FD4: A Framework for Highly Scalable Dynamic Load Balancing and Model Coupling

Symposium on HPC and Data-Intensive Applications in Earth Sciences

13 Nov 2014, Trieste, Italy

Matthias Lieber (matthias.lieber@tu-dresden.de)

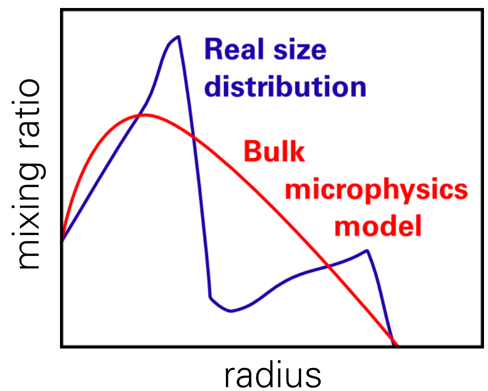
Center for Information Services and High Performance Computing (ZIH)
Technische Universität Dresden, Germany

"Climate models now include more cloud and aerosol processes, and their interactions, than at the time of the AR4, but there remains low confidence in the representation and quantification of these processes in models."

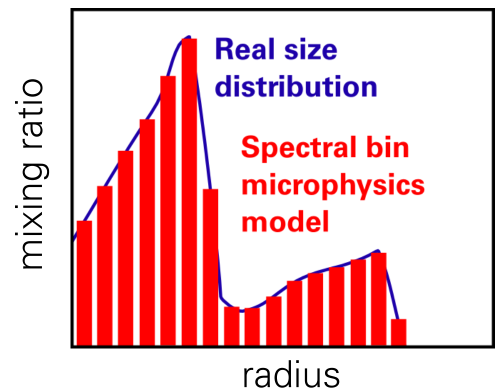
IPCC, 2013: Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Motivation: Spectral Bin Cloud Microphysics Schemes

Widely used bulk models



Spectral bin microphysics



- Bin discretization of cloud particle size distribution
- Allows more detailed modeling of interaction between aerosols, clouds, and precipitation
- Computationally too expensive for forecast
- Only used for process studies up to now

Lynn et al., Mon. Weather Rev., 133:59-71, 2005

Grützun et al., Atmos. Res., 90(2-4):233-242, 2008

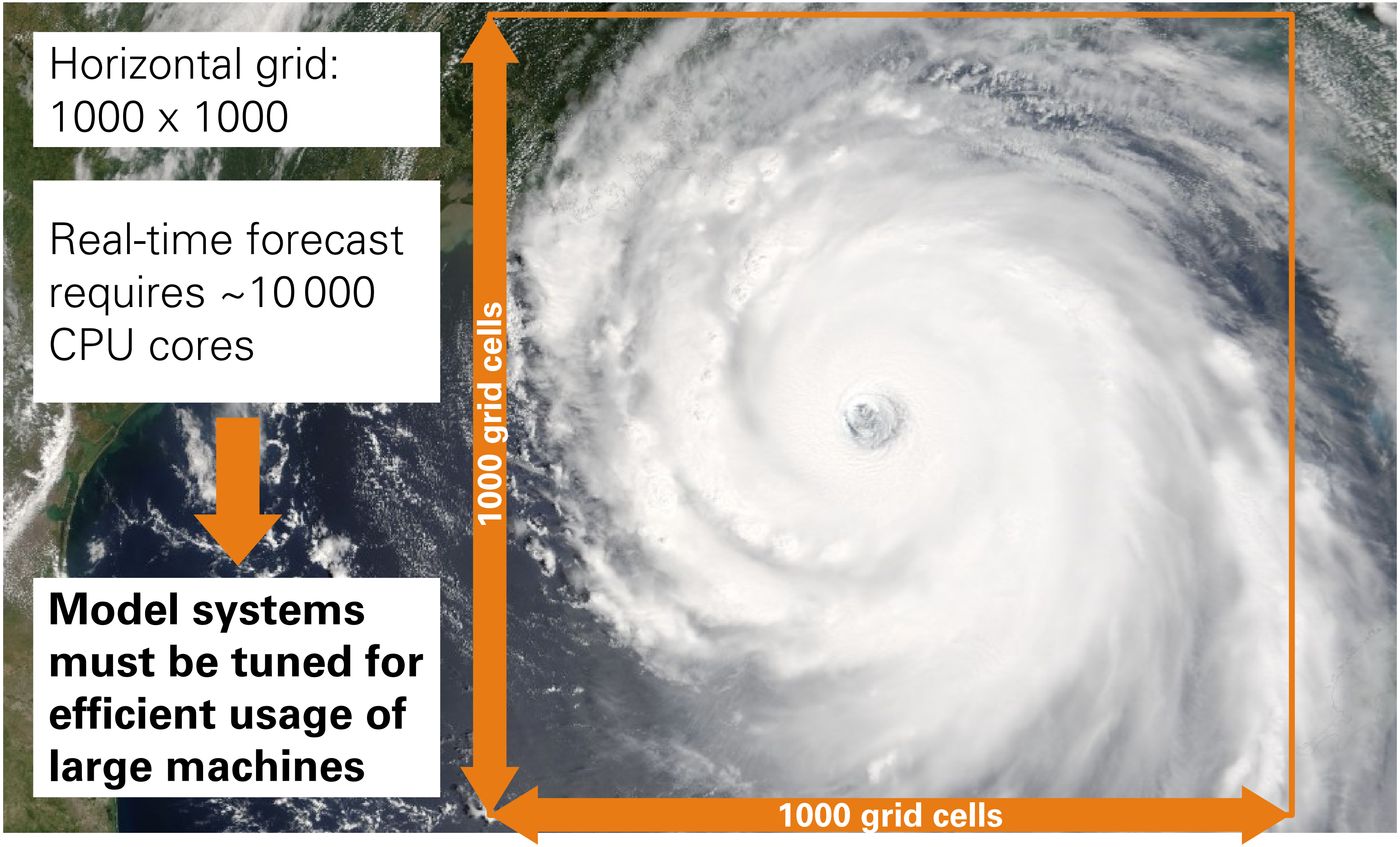
Khain et al., J. Atmos. Sci., 67(2):365-384, 2010

Sato et al., J. Atmos. Sci., 69:2012-2030, 2012

Planche et al., Quart. J. Roy. Meteor. Soc. Vol. 140, No. 683, 2014

Fan et al., Atmos. Chem. Phys., 14:81-101, 2014

Motivation: Tropical Cyclone Forecast with SBM?



Horizontal grid:
1000 x 1000

Real-time forecast requires ~10 000 CPU cores

Model systems must be tuned for efficient usage of large machines

Outline

- Bottleneck Analysis
- Concept of Load-balanced Coupling
- FD4's Features
- Benchmarks
- Conclusion

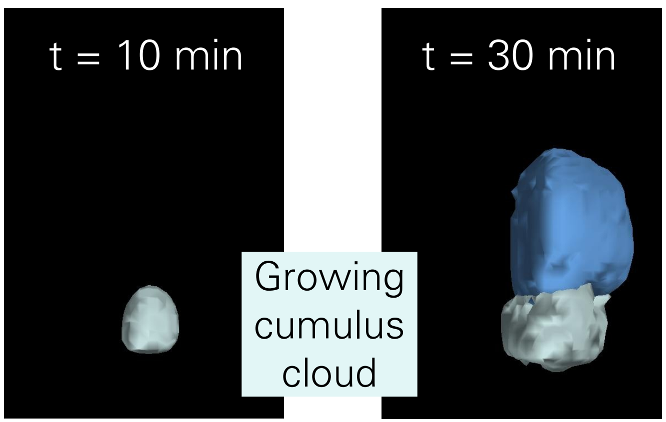
FD4 Motivation: COSMO-SPECS Performance

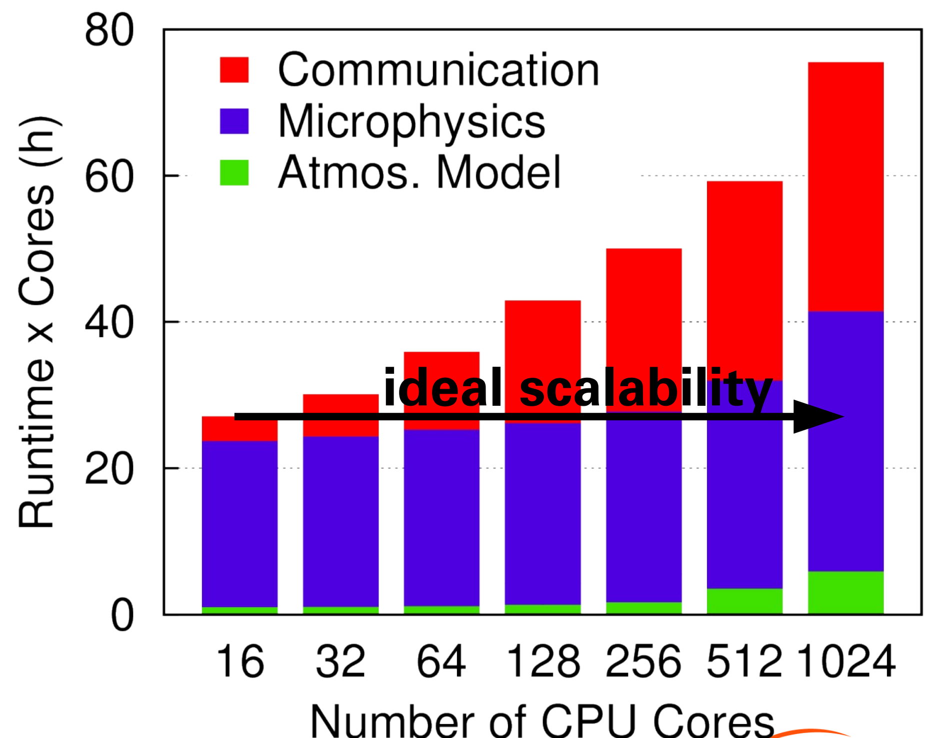
- COSMO-SPECS: Atmospheric model COSMO extended with highly detailed cloud microphysics model SPECS

COSMO CONSORTIUM FOR SMALL SCALE MODELING

TROPOS Leibniz Institute for Tropospheric Research

Small 3D case with 64x64x48 grid



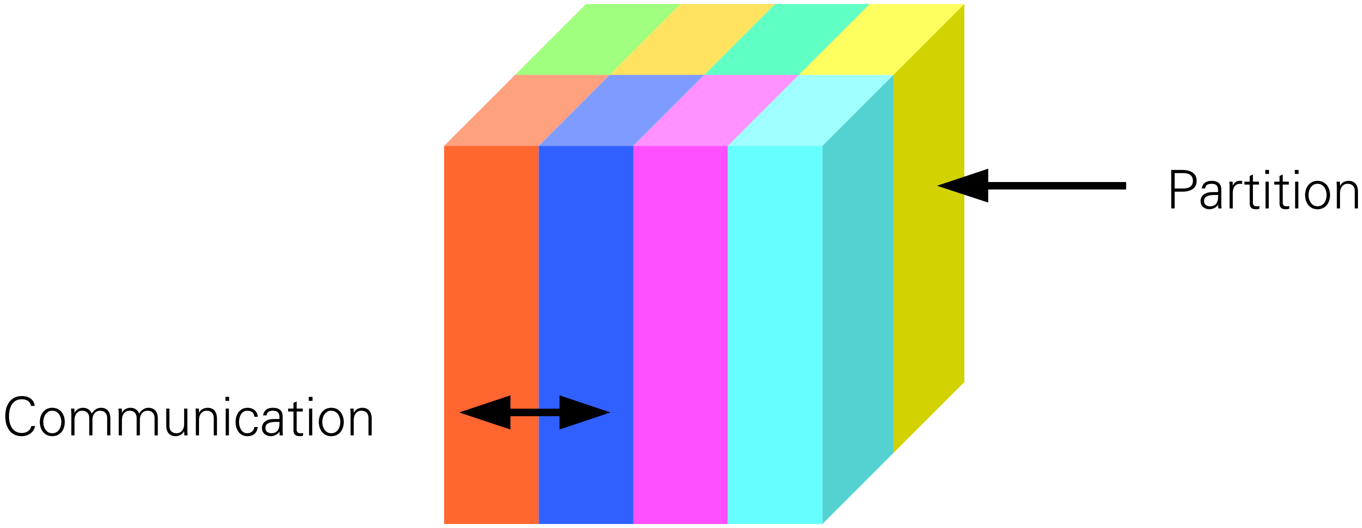


Lieber et al., Highly Scalable Dynamic Load Balancing in the Atmospheric Modeling System COSMO-SPECS+FD4, PARA 2010, 2012

TECHNISCHE UNIVERSITÄT DRESDEN

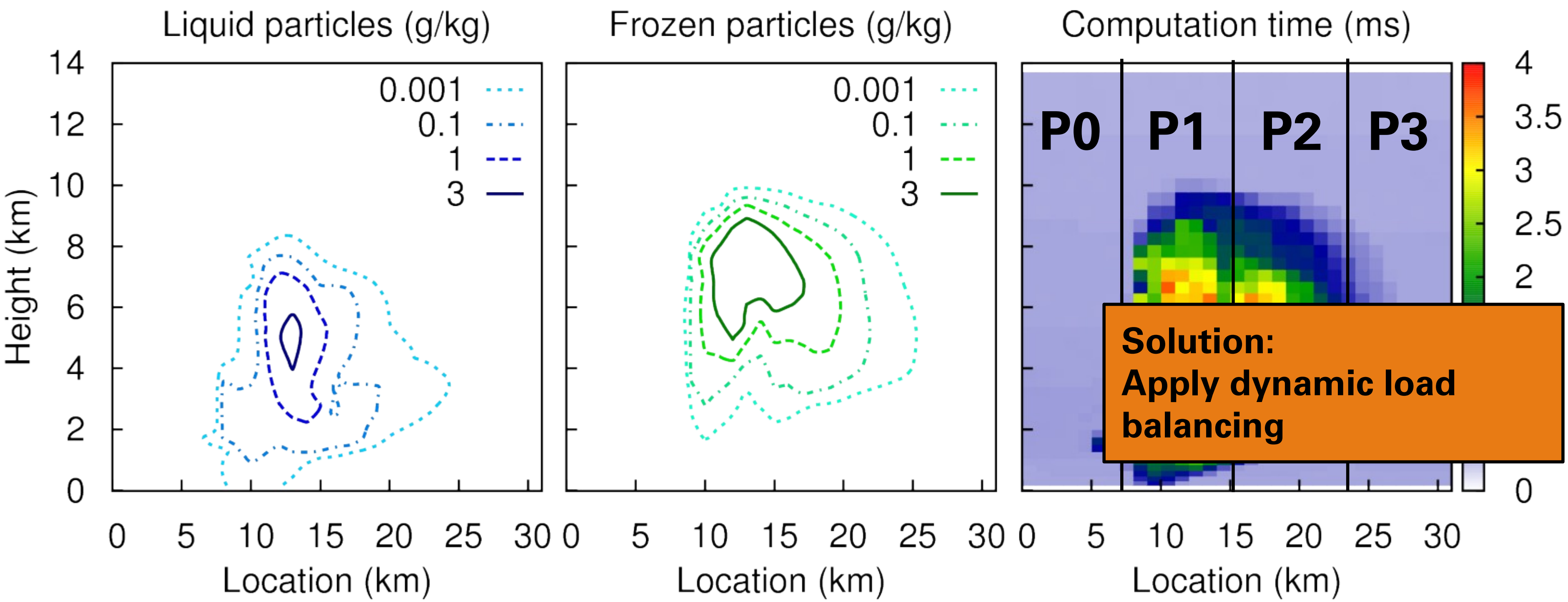
ZIH Center for Information Services & High Performance Computing

Analysis: Common Parallelization Scheme



- 3D domain partitioned into rectangular boxes
- 2D decomposition (horizontal dimensions)
- Regular communication with 4 direct neighbors required (periodic boundary conditions)
- Based on MPI (Message Passing Interface)

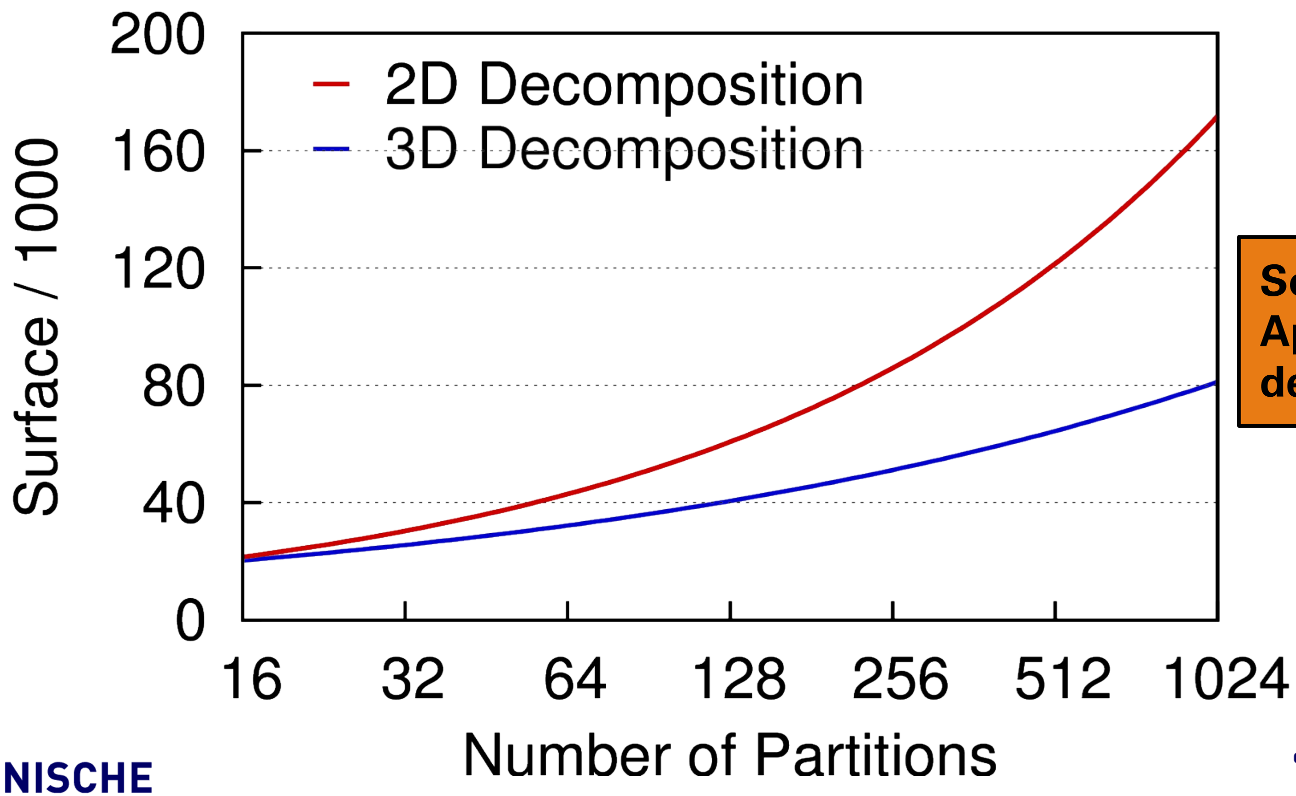
Analysis: Load Imbalance due to Microphysics



- SPECS computing time varies strongly depending on the range of the particle size distribution and presence of frozen particles
- Leads to load imbalances between partitions

Analysis: Increasing Communication Volume

- Surface-to-volume-ratio of partitions grows with number of partitions, in theory (best case):
  - 2D decomposition: $A^{2D}(P) = 4\ G^{2/3}\ P^{1/2} \sim P^{1/2}$
  - 3D decomposition: $A^{3D}(P) = 6\ G^{2/3}\ P^{1/3} \sim P^{1/3}$



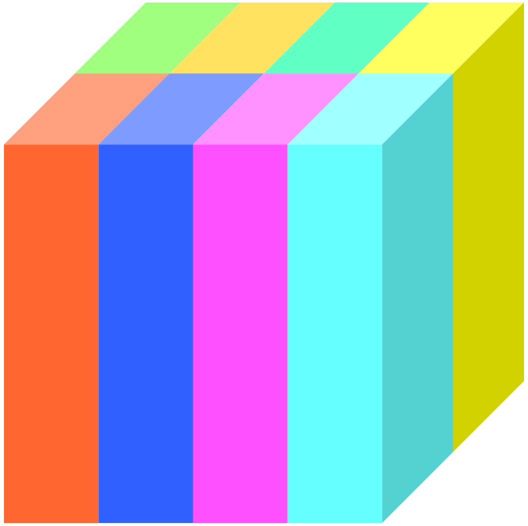
Solution: Apply 3D decomposition

Outline

- Bottleneck Analysis
- **Concept of Load-balanced Coupling**
- FD4's Features
- Benchmarks
- Conclusion

Concept of Load-Balanced Coupling

Atmospheric Model & Spectral Bin Microphysics



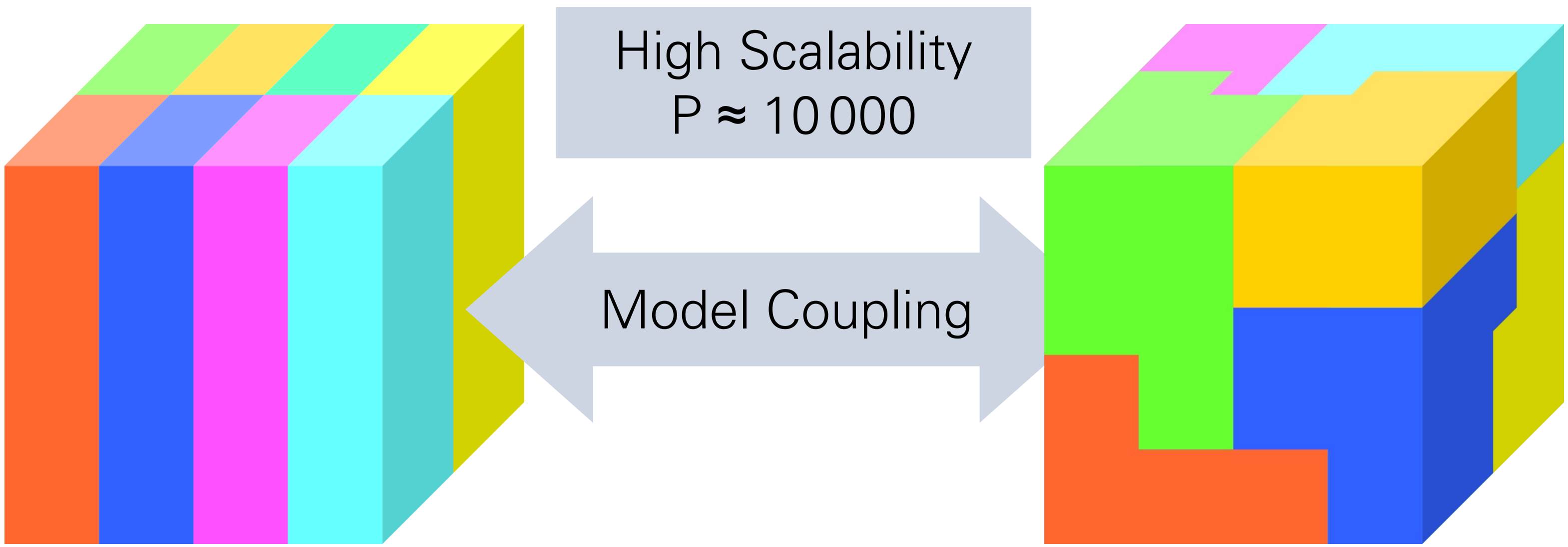
2D Decomposition

Static Partitioning

Lieber et al., Highly Scalable Dynamic Load Balancing in the Atmospheric Modeling System COSMO-SPECS+FD4, PARA 2010, 2012

Concept of Load-Balanced Coupling

Atmospheric Model & Spectral Bin Microphysics **Spectral Bin Microphysics**



Atmospheric Model	Spectral Bin Microphysics
2D Decomposition	Block-based 3D Decomposition
Static Partitioning	Dynamic Load Balancing
	Optimized Data Structures

Lieber et al., Highly Scalable Dynamic Load Balancing in the Atmospheric Modeling System COSMO-SPECS+FD4, PARA 2010, 2012

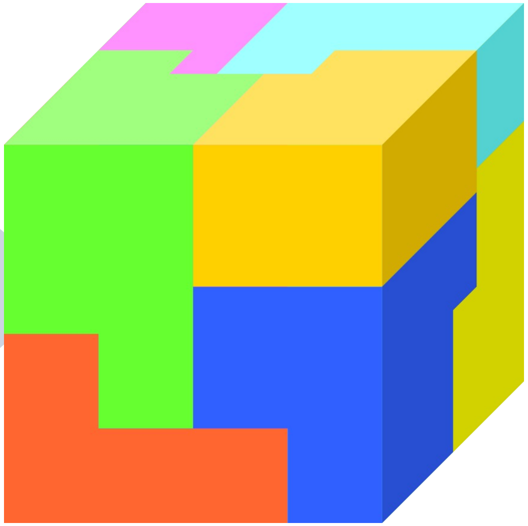
Concept of Load-Balanced Coupling

Implemented as independent framework FD4

FD4:
Four-Dimensional
Distributed
Dynamic
Data structures

High Scalability
$P \approx 10\,000$

Model Coupling



Block-based 3D Decomposition

Dynamic Load Balancing

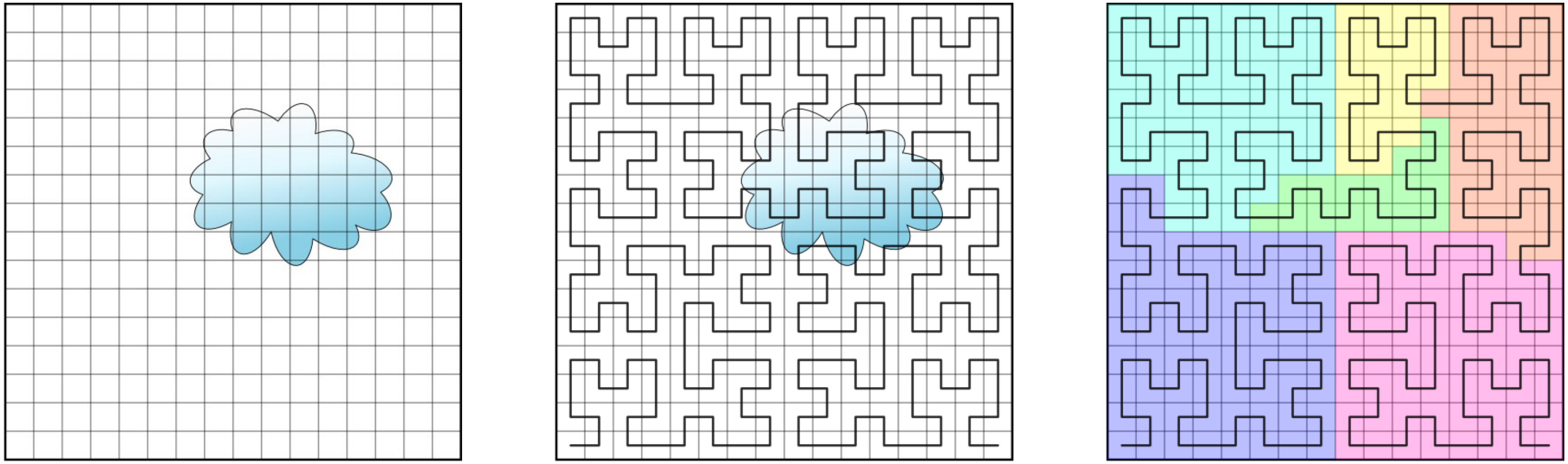
Optimized Data Structures

Lieber et al., Highly Scalable Dynamic Load Balancing in the Atmospheric Modeling System COSMO-SPECS+FD4, PARA 2010, 2012

Outline

- Bottleneck Analysis
- Concept of Load-balanced Coupling
- **FD4's Features**
- Benchmarks
- Conclusion

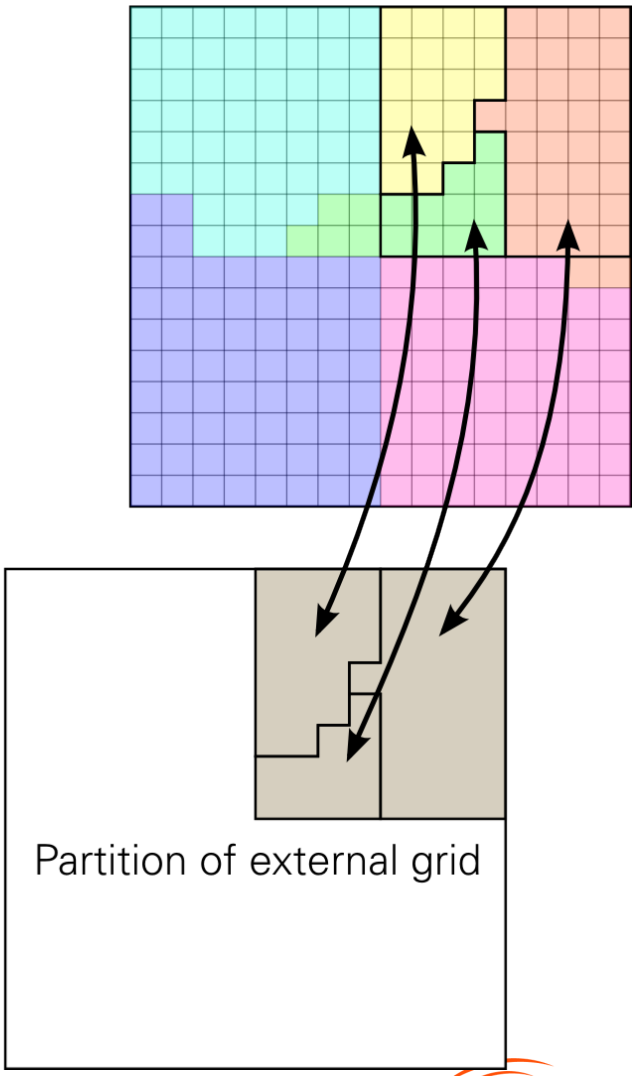
FD4: Dynamic Load Balancing



- 3D block decomposition of rectangular grid
- Space-filling curve (SFC) partitioning to assign blocks to ranks
- SFC reduces 3D partitioning problem to 1D
- High locality of SFC leads to moderate comm. costs
- Developed a highly scalable, hierarchical method for high-quality 1D partitioning of the SFC-indexed blocks

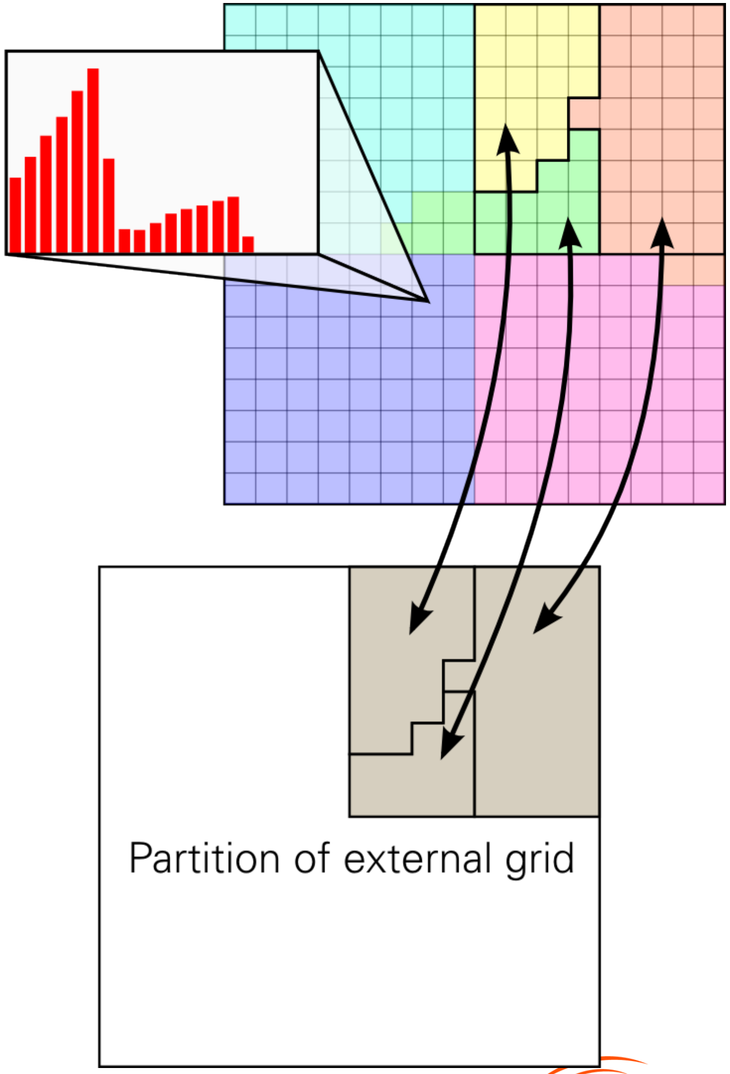
FD4: Model Coupling

- Data exchange between FD4 based model and an external model
  - E.g. weather or CFD model
  - Transfer in both directions
- FD4 computes partition overlaps after each repartitioning of FD4 grid
  - Highly scalable algorithm
- No grid transformation / interpolation
  - External model must provide data matching the FD4 grid
- "Sequential" coupling only
  - Both models run alternately on same set of MPI ranks



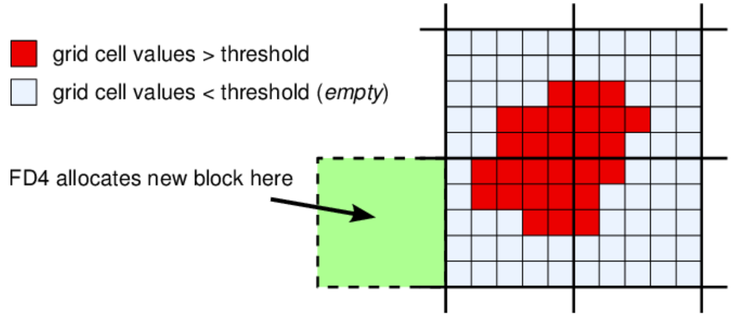
FD4: 4th Dimension

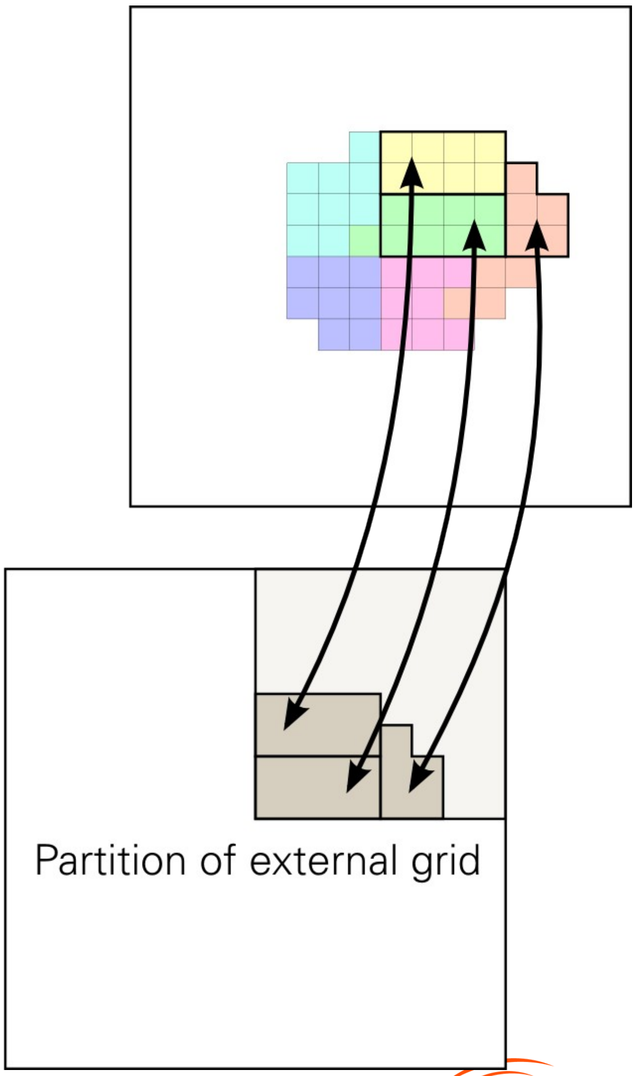
- Extra, non-spatial dimension of grid variables, e.g.
  - Size resolving models
  - Array of gas phase tracers
- FD4 is optimized for a large 4th dimension
- COSMO-SPECS requires
  2 x 11 x 66 ~ 1500 values



FD4: Adaptive Block Mode

- Grid allocation adapts to spatial structure of simulated problem
  - Save memory in case data and computations are required for a subset only
- For multiphase problems like drops, clouds, flame fronts
- FD4 ensures existence of all blocks required for correct stencil operations



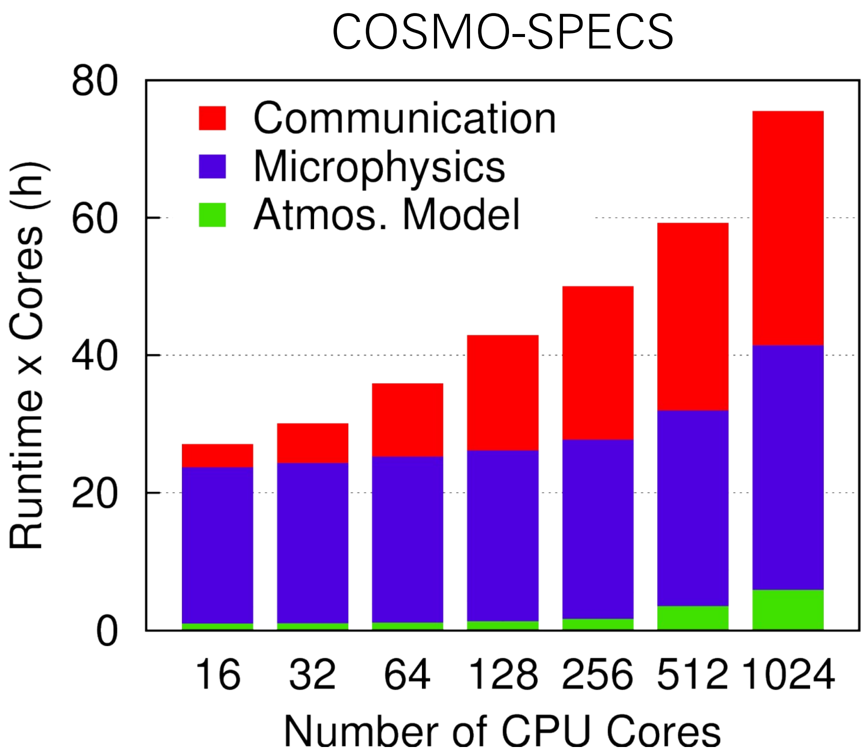


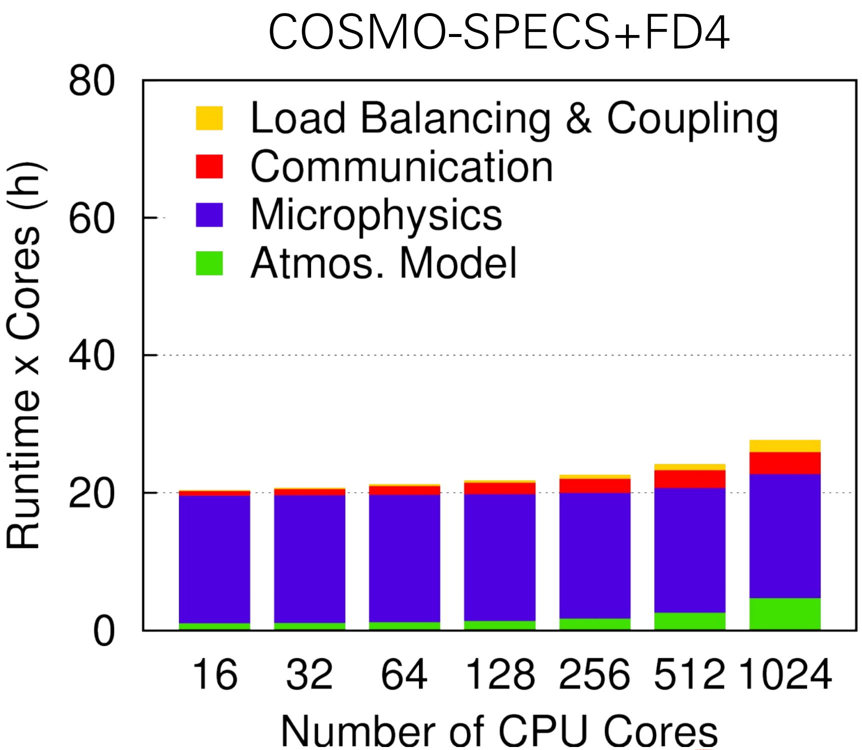
Outline

- Bottleneck Analysis
- Concept of Load-balanced Coupling
- FD4's Features
- **Benchmarks**
- Conclusion

Benchmarks: COSMO-SPECS Performance Comparison

- Almost 3 times faster at 1024 CPU cores
- Load balancing & coupling scale well, but can we reach > 10 000 processes?



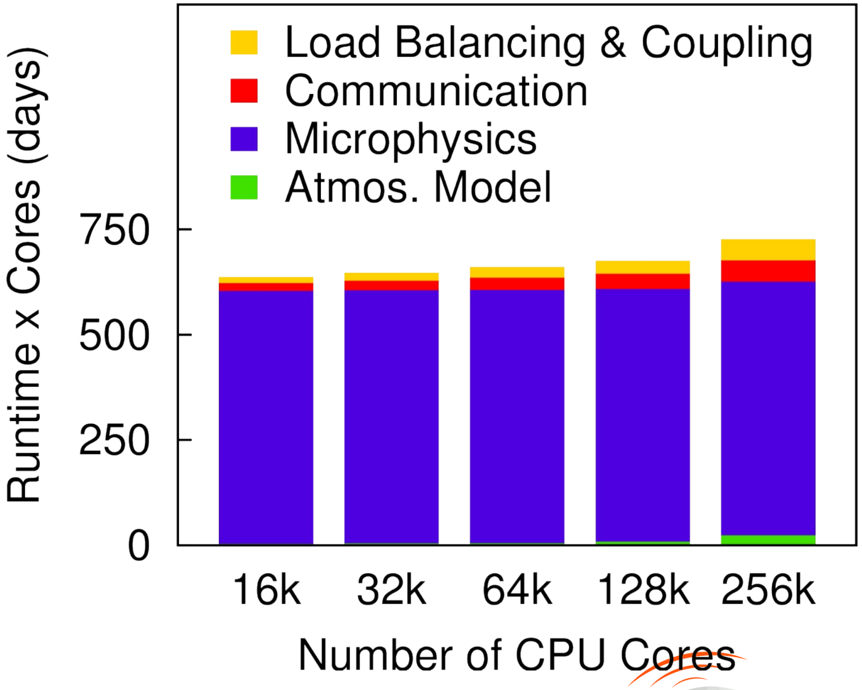


Benchmarks: Scalability on Blue Gene/Q

- Grid size: 1024 x 1024 x 48 grid cells, > 3M blocks
- 256k: 30 min forecast in <5min (w/o init and I/O)
- Runs on Blue Gene/Q with up to 262 144 MPI ranks
- 14 x speed-up from 16k to 256k

Lieber, Nagel, Mix, *Scalability Tuning of the Load Balancing and Coupling Framework FD4*, NIC Symposium 2014, pp. 363-370.

COSMO-SPECS+FD4



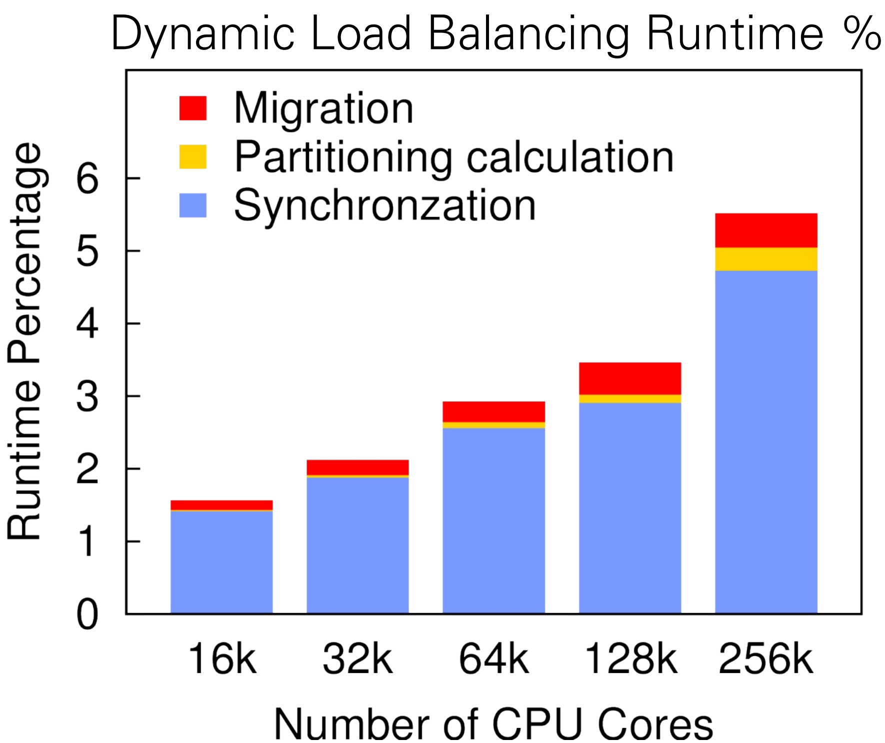
TECHNISCHE UNIVERSITÄT DRESDEN

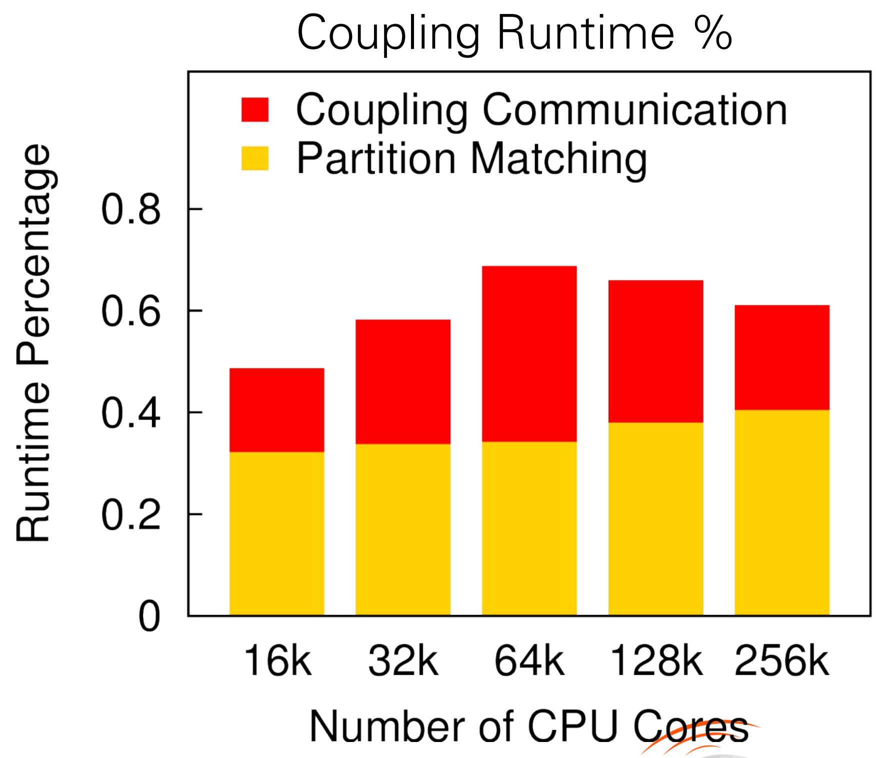
ZIH Center for Information Services & High Performance Computing

Benchmarks: Load Balancing & Coupling Scalability

- Grid size: 1024 x 1024 x 48 grid cells, > 3M blocks
- Load balancing scales comparatively very well
- Coupling scales nearly perfect

Lieber, Nagel, Mix, *Scalability Tuning of the Load Balancing and Coupling Framework FD4*, NIC Symposium 2014, pp. 363-370.





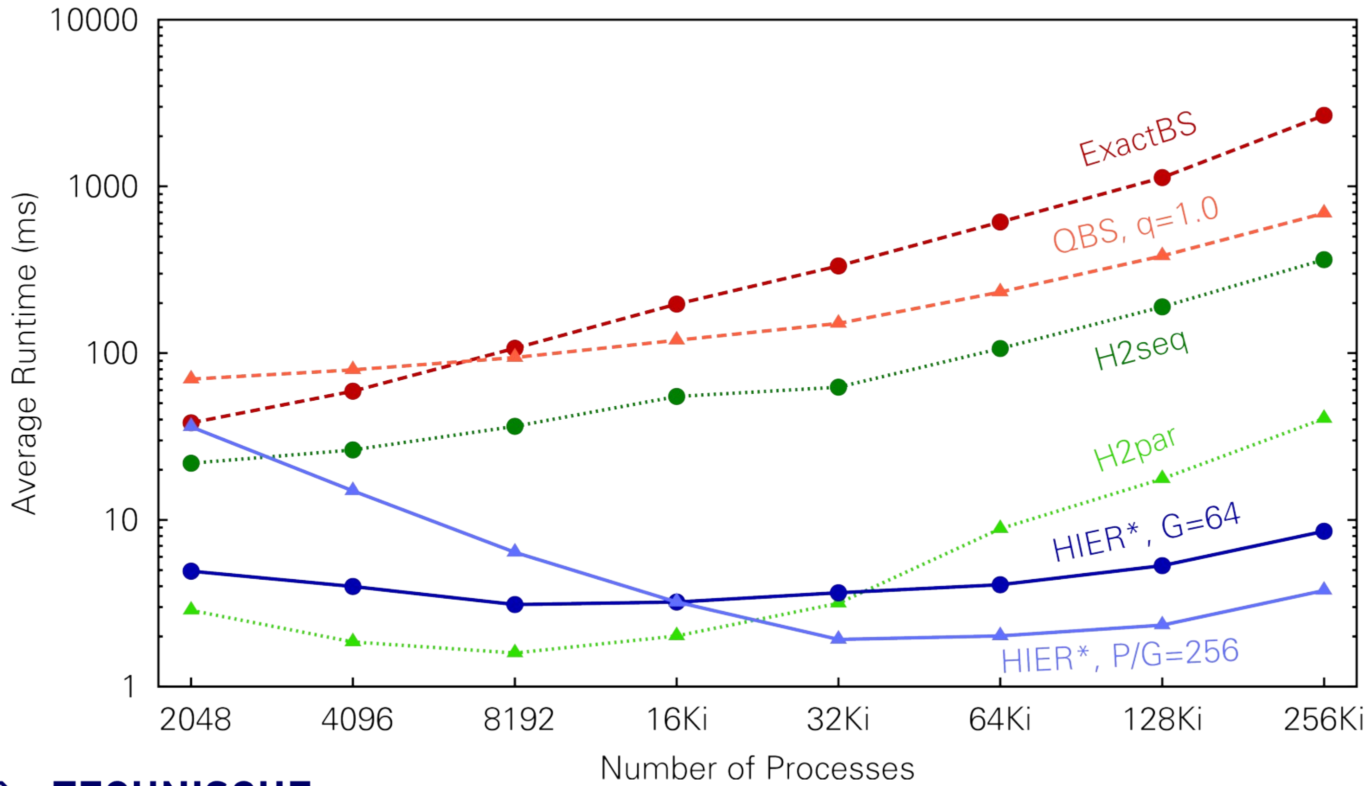
TECHNISCHE UNIVERSITÄT DRESDEN

ZIH Center for Information Services & High Performance Computing

Benchmarks: 1D Partitioning Comparison on Blue Gene/Q

- ExactBS: exact method, but slow and serial
- H2: fast heuristic, but may result in poor load balance
- HIER*: hierarchical algorithm implemented in FD4, achieves nearly optimal load balance

Lieber, Nagel, *Scalable High-Quality 1D Partitioning*, HPCS 2014, pp. 112-119, 2014



ExactBS: 2668 ms

QBS: 692 ms

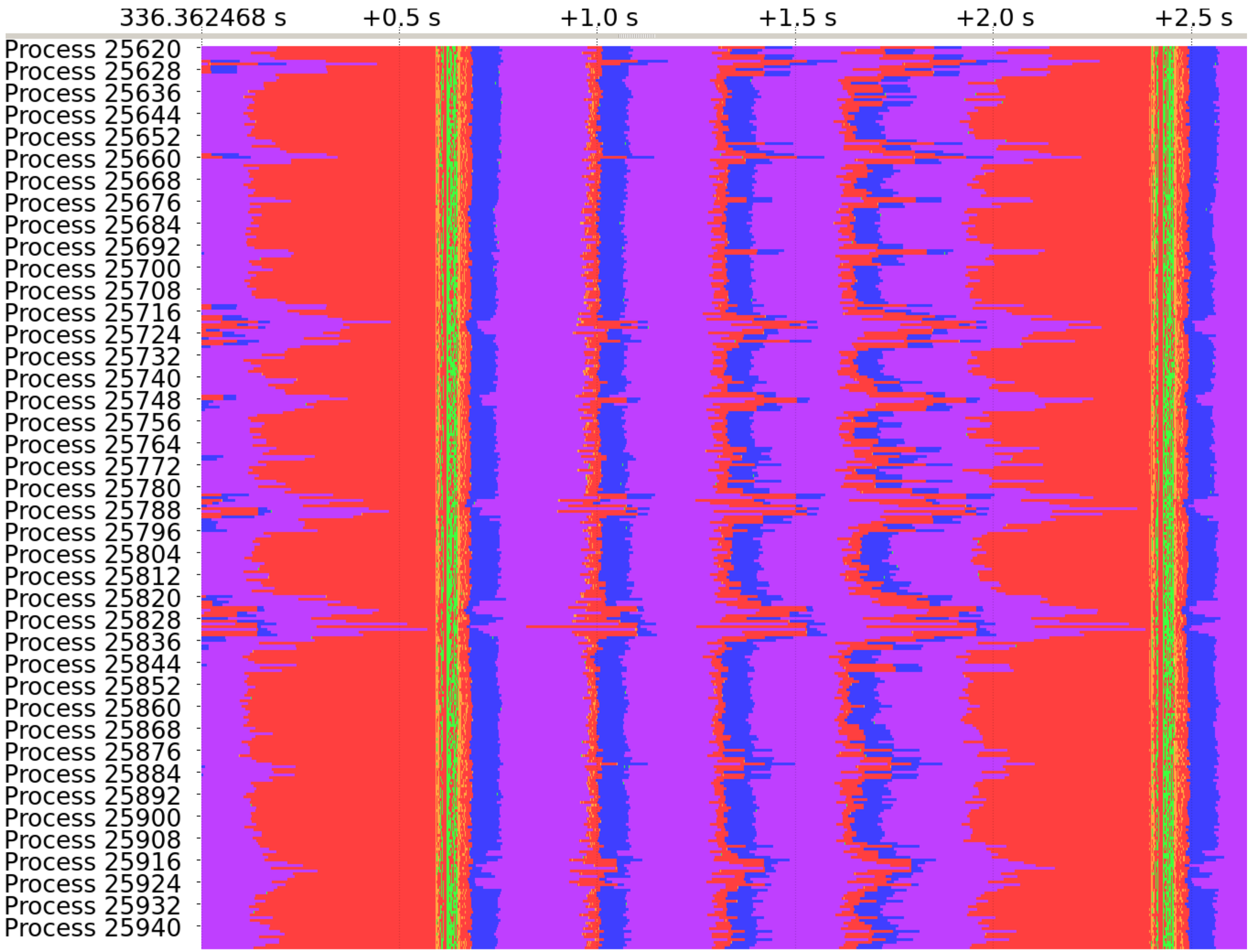
H2seq: 363 ms

H2par: 40.5 ms

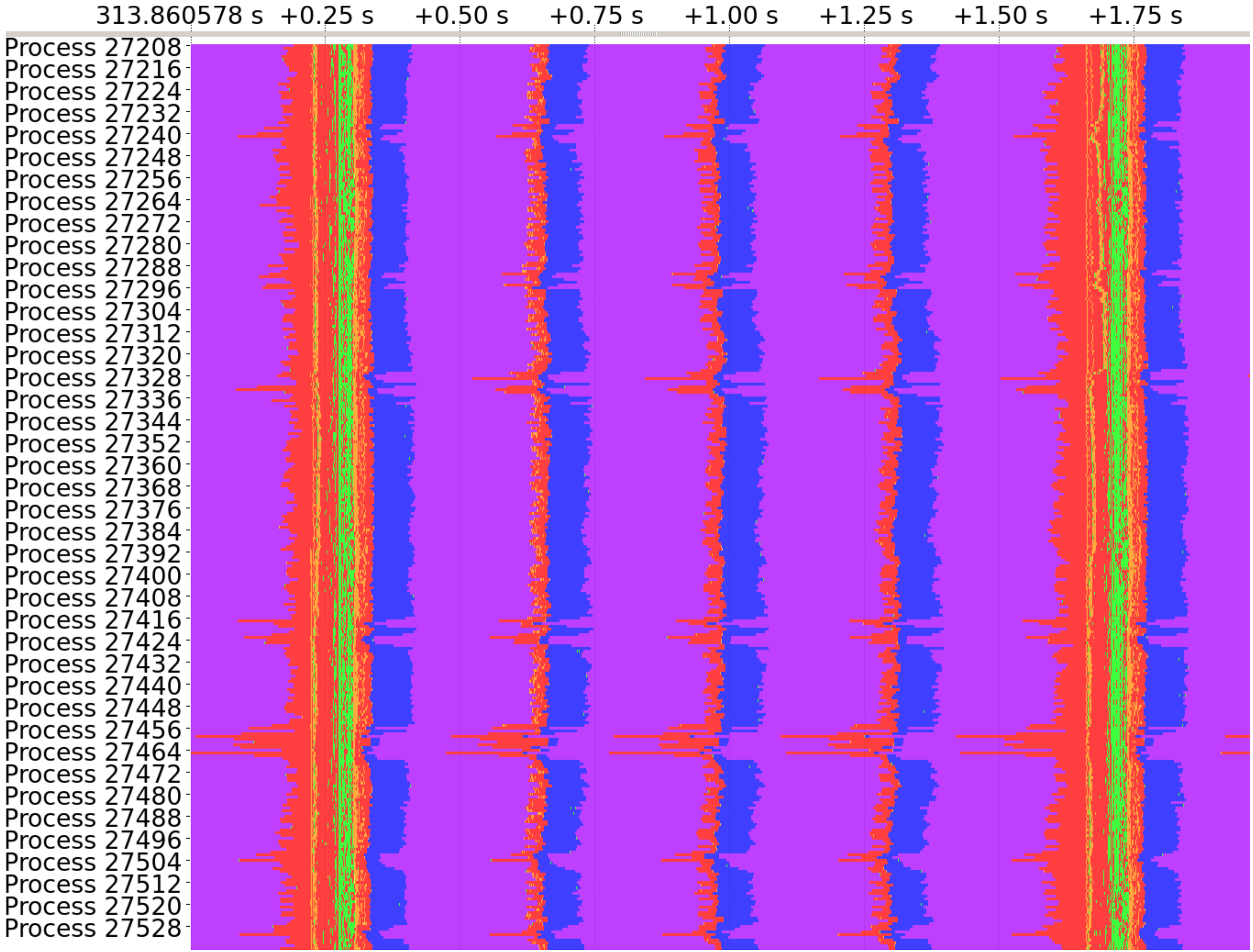
$HIER^{*}_{G=64}$: 8.55 ms

$HIER^{*}_{P/G=256}$: 3.77 ms

Heuristic H2 in Action (COSMO-SPECS+FD4)



HIER* in Action (COSMO-SPECS+FD4)



Outline

- Bottleneck Analysis
- Concept of Load-balanced Coupling
- FD4's Features
- Benchmarks
- **Conclusion**

Conclusions

- FD4 provides for simulation models
  - Parallelization of numerical grid
  - Communication between neighbor partitions
  - Dynamic load balancing
  - Model coupling
  - High scalability
- Initially developed for atmospheric modeling, but generally applicable
- FD4 is available as open source software
  - Fortran 95, MPI-2, NetCDF
  - Tested on many different HPC systems

FD4 website: http://wwwpub.zih.tu-dresden.de/~mlieber/fd4

Lieber, Nagel, *Scalable High-Quality 1D Partitioning*, HPCS 2014, pp. 112-119, 2014

Lieber, Nagel, Mix, *Scalability Tuning of the Load Balancing and Coupling Framework FD4*, NIC Symposium 2014

Lieber et al., Highly Scalable Dynamic Load Balancing in the Atmospheric Modeling System COSMO-SPECS+FD4, PARA 2010, 2012

Lieber et al., *FD4: A Framework for Highly Scalable Load Balancing and Coupling of Multiphase Models*, ICNAAM 2010

Thank you very much for your attention!

Acknowledgments

Verena Grützun, Ralf Wolke,
Oswald Knoth, Martin Simmel,
René Widera, Matthias Jurenz,
Matthias Müller, Wolfgang E. Nagel

lrz Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

VAMPIR www.vampir.eu

TROPOS
Leibniz Institute for
Tropospheric Research

DWD

CONSORTIUM FOR SMALL SCALE MODELING
COSMO

PIConGPU

HZDR
HELMHOLTZ
ZENTRUM DRESDEN
ROSSENDORF

www.tropos.de

www.cosmo-model.org

picongpu.hzdr.de

Funding

DFG Deutsche Forschungsgemeinschaft

cfaed CENTER FOR ADVANCING ELECTRONICS DRESDEN

Europa fördert Sachsen.
ESF
Europäischer Sozialfonds

IMDATA

TECHNISCHE UNIVERSITÄT DRESDEN

ZIH
Center for Information Services & High Performance Computing

Backup Slides

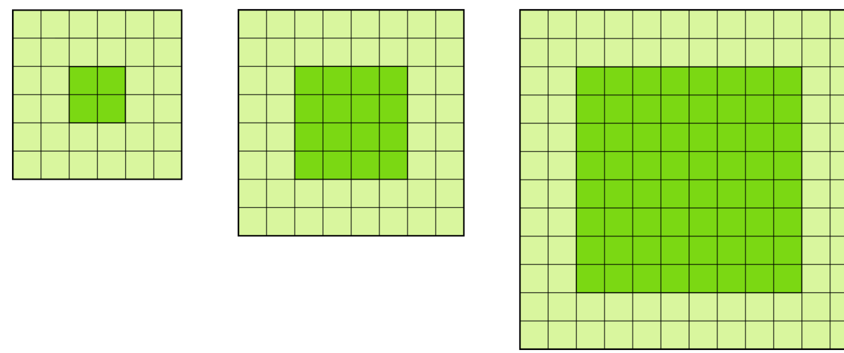
Framework FD4: Optimized Data Structure

- A large number of small blocks are good for performance:
  - Size-resolved approach / ~1000 variables per grid cell: Only small blocks do not exceed processor cache
  - Load balancing: #blocks > #partitions to enable fine-grained balancing
- Additional memory costs for a boundary of ghost cells
  - Too high for small blocks!
- Add ghost blocks at the partition borders only

2^2: 800%
2^3: 2600%

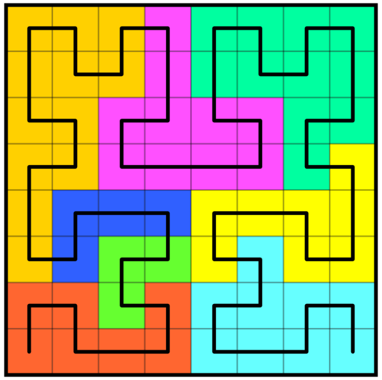
4^2: 300%
4^3: 700%

8^2: 125%
8^3: 237%



From SFC Partitioning to 1D Partitioning

- Space-filling curve (SFC) partitioning widely used
  - nD space is mapped to 1D by SFC
  - Mapping is fast and has high locality
  - Migration typically between neighbor ranks
- 1D partitioning is core problem of SFC partitioning
  - Decomposes task chain into consecutive parts
- Two classes of existing 1D partitioning algorithms:
  - Heuristics: fast, parallel, no optimal solution
  - Exact methods: slow, serial, but optimal

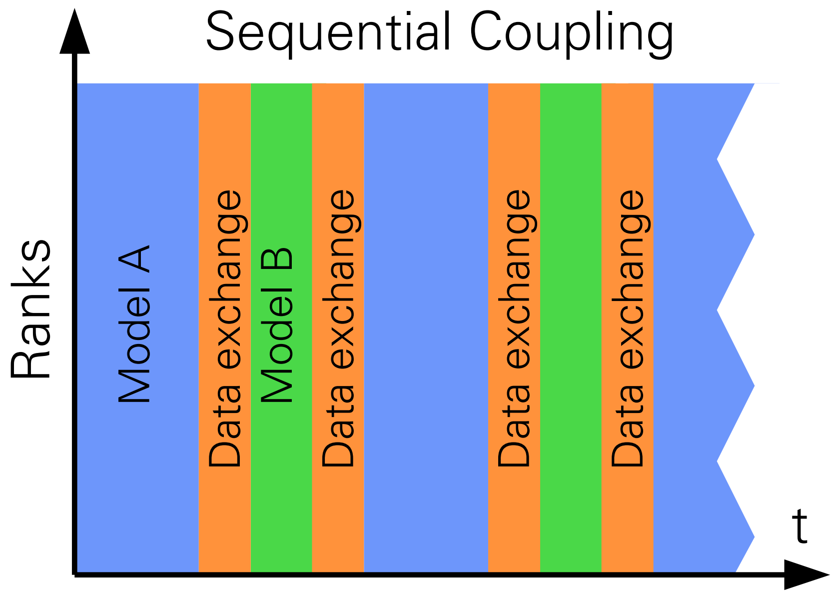


Hilbert SFC

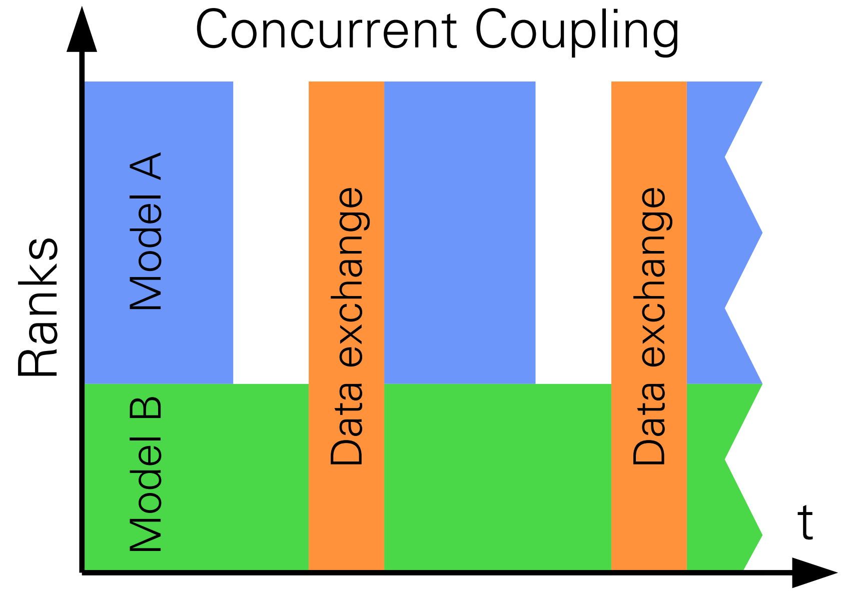
Pilkington, Baden, *Dynamic partitioning of non-uniform structured workloads with spacefilling curves*, IEEE T. Parall. Distr., vol. 7, no. 3, pp. 288-300, 1996.

Pinar, Aykanat, *Fast optimal load balancing algorithms for 1D partitioning*, J. Parallel Distr. Com., vol. 64, no. 8, pp. 974-996, 2004.

Sequential vs. Concurrent Model Coupling



- Both models run alternately on same set of MPI ranks
- Allows tight coupling (data dependencies)
- Avoids load imbalances between models

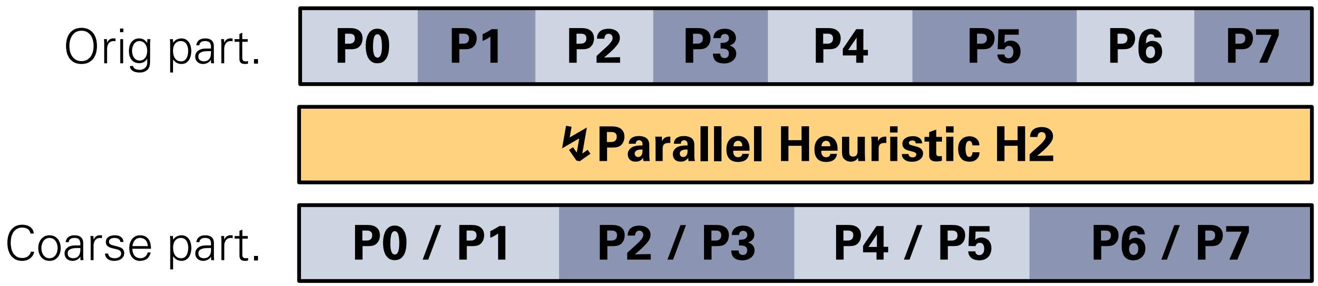


- MPI ranks are split into groups
- Loose coupling, codes may be separate
- Scales to higher total number of ranks

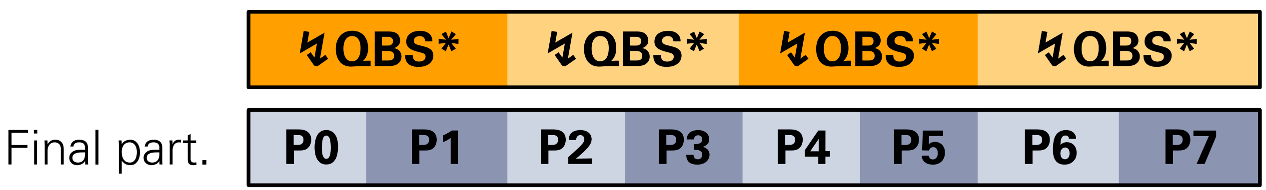
Scalable High-Quality 1D Partitioning: Algorithm HIER*

Large scale applications require a fully parallel method, i.e. without gathering all task weights

- Run parallel H2 to create G < P coarse partitions:



- Run G independent instances of exact QBS* (q=1.0) to create final partitions within each group:



- Parameter G allows trade-off between scalability (high G → heuristic dominates) and load balance (small G → exact method dominates)

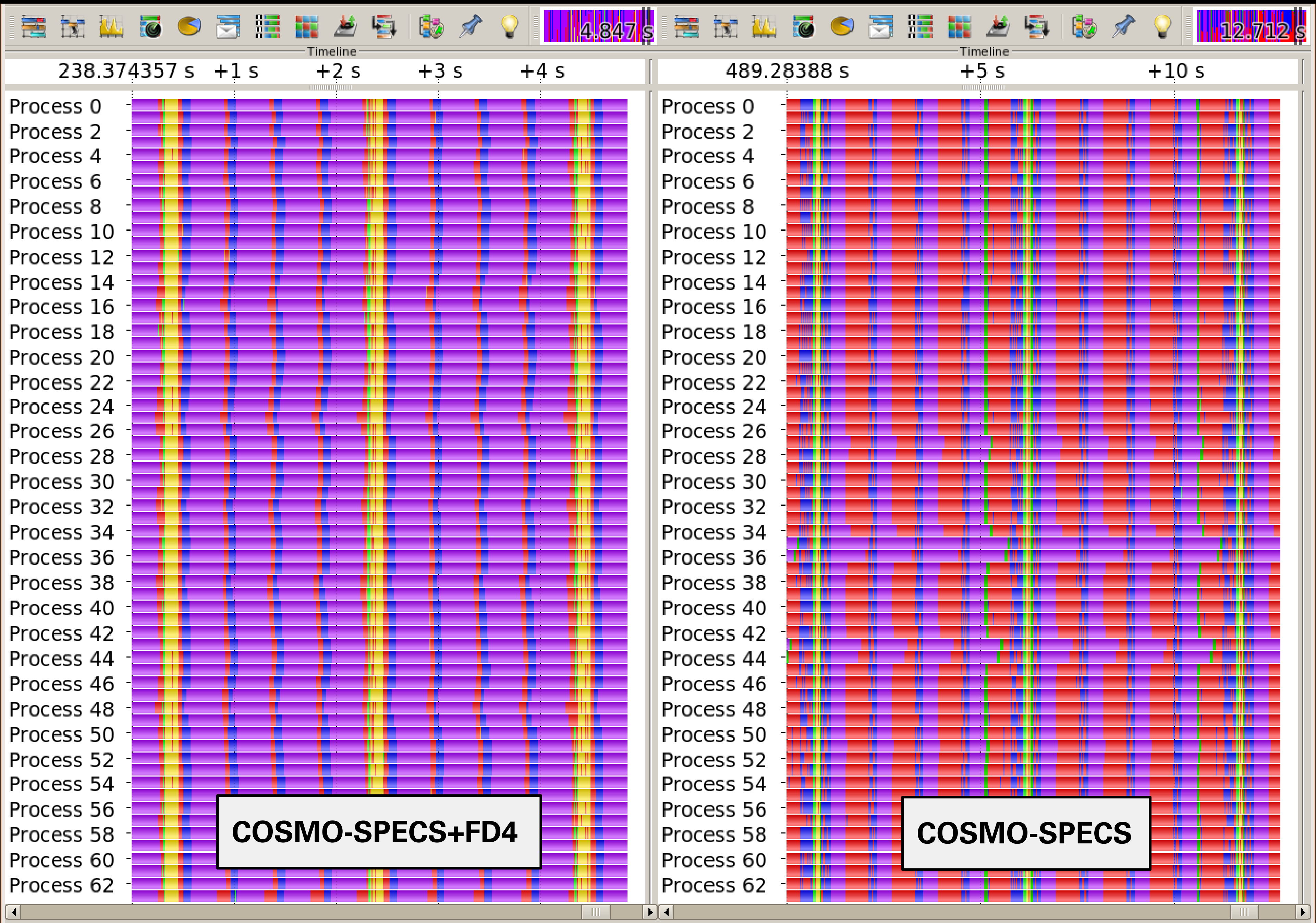
H2 nearly optimal if $w_{max} << W_N / P$: Miguet, Pierson, *Heuristics for 1D rectilinear partitioning as a low cost and high quality answer to dynamic load balancing*, LNCS, vol. 1225, 1997, pp. 550-564.

FD4: Implementation

- Implemented in Fortran 95
- MPI-based parallelization
- Open Source Software
- www.tu-dresden.de/zih/clouds

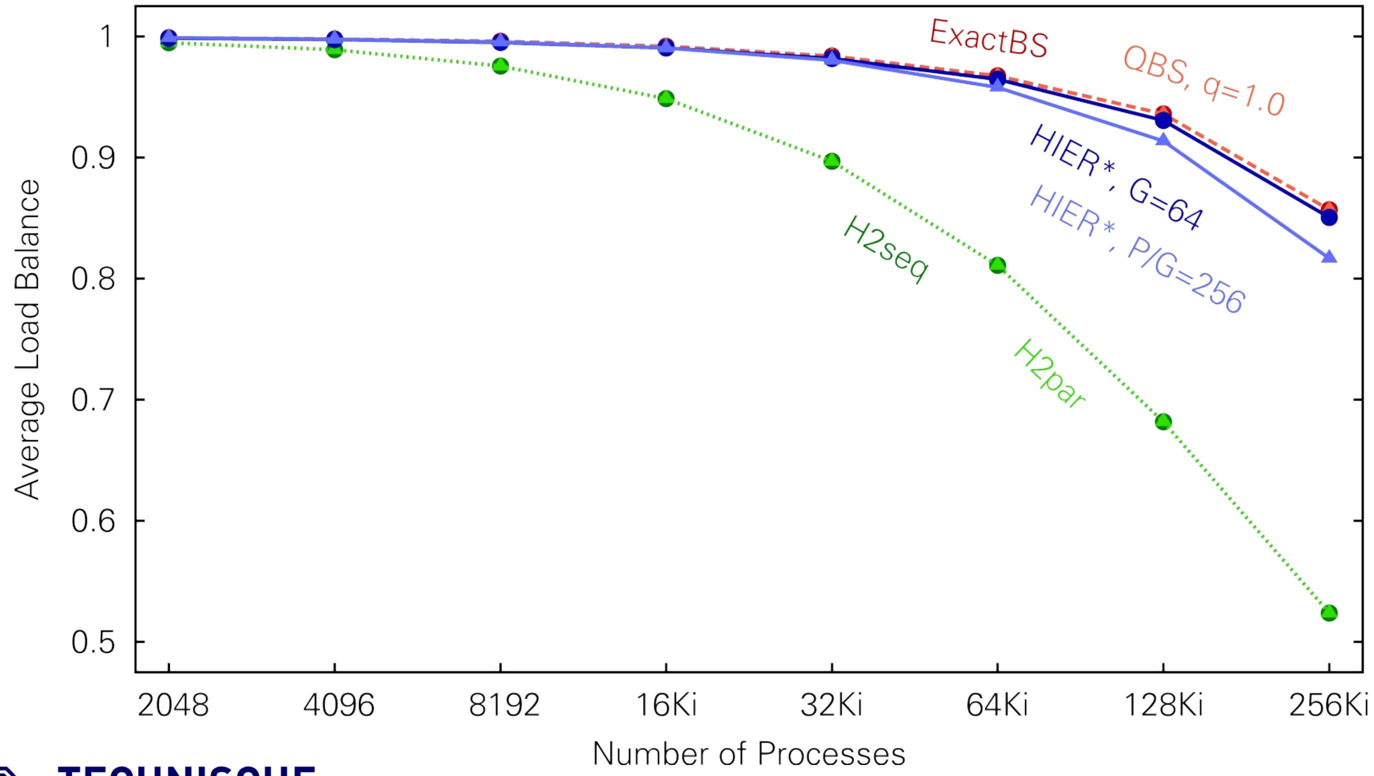
```
! MPI initialization
call MPI_Init(err)
call MPI_Comm_rank(MPI_COMM_WORLD, rank, err)
call MPI_Comm_size(MPI_COMM_WORLD, nproc, err)
! create the domain and allocate memory
call fd4_domain_create(domain, nb, size,      &
     vartab, ng, peri, MPI_COMM_WORLD, err)
call fd4_util_allocate_all_blocks(domain, err)
! initialize ghost communication
call fd4_ghostcomm_create(ghostcomm, domain, &
     4, vars, steps, err)
! loop over time steps
do timestep=1,nsteps
  ! exchange ghosts
  call fd4_ghostcomm_exch(ghostcomm, err)
  ! loop over local blocks
  call fd4_iter_init(domain, iter)
  do while(associated(iter%cur))
    ! do some computations
    call compute_block(iter)
    call fd4_iter_next(iter)
  end do
  ! dynamic load balancing
  call fd4_balance_readjust(domain, err)
end do
```

Benchmarks: COSMO-SPECS Performance Comparison

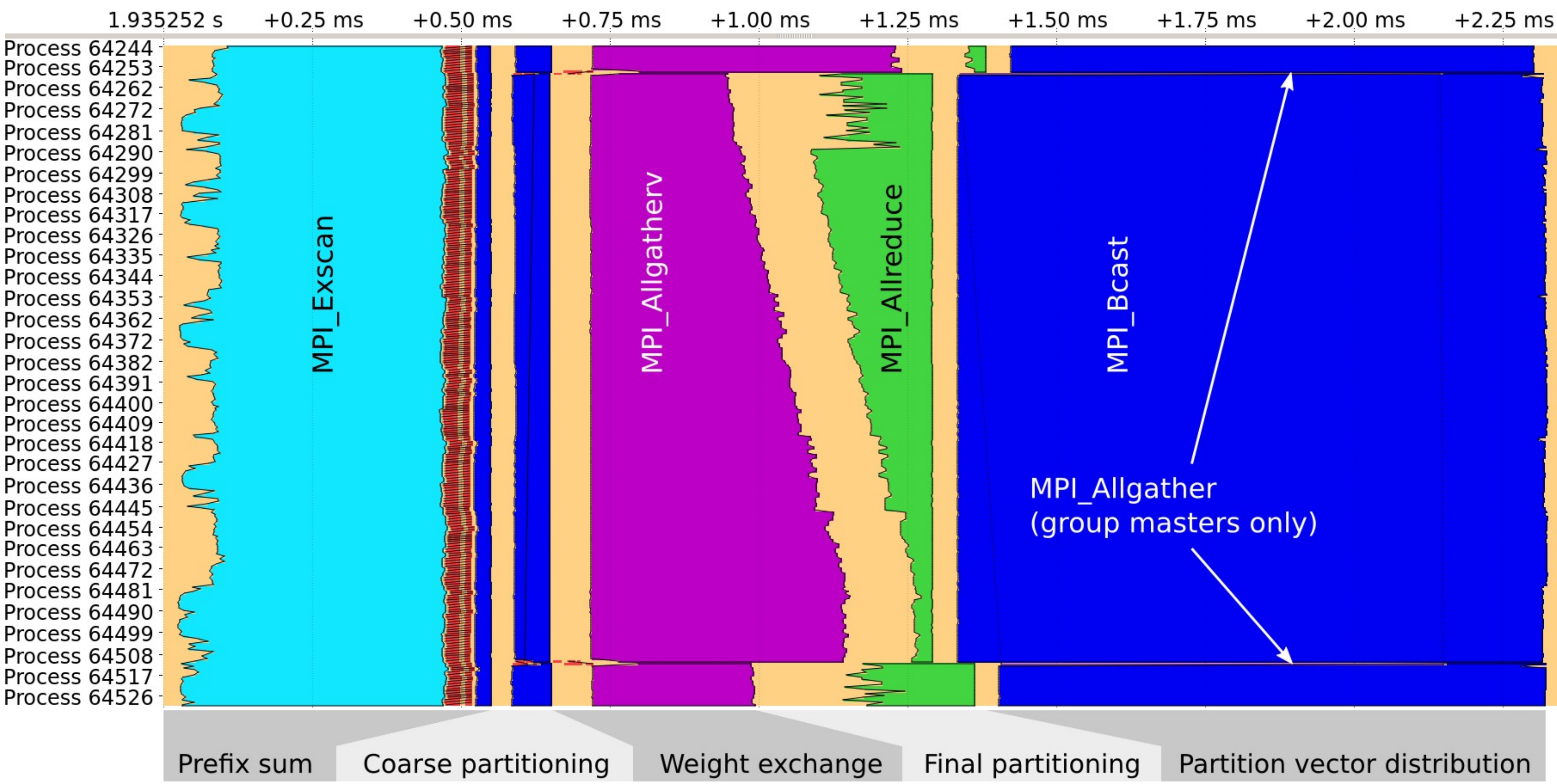


Scalable High-Quality 1D Partitioning: Load Balance

- Cloud simulation, 1 357 824 tasks
- System: JUQUEEN, IBM Blue Gene/Q
- HIER*, G=64 achieves 99.2 % of the optimal load balance at 262 144 processes



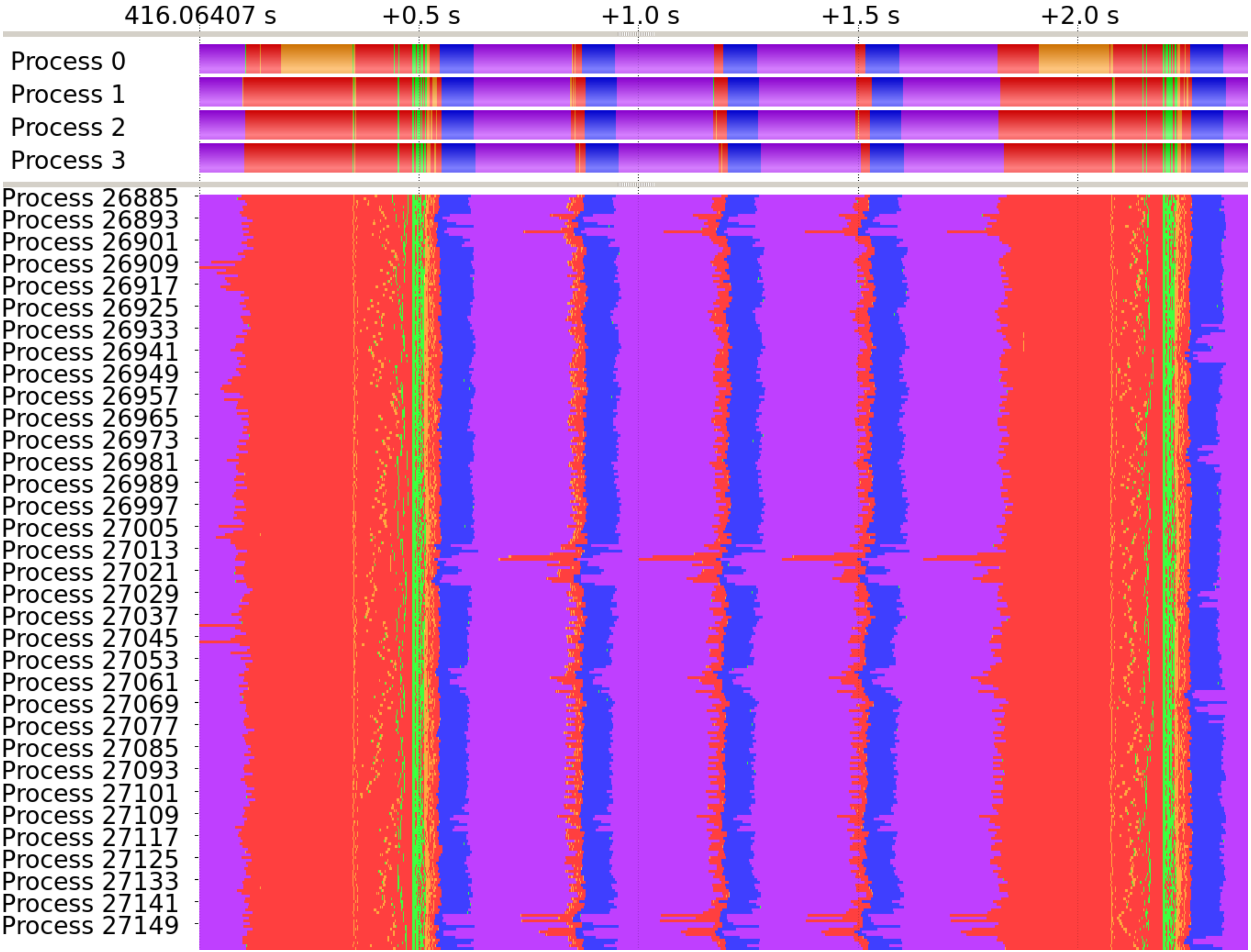
HIER* seen in Vampir (one Group of 256 out of 64Ki)



TECHNISCHE UNIVERSITÄT DRESDEN

ZIH Center for Information Services & High Performance Computing

ExactBS in Action (COSMO-SPECS+FD4)



COSMO-SPECS+FD4: Comparison of Methods

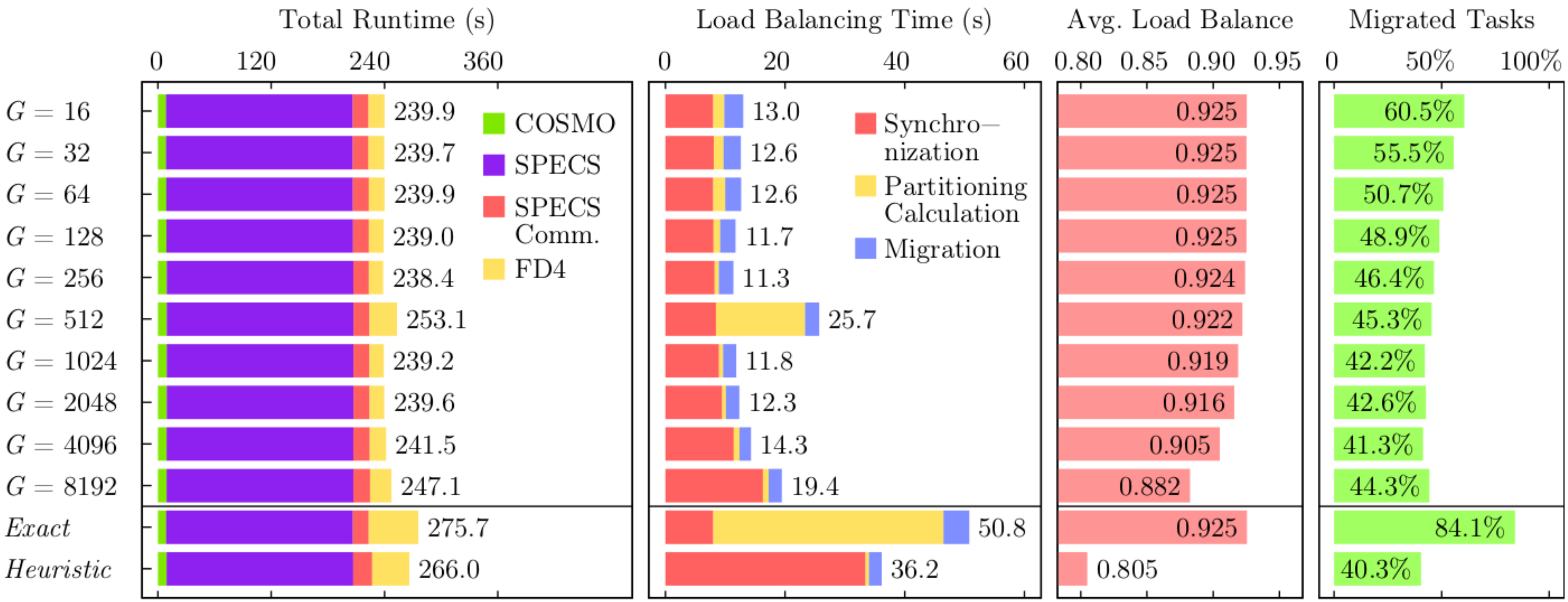
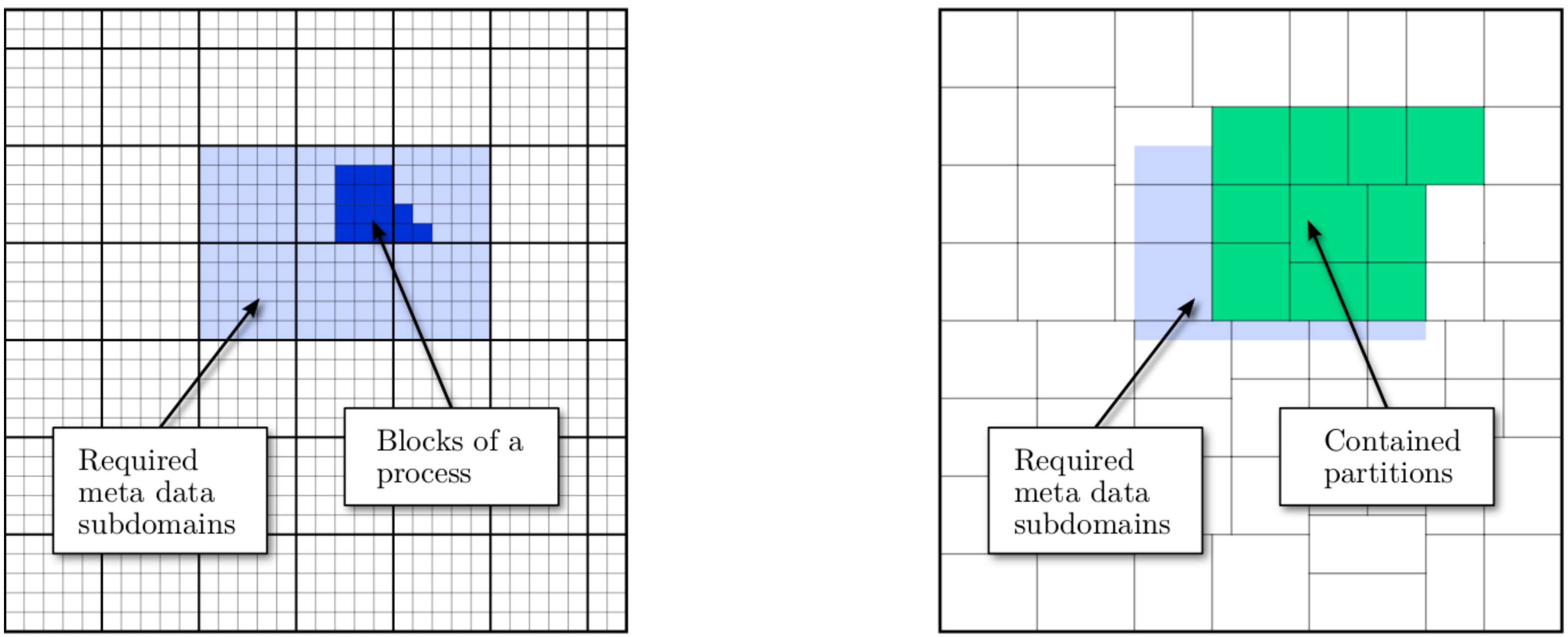


Figure 3. Influence of the group count G on the hierarchical 1D partitioning algorithm in COSMO-SPECS+FD4 with 65 536 processes on BlueGene/Q. The exact method and the heuristic are included as reference.

Lieber, Nagel, Mix, *Scalability Tuning of the Load Balancing and Coupling Framework FD4*, NIC Symposium 2014, pp. 363-370.

Scalable Coupling: Meta Data Subdomains

- "Handshaking" – Identifying partition overlaps between the coupled models – turned out to be the main scalability bottleneck
- Solved with spatially indexed data structure for coupling meta data in FD4
- Time for locating overlap candidates does not depend on number of ranks



(a) Required meta data subdomains. (b) Contained coupled partitions.

Lieber, Nagel, Mix, *Scalability Tuning of the Load Balancing and Coupling Framework FD4*, NIC Symposium 2014, pp. 363-370.