Introduction to Computer Architecture: Supervision 5

Lectures covered by the supervision: [https://www.cl.cam.ac.uk/teaching/2223/IntComArch/](https://www.cl.cam.ac.uk/teaching/2223/IntComArch/)
- Lecture 15: Cuda, OpenCL.
- Summary: Computer design, SoC, and software engineering.


Supervision questions:
1. 2019 Paper 5 Question 3 - parts a, b, c.
2. Discuss an example of a real-world problem that we can solve efficiently by using CUDA.
3. Discuss an example of a problem that we can solve efficiently by using OpenCL.
4. How is it possible that so many different devices, with completely different internal architecture are compatible with OpenCL and enable acceleration of execution?
5. In CUDA, what’s the difference between a thread block and a warp? Create an example describing how execution of code maps to concepts such is thread blocks and warps.
6. Compare and contrast the OpenCL programming model with CUDA.
7. What types of application best suit GPUs and how can you program them to take advantage of the various forms of parallelism available? Try to describe your answer using examples.
8. Why is energy efficiency the “new fundamental limiter of processor performance”, as Borkar and Chien say?
9. Describe Arm’s big.LITTLE system.
10. When might it make sense to implement functionality in a specialised accelerator rather than within a general-purpose core?
11. Discuss a general multicore CPU vs ASIC vs DSP.
    a. What are important aspects to compare?
    b. How would you decide which accelerator(s) to use if you were designing a system?
12. Discuss approximate computing.
14. Discuss the Spectre bug.
15. Create a diagram with all discussed elements of a computer system (e.g., cache, RAM). Discuss common units of time for operations that these elements perform (e.g., time to fetch memory from cache and RAM, time to execute an instruction).
16. Write a 1-page essay (no longer) describing how topics discussed in different lectures are related, and how in your opinion they relate to software engineering. Focus on identifying conceptual solutions discussed in different lectures and explain how these abstract concepts are applied for different concrete problems, on different abstraction levels.
17. Summarize the main message from “Lecture 15: Cuda, OpenCL” in 1-3 sentences?
19. BONUS:
    a. Discuss SYCL.
    b. Create an example of a SoC, on which some processes and threads might execute concurrently. Discuss sources of non-determinism starting from cache and pipelines, to RAM access, network, scheduling, pre-emption, synchronisation between threads... Consider there are also some cloud services. Discuss what range of delays non-determinism can cause.
    c. Discuss Meltdown bug and compare it with Spectre bug.
d. Discuss Misra C in the context of security.

e. Create an example with CUDA and commit it to GitHub.

f. Create an example with OpenCL ("fortune favours the bold") and commit it to GitHub.

Save your answers into MS Teams or email them to me. Please use the following naming pattern:

ICA_Supervision_5_Answers_<last name>_<first name>_Michaelmas_2022

Send your answers as a pdf, doc, image, or any other format of a document for which there exists an easily available software to open.

Jasmin JAHIĆ
jj542@cam.ac.uk
https://www.cl.cam.ac.uk/~jj542/