<https://youtu.be/g0kOLSpdHPI>

Whether you’re a student cramming work before piling deadlines, a dad needing to choose between work and caring for his children, or even the king of England needing to choose between his many vacation homes. none of us are strangers to the idea of having little time to spend, but many things we want to do. Stuck between a rock and a hard place, I turned to my best ally, computers! No, not that homework planner, I’m talking GPUs, the ultimate mundane work cruncher. Surprisingly, many strategies GPUs use reflect strategies we can use in real life, and I’d like to share them with you now.

First, we’ll put all our homework, meetings, holidays into 1 large bundle of GPU tasks. We then serialize our tasks into GPU-native instructions, organized into blocks of 32 tasks, and assigned across multiple GPU sub-processors, known as Streaming Multiprocessors. Each multiprocessor contains a 128kb L1 Instruction Cache, L0 Instruction Cache, a register file containing 32767 32-bit registers, a 128kb L1 Data Cache, and the execution partitions. What’s a cache? Caching is the strategy of storing frequently used data in faster access memory. We’ll be seeing more usage of caches as we go deeper.

Tasks are stored on the L1 or L0 instruction cache for faster retrieval, and then sent to an instruction decoder which interprets the necessary steps to execute an task. To ensure workload is distributed equally across multiprocessors. data can also be assigned to the register file via the L1 Data Cache.

We now look at the processing partitions. We have 2 processing blocks, which perform fast calculations such as simple arithmetic and matrix dot products, and the Memory Input Output partition, which handles slow calculations such as trigonometric and transcendental root functions, and is also responsible for all data allocation in the system.

Now, let’s take a closer look at a processing partition. Our blocks of 32 tasks from earlier will arrive at the warp scheduler, which will dispatch a fixed number of tasks from each block consecutively, a strategy known as Round Robin Scheduling, hence ensuring consistent execution progress across blocks in the same multiprocessor.

Task blocks are sent to the math dispatch unit, which delegates tasks to specialized computation units, each of which process different types of arithmetic and memory operations. The first of which being the ALU, which can perform bitwise operations and is used for half precision to full precision floating point conversion in Ampere GPUs and higher. The CUDA Cores are the powerhouse of the processor, performing most arithmetic operations. Lastly, Tensor Cores, a matrix operation unit that can perform up to 16 dot product computations at a time, bypass the math dispatch unit entirely.

We then fetch data from the register file for processing. Execution of arithmetic operations typically takes many hundreds of clock cycles. However, we can take advantage of the same tasks being executed repeatedly by leveraging the concept of pipelining. Execution of a calculation goes through multiple different circuits, leaving circuits that are not immediately participating in a calculation inactive. We can fill this inactivity by following up with a second calculation, turning our few hundred clock cycle execution into effectively a 1 clock cycle execution time on average. Arithmetic that will take longer will often be prioritized, to overlap operations and maximize use of pipeline stages.

Lastly, the outputs from all computation units are sent to the register cache, and subsequently the register file.

Next, we will discuss the MIO Partition. Slow tasks are sent by the math dispatch unit to a queue, and then subsequently sent to the MIO Scheduler. The scheduler can queue either memory tasks or arithmetic tasks. memory tasks are queued to the Load Store Unit, which resolves memory locations for requested data, and sends lookup requests to memory units outside the Streaming Multiprocessor. The LSU will store our requested operands in the register file. Since fetching data can be slow, we can continue execution while we wait for our data. Our retrieved data can also be stored in caches. We have caches for constant values, arrays and register values among others. Meanwhile, arithmetic tasks are queued to the special function unit. Both the LSU and SFU can communicate data with the register file.

Finally, data written by the LSU to the L1 data cache is migrated upwards to the L2 data cache and then GPU memory, before being transferred to the host device through an IO port, completing the calculation.

It is my hope that these methods of multitasking will help you with your own tasks. Thanks, and goodbye.

**EL5132 Semester 2 2024 Academic Integrity Declaration Form**

NUS High School defines plagiarism as presenting someone else’s work without appropriate acknowledgement of the author or the source which is a form of academic dishonesty. This category of academic misconduct includes but is not limited to the following:

i) acquiring or commissioning a piece of work, which is not your own and representing it as if it were, by:

• purchasing a paper from a commercial service, including websites

• submitting a paper written by another person, either by a fellow student or a person who is not a member of the school;

ii) submission of the same or almost identical work for more than one assessment item;

iii) copying ideas, concepts, research data, images, sounds or text without appropriate acknowledgment;

iv) word for word copying, cutting or pasting statements from multiple sources including from work generated by the Artificial Intelligence (AI) or piecing together work of others and representing them as original work;

v) submitting all or part of another student’s work as one’s own, even with the student’s knowledge or consent.

Any student found to have committed, aided and/or abetted the offence of plagiarism may be subject to academic and disciplinary penalties.

**Guidelines on use of AI**

If you have made use of any sources or AI in your initial research or brainstorming, you must provide clear and accurate information for each source. You must also identify where and how they have been used in your work.

As this is a language module and you are assessed based on your language ability, AI or other generative AI software must not be used in crafting your sentences, organising your ideas or editing the language/flow of the prose. You must not use AI to craft any parts of your final product. AI tools may only be used for the explicit purposes of research and brainstorming. They must NOT be used for the writing of your script

You should not use AI tool as your sole research source. You should fact-check and corroborate the claims with that of other credible sources.

You should seek clarification from your EL5132 teachers if you are unsure whether you are plagiarising the work of another person, other sources, or AI.

**ACKNOWLEDGEMENT BY STUDENT**

I acknowledge that I have read the plagiarism warning and guidelines against plagiarism on page 1. I declare, to the best of my knowledge, that all material in this assignment is my own work and does not involve plagiarism.

|  |  |
| --- | --- |
| **Module Code** | **EL5132** |
| **Module Assignment** | **TED Talk** |
| **Name** | **scrub** |
| **Signature** |  |
| **Date** | **8/6/2024** |

**EL5132**

**Record of AI use**

1. If you have used AI in producing the video that you have submitted for assessment, please fill in table below accordingly.
2. You may add additional rows to reflect the number of prompts used and their output.
3. This table should be completed in addition to in-text and reference citation included in the soft copy of the script you have submitted.
4. If you have not used AI in producing the video, please submit fill in the fields below with ‘Nil’ and submit the document also.

|  |  |  |
| --- | --- | --- |
|  | **AI tool used** | **Prompt & Transcript of output**  Insert full screenshots that capture both the prompt used and the resultant response below |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |