IN5050: Programming heterogeneous multi-core processors

… because the OS-course is just too easy!

Home Exam 2: Video Encoding on Tegra Xavier using nVIDIA CUDA

March 15th, 2024

Håkon Kvale Stensland

[simula . research laboratory]

Video Encoding



- Pål still wants to encode some videos…
- Pål got a brand-new PowerPoint 2022 license, but animations are more resource demanding than he thought, and the email client uses the remaining CPU resources!
- However, Pål remember that his SoC has a GPU, so he needs you to help him use the GPU for video encoding (Use the CPU as little as possible, it is very occupied running Mail and PowerPoint animations).

Precode

- Same precode as Home Exam 1*!
- You are not allowed to change out the Motion Estimation, Motion Compensation or DCT algorithms.
- You are not allowed to paste code from other projects / encoders.
- You only need to optimize the Codec63 encoder!
- Your implementation is supposed to only focus on the GPU, we will not give points for threading on the CPU side.

** There is a CUDA-branch, and some fuctions are moved arround.*

Your task

- Utilize the data parallel GPU architecture and use the *Volta Tuning Guide*, and other tricks to tune your application.
- Start by profiling the encoder to see which parts of the encoder that are the bottlenecks.
- Remember, after optimizing one part of the code, more profiling might be needed to find new bottlenecks. Use the Nvidia profiler to profile the GPU code.
- Write a scientific report with details on which parts of the encoding process that benefited from your optimizations. The report should also explain how your code works.
- Remember to also report on optimization that did not work as you expected!

How are you evaluated?

- (**20%**) Implementation that **compiles, runs**, and can **produce a correct video output** (remember c63pred).
- (**20%**) **Effective use of architecture**: Correct and efficient use of memory on GPU. Efficient data movement between CPU/GPU. Investigate potential of half-precision on GPU.
- (**30%**) **Parallelization potential:** Motion Estimation, DCT/iDCT and Motion Compensation.
- (**20%**) **Report**. Profiling between different steps. What worked, what didn't work and why. Quality of code.
- (**10%**) **Performance** of the solution.
- Bonus points can also be given for non-obvious optimizations.

Formal Information

- **Deadlines:**
  - Design: **Thursday March 21th – 12:00**
  - Code: **Friday April 12th – 23:59**
  - Report: **Tuesday April 16th – 14:00**

- The assignment will be graded and count 33% of the final grade.

- Deliver your design and poster (as .pdf): in5050@ifi.uio.no
- Deliver your code and report : https://devilry.ifi.uio.no/

- Prepare a poster (PDF or PPT) and a short talk (5 minutes) for the class on (April 12th).
  - Best poster & presentation will be awarded!

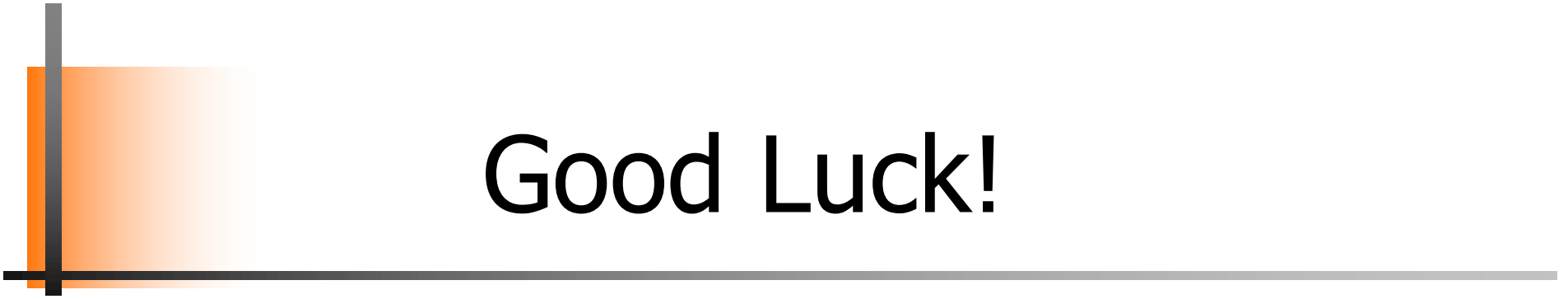
Last but not least!

- Codec63 precode available for download in git. Clone the repository and work on you own local version.

  git clone https://github.com/griwodz/in5050-codec63.git

- Use the CUDA branch.
- Bugs can be reported in Mattermost, or e-mail.

- Use the Mattermost or unofficial Discord to discuss and share tips, tricks and troubles!
  - Everything is allowed, except sharing of code! ☺


Good Luck!

Bonus: Nvidia GTC 2024

- Nvidia's GPU Technology Conference 2024 is this year a hybrid event

  https://www.nvidia.com/gtc/

- Sign up for free, register for events and participate.
- Several recommended talks for IN5050, on CUDA development, debugging and performance analysis.
  - See Mattermost & Discord for more details