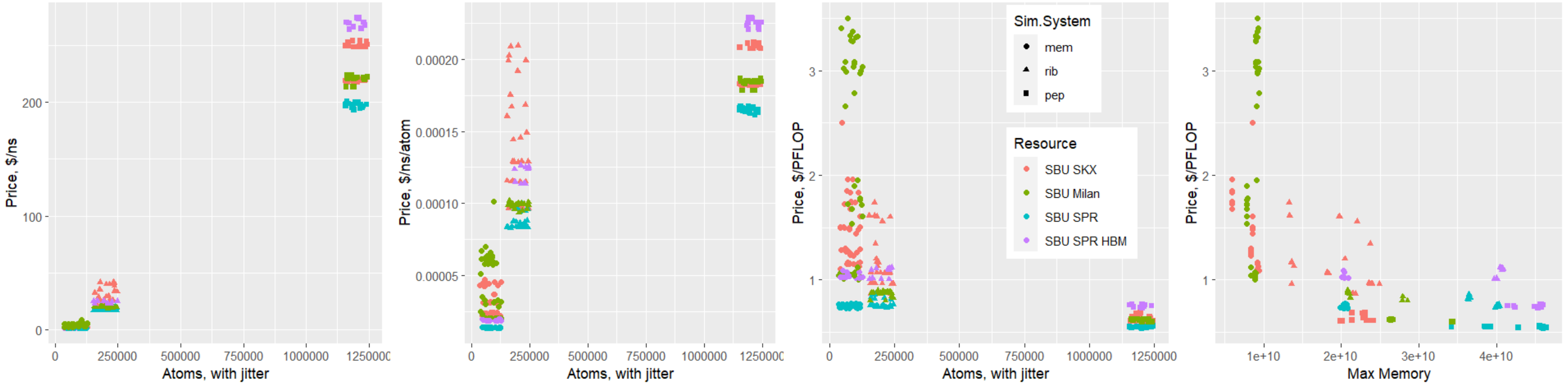
- Something is there, more work is needed
- Need comparison between different system
- Need encourage resource providers to track FLOPS

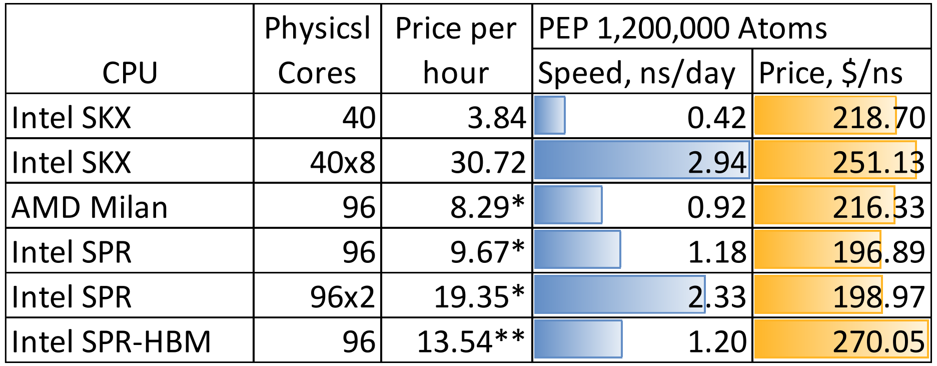
- Challenge on hybrid (CPU-GPU) system: FLOPs from all compute devices, memory from CPU?

- Future plans: Provide FLOP/S and dollars per FLOPs vs Max Memory for user run application on XDMoD and ACCESS-CI portals for resources recommendation.

- Future prognose: economy prioritized job scheduling

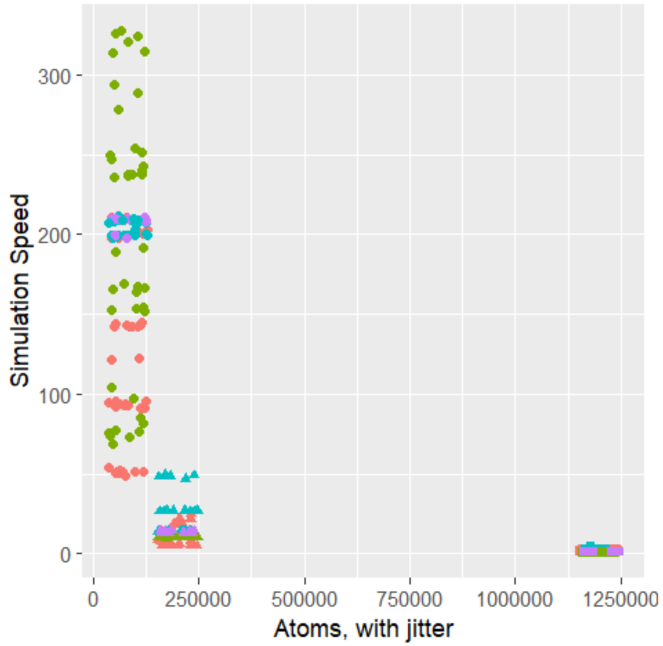
Controlled Gromacs Benchmarks: Cost

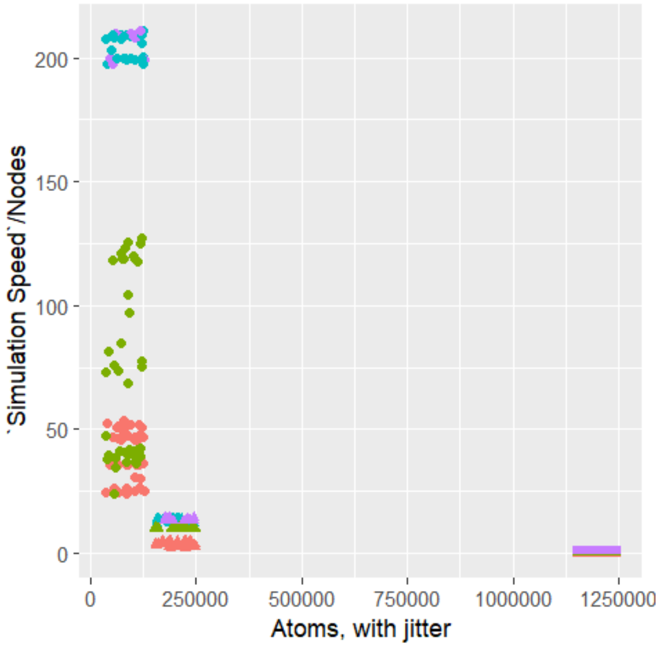


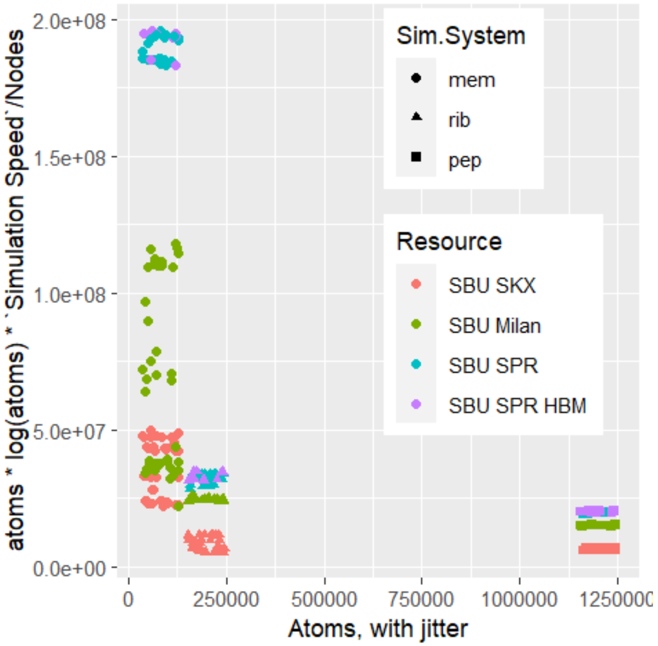


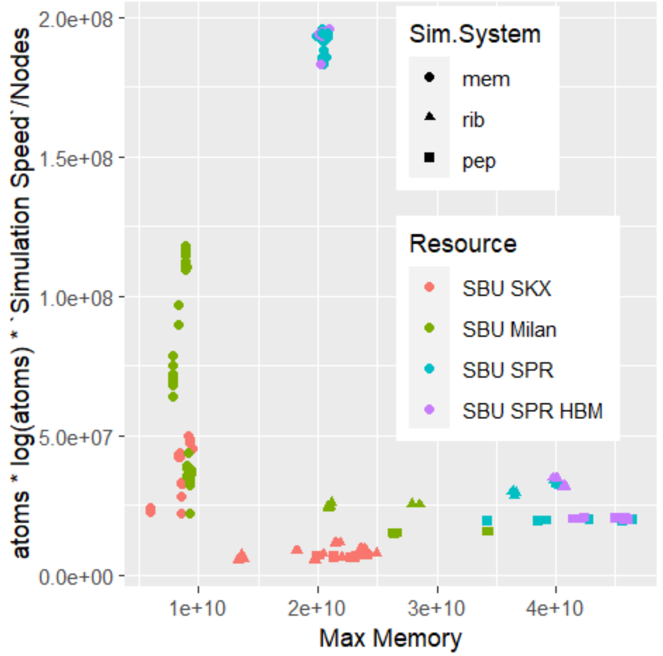
	Physicsl	Price per	PEP 1,200,000 Atoms	
CPU	Cores	hour	Speed, ns/day	Price, $/ns
Intel SKX	40	3.84	0.42	218.70
Intel SKX	40x8	30.72	2.94	251.13
AMD Milan	96	8.29*	0.92	216.33
Intel SPR	96	9.67*	1.18	196.89
Intel SPR	96x2	19.35*	2.33	198.97
Intel SPR-HBM	96	13.54**	1.20	270.05

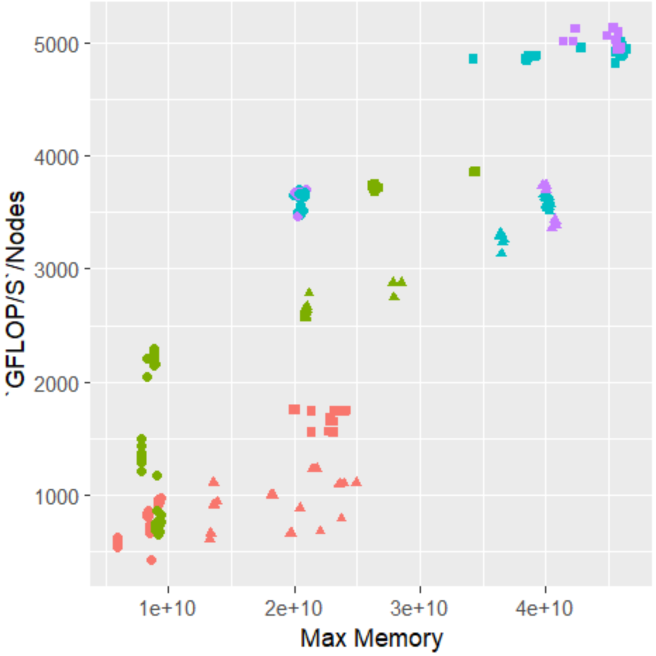
Controlled Gromacs Benchmarks: Speed











total FLOPs ~ sim.time*atoms*log(atoms)

Analog of FLOPs/seconds/Nodes
~ time*atoms*log(atoms)/run time/Nodes
~ Sim.speed*atoms*log(atoms)/Nodes

- Compare only subset of simulated system for each hardware-randomize colors, which one is faster?