Computational Waves in Parallel Programs and Their Impact on Performance Modeling

Ayesha Afzal , Georg Hager, Gerhard Wellein

Erlangen National High Performance Computing Center (NHR@FAU)

SIAM PP22 Minisymposium

Advances in Performance Modeling of Parallel Code (MS51/58)

2022-02-25

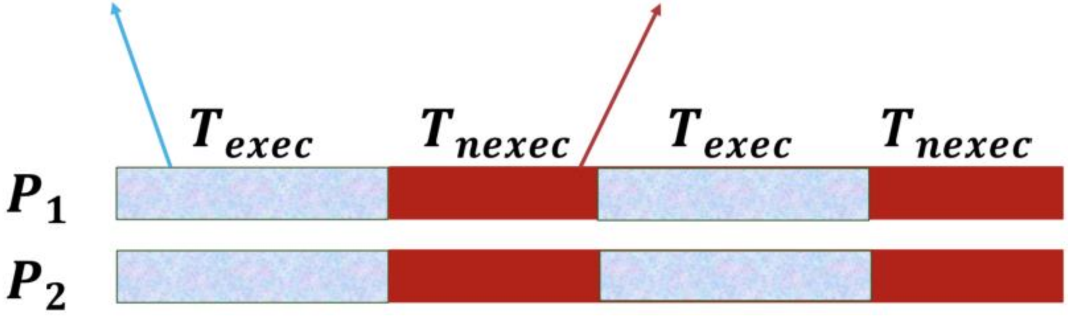
Composite analytic performance models and desynchronization in bulk-synchronous barrier-free programs

Composite analytic models

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

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

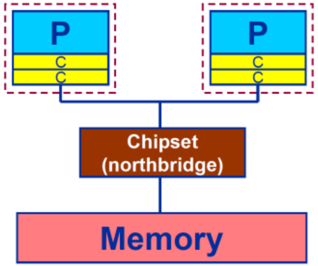
e.g.,
$\lambda + \frac{V}{B}$



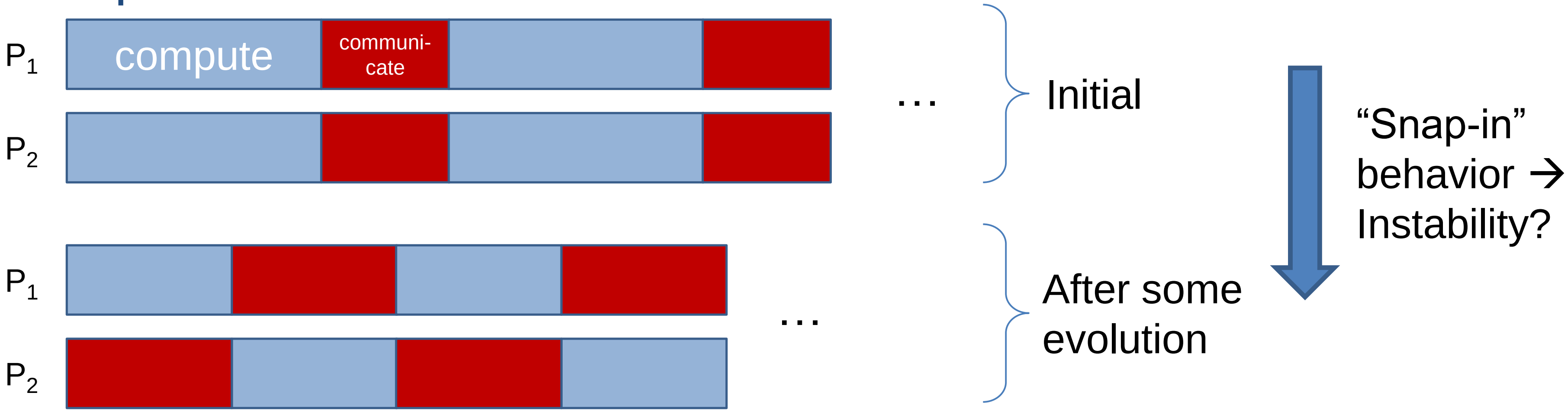
In practice, $T \neq T_{exec} + T_{nexec}$ and it can go in either direction

Initial observation

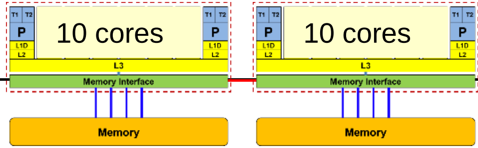
Two-socket single-core Pentium IV "Prescott" node (2004-ish)



MPI-parallel Lattice-Boltzmann solver timeline view:



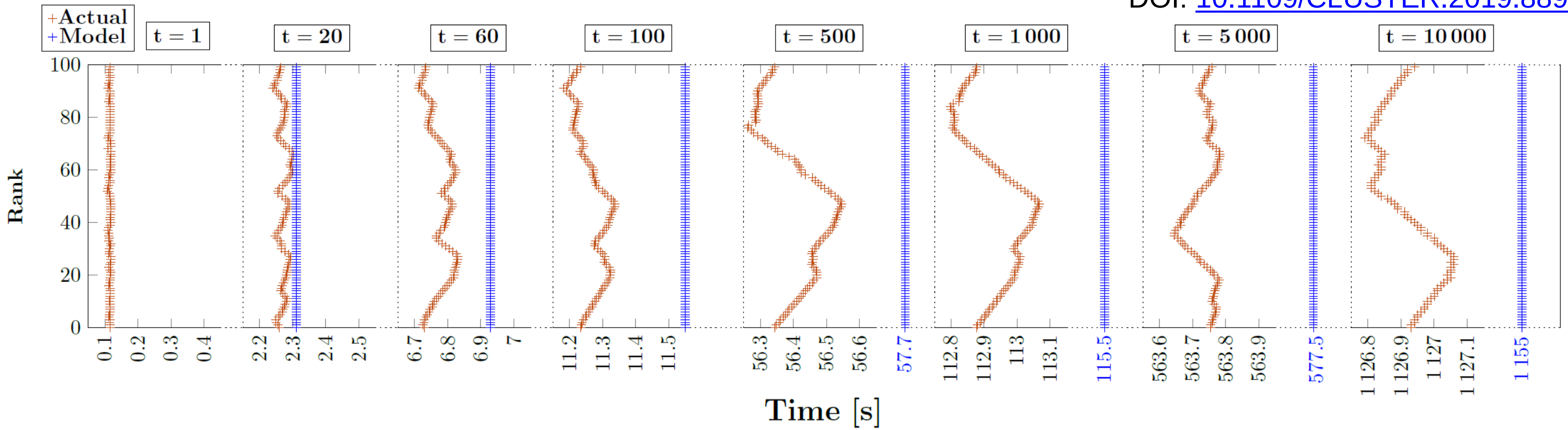
A more modern platform



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

MPI-only Lattice-Boltzmann simulation (empty channel)

DOI: 10.1109/CLUSTER.2019.8890995



→ Spontaneous symmetry breaking, “computational wavefront”

→ Why? Under which conditions? Automatic communication overlap?

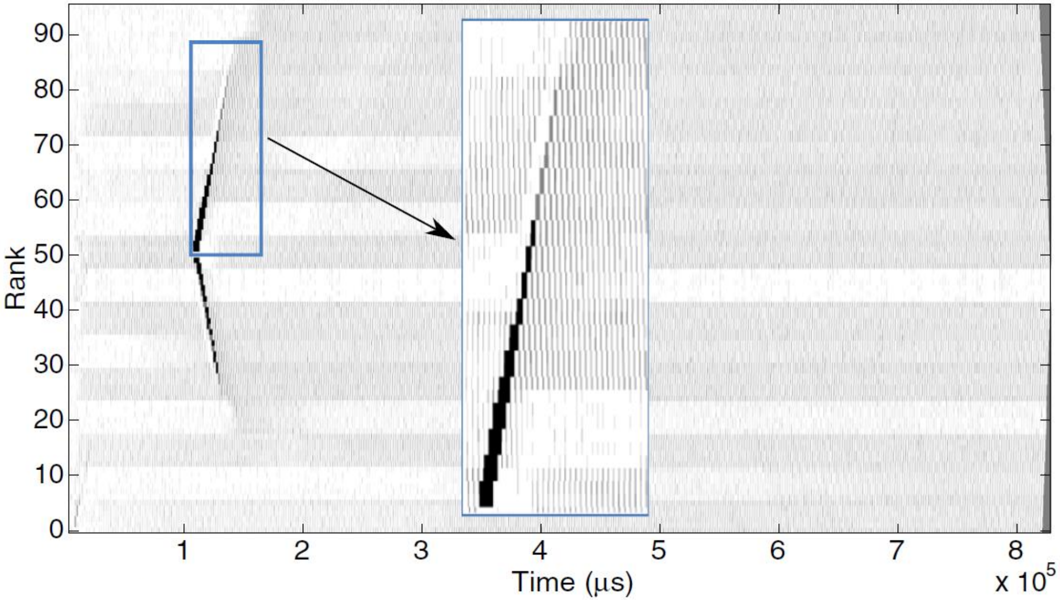
Idle waves in parallel programs

Markidis et al. (2015)

Simulator-based analysis

Idle waves perceived as "damped linear waves"

Classical wave equation postulated for continuum description

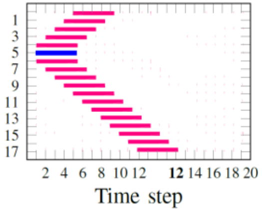


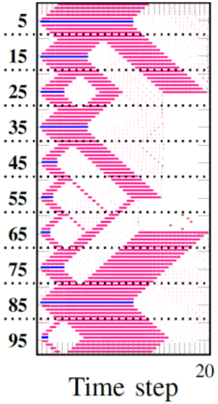
S. Markidis et al.: *Idle waves in high-performance computing*. Phys. Rev. E **91**(1), 013306 (2015).
DOI: [10.1103/PhysRevE.91.013306](https://doi.org/10.1103/PhysRevE.91.013306)

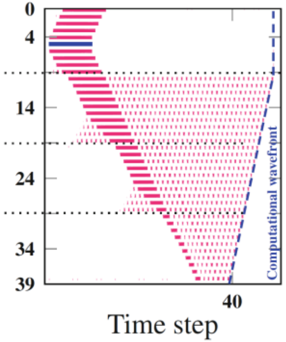
Research questions

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 speed

Assumptions:

- Scalable code (on ccNUMA domain)
- alternates between execution and communication phases
- has inter-process dependencies via point-to-point communication

Simplest case: Next-neighbor (e.g., halo) communication

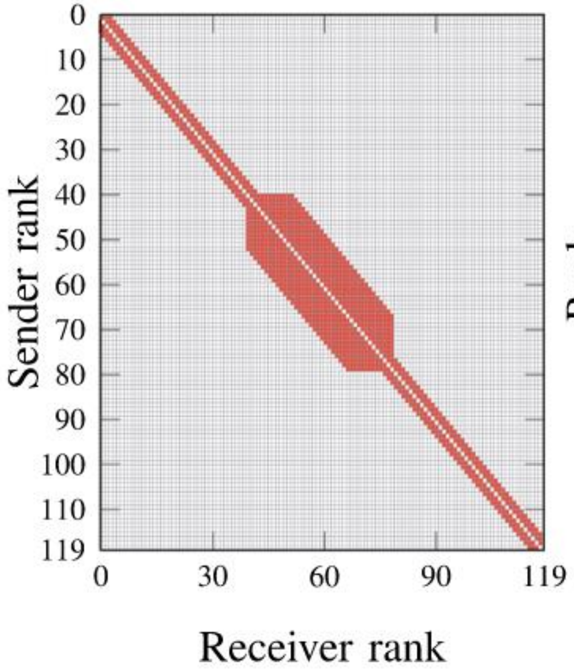
$$v_{\text{silent}}^{\text{min}} = 1\left[\frac{\text{ranks}}{\text{iter}}\right] \times \frac{1}{T_{\text{exec}} + T_{\text{comm}}}\left[\frac{\text{iter}}{\text{s}}\right]$$

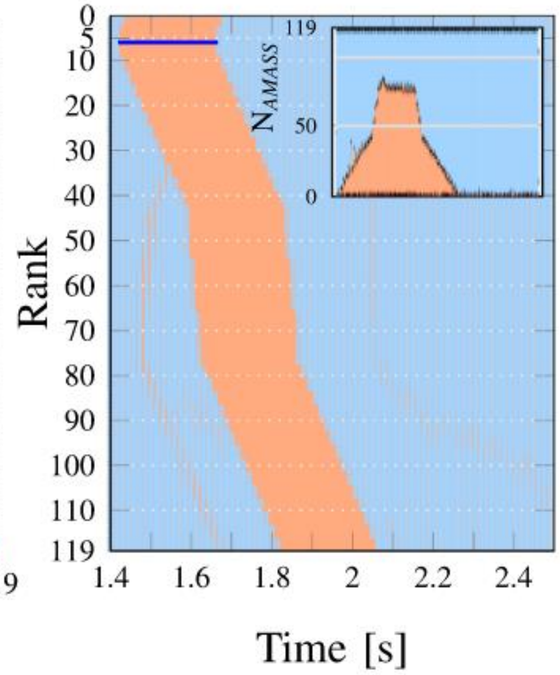
Number of communication partners and details of communication grouping influence the speed

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

Communication topology and idle wave speed

Long-distance point-to-point communication → fast idle waves

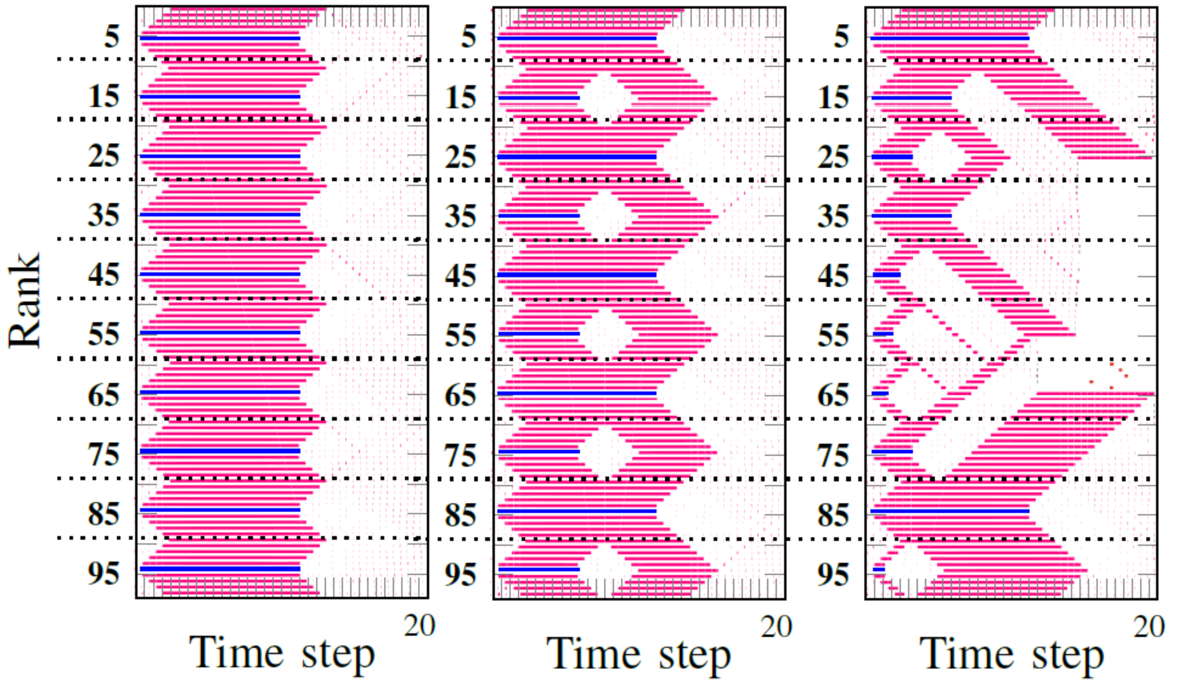




Idle waves interact nonlinearly

DOI: 10.1109/CLUSTER.2019.8890995

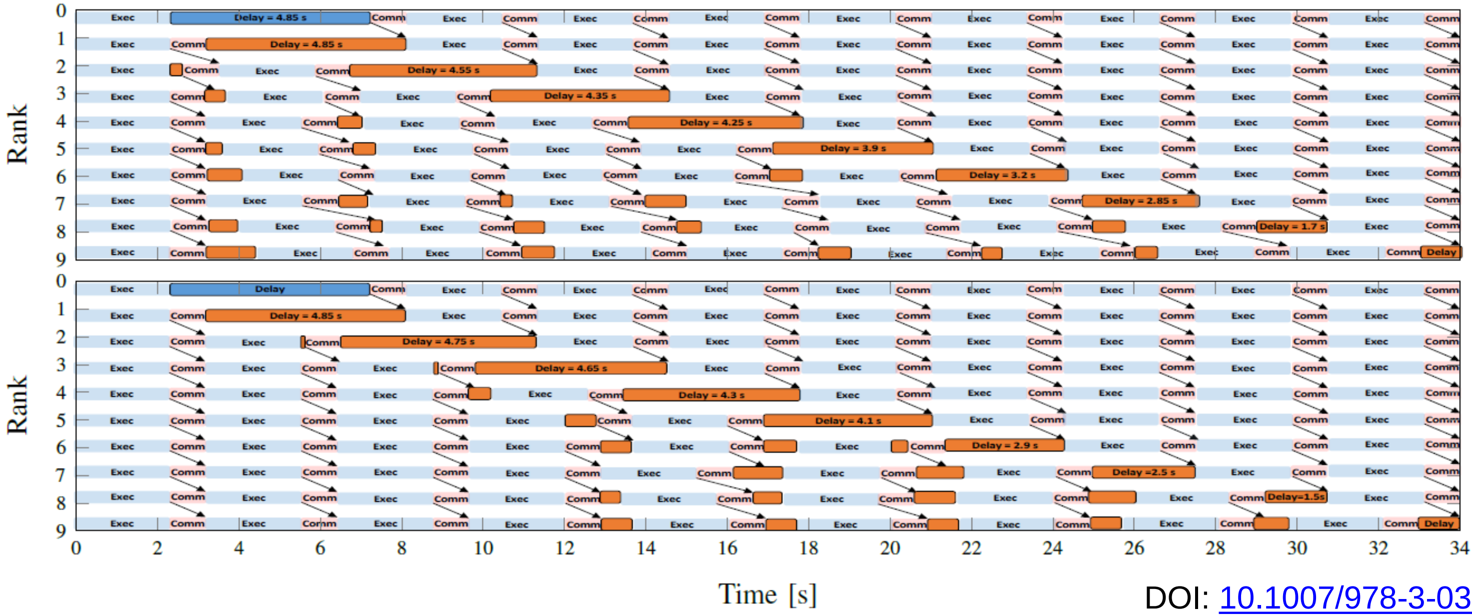
- A wave-like description cannot be based on a linear model
- Basis for noise-induced decay of idle waves



(a) equal (b) half (c) random

Noise-induced idle wave decay

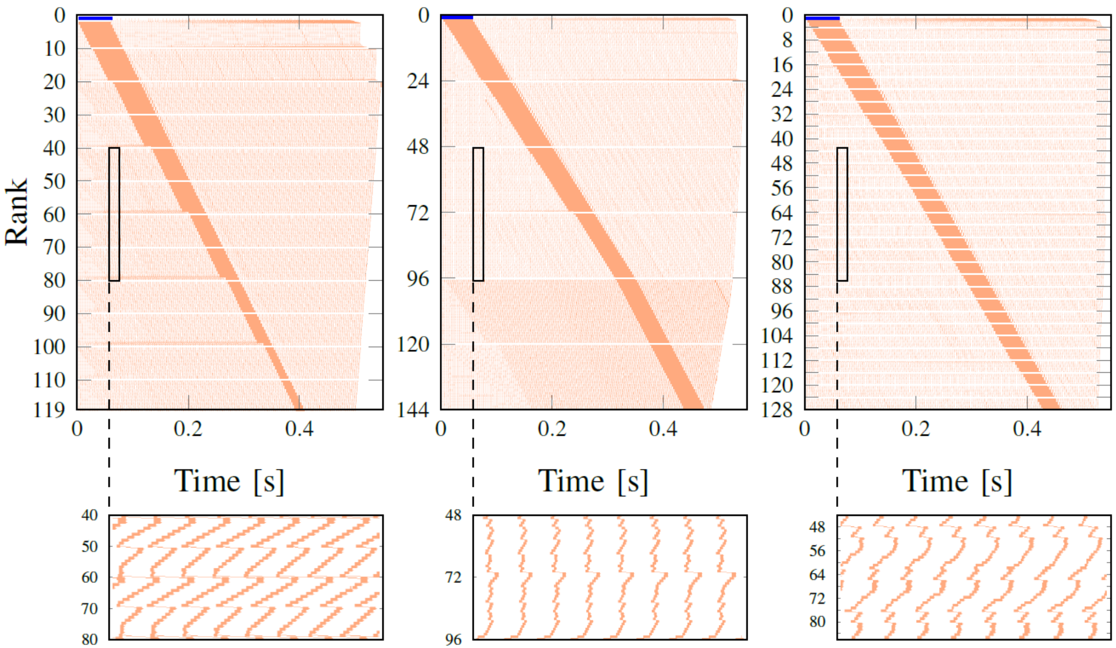
- System or application noise “eats away” on the idle wave
- Decay rate proportional to integrated noise power
- Statistical details do not matter



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)

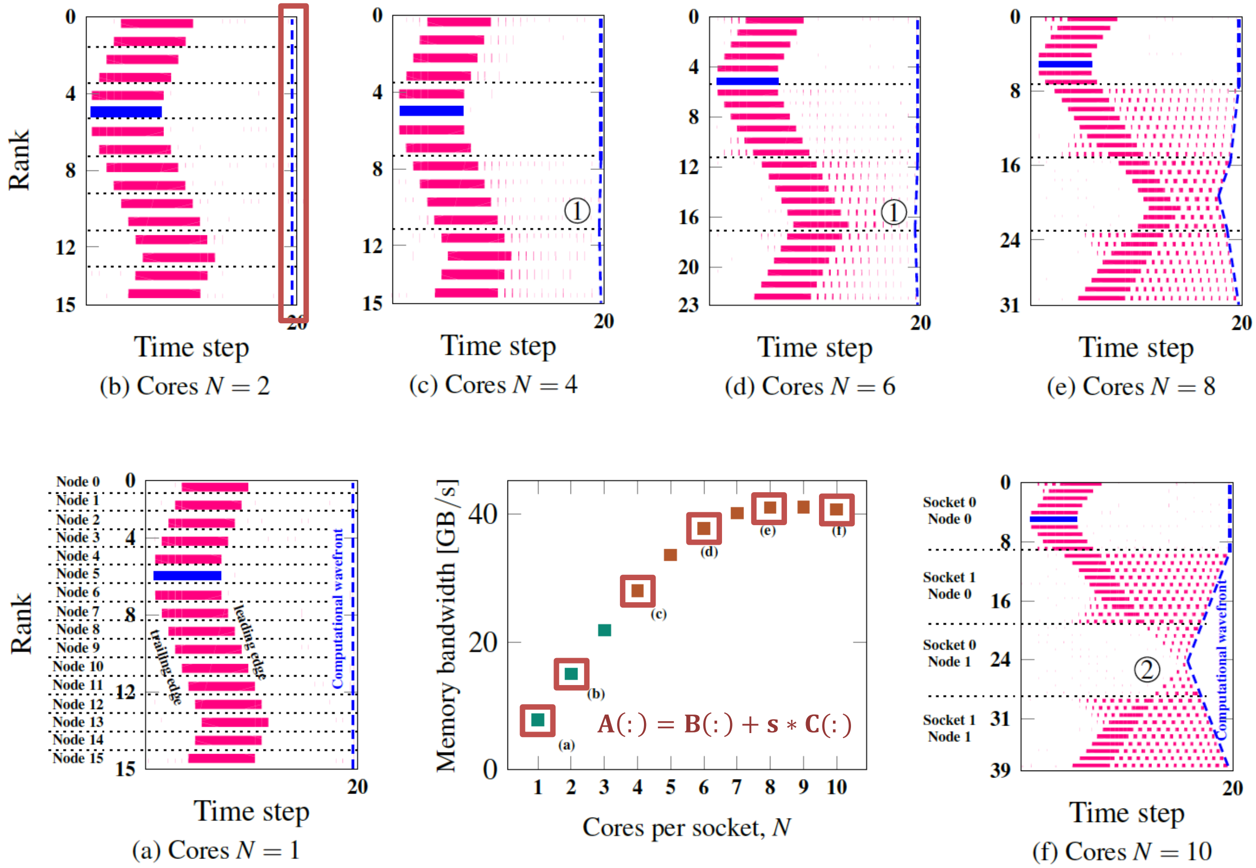


(a) Emmy (b) SuperMUC-NG (c) Hawk

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

Formation of computational wavefronts from idle waves

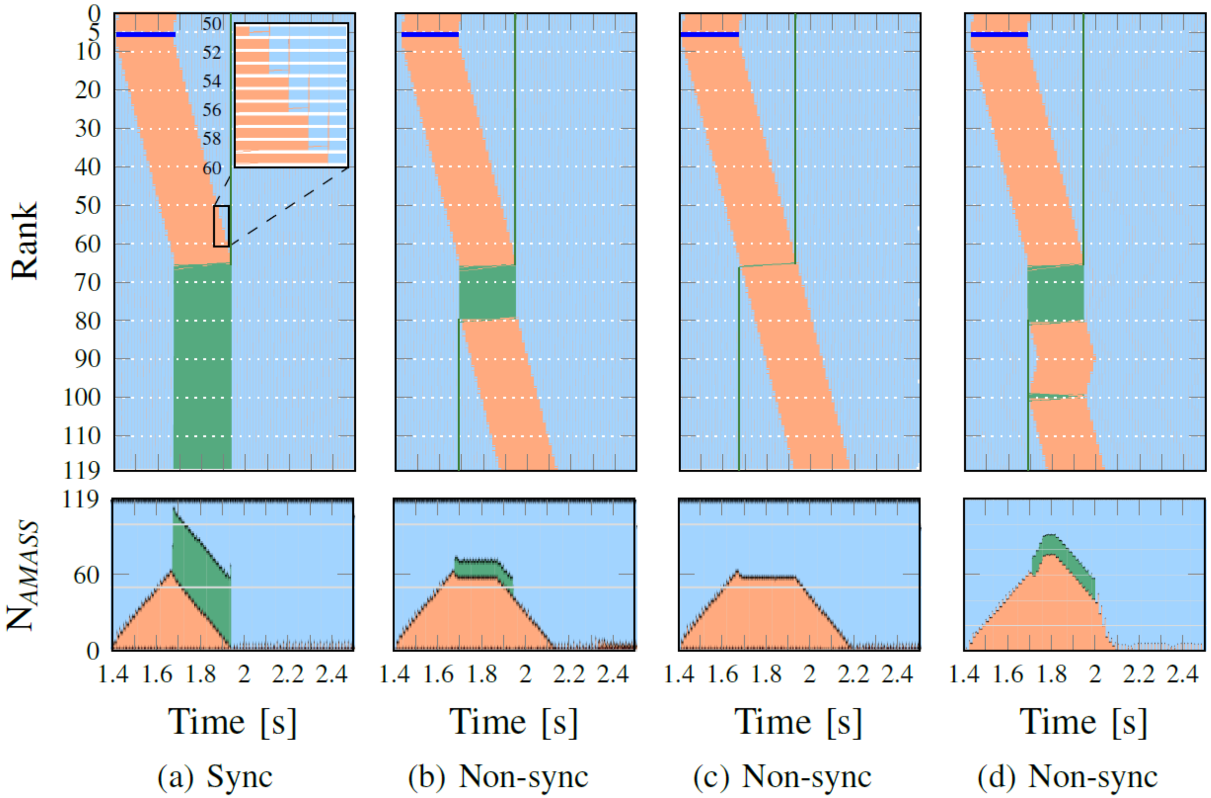
- 2-socket 10-core
- No decay if in non-saturated regime
- Faster decay with stronger saturation
- Comp. wavefront is the "echo" of the idle wave



DOI: [10.1007/978-3-030-50743-5_20](https://doi.org/10.1007/978-3-030-50743-5_20)

Collectives can be permeable to idle waves

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



four different MPI_Reduce() implementations (Intel MPI 19)

DOI: [10.1007/978-3-030-78713-4_19](https://doi.org/10.1007/978-3-030-78713-4_19)

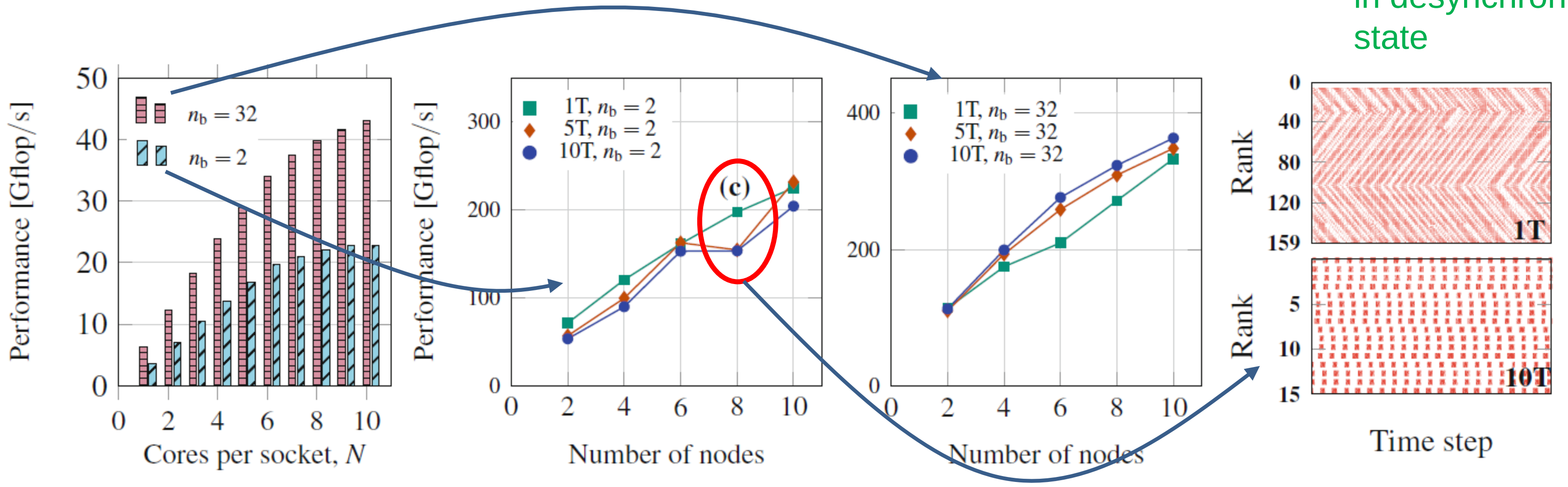
Impact of desynchronization on applications

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

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 wavefronts
- Idle waves can be absorbed by fine-grained system noise
- 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)

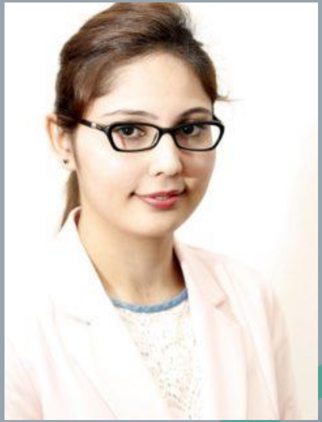
$$\dot{\theta}_i = \omega_i + \alpha \sum_j T_{ij} V(\theta_j - \theta_i)$$

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

NHR FAU

FAU Friedrich-Alexander-Universität Erlangen-Nürnberg

Thank you.



Ayesha Afzal

Georg Hager

KONWIHR

Bundesministerium für Bildung und Forschung