Parallel Approaches to Longtime Atomistic Simulations

Danny Perez

CIRM, Marseille, France

NNSA
National Nuclear Security Administration
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

Acknowledgements

Collaborator:

Art Voter (Los Alamos National Lab)

Richard Zamora (Argonne National Lab)

Rao Huang (Xiamen University)

Funding:

Los Alamos NATIONAL LABORATORY EST. 1943

U.S. DEPARTMENT OF ENERGY

Office of Science

ECP EXASCALE COMPUTING PROJECT

Computing:

Los Alamos NATIONAL LABORATORY EST. 1943

NERSC

Molecular Dynamics

- Molecular Dynamics, the numerical integration of atomistic equations of motion, is the gold standard to investigate the dynamical evolution of atomistic systems
- Essentially a numerical experiment
- MD can in principle be used to compute any atomistic dynamic or thermodynamic property**, if you can afford it…**

What can you simulate with MD?

- **Baseline:**
  - –Force calculation: 3.6 μs/atom/core (~EAM)
  - –dt=1 fs

#atoms	Peak simulation rate (/hour)
10^3	1 ns
10^6	1 ps
10^9	1 fs

What can you simulate with MD?

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , **IBM** DOE/SC/Oak Ridge National Laboratory **United States**	2,282,544	122,300.6	187,659.3	,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , **NRCPC** National Supercomputing Center in Wuxi **China**	10,649,600	93,014.6	125,435.9	5,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , **IBM** DOE/NNSA/LLNL **United States**	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , **NUDT** National Super Computer Center in Guangzhou **China**	4,981,760	61,444.5	100,678.7	8,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , **Fujitsu** National Institute of Advanced Industrial Science and Technology (AIST) **Japan**	391,680	19,880.0	32,576.6	,649
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) **Switzerland**	361,760	19,590.0	25,326.3	,272

What can you simulate with MD?

- Ideal scaling baseline:

#atoms	Peak simulation rate (/hour)
10^3	1 ms
10^6	1 μs
10^9	1 ns

What can you simulate with MD?

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , **IBM** DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , **NRCPC** National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , **IBM** DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , **NUDT** National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , **Fujitsu** National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , **Cray Inc.** Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272

Plan

- **Today:** What is the brute-force baseline for MD?

- Five main classes of solutions:
  - **Decomposition**
  - **Replication**
  - **Bookkeeping**
  - **Speculation**
  - Localization

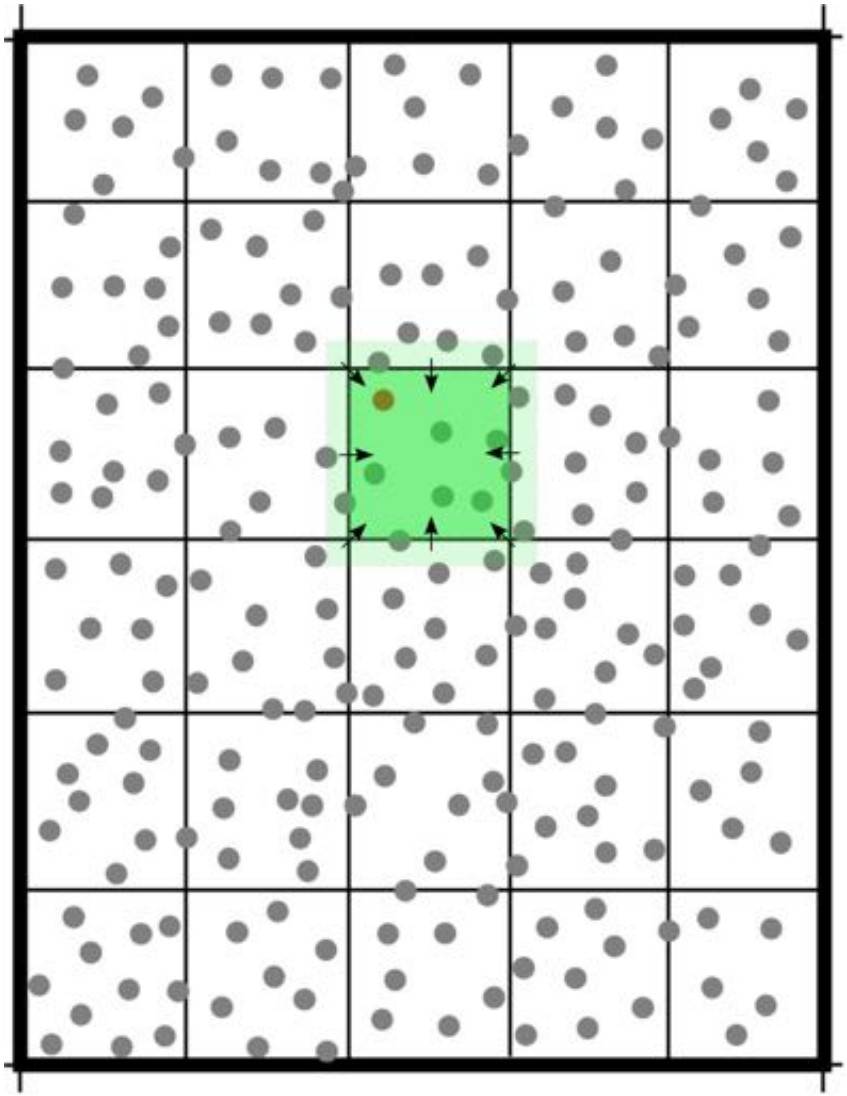
Parallel Trajectory Splicing (ParSplice)

[Perez, Cubuk, Waterland, Kaxiras, Voter, JCTC 12, 18 (2016)]

- **We concentrate on MD-like methods (e.g., goal is to generate a single, dynamically-accurate trajectory)**

Decomposition

- For short-range interactions, natural approach is domain decomposition
- Each core computes forces in its domain
- Communication required at every timestep
- Efficient if computation>>communication



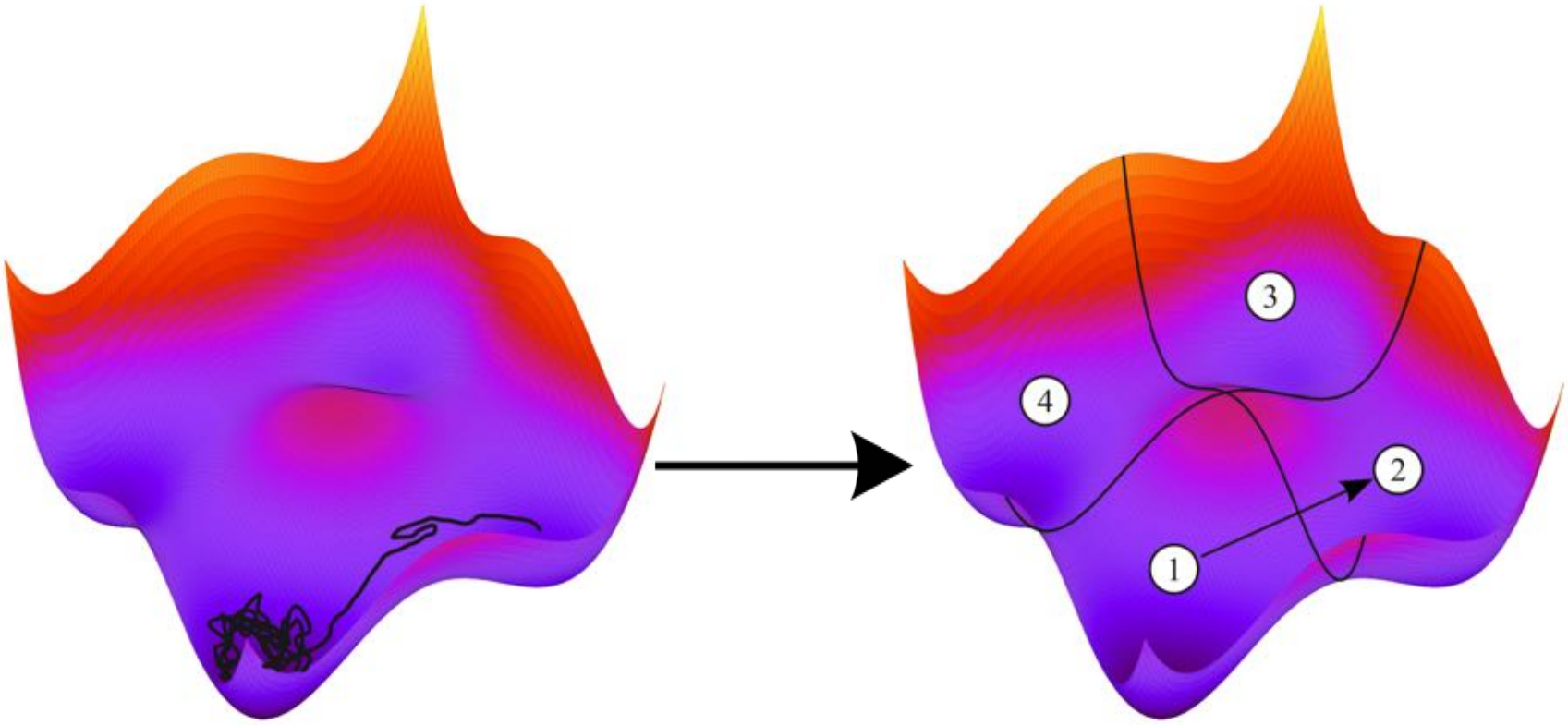
Decomposition

- For EAM: ~10^3 atoms per core to maintain efficiency

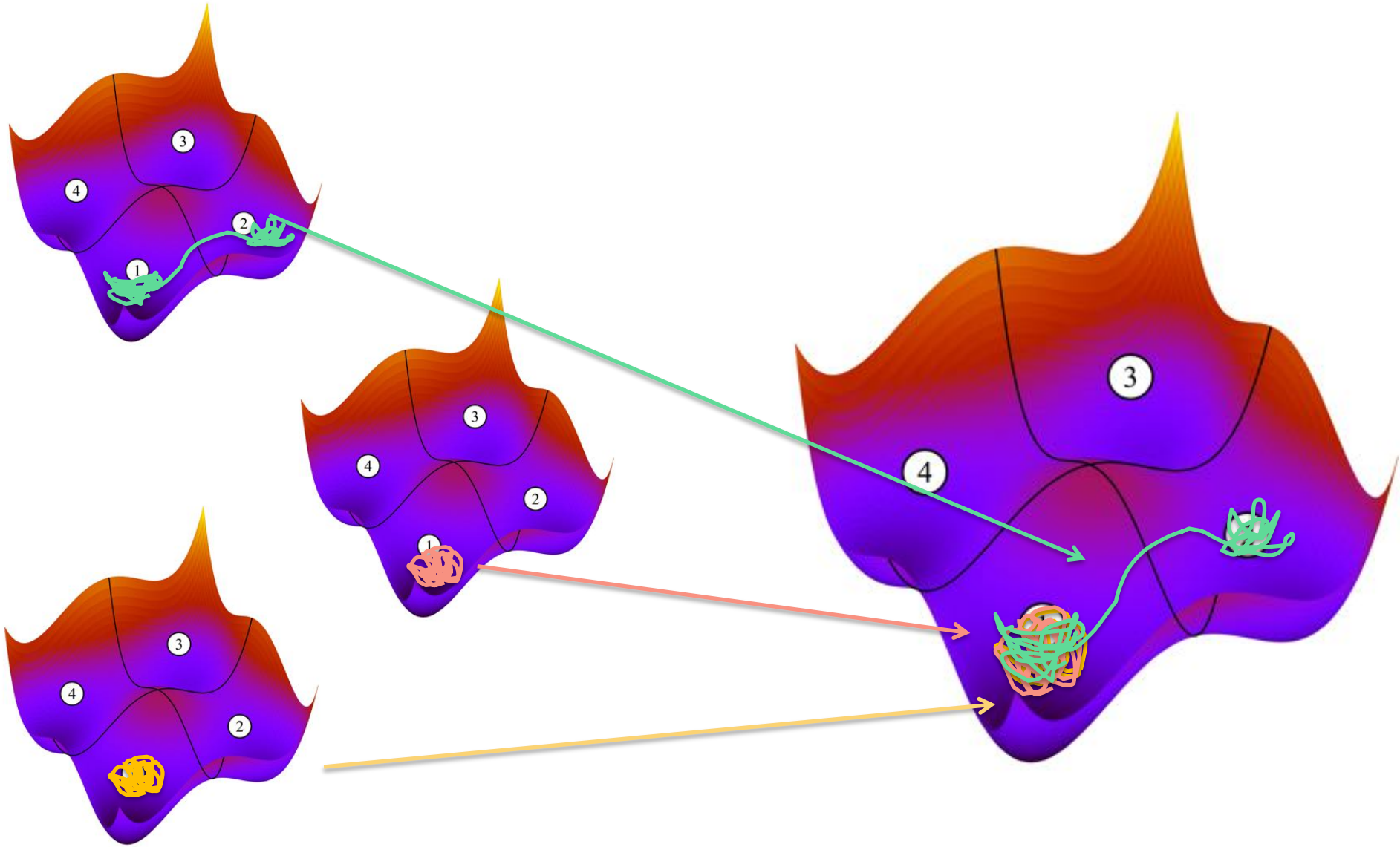
#atoms	Actual simulation rate (/hour)	Peak simulation rate (/hour)
10^3	1 ns	1 ms
10^6	1 ns	1 μs
10^9	1 ns	1 ns

State-to-state dynamics

Assume that we are satisfied by coarser description of the trajectory in terms of a ***state-to-state trajectory***.



Replication



Trajectory Splicing

Consider a Fokker-Planck operator with absorbing boundaries around a state. Let {-λ} be its eigenvalues.

λ_1 : quasi-stationary inter-state transition rate

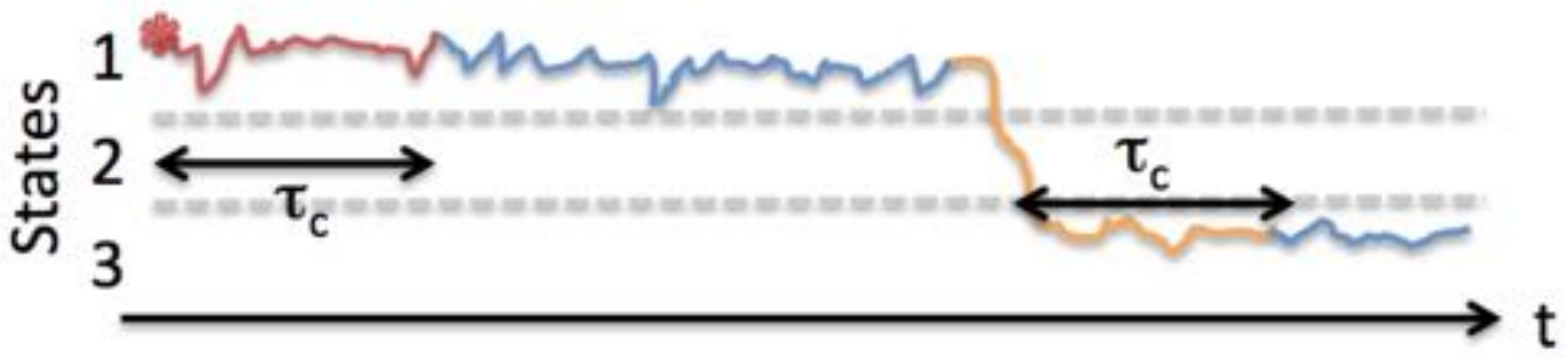
λ_2 : slowest intra-state relaxation rate

- The system looses its memory at a rate $(\lambda_2 - \lambda_1)$
- After a time $\tau_c > 1/(\lambda_2 - \lambda_1)$ without escaping, the system is (almost) quasi-stationary
- Quasi-stationary samples are statistically equivalent w.r.t. predicting state-to-state evolution

[Le Bris, Lelievre, Luskin, and Perez, MCMA 18, 119 (2012)]

Trajectory Splicing

Sections of trajectory that remained **within the same state** for at least τ_c before its beginning **and** before its end can be spliced together.

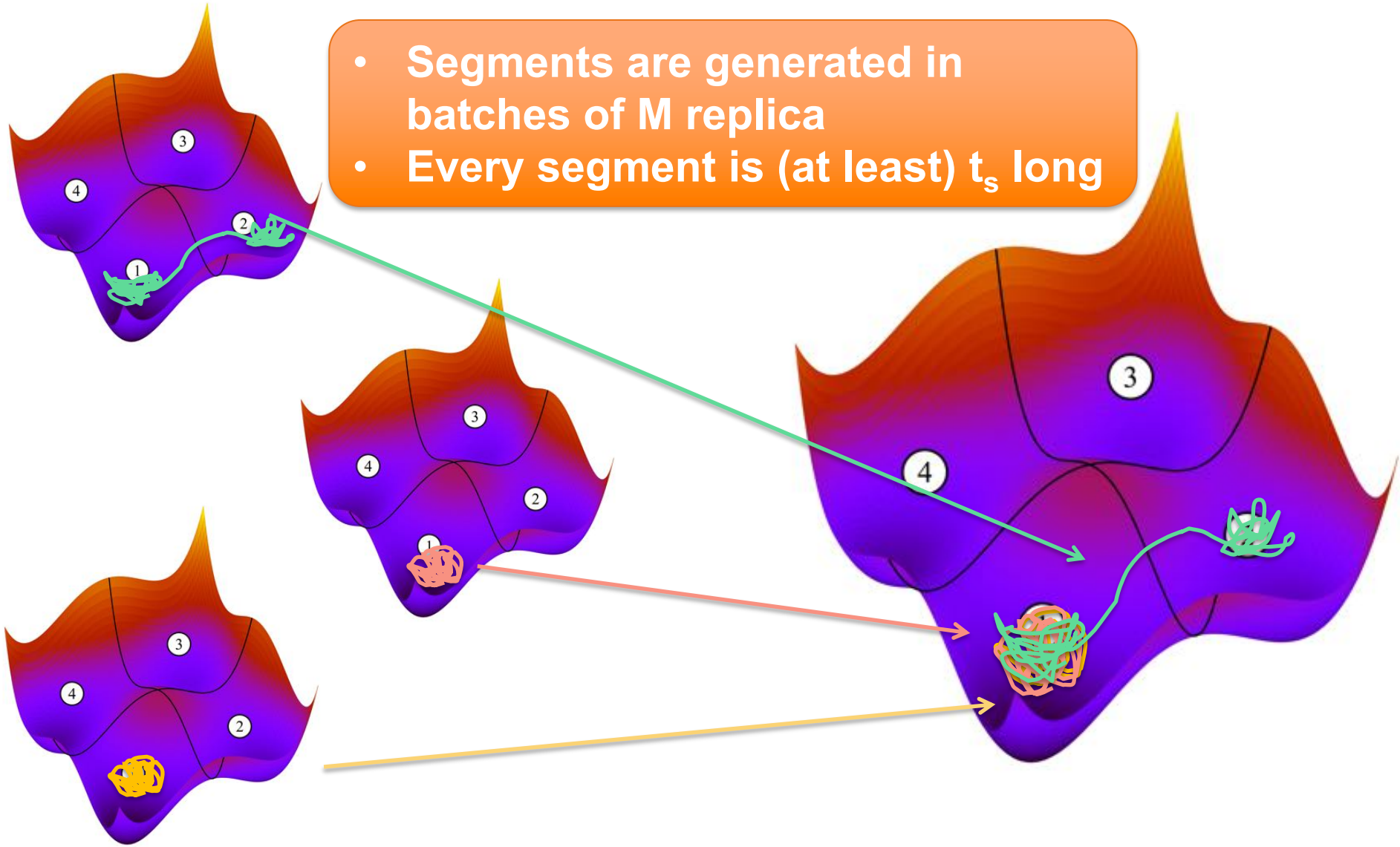


Finite τ_c only introduces $O(\exp[-(\lambda_2 - \lambda_1)\tau_c])$ errors!

[Le Bris, Lelievre, Luskin, and Perez, MCMA 18, 119 (2012)]

Parallel Trajectory Splicing

- Segments are generated in batches of M replica
- Every segment is (at least) t_s long



Baseline system

Assume:

- k=1/μs/atom
- t_c=1 ps
- t_s=1 ps

Decomposition + Replication

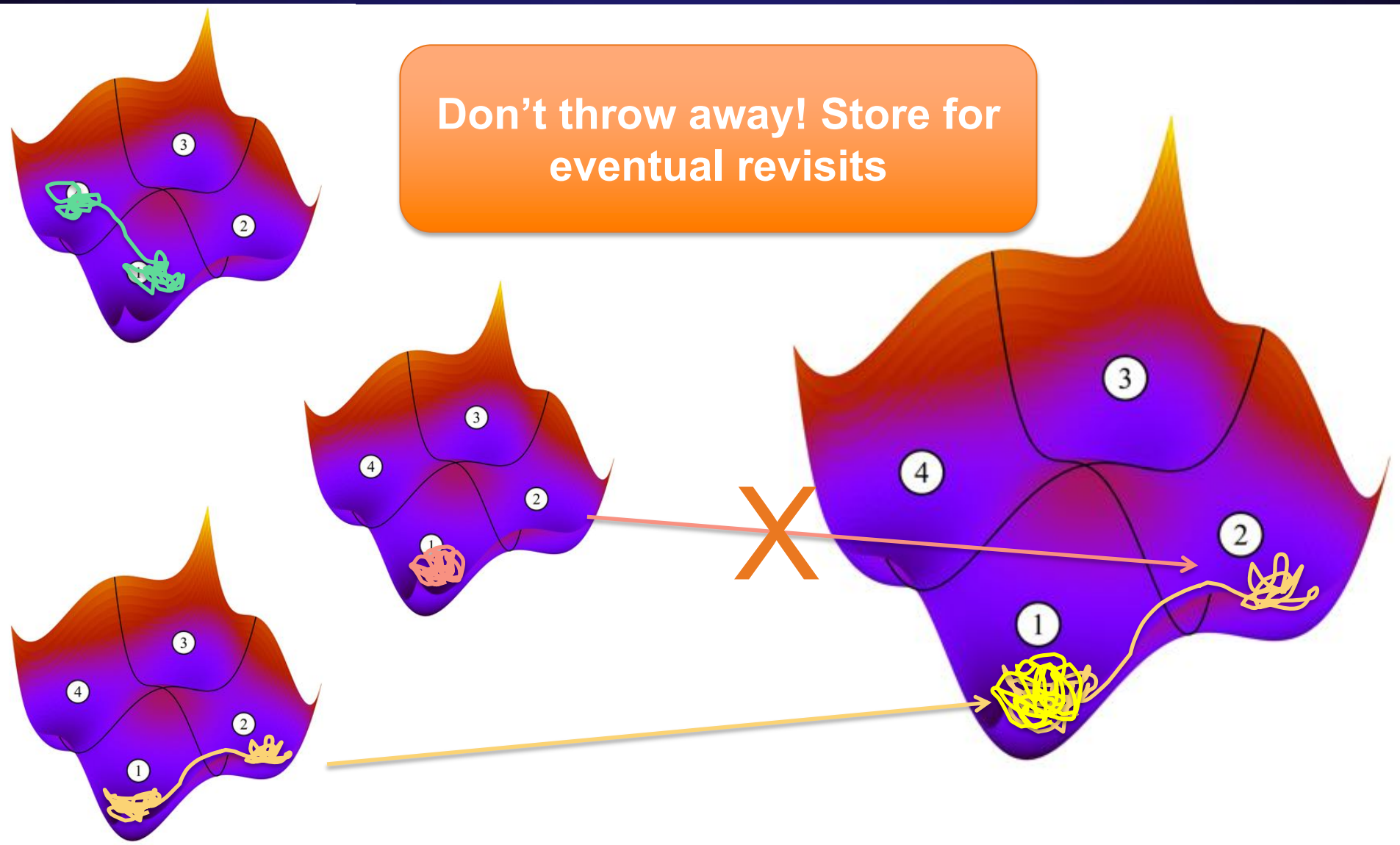
- Parallelize by decomposing first, and then replicate
- Parallel efficiency:

$$\frac{t_{trans}}{\left\lceil \frac{t_{trans}}{Mt_s} \right\rceil Mt_s + Mt_c}$$

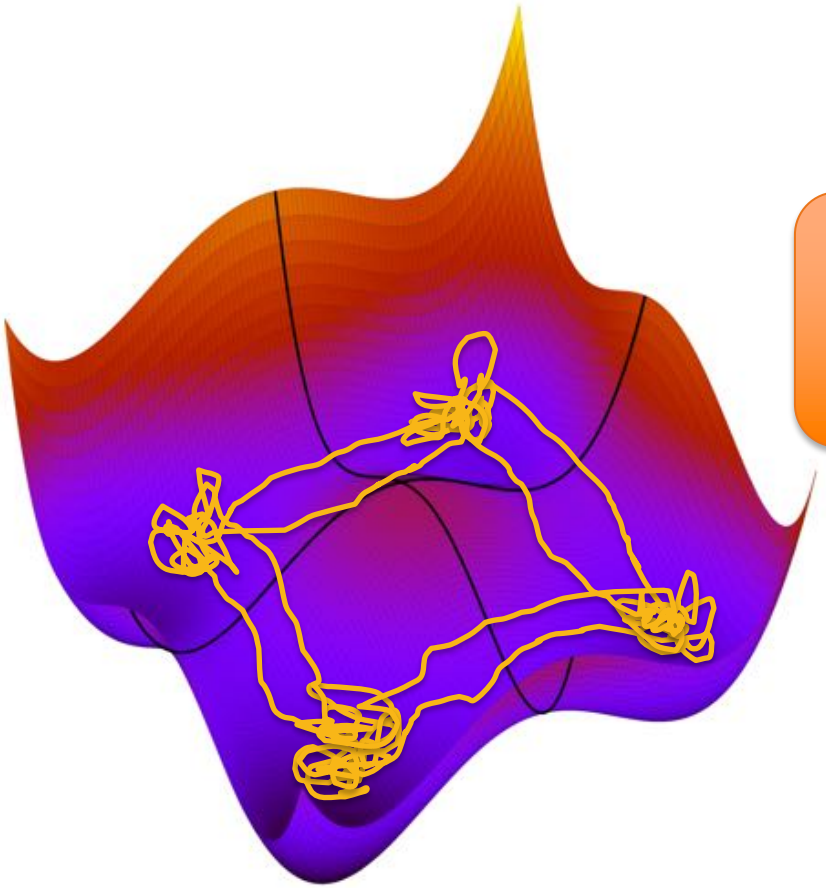
1000 atoms/core

#atoms	Actual simulation rate (/hour)	Peak simulation rate (/hour)
10^3	0.5 μs	1 ms
10^6	0.5 ns	1 μs
10^9	1 ns	1 ns

Bookkeeping



Super-basins



Revisits are extremely common!

Decomposition + Replication + Bookkeeping

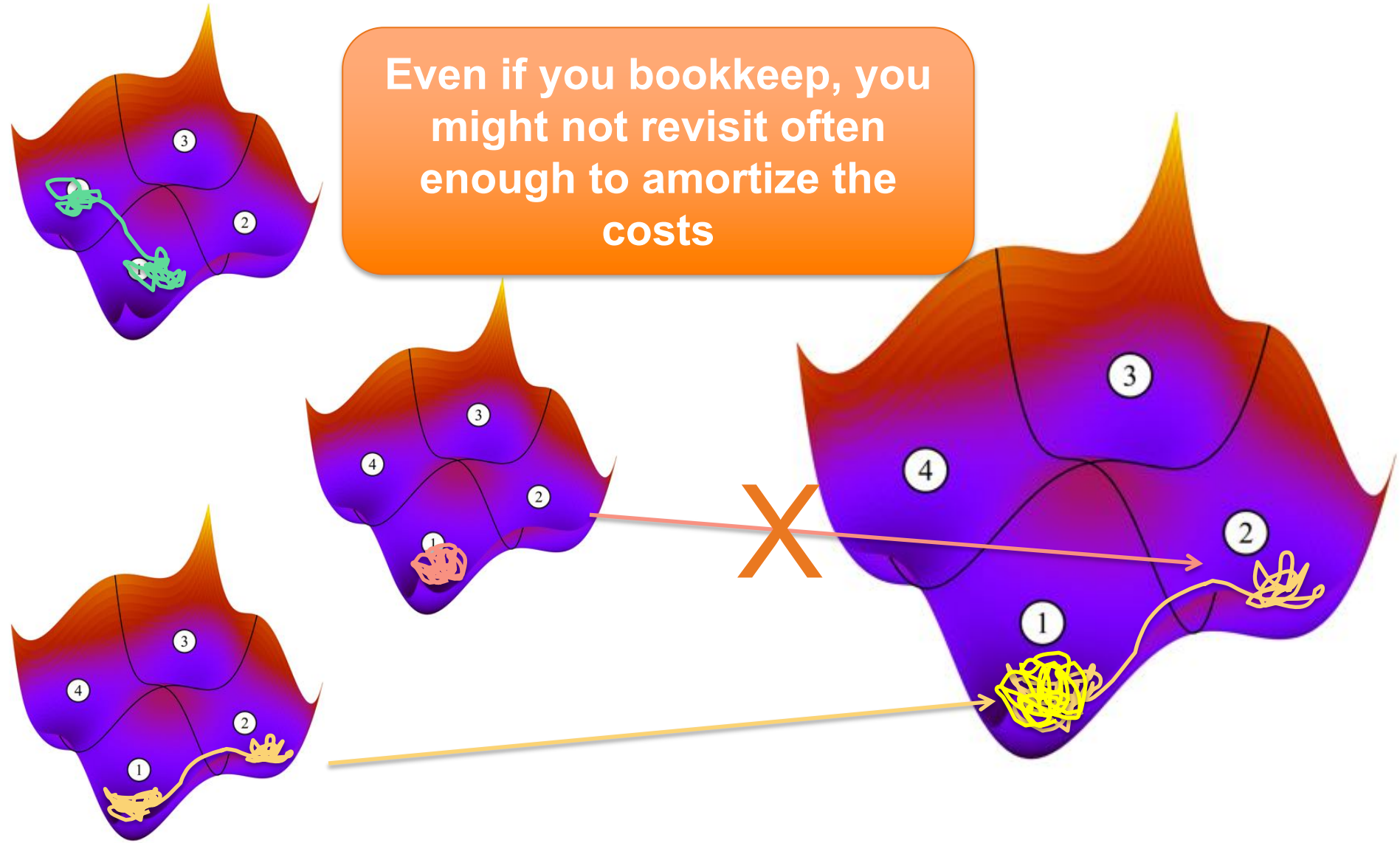
- You can potentially amortize all overhead except for t_c per replica per state, if you revisit often enough
- Parallel efficiency over **R** visits to a state:

$$\frac{Rt_{trans}}{\left\lceil \frac{Rt_{trans}}{Mt_s} \right\rceil Mt_s + Mt_c}$$

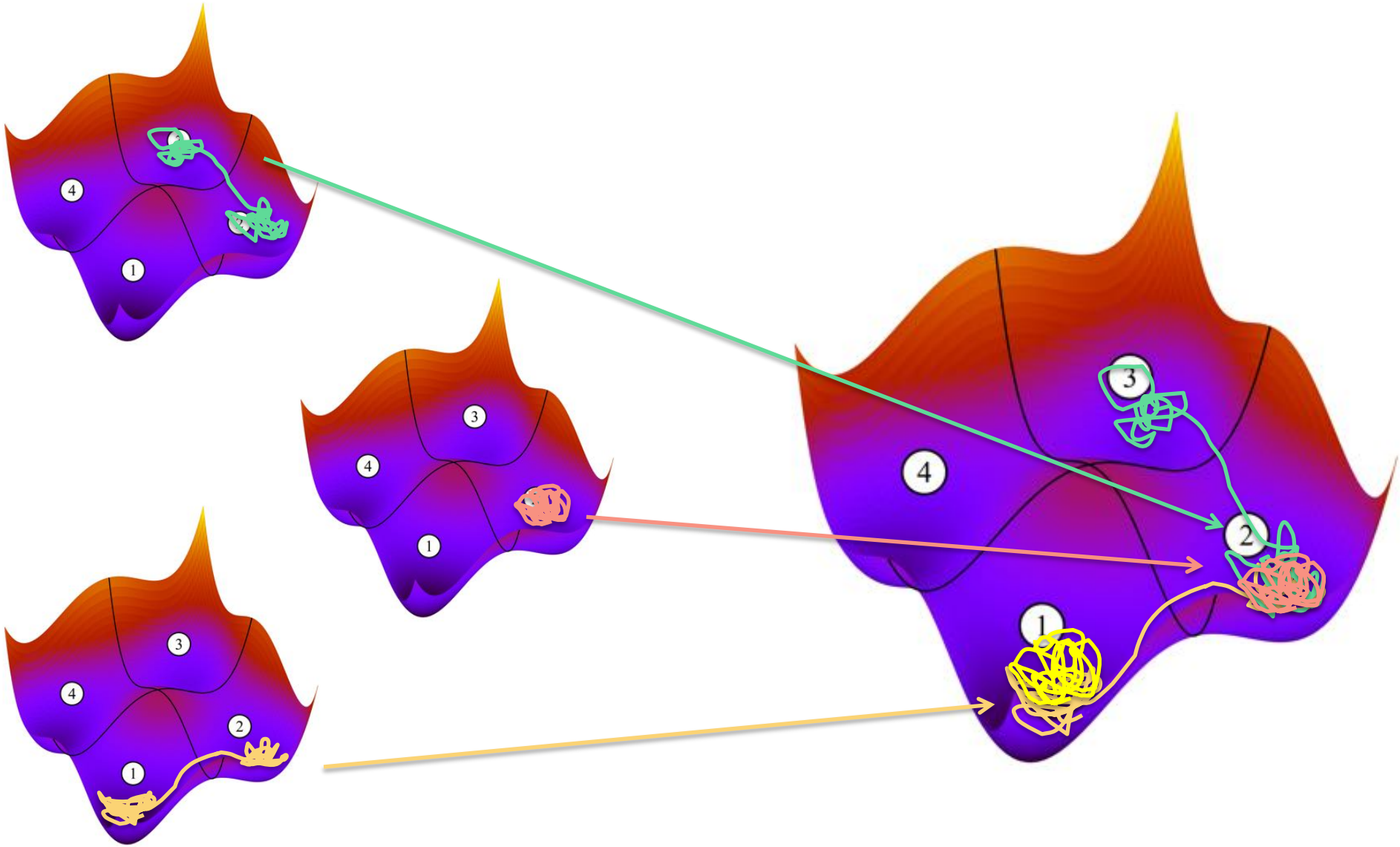
1000 atoms/core, **R=100**

#atoms	Actual simulation rate (/hour)	Peak simulation rate (/hour)
10^3	50 μs	1 ms
10^6	50 ns	1 μs
10^9	1 ns	1 ns

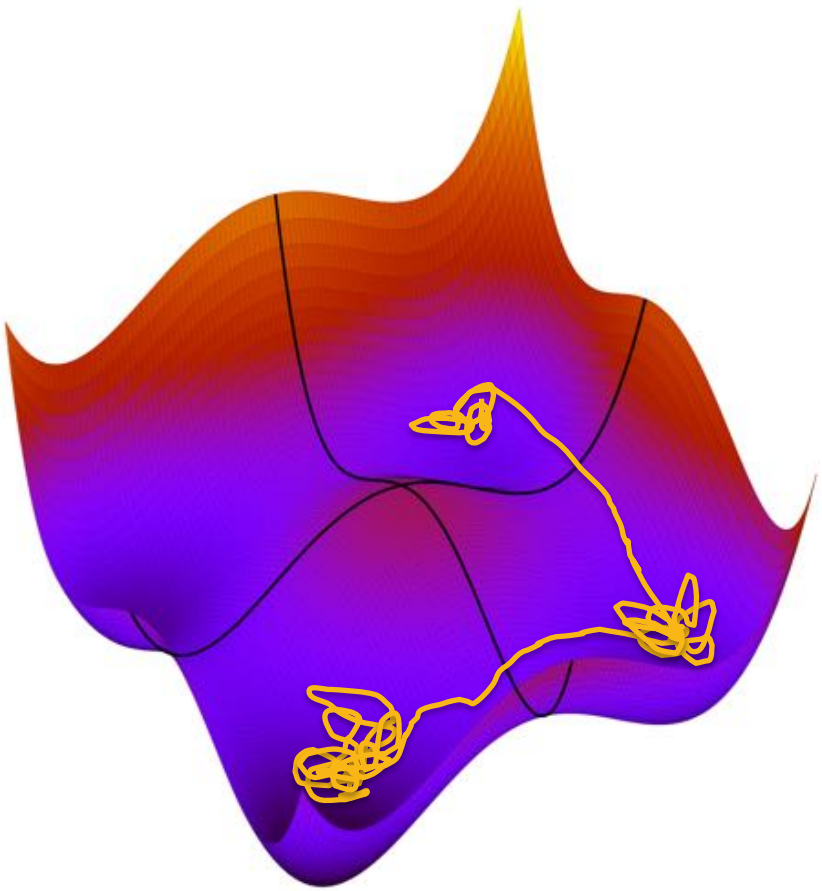
Speculation

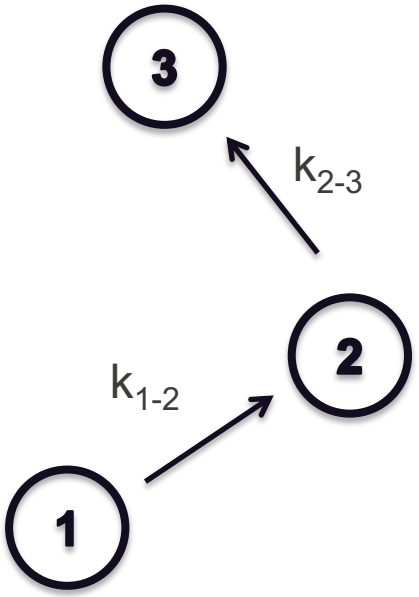


Speculation

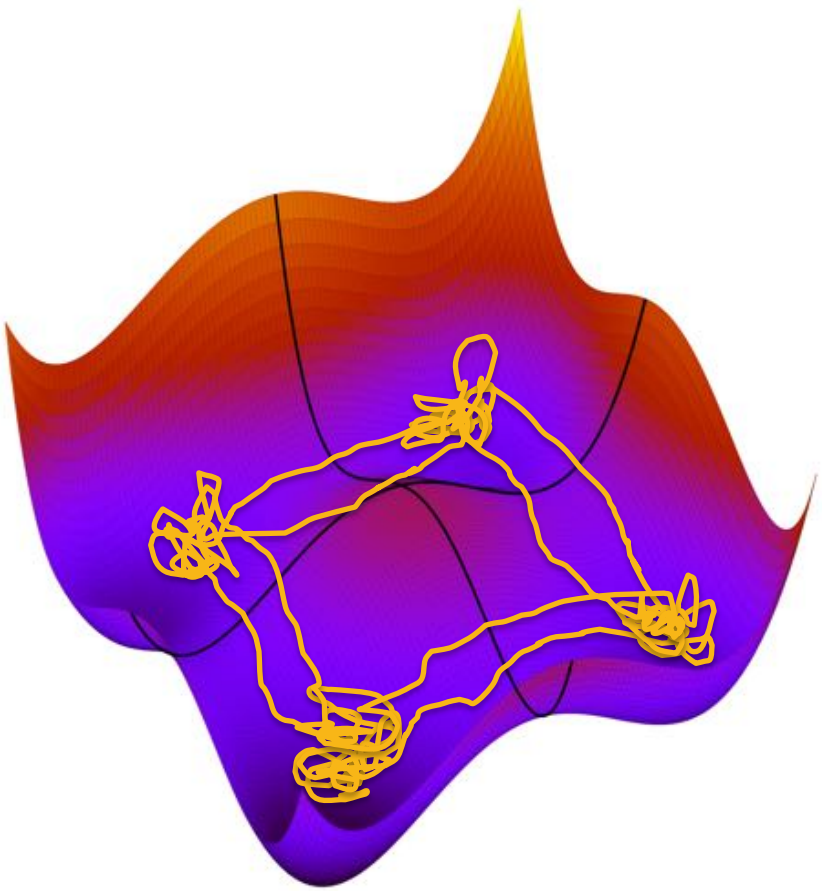


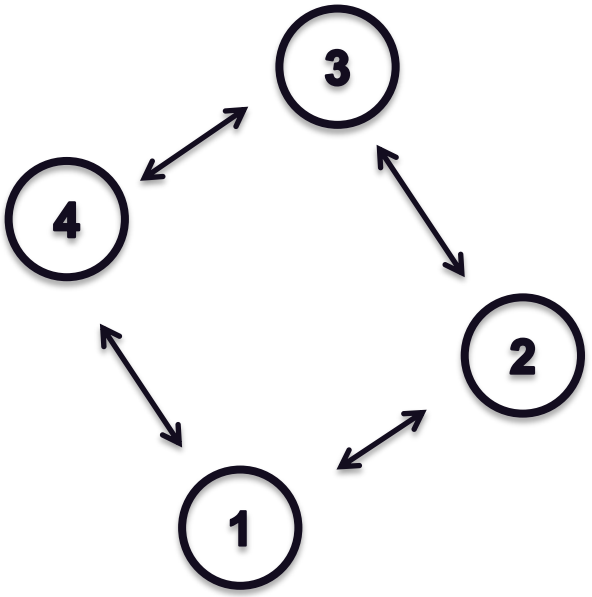
Statistical oracle





Statistical oracle





Decomposition + Replication + Bookkeeping + Speculation

- Akin to a smaller effective batch size: avoid generating excess segments by using speculation
- Parallel efficiency:

$$\frac{Rt_{trans}}{\left\lceil \frac{Rt_{trans}}{fMt_s} \right\rceil fMt_s + fMt_c}$$

1000 atoms/core, R=100, **f=0.25**

#atoms	Actual simulation rate (/hour)	Peak simulation rate (/hour)
10^3	0.6 ms	1 ms
10^6	0.8 μs	1 μs
10^9	1 ns	1 ns

Benchmarks: Metallic nanoparticles

- Voter-Chen EAM Pt
- N=147
- Initial state: quenched liquid

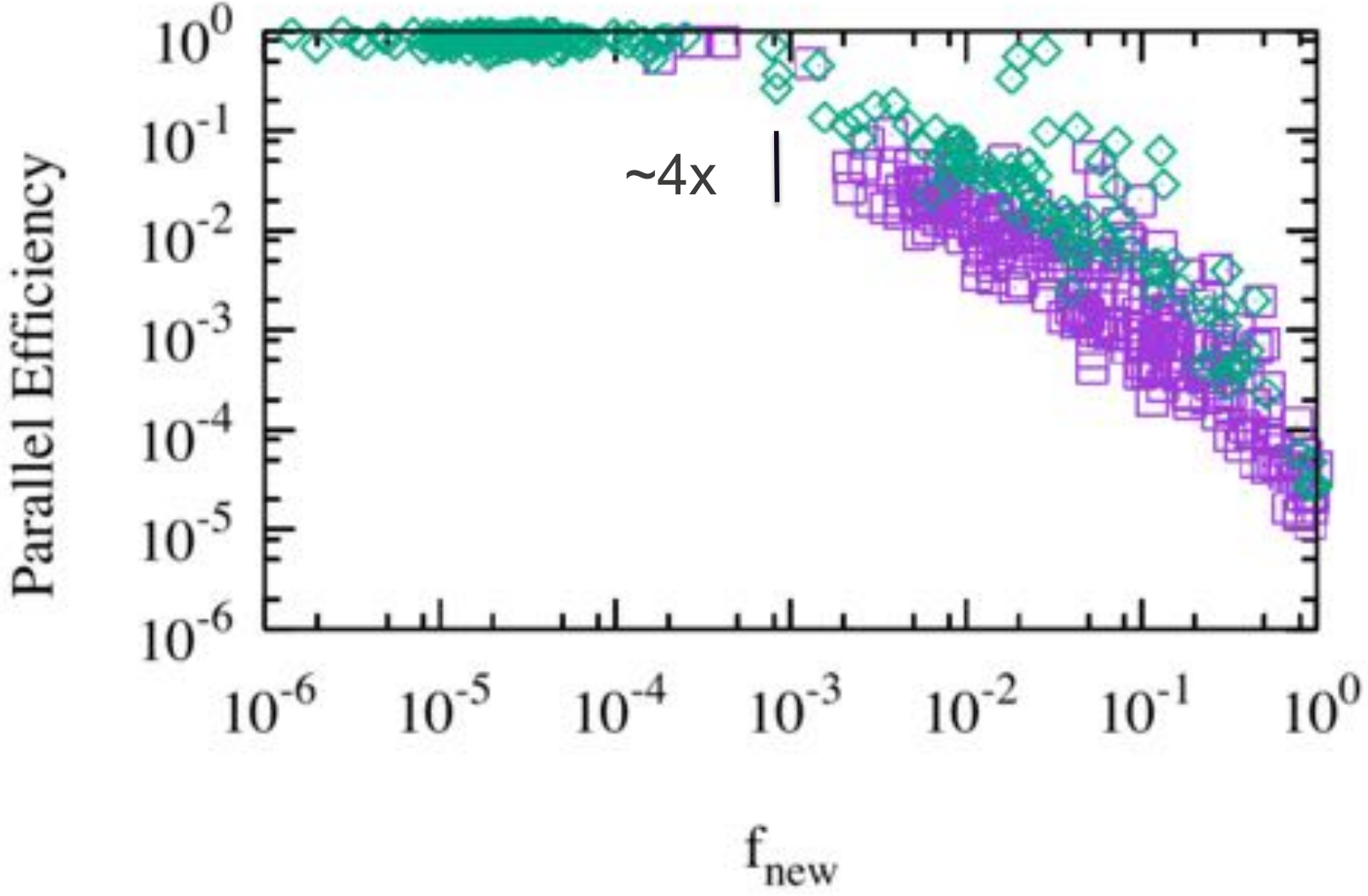
- 3600 workers for 4h on NERSC/Cori

[Huang, Voter, Perez, JCP 147, 152717 (2017)]
[Perez, Huang, Voter, JMR 33, 813 (2018)]

Benchmark results

T (K)	Trajectory length (ps)	Generated segment time (ps)	#Transitions	#States	$\langle t_{trans}/M\ t_c \rangle$	<R>	Simulation rate (μs/hour)
300	340,699,441	341,011,427	2,227	15	21.25	149	85
400	267,608,621	305,631,041	21,629,711	4,785	0.0017	4545	69
600	194,178,176	269,536,048	90,096,511	24,161	0.00029	3846	48
700	70,212,784	228,559,866	33,780,937	250,867	0.00028	135	17
800	1,754,393	220,695,348	738,292	36,613	0.00033	20	0.45
900	169,943	10,673,302	64,208	11,577	0.00030	6	0.043

Speculation



600K: no speculation, speculation

Benchmark results

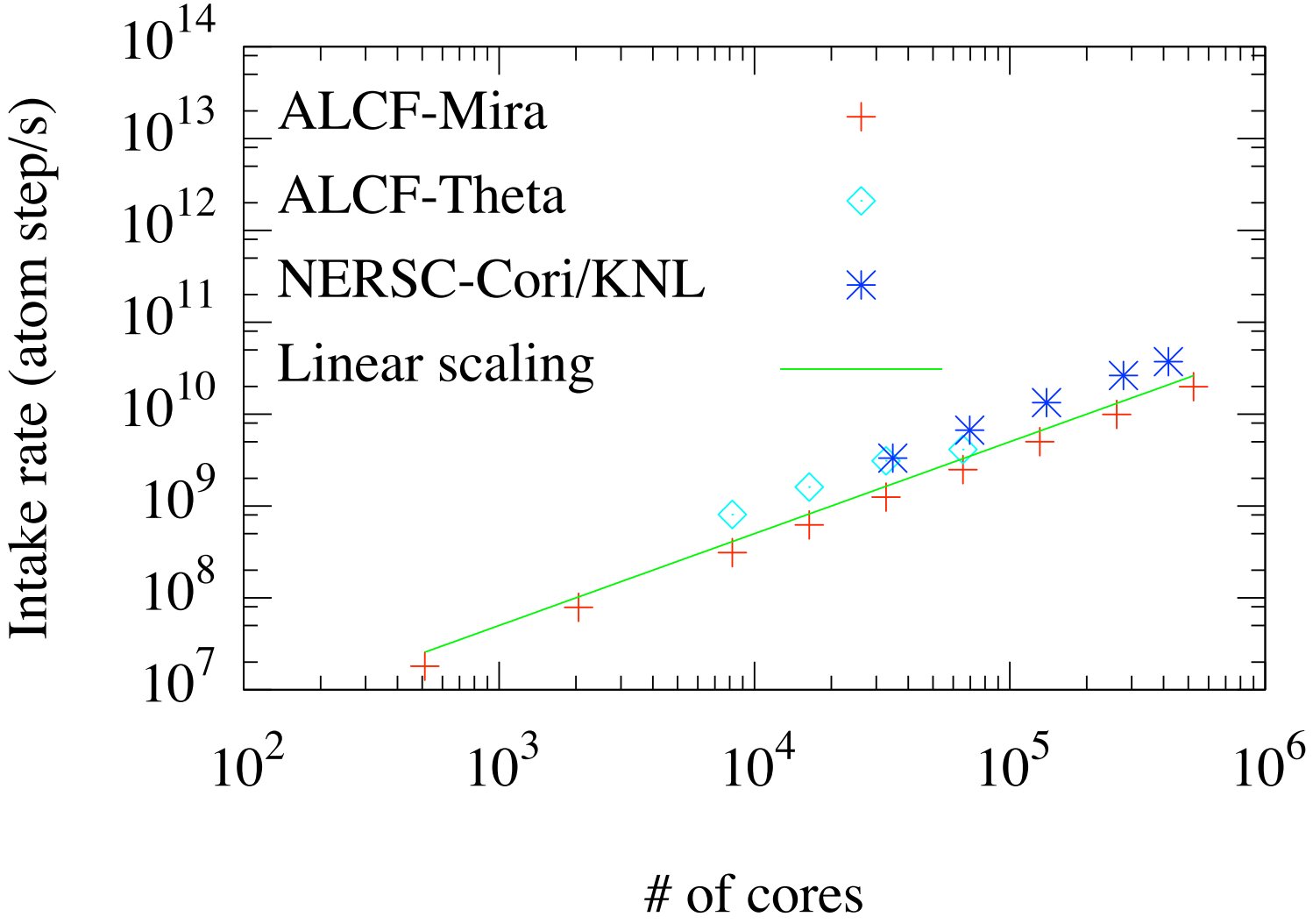
T=300K, LANL Grizzly, 4h runs

N_{cores}	Trajectory length (ps)	Generated segment time (ps)	#Transitions	#States	$\langle t_{trans}/M\ t_c \rangle$	<R>	Simulation rate (μs/hour)
9,000	556,093,988	556,539,980	4,614	28	13.39	166	139
18,000	1,315,941,923	1,346,516,503	24,610	64	2.97	384	333
27,000	2,209,432,238	2,214,868,608	13,479	47	4.55	294	552
36,000	2,291,027,808	2,318,254,470	50,258	60	1.26	909	592

T=400K, LANL Grizzly, 2h runs

36,000	511,059,851	1,090,349,898	35,824	411	0.39	91	257

Benchmark results: towards 10^6 cores



Summary

Efficiency: $$\frac{Rt_{trans}}{\left\lceil \frac{Rt_{trans}}{fMt_s} \right\rceil fMt_s + fMt_c}$$

In order to scale, we need some/all of these:

- **Large t_{trans}:** rare events
- **Small t_c:** short correlation times
- **Large R:** deep super-basins
- **Small M:** strong-scaling MD
- **Small f:** accurate predictions

Physics

Computer Science

Statistics

Summary

To maximize chances of success, need to play every trick in the book

k=1/μs/atom, t_c=1 ps, t_s=1 ps, 1000 atoms/core, R=100, f=0.25

#atoms	Decompose	Replicate	Bookkeep	Speculate
10^3	1x	500x	100x	12x
10^6	1000x	0.5x	100x	16x
10^9	1,000,000x	1x	1x	1x

Conclusion

- Advances in computing present huge opportunities, but also significant challenges
- Breaking off from the natural scaling path of MD requires methods that are specially tailored to massively-parallel hardware
- At very large scales, success depends on exploiting every trick in the book

- **To know more:**
  - Perez, Uberuaga, Voter, Comp. Mat. Sci. 100, 90 (2015)
  - Perez, Cubuk, Waterland, Kaxiras, Voter, JCTC 12, 18 (2016)
  - Perez, Huang, Voter, JMR 33, 813 (2018)
  - Upcoming Section in Handbook of Materials Modeling, 2nd edition

Acknowledgements

Collaborator:

Art Voter (Los Alamos National Lab)

Richard Zamora (Argonne National Lab)

Rao Huang (Xiamen University)

Funding:

Los Alamos NATIONAL LABORATORY EST. 1943

U.S. DEPARTMENT OF ENERGY

Office of Science

ECP EXASCALE COMPUTING PROJECT

Computing:

Los Alamos NATIONAL LABORATORY EST. 1943

NERSC

Applications

- Defect evolution in fusion materials (w. Luis Sandoval, Blas Uberuaga, Art Voter). Up to 100,000 cores, ~10,000 atoms, ms [Sci. Rep. 7, 2522 (2017)]

- Jogs in nickel (w. Lauren Smith, Tom Swinburne, Dallas Trinkle), ~1000 cores, ~10,000 atoms, tens of μs

- Cation defect evolution in pyrochlores (w. Romain Perriot, Blas Uberuaga, Art Voter), ~200 cores, ~1000 atoms, tens of μs [Nature Comm., 8, 681 (2017)]

- Shape evolution of metallic nanoparticles (w. Rao Huang, Art Voter). ~1000 cores, ~100 atoms, ms [JCP 147, 152717 (2017). JMR 33, 813 (2018)]