A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution

Presented by:
Βασιλείου Τερέζα-Άννα
Κουρής Γεώργιος
Ρόζος Μάριος

Τέλος Νόμου Moore & Dennard scaling

- Συνέπειες: Έκλειψη ταχείων βελτιώσεων στην απόδοση προγραμμάτων γενικού σκοπού
- Εναλλακτική Λύση: Μηχανική Μάθηση
    - Τομέας που αξίζει να ερευνηθεί, καθώς εφαρμόζεται σε πολλά πεδία, ενώ παράλληλα έχει δυνατότητα εμβάθυνσης.
    - Απαιτείται συνεργασία ειδικών ML και αρχιτεκτόνων υπολογιστών για να αναδειχθούν όλες οι δυνατότητες του κλάδου.
    - Ανάγκη για σχεδιασμό νέων, προηγμένων συστημάτων ML
- Επανάσταση των Βαθιών Νευρωνικών Δικτύων την τελευταία 5ετία ενδεικτικά στους παρακάτω τομείς:
  Φωνή, Όραση, Κατανόηση Γλώσσας, Μετάφραση φυσικής γλώσσας, Ιατρική Διάγνωση, Αναζήτηση στο Διαδίκτυο, Παιχνίδια

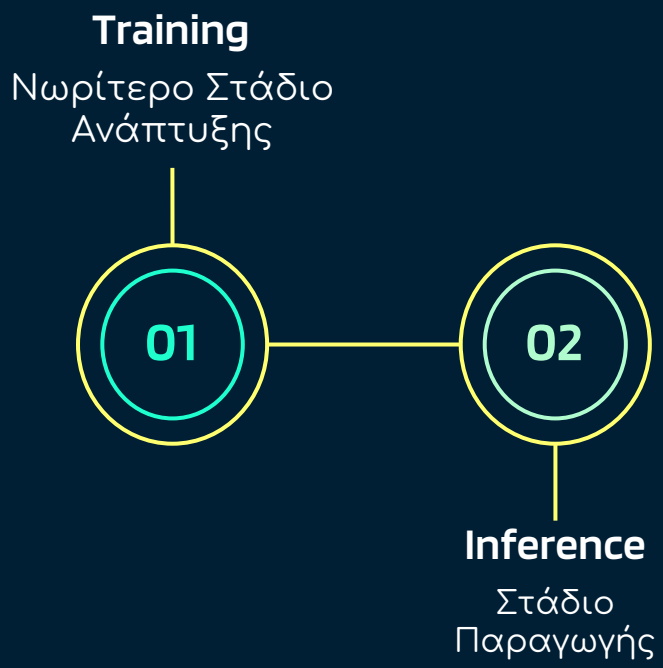
Προκλήσεις σημερινής κοινωνίας

Make solar energy affordable	Reverse-engineer the brain	Provide energy from fusion	Prevent nuclear terror
Develop carbon sequestration methods	Secure cyberspace	Manage the nitrogen cycle	Enhance virtual reality
Provide access to clean water	Advance personalized learning	Restore and improve urban infrastructure	Engineer the tools for scientific discovery
Advance health informatics	Engineer better medicines	Enable universal communication	Build flexible general-purpose AI systems

Επανάσταση ML

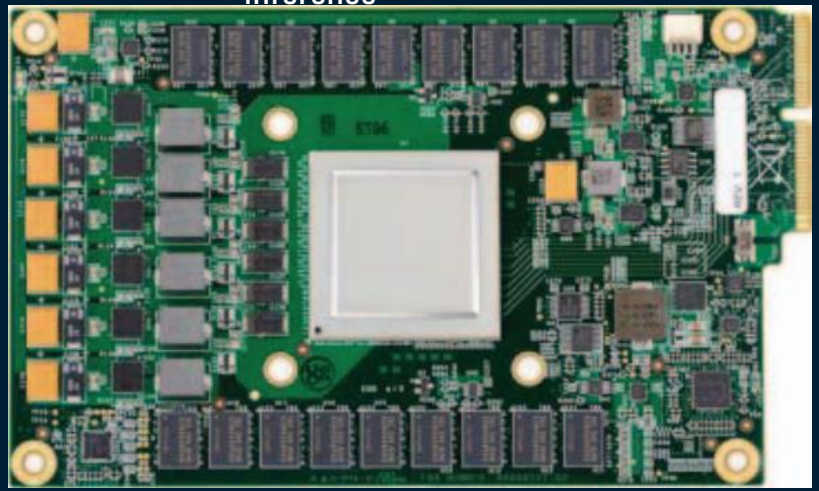
- Κριτήρια:
  - Σύνολα Δεδομένων
  - Υπολογιστικοί Πόροι που απαιτούνται για την ανάλυση των συνόλων
- Μέχρι τώρα οι υπολογισμοί αυτοί γίνονταν με μεγάλα datacenters με σκοπό την επιτάχυνση των γραφικών και όχι το ίδιο το ML.
  → Ανάγκη για datacenters μόνο για ML.

- 2013: Google → σχεδίαση ML ASICs (Application Specific Integrated Circuits).
  2015: χρήση στα υπολογιστικά κέντρα

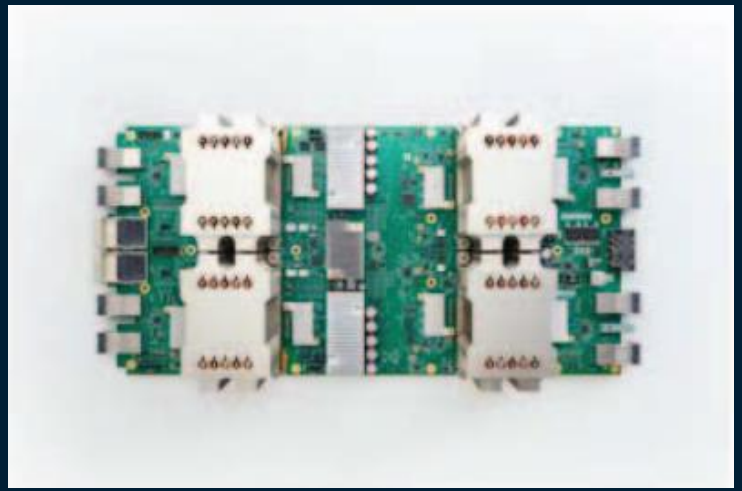


Google's TPU boards

1st gen TPU → Inference



2st gen TPU → Training



Προκλήσεις και Ευκαιρίες για τους Computer Architects

1. Απεριόριστη Ζήτηση για κύκλους ML:
   - Κατώφλι υπομονής για πειραματικά αποτελέσματα
   - Γρηγορότερο hardware επιτρέπει μεγαλύτερου βεληνεκούς πειράματα
2. Το τέλος του Νόμου του Moore:
   - Δεν υπάρχει πλέον εκθετική αύξηση στους πόρους
   - Απαιτούνται αποδοτικές λύσεις για εκθετικά δυσκολότερα προβλήματα
     - Σχεδιασμός υλικού που να είναι relevant σε παράθυρο 5ετίας
3. Προσαρμογή στους διαρκώς μεταβαλλόμενους αλγορίθμους του ML:
   - Ισορροπία μεταξύ ειδίκευσης και ευελιξίας
   - Χρήση high level εργαλείων για τις νέες τεχνολογίες, όπως TensorFlow
4. Ανάγκη για προχωρημένες τεχνικές μεταγλώττισης (compilation techniques)

!! Δεν υπάρχει καθολικά ιδανική λύση → Απαιτείται ευελιξία

6 θέματα για τη Σχεδίαση Υλικού

1) Training
   - Έλλειψη στον αρχιτεκτονικό σχεδιασμό του training.
   - Πιο εύκολα επιλύσιμο το θέμα του inference από αυτού του training.
   - Διατήρηση τιμών ενεργοποίησης για update του accuracy με back-propagation

1) Batch Size
   - Κάποια νευρωνικά και οι GPUs, λειτουργούν καλά για μεγάλα batch sizes.
   - Για μικρά batch sizes πρέπει να βρούμε:
     άλλους αλγορίθμους και διαφορετικές αρχιτεκτονικές

1) Sparsity & Embeddings
   - Fine-grain sparsity & coarse-grain sparsity
   - Το Mixture of Experts αποσυνθέτει το μοντέλο, εκπαιδεύει ένα expert model για κάθε μέρος και συνδυάζει τις προβλέψεις.
   - Τα embeddings μετατρέπουν αραιούς χώρους σε συμπαγείς δομές. Κλειδί για εφαρμογές μετάφρασης και γενικότερα κειμένων, αλλά δε χρησιμοποιούνται πολύ.

...Συνέχεια

4) Quantization & Distillation
- Αριθμητικές Αναπαραστάσεις μειωμένης ακρίβειας ή κβαντισμένες (quantized) είναι αποτελεσματικές σε επιταχυντές inference και χρησιμοποιούνται πλέον στις GPUs.
- Οι μελέτες στη συγκεκριμένη περιοχή αφορούν σε toy-sized datasets.
- Distillation: χρήση μεγαλύτερου μοντέλου για μεταφορά γνώσης σε μικρότερο επιτυγχάνοντας μεγαλύτερη ακρίβεια.

5) Soft Memory (πχ. Neural Turing Machines, memory networks, attention)
- Λειτουργία τεχνικών βαθιάς μάθησης παρόμοιες με RAM. (back-propagation)
- Soft memory: υπολογισμός μέσου όρου όλων των καταχωρήσεων με βάρη. Πιο ακριβό.
- Hard memory: μία μόνο καταχώρηση

6) Learning To Learn (L2L)
- Ανάπτυξη μοντέλων νευρωνικών δικτύων
- Περιορισμός του ανθρώπινου παράγοντα
- Συνδυασμός αναζήτησης hardware και software

Πλάνες & Παγίδες

Πλάνη

Έμφαση στο throughput έναντι του latency

Πλάνη

Θυσία ακρίβειας το βωμό της ταχύτητας

Παγίδα

Σχεδίαση hardware με βάση αναχρονιστικά μοντέλα

Παγίδα

Σχεδίαση hardware αγνοώντας το software

Συμπεράσματα

01

Πολλά υποσχόμενη η σχεδίαση hardware ειδικά για Μηχανική Μάθηση

02

Σημαντικό Πρόβλημα: έλλειψη ML experts
→ η λύση ίσως το L2L

03

Απαραίτητη η συνεργασία computer architects και ML experts

04

Από άποψη αρχιτεκτονικής, οι επόμενες δεκαετίες θα είναι συναρπαστικές

Ευχαριστούμε για την προσοχή σας!

Υπάρχουν ερωτήσεις;