Tuning the tuner: the art of automatic performance optimization

Willemsen, F.Q.

Summary

Modern scientific discoveries and technologies increasingly rely on powerful computers. From predicting the climate to training artificial intelligence, researchers depend on High-Performance Computing (HPC) systems that combine traditional processors with specialized accelerators such as graphics cards (GPUs). These machines are extremely fast, but getting the most out of them is surprisingly difficult. Small choices in how a program is written, such as how data is stored, how computation is organized, or how many threads run in parallel, can make performance vary wildly. Because the number of possible choices is enormous, automatic performance tuning (auto-tuning) is needed: letting the computer automatically test different program configurations until it finds a fast and efficient one. Auto-tuning has proven highly effective, but it also introduces a new problem: how do we best design the tuning process itself? Choosing how to search, which parameters to test, and how to handle constraints can make the difference between success and failure.

This thesis focuses on improving auto-tuning for GPUs, which are widely used in scientific and industrial computing. The research advances the field in six directions:

- Developing a new optimization method based on Bayesian statistics, designed specifically for GPU tuning, which outperforms existing approaches.

- Creating a fair way to compare tuning methods, ensuring that researchers can reliably judge which approach works best and improve upon it.

- Building a fast solver for search spaces, capable of handling billions of possible configurations in less than a second, removing a major bottleneck.

- Showing how tuning the tuners themselves can lead to performance improvements of more than 200%, while reducing energy use through large-scale simulations.

- Making evolutionary algorithms more constraint-aware, so they can better navigate the complex rules that govern GPU programs.

- Exploring whether artificial intelligence can invent new optimizers, showing that machine-generated algorithms can beat human-designed ones.

Together, these contributions show that, as big improvements can come from auto-tuning programs, improving the underlying tuning systems themselves can enable the next generation of technological breakthroughs. All the methods and tools from this thesis have been released as open-source software, enabling researchers and practitioners worldwide to benefit from them.

Looking ahead, the results point to a future where auto-tuning becomes smarter, more efficient, and more adaptive. Optimizers that understand constraints, balance performance with energy use, or even design themselves with the help of AI could make HPC more accessible and sustainable. In this way, this work contributes to enabling faster progress in science and technology, powered by computers that tune themselves to the challenges of tomorrow.