# Integrated Co-Design of Future Exascale Software

[Extended Abstract]

Björn Gmeiner
University
Erlangen-Nuremberg
Cauerstraße 11
91058 Erlangen, Germany
bjoern.gmeiner@fau.de

Markus Huber
Technische Universität
München
Boltzmannstraße 3
85748 Garching, Germany
huber@ma.tum.de

Lorenz John
Technische Universität
München
Boltzmannstraße 3
85748 Garching, Germany
john@ma.tum.de

Ulrich Rüde
University
Erlangen-Nuremberg
Cauerstraße 11
91058 Erlangen, Germany
ulrich.ruede@fau.de

Barbara Wohlmuth
Technische Universität
München
Boltzmannstraße 3
85748 Garching, Germany
wohlmuth@ma.tum.de

## ABSTRACT

The co-design of algorithms for the numerical approximation of partial differential equations is essential to exploit future exa-scale systems. Here, we focus on key attributes such as node performance, ultra scalable multigrid methods, scheduling techniques for uncertain data, and fault tolerant iterative solvers. In the case of a hard fault, we combine domain partitioning with highly scalable geometric multigrid schemes to obtain fast fault-robust solvers. The recovery strategy is based on a hierarchical hybrid concept where the values on lower dimensional primitives such as faces are stored redundantly and thus can be recovered easily. The lost volume unknowns are re-computed approximately by solving a local Dirichlet problem on the faulty subdomain. Different strategies are compared and evaluated with respect to performance, computational cost, and speed up. Locally accelerated strategies resulting in asynchronous multigrid iterations can fully compensate faults.

## Categories and Subject Descriptors

[**Algorithms**]; [**Performance**]

## 1. INTRODUCTION

For incompressible fluid flow, the Stokes system is an essential building block. We consider examples motivated by coupled non-isothermal models in geo-physics and microfluidics. Such flow problems may require extremely fine resolutions with more than $10^{12}$ unknowns. Even on the most advanced supercomputers, the fast solution of such systems of equations is a highly nontrivial and challenging task [3, 4, 5, 7, 10]. We investigate different iterative saddle-point solvers for the Stokes system. These methods are realized in the hierarchical hybrid grids (HHG) framework, which is a carefully designed and implemented high performance finite element geometric multigrid software package. HHG is designed to combine the flexibility of unstructured finite element meshes with the performance advantage of structured grids in a block-structured approach. The operations are inherently local and well-suited for parallel computations on a distributed memory system using message passing with MPI.

## 2. TEXTBOOK MULTIGRID EFFICIENCY

To evaluate the algorithmic and architectural performance of the solver, we extend Achi Brandt's notion of textbook multigrid efficiency (TME) to massively parallel systems [1, 2, 7]. The quantitative ECM performance model [7] shows that the HHG implementation is highly efficient, in the sense that it actually achieves a performance close to the theoretically expected maximum on the architectures considered.

This allows us to predict a narrow corridor for the single-socket performance, and to identify the critical resource. In particular, this proves that HHG reaches an almost optimal efficiency without potential for significant further optimization on current architectures.

To assess the performance and scalability of Stokes solvers, we started with a pressure-correction algorithm in the hierarchical hybrid grid framework [6, 7]. Here, a preconditioned conjugate gradient solver acts on an inexact pressure Schur-complement. Based on performance engineering, as described above, we identify the Stokes solver as a potential component for algorithmic improvement, indicating that it is possible to improve the algorithmic efficiency significantly by selecting a different solver.

Alternatively, we investigate two more types of solvers: a MINRES variant for the saddle point problem and a multigrid method acting on both system variables, i.e., the velocity and the pressure, see [9]. The MINRES is accelerated by a block diagonal preconditioner, where for the velocity components we use a standard geometric multigrid method and for the pressure a lumped mass matrix being spectrally equivalent to the Schur complement. The multigrid for the indefinite system is based on Uzawa-type smoothers. These are found to be an attractive choice, since they are numerically cheap and can be implemented with only collective nearest-neighbor communication.

High node performance together with good strong and weak scaling is achieved within HHG framework which not only illustrates the excellent parallel efficiency of carefully implemented modern multigrid algorithms, but also demonstrates that a co-design of models, discretization, algorithms, and their parallel implementation based on a systematic performance engineering methodology can lead to numerical solvers that are not just asymptotically optimal but also fast as measured by time-to-solution. Further, we present results, illustrating the number of operator applications for the individual solvers for identical termination criteria. Here, we observe that the Uzawa multigrid method profoundly profits from comparatively fewer applications of the viscous block which is considered the most expensive part of the discrete operator. These numbers are also reflected in the time-to-solution for the individual solvers.

Beyond the solution of single deterministic problems, we progress to uncertain systems that require large numbers of system evaluations. Here, the multigrid hierarchy can be exploited by multi-level Monte-Carlo methods. This has a huge potential for advanced scheduling algorithms that operate on three different levels of concurrency. Genetic algorithms form the basis of heterogeneous scheduling techniques that prove to be robust and efficient.

# 3. ALGORITHMIC RESILIENCE

With the increasing number of cores in high-performance systems, the probability of a failure grows and the necessity of fault tolerant methods is reinforced. System level fault tolerance techniques are considered to be too expensive, since they are often based on redundancy. This is inappropriate in the context of future exa-scale systems due to hardware cost and energy consumption. Also rollback-and-restart check-pointing methods are limited by the excessive cost of storing system state to backup memory. Here, we investigate an algorithm-based fault tolerant approach to incorporate a resilience of process failures in our multigrid algorithm. We study the algorithmic influence of the fault within a geometric multigrid method. As test case, we consider the iterative solution of the Poisson equation by multigrid cycles [8]. When a computing node fails, the information of the faulty subdomain is lost and must be recovered efficiently, where we recompute these values by solving a local subproblem in the faulty subdomain. Dirichlet-Neumann strategies well-known from domain partitioning form the starting point. To accelerate the recovery process, we propose a *superman strategy* which consists of an additional level of parallelization. In that case additional iterations and a deterioration of the time-to-solution can be avoided.

# 4. ACKNOWLEDGMENTS

This work was partially supported by the German Research Foundation (DFG) through the Priority Programme 1648 "Software for Exascale Computing" (SPPEXA). The authors gratefully acknowledge the Gauss Centre for Supercomputing (GCS) for providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS share of the supercomputer JUQUEEN at Jülich Supercomputing Centre (JSC).

# 5. ADDITIONAL AUTHORS

Martin Bauer (FAU), email: `martin.bauer@fau.de`, Holger Stengel (RRZE), email: `holger.stengel@fau.de` and Christian Waluga (TUM), email: `waluga@ma.tum.de`

# 6. REFERENCES

[1] A. Brandt. Guide to multigrid development. In *Multigrid methods*, pages 220–312. Springer, 1982.

[2] A. Brandt. *Barriers to achieving textbook multigrid efficiency (TME) in CFD*. Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, 1998.

[3] C. Burstedde, G. Stadler, L. Alisic, L. C. Wilcox, E. Tan, M. Gurnis, and O. Ghattas. Large-scale adaptive mantle convection simulation. *Geophys. J. Internat.*, 192(3):889–906, 2013.

[4] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics.* Oxford University Press, New York, 2005.

[5] R. D. Falgout, J. E. Jones, and U. M. Yang. The design and implementation of hypre, a library of parallel high performance preconditioners. In *Numerical solution of partial differential equations on parallel computers*, volume 51 of *Lect. Notes Comput. Sci. Eng.*, pages 267–294. Springer, Berlin, 2006.

[6] B. Gmeiner, U. Rüde, H. Stengel, C. Waluga, and B. Wohlmuth. Performance and Scalability of Hierarchical Hybrid Multigrid Solvers for Stokes Systems. *SIAM J. Sci. Comput.*, 37(2):C143–C168, 2015.

[7] B. Gmeiner, U. Rüde, H. Stengel, C. Waluga, and B. Wohlmuth. Towards textbook efficiency for parallel multigrid. *Numer. Math. Theory Methods Appl.*, 8, 2015.

[8] M. Huber, B. Gmeiner, U. Rüde, and B. Wohlmuth. Resilience for multigrid software at the extreme scale, 2015. Submitted.

[9] M. Huber, L. John, B. Gmeiner, U. Rüde, and B. Wohlmuth. Massively parallel hybrid multigrid methods for the Stokes system, 2015. In preparation.

[10] H. Sundar, G. Stadler, and G. Biros. Comparison of multigrid algorithms for high-order continuous finite element discretizations. *Numer. Linear Algebra Appl.*, 22(4):664–680, 2015.