Autotuning for GPUs using Orio

Azamat Mametjanov, Daniel Lowell, Ching-Chen Ma, Boyana Norris

Mathematics and Computer Science Division
Argonne National Laboratory
Motivation

- High-throughput architectures provide new programming abstractions
  - Language extensions
  - Library API

- Exploiting new capabilities is obscure
  - Learning curve
  - Copy-paste boilerplate code
  - Low-level thread and data handling

- Can we automate this?
  - Reference implementation
  - Performance hints
Orio autotuning framework

Annotated Code → Annotations Parser → Sequence of (Nested) Annotated Regions

Transformed Code → Code Generator → Code Transformations

Empirical Performance Evaluation → Search Engine

Search Engine → Optimized Code

Tuning Specification

best performing version
for (i=0; i<n; i++)
    y[i] = a*x[i] + b*y[i];
Annotations for transformations

/*@ begin Loop(...

for (i=0; i<n; i++)
    y[i] = a*x[i] + b*y[i];

) @*/

for (i=0; i<n; i++)
    y[i] = a*x[i] + b*y[i];

/*@ end @*/
Annotations for transformations

/*@ begin Loop(transform CUDA(
    threadCount=TC,
    blockCount=BC,
    streamCount=SC, ...
  )
  for (i=0; i<n; i++)
    y[i] = a*x[i] + b*y[i];
) @*/

...
Annotations for autotuning

/*@ begin PerfTuning(
def performance_params{
  param TC[] = range(32,1025,32);
  param BC[] = range(14,113,14);
  param SC[] = range(1,17); ...
}
) @*/

/*@ begin Loop(transform CUDA(
...
/*@ end @*/
 Annotations for autotuning

```c
/*@ begin PerfTuning(
    def input_params {
        param N[] = [1000,10000,100000];
    }
    def input_vars {
        decl double a = random;
        decl double b = random;
        decl static double x[N] = random;
        decl static double y[N] = random;
    }
    ...
) */
```
Annotations for autotuning

/*@ begin PerfTuning(
  def build {
    arg build_command = 'nvcc -arch=sm_20 @CFLAGS';
  }
  def performance_counter {
    arg repetitions = 10;
  }
  ...
) @*/
Time for a demo
Example: Sparse matrix-vector product (7-point stencil) on a GPU

Intel Xeon (dual quad-core E5462 processors), 2.8GHz; GPU: NVIDIA Fermi C2070

Lower is better

Application: Bratu solid fuel ignition problem

Thank you

http://tinyurl.com/OrioTool