Exploiting GPU Hardware Saturation for Fast Compiler Optimization

Alberto Magni, Christophe Dubach, Michael O'Boyle
Introduction

- Wide adoption of GPGPU with OpenCL
- Many application and devices
- Fast tuning of applications is critical in heterogeneous environment
Motivation

- Iterative compilation is a way to get good performance
Motivation

Kernels

.FL

Fermi
Kepler
Tahiti

Large Input Size
Motivation

Quick Tuning with Small Input Size
Motivation

How does the result of **training on small** perform on **large**?
What's next

- Motivation
- Thread Coarsening
- Performance Scalability
- Methodology
- Results
- Conclusion
Thread Coarsening

Original Thread Space

Transformed Thread Space

Reduce thread number

Increase amount of work per Thread
Parameter Space

Static Parameters

- Coarsening
  - Factor
  - Stride
  - Direction

Dynamic Parameters

- Local Work Group Size

~300 configs for One-D benchmarks
~2,000 configs for Two-D benchmarks
What's next

- Motivation
- Thread Coarsening
- Performance Scalability
- Methodology
- Results
- Conclusion
Tuning Across Input Sizes

FloydWashall on Tahiti

Reference Size

Performance

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Scalability

FloydWashall on Tahiti

Execution time [ms]

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Scalability - Metric

Function of **Global Work Size** and **Problem Complexity**

Throughput = \[
\frac{\text{Units of Work}}{\text{Execution Time}}
\]
Scalability

FloydWashall on Tahiti

Throughput [GUnits/s]

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Scalability

FloydWashall on Tahiti

Throughput [GU/tns / s]

Very Low Performance

Global Work Size
Scalability

FloydWashall on Tahiti

Perfect Scalability

Throughput [GUnits / s]

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Scalability

Saturation is very common

binarySearch

blacksholes

convolution

dwtHaar1D

fastWalsh

Global Work Size  Global Work Size  Global Work Size
Scalability – Performance-Throughput Correlation

FloydWashall on Tahiti

**Performance**

**Throughput**

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Scalability – Performance-Throughput Correlation

FloydWashall on Tahiti

Minimum Saturation Point

Performance

Throughput

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
What's next

- Motivation
- Thread Coarsening
- Performance Scalability
- Methodology
- Results
- Conclusion
Methodology

1) Plot Throughput

Throughput [GUnits / s]

Global Work Size

1K  4K  16K  65K  262K  1M  4M  16M  67M
Methodology

2) Find Max

Throughput [GUnits / s]

Global Work Size

Max
Methodology

3) Find input sizes within 10% of max

Throughput [GUnits / s]
Methodology

4) Select smallest input size

Throughput [GUnits/s]

Global Work Size

MSP-Tuning
What's next

- Motivation
- Thread Coarsening
- Performance Scalability
- Methodology
- Results
- Conclusion
Experimental Set-Up

• 16 benchmarks from Nvidia / AMD / Parboil
• 3 Devices:
  – Nvidia Fermi – GTX480
  – Nvidia Kepler – K20
  – AMD Tahiti – 7970

~ 160,000 runs in Total
Results – Fermi

Max Speedup
MSP-Tuning Speedup

Speedup

binarySearch
blackholes
convolution
dwtHaar1D
fastWalsh
floydWarshall
mriQ
ml
mlLocal
mvCoal
mvUncoal
nbody
reduce
sgemm
sobel
stencil
goMean
Results – Fermi

Speedup

Max Speedup
MSP-Tuning Speedup
Search Time Speedup

Search time speedup

binarySearch  blackholes  convolution  dwtHaar1D  fastWalsh  floydWarshall  mriQ  mt  mtLocal  mvCoal  mvUncoal  nbody  reduce  sgemm  sobel  stencil  geoMean
Results – Threshold impact

Throughput [GUnits / s]

Global Work Size

1K 4K 16K 65K 262K 1M 4M 16M 67M
Results – Threshold impact

Average Results on Fermi

Kernel Speedup

Search Speedup

Search Speedup Faster than Kernel Slowdown
Results - Outliers

- Max Speedup
- MSP-Tuning Speedup
- Search Time Speedup

Speedup vs Search time speedup for various benchmarks.
matrixVector Analysis – Fermi

Throughput vs. Global Work Size

Saturation cannot be reached

Too few threads
Too much data
sgemm Analysis - Nvidia

“Good looking” curves

Fermi

Kepler
sgemm Analysis - Tahiti

Throughput

Global Work Size

Major performance drop
sgemm Analysis - Tahiti

Throughput

Global Work Size

MemUnitBusy [%]
What's next

- Motivation
- Thread Coarsening
- Performance Scalability
- Methodology
- Experiments
- Results
- Conclusion
Conclusion

- GPUs show very good scalability for large input sizes
- Stable performance across input sizes for compiler options
Future Work

• Quick identification of the Minimum Saturation Point

• Application of MSP-Tuning to other types of searches