Performance and Accuracy of Lattice-Boltzmann Kernels on Multi- and Manycore Architectures

Dirk Ribbrock, Markus Geveler, Dominik Göddeke, Stefan Turek

Angewandte Mathematik, Technische Universität Dortmund

May 31, 2010

Overview

1. High performance CFD simulations
   - Hardware
   - Software
   - Implementational issues

2. Results

3. Conclusion + Future Work

Overview

technische universität dortmund

1. High performance CFD simulations
   - Hardware
   - Software
   - Implementational issues
2. Results
3. Conclusion + Future Work

From serial to parallel hardware

technische universität dortmund

Different levels of parallelism

1. *microprocessor-level* (and beyond): IL, SIMD, multi-threading on a chip
2. *node-level*: parallel microprocessors in a compute-node
3. *clusters*: parallel compute-nodes in a cluster

Multi- and many-Core architectures

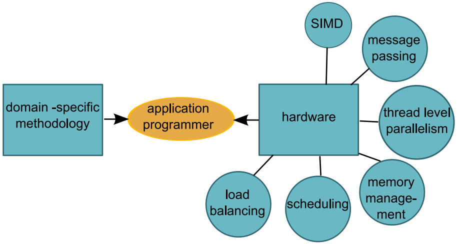
1. doubling the amount of transistors ⇒ doubling the amount of *cores*
2. *multi-core*-architectures have been established in the mass market
3. *many-core*-architectures too (GPUs)
4. ever more *heterogeneities*: Cell BE

The challenge

Application programmers have to deal with...

...both, implementing specific functionality and the paradigm shift of the underlying hardware:

- high-level optimisation on the application level
- hardware-oriented implementation to exploit parallel hardware



today: focus on the method with respect to *multi-level* parallelism

Efficient methods for CFD

Application programmers have to deal with...

- increase in data being acquired for fast ad-hoc processing or future analysis
- applications hunger for ever larger amounts of processing and memory resources
- application design is highly interdisciplinary

The impact on CFD simulation software

- CFD becomes ubiquitous: applications in many different fields
- fast access to result data is an economic issue, software solutions must be:
  - reliable, portable, designed for testability
  - *extendable*, flexible
- real-time constraints become more important

Model Problem

technische universität dortmund

incompressible, laminar fluids with *free surfaces*

Render fluid volumes at *interactive rates* for use in a 3D environment (interactive demonstrations, computer games, tsunami forecast,...)

Quality criteria

- real-time constraint
- embedded in a virtual environment
- qualitative correctness: correct behaviour
- quantitative correctness: numerical quality less important
- stability (!)

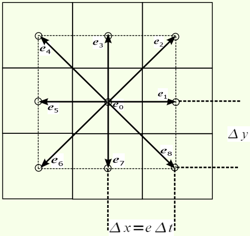
Design goals

- implementation within a building blocks library for maximum reusability
- exploit multi-level parallelism

LBM for SWE

- direct method
- simulates particle propagation
- using a lattice with nine propagation directions

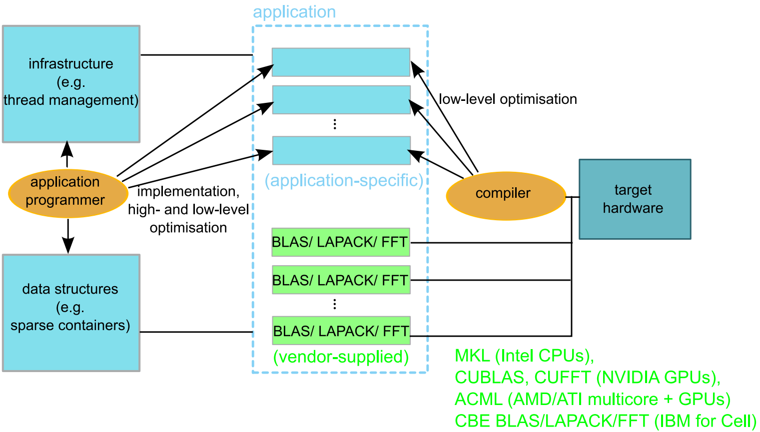
D2Q9 lattice



LBE

$$f_\alpha(\mathbf{x} + \mathbf{e}_\alpha \Delta t, t + \Delta t) = f_\alpha(\mathbf{x}, t) - \frac{1}{\tau}(f_\alpha - f_\alpha^{eq})$$

Implementation

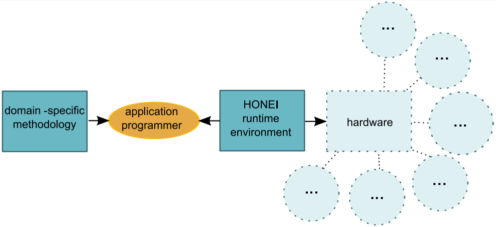


but: Relying upon compilers? Dealing with hardware details? Memory wall? ...?

Framework:HONEI libraries

Primary design-goal:

Abstract from target-hardware!

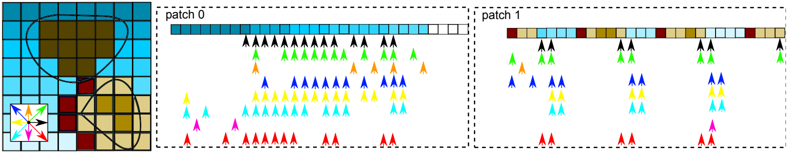


Backends:

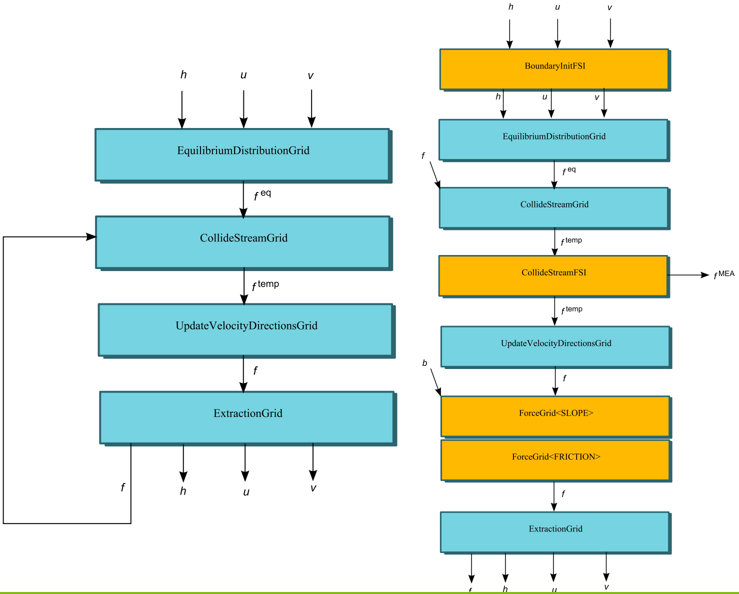
- x86 SIMD (SSE2 intrinsics)
- x86 multi-core (pthreads)
- distributed memory clusters (MPI)
- NVIDIA GPUs (CUDA 2.2) + multiple GPUs
- Cell BE (libspe2)

Packed lattice, domain decomposition

- cut domain in one dimension
- compress lattice: stationary obstacles
- store contiguously in memory
- well suited for SIMD / SIMT



Define building blocks

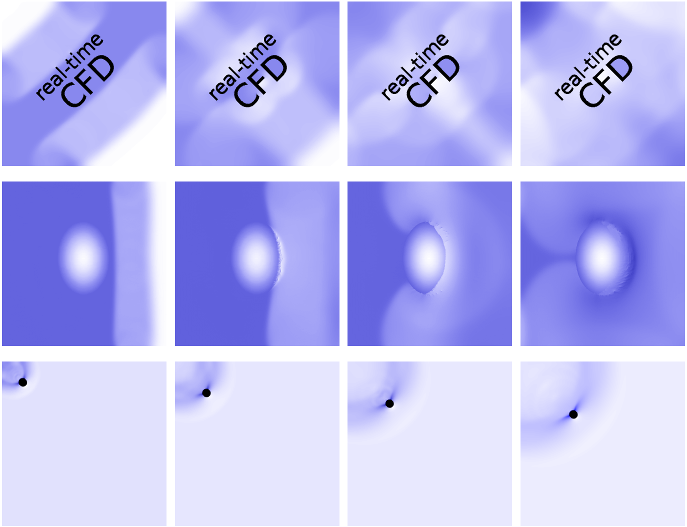


Overview

technische universität
dortmund

1. High performance CFD simulations
   - Hardware
   - Software
   - Implementational issues
2. Results
3. Conclusion + Future Work

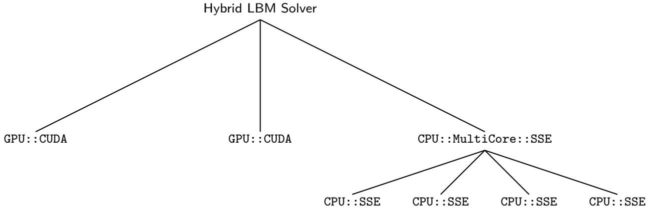
Results



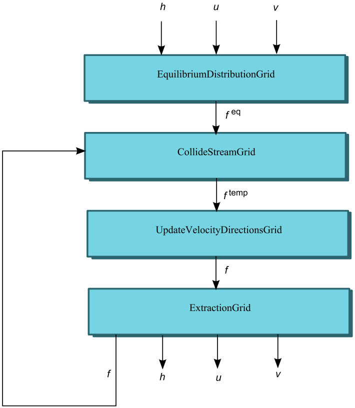
technische universität dortmund

Solver Performance

Type	LBM (1500^2 sites)	
	MFLUPS	Speedup
QS22 Blade	22.36	-
1 Core i7 core	15	-
4 Core i7 cores	40	2.66
1 GTX 285	193	-
2 GTX 285	348	1.80
4 CPU cores + 1 GPU	217	-
4 CPU cores + 2 GPUs	362	1.66



technische universität dortmund



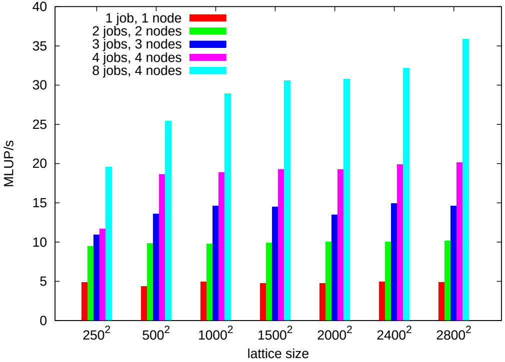
Kernel runtime breakdown

	EquilibriumDistribution	CollideStream	Extraction	Boundaries
Opteron 3 threads	61	24	15	0
Core i7 4 threads	48	35	17	0
QS22 blade	7	72	16	5
GeForce 8800 GTX	19	52	19	10
GeForce GTX 285	32	37	24	7

technische universität dortmund

Opteron cluster, MPI, strong scalability

4 nodes, Opteron 2214



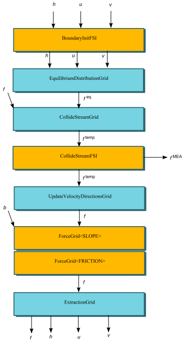
LiDO, weak scalability, n=600

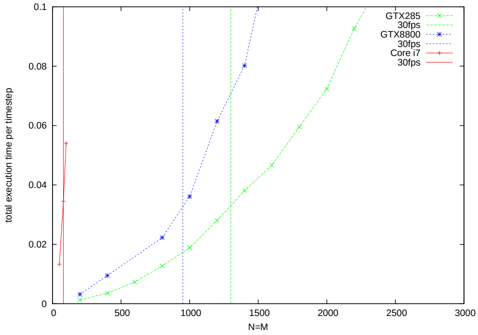
#lattices	time
n^2	9.5
$2n^2$	9.7
$4n^2$	9.7
$8n^2$	9.8
$16n^2$	9.9

technische universität dortmund

Real-time performance on the GPU - hardware

GTX 8800, GTX 285





Overview

technische universität
dortmund

1. High performance CFD simulations
   - Hardware
   - Software
   - Implementational issues

2. Results

3. Conclusion + Future Work

Conclusion + Future Work

Lessons learned

- Advanced CFD applications can be performed in real-time on a single node with reasonable resolution
- The LBM is well suited for this kind of application: it is
  - easy to extend on the discrete level
  - in its basics, well suited for spatial parallelism
- Drawbacks:
  - presented method only first order accurate
  - always: hard to stabilise

Perspectives

- More numerical and performance comparisons with state-of-the-art SWE solvers (FD, FV)
- Large scale faster-than-real-time simulations for tsunami forecast
- Use higher order numerics for real-time simulations (FEM)

Acknowledgements

technische universität dortmund

Supported by DFG, project TU 102/22-2
and BMBF, *HPC Software für skalierbare Parallelrechner: SKALB project 01IH08003D*.
Thanks to IBM for access to Cell hardware.