Υποδομές Cloud

Διάλεξη 1η

Θωμάς Καμαλάκης

Χαροκόπειο Πανεπιστήμιο Αθηνών

Χειμερινό Εξάμηνο, 2024

Περιεχόμενα

1 Πληροφορίες για το μάθημα

2 Ιστορικά Στοιχεία

3 Ορισμοί

4 Η αγορά

Στόχοι του μαθήματος

- Ο στόχος του μαθήματος είναι να μάθουμε τα βασικά δομικά στοιχεία που απαρτίζουν το υπολογιστικό νέφος:
  - Τεχνολογίες Υλικού
  - Τεχνολογίες Λογισμικού

Τεχνολογίες Υλικού

- Τεχνολογίες Υλικού:
  - Διακομιστές (servers)
  - Αποθηκευτικά μέσα (storage)
  - Συσκευές Δικτύου (switching)
  - Ψυκτικά μέσα (cooling)
  - Συσκευές Ασφάλειας (HSM/firewalls)
  - Εξοπλισμός Άκρου (edge hardware)

Τεχνολογίες Λογισμικού

- Τεχνολογίες Λογισμικού:
  - Hypervisors
  - Container Orchestration (Docker/Kubernetes)
  - Πλατφόρμες διαχείρισης (π.χ. Proxmox, Azure, Vcenter)
  - Αυτοματοποίηση (π.χ. Ansible, Terraform)
  - Monitoring και Logging

Φιλοσοφία μαθήματος

- Πως θα προχωρήσουμε;
  - Θα εξερευνουμε κάποιες βασικές έννοιες στο μάθημα.
  - Στο βαθμό του εφικτού θα δούμε και κάποια hands on παραδείγματα.
  - Κυρίως θα χρησιμοποιούμε το Proxmox.
  - Στα παραδείγματα θα προσπαθήσουμε να χρησιμοποιήσουμε `Python`

Αξιολόγηση

- Αξιολόγηση μαθήματος;
  - Με απαλλακτική εργασία στο τέλος του εξαμήνου.

Στοιχεία Διδασκόντων

- Θωμάς Καμαλάκης, Καθηγητής
- thkam@hua.gr
- Αλέξανδρος Δημόπουλος, Επίκουρος Καθηγητής
- alexdem@hua.gr

Links και Βιβλιογραφία

- E-class: `https://eclass.hua.gr/courses/DIT292/`
- Proxmox wiki: `https://pve.proxmox.com/wiki/Main_Page`
- Βιβλία:
  - Erl, Thomas, Ricardo Puttini, and Zaigham Mahmood. Cloud computing: concepts, technology & architecture. Pearson Education.
  - Sosinsky, Barrie. Cloud computing bible. John Wiley & Sons.
  - Kavis, Michael. Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS). John Wiley & Sons, Inc., Hoboken, New Jersey

Project MAC: Mathematics and Computation

- Ίδρυση: Το Project MAC ξεκίνησε στο MIT με στόχο την ανάπτυξη συστήματος χρονομοιρασμού, που θα επέτρεπε την ταυτόχρονη χρήση ενός κεντρικού υπολογιστή από πολλούς χρήστες, από απομακρυσμένες τοποθεσίες (1963).
- Χρηματοδοτήθηκε από την ARPA (Advanced Research Projects Agency) και το Εθνικό Ίδρυμα Επιστημών (NSF) των ΗΠΑ.
- Compatible Time-Sharing System αποτέλεσε τη βάση του Project MAC, επιτρέποντας πολλούς χρήστες να τρέχουν προγράμματα σε έναν κεντρικό υπολογιστή.

Project MAC: Mathematics and Computation

- Μέσα σε έξι μήνες από την έναρξη του Project MAC, 200 χρήστες σε 10 διαφορετικά τμήματα του MIT μπορούσαν να έχουν πρόσβαση στο σύστημα.
- Το Project MAC, σε συνεργασία με τα Bell Labs και την General Electric, ανέπτυξε το Multics, ένα εξελιγμένο σύστημα χρονομοιρασμού που εισήγαγε λειτουργίες όπως διαχείριση αρχείων και ασφάλεια συστήματος.
- Το Multics μπορούσε να υποστηρίξει 300 ταυτόχρονους χρήστες σε 1.000 τερματικούς σταθμούς στο MIT.
- Θεωρείται πρόδρομος του UNIX, ένα πιο απλό σύστημα που αναπτύχθηκε από τα Bell Labs, με τεράστια επίδραση στην ανάπτυξη των λειτουργικών συστημάτων.

Θεωρητική Θεμελίωση και Πρώτες Τεχνολογίες (1960-1990)

- Κύριες Ιδέες: Η ιδέα πίσω από το cloud computing ξεκίνησε από τη δεκαετία του 1960, όταν ο John McCarthy, ένας από τους πατέρες της τεχνητής νοημοσύνης, πρότεινε ότι η πληροφορική θα μπορούσε να προσφέρεται ως δημόσια υπηρεσία, όπως το νερό ή το ηλεκτρικό ρεύμα.
- Τεχνολογίες: Την ίδια περίοδο αναπτύσσονται τα πρώτα time-sharing συστήματα, στα οποία πολλοί χρήστες μπορούσαν να μοιράζονται την ίδια υπολογιστική ισχύ μέσω απομακρυσμένων τερματικών.
- Παράδειγμα: Το ARPANET, το πρώιμο διαδίκτυο, συνδέει πανεπιστήμια και ερευνητικά κέντρα για κοινή χρήση υπολογιστικών πόρων.

Ανάπτυξη Εικονικών Μηχανών (1990-2000)

- Η τεχνολογία της εικονικοποίησης (virtualization) επιτρέπει την εκτέλεση πολλών «εικονικών μηχανών» σε έναν φυσικό υπολογιστή, βελτιώνοντας την αποδοτικότητα της χρήσης υπολογιστικών πόρων. Οι διαχειριστές μπορούσαν πλέον να εκτελούν διαφορετικά λειτουργικά συστήματα και εφαρμογές ταυτόχρονα στον ίδιο εξυπηρετητή.
- VMware ήταν από τους πρωτοπόρους της εικονικοποίησης, λανσάροντας το 1999 το πρώτο εμπορικό λογισμικό που επέτρεπε τη χρήση εικονικών μηχανών σε servers.
- Ο διαμοιρασμός πόρων και η απομόνωση διαδικασιών μέσω virtualization αποτέλεσαν θεμέλιο λίθο για το μέλλον του cloud computing.

Ανάπτυξη Εικονικών Μηχανών (2000-2010)

- Το cloud computing αρχίζει να καθιερώνεται ως εμπορική υπηρεσία. Οι επιχειρήσεις δεν χρειάζεται πλέον να επενδύουν σε δική τους υποδομή, αλλά μπορούν να «νοικιάζουν» υπολογιστική ισχύ, αποθηκευτικό χώρο και λογισμικό από παρόχους.
- Το 2006, η Amazon Web Services (AWS) λάνσαρε το Elastic Compute Cloud (EC2), που επέτρεπε σε προγραμματιστές να «ενοικιάζουν» εικονικούς υπολογιστές για την εκτέλεση των εφαρμογών τους.
- Εμφάνιση των SaaS (Software as a Service) υπηρεσιών όπως το Salesforce, που επέτρεπε στις επιχειρήσεις να χρησιμοποιούν λογισμικό μέσω διαδικτύου αντί να το εγκαθιστούν τοπικά.
- Το AWS αποτέλεσε τον βασικό κινητήριο μοχλό για τη δημιουργία υπηρεσιών υποδομής cloud (Infrastructure as a Service - IaaS).

Ωρίμανση του Cloud (2010-2015)

- Το cloud computing γίνεται πιο προσιτό και υιοθετείται μαζικά από επιχειρήσεις, κυβερνήσεις και ιδιώτες. Εμφανίζονται νέες μορφές υπηρεσιών όπως το Platform as a Service (PaaS) και το Infrastructure as a Service (IaaS), που επιτρέπουν στις επιχειρήσεις να δημιουργούν και να διαχειρίζονται εφαρμογές σε κλίμακα χωρίς να ασχολούνται με την υποδομή.
- Το Microsoft Azure και το Google Cloud Platform (GCP) καθίστανται σημαντικοί παίκτες στο χώρο του cloud, προσφέροντας πλατφόρμες για υπολογιστική ισχύ και αποθήκευση.
- Containerization: Η τεχνολογία των containers (π.χ. Docker) επιτρέπει την απομόνωση εφαρμογών με μεγαλύτερη ευελιξία από την παραδοσιακή εικονικοποίηση.

Hybrid Cloud και Edge Computing (2015-σήμερα)

- Οι επιχειρήσεις αρχίζουν να υιοθετούν υβριδικές και πολυ-νεφικές λύσεις, συνδυάζοντας ιδιωτικά (on-premises) και δημόσια cloud για μεγαλύτερη ευελιξία, ασφάλεια και προσαρμοστικότητα στις ανάγκες τους.
- Ταυτόχρονα, το edge computing γίνεται πιο διαδεδομένο, επιτρέποντας την επεξεργασία δεδομένων κοντά στην πηγή τους.

AI Cloud: Η νέα εποχή του cloud (2023-σήμερα)

- Η GenAI γίνεται ο κύριος μοχλός ανάπτυξης του cloud. Το Q2 2026 οι δαπάνες για cloud υποδομές έφτασαν τα $143 δισ. (+43% σε ετήσια βάση, ο υψηλότερος ρυθμός της τελευταίας 8ετίας), ενώ οι υπηρεσίες GenAI αυξάνονται κατά ∼165%.
- Οι hyperscalers (Amazon, Microsoft, Google, Meta) σχεδιάζουν για το 2026 κεφαλαιακές δαπάνες (capex) άνω των $650 δισ.: Amazon ∼$200 δισ., Google ∼$180 δισ., Meta ∼$125 δισ., Microsoft ∼$120 δισ. Το μεγαλύτερο μέρος αφορά AI data centers.
- Μερίδια αγοράς (Q2 2026): AWS 28%, Azure 20%, Google Cloud 15%. Εμφανίζονται τα *neoclouds* (CoreWeave, Nebius, Crusoe, Oracle OCI) που εξειδικεύονται στην ενοικίαση GPU.
- Το cloud μετατρέπεται από «ενοικίαση εικονικών μηχανών» σε «ενοικίαση υπολογιστικής ισχύος για εκπαίδευση και inference μοντέλων».

AI Cloud: Πώς άλλαξε η υποδομή των hyperscalers

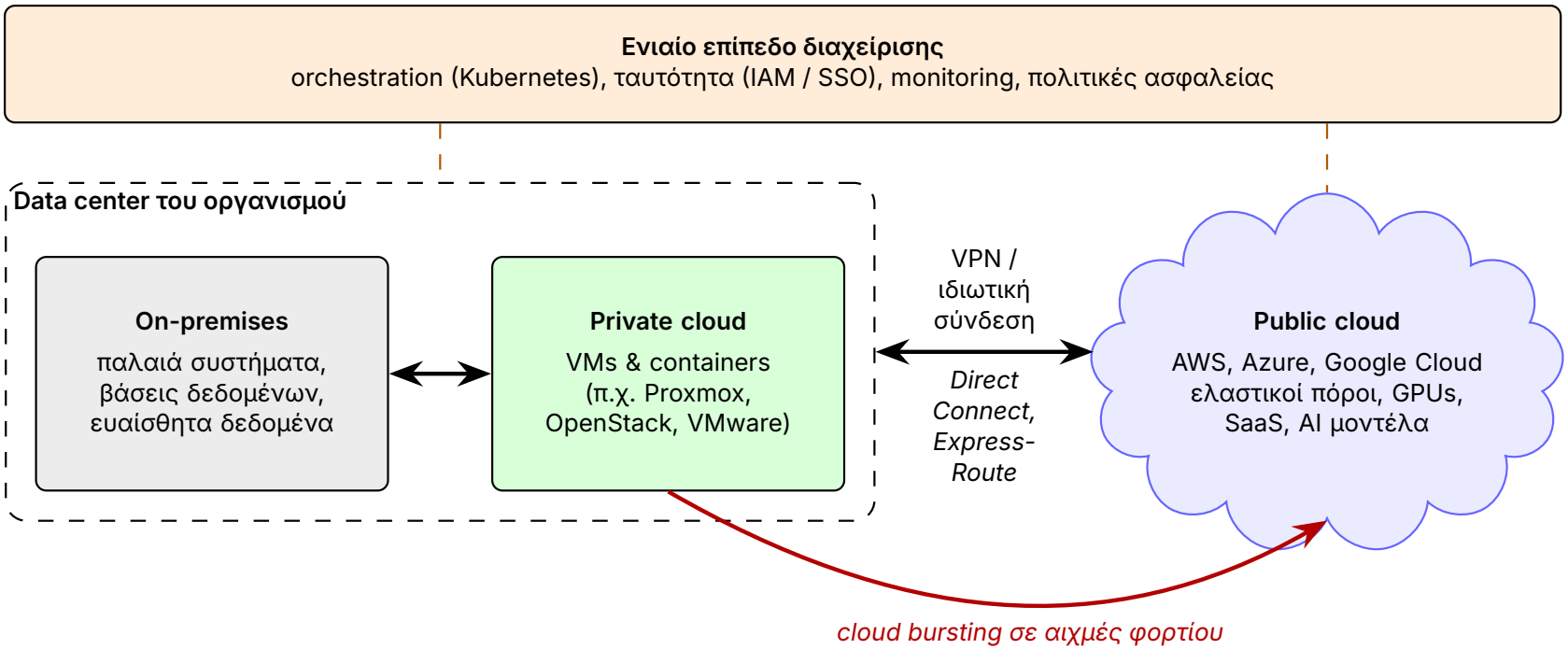
- **Επιταχυντές αντί για CPUs**: Οι νέοι εξυπηρετητές βασίζονται σε GPUs (NVIDIA Blackwell, Vera Rubin) και σε custom chips των ίδιων των παρόχων (Google TPU Ironwood, AWS Trainium, Microsoft Maia).
- **Ισχύς ανά rack**: Από 5–15 kW στα «κλασικά» data centers σε 120–140 kW ανά rack (π.χ. NVIDIA GB300 NVL72), με πορεία προς τα $\sim$600 kW.
- **Υγρή ψύξη**: Η ψύξη με αέρα δεν επαρκεί· γενικεύεται η ψύξη υγρού απευθείας στο chip (direct-to-chip liquid cooling) σε κλειστό κύκλωμα.
- **Δικτύωση**: Εκατοντάδες χιλιάδες GPUs πρέπει να λειτουργούν ως ένας υπερυπολογιστής (π.χ. Microsoft Fairwater), με δίκτυα υψηλού εύρους ζώνης (NVLink, InfiniBand, Ethernet 800G).
- **Ενέργεια**: Τα data centers μετριούνται πλέον σε gigawatts (π.χ. Stargate της OpenAI $>$10 GW). Οι πάροχοι επενδύουν σε δική τους παραγωγή ενέργειας, ακόμη και σε πυρηνικούς σταθμούς.

Cloud computing

Το *cloud computing* είναι η πρόσβαση κατ' απαίτηση σε υπολογιστικούς πόρους — φυσικούς ή εικονικούς διακομιστές, αποθήκευση δεδομένων, δυνατότητες δικτύωσης, εργαλεία ανάπτυξης εφαρμογών, λογισμικό, εργαλεία ανάλυσης με τεχνητή νοημοσύνη και άλλα — μέσω του διαδικτύου, με τιμολόγηση συνήθως βάσει χρήσης.

Hybrid cloud

Το *hybrid cloud* συνδυάζει και ενοποιεί το public cloud, το private cloud και την τοπική υποδομή (on-premises) για να δημιουργήσει μια ενιαία, ευέλικτη και οικονομικά βέλτιστη υποδομή πληροφορικής.



Infrastructure as a Service - IaaS

- Το IaaS (Infrastructure as a Service) είναι ένα μοντέλο υπηρεσίας cloud που παρέχει υπολογιστική υποδομή μέσω του διαδικτύου.
- Οι χρήστες μπορούν να ενοικιάζουν πόρους υποδομής, όπως διακομιστές, αποθηκευτικό χώρο και δίκτυα, αντί να επενδύουν σε φυσικό hardware και να το διαχειρίζονται οι ίδιοι.
- Οι παροχείς IaaS φροντίζουν για τη συντήρηση, την ασφάλεια και την κλιμάκωση της υποδομής, ενώ οι πελάτες μπορούν να δημιουργούν και να διαχειρίζονται τις δικές τους εφαρμογές, βάσεις δεδομένων και λειτουργικά συστήματα.

Infrastructure as a Service - IaaS

- Amazon Web Services (AWS): Με υπηρεσίες όπως το Amazon EC2 (Elastic Compute Cloud) για εικονικούς διακομιστές, το Amazon S3 (Simple Storage Service) για αποθήκευση δεδομένων και το Amazon VPC (Virtual Private Cloud) για απομονωμένα δίκτυα.

Platform as a Service - PaaS

- Το PaaS (Platform as a Service) είναι ένα μοντέλο υπηρεσίας cloud που παρέχει στους χρήστες μια πλατφόρμα για την ανάπτυξη, εκτέλεση και διαχείριση εφαρμογών χωρίς να χρειάζεται να ασχοληθούν με τη διαχείριση της υποκείμενης υποδομής (π.χ. servers, αποθήκευση, δίκτυα).
- Οι προγραμματιστές μπορούν να εστιάσουν στην ανάπτυξη κώδικα και λειτουργιών, ενώ ο πάροχος της υπηρεσίας φροντίζει για την κλιμάκωση, την ασφάλεια και τις ενημερώσεις της πλατφόρμας.
- Google App Engine: Στήνει κανείς εφαρμογές σε πλατφόρμες απευθείας, π.χ. Django, Node.js

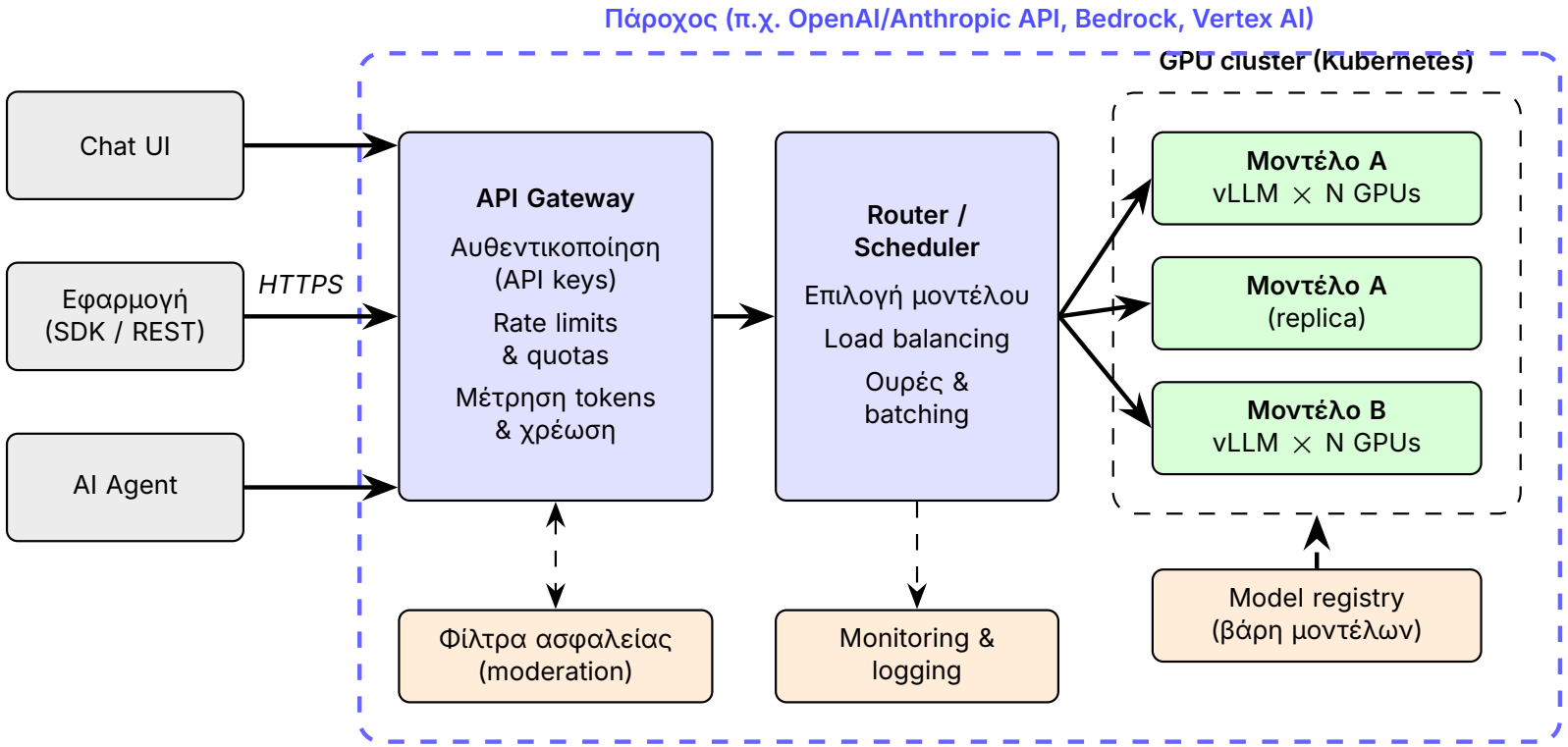
Software as a Service - SaaS

- Το SaaS (Software as a Service) είναι ένα μοντέλο παροχής λογισμικού στο οποίο οι εφαρμογές φιλοξενούνται στο cloud και είναι διαθέσιμες στους χρήστες μέσω του διαδικτύου.
- Αντί να εγκαθιστούν και να διαχειρίζονται το λογισμικό τοπικά στους υπολογιστές τους, οι χρήστες αποκτούν πρόσβαση στις εφαρμογές μέσω ενός web browser, πληρώνοντας συνήθως μια συνδρομή ή με βάση τη χρήση.
- Παραδείγματα SaaS περιλαμβάνουν εφαρμογές όπως το Google Workspace και το Office 365.

AI μοντέλα ως υπηρεσία cloud (Model as a Service)

- Οι περισσότερες εφαρμογές δεν εκπαιδεύουν δικά τους μοντέλα· καλούν ένα μοντέλο μέσω API και πληρώνουν ανά χρήση, συνήθως ανά εκατομμύριο tokens εισόδου/εξόδου.
- Είναι η ίδια λογική με το IaaS/PaaS/SaaS: ο πάροχος αναλαμβάνει GPUs, κλιμάκωση και διαθεσιμότητα, ο χρήστης βλέπει μόνο ένα endpoint.
- Το 2026 διακρίνουμε τρεις βασικούς τρόπους πρόσβασης σε ένα μοντέλο:
  1. **Closed-weight**: τα βάρη του μοντέλου δεν δημοσιεύονται· διατίθεται μόνο μέσω API από την εταιρεία που το ανέπτυξε και τους συνεργάτες της.
  2. **Open-weight μέσω παρόχων**: τα βάρη είναι δημόσια, οπότε πολλοί διαφορετικοί πάροχοι μπορούν να φιλοξενήσουν το ίδιο μοντέλο και να ανταγωνίζονται σε τιμή και ταχύτητα.
  3. **Self-hosted**: ο οργανισμός κατεβάζει ένα open-weight μοντέλο και το τρέχει στους δικούς του servers (on-premises ή σε δικές του GPU VMs).

Αρχιτεκτονική Model as a Service



- Ο χρήστης στέλνει prompt και παίρνει πίσω tokens (συχνά με streaming)· χρεώνεται ανά εκατομμύριο tokens εισόδου/εξόδου.
- Όλα όσα είναι μέσα στο μπλε πλαίσιο (GPUs, κλιμάκωση, διαθεσιμότητα) τα διαχειρίζεται ο πάροχος.

Closed-weight μοντέλα (2026)

- Τα πιο ισχυρά (frontier) μοντέλα παραμένουν κλειστά:
  - **OpenAI**: οικογένεια GPT-5 (π.χ. GPT-5.5).
  - **Anthropic**: οικογένεια Claude (π.χ. Claude Opus 5, Sonnet 5).
  - **Google**: οικογένεια Gemini 3.
- Διατίθενται μέσω του API της εταιρείας, αλλά και μέσω των πλατφορμών των hyperscalers: Amazon Bedrock, Microsoft Foundry, Google Vertex AI. Τον Απρίλιο του 2026 τα μοντέλα της OpenAI ήρθαν για πρώτη φορά στο Amazon Bedrock, τερματίζοντας την αποκλειστικότητα της Microsoft.
- Πλεονεκτήματα: κορυφαίες επιδόσεις, μηδενική διαχείριση υποδομής.
- Μειονεκτήματα: εξάρτηση από τον πάροχο (vendor lock-in), τα δεδομένα φεύγουν από τον οργανισμό, κόστος που αυξάνεται με τη χρήση.

Open-weight μοντέλα μέσω παρόχων (2026)

- Τα open-weight μοντέλα έχουν πλησιάσει σημαντικά τα κλειστά. Παραδείγματα του 2026:
  - DeepSeek V4 (DeepSeek), Qwen 3.5/3.6 (Alibaba), Kimi K2.6 (Moonshot AI), GLM-5 (Zhipu AI)
  - Llama 4 (Meta), gpt-oss (OpenAI), Gemma 4 (Google), Mistral (Γαλλία)
- Το ίδιο μοντέλο προσφέρεται από πολλούς παρόχους: hyperscalers (Bedrock, Foundry, Vertex AI) και εξειδικευμένους inference providers (Together AI, Fireworks, Groq, Cerebras, DeepInfra).
- Aggregators όπως το OpenRouter δίνουν πρόσβαση σε εκατοντάδες μοντέλα από δεκάδες παρόχους με ένα ενιαίο API, επιλέγοντας τον φθηνότερο ή ταχύτερο πάροχο.
- Ο ανταγωνισμός ρίχνει τις τιμές: τα open-weight μοντέλα κοστίζουν συχνά πολλαπλάσια λιγότερο ανά token από τα κλειστά.

Self-hosted μοντέλα

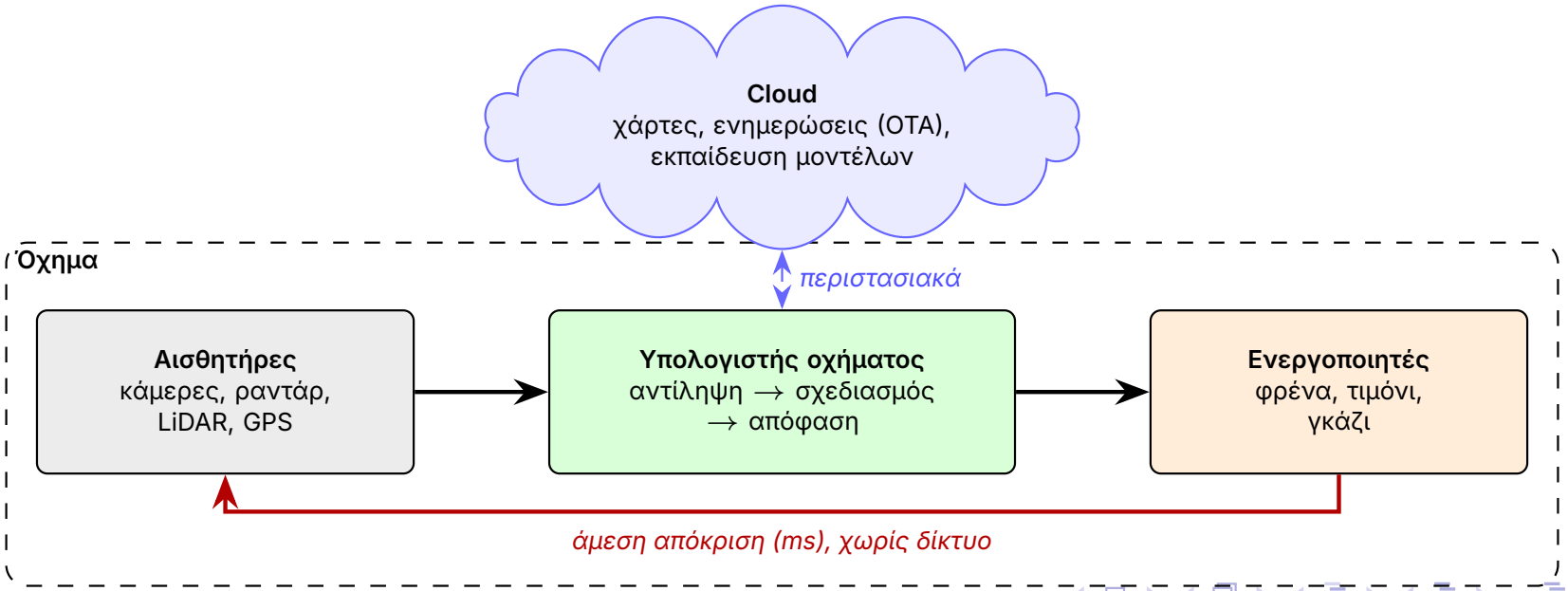
- Ο οργανισμός κατεβάζει τα βάρη (π.χ. από το Hugging Face) και τρέχει το μοντέλο σε δικές του GPUs (on-premises ή GPU VMs).
- Εργαλεία εξυπηρέτησης (inference servers):
  - **vLLM**, **SGLang**: για παραγωγικά περιβάλλοντα σε GPU clusters, συνήθως πάνω σε Kubernetes.
  - **Ollama**, **llama.cpp**: για τοπική εκτέλεση σε workstation ή laptop.
- Μικρότερα μοντέλα (π.χ. Qwen 3.6 27B, Gemma 4, gpt-oss-20b) τρέχουν σε μία GPU ή ακόμη και σε laptop· τα μεγάλα (DeepSeek V4, Kimi K2.6) απαιτούν server με πολλές GPUs.
- Πότε επιλέγεται: ευαίσθητα δεδομένα (υγεία, άμυνα, δημόσιο), κανονιστικές απαιτήσεις (GDPR, EU AI Act), σταθερό και μεγάλο φορτίο, ανάγκη για fine-tuning.
- Είναι στην ουσία η εφαρμογή του hybrid cloud στην AI: ευαίσθητα φορτία τοπικά, τα υπόλοιπα στο δημόσιο cloud.

Edge computing

- Το edge computing είναι μια τεχνολογία που φέρνει την επεξεργασία δεδομένων κοντά στην πηγή τους, δηλαδή στα «άκρα» του δικτύου, αντί να στέλνει όλα τα δεδομένα σε κεντρικούς διακομιστές για επεξεργασία.
- Αυτό μειώνει την καθυστέρηση (latency), βελτιώνει την ταχύτητα απόκρισης και μειώνει το φορτίο του δικτύου, καθώς τα δεδομένα επεξεργάζονται τοπικά ή κοντά στην τοποθεσία όπου δημιουργούνται.

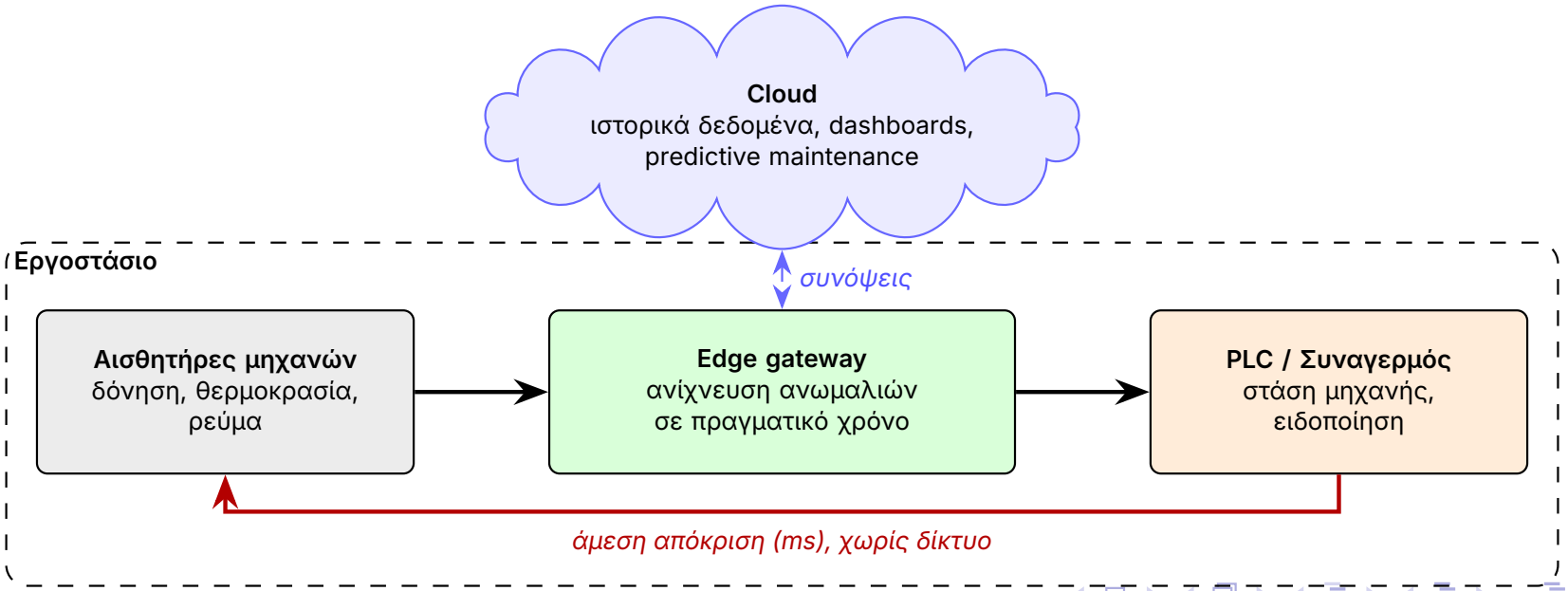
Παραδείγματα edge computing

- Αυτόνομα Οχήματα: Τα αυτόνομα οχήματα χρησιμοποιούν edge computing για να επεξεργάζονται δεδομένα από αισθητήρες (όπως κάμερες και ραντάρ) σε πραγματικό χρόνο, για να λαμβάνουν άμεσα αποφάσεις και να πραγματοποιούν ελιγμούς χωρίς καθυστέρηση που θα προέκυπτε αν τα δεδομένα έπρεπε να σταλούν σε κεντρικό διακομιστή.



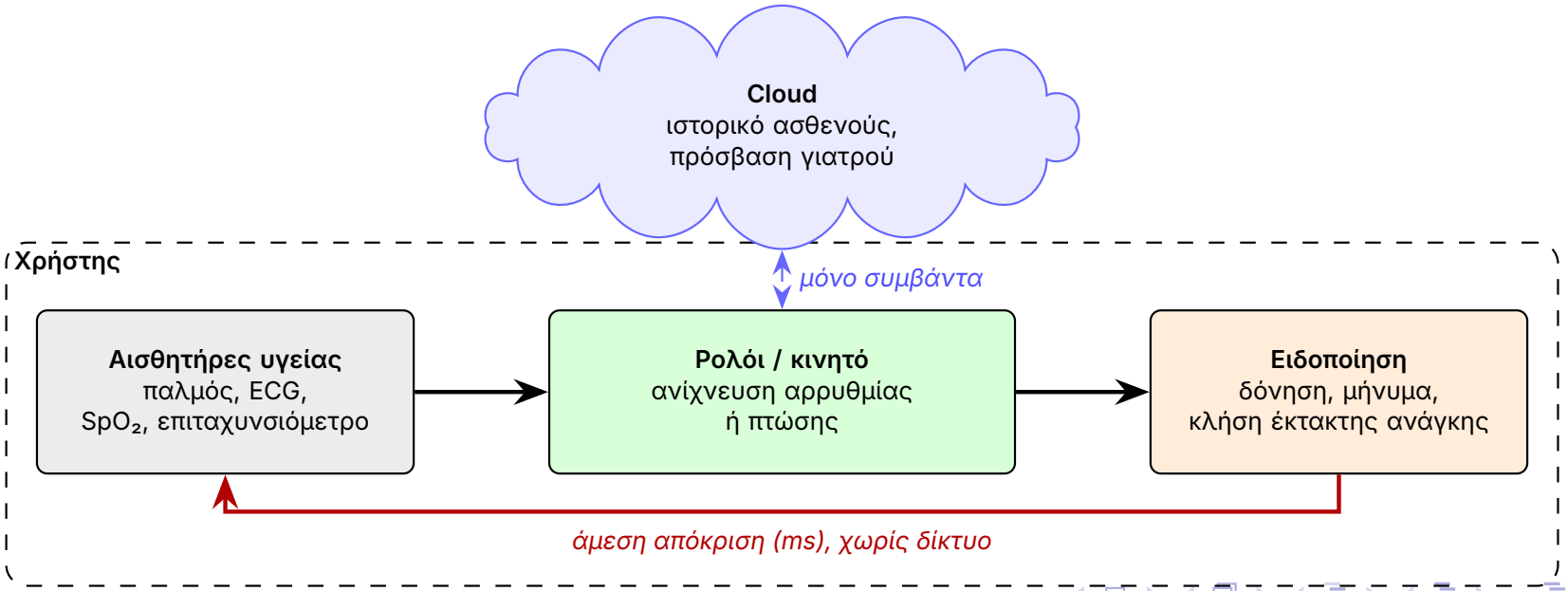
Παραδείγματα edge computing

- Συστήματα Εποπτείας Βιομηχανικών Εγκαταστάσεων: στη βιομηχανία, αισθητήρες που παρακολουθούν την κατάσταση μηχανημάτων ή εγκαταστάσεων μπορούν να χρησιμοποιούν edge computing για την άμεση ανάλυση δεδομένων και την ανίχνευση προβλημάτων, προλαμβάνοντας πιθανά σφάλματα ή διακοπές της λειτουργίας.



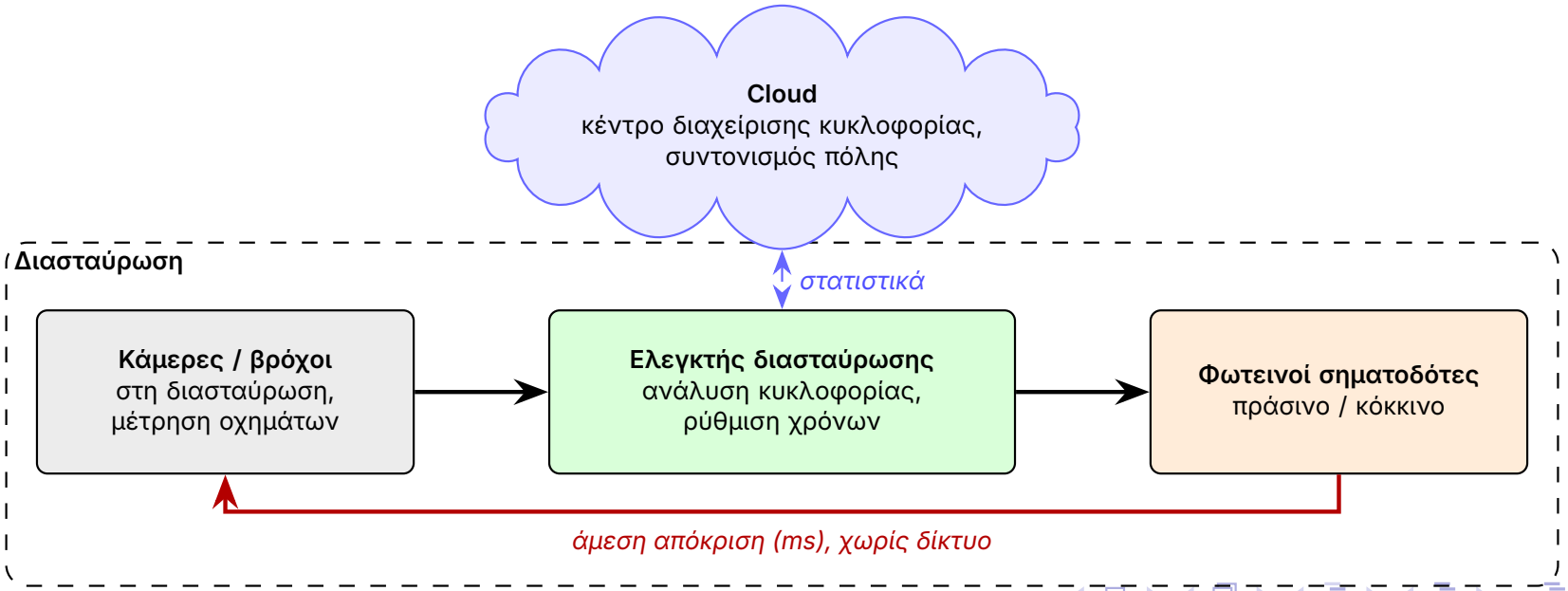
Παραδείγματα edge computing

- Έξυπνα Ρολόγια και Συσκευές Υγειονομικής Παρακολούθησης: Μπορούν να αναλύουν δεδομένα που συλλέγουν από αισθητήρες υγείας τοπικά, παρέχοντας άμεσες πληροφορίες στους χρήστες ή στους γιατρούς χωρίς να απαιτείται η αποστολή όλων των δεδομένων σε κεντρικό διακομιστή.



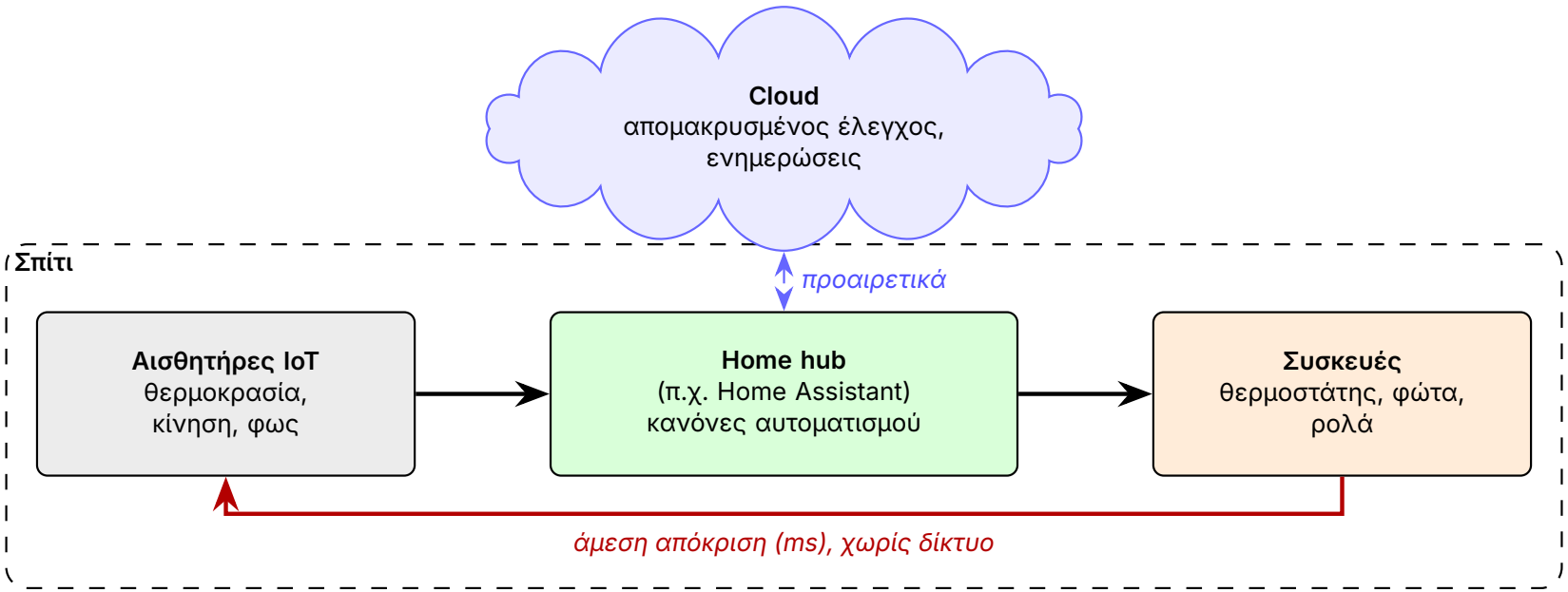
Παραδείγματα edge computing

- Έξυπνες Πόλεις: Συστήματα έξυπνης κυκλοφορίας μπορούν να χρησιμοποιούν edge computing για τη διαχείριση φωτεινών σηματοδοτών, τη συλλογή και ανάλυση δεδομένων κυκλοφορίας σε πραγματικό χρόνο και την προσαρμογή των σημάτων κυκλοφορίας για τη βελτίωση της ροής της κυκλοφορίας και της ασφάλειας.



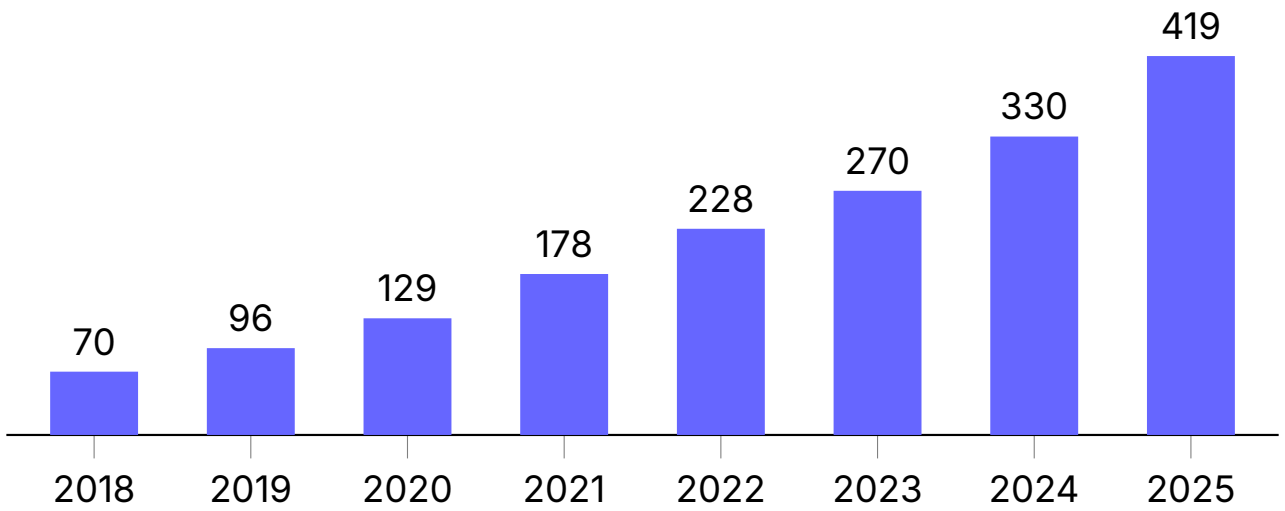
Παραδείγματα edge computing

- Έξυπνα Σπίτια: Συσκευές IoT σε έξυπνα σπίτια, όπως θερμοστάτες ή συστήματα φωτισμού, μπορούν να επεξεργάζονται δεδομένα τοπικά για να προσαρμόζουν αυτόματα τις ρυθμίσεις, χωρίς να χρειάζεται επικοινωνία με κεντρικό cloud.



Το μέγεθος της αγοράς cloud

Ετήσια έσοδα από υπηρεσίες cloud υποδομών (δισ. $)

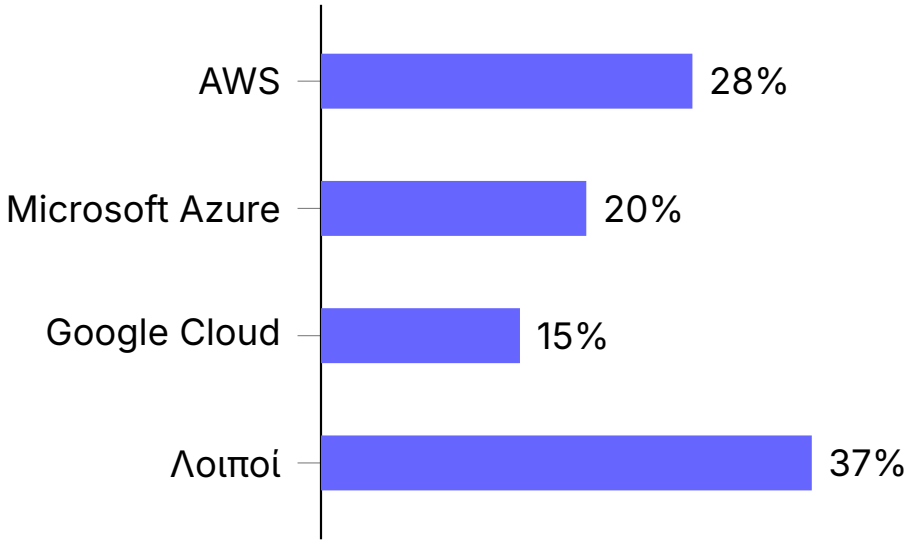


Πηγή: Synergy Research Group (IaaS, PaaS και hosted private cloud)

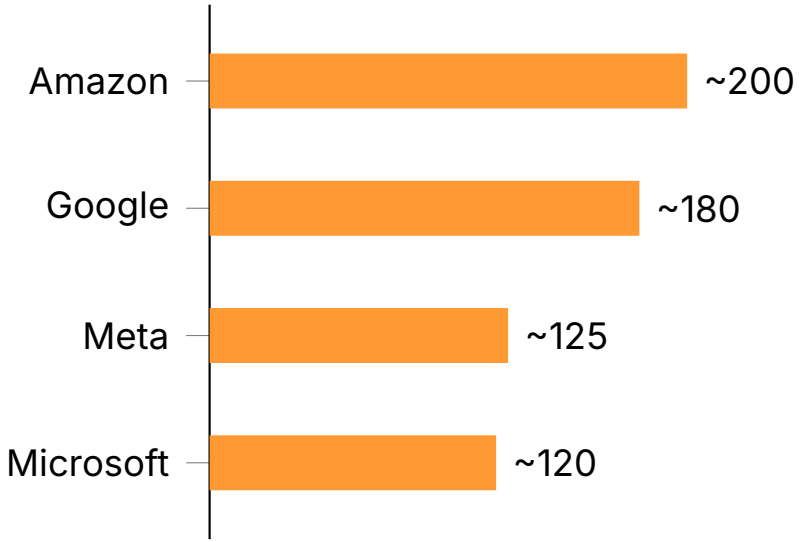
- Η αγορά εξαπλασιάστηκε σε 7 χρόνια: από ~$70 δισ. το 2018 σε $419 δισ. το 2025.
- Από το 2023 η ανάπτυξη επιταχύνεται ξανά λόγω της γενετικής AI (GPU as a service, πλατφόρμες μοντέλων).
- Το 2026 τα τριμηνιαία έσοδα ξεπέρασαν τα $140 δισ. (Q2 2026: $143 δισ.), δηλαδή ετήσιο ρυθμό πάνω από μισό τρισεκατομμύριο δολάρια.

Οι μεγάλοι παίκτες (2026)

Μερίδιο αγοράς cloud υποδομών (Q2 2026)



Capex 2026 (δισ. $, εκτιμήσεις)



- Συνολική αγορά Q2 2026: $143 δισ. (+43% σε ετήσια βάση). Η AWS προηγείται, αλλά Azure και Google Cloud αναπτύσσονται ταχύτερα.
- Στους «λοιπούς» αναπτύσσονται ταχύτατα τα neoclouds (CoreWeave, Nebius, Crusoe) και η Oracle, καθώς και οι OpenAI και Anthropic, που πουλάνε μοντέλα ως υπηρεσία, ενώ τρέχουν κυρίως σε υποδομές hyperscalers και neoclouds.