EEE4120F 2021

Your Own Digital Accelerator (YODA)

Project (Revised)

Introduction

The YODA project has been reworked. There are now two main submissions, the first of which has been completed:

Milestone 1: Proposal Blog (20%)	28 May
Milestone 5a: Final Report Hand in (80%)	18 June by 23h55 on Vula

Milestones 2 – 4 and 5b (code submission) are optional, these have been kept for your groups to assess their progression on the project.

The project has been reworked to foreground the most important aspects of the project and applying of essential aspects from the theory aspects covered in lecture slides. A rubric for the marking of the final report will be given – the report is the main submission counting 80% of the project mark.

Reworking of Project

The focus of the course project is to experience core aspects of the design process and insight into problem-solving approaches in the development of a high-performance embedded system (HPES). As per the (standard) general definition of an embedded system, which is a task-specific computer system built into a larger system for the purpose of monitoring and control of that system, a HPES likewise has the same main purpose. Except that a HPES also incorporates aspects more commonly associated with High Performance Computing (HPC). Both these technologies, and increasingly the combination of these within specialized solutions, are becoming progressively more used in state-of-the-art industry solutions, as the world of work more towards more Industry 4.0 practices.

Accordingly, and to select a project that can be designed around a small and manageable scope, the course project has focused on consideration and activities surrounding the design and testing of a prototyped digital accelerator. The digital accelerator typically encompassing core aspects of embedded system design, data transfer between a host and accelerator, and parallel processing design thinking and exploration.

Specific activities to be performed, and the structure of the final report is provided on the next page.

Report Structure

Considering that there is little time remaining on the course, and the high workload that students have been experiencing in these past weeks, and that MS-1 is only just completed, the project has been refocused towards the essential parts and a fully functional prototype, which can read data and produce results, will not be expected. Instead, the project teams are expected to report on core pieces of the design process, in particular:

- The project description, as a Blog post, providing information as requested earlier in descripting this step (and illustrated by the blog posting provided by the lecturer).
- A high-level design of the system (presented in the report, possibly earlier version uploaded to **YODA MS-2: Conceptual design review** – (the conceptual design review upload is optional, only if you want your design suggestion to be reviewed). The design should show how your digital accelerator will connect to a host, and to other peripherals if needed. It should discuss where computing will be performance, data transfer considerations, how to display results (or how user can confirm it is working), how/where performance measurement would connect (not necessarily explaining in detail how performance measurement and other assessment of the system would be done).
- A golden measure to demonstrate understanding of the project and expected solution. Consideration of how the solution produced by the golden measure might be compared to that produced by the main system (*but note*: not needing any complete implementation of the system proposed, more discussing if that if were to be build how would you compare results / accuracy of the golden measure to the HPES system).
- Select a small non-trivial piece of the HPES to do a small, '*focused investigation*' to gain insight into how well or relevant the proposed solution might. NOTE: this does not even need to implement the main algorithm you are dealing with. This can just be a small test to find out something important to assess the likelihood of the HPES being feasible. For example, if you design were to need 10 multiplers working in parallel and they each need to send data to their own register, you could plan this 'focused investigation' on implementing a multiplier in Vivado, connecting them up to registers and generate a report (may need to make it do something, like multiple a source register to constants 1 – 10, and save in an output register to force Vivado to do a simulation and provide timing, otherwise if you don't have any data being worked on it might optimize out the design and do nothing). This should be something you could do in around 4h. Using the Vivado report on the design (number LUTs etc use) as information to put in your report.
- The report then needs to briefly discuss each of the 8 stages of designing parallel programs, in relation to the proposed HPES system. But not this discussion is not based on a trial implementation of the system, it is based on insight you gain from the *golden measure*, the *focused investigation*, and thinking about what activities would likely be carried our for these steps. (Of course there can be overlap with you answer to first point, the concept design review). You can essentially prepare the report for this project along those steps of the design process, i.e. :
    1. start by explaining the problem (you can reuse what you have in your Blog) and the golden measure (which is essentially helping to explain the core processing the system does);
    2. discuss partitioning, if the solution will be split up in some way, e.g. separating in tasks, maybe one task should happen on the host and another on the accelerator;
    3. discuss the granularity of your data (to develop a result or do a partial computation how much data do you need to access)
    4. communication issues (e.g. are you going to break the data into windows of data, send one windows to the accelerator and get some sort of result back).

5. Identifying data dependencies (can you have multiple chunks of data worked on in parallel or is it going to be more pipelined, can it have parallel pipelines, etc).
6. synchronization (needs, e.g. the host and accelerator might need some synchronization or to external devices, etc).
7. load balancing (if you deploy two kernels to a GPU, do both kernels do a similar amount of work, can they do that simultaneously, are there sometimes lengthy delays while one kernel is waiting on the other?)
8. performance analysis (you can talk about big O of the algorithm, do timing of the golden measure to get a practical sense of the time taken, and then work out, pen on paper perhaps, how good the proposed parallel version might be – but note again that you do not need to implement an actual prototype to do any testing, it can all be done more analytically based on specifying expected functionality characteristics, e.g. if you are doing 10 parallel multiplies on an FPGA compared to a sequence of 10 multiplies on a CPU, maybe the performance improvement on the accelerator would be 6x, due to multiply running a little slower on the FPGA, and time to transfer data back/forth, compared to just doing it on the CPU)

Conclusions and reminder of where to focus effort

In conclusion, it is hope that would reduce a significant amount of the workload. Groups that are willing to attempt an actual implementation, or rather a more substantive 'focused investigation' are welcome to go ahead with that if they have time, but it might not be able to get much additional marks, perhaps getting 100% for the subsection trial and performance testing aspect for the effort and dedication towards doing a thorough investigation.

Note that in this revised plan you are not expected to implement much. The only implementation involves two things: 1) development of a golden measure to help explain the main algorithm involved in your project and to get a baseline for its performance, and 2) performing the 'focused investigation' where you identify just a small piece of the solution and do some analysis (e.g. timing if using OpenCL or LE utilization and/or propagation delay estimates if using FPGA / Verilog).

If you require further clarity on what is involved in this new version of the project please post a query on the 'YODA Project" Vula Forum, or attend the Tuesday 1 June 3pm Zoom session to pose your questions and seek further clarity on this issue.