ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΠΕΛΟΠΟΝΝΗΣΟΥ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε

ΣΥΓΧΡΟΝΕΣ ΤΑΣΕΙΣ ΣΤΗ ΣΧΕΔΙΑΣΗ ΚΑΡΤΑΣ ΓΡΑΦΙΚΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Χριστίνα Στεφανοπούλου

Α.Μ.2012017

Επιβλέπων καθηγητής: Σταύρος Δεληγιαννίδης

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω τον κύριο Δεληγιαννίδη για την επίβλεψη αυτής της πτυχιακής εργασίας, την ευκαιρία που μου έδωσε να την εκπονήσω και την εξαιρετική συνεργασία που είχαμε. Επίσης ευχαριστώ θερμά τους γονείς μου και τους φίλους μου για την υποστήριξη τους.

Σπάρτη,2019

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου, περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάση επιστημονικής παράφρασης.
Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων.
Δηλώνω,συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δε μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας

Όνομα και Επώνυμο Συγγραφέα (Με Κεφαλαία): ……
Υπογραφή (Ολογράφως, χωρίς μονογραφή): ……
Ημερομηνία (Ημέρα – Μήνας – Έτος): ……..

Περίληψη

Αντικείμενο της παρούσας πτυχιακής εργασίας είναι οι σύγχρονες τάσεις στη σχεδίαση κάρτας γραφικών. Θα παρουσιαστούν και θα αναλυθούν οι διάφοροι τύποι ολοκληρωμένων κυκλωμάτων για διαχείριση γραφικών. Θα αναλυθούν τα ιδιαίτερα χαρακτηριστικά τους και οι σχεδιαστικές αρχιτεκτονικές τους. Επίσης θα παρουσιαστούν μέθοδοι αξιολόγησης επιδόσεων και σύγκριση μεταξύ CPU και GPU graphics.

Λέξεις Κλειδιά: Κάρτες γραφικών, ολοκληρωμένα κυκλώματα, παράλληλη επεξεργασία, CPU graphics, GPU graphics.

Abstract

The subject of this thesis is the modern trends in graphic card design. The various types of integrated circuits for graphic management will be presented. They will analyze their specific features and design architectures. They will also present performance evaluation methods and a comparison between CPU and GPU graphics.

Keywords: Graphics Cards, Integrated Circuits, Parallel Processing, CPU Graphics, GPU Graphics

Περιεχόμενα

Εισαγωγή

Οι κάρτες γραφικών αποτελούν σήμερα ένα αναπόσπαστο κομμάτι κάθε υπολογιστικού συστήματος. Λειτουργώντας πάνω στην μητρική κάρτα ή σε ξεχωριστό κύκλωμα διασυνδεμένη με αυτή, οι προηγμένες κάρτες γραφικών επιτρέπουν την εκτέλεση απαιτητικών εφαρμογών, όπως παιχνίδια, επεξεργασίας εικόνας, επεξεργασίας βίντεο και εικονικής πραγματικότητας. Η εξέλιξη στον σχεδιασμό τους και οι νέες τεχνικές που εφαρμόζονται στο πιο σημαντικό συστατικό τους, δηλαδή στον επεξεργαστή της κάρτας γραφικών ή Graphical Processing Unit ή GPU, έχουν οδηγήσει τον επιστημονικό κόσμο να τις αξιοποιεί και σε ένα διαφορετικό πλήθος εφαρμογών, εκμεταλλευόμενοι την ικανότητά τους για γρήγορους μαθηματικούς υπολογισμούς.

Στην παρούσα μελέτη θα παρακολουθήσουμε ποιες είναι αυτές οι σύγχρονες τεχνικές σχεδίασης στις κάρτες γραφικών και πως επηρεάζουν τη λειτουργία του συστήματος, αφαιρώντας αρμοδιότητες από την κεντρική μονάδα επεξεργασίας ή Central Processing Unit ή CPU.

Συγκεκριμένα, στο ΚΕΦΑΛΑΙΟ 1 θα γνωρίσουμε τις βασικές λειτουργίες της κάρτας γραφικών, θα δούμε πως αυτές εξελίχθηκαν στο χρόνο μέχρι να φθάσουμε στις προηγμένες κάρτες γραφικών των ημερών μας και θα παρακολουθήσουμε ποια είναι η κατάσταση σήμερα.

Στο ΚΕΦΑΛΑΙΟ 2 θα αναλύσουμε περισσότερο την κατάσταση που επικρατεί σήμερα στις κάρτες γραφικών και θα επικεντρωθούμε, συγκεκριμένα, στο σχεδιασμό της GPU. Αφού αναλύσουμε τις γενικές αρχές λειτουργίας ενός επεξεργαστή, θα δούμε τι από αυτά ενδιαφέρουν τις GPU και ποιες είναι οι τελευταίες εξελίξεις στον σχεδιασμό τους, αναλύοντας τον σχεδιασμό με κωδική ονομασία Vega και αυτόν με κωδική ονομασία Pascal.

Αμέσως μετά θα δούμε στο ΚΕΦΑΛΑΙΟ 3 τρόπους με τους οποίους μπορούμε να αξιολογήσουμε την επίδοση της κάρτας γραφικών. Θα αναλύσουμε όλα εκείνα τα απαραίτητα στοιχεία που κάνουν μια GPU να είναι αποδοτική και θα παρακολουθήσουμε ένα συγκεκριμένο πείραμα γραμμένο στο MatLab που αναλύει την απόδοσή της.

Στο ΚΕΦΑΛΑΙΟ 4 θα εντοπίσουμε τα όμοια χαρακτηριστικά των GPU και CPU και θα προσπαθήσουμε να εξηγήσουμε γιατί αυτά απλά μοιάζουν μεταξύ τους και για ποιους λόγους δεν μπορούν να είναι ίδια.

Όλα αυτά θα τα συγκεντρώσουμε στο ΚΕΦΑΛΑΙΟ 5 και θα καταλήξουμε στα απαραίτητα συμπεράσματα, όπου θα προσπαθήσουμε να καθορίσουμε ποιο μπορεί να είναι το μέλλον στον σχεδιασμό των GPU και γενικότερα πως αναμένεται να αλλάξει η χρήση στις κάρτες γραφικών.

ΠΡΩΤΟ ΚΕΦΑΛΑΙΟ: ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΚΑΡΤΕΣ ΓΡΑΦΙΚΩΝ

Στη σημερινή ψηφιακή εποχή η εικόνα και γενικότερα τα γραφικά παίζουν σημαντικό ρόλο και συμβάλουν στην έλξη του ανθρώπου σε αυτό τον μαγευτικό κόσμο. Οπότε, παρατηρούμε ότι η ανάγκη για καλύτερα αποτελέσματα στην εικόνα που δημιουργούν οι ηλεκτρονικοί υπολογιστές, οδηγεί στη δημιουργία καρτών γραφικών υψηλών αποδόσεων, αφού είναι αυτές που δημιουργούν την εικόνα που προσφέρει ο υπολογιστής. Στη συνέχεια θα αναφερθούμε στην έννοια της κάρτας γραφικών ως αναπόσπαστο κομμάτι κάθε υπολογιστικού συστήματος, θα παρουσιάσουμε την εξέλιξή της και θα παρακολουθήσουμε σε πιο σημείο βρισκόμαστε σήμερα.

1.1 Τι είναι οι κάρτες γραφικών

Όλοι οι ηλεκτρονικοί υπολογιστές ακολουθούν τις αρχές του John von Neumann. Σύμφωνα με αυτόν (von Neumann, 1945), κάθε υπολογιστική μηχανή χρειάζεται μια μονάδα εισόδου, που θα δίνει δεδομένα σε μια μονάδα που θα διαθέτει μια κεντρική μονάδα επεξεργασίας, με μονάδα ελέγχου και αριθμητική-λογική μονάδα, και μια μονάδα μνήμης και θα εξάγει πληροφορίες σε μια μονάδα εξόδου.

Υλοποιώντας τα παραπάνω, οι υπολογιστές σήμερα διαθέτουν μία οι περισσότερες μονάδες εισόδου και εξόδου και την μητρική κάρτα. Στην τελευταία βρίσκονται η κεντρική μονάδα επεξεργασίας (CPU), οι μνήμες RAM και ROM, οι δίαυλοι επικοινωνίας, τα διάφορα ολοκληρωμένα κυκλώματα, η μπαταρία CMOS και η κάρτα γραφικών, η οποία λαμβάνει τα δεδομένα εξόδου από τον επεξεργαστή και δημιουργεί μια σειρά από εικόνες, τις στέλνει στην οθόνη του υπολογιστή και τις λαμβάνει ο χρήστης.

Παρατηρούμε ότι ο ρόλος της κάρτας γραφικών είναι καίριος για τη λειτουργία κάθε υπολογιστικού συστήματος, αφού είναι αυτή που θα αποτυπώσει τις πληροφορίες εξόδου και χωρίς αυτή δεν μπορεί να λειτουργήσει κανένας υπολογιστής σήμερα.

Οπότε, μία κάρτα γραφικών λαμβάνει τα δεδομένα που τις στέλνει ο επεξεργαστής, εκτελεί πολύπλοκους μαθηματικούς και γεωμετρικούς υπολογισμούς, δημιουργεί πλαίσια εικόνας και τα αποστέλλει.

Για να μπορέσει να ανταποκριθεί σε αυτή τη λειτουργία χρησιμοποιεί μια πλακέτα ολοκληρωμένων κυκλωμάτων, ξεχωριστό επεξεργαστή, δική της μνήμη RAM και κύκλωμα εισόδου και εξόδου. Όλα αυτά τα στοιχεία, ανάλογα με τον τύπο της κάρτας μπορούν να διαφέρουν και θα αναλυθούν στη συνέχεια διεξοδικά, αλλά αποτελούν χαρακτηριστικά σε όλες τις κάρτες γραφικών σήμερα.

Αναλύοντας περισσότερο τα προηγούμενα στοιχεία και σύμφωνα με τους Jeff Tyson & Tracy V. Wilson (Tyson & Wilson), θα έπρεπε να δούμε τι θα συνέβαινε αν ένα υπολογιστικό σύστημα δεν έχει τη δική του κάρτα γραφικών. Αυτό που περιγράψαμε, δηλαδή, ως απαραίτητο, υπήρξε κάποια εποχή που δεν ήταν κομμάτι ενός συστήματος και σε ορισμένες περιπτώσεις εξακολουθεί να μην είναι. Για να λειτουργήσουν τέτοια συστήματα πρέπει, για αρχή, η μητρική κάρτα να έχει και να υποστηρίζει τους κατάλληλους διαύλους επικοινωνίας μεταξύ των συσκευών και το σημαντικότερο όλη την επεξεργασία των γραφικών εικόνων να την αναλαμβάνει η Κεντρική Μονάδα Επεξεργασίας (Central Processing Unit, CPU). Οπότε, τα δυαδικά στοιχεία που παράγει σε κάθε κύκλο της η CPU πρέπει να τα μετατρέπει σε εικόνα και να τα στέλνει στην οθόνη. Η διαδικασία μετατροπής της πληροφορίας από έναν δυαδικό αριθμό σε εικόνα απαιτεί μια ιδιαίτερη διεργασία, ακόμα περισσότερο όταν η πληροφορία αποτυπώνει σχήματα τα οποία πρέπει να συνθέσουν μια εικόνα, διαδικασία που ονομάζεται πιο απλά rendering και απαιτεί πολύπλοκους μαθηματικούς και γεωμετρικούς υπολογισμούς. Η CPU, ενώ είναι έτοιμη να προχωρήσει στην περαιτέρω επεξεργασίας πληροφορίας, θα διακόψει για να ασχοληθεί με την αναπαράσταση της πληροφορίας και θα το κάνει αυτό συνεχώς: υπολογισμούς και οπτικοποίηση, καταναλώνοντας ενέργεια και χρόνο.

Γενικά, το rendering στο χώρο της πληροφορικής είναι η φωτορεαλιστική ή μη-φωτορεαλιστική απεικόνιση εικόνων από ένα δισδιάστατο ή τρισδιάστατο μοντέλο που έχει παραχθεί από κάποιο λογισμικό. Στον επεξεργαστή η εικόνα φθάνει ως πληροφορία η οποία αναφέρει την γεωμετρία του χώρου, την οπτική του θεατή, τον φωτισμό, τις σκιές και τα αντικείμενα. Με βάση αυτά τα στοιχεία που μπορεί να περιγράφουν ένα 2D ή έναν 3D περιβάλλον ο επεξεργαστής πρέπει να δημιουργήσει την ψηφιακή εικόνα.

Ο τελικός στόχος είναι να σταλθεί η εικόνα στην οθόνη και να εμφανιστεί στον χρήστη. Κάθε οθόνη αποτελείται από εικονοστοιχεία (pixels) και ανάλογα το

πόσα pixels εμφανίζει οριζόντια και πόσα κάθετα καταλήγουμε ότι μία οθόνη λειτουργεί σε ανάλυση για παράδειγμα 1024x768 pixels ή μία άλλη σε 1600x900 pixels. Επίσης, το κάθε pixel παρέχει μια πλειάδα χρωμάτων τα οποία στις μέρες χωρούν πληροφορία 32bit, δηλαδή μπορούν να αποτυπώσουν 2^{32} διαφορετικά χρώματα. Τέλος, οι οθόνες των υπολογιστών λειτουργούν με τις ίδιες αρχές του κινηματογράφου, όπου παράγονται 24 καρέ το δευτερόλεπτο για να δημιουργηθεί η αίσθηση στον θεατή ότι η εικόνα που παρακολουθεί είναι κινούμενη.

Από τα τελευταία γίνεται αντιληπτό ότι μία CPU πρέπει να κάνει τους κατάλληλους υπολογισμούς για να μπορεί να δημιουργεί την πληροφορία, ώστε να μπορεί να έχει ένα σύστημα σε λειτουργία και να μπορεί ο χρήστης να λειτουργεί ένα ή περισσότερα προγράμματα. Παράλληλα, αν δεν υπάρχει η κάρτα γραφικών, πρέπει να διακόπτει αυτούς τους υπολογισμούς για να δημιουργεί την εικόνα που θα εμφανιστεί στον χρήστη και αυτή στις μέρες χρειάζεται να καταλαμβάνει αρκετό χώρο στην μνήμη (πολλά pixel με μεγάλη πληροφορία) και να στέλνεται στην οθόνη με μεγάλη συχνότητα δημιουργώντας πολλά καρέ το δευτερόλεπτο (fps, frames per second) σε έναν υπολογιστή γραφείου σήμερα να λειτουργεί η οθόνη του σε συχνότητα εικόνας 60Hz.

Σε αυτό ακριβώς το κομμάτι έρχεται να παρέμβει η λειτουργία της κάρτας γραφικών η οποία έχει το δικό της chipset που ονομάζεται Graphical Processing Unit (GPU) και τη δική της μνήμη RAM, η οποία συχνά λέγεται VRAM (Video RAM), και να απαλλάξει τον επεξεργαστή από την τόση περιττή κατανάλωση πόρων και χρόνου.

Με την κάρτα γραφικών η δυαδική πληροφορία από τον επεξεργαστή στέλνεται στην VRAM και από εκεί την επεξεργάζεται η GPU και εκτελεί το rendering. Όταν η εικόνα είναι έτοιμη στέλνεται από την VRAM στην οθόνη και την παρακολουθεί ο χρήστης.

Η χρήση της κάρτας γραφικών έχει γίνει απαραίτητη όχι μόνο στους υπολογιστές αλλά και σε ένα πλήθος άλλων συσκευών που αποτυπώνουν πληροφορία σε οθόνη, όπως κινητά τηλέφωνα και κονσόλες παιχνιδιών. Ιδιαίτερα, σε συσκευές που εκτελούνται παιχνίδια (κονσόλες και ηλεκτρονικοί υπολογιστές) ο χρήστης απαιτεί από τη συσκευή του υψηλή απόδοση αποτύπωσης εικόνας και τάχιστη εκτέλεση χωρίς κολλήματα και αργή επεξεργασία. Τα video games πάντοτε

προκαλούσαν την λειτουργία της κάρτας γραφικών και είναι αυτά και οι απαιτήσεις τους που έκαναν τους κατασκευαστές να δημιουργούν όλο και πιο προηγμένες κάρτες.

Γενικότερα, η χρήση κάρτας γραφικών σύμφωνα με τον Steven Melendez (Melendez, 2018) είναι απαραίτητη όταν έχουμε να κάνουμε με video games, με εφαρμογές επεξεργασίας video, με προγράμματα αναγνώρισης προσώπου, με συστήματα GIS (Geographical Information System), με εφαρμογές δημιουργίας 2D και 3D γραφικών, όπως το photoshop ή το lightroom, προγράμματα CAD όπως το AutoCAD, το οποίο μπορεί να χρειάζεται από την κάρτα μέχρι και 200 δισεκατομμύρια υπολογισμούς το δευτερόλεπτο για τη δημιουργία μέχρι και 17 εκατομμυρίων πολυγώνων το δευτερόλεπτο, με εξειδικευμένα προγράμματα επιστημονικών μελετών και προσομοιώσεων. Ενδιαφέρον είναι ότι μια καλή κάρτα γραφικών με ισχυρή απόδοση απαιτούν σήμερα και τα κρυπτονομίσματα και η τεχνολογία που τα περιλαμβάνει.

Σημαντικό είναι να τονίσουμε ότι μία κορυφαία κάρτα γραφικών για χρήση σε παιχνίδια απαιτεί κορυφαία χαρακτηριστικά και πολύ υψηλό fps. Ενώ, μία ιδανική κάρτα γραφικών για επαγγελματίες θα πρέπει να έχει πολύ χαμηλό χρόνο για rendering μαζί με όλα τα άλλα κορυφαία χαρακτηριστικά.

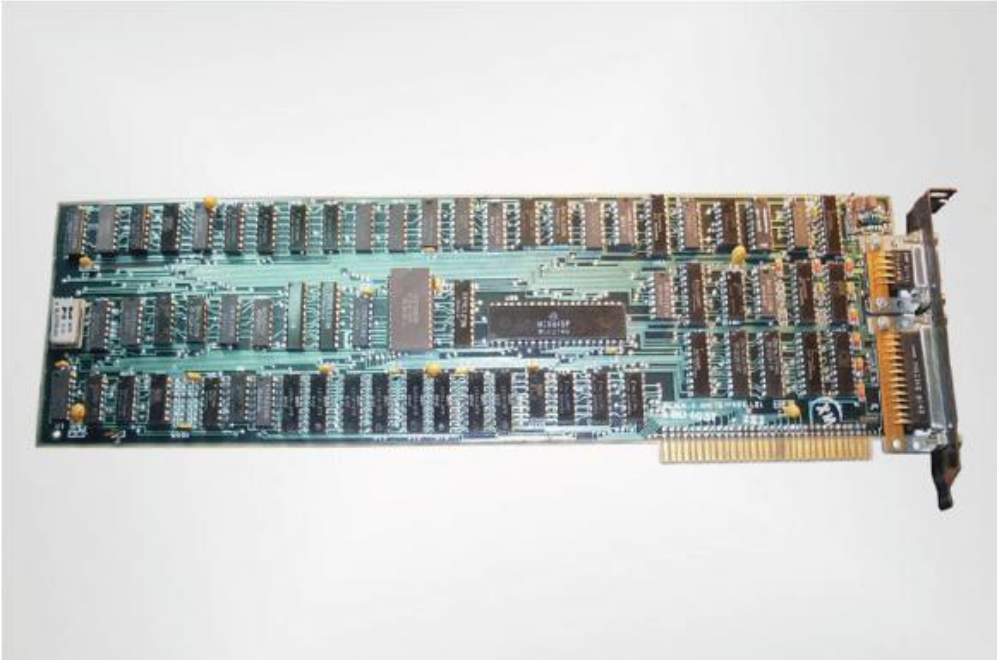
Από όλα τα παραπάνω αντιλαμβανόμαστε ότι η προσθήκη της κάρτας γραφικών σε ένα υπολογιστικό σύστημα αποφορτίζει τον επεξεργαστή από μια σειρά λειτουργίες και σε πάρα πολλές περιπτώσεις το βοηθά να ανταπεξέλθει εκεί που θα αδυνατούσε να το πράξει διαφορετικά. Παρακολουθήσαμε, επίσης, ότι η λειτουργία της κάρτας γραφικών απαιτεί υποστήριξη από την μητρική κάρτα, η οποία θα της παρέχει σύνδεση και ηλεκτρικό ρεύμα για να λειτουργήσει. Σήμερα, όλοι οι τύποι από κάρτες γραφικών υποστηρίζονται από όλες τις σύγχρονες μητρικές κάρτες. Επίσης, η κάρτα γραφικών πρέπει να περιλαμβάνει τον δικό της επεξεργαστή GPU, την δική της μνήμη VRAM και τις κατάλληλες συνδέσεις προς την οθόνη.

Όλα αυτά θα τα αναλύσουμε στη συνέχεια και θα δούμε ποιες είναι οι σύγχρονες τάσεις σχεδίασης στις κάρτες γραφικών, αφού πρώτα παρακολουθήσουμε την εξέλιξή τους, μέσα από μια σύντομη ιστορική αναδρομή.

1.2 Ιστορική αναδρομή

Στην εξέλιξη των τεχνολογιών στις κάρτες γραφικών σημαντικό ρόλο έπαιζαν, όπως γίνεται σε κάθε εμπορικό προϊόν, οι ανάγκες της αγοράς. Αυτές οι ανάγκες έκαναν τους κατασκευαστές αρχικά να δημιουργήσουν την πρώτη κάρτα γραφικών το 1981, να της προσθέσουν αργότερα τον δικό της επεξεργαστή (1999) και να φθάσουμε στο 2012 και την Nvidia να δημιουργεί εικονική GPU σε μια εικονική συσκευή. Στη συνέχεια θα παρακολουθήσουμε τις σημαντικότερες στιγμές σε αυτή την εξέλιξη.

Σύμφωνα με τον Jamie McKane (McKane, 2017) η πρώτη κάρτα γραφικών εμφανίστηκε μαζί με τον πρώτο υπολογιστή το 1981 και ήταν η IBM Monochrome Display Adapter, που παρήγαγε 80 στήλες και 25 γραμμές κειμένου χωρίς να προσφέρει την επεξεργασία που αναλύσαμε προηγουμένως.



Εικόνα 1: IBM Monochrome Display Adapter (1983)

Αργότερα, η Intel iSBX 275 VGCMB προσέφερε το 1983 ανάλυση 256 επί 256 σε οκτώ (8) διαφορετικά χρώματα και το 1988 η ATI VGA Wonder έφθανε τα 16 bit χρώματα και 2D γραφικά.

Εικόνα 2: ATI VGA Wonder (1988)

Η αλλαγή σε αυτό που θεωρούμε εμείς σήμερα κάρτα γραφικών συνέβη το 1996 με την 3dfx Voodoo 1, που έφερε τους υπολογιστές στην εποχή των 3D γραφικών με δική της μνήμη 4MB και δικό της επεξεργαστή 50MHz.

Εικόνα 3: 3dfx Voodoo 1 (1996)

Τις εξελίξεις ακολούθησε η Nvidia Riva 128 το 1997 και αυτή με 3D υποστήριξη γραφικών και η νέα βελτιωμένη 3dfx Voodoo 2 το 1998.

Καινοτόμα ήταν η Nvidia GeForce 256 DDR το 1999, έχοντας την δική της ψύξη και υποστήριξη σε ειδικό λογισμικό το DirectX 7, γεγονότα που θα σηματοδοτούσαν την εξέλιξη στις κάρτες και τα επόμενα χρόνια.

Εικόνα 4: Nvidia GeForce 256 DDR (1999)

Έτσι, το 2002 η ATI Radeon 9700 παρείχε όλες τις παραπάνω καινοτομίες και επιπλέον υποστήριξη στα λογισμικά παραγωγής γραφικών Direct3D 9.0 και OpenGL 2.0.

Εικόνα 5: ATI Radeon 9700 (2002)

Τα επόμενα χρόνια οι απαιτήσεις μεγαλώνουν και οι απαιτήσεις των προγραμμάτων και κυρίως των παιχνιδιών οδηγούν τους κατασκευαστές στη δημιουργία όλο και ταχύτερων επιδόσεων από τις κάρτες γραφικών. Το 2006 η Nvidia GeForce 8800 GTX έχει τις καλύτερες επιδόσεις και θεωρείται η καλύτερη της εποχής της.

Εικόνα 6: Nvidia GeForce 8800 GTX (2006)

Το 2009 εμφανίζεται η ATI παρουσιάζει μία από τις τελευταίες της δημιουργίες πριν εξαγοραστεί από την AMD την ATI Radeon HD 5970 με επεξεργαστή δύο πυρήνων και καταπληκτική απόδοση.

Εικόνα 7: ATI Radeon HD 5970 (2009)

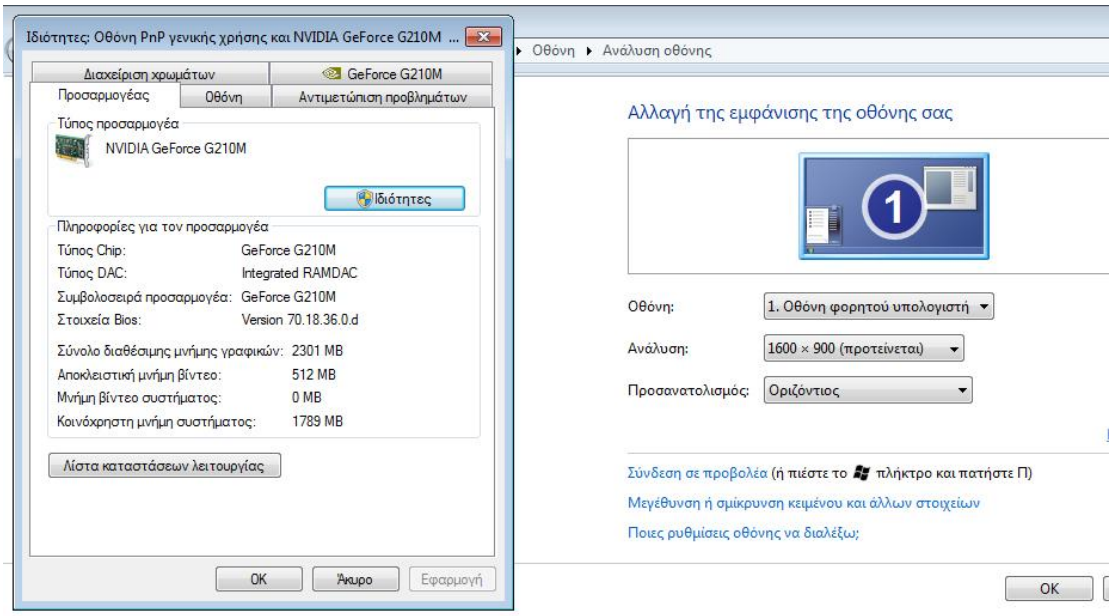
Η συνέχεια θέλει τις AMD και Nvidia, μετά την συγχώνευση της πρώτης με την εταιρία ATI και την εξαγορά από την δεύτερη της εταιρίας 3Dfx, να συναγωνίζονται διαρκώς για την κατάκτηση της αγοράς και να παρουσιάζουν συνεχώς νέες ιδέες. Έτσι, τα επόμενα χρόνια εμφανίζονται όλο και πιο γρήγορες και με καλύτερη απόδοση κάρτες γραφικών με αξιοσημείωτες τις:

- Nvidia GeForce GTX 295 το 2009 με GPU στα 40 νάνο-χιλιοστά.
- AMD Radeon HD 7970 το 2012 με GPU στα 28 νάνο-χιλιοστά και με την αρχιτεκτονική Graphics Core Next (GCN 1.0) στον σχεδιασμό της.
- Nvidia GeForce GTX 680 το 2012 με αρχιτεκτονική Kepler και GPU στα 28 νάνο-χιλιοστά, λογική που κράτησε η εταιρεία και για πολλές από τις επόμενες κάρτες της.
- AMD Radeon R9 290 το 2013, η οποία θεωρείται ισχυρή ακόμα και σήμερα.
- Nvidia GeForce GTX 970 το 2014, η οποία ήταν μία από τις πιο εμπορικά επιτυχημένες σειρές της εταιρίας.
- AMD Radeon RX 480 το 2016 με 14 νάνο-χιλιοστά GPU και πολύ προσφιλής επιλογή για αρκετούς λόγω τον υψηλών της επιδόσεων και της σχετικά χαμηλής τιμής της.
- Nvidia GeForce GTX 1080, το 2016 με 16 νάνο-χιλιοστά και αρχιτεκτονική Pascal στην GPU της, ενώ η μνήμη της έφθανε τα 8GB με την τεχνολογία GDDR5X, που την καθιστούσε εκπληκτικά γρήγορη.

Κάποια από τα παραπάνω φαίνονται χρονολογικά πολύ κοντά στο σήμερα, όμως η ταχύτατη εξέλιξη στον χώρο της πληροφορικής και ιδιαίτερα στο υλικό για ένα ιδιαίτερα εμπορικό προϊόν, το οποίο οι χρήστες συνεχώς ανανεώνουν αναζητώντας τις καλύτερες επιδόσεις, τις καθιστούν αν όχι παρωχημένες, τότε σίγουρα ξεπερασμένες από τις σημερινές τεχνολογίες. Αυτές θα τις παρακολουθήσουμε αναλυτικά στην αμέσως επόμενη ενότητα.

1.3 Οι κάρτες γραφικών σήμερα

Για να ανακαλύψουμε την κάρτα γραφικών που υπάρχει στον υπολογιστή μας, μπορούμε να μεταβούμε στις ρυθμίσεις, να επιλέξουμε σύστημα, να πάμε προχωρημένες ρυθμίσεις και μετά ιδιότητες, αν έχουμε Windows 10 ή απλά να κάνουμε δεξί κλικ στην επιφάνεια εργασίας και να επιλέξουμε ανάλυση οθόνης και έπειτα ρυθμίσεις για προχωρημένους στα Windows 7 ή ανάλογα σε παλαιότερο λειτουργικό σύστημα της Microsoft. Τότε, θα εμφανιστεί μπροστά μας μία εικόνα όπως η επόμενη.



Εικόνα 8: Τυπική κάρτα γραφικών στα Windows 7

Αντίστοιχα, σε συστήματα Mac και για την πιο πρόσφατη έκδοση του OS X, πρέπει να επιλέξουμε το μενού Apple στην πάνω αριστερή γωνία, να διαλέξουμε σχετικά και την αναφορά συστήματος. Εκεί θα περιηγηθούμε στις οθόνες και σε αυτή του υλικού θα επιλέξουμε Γραφικά/Ανάλυση. Η Apple δημιουργεί τις δικές της κάρτες γραφικών. Οπότε στη συνέχεια θα αναφερθούμε στις κάρτες γραφικών των προσωπικών υπολογιστών, όπως είναι αυτή της εικόνας 1.

Σε αυτήν την εικόνα βλέπουμε τον τύπο της κάρτας γραφικών, ο οποίος στη συγκεκριμένη περίπτωση είναι ο NVIDIA GeForce G210M, δηλαδή μία κάρτα γραφικών που υπάρχει στην αγορά εδώ και χρόνια και είναι της εταιρίας NVIDIA. Κάθε κάρτα γραφικών πρέπει να έχει τον δικό της επεξεργαστή GPU, ο οποίος στις μέρες είναι κατασκευασμένος από τις εταιρίες NVIDIA ή AMD και σε ταχύτητες μπορείς να φθάνει τα 1500MHz. Όπως παρακολουθήσαμε η μνήμη της κάρτας

γραφικών ονομάζεται συχνά και VRAM και μας προσφέρει υψηλότερη ανάλυση και περισσότερα χρώματα στην οθόνη, αν η τελευταία τα υποστηρίζει, και καλύτερα ειδικά εφέ. Οι πιο συνηθισμένες σύγχρονες κάρτες διαθέτουν μνήμη από 128MB έως και 2GB, ενώ οι κορυφαίες κάρτες ξεπερνούν τα 10GB, όπως η GeForce RTX 2080Ti με επεξεργαστή στα 1350 MHz και μνήμη 11GB GDDR6, όπου το τελευταίο χαρακτηριστικό υποδηλώνει την γενιά της μνήμης και ουσιαστικά μας δείχνει πόσο γρήγορα λειτουργεί η μνήμη, αφού το βασικό χαρακτηριστικό της VRAM είναι να όχι μόνο το μέγεθος έτσι ώστε να αντέχει τις εφαρμογές μας, αλλά και η ταχύτητά της.

Πέρα από αυτά τα βασικά χαρακτηριστικά, μία σύγχρονη κάρτα γραφικών σύμφωνα με τον Joel Hruska (Hruska, 2018) πρέπει να υποστηρίζει και ένα σύγχρονο πρόγραμμα δημιουργίας γραφικών μέσα από το rendering. Τέτοιο σήμερα θεωρείται το DirectX 11.0 και όπως και η προηγούμενες από αυτό γενιές του, τέτοια προγράμματα τα καλεί το περιβάλλον του λειτουργικού συστήματος και εκτελεί στον επεξεργαστή της κάρτας γραφικών όλες εκείνες τις μαθηματικές πράξεις που απαιτούνται ώστε τα δυαδικά στοιχεία να γίνουν σχήματα και να δημιουργήσουν την τελική εικόνα. Ορισμένα παιχνίδια απαιτούν και την εγκατάσταση και άλλου παρόμοιου προγράμματος όπως το OpenGL για την παραγωγή πλούσιων γραφικών. Από την άποψη του λογισμικού, τέλος, πολλές κάρτες έχουν το δικό τους διαχειριστικό περιβάλλον μέσα από το οποίο ο χρήστης μπορεί να προβεί σε απλές ή ποιο σύνθετες ρυθμίσεις της λειτουργίας τους. Ένα τέτοιο λογισμικό αφορά αποκλειστικά και μόνο την αλληλεπίδραση με τον χρήστη και δεν επηρεάζει σε κανένα βαθμό την απόδοση της κάρτας.

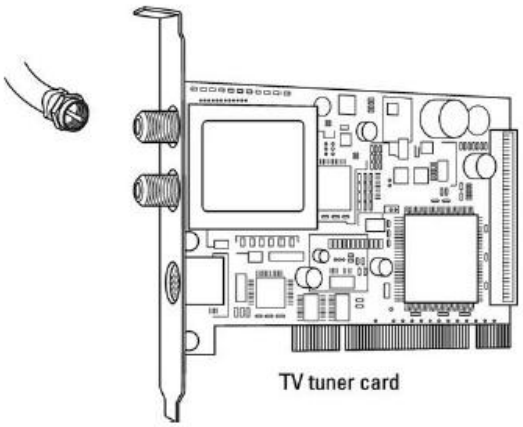
Πέρα από όλα αυτά, μια σύγχρονη κάρτα γραφικών πρέπει να υποστηρίζει και τις κατάλληλες διασυνδέσεις με την συσκευή την οποία ουσιαστικά υπηρετεί, δηλαδή την οθόνη. Έτσι, πρέπει να είναι σε θέση να έχει διαθέσιμες συνδέσεις VGA παλαιότερα και DVI ή HDMI σήμερα ή ότι άλλο απαιτεί η οθόνη για να συνδεθεί μαζί της. Σε αρκετές περιπτώσεις πρέπει να υποστηρίζεται και η υποστήριξη πολλαπλών οθονών, αφού αρκετοί σήμερα εργάζονται πιο αποτελεσματικά σε δύο ή περισσότερες οθόνες.

Σημαντικό, βέβαια, είναι να υποστηρίζει και τις υψηλότερες αναλύσεις που υποστηρίζει μια σύγχρονη οθόνη, δηλαδή ανάλυση 2560x1600, ανάλυση ευρείας

οθόνης 16:9, ανάλυση 21:9 όπως είναι η ανάλυση 2560x1080 και τελευταία υψηλή ανάλυση 4K και 8K.

Για όλες αυτές τις αναλύσεις υπάρχει το πρότυπο VESA (Video Electronics Standards Association) το οποίο ενημερώνει τους κατασκευαστές ποιες είναι οι προτεινόμενες αναλύσεις οθόνης και τον οποίο όλοι ακολουθούν. Δηλαδή, υπάρχει η λίστα με τις επιτρεπόμενες αναλύσεις οθόνης (800x600, 1024x768 κλπ.) την οποία ακολουθούν οι κατασκευαστές οθόνης και οι κατασκευαστές καρτών γραφικών. Οπότε, δεν θα υπάρξει οθόνη που λειτουργεί σε ανάλυση που δεν έχει προβλεφθεί από τον κατασκευαστή της κάρτας γραφικών, αλλά οθόνη που θα μπορούσε να λειτουργήσει σε υψηλότερη ανάλυση και η κάρτα δεν την υποστηρίζει. Από την άλλη, δεν μπορεί να υπάρξει κάρτα γραφικών με ανάλυση που η οθόνη δεν διαθέτει, αλλά μπορεί η κάρτα να υποστηρίζει και υψηλότερες αναλύσεις, αλλά η οθόνη να υποστηρίζει μόνο τις χαμηλότερες.

Ακόμα, αρκετές σύγχρονες κάρτες διαθέτουν TV output και υποστήριξη TV Tuner ώστε ο χρήστης να μπορεί να συνδέσει το σήμα της τηλεόρασης στην οθόνη του. Ενώ μια σύγχρονη κάρτα γραφικών πρέπει να υποστηρίζει κωδικοποίηση βίντεο mpeg, avi και mov. Με σημαντικότερη την κωδικοποίηση mpeg, αυτές είναι οι συνηθέστερες κωδικοποιήσεις αρχείων βίντεο και όταν ένα σύστημα τα δημιουργεί και κωδικοποιεί ή τα αναπαράγει και αποκωδικοποιεί, είναι σημαντικά ταχύτερο αυτές οι διαδικασίες να υποστηρίζονται από την κάρτα γραφικών, να εκτελούνται από αυτήν και να γίνονται αρκετά ταχύτερα από το να τις εκτελούσε η CPU.



Εικόνα 9: Κάρτα TV tuner

Από την πλευρά των κατασκευαστών, στις μέρες υπάρχουν δύο που καλύπτουν το σύνολο της αγοράς, η AMD με μερίδιο 35,5% και η Nvidia με το υπόλοιπο 64,5% (techpowerup, 2013). Αυτές κατασκευάζουν το σύνολο των επεξεργαστών γραφικών και προϊστανται στην έρευνα για την περαιτέρω εξέλιξη των τεχνολογιών τους. Όμως, πέρα από τον επεξεργαστή και φυσικά την μνήμη της κάρτας γραφικών, για να λειτουργήσει η όλη διαδικασία απαιτείται η κάρτα να λαμβάνει ρεύμα, να έχει σύστημα ψύξης και άσχετα με την απόδοσή της να έχει ξεχωριστή εμφάνιση για να προσελκύσει τον αγοραστή. Οι εταιρίες που αναφέραμε δίνουν σε άλλες, που τις ονομάζουν συνεργάτες τους, τη δυνατότητα να δημιουργήσουν τη δική τους σύνθεση, να την προωθήσουν και να την προσφέρουν στον καταναλωτή. Χαρακτηριστικό είναι ότι υπάρχουν εταιρίες που συνεργάζονται και με τους δύο κατασκευαστές, προωθώντας τα προϊόντα τους. Έτσι, έχουμε την εταιρία Palit Microsystems, την PC Partner, που κυκλοφορεί κάρτες με επεξεργαστή AMD κάτω από το όνομα Sapphire και κάρτες με επεξεργαστή Nvidia κάτω από την ονομασία Zotac, την Asus, την MSI (Micro-Star International), την Gigabyte Technology, την Brea, την EVGA και την XFX όλες να κυκλοφορούν διαφορετικές κάρτες γραφικών με επεξεργαστές από τις ίδιες δύο εταιρίες.

Καταλήγοντας, οι κάρτες γραφικών είναι σημαντικότατο κομμάτι στη λειτουργία των ηλεκτρονικών υπολογιστών και σήμερα η εξέλιξή τους εμφανίζεται ραγδαία, με σημαντική έρευνα στις τεχνολογίες που αξιοποιούν και συνεχώς βελτιούμενες κυκλοφορίες νέων επεξεργαστών. Στα επόμενα κεφάλαια θα αναλύσουμε τα χαρακτηριστικά αυτών των καρτών, εστιάζοντας στο πιο σημαντικό από αυτά, δηλαδή τον επεξεργαστή τους.

ΔΕΥΤΕΡΟ ΚΕΦΑΛΑΙΟ: ΣΥΓΧΡΟΝΕΣ ΤΑΣΕΙΣ ΣΧΕΔΙΑΣΗΣ ΣΤΙΣ ΚΑΡΤΕΣ ΓΡΑΦΙΚΩΝ

Στο προηγούμενο κεφάλαιο παρακολουθήσαμε βασικές αρχές λειτουργίας και την εξέλιξη στις κάρτες γραφικών ανά τα χρόνια. Ολοκληρώνοντας είδαμε ποιες είναι οι βασικές χρήσεις τους και σε ποιο σημείο βρίσκεται η τεχνολογία τους μέσα από μερικά βασικά χαρακτηριστικά τους.

Στη συνέχεια θα αναλύσουμε περισσότερο τη λειτουργία στις κάρτες γραφικών, αναλύοντας περαιτέρω τα χαρακτηριστικά σύνδεσης με τη μητρική κάρτα και την οθόνη, ώστε να καταλήξουμε στο σημαντικότερο χαρακτηριστικά της κάρτας γραφικών, τον επεξεργαστή της (Graphical Processing Unit, GPU) και για την οποία θα παρουσιάσουμε τις σημαντικότερες αρχιτεκτονικές της.

2.1 Χαρακτηριστικά στις κάρτες γραφικών

Οι κάρτες γραφικών είναι μια πλακέτα ολοκληρωμένου κυκλώματος και συνδέεται σε ένα slot της μητρική κάρτας. Μέσα από αυτό το slot ανταλλάσσει τα δεδομένα με την μητρική και αυτή της παρέχει και μια ελάχιστη παροχή ρεύματος μέχρι τα 75 Watt για να λειτουργήσει. Στις μέρες μας, για να συνδεθεί μια κάρτα γραφικών χρειάζεται από την μητρική ένα slot τύπου PCI Express (PCI-E), το οποίο όλες οι σύγχρονες μητρικές διαθέτουν. Το PCI Express λειτουργεί ως ενδιάμεσο και σαν τεχνολογία χρησιμοποιείται από το 2004.

Σαν αποτέλεσμα όλων αυτών, όλες οι σύγχρονες κάρτες γραφικών μπορούν να ταιριάξουν με όλες τις σύγχρονες μητρικές κάρτες. Μοναδική εξαίρεση αποτελεί η επιλογή να λειτουργήσουμε δύο ή περισσότερες κάρτες γραφικών ταυτόχρονα, ώστε να συνδυάσουμε την ισχύ τους και να πετύχουμε καλύτερες αποδόσεις. Αυτή η τεχνική απαιτεί και έναν δίαυλο μεταξύ των συνεργαζόμενων καρτών και ονομάζεται Crossfire στην περίπτωση των GPU τύπου Nvidia ή SLI στην περίπτωση των AMD. Είναι προφανές ότι σε τέτοιες περιπτώσεις απαιτείται από την μητρική κάρτα η ύπαρξη συνεχόμενων ελευθέρων slot ώστε να κουμπώσουν οι συνεργαζόμενες κάρτες γραφικών.

Γενικότερα, έχουν υπάρξει κατά καιρούς διαφορετικές τεχνολογίες στον δίαυλο επικοινωνίας μεταξύ μητρικής και κάρτας γραφικών. Στην επόμενη λίστα τις παραθέτουμε χρονολογικά:

- ISA XT
- ISA AT
- MCA
- NUBUSEISA 32 8.33 32 Parallel
- VESA
- PCI
- AGP
- AGP
- AGP
- AGP
- PCIe x1
- PCIe x4
- PCIe x8
- PCIe x16
- PCIe x1 2.0
- PCIe x4 2.0
- PCIe x8 2.0
- PCIe x16 2.0
- PCIe X1 3.0
- PCIe X4 3.0
- PCIe X8 3.0
- PCIe X16 3.0

Προηγουμένως, παρακολουθήσαμε ότι μέσω της PCI-E η μητρική δίνει μια παροχή ρεύματος στην κάρτα γραφικών ως 75 Watt. Όμως, πολλές σύγχρονες κάρτες γραφικών απαιτούν πολύ περισσότερο, δηλαδή ως και 250 Watt, γεγονός που τις καθιστά τα μέρη του υπολογιστή με την μεγαλύτερη κατανάλωση. Για να καλύψουν τις ανάγκες τους μερικές κάρτες έχουν δική τους σύνδεση στην παροχή ρεύματος.

Όμως, επειδή αυξάνουμε συνεχώς την παροχή ρεύματος και ακόμα περισσότερο επειδή αυξάνονται συνεχώς οι ταχύτητες λειτουργίας της GPU, απαιτείται και ένα σύστημα ψύξης. Αυτό δεν ήταν απαραίτητο μέχρι τη δεκαετία του ’90, αλλά οι πιο σύγχρονες κάρτες απαιτούν το δικό τους σύστημα εξισορρόπησης θερμοκρασίας, το οποίο μπορεί να είναι ανεμιστήρας, ψήκτρα ή υδρόψυξη. Ακόμα

και αν αυτό δεν παρέχεται από τον κατασκευαστή, συνήθως προστίθεται ως επιπλέον εξάρτημα πάνω από την GPU.

Πριν προχωρήσουμε βαθύτερα στον τρόπο λειτουργίας μια κάρτας γραφικών, πρέπει να τονίσουμε ότι οι σύγχρονες απαιτήσεις αποτύπωσης από 3D περιβάλλοντα σε 2D εικόνες είναι μια περίπλοκη διαδικασία, που αναφέρθηκε προηγουμένως ως rendering και απαιτεί ειδικά προγράμματα. Αυτά τα καλεί το λειτουργικό σύστημα και τα πιο γνωστά στο λειτουργικό Windows είναι το Direct X και το OpenGL, ενώ άλλα που λειτουργούν σε περιβάλλον Windows, στο λειτουργικό της Apple και σε κονσόλες παιχνιδιών όπως το Nintendo ή το Xbox είναι τα Vulkan, GNMX, Metal και OpenGL ES.

Προχωρώντας παραπέρα θα εξετάσουμε τα υπόλοιπα χαρακτηριστικά της κάρτας γραφικών, πέρα από τον επεξεργαστή της, τον οποίο θα εξετάσουμε εκτενέστερα σε ειδική ενότητα, μιας και αποτελεί το ουσιαστικότερο κομμάτι της.

Ένα από αυτά τα χαρακτηριστικά αποτελεί και το BIOS της κάρτας γραφικών ή το video BIOS ή VBIOS. Αυτό αποτελεί πρακτικά το firmware της κάρτας και είναι αυτό που θα την εκκινήσει, θέτοντας της αρχικές ρυθμίσεις ώστε να λειτουργήσει. Ο ρόλος του είναι σχετικά περιορισμένος.

Σύμφωνα με τον Mark Coppock (Coppock, 2018), πιο σημαντικό ρόλο έχει η μνήμη της κάρτας, η οποία λειτουργεί ως μια τυπική μνήμη RAM, αποθηκεύοντας την πληροφορία ώστε να επεξεργαστεί από την GPU και λειτουργώντας ως ενδιάμεσος μετά την μετατροπή της σε εικόνα. Πρέπει να είναι αρκετά μεγάλη ώστε να ανταποκρίνεται στις απαιτήσεις του χρήστη και των προγραμμάτων του και αρκετά γρήγορη ώστε να μην καθυστερεί την GPU. Το θέμα του χώρα είναι εύκολα υπολογίσιμο, αφού για την αγορά κάθε κάρτας γραφικών ο πελάτης ενημερώνεται για το μέγεθός της. Οι σύγχρονες κάρτες έχουν μεγέθη RAM ή VRAM, όπως ονομάζεται, από 1GB ως και 12GB. Όσον αφορά την ταχύτητα, αυτή έχει να κάνει με την τεχνολογία που υλοποιούν και μπορεί να είναι τύπου DDR3 ή η πιο σύγχρονη GDDR5X. Όλοι οι τύποι μνήμης VRAM και τα χαρακτηριστικά τους μπορούν να φανούν στον επόμενο πίνακα. Η ένδειξη ρολόι αναφέρεται στον κύκλο ρολογιού της μνήμης και αποτελεί την ουσιαστικότερη ένδειξη της ταχύτητάς της, αφού κάθε VRAM μπορεί να μεταφέρει ένα πλαίσιο πληροφορίας σε κάθε κύκλο, οπότε όσο πιο γρήγορα εκτελούνται αυτοί οι κύκλοι, τόσο πιο γρήγορα μεταφέρεται η πληροφορία.

Αντίστοιχο ρόλο σε αυτό έχει και το μέγεθος της πληροφορίας που επεξεργάζεται σε μια μονάδα χρόνου, αφού αν είναι αρκετά μεγάλο επεξεργάζεται περισσότερη πληροφορία και η διαδικασία ολοκληρώνεται πιο σύντομα και με μεγαλύτερη ταχύτητα. Αυτό μετριέται σήμερα σε GB ανά δευτερόλεπτο. Όλα αυτά τα χαρακτηριστικά για κάθε τύπου τεχνολογίας μνήμης VRAM φαίνονται στον επόμενο πίνακα.

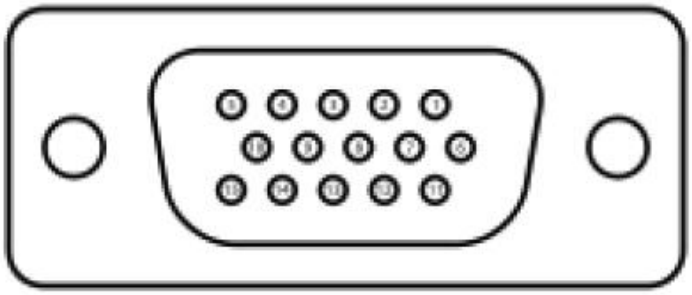
Πίνακας 1: Τύποι μνήμης VRAM και χαρακτηριστικά τους

Τύπος	Ρολόι (σε MHz)	Bandwidth (σε GB/s)
DDR	200 - 400	1,6 – 3,2
DDR2	400 - 1066.67	3,2 – 8,533
DDR3	800 - 2133.33	6,4 – 17,066
DDR4	1600 - 4866	12,8 – 25,6
GDDR4	3000 - 4000	160 – 256
GDDR5	1000 - 2000	288 – 336,5
GDDR5X	1000 - 1750	160 – 673
HBM	250 - 1000	512 – 1024

Είναι φανερό ότι η ταχύτητα μιας μνήμης τύπου GDDR5 είναι δύο φορές πιο γρήγορη από μία τύπου DDR3. Οπότε, αν έχουμε 1 GB GDDR5 είναι προτιμότερο, τις περισσότερες φορές, από 4 GB DDR3.

Άλλο ένα χαρακτηριστικό, το οποίο τείνει να μην μας απασχολεί πλέον είναι η μνήμη αποκωδικοποίησης (RAMDAC, Random Access Memory Digital to Analog Converter), η οποία λαμβάνει την εικόνα από την κάρτα γραφικών την μετατρέπει σε αναλογικό σήμα και την αποστέλλει στην οθόνη. Όλη αυτή η διαδικασία δεν έχει καμία χρησιμότητα στις μέρες μας, αφού ελάχιστες οθόνες λειτουργούν με αναλογικό σήμα και τεχνολογία καθοδικού σωλήνα. Παρόλα αυτά υπάρχουν οθόνες ακόμα και σήμερα στις οποίες το σήμα μεταδίδεται με VGA, όπως θα δούμε ακριβώς στη συνέχεια, αναλογικά, το ξαναμετατρέπουν στην ψηφιακή του μορφή και το προβάλλουν στον χρήστη. Αυτές οι οθόνες, ακόμα και σήμερα, αξιοποιούν το chip της μνήμης RAMDAC.

Τέλος, σημαντικό χαρακτηριστικό μιας κάρτας γραφικών είναι η έξοδος που δίνει προς την οθόνη και την τεχνολογία που υποστηρίζει για αυτή τη διαδικασία. Από την δεκαετία του ’80 και μέχρι σήμερα χρησιμοποιείται αναλογικό σήμα και έξοδος με το πρότυπο VGA (Video Graphics Array), όπως της εικόνας. Αυτό στις ημέρες μας υποστηρίζει υψηλές αναλύσεις, μεγάλη ευκρίνεια και χρησιμοποιείται αρκετά ακόμα και από σύγχρονες LCD οθόνες που λειτουργούν με ψηφιακό σήμα.

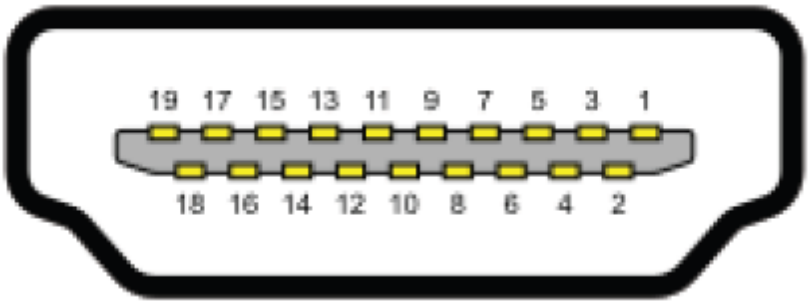


Εικόνα 10: Πρότυπο VGA

Οι πιο σύγχρονες οθόνες υποστηρίζουν το πρότυπο DVI (Digital Visual Interface), με το οποίο μεταφέρεται ψηφιακό το σήμα και η εικόνα αντιστοιχίζεται pixel προς pixel από τον υπολογιστή στην οθόνη. Είναι χαρακτηριστικό ότι αν η οθόνη ή η κάρτα γραφικών δεν υποστηρίζει αυτό το πρότυπο, τότε μπορεί να προστεθεί ειδικός προσαρμογέας και να μετατρέψει την έξοδο DVI σε αναλογική VGA.

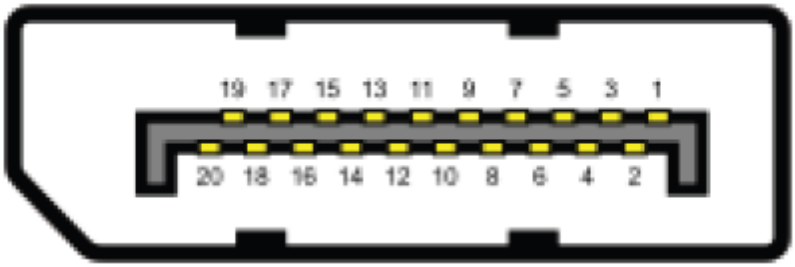
Εικόνα 11: Πρότυπο DVI

Αυτά τα πρότυπα οδεύουν προς αντικατάσταση από δύο νέα ψηφιακά. Το πρώτα από αυτά είναι το HDMI (High Definition Multimedia Interface), το οποίο χρησιμοποιούν σήμερα οι περισσότερες κάρτες γραφικών και το άλλο, επίσης ψηφιακό είναι το DisplayPort.



Εικόνα 12: Το πρότυπο HDMI

Αξίζει να σημειώσουμε ότι οι σύγχρονες κάρτες έχουν εξόδους σε διάφορα πρότυπα αν και το HDMI και ίσως το DisplayPort είναι αυτά τα οποία δείχνουν ότι θα κυριαρχήσουν στο μέλλον.



Εικόνα 13: Το πρότυπο DisplayPort

Όλα αυτά τα χαρακτηριστικά πρέπει να τα λάβουμε υπόψη μας όταν είναι να αξιολογήσουμε μια κάρτα γραφικών. Το σπουδαιότερο όμως από όλα είναι αυτό το οποίο θα αναφέρουμε στη συνέχεια και είναι η κεντρική μονάδα επεξεργασίας της κάρτας γραφικών (GPU). Αυτήν θα την αναλύσουμε στη συνέχεια και θα παρουσιάσουμε τις αρχιτεκτονικές που προτείνουν οι δύο κατασκευαστές GPU σήμερα, η Nvidia και η AMD.

2.2 Αρχιτεκτονικές και ολοκληρωμένα κυκλώματα στη διαχείριση γραφικών

Αφού παρακολουθήσαμε το σύγχρονο σχεδιασμό μιας κάρτας γραφικών, ήρθε η στιγμή να ανακαλύψουμε και τη δομή του κύριου συστατικού της, του επεξεργαστή της. Μέχρι τώρα παραλείψαμε εσκεμμένα να αναφερθούμε σε αυτόν, ώστε να τον αναλύσουμε διεξοδικά σε ένα ενιαίο μέρος της παρούσας εργασίας, όντας και το πιο σημαντικό χαρακτηριστικό αυτών των καρτών.

Καταρχήν, με την έννοια της αρχιτεκτονικής στους επεξεργαστές εννοούμε την μικροαρχιτεκτονική του επεξεργαστή, δηλαδή την οργάνωσή του μέσα από το μοντέλο Instruction Set Architecture ή ISA και πως αυτό υλοποιείται κάθε φορά. Αυτό το μοντέλο περιγράφει τη δομή που πρέπει να έχει κάθε μέρος του υπολογιστή και η μικροαρχιτεκτονική είναι η υλοποίησή του. Οπότε, θα παρακολουθήσουμε την δομή της GPU και τις σύγχρονες υλοποιήσεις της.

Για να κατανοήσουμε πληρέστερα την λειτουργία της θα ξεκινήσουμε την περιγραφή της από την ανάλυση της λειτουργίας μιας κεντρικής μονάδας επεξεργασίας, CPU.

Σύμφωνα με τους Christopher Cullinan, Christopher Wyant και Timothy Frattesi (Cullinan, Wyant, Frattesi) ο σχεδιασμός μιας CPU πρέπει να περιλαμβάνει έξι συγκεκριμένους τομείς:

- τη μονάδα δεδομένων,
- τη μονάδα ελέγχου,
- την μνήμη,
- τον κύκλο του ρολογιού,
- την πλακέτα λειτουργίας και
- την λογική μονάδα.

Ξεκινώντας από το πρώτο, αυτό πρέπει να προσφέρει τους διαύλους επικοινωνίας μεταξύ των επιμέρους μερών και μαζί με τη μονάδα ελέγχου να πραγματοποιεί τους αριθμητικούς υπολογισμούς, με την μονάδα ελέγχου να εξειδικεύεται στον έλεγχο αυτών των πράξεων μαζί με τον έλεγχο της μνήμης.

Οι περισσότερες μοντέρνες CPU δεν έχουν έναν, αλλά έχουν περισσότερους τύπους από μνήμες στο ίδιο chip. Οι πιο σημαντικές από αυτές είναι οι μνήμες cache και register, οι οποίες είναι υψηλής ταχύτητας μνήμες τύπου SRAM. Οι μνήμες register επικοινωνούν απευθείας με την CPU και προσφέρουν σε αυτή τα δεδομένα που χρειάζεται για να πραγματοποιήσει τους υπολογισμούς. Από την άλλη, η μνήμη cache είναι η αμέσως επόμενη μνήμη που χρησιμοποιεί ο πυρήνας για τους υπολογισμούς και συνήθως συναντάται σε cache δύο κατηγοριών, της κατηγορίας ένα ή L1 (Level 1) και της κατηγορίας δύο ή L2 (Level 2).

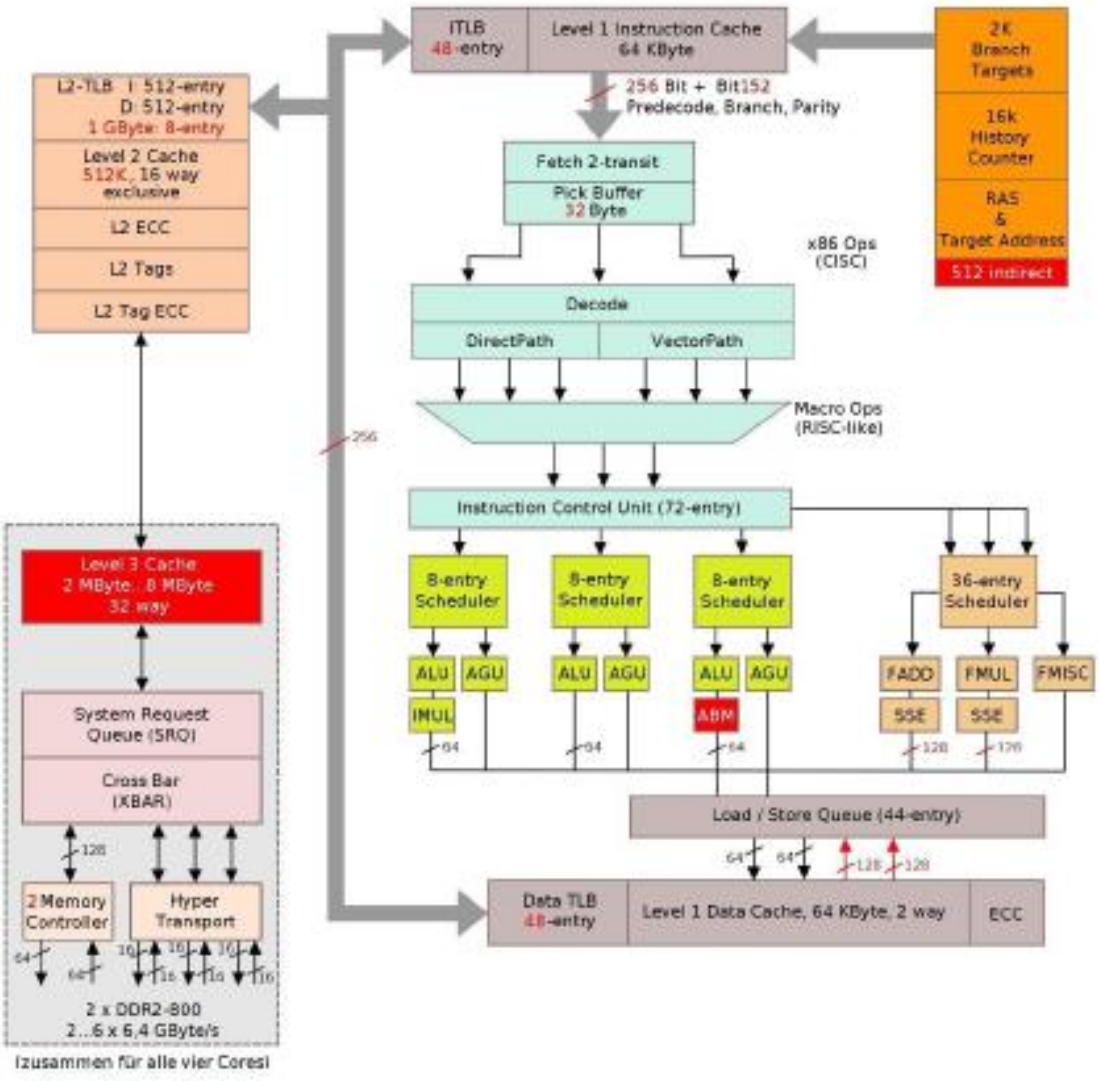
Το ρολόι της CPU είναι ένα περιοδικό σήμα, το οποίο δημιουργείται από έναν κρύσταλλο και χρησιμοποιείται για να συγχρονίσει τις ενέργειες των επιμέρους κομματιών σε μία CPU. Οπότε, αυτά γνωρίζουν λαμβάνουν το σήμα μέσα από ένα ειδικό δίκτυο μεταφοράς και μπορούν να γνωρίζουν πόσο χρόνο έχουν για να εκτελέσουν μία διαδικασία και το σημαντικότερο σε πόσο χρόνο θα έχουν εκτελέσει τα άλλα μέρη την δική τους.

Η λογική μονάδα είναι μια συλλογή από λογικές πύλες και χρησιμοποιείται για να πραγματοποιήσει η CPU τις λογικές της πράξεις. Σε αυτήν υπάρχουν οι πύλες AND, OR και NOT, τα flip-flops και οι buffers. Είναι χαρακτηριστικό ότι η λογική μονάδα υλοποιείται στο φυσικό χώρο σε περιορισμένο ύψος, αλλά μεταβλητό πλάτος και έτσι απλώνεται στον χώρο της πλακέτας και λειτουργεί πιο αποτελεσματικά.

Τέλος, η πλακέτα λειτουργίας είναι αυτή η πλακέτα στην οποία λειτουργούν όλα τα παραπάνω.

Μια υλοποίηση όλων των παραπάνω φαίνεται στην επόμενη εικόνα, όπου εμφανίζονται τα χαρακτηριστικά μέρη μιας CPU της AMD και αρχιτεκτονικής K10 του 2007, ο οποίος στις μέρες μας θεωρείται ξεπερασμένος, αλλά μπορούμε μέσα στην απλότητά του να παρατηρήσουμε τα επιμέρους μέρη του. Σε αυτή την, ιδιαίτερα, φαίνονται οι αλλαγές με κόκκινα γράμματα που έφερε η αρχιτεκτονική K10 σε σχέση με την παλαιότερη αρχιτεκτονική K8.

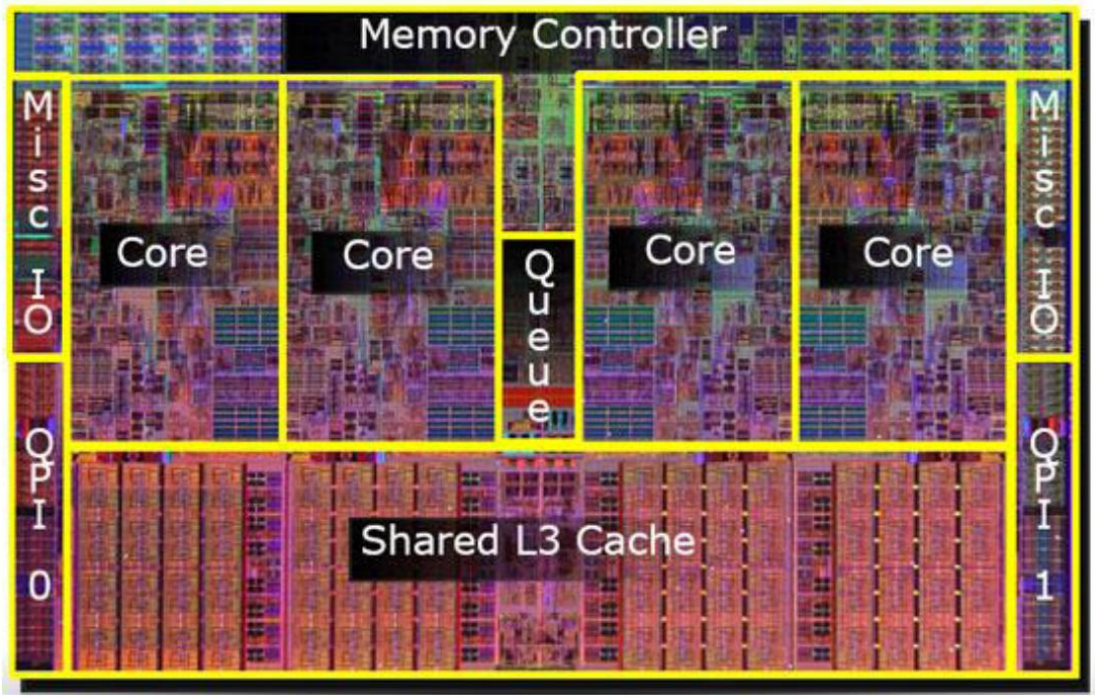
Μία εξέλιξη όλων αυτών είναι οι πολυπύρηνοι επεξεργαστές που κυριαρχούν σήμερα. Αυτοί είναι πολλοί διαφορετικοί επεξεργαστές, οι οποίοι συνεργάζονται και λειτουργούν ως ένας για να εξυπηρετήσουν έναν μοναδικό υπολογιστικό σύστημα, μέσα από τη λογική της παράλληλης επεξεργασίας, δηλαδή, των προσομοιωμένων ταυτόχρονων υπολογισμών.



Εικόνα 14: Μίκρο-αρχιτεκτονική CPU τύπου AMD K10.

Η παράλληλη επεξεργασία βασίζεται στην απλή σκέψη ότι τα δυσκολότερα προβλήματα λύνονται ευκολότερα αν διαιρεθούν σε απλούστερα μέρη και έχει γίνει ο κυρίαρχος τρόπος υλοποίησης CPU σήμερα.

Ένα τέτοιο παράδειγμα φαίνεται στην επόμενη εικόνα με τον τετραπύρηνο επεξεργαστή της Intel, τον I7 950 Quad Core Processor.



Εικόνα 15: Τετραπύρηνη CPU τύπου Intel I7 950

Αυτές είναι και οι αρχές στις οποίες βασίζεται η CPU.

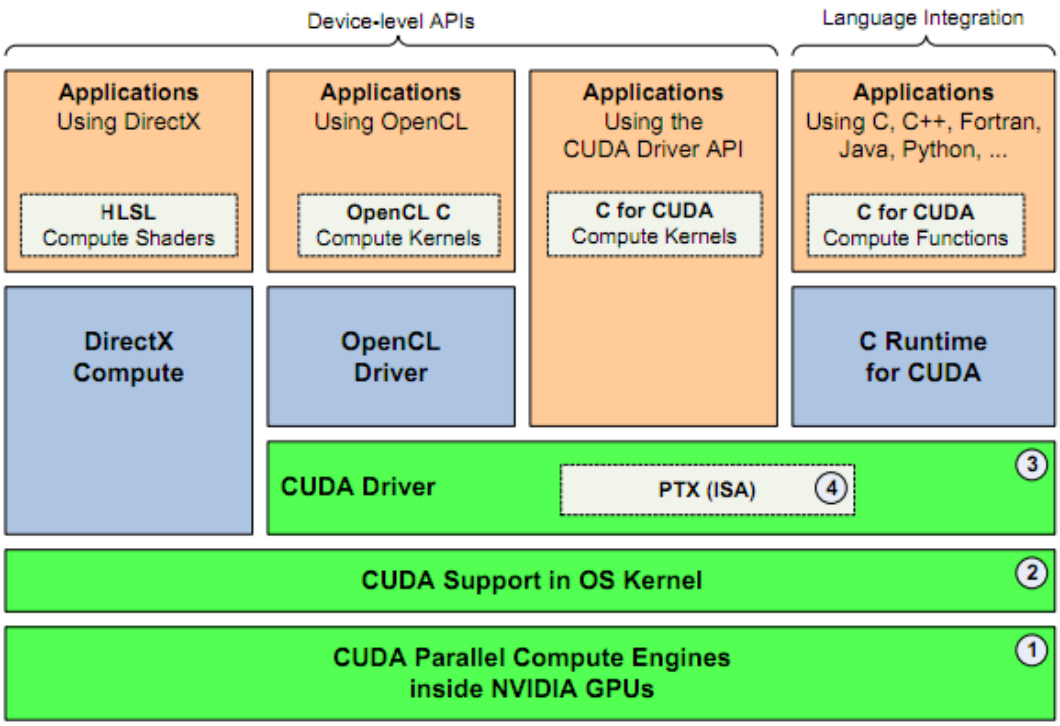
Εικόνα 16: Εικόνα 16: Streaming Multiprocessor της GPU.

Από την άλλη η GPU σχεδιάζονται σήμερα ώστε να ικανοποιούν τις ανάγκες τους, οι οποίες κυρίως είναι να μπορούν να εκτελούν όσο το δυνατό περισσότερους

υπολογισμούς και να ολοκληρώνουν όσο πιο πολλούς από αυτούς ταυτόχρονα. Για το λόγο αυτό, το βασικό δομικό τους στοιχείο είναι ο Streaming Multiprocessor ή SM.

Ο SM με τη σειρά του αποτελείται από ALUs (Αριθμητικές Λογικές Μονάδες) ή CUDA-cores, όπως τις ονομάζει η Nvidia και οι οποίες εκτελούν στην πράξη τις μαθηματικές πράξεις που απαιτούνται. Το μοντέλο αυτό της λειτουργίας το ονομάζει η Nvidia Single Instruction Multiple Threads (SMTP), ενώ η ίδια εταιρία έχει δημιουργήσει με το ίδιο όνομα λογισμικό για τη δημιουργία προγραμμάτων που έχουν απευθείας πρόσβαση στον ALU και μία γλώσσα προγραμματισμού CUDA παραλλαγή της γλώσσας προγραμματισμού C.

Στην επόμενη εικόνα φαίνεται η αρχιτεκτονική του λογισμικού και του υλικού CUDA που το υποστηρίζει.

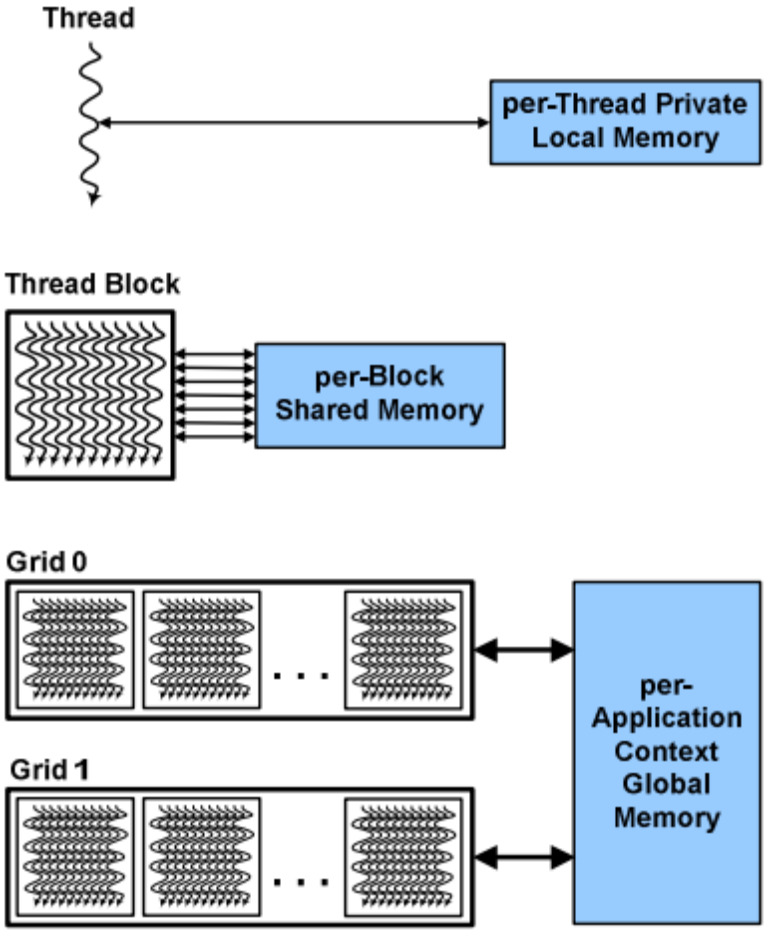


Εικόνα 17: Αρχιτεκτονική λογισμικού και υλικού CUDA

Στις GPU η παράλληλη επεξεργασία υλοποιείται με την χρήση νημάτων, αντί για τη χρήση επιπλέων πυρήνων, όπως συμβαίνει στις σύγχρονες CPU. Για παράδειγμα, μια σύγχρονη κάρτα γραφικών Nvidia μπορεί να έχει 1024 πυρήνες CUDA σε 65.535 μπλοκ και να έχει 1024 νήματα ο καθένας. Το σύνολο μπορεί να φθάνει στα 65.535x1024x1024, δηλαδή πάνω από 68 εκατομμύρια παράλληλους υπολογισμούς.

Για να εξυπηρετηθούν όλοι αυτοί οι παράλληλοι χρειάζεται και η υποστήριξη από πιο εξεζητημένο σύστημα μνημών, αφού κάθε νήμα απαιτεί και το δικό του αποκλειστικό κομμάτι μνήμης, ώστε να αποθηκεύει τους μετρητές προγραμμάτων και τις τιμές των registers. Κάθε πρόγραμμα φορτώνεται σε ένα μπλοκ του επεξεργαστεί και τα παράλληλα νήματα το υλοποιούν χρησιμοποιώντας το ίδιο κομμάτι μνήμης, το οποίο εξυπηρετεί ξεχωριστά το κάθε νήμα (thread) και συνολικά την εφαρμογή, όπως φαίνεται και στην επόμενη εικόνα.

Η διαφορά αυτής της ιεραρχίας μνήμης, όπως περιγράφεται στις GPU και όπως την παρακολουθήσαμε προηγουμένως με τις μνήμες cache επιπέδων L στην CPU, είναι ότι η επεξεργασία των δεδομένων γίνεται με δική της αποκλειστική μνήμη και όχι με αργή cache.



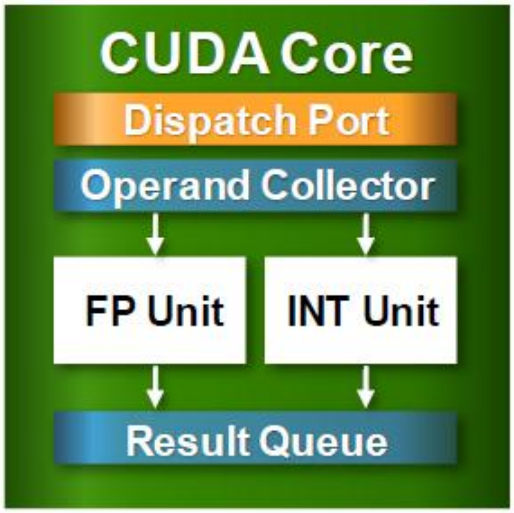
Εικόνα 18: Ιεραρχία νημάτων.

Από την άλλη, για να χειριστεί όλα αυτά τα θέματα η GPU χρησιμοποιεί ένα χαρακτηριστικό 32 νημάτων που ονομάζεται warp handler. Ο προγραμματιστής ενός

μπλοκ ή ενός CUDA δεν ασχολείται με το πως θα εργαστούν τα νήματα του μπλοκ για την ολοκλήρωση των υπολογισμών, αλλά αναλαμβάνει ο warp handler.

Συμπερασματικά, η αρχιτεκτονική μιας GPU περιλαμβάνει την απαραίτητη μνήμη και τα μπλοκ SP ή Streaming Processors. Τα SM αποτελούνται από πολλά CUDA, τις ειδικές μαθηματικές μονάδες υπολογισμού ημίτονου, συνημίτονου και άλλες, την μνήμη επιπέδου L1 και εκατοντάδες 32 bit καταχωρητές, με σημαντικότερο από όλα τα παραπάνω τα CUDA.

Αυτά, πραγματοποιούν όλες τις πράξεις με δεκαδικούς και ακέραιους αριθμούς, περιλαμβάνουν την λογική μονάδα και κάνουν σύγκριση και swift (μετακίνηση) σε αριθμούς.



Εικόνα 19: Αναπαράσταση πυρήνα CUDA.

Στην περίπτωση των GPU της εταιρίας AMD δεν συναντιέται η ορολογία CUDA, αλλά υπάρχει παρόμοια αρχιτεκτονική. Δηλαδή, η GPU διαιρείται σε μπλοκ επεξεργασίας, με όμοια λειτουργία με τα SM και αυτά αποτελούνται από μικρότερες μονάδες επεξεργασίας, όμοιες με τα CUDA.

Από την άποψη του λογισμικού, υπάρχει παρόμοια διαίρεση. Οπότε, κάθε μονάδα επεξεργασίας εκτελεί ένα νήμα διεργασίας, κάθε μπλοκ επεξεργασίας εκτελεί ένα μπλοκ νημάτων και μερικά μπλοκ νημάτων δημιουργούν τα warp και όλη η διεργασία γίνεται σε τμήματα τέτοιων warp νημάτων.

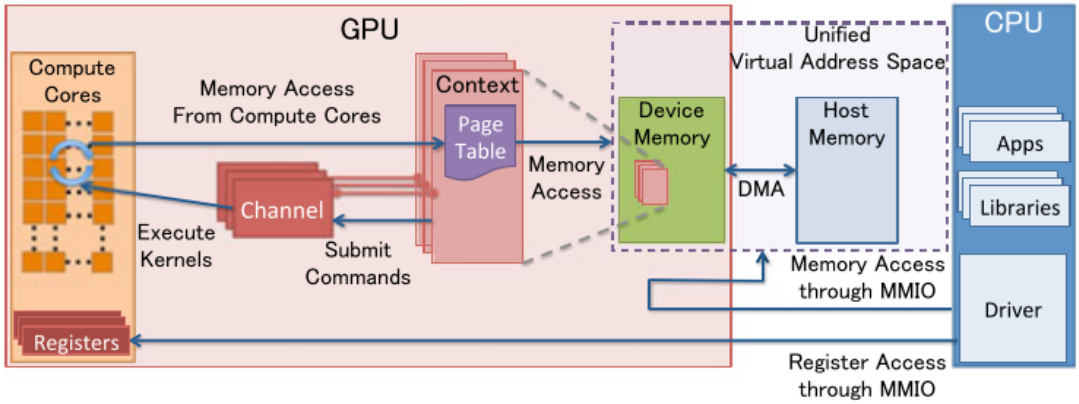
Για να λειτουργήσουν όλα αυτά χρειάζεται να αποθηκεύουν τα δεδομένα τους σε μονάδες μνήμης. Τέτοιες μονάδες στην GPU αποτελούν οι γρήγορες, αλλά μικρές

μνήμες καταχωρητών και cache επιπέδου L1, η γρήγορή, αλλά με πρόσβαση από όλους cache επιπέδου L2 και φυσικά η μνήμη VRAM. Σε φυσικό επίπεδο, πολλές φορές, η μνήμη VRAM μπορεί να περιλαμβάνει και τη μνήμη RAM, αλλά οι εφαρμογές να της αντιμετωπίζουν ως ενοποιημένη μνήμη και να την αξιοποιούν με αυτό τον τρόπο.

Γενικότερα, η GPU μπορεί να βρίσκεται σε οποιαδήποτε συσκευή, από μικρές φορητές συσκευές και κινητά τηλέφωνα έως μεγάλους υπέρ-υπολογιστές. Στην επόμενη εικόνα βλέπουμε σε ποιο σημείο εντοπίζεται η GPU σε ένα έξυπνο τηλέφωνο iPhone της εταιρίας Apple.

Εικόνα 20: : GPU ενός iPhone.

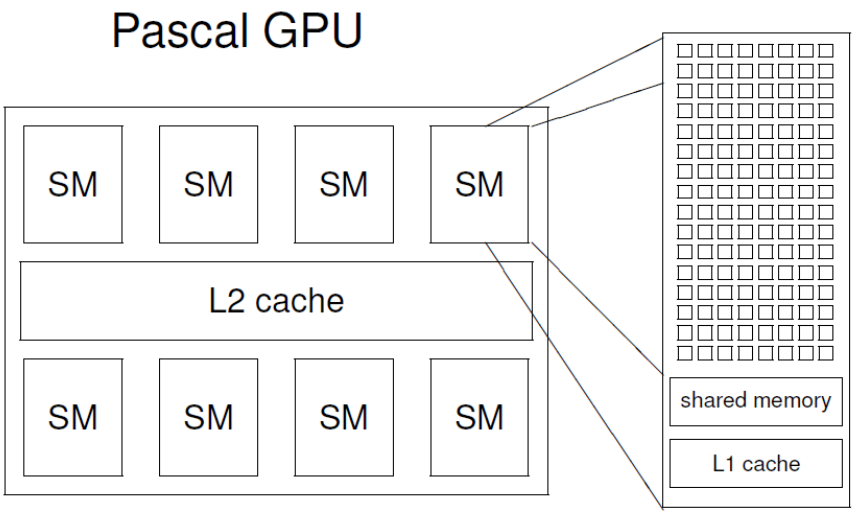
Ενδιαφέρον είναι ότι πολύ σύγχρονες υπολογιστές χρησιμοποιούν την GPU για να πραγματοποιούν τους υπολογισμούς τους, μια διαδικασία που θα αναλύσουμε περισσότερο στη συνέχεια. Σύμφωνα με στοιχεία της Sylvain Collange (Collange, 2017) το 2010 μόλις το 2% των κορυφαίων 500 υπέρ-υπολογιστών χρησιμοποιούσαν GPU για υπολογισμούς, ενώ το 2017 το ποσοστό έφθασε στο 18%.



Εικόνα 21: Σύστημα διαχείρισης μιας GPU.

Στις μέρες μας, οι πιο νέες προτάσεις στη σχεδίαση αρχιτεκτονικών GPU έρχονται από τις AMD και Nvidia. Η πρώτη είναι αυτή που παρουσίασε την πιο σύγχρονη αρχιτεκτονική, την GCN 5^{ης} γενιάς με κωδικό όνομα Vega και η πρώτη της υλοποίηση έγινε στην κάρτα γραφικών AMD Randeon VII με επεξεργαστή Vega μόλις 7 νάνο-χιλιοστών. Η δεύτερη εξέλιξε τις αρχιτεκτονικές Kepler και Maxwell στην νεότερη αρχιτεκτονική Pascal. Οπότε, μετά τους Kepler GK110 και Maxwell GM200, η Nvidia παρουσίαση τη νεότερη Pascal GP100.

Στη συνέχεια θα δούμε τα χαρακτηριστικά αυτών των αρχιτεκτονικών, δηλαδή των Vega και Pascal.



Εικόνα 22: Απλοποιημένη μορφή GPU αρχιτεκτονικής Pascal.

Αρχιτεκτονική Vega

Την τελευταία δεκαετία η εταιρία AMD εφαρμόζει στις GPU που παράγει την μικροαρχιτεκτονική και την υλοποίηση αυτής με την ονομασία GCN ή Graphics Core Next. Σήμερα έχουμε φθάσει στην 5ης γενιάς GCN με την ειδική ονομασία Vega. Σύμφωνα με τον Ryan Smith (Smith, 2018), η πιο σύγχρονη γενιά GPU της AMD είναι η Vega μόλις των 7 νάνο-χιλιοστών που θα κυκλοφορήσουν στις κάρτες γραφικών Radeon Instinct MI60 και Radeon Instinct MI50.

Σύμφωνα με τον επίσημο οδηγό της AMD για την αρχιτεκτονική Vega (AMD, 2017) υλοποιείται σε ένα μεγάλης κλίμακας chip με σκοπό να εξυπηρετήσει τις πιο απαιτητικές εφαρμογές, όπως τα πιο υψηλής ευκρίνειας βιντεοπαιχνίδια, τις εφαρμογές εικονικής πραγματικότητας και επιστημονικές εφαρμογές.

Στις τελευταίες γενιές, το chip Vega κυκλοφόρησε σε 14 νάνο-χιλιοστά για να γίνει στις κάρτες που θα κυκλοφορήσουν στις μέρες μας σε 7 νάνο-χιλιοστά με 12,5 εκατομμύρια τρανζίστορς πάνω σε μια πλακέτα 486 mm. Ο σκοπός του chip είναι να ξεπεράσει σε ταχύτητες τα 1,67 GHz και μοντέλο του φαίνεται στην επόμενη εικόνα.

Σε αυτήν βλέπουμε τον επεξεργαστή Vega Να έχει 64 τελευταία γενιάς μονάδες υπολογισμού τις NCU, οι οποίες με τη σειρά τους να προσφέρουν μια υπολογιστική δύναμη 4.096 stream processors, που είναι και η καρδιά των υπολογισμών στην GPU. Σε αυτό το σημείο δεν υπάρχουν ιδιαίτερες διαφορές από τις αμέσως προηγούμενες AMD GPU, αλλά όλα αυτά υλοποιούνται με σαφώς μεγαλύτερες ταχύτητες στον κύκλο του ρολογιού, με αποτέλεσμα από τα 13,7 teraflops υπολογισμών για απλές αριθμητικές πράξεις, να φθάσουμε στα 27,4 teraflops ακόμα και για πιο σύνθετες πράξεις. Στην ουσία αυτό σημαίνει ότι η ταχύτητα δεν έχει απλά διπλασιαστεί, αλλά ουσιαστικά τετραπλασιαστεί στις τελευταίας γενιάς Vega.

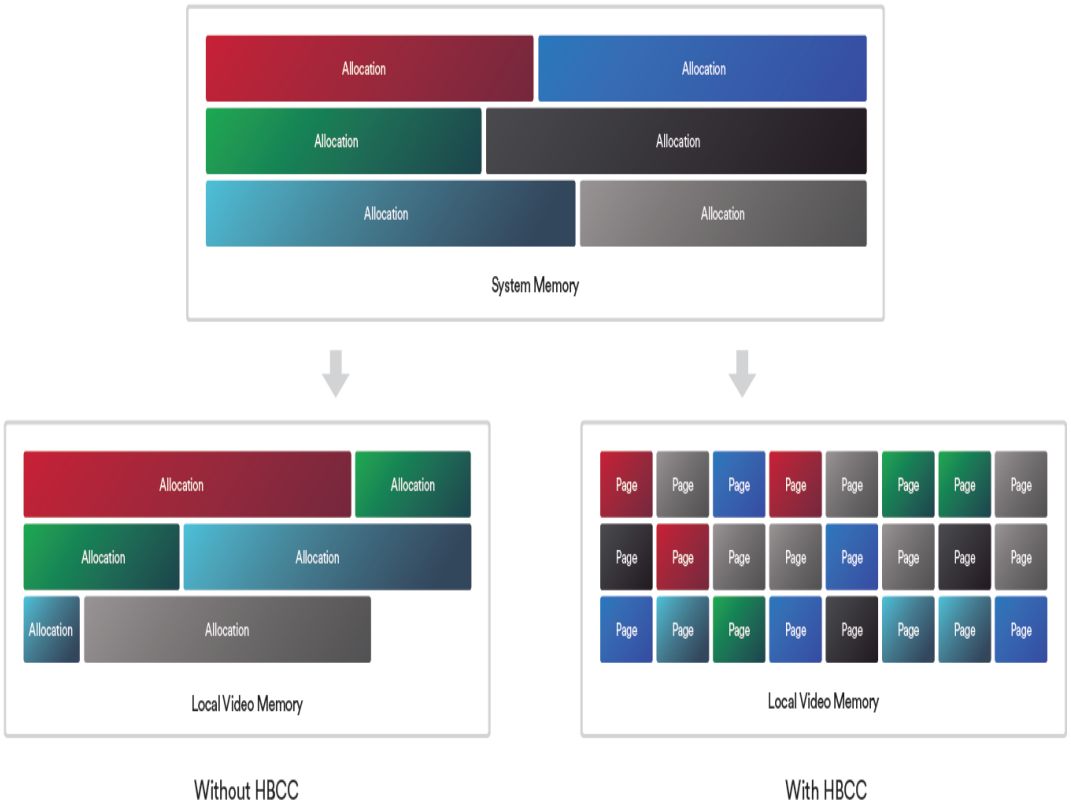
Στο σημείο αυτό να σημειώσουμε ότι τα FLOPS ή Floating Point Operations Per Second αποτελούν μια συνηθισμένη μονάδα μέτρησης της ταχύτητας του επεξεργαστή.

Σημαντική είναι η εφαρμογή στη διασύνδεση των επιμέρους μερών του επεξεργαστή με την τεχνολογία Infinity Fabric, με την οποία συνδέεται η κεντρική μονάδα με τις άλλες λογικές μονάδες, επιτυγχάνοντας μηδενική διάρκεια απόκρισης.

Συνδέοντας όλα τα μέρη με διαύλους αυτής της τεχνολογίας επιτυγχάνουμε την ταχύτερη επικοινωνίας όχι μόνο μεταξύ κεντρικής μονάδας και λογικών μονάδων, αλλά και μεταξύ των διαφόρων ειδών μνήμης, της διασύνδεσης PCI Express και των μπλοκ επιτάχυνσης.

Εικόνα 23: Μοντέλο Vega για τον σχεδιασμό GPU.

Είδαμε ότι την μεγάλη ανάγκη για γρήγορη μνήμη που απαιτεί η παράλληλη επεξεργασία δεδομένων, όπως αυτή υλοποιείται με τη χρήση νημάτων στην GPU, καλύπτεται μερικώς παραλείποντας την αργή μνήμη cache και αξιοποιώντας μόνο γρήγορη μνήμη. Πέρα από αυτό βασίζονται και σε ένα συνδυασμό προηγμένων συσκευών μνήμης και συστημάτων πολύ-επίπεδης κρυφής μνήμης για την κάλυψη αυτής της ανάγκης. Σε μια τέτοια τυπική διάταξη, μπλοκ επεξεργασίας αντλούν τα δεδομένα τους από ένα σύνολο μνημών cache επιπέδου L1, το οποίο με τη σειρά του έχει πρόσβαση σε μια ενιαία μνήμη cache επιπεδου L2. Η τελευταία παρέχει στη συνέχεια πρόσβαση υψηλής ταχύτητας και χαμηλής καθυστέρησης στην ειδική μνήμη βίντεο της GPU (local video memory) της επόμενης εικόνας.



Εικόνα 24: Σύγκριση μνήμης HBCC και μνήμης και κλασική κατανομή.

Η εξυπηρέτηση κάθε νήματος από την μνήμη δεν επαρκεί σε αρκετές περιπτώσεις, όπου απαιτείται η υποστήριξη βίντεο μεγάλου μεγέθους και η εναλλακτική για χρήση της μνήμης RAM δεν επαρκεί, αφού η επικοινωνία με αυτήν γίνεται μέσα από διαύλους PCI Express και δεν ολοκληρώνεται στην απαιτούμενη υψηλή ταχύτητα. Δηλαδή, απαιτείται μνήμη αρκετά μεγάλη σε μέγεθος και ταχύτητα.

Η αρχιτεκτονική "Vega" ξεπερνά αυτόν τον περιορισμό, επιτρέποντας στην τοπική μνήμη βίντεο να συμπεριφέρεται σαν μνήμη cache τελευταίου επιπέδου. Έτσι, αν η GPU προσπαθήσει να αποκτήσει πρόσβαση σε ένα κομμάτι δεδομένων που δεν είναι αποθηκευμένο στην τοπική μνήμη, μπορεί να τραβήξει μόνο τις απαραίτητες σελίδες μνήμης μέσα από τον δίαυλο PCI Express και να τις αποθηκεύσει στη μνήμη cache υψηλού bandwidth. Από τη στιγμή που οι σελίδες μνήμης είναι πολύ μικρότερες μπορούν να αντιγραφούν πολύ πιο γρήγορα. Μόλις ολοκληρωθεί η μεταφορά, οι μνήμες θα βρίσκονται στην μνήμη cache και αν της χρειαστεί κάποιο μπλοκ, μπορεί να τις ζητήσει πλέον από εκεί.

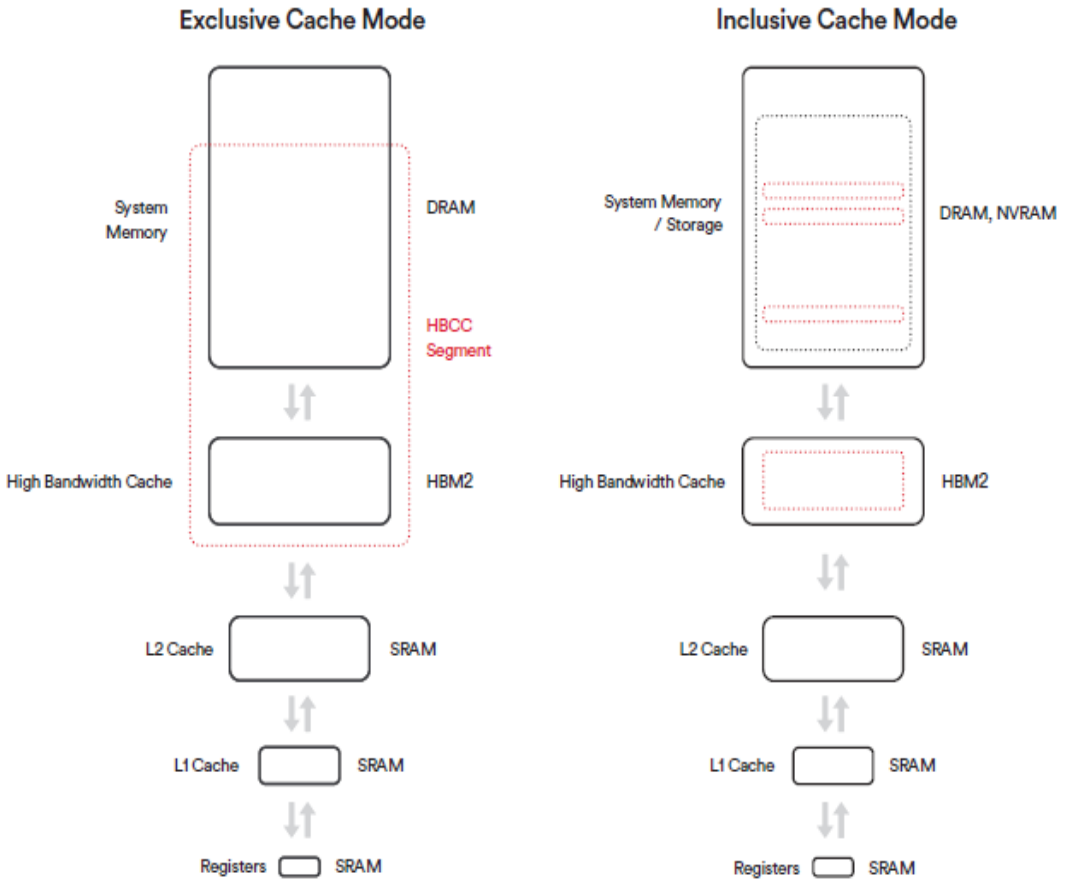
Για να συμβεί η παραπάνω διαδικασία απαιτείται ένας ειδικός ελεγκτής, που στην αρχιτεκτονική Vega ονομάζεται High-Bandwidth Controller Cache (HBCC) και παρέχει ένα σύνολο χαρακτηριστικών που επιτρέπουν στην απομακρυσμένη μνήμη να συμπεριφέρεται σαν τοπική μνήμη βίντεο και τοπική μνήμη βίντεο, όπως περιγράψαμε προηγουμένως, να συμπεριφέρεται σαν μνήμη cache τελευταίου επιπέδου. Το HBCC υποστηρίζει διευθυνσιοδότηση 49 bit, παρέχοντας έως και 512 terabytes εικονικού χώρου διευθύνσεων. Αυτό αρκεί για να καλύψει τον χώρο διευθύνσεων 48-bit που είναι προσβάσιμος από τις σύγχρονες.

Το HBCC θεωρείται από την Vega μια επαναστατική τεχνολογία για servers και επαγγελματικές εφαρμογές. Οι GPU που είναι βασισμένες στην αρχιτεκτονική Vega έχουν τη δυνατότητα να παρέχουν στις εφαρμογές αυτές αποτελεσματική απόδοση μνήμης συγκρίσιμη με την τοπική μνήμη βίντεο ενώ επεξεργάζονται σύνολα δεδομένων πιο κοντά στην χωρητικότητα της μνήμης του συστήματος.

Η τεχνολογία HBCC μπορεί επίσης να αξιοποιηθεί και από τις υπόλοιπες εφαρμογές, αν και τα περισσότερα συστήματα δεν διαθέτουν αρκετά μεγάλη μνήμη VRAM. Σε αυτή την περίπτωση, το HBCC την επεκτείνει για να συμπεριλάβει ένα μέρος της RAM. Οι εφαρμογές που θα τρέξουν σε ένα τέτοιο σύστημα θα δουν αυτήν την χωρητικότητα αποθήκευσης ως ένα ενιαίο χώρο μνήμης και όταν ζητήσουν δεδομένα από την RAM, τότε το HBCC θα μεταφέρει τις κατάλληλες σελίδες μνήμης. Αυτό το σύστημα διαχείρισης της μνήμης HBCC το ονομάζει η AMD ως HBCC Memory Segment ή HMS.

Η ιεραρχία και η επιμέρους επικοινωνία της μνήμης, της υψηλής ταχύτητας cache, της μνήμης επιπέδου L2 και L1 και των καταχωρητών register φαίνονται στα

αριστερά της επόμενης εικόνας. Στην ίδια εικόνα και στα δεξιά της φαίνεται η ίδια ιεραρχία, χωρίς την τεχνολογία HBCC.



Εικόνα 25: Ιεραρχία μνήμης και μνήμης cache.

Η διαθεσιμότητα μιας μεγάλης μνήμης μπορεί να βοηθήσει τους προγραμματιστές απαιτητικών εφαρμογών, όπως είναι οι προγραμματιστές παιχνιδιών, να δημιουργήσουν εικονικούς κόσμους με υψηλότερες λεπτομέρειες, πιο ρεαλιστικά κινούμενα σχέδια και πιο περίπλοκα συστήματα φωτισμού χωρίς να ανησυχούν για την υπέρβαση των παλαιότερων περιορισμών της χωρητικότητας μνήμης. Στην αρχιτεκτονική Vega όλα τα μπλοκ γραφικών εξυπηρετούνται από την cache L2, σε αντίθεση από τις προηγούμενες αρχιτεκτονικές που βασίζονται στο GCN, όπου κάθε μπλοκ είχε τις δικές του ανεξάρτητες κρυφές μνήμες. Εξαιτίας αυτού του κεντρικού ρόλου της L2, οι "Vega" GPU διαθέτουν μια μνήμη cache L2 χωρητικότητας 4 MB, η οποία είναι διπλάσια από το μέγεθος της μνήμης L2 σε προηγούμενες GPU υψηλής τεχνολογίας της εταιρίας AMD.

Άλλη διαφορά στην μνήμη HBM2 είναι ότι αυτή είναι ενσωματωμένη απευθείας στο πακέτο GPU και χρησιμοποιεί μια διασύνδεση πυριτίου για τη φυσική επικοινωνία. Οπότε, στον ίδιο περιορισμένο χώρο μπορούμε να έχουμε μεγέθη μνήμης 8GB και να έχουμε εξαιρετικά συμπαγή σχέδια για επιτραπέζιους και φορητούς υπολογιστές μικρού μεγέθους χωρίς να θυσιάζεται η χωρητικότητα μνήμης.

Τέλος, η μνήμη HBM2 έχει αυξήσει τις ταχύτητες δεδομένων κατά σχεδόν δύο φορές. Οι ευρείες διεπαφές επιτρέπουν σε κάθε συσκευή να λειτουργεί σε χαμηλότερες ταχύτητες ρολογιού όταν παρέχει ένα δεδομένο εύρος ζώνης, με συνέπεια να μειώνεται και η απαιτούμενη ενέργεια ανά μεταφερόμενο bit. Ο συνδυασμός της χαμηλότερης κατανάλωσης ενέργειας και πολύ υψηλής ταχύτητας σε ένα εξαιρετικά συμπαγές φυσικό αποτύπωμα κάνει το HBM2 να φαντάζει τον κυρίαρχο στον χώρο των τεχνολογιών μνήμης, ακόμα και έναντι στην πολύ αποδοτική GDDR5.

Άλλες λεπτομέρειες στον σχεδιασμός μιας Vega κάρτας γραφικών είναι ο νέος μηχανισμός γεωμετρίας επόμενης γενιάς, όπου η δημιουργία πολυγώνων γίνεται σε μεγαλύτερες ταχύτητες και χωρίς περιττή επεξεργασία, οπότε και η εικόνα δημιουργείται γρηγορότερα. Με αυτό τον τρόπο μια GPU τεχνολογίας Vega τέσσερις μηχανές γεωμετρίας, οι οποίες κανονικά θα περιορίζονται σε μια μέγιστη απόδοση τεσσάρων παραγωγών ανά κύκλο ρολογιού, αλλά αυτό το όριο αυξάνεται σε περισσότερα από 17 παραγωγές σχημάτων όταν συνεργάζονται μεταξύ τους.

Γενικότερα, οι GPU αρκετά χρησιμοποιούν περισσότερη μαθηματική ακρίβεια από ότι απαιτείται για τους υπολογισμούς που εκτελούν. Οι σύγχρονες κάρτες γραφικών λειτουργούν έτσι ώστε να μπορούν να υποστηρίζουν υπολογισμούς με 32-bit αριθμούς κινητής υποδιαστολής, οι οποίοι αποτελούν στις μέρες μας τον τρόπο με τον οποίο γίνονται τα 3D γραφικά, δηλαδή ικανοποιείται ο βασικός λόγος ύπαρξης μια κάρτας γραφικών. Παρόλα αυτά, οι χρήσεις μια σύγχρονης κάρτας υψηλής απόδοσης έχει εξελιχθεί και έχει περιλάβει και επιστημονικές εφαρμογές, που απαιτούν κυρίως αριθμούς 16-bit και η Vega έχει συμπεριλάβει μηχανισμούς, που υποστηρίζουν τον υπολογισμών τέτοιων αριθμών, μειώνοντας τον χρόνο στο μισό.

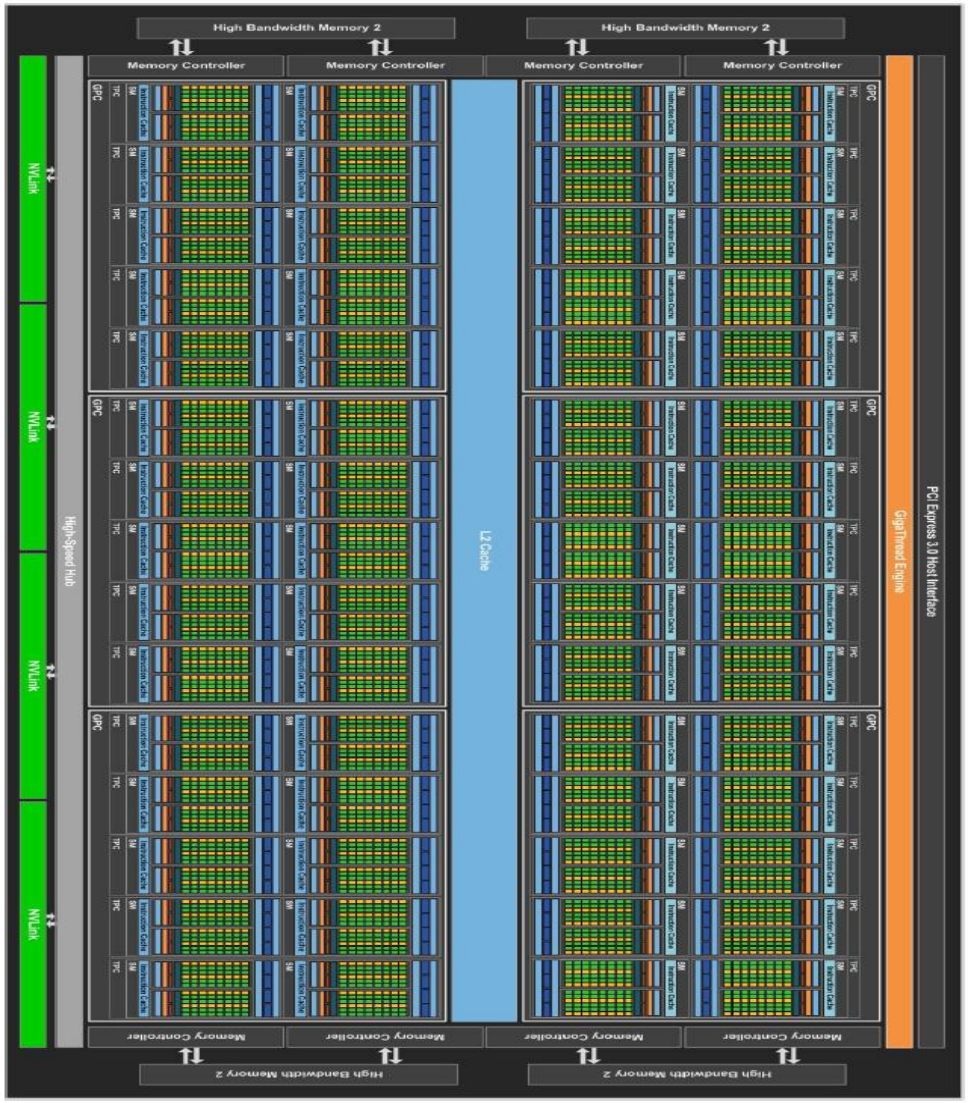
Αρχιτεκτονική Pascal

Η αρχιτεκτονική Pascal είναι η αρχιτεκτονική με την οποία η εταιρία Nvidia δημιουργεί τις δικές της GPU τα τελευταία χρόνια και είναι αυτή που έχει αποτελέσει

την βάση για τη σειρά GeForce 10, όπως είναι οι κάρτες GeForce GTX 1080, GTX 1070 και άλλες.

Αυτή η αρχιτεκτονική συνεχώς βελτιώνεται και τα κυριότερα χαρακτηριστικά της σύμφωνα με την Nvidia (Nvidia, 2016) είναι ο 16 νάνο-χιλιοστών πυρήνας, οι μαθηματικοί υπολογισμοί υψηλής ταχύτητας, η υψηλής ταχύτητας διασύνδεση, η μνήμη HBM2 και ο Streaming Multiprocessor (SM).

Η αρχιτεκτονική του επεξεργαστή φαίνεται στην επόμενη χαρακτηριστική εικόνα, όπου αυτός διαιρείται σε μπλοκ γραφικής επεξεργασίας (GPC) και αυτά με τη σειρά τους διαιρούνται σε Texture Processing Clusters (TPC), Streaming Multiprocessors (SM) και φυσικά σε πολλούς διαχειριστές μνήμης, με τα SM να λειτουργούν καθοριστικά στην απόδοση του συστήματος.



Εικόνα 26: GPU αρχιτεκτονικής Pascal.

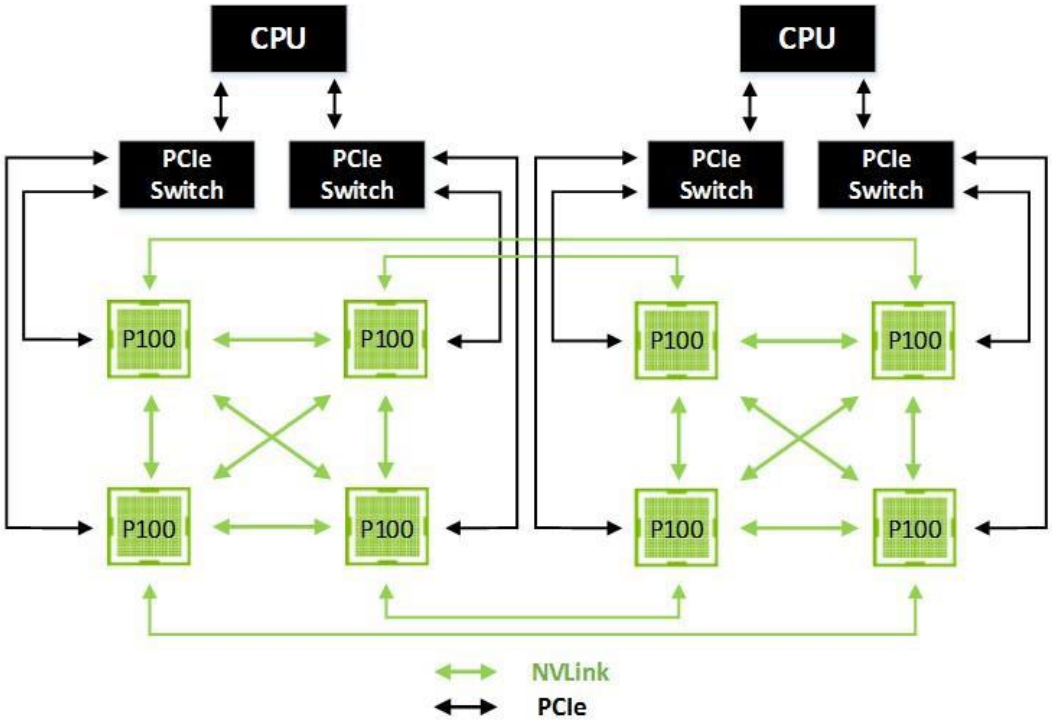
Στην συγκεκριμένη εικόνα, κάθε GPC έχει δέκα SM. Στην αρχιτεκτονική Pascal κάθε SM περιέχει 64 CUDA και τέσσερες μονάδες επεξεργασίας σκιάσεων (τα μπλε κουτάκια της προηγούμενης εικόνας). Δηλαδή, μια τέτοια GPU περιέχει 3850 πυρήνες CUDA. Όσον αφορά τη μνήμη, διαχειριστής μνήμης ή memory controller (στις άκρες τις εικόνας) συνδέεται με 512 KB μνήμης cache L2 και κάθε μνήμη HBM2 ελέγχεται από δύο διαχειριστές μνήμης. Η GPU της εικόνας έχει ένα σύνολο μνήμης cache L2 μεγέθους 4096 KB, ενώ παλαιότερες εκδόσεις είχαν 1536 KB. Αυτή η επιπλέον μνήμη cache σε κάθε chip μειώνει την επικοινωνία με την VRAM και κατά συνέπεια μειώνει τον χρόνο απόκρισης της μνήμης συνολικά και αυξάνει την απόδοση.

Η επόμενη εικόνα δείχνει αναλυτικότερα τη δομή κάθε SM.

Εικόνα 27: Streaming Multiprocessor στην αρχιτεκτονική Pascal.

Όσον αφορά τους μαθηματικούς υπολογισμούς, η Pascal επιτυγχάνει ταχύτητες της τάξης των 5,3 TFLOPS για 64-bit αριθμούς διπλής ακρίβειας κινητής υποδιαστολής, 10,6 TFLOPS για τους 32-bit με κινητή υποδιαστολή και 21,2 TFLOPS για τους 16-bit αριθμούς.

Επίσης, χρησιμοποιεί διασύνδεση τεχνολογίας NVLink για την επικοινωνία με άλλες μονάδες επεξεργασίας. Αυτές μπορεί να είναι άλλες GPU σε σύστημα πολλαπλών GPU, όπου η επικοινωνία γίνεται με ταχύτητες 160GB/δευτερόλεπτο, δηλαδή σχεδόν πέντε φορές περισσότερο από την κλασική επικοινωνία με PCI Express. Η υλοποίηση μιας τέτοιας επικοινωνίας φαίνεται στην επόμενη εικόνα, με την επικοινωνία πυρήνων Tesla P100 αρχιτεκτονικής Pascal και διασύνδεσης NVLink.



Εικόνα 28: Διασύνδεση NVLink σε GPU Tesla P100 αρχιτεκτονικής Pascal.

Η αρχιτεκτονική Pascal ήταν η πρώτη που χρησιμοποίησε μνήμες υψηλής ταχύτητας τύπου HBM2, οι οποίες βρίσκονται στον ίδιο φυσικό χώρο με την GPU και λειτουργούν σε μεγαλύτερες ταχύτητες από τις μνήμες προηγούμενης γενιάς τύπου GDDR5.

ΤΡΙΤΟ ΚΕΦΑΛΑΙΟ: ΜΕΘΟΔΟΙ ΑΞΙΟΛΟΓΙΣΗΣ ΕΠΙΔΟΣΕΩΝ

Παρακολουθήσαμε ότι το σημαντικότερο μέρος μιας κάρτας γραφικών είναι ο επεξεργαστής της. Οπότε, όποια και αν είναι η εταιρία από την οποία αγοράσαμε την κάρτα, δηλαδή η ASUS, η Gigabyte, η MSI ή κάποια άλλη, αυτό που θα αλλάζει θα είναι όλα τα επιπλέον χαρακτηριστικά που είδαμε στο προηγούμενο κεφάλαιο, σύστημα ψύξης, δίαυλοι επικοινωνίας ακόμα και λαμπάκια led, εκτός από το πιο σημαντικό. Αυτό εξαρτάται από την εταιρία κατασκευής της GPU και την γενιά του επεξεργαστή. Στις μέρες μας, θα είναι της εταιρίας Nvidia και αρχιτεκτονικής Pascal ή της εταιρίας AMD και αρχιτεκτονικής μέχρι πρόσφατα RX 500 και τελευταία Vega. Στο προηγούμενο κεφάλαιο έγινε μια προσπάθεια αξιολόγησης αυτών των χαρακτηριστικών, των δευτερευόντων και των πιο σημαντικών, ενώ σε αυτό το κεφάλαιο θα προσπαθήσουμε να δούμε πως όλα αυτά τα λαμβάνει ο τελικός χρήσης και αν υπάρχει κάποια έγκυρη μέθοδο αξιολόγησης των επιδόσεων.

Πριν από όλα αυτά θα ξεκινήσουμε από την γενική εικόνα και θα ξεκαθαρίσουμε μερικά επιπλέον των επεξεργαστών στοιχείων και χαρακτηριστικών που περιβάλλον αυτές τις κάρτες.

Θα αποσαφηνίσουμε τον ρόλο της μνήμης τυχαίας προσπέλασης της κάρτας γραφικών, της VRAM. Αυτές πρέπει να έχουν ικανό μέγεθος για να μπορούν να αποθηκεύουν την παραγόμενη εικόνα πριν αυτή αποσταλεί στην οθόνη. Στα όριά τους φθάνουν όταν η ανάλυση της οθόνης είναι στα 4K ή την ξεπερνά ή ακόμα περισσότερο όταν χρησιμοποιούμε δύο ή περισσότερες οθόνες, οπότε οι απαιτήσεις προστίθενται και μεγιστοποιούνται. Βασικότερο από το μέγεθος της μνήμης είναι, ίσως, η τεχνολογία που αυτή χρησιμοποιεί. Έτσι, έχουμε μνήμες τεχνολογίας GDDR5 να είναι δύο φορές πιο γρήγορες από τις μνήμες τεχνολογίας DDR3 και χαρακτηριστικό είναι ότι στις περισσότερες περιπτώσεις είναι προτιμότερο να έχουμε 1 GB μνήμης GDDR5 από 4GB μνήμης DDR3.

Αναφέραμε, επίσης, την τεχνολογία ύπαρξης δύο ή περισσότερων καρτών στον ίδιο υπολογιστή. Την τακτική αυτή η Nvidia την ονομάζει CrossFire και η AMD την αποκαλεί SLI. Παρόλα αυτά, η συνεργασία των επιπλέων καρτών συνήθως δεν έχει τα επιθυμητά αποτελέσματα και σύμφωνα με τον Julian Vernon (Vernon, 2014) κάθε επιπλέον κάρτα δεν προσφέρει περισσότερο από 25 ως 50% σε απόδοση, με τα

νούμερα να είναι ακόμα μικρότερα όταν ξεπερνούμε τις δύο κάρτες και τις ανάγκες για ρεύμα και τον επιπλέον θόρυβο να είναι πολλαπλάσια.

Γενικότερα, αυτό το χαρακτηριστικό των καρτών γραφικών, δηλαδή οι ανάγκες για σωστή τροφοδοσία ρεύματος και σωστή ψύξη είναι για τον Mark Coppock (Coppock, 2018) είναι ένα από τα βασικά ζητήματα στην αγορά μιας κάρτας και αποτελεί το πρόβλημα της TDP ή Thermal Design Power. Οπότε, για μερικές από τις κάρτες γραφικών με υψηλή απόδοση να θέλω επιπλέον τροφοδοσία. Η πιο σωστή αντιμετώπιση για αυτόν είναι η αγορά κάρτας με τις επιδόσεις που πραγματικά χρειαζόμαστε και η αγορά κάρτας με δικό της ψύκτη.

Πέρα από όλα αυτά δεν πρέπει να αγνοούμε το γεγονός ότι η κάρτα γραφικών είναι απλά ένα μέρος του υπολογιστικού μας συστήματος και όσο βασικό και αν θεωρείται για εμάς, δεν παύει να εξαρτά την επίδοσή της και από τα υπόλοιπα μέρη του. Έτσι, αν έχουμε μια πολύ γρήγορη και τελευταίας γενιάς κάρτα σε ένα μηχάνημα ξεπερασμένης τεχνολογίας, τότε θα έχουμε να κάνουμε με το φαινόμενο, που συχνά εμφανίζεται στην βιβλιογραφία ως «λαιμός του μπουκαλιού», όπου το υγρό κυλά γρήγορα μέσα στο μπουκάλια, αλλά όταν προσπαθεί να περάσει από τον λαιμό του μπουκαλιού έχει λιγότερο χώρο να κινηθεί και χύνεται πιο αργά από το στόμιο. Όμοια, η κάρτα γραφικών μπορεί να επεξεργάζεται πολύ γρήγορα τις πληροφορίες που καταφθάνουν, αλλά αυτές να έρχονται με πολύ μικρό ρυθμό, αν υπάρχει χαμηλής ταχύτητας επεξεργαστής, ή να είναι μικρές σε μέγεθος και να έρχονται σε μικρά μέρη, αν υπάρχει λίγη μνήμη RAM.

Για το 2018 οι σημαντικότερες κάρτες της εταιρίας Nvidia, σύμφωνα με τον Mark Coppock (Coppock, 2018) ήταν οι:

Πίνακας 2: Σημαντικότερες Nvidia κάρτες γραφικών το 2018

GPU	Ρολόι GPU (MHz)	Τύπος RAM	Μέγεθος RAM (GB)	Ταχύτητα RAM (GB/s)	TDP (watts)
GeForce GTX 1050	1354	GDDR5	2	112	75

GeForce GTX 1050 Ti	1290	GDDR5	4	112	75
GeForce GTX 1060	1506	GDDR5	6	192	120
GeForce GTX 1070	1506	GDDR5	8	256	150
GeForce GTX 1070 Ti	1683	GDDR5	8	256	180
GeForce GTX 1080	1607	GDDR5X	8	352	180
GeForce GTX 1080 Ti	1480	GDDR5X	11	484	250

Όμοια οι σημαντικότερες κάρτες της AMD για το 2018 είχαν RX τεχνολογία επεξεργαστή και οι πιο γρήγορες από τα μέσα το 2018 και ύστερα την Vega αρχιτεκτονική που αναλύσαμε σε προηγούμενη ενότητα και ήταν οι:

Πίνακας 3: Σημαντικότερες Nvidia κάρτες γραφικών το 2018

GPU	Ρολόι GPU (MHz)	Τύπος RAM	Μέγεθος RAM (GB)	Ταχύτητα RAM (GB/s)	TDP (watts)
Radeon RX 570	1168	GDDR5	4	224	150
Radeon RX 580	1257	GDDR5	4/8	256	185
Radeon RX Vega 56	1156	HBM2	8	410	210
Radeon RX Vega 64	1247	HBM2	8	484	295
Radeon RX Vega 64 Liquid	1406	HBM2	8	484	345

Από τα παραπάνω στοιχεία αντιλαμβανόμαστε ότι η τεχνολογία σε αυτό τον τομέα εξελίσσεται ραγδαία και ότι αν μια κάρτα θεωρείται κορυφαία είναι βέβαιο ότι μέσα στην ίδια χρονιά θα κυκλοφορήσει μία αρκετά ανώτερή της. Πέρα από τα παραπάνω τεχνικά χαρακτηριστικά που μας δίνουν οι εταιρίες, υπάρχει και ο κόσμος της πληροφορικής που παρακολουθεί τις εξελίξεις, ελέγχει τις επιδόσεις των καρτών γραφικών και καταλήγει στα δικά του συμπεράσματα. Αν προσπαθούσαμε να ομαδοποιήσουμε τους ελέγχους επιδόσεων που μπορούμε να αναζητήσουμε, θα λέγαμε ότι αυτοί μπορεί να είναι:

- εγκατάσταση ειδικών προγραμμάτων ελέγχου απόδοσης,
- αναφορά σε έγκυρες κριτικές τρίτων και
- προσωπικός έλεγχος της κάρτας γραφικών.

Ξεκινώντας με το πρώτο, υπάρχει μία πληθώρα ελεύθερης πρόσβασης εφαρμογών για τον έλεγχο της απόδοσης της κάρτας γραφικών. Σύμφωνα με το έγκυρο uk gaming computers (uk gaming computers, 2018) η καλύτερη εφαρμογή για αυτό τον έλεγχο είναι η Furmark στη διεύθυνση http://www.ozone3d.net/benchmarks/fur. Αυτή, μετά την εγκατάστασή της, την εκτελούμε, επιλέγουμε την ανάλυση που χρησιμοποιούμε τη συσκευή μας, επιλέγουμε εκτέλεση σε πλήρη οθόνη και την τρέχουμε. Η εφαρμογή ουσιαστικά εμφανίζει ένα τεράστιο ντόνατ να περιστρέφεται στην οθόνη, αλλά αυτό που κάνει ουσιαστικά είναι να οδηγεί την κάρτα γραφικών να λειτουργεί στα άκρα και να καταγράφει τις επιδόσεις της. Αν η θερμοκρασία της κάρτας που εμφανίζει η εφαρμογή ανέβει αρκετά και ξεπεράσει το συνηθισμένο φθάνοντας για παράδειγμα τους 95^{o}C, τότε πρέπει να πατήσουμε Esc και να τερματίσουμε την διαδικασία. Αυτό είναι ένα ξεκάθαρο σημάδι ότι η κάρτα μας δεν μπορεί να ανταπεξέλθει στις οριακές συνθήκες που περιγράφει ο κατασκευαστής και θα πρέπει να αντικατασταθεί. Επίσης, αν το ντόνατ, που εμφανίζεται στην οθόνη, αρχίσει μετά από μερικά λεπτά να παραμορφώνεται και να μην έχει ευδιάκριτο σχήμα, πάλι σημαίνει ότι η κάρτα μας δεν μπορεί να λειτουργήσει σωστά. Αν παρόλα αυτά περάσουν πάνω από δεκαπέντε λεπτά και δεν έχει εμφανιστεί κάτι από τα παραπάνω, τότε μπορούμε να τερματίσουμε την εκτέλεση της εφαρμογής και να θεωρήσουμε ότι όλα πήγαν καλά και η κάρτα μας λειτουργεί σωστά στις προδιαγραφές που μας περιέγραψε ο κατασκευαστής της.

Άλλες τέτοιες εφαρμογές μπορούν να βρεθούν στο https://novabench.com/, όπου εκτελούνται ξανά κάποιοι έλεγχοι και εμφανίζεται τελικά μια τελική βαθμολογία, την οποία μπορούμε να την συγκρίνουμε στην διεύθυνση https://novabench.com/parts/gpu με την βαθμολογία που θα έπρεπε να έχει το μοντέλο της κάρτας γραφικών που έχουμε στην κατοχή μας.

Από την άλλη, αν προσπαθούσαμε να ανατρέξουμε σε έγκυρες κριτικές τρίτων θα καταλήγαμε σε μια πληθώρα ιστοσελίδων με αναφορά στο Videocard Benchmarks με την πιο έγκυρη από αυτές να είναι του PassMark Software με κριτικές για περίπου 1.000.000 κάρτες γραφικών, οι οποίες ανανεώνονται συνεχώς. Στην πιο τελευταία λίστα κριτικών (PassMark Software, 2019) οι δέκα κάρτες με τις υψηλότερες επιδόσεις εμφανίζονται να είναι αυτές του επόμενου πίνακα.

Πίνακας 4: Οι 10 καλύτερες επιδόσεις το 2019

Τύπος Κάρτας Γραφικών	Επίδοση
Quadro RTX 6000	21.367
TITAN RTX	20.091
TITAN V CEO Edition	16.988
GeForce RTX 2080 Ti	16.909
GeForce RTX 2080	15.580
TITAN Xp COLLECTORS EDITION	14.899
TITAN V	14.715
GeForce RTX 2070	14.412
NVIDIA TITAN Xp	14.205
GeForce GTX 1080 Ti	14.167

Τέλος, στο Math Works (mathworks, 2018) παρουσιάζεται μια απλή εφαρμογή του προγράμματος Mat Lab, με την οποία μπορούμε μόνοι μας να ελέγξουμε την επίδοση της κάρτας γραφικών. Στη συνέχεια θα παρουσιάσουμε τον κώδικα της εφαρμογής και θα αναλύσουμε τα συμπεράσματα του παραδείγματος, που το Math Works παρουσιάζει.

Αυτός ο κώδικας ελέγχει μία κάρτα γραφικών και κατά πόσο γρήγορα αυτή:

- πόσο γρήγορα τα δεδομένα επιστρέφονται από την GPU,
- πόσο γρήγορα ο πυρήνας της GPU μπορεί να διαβάσει και να επεξεργαστεί τα δεδομένα και
- πόσο γρήγορα η GPU μπορεί να κάνει υπολογισμούς.

Όλες αυτές οι μετρήσεις είναι χρήσιμες για να εξετάσουμε την ταχύτητα της GPU, που είναι και το θέμα της παρούσας ενότητας, αλλά και να τη συγκρίνουμε με την ταχύτητα και τις επιδόσεις της CPU, ερώτημα που θα μας απασχολήσει στην αμέσως επόμενη ενότητα, εξετάζοντας, μάλιστα, από πιο σημείο και ύστερα ο όγκος της πληροφορίας και οι ανάγκες για γρήγορους υπολογισμούς είναι τέτοιες ώστε η GPU να υπερέχει της CPU.

Ξεκινώντας, το πρόγραμμα τρέχει τις αρχικές ρυθμίσεις:

```
gpu = gpuDevice();

fprintf('Using a %s GPU.\n', gpu.Name)

sizeOfDouble = 8; % Each double-precision number needs 8 bytes of storage

sizes = power(2, 14:28);
```

Αποτέλεσμα του παραπάνω είναι κάποιες τιμές να πάρουν αρχικές τιμές και στην οθόνη να εμφανιστεί το όνομα της GPU. Στα παράδειγμα θα έχουμε:

```
Using a Tesla K40c GPU.
```

Στη συνέχεια γράφουμε τον κώδικα για να υπολογίσουμε πόσο γρήγορα η GPU λαμβάνει και διαβάζει τα δεδομένα. Για να συμβεί αυτό αρχικοποιούμε την μνήμη και στέλνουμε τα δεδομένα με τη συνάρτηση gpuArray και η μνήμη αρχικοποιείται και στέλνει πίσω τα δεδομένα με τη συνάρτηση gather. Οπότε, ο κώδικας έχει ως εξής:

```
sendTimes = inf(size(sizes));

gatherTimes = inf(size(sizes));

for ii=1:numel(sizes)
```

```
        numElements = sizes(ii)/sizeOfDouble;
        hostData = randi([0 9], numElements, 1);
        gpuData = randi([0 9], numElements, 1,
'gpuArray');
        % Time sending to GPU
        sendFcn = @() gpuArray(hostData);
        sendTimes(ii) = gputimeit(sendFcn);
        % Time gathering back from GPU
        gatherFcn = @() gather(gpuData);
        gatherTimes(ii) = gputimeit(gatherFcn);
    end
    sendBandwidth = (sizes./sendTimes)/1e9;
    [maxSendBandwidth,maxSendIdx] = max(sendBandwidth);
    fprintf('Achieved peak send speed of %g
GB/s\n',maxSendBandwidth)
    gatherBandwidth = (sizes./gatherTimes)/1e9;
    [maxGatherBandwidth,maxGatherIdx] =
max(gatherBandwidth);
    fprintf('Achieved peak gather speed of %g
GB/s\n',max(gatherBandwidth))
```

Όπου, η διαίρεση με το 8 γίνεται αρχικά για να μετατραπούν τα bit σε bytes και στη συνέχεια η διαίρεση με το 1ε9, δηλαδή το 1.000.000.000, για να μετατραπούν τα bytes σε gigabytes. Η τελική απάντηση μας δείχνει την ταχύτητα αποστολής και λήψης σε gigabyte ανά δευτερόλεπτο και για το συγκεκριμένο παράδειγμα θα είναι:

```
    Achieved peak send speed of 6.18519 GB/s
```

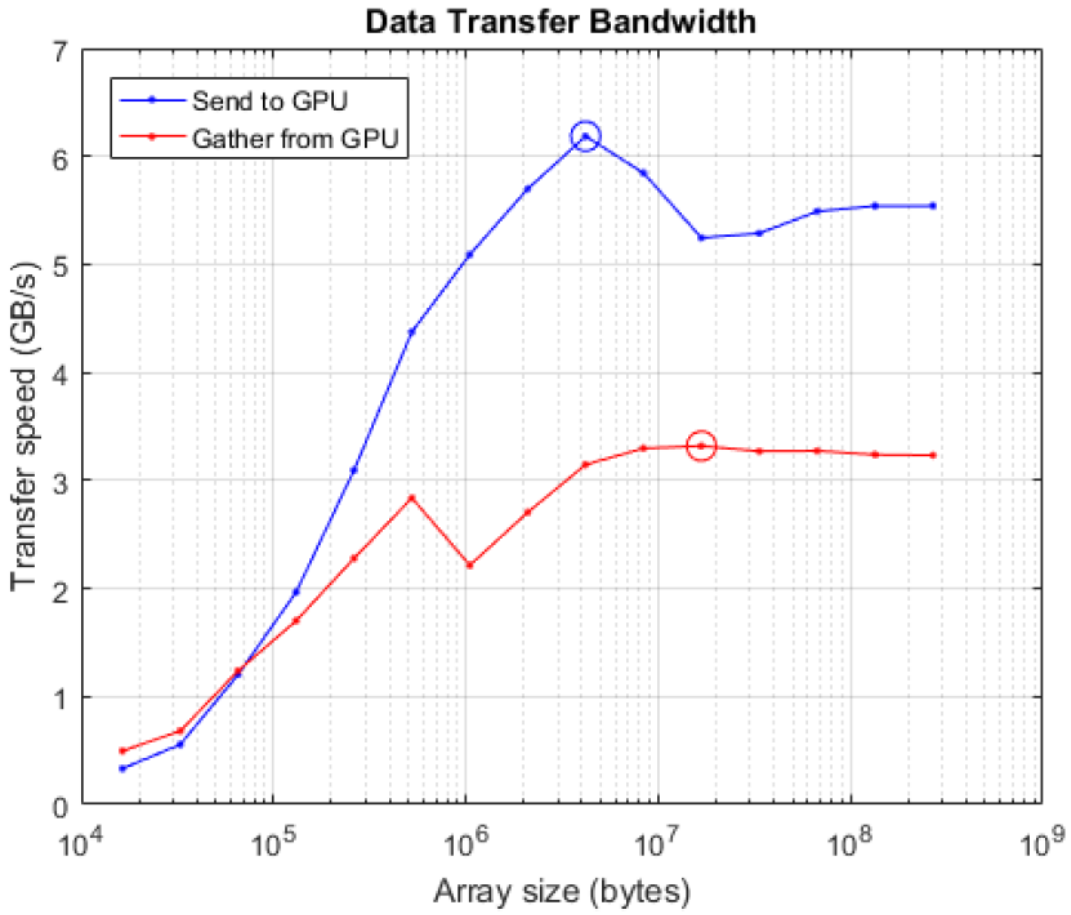
```
    Achieved peak gather speed of 3.31891 GB/s
```

Για να αντιληφθούμε τι σημαίνουν αυτά τα αποτελέσματα δημιουργούμε ένα plot στο MatLab όπου το μέγεθος του πίνακα αυξάνεται συνεχώς και ο κώδικας είναι ο επόμενος:

```
    hold off

    semilogx(sizes, sendBandwidth, 'b.-', sizes,
gatherBandwidth, 'r.-')

    hold on

    semilogx(sizes(maxSendIdx), maxSendBandwidth, 'bo-',
'MarkerSize', 10);

    semilogx(sizes(maxGatherIdx), maxGatherBandwidth,
'ro-', 'MarkerSize', 10);

    grid on

    title('Data Transfer Bandwidth')

    xlabel('Array size (bytes)')

    ylabel('Transfer speed (GB/s)')

    legend('Send to GPU', 'Gather from GPU', 'Location',
'NorthWest')
```

Το αποτέλεσμα του τελευταίου κώδικα θα εμφανιστεί στην οθόνη μας με τη μορφή του γραφήματος της επόμενης εικόνας.

Να σημειώσουμε ότι στο συγκεκριμένο παράδειγμα αναφέρεται ότι χρησιμοποιείται PCI express v3 με θεωρητικό bandwidth 0,99GB/sec και για την κάρτα γραφικών Nvidia με 16 slots PCIe3 αυτός μας δίνει 0,99GB/sec * 16 = 15,75GB/sec bandwidth θεωρητικά.



Εικόνα 29: Εύρος μεταφοράς δεδομένων στην GPU

Παρόλα αυτά, όπως φαίνεται στην τελευταία εικόνα, το ανώτερο σημείο αποστολής μόλις που ξεπερνά τα 6GB/sec, ενώ της λήψης φθάνει το ανώτερο λίγο πριν τα 3,5GB/sec. Δηλαδή, όσο περισσότερα δεδομένα αποστέλλουμε στην GPU, τόσο πιο γρήγορα αυτά εξυπηρετούνται, αλλά από ένα, σχετικά μικρό, σημείο και ύστερα η ταχύτητα αυτή πέφτει. Αντίστοιχα, η λήψη των δεδομένων ακολουθεί την ίδια πορεία με ακόμα μικρότερες τιμές.

Κρατώντας τα παραπάνω σαν πρώτη σκέψη, θα περάσουμε και στους επόμενους ελέγχους για να καταλήξουμε αν ο ενδιάμεσος δίαυλος είναι ο κρίσιμος παράγοντας που καθυστερεί τη διαδικασία.

Στη συνέχεια θα παρακολουθήσουμε πόσο γρήγορα η GPU μπορεί να διαβάζει τα προς επεξεργασία δεδομένα. Για να το αντιληφθούμε αυτό θα δημιουργήσουμε μια συνάρτηση plusFcn που θα εκτελεί ένα απλό διάβασμα και εκτέλεση ενός δεκαδικού. Η επιλογή διαβάσματος ενός αριθμού έχει να κάνει με την ελάχιστη ενέργεια που χρειάζεται να καταναλώσει ο επεξεργαστής για να φέρει σε

πέρας τη διαδικασία. Οπότε, αυτό που θα κριθεί θα είναι η ταχύτητα που λαμβάνει τα δεδομένα από την μνήμη, αφού η διάρκεια επεξεργασίας θα είναι ουσιαστικά μηδέν.

Μια τέτοια συνάρτηση φαίνεται στη συνέχεια.

```
memoryTimesGPU = inf(size(sizes));

for ii=1:numel(sizes)

  numElements = sizes(ii)/sizeOfDouble;

  gpuData = randi([0 9], numElements, 1, 'gpuArray');

  plusFcn = @() plus(gpuData, 1.0);

  memoryTimesGPU(ii) = gputimeit(plusFcn);

end

memoryBandwidthGPU = 2*(sizes./memoryTimesGPU)/1e9;

[maxBWGPU, maxBWIdxGPU] = max(memoryBandwidthGPU);

fprintf('Achieved peak read+write speed on the GPU: %g GB/s\n',maxBWGPU)
```

Με την συγκεκριμένη κάρτα γραφικών η έξοδος θα ήταν:

```
Achieved peak read+write speed on the GPU: 186.494 GB/s
```

Δηλαδή, το πέρασμα των δεδομένων από την μνήμη VRAM στον επεξεργαστή GPU γίνεται αρκετά γρήγορα. Με τι όμως θα συγκριθεί η έννοια της ταχύτητας; Στο επόμενο κώδικα υπολογίζουμε τις ίδιες πράξεις για την CPU:

```
memoryTimesHost = inf(size(sizes));

for ii=1:numel(sizes)

  numElements = sizes(ii)/sizeOfDouble;

  hostData = randi([0 9], numElements, 1);
```

```
        plusFcn = @() plus(hostData, 1.0);

        memoryTimesHost(ii) = timeit(plusFcn);

    end

    memoryBandwidthHost                                    =
2*(sizes./memoryTimesHost)/1e9;

    [maxBWHost,              maxBWIdxHost]                 =
max(memoryBandwidthHost);

    fprintf('Achieved peak read+write speed on the host:
%g GB/s\n',maxBWHost)
```

Η έξοδος του τελευταίου κώδικα θα ήταν:

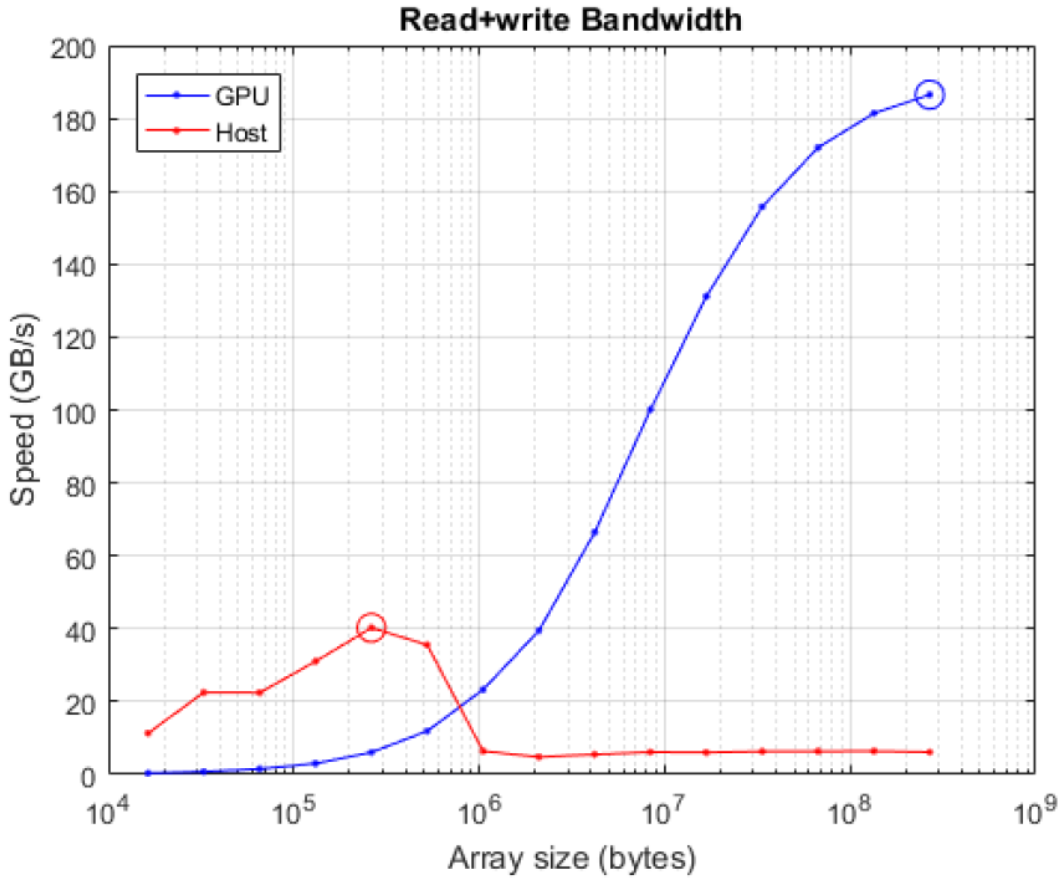
```
    Achieved peak read+write speed on the host: 40.2573
GB/s
```

Δηλαδή, συγκριτικά πιο αργός από την GPU. Ιδιαίτερα συμπεράσματα έχουμε αν δημιουργήσουμε το επόμενο plot.

```
    % Plot CPU and GPU results.

    hold off

    semilogx(sizes, memoryBandwidthGPU, 'b.-', ...

    sizes, memoryBandwidthHost, 'r.-')

    hold on

    semilogx(sizes(maxBWIdxGPU),    maxBWGPU,    'bo-',
'MarkerSize', 10);

    semilogx(sizes(maxBWIdxHost),   maxBWHost,   'ro-',
'MarkerSize', 10);

    grid on

    title('Read+write Bandwidth')
```

```
xlabel('Array size (bytes)')

ylabel('Speed (GB/s)')

legend('GPU', 'Host', 'Location', 'NorthWest')
```

Αν τρέξουμε το τελευταίο θα έχουμε το γράφημα της επόμενης εικόνας, όπου συγκρίνεται η ταχύτητα περάσματος δεδομένων από την VRAM στην GPU σε σχέση με το αντίστοιχο πέρασμα από την RAM στην CPU.



Εικόνα 30: Διάβασμα δεδομένων με GPU και CPU

Το ολοφάνερο συμπέρασμα είναι ότι η GPU συνεργάζεται ταχύτερα με την μνήμη της από ότι η CPU.

Στη συνέχεια θα τι πραγματικά συμβαίνει στον πυρήνα των δύο επεξεργαστών, δημιουργώντας ένα πρόγραμμα που εκτελεί πιο περίπλοκους υπολογισμούς από το πέρασμα ενός δεκαδικού από και προς την μνήμη και θα εκτελέσουμε έναν πολλαπλασιασμό πινάκων. Θεωρητικά, αν πολλαπλασιάσουμε δύο πίνακες NxN, χρειαζόμαστε $2N^3 - N^2$ υπολογισμούς ή FLOPS, όπως συχνά

ονομάζουμε για να δείξουμε ότι αυτοί συμβαίνουν στον πυρήνα ενός επεξεργαστή. Άρα, για πολλαπλασιασμό δύο πινάκων και αποτύπωση αποτελεσμάτων σε ένα τρίτο πίνακα, θα κάνουμε υπολογισμούς για $3N^2$ στοιχεία, αφού είναι τρεις οι πίνακες μεγέθους NxN ο κάθε ένας. Οπότε, η αναλογία είναι $(2N^3 – N^{2)}/3N^2$ ή (2N – 1)/3 FLOP ανά στοιχείο. Αντίθετα με το προηγούμενο παράδειγμα όπου ουσιαστικά χρειαζόμασταν ½ FLOP ανά στοιχείο.

Συμπερασματικά, θα τρέξουμε τον ίδιο κώδικα στην GPU και στην CPU και τα αποτελέσματα θα δείχνουν ξεκάθαρα την ταχύτητα των δύο επεξεργαστών, αφού η διαφορά διαβάσματος από την μνήμη σε όποιον από τους δύο επεξεργαστές, θεωρείται αμελητέα και αυτό που θα μας καθυστερεί είναι οι ίδιοι οι υπολογισμοί.

Ένας τέτοιος κώδικας είναι ο επόμενος:

```
sizes = power(2, 12:2:24);

N = sqrt(sizes);

mmTimesHost = inf(size(sizes));

mmTimesGPU = inf(size(sizes));

for ii=1:numel(sizes)

  % First do it on the host

  A = rand( N(ii), N(ii) );

  B = rand( N(ii), N(ii) );

  mmTimesHost(ii) = timeit(@() A*B);

  % Now on the GPU

  A = gpuArray(A);

  B = gpuArray(B);

  mmTimesGPU(ii) = gputimeit(@() A*B);

end
```

```
    mmGFlopsHost = (2*N.^3 - N.^2)./mmTimesHost/1e9;

    [maxGFlopsHost,maxGFlopsHostIdx]                    =
max(mmGFlopsHost);

    mmGFlopsGPU = (2*N.^3 - N.^2)./mmTimesGPU/1e9;

    [maxGFlopsGPU,maxGFlopsGPUIdx] = max(mmGFlopsGPU);

    fprintf(['Achieved peak calculation rates of ', ...

    '%1.1f GFLOPS (host), %1.1f GFLOPS (GPU)\n'], ...

    maxGFlopsHost, maxGFlopsGPU)
```

Η έξοδος του προγράμματος θα ήταν:

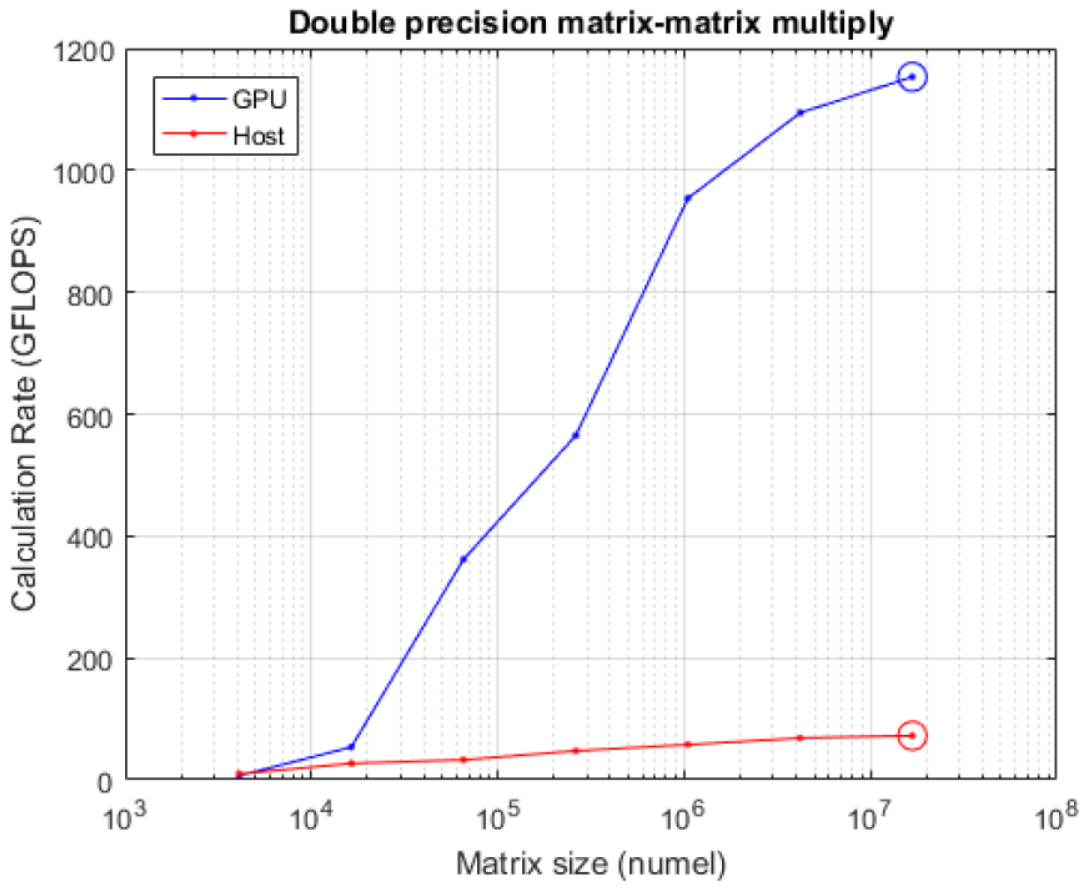
```
    Achieved peak calculation rates of 72.5 GFLOPS
(host), 1153.3 GFLOPS (GPU)
```

Σε αυτήν παρατηρούμε μια αξιοσημείωτη διαφορά υπέρ της GPU, ιδιαίτερα αν τρέξουμε το επόμενο plot:

```
    hold off

    semilogx(sizes,    mmGFlopsGPU,    'b.-',    sizes,
mmGFlopsHost, 'r.-')

    hold on

    semilogx(sizes(maxGFlopsGPUIdx), maxGFlopsGPU, 'bo-
', 'MarkerSize', 10);

    semilogx(sizes(maxGFlopsHostIdx),     maxGFlopsHost,
'ro-', 'MarkerSize', 10);

    grid on

    title('Double precision matrix-matrix multiply')

    xlabel('Matrix size (numel)')

    ylabel('Calculation Rate (GFLOPS)')
```

```
legend('GPU', 'Host', 'Location', 'NorthWest')
```

Το αποτέλεσμα του plot θα μας δείχνει ξεκάθαρα την διαφορά στην ταχύτητα των δύο επεξεργαστών για διάφορους όγκους δεδομένων. Ένα τέτοιο αποτέλεσμα είναι αυτό της επόμενης εικόνας.



Εικόνα 31: Ταχύτητα εκτέλεσης πολλαπλασιασμού πινάκων σε CPU και GPU

Μετά την ολοκλήρωση και των τριών ελέγχων, μπορούμε να καταλήξουμε σε μερικά ασφαλή συμπεράσματα.

Καταρχήν, η αμφιβολία που υπήρχε για το πέρασμα των δεδομένων από την RAM στην VRAM και στην GPU, δεν υπάρχει πια. Αυτό που καθυστερεί την όλη διαδικασία είναι οι δίαυλοι επικοινωνίας της κάρτας γραφικών με την μητρική κάρτα. Γενικά, αυτή η επικοινωνία εύκολα χαρακτηρίζεται ως αργή.

Σημαντικό είναι και το συμπέρασμα ότι η GPU διαβάζει και γράφει στην VRAM πολύ πιο γρήγορα από ότι η CPU στην RAM. Το γεγονός αυτό, μαζί με το

προηγούμενο μας δείχνουν ότι μία εφαρμογή θα εκτελούταν πολύ πιο γρήγορα αν μετέφερε την λειτουργία της στην VRAM.

Ιδιαίτερα μετά και από το τρίτο συμπέρασμα, όπου ολοφάνερα σε μεγάλους όγκους δεδομένων η GPU εκτελεί υπολογισμούς πολύ πιο γρήγορα από την CPU. Τα σημεία που πρέπει να παρατηρηθούν, βέβαια, είναι ότι αυτό συμβαίνει για μεγάλους όγκους δεδομένων και για μαθηματικές πράξεις, αφού ο προορισμός της CPU είναι να εξυπηρετεί την καθημερινή λειτουργία ενός συστήματος, οπότε έχει σχεδιαστεί για να εκτελεί γρήγορα όλων των ειδών τους υπολογισμούς που απαιτούνται και για όγκο δεδομένων σε συγκεκριμένα πλαίσια.

Στο επόμενο κεφάλαιο θα συγκρίνουμε τον τρόπο σχεδιασμού GPU και CPU και θα δούμε πως ο διαφορετικός σκοπός για τον οποίο έχουν σχεδιαστεί επηρεάζει την αρχιτεκτονική τους.

ΤΕΤΑΡΤΟ ΚΕΦΑΛΑΙΟ: ΣΥΓΚΡΙΣΗ CPU ΚΑΙ GPU

Στα προηγούμενα κεφάλαια παρακολουθήσαμε διάφορα στοιχεία για τις κάρτες γραφικών και επικεντρωθήκαμε στην μελέτη του βασικού της στοιχείου, δηλαδή στην μονάδα επεξεργασίας της ή την GPU.

Στη συνέχεια αυτού του κεφαλαίου θα συγκεντρώσουμε όλα τα στοιχεία που παρακολουθήσαμε για την αρχιτεκτονική επεξεργαστών CPU και GPU, θα δούμε περιπτώσεις που δε χρησιμοποιούμε την μία από τις δύο, περιπτώσεις που αυτές συνεργάζονται μεταξύ τους και θα καταλήξουμε, τελικά, στην σύγκρισή τους.

4.1 Στοιχεία σχεδιασμού CPU και GPU

Συγκεντρώνοντας όλα όσα μελετήσαμε μέχρι τώρα και αφορούν την αρχιτεκτονική των GPU όλα όσα αναφέραμε για τις CPU θα μπορούσαμε να καταλήξουμε ότι η παράλληλη επεξεργασία είναι το μέλλον στην επεξεργασία των δεδομένων και ότι υπάρχουν καλές και κακές αρχιτεκτονικές στον σχεδιασμό και στην υλοποίηση επεξεργαστών, είτε για την κεντρική μονάδα επεξεργασίας είτε για την κάρτα γραφικών. Αυτά, όμως, δεν αποτελούν την αλήθεια όσον αφορά στο σχεδιασμό επεξεργαστών.

Για αρχή θα πούμε ότι οι αρχές των συστημάτων παράλληλης επεξεργασίας βρίσκονται στη δεκαετία του ’90 και στην μελέτη του Ralph Duncan (Duncan, 1990), οπότε σαν ιδέα δεν είναι καθόλου σύγχρονη.

Επίσης, ήδη από το 1972 ο Michael Flynn (Flynn, 1972) δημιούργησε τη βασική ταξινόμηση στην αρχιτεκτονική των επεξεργαστών. Αυτή παρέμεινε για χρόνια και φθάνει να ακολουθείται ακόμα και στις μέρες μας. Η ταξινόμηση αυτή γίνεται πάνω στην απλή λογική που εξετάζει από τις πόσες ροές επεξεργασίας δημιουργεί η μονάδα ελέγχου για να επεξεργαστεί κάποια ή κάποιες ροές δομένων. Έτσι, μπορούμε να έχουμε μία ή πολλές ροές επεξεργασίας για μία ή πολλές ροές δεδομένων, σε όλους τους δυνατούς συνδυασμούς.

Ακολουθώντας το τελευταίο έχουμε τις επόμενες κατηγορίες αρχιτεκτονικών:

- Single instruction stream – single data stream (SISD), με μία ροή επεξεργασίας για μία ροή δεδομένων.

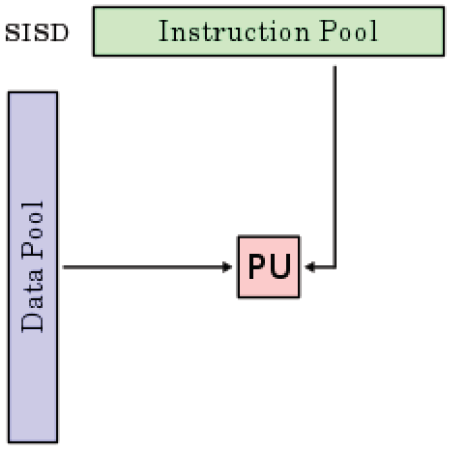
- Single instruction stream – multiple data streams (SIMD), με πολλές ροές επεξεργασίας να επεξεργάζονται μία ροή δεδομένων.

- Multiple instruction streams, single data stream (MISD), με πολλές ροές επεξεργασίας να επεξεργάζονται μία ροή δεδομένων, κυρίως σε περιπτώσεις επαλήθευσης.

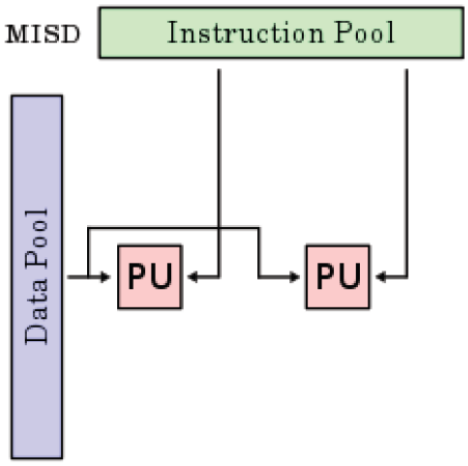
- Multiple instruction streams, multiple data streams (MIMD), με πολλές ροές επεξεργασίας να εξυπηρετούν πολλές ροές δεομένων.

Όλες οι παραπάνω αρχιτεκτονικές ακολουθούνται και στις μέρες μας. Καμία δεν είναι ξεπερασμένη και όλες υλοποιούνται σε συγκεκριμένες περιπτώσεις. Βέβαια, οι επεξεργαστές σε CPU αιχμής χρησιμοποιούν αρχιτεκτονική MIMD, οι περισσότερες CPU σήμερα ακολουθούν αρχιτεκτονική SIMD και η πολύ-νηματική επεξεργασία Single instruction – multiple threads ή SIMT των GPU της Nvidia θα μπορούσαμε να αναφέρουμε ότι αποτελούν πρακτικά υποκατηγορία της SIMD. Οπότε, η αρχιτεκτονική SIMD θα λέγαμε ότι αποτελεί σήμερα την τάση στην κατασκευή και υλοποίηση επεξεργαστών.

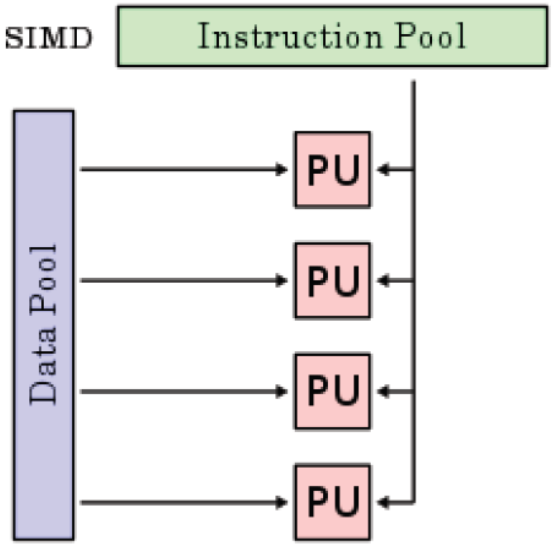
Στις επόμενες εικόνες φαίνονται σχηματικά οι τέσσερεις αρχιτεκτονικές.



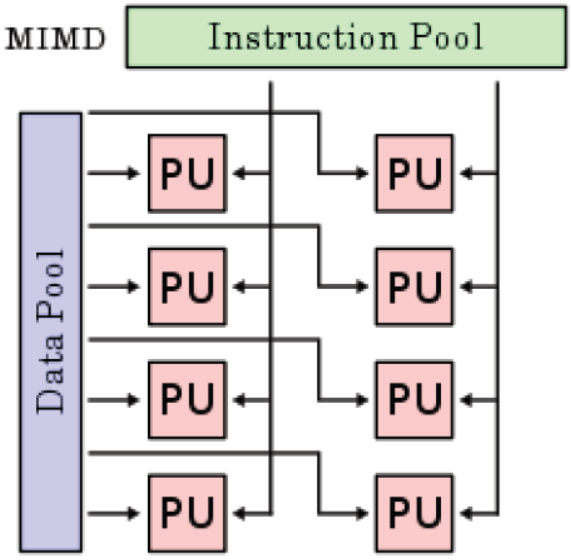
Εικόνα 32: Αρχιτεκτονική SISD.



Εικόνα 33: Αρχιτεκτονική MISD.



Εικόνα 34: Αρχιτεκτονική SIMD.



Εικόνα 35: Αρχιτεκτονική MIMD.

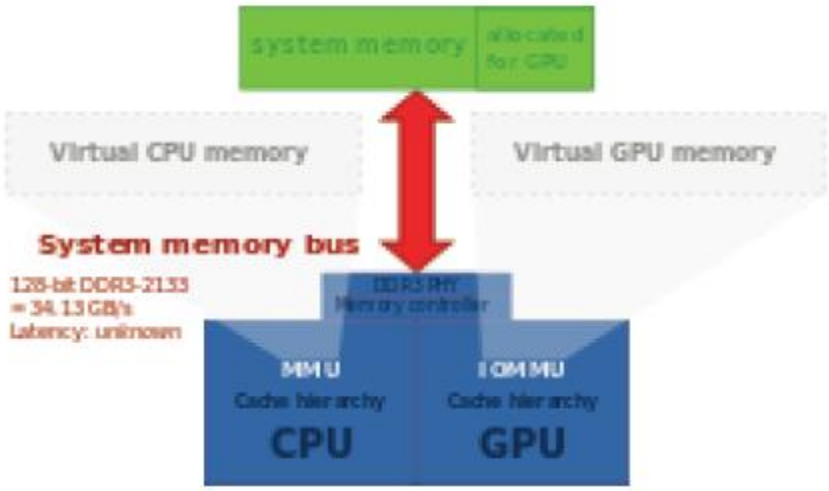
4.2 Dedicated και Integrated κάρτες γραφικών

Σύμφωνα με τον Andy Betts (Betts, 2018) υπάρχουν δύο τρόποι που οι ηλεκτρονικοί υπολογιστές χειρίζονται τα γραφικά τους, με dedicated κάρτα γραφικών ή με integrated κάρτα γραφικών.

Ο πρώτος τρόπος είναι αυτό που περιγράψαμε στην μελέτη μας μέχρι τώρα και αφορά μια ενσωματωμένη κάρτα στο σύστημά μας, που ονομάζεται κάρτα γραφικών.

Ο δεύτερος τρόπος αφορά τις κεντρικές μονάδες επεξεργασίας με δική τους GPU, δηλαδή με δικό τους επεξεργαστή για τη δημιουργία γραφικών. Για να λειτουργήσουν, χρησιμοποιούν την μνήμη RAM και για αυτό αναφέρονται και ως shared κάρτες. Οπότε, αν ένα σύστημα διαθέτει 4GB RAM και έχει Integrated κάρτα, τότε είναι πιθανό η VRAM του να είναι 1GB και η πραγματική του RAM να είναι 3GB. Όλα αυτά φαίνονται στην επόμενη εικόνα.

Είναι προφανές ότι πιο αποδοτικό είναι ένα σύστημα με dedicated κάρτα γραφικών, αλλά η πραγματικότητα είναι ότι οι περισσότερες σύγχρονες μητρικές κάρτες διαθέτουν dedicated κάρτα γραφικών και είναι κάθε φορά στην επιλογή του χρήστη αν θα παραμείνει με αυτόν τον φθηνό τρόπο ή θα προσθέσει κάρτα γραφικών της δικής του επιλογής, ξοδεύοντας επιπλέον χρήματα, αλλά κερδίζοντας σε απόδοση και αποκτώντας τη δυνατότητα να τρέξει απαιτητικές εφαρμογές.



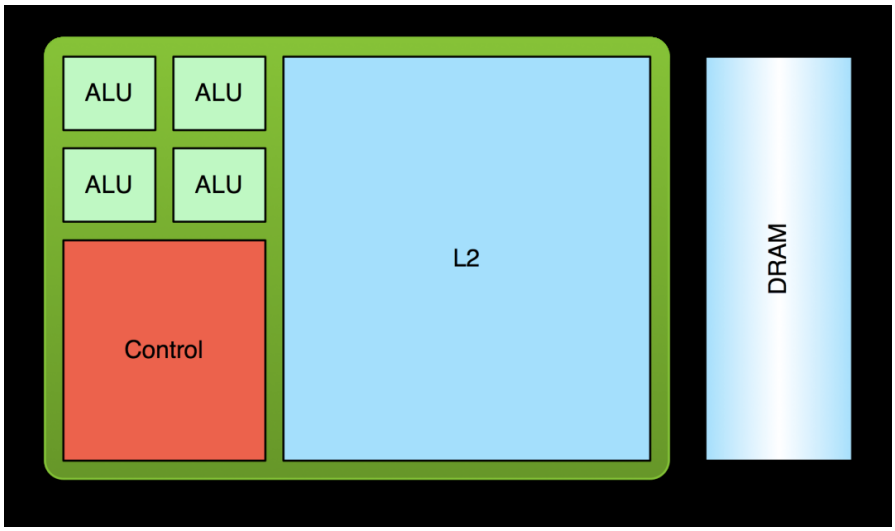
Εικόνα 36: Integrated κάρτα γραφικών.

4.3 Διαφορές GPU και CPU

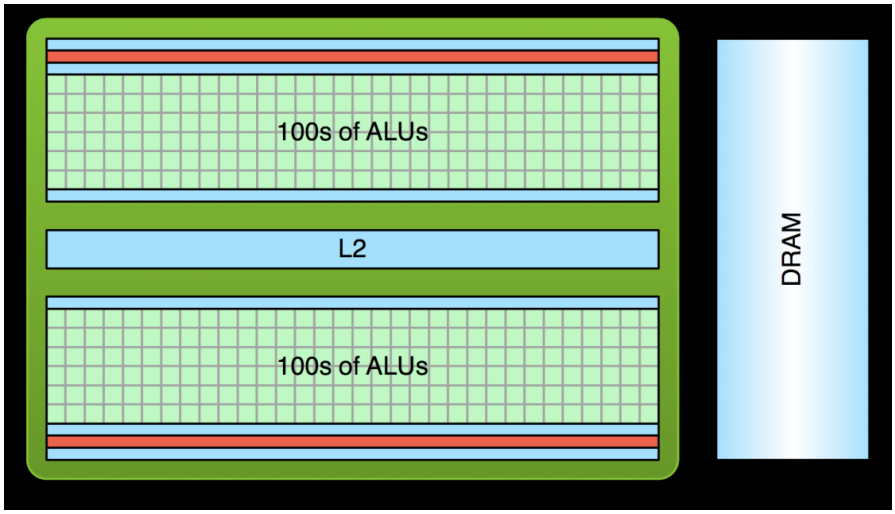
Οι διαφορές στην αρχιτεκτονική GPU και CPU έχουν να κάνουν με τον σκοπό για τον οποίο τις χρειαζόμαστε. Η GPU χρησιμοποιείται για γρήγορους μαθηματικούς υπολογισμούς, ενώ η CPU για υποστήριξη όλων των ειδών των εφαρμογών.

Οπότε, η CPU χρειάζεται αρκετή cache μνήμη επιπέδου L2 και τρέχει μερικές δεκάδες νήματα, ενώ η GPU έχει πολλές μονάδες υπολογισμού και τρέχει δεκάδες χιλιάδες από νήματα.

Στην επόμενη εικόνα φαίνεται ένα τυπικό σχέδιο μιας CPU και στην αμέσως επόμενη εικόνα βλέπουμε ένα αντίστοιχο σχέδιο για την GPU.



Εικόνα 37: Τυπικός πυρήνας CPU.

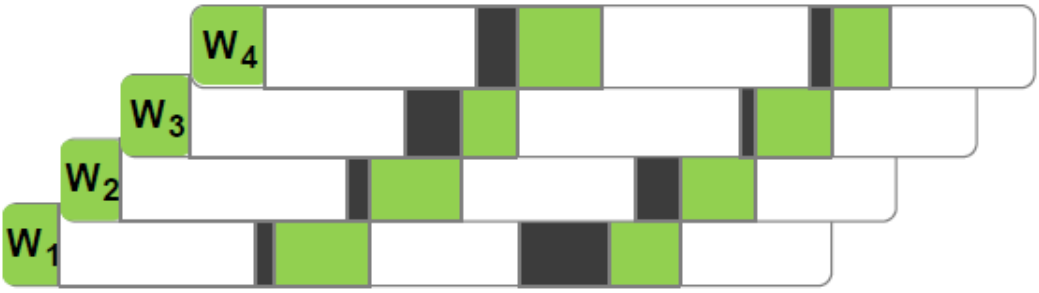


Εικόνα 38: Τυπικός πυρήνα GPU.

Σημαντικές διαφορές βλέπουμε και στον τρόπο που διαχειρίζονται την διαθέσιμη μνήμη. Στην επόμενη εικόνα έχουμε διεργασίες σε μια CPU, οι οποίες βρίσκονται σε κατάσταση εξυπηρέτησης (πράσινο χρώμα), αναμονής για δεδομένα από την μνήμη (λευκό χρώμα) και εναλλαγής μεταξύ τους (γκρίζο χρώμα).

Εικόνα 39: Κενά επεξεργασίας στην CPU.

Η αντίστοιχη εικόνα για την GPU φαίνεται στην συνέχεια. Τα πολλά νήματα επιτρέπουν την εξυπηρέτηση πολλών διεργασιών, οι οποίες θα εξυπηρετούνται (πράσινο χρώμα), θα αναμένουν για δεδομένα (άσπρο χρώμα), θα είναι έτοιμες για επεξεργασία (μαύρο χρώμα). Όμως, σε κάθε κβάντο χρόνου μία από αυτές θα εξυπηρετείται, με τον συνολικό χρόνο εξυπηρέτησης να μειώνεται σημαντικά.



Εικόνα 40: Κενά επεξεργασίας στην GPU.

Οπότε, οι GPU είναι κατάλληλες για μαθηματικούς υπολογισμούς και εξυπηρετούν τις διεργασίες τους πιο γρήγορα, λόγε αρχιτεκτονικού σχεδιασμού και χρήσης πολλών νημάτων.

Από το τελευταίο καταλαβαίνουμε τη χρήση της GPU έναντι της CPU στις επιστημονικές μελέτες. Είναι χαρακτηριστική η εργασία του Daniel Schlegel (Schlegel, 2015), στην οποία τονίζει την αύξηση της χρήσης της υπολογιστικής ισχύς της GPU στην επιστήμη και τονίζει την χρήση της στην τεχνητή νοημοσύνη, μέσα από τη εκτέλεση νευρωνικών δικτύων στην GPU, με τις Λογικές Αριθμητικές Μονάδες να εκτελούν ταυτόχρονα σε πολλά νήματα.

Στο επόμενο κεφάλαιο θα παρακολουθήσουμε τι σημαίνουν όλα αυτά για το μέλλον στις κάρτες γραφικών και θα δούμε πως θα επηρεάσουν τον τρόπο με τον οποίο θα χρησιμοποιούνται οι GPU στο μέλλον.

ΠΕΜΤΟ ΚΕΦΑΛΑΙΟ: ΣΥΜΠΕΡΑΣΜΑΤΑ

Σύμφωνα με τα όσα παρακολουθήσαμε μέχρι τώρα μπορούμε να καταλήξουμε σε ορισμένα χρήσιμα συμπεράσματα, μέσα από τα οποία θα προσπαθήσουμε να καταλάβουμε τον τρόπο με τον οποίο θα αξιοποιούνται οι GPU στο μέλλον.

Το βασικό στοιχείο που προκύπτει είναι ότι οι κάρτες γραφικών είναι απολύτως απαραίτητες για ένα πλήθος απαιτητικών εφαρμογών, όπως αυτές παρουσιάστηκαν στο ΚΕΦΑΛΑΙΟ 1. Αυτό το γεγονός μας δείχνει ότι θα συνεχίσουν να αναπτύσσονται για να καλύψουν την όλο και πιο απαιτητική αγορά.

Σε αντίθεση με αυτό, η CPU είναι απαραίτητο κομμάτι κάθε υπολογιστικού συστήματος, ακόμα και απλών οικιακών υπολογιστών. Οπότε, η ανάγκη της αγοράς δεν ωθεί την εξέλιξή τους στην απόκτηση μεγαλύτερων ταχυτήτων απαραίτητα, αλλά αυτές οι νέες ταχύτητες να ενσωματώνονται στην τεχνολογία τους και να φθάνουν στον καταναλωτή σε χαμηλή τιμή.

Από την άλλη, παρακολουθήσαμε στο ΚΕΦΑΛΑΙΟ 4 ότι από άποψη σχεδιασμού η GPU υπερέχει στην ταχύτητα εκτέλεσης μαθηματικών υπολογισμών, χρησιμοποιεί μνήμη που ανταποκρίνεται γρήγορα και με χαμηλότερο bandwidth.

Όλα αυτά μας δείχνουν πως οι GPU θα εξελίσσονται πιο γρήγορα από τις CPU και θα μεγαλώσουν το χάσμα στην ταχύτητα υπολογισμών, που ήδη έχει δημιουργηθεί μεταξύ τους.

Αποτέλεσμα αυτού θα είναι να βλέπουμε όλο και πιο συχνά servers και super υπολογιστές να εκτελούν τις εφαρμογές και τα πειράματά τους στην GPU και όχι στην CPU. Όλα αυτά μένει να επαληθευτούν στο μέλλον.

Βιβλιογραφία

Advanced Micro Devices, Inc, *Radeon's next-generation Vega architecture, Vega Whitepaper*, 2017.

Collange, Sylvain, *Introduction to GPU architecture*, ADA, 2017.

Coppock, Mark, "How to Choose a Graphic Card", Newegg Insider, 4-6-2018, [προσβάσιμο στο https://www.newegg.com/insider/how-to-choose-graphics-card/]

Coppock, Mark, "How to Choose a Graphics Card", Newegg Insider, 4-6-2018, [προσβάσιμο στο https://www.newegg.com/insider/how-to-choose-graphics-card/]

Cullinan, Christopher, Christopher Wyant, Timothy Frattesi, *Computing Performance Benchmarks among CPU, GPU, and FPGA*, MathWorks.

Duncan, Ralph, "A Survey of Parallel Computer Architectures", Survey & Tutorial Series, Φεβρουάριος 1990.

Flynn, Michael J., "Some Computer Organizations and Their Effectiveness", IEEE Transactions, Σεπτέμβριος 1972, σελ. 948.

Hruska, Joel, "How Graphics Cards Work", ExtremeTech, 16-5-2018, [προσβάσιμο στο https://www.extremetech.com/gaming/269335-how-graphics-cards-work]

Math Works, "Measuring GPU Performance", mathworks.com, [προσβάσιμο στο https://www.mathworks.com/help/distcomp/examples/measuring-gpu-performance.html]

McKane, Jamie, "How graphics cards have evolved over the years", MyBroadBand, 18-2-2017 [προσβάσιμο στο https://mybroadband.co.za/news/hardware/196964-how-graphics-cards-have-evolved-over-the-years.html]

Melendez, Steven, "The Purpose of a Graphics Card", Chron.com, 3-8-2018, [προσβάσιμο στο https://smallbusiness.chron.com/purpose-graphics-card-55327.html]

nVIDIA, *NVIDIA Tesla P100: The Most Advanced Datacenter Accelerator Ever Built, Featuring Pascal GP100, the World's Fastest GPU, Pascal Whitepaper*, NVIDIA Corporation, 2016.

Schlegel, Daniel, *Deep Machine Learning on GPUs*, Seminar Talk, University of Heidelberg, ZITI, 2015.

Smith, Ryan, "AMD Announces Radeon Instinct MI60 & MI50. Accelerators: Powered By 7nm Vega", AnandTech, 6-11-2018, [προσβάσιμο στο https://www.anandtech.com/show/13562/amd-announces-radeon-instinct-mi60-mi50-accelerators-powered-by-7nm-vega]

Techpowerup, "Graphics Card Market Up Sequentially in Q3, NVIDIA Gains as AMD Slips", [προσβάσιμο στο http://www.techpowerup.com/194979/graphics-card-market-up-sequentially-in-q3-nvidia-gains-as-amd-slips.html]

Tyson, Jeff & Wilson, V. Tracy, "How Graphics Cards Work", HowStuffWorks, [προσβάσιμο στο https://computer.howstuffworks.com/graphics-card.htm]

UK gaming computers, "How To Test Your Graphics Card", 9-2-2018, ukgamingcomputers.co.uk, [προσβάσιμο στο https://www.ukgamingcomputers.co.uk/blog/how-to-test-your-graphics-card/]

Vernon, Julian, "How to buy a graphics card—Six things you must know about GPUs", PC Gamer, 3-6-2014, [προσβάσιμο στο https://www.pcgamer.com/how-to-buy-a-graphics-card-six-things-you-must-know-about-gpus/]

von Neumann, John (1945), "First Draft of a Report on the EDVAC", Pennsylvania, 30-6-1945.

Betts, Andy, "Integrated vs. Dedicated Graphics Card: 7 Things You Need to Know", MUO, 20-9-2018, [προσβάσιμο στο https://www.makeuseof.com/tag/can-shared-graphics-finally-compete-with-a-dedicated-graphics-card/]