Maksim Levental

maksim.levental@gmail | makslevental.github.io | github.com/makslevental

INTERESTS

I'm interested in using, improving, and designing application-specific hardware accelerators for Deep Neural Networks (DNNs), for science applications. Specifically, using compiler techniques such as symbolic execution, and mathematical optimization techniques such as integer programming, to improve memory consumption and inference latencies of DNNs on GPUs, FPGAs, and ASICs.

EDUCATION

Ph.D., **Computer Science**, *University of Chicago*, GPA: 4.0 — **Oct. 2020 — Apr. 2024**
M.S., **Computer Science**, *University of Florida*, GPA: 3.7 — **Aug. 2014 — Dec. 2015**
B.S., **Pure Math**, **Honors**, *Florida State University*, GPA: 3.7 — **Aug. 2007 — Dec. 2010**

EXPERIENCE

Senior Member of Technical Staff, *AMD* — **May 2024 — Present**

- IREE implementation for AI Engines; wrote/owned the entire runtime (HAL and kernel module interface).
- Triton codegen for AMDGPU.

Compiler Research Intern, *AMD* — **June 2023 — May 2024**

- eDSL design for tiled, spatial, architecture with programmable data movement (AI Engine).
- ILP for loop tiling, memory allocation, and congestion-aware routing.
- MLIR development and upstream contribution.

Compiler Research Intern, Torch-MLIR, *nod.ai* — **Feb. 2022 — June 2022, Sept. 2022 — June 2023**

- Built an end-to-end compiler for PyTorch models (through Torch-MLIR) which functions independently of PyTorch (https://github.com/nod-ai/PI).
- Built out first implementation of eager-mode for Torch-MLIR using `torch_dispatch` and JITing techniques.
- Implemented various new "ops" and infrastructure (including, extending C bindings).

Distributed Systems Research Intern, PyTorch, *Facebook* — **June 2022 — Sept. 2022**

- Researched automated tensor sharding for distributed training of large models.
- Studied DP formulation of inter-node communication (`alpa`'s) overhead and contributed a ∼200x improvement in performance for large tensors.
- Used MLIR to explore statically inferring tensor sharding and compute paralellization strategies (based on shape analysis/refinement).

Compiler Research Intern, PyTorch, *Facebook* — **June 2021 — Feb. 2022**

- Developed functionality for statically allocating memory for reduced inference latency in production and OSS PyTorch models.
- Developed symbolic analysis techniques for inferring memory requirements of intermediate tensors in dynamic neural networks. Specifically, used and extended shape analysis in TorchScript, combined with MIP formulation of storage allocation, to derive upper bounds on sizes of intermediate tensors.

Ranking Engineer Intern, Groups, *Facebook* — **June 2020 — Sep. 2020**

- Investigated effects of repeated recommendations on conversion for Groups You Should Join (GYSJ). Specifically, developed and shipped pipelines and metrics for measuring repetition.
- Developed features used in scoring recommendations, including collecting, cleaning, and transforming data.
- Ran A/B tests on 100MM population to measure efficacy of aforementioned features in reducing repetition and increasing conversion. Effects were marginal (i.e., not statistically significant).

Graduate Research Assistant, Globus Labs, *University of Chicago* — **Oct. 2020 — Present**
Graduate Research Assistant, *University of Florida* — **Aug. 2014 — Dec. 2015, Aug. 2018 — Dec. 2019**
Peace Corps Education Volunteer, *Mbale, Uganda* — **March 2011 — March 2013**

SKILLS

Formal Languages	Python, C++, Rust, SQL, CUDA, Scala
Human Languages	English (native), Russian (native)
Platforms	LLVM, MLIR, PyTorch, TensorFlow, PostgreSQL
Skills	Deep Learning, Compilers, Hardware, Baking
Coursework	*Computer Science*: Quantum Computing, Deep Learning Systems, Analysis of Algorithms, Consensus and Economics, Automata *Math*: Statistics, Numerical Linear Algebra, Optimization, Real Analysis, Topology, Algebra, PDEs *Physics*: Computational Physics, Electricity and Magnetism, Quantum Mechanics, Statistical Mechanics, Nuclear Physics, Waves and Optics

PUBLICATIONS

Levental M., Khan A., Chard R., Chard K., Neuendorffer S., Foster I., *An End-to-End Programming Model for AI Engine Architectures*, Proceedings of the 14th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies.

Levental M., Khan A., Chard R., Yoshi K., Chard K., Foster I., *OpenHLS: High-Level Synthesis for Low-Latency Deep Neural Networks for Experimental Science*, Proceedings of the 14th International Symposium on Highly Efficient Accelerators and Reconfigurable Technologies.

Levental M., Kamatar A., Chard R., Chard K., Foster I., *nelli: a lightweight frontend for MLIR*, **In Pre-print**.

Levental M., Chard R., Chard K., Foster I., Wildenberg G., *Ultrafast Focus Detection for Automated Microscopy* eScience IEEE 17th International Conference on eScience (2021).

Huerta E., Khan A., Huang X., Tian M., **Levental M.**, Chard R., Wei W., Heflin M., Katz D., Kindratenko V., Mu D., Blaiszik B., Foster I., *Accelerated, Scalable and Reproducible AI-driven Gravitational Wave Detection.* Nature Astronomy volume 5, pages 1062-1068 (2021).

Yoshii K., Sankaran R., Strempfer S., **Levental M.**, Hammer M., Miceli A., *A Hardware Co-design Workflow for Scientific Instruments at the Edge.* Accepted to the Smoky Mountains Computational Sciences and Engineering Conference (SMC '21).

Levental M., Chard R., Libera J. A., Chard K., Koripelly A., Elias J., Schwarting M., Blaiszik B., Stan M., Chaudhuri S., Foster I., *Towards Online Steering of Flame Spray Pyrolysis Nanoparticle Synthesis.* **Best Paper at 2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP '20)**.

Wilson J., Toska F., **Levental M.**, Dobbins P., *A Deep Neural Network Model for Hazard Classification.* Artificial Intelligence and Machine Learning in Defense Applications (2019).

TECHNICAL REPORTS

Levental M., *PhD Thesis: An End-to-End Programming Model for AI Engine Architectures* uchicago.edu (2024).

Levental M., *MS Thesis: Memory Planning for Deep Neural Networks* arXiv:4178464 (2022).

Levental M., *Tensor Networks for Simulating Quantum Circuits on FPGAs.* arXiv:2108.06831 (2021).

Levental M., Orlova E. *Comparing the Costs of Abstraction for DL Frameworks.* arXiv:2012.07163 (2020).