A Cloud-Scale Acceleration Architecture

- **Abstract**
- Introduction
- Hardware Architecture
- Local Acceleration
- Network Acceleration

ABSTRACT

Σήμερα οι hyperscale datacenter providers (π.χ. Google, Facebook, Amazon) προσπαθούν μεσα απο τα οικονομικά ωφέλη της ομοιγένειας (homogeneity) να εξισορροπίσουν την αυξανόμενη ανάγκη για εξειδικευμένο υλικό(specialized hardware).Την λύση έρχεται να δώσει η Microsoft όπου προτείνει μια νέα cloud αρχιτεκτονική , την λεγόμενη Cloud Configuration Architecture , όπου τοποθετεί αναμεσα στουσ διακόπτες δικτύου(network switches) και τους servers ένα επίπεδο με επαναρυθμίσημη λογική(reconfigurable logic-FPGAs).Με αυτό το τρόπο μπορεί να προγραμματίσει την ροή του δικτύου(network flow) , επιτρέποντας έτσι την επιτάχυνση τοπικών εφαρμογών που τρέχουν σε servers(service acceleration) αλλά και την συλλογή FPGAs που δεν χρησημοποιηούνται απο τους τοπικούς servers (network accelaration).

- Abstract
- **Introduction**
- Hardware Architecture
- Local Acceleration
- Network Acceleration

INTRODUCTION

- Πολλόι σύχρονοι hyperscale datacenters έχουν βελτιωθοί σε πολλού τομείς διατηρώντας όμως την βασική δομή που υπάρχει εδω και χρόνια(servers με CPUs, DRAM, and τοπική αποθήκευση,ενώνονται μεταξύ τους με τις NIC (Network Interface Card) μεσω των διακοπτών Ethernet (Ethernet witches).

- Αυξάνοντας την ομοιγένεια έχουμε σημαντικά ωφέλη όπως του ότι ο φόρτος εργασίας μοιράζεται σε όλη την υποδομή και η διαχείρηση γίνεται πιο απλή(μειώνει κόστος και configuration errors).

- Μπορούμε όμως να εξισσοροπήσουμε αυτή την επιθυμία για ομοιγένεια με την τοποθέτηση ειδικών επιταχυντών(specialized accelerators-FPGAs) σε μέρη της υποδομής.

- Ενας ειδικός επιταχυντής πρέπει να είναι προγραμματήσιμος έτσι ώστε κατα την δίαρκεια(μπορεί και χρόνια) που τον αναπτύσουμε να μπορεί να ανταπεξέλθει σε διαφορετικού φόρτους εργασίας που του ανατήθονται.

INTRODUCTION

Medium-scale FPGA deployment

(48 FPGAs(directly-connected) connected by a secondary network)

Περιορισμοί:

- Το δευτερεύον δίκτυο χρειάζεται περίπλοκη και ακριβή καλωδίωση καθώς και γνώση της φυσικής τοποθεσίας της μηχανής.
- Περιπλοκη επαναδρομολόγηση σε περίπτωση αποτυχίας(πέφτει απόδοση και χάνονται κόμποι του δικτύου)
- Ο αριθμός των FPGAs που μπορούν να επικοινωνήσουν απευθείας μεταξύ τους (χωρίσ χηση λογισμοικού) είναι περιορισμένα.
- Δεν προσφέρει τόσο στην δικτύωση και αποθήκευση ροών δεδομένων.

INTRODUCTION

Cloud-scale FPGA deployments

Configurable Cloud

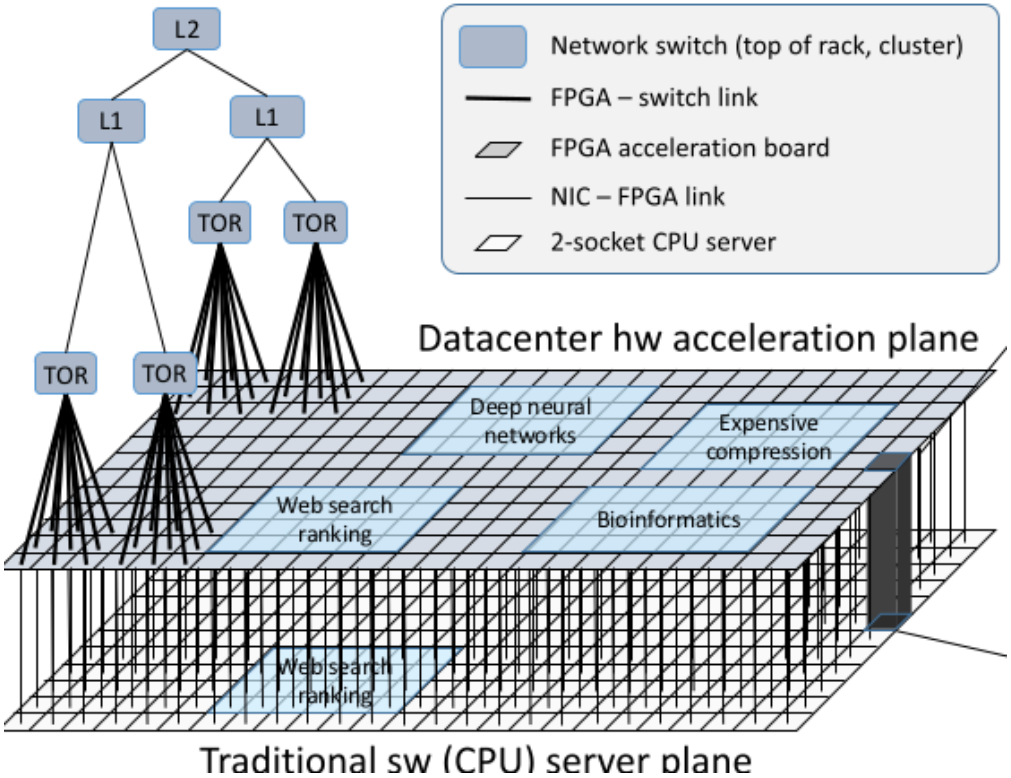
Βασικη διαφορά με τη προήγουμενη ανάπτυξη είναι οτι τα FPGAs είναι απευθείας συζευγμένα με τους servers.

Η τοπολογία των διακοπτών του δικτύου είναι τέτοια έτσι ώστε:

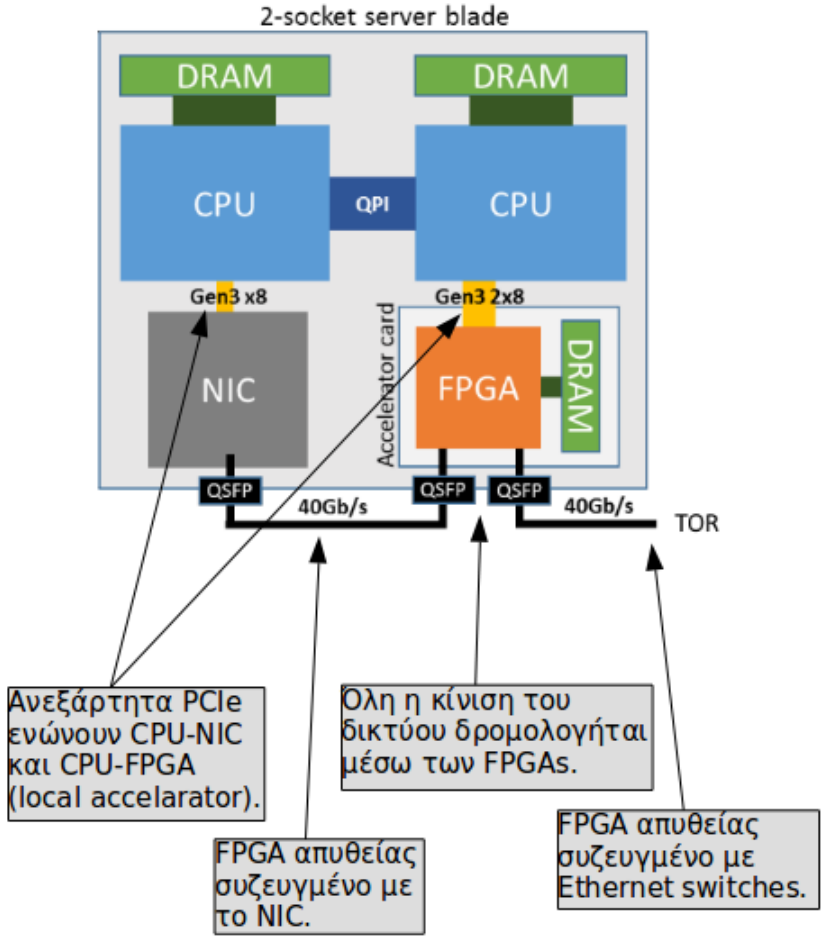
- Να μην υπάρχει επίδραση σε γειτονικό server
- Να μην χρειάζεται περιπλοκη καλωδίωση
- Να μην είναι απαραίτητο να ξέρουμε τη φυσική τοποθεσία της μηχανής.

INTRODUCTION

(a)Decoupled Programmable Hardware Plane



(b) Server + FPGA schematic.



INTRODUCTION

FPGAs Global pool

- Τα FPGAs παράγουν και καταναλώνουν δικα τους networking packets ανεξάρτητα απο το host server.
- Κάθε FPGA στο datacenter μπορει να αποκτήσει networking packets απο τα υπολοιπα FPGAs(της κλίμακας 100-1000) σε πολλή λίγα ms χωρίς την παρέμβαση του λογισμικού.
- Servers που βρίσκονται στο datacenter και δεν χρησιμοποιούν το FPGA τους,μπορούν να το παραχωρήσουν σε ένα **global pool** όπου θα το αιτηθεί ένας άλλος server που τρέχει μια μεγάλη υπηρεσία(π.χ. machine learning).

- Abstract
- Introduction
- **Hardware Architecture**
- Local Acceleration
- Network Acceleration

HARDWARE ARCHITECTURE

Περιορισμοί Σχεδίασης

- Η ανάγκη για ομοιγένεια και ειδίκευση υλικού οδήγησε στην ανάπτυξη προγραμματισμένων επιταχυντών(programmable accelerators) όπως είναι τα FPGAs και οι GPUs

- Οι προγραμματισμένοι επιταχυντές πρέπει να είναι ευέλικτοι έτσι ώστε να μπορούν να καλύψουν τα πιο κάτω σενάρια :

    - Τοπική επιτάχυνση **(local compute accelaration)**
      Κάθε server να μπορεί να χειριστεί σενάρι όπως search ranking μονο με το δίκο του FPGA.
    - Δικτυακή επιτάχυνση**(network accelaration)**
      Λόγω του ότι έχουμε ποικιλομορφία στους χρήστες δεν μας συμφέρει οικονομικά να έχουμε μόνο τοπικους επιταχυντές.
    - Επιτάχυνση εφαρμογών(**Global Application accelaration**)
      Servers παραχωρούν τα αχρησημοποιήτα FPGAs στο global pool για να τα πάρουν αυτοί που το αιτήθηκαν.

HARDWARE ARCHITECTURE

Φυσικοί Περιορισμοί

- **Περιορισμένη Ισχύ**
- **Μικρο φυσικό μέγεθος**
- **Αντοχή σε υψηλές θερμοκρασίες**

256 MBFlash memory (κρατά το golden image)

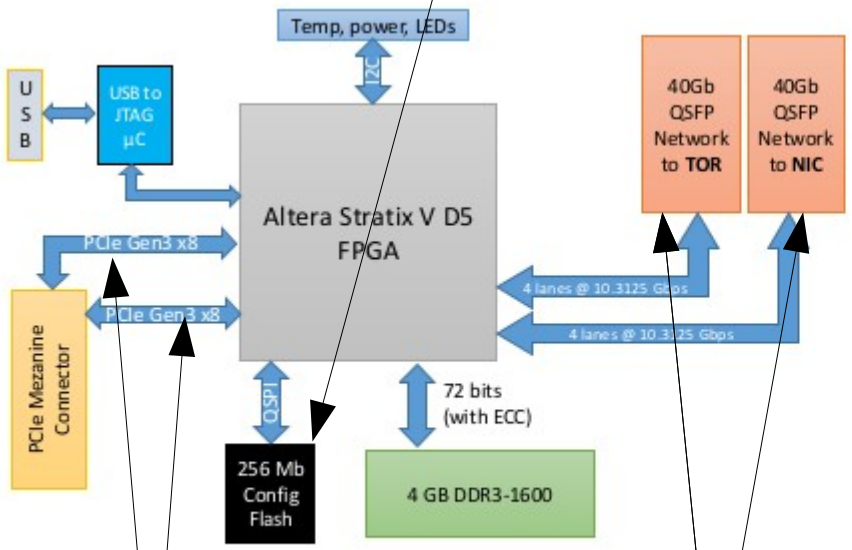


Fig. 2. Block diagram of the major components of the accelerator board.

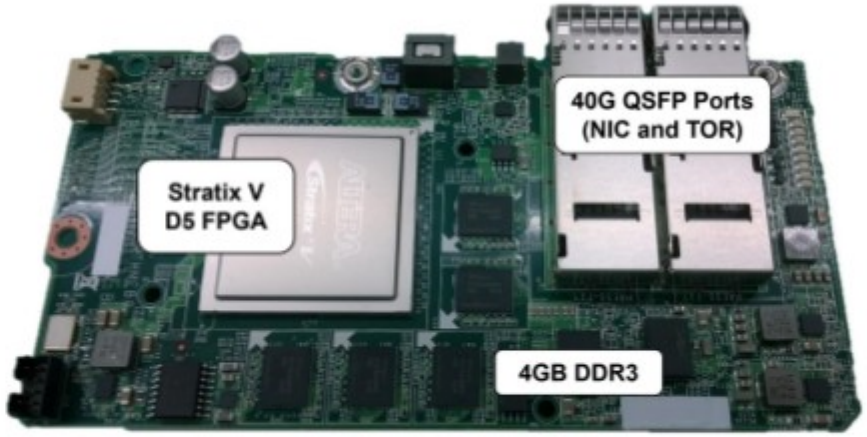


Fig. 3. Photograph of the manufactured board. The DDR channel is implemented using discrete components. PCIe connectivity goes through a mezzanine connector on the bottom side of the board (not shown).

2 x PCIe (ανεξάρτητοι) (16GB/s aggregate to CPU)

2 x Ethernet interfaces (ανεξάρτητα) (40GB) με QSFP+ connectors

HARDWARE ARCHITECTURE

Shell Architecture

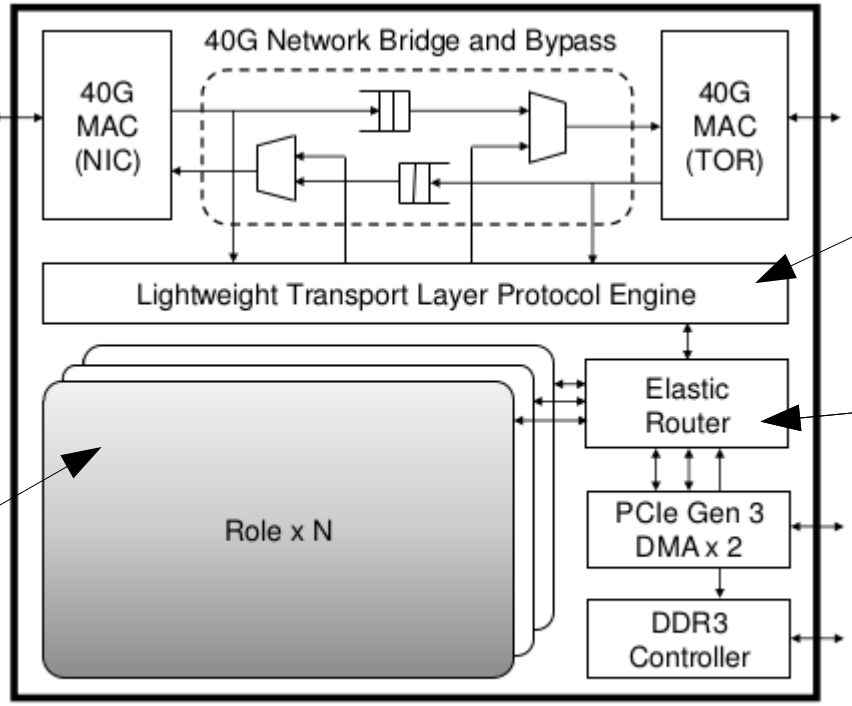


Fig. 4. The Shell Architecture in a Single FPGA.

LTL: ελέγχει την επικοινωνία μέσα στο FPGA.

Elastic router: είναι ένα message router που μέσω των virtual channels επιτρέπει σε πολλαπλά roles να μπούν στο δίκτυο.

Role : η λογική της εφαρμογής

HARDWARE ARCHITECTURE

Datacenter Deployment

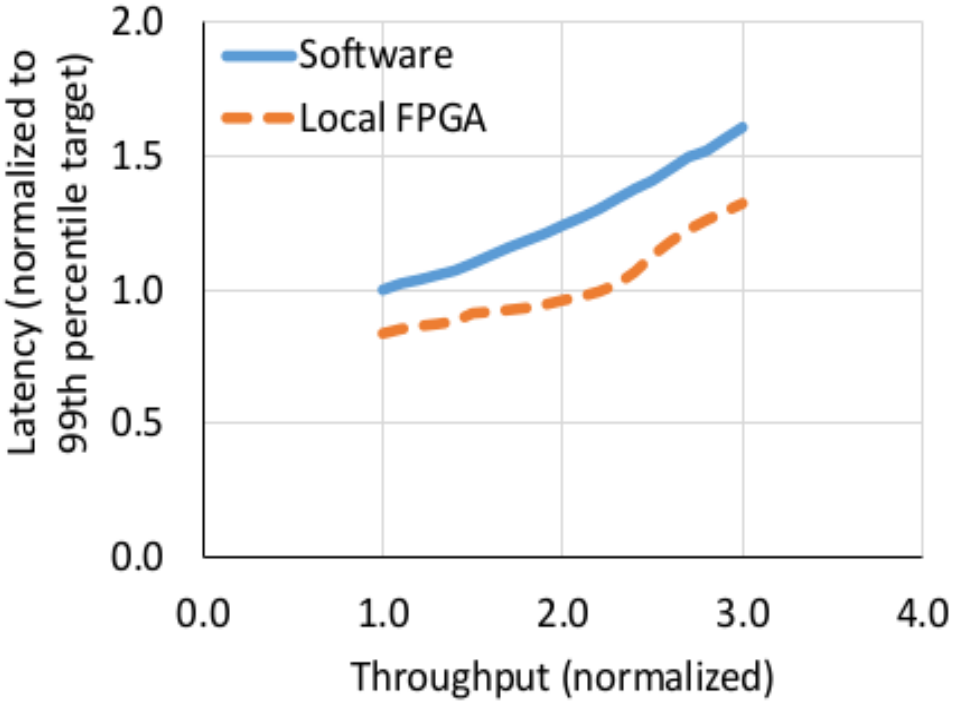
- Σε μεγάλα datacenters (5,760 servers) με accelerators (FPGAs),έγιναν δοκιμές για την αντοχή τους σε μεγάλο φόρτο εργασίας και ακραίες συνθήκες έδειξαν ό.Οι servers πέρασαν την δοκιμασία και εγκρίθηκαν για χρήση σε datacenters.

- Σε μηχανές αναζήτης (3,081 servers) με accelerators (FPGAs), έγιναν δοκιμές για την χρήση τους ως local accelerators, παρατηρώντας ελάχιστα αλλά αποδεχτά σφάλματα.

- Abstract
- Introduction
- Hardware Architecture
- **Local Acceleration**
- Network Acceleration

LOCAL ACCELERATION

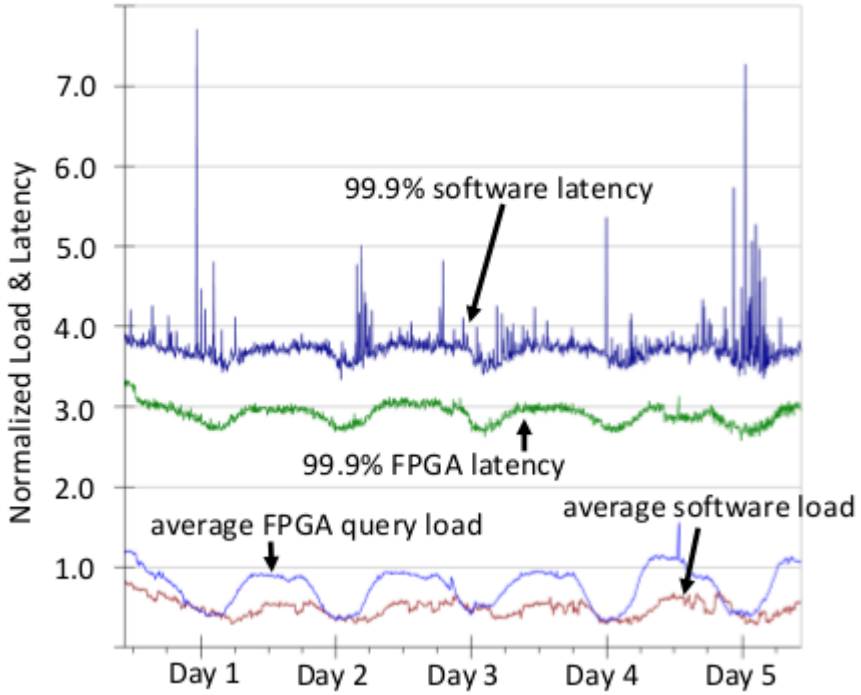
Σε μηχανή αναζήτησης ,με τη χρήση του single-box test έγινε μέτρηση του query latency με το throughput, για 200,000 queries που είχαν διαφορετική ώρα άφιξης.Στην πιο κάτω γραφική βλέπουμε το 99% των πιο αργών queries που μετρήθηκαν.



LOCAL ACCELERATION

(συνέχεια...)

Πιο κάτω βλέπουμε την αποδοση ενός ranking service σε δυο production datactenters σε περίοδο 5 ημερών.Ο ένας datacenter έχει FPGAs και άλλος όχι.



LOCAL ACCELERATION

(συνέχεια...)

Πιο κάτω βλέπουμε το φόρτο εργασίας με το latency την ίδια ημερα (5η) στους δυο προηγούμενους datacenters.

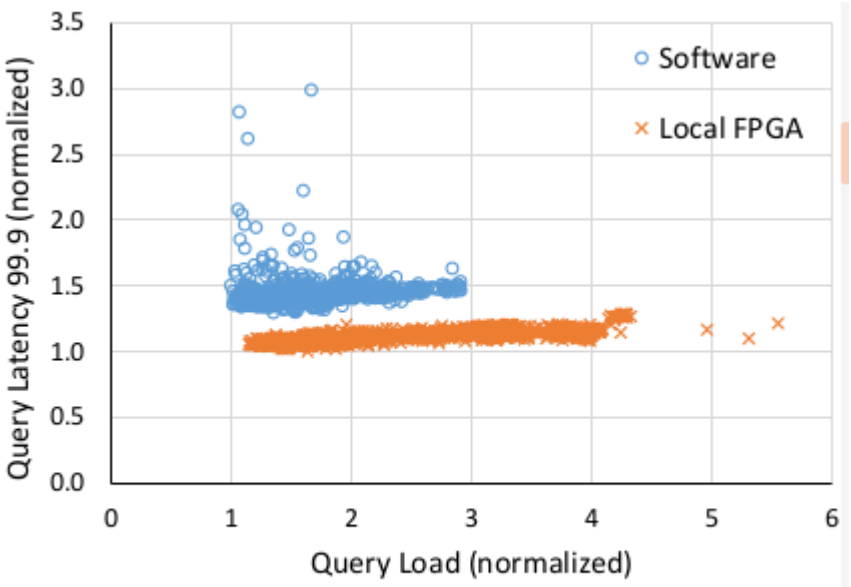


Fig. 8. Query 99.9% Latency vs. Offered Load.

LOCAL ACCELERATION

(συνέχεια...)

Συμπέρασμα:

- Όπως βλέπουμε στις δυο προηγούμενες γραφικές, το FPGA μπορεί να επεξεργαστεί αιτήματα σε χαμηλά latencies.

- Μπορεί να απορροφήσει διπλάσιο φόρτο εργασίας εκτελώντας τα queries σε latency που δεν υπάρχει για κανένα load στο server που χρησιμοποιεί software.

- Abstract
- Introduction
- Hardware Architecture
- Local Acceleration
- **Network Acceleration**

NETWORK ACCELERATION

- Κάθε πακέτο ελέγχεται κατά το πέρασμά του από το NIC (Network Interface Card) στην FPGA και τέλος στο ToR (top of rack switch) για το αν είναι encrypted από software, και αν είναι, τότε το encryption key διαβάζεται από την SRAM ή την DRAM της FPGA για να γίνει το decryption. Έτσι δεν έχουμε load στην CPU για το encryption-decryption.

- Ανάλογα το clock frequency, μπορεί να υπάρξει η ανάγκη για hash functions που βοηθούν στο encryption-decryption (2.4 GHz με 40 GB/s encryption/decryption απαιτούν 5 πυρήνες).

REMOTE ACCELERATION

Για την υποστήριξη υπηρεσιών που χρειάζονται περισσότερες των 1 FPGA, η επικοινωνία μεταξύ των FPGA είναι απαραίτητη.

Έχοντας τις FPGA να επικοινωνούν απευθείας μέσω του Ethernet του datacenter επιτυγχάνεται low latency και δυνατότητα επίλυσης μεγάλων προβλημάτων.

Απαιτήσεις καναλιών :

- Τα κανάλια επικοινωνίας δεν μπορούν να περάσουν μέσω software πρωτοκόλλου στην CPU λόγο μεγάλου latency.
- Θα πρέπει να είναι ελαστικά σε αποτυχίες και packet loss, αλλά να κάνουν drop ένα packet σπάνια.
- Δεν πρέπει να χρησιμοποιούν πολλούς πόρους των FPGA.

FPGA Functions

Υπάρχει ανάγκη και για cross – datacenter υλοποίηση και inter – FPGA communication. Οι δύο βασικές λειτουργίες που πρέπει να υλοποιηθούν στην FPGA είναι :

- Inter - FPGA communication engine.

- Intra - FPGA router, που συγχρονίζει το traffic στην FPGA στο δίκτυο.

Lightweight Transport Layer

Το LTL είναι το inter – FPGA protocol. Χρησιμοποιώντας traffic κλάσεις χωρίς απώλειες αποφεύγονται τα packet drops και τα reorders. Σε μείωση των packet drops οδηγούμαστε και από την ταχύτερη απόκριση των FPGAs λόγο της επικοινωνίας τους.

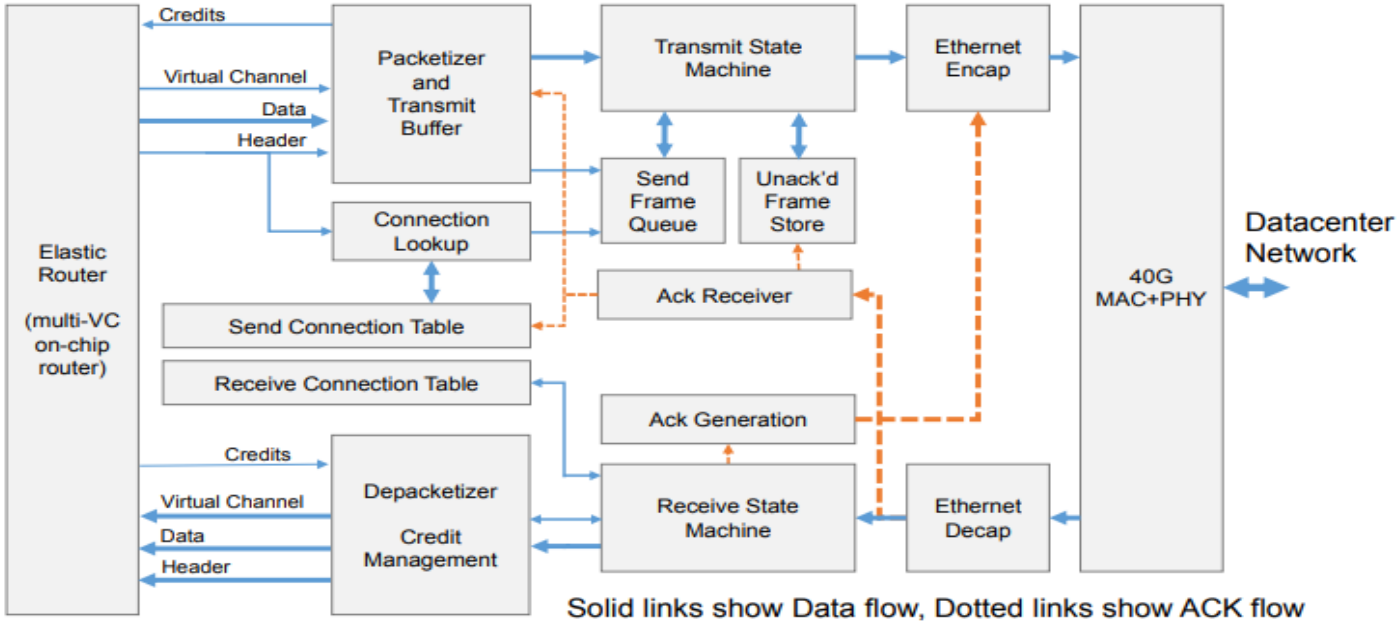


Fig. 9. Block diagram of the LTL Protocol Engine.

Lightweight Transport Layer

Λόγο του static allocation μειώνεται επίσης το latency για το inter – FPGA και για την εσωτερική επικοινωνία μηνυμάτων.

Γίνεται χρήση timeouts τα οποία βοηθούν στην επαναμετάδοση packets και γρήγορη εύρεση προβληματικών nodes.

Η συγκεκριμένη υλοποίηση LTL δίνει την δυνατότητα στις FPGA για ασφαλή είσοδο και διαγραφή packets από το δίκτυο χωρίς να επηρρεάζει την υπάρχουσα ροή.

Γενικά το LTL Protocol Engine υπάρχει με στόχο την υποστήριξη των υπάρχοντων datacenter protocols, χωρίς όμως να τα υποβαθμίζει λειτουργικά.

Elastic Router

Το Elastic Router είναι ένα switch το οποίο τροφοδοτείται με εισόδους και έχει στόχο την υποστήριξη της επικοινωνίας μεταξύ πολλαπλών εξόδων μιας FPGA σε πολλά Virtual Channels. Απευθύνεται στο intra – FPGA communication.

Η σχεδίαση μπορεί να παραμετροποιηθεί για όποιον αριθμό ports χρειάζεται. Ακόμα πολλά ER μπορούν να συμπτιχθούν δημιουργώντας ένα μεγαλύτερο on-chip δίκτυο.

Το ER χρησιμοποιεί credits για τον έλεγχο της ροής πληροφοριών, τα οποία μπορεί να τα ελέγχει δυναμικά για τα Vitrual Channels που τα απαιτούν.

Το LTL και το ER είναι απαραίτητα για την FPGA, καθώς επιτρέπουν την οργάνωσή της σε multi – FPGA services και την απομακρυσμένη χρήση της αν δεν χρησιμοποιείται από το δίκτυό της.

LTL Communication Evaluation

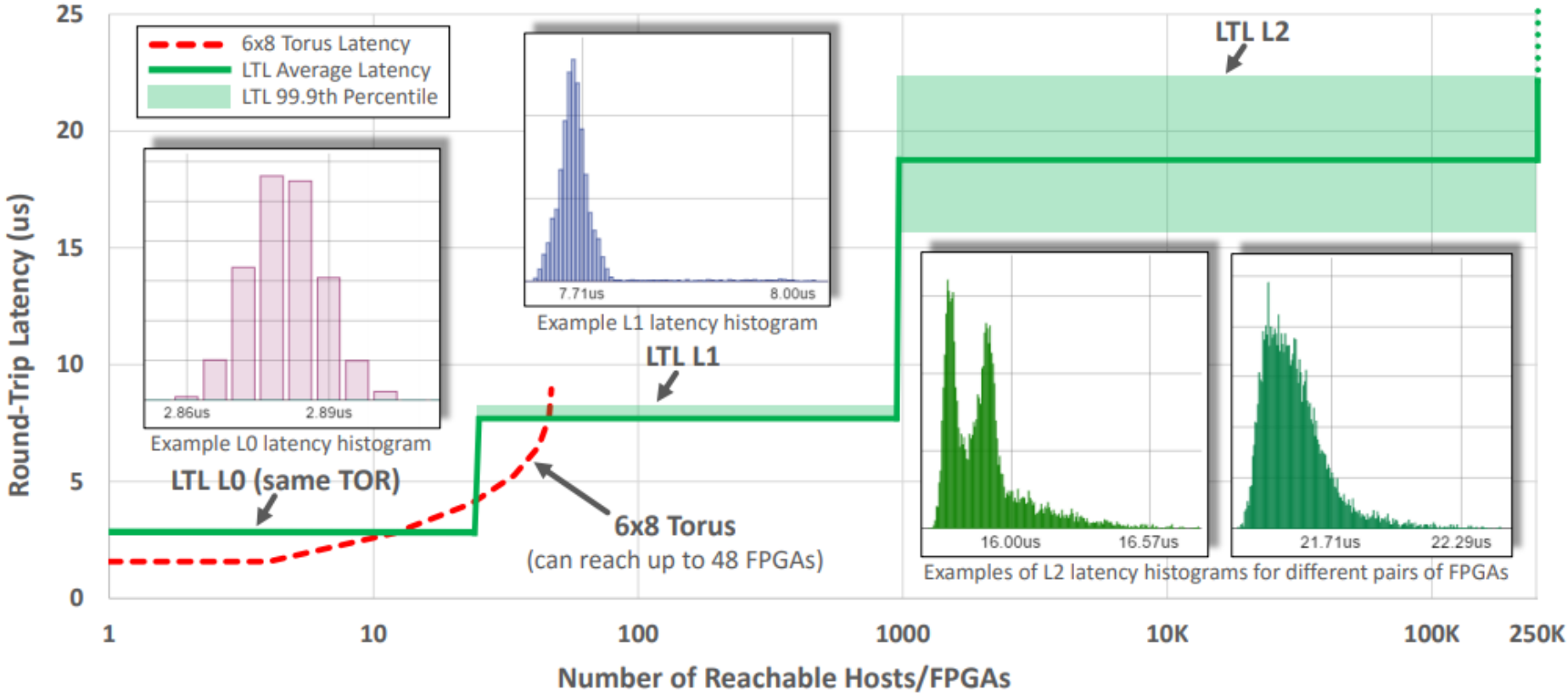


Fig. 10. Round-trip latency of accesses to remote machines with LTL compared to the 6x8 torus from [4]. Shaded areas represent range of latencies up to the 99.9th percentile. LTL enables round-trip access to 100,000+ machines in under 23.5 μs.

Remote Acceleration Evaluation

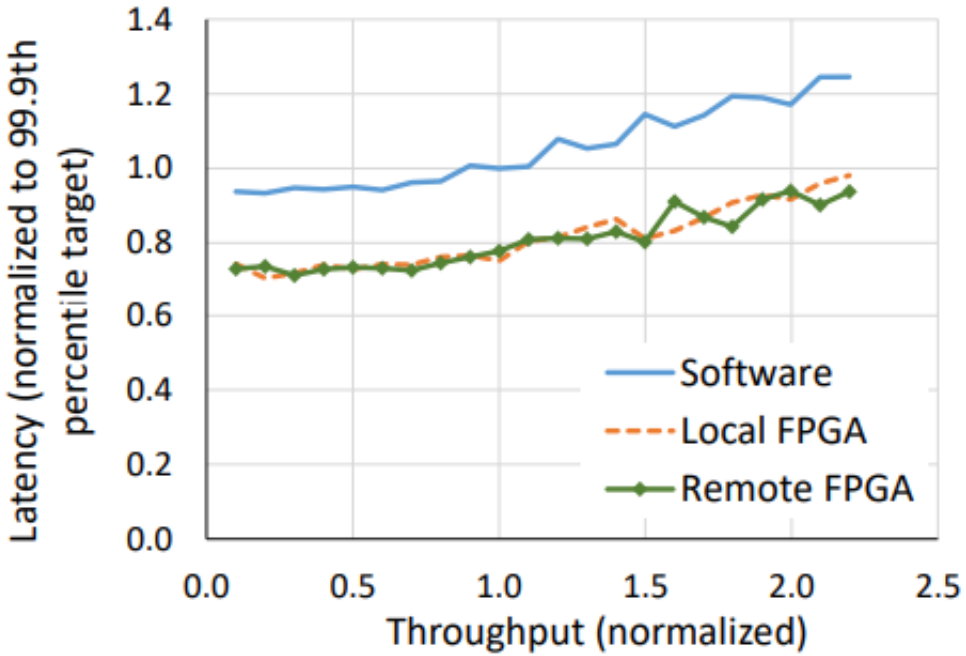


Fig. 11. Latencies of software ranking, locally accelerated ranking, and remotely accelerated ranking. All data are normalized to software 99.9th percentile latency target.

Η επίδραση στον host server βλέπουμε ότι είναι μικρή όταν εξυπηρετεί remote requests, καθώς η FPGA απευθείας αναλαμβάνει τον υπολογιστικό φόρτο, και τον φόρτο του δικτύου.

Ο κεντρικός server δεν βλέπει αύξηση – πίεση της CPU ή την χρησιμοποίηση της μνήμης, παρά μόνο μία μικρή αύξηση στην συνολική κατανάλωση ενέργειας.

Oversubscribing Evaluation

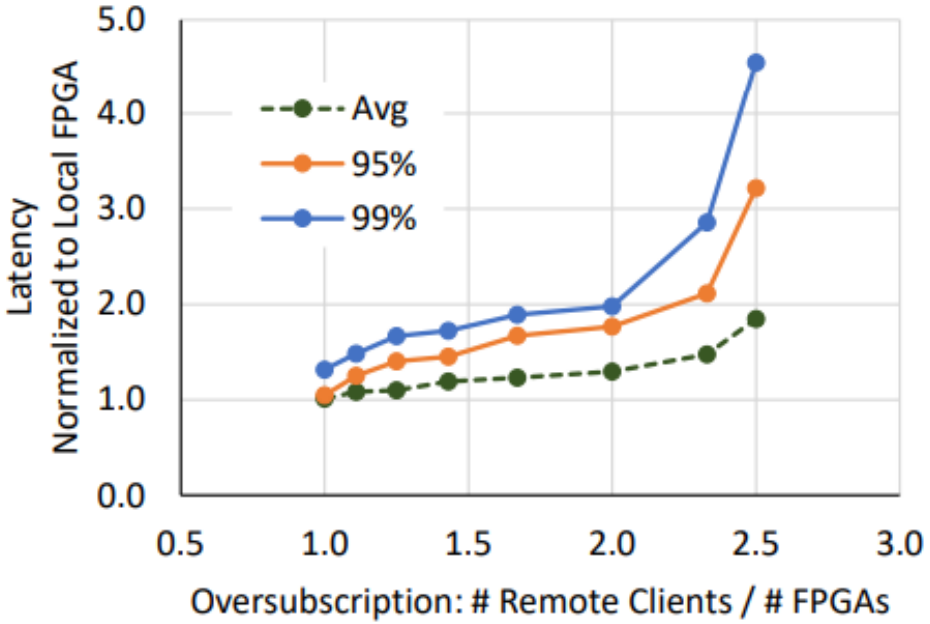


Fig. 12. Average, 95th, and 99th percentile latencies to a remote DNN accelerator (normalized to locally-attached performance in each latency category).

Υπάρχουν services που χρειάζονται περισσότερες FPGA ή που μπορούν να τις εκμεταλλευτούν αν δεν αξιοποιούνται.

Με χρήση latency - sensitive DNN accelerators δοκιμάστηκε η επίδραση του latency από oversubscription.

Οι συγκρούσεις πακέτων και η καθυστέρηση στην ουρά αυξάνεται όσο και το oversubscription αυξάνεται.

Hardware-as-a-Service Model

Μερικά servises που τρέχουν κάτων από το HaaS model.

Ο Resource Manager από κάθε service, εντοπίζει τους πόρους της κάθε FPGA από το datacenter.

Ο lightweight Manager κάθε FPGAs διαχειρίζεται τα calls από τους Resource Managers.

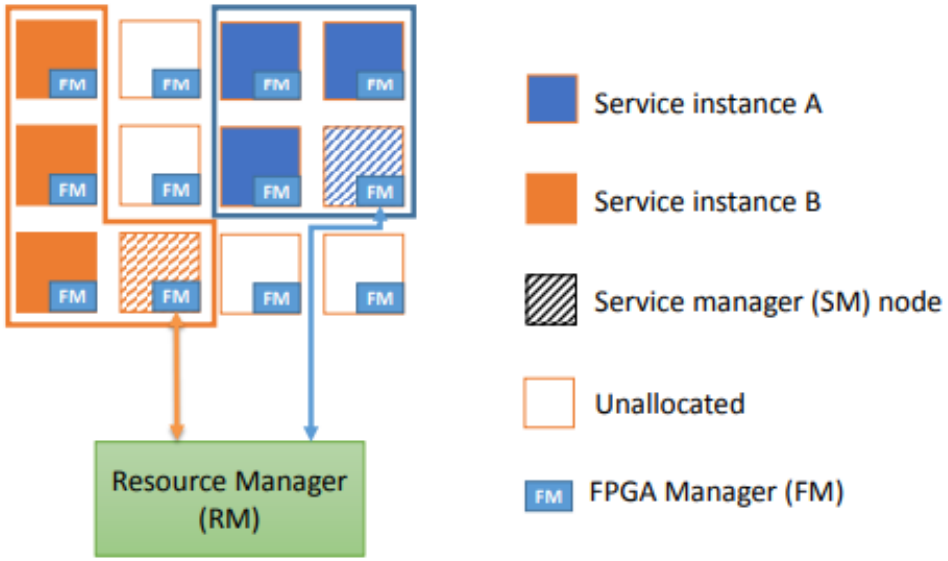


Fig. 13. Two Hardware-as-a-Service (HaaS) enabled hardware accelerators are shown running under HaaS. FPGAs are allocated to each service from Resource Manager's resource pool. Each service has a Service Manager node to administer the service on the allocated resources. Each FPGA has a lightweight FPGA Manager for managing calls from the Service Manager.

Accelerator Incorporation

Τρόποι ενσωμάτωσης accelerator σε μεγάλου μεγέθους συστήματα :

- Διαφορετικοί τύποι accelerator (FPGAs, GPUs, ASICs).
- Τρόποι αλληλεπίδρασης με CPU. Όσο πιο “κοντά” είναι η ενσωμάτωση στη CPU, τόσο πιο κοντά σε fine – grain είναι το πρόβλημα για την διαχείρησή του (I/O level, network level).
- Τρόποι επικοινωνίας accelerators μεταξύ τους (single server, datacenter scale).

Γενικά Συμπεράσματα

Με την συνεχόμενη αύξηση των αναγκών για υποδομές datacenter, η σχεδίαση scalable accelerators είναι ιδιαίτερα σημαντική.

Σχολιάσαμε μία datacenter - scale acceleration αρχικεκτονικήβασισμένη σε FPGAs η οποία είναι και scalable και flexible.

Σε αυτήν την αρχιτεκτονική οι FPGA “μιλούν” απευθείας στο network switch και άρα επικοινωνούν μεταξύ τους χωρίς κάποιο CPU software.

Η αρχιτεκτονική αυτή έχει δοκιμαστεί επιτυχώς σε datacenter, σαν local offload engine, local network acceleration engine και remote acceleration service για web search.

Η Catapult v2 αρχιτεκτονική έχει ήδη εφαρμοστεί στους περισσότερους καινούργιους server στα data centers της Microsoft.

Ευχαριστούμε για το χρόνο σας!