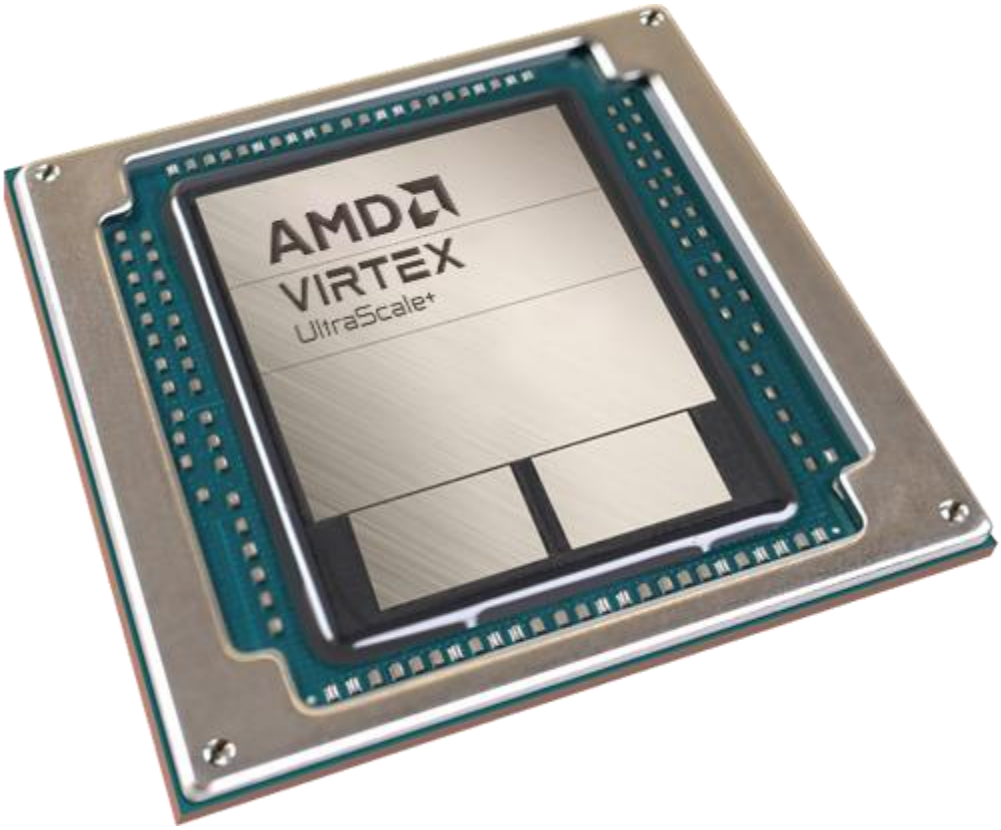
# AMD Deep Learning Processor Unit και Νευρωνικά Δίκτυα Βαθιάς Μάθησης




Ανεστόπουλος Κωνσταντίνος, Προπτυχιακός Φοιτητής
Λένης Αλέξανδρος, Προπτυχιακός Φοιτητής
Αυλωνίτης Κωνσταντίνος- Οδυσσέας, Προπτυχιακός Φοιτητής

Τμήμα ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ

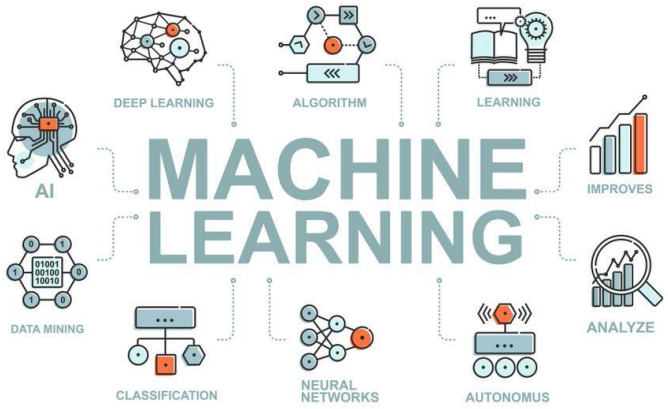
Ημερομηνία: 10/12/2023

# Περιεχόμενα

# 1. Εισαγωγή

## 1.1 Τεχνητή Νοημοσύνη και Βαθιά Μάθηση

Ο όρος Μηχανική Μάθηση (Machine Learning), αναφέρεται σε ένα υπολογιστικό σύστημα το οποίο έχει την δυνατότητα να αυτοεκπαιδεύεται και να παράγει πληροφορία με βάση την εκπαίδευση του και όχι απλώς χρησιμοποιώντας έναν μη μεταβλητό αλγόριθμο επεξεργασίας των δεδομένων, ο οποίος του δίνεται από τον χρήστη.



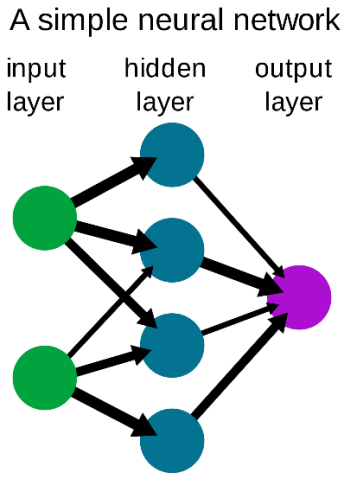

**Εικόνα 1:** Τομείς Μηχανικής Μάθησης

Ένα σύστημα που εφαρμόζει Μηχανική Μάθηση, είναι προγραμματισμένο με τέτοιον τρόπο ώστε να αντιδρά στα δεδομένα εισόδου του, όπως ένας άνθρωπος αντιδρά σε ερεθίσματα από το περιβάλλον του. Δηλαδή, αναλύει τα δεδομένα ψάχνοντας για μοτίβα ή δομές, μέσα στις οποίες μπορεί να τα κατατάξει. Το κυριότερο χαρακτηριστικό ενός τέτοιου συστήματος, είναι η ικανότητα του να βελτιώνει τις επιδόσεις του, βελτιστοποιώντας τον αλγόριθμο που χρησιμοποιεί, όσο του δίνονται περισσότερα δεδομένα προς επεξεργασία.

Οι αλγόριθμοι που χρησιμοποιούνται στην Μηχανική Μάθηση, αξιοποιούν Νευρωνικά Δίκτυα (Neural Networks). Συγγενή με την ανθρώπινη προσέγγιση της διαδικασίας της εκμάθησης, τα Νευρωνικά Δίκτυα κατηγοριοποιούν δεδομένα με βάση στοιχεία τα οποία μπορούν να αναγνωρίσουν, όπως ένας άνθρωπος θα μπορούσε να ξεχωρίσει σε ένα άλμπουμ φωτογραφιών, ποιες από τις φωτογραφίες περιέχουν σκύλους.

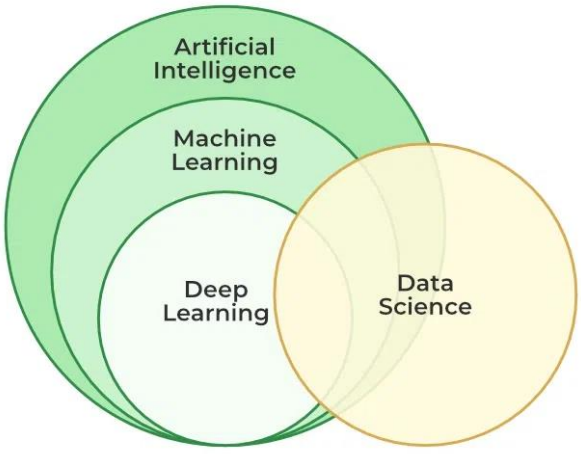
Το ποσοστό επιτυχίας των νευρωνικών δικτύων βέβαια, δεν φτάνει σε ανθρώπινα επίπεδα. Μέσα όμως από ανάδραση, η οποία προέρχεται από ανθρώπινο παράγοντα και είτε επιβεβαιώνει την ορθότητα της κατάταξης που έγινε, είτε επισημαίνει την λανθασμένη, ένα νευρωνικό δίκτυο είναι ικανό να μάθει και να αυξήσει το ποσοστό επιτυχίας του, όπως και να

μειώσει τον χρόνο που χρειάζεται για την επεξεργασία. Αυτή η διαδικασία ουσιαστικά μιμείται την ανθρώπινη εκμάθηση και μπορούμε κατά αυτόν τον τρόπο να εκπαιδεύσουμε ένα νευρωνικό δίκτυο να παράγει το επιθυμητό αποτέλεσμα.




**Εικόνα 2:** Τομείς Μηχανικής Μάθησης

Η Βαθιά Μάθηση (Deep Learning) εστιάζει σε ένα υποσύνολο της Μηχανική Μάθησης, αντλώντας έμπνευση από το πως ο ανθρώπινος εγκέφαλος αναγνωρίζει και ανακαλεί πληροφορίες και έχοντας ως σκοπό να εξαλείψει τον ανθρώπινο παράγοντα ανάδρασης, που απαιτείται για την βελτίωση των κλασικών Νευρωνικών Δικτύων. Πλέον αναφερόμαστε σε εκμάθηση και εκπαίδευση της μηχανής, από την ίδια την μηχανή.




**Εικόνα 3:** Τομείς Μηχανικής Μάθησης

Λόγω της πολυπλοκότητας και της φύσης του αλγορίθμου Βαθιάς Μάθησης, οι εφαρμογές που την αξιοποιούν χρήζουν πρόσβασης σε εξαιρετικά μεγάλο όγκο δεδομένων, τα οποία θα χρησιμοποιήσουν για να αυτοεκπαιδευτούν, τρέχοντας παράλληλα πολλαπλά εμβαθυμένα

νευρωνικά δίκτυα. Ένα παράδειγμα μίας τέτοιας εφαρμογής, είναι η συνεχής πρόταση νέων κομματιών μουσικής στους χρήστες εφαρμογών όπως το Spotify ή Apple Music, η οποία γίνεται μέσα από εξερεύνηση και ανάλυση όλης της μουσικής που υπάρχει διαθέσιμη και εντοπίζοντας κοινά χαρακτηριστικά με την μουσική που ο χρήστης φαίνεται να προτιμά και άλλα δεδομένα τηλεμετρίας.

Άλλα παραδείγματα Βαθιάς Μάθησης τα οποία είναι κοινά πλέον στην καθημερινότητά μας, είναι:

- **Αυτόνομη Οδήγηση:** Συνδυασμός δεδομένων από χάρτες, GPS, περιβαλλοντικά στοιχεία, οπτική αναγνώριση σήμανσης και λήψη αποφάσεων για την ταχύτητα και την διεύθυνση του οχήματος.
- **Ιατρικές Εφαρμογές:** Οπτική αναγνώριση παθήσεων μέσα από εικόνα (αξονική/μαγνητική τομογραφία, αγγειογραφία, κ.α)
- **Έξυπνες Συσκευές:** Αναγνώριση φωνής, ανταπόκριση σε τηλεμετρία παρεχόμενη από τον χρήση.

## 1.2 Χρήση Υλικού για Επιτάχυνση Αλγορίθμων Βαθιάς Μάθησης

Όπως είναι εμφανές, οι απαιτήσεις των αλγορίθμων Βαθιάς Μάθησης, είναι ιδιαίτερα υψηλές και σε δεδομένα, αλλά και σε επεξεργαστική ισχύ. Για να κρατηθούν οι χρόνοι εκτέλεσης σε ικανοποιητικά επίπεδα, οι ανάγκες παραλληλισμού και εισαγωγής/εξαγωγής δεδομένων είναι σχεδόν ακόρεστες. Γι’ αυτόν τον λόγο, οι αλγόριθμοι αυτοί σπάνια χρησιμοποιούν την CPU για να εκτελεστούν, επιλέγοντας να αξιοποιήσουν της αυξημένες δυνατότητες παράλληλης επεξεργασίας της GPU, όπως και το διευρυμένο εύρος ζώνης που προσφέρει συνήθως η μνήμη της.

Η τελευταία λέξη της τεχνολογίας όμως, είναι εξειδικευμένες μονάδες υλικού, φτιαγμένες αποκλειστικά για την επιτάχυνση των αλγορίθμων βαθιάς μάθησης και είναι γνωστές ως Επιταχυντές Βαθιάς Μάθησης ή DLP’s (Deep Learning Processor).

## 1.3 Η Προσέγγιση της AMD/Xilinx: Η μονάδα DPU

Η μονάδα DPU (Deep Learning Processor Unit) είναι ένα ηλεκτρονικό κύκλωμα, σχεδιασμένο για την εκτέλεση αλγορίθμων βαθιάς μάθησης, η οποία συνοδεύεται συνήθως από ξεχωριστή/αποκλειστική μονάδα μνήμης και εξειδικευμένη αρχιτεκτονική σετ εντολών για επιτάχυνση τέτοιων αλγορίθμων. Αποτελείται από ένα σύνολο Πυρήνων Πνευματικής Ιδιοκτησίας (τα λεγόμενα IP Cores), οι οποίοι είναι παραμετροποιήσιμοι και έρχονται

προεγκατεστημένοι στο υλικό, εξαλείφοντας την ανάγκη να ασχοληθεί ο σχεδιαστής του συστήματος με την τοποθέτηση και διασύνδεσή τους.

Κύρια λειτουργία τους και βάση του σχεδιασμού τους αποτελεί η επιτάχυνση αλγορίθμων που συναντώνται σε πολλαπλές εφαρμογές υπολογιστικής όρασης, όπως κατηγοριοποίηση εικόνας/βίντεο, εννοιολογική τμηματοποίηση εικόνας (semantic segmentation) και αναγνώριση/παρακολούθηση αντικειμένων.

Ο σκοπός τους είναι να παρέχουν υψηλότερη αποδοτικότητα, ενεργειακή και υπολογιστική, σε σχέση με CPU και κυρίως σε σχέση με GPU, οι οποίες έστω και αν προσφέρουν αρκετά καλές επιδόσεις, το ενεργειακό τους αποτύπωμα είναι τρομερά υψηλό.

# 2. Ανάλυση του DPU

## 2.1 Βασικά Χαρακτηριστικά

Το DPU είναι μια διαμορφώσιμη υπολογιστική μηχανή, ειδικά βελτιωμένη για νευρωνικά δίκτυα. Ο βαθμός παραλληλισμού στην μηχανή είναι μια σχεδιαστική παράμετρος, που μπορεί να επιλεγεί από την συσκευή και την εφαρμογή. Σε υψηλό επίπεδο, το DPU είναι μία μικρο-κωδικοποιημένη υπολογιστή μηχανή, η οποία διαθέτει ένα αποτελεσματικό και βελτιωμένο σύνολο εντολών, με αποτέλεσμα να υποστηρίζει περισσότερα νευρωνικά δίκτυα.

**Γενικότερα:**

- Χρησιμοποιείται διεπαφή AXI ως slave για την πρόσβαση σε καταχωρητές κατάστασης και ρυθμίσεων.
- Μία κύρια διεπαφή για πρόσβαση σε εντολές
- Υποστήριξη ρυθμιζόμενης κύριας διεπαφής AXI με 64 ή 128 Bits για πρόσβαση δεδομένων.
- Υποστήριξη ατομικής διαμόρφωση για κάθε κανάλι.
- Υποστήριξη προαιρετικής δημιουργίας αιτημάτων διακοπής(interrupt).

**Μερικά κύρια σημεία της λειτουργίας της DPU :**

- Αρχιτεκτονική υλικού με δυνατότητα διαμόρφωσης: B512, B800, B1024, B1152, B1600, B2304, B3136, and B4096
- Αλληλουχία τανυστών (Tensor Concatenation).
- Δυνατότητα διαμόρφωσης έως τρεις πυρήνες.

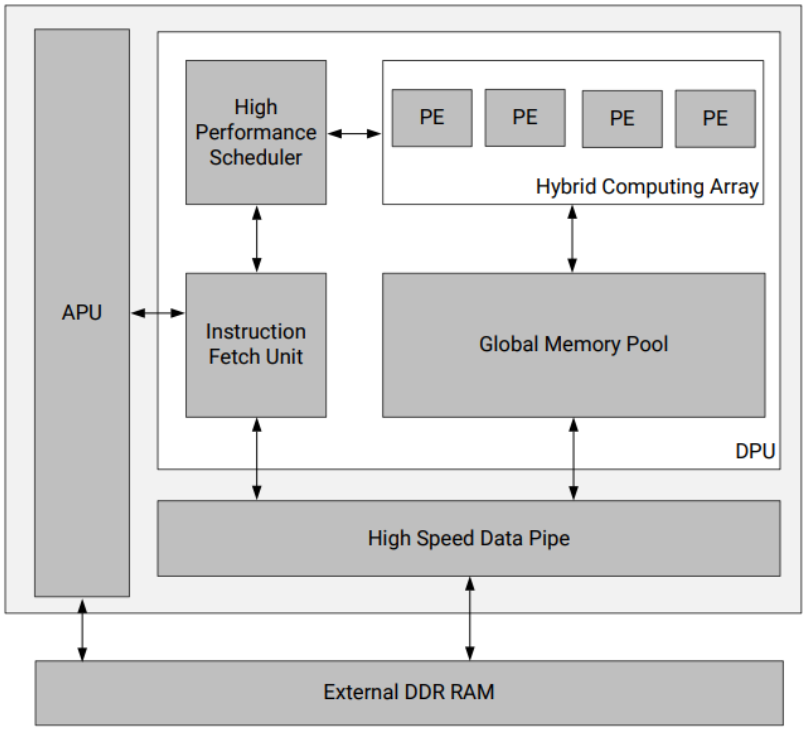
- Συνέλιξη και Αποσυνέλιξη.
- Συνάρτηση ενεργοποίησης ReLu (Rectified Linear Unit) & Leaky ReLu.
- Στοιχειακές πράξεις τανυστών.
- Διεσταλμένη συνέλιξη.
- Αναδιάταξη δεδομένων (Reorganization).
- Πλήρως συνδεδεμένα νευρωνικά δίκτυα σε όλα τα επίπεδα.
- Κανονικοποίηση των εισόδων μεγάλων παρτίδων (batch).

Ένα από τα κύρια χαρακτηριστικά που δίνουν στο DPU την απαραίτητη ταχύτητα στην εκτέλεση των αλγορίθμων βαθιάς μάθησης, είναι το ειδικά κατασκευασμένο σετ εντολών που χρησιμοποιούν Vitis AI. Αυτό, σε συνδυασμό με έναν χρονοπρογραμματιστή (scheduler) υψηλών επιδόσεων, μία συστοιχίας υβριδικών υπολογιστικών μηχανών και ένα εξελιγμένο front-end με βελτιστοποιημένη διαδικασία προσκόμισης εντολών (instruction fetch), παρέχει τις δυνατότητες για να επιταχύνει και να υποστηρίξει πληθώρα δημοφιλών συνελικτικών νευρωνικών δικτύων (convolutional neural networks), όπως τα παρακάτω:

- VGGNet της Viso AI
- Residual Networks (ResNets)
- GoogLeNet
- YOLO
- SSD
- MobileNet

## 2.2 Επισκόπηση Πυρήνα

Το DPU εκτελεί συγκεντρωμένο μικροκώδικα που παράγεται από τον γράφο του νευρωνικού δικτιού και απαιτεί προσβάσιμες θέσης μνήμης για τις εισαγόμενες εικόνες, καθώς και για προσωρινά εξαγόμενα δεδομένα. Επίσης ένα πρόγραμμα το οποίο εκτελείται στη μονάδα επεξεργασίας εφαρμογής (APU, Application Processing Unit), πρέπει να μπορεί να εξυπηρετήσει τυχών διακοπές και να συντονίσει τις μεταφορές δεδομένων.

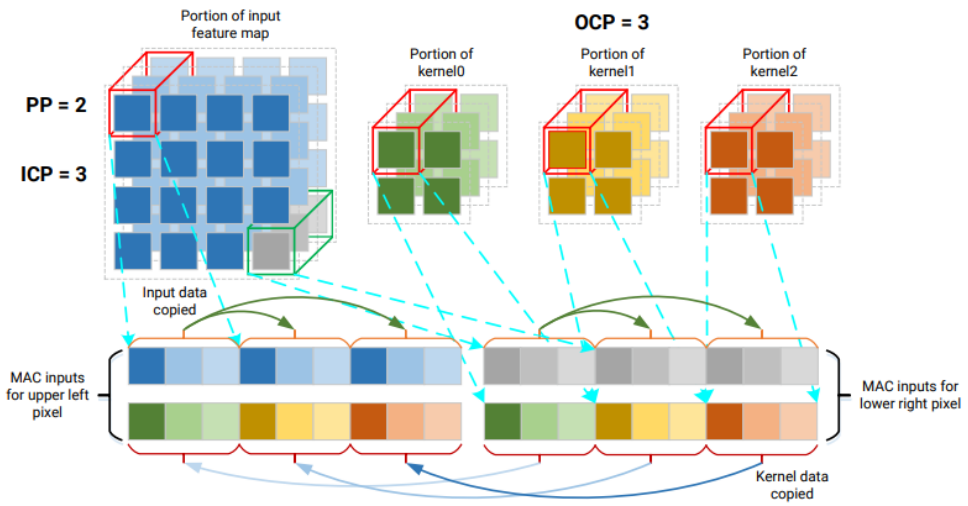



**Εικόνα 4:** Μπλοκ Διάγραμμα του Πυρήνα

Παρέχει επίσης ορισμένες διαμορφώσιμες παραμέτρους χρήστη για την βελτιστοποίηση της χρήσης πόρων και προσαρμογής διάφορων εφαρμογών. Διαφορετικές διαμορφώσεις μπορούν να επιλεγούν μέσω για τμήματα DSP, LUT, μπλοκ RAM, και UltraRAM βασισμένες στον αριθμό των διαθέσιμων πόρων προγραμματιστικής λογικής. Επιπλέων υπάρχει επιλογή για πρόσθετες λειτουργείες, όπως ενίσχυση καναλιού, συγκέντρωση μέσων (average pooling), εις-βάθος συνέλιξη (depthwise convolution) και softmax. Τέλος υπάρχει η επιλογή να καθοριστεί ο αριθμός των πυρήνων DPU που θα δημιουργηθούν σε ένα μόνο DPU IP.

## 2.3 Αρχιτεκτονική του DPU

Το DPU μπορεί να ρυθμιστεί με διάφορες αρχιτεκτονικές συνέλιξης που σχετίζονται με τον παραλληλισμό την μονάδας συνέλιξης. Η αρχιτεκτονική της DPU περιλαμβάνει B512, B800, B1024, B1152, B1600, B2304, B3136, and B4096. Υπάρχουν τρεις διαστάσεις παραλληλισμού της αρχιτεκτονικής συνέλιξης : παραλληλισμός πίξελ, παραλληλισμός καναλιών εισόδου και εξόδου. Ο παραλληλισμός καναλιών εισόδου είναι πάντα ίσος με αυτόν της εξόδου (αυτό είναι ισοδύναμο με το channel_parallel στον παρακάτω πίνακα).




**Εικόνα 5:** *Απεικόνιση των τριών διαστάσεων του παραλληλισμού*

Στη παραπάνω εικόνα, απεικονίζεται ο παραλληλισμός καναλιών εισόδου (ICP = 3), ο παραλληλισμός καναλιών εξόδου (OCP = 3) και ο παραλληλισμός πίξελ (PP = 2). Ο παραλληλισμός καναλιών εξόδου είναι ισοδύναμος με τον αριθμό των πυρήνων που χρησιμοποιούνται κατά τη διάρκεια μιας υπολογιστικής συνέλιξης.

Η διαφορετική αρχιτεκτονική απαιτεί διαφορετικό προγραμματισμό λογικών πόρων. Οι μεγαλύτερες αρχιτεκτονικές μπορούν να επιτύχουν υψηλότερη απόδοση με περισσότερους πόρους. Ο παραλληλισμός για τις διαφορετικές αρχιτεκτονικές αναφέρεται στον παρακάτω πίνακα.

| DPUCZDX8G Architecture | Pixel Parallelism (PP) | Input Channel Parallelism (ICP) | Output Channel Parallelism (OCP) | Peak Ops (operations/per cycle) |
|---|---|---|---|---|
| B512 | 4 | 8 | 8 | 512 |
| B800 | 4 | 10 | 10 | 800 |
| B1024 | 8 | 8 | 8 | 1024 |
| B1152 | 4 | 12 | 12 | 1152 |

**Πίνακας 1:** Παραλληλισμός για διαφορετικές αρχιτεκτονικές συνέλιξης

### 2.3.1 Διαχείριση RAM

Τα βάρη, bias και ενδιάμεσοι χαρακτηριστικοί χάρτες αποθηκεύονται προσωρινά στην μνήμη εντός του κυκλώματος. Η μνήμη αυτή αποτελείται από RAM η οποία μπορεί να αρχικοποιηθεί ως μπλοκ RAM και UltraRAM. Η επιλογή της χρήσης της RAM καθορίζει το σύνολο της μνήμης εντός του κυκλώματος και γίνεται χρήση του σε διαφορετικές DPUCZDX8G αρχιτεκτονικές.

Η ρύθμιση αυτή αφορά όλους τους DPUCZDX8G πυρήνες στο DPUCZDX8G IP. Η υψηλή χρήση της RAM σημαίνει ότι η μνήμη εντός του κυκλώματος θα είναι μεγαλύτερη, επιτρέποντας στον DPUCZDX8G περισσότερη ευελιξία στον χειρισμό των ενδιάμεσων δεδομένων. Υψηλότερη χρήση RAM τείνει σε υψηλότερη απόδοση σε κάθε πυρήνα.

### 2.3.2 Channel Augmentation

Το channel augmentation είναι ένα προαιρετικό χαρακτηριστικό για την βελτίωση της αποδοτικότητας του DPUCZDX8G όταν ο αριθμός των καναλιών εισόδου δεδομένων είναι αρκετά χαμηλότερος από την διαθέσιμη ικανότητα παραλληλισμού καναλιών. Για παράδειγμα, τα κανάλια εισόδου δεδομένων του πρώτου στρώματος στα περισσότερα συνελικτικά νευρωνικά δίκτυα είναι συνήθως τρία, πράγμα το οποίο σημαίνει πως δεν αξιοποιούνται όλα τα διαθέσιμα κανάλια υλικού.

Εάν ο αριθμός των καναλιών εισόδου είναι μεγαλύτερος από την ικανότητα παραλληλισμού καναλιών, τότε το channel augmentation μπορεί να αξιοποιηθεί. Συνεπώς, το channel augmentation μπορεί να βελτιώσει την αποδοτικότητα για τα περισσότερα συνελικτικά νευρωνικά δίκτυα, ωστόσο θα κοστίσει παραπάνω λογικούς πόρους.

| DPUCZDX8G Architecture | Extra LUTs with Channel Augmentation |
|---|---|
| B512 | 3121 |
| B800 | 2624 |
| B1024 | 3133 |
| B1152 | 1744 |
| B1600 | 2476 |
| B2304 | 1710 |
| B3136 | 1946 |
| B4096 | 1701 |

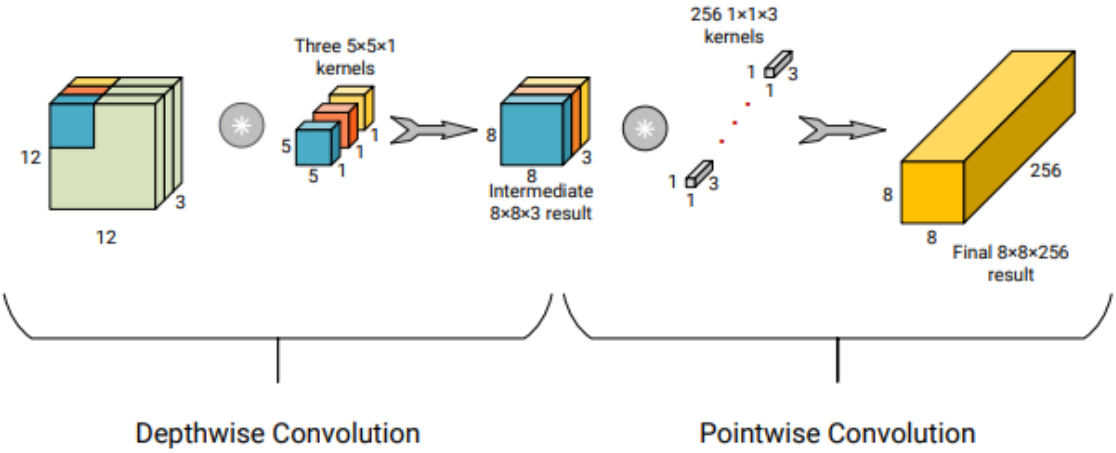
**Πίνακας 2:** Αριθμός έξτρα LUT με χρήση Channel Augmentation

### 2.3.3 Depthwise Convolution

Στην συμβατική συνέλιξη, κάθε κανάλι εισόδου χρειάζεται να πραγματοποιήσει την πράξη με ένα συγκεκριμένο πυρήνα (kernel) και μετέπειτα το αποτέλεσμα λαμβάνεται συνδυάζοντας τα αποτελέσματα όλων των καναλιών. Στην διαχωρίσιμη εις βάθος (depthwise) συνέλιξη, η λειτουργία πραγματοποιείται σε δύο βήματα: στην εις βάθος συνέλιξη και στην κατά σημείον (pointwise) συνέλιξη.

Η εις βάθος συνέλιξη πραγματοποιείται για κάθε χαρακτηριστικό χάρτη ξεχωριστά, όπως φαίνεται στο αριστερό μέρος της παρακάτω εικόνας. Το επόμενο βήμα είναι να πραγματοποιηθεί η κατά σημείον συνέλιξη, η οποία είναι ίδια με την συμβατική συνέλιξη με μέγεθος πυρήνα 1x1. Ο παραλληλισμός της εις βάθους συνέλιξης είναι μισός από τον παραλληλισμό πίξελ (PP).

Στον DPUCZDX8G, η εις βάθος συνέλιξη πραγματοποιείται από την μηχανή ALU σε συνδυασμό με την συσσώρευση. Το εύρος παραλληλισμού της ALU εκτείνεται από το 1 ως το PP και προτείνεται να τεθεί ως PP/2.




**Εικόνα 6:** *Απεικόνιση της διαχωρίσιμης εις βάθος συνέλιξης*

| ALU Parallel | LUTs | FF | Block RAMs | DSPs |
|---|---|---|---|---|
| 1 | 44212 | 88250 | 255 | 662 |
| 2 | 46599 | 92380 | 255 | 678 |
| 4 (recommended) | 51388 | 98525 | 255 | 710 |
| 8 | 60751 | 111329 | 255 | 774 |

**Πίνακας 3:** Πόροι του DPUCZDX8G με διαφορετικά επίπεδα παραλληλισμού ALU

### 2.3.4 Elementwise Multiply και AveragePool

Ο πολλαπλασιασμός κατά στοιχείο (Elementwise) υπολογίζει το γινόμενο Ανταμάρ μεταξύ δυο εισαγόμενων χαρακτηριστικών χαρτών. Το κανάλι εισόδου του κατά στοιχείο πολλαπλασιασμού εκτείνεται από το 1 έως το 256* channel_parallel. Αυτό το χαρακτηριστικό είναι πάντα ενεργό.

Η επιλογή AveragePool καθορίζει το αν η πράξη της συσσώρευσης μέσων θα πραγματοποιηθεί στο DPUCZDX8G ή όχι. Αυτό το χαρακτηριστικό είναι πάντα ενεργοποιημένο.

### 2.3.5 Τύποι ReLU

Η επιλογή τύπου ReLU καθορίζει την λειτουργία ενεργοποίησης συγκεκριμένων ReLU, η οποία μπορεί να χρησιμοποιηθεί με το DPUCZDX8G. Τα ReLU και ReLU6 υποστηρίζονται από προεπιλογή. Η χρήση της επιλογής “ReLU + LeakyReLU + ReLU6“ θα ενεργοποιήσει το LeakyReLU, ως συνάρτηση ενεργοποίησης.

| DPUCZDX8G Architecture | Extra LUTs from Conv | Extra LUTs from ALU |
|---|---|---|
| B512 | 465 | 493 |
| B800 | 502 | 747 |
| B1024 | 636 | 643 |
| B1152 | 1038 | 896 |
| B1600 | 831 | 658 |
| B2304 | 460 | 439 |
| B3136 | 812 | 1051 |
| B4096 | 746 | 947 |

**Πίνακας 4:** Αριθμός έξτρα LUT με χρήση ReLU + Leaky ReLU + ReLU6 εν συγκρίσει με χρήση ReLU + ReLU6

### 2.3.6 Argmax, Max και Softmax

Η επιλογή Save Argmax ενεργοποιεί τις λειτουργείες argmax και max καθ’ όλη τη διάσταση του καναλιού, όταν ανακτά την έξοδο στον χώρο της μνήμης. Σε μερικές περιπτώσεις, όπως αυτή του καταμερισμού, μόνο ο σελιδοδείκτης της μέγιστης τιμής είναι απαραίτητος. Στην συνέχεια είναι χρήσιμο το argmax να αντικαταστήσει το softmax, στο μοντέλο, έτσι ώστε να αφαιρεθούν οι εκθετικοί υπολογισμού και να μειωθεί ο χρόνος καθυστέρησης.

| DPUCZDX8G Architecture | Extra LUTs | Extra Registers |
|---|---|---|
| B512 | 422 | 556 |
| B800 | 399 | 547 |
| B1024 | 460 | 546 |
| B1152 | 503 | 631 |
| B1600 | 590 | 640 |
| B2304 | 803 | 442 |
| B3136 | 832 | 758 |
| B4096 | 735 | 389 |

**Πίνακας 5:** Έξτρα διαθέσιμοι πόροι με την επιλογή "Save Argmax"

Υπάρχει επίσης η επιλογή να ενεργοποιειθεί μία εφαρμογή υλικού του τελεστή softmax. Ο επιταχυντής υλικού softmax είναι τοποθετημένος μέσα στον wrapper του DPU IP, αλλά είναι ένας ξεχωριστός επιταχυντής, με την δική του διεπαφή, ο οποίος ενσωματώνει μορφή δεδομένων ακεραίων 8-bit στην είσοδο και αριθμών κινητής υποδιαστολής στην έξοδο.

Η υλοποίηση υλικού του softmax μπορεί να αποβεί έως και 160 φορές πιο γρήγορη σε σχέση με μία υλοποίηση λογισμικού, σε συσκευές MPSoC. Οι χρήστες μπορούν να ενεργοποιήσουν αυτό το χαρακτηριστικό, εάν το μοντέλο/δίκτυό τους περιλαμβάνει ένα υπολογιστικό στρώμα softmax και επιθυμούν να βελτιώσουν την παροχέτευση.

| IP Name | Extra LUTs | Extra FFs | Extra BRAMs | Extra DSPs |
|---|---|---|---|---|
| Softmax | 9580 | 8019 | 4 | 14 |

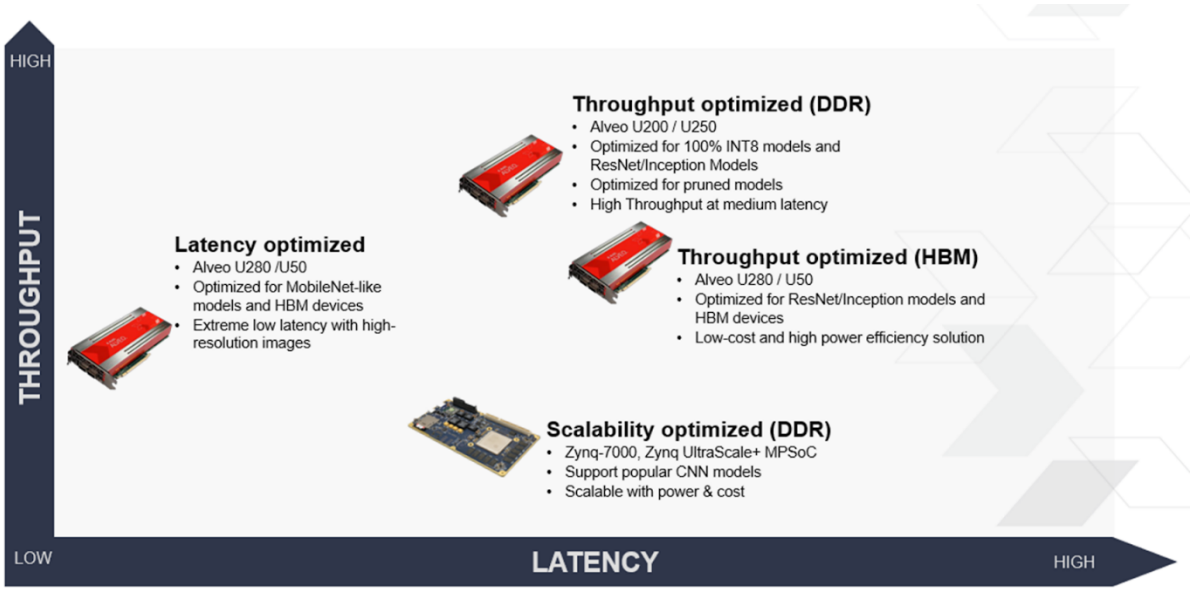
**Πίνακας 6:** Έξτρα πόροι με χρήση Softmax

# 3. Εφαρμογές και χρήση του DPU

Πέραν της ισχυρής επεξεργαστικής δύναμης που προσφέρει, το DPU έχει τεράστιες δυνατότητες επεκτασιμότητας, εφόσον μπορεί να υλοποιηθεί σε οποιοδήποτε μέγεθος απαιτεί η εφαρμογή και υποστηρίζεται από τους παρακάτω επιταχυντές ή FPGA Boards:

- Xilinx Zynq®-7000 devices
- Zynq UltraScale+ MPSoCs
- Xilinx Kria KV260
- Versal ACAP cards
- Alveo Accelerator Cards

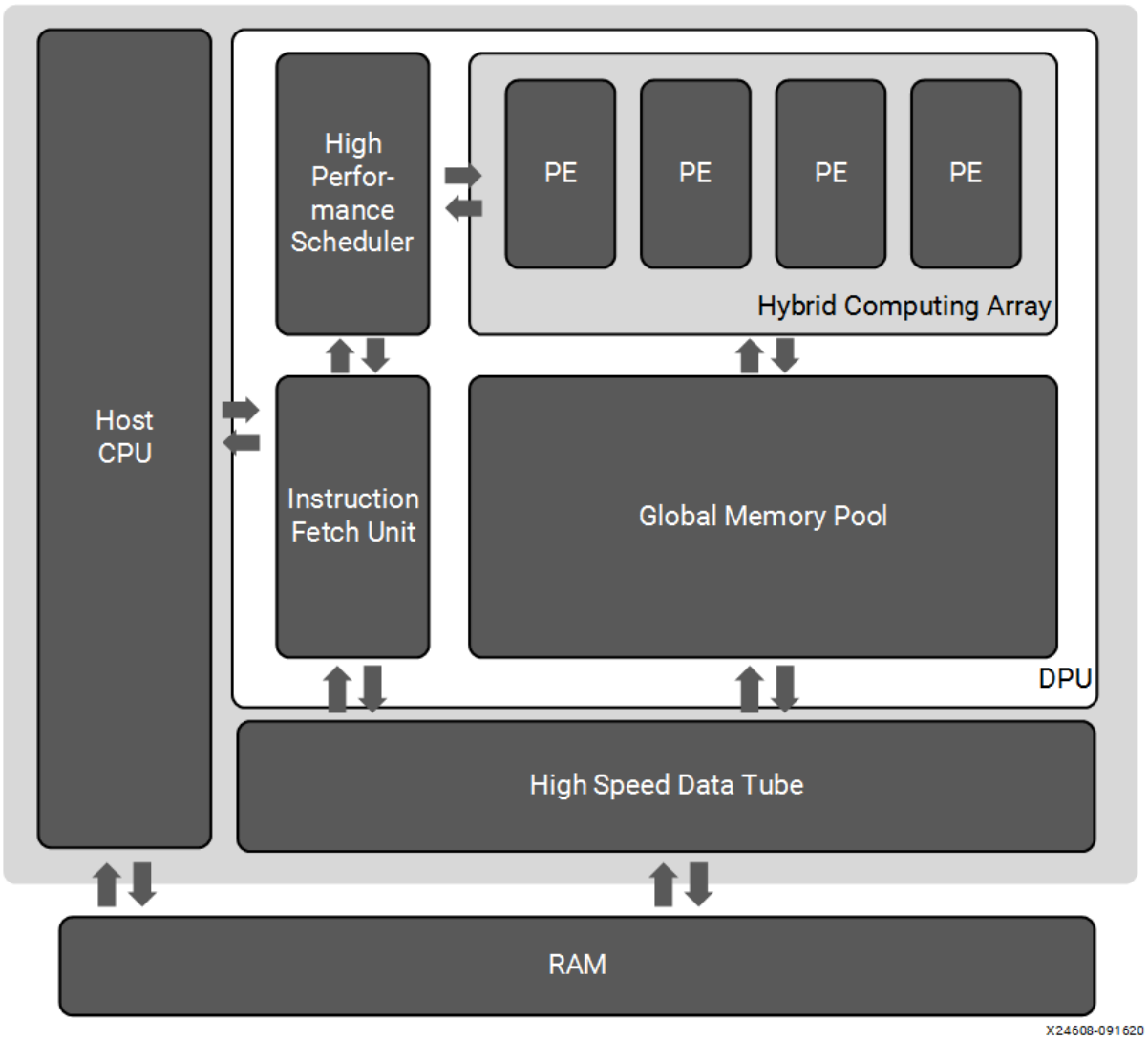
Ο πυρήνας του DPU μπορεί να ενσωματωθεί στα μπλοκ προγραμματιζόμενης λογικής (Programmable Logic) των παραπάνω, με απευθείας σύνδεση στο σύστημα επεξεργασίας (Processing System), μειώνοντας έτσι την χρονοκαθυστέρηση για την μεταξύ τους επικοινωνία και αυξάνοντας τον ρυθμό ροής δεδομένων.




**Εικόνα 7:** *Απεικόνιση κατηγοριών βελτιστοποίησης DPU*

## 3.1 Υλοποίηση Zynq UltraScale+ MPSoC: DPUCZDX8G

Ο πυρήνας DPUCZDX8G έχει βελτιστοποιηθεί για χρήση στο Zynq UltraScale+ MPSoC. Μπορεί να ενσωματωθεί ως ένα μπλοκ στην προγραμματιζόμενη λογική των επιλεγμένων Zynq UltraScale+ MPSoCs με άμεσες συνδέσεις στο επεξεργαστικό σύστημα. Το DPU μπορεί να παραμετροποιηθεί από τον χρήστη, παρουσιάζοντας πολλαπλές παραμέτρους που μπορούν να προσδιοριστούν έτσι ώστε να βελτιστοποιούν τους πόρους της προγραμματιζόμενης λογικής ή να προσαρμόσουν κατάλληλα τις λειτουργίες που έχουν ενεργοποιηθεί.

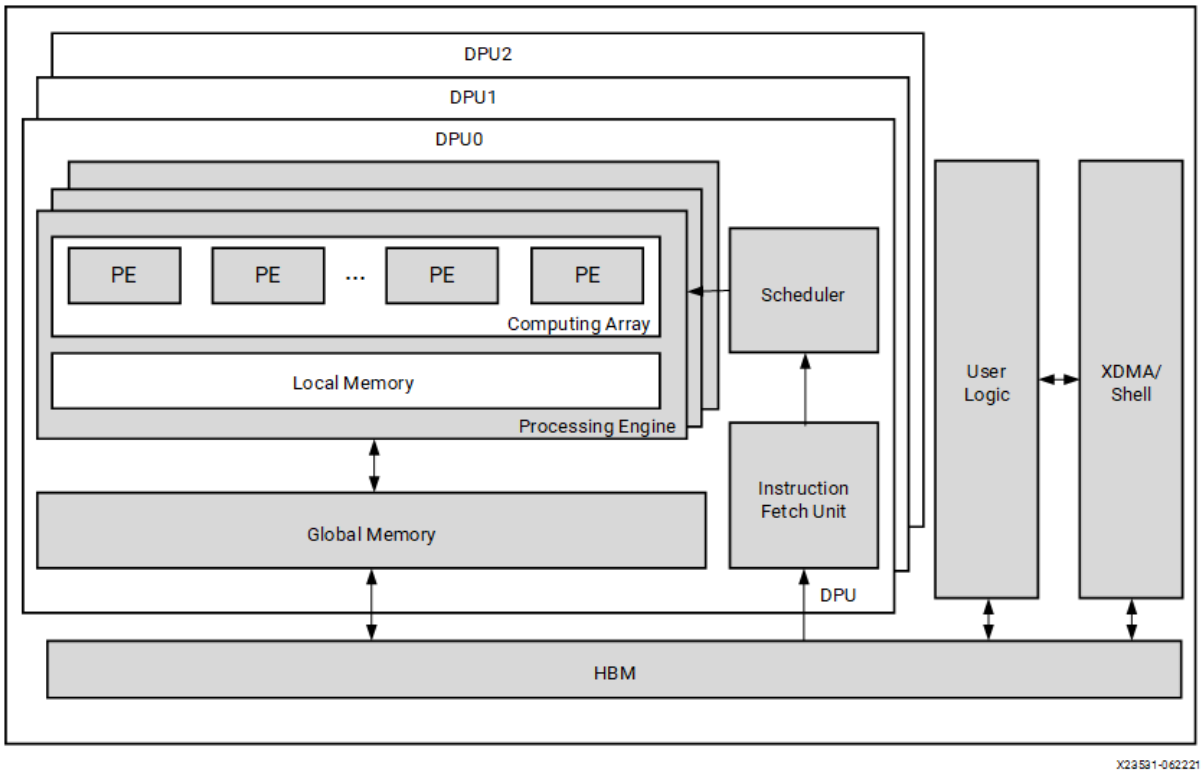



**Εικόνα 8:** *Απεικόνιση αρχιτεκτονικής DPUCZDX8G*

## 3.2 Υλοποίηση Alveo U50/U280 Card: DPUCAHX8H

Το DPUCAHX8H DPU είναι μια προγραμματιζόμενη μηχανή, βελτιστοποιημένη για συνελικτικά νευρωνικά δίκτυα, κυρίως για εφαρμογές υψηλής απόδοσης. Αυτή η μονάδα περιλαμβάνει μια μονάδα χρονοπρογραμματιστή υψηλής απόδοσης, μια μονάδα υβριδικής συστοιχίας υπολογιστικών μηχανών, μια μονάδα ανάκτησης εντολών και μια μονάδα καθολικής δεξαμενής μνήμης. Το DPU χρησιμοποιεί ένα εξειδικευμένο σύνολο εντολών, το οποίο επιτρέπει την αποτελεσματική υλοποίηση πολλών συνελικτικών νευρωνικών δικτύων. Μερικά παραδείγματα συνελικτικών νευρωνικών δικτύων που αναπτύσσονται περιλαμβάνουν τα VGG, ResNet, GoogLeNet, YOLO, SSD, MobileNet και FPN.

Ο χώρος της μνήμης HBM (High Bandwidth Memory) χωρίζεται σε εικονικές τράπεζες (virtual banks) και στην μνήμη συστήματος. Οι εικονικές τράπεζες χρησιμοποιούνται για την αποθήκευση προσωρινών δεδομένων και η μνήμη του συστήματος χρησιμοποιείται για την αποθήκευση εντολών, εικόνων εισόδου, αποτελεσμάτων εξόδου και δεδομένων χρήστη. Μετά την εκκίνηση, το DPU ανακτά οδηγίες μοντέλου από τη μνήμη του συστήματος για να ελέγχει τη λειτουργία του υπολογιστικού μηχανισμού.




**Εικόνα 9:** *Μπλοκ διάγραμμα κορυφαίου επιπέδου του DPUCAHX8H*

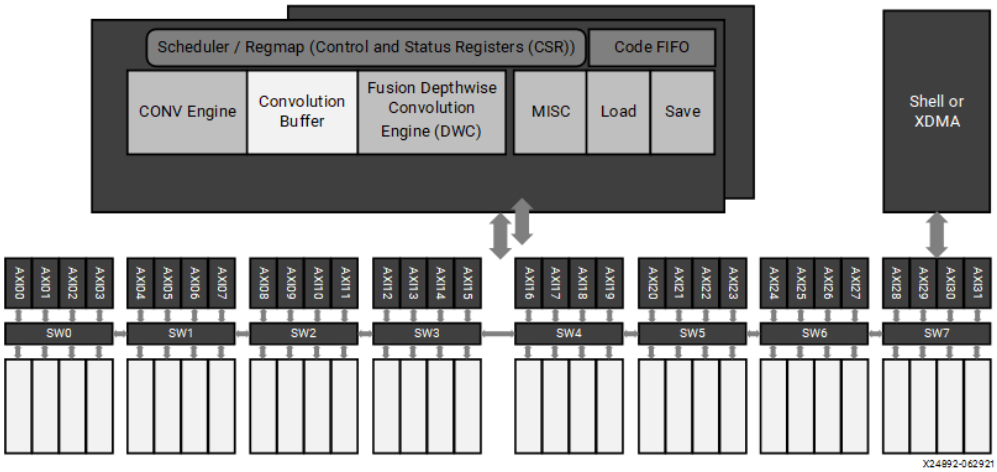
Τα εργαλεία τεχνητής νοημοσύνης χρησιμοποιούνται για την εκτέλεση βελτιστοποιήσεων μοντέλων για αποτελεσματική χρήση της DPU και στη συνέχεια δημιουργούν τις οδηγίες χρόνου εκτέλεσης που απαιτούνται για την υλοποίηση του μοντέλου στον πυρήνα της DPU. Η μνήμη HBM χρησιμοποιείται για την προσωρινή αποθήκευση βαρών (weights), του bias, ενδιάμεσων δεδομένων (intermediate data) και δεδομένων εξόδου για την επίτευξη υψηλής.

Ο πυρήνας του DPU μπορεί να ενσωματωθεί στο κομμάτι προγραμματιζόμενης λογικής της επιλεγμένης πλακέτας Alveo. Το DPU απαιτεί κατάλληλες εντολές για να εφαρμόσει ένα νευρωνικό δίκτυο και τις προσβάσιμες θέσεις μνήμης για την εισαγωγή εικόνων, όπως επίσης προσωρινά δεδομένα και δεδομένα εξόδου. Μία μονάδα καθορισμένη από τον χρήστη, η οποία τρέχει στο κομμάτι προγραμματιζόμενης λογικής πρέπει να μπορεί να πραγματοποιήσει κατάλληλες ρυθμίσεις, να εισάγει εντολές, να διαχειριστεί διακοπές (interrupts) και να συντονίσει μεταφορές δεδομένων.

## 3.3 Υλοποίηση Alveo U50/U50LV/U280 Card: DPUCAHX8L

Το DPUCAHX8L είναι ένας νέος επιταχυντής γενικής χρήσης, συνελικτικών νευρωνικών δικτύων, ο οποίος είναι βελτιστοποιημένος για χρήση σε κάρτες με HBM, όπως οι Alveo U50/U50LV και U280, οι οποίες είναι σχεδιασμένες για εφαρμογές χαμηλής καθυστέρησης (latency). Περιέχει ένα νέο DPU χαμηλής καθυστέρησης, συνδυασμένο με ένα υποσύστημα μνήμης HBM και υποστηρίζει 4 TOPs (Terra-Operations) έως 5.3TOPs συστοιχία MAC.

Υποστηρίζει επίσης αλλεπάλληλη συνέλιξη και συνέλιξη εις-βάθος (depthwise convolution), με σκοπό να αυξήσει τον υπολογιστικό παραλληλισμό. Πέραν αυτού, παρέχει υποστήριξη για ιεραρχικό σύστημα μνήμης UltraRAM και HBM, για να μεγιστοποιήσει την μετακίνηση δεδομένων. Με το συγκεκριμένο DPU, ο μεταγλωττιστής Vitis AI υποστηρίζει την διεπαφή υπερ-στρώματος (super layer interface) και πολλές άλλες νέες στρατηγικές μεταγλωττισμού για συγχώνευση πυρήνων (kernels) και διαμερισματοποίηση γράφων.

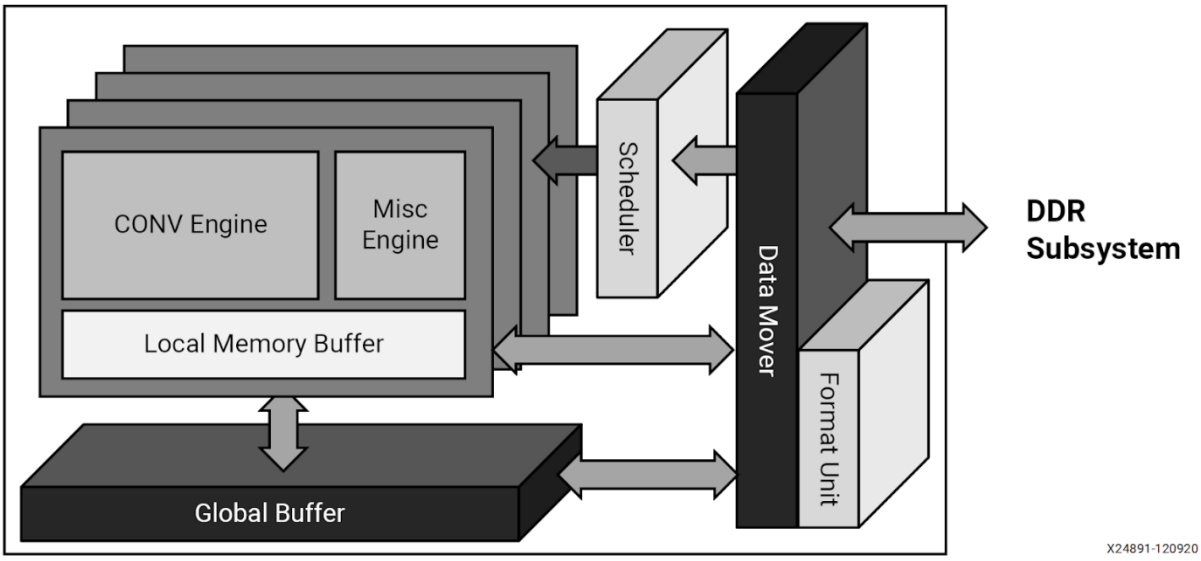



**Εικόνα 10:** *Αρχιτεκτονική του DPUCAHX8L*

## 3.4 Υλοποίηση Alveo U200/U250 Card: DPUCADF8H

Το DPUCADF8H είναι ένα DPU βελτιστοποιημένο για κάρτες Alveo U200/U250 και στοχεύει για εφαρμογές υψηλών επιδόσεων και παροχέτευσης. Τα κυριότερα χαρακτηριστικά του είναι:

- Οι υπολογιστικές μηχανές ειδικά σχεδιασμένες για υψηλή αποδοτικότητα και παροχέτευση, με τις ικανότητες παροχέτευσης να είναι βελτιωμένες κατά 1.5X έως 2.0X σε διαφορετικά σενάρια χρήσης.
- Μεγάλο εύρος υποστήριξης για συνελικτικά νευρωνικά δίκτυα
- Φιλικό σε κλαδεμένα (pruned) συνελικτικά νευρωνικά δίκτυα
- Βελτιστοποιημένο για ανάλυση εικόνων υψηλής ανάλυσης

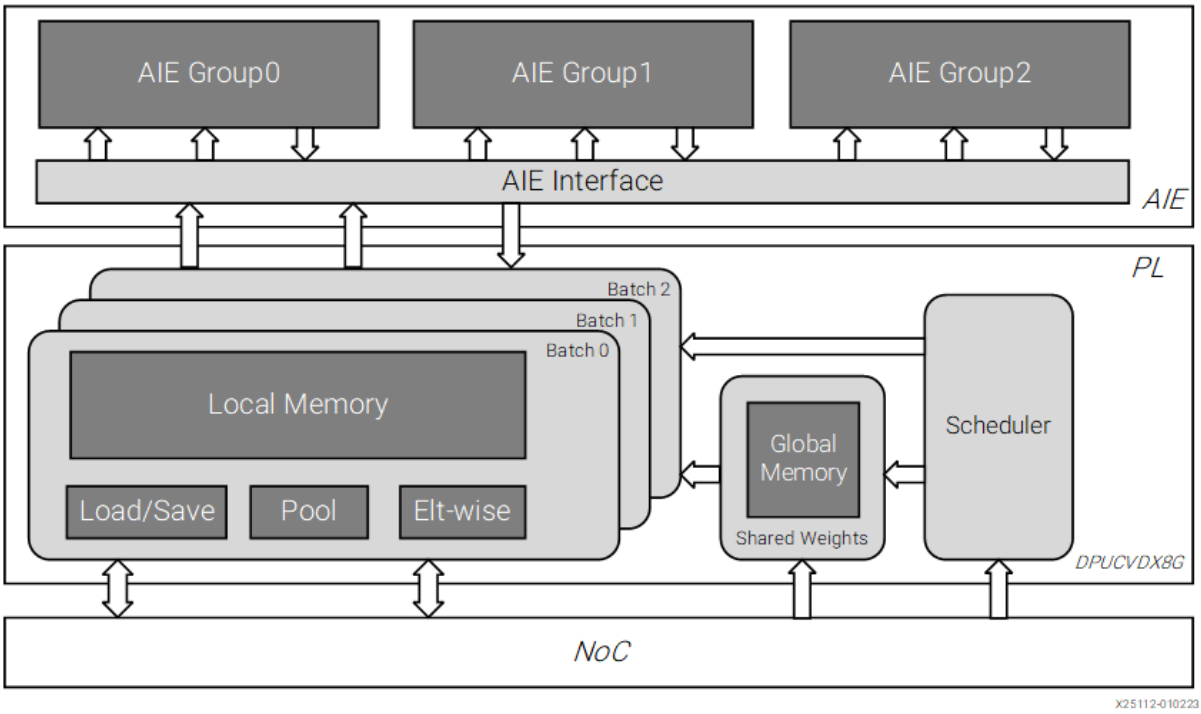



**Εικόνα 11:** *Αρχιτεκτονική του DPUCADF8H*

## 3.5 Υλοποίηση Versal AI Core Series: DPUCVDX8G

Το DPUCVDX8G είναι μία υψηλών επιδόσεων υπολογιστική μηχανή γενικευμένων συνελικτικών νευρωνικών δικτύων, βελτιστοποιημένων για τους πυρήνες τεχνητής νοημοσύνης Versal. Οι πυρήνες Versal μπορούν να παρέχουν επιδόσεις/βατt ανώτερης κλάσης, σε σχέση με τα συμβατικά FPGA, CPU και GPU. Το DPUCVDX8G συντίθεται από Μηχανές AI και κυκλώματα προγραμματιζόμενης λογικής.

Ο συγκεκριμένος πυρήνας είναι παραμετροποιήσιμος από τον χρήστη και εκθέτει πολλές παραμέτρους, οι οποίες μπορούν να αξιοποιηθούν για να βελτιστοποιηθούν οι μηχανές AI και οι πόροι προγραμματιζόμενης λογικής ανά σενάριο χρήσης, ή να εξατομικευτούν διάφορα χαρακτηριστικά τους.



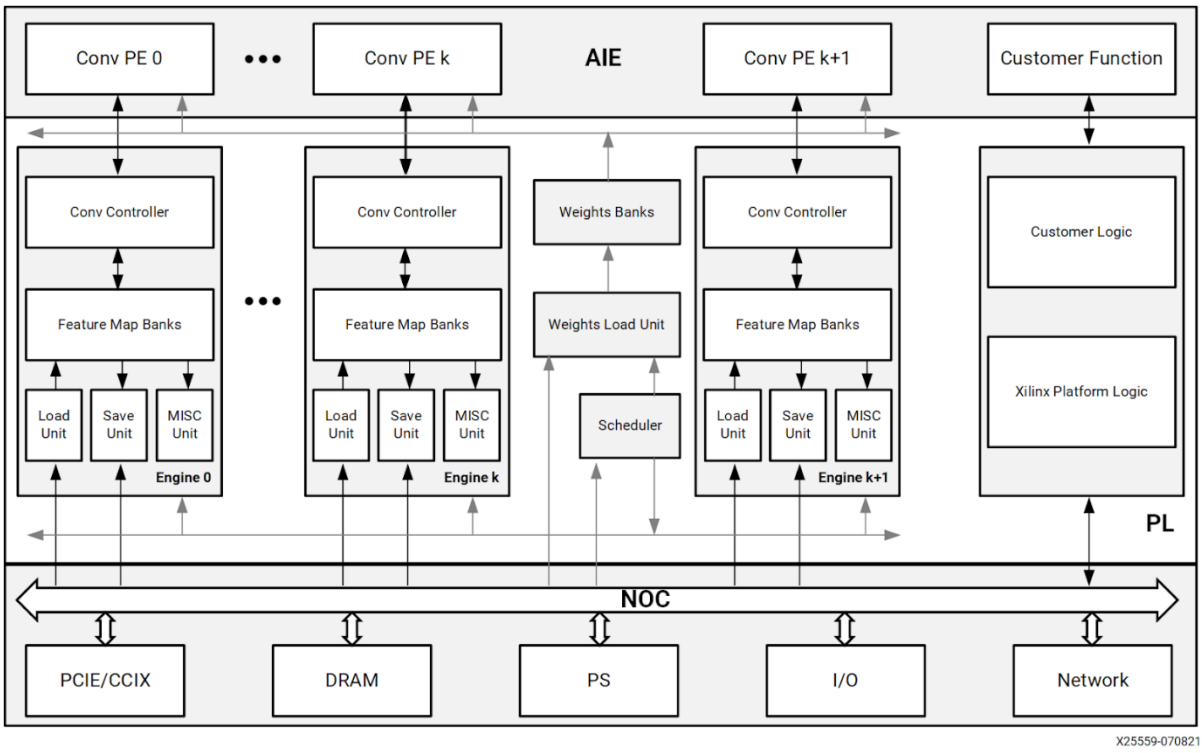

**Εικόνα 12:** *Αρχιτεκτονική του DPUCVDX8G*

## 3.6 Υλοποίηση Versal AI Core Series: DPUCVDX8H

Το DPUCVDX8H είναι μία μηχανή επεξεργασίας υψηλών επιδόσεων και παροχέτευσης για γενικευμένα συνελικτικά νευρωνικά δίκτυα, με χρήση στην σειρά πυρήνων AI Versal.

Πέραν της παραδοσιακής λογικής προγράμματος, οι συσκευές Versal ενσωματώνουν συστοιχίες μηχανών AI υψηλών επιδόσεων, NoC (Network on Chip) υψηλού εύρους ζώνης, ελεγκτές DDR (Dual Data Rate) και LPDDR (Low Power DDR) και άλλες διεπαφές υψηλής ταχύτητας, που μπορούν να προσφέρουν ανώτερη κλάση επιδόσεων/βατ σε σχέση με συμβατικά FPGA, CPU και GPU. Το DPUCVDX8H ενσωματώνεται σε συσκευές Versal για να αξιοποιήσει αυτά τα προτερήματα.

Ο χρήστης μπορεί να το παραμετροποιήσει κατάλληλα για να ανταποκριθεί στις απαιτήσεις της εφαρμογής στο κέντρου δεδομένων (data center) του.




**Εικόνα 13:** *Αρχιτεκτονική του DPUCVDX8H*

# 4. Πηγές

https://docs.xilinx.com/r/1.4.1-English/ug1414-vitis-ai/Deep-Learning-Processor-Unit

https://en.wikipedia.org/wiki/Deep_learning_processor

https://docs.xilinx.com/r/en-US/pg366-dpucahx8l/

https://design.google/library/ux-ai.

Louis Columbus, “Roundup Of Machine Learning Forecasts And Market Estimates, 2018,” Forbes, February 18, 2018.

https://www.xilinx.com/products/intellectual-property/dpu.html#overview

https://docs.xilinx.com/r/en-US/pg338-dpu?tocId=3xsG16y_QFTWvAJKHbisEw