Carsten Kutzner
Theoretical & Computational Biophysics
MPI for biophysical Chemistry

“BEST BANG FOR YOUR BUCK”

Cost-efficient MD simulations

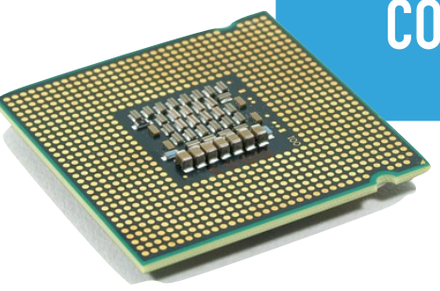
COST-EFFICIENT MD SIMULATIONS

TASK 1: CORE-H → .XTC

HOW TO GET OPTIMAL GROMACS PERFORMANCE?

fixed amount of core-h

YOU HAVE ACCESS TO COMPUTE RESOURCES



TASK 2: € → .XTC

WHAT IS THE OPTIMAL HARDWARE TO RUN GROMACS ON?

fixed amount of money

YOU WANT TO BUY A CLUSTER

HOW CAN I PRODUCE AS MUCH TRAJECTORY AS POSSIBLE FOR MY SCIENCE?

WHAT IS THE ‘OPTIMAL’ HARDWARE TO BUY?

WHAT DO WE WANT?

general-purpose cluster for all kinds of applications

- large RAM
- high-throughput, low-latency interconnect
- double-prec. GPU performance
- large GPU memory

WHAT IS THE ‘OPTIMAL’ HARDWARE TO BUY?

WHAT DO WE WANT?

general-purpose cluster for all kinds of applications

- large RAM
- high-throughput, low-latency interconnect
- double-prec. GPU performance
- large GPU memory

WHAT CAN WE SPARE?

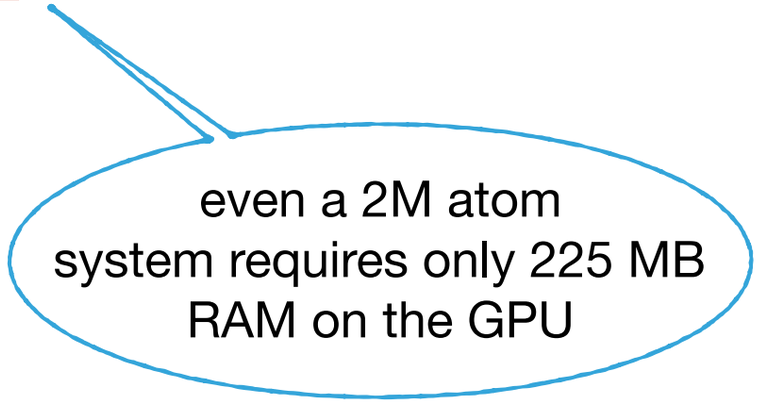
specialization maximizes cost-efficiency

WHAT IS THE ‘OPTIMAL’ HARDWARE TO BUY?

WHAT DO WE WANT?

general-purpose cluster for all kinds of applications

- large ~~RAM~~
- high-throughput, low-latency interconnect
- double-prec. ~~GPU~~ performance
- large ~~GPU~~ memory



WHAT CAN WE SPARE?

specialization maximizes cost-efficiency

- GROMACS only

WHAT IS THE 'OPTIMAL' HARDWARE TO BUY?

WHAT DO WE WANT?

general-purpose cluster for all kinds of applications

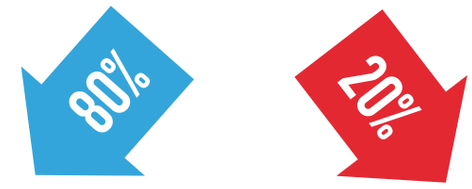
- ~~large RAM~~
- ~~high-throughput, low-latency interconnect~~
- ~~double-prec. GPU performance~~
- ~~large GPU memory~~

even a 2M atom system requires only 225 MB RAM on the GPU

WHAT CAN WE SPARE?

specialization maximizes cost-efficiency

- GROMACS only



max. sampling, many separate simulations

What we optimize our cluster for!

~~single long trajectories~~

run these @ national HPC centers

WHAT IS THE ‘OPTIMAL’ HARDWARE TO BUY?

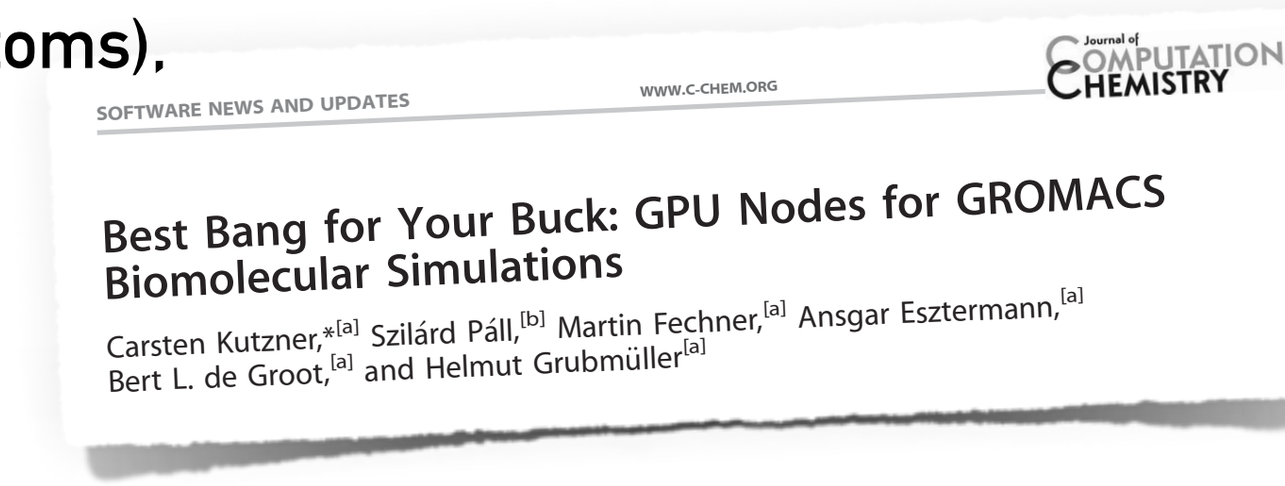
For us:

1. high performance-to-price ratio
   → maximize trajectory output per invested €
2. low energy consumption
3. good single-node performance
4. low rack space requirements
5. ~~scaling across many cluster nodes~~
   → HPC centers

FINDING THE OPTIMAL HARDWARE

- get prices + benchmark GROMACS performance for all reasonable hardware configurations

- 'Best bang for your buck' (2015):
  2 benchmark systems (80k / 2 M atoms),
  12 CPU types
  13 GPU types
  >50 hardware configurations



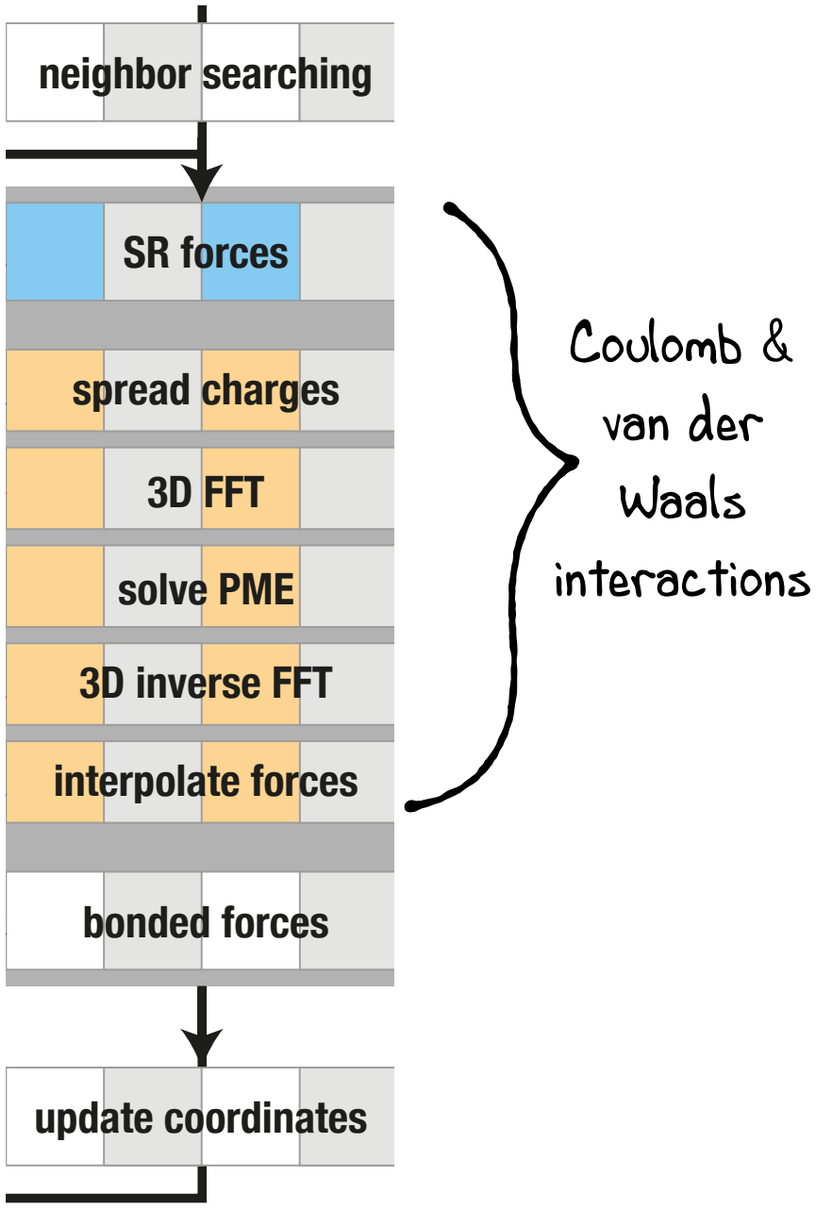
- on each hardware try to get optimal GROMACS performance

COMPILATION: COMPILER, SIMD INSTRUCTIONS, MPI LIB

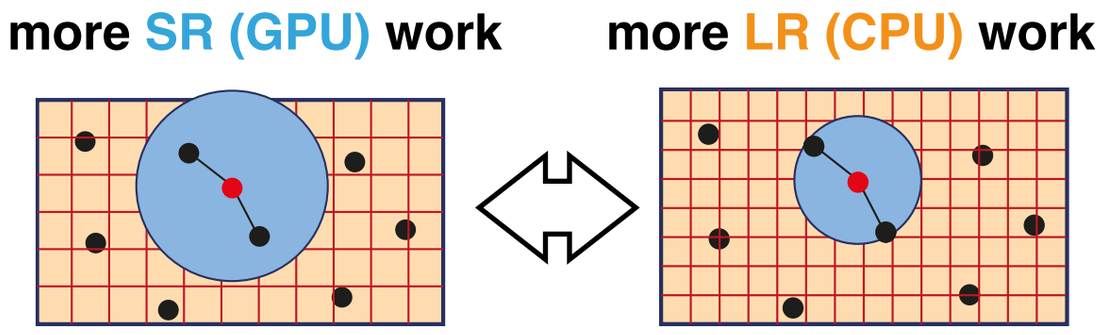
SYSTEM SETUP: V-SITES, BOX TYPE

MDRUN: FIND OPTIMAL RUN-TIME PARAMETERS

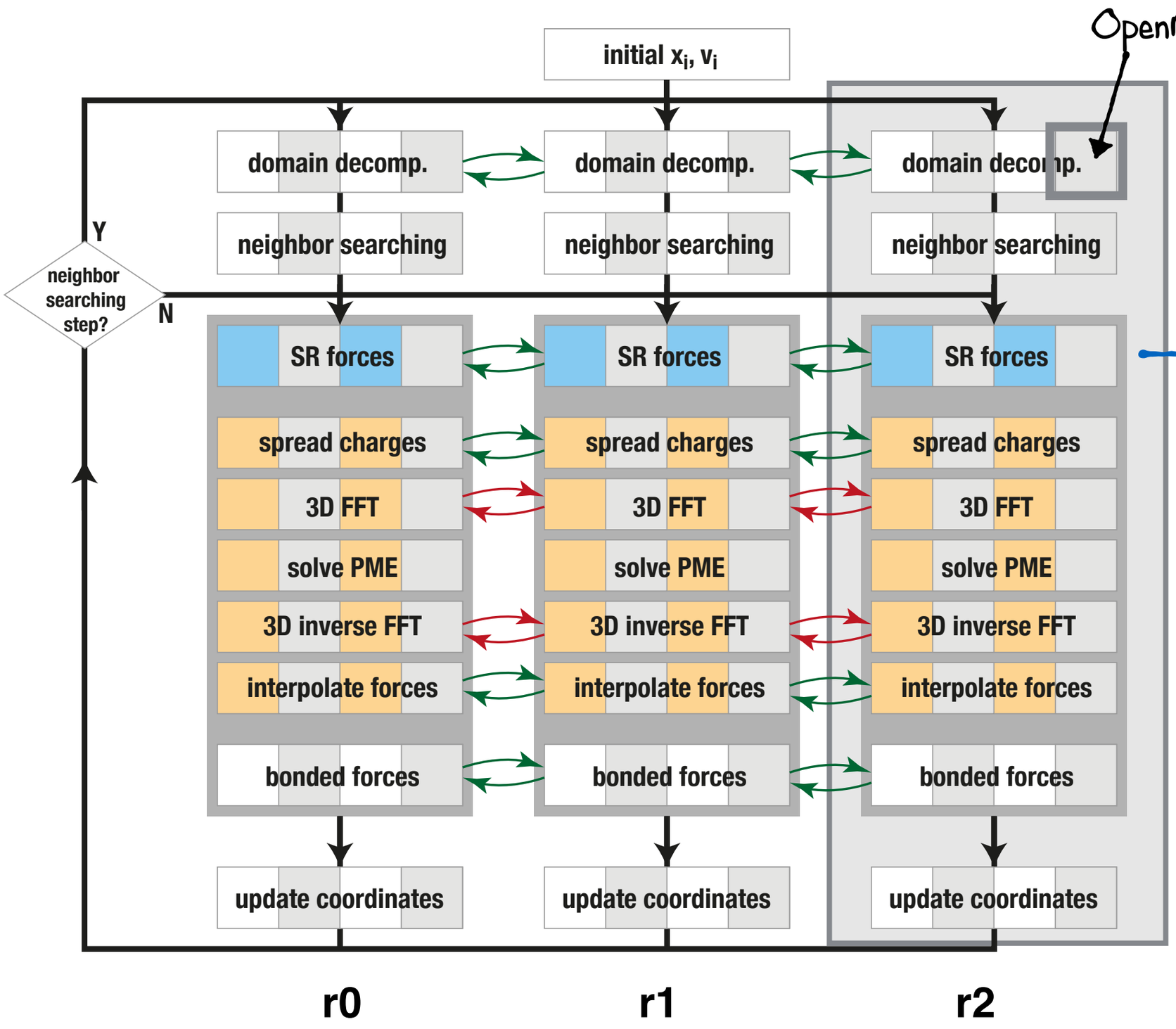
GROMACS TIME STEP

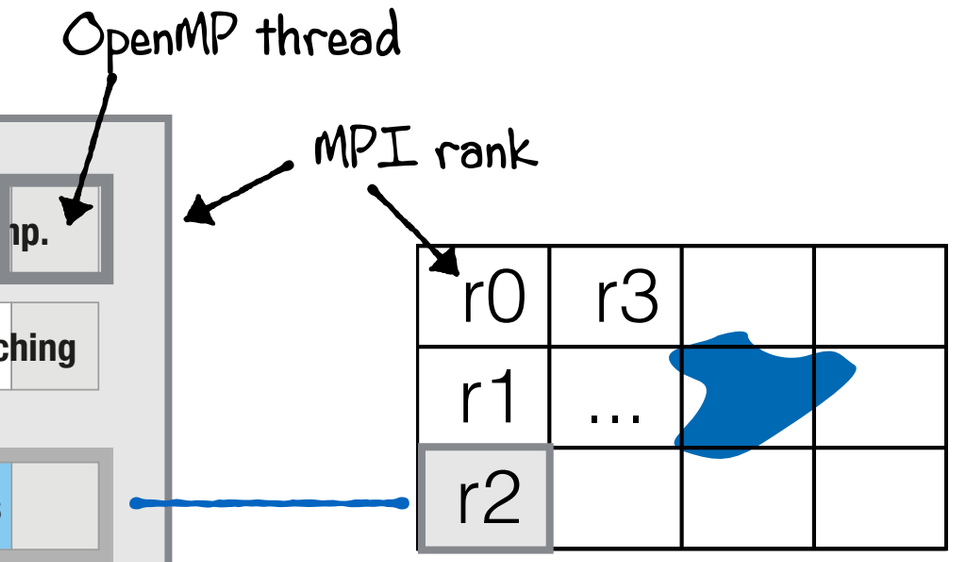


- Coulomb + vdW make up for most of the time step
- PME decomposes these into SR (direct) and LR (grid) contributions
- PME allows to shift work between real, SR (PP), and reciprocal, LR (PME), space parts (balance cutoff : grid spacing)

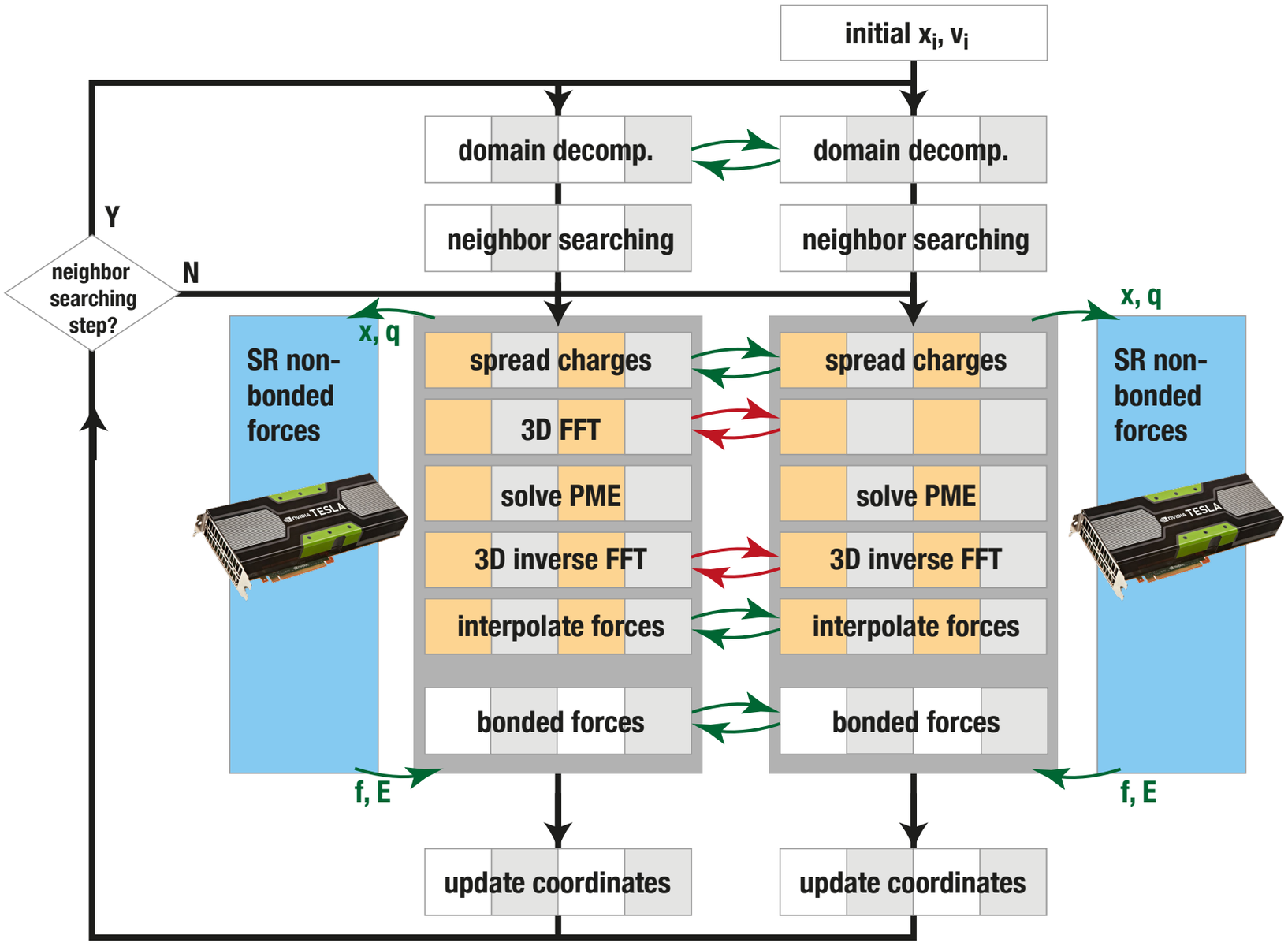


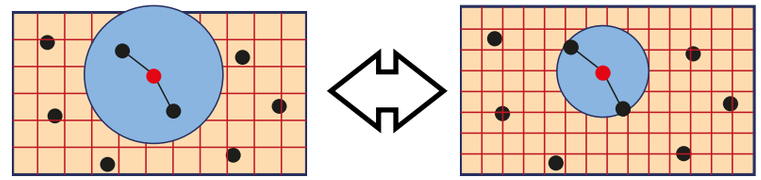
GROMACS TIME STEP / PARALLEL





direct space interactions decomposed into domains





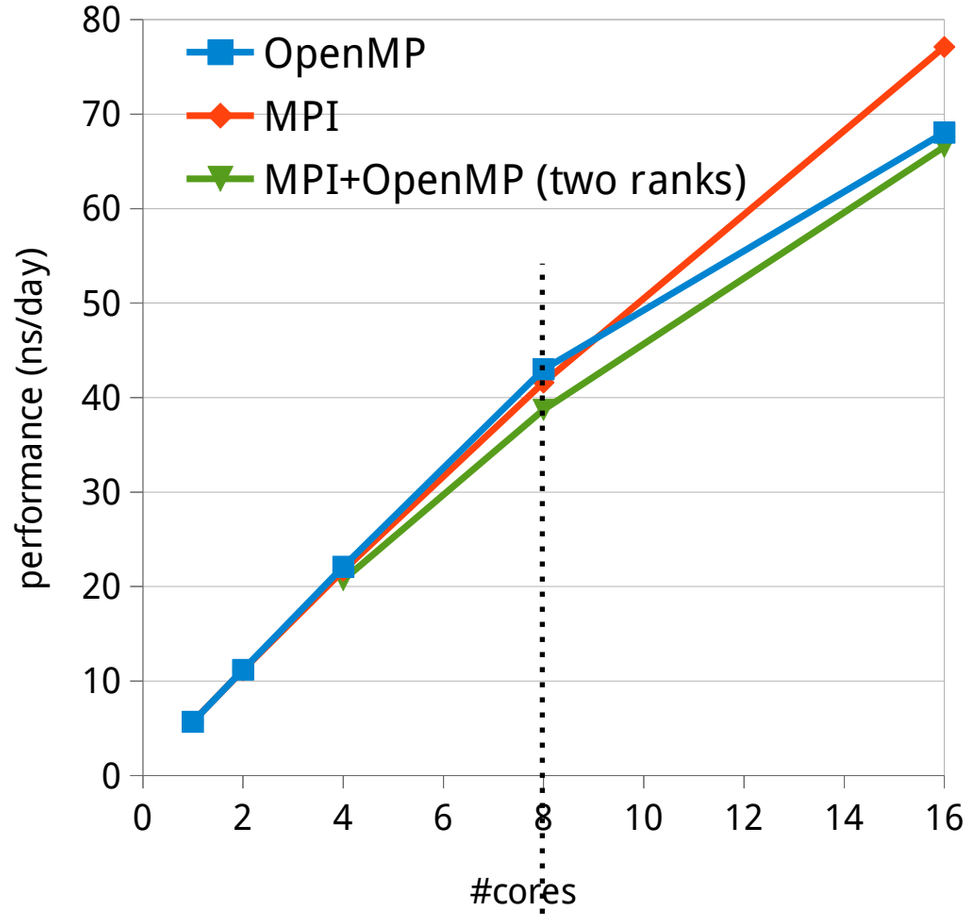
more SR (GPU) work

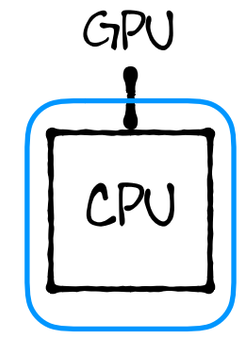
more LR (CPU) work

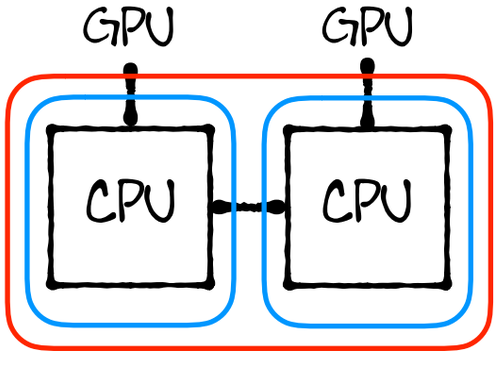
SR NON-BONDED FORCES ARE OFFLOADED TO GPUS, WITH AUTOMATIC BALANCING

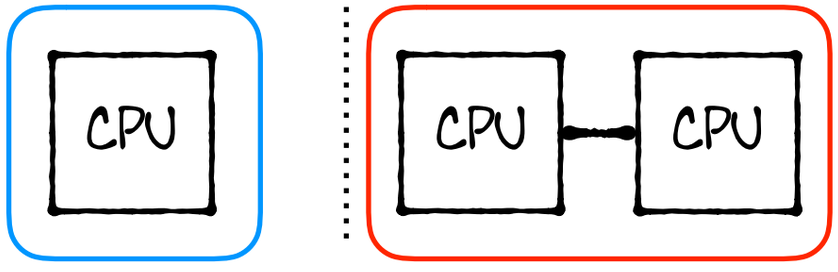
THE OPTIMAL MIX OF THREADS & RANKS

- MPI + OpenMP
  → work can be distributed in various ways
- **pure OpenMP** performs well on single CPUs, but does not scale well across sockets
- → on multi-socket nodes **pure MPI** is best
- OpenMP+MPI adds overhead
- With **GPUs** it is beneficial to have few large domains offloading their data to the GPU
  → use pure OpenMP unless multi-socket
- Multi-socket GPU nodes
  → find optimum!



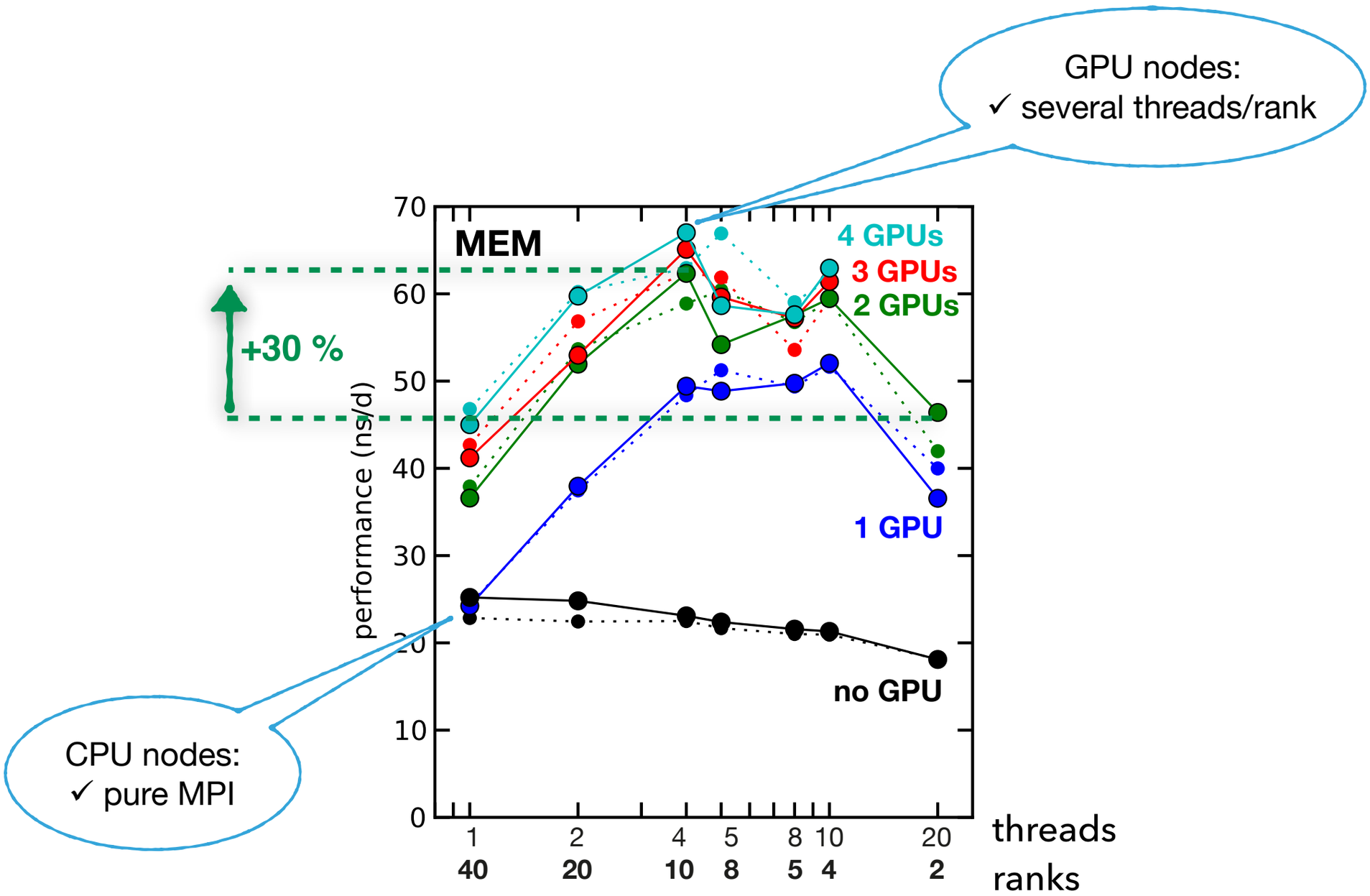






2x 8-core E5-2690 (Sandy Bridge), RNAse protein, solvated, 24k atoms, PME, 0.9 nm cutoffs (Fig. taken from S Pall, MJ Abraham, C Kutzner, B Hess, E Lindahl, EASC 2014, Springer, 2015)

THE OPTIMAL MIX OF THREADS & RANKS



2x E5-2680v2 (2x 10 cores) processors with 4x GTX 980 GPUs

GPU MODELS

2014

NVIDIA model	architec-ture	CUDA cores	clock rate (MHz)	memory (GB)	SP throughput (Gflop/s)	≈ price (€) (net)
Tesla K20X[a]	Kepler GK110	2,688	732	6	3,935	2,800
Tesla K40[a]	Kepler GK110	2,880	745	12	4,291	3,100
GTX 680	Kepler GK104	1,536	1,058	2	3,250	300
GTX 770	Kepler GK104	1,536	1,110	2	3,410	320
GTX 780	Kepler GK110	2,304	902	3	4,156	390
GTX 780Ti	Kepler GK110	2,880	928	3	5,345	520
GTX Titan	Kepler GK110	2,688	928	6	4,989	750
GTX Titan X	Maxwell GM200	3,072	1,002	12	6,156	
GTX 970	Maxwell GM204	1,664	1,050	4	3,494	250
GTX 980	Maxwell GM204	2,048	1,126	4	4,612	430
GTX 980^{+}	Maxwell GM204	2,048	1,266	4	5,186	450
GTX 980‡	Maxwell GM204	2,048	1,304	4	5,341	450

2017

NVIDIA model	architec-ture	CUDA-cores	clock rate (MHz)	memory (GB)	SP throughput (GFlop/s)	≈ price (€ net)
Tesla K40	Kepler GK110B	2 880	745	12	4 291	2 500
Tesla P100	Pascal P100	3 584	1328	16	9 519	3 200
GTX 1060	Pascal GP106-400	1 280	1506	3	3 855	152
GTX 1070	Pascal GP104-200	1 920	1506	8	5 783	330
GTX 1080	Pascal GP104-400	2 560	1607	8	8 228	420
GTX 1080Ti	Pascal GP102-350-K1	3 584	1480	11	10 609	625

CONSUMER GPU ERROR RATES

- consumer GPUs do not have ECC memory, thus cannot correct for rare bit-flips
- however, GPU stress tests can be used to sort out problematic GPUs

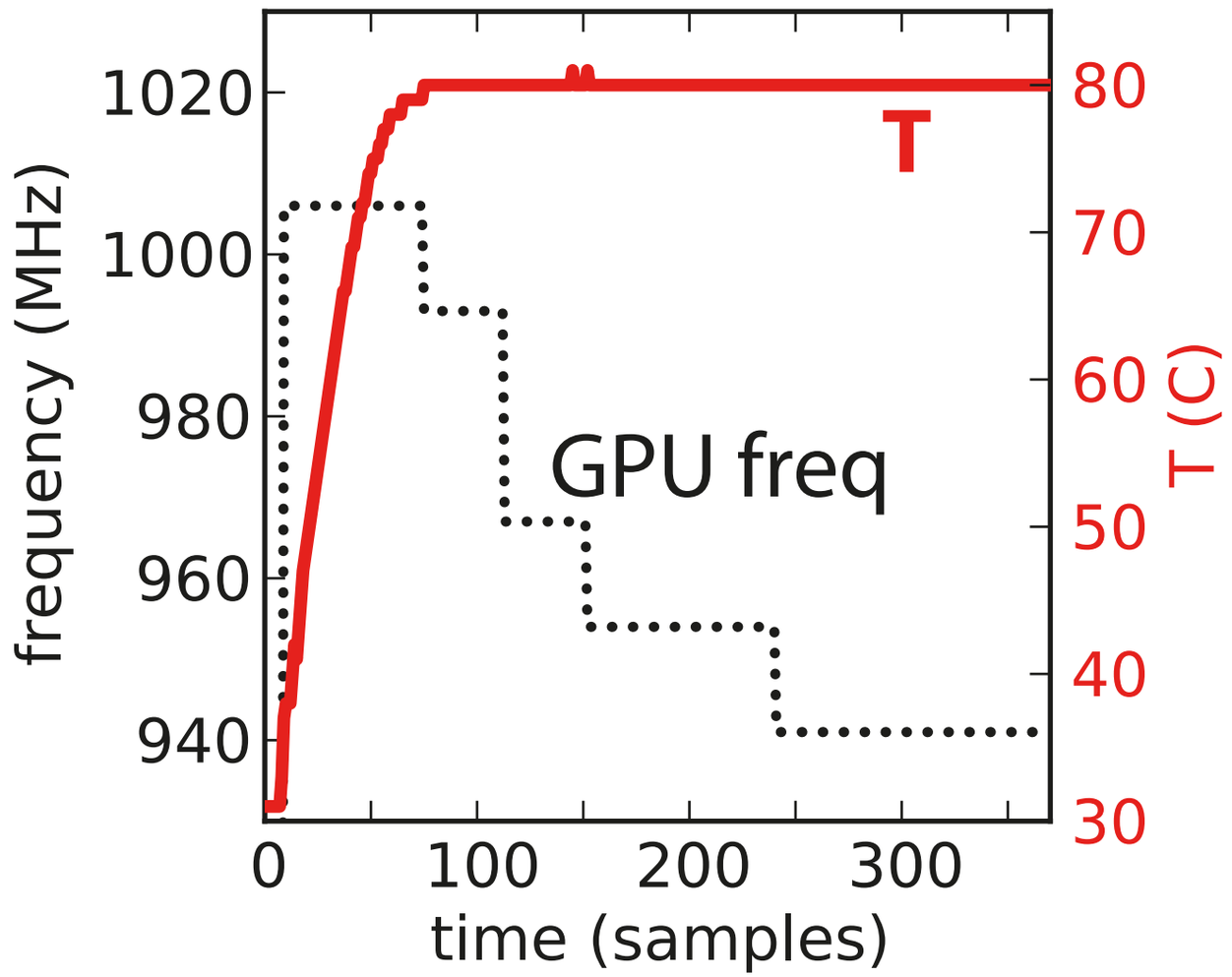
NVIDIA model	GPU memory checker[13]	# of cards tested	# memtest iterations	# cards with errors
GTX 580	`memtestG80`	1	$10,000$	–
GTX 680	`memtestG80`	50	$4,500$	–
GTX 770	`memtestG80`	100	$4,500$	–
GTX 780	`memtestCL`	1	$50,000$	–
GTX Titan	`memtestCL`	1	$50,000$	–
GTX 780Ti	`memtestG80`	70	$4 \times 10,000$	6
GTX 980	`memtestG80`	4	$4 \times 10,000$	–
GTX 980^{+}	`memtestG80`	70	$4 \times 10,000$	2

- newer GTX 1060/70/80 GPUs seem to have comparable error rates

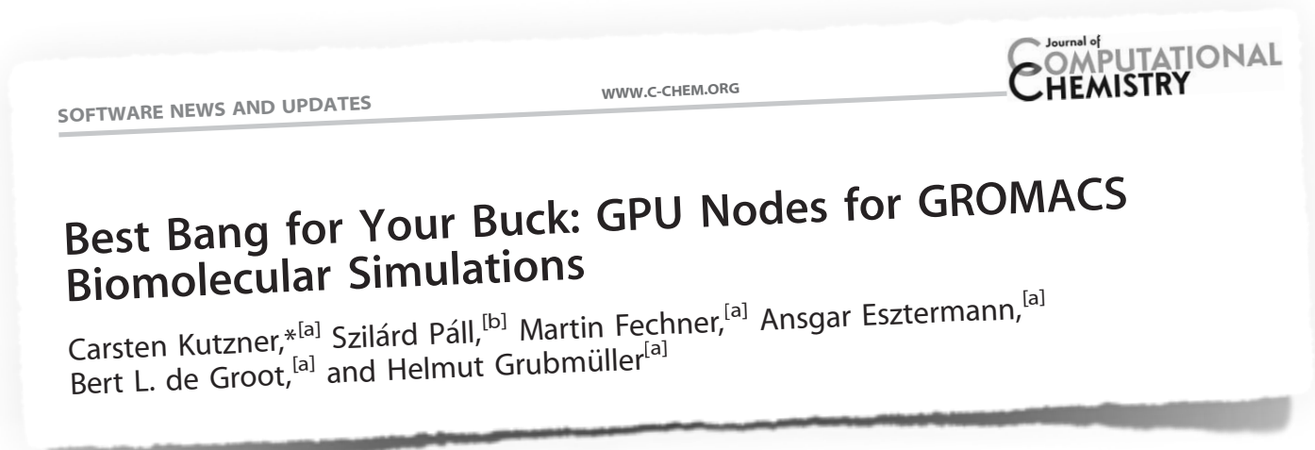
(13) I. S. Hague, V. S. Pande, In 10th IEEE/ACM International conference on cluster, cloud and grid computing: Stanford University, 2010

GPU FREQUENCY THROTTLING

- Consumer GPUs are optimized for acoustics:
- their fan speed is limited to 60% of max
- they reduce GPU frequency if too hot
- affects performance!
- see suppl. for how to fix GPU fan speed



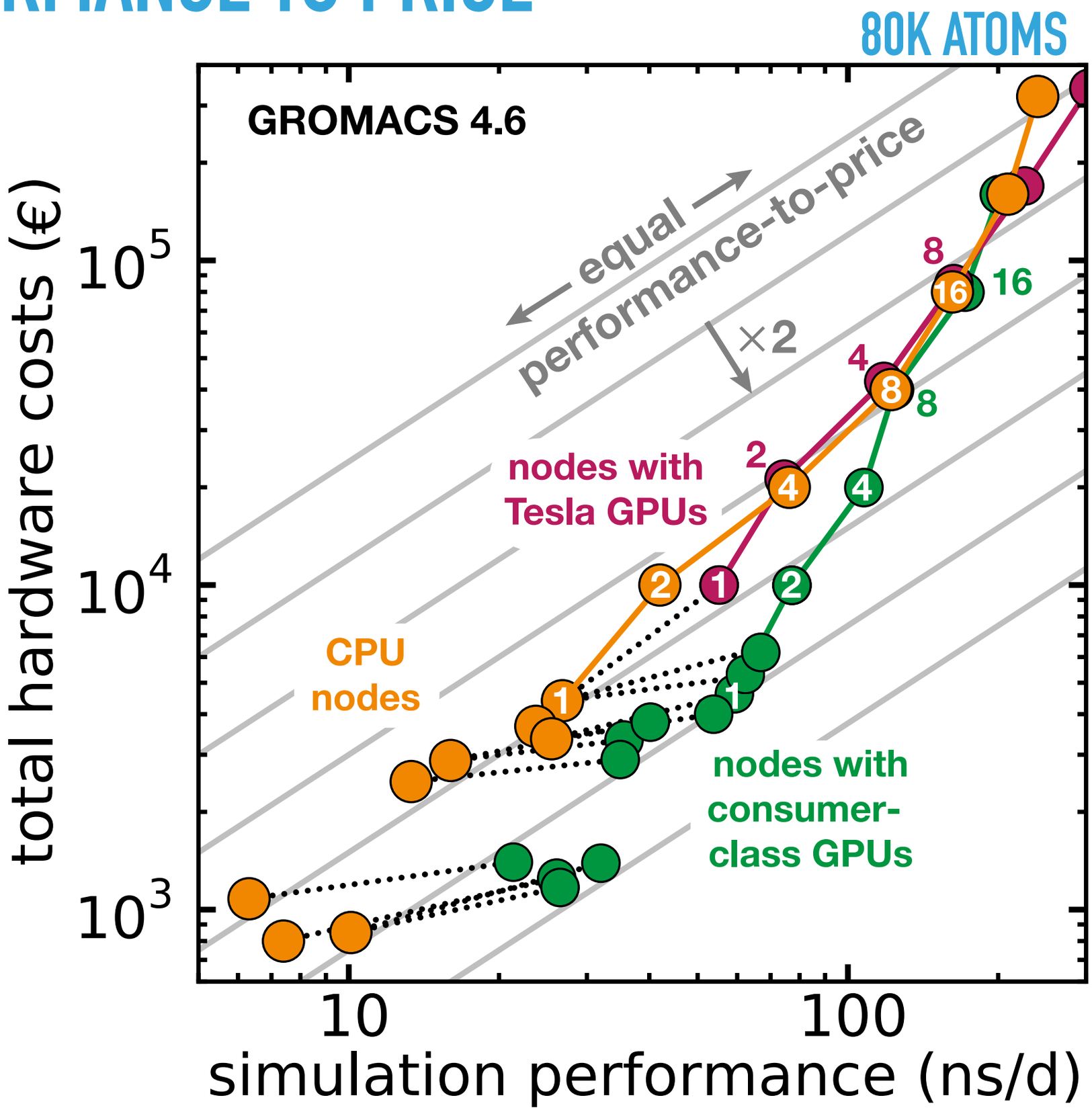
GeForce GTX TITAN


SOFTWARE NEWS AND UPDATES WWW.C-CHEM.ORG Journal of COMPUTATIONAL CHEMISTRY

Best Bang for Your Buck: GPU Nodes for GROMACS Biomolecular Simulations

Carsten Kutzner,*[a] Szilárd Páll,[b] Martin Fechner,[a] Ansgar Esztermann,[a] Bert L. de Groot,[a] and Helmut Grubmüller[a]

PERFORMANCE TO PRICE



PERFORMANCE TO PRICE

GROMACS 4.6

2M ATOMS

80K ATOMS

hardware costs (€)

performance (ns/d)

10^4

10^3

1

10

2 E5-2680v2 +4x 980

2 E5-2680v2 +2x 980

2 E5-2670v2 +2x 770

2 E5-2680v2

4 AMD 6272

2 E5-2670v2

2 AMD 6380

E5-2670v2

E3-1270v2 + 770

E3-1270v2

i7-5820K

i7-4770K

2 E5-2670v2 +2x 780Ti

2 E5-2680v2 +2 K20X

2 AMD 6380 +2x 980

2 AMD 6380 + 980

E5-2670v2 + 980

i7-5820K + 980

i7-5820K + 770

i7-4770K + 980

2x as good

2x E5-2680v2

4 AMD 6272

2 E5-2670v2

E5-2670v2

2 AMD 6380

E3-1270v2 + 770

E3-1270v2

i7-5820K

i7-4770K

E5-1620 + 980

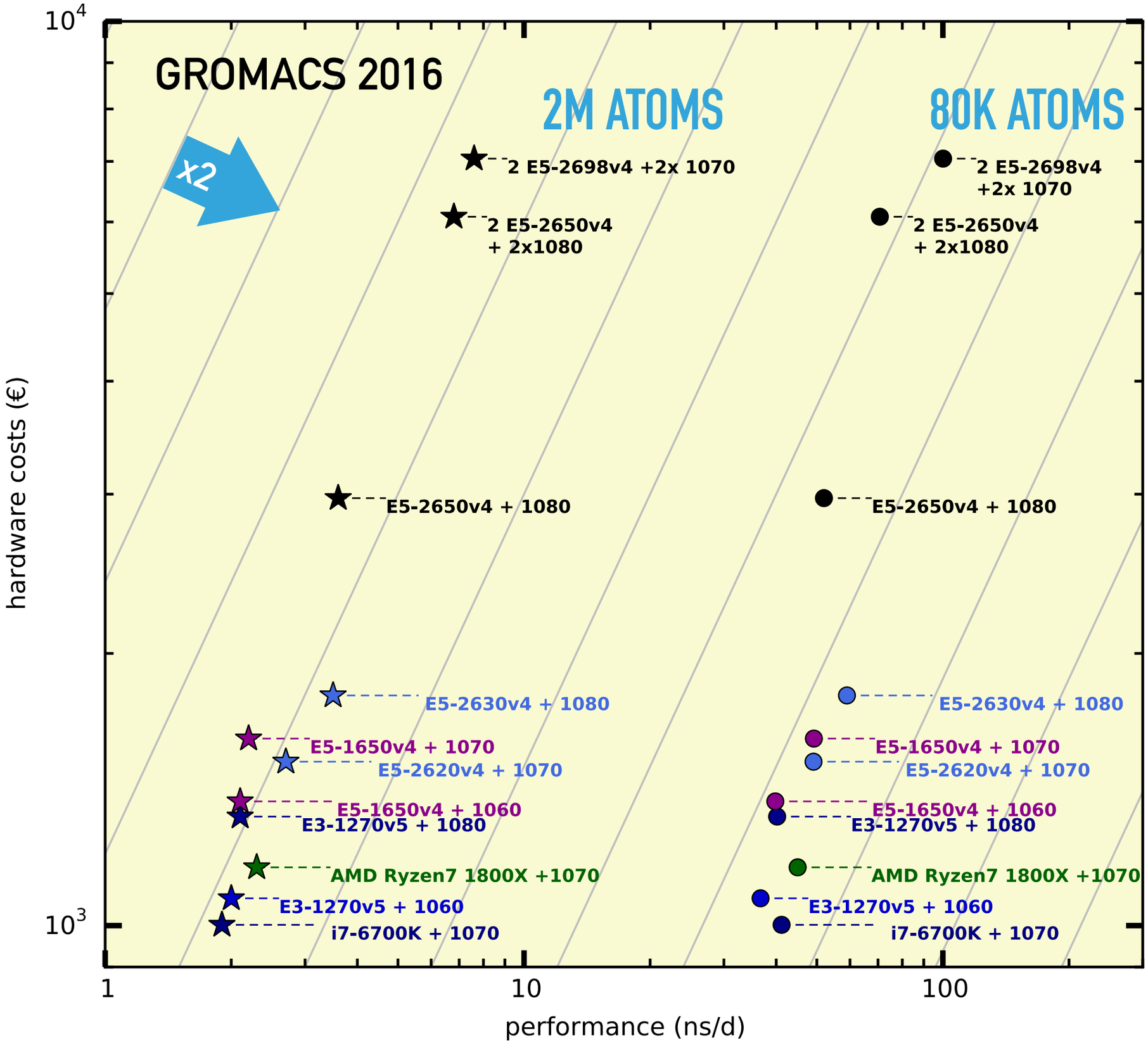
i7-5820K + 980

i7-4770K + 980

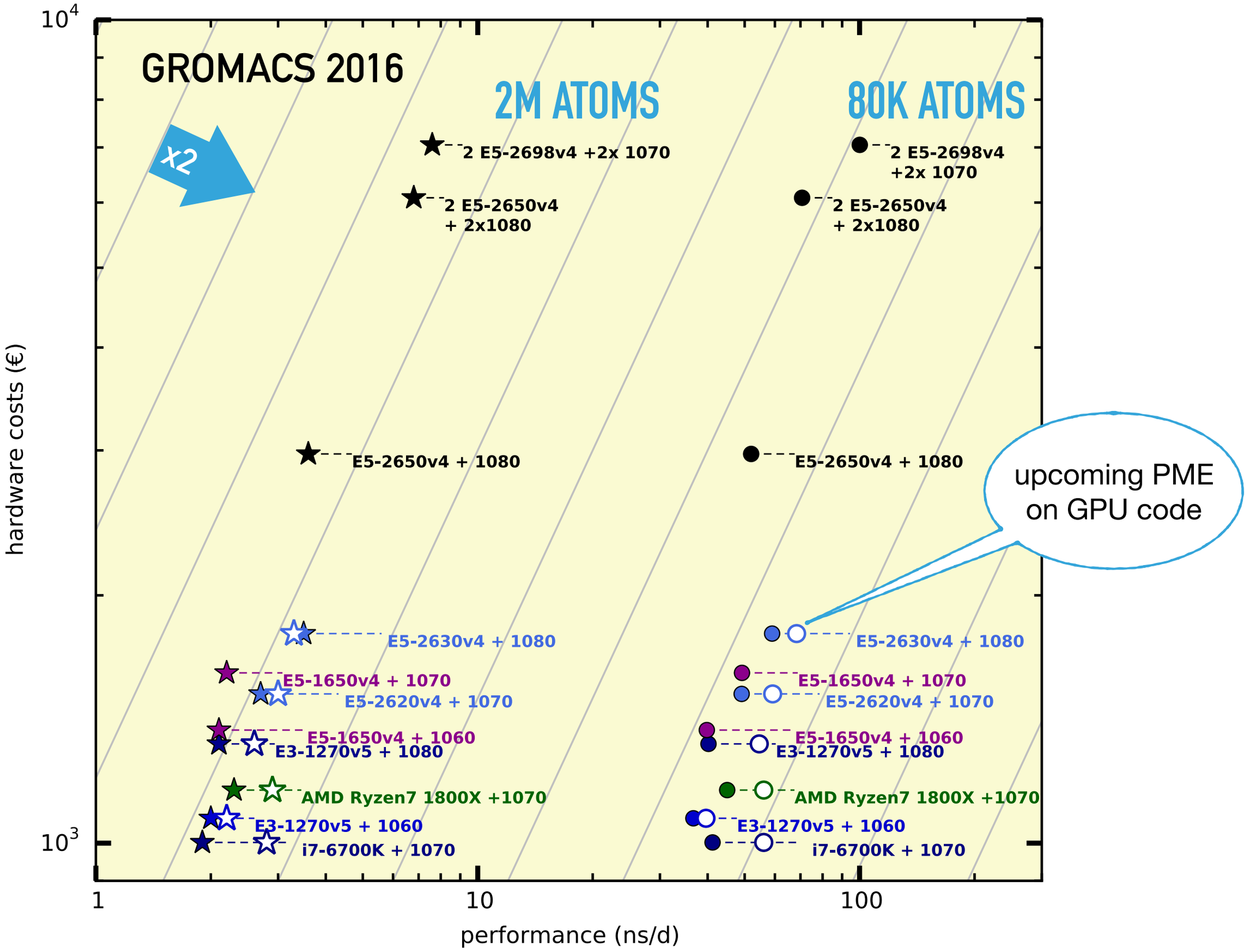
i7-5820K

adding the first GPU yields the largest performance benefit

PERFORMANCE TO PRICE 2017

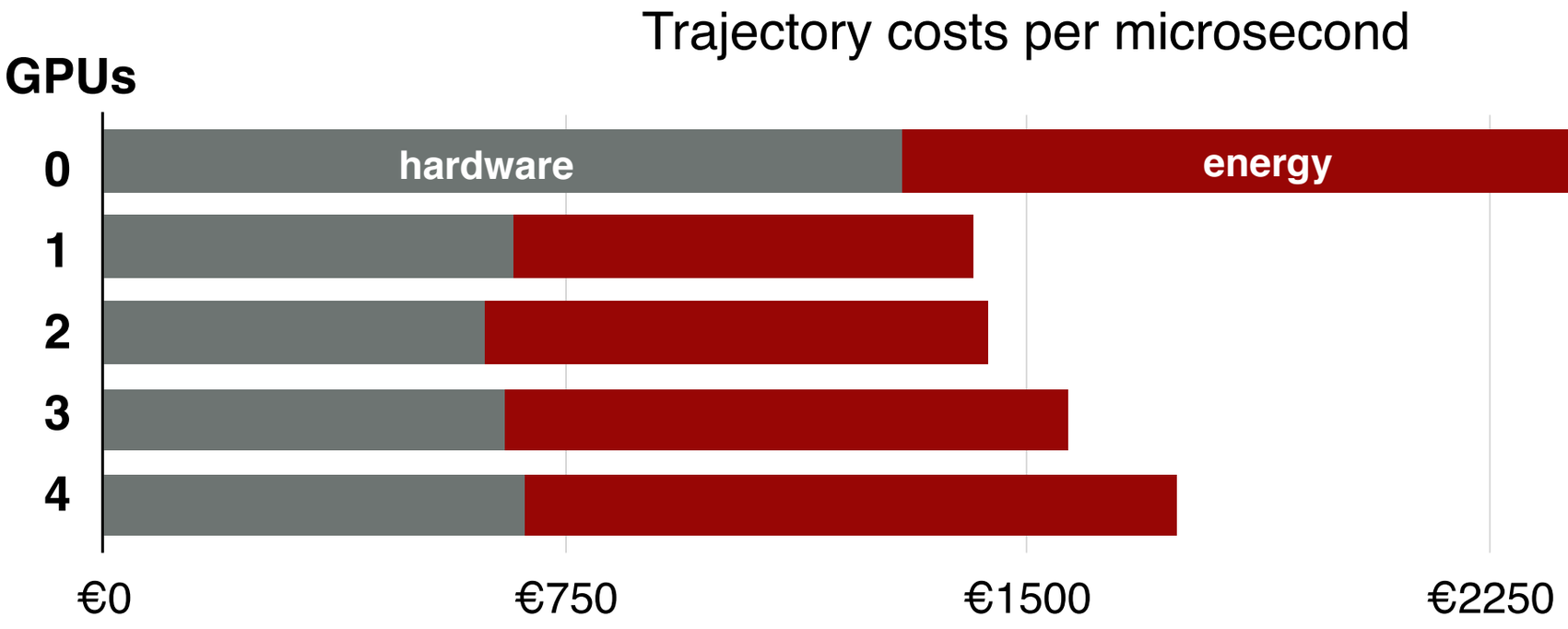


PERFORMANCE TO PRICE 2017



ENERGY EFFICIENCY

- Over cluster lifetime, energy costs become comparable to hardware costs
- assuming 5 yr of operation and 0.2 EUR / kWh (incl. cooling)

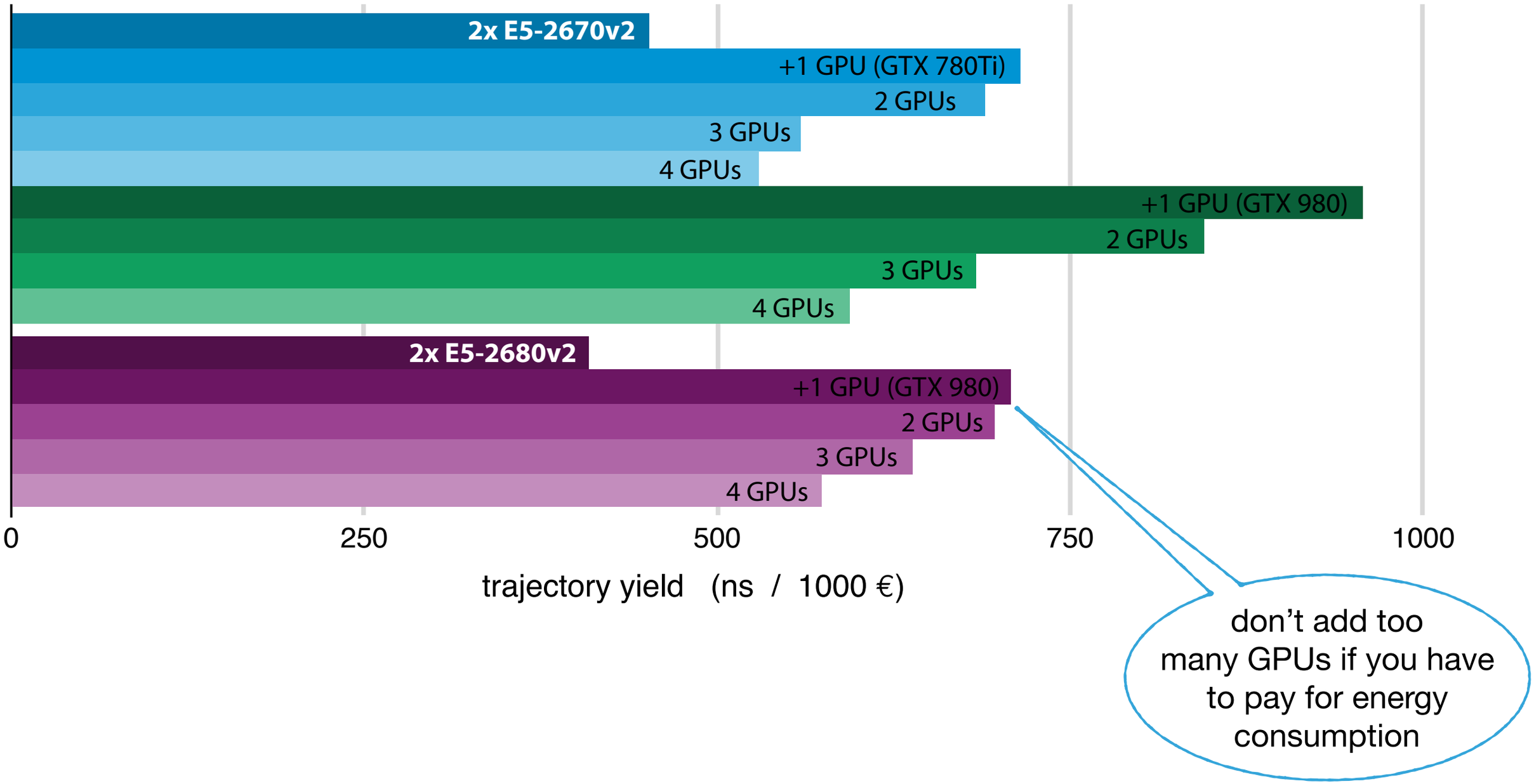


2x E5-2680v2 (2x 10 core) with GTX 980 GPUs, RIB benchmark

- balanced CPU/GPU resources keep energy costs low

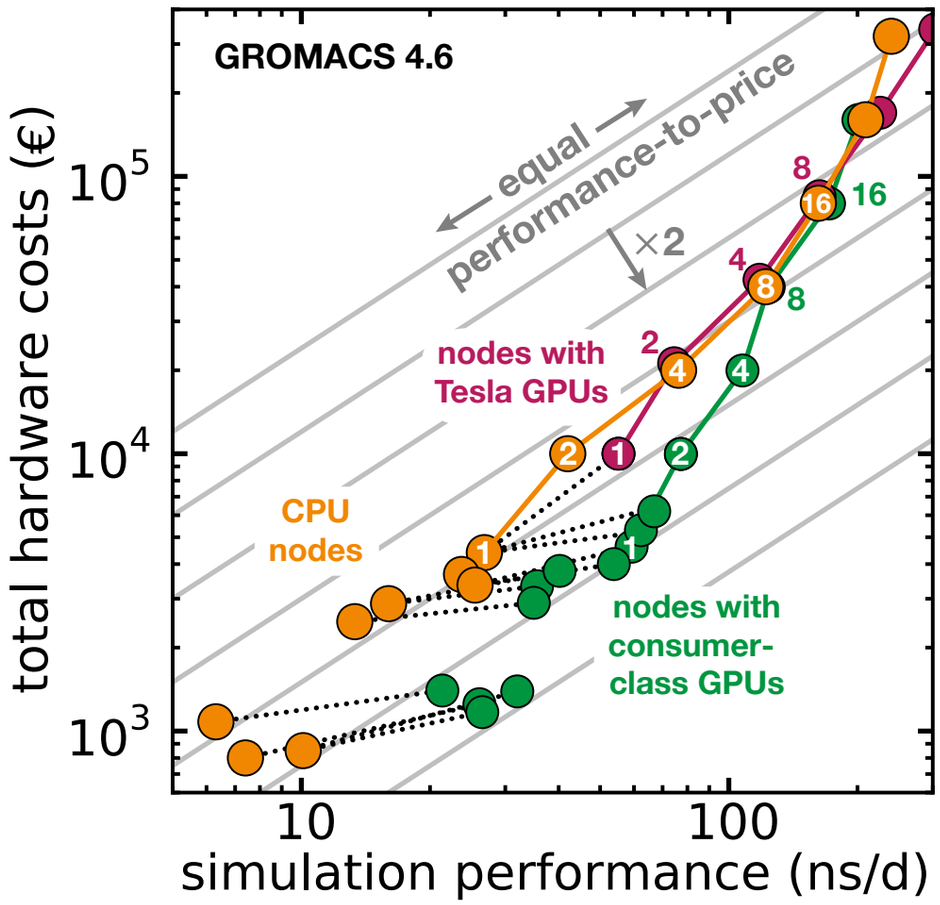
ENERGY EFFICIENCY

- Fixed budget trajectory yield taking into account energy + cooling (0.2 EUR / kWh) RIB



CONCLUSIONS

- buying dedicated MD nodes boosts the performance to price ratio
- Nodes with 1–2 consumer-class GPUs produce >2x as much trajectory as CPU nodes or nodes with “professional” Tesla GPUs
- consumer GPUs with memory errors can be replaced, GPU throttling can be prevented by proper ventilation
- Energy efficiency can be optimized by balancing the GPU to CPU compute power
- upcoming PME-GPU code further enhances performance to price ratio, as it allows for cheaper CPUs



THANKS FOR YOUR ATTENTION!

PEOPLE INVOLVED

Martin Fechner, Szilard Pall,
Timo Graen, Ansgar Esztermann,
Markus Rampp, Aleksei Yupinov,
Bert L de Groot, Helmut Grubmüller