

# CUDA optimizations

## Felipe A. Cruz

Nagasaki Advanced Computing Center Nagasaki University, Japan





Banda Ital Banda Ital	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Process delivers supercomputing power at 1/20th the p consumption and 1/10th the cost, bringing the performance of a small cluster to the desktop.            • Buy from participating partners.	or iower		
▶ Overview	Specifications	+	Drivers & Downloads	▶ Support
Form Factor			9.75" PCIe x16 form	n factor
# of Tesla GPUs			1	
# of CUDA Core			448	
Frequency of CUDA	Cores		1.15 GHz	
Double Precision flo	ating point performance (peak)		515 Gflops	
Single Precision floa	ating point performance (peak)		1.03 Tflops	
Total Dedicated Me	mory* Tesla Tesla	C2050 C2070	3GB GDDR5 6GB GDDR5	
Memory Speed			1.5 GHz	
Memory Interface			384-bit	
Memory Bandwidth			144 GB/sec	



and total	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Processor delivers supercomputing power at 1/20th the power consumption and 1/10th the cost, bringing the performance of a small cluster to the desktop.	44	448*		
► Overview	Specifications	Drivers & Downloads	▶ Support		
Form Factor		9.75" PCle x16 for	m factor		
# of Tesla GPUs	# of Tesla GPUs		1		
# of CUDA Core	# of CUDA Core		448		
Frequency of CUDA	Cores	1.15 GHz			
Double Precision flo	ating point performance (peak)	515 Gflops			
Single Precision floa	ting point performance (peak)	1.03 Tflops	5		
Total Dedicated Mer	Total Dedicated Memory* Tesla C20 Tesla C20		3GB GDDR5 6GB GDDR5		
Memory Speed		1.5 GHz			
Memory Interface	Memory Interface		384-bit		
Memory Bandwidth		144 GB/sec	:		



oner total	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Processor delivers supercomputing power at 1/20th the power consumption and 1/10th the cost, bringing the performance of a small cluster to the desktop.	r	448*	<b>'1.15</b>
▶ Overview	Specifications	⊧ D	rivers & Downloads	▶ Support
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 flo	ating point performance (peak)		515 Gflops	
Single Precision floa	ating point performance (peak)		1.03 Tflops	
Total Dedicated Me	mory* Tesla C Tesla C	2050 2070	3GB GDDR5 6GB GDDR5	
Memory Speed			1.5 GHz	
Memory Interface			384-bit	
Memory Bandwidth			144 GB/sec	



Byron teat Byron teat	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Processor delivers supercomputing power at 1/20th the power consumption and 1/10th the cost, bringing the performance of a small cluster to the desktop.         Image: Processor       Buy from participating partners.	Pearf = 1	.03Tfl	
► Overview	▶ Specifications	Drivers & Downloads	▶ Support	
Form Factor		9.75" PCIe x16 form	n factor	
# of Tesla GPUs		1	1	
# of CUDA Core		448		
Frequency of CUDA	Cores	1.15 GHz	1.15 GHz	
Double Precision floating point performance (peak)		515 Gflops	515 Gflops	
Single Precision floating point performance (peak)		1.03 Tflops	1.03 Tflops	
Total Dedicated Memory* Tesla C2050 Tesla C2070		2050 3GB GDDR5 2070 6GB GDDR5	3GB GDDR5 6GB GDDR5	
Memory Speed		1.5 GHz	1.5 GHz	
Memory Interface		384-bit		
Memory Bandwidth		111 60 / 100	144 GB/sec	



Brown ton Brown ton Additional View	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Processor         delivers supercomputing power at 1/20th the power         consumption and 1/10th the cost, bringing the performance of a small cluster to the desktop.         Image: Process of the performance of a small cluster to the desktop.         Image: Process of the performance of a small cluster to the desktop.	Bandwidth	n = 144	
▶ Overview	Specifications	Drivers & Downloads	▶ Support	
Form Factor		9.75" PCle x16 fo	orm factor	
# of Tesla GPUs		1	1	
# of CUDA Core		448	448	
Frequency of CUD	A Cores	1.15 GH	1.15 GHz	
Double Precision floating point performance (peak)		515 Gflop	515 Gflops	
Single Precision floating point performance (peak)		1.03 Tflo	1.03 Tflops	
Total Dedicated Memory* Tesla C2050 Tesla C2070		2050 3GB GDDF 2070 6GB GDDF	3GB GDDR5 6GB GDDR5	
Memory Speed		1.5 GHz	:	
Memory Interface		384-bit	384-bit	
Memory Bandwidth		144 GB/se	144 GB/sec	



Banera teat	Tesla C2050 / C2070 GPU Computing Processor         NVIDIA® Tesla™ C2050/2070 Computing Processor         delivers supercomputing power at 1/20th the power         consumption and 1/10th the cost, bringing the         performance of a small cluster to the desktop.            • Buy from participating partners.	work/data = ~ 29 (p		
▶ Overview	▶ Specifications	Drivers & Downloads	▶ Support	
Form Factor		9.75" PCIe x16 form	n factor	
# of Tesla GPUs		1		
# of CUDA Core		448		
Frequency of CUDA Cores		1.15 GHz		
Double Precision floa	ting 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		



# **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.**