THESIS SUPERVISOR'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Computation scheduling in neural network inference on embedded hardware**
Author's name:	**Eldar Iosip**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Dept. of Computer Science
Thesis supervisor:	Michal Sojka
Supervisor's department:	CIIRC, IID

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

I consider the assignment as challenging because I have only a little experience with neural networks and modern ML frameworks so I could not guide the student strongly. It was the purpose of the thesis to explore this area and help me with applying our GPU scheduling algorithms to it. As ML frameworks are being rapidly developed, it is challenging to work with the latest versions due to many dependencies needed to run them.

Fulfilment of assignment **fulfilled with major objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

Unfortunately, the stated goal of applying our GPU scheduling algorithms to neural network inference was not fulfilled. However, this work has paved the road for doing this in the future. In order to extend the framework, it was necessary to find which pieces of the code to modify. I expected the student the associate the nodes in the data-flow graphs with the code implementing the computations. This information is not included in the thesis, but based on the provided description of TensorFlow internals, now it is much easier for me to find it myself. The student needed more time than expected to understand the TensorFlow architecture, and very little time was left to work on the follow-up tasks.

Activity and independence when creating final thesis **C - good.**

Assess whether the student had a positive approach, whether the time limits were met, whether the conception was regularly consulted and whether the student was well prepared for the consultations. Assess the student's ability to work independently.

The student worked on the thesis steadily since beginning with short exception in the middle. He started with surveying the ML frameworks and then, we have roughly agreed on which frameworks to cover in the thesis. TensorFlow was unsurprisingly one of them and a new version TensorFlow 2 (TF2) was expected to be released soon. Its new features suggested that it would be easier to modify this version instead of the old one. This might be true, but we underestimated the time needed to compile and run and debug this version on our target platforms. I think that the student spent most of its time on this task.

During the work, we had semi-regular meetings and on-line communications. The student tried to work independently, but sometimes, it was counterproductive. For example, he invested a lot of time (and even some money) to compile TF2 on a rented cloud platform, but that platform would be hardly suitable for the planned precise benchmarking. Then, another big effort was spent on compilation on the embedded Tegra X2 platform, where the TF2 compilation was measured in tens of hours, but it would be probably less painful to cross-compile it.

Technical level	**D - satisfactory.**
Is the thesis technically sound? How well did the student employ expertise in his/her field of study? Does the student explain clearly what he/she has done?	
The description of the TensorFlow framework in the theses is good and the final version is significantly better than previous drafts. This is the most valuable part of the thesis despite it lacking the description of the lowest level (GPU kernels) needed for integration with our GPU scheduling. Rather than the lowest level, the thesis describes a higher level – TensorFlow Serving, which is irrelevant for the planned scheduling and just adds overhead caused by the HTTP protocol and data decoding. Given that the student likely spent most of its time with compilation, it is unfortunate that the thesis does not mention this part and the description of compilation commands and configuration is missing even in the attached data. The student provided me with this information only after thesis submission. The benchmarks and their results presented in the thesis are hard, if not impossible, to reproduce, because the thesis provides almost no details about what was benchmarked and how.	

Formal level and language level, scope of thesis	**B - very good.**
Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?	
The thesis is written in good English with just a few typos and grammatical errors. Formatting and typography in at good level thank to the use of recommended LaTeX template. I have minor objections to the size of letters in some graphs in Section 6.	

Selection of sources, citation correctness	**B - very good.**
Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?	
The student was actively finding relevant literature and I'm satisfied with his selection.	

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Although the goals of the thesis were not completely fulfilled, the thesis has a value for me. Its description of the Tensor Flow 2 internals, although not perfect, is a good way to start understand the code. In addition, the instructions for compiling and reproducing author's work are useful – it is unfortunate that they are not a part of the submitted thesis.

My impression from collaboration with the student is that he was used to work at higher levels of software frameworks and this work, which required digging down through the code across multiple programming languages (Python, C++, CUDA), was challenging for him. Working lower in the software stack was author's conscious intention declared during discussion about the thesis assignment. Even though this experience could have been painful for him, I think he has learned many things.

The grade that I award for the thesis is **D - satisfactory.**

Date: **23.1.2020** Signature: