Automatic Tuning of Scientific Applications

Apan Qasem  Ken Kennedy
Rice University
Houston, TX
Recap from Last Year

• A framework for automatic tuning of applications
  - Fine grain control of transformations
  - Feedback beyond whole program execution time
  - Parameterized Search Engine
  - Target: Whole Applications
  - Search Space
    • Multi-loop transformations: e.g. Loop Fusion
    • Numerical Parameters
Recap from Last Year

- Experiments with Direct Search
  - Direct Search able to find suitable tile sizes and unroll factors by exploring only a small fraction of the search space
    - 95% of best performance obtained by exploring 5% of the search space
  - Search space pruning needed for making search more efficient
    - Wandered into regions containing mostly bad values
    - Direct search required more than 30 program evaluations in many cases
Today’s talk

Search Space Pruning
Search Space Pruning

• Key Idea:

  Search for architecture-dependent model parameters rather than transformation parameters

• Fundamentally different way of looking at the optimization search space

• Implemented for loop fusion and tiling
  [Qasem and Kennedy, ICS06]
Architectural Parameters

- Register Set
- L1 Cache

Effective Register Set

Effective Cache Capacity

Cost Model

Tile Size

Fusion Config.

Search Space

\( T_{01} \ldots \ldots \ldots T_{0N} \)

\( T_{11} \ldots \ldots \ldots T_{1N} \)

\( T_{21} \ldots \ldots \ldots T_{2N} \)

\( T_{31} \ldots \ldots \ldots T_{3N} \)

\( F_0 \ldots \ldots \ldots F_{2^L} \)

Gray Code Representation

\((N - \frac{pL}{N} \times 2^L)\) dimensions

\((L + 1)\) dimensions

Estimates of architectural parameters

New Search Space has only two dimensions!

Effective Register Set

Effective Cache Capacity
Our Approach

• Build a *combined cost model* for fusion and tiling to capture the interaction between the two transformations
  - Use reuse analysis to estimate trade-offs

• Expose *architecture–dependent parameters* within the model for tuning through empirical search
  - Pick T such that
    • Working Set < Effective Cache Capacity
  - Search for suitable Effective Cache Capacity
Tuning Parameters

- Use a *tolerance term* to determine how much of a resource we can use at each tuning step

\[
\text{Effective Register Set} = \left\lfloor T \times \text{Register Set Size} \right\rfloor \\
[0 < T \leq 1]
\]

\[
\text{Effective Cache Capacity} = E(a, s, T) \\
[0.01 \leq T \leq 0.20]
\]
Search Strategy

- Start off conservatively with a low tolerance value and increase tolerance at each step
- Each tuning parameter constitutes a single search dimension
- Search is **sequential** and **orthogonal**
  - stop when performance starts to worsen
  - use reference values for other dimensions when searching a particular dimension
Benefits of Pruning Strategy

• Reduce the size of the exploration search space
  – Single parameter captures the effects of multiple transformations
    • Effective cache capacity for fusion and tiling choices
    • Register pressure for fusion and loop unrolling
  – Search Space does not grow with program size
    • One parameter for all tiled loops in the application

• Correct for inaccuracies in model
Performance Across Architectures

Mean Speedup over baseline

Platforms

MIPS  Itanium  Alpha  PowerPC  Pentium III  Geo Mean

model-based  native
Performance Comparison with Direct Search

- advect3d
- erle
- liv18
- mgrid

- model-based
- direct
Tuning Time Comparison with Direct Search

- **advecc3d**
- **erle**
- **liv18**
- **mgrid**

Legend:
- **model-based**
- **direct**
Conclusions and Future Work

- Approach of tuning for architectural parameters can significantly reduce the optimization search space, while incurring only a small performance penalty.

- Extend pruning strategy to cover more transformations:
  - Unroll-and-jam
  - Array Padding

- Extend pruning strategy to cover more architectural parameters:
  - TLB
Questions
Extra Slides Begin Here
Performance Improvement Comparison

![Graph showing performance improvement comparison. The x-axis represents the number of iterations, and the y-axis represents the fraction of best speedup. The graph compares three methods: Exhaustive Search (red), Direct Search (yellow), and Random Search (green).](image-url)
Tuning Time Comparison

The diagram compares the fraction of tuning time across various benchmarks. The benchmarks include `advevt3d`, `lud`, `mm`, `vpenta`, `mgrid`, and `swim`. The lines represent different data sets: DS 30, DS 60, DS 90, and DS 120. Each line shows the performance trend across the benchmarks for the respective data set.
Framework Overview

User

Source

LoopTool

Architectural Specs

Native Compiler

Performance Measurement Tools

Transformed Source

Binary

Search Space

Parameterized Search Engine

Feedback

Next Iteration Parameters

Tuning Level

AutoTuning Workshop 06

Rice University
Why Direct Search?

- Search decision based solely on function evaluations
  - No modeling of the search space required
- Provides approximate solutions at each stage of the calculation
  - Can stop the search at any point when constrained by tuning time
- Flexible
  - Can tune step sizes in different dimensions
- Parallelizable
- Relatively easy to implement
\( L_A: \) do \( j = 1, N \) 
\hspace{1cm} do \( i = 1, M \) 
\hspace{2cm} \( b(i,j) = a(i,j) + a(i,j-1) \) 
\hspace{1cm} enddo 
\hspace{1cm} enddo 

\( L_B: \) do \( j = 1, N \) 
\hspace{1cm} do \( i = 1, M \) 
\hspace{2cm} \( c(i,j) = b(i,j) + d(j) \) 
\hspace{1cm} enddo 
\hspace{1cm} enddo 

(a) code before transformations
\[ L_{AB} : \]
\[
\begin{align*}
do & \ j = 1, \ N \\
do & \ i = 1, \ M \\
& \quad b(i,j) = a(i,j) + a(i,j-1) \\
& \quad c(i,j) = b(i,j) + d(j) \\
\end{align*}
\]

- lost reuse of \(a()\)
- saved loads of \(b()\)

*increased potential for conflict misses*

(b) code after two-level fusion
\[ \begin{align*}
\text{do } & i = 1, \ M, \ T \\
\text{do } & j = 1, \ N \\
\text{do } & \text{ii} = i, \ i + T - 1 \\
& b(\text{ii}, j) = a(\text{ii}, j) + a(\text{ii}, j - 1) \\
& c(\text{ii}, j) = b(\text{ii}, j) + d(j) \\
\text{enddo} \\
\text{enddo} \\
\text{enddo}
\end{align*} \]

How do we pick T?

Not too difficult if caches are fully associative

Can use models to estimate effective cache size for set-associated caches

Model unlikely to be totally accurate
  - Need a way to correct for inaccuracies
Register Set

L1 Cache

Cost Models

Tile Size

Fusion Config.

\( T_0, \ldots, T_0^N \)

\( T_1, \ldots, T_1^N \)

\( T_2, \ldots, T_2^N \)

\( T_3, \ldots, T_3^N \)

\( F_0, \ldots, F_2^L \)

\((N-p) \times (2^L-q)\) points

\((L+1)\) dimensions
Effective Register Set

Effective Cache Capacity

Cost Model

Tile Size

Fusion Config.

Estimates of machine parameters