



هدف از این تمرین پیاده‌سازی عمل **reduction** بر روی GPU است. این عمل یک آرایه را از ورودی گرفته و مجموع درایه‌های آن را محاسبه می‌کند. در این تمرین برای پیاده‌سازی عمل **reduction** بر روی GPU، ابتدا با یک **kernel** ساده شروع می‌کنیم و گام به گام با بهبود **kernel** تلاش می‌کنیم تا تسریع بیشتری حاصل شود. به عنوان راهنمای انجام کار، یک سند راهنما از شرکت **NVIDIA** به پیوست تمرین در دسترس قرار دارد. لطفاً با مطالعه‌ی این سند، پیاده‌سازی‌ها را تا صفحه 23 انجام داده و سپس نتایج و کد هر گام را جداگانه گزارش کنید. اگرچه اکیدا توصیه به مطالعه و پیاده‌سازی تمام سند می‌شود اما مطالب بیشتر آن هدف این تمرین نیست. کد سریال عمل **reduction** نیز به پیوست ضمیمه شده است تا در نهایت نتایج به دست آمده در هر مرحله با کد سریال مقایسه و گزارش شود. بنابراین گزارش نهایی شما شامل موارد ذیل خواهد بود:

- فایل کد هر گام
- فایل گزارش

○ محاسبه‌ی **peak bandwidth** برای کارت گرافیک مورد استفاده (همانند صفحه 6 سند)

○ توضیح **kernel** در هر گام

○ جدول نتایج (عینا مانند جدول صفحه 23 سند)

- برای محاسبه‌ی پهنای باند هر **int** را چهار **byte** فرض کنید.
- زمان‌های جدول صرفاً زمان محاسبات هستند. زمان کل (محاسبات + انتقال داده‌ها) را نیز گزارش کنید.
- مقایسه بهترین زمان اجرا روی GPU با کد سریال CPU (گزارش زمان‌ها و محاسبه‌ی تسریع)
 - آیا برای حجم داده‌ی بالا اختلاف تغییر می‌کند؟ تحلیل کنید.
 - با زمان محاسبات و کل مقایسه کنید. کدام مقایسه منصفانه است؟
- برای هر مرحله علاوه بر ورودی پارامترهای **kernel launch** (تعداد بلوک‌ها و نخ‌ها) نیز قید شوند.