

Analysis Report

evolvi(configurazione, curandStateXORWOW*)

Duration	1.467 ms (1,466,957 ns)
Grid Size	[128,1,1]
Block Size	[1024,1,1]
Registers/Thread	26
Shared Memory/Block	0 B
Shared Memory Requested	64 KiB
Shared Memory Executed	64 KiB
Shared Memory Bank Size	4 B

[0] GeForce 840M

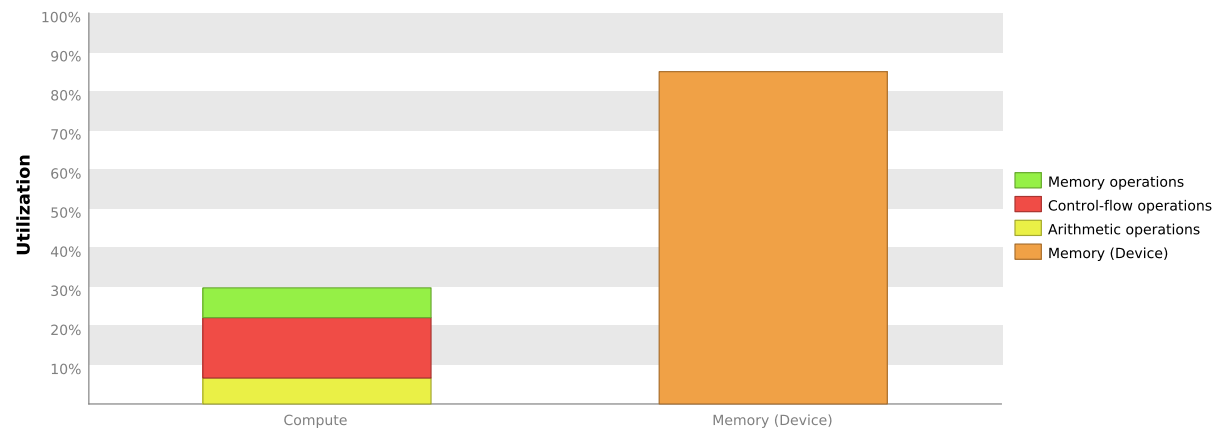
GPU UUID	GPU-4a39874c-8303-b7e3-9758-a265118f0297
Compute Capability	5.0
Max. Threads per Block	1024
Max. Shared Memory per Block	48 KiB
Max. Registers per Block	65536
Max. Grid Dimensions	[2147483647, 65535, 65535]
Max. Block Dimensions	[1024, 1024, 64]
Max. Warps per Multiprocessor	64
Max. Blocks per Multiprocessor	32
Single Precision FLOP/s	863.232 GigaFLOP/s
Double Precision FLOP/s	26.976 GigaFLOP/s
Number of Multiprocessors	3
Multiprocessor Clock Rate	1.124 GHz
Concurrent Kernel	true
Max IPC	6
Threads per Warp	32
Global Memory Bandwidth	14.4 GB/s
Global Memory Size	2 GiB
Constant Memory Size	64 KiB
L2 Cache Size	1 MiB
Memcpy Engines	1
PCIe Generation	2
PCIe Link Rate	5 Gbit/s
PCIe Link Width	4

1. Compute, Bandwidth, or Latency Bound

The first step in analyzing an individual kernel is to determine if the performance of the kernel is bounded by computation, memory bandwidth, or instruction/memory latency. The results below indicate that the performance of kernel "evolvi" is most likely limited by memory bandwidth. You should first examine the information in the "Memory Bandwidth" section to determine how it is limiting performance.

1.1. Kernel Performance Is Bound By Memory Bandwidth

For device "GeForce 840M" the kernel's compute utilization is significantly lower than its memory utilization. These utilization levels indicate that the performance of the kernel is most likely being limited by the memory system. For this kernel the limiting factor in the memory system is the bandwidth of the Device memory.



2. Memory Bandwidth

Memory bandwidth limits the performance of a kernel when one or more memories in the GPU cannot provide data at the rate requested by the kernel. The results below indicate that the kernel is limited by the bandwidth available to the device memory.

2.1. Global Memory Alignment and Access Pattern

Memory bandwidth is used most efficiently when each global memory load and store has proper alignment and access pattern.

Optimization: Each entry below points to a global load or store within the kernel with an inefficient alignment or access pattern. For each load or store improve the alignment and access pattern of the memory access.

`/home/giuseppe/Documents/myCUDA/ostacoli-2D/poligoni/improved/cuda_impr/const_memory/one_more/inside/inside.cu`

Line 153	Global Load L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 153	Global Load L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 153	Global Load L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 153	Global Load L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 161	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32750 L2 transactions for 4096 total executions]
Line 161	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32750 L2 transactions for 4096 total executions]
Line 167	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 167	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 167	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]
Line 167	Global Store L2 Transactions/Access = 8, Ideal Transactions/Access = 4 [32768 L2 transactions for 4096 total executions]

`/usr/local/cuda-7.5/bin/./include/curand_normal.h`

Line 310	Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [131072 L2 transactions for 4096 total executions]
Line 310	Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]
Line 312	Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]
Line 312	Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]
Line 313	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]
Line 313	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]

/usr/local/cuda-7.5/bin/./include/curand_normal.h

Line 313	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 8 [131072 L2 transactions for 4096 total executions]
Line 315	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [131072 L2 transactions for 4096 total executions]
Line 316	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [131072 L2 transactions for 4096 total executions]
Line 319	Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [131072 L2 transactions for 4096 total executions]

2.2. GPU Utilization Is Limited By Memory Bandwidth

The following table shows the memory bandwidth used by this kernel for the various types of memory on the device. The table also shows the utilization of each memory type relative to the maximum throughput supported by the memory. The results show that the kernel's performance is potentially limited by the bandwidth available from one or more of the memories on the device.

Optimization: Try the following optimizations for the memory with high bandwidth utilization.







Shared Memory - If possible use 64-bit accesses to shared memory and 8-byte bank mode to achieved 2x throughput.

L2 Cache - Align and block kernel data to maximize L2 cache efficiency.

Unified Cache - Reallocate texture data to shared or global memory. Resolve alignment and access pattern issues for global loads and stores.

Device Memory - Resolve alignment and access pattern issues for global loads and stores.

System Memory (via PCIe) - Make sure performance critical data is placed in device or shared memory.

Transactions	Bandwidth	Utilization	
Shared Memory			
Shared Loads	0	0 B/s	
Shared Stores	0	0 B/s	
Shared Total	0	0 B/s	
L2 Cache			
Reads	655424	14.019 GB/s	
Writes	991159	21.199 GB/s	
Total	1646583	35.218 GB/s	
Unified Cache			
Local Loads	0	0 B/s	
Local Stores	0	0 B/s	
Global Loads	655360	14.017 GB/s	
Global Stores	991153	21.199 GB/s	
Texture Reads	180224	3.855 GB/s	
Unified Total	1826737	39.071 GB/s	
Device Memory			
Reads	262551	5.616 GB/s	
Writes	299165	6.399 GB/s	
Total	561716	12.014 GB/s	
System Memory			
[PCIe configuration: Gen2 x4, 5 Gbit/s]			
Reads	0	0 B/s	
Writes	5	106.942 kB/s	

3. Instruction and Memory Latency

Instruction and memory latency limit the performance of a kernel when the GPU does not have enough work to keep busy. The results below indicate that the GPU does not have enough work because instruction execution is stalling excessively.

3.1. Instruction Latencies May Be Limiting Performance

Instruction stall reasons indicate the condition that prevents warps from executing on any given cycle. The following chart shows the break-down of stalls reasons averaged over the entire execution of the kernel. The kernel has good theoretical and achieved occupancy indicating that there are likely sufficient warps executing on each SM. Since occupancy is not an issue it is likely that performance is limited by the instruction stall reasons described below.

Memory Dependency - A load/store cannot be made because the required resources are not available or are fully utilized, or too many requests of a given type are outstanding. Data request stalls can potentially be reduced by optimizing memory alignment and access patterns.

Synchronization - The warp is blocked at a `__syncthreads()` call.

Memory Throttle - Large number of pending memory operations prevent further forward progress. These can be reduced by combining several memory transactions into one.

Texture - The texture sub-system is fully utilized or has too many outstanding requests.

Not Selected - Warp was ready to issue, but some other warp issued instead. You may be able to sacrifice occupancy without impacting latency hiding and doing so may help improve cache hit rates.

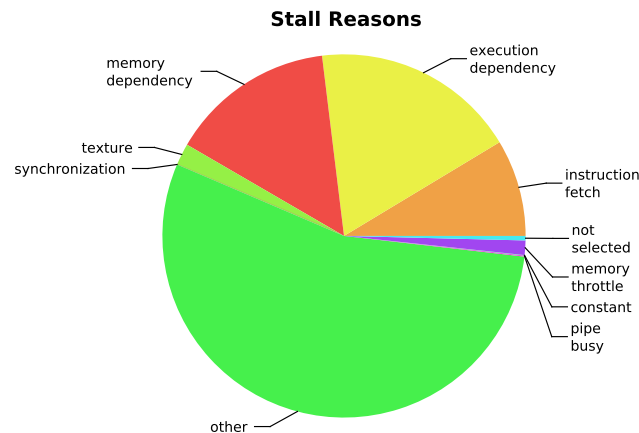
Pipeline Busy - The compute resource(s) required by the instruction is not yet available.

Constant - A constant load is blocked due to a miss in the constants cache.

Execution Dependency - An input required by the instruction is not yet available. Execution dependency stalls can potentially be reduced by increasing instruction-level parallelism.

Instruction Fetch - The next assembly instruction has not yet been fetched.

Optimization: Resolve the primary stall issue; other.



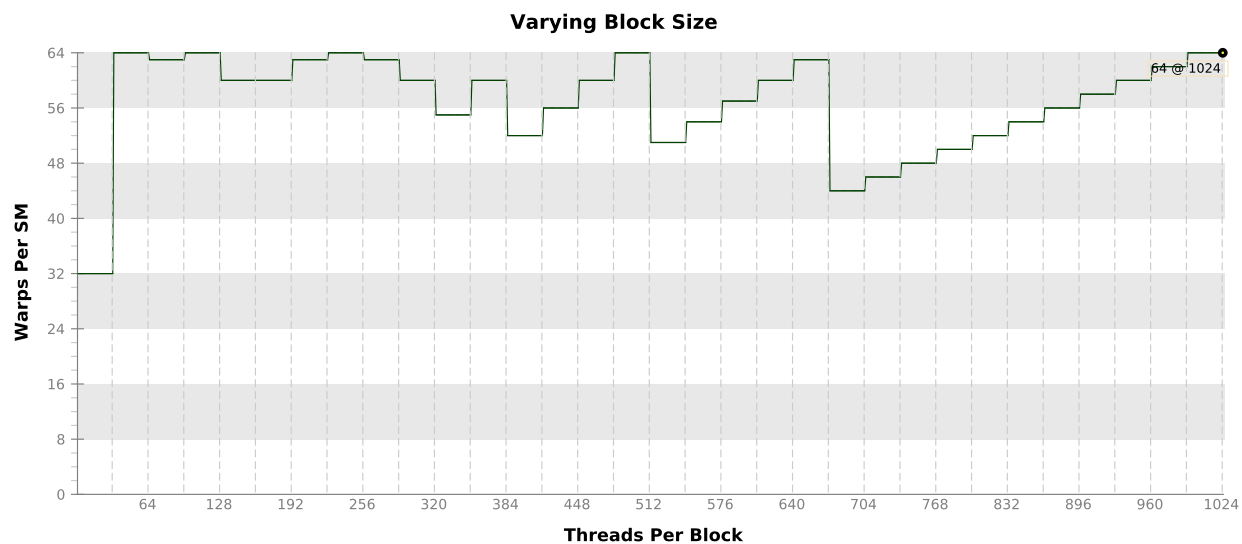
3.2. Occupancy Is Not Limiting Kernel Performance

The kernel's block size, register usage, and shared memory usage allow it to fully utilize all warps on the GPU.

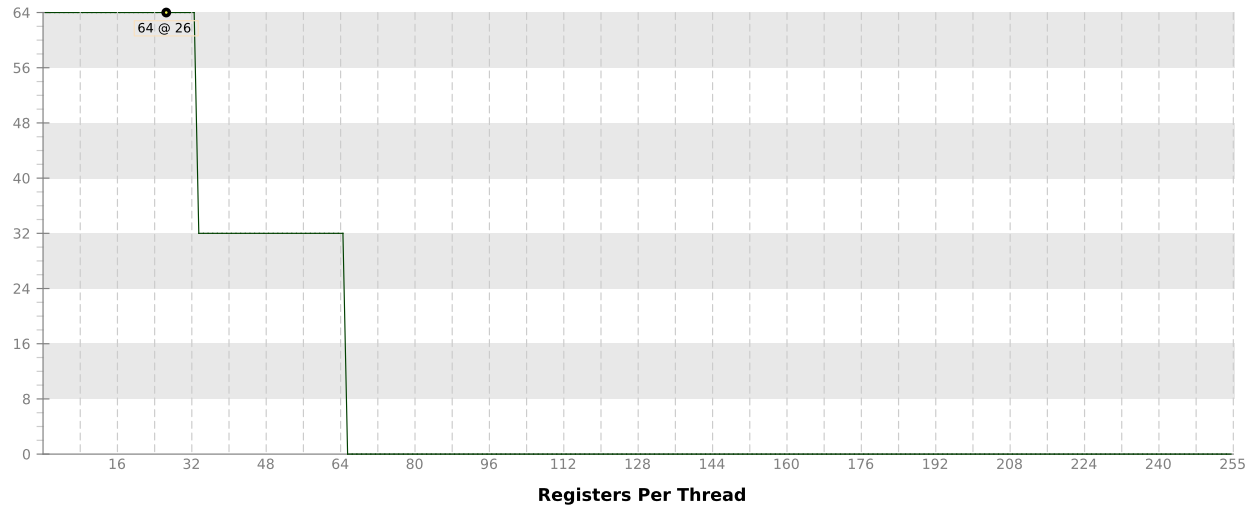
Variable	Achieved	Theoretical	Device Limit	Grid Size: [128,1,1] (128 blocks) Block Size: [1024,1,1] (1024 threads)
Occupancy Per SM				
Active Blocks		2	32	
Active Warps	52.63	64	64	
Active Threads		2048	2048	
Occupancy	82.2%	100%	100%	
Warps				
Threads/Block		1024	1024	
Warps/Block		32	32	
Block Limit		2	32	
Registers				
Registers/Thread		26	255	
Registers/Block		32768	65536	
Block Limit		2	32	
Shared Memory				
Shared Memory/Block		0	65536	
Block Limit			32	

3.3. Occupancy Charts

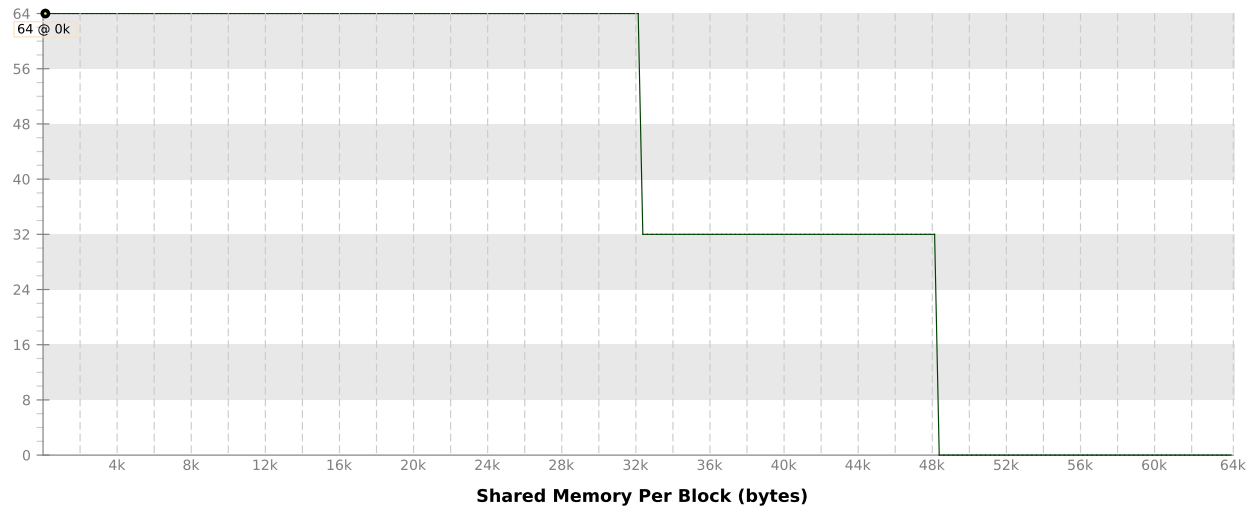
The following charts show how varying different components of the kernel will impact theoretical occupancy.



Varying Register Count



Varying Shared Memory Usage



4. Compute Resources

GPU compute resources limit the performance of a kernel when those resources are insufficient or poorly utilized. Compute resources are used most efficiently when all threads in a warp have the same branching and predication behavior. The results below indicate that a significant fraction of the available compute performance is being wasted because branch and predication behavior is differing for threads within a warp.

4.1. Low Warp Execution Efficiency

Warp execution efficiency is the average percentage of active threads in each executed warp. Increasing warp execution efficiency will increase utilization of the GPU's compute resources. The kernel's warp execution efficiency of 53.7% is less than 100% due to divergent branches and predicated instructions. If predicated instructions are not taken into account the warp execution efficiency for these kernels is 56.9%.

Optimization: Reduce the amount of intra-warp divergence and predication in the kernel.

4.2. Divergent Branches

Compute resource are used most efficiently when all threads in a warp have the same branching behavior. When this does not occur the branch is said to be divergent. Divergent branches lower warp execution efficiency which leads to inefficient use of the GPU's compute resources.

Optimization: Each entry below points to a divergent branch within the kernel. For each branch reduce the amount of intra-warp divergence.

/home/giuseppe/Documents/myCUDA/ostacoli-2D/poligoni/improved/cuda_impr/const_memory/one_more/inside/inside.cu

Line 111	Divergence = 81.1% [166118 divergent executions out of 204800 total executions]
Line 111	Divergence = 81.1% [166020 divergent executions out of 204800 total executions]
Line 160	Divergence = 99% [4053 divergent executions out of 4096 total executions]

4.3. Function Unit Utilization

Different types of instructions are executed on different function units within each SM. Performance can be limited if a function unit is over-used by the instructions executed by the kernel. The following results show that the kernel's performance is not limited by overuse of any function unit.

Load/Store - Load and store instructions for shared and constant memory.

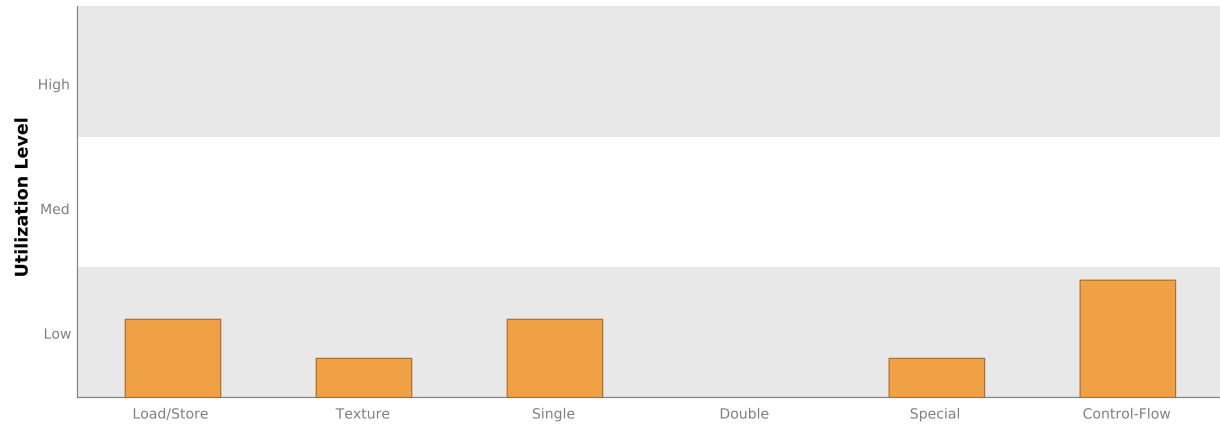
Texture - Load and store instructions for local, global, and texture memory.

Single - Single-precision integer and floating-point arithmetic instructions.

Double - Double-precision floating-point arithmetic instructions.

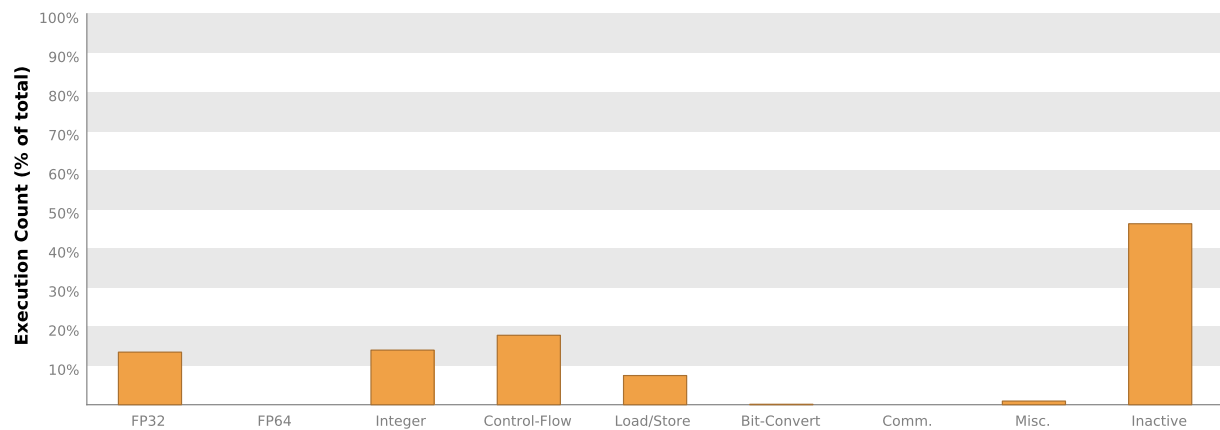
Special - Special arithmetic instructions such as sin, cos, popc, etc.

Control-Flow - Direct and indirect branches, jumps, and calls.



4.4. Instruction Execution Counts

The following chart shows the mix of instructions executed by the kernel. The instructions are grouped into classes and for each class the chart shows the percentage of thread execution cycles that were devoted to executing instructions in that class. The "Inactive" result shows the thread executions that did not execute any instruction because the thread was predicated or inactive due to divergence.



4.5. Floating-Point Operation Counts

The following chart shows the mix of floating-point operations executed by the kernel. The operations are grouped into classes and for each class the chart shows the percentage of thread execution cycles that were devoted to executing operations in that class. The results do not sum to 100% because non-floating-point operations executed by the kernel are not shown in this chart.

