

REVIEW

of the Academic Supervisor
on the Doctoral Dissertation submitted by

Makhmut Yerlan

for the degree of Doctor of Philosophy (PhD)
in Specialty 8D06105 — Systems Engineering

Dissertation title:

“Development of Hybrid High-Performance Algorithms for Numerical Simulation of Oil Displacement Problems”

The present review is submitted with the purpose of attesting that doctoral candidate Makhmut Yerlan has comprehensively and rigorously fulfilled all stipulated requirements for the conferral of the degree of Doctor of Philosophy (PhD).

1. General Characterization and Research Objectives

The dissertation of Makhmut Yerlan addresses a compelling and timely intersection of computational science and systems engineering, namely the conception, formulation, and rigorous benchmarking of numerically stable, highly scalable high-performance computing (HPC) algorithms tailored to the simulation of polymer-augmented waterflooding in subsurface petroleum reservoirs. Polymer flooding constitutes one of the most extensively deployed chemical enhanced oil recovery (EOR) methodologies; nevertheless, the mathematical apparatus governing this process — characterized by pronounced nonlinearity, multi-physics coupling, and substantial computational expenditure — has historically impeded the adoption of high-resolution, field-scale simulation in practical engineering workflows. The overarching objective of the dissertation is to conceive and systematically evaluate parallel algorithms that traverse the entirety of the computational hierarchy: from a sequential single-core CPU reference implementation, through shared-memory and distributed-memory CPU parallelism and their hybridization, to single-GPU and multi-GPU accelerated configurations. At each hierarchical level the author endeavors to identify the dominant hardware bottleneck and to demonstrate attainment of performance that approaches the theoretical ceiling of the deployed hardware.

The governing mathematical model consists of three tightly interlocked nonlinear partial differential equations that encode mass conservation principles for the oil phase, the aqueous phase, and the dissolved polymer species. This system is augmented by Darcy's constitutive law for porous-medium flow, the Brooks-Corey empirical formulation for relative permeabilities, a power-law rheological modification accounting for polymer-induced viscosity alteration, and the Langmuir isotherm describing polymer adsorption onto the rock matrix. The coupled system is spatially discretized on a one-dimensional structured Cartesian grid and temporally integrated by the Implicit-Pressure – Explicit-Saturation (IMPES) operator-splitting scheme, which exploits the elliptic character of the pressure equation to admit an unconditionally stable implicit pressure solve while preserving the computational simplicity of explicit hyperbolic transport updates.

Comprising an introduction, eight substantive chapters, and a synthesizing conclusion, the dissertation pursues a methodologically coherent, simulation-driven scientific inquiry conducted entirely without recourse to commercial reservoir simulators. The scholarly merit and translational potential of the proposed framework are independently corroborated by its dissemination in peer-reviewed journals indexed within the Scopus bibliographic database and through oral presentations at international scientific conferences.

2. Principal Scientific Contributions and Key Findings

Among the most consequential contributions of the dissertation is the systematic conception, implementation, and cross-comparison of seven distinct parallel algorithmic strategies: a shared-memory OpenMP implementation, a distributed-memory MPI implementation, an MPI+OpenMP hybrid, a CUDA global-memory variant, a CUDA shared-memory variant employing explicit on-chip data staging, an MPI+CUDA two-level hybrid, and a three-level MPI+OpenMP+CUDA hybrid that simultaneously exploits all three principal parallelism paradigms. All seven implementations are evaluated against a common sequential baseline, employing identical governing equations, convergence tolerances, and validation criteria throughout. This methodologically controlled experimental design affords unambiguous attribution of observed performance differentials to the parallelization strategy rather than to any alteration of the underlying mathematical formulation.

Two original intellectual contributions of the work merit particular commendation. The first constitutes a fully pipelined CUDA shared-memory tiling methodology in which the pressure, saturation, and polymer concentration computational kernels are concurrently

staged into GPU on-chip shared memory and processed as an integrated, coherently coupled computational pipeline rather than as operationally isolated kernels. This design eliminates redundant global-memory traversals and substantially improves effective arithmetic intensity within the device memory hierarchy. The second contribution is the pioneering academic realization of a three-level MPI+OpenMP+CUDA hybrid algorithm that cohesively unifies distributed-memory inter-node communication, intra-node shared-memory thread-level parallelism, and GPU acceleration within a single, algorithmically consistent implementation. This formulation attains a peak wall-clock speedup of $34.70\times$ relative to the sequential baseline and surpasses all competing implementations at five of the six problem sizes investigated. The author furnishes rigorous, hardware-grounded explanations for the observed performance characteristics, examining in detail the FP64 arithmetic pipeline ceiling imposed by the consumer-class Turing GPU, the ameliorative influence of L2 cache residency on effective memory throughput, and the fourfold attenuation of MPI collective communication latency achieved through host-side asynchronous execution overlapping. Every reported speedup figure is independently substantiated by quantitative profiling data acquired with Intel VTune Profiler, the mpiP lightweight MPI profiling library, NVIDIA Nsight Compute, and NVIDIA Nsight Systems. The numerical fidelity of all seven parallel implementations is rigorously established through cell-by-cell comparison against the sequential reference: the maximum observed L_∞ relative error remains below 2×10^{-8} across all problem sizes, and global mass conservation is upheld to the level of IEEE-754 double-precision machine epsilon.

3. Evaluation of Relevance, Scientific Novelty, and Practical Significance

The present investigation addresses a problem domain of profound and enduring relevance: the efficient, high-fidelity simulation of multiphase subsurface flow occupies a central position in the resolution of pressing engineering challenges, most notably in the optimization of enhanced oil recovery operations and the strategic management of hydrocarbon reservoirs. The scientific novelty of the dissertation resides principally in two advances: the extension of GPU shared-memory tiling from isolated pressure solvers to a fully coupled, multi-physics polymer-flooding computational pipeline; and the inaugural academic realization of a three-tiered MPI+OpenMP+CUDA algorithm specifically tailored to this class of reservoir simulation problems. The author provides compelling evidence that three orthogonal levels of parallelism — inter-node distributed memory, intra-node shared memory, and GPU data parallelism — may be cohesively integrated without algorithmic degradation or loss of

numerical fidelity, and that the resultant performance benefits are not fortuitous but are mechanistically attributable to a precise and nuanced comprehension of the underlying hardware architecture.

The experimental evidence substantiates the comparative superiority of the proposed methodology. The three-level hybrid algorithm achieves its peak acceleration of $34.70\times$ at an intermediate problem scale of 1,048,576 grid cells and demonstrably outperforms the two-level MPI+CUDA hybrid across the majority of tested configurations. Profiling attribution conclusively traces this performance advantage to a fourfold diminution of MPI collective communication overhead, realized through carefully orchestrated host-side concurrent execution. These outcomes substantiate both the theoretical soundness and the operational utility of the advocated approach; given that all reported results are reproducible on commercially accessible workstation-class hardware, the methodology is directly transferable to the computational environments habitually available to academic research laboratories and small-to-medium engineering organizations.

4. Critical Observations and Recommendations

The dissertation is distinguished by a lucid and logically coherent expository structure. The work furnishes compelling testimony to the author's thorough command of both the physical phenomenology of polymer-augmented waterflooding and the architectural principles governing modern high-performance computing platforms, and to his capacity to synthesize these two disciplines through meticulous algorithmic design, comprehensive quantitative validation, and systematic hardware-level profiling. The reported findings are robust and replicable, and have received independent scholarly endorsement through publication in internationally recognized peer-reviewed venues and through conference presentations before specialist audiences. As the author himself candidly acknowledges, the investigation is circumscribed by several inherent constraints: the restriction to a one-dimensional structured-grid spatial formulation, the reliance on consumer-class GPU hardware operating without the benefit of NVLink high-bandwidth interconnect, the adoption of homogeneous and isothermal reservoir property assumptions, and the confinement of performance benchmarking to a single workstation node. Each of these delimitations is articulated with commendable transparency and is systematically mapped to a well-motivated avenue of prospective inquiry — prominently including the generalization to three-dimensional unstructured spatial discretization and the deployment across multi-node cluster architectures exploiting NVLink and GPUDirect communication technologies. Pursuit of these extensions

would further amplify the impact of an already accomplished and substantive scholarly contribution.

5. Concluding Assessment

Upon comprehensive evaluation, the dissertation is adjudged to have been executed at an exemplary academic standard commensurate with the requirements of doctoral-level scholarship. The methodological framework is expounded with precision and intellectual rigor; the experimental program is systematic, reproducible, and internally consistent; and the findings generated are of demonstrable theoretical depth and tangible practical consequence. The dissertation of Makhmut Yerlan satisfies in full the stipulated criteria for the award of the degree of Doctor of Philosophy (PhD). It constitutes a substantive and original contribution to the contemporary state of knowledge in high-performance petroleum reservoir simulation and holds considerable promise for advancing the development of computationally efficient simulation tools in the domain of chemical enhanced oil recovery research.

In my capacity as Academic Supervisor, I unreservedly endorse this dissertation and formally confirm that the candidate fully merits the award of the degree of Doctor of Philosophy (PhD).

Academic Supervisor:

Timur Imankulov

Al-Farabi Kazakh National University,

Faculty of Information Technologies

and Artificial Intelligence,

PhD, Associate Professor

Signature: _____

Date: "01" June 2026

Almaty, Republic of Kazakhstan



РАСТАЙМЫН

Би атындағы Қазақ Ғылыми кадрларды
жарындау және аттестаттау басқармасының басшысы

ЗАВЕРЯЮ

Назарбаев университетінің
Ғылыми кадрларды
қалыптастыру және аттестация
басқармасының басшысы

«01» 06 2026 ж.г.