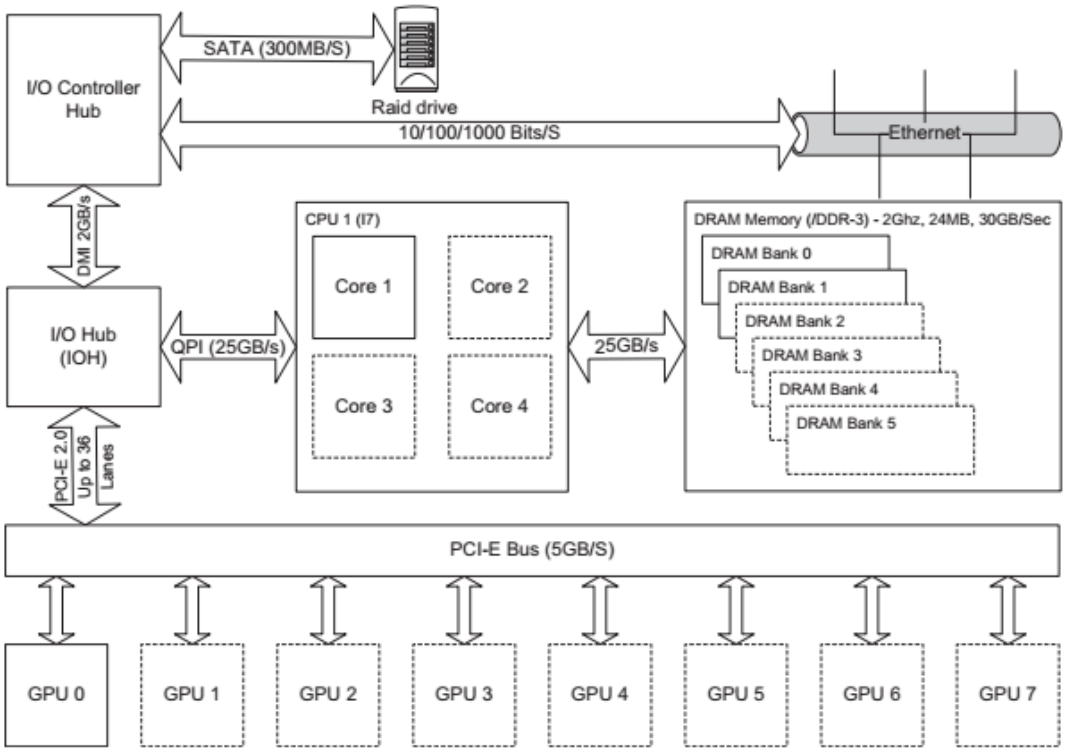
به نام خدا

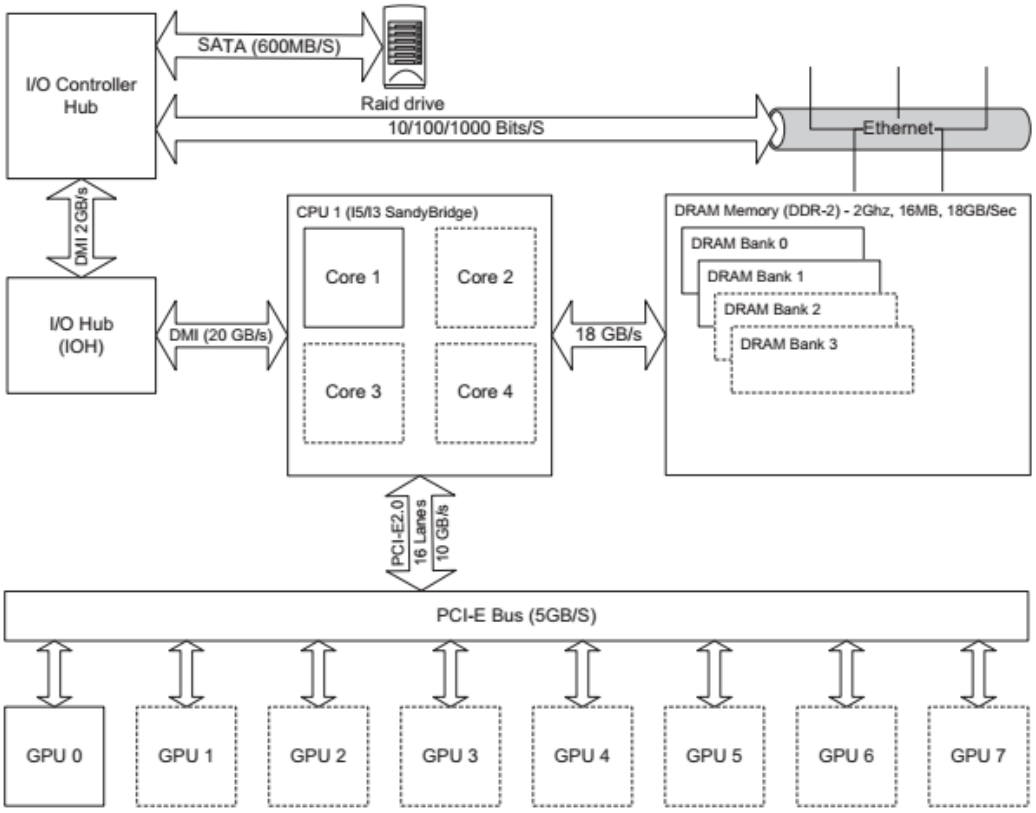
برنامه‌نویسی چندهسته‌ای

1. یک سیستم اجزای متفاوتی دارد. اجزای تشکیل دهنده یک سیستم -به‌طور نمونه- عبارت‌اند از یک یا تعدادی CPU، GPU، RAM، HDD، SDD و ... . این اجزاء در نسل‌های مختلف معماری به شیوه‌های متفاوتی به یکدیگر متصل می‌شوند که در نهایت کارایی سیستم را تعیین می‌کنند. معماری چند نسل از سیستم‌های کامپیوتری در شکل‌های 1 و 2 نشان داده شده است. لطفاً از کتاب [Cook12][1] از فصل 3، بخش PC Architecture را مطالعه کنید. سپس با استدلال به سؤالات خواسته شده پاسخ دهید.
   - سرعت اتصال حافظه‌های جانبی در دو معماری چه تفاوتی دارد؟
   - آیا سرعت اتصال حافظه‌های جانبی اهمیت دارد؟ چرا؟
   - برای اتصال GPU به سیستم از چه درگاهی استفاده می‌شود؟ ویژگی‌های این درگاه چیست؟
   - با توجه به شکل‌های نشان داده شده چند GPU و با چه پهنای باندی می‌توان در هر معماری به سیستم متصل کرد؟



شکل 1 Nehalem/X58 system

[1] Cook, S., 2012. *CUDA programming: a developer's guide to parallel computing with GPUs*.



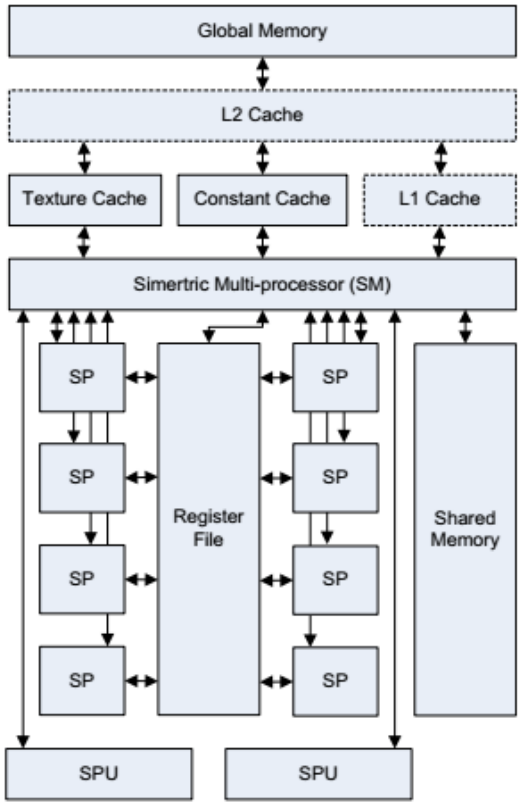
شکل 2 Sandybridge design

2. اجزای اصلی پردازنده‌ی گرافیکی عبارت‌اند از

I. Memory (global, constant, shared)
II. Streaming multiprocessors (SMs)
III. Streaming processors (SPs)

a. هر یک از این اجزا را شرح داده و ارتباط میان آن‌ها را توصیف کنید.
b. در GPU از حافظه GDDR استفاده می‌شود. تفاوت آن با DDR در چیست؟ چرا مانند CPU از DDR استفاده نمی‌شود؟
c. شکل 3 درون یک SM را نشان می‌دهد. اجزای درونی نشان داده شده را شرح دهید.
d. حافظه‌های Global، Texture و Constant چه فرقی با یکدیگر دارند؟ آیا این حافظه‌ها از نظر فیزیکی مجزا هستند؟

راهنمایی: می‌توانید از کتاب [Cook12] از فصل 3، بخش GPU Hardware را مطالعه کنید.



شکل 3 Inside an SM

3. مفهوم Compute Level (Compute Capability) در CUDA چیست؟

4. مفهوم Occupancy در CUDA چیست؟ با در نظر گرفتن چه مؤلفه‌هایی محاسبه می‌شود؟

5. می‌خواهیم دو بردار را به 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;

6. برای جمع دو بردار به طول 8000 عنصر، هر نخ یک خروجی را تولید می‌کند و اندازه بلوک 1024 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