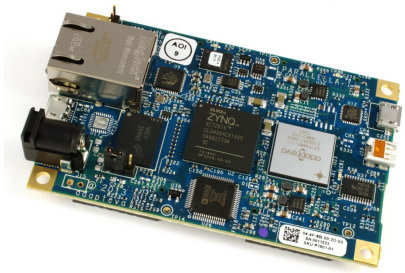
M.Sc. thesis project proposal: Embedded streaming text analytics

Machine learning methods for automated extraction of information in streaming data are important tools that enable new innovations and applications of information technology. Streaming text on the web, and in telecommunication networks in general, is one particular type of data with rich semantic structure, which require efficient methods for information representation in order to enable online storage and analysis using reasonable computing resources. Semantic analysis of such data requires non-trivial operations, which make it necessary for the task to be carried out in a centralised fashion using computer centers, where vast amounts of text is collected and processed.

The need to centralize data is a bottleneck in some potential applications of text analytics, calling for a way to decentralize the system. Therefore, we propose a master's thesis project investigating the possibility to embed algorithms for semantic text representation in resource-efficient hardware that can be distributed in communication networks, for example for deep packet inspection purposes. The proposed project will focus on studying the feasibility to implement the Random Indexing text representation method in a modern system on chip (SoC) with an FPGA, see Figure 1, using high-level synthesis tools. In particular it is interesting to investigate possibilities to improve the efficiency by implementing part of the algorithm on the FPGA.



parallella

Figure 1. Parallella development board with a Xilinx Zynq system on chip including an FPGA and two ARM A9 cores combined with 1GB SDRAM and a 16-core Adapteva coprocessor optimized for parallel computation.

The project work can be carried out either at Gavagai AB in Stockholm, or at EISLAB at the Luleå University of Technology.

Requirements: Background in Computer Science, Engineering Physics or Electrical Engineering. Knowledge of C is necessary. Knowledge of embedded system programming or digital design, and a general interest in mathematics and machine learning is beneficial. Participation in writing of a scientific paper summarising the results of the project is encouraged.

Contact: Magnus Sahlgren (mange@gavagai.se), Fredrik Sandin (Fredrik.Sandin@ltu.se)
Links: www.gavagai.se