The surprising dynamics of non-lockstep execution

Georg Hager, Ayesha Afzal

Erlangen National High Performance Computing Center (NHR@FAU)

Friedrich-Alexander University Erlangen-Nürnberg

ScalPerf 2021 Workshop

September 19-23, 2021

Bertinoro, Italy

Idle wave propagation and (de)synchronization phenomena

- A. Afzal, G. Hager, and G. Wellein: *Propagation and Decay of Injected One-Off Delays on Clusters: A Case Study.* Proc. 2019 IEEE International Conference on Cluster Computing (CLUSTER), Albuquerque, NM, September 23-26, 2019. DOI: 10.1109/CLUSTER.2019.8890995
- A. Afzal, G. Hager, and G. Wellein: *Desynchronization and Wave Pattern Formation in MPI-Parallel and Hybrid Memory-Bound Programs*. In: P. Sadayappan, B. Chamberlain, G. Juckeland, H. Ltaief (eds): High Performance Computing. ISC High Performance 2020. Lecture Notes in Computer Science, vol 12151. Springer, Cham. **Available with Open Access.** DOI: 10.1007/978-3-030-50743-5_20
- A. Afzal, G. Hager, and G. Wellein: *Delay Flow Mechanisms on Clusters*. Poster at EuroMPI 2019. EuroMPI2019_AHW-Poster.pdf EuroMPI2019-AHW-Summary.pdf
- A. Afzal, G. Hager, and G. Wellein: *Analytic Modeling of Idle Waves in Parallel Programs: Communication, Cluster Topology, and Noise Impact*. ISC High Performance 2021 Digital, June 24 – July 2, 2021, Frankfurt, Germany. DOI: 10.1007/978-3-030-78713-4_19

Ayesha Afzal

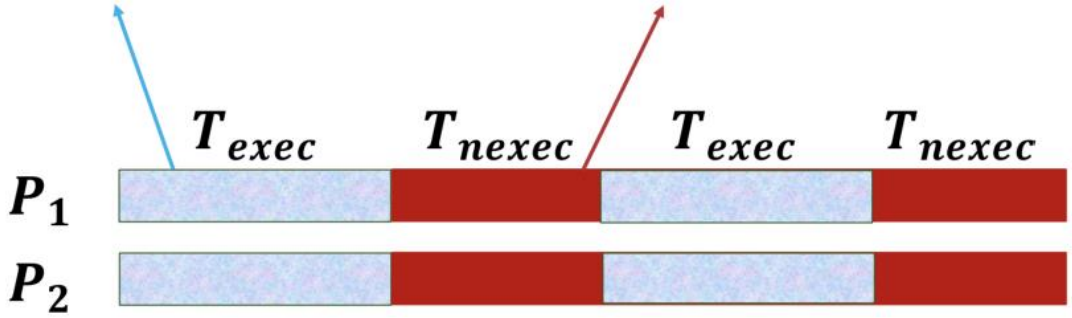
- A. Afzal, G. Hager, and G. Wellein: *An analytic performance model for overlapping execution of memory-bound loop kernels on multicore CPUs*. Submitted. Preprint: arXiv:2011.00243

Composite analytic models

Plausible assumption: $T = T_{exec} + T_{nexec}$

e.g.,
$\max(T_{BW}, T_{flops})$

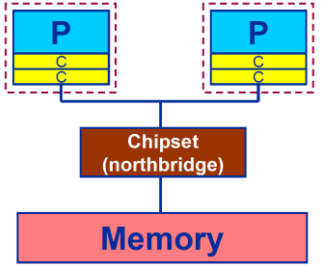
e.g.,
$\lambda + \frac{V}{B}$



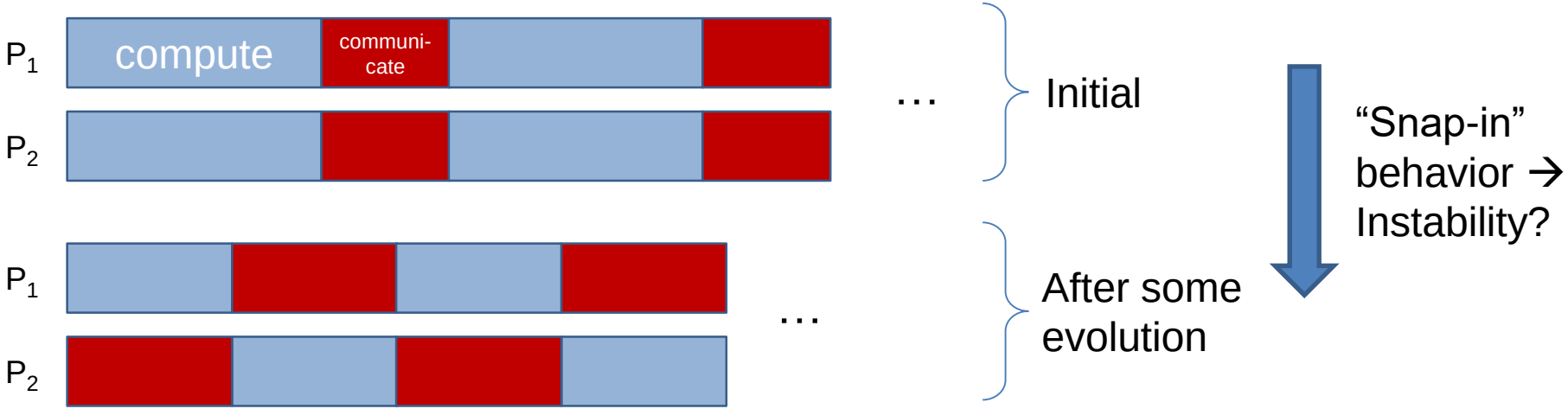
In practice, $T \neq T_{exec} + T_{nexec}$ and it can go in either direction

Initial observation

Two-socket single-core Pentium IV “Prescott” node (2004-ish)



MPI-parallel Lattice-Boltzmann solver timeline view:

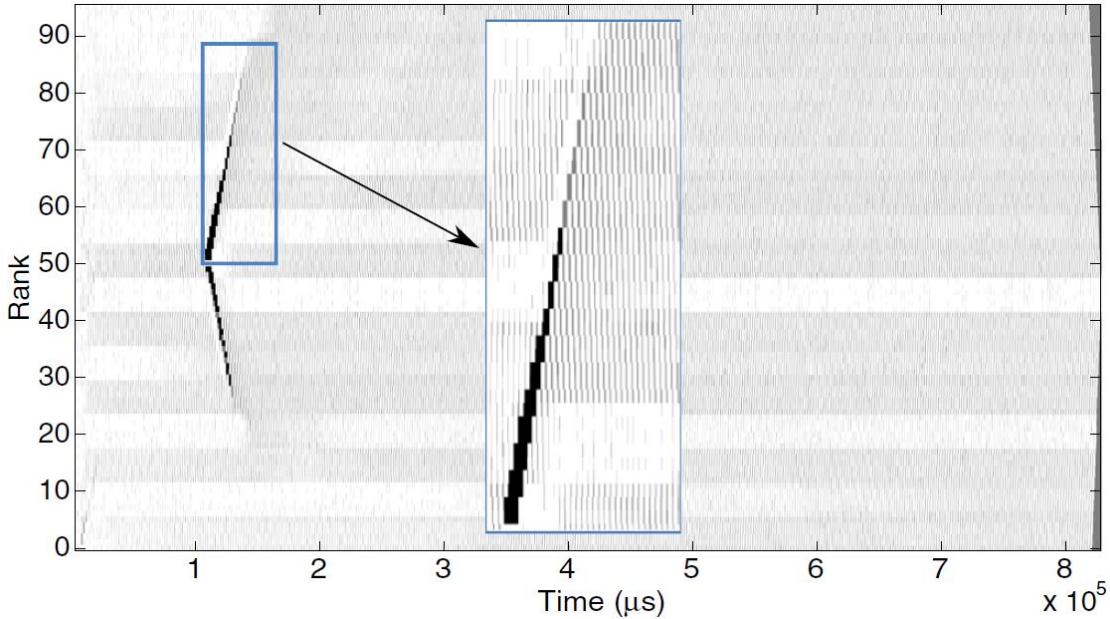


Markidis et al. (2015)

Simulator-based analysis

Idle waves perceived as “damped linear waves”

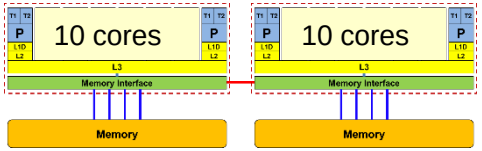
Classical wave equation postulated for continuum description

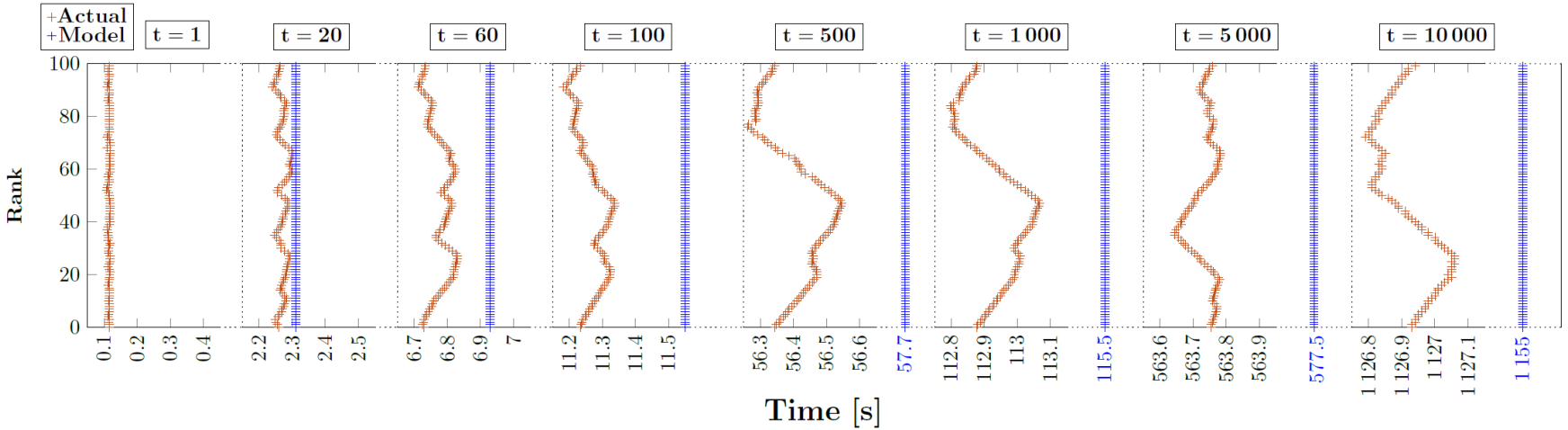


S. Markidis et al.: *Idle waves in high-performance computing*. Phys. Rev. E **91**(1), 013306 (2015). DOI: 10.1103/PhysRevE.91.013306

A more modern platform

RRZE “Emmy” cluster, 10 cores/socket,
2 sockets/node





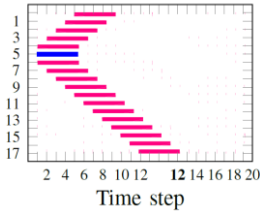
→ spontaneous symmetry breaking, “computational wave”

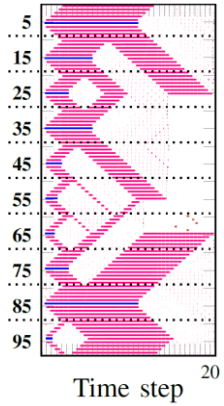
Why? Under which conditions?

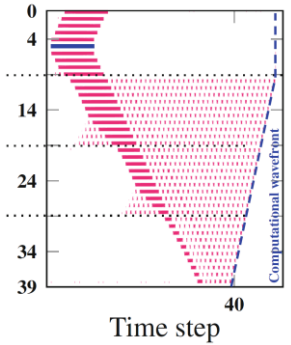
Research questions

Setting: MPI- or hybrid-parallel bulk-synchronous barrier-free programs

- How do “disturbances” propagate?
  - Injected idle periods
  - Dependence on communication characteristics
- How do idle waves interact with each other, with noise, and with the hardware?
  - Idle wave decay
    (noise-induced, bottleneck-induced, topology-induced)
- How do computational waves form? Instabilities?
  - Core-bound vs. memory-bound
  - Amplitude of the computational wave?
- Continuum description?



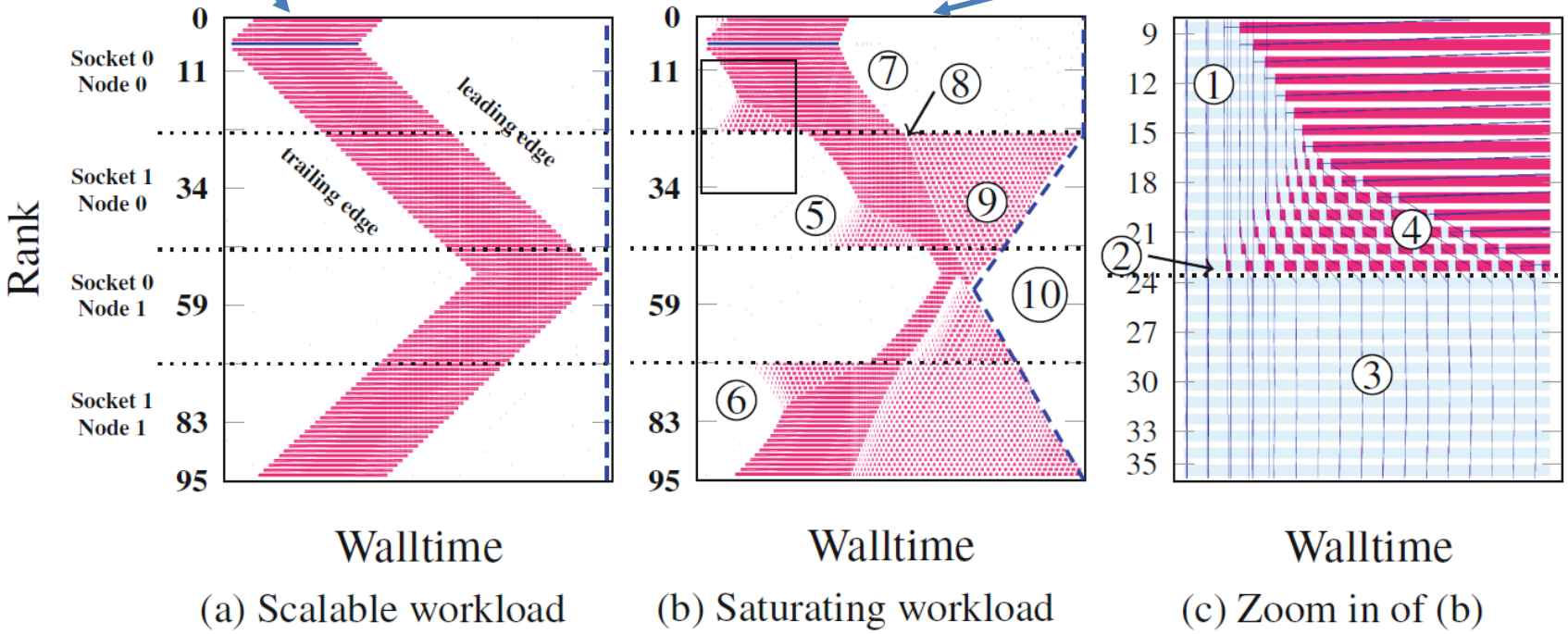




Idle wave propagation and bandwidth-bottleneck-induced decay

Analytical model for idle wave speed with scalalble workload:
DOI: 10.1007/978-3-030-78713-4_19

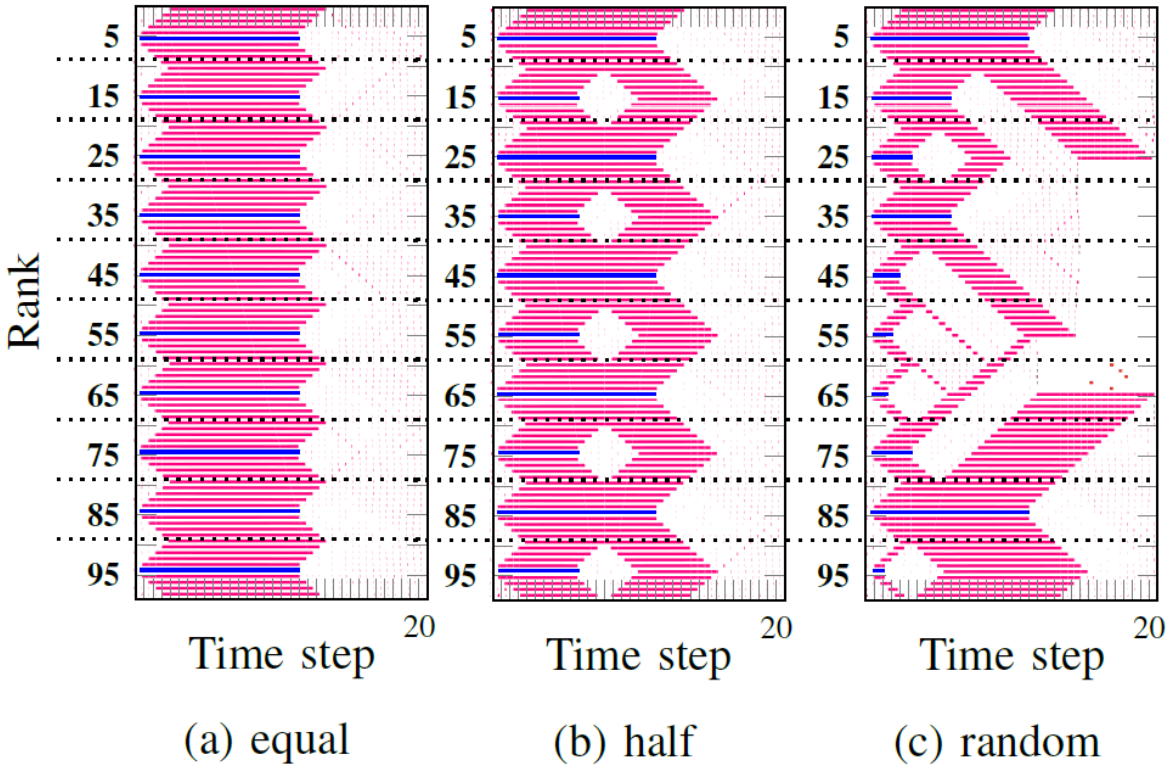
Decay even on silent system:
DOI: 10.1007/978-3-030-50743-5_20



Idle waves interact nonlinearly

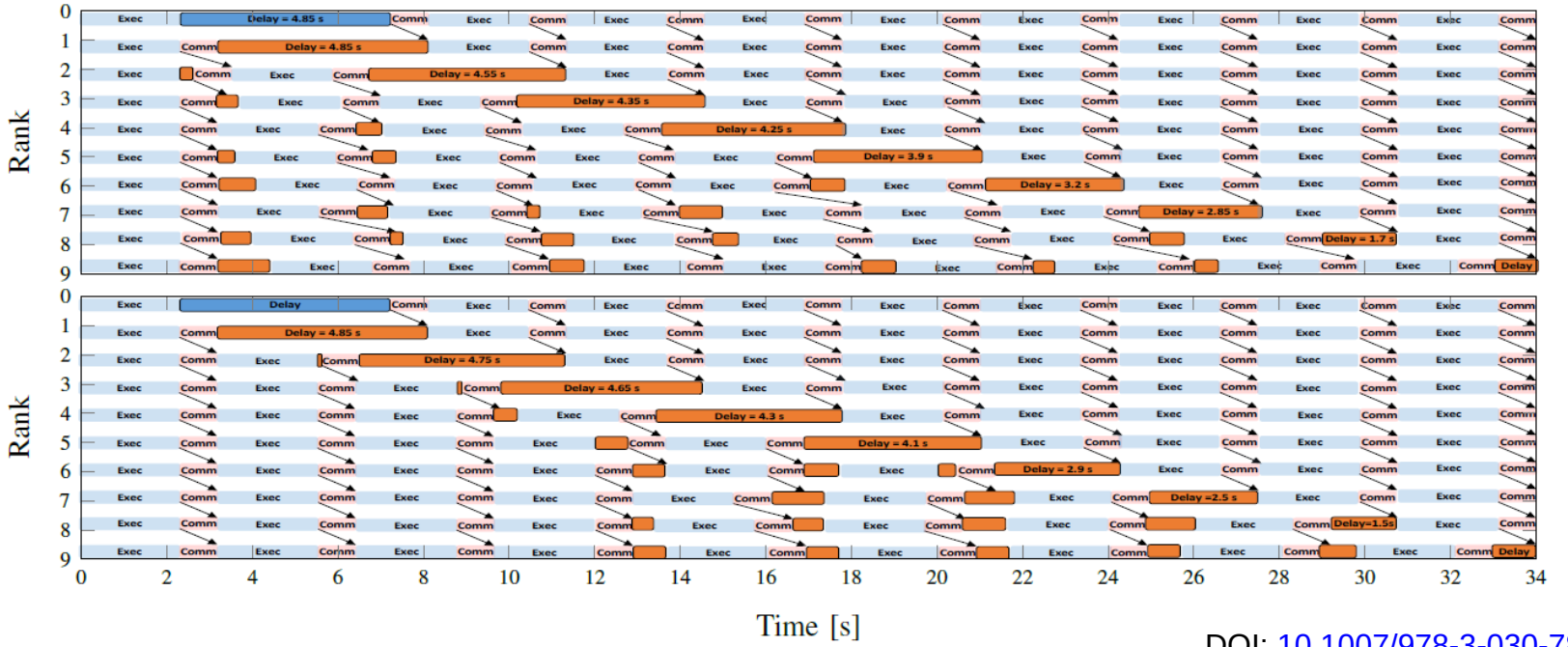
- A wave-like description cannot be based on a linear model
- Basis for noise-induced decay of idle waves

DOI: 10.1109/CLUSTER.2019.8890995



Noise-induced idle wave decay

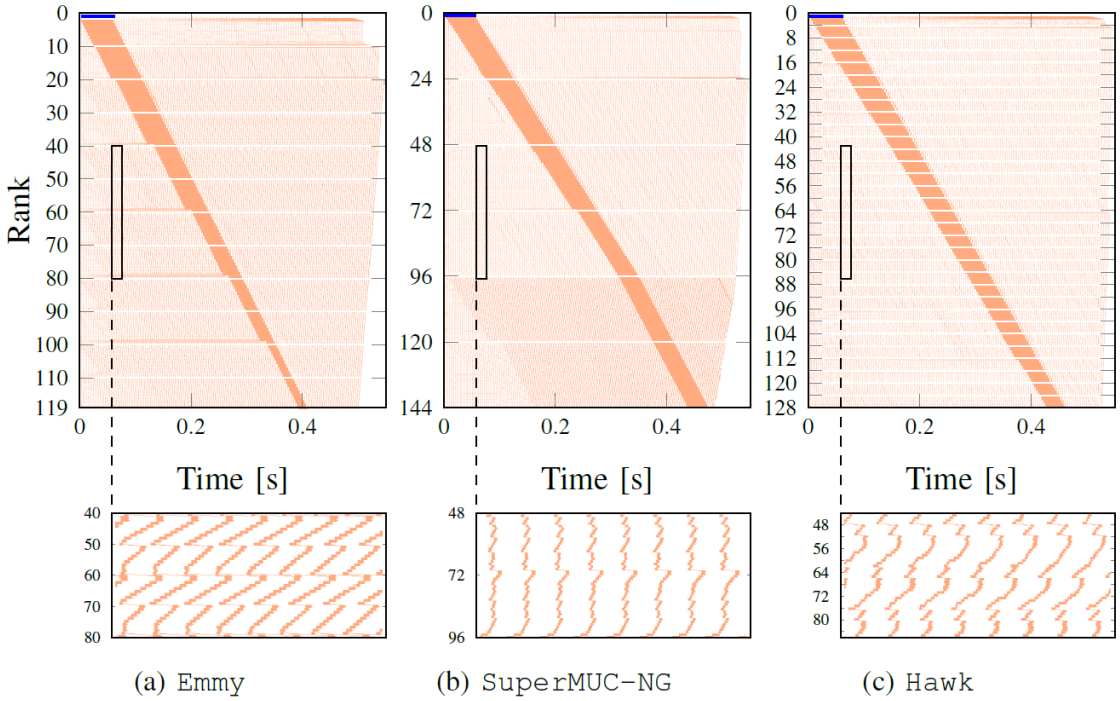
- System or application noise “eats away” on the idle wave
- Statistical details do not matter (only integrated noise power)



DOI: 10.1007/978-3-030-78713-4_19

Topological idle wave decay

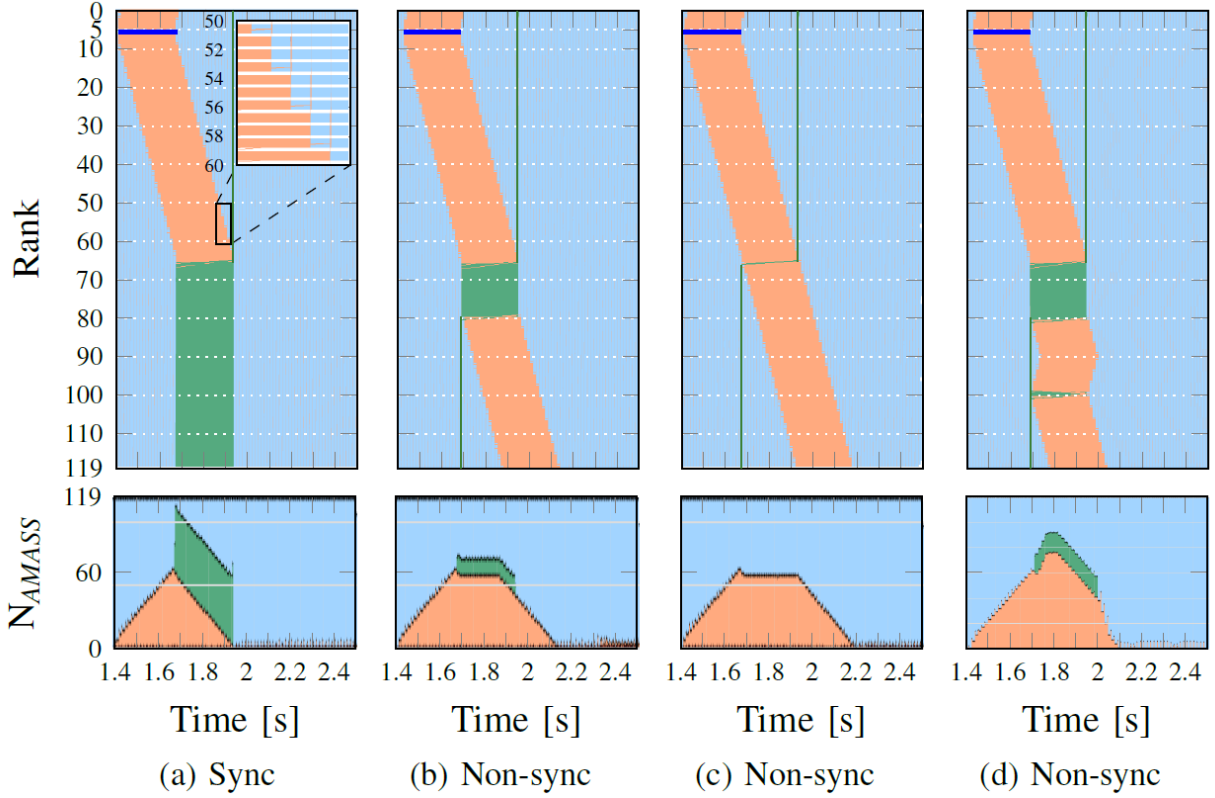
- Topological boundaries (ccNUMA domains, sockets, nodes) cause fine-grained noise which dampens the idle wave
- Highly system dependent
- No decay in homogeneous situation (round-robin placement)



DOI: 10.1007/978-3-030-78713-4_19

Collectives can be permeable to idle waves

- Some collectives are not necessarily synchronizing
- Many implementations let idle waves pass through

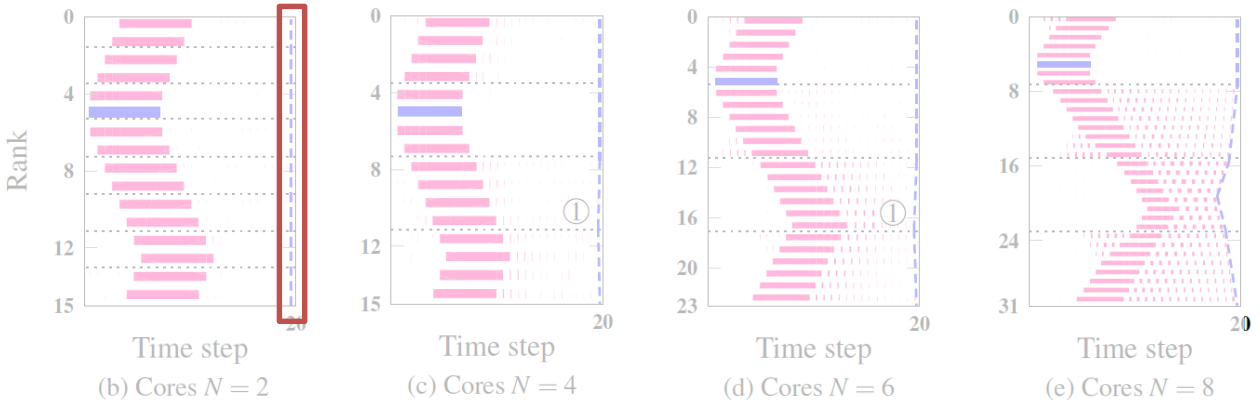


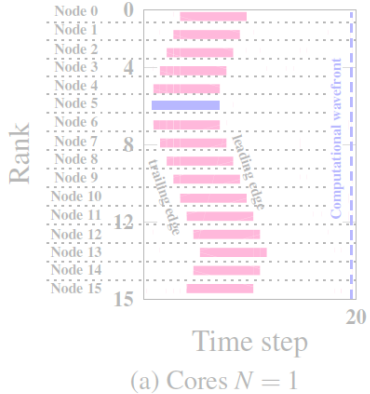
4 different MPI_Reduce() implementations (Intel MPI 19)

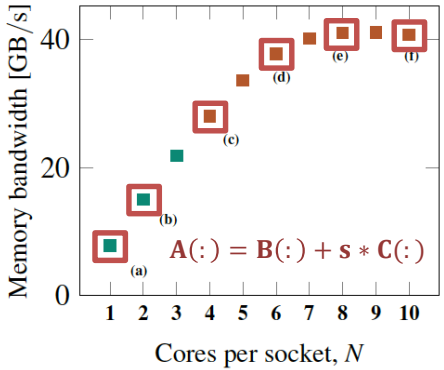
DOI: 10.1007/978-3-030-78713-4_19

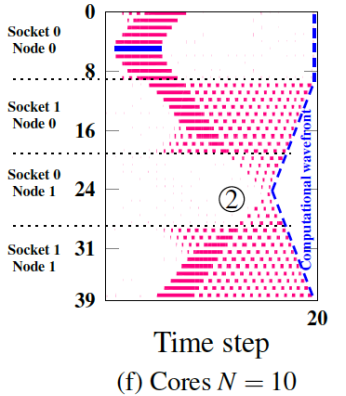
Formation of computational wavefronts

- 2-socket 10-core
- No decay if in non-saturated regime
- Faster decay with stronger saturation





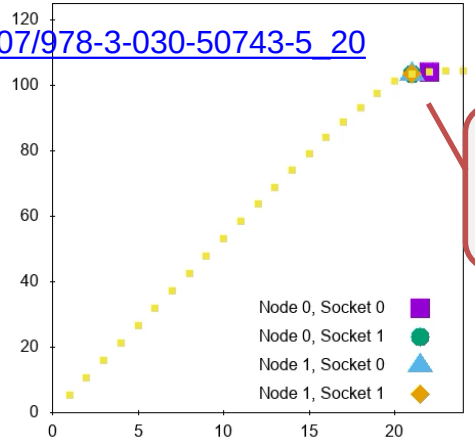




DOI: 10.1007/978-3-030-50743-5_20

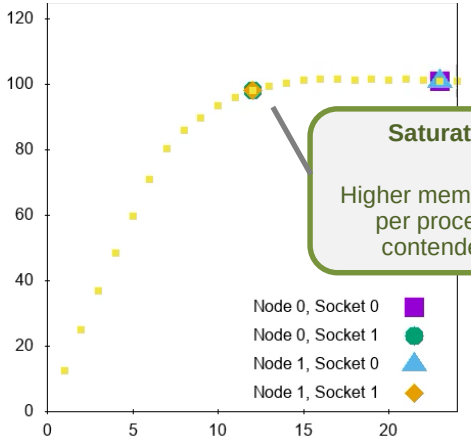
Computational wave settles at the saturation point (sometimes)

DOI: 10.1007/978-3-030-50743-5_20



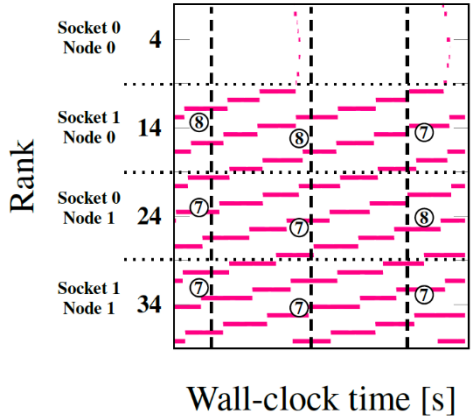
Saturation point

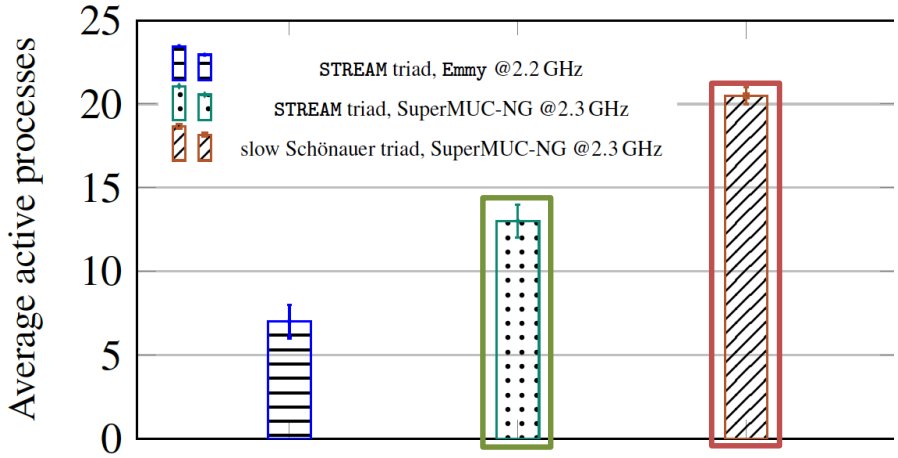
Higher memory bandwidth per process in each contended domain



Saturation point

Higher memory bandwidth per process in each contended domain



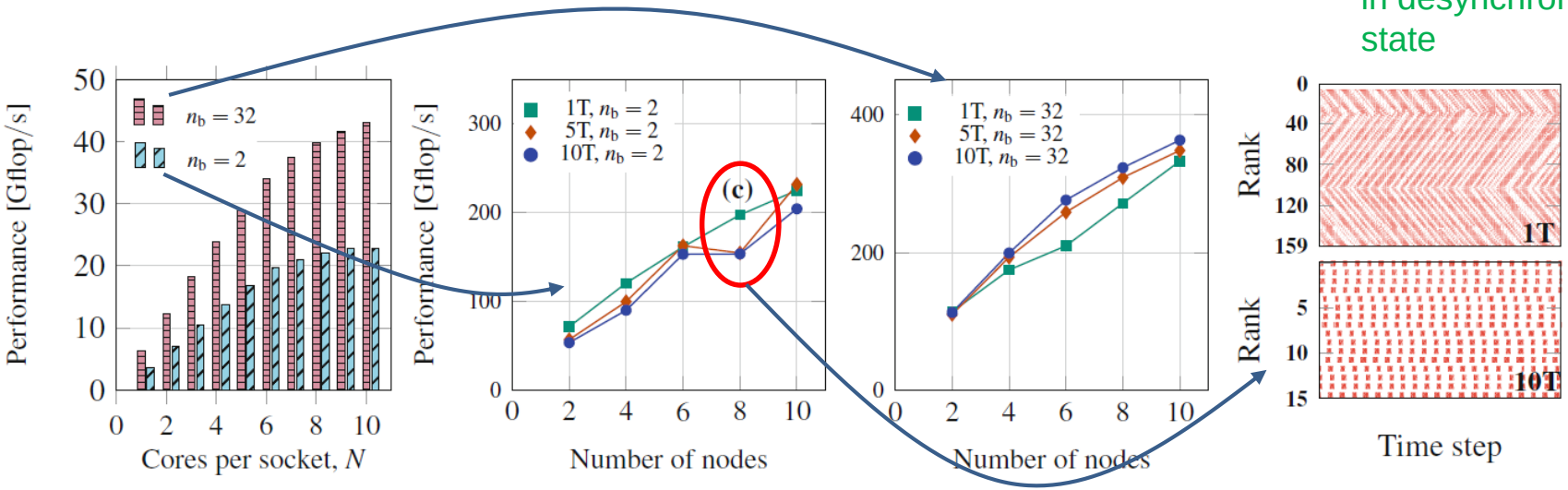


Application: Chebyshev Filter Diagonalization (ChebFD)

DOI: 10.1007/978-3-030-50743-5_20

- Computes inner eigenvalues of a large sparse matrix
- Blocking optimization: M. Kreutzer, G. H., D. Ernst, H. Fehske, A.R. Bishop, G. Wellein, DOI: 10.1007/978-3-319-92040-5_17
- MPI+OpenMP hybrid, topological insulator matrix, Emmy@RRZE

Computes faster in desynchronized state

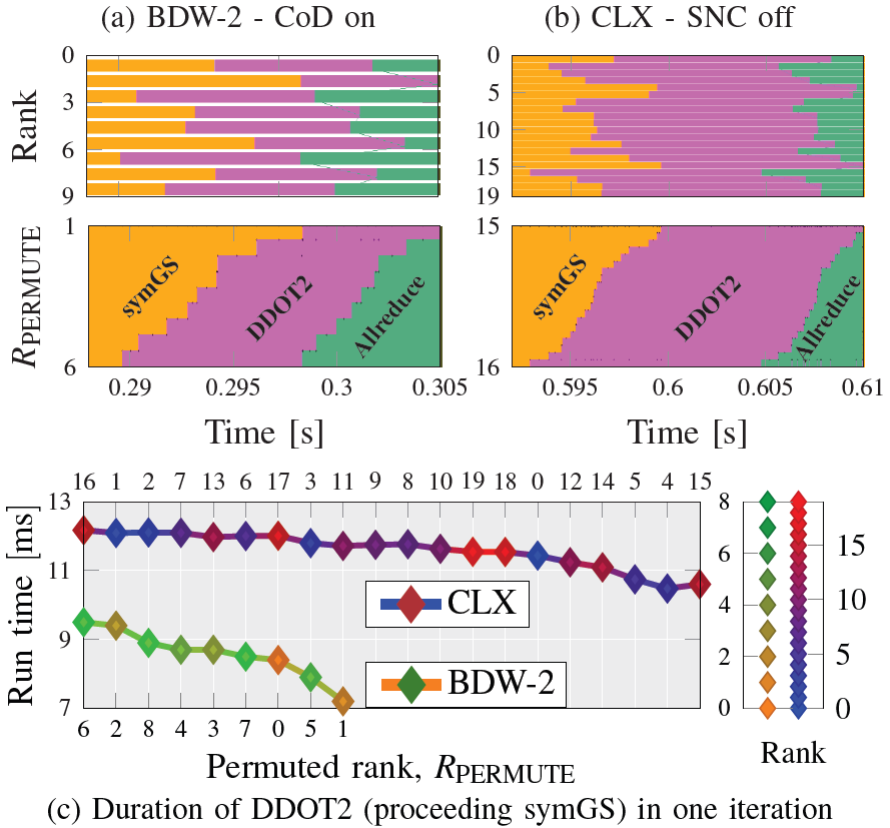


(a) Single socket performance (b) MPI only vs. hybrid communication, $n_b = 2$, $n_b = 32$ (c) Timeline visulization

Dynamics of desynchronized overlapping kernels

- Single-core memory bandwidth utilization f of kernels determines re-/desync behavior

- $f_{before} > f > f_{after}$: increase desync
- $f_{before} < f < f_{after}$: reduce desync



(c) Duration of DDOT2 (proceeding symGS) in one iteration

arXiv:2011.00243

Current results

- Instability of bulk-synchronous barrier-free programs is bound to the presence of a resource bottleneck
- Desynchronized bottlenecked programs can exhibit automatic communication/execution overlap via formation of computational waves
- Idle waves can be absorbed by fine-grained system noise, and the mechanism behind this is well understood
- Idle waves can decay via topological noise caused by inhomogeneous communication characteristics
- Proof that noise statistics is largely irrelevant for idle wave decay rate
- Analytic model for idle wave velocity w.r.t. communication topology and characteristics
- Experimental evidence that MPI collectives can be transparent to idle waves

Future directions

- Development of a comprehensive, bottleneck-aware simulator framework for message-passing programs
- Analytic description of decaying wave for bottleneck-triggered decay
- Bottlenecks other than memory bandwidth
- Analytic understanding of computational wave amplitude w.r.t. communication characteristics and bottleneck saturation
- Idle wave phenomena in irregular programs
- Physical model for coupled processes (Kuramoto-like)

$$\dot{\theta}_i = \omega_i + \alpha \sum_j T_{ij} V(\theta_j - \theta_i)$$

- Continuum description of parallel system as a nonlinear (dissipative?) medium

NHR FAU

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

THANK YOU.

OMI4papps

ESSEX, ExaSteel

Metacca/SeaSiTE/SKAMPY