

برنامه‌نویسی چند هسته‌ای

تمرین ۵

۱. در برنامه‌نویسی کودا منظور از Compute Capability چیست؟
۲. PTX چیست و چگونه استفاده‌ای از آن را توضیح دهید. ساختار PTX به چه صورت است
۳. 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؟
۴. هسته‌های tensor چیست و چه استفاده‌ای دارند؟ آیا امکان استفاده‌ای آنها به صورت مستقیم وجود دارد؟
۵. آیا امکان استفاده تواما OpenMP و کودا با هم وجود دارد؟ در صورت مثبت بودن جواب به نظر شما در چه مواردی کاربرد دارد؟
۶. 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 Y
۷. هنگام ساخت پروژه کودا در Visual Studio برای مثال و امتحان کارکرد صحیح مجموعه‌ی کودا، کد پیش فرضی 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 بنویسید.
 - b. برنامه‌ی جمع بردار را با استفاده از OpenMP موازی کنید.
 - c. برنامه‌ی پیش فرض را با تغییر grid size و block size، برای انجام هر جمع توسط یک نخ آماده کنید.
 - i. اندازه‌ی grid بزرگ‌تر باعث تسریع بیشتر می‌شود یا اندازه‌ی block بزرگ‌تر؟ چرا؟
 - d. برنامه‌ی پیش فرض تولیدشده را به گونه‌ای تغییر دهید که هر نخ بیشتر از یک المان را جمع کند.
 - i. ریزدانه‌ی (تعداد نخ بیشتر) و درشت‌دانه‌ی (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؟
۸. سؤال‌های زیر را به صورت کوتاه جواب بدهید.
 - a. تفاوت PTX و SASS چیست؟
 - b. دلیل وجود اشاره گر دوگانه در آرگومان اول CudaMalloc چیست؟

با آرزوی موفقیت