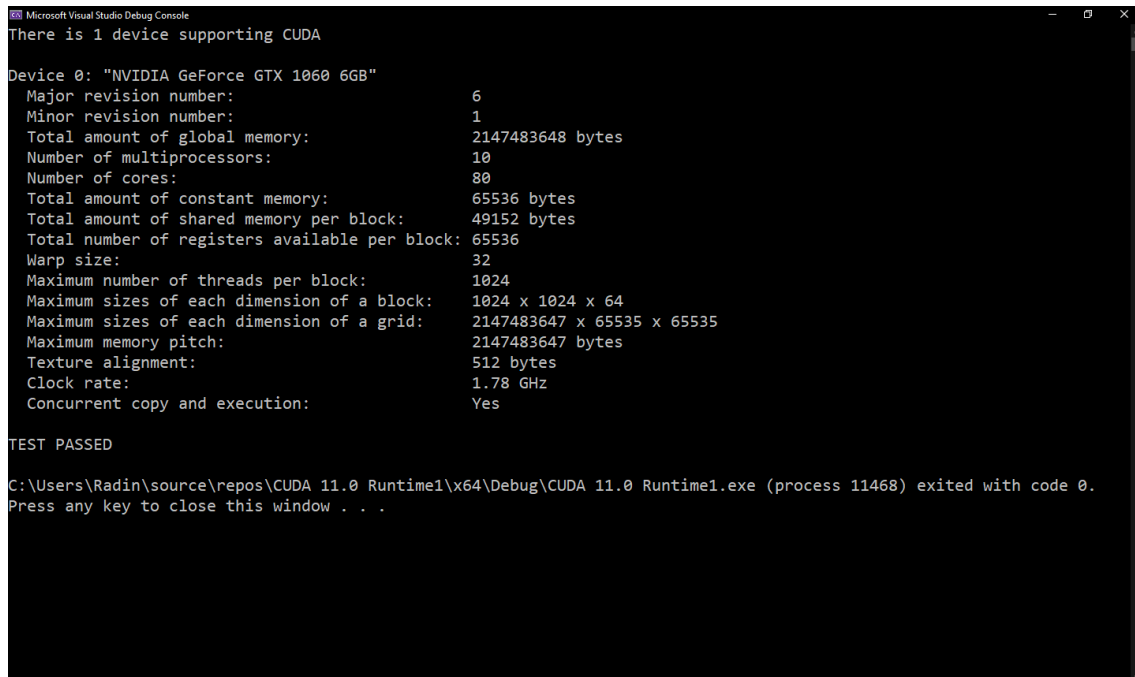


۱ گام اول

کد deviceQuery.cu داده شده را کامپایل و اجرا می‌کنیم. مشخصات دستگاه را در شکل ۱ مشاهده می‌کنیم.



```
Microsoft Visual Studio Debug Console
There is 1 device supporting CUDA

Device 0: "NVIDIA GeForce GTX 1060 6GB"
Major revision number:          6
Minor revision number:          1
Total amount of global memory:  2147483648 bytes
Number of multiprocessors:       10
Number of cores:                 80
Total amount of constant memory: 65536 bytes
Total amount of shared memory per block: 49152 bytes
Total number of registers available per block: 65536
Warp size:                      32
Maximum number of threads per block: 1024
Maximum sizes of each dimension of a block: 1024 x 1024 x 64
Maximum sizes of each dimension of a grid: 2147483647 x 65535 x 65535
Maximum memory pitch:            2147483647 bytes
Texture alignment:               512 bytes
Clock rate:                      1.78 GHz
Concurrent copy and execution:    Yes

TEST PASSED

C:\Users\Radin\source\repos\CUDA 11.0 Runtime1\x64\Debug\CUDA 11.0 Runtime1.exe (process 11468) exited with code 0.
Press any key to close this window . . .
```

شکل ۱: مشخصات دستگاه با استفاده از کد deviceQuery

۲ گام دوم

زمان‌های اجرا میانگین ۱۰ بار اجرا است. همچنین زمان پر و کپی کردن بردارها و آزاد کردن حافظه در نظر گرفته نشده است.

برنامه جمع دو بردار را در حالت سریال بر روی CPU اجرا می‌کنیم. زمان اجرا در حالت سریال به علت کوچک بودن برابر صفر گزارش می‌شود.

با اضافه کردن تابع addWithCuda و بردن محاسبات بر روی GPU زمان اجرا به ۰/۰۰۰۰۴۸ ثانیه می‌رسد. دلیل کاهش سرعت اجرا، بالا بودن سربارهای بردن محاسبات بر روی GPU نسبت به اندازه مسئله است.

۳ گام سوم

به کد قسمت قبل متغیرهای NUM_THREADS، ELEMENTS_PER_THREAD و NUM_BLOCKS را اضافه می‌کنیم و تغییرات لازم را در کرنل انجام می‌دهیم. زمان‌های اجرای آزمایش شده در جدول ۱ آمده است (تسریع با میانگین‌گیری تسریع دو ستون انتهایی محاسبه شده است).

جدول ۱: زمان‌های اجرا (ثانیه) به ازای اندازه ورودی‌های مختلف

تسریع	اندازه ورودی			موازی‌سازی
	2^{28}	2^{27}	2^{26}	
–	۰/۲۸۱۴۲۶	۰/۱۴۱۷۹۰	۰/۰۷۱۸۰۲	سریال
۰/۳۶	۱/۰۳۸۵۷۰	۰/۳۰۲۵۶۹	۰/۱۵۳۹۵۹	پردازش n المان توسط هر نخ
۱۳/۰۰	۰/۰۲۱۴۷۱	۰/۰۱۰۹۹۰	۰/۰۰۵۵۶۱	پردازش با n بلوک

در حالتی که هر نخ یک المان را پردازش می‌کند متغیر ELEMENTS_PER_THREAD برابر یک قرار داده شده است. در حالت پردازش n المان توسط هر نخ، مقدار این متغیر به شکل زیر محاسبه شده است.

```
int ELEMENTS_PER_THREAD = size / 1024;
```

در شکل‌های ۲ و ۳ به ترتیب خروجی برنامه در حالت پردازش n المان توسط هر نخ و پردازش یک المان توسط n بلوک برای اندازه ورودی 2^{28} آمده است.

همانطور که می‌بینیم در حالت اول تنها یک بلوک نخ ۱۰۲۴ تایی داریم و هر نخ ۲۶۲۱۴۴ المان را پردازش می‌کند. این حالت به دلیل شباهت نحوه محاسبه به محاسبات روی CPU، روی هسته‌های کوچک و ضعیف GPU به خوبی جواب نمی‌دهد و اجرای آن بسیار کندتر است.

در حالت دوم هر نخ تنها یک المان را پردازش می‌کند اما تعداد بلوک‌های ۱۰۲۴ تایی ۲۶۲۱۴۴ است. این کار (دادن کارهای کوچک به هر نخ و زیاد کردن تعداد نخ‌ها با افزایش تعداد بلوک‌ها) باعث استفاده حداکثری از قدرت GPU می‌شود و زمان اجرا به شدت کاهش می‌یابد.

```
Microsoft Visual Studio Debug Console
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.037100 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.040595 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.038132 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.037942 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.037367 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.039149 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.038575 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.039040 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.038712 Secs
elements per thread: 262144, threads per blocks: 1024, blocks: 1
[-] Time Elapsed: 1.039093 Secs

[-] The average running time was: 1.038570

C:\Users\Radin\source\repos\CUDA 11.0 Runtime1\x64\Release\CUDA 11.0 Runtime1.exe (process 11236) exited with code 0.
Press any key to close this window . . .
```

شکل ۲: پردازش n المان توسط هر نخ (درشت دانگی) منجر به کاهش شدید سرعت روی GPU می‌شود.

```
Microsoft Visual Studio Debug Console
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021675 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021613 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021651 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021338 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021634 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021354 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021325 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021391 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021387 Secs
elements per thread: 1, threads per blocks: 1024, blocks: 262144
[-] Time Elapsed: 0.021345 Secs

[-] The average running time was: 0.021471

C:\Users\Radin\source\repos\CUDA 11.0 Runtime1\x64\Release\CUDA 11.0 Runtime1.exe (process 21024) exited with code 0.
Press any key to close this window . . .
```

شکل ۳: پردازش یک المان توسط هر نخ (ریز دانگی) منجر به کاهش شدید زمان اجرا و تسريع مناسب روی GPU می‌شود.