Towards Exascale Computing Simulation on Millions of Cores

N. Thürey, H. Köstler, J. Götz, M. Stürmer, S. Donath, F. Deserno, C. Mihoubi, K. Pickl, B. Gmeiner, T. Gradl, D. Ritter, T. Pohl, J. Treibig, T. Dreher, C. Feichtinger, K. Iglberger, T. Preclik, S. Bogner

U. Rüde (LSS Erlangen, ruede@cs.fau.de)

Lehrstuhl für Informatik 10 (Systemsimulation)

Universität Erlangen-Nürnberg

www10.informatik.uni-erlangen.de

COSSE Workshop

Mathematics in Waterland

Delft, Feb. 7-10, 2011

Friedrich-Alexander University Erlangen-Nürnberg

- Founded 1743
- 5 Schools, 22 Departments, 24 Clinics
- 28,000 students in 132 programs
- ~ 500 tenured faculty (full and associate profs)
- ~ 900 PhD and Dr. habil. degrees in 2007
- ~ 92 Mio Euro external funding/ year
- School of Engineering founded in 1966
  - Currently about 6000 students

Overview

- Motivation: How fast are computers today (and tomorrow)
- Scalable Parallel Multigrid Algorithms for PDE
  - Matrix-Free FE solver: Hierarchical Hybrid Grids
  - Experiments with Multigrid on GPUs
- A Multi-Scale & Multi-Physics Simulation
  - Rigid Body Dynamics for Granular Media
  - Flow Simulation with Lattice Boltzmann Methods
  - Fluid-Structure Interaction with Moving Rigid Objects
    - particle ladden flows
    - preliminary GPU performance comparison
- Conclusions

Motivation

Example Peta-Scale System: Jugene @ Jülich

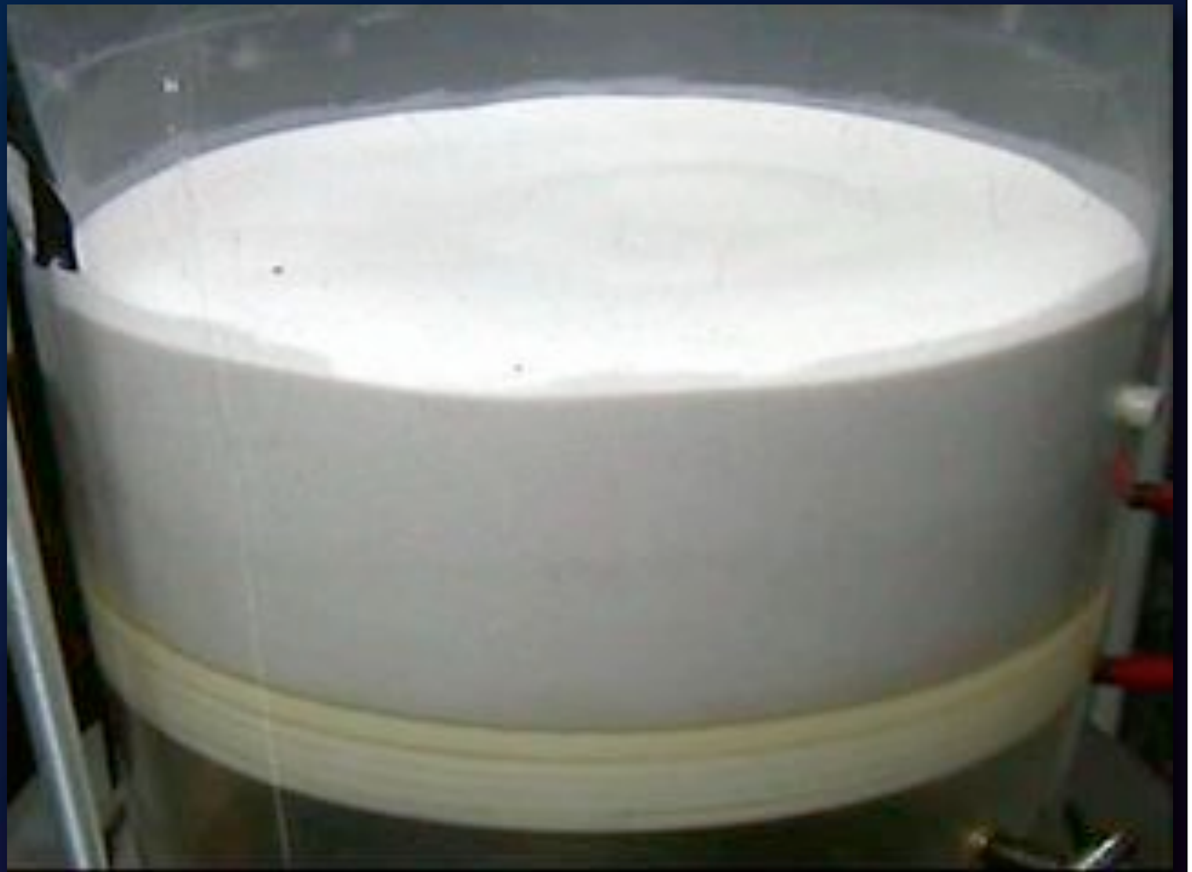
Extreme Scaling Workshop 2010 at Jülich Supercomputing Center

- PetaFlops = 10^{15} operations/second
- IBM Blue Gene
- Theoretical peak performance: 1.0027 Petaflop/s
- 294 912 cores
- 144 TBytes = $1.44\ 10^{14}$
- #9 on TOP 500 List in Nov. 2010

- For comparison: Current fast desktop PC is ~ 20.000 times slower
- > 1 000 000 cores expected 2011
- Exa-Scale System expected by 2018/19

What can we do with Exa-Scale Computers?

- Even if we want
  - to simulate a billion objects (particles): we can do a billion operations for each of them in each second
  - a trillion finite elements (finite volumes) to resolve a PDE, we can do a million operations for each of them in each second
- Most existing software dramatically underperforms on contemporary HPC architectures
- This will get more dramatic on future exa-scale systems



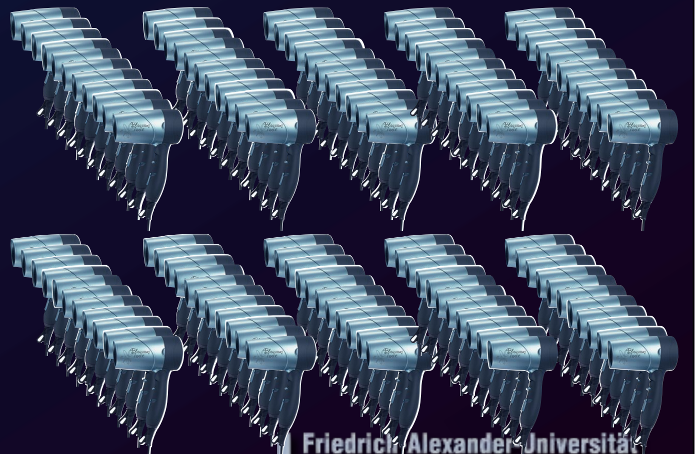
Fluidized Bed
(movie: thanks to K.E. Wirth)

What‘s the problem?

with four strong jet engines
(not those of Rolls-Royce of course)

Would you want to propel a Super Jumbo

or with 300,000 blow dryer fans?



How fast are our algorithms (multigrid) on current CPUs

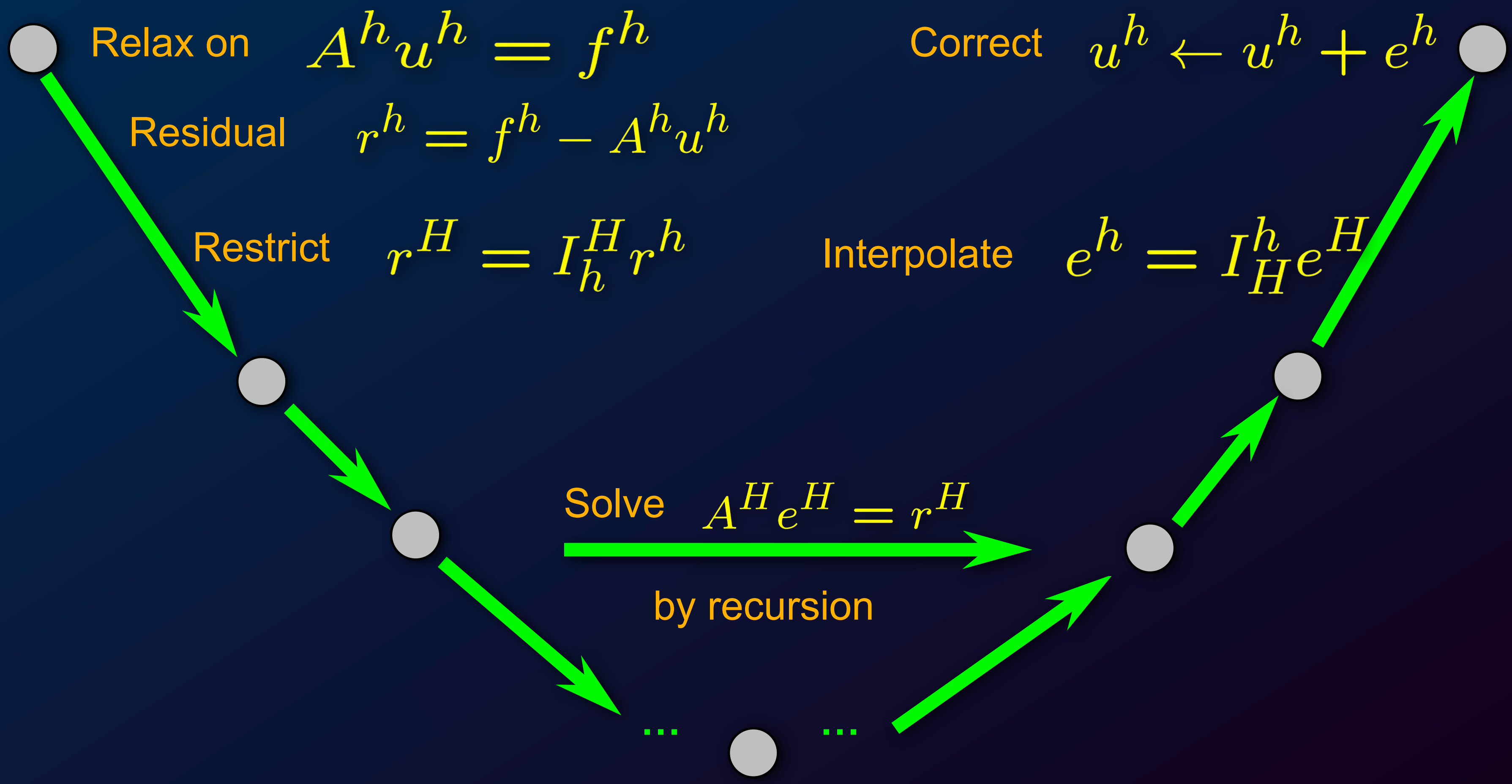
- Assumptions:
  - Multigrid requires 27.5 Ops/unknown to solve an elliptic PDE (Griebel ´89 for 2-D Poisson)
  - A modern laptop CPU delivers >10 GFlops peak
- Consequence:
  - We should solve one million unknowns in 0.00275 seconds
  - ~ 3 ns per unknown
- Revised Assumptions:
  - Multigrid takes 500 Ops/unknown to solve your favorite PDE
  - you can get 5% of 10 Gflops performance
- Consequence: On your laptop you should
  - solve one million unknowns in 1.0 second
  - ~ 1 microsecond per unknown
- Consider Banded Gaussian Elimination on the Play Station (Cell Processor), single Prec. 250 GFlops, for 1000 x 1000 grid unknowns
  - ~2 Tera-Operations for factorization - will need about 10 seconds to factor the system
  - requires 8 GB Mem.
  - Forward-backward substitution should run in about 0.01 second, except for bandwidth limitations

Towards Scalable FE Software

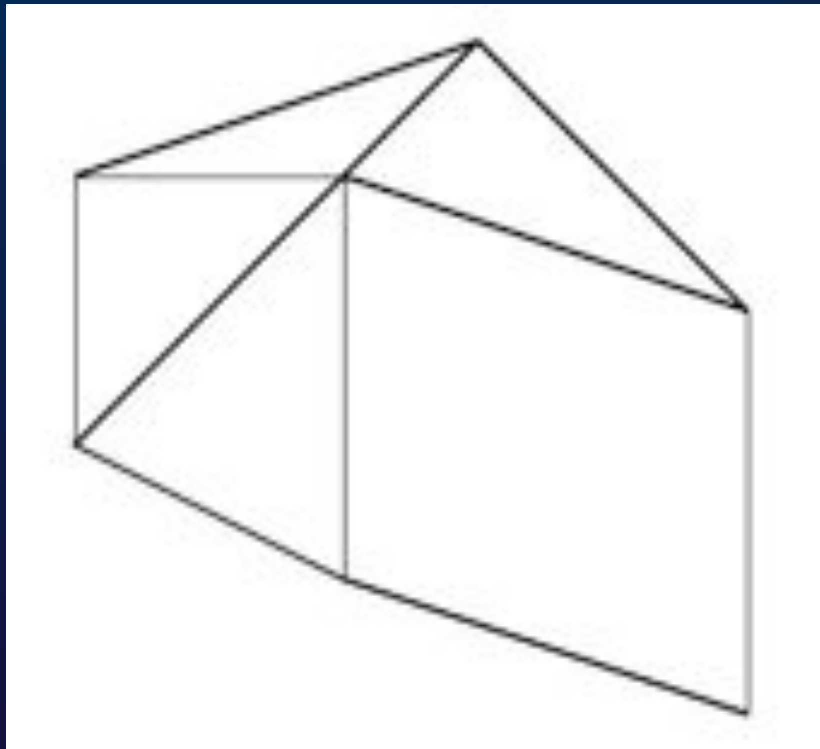
Scalable Algorithms
and Data Structures

Multigrid: V-Cycle

Goal: solve $A^h u^h = f^h$ using a hierarchy of grids

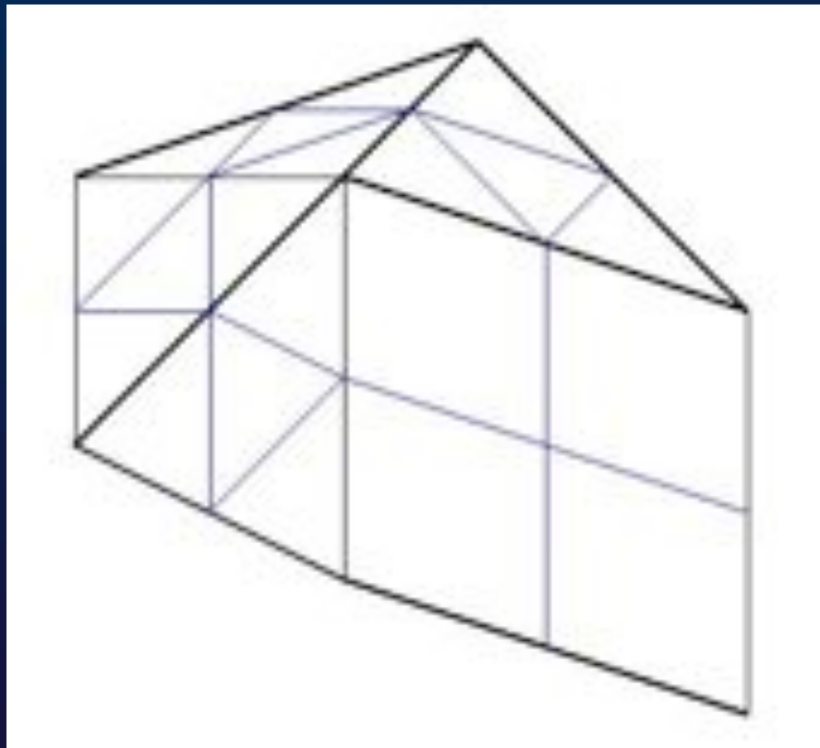


HHG refinement example



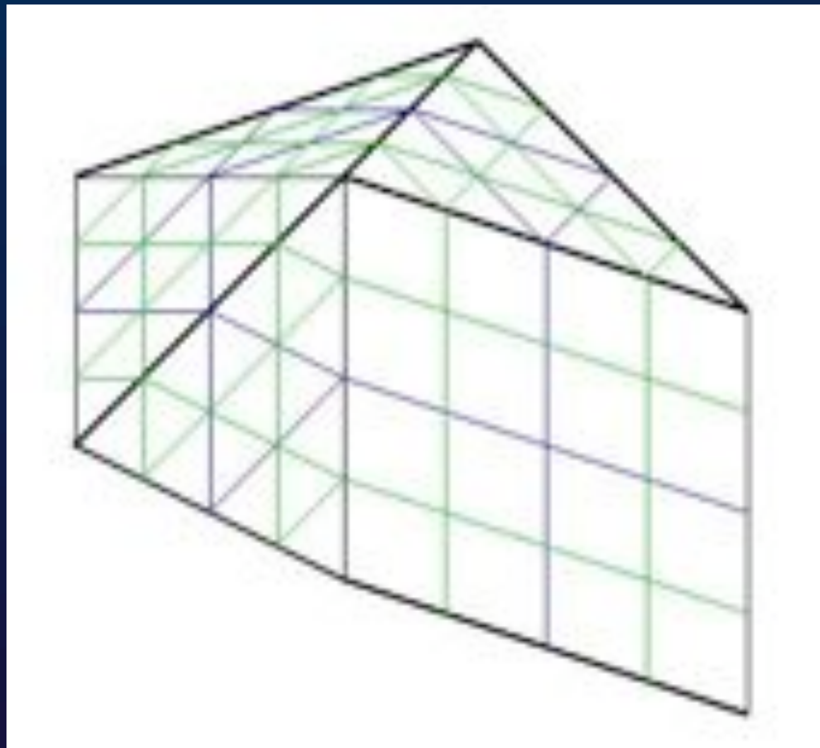
Input Grid

HHG Refinement example



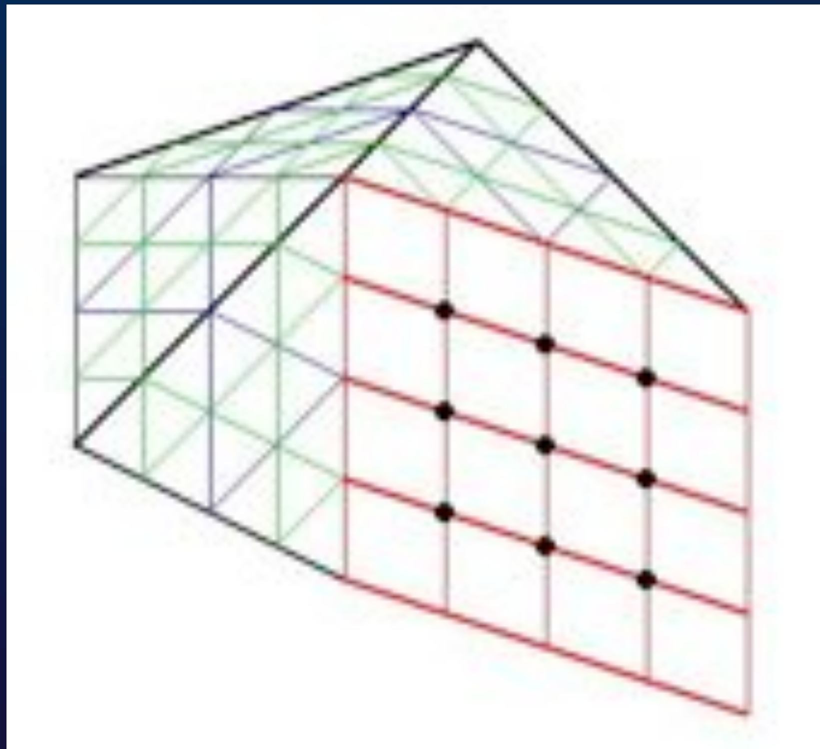
Refinement Level one

HHG Refinement example



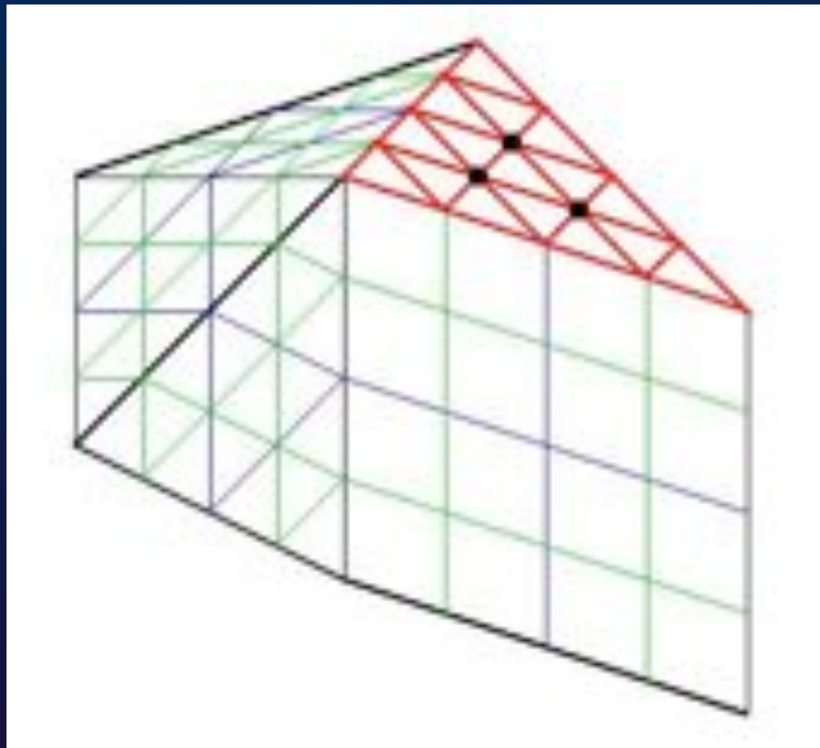
Refinement Level Two

HHG Refinement example



Structured Interior

HHG Refinement example



Structured Interior

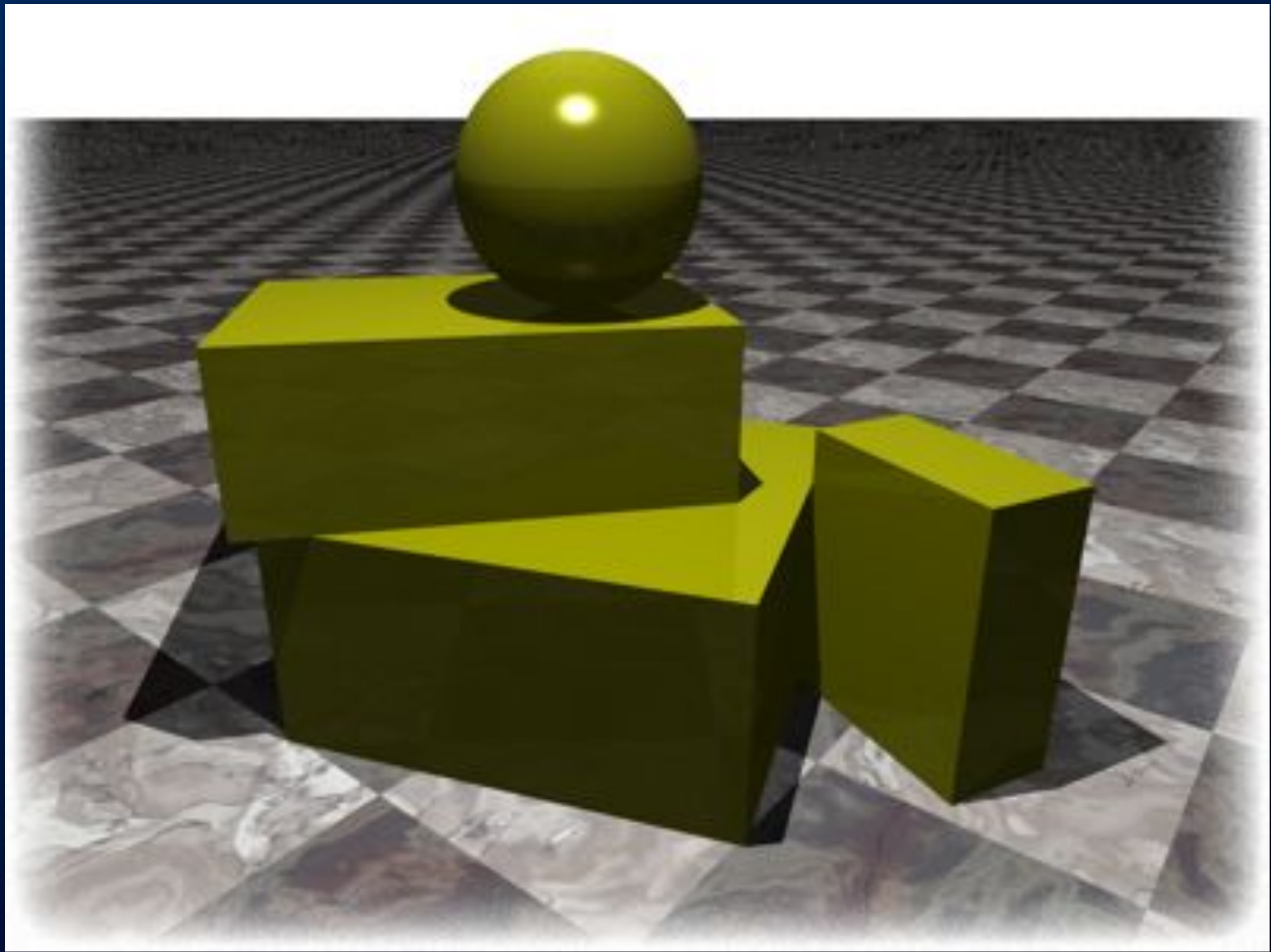
HHG Scalability on Jugene

Cores	Struct. Regions	Unknowns	CG	Time
128	1 536	534 776 319	15	5.64
256	3 072	1 070 599 167	20	5.66
512	6 144	2 142 244 863	25	5.69
1024	12 288	4 286 583 807	30	5.71
2048	24 576	8 577 357 823	45	5.75
4096	49 152	17 158 905 855	60	5.92
8192	98 304	34 326 194 175	70	5.86
16384	196 608	68 669 157 375	90	5.91
32768	393 216	137 355 083 775	105	6.17
65536	786 432	274 743 709 695	115	6.41
131072	1 572 864	549 554 511 871	145	6.42
262144	3 145 728	1 099 176 116 223	180	6.81
294912	294 912	824 365 314 047	110	3.80

An Example of High Performance Multi-Scale and Multi-Physics Simulation

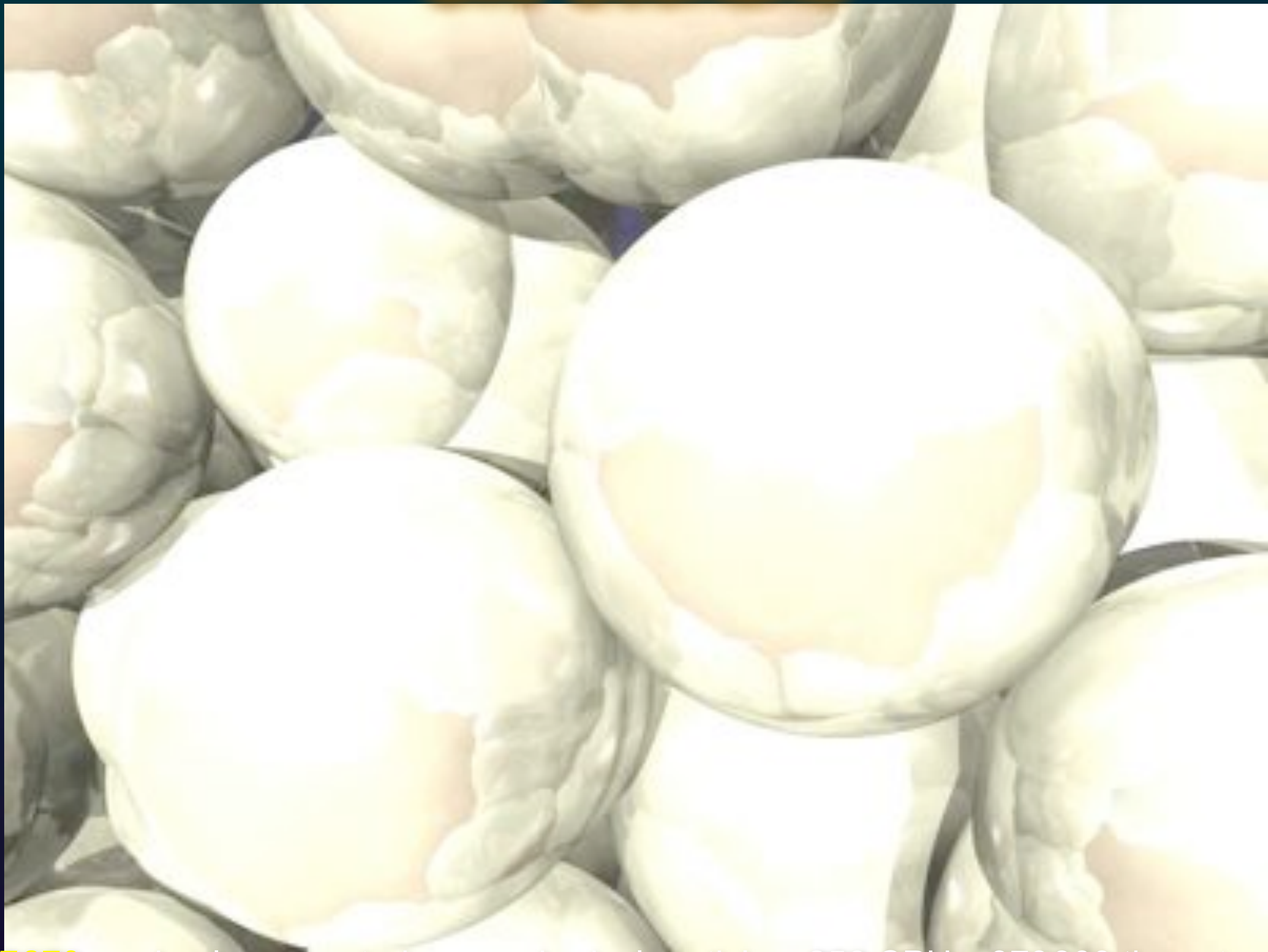
(without multigrid)

Collisions & Contacts between Rigid Objects



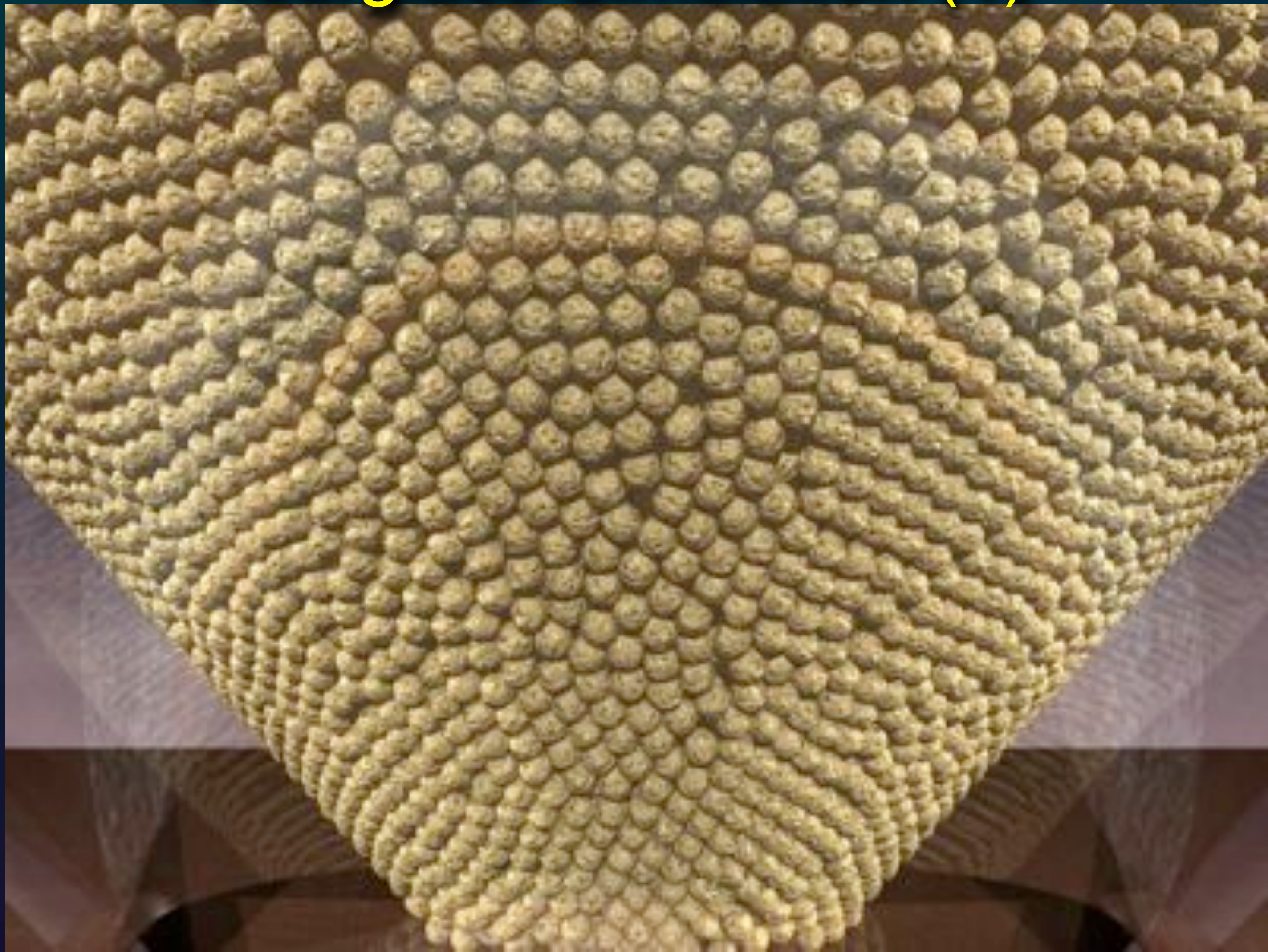
Granular Media Simulations

Silo Scenario



27270 randomly generated, non-spherical particles, 256 CPUs, 379300 time steps, runtime: 16.4 h (including data output), 0.154 s per time step

Hourglass Simulation (1)



1 250 000 spherical particles, 256 CPUs, 300 300 time steps, runtime: 48 h (including data output)

How far can we go? Scaling Results!

# Cores	# Particles	Partitioning	Runtime [s]
128	2 000 000	8 x 4 x 4	727.096
256	4 000 000	8 x 8 x 4	726.991
512	8 000 000	8 x 8 x 8	727.150
1 024	16 000 000	16 x 8 x 8	727.756
2 048	32 000 000	16 x 16 x 8	727.893
4 096	64 000 000	16 x 16 x 16	728.593
8 192	128 000 000	32 x 16 x 16	728.666
16 384	256 000 000	32 x 32 x 16	728.921
32 768	512 000 000	32 x 32 x 32	729.094
65 536	1 024 000 000	64 x 32 x 32	728.674
131 072	2 048 000 000	64 x 64 x 32	728.320

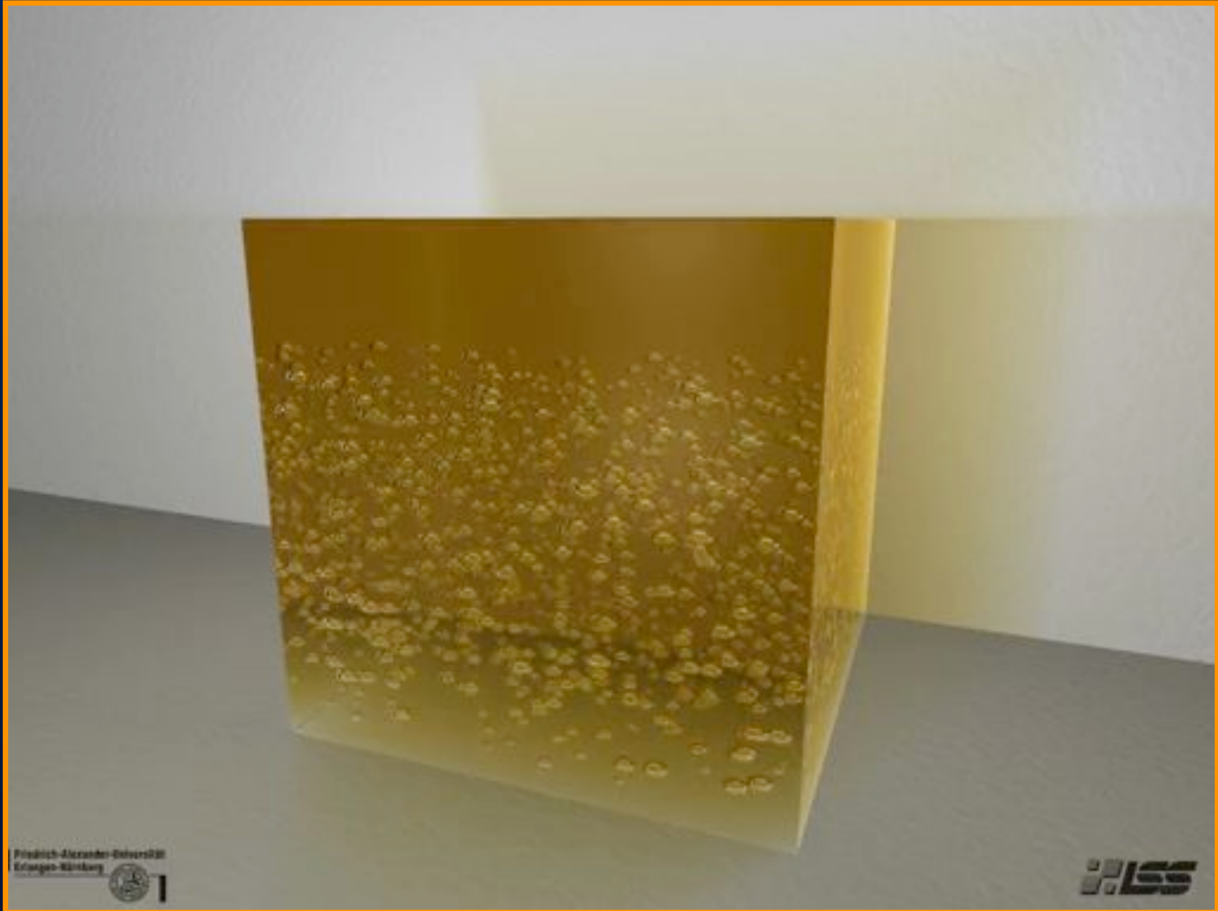
* Jugene simulation results of 1000 time steps of a dense granular gas contained in an evacuated box without external forces. Klaus Iglberger, PRACE Award, 2010.

Flow Simulation with Lattice Boltzmann Methods

Simulation of Metal Foams

- Example application:
  - Engineering: metal foam simulations
- Based on LBM:
  - Free surfaces
  - Surface tension
  - Disjoining pressure to stabilize thin liquid films
  - Parallelization with MPI and load Balancing
- Collaboration with C. Körner (Dept. of Material Sciences, Erlangen)
- Other applications:
  - Food processing
  - Fuel cells

Larger-Scale Computation: 1000 Bubbles



Simulation

1000 Bubbles
510x510x530 =
$1.4 \cdot 10^8$ lattice cells
70,000 time steps
77 GB
64 processes
72 hours
4,608 core hours

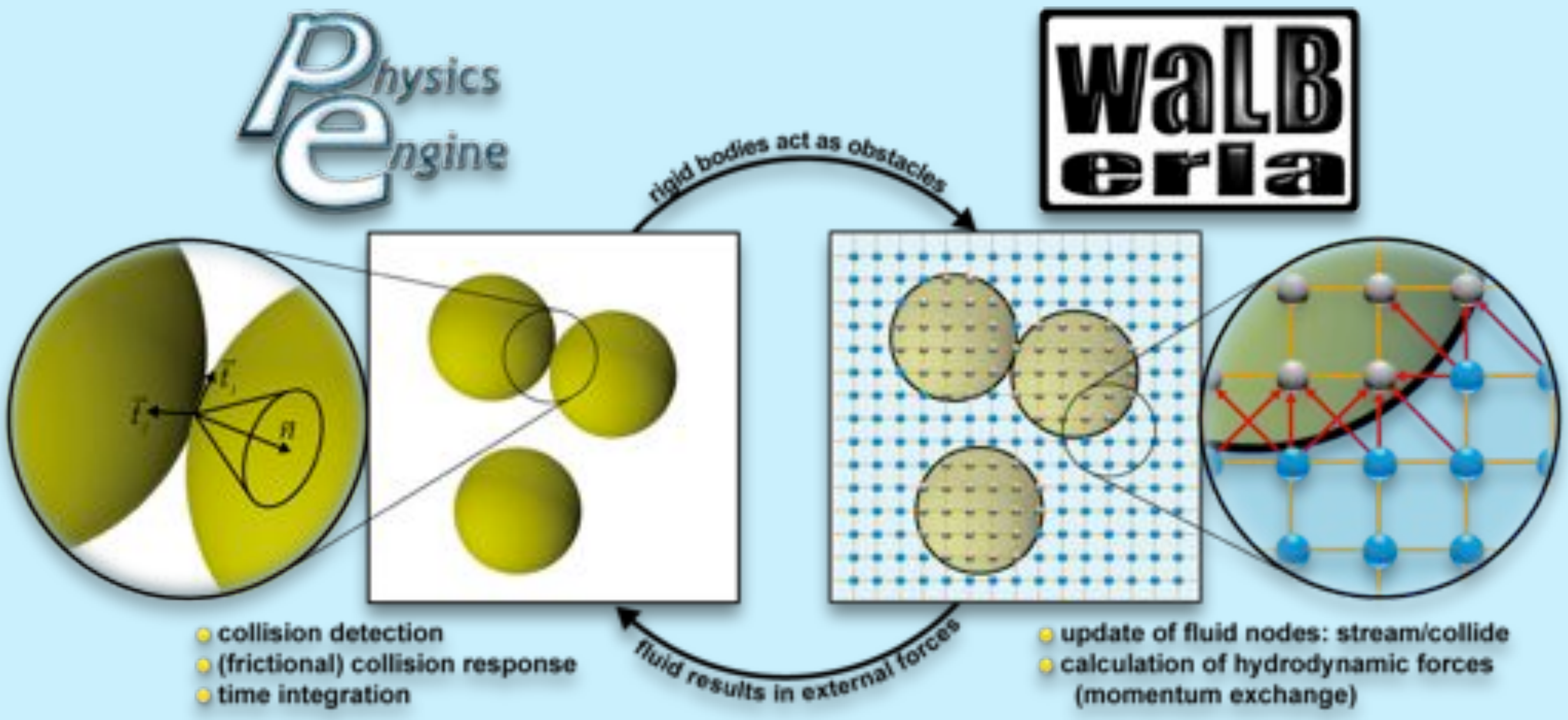
Visualization

770 images
Approx. 12,000 core hours for rendering

Best Paper Award for Stefan Donath (LSS Erlangen) at ParCFD, May 2009 (Moffett Field, USA)

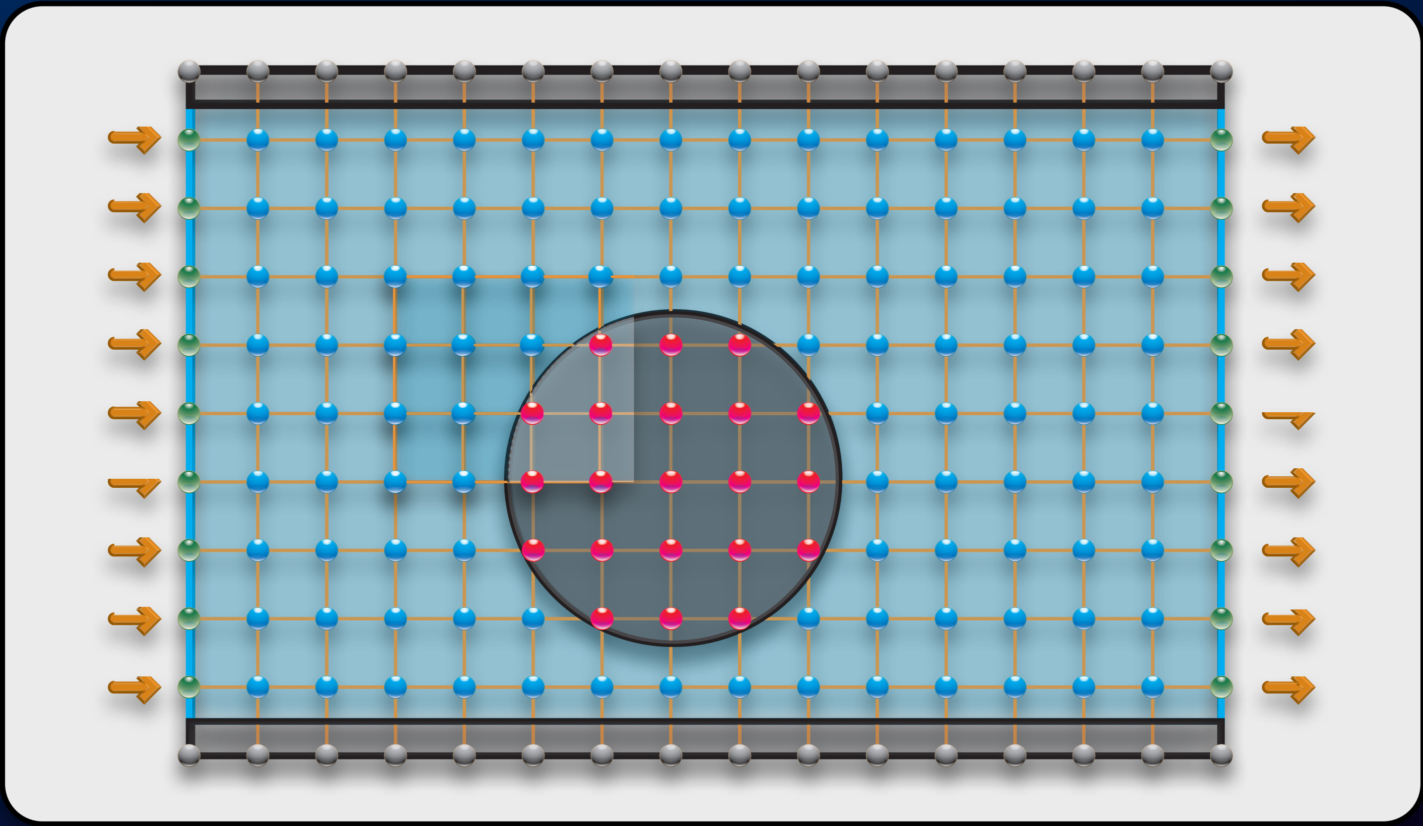
Fluid-Structure Interaction with Moving Rigid Bodies

Fluid-Structure Interaction

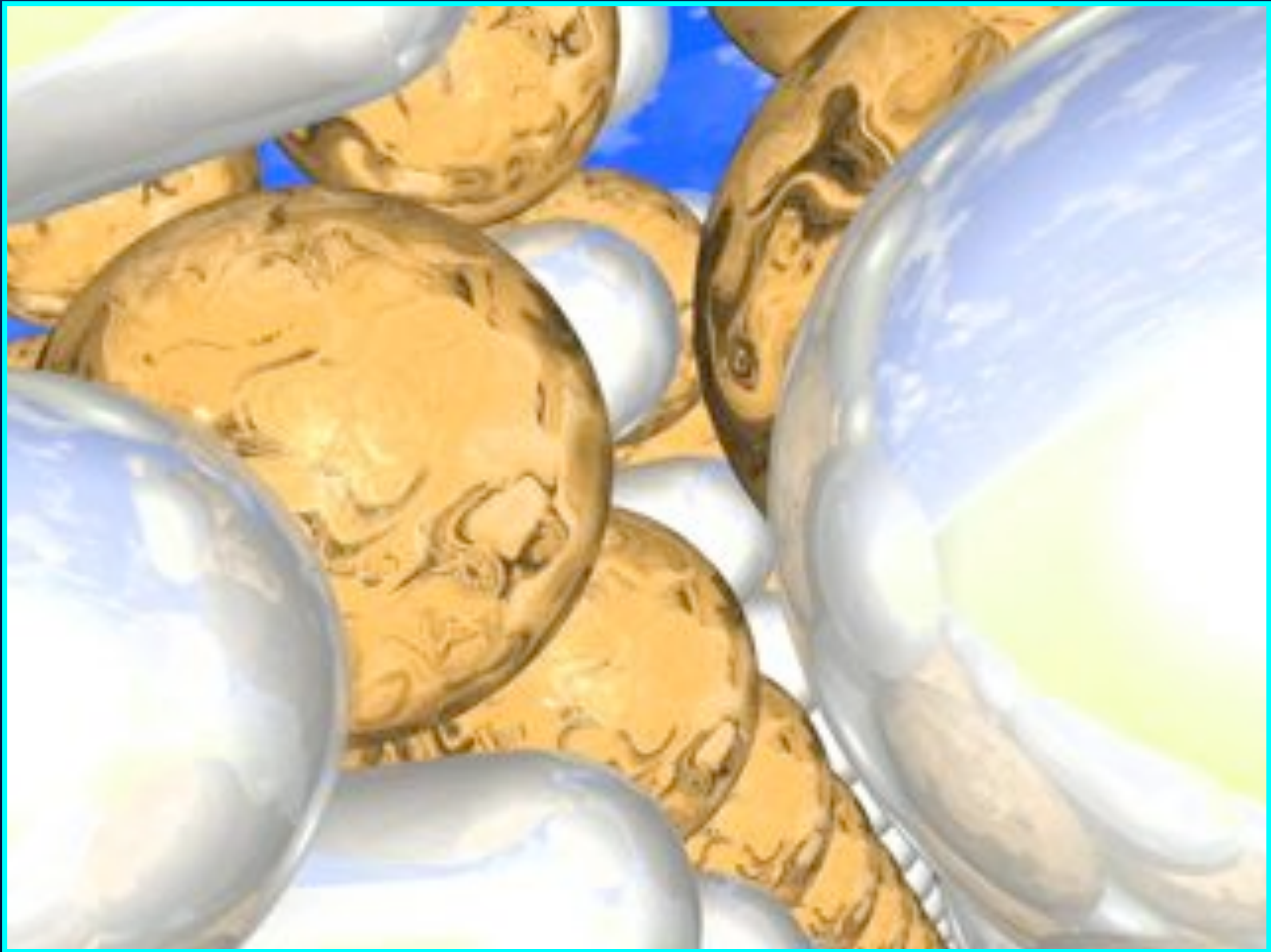


Mapping Moving Obstacles into the LBM Fluid Grid

An Example



Virtual Fluidized Bed



512 processors

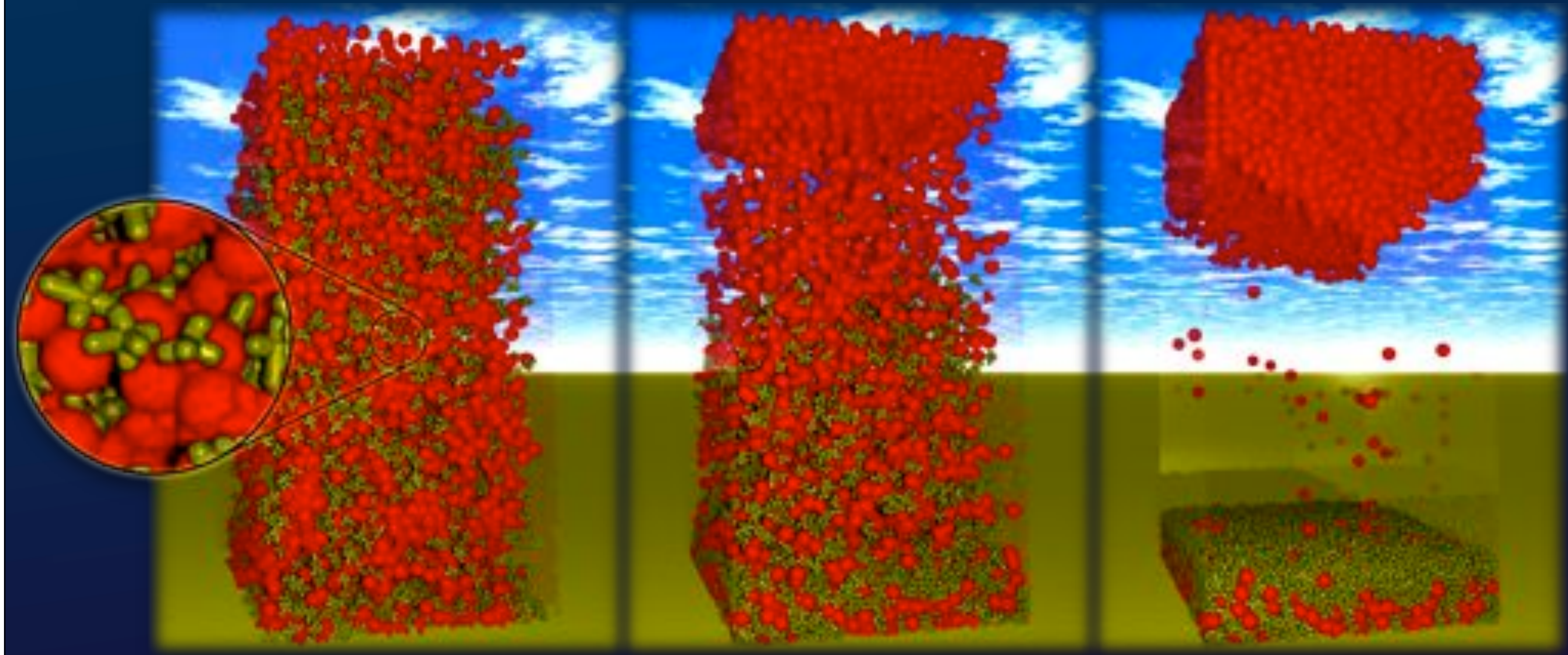
Simulation Domain Size: 180x198x360 cells of LBM

900 capsules and 1008 spheres
= 1908 objects

Number time steps:
252,000

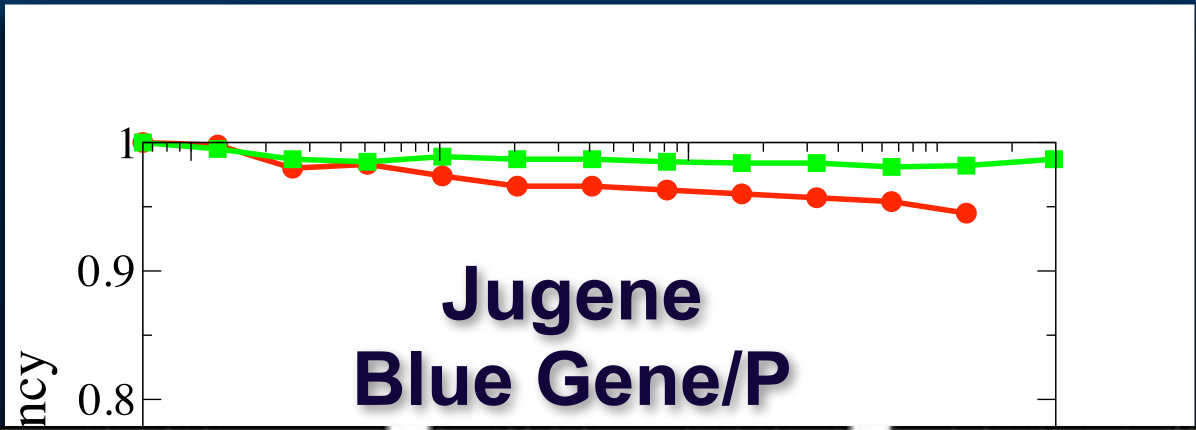
Run Time:
07h 12 min

Simulation of a Segregation Process

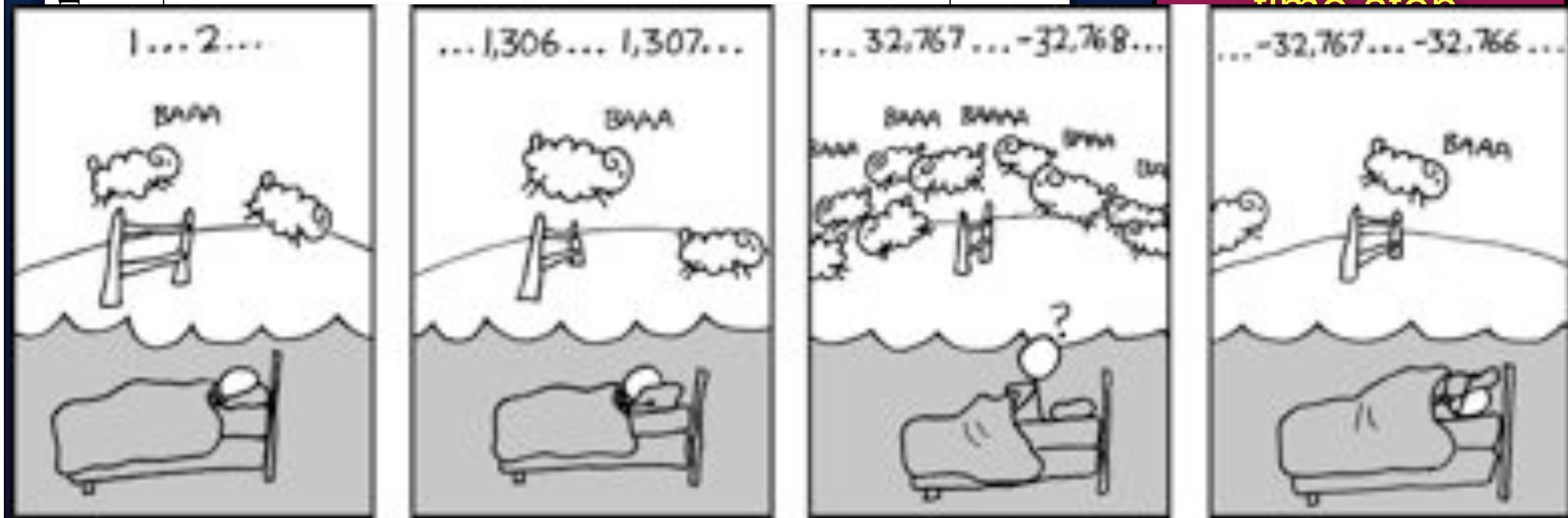


Segregation simulation of 12 013 objects. Density values of 0.8 kg/dm^3 and 1.2 kg/dm^3 are used for the objects in water.

Weak Scaling



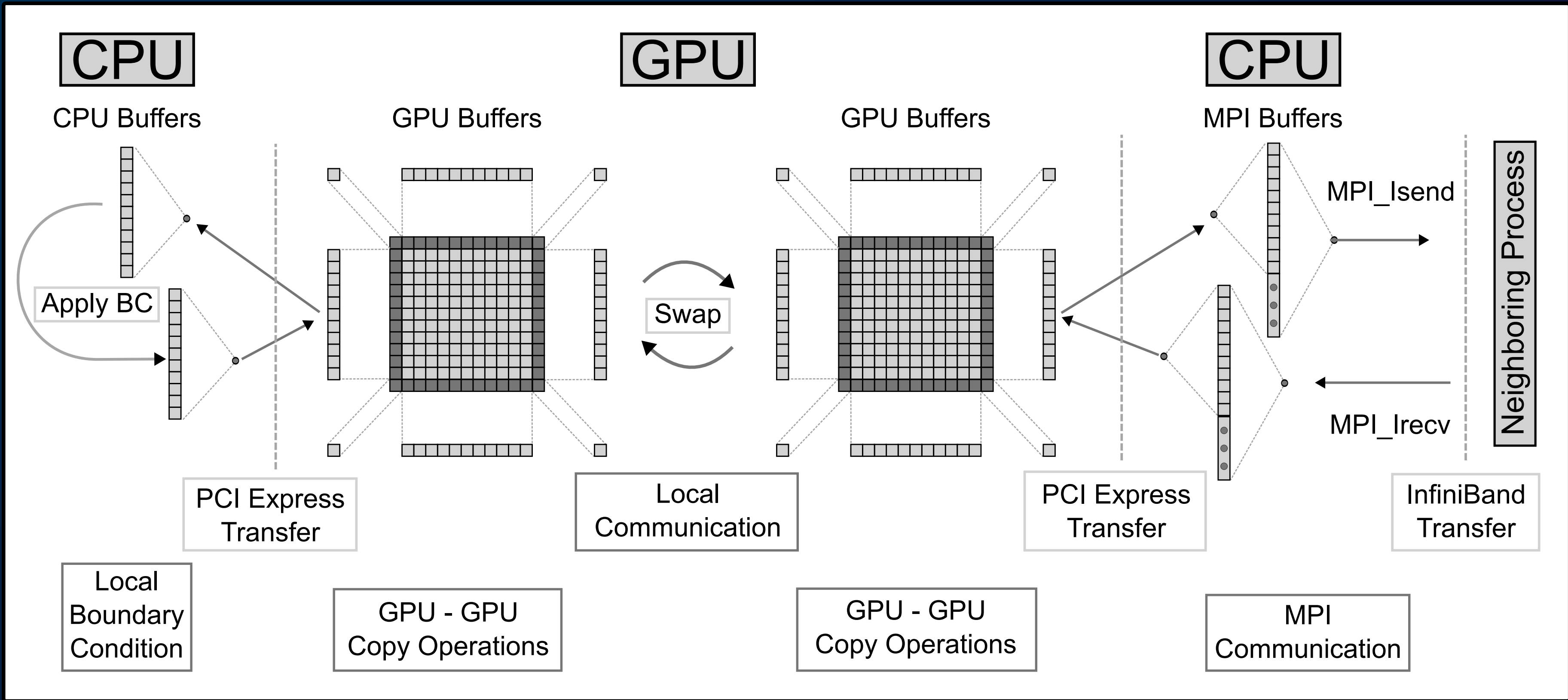
Largest simulation to date :
8 Trillion (10^{12}) variables per
time step



- Densely packed particles
- 264 331 905 rigid spherical objects

LBM
on
Clusters with GPUs

waLBerla Software Architecture for GPU Usage

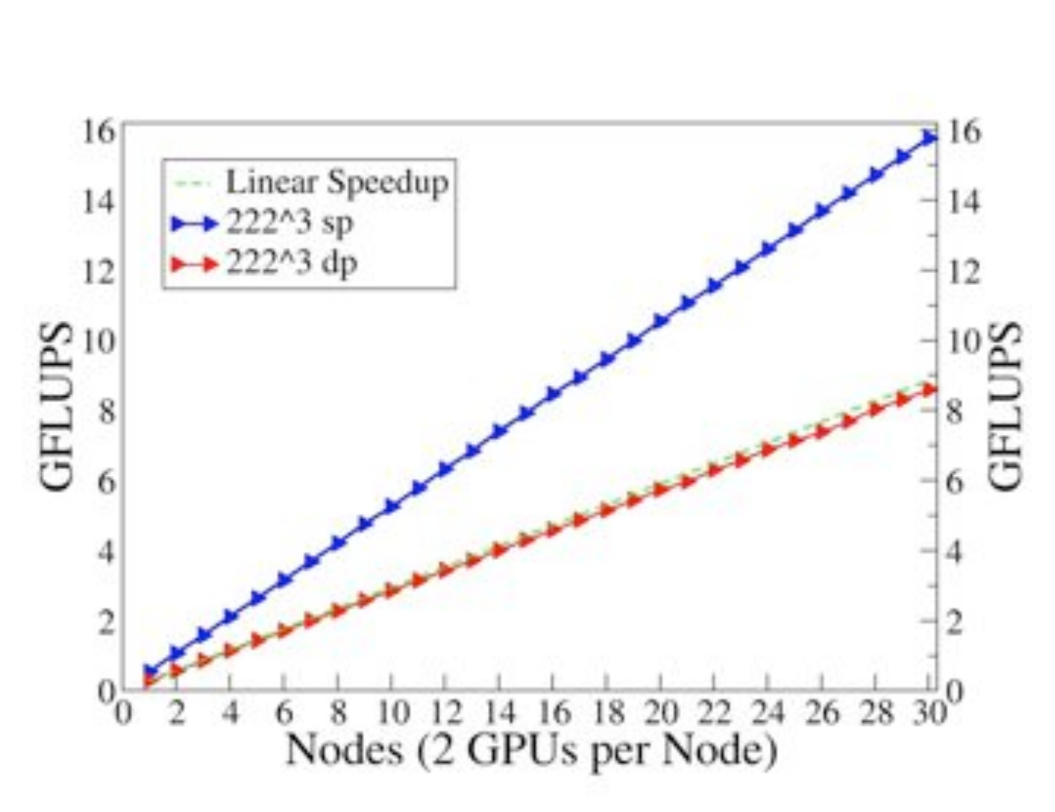


- Patch Architecture
- Only LBM on GPU
  - no free surfaces
  - no FSI

- NEC Nehalem
  - Xeon E5560
  - 2.8 GHz
  - 12 GB per Node
  - 2 GPUs per Node

- nVIDIA TESLA S1070
- 30 Nodes
  - up to 60 GPUs

GPU Performance Results and Comparison



How far is it to do „Real Time CFD"?

25 GLups would compute

- 25 Frames per second for a LBM grid with
- resolution 1000 x 1000 x 1000

- Up to 500 MLup/s on a single GPU for plain LBM kernel (SP)
- 250 MLups/s for GPU in cluster
- Compares to 75 MLup/s for Nehalem Node (8 cores)
- A GPU node (2 GPUs) delivers performance like
  - 6 Nehalem Nodes (48 cores)
  - 75 IBM Blue Gene/P Nodes
- 30 GPU nodes (60 GPUs) are equivalent to
  - 137 Nehalem nodes (1096 cores)
  - 1275 Jugene/P nodes (5100 cores)

Conclusions

The ~~Two~~ Three Principles of Science

Theory

Mathematical Models,
Differential Equations,
Newton

Experiments

Observation and prototypes

empirical Sciences

Computational Science

Simulation, Optimization

(quantitative) virtual Reality

Thank you for your attention!

Slides, reports, thesis, animations available for download at:

www10.informatik.uni-erlangen.de