Algorithms and Computation in Signal Processing

special topic course 18-799B
spring 2005
8th Lecture Feb. 3, 2005

Instructor: Markus Pueschel

TA: Srinivas Chellappa

ATLAS MMM Code Generation (cont'd)

Last Time: From Triple Loop to …

```
// MMM loop-nest
for i=0:N_B:N-1
 for j=0:N_B:M-1
  for k=0:N_B:K-1
   // mini-MMM loop nest
   for i'=i:M_U:i+N_B-1
    for j'=j:N_U:j+N_B-1
     for k'=k:K_U:k+N_B-1
      // micro-MMM loop nest
      for k''=k':1:k'+K_U-1
       for i''=i':1:i'+M_U-1
        for j''=j':1:j'+N_U-1
```

ij or ji depending on N and M

Blocking for cache

Blocking for registers

unrolling

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

Search parameters: N_B, M_U, N_U, K_U, L

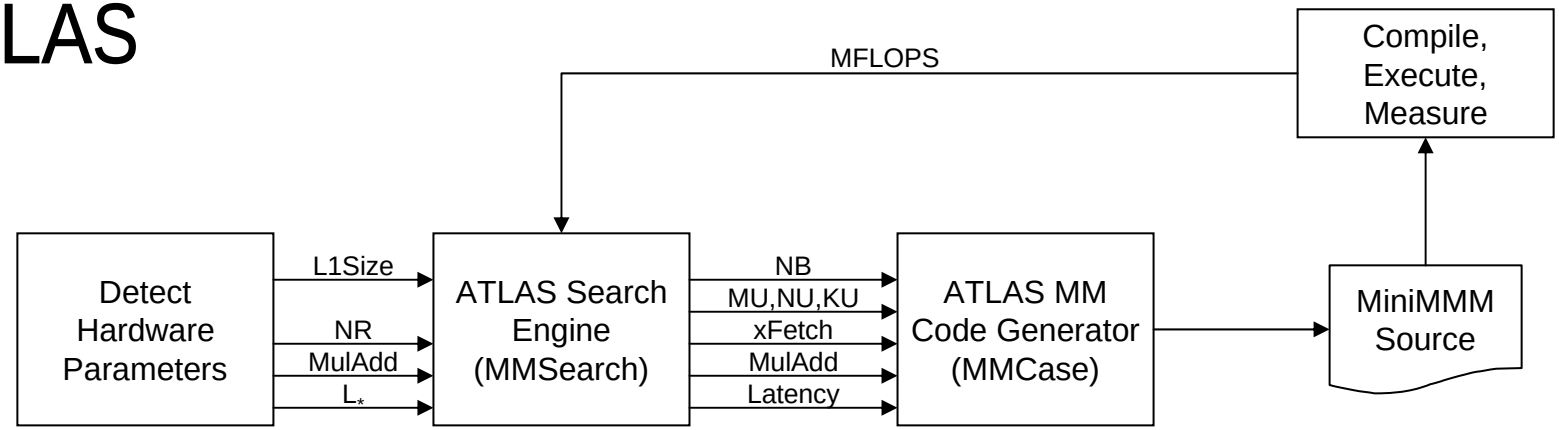
Principles in ATLAS Code Generation

- Optimization for memory hierarchy = increasing locality (Blocking for cache, blocking for registers)
- Fast basic blocks for small sizes (micro-MMM):
  - Loop unrolling (reduce loop overhead)
  - Scalar replacement (enables better compiler optimization)
  - Add/mult interleaving (better throughput)
  - Skewing (better instruction level parallelism)
- Search for the fastest over a relevant set of algorithm/implementation alternatives

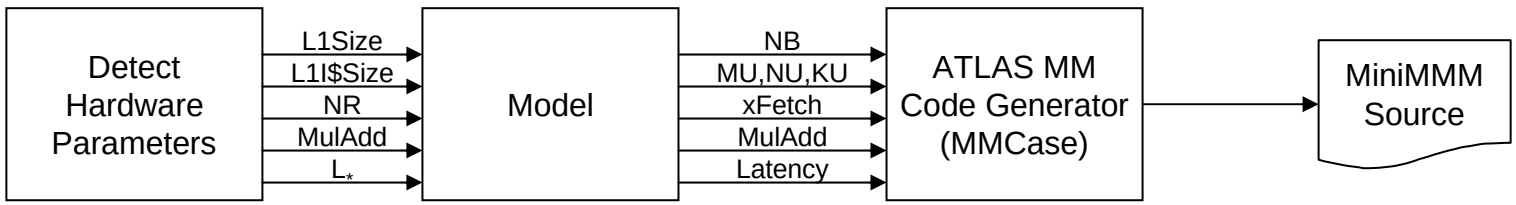
Model-Based ATLAS

- Paper: “Is Search Really Necessary to Generate High-Performance BLAS?,” Kamen Yotov et al.
- Goal: Instead of searching, find MMM parameters through model

ATLAS



Model-Based ATLAS



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

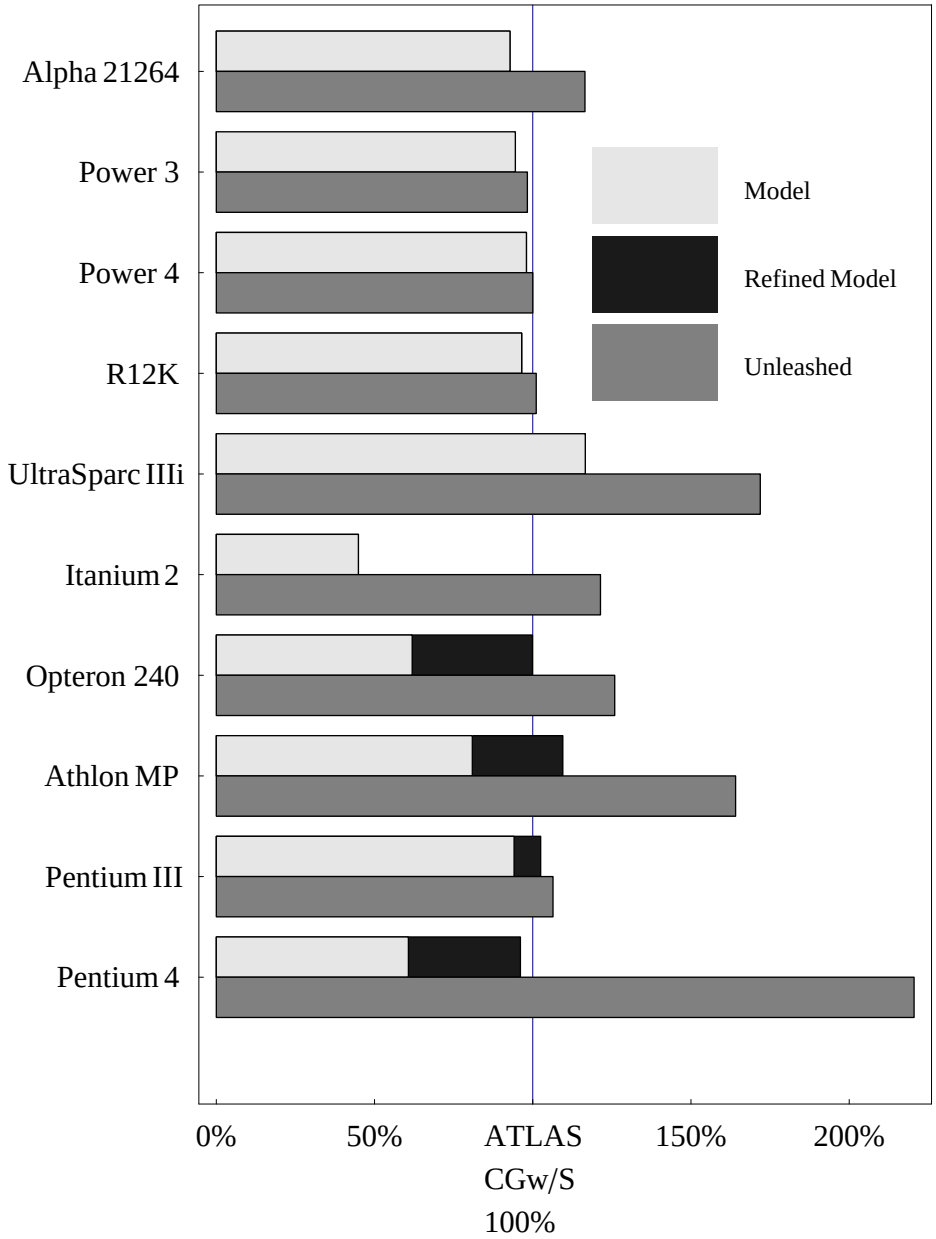
source: Pingali, Yotov, Cornell U.

Model-Based ATLAS: Details

- Blackboard

ATLAS: Experiments

- Hand-written code often substantially faster (e.g., vector instructions)
- Model-based comparable to search-based (except Itanium)



graph: Pingali, Yotov, Cornell U.