TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

FD4: A Framework for Highly Scalable Load Balancing and Coupling of Multiphase Models

ZIH HPC User Forum, 6. December 2010, Dresden

Matthias Lieber

matthias.lieber@tu-dresden.de

Center for Information Services and High Performance Computing (ZIH), TU Dresden

ZIH
Center for Information Services & High Performance Computing

Outline

- Introduction
  - Project Motivation
  - Basic Idea of FD4
- FD4 Key Features
- Application of FD4
  - COSMO-SPECS+FD4
  - Performance Comparison
- Conclusion & Outlook

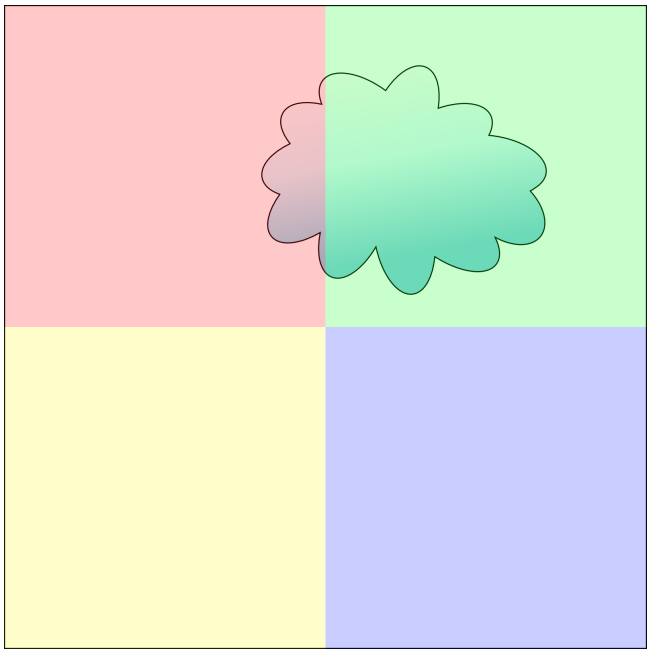
Introduction: Detailed Simulation of Clouds

- *"Parallel coupling framework and advanced time integration methods for detailed cloud processes in atmospheric models"*
- Cooperation with Leibniz Institute for Tropospheric Research (IfT), Leipzig, Germany

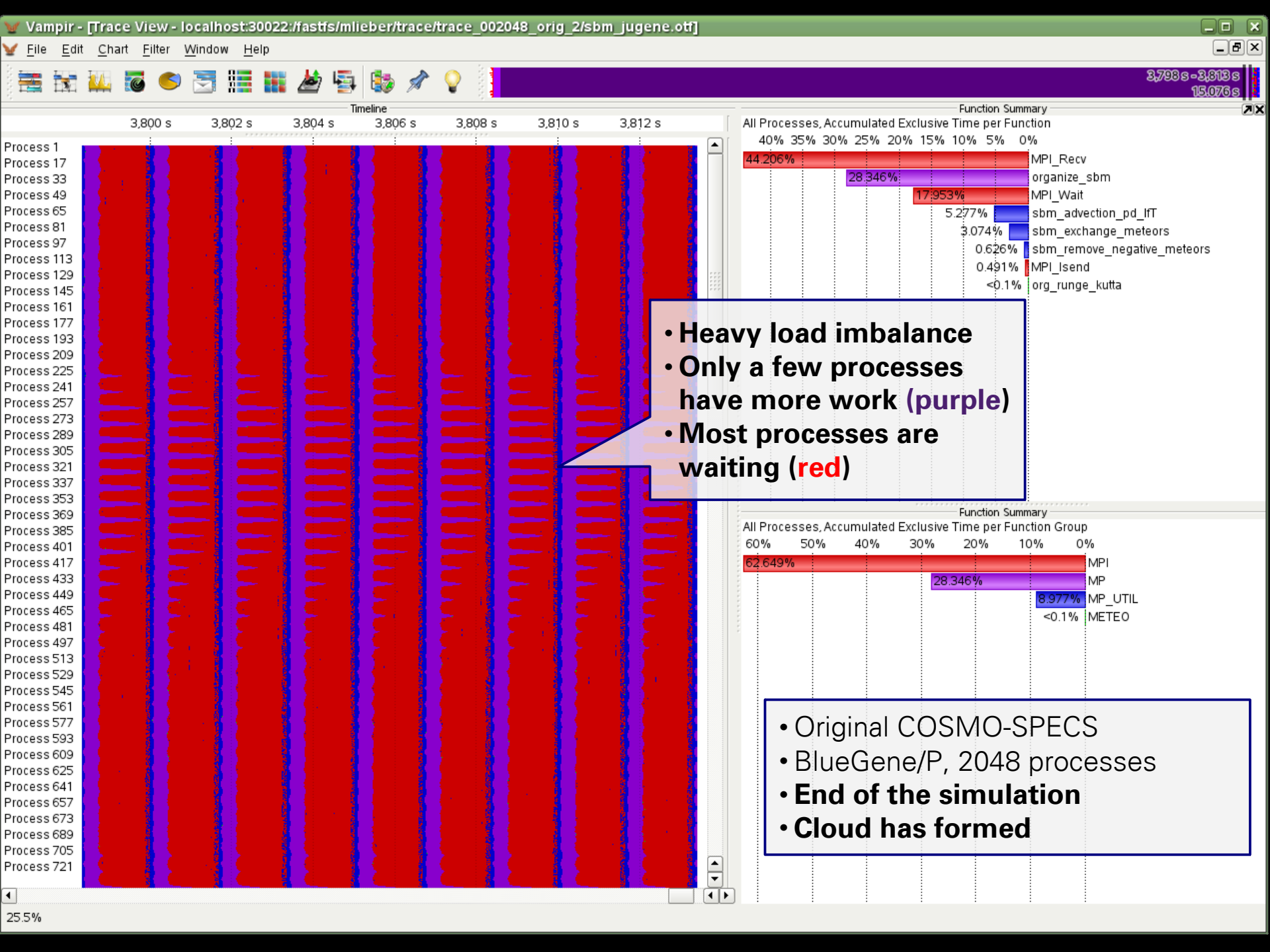
- Performance improvement of the model system COSMO-SPECS
  - Detailed modeling of interactions between aerosol particles, clouds, and precipitation
  - COSMO Model: non-hydrostatic limited-area atmospheric model (www.cosmo-model.org)
  - SPECS: Cloud parameterization scheme of COSMO replaced by the detailed cloud model SPECS (SPECtral bin microphysicS) [Simmel06, Grützun08]

Introduction: COSMO-SPECS Performance

- SPECS is very costly
  - > 99% of total runtime
- SPECS runtime varies strongly
  - Depending on range of droplet size distribution and the presence of frozen particles
- This leads to severe load imbalance
  - COSMO's parallelization is based on static 2D partitioning

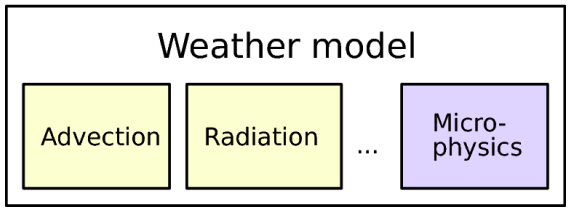


Dynamic load balancing needed to run realistic cases on large HPC systems

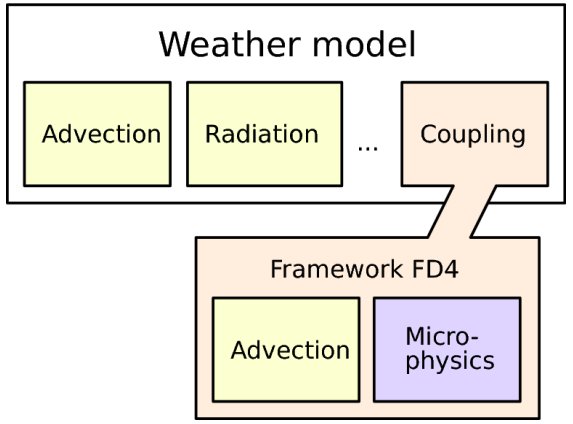

Vampir - [Trace View - localhost:30022:/fastfs/mlieber/trace/trace_002048_orig_2/sbm_jugene.otf]
File Edit Chart Filter Window Help
3,798 s - 3,813 s
15.076 s
Timeline
3,800 s
3,802 s
3,804 s
3,806 s
3,808 s
3,810 s
3,812 s
Process 1
Process 17
Process 33
Process 49
Process 65
Process 81
Process 97
Process 113
Process 129
Process 145
Process 161
Process 177
Process 193
Process 209
Process 225
Process 241
Process 257
Process 273
Process 289
Process 305
Process 321
Process 337
Process 353
Process 369
Process 385
Process 401
Process 417
Process 433
Process 449
Process 465
Process 481
Process 497
Process 513
Process 529
Process 545
Process 561
Process 577
Process 593
Process 609
Process 625
Process 641
Process 657
Process 673
Process 689
Process 705
Process 721
25.5%
Function Summary
All Processes, Accumulated Exclusive Time per Function
40% 35% 30% 25% 20% 15% 10% 5% 0%
44.206% MPI_Recv
28.346% organize_sbm
17.953% MPI_Wait
5.277% sbm_advection_pd_lfT
3.074% sbm_exchange_meteors
0.626% sbm_remove_negative_meteors
0.491% MPI_Isend
<0.1% org_runge_kutta
• Heavy load imbalance
• Only a few processes have more work (purple)
• Most processes are waiting (red)
Function Summary
All Processes, Accumulated Exclusive Time per Function Group
60% 50% 40% 30% 20% 10% 0%
62.649% MPI
28.346% MP
8.977% MP_UTIL
<0.1% METEO
• Original COSMO-SPECS
• BlueGene/P, 2048 processes
• End of the simulation
• Cloud has formed

Introduction: Basic Idea of FD4

- Present approaches:
  - Cloud model is implemented as a submodule within the weather model
  - Uses (static) data structures of the weather model

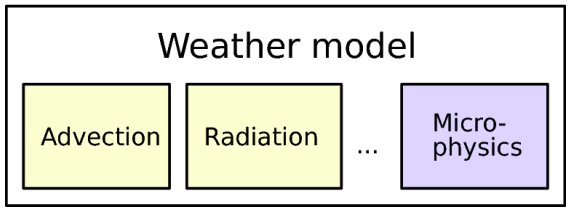


- Our idea:
  - Separate cloud model data from weather model data structures
  - Independent domain decompositions
  - Dynamic load balancing for the cloud model
  - (Re)couple weather and cloud model



Introduction: Basic Idea of FD4

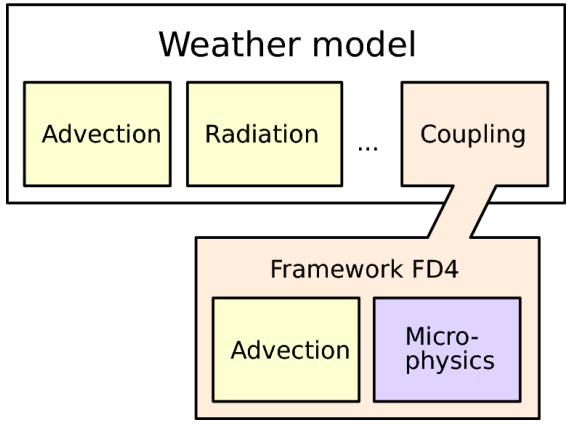
- Present approaches:
  - Cloud model is implemented as a submodule within the weather model
  - Uses (static) data structures of the weather model



- Our idea:

Functionality provided by FD4

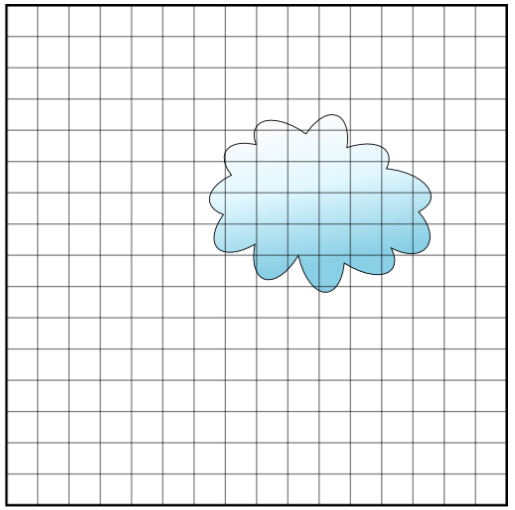
  - Separate cloud model data from weather model data structures
  - Independent domain decompositions
  - Dynamic load balancing for the cloud model
  - (Re)couple weather and cloud model



Framework FD4: Key Features

FD^4 = Four-Dimensional Distributed Dynamic Data structures

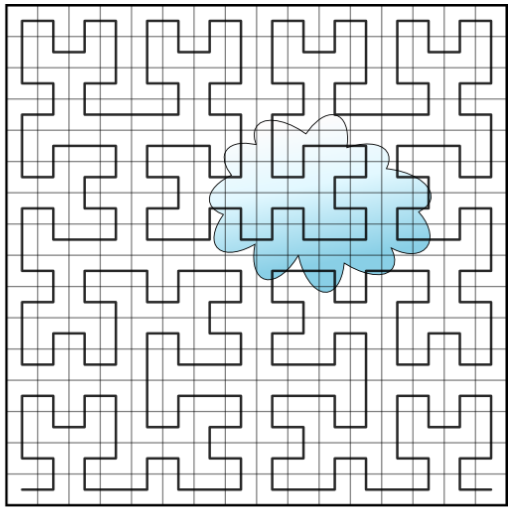
- Dynamic load balancing
  - Regular grid managed by FD4
  - Block-based 3D decomposition
  - Hilbert space-filling curve [Sagan94] partitioning
- Model coupling
- Adaptive block mode
- 4th dimension



Framework FD4: Key Features

FD4 = Four-Dimensional Distributed Dynamic Data structures

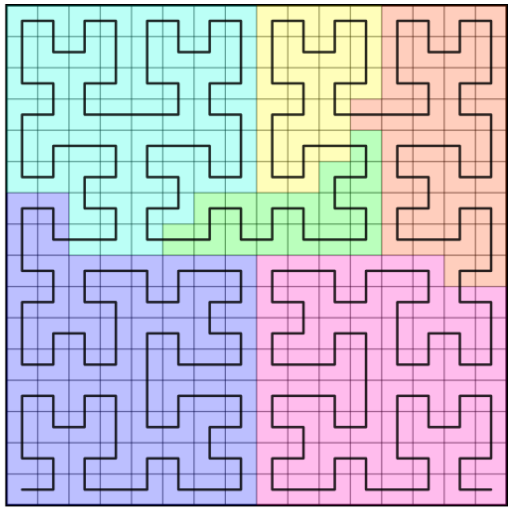
- Dynamic load balancing
  - Regular grid managed by FD4
  - Block-based 3D decomposition
  - Hilbert space-filling curve [Sagan94] partitioning
- Model coupling
- Adaptive block mode
- 4th dimension



Framework FD4: Key Features

FD^4 = Four-Dimensional Distributed Dynamic Data structures

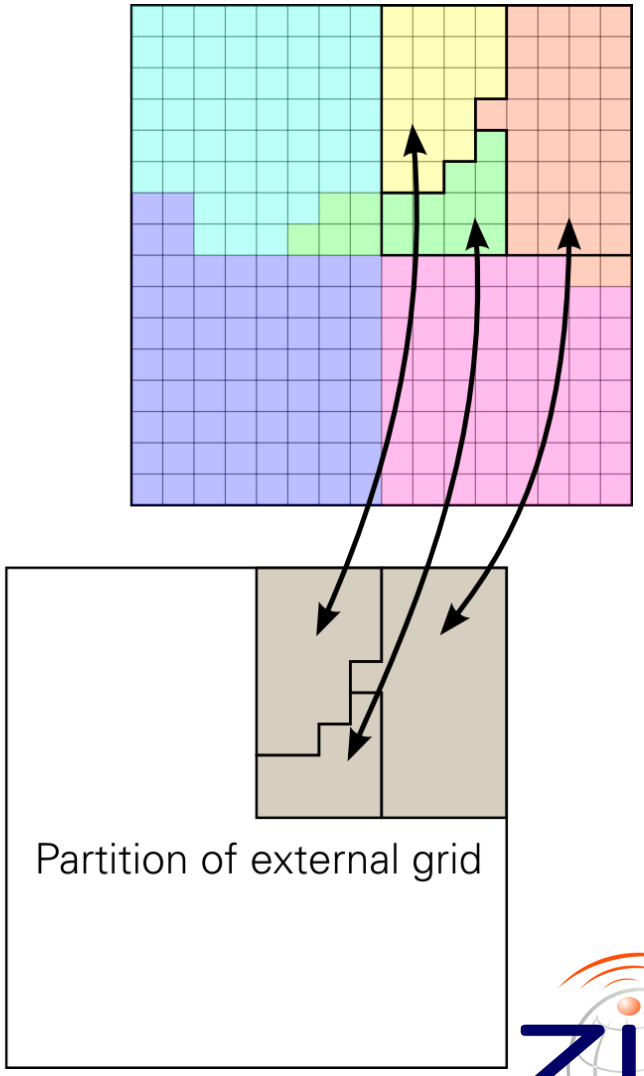
- Dynamic load balancing
  - Regular grid managed by FD4
  - Block-based 3D decomposition
  - Hilbert space-filling curve [Sagan94] partitioning
- Model coupling
- Adaptive block mode
- 4th dimension



Framework FD4: Key Features

FD^4 = Four-Dimensional Distributed Dynamic Data structures

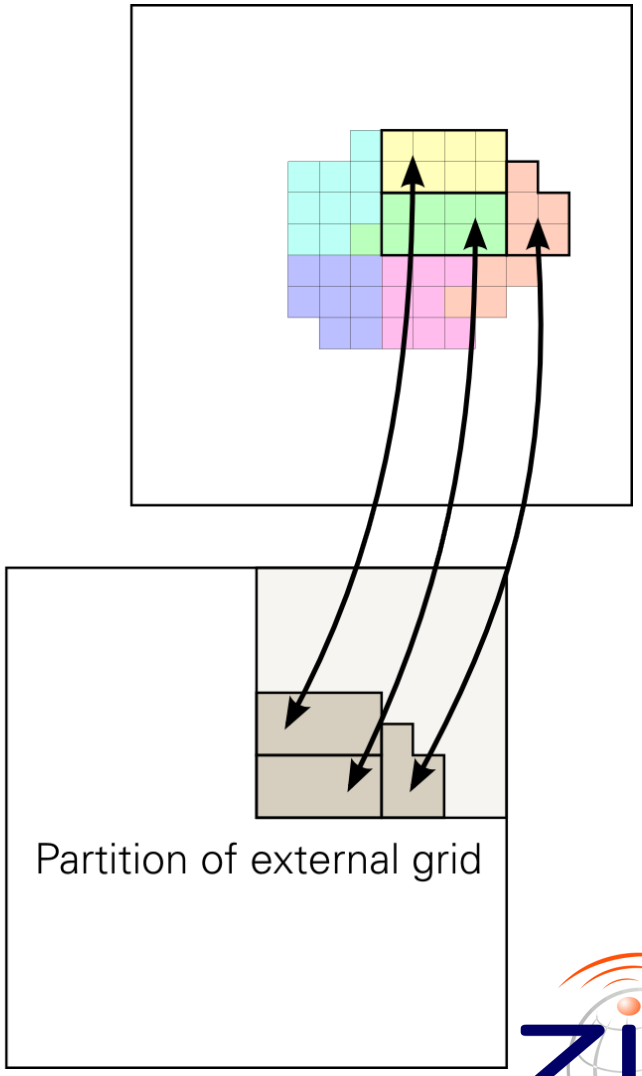
- Dynamic load balancing
- Model coupling
  - Data exchange between FD4 based model and external model
  - E.g. CFD or weather model
  - Direct data transfer between overlapping parts of of the partitions
- Adaptive block mode
- 4th dimension



Framework FD4: Key Features

FD⁴ = Four-Dimensional Distributed Dynamic Data structures

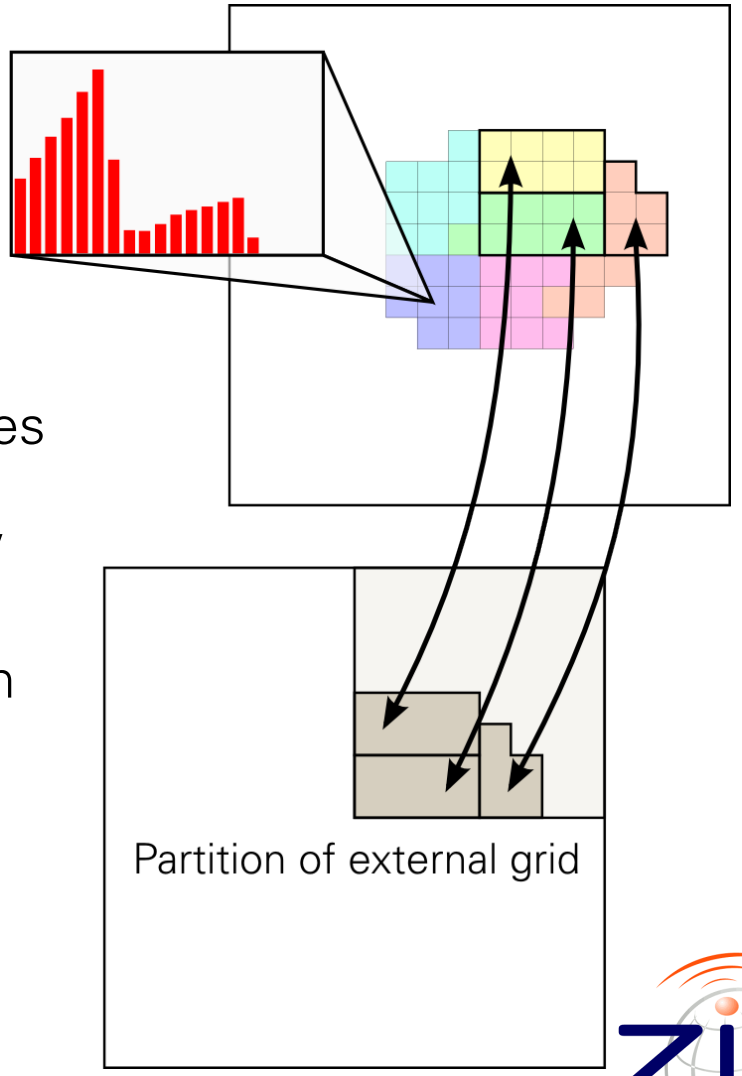
- Dynamic load balancing
- Model coupling
- Adaptive block mode
  - Save memory in case data and computations are required for a spatial subset only
  - Suitable for multiphase problems like drops, clouds, flame fronts
- 4th dimension



Framework FD4: Key Features

FD4 = Four-Dimensional Distributed Dynamic Data structures

- Dynamic load balancing
- Model coupling
- Adaptive block mode
- 4th dimension
  - Extra dimension of grid variables
  - E.g. array of gas phase tracers, size resolving models
  - FD4 is optimized for a large 4th dimension
  - COSMO-SPECS requires 2 x 11 x 66 ~ 1500 values



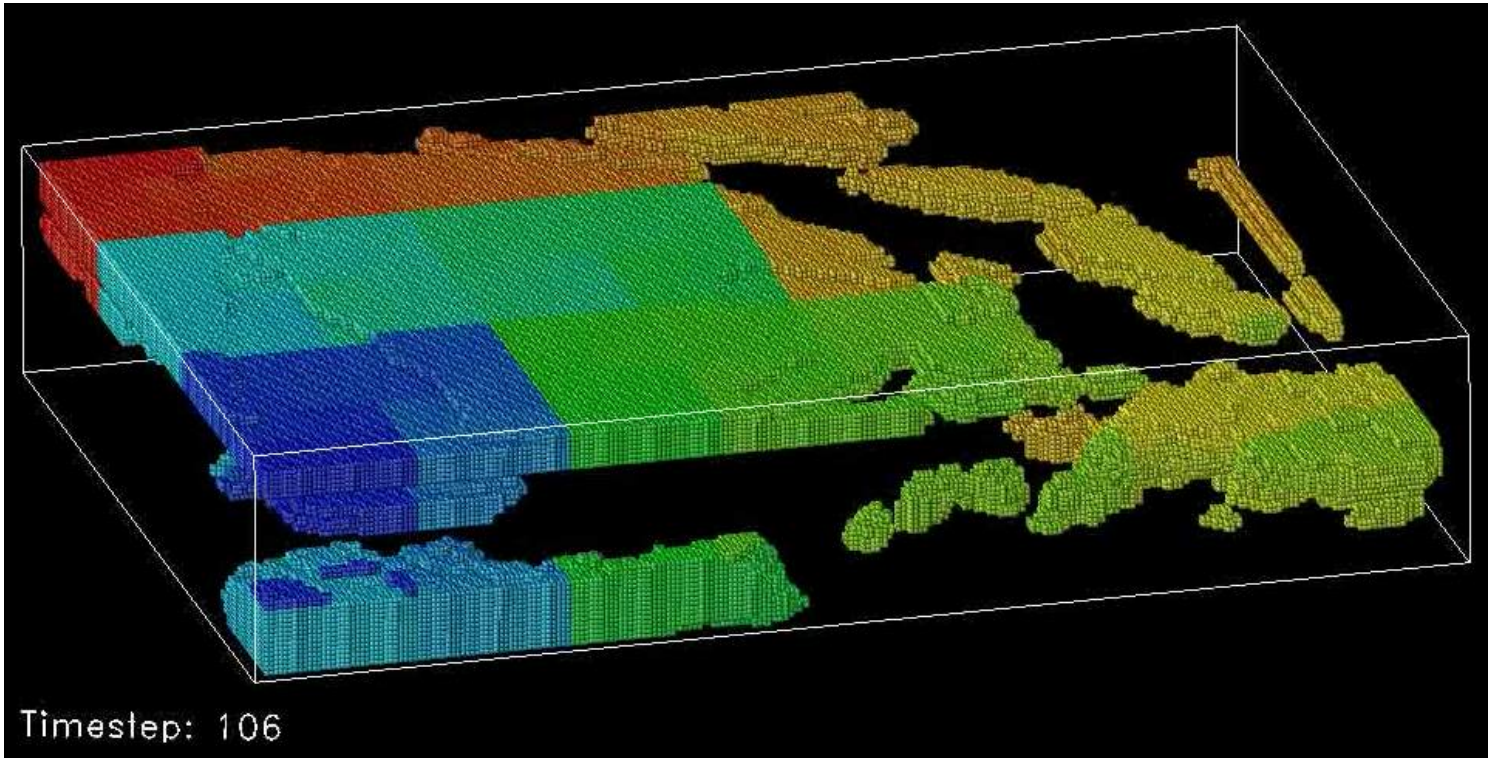
Framework FD4: Implementation

- FD4 is written in Fortran 95
- MPI based parallelization
- (Simple) I/O interfaces to
  - NetCDF
  - Vis5D
- Open source software

```
! 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
```

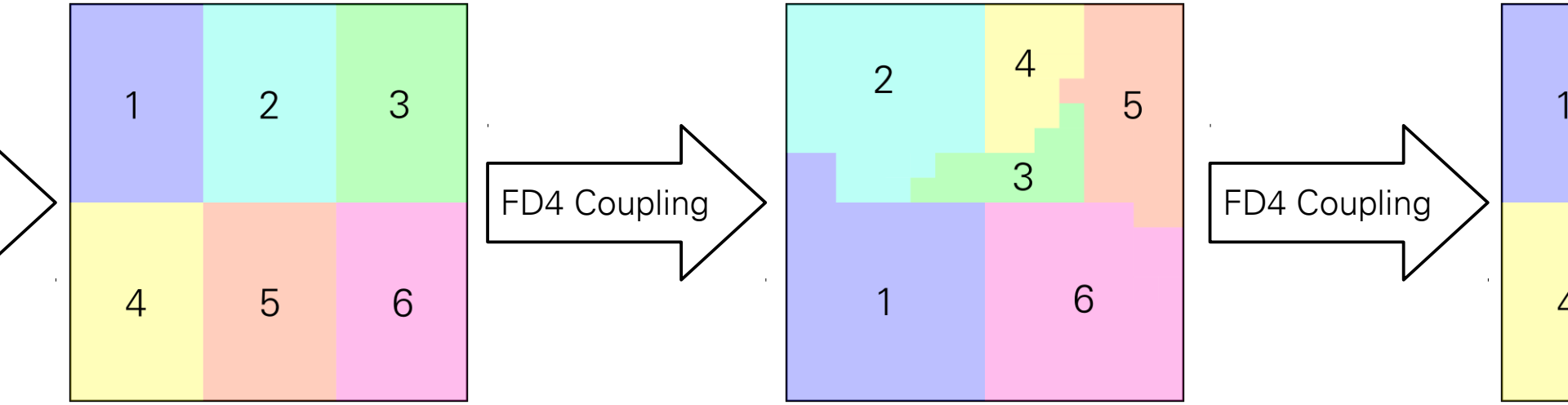
Available at http://www.tu-dresden.de/zih/clouds

Framework FD4: Dynamic Load Balancing Movie



- Overhead test of adaptive block mode and load balancing [Lieber10]
- FD4 adapts to cloud formation in COSMO weather model
- Real-life scenario, 249 x 174 x 50 grid, 256 processes

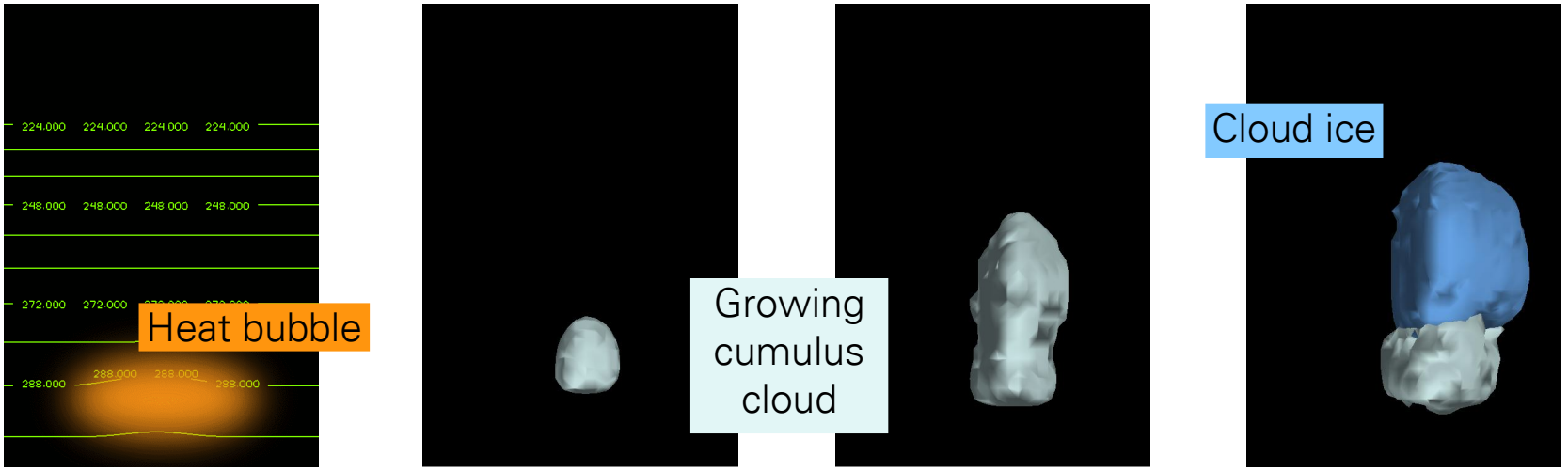
Application of FD4: COSMO-SPECS+FD4



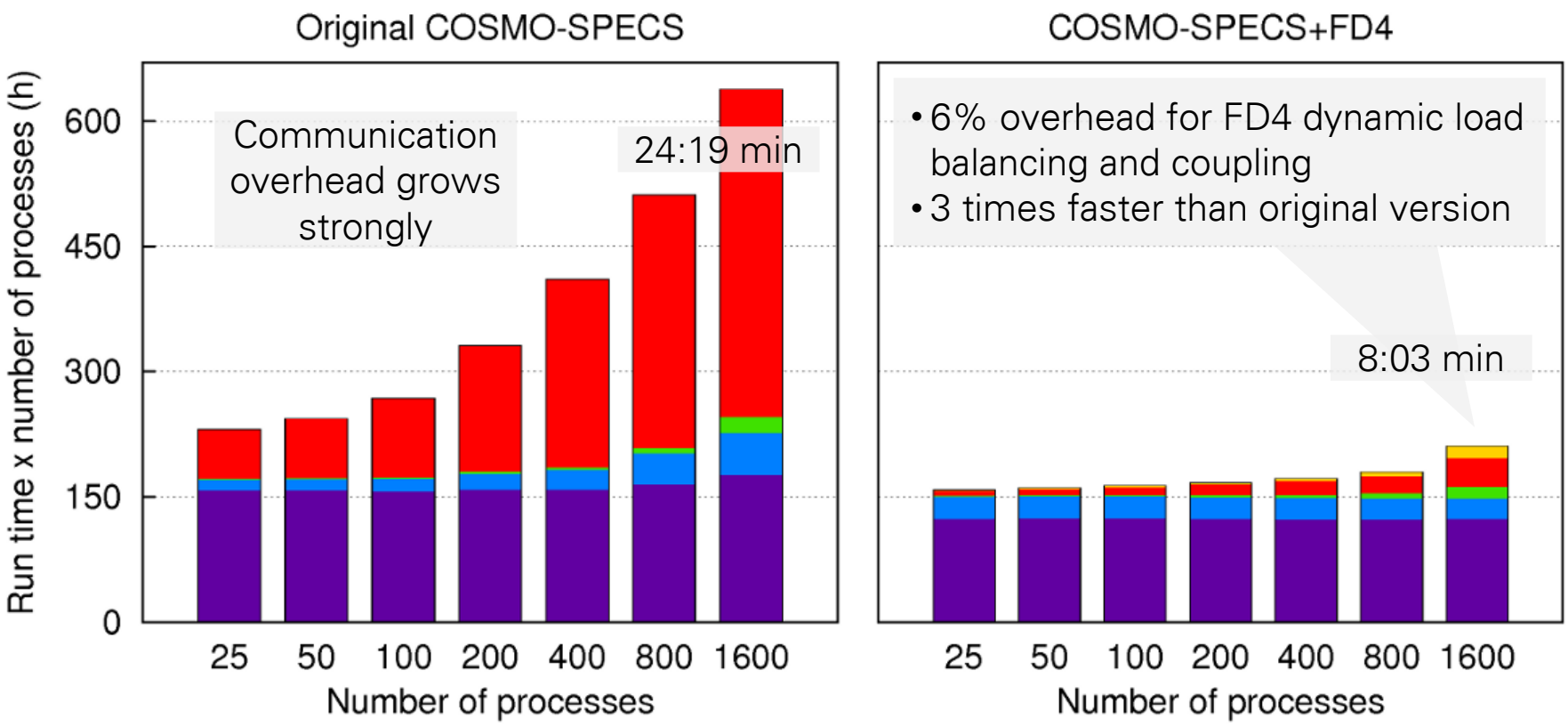
COSMO	**FD4**	**SPECS**	**FD4**
Computes dynamics	Send data to SPECS grid:	Computes Microphysics	Receive data from SPECS grid:
Static MxN partitioning	u, v, w, T, p, ρ, q_v	Data dynamically balanced by FD4	ΔT, q_v, q_c, q_i

Application of FD4: Scalability Benchmark Case

- Comparing original COSMO-SPECS with COSMO-SPECS+FD4
- Test scenario: heat bubble results in growth of cumulus cloud
- 30 min forecast time
- Vertical grid: 48 nonuniform height levels (up to 18 km)
- Horizontal grid: 80 x 80, 1km resolution
- 2 x 2 x 4 = 16 grid cells per FD4 block, 19 200 blocks



Application of FD4: SGI Altix 4700 (mars) Scalability



- FD4 load balancing and coupling
- Ghost exchange for SPECS including waiting times due to imbalance
- COSMO computations
- SPECS advection
- SPECS microphysics

Application of FD4: BlueGene/P Scalability Benchmark

- How far can we scale? Are we prepared for HRSK-2?
- IBM BlueGene/P System at Jülich Supercomputing Centre
  - 294 912 IBM PowerPC 450 processors, #9 in the Top500
- Weak scaling test: problem size per process = constant
  - Replicated heat bubble for each 32 x 32 grid section

# Proc.	Grid size	# Replicated clouds	# FD4 blocks
256	32x32	1x1	3072
512	64x32	2x1	6144
1024	64x64	2x2	12 288
…			
32768	512x256	16x8	393 216
65 536	512x512	16x16	786 432

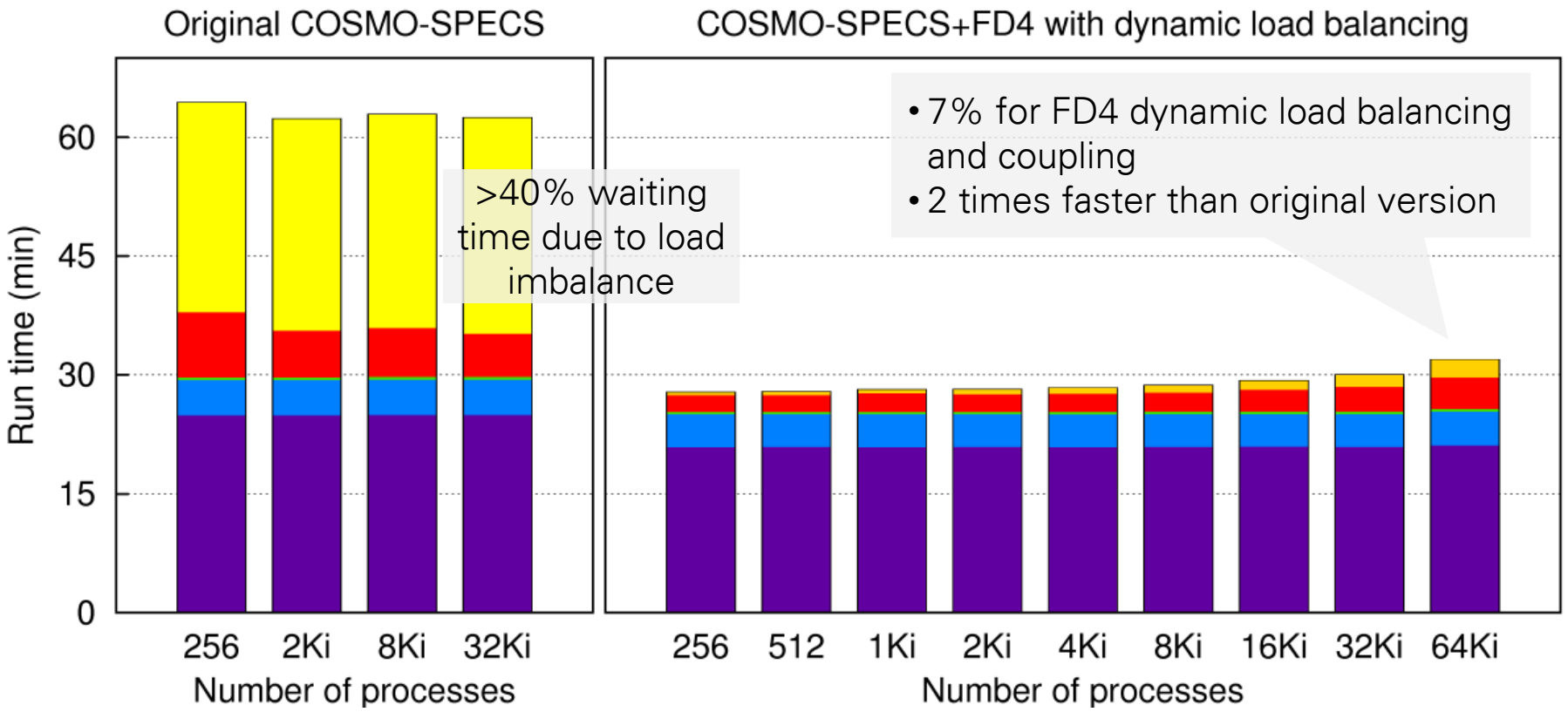
http://www.fz-juelich.de/jsc/

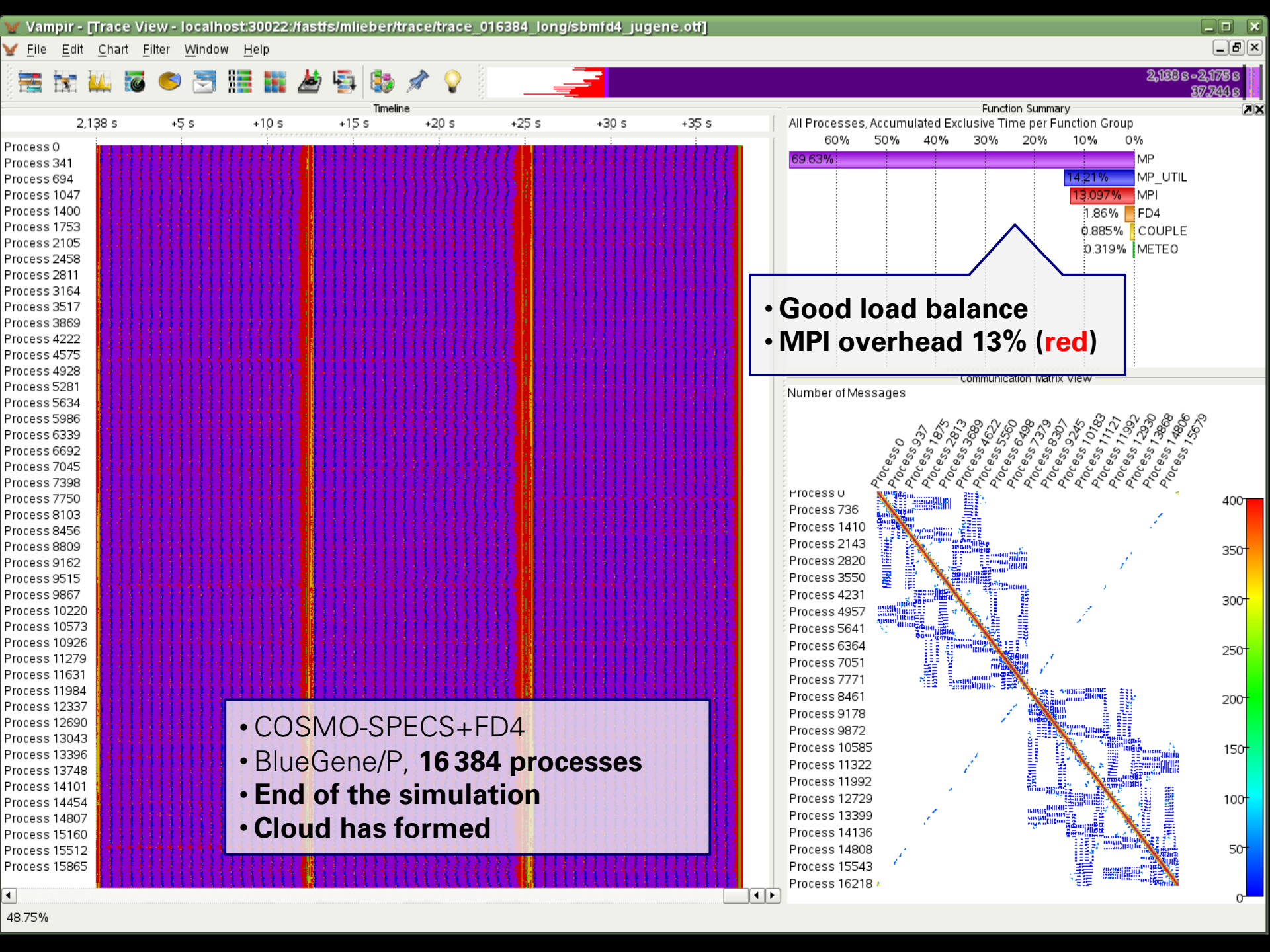
JÜLICH FORSCHUNGSZENTRUM

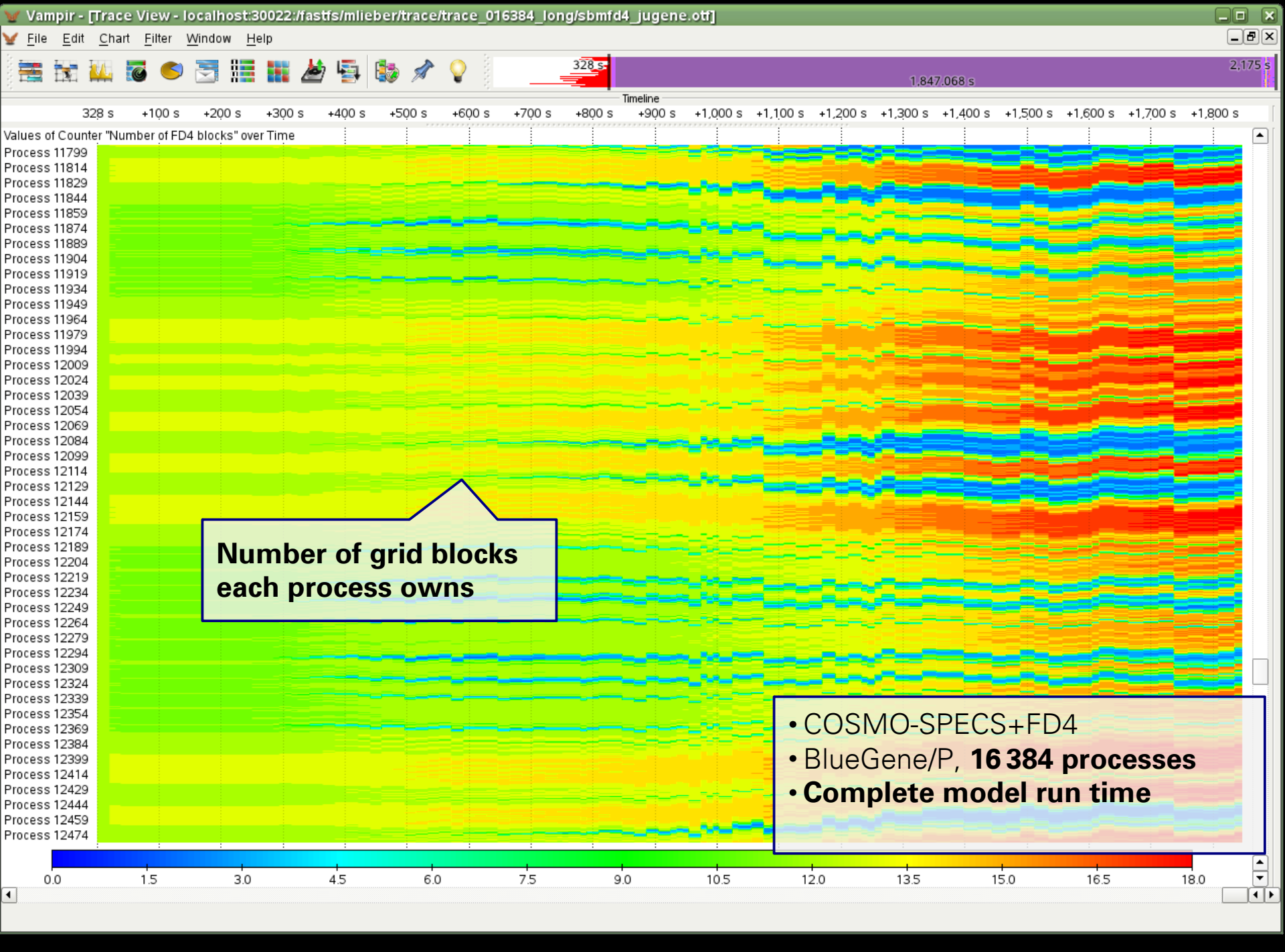
TECHNISCHE UNIVERSITÄT DRESDEN

ZIH Center for Information Services & High Performance Computing

Application of FD4: BlueGene/P Scalability Benchmark




Vampir - [Trace View - localhost:30022:/fastfs/mlieber/trace/trace_016384_long/sbmfd4_jugene.otf]
File Edit Chart Filter Window Help
2,138 s - 2,175 s
37.744 s
Timeline
2,138 s +5 s +10 s +15 s +20 s +25 s +30 s +35 s
Process 0
Process 341
Process 694
Process 1047
Process 1400
Process 1753
Process 2105
Process 2458
Process 2811
Process 3164
Process 3517
Process 3869
Process 4222
Process 4575
Process 4928
Process 5281
Process 5634
Process 5986
Process 6339
Process 6692
Process 7045
Process 7398
Process 7750
Process 8103
Process 8456
Process 8809
Process 9162
Process 9515
Process 9867
Process 10220
Process 10573
Process 10926
Process 11279
Process 11631
Process 11984
Process 12337
Process 12690
Process 13043
Process 13396
Process 13748
Process 14101
Process 14454
Process 14807
Process 15160
Process 15512
Process 15865
• COSMO-SPECS+FD4
• BlueGene/P, 16 384 processes
• End of the simulation
• Cloud has formed
48.75%
Function Summary
All Processes, Accumulated Exclusive Time per Function Group
60% 50% 40% 30% 20% 10% 0%
69.63% MP
14.21% MP_UTIL
13.097% MPI
1.86% FD4
0.885% COUPLE
0.319% METEO
• Good load balance
• MPI overhead 13% (red)
Communication Matrix View
Number of Messages
Process 0 Process 937 Process 1875 Process 2813 Process 3689 Process 4622 Process 5560 Process 6498 Process 7379 Process 8307 Process 9245 Process 10183 Process 11121 Process 11992 Process 12930 Process 13868 Process 14806 Process 15679
Process 0
Process 736
Process 1410
Process 2143
Process 2820
Process 3550
Process 4231
Process 4957
Process 5641
Process 6364
Process 7051
Process 7771
Process 8461
Process 9178
Process 9872
Process 10585
Process 11322
Process 11992
Process 12729
Process 13399
Process 14136
Process 14808
Process 15543
Process 16218
400
350
300
250
200
150
100
50
0


Vampir - [Trace View - localhost:30022:/fastfs/mlieber/trace/trace_016384_long/sbmfd4_jugene.otf]
File Edit Chart Filter Window Help
328 s
1,847.068 s
2,175 s
Timeline
328 s +100 s +200 s +300 s +400 s +500 s +600 s +700 s +800 s +900 s +1,000 s +1,100 s +1,200 s +1,300 s +1,400 s +1,500 s +1,600 s +1,700 s +1,800 s
Values of Counter "Number of FD4 blocks" over Time
Process 11799
Process 11814
Process 11829
Process 11844
Process 11859
Process 11874
Process 11889
Process 11904
Process 11919
Process 11934
Process 11949
Process 11964
Process 11979
Process 11994
Process 12009
Process 12024
Process 12039
Process 12054
Process 12069
Process 12084
Process 12099
Process 12114
Process 12129
Process 12144
Process 12159
Process 12174
Process 12189
Process 12204
Process 12219
Process 12234
Process 12249
Process 12264
Process 12279
Process 12294
Process 12309
Process 12324
Process 12339
Process 12354
Process 12369
Process 12384
Process 12399
Process 12414
Process 12429
Process 12444
Process 12459
Process 12474
Number of grid blocks each process owns
0.0 1.5 3.0 4.5 6.0 7.5 9.0 10.5 12.0 13.5 15.0 16.5 18.0
• COSMO-SPECS+FD4
• BlueGene/P, 16 384 processes
• Complete model run time

Conclusion & Outlook

- FD4 provides highly scalable dynamic load balancing and coupling for multiphase models
- Scalability to 10 000s of processes
- COSMO-SPECS performance increased significantly by FD4
- FD4 not limited to meteorology
- Freely available at http://www.tu-dresden.de/zih/clouds

- Outlook:
  - Multirate time stepping in COSMO-SPECS+FD4
  - Parallel I/O in FD4
  - Simulate a real-case scenario with COSMO-SPECS+FD4

Thank you for your attention!

Acknowledgments

- COSMO Model: German Weather Service (DWD)
- Access to IBM BlueGene/P: Jülich Supercomputing Centre (JSC)
- Funding: German Research Foundation (DFG)

[Grützun08] V. Grützun, O. Knoth, and M. Simmel. *Simulation of the influence of aerosol particle characteristics on clouds and precipitation with LM–SPECS: Model description and first results.* Atmos. Res., 90:233–242, 2008.

[Lieber08] M. Lieber and R. Wolke. *Optimizing the coupling in parallel air quality model systems*, Environ. Modell. Softw., 23:235-243, 2008

[Lieber10] M. Lieber, R. Wolke, V. Grützun, M.S. Müller, and W.E. Nagel. *A framework for detailed multiphase cloud modeling on HPC systems*, in Parallel Computing Vol. 19, 281-288, IOS Press, 2010.

[Pinar04] A. Pinar and C. Aykanat. *Fast optimal load balancing algorithms for 1D partitioning.* J. Parallel Distrib. Comput., 64(8):974-996, 2004.

[Sagan94] H. Sagan. *Space-filling curves*, Springer, 1994

[Simmel06] M. Simmel and S. Wurzler. *Condensation and activation in sectional cloud microphysical models*, Atmos. Res., 80:218-236, 2006.

[Teresco06] J.D. Teresco, K.D. Devine, and J.E. Flaherty. *Partitioning and Dynamic Load Balancing for the Numerical Solution of Partial Differential Equations,* in Numerical Solution of Partial Differential Equations on Parallel Computers, pages 55-88, Springer, 2006.