OpenEvolve-Driven Selection and Tuning of Ozaki-Scheme GPU Matrix Multiplication

Connor Vitz, Aaron Jarmusch, and Sunita Chandrasekaran
University of Delaware
Contact: vitzc@udel.edu

CRPL

UNIVERSITY OF DELAWARE
Department of Computer & Information Sciences

Can an optimizing system built around OpenEvolve [1] support the development, verification, and performance tuning of an Ozaki-Scheme GPU matrix multiplication implementation?

System Pipeline

1. INPUT
Shape · GPU · accuracy target · distribution · priority mode

2. Osaki Scheme SELECTOR
Decision tree and classifier picks a OS variant + bounds

3. OpenEvolve (OE) LOOP
Gemini Flash 2.5 [7]-driven search tunes params, GPU evaluates

4. OUTPUT
Reports optimal parameters; validated on a QE workload

Contributions

- Many GPUs are focusing on INT8/FP8 capabilities
- Ozaki-Scheme [2] [3] emulates FP64 via INT8/FP8
- Previous poster [4]: 6 variant system and initial benchmarks
- **This poster:**
  - Throughput-priority mode: +106.6% GFLOPS
  - Feasibility gate: 88.7% accuracy, filters wasted time
  - Quantum ESPRESSO [5] workload: improved with fix
  - New GPU: 20.6× throughput (GB10) vs 1.42× (GH200)

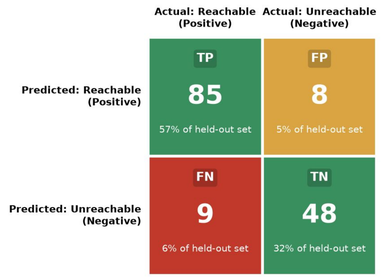
Throughput Mode

+106.6% GFLOPS
with --throughput-priority



- Same shape (4096^3), same code, one CLI flag flipped
- Normal: **2.25e-11** error
- Throughput mode: **3.02e-04** error
- Tradeoff of error for 2× the throughput

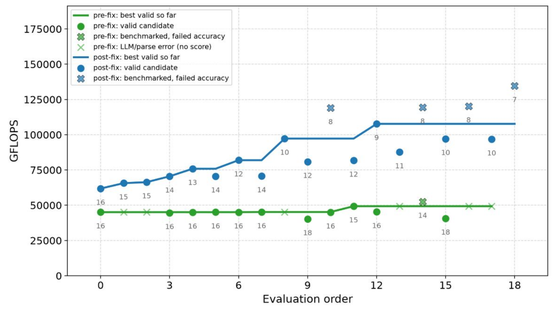
Feasibility Classifier



88.7% accuracy

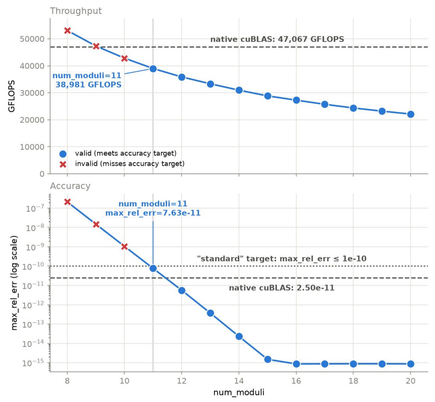
- Predict reachability before spending GPU time
- Model:
  - Monotonicity-constrained **XGBoost**
- Candidates:
  - 600-row dataset, 150 held out
  - Deliberately biased toward the decision boundary
- Results:
  - **Precision: 91.4%** – 95% CI [83.9%, 95.6%]
  - **Recall: 90.4%** – 95% CI [82.8%, 94.9%]

OE Convergence



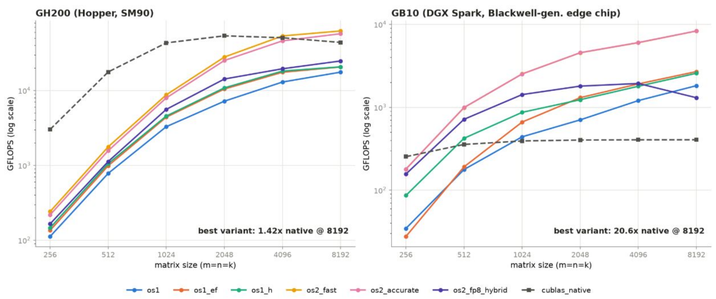
- Examination of OE Loop finding best candidate
- Only the per-island-dilution fix changed
- **Pre-fix:**
  - stuck repeating num_moduli=16
  - ~45K to ~50K GFLOPS best-valid
- **Post-fix:**
  - Explores number of module steadily downward
  - ~61K to ~108K GFLOPS best-valid

AUSURF Benchmark



- Quantum ESPRESSO AUSURF112
  - 50 GEMM shapes
  - Complex/ZGEMM
- Best candidate: **38,980.55 GFLOPS** and **7.63e-11** error
- Native cuBLAS: **47,067.14 GFLOPS** (1.21× faster)

Hardware Analysis – GH200 vs GB10



- Sweep methodology on both:
  - narrow distribution
  - 7 trials × 20 repeats
  - median
- **GH200 throughput:**
  - Best achieves **1.42×** native cuBLAS
  - Most achieve worse
- **GB10 throughput:**
  - Best achieves **20.6×** native cuBLAS
- **GB10's** native FP64 throughput is nearly flat
  - Tracks how FP64-starved the chip is
- **GB10** is DGX Spark's [6] compact/edge chip (Different from a full datacenter Blackwell part such as B100/B200/B300)

Conclusion

Our OpenEvolve-driven optimizing system can develop, verify, and tune Ozaki-Scheme GPU matrix multiplication implementations

- Uses same pipeline design as previous poster
  - Input, Selector, OE Loop, Output
- Adapts to different workloads and objectives
- Competitive with native cuBLAS
  - Observed improvement in exception (AUSURF112)
- Future work:
  - Further close throughput gap on AUSURF112
  - Analyze performance on other hardware
  - Deploy system for real scientific application

References

[1] Algorithmic Superintelligence, "OpenEvolve: Open-source implementation of AlphaEvolve," Software repository, 2026. [Online]. Available: https://github.com/algorithmicsuperintelligence/openevolve
[2] H. Ootomo, K. Ozaki, and R. Yokota, "DGEMM on integer matrix multiplication unit," 2023. [Online]. Available: https://arxiv.org/abs/2306.11975
[3] K. Ozaki, Y. Uchino, and T. Imamura, "Ozaki Scheme II: A GEMM-oriented emulation of floating-point matrix multiplication using an integer modular technique," 2025. [Online]. Available: https://arxiv.org/abs/2504.08009
[4] C. Vitz, A. Jarmusch, and S. Chandrasekaran, "Preliminary Evaluation of a Multi-Agent Optimization System for Ozaki Scheme GPU Matrix Multiplication," poster abstract, IEEE CLUSTER 2026, 2026. [Online]. Available: https://github.com/WhizZarD31415/IEEE26-Evaluating-Multi-Agent-for-OzakiScheme-Poster
[5] P. Giannozzi, S. Baroni, N. Bonini et al., "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials," Journal of Physics: Condensed Matter, vol. 21, no. 39, p. 395502, 2009. [Online]. Available: https://arxiv.org/abs/0906.2569
[6] NVIDIA Corporation, "NVIDIA DGX Spark: Personal AI Supercomputer Powered by Blackwell," 2025. [Online]. Available: https://www.nvidia.com/en-us/products/workstations/dgx-spark/
[7] Google DeepMind, "Gemini 2.5 Flash Model Card," 2025. [Online]. Available: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf

U.S. DEPARTMENT OF ENERGY

NSF

This work used and is supported by...

DeltaAI (NCSA) — supported by NSF award OAC 2320345, allocation CIS250803 via ACCESS, with support from NSF grants #2138259, #2138286, #2138307, #2137603, #2138296.

DOE Office of Science — supported under prime contract DE-AC05-000R22725 through the S4PST project, Next Generation Scientific Software Technologies program.