ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Μαγειρόπουλος Ευάγγελος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Μ. Κατεβαίνης

Ν. Χρυσός (επιβλέπων)

Πέμπτη, 8 Οκτωβρίου 2020 ,ώρα 09:00 π.μ.

Τηλεδιάσκεψη (μέσω του συστήματος e:Presence), Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Διεύθυνση μετάδοσης (url): http://video.ucnet.uoc.gr/live/show/319

Κανάλι YouTube του Τμήματος
https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

" Υλοποίηση Συνελικτικών Νευρωνικών Δικτύων σε Ομάδες Διασυνδεδεμένων FPGAs Χρησιμοποιώντας το Vivado HLS"

Περίληψη

Τα Συνελικτικά Νευρωνικά Δίκτυα χρησιμοποιούνται ευρέως για να βελτιώσουν την καθημερινή μας εμπειρία με τον κόσμο, κατηγοριοποιώντας αυτόματα ψηφιακά δεδομένα, όπως εικόνες, ηχητικές καταγραφές και βίντεο, βοηθώντας έτσι στις διαδικτυακές αναζητήσεις και την κατανόηση των δεδομένων που είναι διαθέσιμα στον ψηφιακό κόσμο. Σε αυτή την εργασία,

εξερευνούμε τη δυνατότητα να κατανείμουμε ολόκληρα Συνελικτικά Νευρωνικά Δίκτυα σε ομάδες πολλαπλών διασυνδεδεμένων συστημάτων υπολογισμού, όπως ASICs και FPGAs. Επιδιώκουμε να ορίσουμε μια κλιμακούμενη αρχιτεκτονική όπου μια ομάδα από FPGAs (τμήμα του προτύπου HPC που βασίζεται στο ExaNeSt) δουλεύει ταυτόχρονα σε ροές αιτημάτων χρηστών για κατηγοριοποιήσεις. Βασίζουμε την εργασία μας σε προϋπάρχοντα εργαλεία που απλοποιούν την κατανομή διαφόρων δικτύων, όπως το Keras για να ορίσουμε το Συνελικτικό Νευρωνικό Δίκτυο και το hls4ml, ένα εργαλείο που έχει αναπτυχθεί στο CERN, που υλοποιεί ένα νευρωνικό δίκτυο σε RTL χρησιμοποιώντας τη δομή λογισμικού Vivado High Level Synthesis (HLS). Συστήνουμε ένα σύνολο από βελτιστοποιήσεις στον κώδικα και βασιζόμενες σε οδηγίες, ώστε να πετύχουμε επιτάχυνση άνω του 700x σε αυτούσιους υπολογιστικούς πυρήνες, και καταφέρνουμε να κατανείμουμε όλες τις παραμέτρους σε BRAMs των FPGA. Επιπλέον, διαχωρίζουμε και επανασχεδιάζουμε το δίκτυο, ώστε να ελαχιστοποιήσουμε τις μεταφορές δεδομένων και εξισορροπούμε την εργασία στο σύνολο των FPGAs. Εν τέλει, σχεδιάζουμε ιδιόχειρα κομμάτια RTL, στα οποία ενσωματώνουμε οδηγίες του HLS, ώστε να χρησιμοποιήσουμε ένα δίκτυο HPC για επικοινωνία μεταξύ των FPGAs. Η τελική μας υλοποίηση του Συνελικτικού Νευρωνικού Δικτύου SqueezeNet που απαιτεί 800 εκατομμύρια πράξεις ανά εργασία κατηγοριοποίησης, σε 5 FPGAs, προσφέρει διεκπεραιωτική ικανότητα 303 κατηγοριοποιήσεων εικόνων ανά δευτερόλεπτο και συνολική καθυστέρηση κατηγοριοποίησης 24 χιλιοστά του δευτερολέπτου, μια τάξη μεγέθους μικρότερη από τυπικές Συμφωνίες Επιπέδου Υπηρεσίας.

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Mageiropoulos Evagelos

Master's Thesis Supervisor: Professor M. Katevenis

N. Chrysos (Thesis Co- Advisor)

Thursday, 8 October 2020, 09:00 a.m

Teleconference (will use the e: Presence system), Computer Science Department, University of Crete

(url) : http://video.ucnet.uoc.gr/live/show/319

YouTube channel :

https://www.youtube.com/channel/UC7uE3QiMTQjkrpByB_Gnt6Q/live

“Implementing Convolutional Neural Networks in a Cluster of Interconnected FPGAs Using Vivado HLS”

Abstract

Convolutional Neural Networks (CNNs) are extensively used to augment our everyday experience of the world by automatically labeling and categorizing digital data, such as images, voice records, and video, thus helping in web search and in comprehension of data available in the digitaql world. In this thesis, we explore the possibility to map complete CNNs in clusters of multiple interconnected computing devices, such as ASICs or FPGAs. We seek to define a scalable architecture where a cluster of FPGAs (a segment of the ExaNeSt-based HPC prototype) works concurrently on user streams of inference requests. We base our work on existing tools that simplify the mapping of arbitrary networks, such as using Keras to define the Convolutional Neural Network and hls4ml, a tool developed at CERN, that implements a convolutional neural network into RTL using the Vivado High Level Synthesis (HLS) framework. We introduce a number of code and directive-based optimizations in order to achieve speedups in excess of 700x for the individual kernels, and manage to map all parameters inside FPGA BRAMs. Furthermore, we split and redesign the network in order to minimize the data transfers and balance the work across FPGAs. Finally, we design custom RTL blocks which we integrate with HLS directives in order to use an HPC network for inter-FPGA communication. Our final implementation of the SqueezeNet CNN network which needs 800 million operations per inference task, in 5 FPGAs, offers a throughput of 303 image classifications per second (CPS), and a total inference latency of 24 ms, one order of magnitude smaller than typical user Service Level Aggreements (SLAs).