Епоха штучного інтелекту: AI-чипи до 2034 року

Переклад та редагування: Юлія Ямненко, д.т.н., професор, кафедра ЕПС, ФЕЛ, КПІ ім. Ігоря Сікорського

Штучний інтелект (ШІ) змінює світ, яким ми його знаємо: від успіху DeepMind над чемпіоном світу з гри в го Лі Седолом у 2016 році до надійних можливостей прогнозування ChatGPT від OpenAI, складність алгоритмів навчання ШІ зростає надзвичайно швидкими темпами, а обсяг обчислень, необхідних для запуску нових алгоритмів навчання, подвоюється приблизно кожні чотири місяці. Щоб іти в ногу з часом, апаратне забезпечення для додатків ШІ повинно бути не лише масштабованим — забезпечувати довговічність у міру появи нових алгоритмів, зберігаючи при цьому низькі операційні витрати — але й бути здатним обробляти все складніші моделі.

Компанія IDTechEx, спираючись на технологічні звіти та ринкові прогнози «AI Chips: 2023-2033» і «AI Chips for Edge Applications 2024-2034: Artificial Intelligence at the Edge», прогнозує, що зростання використання штучного інтелекту, як для навчання і обробки даних як в хмарі, так і на периферії, продовжиться протягом наступних десяти років, оскільки наш світ і пристрої, які його наповнюють, стають все більш автоматизованими і взаємопов'язаними.

НАВІЩО І ДЛЯ ЧОГО ПОТРІБНІ ЧИПИ ЗІ ШТУЧНИМ ІНТЕЛЕКТОМ

Ідея розробки апаратного забезпечення для виконання певної функції, особливо якщо ця функція полягає у прискоренні певних типів обчислень, перебираючи контроль над ними від основного (хост) процесора, не нова: на початку розвитку обчислювальної техніки з'явилися центральні процесори ЦП (*Central Pocessing Unit, CPU*) в парі з математичними співпроцесорами, відомими як процесори з плаваючою комою (*Floating-Point Units, FPU*). Мета полягала в тому, щоб перекласти складні математичні операції з плаваючою комою з центрального процесора на цей спеціалізований чип, оскільки останній міг би виконувати обчислення більш ефективно, тим самим звільняючи центральний процесор, щоб той міг зосередитися на інших речах.

З розвитком ринків і технологій зростали й робочі навантаження, а отже, потрібні були нові апаратні засоби, які б могли впоратися з цими навантаженнями. Особливо примітним прикладом одного з таких спеціалізованих робочих навантажень є створення комп'ютерної графіки, де прискорювач, про який йде мова, став чимось на кшталт прозивного (загального) імені: графічний процесор (*Graphics Processing Unit, GPU*).

Так само, як комп'ютерна графіка потребувала іншого типу архітектури чипів, поява машинного навчання викликала попит на інший тип прискорювачів, здатних ефективно справлятися з робочими навантаженнями машинного навчання. Машинне навчання (*Machine learning, ML*) — це процес, за допомогою якого комп'ютерні програми використовують дані для прогнозування на основі моделі, а потім оптимізують модель для кращого узгодження з наданими даними шляхом коригування вагових коефіцієнтів. Обчислення, таким чином, включає два етапи: Навчання та Формування Висновків.

Першим етапом реалізації алгоритму штучного інтелекту є етап Навчання, на якому дані подаються в модель, і модель коригує свої вагові коефіцієнти до тих пір, поки вони не будуть належним чином відповідати наданим даним. Другий етап — етап Формування Висновків, на якому виконується навчений алгоритм ШІ, і нові дані (не надані на етапі Навчання) класифікуються у спосіб, що відповідає отриманим даним.

З цих двох етапів етап Навчання є більш інтенсивним з точки зору обчислень, оскільки він передбачає виконання одних і тих самих обчислень мільйони разів (навчання деяких провідних алгоритмів ШІ може тривати кілька днів). Тому етап Навчання відбувається в хмарних обчислювальних середовищах (тобто в датацентрах), де використовується велика кількість чипів, які можуть виконувати паралельну обробку, необхідну для ефективного навчання алгоритмів (процесори обробляють завдання в послідовному режимі, коли один цикл виконання починається після завершення попереднього). Для того, щоб мінімізувати затримку, використовуються великі та численні кеші пам'яті, так що більша частина часу роботи циклів виконання присвячується обробці. Для порівняння, паралельна обробка передбачає одночасне виконання декількох обчислень, де легкі потоки обчислень перекриваються таким чином, що затримка ефективно маскується. Можливість розділення і одночасного виконання декількох обчислень є основною перевагою для навчання алгоритмів ШІ. На відміну від цього, етап Формування Висновків може відбуватися як у хмарних, так і в периферійних обчислювальних середовищах. У вищезгаданих доповідях детально описані відмінності між архітектурами CPU, GPU, FPGA (*Field Programmable Gate Array*) і ASIC (*Application-Specific Integrated Circuit*), а також їх відносна ефективність в обробці робочих навантажень машинного навчання.

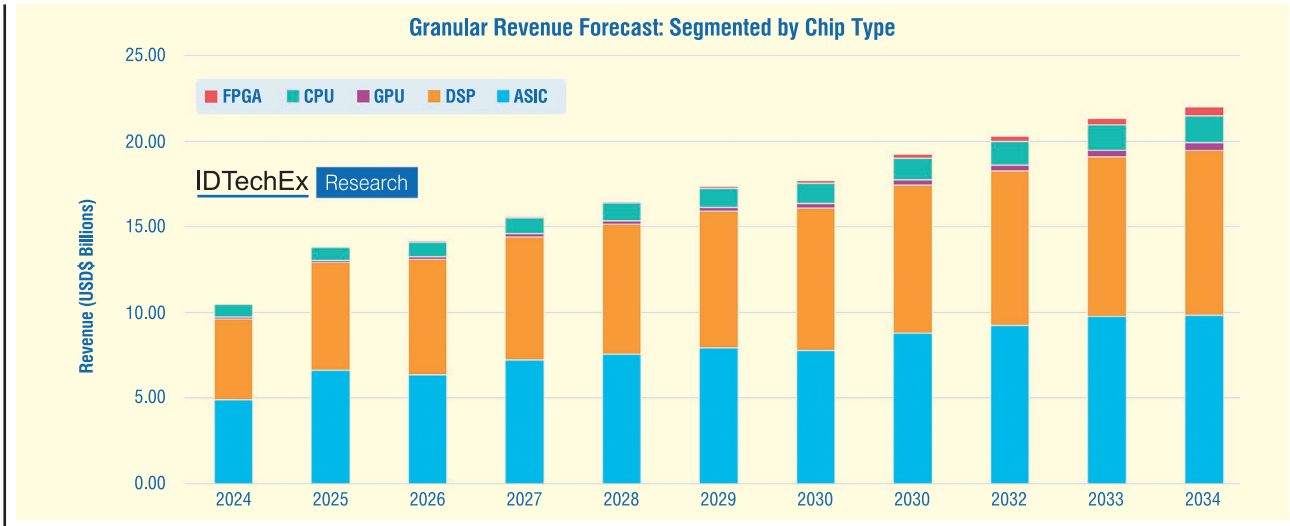

Рис. 1. Доходи, що генеруються різними архітектурами чипів для периферійних пристроїв, 2024-2034 роки. Джерело: «AI Chips for Edge Applications 2024–2034: Artificial Intelligence at the Edge», IDTechEx

У середовищі хмарних обчислень наразі домінують графічні процесори, і, за прогнозами, це збережеться протягом наступного десятиріччя, враховуючи домінування Nvidia у сфері навчання ШІ. Для ШІ на периферії (*AI at the edge*) перевага надається ASIC, оскільки мікросхеми частіше розробляються з урахуванням конкретних завдань (наприклад, для виявлення об'єктів у системах камер спостереження). Як показано на рисунку 1, на цифрові сигнальні процесори (*Digital Signal Processors, DSP*) також припадає значна частка спільної обробки ШІ на периферії, хоча слід зазначити, що ця велика кількість в першу чергу пов'язана з тим, що процесор HTP (Hexagon Tensor Processor) від Qualcomm (який можна знайти в їхніх сучасних продуктах Snapdragon) також є DSP. Якщо Qualcomm змінить дизайн HTP таким чином, що він перестане бути DSP, то прогноз сильно зміститься на користь ASIC.

ШІ ЯК ДРАЙВЕР ДЛЯ ВИРОБНИЦТВА НАПІВПРОВІДНИКІВ

Мікросхеми для навчання штучного інтелекту зазвичай виробляються з використанням найсучасніших технічних процесів (під технічним процесом мається на увазі транзисторна технологія, що використовується у виробництві напівпровідникових мікросхем), зважаючи на те, наскільки обчислювально інтенсивним є етап Навчання для реалізації алгоритму ШІ. Компанії Intel, Samsung і TSMC — це єдині компанії, які можуть виробляти мікросхеми за 5-нм технологією. З них TSMC просунулася найдалі, освоївши 3-нм технологію виготовлення мікросхем. Частка TSMC на світовому ринку виробництва напівпровідників наразі коливається на рівні 60%. Для більш просунутих технологій цей показник наближається до 90%. З шести 12-дюймових і шести 8-дюймових заводів TSMC лише два знаходяться в Китаї, а один — у США. Решта — на Тайвані. Таким чином, виробництво напівпровідників у глобальному ланцюжку постачання значною мірою зосереджене в регіоні ATP (Азійсько-Тихоокеанський регіон), головним чином на Тайвані.

Така концентрація пов'язана з великим ризиком, якщо ця частина ланцюжка постачання опиниться під загрозою. Саме це і сталося у 2020 році, коли низка взаємодоповнюючих факторів (про які докладніше йдеться у звіті «AI Chips: 2023-2033») призвела до глобального дефіциту мікросхем. Відтоді найбільші учасники світового ринку напівпровідників (США, ЄС, Південна Корея, Японія та Китай, за винятком Тайваню) намагаються зменшити свою залежність від дефіциту чипів (якщо, раптом, знову виникнуть якісь обставини, які можуть призвести до його появи). Про це свідчить збільшене державне фінансування, оголошене всіма основними зацікавленими в цьому сторонами у зв'язку з загрозою глобального дефіциту мікросхем (рис. 2).

Ініціативи урядів цих країн спрямовані на залучення додаткових приватних інвестицій шляхом надання податкових пільг та часткового фінансування у вигляді грантів і кредитів. Хоча багато з приватних інвестицій (рис. 3) були здійснені ще до оголошення таких урядових ініціатив, інші додаткові та/або нові приватні інвестиції були оголошені вже пізніше, бо були заохочені запропонованими стимулами.

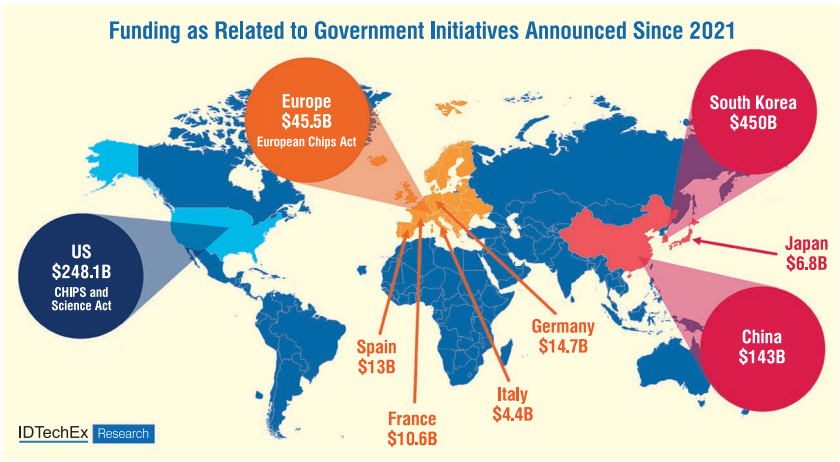

Рис. 2. Фінансування в рамках урядових ініціатив, оголошених з 2021 року. Джерело: IDTechEx

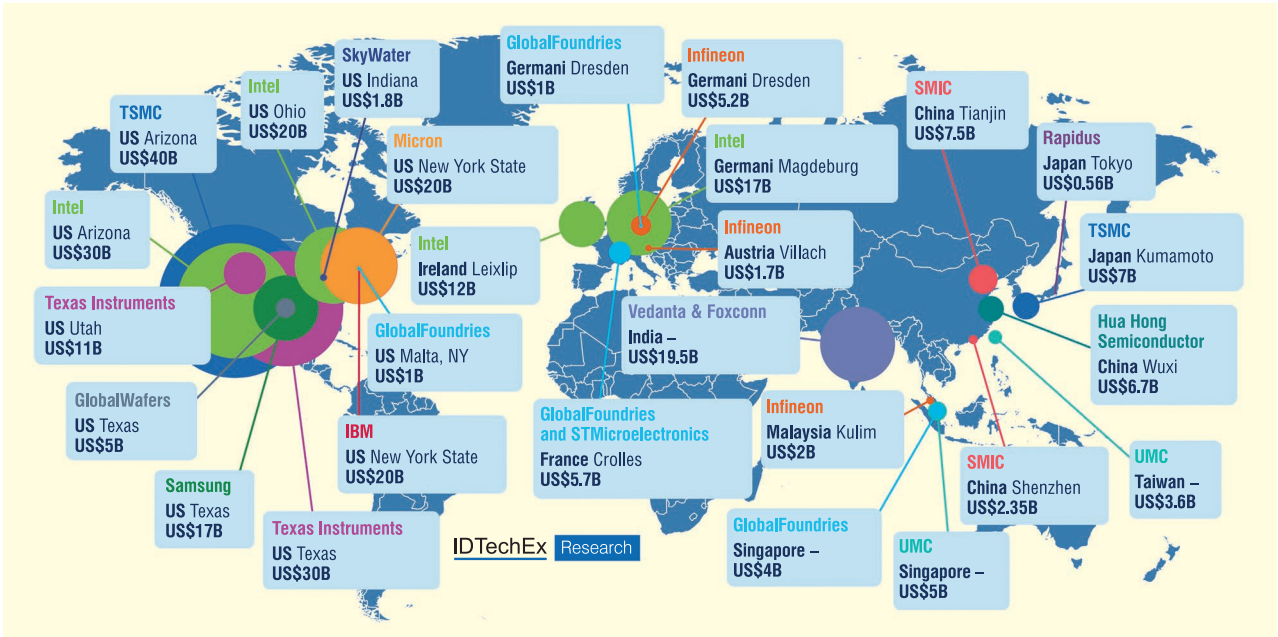

Рис. 3. Запропоновані та підтверджені інвестиції в напівпровідникові потужності від виробників з 2021 року. Там, де валюта вказана не в доларах США, вона була перерахована в долари США станом на травень 2023 року. Джерело: « AI Chips: 2023-2033», IDTechEx

Основною причиною цих урядових ініціатив і додаткових приватних витрат є потенціал реалізації передових технологій, до яких можна віднести і штучний інтелект. Виробництво сучасних напівпровідників сприяє розвитку національних/регіональних можливостей ШІ, де можливість автономного виявлення і аналізу об'єктів, зображень і мови настільки важлива для ефективності певних продуктів (наприклад, автономних транспортних засобів і промислових роботів), а також для моделей національного управління і безпеки, що розробка апаратного і програмного забезпечення для ШІ стала першочерговим завданням для державних органів, які хочуть бути в авангарді технологічних інновацій і впровадження перспективних технологій.

ЗРОСТАННЯ РИНКУ ЧИПІВ ШІ В НАСТУПНОМУ ДЕСЯТИЛІТТІ

Очікується, що до 2034 року дохід від продажу чипів ШІ (включаючи продаж фізичних чипів і оренду чипів через хмарні сервіси) зросте щонайменше до 300 мільярдів доларів США при середньорічному темпі зростання у 22% за період з 2024 по 2034 роки (рис. 4а). Ця цифра включає використання чипів для прискорення машинного навчання на периферії мережі, для телекомунікаційної периферії та в хмарних центрах обробки даних. Станом на 2024 рік на чипи для реалізації другого етапу машинного навчання — Формування Висновків (як на периферії, так і в хмарі) припадає 63% отриманого доходу, а до 2034 року ця частка зросте до більш як двох третин від загального доходу (рис. 4б).

Це значною мірою пов'язано зі значним зростанням використання ШІ на периферії та в телекомунікаційному секторі, оскільки можливості ШІ використовуються ближче до кінцевого споживача. З точки зору різних галузей, очікується, що IT та телекомунікації будуть лідерами у використанні мікросхем ШІ протягом наступного десятиліття.

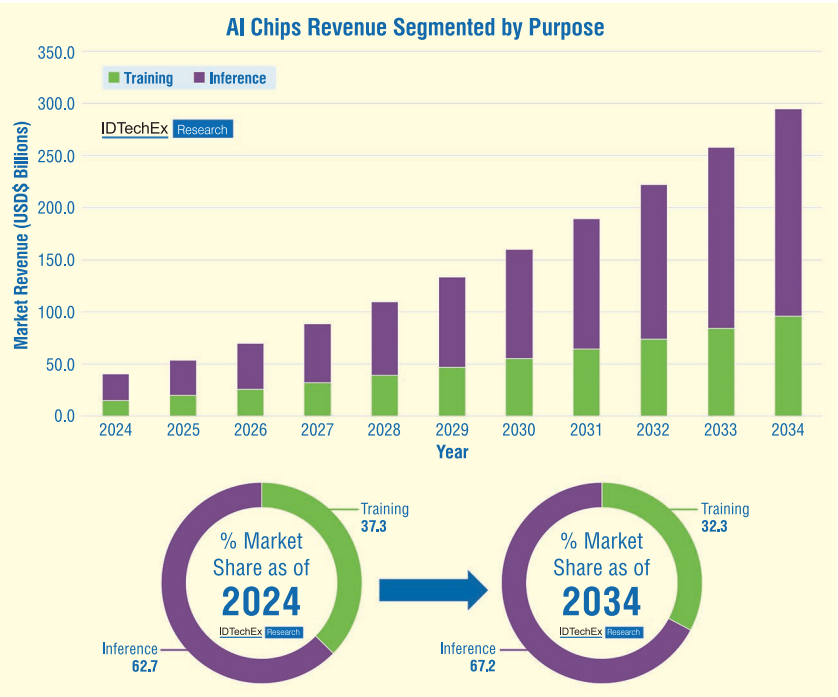

Рис. 4. Доходи, що генеруються чипами ШІ, зростатимуть на 22% в середньорічному обчисленні протягом наступних десяти років, до 2034 р. У цей час дохід, отриманий від чипів штучного інтелекту, буде домінувати над доходом від навчання ШІ, оскільки ШІ все активніше мігрує до розгортання на периферії мережі. Джерело: IDTechEx