ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Κολοκάσης Ιάκωβος
Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Επικ. Καθηγητής, Π. Πρατικάκης & Α. Μπίλας

Παρασκευή , 30 Οκτωβρίου 2020 ,ώρα 10:00 π.μ.

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

Διεύθυνση μετάδοσης (url): http://video.ucnet.uoc.gr/live/show/327

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

" Tera Cache: Αποτελεσματική αποθήκευση ενδιάμεσων δεδομένων στο SPARK σε συσκευές γρήγορης αποθήκευσης "

Περίληψη

Οι εφαρμογές που εκτελούν αλγόριθμους μηχανικής μάθησης αποτελούνται από μεγάλος πλήθος επαναληπτικών υπολογισμών επεξεργασίας δεδομένων που εκτελούνται μέχρι να ικανοποιήσουν μια συνθήκη σύγκλισης. Για να εκτελεί τους υπολογισμούς μηχανικής μάθησης σε μικρό χρόνο εκτέλεσης συμβαδίζοντας παράλληλα με την εκθετική αύξηση του μεγέθους των δεδομένων καθώς και την αργή αύξηση της

κλιμακοσιμότητητας της μνήμης τυχαίας προσπέλασης (DRAM), το Spark χρησιμοποιεί γρήγορες συσκευές αποθήκευσης για την προσωρινή αποθήκευση των ενδιάμεσων αποτελεσμάτων εκτός της μνήμης. Ωστόσο, η προσωρινή αποθήκευση εκτός της μνήμης απαιτεί τη σειριοποίηση και την αποσειριοποίηση (serdes) των δεδομένων, το οποίο προσθέτει σημαντική επιβάρυνση στο χρόνο εκτέλεσης ειδικά όσο αυξάνεται το συνολικό μέγεθος των δεδομένων επεξεργασίας.

Αυτή η διατριβή προτείνει το μηχανισμό TeraCache, μια επέκταση της προσωρινής μνήμης αποθήκευσης ενδιάμεσων δεδομένων του συστήματος ανάλυσης δεδομένων Spark που αποφεύγει την ανάγκη για σειριοποίηση/αποσειριοποίηση διατηρώντας όλα τα αποθηκευμένα δεδομένα στο σωρό (heap) αλλά εκτός μνήμης, χρησιμοποιώντας χαρτογραφημένη μνήμη εισόδου/εξόδου (mmio). Για να επιτευχθεί αυτό, η TeraCache επεκτείνει το σωρό της JAVA εικονικής μηχανής (JVM) με έναν διαχειριζόμενο σωρό που βρίσκεται σε μια γρήγορη χαρτογραφημένη στη μνήμη συσκευή αποθήκευσης και χρησιμοποιείται αποκλειστικά για την αποθήκευση ενδιάμεσων αποτελεσμάτων. Προκαταρκτικά αποτελέσματα δείχνουν ότι η πρωτότυπη υλοποίηση της TeraCache μπορεί να επιταχύνει τα εφαρμογές μηχανικής μάθησης που αποθηκεύουν ενδιάμεσα αποτελέσματα έως και 37% σε σύγκριση με τις υφιστάμενες μεθόδους αποθήκευσης.

University of Crete

Computer Science Department

M.Sc. Thesis presentation / examination

Kolokasis Iakovos

Master's Thesis Supervisor: Assistant Professor, P. Pratikakis & Prof. Angelos Bilas

Friday, 30 October 2020, 10:00 a.m

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

(url) : http://video.ucnet.uoc.gr/live/show/327

YouTube channel :

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

“TeraCache: Efficient Spark Caching Over Fast Storage Devices”

Abstract

Many analytics computations are dominated by iterative processing stages, executed until a convergence condition is met. To accelerate such workloads while keeping up with the exponential growth of data and the slow scaling of DRAM capacity, Spark employs off-heap caching of intermediate results. However, off-heap caching requires serialization and deserialization (serdes) of data, that add significant overhead especially with growing datasets.

This thesis proposes TeraCache, an extension of the Spark data cache that avoids the need of serdes by keeping all cached data on-heap but off-memory, using memory-mapped I/O (mmio). To achieve this, TeraCache extends the original JVM heap with a managed heap that resides on a memory-mapped fast storage device and is exclusively used for cached data. Preliminary results show that the TeraCache prototype can speed up Machine Learning (ML) workloads that cache intermediate results by up to 37% compared to the state-of-the-art serdes approach.