محاسبات عددی **سری شمارهٔ ۱**

نیم‌سال اول ۱۴۰۲-۱۴۰۳ **موعد تحویل: ۲۵ آبان**

سوال ۱. در این تمرین لازم است تا برنامه‌ی ضرب دو ماتریس دوبعدی را با استفاده از CUDA در زبان برنامه‌نویسی C/C++ بنویسید. داده‌ها را از نوع float و ابعاد آن‌ها را ۱۰۲۴۰ * ۱۰۲۴۰ فرض کنید.

لطفا گزارشی از آن‌چه کردید و تحلیل آن و هم‌چنین سورس‌کد پاسخ خود را آپلود کنید.

الف) ابتدا این برنامه را به C/C++ عادی بنویسید و زمان اجرای برنامه در CPU را محاسبه کنید. درصورتی‌که ابعاد ۱۰۲۴۰ بیش‌ازحد بزرگ بود و اجرای آن زیاد طول می‌کشید مجازید برنامه‌ی خود را در ابعاد ۱۰۲۴ اجرا کنید اما در تحلیل‌های بعدی خود این مسأله را در نظر بگیرید.

ب) محاسبات برنامه را در CUDA بازنویسی کنید. در این مرحله نیازی به استفاده از امکانات GPU مانند Shared Memory ندارید.

ج) برنامه‌ی پیشین را با استفاده از Shared Memory بازنویسی کنید. بررسی کنید که Shared Memory حدودا در کدام ابعاد ماتریس‌ها پر می‌شود.

د) راجع‌به تکنیک tiling تحقیق کنید. با استفاده از tiling برنامه‌ی پیشین خود را بهبود ببخشید و توضیح بدهید که چرا این برنامه نسبت‌به برنامه‌ی قبلی عملکرد بهتری دارد.

ه) راجع‌به memory coalescing در GPU تحقیق کنید. اکنون کدام دستورات load در برنامه‌ی شما از memory coalescing استفاده می‌کنند و کدام‌ها نه؟ چه روشی برای استفاده از memory coalescing در این دستورات پیشنهاد می‌دهید؟ آن را پیاده‌سازی کنید و overhead پیاده‌سازی و هم‌چنین بهبود عملکرد حاصل از آن را گزارش دهید.

راهنمایی: می‌توانید راجع‌به نمایش‌های row major order و column major order در آرایه‌های دوبعدی تحقیق کنید.

موفق باشید.