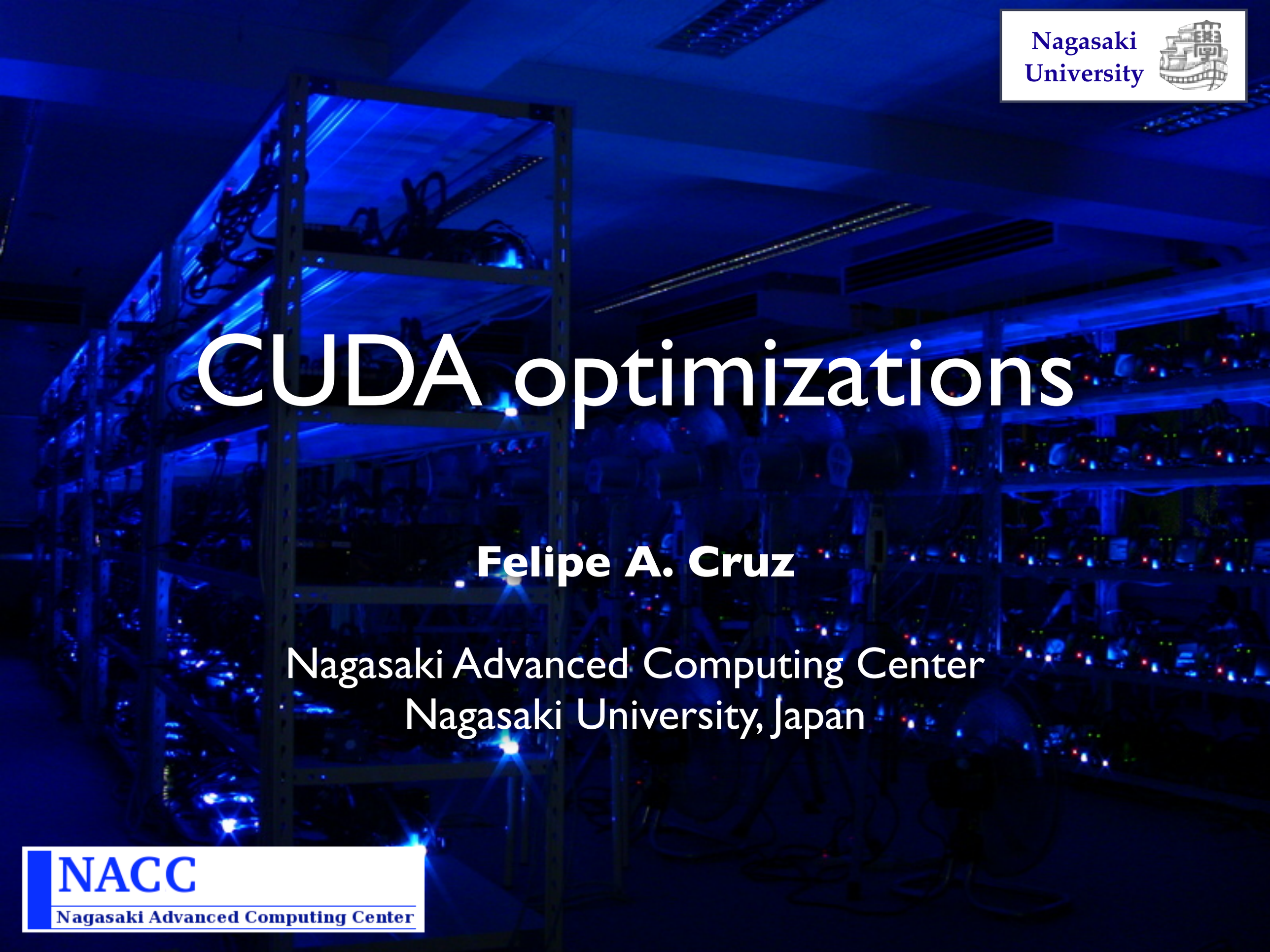

Nagasaki University

CUDA optimizations

Felipe A. Cruz

Nagasaki Advanced Computing Center
Nagasaki University, Japan

NACC
Nagasaki Advanced Computing Center

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 Tesla C2070	3GB GDDR5 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

448*

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 Tesla C2070	3GB GDDR5 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

448*1.15

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 / Tesla C2070	3GB GDDR5 / 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 / Tesla C2070	3GB GDDR5 / 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

Pearf = 1.03Tflops

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 / Tesla C2070	3GB GDDR5 / 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

Bandwidth = 144 GB/s

Form Factor	9.75" PCIe x16 form factor
# of Tesla GPUs	1
# of CUDA Core	448
Frequency of CUDA Cores	1.15 GHz
Double Precision floating point performance (peak)	515 Gflops
Single Precision floating point performance (peak)	1.03 Tflops
Total Dedicated Memory* Tesla C2050 / Tesla C2070	3GB GDDR5 / 6GB GDDR5
Memory Speed	1.5 GHz
Memory Interface	384-bit
Memory Bandwidth	144 GB/sec

work/data = ~ 29 (peak)

Target metrics

Throughput (measured in FLOP/s):

Average number of floating point operations per second than can be executed on the GPU.

Bandwidth (measured in GigaBytes/s):

Rate at which data is transferred between memory and the processor per second. All read and write memory transactions must be considered.

Design notions

Computational intensity:

Ratio of floating point operations to memory accesses.

Concurrency:

Sections of the algorithm that can be executed concurrently.
Can be organized into levels: fine to coarse grained concurrency.

Homogeneity of calculations:

Degree at which concurrent computations are the same, input independent.

Data-locality:

The way in which physically stored data is accessed by the algorithm.
Spatial data locality: data is physically adjacent.
Temporal data locality: data is temporally adjacent.

Implementation discussion

Thread execution branching:

Warp branching has direct impact on thread performance.

Multithreading:

You must be able to load enough threads for hiding memory latency.
GPU occupancy can tell you the number of active threads.

Memory management:

Small and fast shared memory and registers.
Large and slow global memory.
Avoid shared memory conflicts.
Efficient global memory access (coalesced).
Memory camping.
Collaborative memory transactions (balance across threads).

Loop unrolling.

Implementation discussion

Thread execution branching:

Warp branching has direct impact on thread performance.

Multithreading:

You must be able to load enough threads for hiding memory latency.
GPU occupancy can tell you the number of active threads.

Memory management:

Small and fast shared memory and registers.
Large and slow global memory.
Avoid shared memory conflicts.
Efficient global memory access (coalesced).
Memory camping.
Collaborative memory transactions (balance across threads).

Loop unrolling.

More tips

Keep the kernel complexity low.

Use many threads per block.

Interleave computations and memory transfers.

On-the-fly calculations.