

REVIEW

of the foreign scientific supervisor for the dissertation thesis submitted by
Mr. Yerlan Makhmut for the degree of Doctor of Philosophy (PhD)
in the specialty 8D06105 – System Engineering, entitled

**“Development of hybrid high-performance algorithms
for numerical simulation of oil displacement problems”**

This review is presented to certify that the doctoral student, Mr. Yerlan Makhmut, has successfully fulfilled all the requirements needed for the conferment of the Ph.D. degree.

Overview and Objectives

The dissertation by Yerlan Makhmut addresses the important topic of modern computer science and system engineering -- development of numerically stable and scalable high-performance computing (HPC) algorithms for simulating the polymer-flooding problem in oil reservoirs. Polymer flooding is a widely applied, enhanced oil recovery technique whose governing equations are strongly non-linear and computationally demanding, which in practice limits the use of high-resolution simulations at field scale. The goal of the work is to design parallel algorithms that span the full computational hierarchy – from a sequential CPU baseline, through shared-memory and distributed-memory CPU parallelism, to both single-GPU acceleration and multi-GPU hybrid execution – and, at each level, to identify the binding hardware constraint and attempt to reach the relevant hardware ceiling, rather than merely demonstrate some speedup.

The mathematical model consists of three strongly coupled, non-linear partial differential equations describing the conservation of the oil phase, the water phase, and the dissolved polymer species, supplemented by Darcy's law, the Brooks-Corey relative permeability model, a power-law polymer viscosity modification, and a Langmuir adsorption isotherm. The coupled system is solved on a one-dimensional structured Cartesian grid using the IMPES (Implicit Pressure – Explicit Saturation) time-stepping scheme.

The dissertation comprises an introduction, eight chapters, and a conclusion. It presents a systematic research study based on numerical modelling, without relying on commercial simulators. The author has disseminated the results through several peer-reviewed journal and conference publications, including Scopus-indexed outputs and presentations at international conferences, which confirms the scientific relevance and practical potential of the proposed methodology.

Key Achievements

One of the central results of the thesis is the systematic development and comparison of seven parallel computational algorithm implementations – OpenMP, MPI, hybrid MPI+OpenMP, CUDA global-memory, CUDA shared-memory, MPI+CUDA, and a triple-hybrid MPI+OpenMP+CUDA – all evaluated against the same serial baseline, on the same governing equations, using the same validation metrics. This controlled experimental design ensures that the performance conclusions are attributable to the chosen parallelization strategy itself, rather than to changes in the underlying mathematical problem.

I would like to stress the following two original scientific contributions of this work. The first is a full-pipelined, shared-memory CUDA tiling approach in which the pressure, saturation, and polymer-concentration kernels are tiled into on-chip memory simultaneously and treated as a coupled computational pipeline, rather than as isolated kernels. The second is a triple-hybrid MPI+OpenMP+CUDA algorithm that combines distributed-memory, shared-memory, and GPU parallelism within a single coherent implementation; it achieves a peak speedup of $34.70\times$ over the serial baseline and wins on five of the six tested grid sizes. The author explains the observed performance through a principled hardware analysis, including the FP64-pipeline ceiling and the L2-cache-residency effect, and corroborates every speedup claim with independent profiling evidence from Intel VTune, mpiP, NVIDIA Nsight Compute, and NVIDIA Nsight Systems. The numerical correctness of all algorithms is confirmed against the serial baseline to a relative error below 2×10^{-8} , with mass conservation preserved to machine precision.

Evaluation of Relevance, Novelty, and Practical Value

This research is highly relevant, as the efficient simulation of multiphase flow plays a crucial role in addressing significant engineering challenges, particularly in enhanced oil recovery and reservoir management. The scientific novelty of the dissertation lies in extending shared-memory tiling from the pressure solver to the complete coupled polymer-flooding pipeline, and in the first academic implementation of a three-level MPI+OpenMP+CUDA algorithm for this class of problems. The author convincingly demonstrates that three independent levels of parallel-

lism can be combined without algorithmic compromise, and the resulting gains are not incidental but follow from a clear understanding of the hardware limits.

The experimental results confirm the comparative effectiveness of the studied methods. The triple-hybrid algorithm delivers its peak speedup of 34.70× at the intermediate grid size and outperforms the pure MPI+CUDA hybrid by up to 5.75×, a margin that the profiling analysis attributes to a four-fold reduction in MPI collective-communication overhead achieved through host-side overlap. These results support the scientific soundness and the practical value of the proposed methods, which are obtained on widely accessible workstation-class hardware and are therefore directly applicable to the computing environments typical for academic research groups and small engineering teams.

Observations and Recommendations

The thesis is well structured and clearly written. It shows that the author fully understands both the physical processes of polymer flooding and the principles of high-performance computing, and is able to connect the two through careful algorithm design, validation, and profiling. The results are convincing and are supported by international publications and conference presentations. As the author himself states, the study has natural limitations – the one-dimensional structured-grid formulation, the use of consumer-class GPUs without NVLink, the assumption of homogeneous and isothermal reservoir properties, and benchmarking on a single workstation. These are stated explicitly and mapped to concrete future-work directions, such as extension to three-dimensional unstructured grids and multi-node deployment with NVLink and GPUDirect. Pursuing these directions would further strengthen this already solid body of work.

Conclusion

Overall, the dissertation is carried out at a high academic level. The methods are well explained, the experiments are solid and reproducible, and the results are useful for both theory and practice. Yerlan Makhmut's work meets the expected standards for a PhD thesis. It offers a meaningful contribution to current research in high-performance reservoir simulation and can help improve modeling tools for enhanced oil recovery studies.

As a foreign scientific supervisor for this dissertation, I strongly recommend the dissertation for defense.



Prof. Dr. habil. Sergei Gorlatch
Foreign Scientific Supervisor for the Dissertation

Münster, Germany, 29.05.2026

