هدف: آشنایی با مفاهیم Cuda

۱. میخواهیم دو بردار را با یکدیگر جمع کنیم. اگر بخواهیم هر نخ یک خروجی را تولید کند، اندیس مناسب برای بردار خروجی کدام است؟

a. i = threadIdx.x + threadIdx.y;
b. i = blockIdx.x + threadIdx.x;
c. i = blockIdx.x*blockDim.x + threadIdx.x;
d. i = blockIdx.x * threadIdx.x;

۲. برای جمع دو بردار به طول ۸۰۰۰ عنصر، هر نخ یک خروجی را تولید می‌کند و اندازه بلوک ۱۰۲۴ نخ می‌باشد. برنامه نویس kernel launch را به گونه‌ای تنظیم می‌کند که با کمترین تعداد بلوک نخ همه‌ی عناصر بردار پوشش داده شوند. در این شرایط چند نخ در grid وجود دارد؟

a. 8000
b. 8196
c. 8192
d. 8200

۳. در برنامه‌نویسی کودا منظور از Compute Capability چیست؟

۴. PTX چیست و چگونگی استفاده‌ی از آن را توضیح دهید. ساختار PTX به چه صورت است؟

۵. ساختار یک CUDA core را توضیح دهید. آیا در یک CUDA core امکان بکارگیری همزمان واحد ممیز شناور و عدد صحیح وجود دارد؟ اگر جواب مثبت است از چه معماری این قابلیت اضافه شده است؟

۶. هسته‌های tensor چیست و چه استفادهای دارند؟ آیا امکان استفاده‌ی آن‌ها به صورت مستقیم وجود دارد؟

۷. آیا امکان استفاده تواما OpenMP و کودا با هم وجود دارد؟ در صورت مثبت بودن جواب به نظر شما در چه مواردی کاربرد دارد؟

۸. برنامه‌ای بنویسید که هر نخ grid و block خود را به صورت زیر چاپ کند.

Hello CUDA I’m a thread from grid X and block

۹. هنگام ساخت پروژه کودا در Visual Studio برای مثال و امتحان کارکرد صحیح مجموعه‌ی کودا، کد پیشفرضی برای جمع دو بردار تولید میکند. قسمت‌های زیر را انجام داده و زمان اجرا را برای اندازه‌ی ده میلیون المان گزارش کنید و به سؤال‌های موجود پاسخ دهید (زمان پر و کپی کردن آرایه‌ها را در نظر نگیرید و فقط زمان اجرای kernel و عملیات جمع در نظر گرفته شود)

A. برنامه‌ی جمع بردار را به صورت سریال بنوسید.
B. برنامه‌ی جمع بردار را با استفاده از OpenMP موازی کنید.
C. برنامه‌ی پیشفرض را با تغییر grid size و ،block size برای انجام هر جمع توسط یک نخ آماده کنید
   i. اندازه‌ی grid برگتر باعث تسریع بیشتر می‌شود یا اندازه‌ی block بزرگتر؟ چرا؟
D. برنامه‌ی پیشفرض تولیدشده را به‌گونهای تغییر دهید که هر نخ بیشتر از یک المان را جمع کند.

   i. ریزدانگی (تعداد نخ بیشتر) و درشت‌دانگی (تعداد نخ کمتر) کدامیک برای این مسئله مناسبتر است؟ چرا؟

   ii. اندازه‌ی مناسب grid و block برای حداکثر تسریع چه رابطهای با معماری GPU داشته است؟

۱۰. سؤال‌های زیر را به صورت کوتاه جواب بدهید.

A. تفاوت PTX و SASS چیست؟
B. دلیل وجود اشاره‌گر دوگانه در آرگومان اول CudaMalloc چیست؟

نکات مربوط به ارزیابی

گزارش: پاسخ سوالات مطرح شده را در یک فایل پی‌دی‌اف بنویسید و فایل پی‌دی‌اف را به همراه کدهای پیاده‌سازی شده در یک فایل زیپ قرار دهید. **توجه نمایید در پاسخ به سوالات، تمامی استدلالات خود شامل توضیحات یا نتایج آزمایشات انجام شده را ذکر نمایید.**

تذکر: مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیداً برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری بلامانع است، اما کپی کردن غیرمجاز است.

راهنمایی: در صورت نیاز می‌توانید سوالات خود را در خصوص پروژه از تدریس یار درس، از طریق ایمیل زیر بپرسید.
E-mail: multicore.ceit.aut@gmail.com

ارسال: فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW05.zip ارسال نمایید. شایان ذکراست هر روز تاخیر باعث کسر ۱۰٪ نمره خواهد شد. (تحویل تمرین‌ها از طریق سایت کورسز انجام خواهد شد.)

موفق باشید

سید امیرحسین سعیدی