

How to Write Fast Numerical Code

Spring 2011

Lecture 10

Instructor: Markus Püschel

TA: Georg Ofenbeck

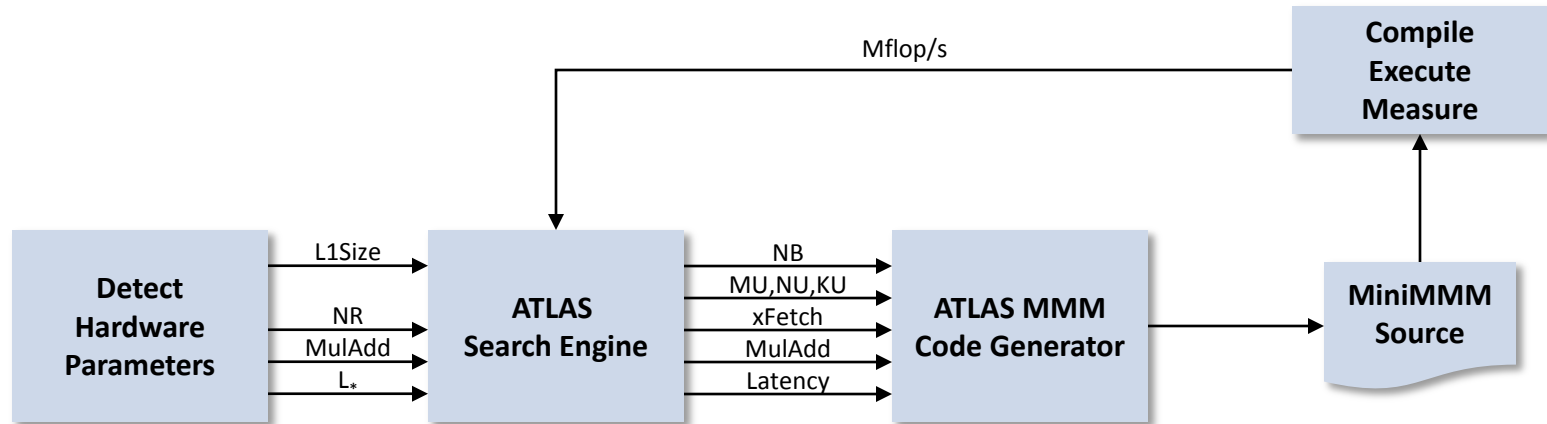


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Today

- **ATLAS: Principles**
- **Model-based ATLAS**
- 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](#).

Last Time: ATLAS



- Blocks MMM into mini-MMMs
- Searches for fastest (highest-performance) mini-MMM
- Mini-MMM choices encoded by parameters (N_B , M_U , N_U , ...)
- Parameter space bounded through microarchitecture parameters for example: $N_B \leq \text{sqrt}(\text{cache size})$

How it Worked: From Triple Loop to ...

// MMM loop-nest

```
for i = 0:NB:N-1
  for j = 0:NB:M-1
    for k = 0:NB:K-1
```

- *ij or ji depending on N and M*
- *Blocking for cache*

// mini-MMM loop nest

```
for i' = i:MU:i+NB-1
  for j' = j:NU:j+NB-1
    for k' = k:KU:k+NB-1
```

- *Blocking for registers*

// micro-MMM loop nest

```
for k'' = k':1:k'+KU-1
  for i'' = i':1:i'+MU-1
    for j'' = j':1:j'+NU-1
```

- *unrolling*
- *scalar replacement*
- *add/mult interleaving*
- *skewing*

Search parameters: N_B , M_U , N_U , K_U , L_S , ...

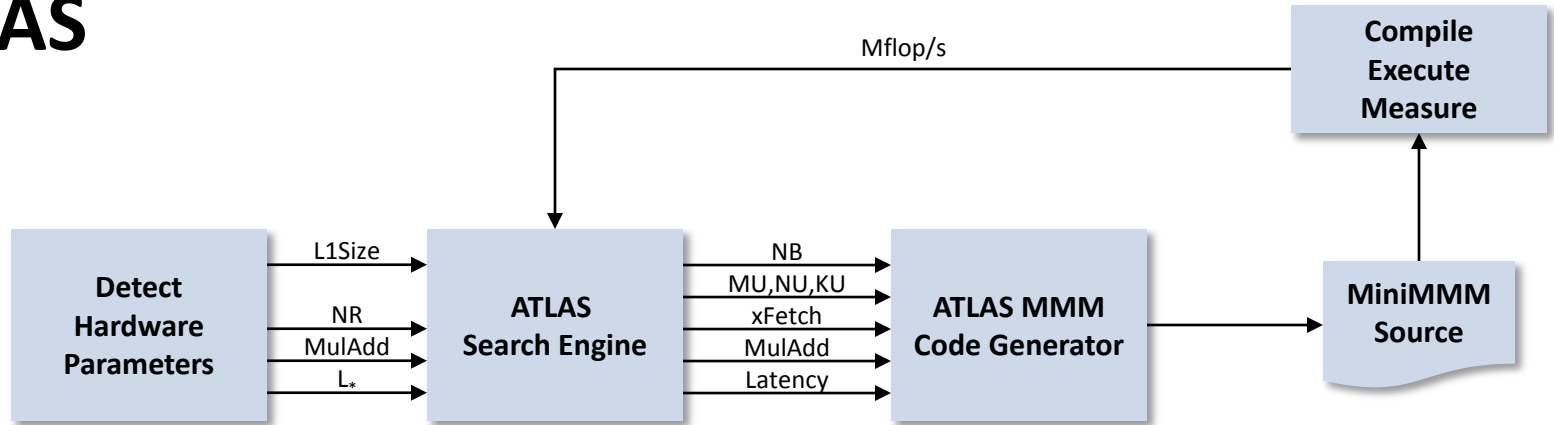
Principles used in ATLAS Optimization

- **Optimization for memory hierarchy = increasing locality**
 - Blocking for cache, blocking for registers
 - Done by loop tiling and loop exchange
- **Fast basic blocks for small sizes (micro-MMM):**
 - Loop order chosen for ILP
 - Loop unrolling to enable scalar replacement (enables register allocation and instruction scheduling)
 - Add/mult interleaving and skewing (improving ILP)
- **Search for the fastest (mini-MMM) over a relevant set of algorithm/implementation alternatives**

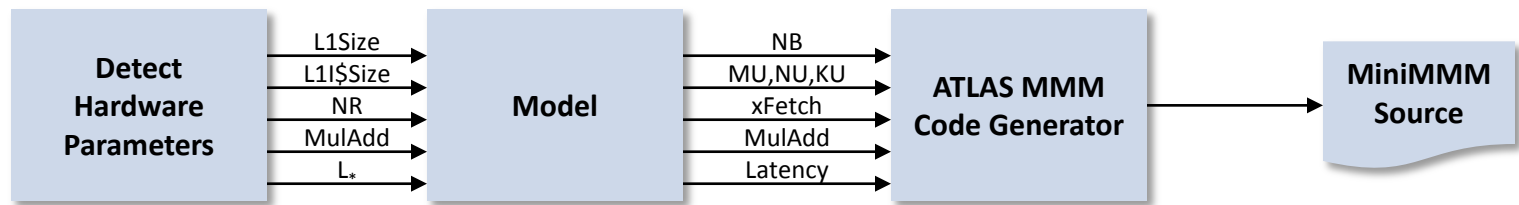
MMM: So far

- We learned a set of optimization techniques for the memory hierarchy
- But there are degrees of freedom
- *Practical problem:* How to choose them without using search?
- *Scientific problem:* How to choose them from an understanding of the microarchitecture?

ATLAS



Model-Based ATLAS



- Search for parameters replaced by model to compute them
- More hardware parameters needed

Model-Based ATLAS: Details

- Blackboard

Experiments

- **Unleashed:** Not generated = hand-written contributed code
- **Refined model** for computing register tiles on x86
- **Blocking** is for L1 cache
- **Result:** Model-based is comparable to search-based (except Itanium)

