

TFlop, Tbit, TWatt – The need for a unified strategy in Datacenter, Telecom and Energy

Η ανάγκη υιοθέτησης μιας ενιαίας στρατηγικής στα Datacenters, τις Τηλεπικοινωνίες και την Ενέργεια

Κώστας Μασσέλος

Πρόεδρος Εθνικής Επιτροπής Τηλεπικοινωνιών και Ταχυδρομείων

Αντιπρόεδρος Body of European Regulators for Electronic Communications (BEREC) 2024

Καθηγητής Παν. Πελοποννήσου

Εμείς οι Έλληνες, με την χαρακτηριστική μας, 'ήπια' θα έλεγα ευγενικά, τάση προς την υπερβολή, συνήθως δυσκολευόμαστε να βάλουμε φρένο στην αισιοδοξία μας, αλλά και την απαισιοδοξία μας κατά περιόδους. Συχνά ψάχνουμε να βρούμε ένα 'σχέδιο' πίσω από τα θετικά ή τα αρνητικά, στο οποίο τυπικά κάποιος κατέχει κεντρικά τον ρόλο του 'σχεδιαστή', του 'καλού' ή του 'κακού'.

Ίσως αυτό να αποτελεί την πιο απτή απόδειξη ότι ακόμα και σήμερα, μάλλον λίγο πιστεύουμε στις ελεύθερες αγορές και τη δύναμη των ανθρώπων που συμμετέχουν σε αυτές. Των ανθρώπων που κάνουν μικρές επιλογές, με βάση την αντίληψή τους, εκείνη τη στιγμή, για το ατομικό τους συμφέρον, που όμως αθροίζονται σε κύματα και δημιουργούν μεγάλες Τάσεις, που κινούν την παγκόσμια οικονομία, συμπαρασύροντας και μας, θέλοντας ή μη.

Μια τέτοια Τάση είναι οι ΑΠΕ.

Η Ελλάδα βρίσκεται σήμερα ανάμεσα στις πλέον προηγμένες χώρες σε ότι αφορά στη χρήση των ΑΠΕ στο ηλεκτροπαραγωγικό τους μείγμα:

Το 32% της ηλεκτρικής ενέργειας που καταναλώνουμε στη χώρα μας παράγεται από ηλιακή και αιολική ενέργεια (16TWh), ενώ αν προσθέσουμε και την υδροηλεκτρική (4.73TWh), η ηλεκτροπαραγωγή στην Ελλάδα υποστηρίζεται σε ποσοστό που ξεπερνά το 41% από ανανεώσιμες πηγές. (στοιχεία 2023)

Στην χώρα, η εγκατεστημένη ισχύς (αθροιστικά) φωτοβολταϊκών και αιολικών πάρκων ξεπερνά τα 12GW (2023) ενώ μέχρι το 2030, από ήδη προγραμματισμένες επενδύσεις, η εγκατεστημένη αυτή ισχύς προβλέπεται να ξεπεράσει τα 23-24GW (στοιχεία ΕΣΕΚ – Εθνικό Σχέδιο για την Ενέργεια και το Κλίμα) ή 35GW αν αθροίσουμε και τα μεγάλα υδροηλεκτρικά, όταν η ζήτηση κατά μέσο όρο κινείται στα 6GW και δεν ξεπερνά τα 10GW σε ώρες αιχμής με πτωτικές (από το 2008) τάσεις.

Μέσα στην Τάση των ΑΠΕ, έχουμε την 'υπο-τάση' των φωτοβολταϊκών

Το 2004, χρειάστηκε ένα ολόκληρο έτος για να εγκατασταθεί (ανά τον κόσμο) 1GW ισχύος με ηλιακά πάνελ. Για 1GW το 2010 χρειάστηκε ένας μήνας, το 2016 μία εβδομάδα και το 2023 μια μέρα. Ενώ για το 2024, αναλυτές του Bloomberg αναμένουν ότι 1GW θα εγκαθίσταται κάθε 15 ώρες.

Σύμφωνα με την Διεθνή Υπηρεσία Ενέργειας (International Energy Agency – IEA), η αγορά και η εγκατάσταση ηλιακών πάνελ είναι σήμερα η μεγαλύτερη κατηγορία επενδύσεων στην παραγωγή

ηλεκτρικής ενέργειας. Η Διεθνής Υπηρεσία Ενέργειας (IEA) εκτιμά ότι το τρέχον έτος θα επενδυθούν περισσότερα από \$500 δις δολάρια (στη συγκεκριμένη αγορά), και η εγκατεστημένη ισχύς θα διπλασιάζεται κάθε τρία χρόνια.

Μέχρι το 2040 η (απολύτως περιοδική) παραγωγή ενέργειας από φωτοβολταϊκά πάνελ θα είναι η μεγαλύτερη παραγωγός ενέργειας -όχι μόνο ηλεκτρικής- της ανθρωπότητας (Πηγή IEA)

Στην Ελλάδα, έχοντας ήδη μεγαλύτερη παραγωγή σε σχέση με τη ζήτηση, σκεφτόμαστε την εξαγωγή του 'πράσινου' ενεργειακού πλεονάσμάτος μας. Επειδή όμως α) δεν υπάρχουν ακόμα οι υποδομές για να το κάνουμε και β) οι 'γείτονες' υποψήφιοι πελάτες' επηρεάζονται και αυτοί, όπως εμείς, από τις Τάσεις των ΑΠΕ και των φωτοβολταϊκών (έχουν δηλαδή ήδη ή θα έχουν σύντομα και εκείνοι ενεργειακό πλεόνασμα ΑΠΕ), προς το παρόν αναγκάζομαστε να απορρίπτουμε πράσινη ενέργεια!

Από τις αρχές του έτους, χάθηκαν περισσότερες από 500GWh (πηγή ΑΔΜΗΕ) στην χώρα μας από την αδυναμία του συστήματος να απορροφήσει την πλεονάζουσα παραγωγή, πάντα στις ώρες αιχμής των φωτοβολταϊκών.

Λογικά, θα έλεγε κάποιος, επείγει να διαφοροποιήσουμε το μίγμα μας στις ΑΠΕ, με περισσότερα αιολικά και υδροηλεκτρικά, αντί φωτοβολταϊκών, και να επενδύσουμε σε μπαταρίες για την αποθήκευση ενέργειας.

Αν δεν ήταν τόσο αυστηρό το χρονικό περιθώριο που έχω γι' αυτή την ομιλία, θα μπορούσα να μιλήσω περισσότερο σχετικά με το που κατέληξαν (δλδ στην αποτυχία) όλα τα σχέδια διεθνώς, που προσπάθησαν να αντιμετωπίσουν 'λογικά' τις αγορές – οι οποίες παρεμπιπτόντως σήμερα, στην Ελλάδα, προτείνουν 11.000 νέες επενδύσεις με συνολική ισχύ 48GW (πηγή ΑΔΜΗΕ) σε ανανεώσιμες πηγές ενέργειας.

Ας συνοψίσουμε λοιπόν προς το παρόν εδώ ότι το πρώτο εξάμηνο του '24 απορρίψαμε 500 πράσινες GWh, οι οποίες θα μπορούσαν να τροφοδοτήσουν περισσότερες από 3 φορές, στη μέγιστη κατανάλωσή τους, όλα τα Datacenters της χώρας (6-7 μεγάλα και μικρότερα, συνολικής ισχύος περίπου 35MW) για το σύνολο της ίδιας περιόδου (δηλαδή για 6 ολόκληρους μήνες).

Ας ρίξουμε τώρα μια ματιά σε μια άλλη Τάση, αυτή της «Τεχνητής Νοημοσύνης», που πρακτικά είναι η εξέλιξη των Large Language Models, μιας κατηγορίας νευρωνικών δικτύων (deep neural networks), με όλες τις τελευταίες εξελίξεις από το 2020 και έπειτα ενσωματωμένες. Δηλαδή: Transformers, τεχνικές chain-of-thought, οργάνωση σε agents κ.ο.κ. – ότι δηλαδή δουλεύει πίσω από τις οθόνες μας όταν ο web browser μας δείχνει OpenAI, Midjourney κλπ.

Το ότι αποτελεί Τάση, νομίζω δεν χρειάζεται ιδιαίτερη ανάλυση, αρκεί να δει κάποιος την κερδοφορία της NVIDIA τα τρία τελευταία εξάμηνα ή/και τα χρήματα που επενδύονται στον συγκεκριμένο χώρο.

Εδώ, θα εγκαταλείψουμε προσωρινά τα Watts και θα μιλήσουμε για τα FLOPs (Floating Point Operations) που είναι στην ουσία απλές αριθμητικές πράξεις (προσθέσεις, πολλαπλασιασμοί) στις οποίες το πλήθος των ακεραίων και δεκαδικών ψηφίων που εξετάζεται είναι περιορισμένο (όχι απεριόριστο) προκειμένου να μπορεί να τις εκτελεί με ευκολία ένας υπολογιστής για να 'εκπαιδεύσει' (train) ένα μοντέλο, όπως αυτά της Meta ή της OpenAI, ή να κάνει 'προβλέψεις' (inference) χρησιμοποιώντας το.

Ως υπολογιστικά φορτία, τα μοντέλα αυτά, αν πάρουμε το δημοφιλές ChatGPT ως αναφορά, στην έκδοση 3 ξεκίνησε έχοντας 175 Δισεκατομμύρια παραμέτρους, ενώ η τρέχουσα έκδοση 4 θεωρείται ότι έχει περίπου 1.5-1.8 Τρισεκατομμύρια παραμέτρους. Στην επερχόμενη έκδοση 5, που δεν αποκλείεται να δούμε πριν το τέλος του 2024, πιθανότατα θα ξεπεράσει τις 10 Τρισεκατομμύρια παραμέτρους, με πολλούς να εικάζουν για ένα σημαντικά μεγαλύτερο πλήθος. Βλέπουμε μια τάση δηλαδή το πλήθος των παραμέτρων σε κάθε νέα έκδοση να δεκαπλασιάζεται.

Τί είναι οι παράμετροι και τι μας ενδιαφέρουν;

Οι παράμετροι ενός μοντέλου είναι οι μεταβλητές των οποίων τις τιμές προσδιορίζουμε εκπαιδεύοντάς το ή χρησιμοποιούμε όταν κάνουμε προβλέψεις με αυτό.

Το πλήθος των παραμέτρων μας ενδιαφέρει γιατί με βάση αυτές, μπορούμε να έχουμε μια σχετική εικόνα για το πλήθος των δεδομένων που χρειαζόμαστε για να τις 'εκπαιδεύουμε', ενώ υπάρχει και μια γενικά αποδεκτή από την ακαδημαϊκή κοινότητα σχέση μεταξύ των παραμέτρων, των δεδομένων προς εκπαίδευση και του υπολογιστικού φορτίου που δημιουργείται κατά την διαδικασία εκπαίδευσης του μοντέλου – κάτι που τελικά μας συνδέει το μοντέλο με την ηλεκτρική ενέργεια που χρειάζεται να δαπανήσουμε για να το εκπαιδεύσουμε.

Μπορούμε δηλαδή να πούμε ότι ένα μοντέλο της 'κατηγορίας' του GPT3 θέλει 10^{23} πράξεις, ενώ ένα της 'κατηγορίας' του GPT4 10^{25} πράξεις – αυτές είναι πολλές πράξεις!

Το 1 ακολουθούμενο από 25 μηδενικά – για να ας δώσω έναν άλλο αριθμό να βάλετε δίπλα για αναφορά, η εκτίμηση των αστρονόμων για το πόσα αστέρια υπάρχουν στο σύμπαν έχει 24 μηδενικά.

Βλέπουμε δηλαδή ότι ενώ από 'γενιά' σε 'γενιά' τα μοντέλα έχουν μια τάξη μεγέθους περισσότερες παραμέτρους, έχουν δυο τάξεις μεγέθους μεγαλύτερες ανάγκες σε υπολογιστική ισχύ.

Τι γίνεται με τους επεξεργαστές κάθε τύπου (CPUs, GPUs, ASICs, FPGAs), που χρησιμοποιούμε για να καλύψουμε τις υπολογιστικές αυτές ανάγκες; Ας πάρουμε εδώ για αναφορά τις GPUs της εταιρείας NVIDIA.

Εκεί, μελετώντας τους επεξεργαστές που κυκλοφόρησε η συγκεκριμένη εταιρεία τα τελευταία 6 χρόνια, βλέπουμε ότι η υπολογιστική ικανότητα διπλασιάζεται κάθε δυο χρόνια ή, αν θέλουμε να συμπεριλάβουμε και της μειωμένης ακρίβειας πράξεις (FP8, FP4), εξαπλασιάζεται κάθε δυο χρόνια. Σε καμία περίπτωση βέβαια δεν δεκαπλασιάζεται κάθε χρόνο, άρα έχουμε ήδη μια σημαντική υστέρηση της προσφοράς σε σχέση με τη ζήτηση.

Η κατανάλωση; Εκεί τα πράγματα χρειάζονται δυο αναγνώσεις:

Η μια ανάγνωση λέει ότι η κατανάλωση ανά πράξη μειώνεται με ένα ρυθμό που μας δίνει περίπου 3-4x το πλήθος των πράξεων ανά W κατανάλωσης τα τελευταία 6 χρόνια – βλέπετε και πάλι σημαντική υστέρηση σε σχέση με την ταχύτητα που μεγαλώνουν οι ανάγκες, 10x ή 100x –

Οι πράξεις ανά Watt έχουν ενδιαφέρον όταν εκπαιδεύεις ένα μοντέλο άπαξ. Μας λένε ότι αν ήθελες να εκπαιδεύσεις το ChatGPT 4 το 2016 θα χρειαζόσουν 1000GWh ενέργειας (γί' αυτό και πρακτικά δεν μπορούσες να το εκπαιδεύσεις για όσους αναρωτιούνται γιατί συμβαίνουν τώρα αυτά που συμβαίνουν), ενώ το 2022 ήθελες 'μόλις' 13GWh. Τι γίνεται όμως όταν έχεις συχνά και πολλά μοντέλα να εκπαιδεύσεις ή αξιοποιείς τέτοια μοντέλα για να κάνεις προβλέψεις συνεχώς;

Εδώ χρειαζόμαστε τη δεύτερη ανάγνωση που αναφέραμε προηγουμένως.

Η δεύτερη αυτή ανάγνωση λέει, ότι ένα state-of-the-art AI Datacenter έχει 100.000 ειδικούς 'ΑΙ' (GPU based) servers, που ο καθένας καταναλώνει, ας πούμε προσεγγιστικά, 1.4kW, άρα στο σύνολο 140MW. Αν αυτό το Data Center είναι πάντα απασχολημένο εκπαιδεύοντας μοντέλα και κάνοντας προβλέψεις, ναι, μπορεί να εκπαιδεύσει συνεχώς καλύτερα και μεγαλύτερα μοντέλα, αλλά στο τέλος της ημέρας έχουμε μια κατανάλωση 600GWh το εξάμηνο – κλείνοντας το κύκλο: περίπου όση απορρίψαμε στην Ελλάδα το πρώτο εξάμηνο του 2024!

Μήπως η Ελλάδα, σε αυτή τη νέα πραγματικότητα που φαίνεται να καταφθάνει, μπορεί να συμμετέχει απλά αξιοποιώντας τις ΑΠΕ που σε κανονικές συνθήκες θα απέρριπτε, μαζί ίσως με μια στρατηγική ευελιξία στο πως σχεδιάζει την υποδομή της;

Μήπως αντί να ανησυχούμε για την 'υπερθέρμανση' της αγοράς στις ΑΠΕ, αποθαρρύνοντας νέες επενδύσεις, πρέπει να επιταχύνουμε ακόμα περισσότερο, ξεπερνώντας κατά πολύ τον ευρωπαϊκό μέσο όρο και εδραιώνοντας μια δεύτερη πρωτοπορία (μετά την ναυτιλία) για την χώρα μας;

Μήπως τα διεθνή καλώδια οπτικών ινών που όλο και πιο συχνά διασχίζουν τα σύνορά μας, δεν θα πρέπει να κάνουν 'απλή διέλευση' αλλά θα πρέπει και να τροφοδοτούνται με δεδομένα εδώ;

Μήπως επιπλέον η Ελλάδα σε αυτή τη διαδικασία θα μπορούσε να αξιοποιήσει και το 'brand' της; Ενός ξεκάθαρα Ευρωπαϊκά προσανατολισμένου κράτους, που όμως δεν αντιμετωπίζει δογματικά την Ασία και έχει (πάλι από το πεδίο της ναυτιλίας) ισχυρές συνεργασίες δεκαετιών εκεί;

Μήπως αυτό θα επέτρεπε στην Ελλάδα να τοποθετηθεί πιο ανταγωνιστικά, σε σχέση με άλλες χώρες, στο κόστος ανάπτυξης τέτοιων data centers; Άλλωστε η εκπαίδευση και η χρήση ΑΙ μοντέλων πραγματοποιείται σε air-tight υπολογιστικά περιβάλλοντα, που διευκολύνουν τον έλεγχο της ασφάλειας της διαδικασίας.

Σύμφωνα με τις εκτιμήσεις των Bloomberg, Goldman Sachs και IDC, το 2025 θα επενδυθούν περισσότερα από 200\$ δις σε ΑΙ υποδομές μόνο από τις 7 μεγαλύτερες εταιρείες τεχνολογίας των ΗΠΑ. Περίπου όσο το Ελληνικό ΑΕΠ. Δηλαδή με κάθε ποσοστιαία μονάδα από αυτές τις επενδύσεις που θα καταφέρουμε να προσελκύσουμε στην Ελλάδα, θα ανεβάζουμε αντίστοιχα κατά μια σχεδόν μονάδα και το Ακαθάριστο Εθνικό μας Προϊόν.

Μήπως χρειαζόμαστε κοινές στρατηγικές στο χώρο των data centers και της ενέργειας;

..και ναι, κλείνοντας, προφανώς ναι, όλα τα προηγούμενα έχουν και ένα τηλεπικοινωνιακό στοιχείο.

Χρειαζόμαστε δίκτυα για να διακινούμε πολλά δεδομένα.

Στο πλαίσιο που εξετάζουμε το τηλεπικοινωνιακό στοιχείο όμως, ο όγκος, τα Tbits που απαιτούνται, είναι διαχειριζόμενος. Η πρόκληση είναι η ασφάλεια του δικτύου. Όταν υλοποιείς μια τέτοια στρατηγική, η ασφάλεια των υποβρυχίων και επίγειων καλωδίων σου γίνεται πρόκληση και για αυτό, χρειαζόμαστε

μια ξεχωριστή κουβέντα. Ας εστιάσουμε πρώτα όμως στο να 'δούμε' την ευκαιρία και να τοποθετηθούμε στρατηγικά απέναντι σε αυτή:

Μπορούμε να φανταστούμε μια Ελλάδα που 09:00 με 16:00 (στη ζώνη GMT/Greenwich Mean time +2/+3) καταναλώνει TWatt πράσινης ενέργειας και εξάγει Tbits δεδομένων; Η γεωγραφίας μας, για άλλη μια φορά, μας ευνοεί.