

The NASTJA Framework: Non-Collective Scalable Global Communications

Marco Berghoff and Ivan Kondov

Enabling High-Performance Computing

High Priority for Efficient Use of Current and Future HPC-Systems

- high node level performance
- scalable parallel communication
- omitting unnecessary calculations

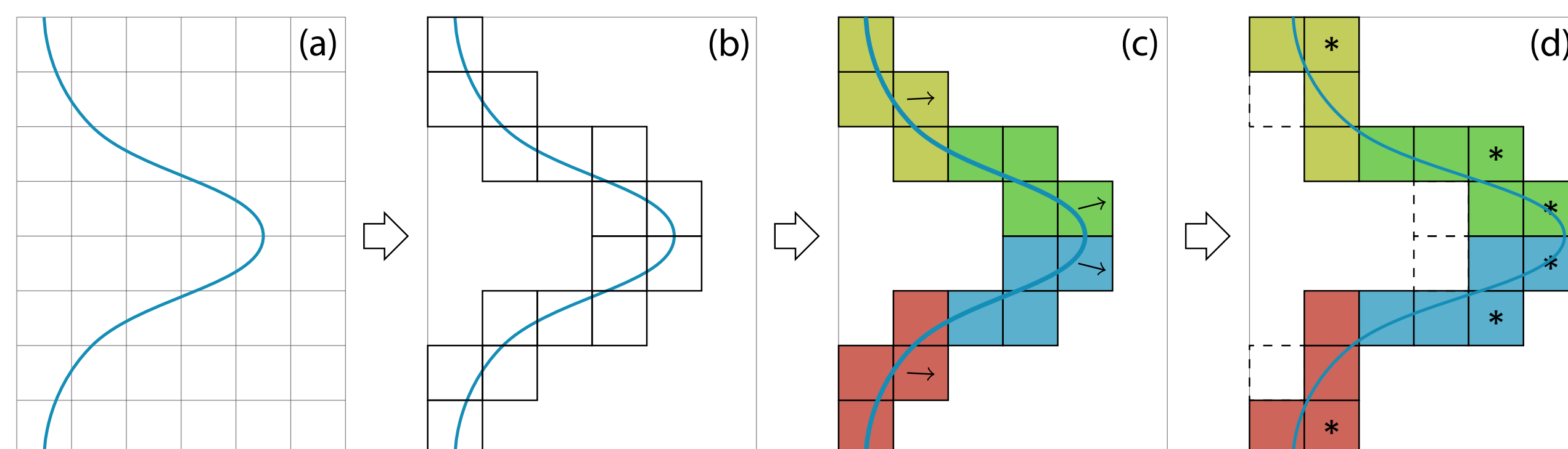
Design Goals and Development Guidelines of NASTJA*

- high level object oriented programming in C++ for flexibility
- implement calculation sweeps in plain C-like fashion
- optimized calculation sweeps using AVX-Intrinsics
- domain decomposition in a block structured grid for: cache efficiency, simple load balance, offload to Xeon Phi coprocessors or graphic cards
- local groups decide autonomously and consistent without any collective communications
- version control, issues tracker and continuous unit and simulation testing
- automatic build and deployment of Doxygen documentation and coverage reports

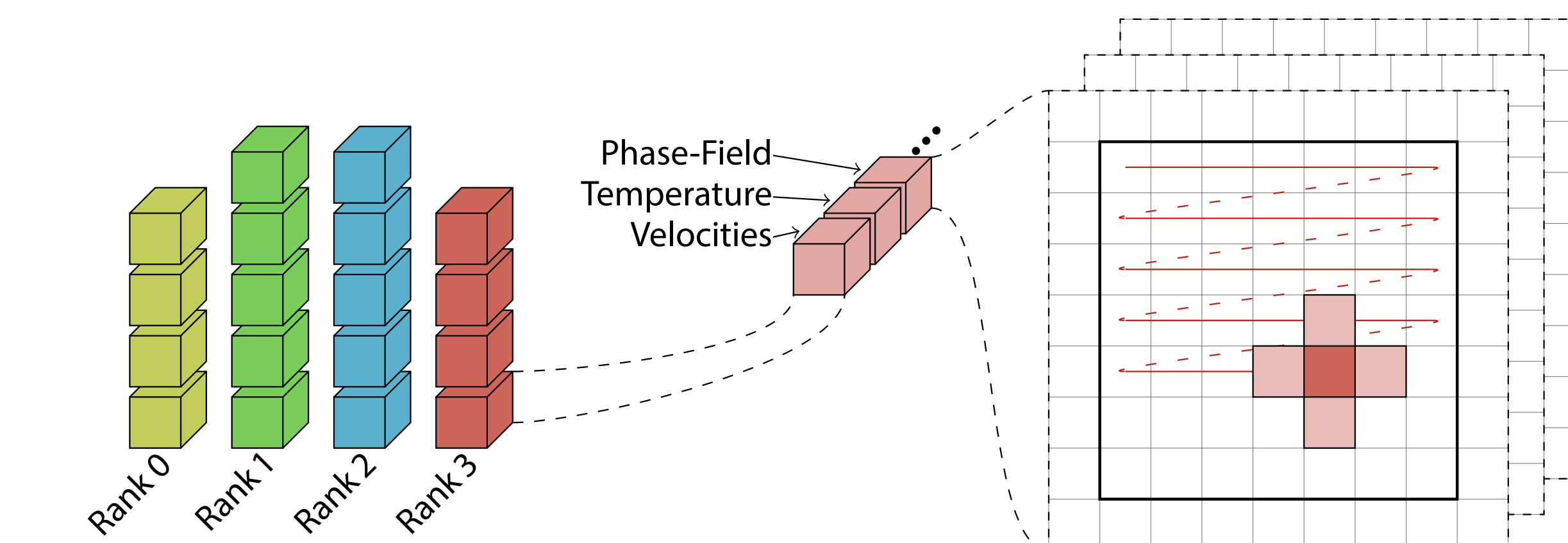
* Neoteric Autonomous Stencil code for Jolly Algorithms

Techniques behind NASTJA

Autonomous Adaptive Blocks in a Block Structured Grid



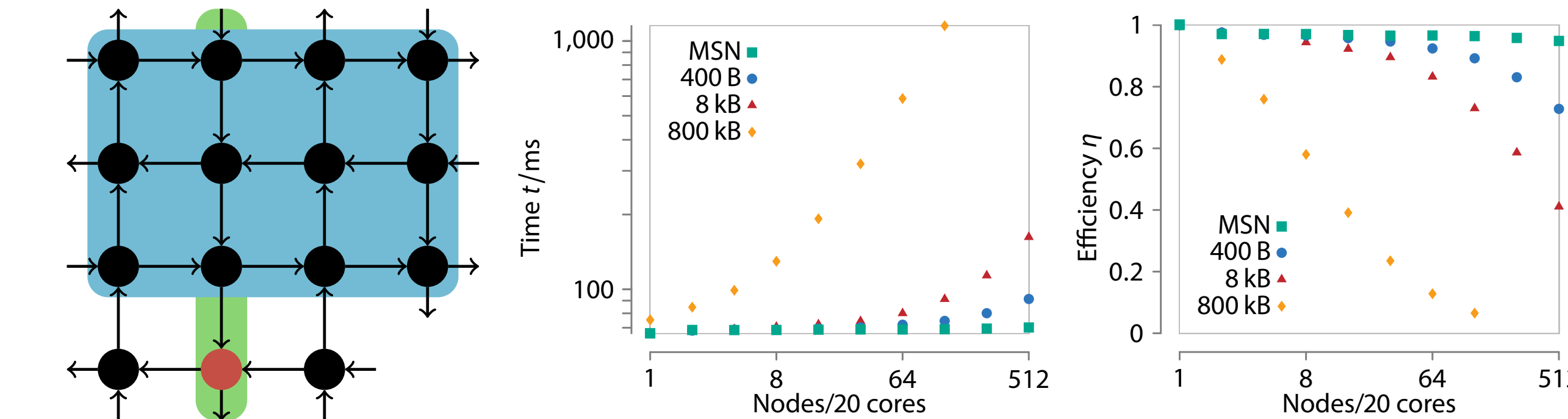
Data Structure and Sweeps



- blocks contain different fields, stencils sweep over fields and do the calculations

Global Communication Network

Manhattan Street Network



- avoiding global collective communications by multi-hop network (efficiency > 95 %)
- diameter d decrease with increased degree $k \rightarrow d \leq \sum_{i=0}^{k-1} \frac{N_i}{2} + 2$

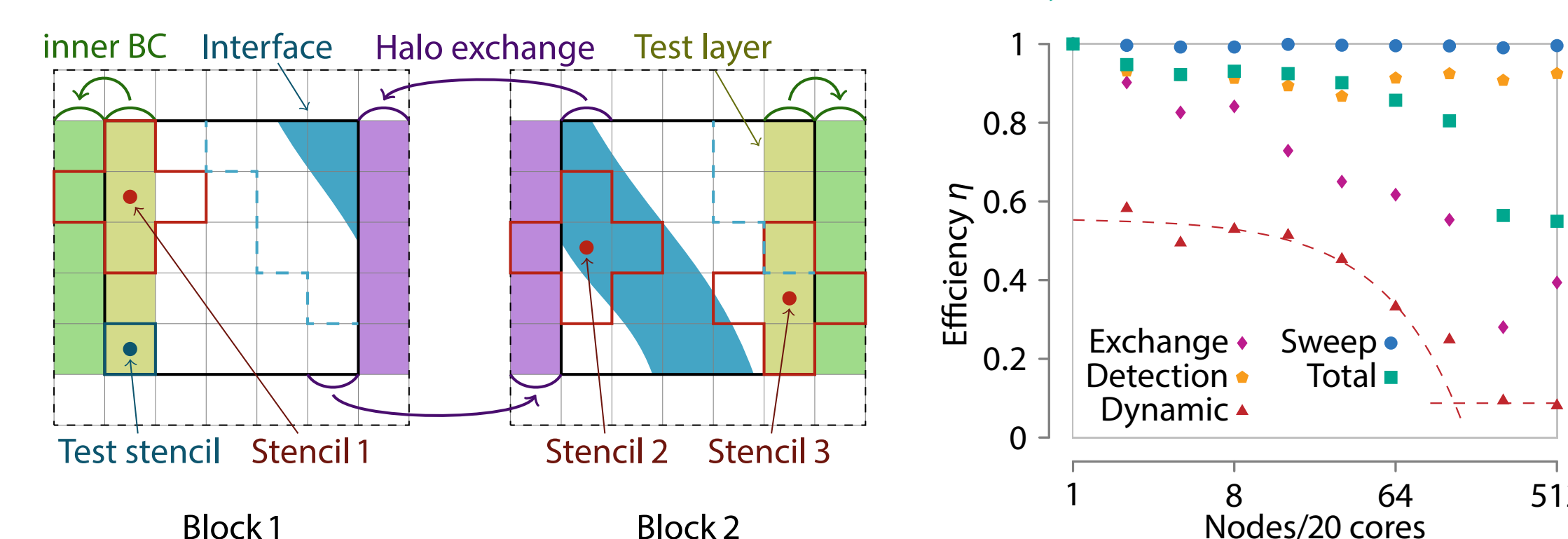
Diameter for Given Network Degree

Number of processes	System	2	3	4	5	6	7	8	9	10
10 240	ForHLR II prod.	102	33	21	16	15	14	13	13	13
22 960	ForHLR II whole	152	43	25	19	17	15	15	14	14
114 486	JUWELS Module 1	339	74	37	26	21	19	18	17	17
185 088	Hazel Hen	431	86	42	29	23	20	19	18	17
361 760	Piz Daint	602	107	50	33	26	22	20	19	19
2 282 544	Summit	1511	198	78	48	35	29	25	23	22
10 649 600	Sunway TaihuLight	3264	331	115	64	45	36	31	28	26

- the minimum block size requires additional 7 hops to establish the communication

Block Adaption

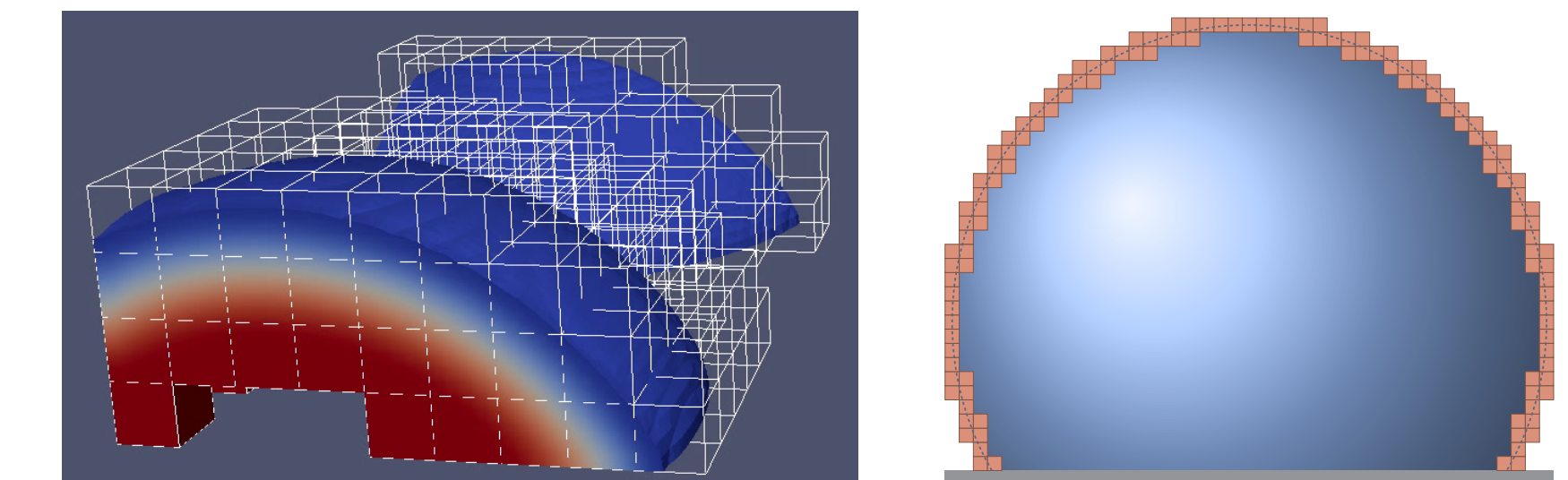
Dynamic Block Adaption with Excellent Scalability



- a maximum of 125 connections per block $\rightarrow$ constant communications for many cores
- results in a scalability with an efficiency of 60 % on 10 240 cores
- by saving a total of 96.5 % unnecessary blocks

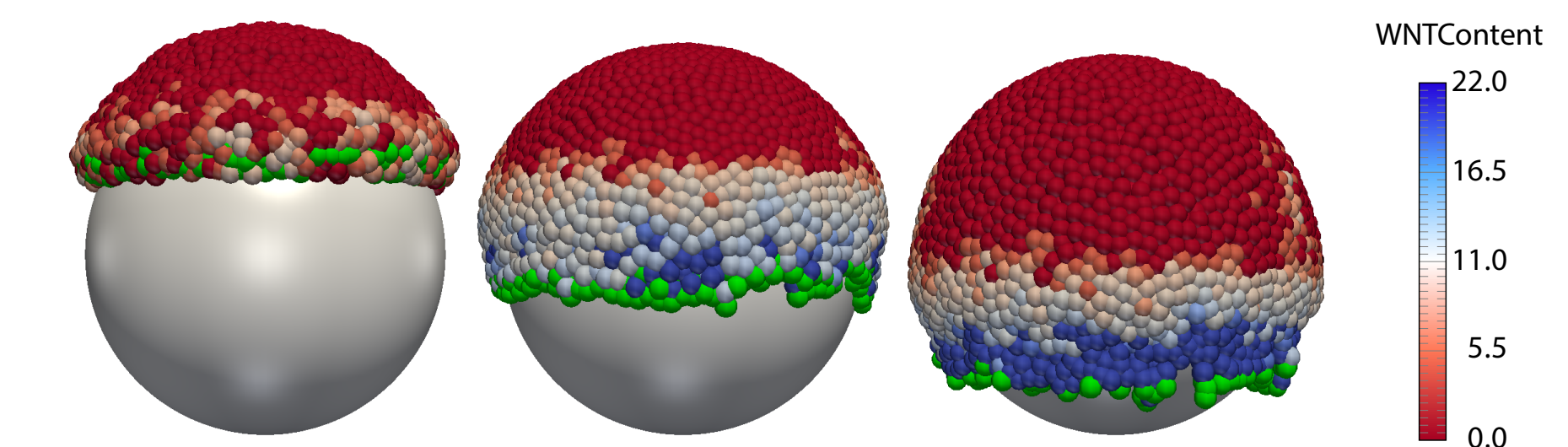
Applications

Phase-Field Method



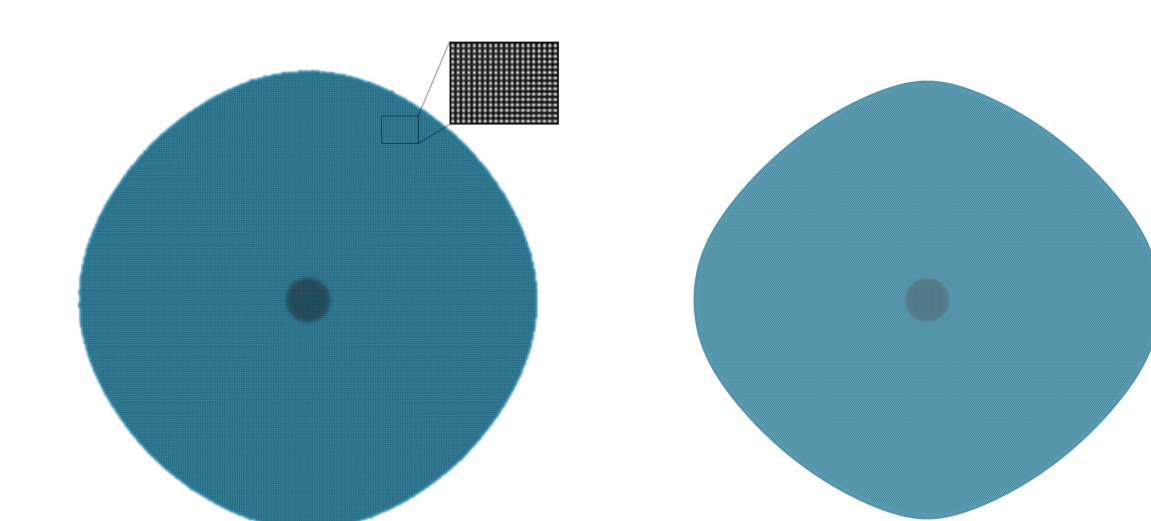
- water droplet on 700 lamellae of about 5 μm width each
- diffuse phase-field interface needs 15 cells resolution per lamella: 10^{12} grid points in total
- for a 2D cut, 177 of 1462 blocks of the plane are enough to cover the interface
- only 3.5 % of blocks with an edge length of 25 are needed to cover the interface
- cooperation with M. Ben Said (IBG, KIT)

Cellular Potts Model



- dynamic simulation of biological cells: tissues development and growth
- applicable for cancer, artery and stem cell specialization
- example epiboly movement of zebrafish: 1000^3 voxel on 1000 cores in 80 min
- cooperation with J. Rosenbauer and A. Schug (FZ Jülich)

Phase-Field Crystal



- anisotropy of crystal growth in comparison to molecular dynamics and phase-field method
- parallel semi-implicit spectral method for phase-field crystal models up to the eighth order
- cooperation with M. Guerdane (IAM, KIT)

Publications

- [1] M. Berghoff, I. Kondov, J. Hötzer, *Massively Parallel Stencil Code Solver with Autonomous Adaptive Block Distribution*, IEEE Transactions on Parallel and Distributed Systems 29(10), 2282–2296 (2018).
- [2] M. Guerdane, M. Berghoff, *Crystal-Melt Interface Mobility and its Anisotropy in bcc Fe: Linking Molecular Dynamics to Phase-Field and Phase-Field Crystal Modeling* Physical Review B 97, 144105 (2018).
- [3] M. Berghoff, I. Kondov, *Scalable Global Network Based on Local Non-Collective Communications*, SC Workshop: Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA'18).

Contact
 marco.berghoff@kit.edu
 @NASTJAsolver

