How to Write Fast Code

18-645, spring 2008
7th Lecture, Feb. 6th

Instructor: Markus Püschel

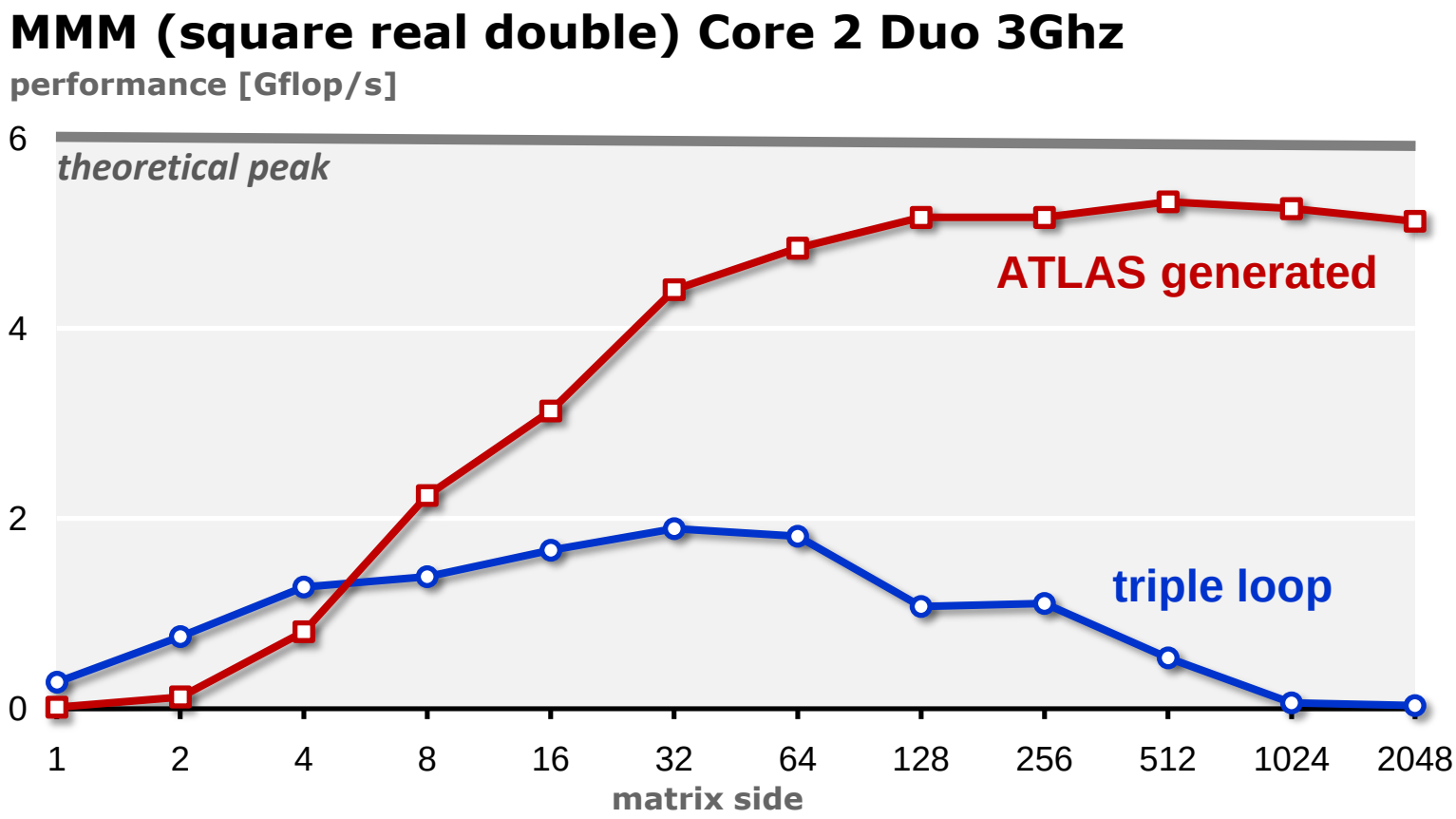
TAs: Srinivas Chellappa (Vas) and Frédéric de Mesmay (Fred)

Today

- **ATLAS: MMM program generator**

- **ATLAS homepage**
- C. Whaley, A. Petitet, J. Dongarra, **Automated empirical optimizations of software and the ATLAS project**, Parallel Computing 27(2), pp. 3–35, 2001. Link.
- Our presentation follows: K. Yotov, X. Li, G. Ren, M. Garzaran, D. Padua, K. Pingali, P. Stodghill, **Is Search Really Necessary to Generate High-Performance BLAS?,** Proceedings of the IEEE, 93(2), pp. 358–386, 2005. Link.

MMM: Memory Hierarchy Optimization

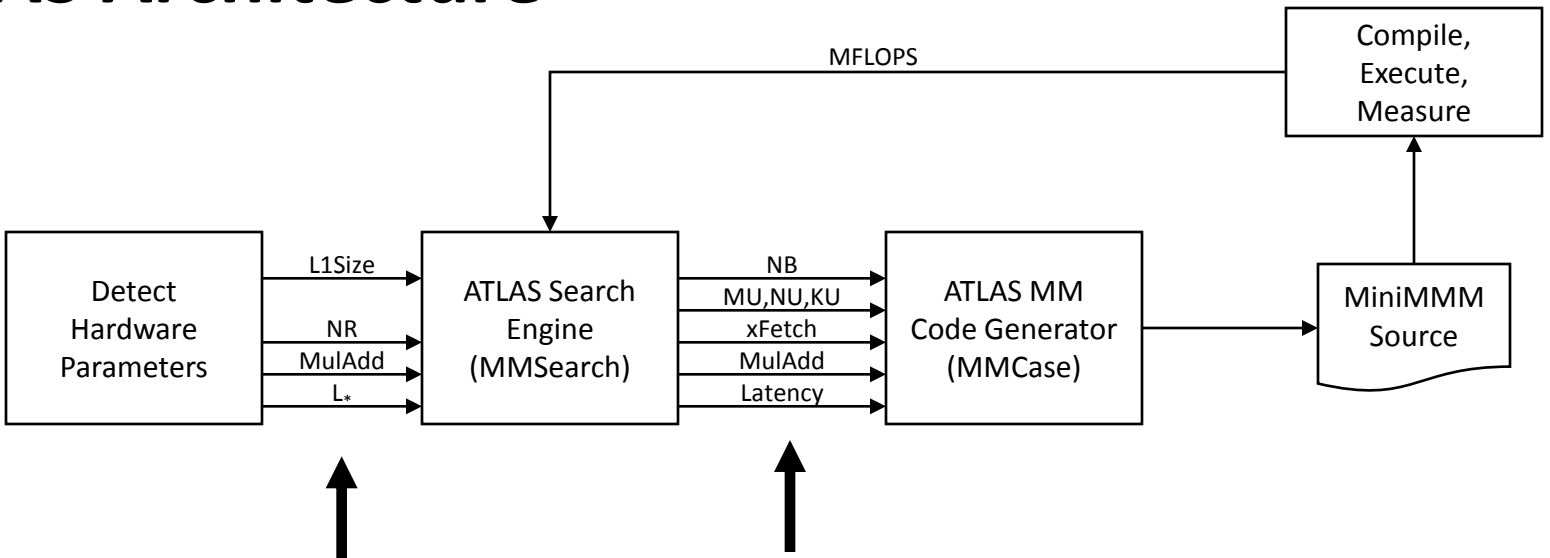


- Intel compiler icc –O2
- Huge performance difference for large sizes
- Great case study to learn memory hierarchy optimization

ATLAS

- Successor of PhiPAC, BLAS program generator (web)
- People can also contribute handwritten code
- The generator uses empirical search over implementation alternatives to find the fastest implementation
  no vectorization or parallelization
- We focus on BLAS3 MMM
- Search only over $2n^3$ algorithms
  (cost equal to triple loop)

ATLAS Architecture



Search parameters:

- **span search space**
- **specify code**
- **found by orthogonal line search**

Hardware parameters:

- **L1Size: size of L1 data cache**
- **NR: number of registers**
- **MulAdd: fused multiply-add available?**
- **L_* : latency of FP multiplication**

source: Pingali, Yotov, Cornell U.

How ATLAS Works

- Blackboard

Search in ATLAS

- **It is all about finding the highest performance mini-MMM**
- **Orthogonal line search:**
  - Choose parameter p to search over, fix all other to reasonable values of the values that have been found already
  - Search for best p within bounds
  - Repeat until done
  - ATLAS starts with mini-MMM block size NB
- **Other searches would be possible**

- **Example: generated mini-MMM**