
เอไอ และ
เซมิคอนดักเตอร์
มุ่งสู่ เอไอชิป
TSMC
TSMC
AI
Processing
Unit
Samsung
AI

เอไอ และ เซมิคอนดักเตอร์
มุ่งสู่ เอไอชิป

AI and Semiconductor
towards AI Chip

คำนำ

โลกในยุคปัจจุบันกำลังขับเคลื่อนด้วยคลื่นลูกใหญ่สองลูกที่หมุนมาบรรจบกันอย่างลงตัว ลูกแรกคือ ปัญญาประดิษฐ์ (Artificial Intelligence, AI) ซอฟต์แวร์อัจฉริยะที่เข้ามาปฏิวัติวิถีชีวิตและการทำงานของมนุษย์ และลูกที่สองคือ เทคโนโลยีเซมิคอนดักเตอร์ (Semiconductors) ฮาร์ดแวร์อันเป็นรากฐานและฟันเฟืองชิ้นสำคัญของระบบคอมพิวเตอร์โลก เมื่อเอไอเติบโตอย่างก้าวกระโดด ความต้องการพลังในการประมวลผลจึงมหาศาลเกินกว่าที่สถาปัตยกรรมชิปแบบดั้งเดิมจะรับมือได้ จุดตัดนี้เองที่นำพาเราไปสู่ยุคสมัยใหม่ นั่นคือยุคของ เอไอชิป (AI Chip)

แรงบันดาลใจและจุดเริ่มต้นในการเขียนหนังสือเล่มนี้ เกิดขึ้นจากการที่ผู้เขียนได้มีโอกาสเข้าร่วมงานประชุมระดับนานาชาติ SNU AI Semiconductor Forum (SAISF 2026): International Collaboration of Research, Education, and Strategy in Southeast and East Asia ซึ่งจัดขึ้นระหว่างวันที่ 1–2 มิถุนายน พ.ศ. 2569 ณ มหาวิทยาลัยแห่งชาติโซล (Seoul National University) ประเทศสาธารณรัฐเกาหลี งานฟอรัมครั้งนี้จัดขึ้นโดย Graduate School of AI Semiconductor ซึ่งเป็นหนึ่งในสถาบันชั้นนำระดับโลกที่มุ่งเน้นการวิจัยและพัฒนาบุคลากรด้านชิปปัญญาประดิษฐ์โดยเฉพาะ บรรยากาศในงานประชุม SAISF 2026 เต็มไปด้วยการแลกเปลี่ยนวิสัยทัศน์ นวัตกรรม งานวิจัยล่าสุด รวมถึงยุทธศาสตร์ระดับชาติของประเทศต่างๆ ที่เข้าร่วม (คือ เกาหลีใต้ ไทย เวียดนาม และ มาเลเซีย) การได้ร่วมหารือกับเหล่านักวิจัยและผู้เชี่ยวชาญชั้นนำ ทำให้ผู้เขียนเห็นภาพชัดเจนว่าอุตสาหกรรมเทคโนโลยีกำลังเปลี่ยนผ่านไปสู่การผสานรวมทางยุทธศาสตร์อย่างลึกซึ้งระหว่างองค์ความรู้ฝั่งวิทยาการคอมพิวเตอร์ (Computer Science) และวิศวกรรมเซมิคอนดักเตอร์ (Semiconductor Engineering) การพัฒนาทั้งสองส่วนนี้ไม่สามารถแยกออกจากกันได้อีกต่อไป หากต้องการให้เทคโนโลยีเอไอบรรลุขีดสุด

ความสามารถ ด้วยเหตุนี้ ผู้เขียนจึงตั้งใจถ่ายทอดความรู้ มุมมอง และรากฐานที่สำคัญออกมาเป็นหนังสือ "เอไอ และ เซมิคอนดักเตอร์ มุ่งสู่ เอไอชิป" เล่มนี้ โดยแบ่งเนื้อหาออกเป็น 3 บทหลัก คือ **บทที่ 1 เอไอ (ปัญญาประดิษฐ์)** ปูพื้นฐานความเข้าใจในฝั่งซอฟต์แวร์และอัลกอริทึม ตั้งแต่ยุคเริ่มต้นจนถึงยุคการเรียนรู้เชิงลึก (Deep Learning) และ ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) **บทที่ 2 พื้นฐานเซมิคอนดักเตอร์ (Semiconductor Foundation)** ที่กล่าวถึงเทคโนโลยีสารกึ่งตัวนำ และห่วงโซ่อุปทานของอุตสาหกรรมสารกึ่งตัวนำ รวมถึงกระบวนการผลิตชิปในระดับนาโนเมตรที่เป็นกระดูกสันหลังของเทคโนโลยีโลก และ **บทที่ 3 มุ่งสู่เอไอชิป (Towards AI Chip)** เป็นจุดเชื่อมโยงที่นำซอฟต์แวร์เอไอมาหลอมรวมกับฮาร์ดแวร์ที่มนุษย์สามารถสร้างขึ้นได้ การวิเคราะห์สถาปัตยกรรมของชิปยุคใหม่ เช่น จีพียู (GPU) ทีพียู (TPU) เอ็นพียู (NPU) และส่องอนาคตของเทคโนโลยีที่จะขับเคลื่อนสมองกลอัจฉริยะต่อไป

ผู้เรียบเรียงหวังเป็นอย่างยิ่งว่า หนังสือเล่มนี้จะเป็นสะพานเชื่อมองค์ความรู้ และสร้างแรงบันดาลใจให้กับนิสิต นักศึกษา นักวิจัย ตลอดจนผู้ที่สนใจในอุตสาหกรรมเทคโนโลยี ได้มองเห็นโอกาสและการเตรียมความพร้อมรับมือกับคลื่นความเปลี่ยนแปลงครั้งใหญ่ที่กำลังเกิดขึ้นในโลกของเอไอและเซมิคอนดักเตอร์ในปัจจุบันและอนาคต

สุวิทย์ กิระวิทยา

6 มิถุนายน 2569

สารบาญ

1. เอไอ (ปัญญาประดิษฐ์)
AI (Artificial Intelligence)

1.1 ประวัติและวิวัฒนาการของเอไอ (History of AI)

การเดินทางของปัญญาประดิษฐ์ไม่ได้เริ่มต้นขึ้นในห้องปฏิบัติการคอมพิวเตอร์ยุคใหม่ แต่ร่องรอยของมันถูกถักทออยู่ในความนึกคิดของมนุษย์มานานนับพันปี ตั้งแต่เรื่องเล่าปรัมปรา ความปรารถนาที่จะสร้างสิ่งมีชีวิตจำลอง จนกระทั่งหลอมรวมเข้ากับหลักการทางคณิตศาสตร์ ฟิสิกส์ และวิศวกรรมคอมพิวเตอร์ในปัจจุบัน วิวัฒนาการของเอไอสามารถแบ่งออกเป็นยุคสมัยที่สำคัญได้ดังนี้

1. ยุคก่อนประวัติศาสตร์คอมพิวเตอร์: รากฐานทางตรรกะและหุ่นยนต์กลไก

มนุษยชาติต่างมีความฝันเกี่ยวกับ “สิ่งประดิษฐ์ที่มีชีวิตจิตใจ” ปรากฏให้เห็นในตำนานกรีกโบราณ เช่น ยักษ์ทาลอส (Talos) หรือหุ่นยนต์ทองคำของเทพเฮเฟสตัส ต่อมาในคริสต์ศตวรรษที่ 17–19 นักปรัชญาและนักคณิตศาสตร์ เช่น ก๊อตฟรีด ไลบ์นิซ (Gottfried Leibniz) โธมัส ฮอบส์ (Thomas Hobbes) และเรเน เดส์การ์ตส์ (René Descartes) ได้เริ่มตั้งสมมติฐานว่า กระบวนการคิดของมนุษย์สามารถจัดระบบและคำนวณออกมาเป็นระบบเชิงกลได้ โดยฮอบส์ได้บันทึกไว้อย่างคมคายในหนังสือลิไวอาทาน (Leviathan) ว่า “การใช้เหตุผล... ก็คือการคำนวณ นั่นคือการบวกและการลบ” รากฐานนี้ถูกพิสูจน์ให้เป็นรูปธรรมมากขึ้นเมื่อ จอร์จ บูล (George Boole) เสนอตรรกะบูลีน (Boolean Logic) ในปี ค.ศ. 1854 และตามมาด้วยเครื่องจักรคำนวณทางทฤษฎีของ อลัน ทิวริง (Turing Machine) ในปี ค.ศ. 1936 ซึ่งพิสูจน์ว่า อุปกรณ์เชิงกลที่ประมวลผลสัญลักษณ์ง่ายๆ อย่าง ‘0’ และ ‘1’ สามารถจำลองกระบวนการคำนวณทางคณิตศาสตร์ใด ๆ ก็ได้บนโลกนี้

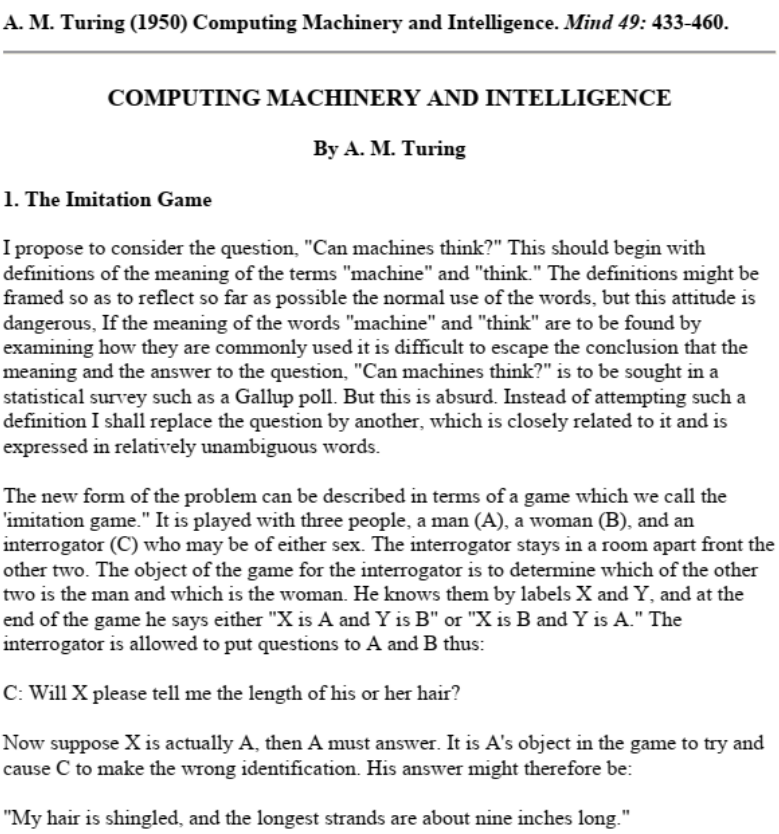
รูปที่ 1 ภาพวาดของก๊อตฟรีด ไลบ์นิซ นักปรัชญาผู้คิดค้นระบบเลขฐานสอง และคาดการณ์ว่าความคิดมนุษย์ลดรูปเป็นเครื่องจักรคำนวณได้
(ที่มา https://en.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz)

2. ยุคกำเนิดปัญญาประดิษฐ์และการทดสอบทิวริง (ค.ศ. 1941–1956)

ในช่วงสงครามโลกครั้งที่ 2 คอมพิวเตอร์ดิจิทัลเครื่องแรกๆ ของโลกได้ถูกสร้างขึ้นทำให้นักวิทยาศาสตร์เริ่มถกเถียงกันอย่างจริงจังเกี่ยวกับโอกาสในการสร้าง “สมองอิเล็กทรอนิกส์” ในปี ค.ศ. 1950 อลัน ทิวริง (Alan Turing) ได้ตีพิมพ์ผลงานวิชาการระดับประวัติศาสตร์ชื่อ “Computing Machinery and Intelligence” ที่อาจแปลว่า เครื่องจักรคำนวณและสติปัญญา ทิวริงตระหนักดีว่าคำว่า “ความคิด” นั้นนิยามยากเกินไป เขาจึงเสนอ “การทดสอบทิวริง” (Turing Test) หากเครื่องจักรสามารถสนทนาโต้ตอบผ่านตัวอักษรกับมนุษย์ แล้วทำให้มนุษย์คู่สนทนาแยกไม่ออกว่ากำลังคุยกับคนหรือคอมพิวเตอร์ ก็ถือได้ว่าเครื่องจักรนั้นมีสติปัญญา

คำว่า ปัญญาประดิษฐ์ (Artificial Intelligence, AI) หรือเอไอ ถูกนิยามและใช้เป็นทางการครั้งแรกในฐานะสาขาวิชาการใหม่ในการประชุมเวิร์กชอปที่วิทยาลัยดาร์ทมัธ (Dartmouth College) ประเทศสหรัฐอเมริกา ในปี ค.ศ. 1956 นำโดย

บุคคลสำคัญอย่าง จอห์น แม็กคาร์ธี (John McCarthy) มาร์วิน มินสกี (Marvin Minsky) อัลเลน นิวเวลล์ (Allen Newell) และ เฮอร์เบิร์ต ไซมอน (Herbert Simon) ซึ่งพวกเขามองโลกในแง่ดีอย่างยิ่งว่า เครื่องจักรที่มีความฉลาดเท่ามนุษย์จะเกิดขึ้นภายในหนึ่งชั่วอายุคน



A. M. Turing (1950) Computing Machinery and Intelligence. ***Mind 49:*** **433-460.**

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

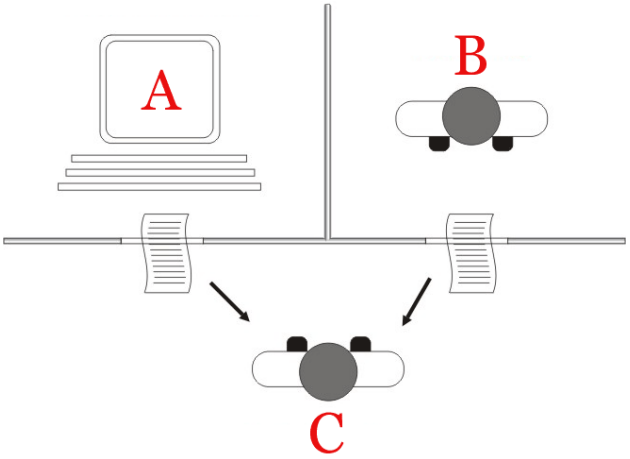
The new form of the problem can be described in terms of a game which we call the 'imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:

"My hair is shingled, and the longest strands are about nine inches long."

รูปที่ 2 ภาพส่วนต้นของสำเนาบทความ โดย อลัน ทิวริง และเผยแพร่ในปี ค.ศ. 1950



รูปที่ 3 ภาพลักษณะการทดสอบทิวริง

(ที่มา https://en.wikipedia.org/wiki/Turing_test)

3. ยุคทองครั้งแรก สู่ ฤดูหนาวของเอไอ (ค.ศ. 1956–1980)

หลังจากการประชุมที่ดาร์ทมัธ รัฐบาลสหรัฐอเมริกา ได้ทุ่มเงินทุนมหาศาลให้กับงานวิจัยเอไอ คอมพิวเตอร์ในยุคนั้นสามารถแก้โจทย์พีชคณิต พิสูจน์ทฤษฎีบทเรขาคณิต และหัดพูดภาษาอังกฤษได้ รวมถึงการกำเนิดของโครงข่ายประสาทเทียมยุคแรก ที่เรียกว่า เพอร์เซปตรอน (Perceptron)

อย่างไรก็ตาม นักวิจัยประเมินความยากของการพัฒนาเอไอต่ำเกินไป เมื่อคอมพิวเตอร์ต้องเจอกับปัญหาในโลกความเป็นจริงที่มีความซับซ้อน มหาศาล และข้อมูลจำกัด คอมพิวเตอร์ยุคนั้นที่มีหน่วยความจำและพลังประมวลผลต่ำมากไม่สามารถทำงานได้ ในปี ค.ศ. 1974 รัฐบาลสหรัฐฯ และอังกฤษจึงสั่งระงับทุนวิจัยทั้งหมด นำไปสู่สภาวะท้อแท้และขาดเงินทุนที่เรียกว่า “ฤดูหนาวของเอไอครั้งที่ 1” (First AI Winter: ค.ศ. 1974–1980)

4. ยุคระบบผู้เชี่ยวชาญ และฤดูหนาวครั้งที่สอง (ค.ศ. 1980–1990s)

ในทศวรรษ ค.ศ. 1980 เอไอกลับมาบูมอีกครั้งด้วยแนวคิด “ระบบผู้เชี่ยวชาญ” (Expert Systems) ซึ่งเน้นการป้อนฐานข้อมูลความรู้และเงื่อนไข (If-Then Rules) ของผู้เชี่ยวชาญเฉพาะทางลงไปในคอมพิวเตอร์ ธุรกิจขนาดใหญ่หันมาใช้ระบบนี้จนเกิดอุตสาหกรรมมูลค่าพันล้านดอลลาร์ แต่ในที่สุด ระบบนี้ก็เจอกับทางตันเนื่องจากบำรุงรักษาได้ยาก อัปเดตข้อมูลลำบาก และขาดความยืดหยุ่น ตลาดหุ้นและนักลงทุนผิดหวังอีกครั้งจนเกิด “ฤดูหนาวของเอไอครั้งที่ 2” (Second AI Winter) ในช่วงปลายทศวรรษ ค.ศ. 1980 ถึง ค.ศ. 1990

5. ยุคเบื้องหลังการถ่ายทำ และ กฎของมัวร์ (ยุค 90s – 2005)

แม้จะถูกสื่อมวลชนและนักลงทุนหันหลังให้ แต่นักวิจัยเอไอยุค 90 ได้เปลี่ยนแนวทางอย่างเงียบ ๆ จากการพยายามสร้างสมองกลที่คิดได้ทุกเรื่องมาเป็นการใช้ “คณิตศาสตร์ ตรรกะความน่าจะเป็น และสถิติ” มาแก้ปัญหาเฉพาะด้าน เกิดเป็นรากฐานของการเรียนรู้ของเครื่อง (Machine Learning)

เหตุการณ์ประวัติศาสตร์ที่โลกต้องจารึกคือในปี ค.ศ. 1997 เมื่อซูเปอร์คอมพิวเตอร์ดีปบลู (Deep Blue) ของบริษัทไอบีเอ็ม (IBM) สามารถเอาชนะ แกรีคาสปารอฟ (Garry Kasparov) แชมป์โลกหมากรุกสากลได้สำเร็จ ความสำเร็จนี้ไม่ได้มาจากอัลกอริทึมที่คิดเหมือนมนุษย์ แต่เกิดจาก “พลังคำนวณของฮาร์ดแวร์” ที่เร็วขึ้นจากการพัฒนาอย่างต่อเนื่องของเซมิคอนดักเตอร์ตามกฎของมัวร์ (Moor’s law) ทำให้คอมพิวเตอร์สามารถคำนวณตาเดินล่วงหน้าได้นับล้านรูปแบบในเสี้ยววินาที

รูปที่ 4 (ซ้าย) ภาพคอมพิวเตอร์ดีฟบลู และ (ขวา) ภาพการแข่งขันระหว่างดีฟบลู กับ แกรี คาสปารอฟ ในปี ค.ศ. 1997

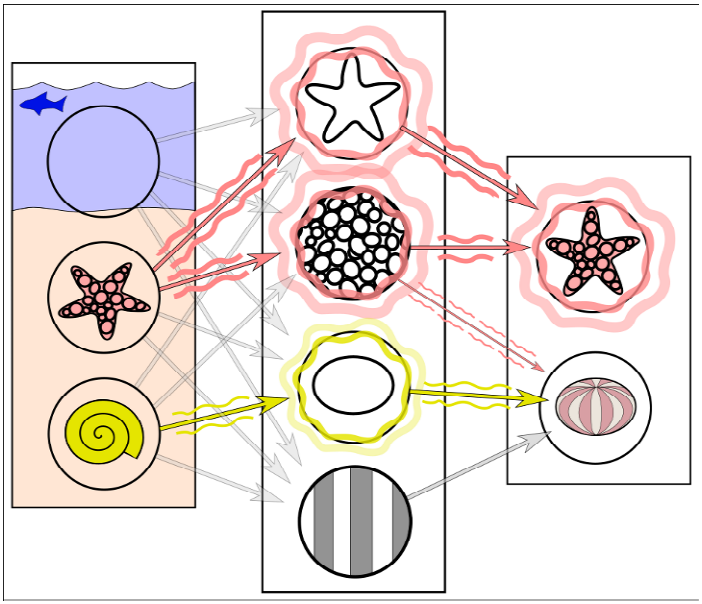
(ที่มา https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer) และ https://www.computerhistory.org/chess)

6. ยุคบิ๊กดาต้า (Big Data) การเรียนรู้เชิงลึก (Deep Learning) และการระเบิดของเอไอ (ค.ศ. 2005–ปัจจุบัน)

ตั้งแต่ปี ค.ศ. 2012 เป็นต้นมา โลกได้เข้าสู่ยุคที่สามของการเกิดใหม่เอไออย่างแท้จริง ขับเคลื่อนด้วย 3 ปัจจัยหลัก

1. **บิ๊กดาต้า (Big Data)**: การเกิดของอินเทอร์เน็ตและโซเชียลมีเดียทำให้มีข้อมูลมหาศาลสำหรับฝึกฝนเอไอ
2. **อัลกอริทึมขั้นสูง**: การพัฒนาโครงข่ายประสาทเทียมหลายชั้นลึกที่เรียกว่า การเรียนรู้เชิงลึก (Deep Learning)
3. **ฮาร์ดแวร์ที่ทรงพลัง**: การนำชิปประมวลผลกราฟิกหรือจีพียู (Graphics Processing Unit, GPU) มาใช้คำนวณงานเอไอ แทนหน่วยประมวลผลกลางหรือซีพียู (Central Processing Unit, CPU) แบบเดิม

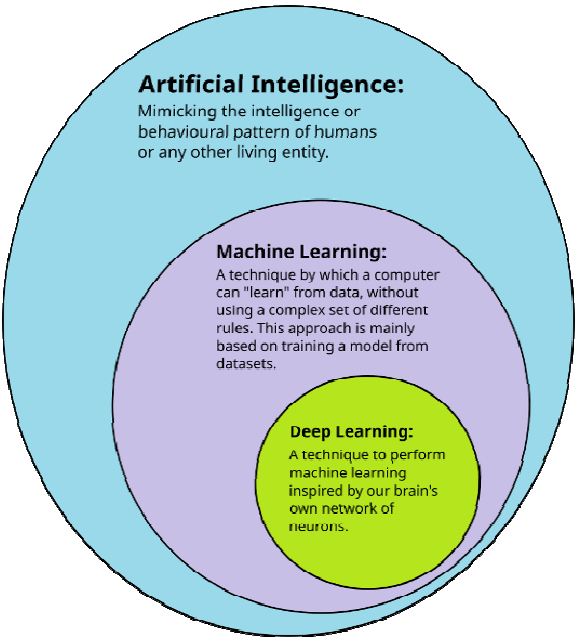
จุดเปลี่ยนครั้งใหญ่เกิดขึ้นในปี ค.ศ. 2017 เมื่อสถาปัตยกรรมที่ชื่อว่า แบบจำลองทรานซ์ฟอร์เมอร์ (Transformer) ถูกนำเสนอ มันกลายเป็นหัวใจหลักที่ขับเคลื่อนระบบปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) และ แบบจำลองภาษาขนาดใหญ่ (Large Language Models, LLMs) เช่น แชทจีพีที (ChatGPT) กูเกิ้ล เจมิไน (Google Gemini) และ คลอดด์ (Claude) ในปัจจุบัน โมเดลเหล่านี้แสดงให้เห็นถึงความสามารถที่ใกล้เคียงกับมนุษย์ ทั้งในด้านความรู้ ความเข้าใจ และความคิดสร้างสรรค์ นำไปสู่ยุคการลงทุนและแข่งขันทางเทคโนโลยีที่ร้อนแรงที่สุดในประวัติศาสตร์



รูปที่ 5 ภาพตัวอย่างของลักษณะการใช้โครงข่ายปราสาทเทียม (Artificial Neural Networks) ซึ่งเป็นการเรียนรู้เชิงลึกแบบหนึ่ง ในการตรวจสอบภาพวัตถุ (ที่มา https://en.wikipedia.org/wiki/Deep_learning)

1.2 แกนหลักของเอไอยุคปัจจุบัน (Core Technologies)

ปัญญาประดิษฐ์ในยุคปัจจุบันไม่ได้ทำงานด้วยระบบกฎเกณฑ์เงื่อนไข (Rule-based) ที่มนุษย์เขียนขึ้นทีละข้ออีกต่อไป แต่ความอัจฉริยะที่เรารู้จักกันในวันนี้ขับเคลื่อนด้วยระบบสถิติ คณิตศาสตร์ชั้นสูง และการประมวลผลข้อมูลมหาศาล โดยมีแกนหลักทางเทคโนโลยี 3 ระดับที่ซ้อนทับและพัฒนาต่อยอดซึ่งกันและกัน ได้แก่ การเรียนรู้ของเครื่อง (Machine Learning), การเรียนรู้เชิงลึก (Deep Learning) และ ปัญญาประดิษฐ์เชิงสร้างสรรค์ ที่มักเรียกว่า เจเนอเรทีฟเอไอ (Generative AI) และ สถาปัตยกรรมแบบจำลองทรานส์ฟอร์เมอร์



รูปที่ 6 แผนผังความสัมพันธ์ของเอไอ การเรียนรู้ของเครื่อง และ การเรียนรู้เชิงลึก (ที่มา https://en.wikipedia.org/wiki/Deep_learning)

1. การเรียนรู้ของเครื่อง (Machine Learning)

อาร์เธอร์ ซามูเอล (Arthur Samuel) ได้เคยนิยามคำนี้ไว้ในปี ค.ศ. 1959 ว่า การเรียนรู้ของเครื่อง คือ “ศาสตร์ที่ทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ได้โดยไม่ต้องอาศัยการเขียนโปรแกรมกำหนดคำสั่งไว้เป็นการเฉพาะ” แทนที่มนุษย์จะบอกวิธีแก้ปัญหาให้คอมพิวเตอร์อย่างตายตัว อัลกอริทึมการเรียนรู้ของเครื่อง จะทำหน้าที่สร้าง “โมเดล” ขึ้นมาจากชุดข้อมูลตัวอย่าง (Training Set) เพื่อค้นหารูปแบบและพยากรณ์ผลลัพธ์ในข้อมูลชุดใหม่ได้อย่างยืดหยุ่น

ในสารานุกรม Wikipedia (https://en.wikipedia.org/wiki/Machine_learning) ได้แบ่งกระบวนการของการเรียนรู้นี้ ออกเป็น 3 รูปแบบหลัก ได้แก่ *การเรียนรู้แบบมีผู้ชี้แนะ (Supervised Learning)* เป็นการฝึกโมเดลด้วยข้อมูลที่มีคำเฉลยกำกับไว้ล่วงหน้า (Labeled Data) เช่น การป้อนรูปภาพสุนัขและแมวที่มีป้ายกำกับชัดเจน เพื่อให้ระบบแยกแยะลักษณะเด่น *การเรียนรู้แบบไม่มีผู้ชี้แนะ (Unsupervised Learning)* เป็นการปล่อยให้โมเดลค้นหาโครงสร้างและรูปแบบที่ซ่อนอยู่ในชุดข้อมูลดิบที่ไม่มีการระบุป้ายกำกับ (Unlabeled Data) ด้วยตัวเอง เช่น การจัดกลุ่มพฤติกรรมลูกค้า และ *การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning)* เป็นการสอนให้ตัวแทน (Agent) เรียนรู้ที่จะตัดสินใจด้วยการทดลองถูกผิดในสิ่งแวดล้อมจำลอง โดยมีระบบการให้รางวัล (Rewards) หรือการลงโทษ (Penalties) เพื่อนำทางไปสู่เป้าหมายที่ดีที่สุด

2. การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกเป็นสาขาย่อยที่ทรงพลังที่สุดของการเรียนรู้ของเครื่อง ในปัจจุบัน โดยโมเดลประเภทนี้จะจำลองโครงสร้างและกลไกการรับส่งกระแสประสาทในสมองของสิ่งมีชีวิต เรียกว่า โครงข่ายประสาทเทียม (Artificial Neural Networks)

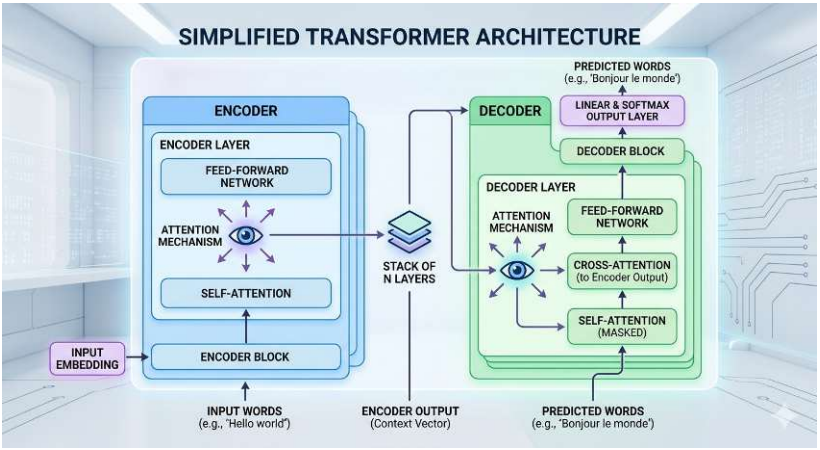
คำว่า “Deep” (เชิงลึก) สะท้อนถึงการนำ “นิวรอนเทียม” (Artificial Neurons) มาเรียงต่อทับซ้อนกันเป็นชั้น ๆ (Layers) ตั้งแต่สามชั้นไปจนถึงหลายร้อยหรือหลายพันชั้น ข้อมูลดิบจะถูกส่งผ่านชั้นเหล่านี้ โดยแต่ละชั้นจะทำหน้าที่สกัดรูปแบบหรือคุณลักษณะที่ซับซ้อนขึ้นเรื่อย ๆ (เช่น ชั้นแรกจับคู่เส้นตรง ชั้นถัดมาจับคู่รูปทรง และชั้นสุดท้ายสามารถบอกได้ว่าเป็นใบหน้าของมนุษย์) สถาปัตยกรรมที่สำคัญของการเรียนรู้เชิงลึกประกอบด้วย *โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network, CNN)* โครงข่ายที่ใช้ตัวกรองแบบคอนโวลูชันออกแบบมาเพื่อตรวจจับรูปแบบในข้อมูลเชิงมิติ เช่น ภาพถ่ายและวิดีโอ และ *โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network, RNN)* โครงข่ายนี้มีกลไกย้อนกลับเพื่อประมวลผลข้อมูลแบบอนุกรมตามเวลา (Sequential/Time-series Data) เช่น ข้อความยาว ๆ หรือคลิปเสียง

3. ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) และ สถาปัตยกรรมแบบจำลองทรานส์ฟอร์เมอร์ (Transformer)

จุดสูงสุดของเทคโนโลยีเอไอยุคนี้เริ่มขึ้นในปี ค.ศ. 2017 เมื่อนักวิจัยของกูเกิ้ลได้เสนองานวิจัยพลิกโลกชื่อ “Attention Is All You Need” ที่อาจแปลได้ว่า “สิ่งที่คุณต้องการทั้งหมดคือความสนใจ” แนะนำสถาปัตยกรรมโครงข่ายประสาทเทียมรูปแบบใหม่ที่เรียกว่า ทรานส์ฟอร์เมอร์

ข้อจำกัดเดิมของโครงข่ายปราสาทเทียมเดิม คือต้องประมวลผลข้อมูลข้อความทีละคำอย่างเป็นลำดับ ทำให้ใช้เวลาฝึกฝนยาวนานและลืมเนื้อหาในส่วนแรก ๆ แต่ทรานส์ฟอร์เมอร์เข้ามาแก้ปัญหานี้โดยใช้กลไกที่เรียกว่า กลไกการใส่ใจในตัวเอง (Self-Attention Mechanism) ซึ่งเปิดโอกาสให้โมเดลสามารถมองเห็นและประมวลผลทุกคำในประโยคพร้อมกันแบบขนาน (Parallel Processing) และคำนวณได้อย่างแม่นยำว่าคำไหนมีความสัมพันธ์กันมากที่สุด

สถาปัตยกรรมแบบจำลองของทรานส์ฟอร์เมอร์นี้เองที่เป็นหัวใจขับเคลื่อนปัญญาประดิษฐ์เชิงสร้างสรรค์และแบบจำลองภาษาขนาดใหญ่ ในอุตสาหกรรมสารสนเทศปัจจุบัน เช่นซีรีส์จีพีที (อันเป็นรากฐานของแชทจีพีที) รวมถึง เจมิไน และคลอดด์ โมเดลเหล่านั้นจะถูกจัดเตรียมล่วงหน้าด้วยข้อมูลมหาศาลจากอินเทอร์เน็ต จนเข้าใจโครงสร้างภาษาและตรรกะในระดับลึก จากนั้นจึงสร้างสรรค์ข้อมูลใหม่ ๆ ออกมา ไม่ว่าจะเป็นข้อความ รหัสโปรแกรมคอมพิวเตอร์ ภาพวาด หรือแม้แต่เสียงและภาพเคลื่อนไหว



รูปที่ 7 ภาพแสดงสถาปัตยกรรมแบบจำลอง ทรานส์ฟอร์เมอร์ โดยภายในประกอบด้วย ฝั่งของตัวเข้ารหัส (Encoder) และ ตัวถอดรหัส (Decoder) พร้อมแสดงกลไกการใส่ใจ (Attention Mechanism) ซึ่งเป็นหัวใจหลักของเอไอตัวนี้

1.3 ข้อจำกัดของฮาร์ดแวร์แบบดั้งเดิม (Conventional Hardware Limitations)

เมื่อโมเดลปัญญาประดิษฐ์ในปัจจุบันเปลี่ยนผ่านเข้าสู่ยุคการเรียนรู้เชิงลึกและสถาปัตยกรรมทรานส์ฟอร์เมอร์ ความต้องการในการประมวลผลข้อมูลก็ไม่ได้เพิ่มขึ้นเป็นเส้นตรงอีกต่อไป แต่ขยายตัวเป็นทวีคูณ (Exponential Growth) ทว่าสถาปัตยกรรมฮาร์ดแวร์ของคอมพิวเตอร์ที่เราใช้งานกันมาเกือบศตวรรษกลับไม่ได้ถูกออกแบบมาเพื่อรองรับคณิตศาสตร์ของเอไอในลักษณะนี้ จนนำไปสู่ปัญหาคอขวดครั้งใหญ่ที่เรียกว่า ปัญหาคอขวดของการคำนวณ (The Compute Bottleneck)

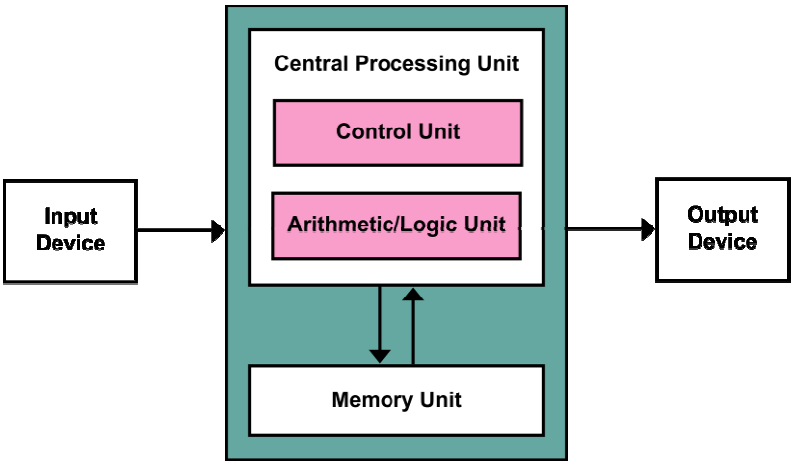
ข้อจำกัดสำคัญของฮาร์ดแวร์ดั้งเดิมสามารถจำแนกออกเป็น 3 ประเด็นหลักตามหลักการของสถาปัตยกรรมคอมพิวเตอร์ ดังนี้

1. ปัญหาคอขวดฟอนนอยมันน์ (Von Neumann Bottleneck)

คอมพิวเตอร์เกือบทั้งหมดในโลก ตั้งแต่สมาร์ตโฟนไปจนถึงซูเปอร์คอมพิวเตอร์ล้วนใช้ สถาปัตยกรรมฟอนนอยมันน์ (Von Neumann Architecture) ซึ่งคิดค้นโดยจอห์น วอน นอยแมน (John von Neumann) มาตั้งแต่ปี ค.ศ. 1945

หลักการพื้นฐานของระบบนี้คือการแยกส่วน หน่วยประมวลผลกลางและหน่วยความจำชนิดแรม (Random Access Memory, RAM) ออกจากกันอย่างเด็ดขาด ทุกครั้งที่คอมพิวเตอร์จะคำนวณอะไรบางอย่าง หน่วยประมวลผลกลางจะต้องส่งสัญญาณไป “ดึง” ข้อมูลจากหน่วยความจำ ผ่านช่องทางส่งข้อมูลที่เรียกว่า “บัส” (Bus) มาคำนวณ เมื่อคำนวณเสร็จก็ต้องส่งกลับไปเก็บที่หน่วยความจำ

ปัญหาสำหรับเอไอคือ โมเดลการเรียนรู้เชิงลึก มีพารามิเตอร์ (Parameters) และค่าถ่วงน้ำหนัก (Weights) ซึ่งเป็นตัวแปรนับพันล้านถึงแสนล้านตัว ข้อมูลมหาศาลเหล่านี้ต้องถูกดึงเข้า-ออกระหว่างหน่วยประมวลผลกลางและหน่วยความจำ



รูปที่ 8 แผนผังโครงสร้างสถาปัตยกรรมวอนนอยมันน์
(ที่มา https://en.wikipedia.org/wiki/Von_Neumann_architecture)

ตลอดเวลา ส่งผลให้ความเร็วของบัสกลายเป็นข้อจำกัดที่คอยฉุดรั้งความเร็วทั้งหมด ต่อให้หน่วยประมวลผลกลางจะประมวลผลได้เร็วแค่ไหน แต่ถ้าต้องนั่งรอข้อมูลส่งมาจากหน่วยความจำ ระบบก็ยังคงมีคอขวดและเกิดการสิ้นเปลืองพลังงานอย่างมาก

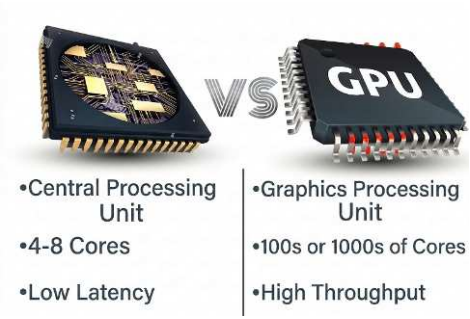
2. ความแตกต่างระหว่างเวลาแฝงและปริมาณงาน

หน่วยประมวลผลกลาง ถูกออกแบบมาให้เป็น “พ่อมดสารพัดประโยชน์” ที่มีประสิทธิภาพอย่างมากในการทำงานที่มีเงื่อนไขซับซ้อนและต้องทำเสร็จอย่างรวดเร็วทีละงาน (เน้นความเร็วต่อหนึ่งคำสั่ง) หรือเรียกว่า มีเวลาแฝงต่ำ (Low Latency) โดยส่วนใหญ่หน่วยประมวลผลกลาง จะมีแกนหรือคอร์ (Core) ในการประมวลผลที่ทรงพลังเพียงไม่กี่แกน (เช่น 4 หรือ 8 แกน)

ปัญหาสำหรับเอไอ คือ งานของเอไอไม่ใช่งานที่ซับซ้อนในเชิงเงื่อนไข แต่เป็นงานคณิตศาสตร์พื้นฐานอย่างการคูณและการบวกเมทริกซ์ (Matrix Multiplication and Addition) ที่ซ้ำ ๆ กันปริมาณมหาศาลพร้อมกัน การเอาหน่วยประมวลผลกลางที่มีไม่กี่แกนมาประมวลผลเอไอ จึงเหมือนกับการเอา “ศาสตราจารย์ปริญญาเอก 4 หรือ 8 คน” มานั่งบวกเลขระดับประถมจำนวน 1 ล้านข้อ ซึ่งทำงานได้ช้ากว่า “เด็กประถม 1,000 คน” ที่ช่วยกันบวกเลขพร้อมกันอย่างแน่นอน เอไอต้องการปริมาณงานต่อเวลาที่สูง (High Throughput)

3. ขีดจำกัดของพลังงานและความร้อน (The Thermal Wall)

ในอดีต ยิ่งเราอยากให้ชิปแรงขึ้น เราก็แค่อัดความเร็วสัญญาณนาฬิกา (Clock Speed) ให้เร็วขึ้น (เช่น จาก 1 GHz เป็น 3 GHz) แต่เมื่อเข้าสู่ช่วงปี ค.ศ. 2000 เป็นต้นมา อุตสาหกรรมชิปต้องเผชิญกับสภาวะตันเนื่องจากหากเร่งความเร็วสัญญาณนาฬิกาเกินกว่านี้ ทรานซิสเตอร์จะปล่อยความร้อนออกมามหาศาลจนชิปอาจละลายได้ คอมพิวเตอร์จึงไม่สามารถเพิ่มประสิทธิภาพให้มากขึ้นได้ง่าย ๆ ด้วยการเร่งความเร็วตัวสัญญาณนาฬิกาของประมวลผลเดี่ยว ๆ อีกต่อไป



รูปที่ 9 การเปรียบเทียบคุณสมบัติชิปหน่วยประมวลผลกลาง (CPU) และชิปประมวลผลกราฟิก (GPU)

เมื่อซอฟต์แวร์เอไอเดินหน้าไปไกล แต่สถาปัตยกรรมคอมพิวเตอร์แบบวอนนอยมันน์และหน่วยประมวลผลกลาง ขับเคลื่อนต่อได้ยากเนื่องจากติดกำแพงความร้อนและปัญหาคอขวดของหน่วยความจำ ทางออกเดียวของมนุษยชาติคือการกลับไปรื้อวิธีคิดในฝั่งฮาร์ดแวร์ใหม่ ตั้งแต่ระดับวัสดุศาสตร์และสารกึ่งตัวนำ ใน บทที่ 2: พื้นฐานเซมิคอนดักเตอร์ เราจะไปทำความเข้าใจร่วมกันว่า แผ่นซิลิคอนและทรานซิสเตอร์ตัวเล็ก ๆ ถูกสร้างขึ้นมาอย่างไร และกฎของมัวร์จะช่วยให้เราก้าวข้ามขีดจำกัดนี้ไปได้อย่างไร

2. พื้นฐานเซมิคอนดักเตอร์ (Semiconductor Foundation)

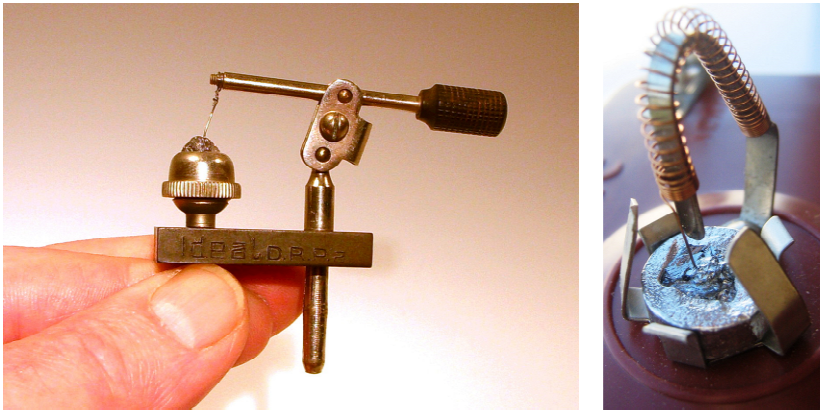
2.1 จุดเริ่มต้น: ฟิสิกส์ของวัสดุสารกึ่งตัวนำ (The Beginning: Semiconductor Physics)

ก่อนที่อุตสาหกรรมคอมพิวเตอร์จะสามารถสร้างชิปประมวลผลที่มีทรานซิสเตอร์นับแสนล้านตัวได้ มนุษยชาติต้องใช้เวลาศึกษาคุณสมบัติอันแปลกประหลาดของวัสดุกลุ่มหนึ่งนานกว่าศตวรรษ วัสดุกลุ่มนี้ไม่ใช่อโลหะที่เป็นฉนวน และไม่ใช่โลหะที่นำไฟฟ้าได้ดี แต่คุณสมบัติทางไฟฟ้าของมันแปรผันตามอุณหภูมิ แสง และการกระตุ้นทางกล ซึ่งขัดกับสามัญสำนึกของนักฟิสิกส์ในยุคดั้งเดิม (ยุคก่อนปี 1900) อย่างสิ้นเชิง

1. ประวัติศาสตร์ยุคแรกและการค้นพบคุณสมบัติสารกึ่งตัวนำ (Early History)

ในหน้าประวัติศาสตร์ของสารกึ่งตัวนำช่วงแรก ๆ ปรากฏบันทึกการค้นพบคุณสมบัติพิเศษเป็นครั้งแรกโดย ไมเคิล ฟาราเดย์ (Michael Faraday) นักเคมีและนักฟิสิกส์ชาวอังกฤษในปี ค.ศ. 1833 ฟาราเดย์พบว่า เมื่อเขาทำการทดลองเพิ่มอุณหภูมิให้กับสาร ซิลเวอร์ซัลไฟด์ (Silver Sulfide, Ag_2S) ค่าความต้านทานไฟฟ้ากลับลดลง ซึ่งขัดแย้งกับพฤติกรรมของโลหะทั่วไป เช่น ทองแดง หรือเงิน ที่เมื่อร้อนขึ้นจะมีความต้านทานไฟฟ้าเพิ่มขึ้น ปรากฏการณ์นี้เป็นสัญญานแรกที่บ่งชี้ว่าสสารชนิดนี้มีกลไกภายในที่ต่างออกไปต่อมาในปี ค.ศ. 1874 คาร์ล เฟอร์ดินานด์ บราวน์ (Karl Ferdinand Braun) นักฟิสิกส์ชาวเยอรมัน ได้ค้นพบคุณสมบัติสำคัญอีกประการหนึ่ง เรียกว่า การเรียงกระแส (Rectification) หรือผลของตัวตรวจจับจุด

สัมผัส (Point-contact effect) บราวน์สังเกตเห็นว่าเมื่อนำปลายลวดโลหะแหลมไปสัมผัสกับผลึกของสารคริสตัลอย่าง แร่กาเลนา (Galena หรือ Lead Sulfide: PbS) กระแสไฟฟ้าจะไหลผ่านได้ง่ายใน ทิศทางเดียว แต่จะถูกบล็อกเมื่อไหลย้อนศร การค้นพบนี้ถูกนำไปพัฒนาต่อยอดโดย เซอร์ จักดิก จันทรา โบส (Sir Jagadish Chandra Bose) นักวิทยาศาสตร์ชาวอินเดีย ในปี ค.ศ. 1894 เพื่อใช้เป็นตัวตรวจจับคลื่นวิทยุตัวแรกของโลก และได้รับการจดสิทธิบัตรในปี ค.ศ. 1901 ในชื่ออุปกรณ์ ตัวตรวจจับหนวดแมว (Cat's-whisker detector) ซึ่งถือเป็นอุปกรณ์ไดโอดสารกึ่งตัวนำ (Semiconductor Diode) ชิ้นแรกที่มนุษย์สร้างขึ้นมาใช้งานจริงในเครื่องรับวิทยุแร่ยุคแรก ทว่า ในยุคนั้นยังไม่มีใครสามารถอธิบายได้เลยว่า เพราะเหตุใดอิเล็กตรอนจึงไหลในแร่กาเลนาได้เพียงทิศทางเดียว จนกระทั่ง มีการคิดค้นกลศาสตร์ควอนตัม (Quantum Mechanics) ในปี ค.ศ. 1925 และ มีการพัฒนาศาสตร์ด้าน ฟิสิกส์สถานะของแข็ง (Solid-state Physics) ในทศวรรษ ค.ศ. 1930 ความลับระดับอะตอมนี้จึงถูกเปิดเผย



รูปที่ 10 อุปกรณ์ตรวจจับสัญญาณวิทยุยุคบุกเบิก ตัวตรวจจับหนวดแมว
(ที่มา https://en.wikipedia.org/wiki/Crystal_detector)

2. ฟิสิกส์ของโครงสร้างแถบพลังงาน (Energy Band Structure)

ตำราฟิสิกส์สารกึ่งตัวนำอธิบายว่า พฤติกรรมทางไฟฟ้าของวัสดุเกิดจากปฏิสัมพันธ์ของอิเล็กตรอนตาม หลักการกีดกันของเพาลี (Pauli Exclusion Principle) เมื่ออะตอมของธาตุมาอยู่รวมกันเป็นผลึกแข็ง ระดับพลังงานของอิเล็กตรอนที่เคยแยกกันจะรวมตัวกันกลายเป็น แถบพลังงาน (Energy Bands) ซึ่งแบ่งออกเป็น 2 แถบหลักที่สำคัญต่อการนำไฟฟ้า

- *แถบวาเลนซ์ (Valence Band)* แถบพลังงานระดับต่ำที่อัดแน่นไปด้วยอิเล็กตรอนที่ยึดเหนี่ยวอยู่กับอะตอม หากแถบนี้เต็มและไม่มีอิเล็กตรอนใดขยับได้ วัสดุนั้นจะไม่นำไฟฟ้า

- *แถบนำไฟฟ้า (Conduction Band)* แถบพลังงานระดับสูงกว่า ซึ่งเป็นพื้นที่ของอิเล็กตรอนอิสระที่สามารถเคลื่อนที่ไปมาทั่วทั้งผลึกเพื่อนำกระแสไฟฟ้าได้

สิ่งที่กำหนดว่าวัสดุใดจะเป็นตัวนำ ฉนวน หรือสารกึ่งตัวนำ คือระยะห่างระหว่างสองแถบนี้ ซึ่งเรียกว่า ช่องว่างพลังงาน (Band gap: E_g)

ตัวนำ (Conductors) แถบวาเลนซ์และแถบนำไฟฟ้าจะซ้อนทับกัน ทำให้อิเล็กตรอนวิ่งไปนำไฟฟ้าได้ทันทีโดยไม่ต้องใช้พลังงานกระตุ้น

ฉนวน (Insulators) มีช่องว่างพลังงาน กว้างมาก (โดยทั่วไป $E_g > 5$ eV) อิเล็กตรอนไม่สามารถรับพลังงานเพียงพอที่จะกระโดดข้ามไปได้

สารกึ่งตัวนำมีช่องว่างพลังงาน ที่ขนาดพอดี (โดยทั่วไป $E_g = 0.5 – 3$ eV) ตัวอย่างเช่น ซิลิคอน (Silicon, Si) มีค่าช่องว่างพลังงานอยู่ที่ประมาณ 1.12 อิเล็กตรอนโวลต์ ณ อุณหภูมิห้อง ซึ่งแคบพอที่พลังงานความร้อนหรือสนามไฟฟ้าจะสามารถกระตุ้นให้อิเล็กตรอนบางส่วน กระโดดขึ้นจากแถบวาเลนซ์ไปสู่แถบนำไฟฟ้าได้ เมื่ออิเล็กตรอนกระโดดขึ้นไปแล้ว มันจะทิ้ง “ช่องว่างที่เสมือนว่ามีประจุบวก” ไว้

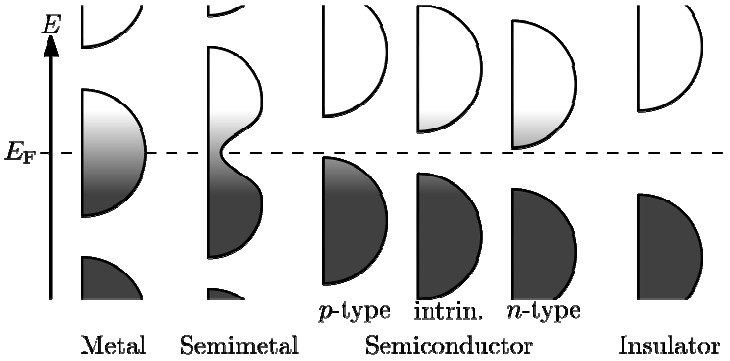
ในแถบวาเลนซ์ เรียกว่า “โฮล” (Hole) ทั้งอิเล็กตรอนอิสระและโฮลนี้เองที่ทำหน้าที่เป็นพาหะนำไฟฟ้า (Charge Carriers)

3. กลไกการพาประจุและการโด๊ปสาร (Carrier Transport and Doping)

ซิลิคอนบริสุทธิ์จะมีจำนวนอิเล็กตรอนอิสระและโฮลเท่ากันพอดีทำให้นำไฟฟ้าได้ต่ำมาก ในทางวิศวกรรมเซมิคอนดักเตอร์ เราต้องการควบคุมปริมาณและประเภทของประจุให้ได้อย่างแม่นยำ จึงต้องใช้กระบวนการโด๊ปสาร (Doping) หรือการผสมธาตุเจือปนระดับอะตอมเข้าไปในโครงสร้างผลึก เพื่อปรับตำแหน่งของ ระดับพลังงานของแถบพลังงาน ซึ่งวิศวกรจะสร้างสารกึ่งตัวนำที่มีการโด๊ปขึ้นมา 2 ประเภท คือ

- *สารกึ่งตัวนำชนิดเอ็น (N-type Semiconductor)* เกิดจากการเติมธาตุหมู่ 5 เช่น ฟอสฟอรัส (P) หรือสารหนู (As) ลงไปในผลึกซิลิคอน เนื่องจากธาตุเหล่านี้มีวาเลนซ์อิเล็กทรอนิกส์ 5 ตัว เมื่อจับคู่กับซิลิคอนที่มี 4 ตัว จะทำให้มีอิเล็กตรอนเหลือเกินมา 1 ตัว กลายเป็นอิเล็กตรอนอิสระทันทีโดยไม่ต้องรอกระตุ้น แถบพลังงานทั้งสองจะขยับลง สารชนิดนี้จึงมี อิเล็กตรอนเป็นพาหะส่วนใหญ่ และมีโฮลเป็นพาหะส่วนน้อย

- *สารกึ่งตัวนำชนิดพี (P-type Semiconductor)* เกิดจากการเติมธาตุหมู่ 3 เช่น โบรอน (B) หรือแกลเลียม (Ga) ซึ่งมีวาเลนซ์อิเล็กทรอนิกส์เพียง 3 ตัว ทำให้เมื่อจับคู่กับผลึกซิลิคอนแล้วจะขาดอิเล็กตรอนไป 1 ตำแหน่ง เกิดเป็นโฮลทันที แถบพลังงานทั้งสองจะขยับขึ้น สารชนิดนี้จึงมี โฮลเป็นพาหะส่วนใหญ่ และมีอิเล็กตรอนเป็นพาหะส่วนน้อย



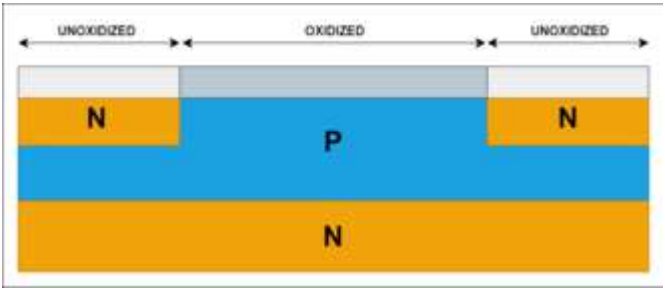
รูปที่ 11 (จากซ้ายไปขวา) แผนภาพแถบพลังงานของโลหะ สารกึ่งโลหะ สารกึ่งตัวนำชนิดพี สารกึ่งตัวนำที่ไม่มีการโด๊ป สารกึ่งตัวนำชนิดเอ็น และ ฉนวน
(ที่มา https://en.wikipedia.org/wiki/Semiconductor)

เมื่อพาหะที่มีประจุเหล่านี้ (คือ อิเล็กตรอนและโฮล) เคลื่อนที่ ภายใต้หลักการฟิสิกส์ภายในชิป จะเกิดปรากฏการณ์เคลื่อนที่ของพาหะสองรูปแบบคือ กระแสไหลแบบเลื่อน (Drift Current) จากแรงผลักของสนามไฟฟ้า และ การไหลแบบแพร่ (Diffusion Current) จากการกระจายตัวของความหนาแน่นประจุ ซึ่งพฤติกรรมคณิตศาสตร์และฟิสิกส์เหล่านี้เองที่จะกลายไปเป็นลักษณะการเรียงกระแสของไอโอดและสวิตช์ปิด-เปิดของทรานซิสเตอร์ทั้งแบบบีเจที (Bipolar Junction Transistor, BJT) และแบบมอสเฟต (Metal Oxide Semiconductor Field Effect Transistor, MOSFET) และเป็นรากฐานในการรวมตัวกันเป็นวงจรลอจิกเกตในที่สุด

4. การประยุกต์สู่ทรานซิสเตอร์และลอจิกเกต (Application to Transistors and Logic Gates)

เมื่อเราสามารถควบคุมจำนวนพาหะนำไฟฟ้าทั้งอิเล็กตรอนและโฮลผ่านการโด๊ปแล้ว ก้าวสำคัญต่อมาที่เปลี่ยนแปลงประวัติศาสตร์มนุษยชาติคือการนำสารกึ่งตัวนำเหล่านั้นมาประกอบกันเพื่อสร้างทรานซิสเตอร์ ซึ่งเป็นอุปกรณ์โซลิดสเตตที่ทำหน้าที่เป็นสวิตช์อิเล็กทรอนิกส์ที่ไม่มีชิ้นส่วนเคลื่อนไหว และกลายเป็นบล็อกอิฐก้อนเล็กที่สุดในชิปคอมพิวเตอร์ทุกชนิดบนโลก

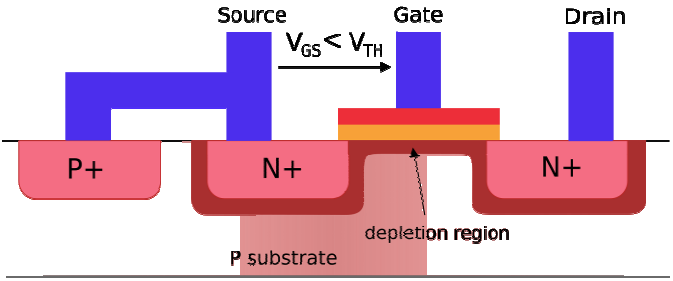
ในยุคแรกเริ่ม ทรานซิสเตอร์ถูกคิดค้นขึ้นในรูปแบบบีเจที ในปี ค.ศ. 1947 ณ ห้องวิจัยเบล (Bell Labs) โดยจอห์น บาร์ดีน, วอลเตอร์ แบรตเทน และวิลเลียม ช็อคลีย์ ซึ่งใช้กระแสไฟฟ้าควบคุมกระแสไฟฟ้า แต่จุดเปลี่ยนสำคัญที่นำไปสู่สถาปัตยกรรมชิปในปัจจุบันคือการประดิษฐ์ทรานซิสเตอร์ชนิดมอสเฟต ในปี ค.ศ. 1959 โดยโมฮาเหม็ด อตัลลา และดาวอน คาง ซึ่งเปลี่ยนมาใช้ ศักย์หรือแรงดันไฟฟ้าในการสร้างสนามไฟฟ้าเพื่อควบคุมการเปิด-ปิดแทน ทำให้อุปกรณ์กินไฟฟ้าน้อยลงอย่างมหาศาลและสามารถย่อขนาดให้เล็กลงได้ง่ายขึ้นโครงสร้างพื้นฐานของเอ็นแชนแนลมอสเฟต (n-channel MOSFET, NMOS) จะประกอบด้วยฐานรองที่เป็นสารกึ่งตัวนำชนิดพี และมีขารับ-ส่งประจุไฟฟ้าที่เรียกว่า ซอส (Source, S) และเดรน (Drain, D) ซึ่งเป็นสารกึ่งตัวนำชนิดเอ็นฝังอยู่สองฝั่ง โดยมีขาควบคุมที่เรียกว่าเกต (Gate, G) วางอยู่ตรงกลาง ชั้นระหว่างขาเกตและฐานรองจะถูกคั่นแยกออกจากกันอย่างเด็ดขาดด้วยฉนวนที่บางเฉียบซึ่งทำจากซิลิคอนไดออกไซด์ (SiO_2)



รูปที่ 12 ภาพไดอะแกรมแสดงหนึ่งในอุปกรณ์ทรานซิสเตอร์แบบซิลิคอนไดออกไซด์ (SiO_2) ที่สร้างขึ้นโดย ฟรอสช์ และ เดอริก ในปี ค.ศ. 1957
(ที่มา https://en.wikipedia.org/wiki/MOSFET)

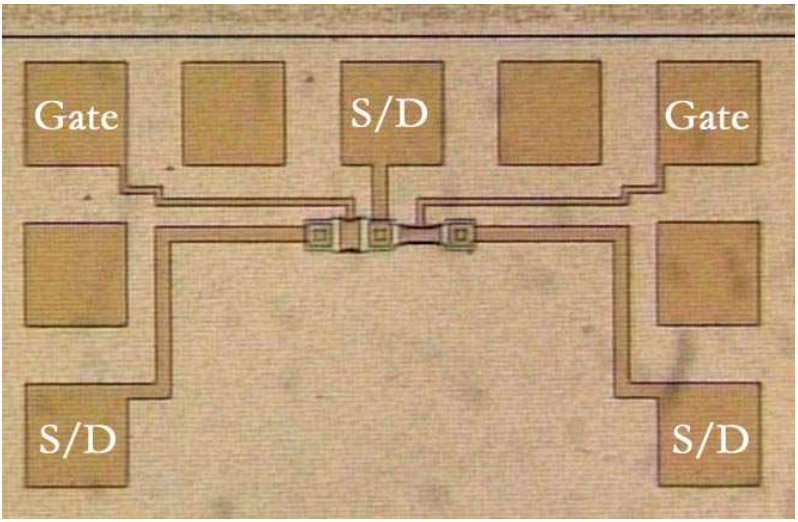
ในกลไกเชิงฟิสิกส์ เมื่อไม่มีการป้อนแรงดันไฟฟ้าเข้าที่ขาเกต (มีสถานะเป็นสวิตช์ปิด หรือลอจิก ‘0’) อิเล็กตรอนจะไม่สามารถวิ่งจากซอสไปยังเดรนได้ เนื่องจากติดแนวกั้นของสารชนิดพีตรงกลาง แต่เมื่อเราป้อนแรงดันไฟฟ้าค่าบวกเข้าที่ขาเกต สนามไฟฟ้าจะทะลุผ่านชั้นฉนวนขจัดโฮลออกไปและดึงดูดอิเล็กตรอนให้มารวมตัวกันหนาแน่นใต้ชั้นฉนวน เกิดปรากฏการณ์ที่เรียกว่า การผกผัน (Inversion) จนกลายเป็นช่องทางเดินประจุเชื่อมระหว่างสองฝั่ง เรียกว่า เอ็นแชนแนล (n-channel) ทำให้อิเล็กตรอนไหลผ่านได้อย่างสะดวก (มีสถานะเป็นสวิตช์เปิด หรือลอจิก ‘1’)

ในกระบวนการออกแบบชิป วิศวกรจะนำมอสเฟตชนิดที่เปิดด้วยไฟบวก (คือ เอ็นแชนแนลมอสเฟต) และชนิดเปิดด้วยไฟลบ (พีแชนแนลมอสเฟต) มาทำงานร่วมกันเป็นคู่ในเทคโนโลยีที่เรียกว่าซีมอส (Complementary MOS, CMOS) เพื่อนำมาต่อพ่วงกันสร้างเป็น ลอจิกเกต (Logic Gates) พื้นฐานทางคณิตศาสตร์ เช่น นิเสธ (NOT), แอนด์ (AND), ออร์ (OR), แนนด์ (NAND) และ นอร์ (NOR) จากนั้นจึงเชื่อมต่อวงจรลอจิกเกตเหล่านี้กัน จนกลายเป็นวงจรคำนวณที่ซับซ้อน เช่น วงจรบวกเลข (Adders) และวงจรคูณเลข (Multipliers)



รูปที่ 13 ภาพตัดขวางของมอสเฟต เมื่อแรงดันไฟฟ้าที่เกตต่ำกว่าระดับแรงดันที่ใช้เปิดแชนแนล (= ช่องนำกระแส) ทำให้ไม่มีการนำกระแสระหว่างขั้วเดรนและขั้วซอร์ส สวิตช์จึงอยู่ในสถานะ ‘ปิด’ และเมื่อเกตมีความเป็นบวกมากขึ้น มันจะดึงดูดอิเล็กตรอนให้เข้ามาใกล้เกต จนเกิดเป็นช่องนำกระแสชนิดเอ็น (n-channel) บริเวณใต้ชั้นออกไซด์ ซึ่งช่วยให้อิเล็กตรอนสามารถไหลระหว่างขั้วเดรนและซอร์สที่ผ่านการเจือสารชนิดเอ็นได้ สวิตช์จึงอยู่ในสถานะ ‘เปิด’

(ที่มา https://en.wikipedia.org/wiki/MOSFET)



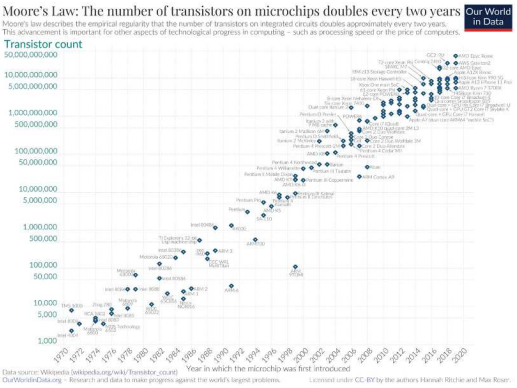
รูปที่ 14 ภาพถ่ายผ่านกล้องจุลทรรศน์ของมอสเฟตแบบเกตโลหะสองตัว โดยมีการระบุข้อความกำกับขั้วสำหรับต่อหัวโพรบ (Probe pads) สำหรับเกตสองตำแหน่ง และจุดต่อซอร์ส/เดรนสามตำแหน่ง

(ที่มา https://en.wikipedia.org/wiki/MOSFET)

สถาปัตยกรรมของเอไอชิปในปัจจุบันก็ตั้งอยู่บนรากฐานนี้ โดยโครงสร้างภายในจะเน้นการจัดวางตำแหน่งวงจรลอจิกเกตเหล่านี้ เพื่อสร้างหน่วยคำนวณคณิตศาสตร์เฉพาะทางปริมาณมหาศาลที่ทำหน้าที่ “คูณแล้วสะสมผลบวก” (Multiply-Accumulate) ของเมทริกซ์ตัวแปรเอไอ พร้อมๆ กันแบบขนาน ดังนั้น ทรานซิสเตอร์นับหมื่นล้านตัวบนเอไอชิปจึงเปรียบเสมือนสวิตช์จิ๋วที่เปิด-ปิดพร้อมกันเพื่อขับเคลื่อนสมองกลอัจฉริยะนั่นเอง

2.2 กฎของมัวร์ในมุมมองปัจจุบัน (Moore's Law in Modern Perspective)

ในโลกของเทคโนโลยีเซมิคอนดักเตอร์ ไม่มีแนวคิดใดที่จะทรงอิทธิพลและสามารถกำหนดทิศทางอุตสาหกรรมได้ยาวนานเท่ากับ กฎของมัวร์ (Moore's Law) แนวคิดนี้ไม่ได้เป็นกฎเกณฑ์ที่ตายตัวทางฟิสิกส์เหมือนกับกฎของนิวตันหรือสมการของแมกซ์เวลล์ แต่เริ่มต้นจากการเป็นเพียงข้อสังเกตเชิงพยากรณ์ทางเศรษฐศาสตร์และวิศวกรรม โดยกอร์ดอน มัวร์ (Gordon Moore) ผู้ร่วมก่อตั้งบริษัทอินเทล (Intel) ได้ตีพิมพ์บทความลงในวารสารวิชาการชื่อ Electronics เมื่อปี ค.ศ. 1965 เขาตั้งข้อสังเกตว่า จำนวนของทรานซิสเตอร์และส่วนประกอบอื่นๆ บนชิปวงจรรวม (Integrated Circuit, IC) ขนาดเท่าเดิม จะเพิ่มขึ้นเป็นสองเท่าในทุก ๆ ปี ต่อมาในปี ค.ศ. 1975 เขาได้ปรับแก้ตัวเลขคำทำนายว่า จำนวนทรานซิสเตอร์จะเพิ่มขึ้นเป็นสองเท่าในทุกๆ สองปี (24 เดือน)



รูปที่ 15 กราฟแสดงจำนวนทรานซิสเตอร์ในไมโครโพรเซสเซอร์เทียบกับปีที่เปิดตัว โดยมีจำนวนเพิ่มขึ้นเกือบเป็นสองเท่าในทุกๆ สองปี

(ที่มา https://en.wikipedia.org/wiki/Moore%27s_law)

ตลอดระยะเวลากว่าครึ่งศตวรรษที่ผ่านมา อุตสาหกรรมคอมพิวเตอร์และเซมิคอนดักเตอร์โลกได้ใช้กฎของมัวร์เป็น “เข็มทิศและเป้าหมาย” ในการวางยุทธศาสตร์ทางธุรกิจและวิศวกรรม ความสามารถในการย่อขนาดทรานซิสเตอร์ให้เล็กลงครึ่งหนึ่งในทุกๆ สองปี ส่งผลให้ชิปประมวลผลมีพละกำลังสูงขึ้นอย่างเป็นทวีคูณ ในขณะที่ต้นทุนการผลิตต่อหน่วยทรานซิสเตอร์ลดต่ำลงอย่างมหาศาล ทำให้คอมพิวเตอร์ที่เคยมีขนาดเท่าห้องโถงย่อส่วนลงมาอยู่ในสมาร์ตโฟนบนฝ่ามือของเราได้ และกลายเป็นตัวเร่งปฏิกิริยาสำคัญที่ทำให้ระบบเอไอสามารถฟื้นคืนชีพขึ้นมาจากยุคฤดูหนาวได้สำเร็จเนื่องจากฮาร์ดแวร์มีพลังการคำนวณที่สูงพอ

ทว่า ในมุมมองปัจจุบันและอนาคตอันใกล้ กฎของมัวร์กำลังเดินทางมาถึงทางตันเนื่องจากต้องเผชิญหน้ากับ ขีดจำกัดทางฟิสิกส์ขั้นรุนแรง ซึ่งนักวิเคราะห์และวิศวกรในอุตสาหกรรมเซมิคอนดักเตอร์ต่างยอมรับว่าเราไม่สามารถย่อขนาดทรานซิสเตอร์ด้วยสัดส่วนเดิมได้อีกต่อไป โดยมีปัจจัยความท้าทายหลักอยู่ 3 ประการ

1. การพังทลายของกฎการย่อขนาดเดนนาร์ด (The Breakdown of Dennard Scaling)

โรเบิร์ต เดนนาร์ด (Robert Dennard) ได้เสนอกฎทางฟิสิกส์ควบคู่กับกฎของมัวร์ไว้ในปี ค.ศ. 1974 ว่า ทุกครั้งที่เราย่อขนาดทรานซิสเตอร์ให้เล็กลง ความหนาแน่นของพลังงานต่อพื้นที่ของชิปจะคงเดิมเสมอ หมายความว่า ยิ่งทรานซิสเตอร์เล็กลง มันจะยิ่งใช้แรงดันไฟฟ้าน้อยลงและกินไฟน้อยลง ทำให้เราสามารถเร่งความเร็วของสัญญาณนาฬิกาให้สูงขึ้นได้โดยที่ชิปไม่ร้อนจนเกินไป ทว่า กฎนี้ได้พังทลายลงอย่างสมบูรณ์ในช่วงปี ค.ศ. 2005–2006 เนื่องจากเมื่อทรานซิสเตอร์เล็กเกินไป สารกึ่งตัวนำจะไม่สามารถกักเก็บประจุไฟฟ้าได้ดีพอ เกิดกระแสไฟฟ้ารั่วตลอดเวลา ส่งผลให้ความร้อนสะสมบนชิปสูงพุ่งสูงจนชนขีดจำกัดด้านความร้อนของวัสดุ วิศวกรจึงไม่สามารถเพิ่มความเร็วสัญญาณนาฬิกาของตัวประมวลผลเดี่ยวให้สูงเกินกว่าระดับ 4-5 GHz ได้อีกต่อไป

2. ปรากฏการณ์ควอนตัมทันเนลลิ่ง (Quantum Tunnelling)

เมื่ออุตสาหกรรมก้าวเข้าสู่กระบวนการผลิตระดับล้ำหน้าในปัจจุบัน (เช่น ระดับ 3 นาโนเมตร หรือ 2 นาโนเมตร) ระยะห่างระหว่างขารับประจุ (ขาซอส) และขาส่งประจุ (ขาเดรน) ของทรานซิสเตอร์ รวมถึงความหนาของชั้นฉนวนใต้ขาเกต จะมีขนาดบางเฉียบเพียงไม่กี่สิบอะตอมของธาตุซิลิคอน ณ จุดนี้ กฎฟิสิกส์คลาสสิกที่มนุษย์คุ้นเคยจะเริ่มใช้ไม่ได้ผล อิเล็กตรอนจะเริ่มแสดงพฤติกรรมตามหลักกลศาสตร์ควอนตัม โดยมีแนวโน้มที่จะ “วาร์ป” หรือทะลุผ่านกำแพงฉนวนที่กั้นอยู่ได้เองตามธรรมชาติ ปรากฏการณ์นี้เรียกว่า ปรากฏการณ์ควอนตัมทันเนลลิ่ง ส่งผลให้ทรานซิสเตอร์ไม่สามารถทำหน้าที่เป็นสวิตช์ปิด-เปิดที่สมบูรณ์ได้ เพราะต่อให้ปิดสวิตช์ (สั่งการให้ลอจิกเป็น ‘0’) แต่อิเล็กตรอนก็ยังคงรั่วไหลข้ามแนวกั้นไปได้ ทำให้สิ้นเปลืองพลังงานและชิปเกิดความร้อนในระบบการคำนวณ

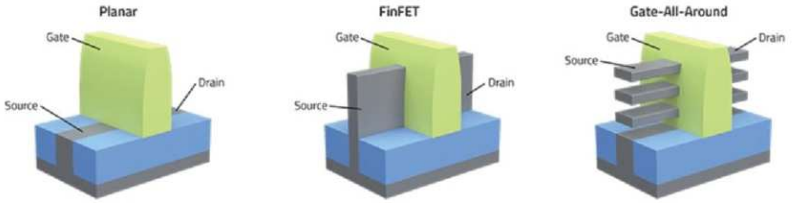
3. ต้นทุนทางเศรษฐศาสตร์ที่สวนทาง (Economic Bottleneck)

นอกเหนือจากขีดจำกัดทางฟิสิกส์แล้ว ปัจจุบันเรายังติดขัดในเรื่องของต้นทุน ในอดีตการย่อขนาดทรานซิสเตอร์จะทำให้ราคาต่อแกนถูกลง แต่ในปัจจุบัน การจะสร้างชิปที่มีขนาดสถาปัตยกรรมระดับนาโนเมตรเดี่ยว ๆ จำเป็นต้องพึ่งพาเทคโนโลยีที่ซับซ้อนและมีราคาสูงลิบลิ่ว เช่น เครื่องพิมพ์ลายวงจรด้วยแสงความยาวคลื่นสั้นพิเศษย่านอียูวี (Extreme Ultraviolet, EUV) ราคาเครื่องละหลายร้อยล้านดอลลาร์ ทำให้ค่าวิจัยและการตั้งโรงงานผลิต (Foundry) พุ่งสูงขึ้นเป็นเงาตามตัว จนกระทั่งเจนเซน หวง (Jensen Huang) ซีอีโอของเอ็นวีเดีย (NVIDIA) ได้เคยออกมากล่าวประโยคประวัติศาสตร์ว่า “กฎของมัวร์ได้ตายไปแล้ว” (Moore's Law is dead) ในแง่ที่ว่าเราไม่สามารถได้ฮาร์ดแวร์ที่แรงขึ้นเป็นเท่าตัวในราคาที่ถูกลงเท่าเดิมได้อีกต่อไป

ยุคสมัยแห่งมอร์แดนมัวร์ (“More than Moore”) และทางรอดของเอไอชิป

เพื่อก้าวข้ามผ่านวิกฤตการณ์ทางตันนี้ อุตสาหกรรมเซมิคอนดักเตอร์จึงปรับเปลี่ยนกระบวนทัศน์เข้าสู่ยุคมอร์แดนมัวร์ (More than Moore) หรือหลังจากมัวร์ ในยุคนี้วิศวกรจะมีการมองหาความล้ำหน้าผ่านช่องทางอื่น ๆ แทนที่จะมุ่งเน้นการบีบอัดทรานซิสเตอร์ในแนวราบเพียงอย่างเดียว นวัตกรรมในมุมมองปัจจุบันจึงเปลี่ยนไปสู่ การปฏิวัติโครงสร้างสามมิติ และ สถาปัตยกรรมเฉพาะทาง

การปฏิวัติโครงสร้างสามมิติ ซึ่งเป็นเปลี่ยนจากทรานซิสเตอร์แนวราบแบบเดิมไปสู่โครงสร้างครีบสามมิติอย่างฟินเฟต (Fin Field-Effect Transistor, FinFET) และล่าสุดคือโครงสร้างที่ขาเกตโอบล้อมช่องทางเดินประจุทุกทิศทางอย่าง ที่เรียกว่า จีเอเอเฟต (Gate-All-Around FET, GAAFET) เพื่อควบคุมการรั่วไหลของกระแสไฟให้ดีขึ้น



รูปที่ 16 ลักษณะโครงสร้างมอสเฟต ฟินเฟตและจีเอเอเฟต

(ที่มา Duan, H. (2024). From MOSFET to FinFET to GAAFET: The evolution, challenges, and future prospects. Proceedings of the 4th International Conference on Signal Processing and Machine Learning, 50, 113–120.)

สถาปัตยกรรมเฉพาะทาง (Domain-Specific Architecture) คนส่วนมากยอมรับแล้วว่าชิปหน่วยประมวลผลกลางที่ทำได้ทุกหน้าที่นั้นหมดเวลาการเป็นผู้นำเดี่ยวของอุตสาหกรรมเซมิคอนดักเตอร์แล้ว เราจึงหันมาสร้างชิปที่เกิดมาเพื่อทำงานบางประเภทโดยเฉพาะ เช่น การสร้างชิปเร่งความเร็วเอไอ (AI Accelerators) ที่ออกแบบวงจรภายในมาให้เก่งเฉพาะเรื่องการคูณและบวกเมทริกซ์คณิตศาสตร์ของแบบจำลองภาษาขนาดใหญ่ และการเรียนรู้เชิงลึก ซึ่งให้ประสิทธิภาพการทำงานต่อวัตต์ (Performance-per-Watt) สูงกว่าการใช้หน่วยประมวลผลกลางหลายเท่าตัว

2.3 ห่วงโซ่อุปทานและกระบวนการผลิต (Semiconductor Supply Chain & Fabrication)

กระบวนการสร้างชิปเซมิคอนดักเตอร์ โดยเฉพาะชิปประมวลผลสำหรับปัญญาประดิษฐ์หรือเอไอชิปที่มีทรานซิสเตอร์ระดับแสนล้านตัวอัดแน่นอยู่ในพื้นที่ขนาดเท่าเหรียญสิบบาท ถือเป็นหนึ่งในกิจกรรมทางวิศวกรรมที่ซับซ้อน ละเอียดอ่อน และใช้เงินทุนสูงที่สุดในประวัติศาสตร์ของมนุษยชาติ ความซับซ้อนนี้ส่งผลให้อุตสาหกรรมนี้ไม่สามารถเบ็ดเสร็จได้ภายในองค์กรเดียวหรือประเทศเดียว แต่ต้องพึ่งพาระบบห่วงโซ่อุปทานโลก (Global Supply Chain) ที่มีความเชี่ยวชาญเฉพาะทางขั้นสูงและตัดขาดจากกันไม่ได้

1. โครงสร้างและโมเดลธุรกิจในห่วงโซ่อุปทานเซมิคอนดักเตอร์โลก

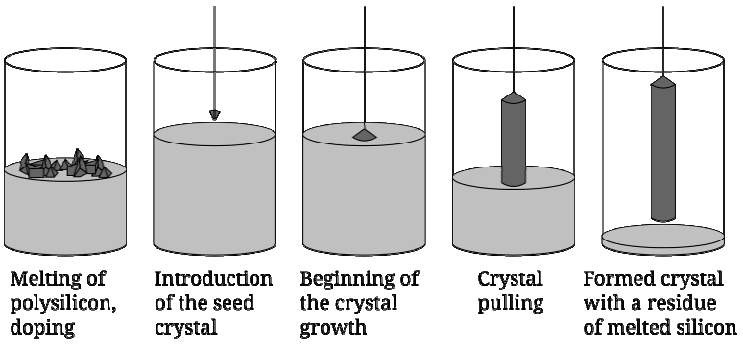
ในอดีต บริษัทเซมิคอนดักเตอร์มักเป็นรูปแบบไอดีเอ็ม (Integrated Device Manufacturer, IDM) คือทำทุกอย่างตั้งแต่ต้นน้ำถึงปลายน้ำ ตั้งแต่ออกแบบ สร้างโรงงานผลิต ไปจนถึงทดสอบและบรรจุภัณฑ์ (packaging) ด้วยตัวเอง ตัวอย่างที่ชัดเจนคือ บริษัท อินเทล และ ซัมซุง) แต่เมื่อเทคโนโลยีสถาปัตยกรรมชิปเล็กลงในระดับนาโนเมตรเดี่ยว ค่าใช้จ่ายในการสร้างและดูแลโรงงานผลิตชิป (Fab หรือ

Foundry) พุ่งสูงขึ้นแตะระดับหมื่นล้านดอลลาร์สหรัฐต่อหนึ่งโรงงาน ทำให้อุตสาหกรรมเกิดการแบ่งขั้วห่วงโซ่อุปทานออกเป็นโมเดลเฉพาะทางหลัก ๆ 3 ส่วน คือ (1) แฟบเลส (Fabless) หรือ บริษัทผู้ออกแบบชิป คือบริษัทที่ไม่มีโรงงานผลิตของตัวเอง แต่มุ่งเน้นไปที่การวิจัย พัฒนา และออกแบบสถาปัตยกรรมวงจรชิปตามโจทย์ทางธุรกิจ จากนั้นจึงส่งไฟล์พิมพ์เขียวดิจิทัลของชิป ไปให้โรงงานภายนอกผลิตให้ บริษัทกลุ่มนี้ได้แก่ เอ็นวีเดีย , เอเอ็มดี (AMD), ควอลคอม (Qualcomm) และแอปเปิ้ล (2) โรงงานรับจ้างผลิต (Foundry) คือบริษัทที่ไม่ทำการออกแบบชิปแข่งกับลูกค้า แต่ลงทุนมหาศาลเพื่อสร้างโรงงานและซื้อเครื่องจักรล้ำสมัยมารับจ้างผลิตชิปตามสั่งอย่างเดียว โดยผู้นำเบอร์หนึ่งของโลกที่มีเทคโนโลยีล้ำหน้าที่สุดคือ ทีเอสเอ็มซี (Taiwan Semiconductor Manufacturing Company, TSMC) จากไต้หวัน ซึ่งควบคุมสัดส่วนการผลิตชิปขั้นสูงของโลกไว้มากกว่า 90% และ (3) โอแสด (Outsourced Semiconductor Assembly and Test, OSAT) เป็นกลุ่มบริษัทปลายน้ำที่รับชิปที่ผลิตเสร็จเป็นแผ่นเวเฟอร์จากโรงงานผลิตชิปมาทำหน้าที่ตัดแบ่ง ทดสอบระบบ และบรรจุภัณฑ์ ออกมาเป็นตัวชิปที่พร้อมเชื่อมต่อบนบอร์ด บริษัทจำนวนมากในประเทศไทยทำงานในส่วนโอแสด (ประเทศไทยมีโรงงานรับจ้างผลิต 1 แห่งคือ ศูนย์เทคโนโลยีไมโครอิเล็กทรอนิกส์ (Thai Microelectronics Center, TMEC) และ บริษัทผู้ออกแบบชิป 1 แห่ง คือ ซิลิกอนคราฟท์ (Silicon Craft Technology PLC))

2. กระบวนการผลิตชิปจากเม็ดทรายสู่เวเฟอร์ (Silicon Ingot to Wafer)

วัสดุกำเนิดของชิปเอไอทั้งหมดเริ่มต้นจากสิ่งที่มีอยู่ล้นเหลือบนโลก นั่นคือทรายซิลิกาหรือ ซิลิคอนไดออกไซด์ (SiO_2) โดยเริ่มจาก **การสกัดและทำให้บริสุทธิ์** ทรายจะถูกนำไปผ่านกระบวนการทางเคมีและความร้อนสูงในเตาหลอมอาร์กไฟฟ้าเพื่อขจัดออกซิเจนออก จนได้ซิลิคอนเกรดโลหะวิทยา จากนั้นนำไปทำปฏิกิริยาเคมี

ซ้ำๆ จนได้ซิลิคอนที่มีความบริสุทธิ์สูงเป็นพิเศษในระดับอิเล็กทรอนิกส์ (Electronic-grade Silicon) ซึ่งมีความบริสุทธิ์สูงถึง 99.999999999% (หรือที่เรียกว่า Eleven Nines คือมีอะตอมแปลกปลอมปนเปื้อนไม่เกิน 1 อะตอมในซิลิคอน 1 แสนล้านอะตอม) จากนั้นจึงทำ**การดึงแท่งผลึกเดี่ยวด้วยกระบวนการโชชราล์สกี (Czochralski Process)** ซิลิคอนบริสุทธิ์จะถูกหลอมละลายอีกครั้ง จากนั้นจะใช้ “หัวเชื้อผลึกซิลิคอน” จุ่มลงไปและค่อยๆ หมุนดึงขึ้นมาอย่างช้า ๆ ภายใต้การควบคุมอย่างแม่นยำ กระบวนการนี้จะทำให้ซิลิคอนเย็นตัวลงและก่อตัวเป็นแท่งผลึกเดี่ยวรูปทรงกระบอกขนาดใหญ่ เรียกว่า ซิลิคอนอินก็อต (Silicon Ingot) แล้วจึงทำ**การหั่นแผ่นเวเฟอร์ (Wafer Slicing)**: แท่งอินก็อตจะถูกนำมาขัดผิวให้กลมมนได้มาตรฐาน (ส่วนใหญ่มีเส้นผ่านศูนย์กลาง 300 มิลลิเมตร หรือ 12 นิ้ว) จากนั้นจะใช้ลวดเพชรความเร็วสูงหั่นแท่งซิลิคอนนี้ออกเป็นแผ่นดิสก์บางเฉียบที่มีความหนาน้อยกว่า 1 มิลลิเมตร เรียกว่า เวเฟอร์ (Wafer) แผ่นเวเฟอร์นี้จะถูกนำไปขัดเงาจนเรียบเนียนดั่งกระจกเงาเพื่อเตรียมเข้าสู่โรงงานผลิตขั้นต่อไป



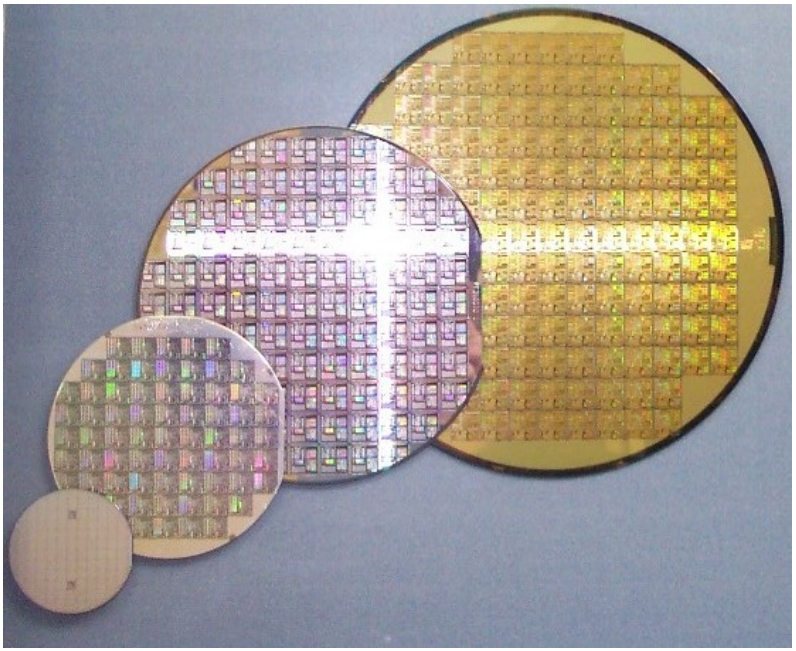
รูปที่ 17 แผนภาพลักษณะการการดึงแท่งผลึกเดี่ยวด้วยกระบวนการโชชราล์สกี (ที่มา https://en.wikipedia.org/wiki/Wafer_(electronics))

3. กระบวนการพิมพ์ลายวงจรด้วยแสงความยาวคลื่นสั้นพิเศษอียูวี (Photolithography & EUV)

ภายในโรงงานผลิตชิปซึ่งเป็นห้องสะอาดควบคุมฝุ่นละอองอย่างเข้มงวดขั้นสูงสุด (Cleanroom Class 100 คือมีฝุ่นขนาดเล็กกว่า 0.5 ไมครอนไม่เกิน 100 เม็ดในอากาศหนึ่งลูกบาศก์ฟุต) แผ่นเวเฟอร์จะถูกนำมาผ่านกระบวนการซ้ำๆ นับร้อยรอบเพื่อพิมพ์ลายวงจร โดยขั้นตอนที่เป็นดั่ง “หัวใจและคอขวด” ของอุตสาหกรรม คือ การพิมพ์ลวดลายด้วยแสง (Photolithography) [มีผู้แปลคำนี้ไว้หลายคำ เช่น การพิมพ์หิน หรือ การสลักลายด้วยแสง]

เพื่อให้สามารถพิมพ์ลายวงจรสวิตช์ทรานซิสเตอร์ที่มีขนาดเล็กระดับนาโนเมตรลงบนเวเฟอร์ได้ โรงงานจำเป็นต้องใช้เทคโนโลยีขั้นสุดยอดที่ใช้แสงอียูวี (Extreme Ultraviolet Lithography) ซึ่งเป็นเครื่องจักรที่ผลิตได้โดยบริษัทเดียวในโลกคือ เอเอสเอ็มแอล (ASML) ที่ผลิตจากประเทศเนเธอร์แลนด์ เครื่องจักรนี้จะใช้ลำแสงเลเซอร์คาร์บอนไดออกไซด์ยิงใส่หยดดีบุกเหลวที่กำลังร่วงหล่นด้วยความเร็วสูง เพื่อสร้างพลาสม่าที่แผ่รังสีแสงช่วงคลื่นสั้นพิเศษที่มีความยาวคลื่นเพียง 13.5 นาโนเมตร ออกมา

แสงย่านอียูวีนี้จะถูกสะท้อนผ่านระบบกระจกเงาที่มีความเรียบเนียนระดับอะตอม (เนื่องจากแสงช่วงคลื่นนี้จะถูกดูดกลืนโดยแก้วที่ใช้ในเลนส์ทั่วไปหมด) เพื่อฉายผ่านแผ่นพิมพ์ลาย (Mask) ที่ย่อส่วนวงจรลงมา จากนั้นแสงจะตกกระทบลงบนแผ่นเวเฟอร์ที่ถูกเคลือบไว้ด้วยสารไวแสง (Photoresist) บริเวณที่ถูกแสงจะเปลี่ยนคุณสมบัติทางเคมี ทำให้สามารถล้างออกได้ในขั้นตอนการล้าง (Developing) ทิ้งไว้เพียงลายวงจรตามที่ออกแบบไว้ จากนั้นแผ่นเวเฟอร์จะถูกส่งไปผ่านกระบวนการกัด (Etching) ด้วยแก๊สพลาสม่า และกระบวนการโด๊ปฝังไอออนเคมี (Ion Implantation) เพื่อสร้างชั้นทรานซิสเตอร์และสายไฟเชื่อมต่อโลหะทับซ้อนกันขึ้นไปหลายสิบชั้น



รูปที่ 18 เวเฟอร์ขนาด 2, 4, 6 และ 8 นิ้ว

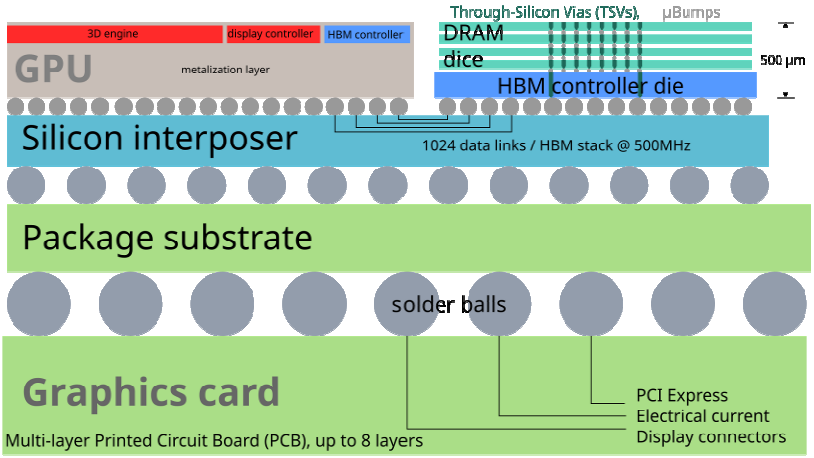
(ที่มา https://en.wikipedia.org/wiki/Wafer_(electronics))

4. การบรรจุภัณฑ์ขั้นสูง: กลไกทำลายคอขวดของเอไอชิป

เมื่อกระบวนการบนเวเฟอร์เสร็จสิ้นลง เราจะได้ชิปสี่เหลี่ยมชิ้นเล็กๆ นับร้อยนับพันชิ้นอยู่บนแผ่นวงกลม ซึ่งชิปแต่ละชิ้นจะถูกตัดแบ่งออกมา เรียกว่า ดาย (Die) ในอดีต การบรรจุภัณฑ์หรือแพ็คเกจจิ้งคือการนำดายชิ้นนี้ไปเชื่อมต่อสายไฟแล้วครอบพลาสติกสีดำติดขาพินธรรมดา แต่สำหรับเอไอชิปสมัยใหม่ กระบวนการแบบเดิมจะก่อให้เกิดปัญหาคอขวดของข้อมูลทันที เพราะหน่วยประมวลผลคำนวณได้เร็วมาก แต่สายไฟแบบเดิมส่งข้อมูลจากหน่วยความจำมาให้ไม่ทัน อุตสาหกรรมจึงต้องหันมาพึ่งพาการบรรจุภัณฑ์ขั้นสูง (Advanced Packaging) ซึ่งถือเป็นเทคโนโลยีชี้ขาดแพ้-

ชนะของเอไอชิป ในปัจจุบัน ตัวอย่างระบบที่เป็นมาตรฐานอุตสาหกรรมคือ โควอส (Chip-on-Wafer-on-Substrate, CoWoS) ของทีเอสเอ็มซี

เทคนิคโควอส คือการนำดายประมวลผลลอจิก (Logic Die) เช่น จีพียู/ทีพียู มาวางระนาบเคียงข้างติดกับ หน่วยความจำแบนด์วิดท์สูง (High Bandwidth Memory, HBM) บนแผ่นเชื่อมต่อซิลิคอนตัวกลางชิ้นเดียวกันที่เรียกว่าซิลิคอนอินเตอร์โพเซอร์ (Silicon Interposer) แผ่นตัวกลางนี้จะมีเส้นลวดเชื่อมต่อขนาดไมโครเมตรนับหมื่น ๆ เส้นเชื่อมต่อกันอยู่ภายใน ทำให้อัตราการรับส่งข้อมูลระหว่างสมองกลเอไอและหน่วยความจำมีความเร็วสูงขึ้นอย่างมหาศาลและมีระยะทางที่สั้นลงจนเกือบเป็นศูนย์ กระบวนการแพ็คเกจจิ้งขั้นสูงนี้เองที่เป็นหนึ่งในชิ้นส่วนจิ๊กซอว์สำคัญที่สุดที่ทำให้ซูเปอร์ชิปอย่าง NVIDIA H100, B200 หรือ Google TPU สามารถทำงานและประมวลผลข้อมูลโมเดลภาษาขนาดใหญ่ได้อย่างมีประสิทธิภาพ



รูปที่ 19 ลักษณะภาพตัดขวางของกราฟิกการ์ดที่ใช้หน่วยความจำแบนด์วิดท์สูง โดยใช้เทคโนโลยีชิปสามมิติแบบรูเจาะทะลุแผ่นซิลิคอน

(ที่มา https://en.wikipedia.org/wiki/Three-dimensional_integrated_circuit)

เมื่อเรามองย้อนกลับไปตั้งแต่จุดเริ่มต้นในระดับอะตอมของฟิสิกส์สารกึ่งตัวนำ กลไกสวิตช์เปิด-ปิดทางไฟฟ้าของทรานซิสเตอร์มอสเฟต โครงสร้างสามมิติแบบจีเอเอเฟต ตลอดจนความก้าวหน้าทางวิศวกรรมในห่วงโซ่อุปทานโลกที่ต้องใช้แสงความยาวคลื่นสั้นพิเศษอียูวี พิมพ์ลายวงจรลงบนแผ่นซิลิคอนที่มีความบริสุทธิ์สูงมาก สิ่งเหล่านี้คือข้อพิสูจน์ว่า “พลังคำนวณ” ไม่ใช่สิ่งที่เป็นนามธรรม หากแต่เป็นผลลัพธ์ที่จับต้องได้จากการการใช้กฎทางฟิสิกส์และเคมี ทว่าเมื่อการย่อขนาดตามกฎของมัวร์ต้องเผชิญหน้ากับกำแพงความร้อนและปรากฏการณ์ควอนตัมทันเนลลิ่งอันเป็นขีดจำกัดทางฟิสิกส์ที่ไม่สามารถก้าวข้ามได้ ทางรอดเดียวของอุตสาหกรรมจึงไม่ใช่การสร้างสวิตช์ที่เล็กลงในหน่วยประมวลผลทั่วไปแบบเดิมอีกต่อไป แต่คือการปฏิวัติวิธีคิดในการจัดวางผังเมืองทรานซิสเตอร์เหล่านั้นใหม่ทั้งหมด นวัตกรรมการบรรจุภัณฑ์ขั้นสูงอย่าง โควอสและการออกแบบสถาปัตยกรรมเฉพาะทางเพื่อเร่งความเร็วการคำนวณแบบขนาน จึงกลายเป็นสะพานเชื่อมชิ้นสำคัญที่เปลี่ยนผ่านจากยุคกายภาพของวัสดุศาสตร์ เข้าสู่ยุคสมัยแห่งการสร้างสถาปัตยกรรมสมองกลที่จะถูกจารึกไว้ในหน้าประวัติศาสตร์ถัดไป ซึ่งเรากำลังจะร่วมออกเดินทางไปเจาะลึกพร้อมกันใน บทที่ 3 มุ่งสู่เอไอชิป

3. มุ่งสู่เอไอชิป (Towards AI Chip)

3.1 ทำไมต้อง เอไอชิป? (Why AI Chips?)

ในการประมวลผลคอมพิวเตอร์ยุคดั้งเดิม หน่วยประมวลผลกลาง หรือ ซีพียู (Central Processing Unit, CPU) ถูกออกแบบมาในฐานะ ผู้รอบรู้ สถาปัตยกรรมภายในของซีพียูจึงเน้นการประมวลผลคำสั่งแบบเรียงลำดับทีละคำสั่งด้วยความเร็วสูง มีหน่วยความจำแคชขนาดใหญ่ และมีหน่วยควบคุมตรรกะที่ซับซ้อน เพื่อให้สามารถรองรับการทำงานของระบบปฏิบัติการ สลับโปรแกรมไปมา และคำนวณตรรกะที่มีเงื่อนไขซับซ้อน (เช่น คำสั่ง If-Else หรือ Branch Prediction) ได้อย่างราบรื่น ทว่าเมื่อโลกก้าวเข้าสู่ยุคของปัญญาประดิษฐ์ โดยเฉพาะอัลกอริทึม การเรียนรู้เชิงลึก และแบบจำลองภาษาขนาดใหญ่ (Large Language Models) พฤติกรรมการคำนวณของซอฟต์แวร์ได้เปลี่ยนไปอย่างสิ้นเชิง อัลกอริทึมเอไอไม่ได้ต้องการตัวประมวลผลที่เก่งตรรกะซับซ้อน แต่ต้องการการคำนวณคณิตศาสตร์พื้นฐานอย่าง พีชคณิตเชิงเส้น (Linear Algebra) โดยเฉพาะการคูณและการบวกสะสมค่าของเมทริกซ์และเวกเตอร์ (Matrix Multiplication) ปริมาณมหาศาลพร้อมๆ กัน พฤติกรรมนี้เรียกว่า การประมวลผลแบบขนานขนาดใหญ่ (Massively Parallel Processing) ซึ่งซีพียูธรรมดาที่มีแกนการประมวลผลเพียงไม่กี่แกน (โดยทั่วไป 4 ถึง 24 แกน) ไม่สามารถตอบสนองความต้องการนี้ได้ทันเวลา ส่งผลให้เกิดคอขวดทั้งในแง่ของเวลาและความร้อนสะสม อุตสาหกรรมจึงจำเป็นต้องพัฒนาสถาปัตยกรรมฮาร์ดแวร์เฉพาะทางที่เรียกว่า เอไอชิป ขึ้นมาเพื่อทำหน้าที่นี้โดยเฉพาะ

เพื่อความเข้าใจที่ชัดเจน เราสามารถจำแนกเหตุผลความจำเป็นและความแตกต่างเชิงสถาปัตยกรรมที่ทำให้ต้องเปลี่ยนมาใช้ เอไอชิป ออกเป็น 3 มิติหลัก ดังนี้

1. ความหนาแน่นของหน่วยคำนวณ (ALU Density)

ภายในชิปประมวลผลจะมีหน่วยที่ทำหน้าที่คำนวณคณิตศาสตร์เรียกว่าหน่วยคำนวณหรือเอแอลยู (Arithmetic Logic Unit, ALU) หากเปรียบเทียบโครงสร้างภายใน ซีพียูจะสละพื้นที่ส่วนใหญ่บนเวเฟอร์ให้กับหน่วยควบคุมและหน่วยความจำแคชขนาดใหญ่ ทำให้มีพื้นที่เหลือให้หน่วยคำนวณเพียงไม่กี่ตัวต่อหนึ่งแกน ในทางกลับกัน เอไอชิป (รวมถึงจีพียูที่ถูกนำมาประยุกต์ใช้) จะตัดหน่วยควบคุมและแคชที่ไม่จำเป็นออกไป แล้วถมพื้นที่ว่างทั้งหมดบนชิปด้วยหน่วยคำนวณขนาดเล็กนับพันนับหมื่นตัว การจัดวางผังเมืองทรานซิสเตอร์แบบนี้ทำให้ เอไอชิป สามารถคำนวณคณิตศาสตร์พร้อมกันในเวลาเดียว (Parallelism) ได้มากกว่าซีพียูหลายร้อยเท่าตัว

2. รูปแบบชุดคำสั่งคำนวณขนาน (SIMD และ Matrix Core)

ซีพียูแบบดั้งเดิมคำนวณในลักษณะคำสั่งเดียวข้อมูลเดียว (Single Instruction Single Data, SISD) คือหนึ่งคำสั่งจัดการหนึ่งข้อมูล แต่ เอไอชิป จะใช้สถาปัตยกรรมแบบคำสั่งเดียวหลายข้อมูล (Single Instruction Multiple Data, SIMD) หรือพัฒนาไปถึงขั้นแกนเมทริกซ์หรือเทนเซอร์ (Matrix/Tensor Core) ซึ่งเป็นบล็อกคำนวณฮาร์ดแวร์ที่ถูกสร้างมาเพื่อทำคำสั่งการ “คูณแล้วบวกสะสมผลรวม” ของข้อมูลทั้งแถวและคอลัมน์ในเมทริกซ์ให้เสร็จสิ้นภายในสัญญาณนาฬิกาเดียว (Single Clock Cycle) การเปลี่ยนผ่านชุดคำสั่งนี้ช่วยลดภาระการดึงคำสั่งซ้ำ ๆ และเพิ่มประสิทธิภาพในการประมวลผลโครงข่ายประสาทเทียมได้อย่างก้าวกระโดด

3. ประสิทธิภาพต่อพลังงานที่ได้รับ (Performance-per-Watt)

เมื่อโมเดลเอไอขยายใหญ่ขึ้นจนมีพารามิเตอร์นับแสนล้านตัว หากเรายังคงพยายามใช้หน่วยประมวลผลกลางในการประมวลผล เราจะต้องใช้ชิปนับหมื่นตัวในศูนย์ข้อมูล (Data Center) ซึ่งจะส่งผลให้เกิดการใช้กระแสไฟฟ้าในปริมาณมหาศาล

จนโรงไฟฟ้าไม่สามารถจ่ายไฟได้ทัน และระบบจะพังทลายลงจากความร้อนสะสม เอไอชิปจึงเข้ามาตอบโจทย์ในด้านเศรษฐศาสตร์ เนื่องจากชิปนี้ถูกปรับแต่งวงจรภายในให้ทำงานคณิตศาสตร์ของเอไอ ได้อย่างมีประสิทธิภาพสูงสุด ทำให้ได้ผลลัพธ์การคำนวณที่เร็วกว่า แต่บริโภคพลังงานไฟฟ้าน้อยกว่าการใช้ซีพียูอย่างมหาศาล

ด้วยเหตุผลเหล่านี้ เอไอชิปจึงไม่ได้เป็นเพียงแค่ทางเลือกเสริม แต่กลายเป็น “โครงสร้างพื้นฐานภาคบังคับ” ที่ขาดไม่ได้ในการขับเคลื่อน รันระบบ และฝึกฝนปัญญาประดิษฐ์ในยุคปัจจุบัน

3.2 เทคโนโลยีเบื้องหลังชิประดับโลก (Technologies Behind World-Class AI Chips)

เมื่อความต้องการพลังคำนวณของอัลกอริทึมปัญญาประดิษฐ์พุ่งทะยานเกินกว่าที่หน่วยประมวลผลทั่วไปจะรับไหว อุตสาหกรรมเทคโนโลยีโลกจึงเกิดการแข่งขันพัฒนาสถาปัตยกรรมเฉพาะทางเพื่อก้าวขึ้นมาเป็นผู้นำ ในปัจจุบัน เทคโนโลยีเบื้องหลัง เอไอชิป ระดับโลกสามารถแบ่งออกเป็น 3 ตระกูลหลักตามโครงสร้างทางวิศวกรรมและรูปแบบการนำไปใช้งาน ได้แก่ จีพียู ทีพียู และ เอ็นพียู ซึ่งต่างก็มีจุดเด่นและกลไกภายในที่แตกต่างกันอย่างสิ้นเชิง

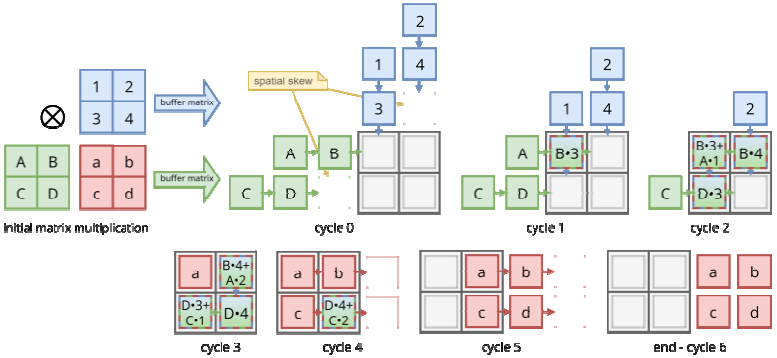
1. จีพียู (Graphics Processing Unit, GPU)

ชิปประมวลผลกราฟฟิก หรือ จีพียู เปลี่ยนผ่านสู่อุตสาหกรรมเอไอ เดิมที จีพียูถูกสร้างขึ้นมาเพื่อประมวลผลภาพกราฟิกสามมิติและวิดีโอเกม แต่เนื่องจากโครงสร้างภายในประกอบด้วยคอร์ประมวลผลขนาดเล็กจำนวนหลายพันคอร์ที่ทำงานขนานพร้อมกันได้ดีมาก นักวิจัยจึงพบว่าสถาปัตยกรรมนี้เหมาะอย่างยิ่งสำหรับการคูณเมทริกซ์ในโครงข่ายประสาทเทียม ผู้นำตลาดโลกอย่างเด็ดขาดในกลุ่มนี้คือ เอ็นวีเดีย ซึ่งก้าวขึ้นมาครองตลาดเอไอผ่านการพัฒนาฮาร์ดแวร์ตระกูลสถาปัตยกรรมล้ำ

หน้า เช่น ชิปสถาปัตยกรรมแบล็คเวล (Blackwell) ความลับทางเทคโนโลยีของจีพียูยุคปัจจุบันคือการใส่บล็อกวงจรเฉพาะทางที่เรียกว่า “เทนเซอร์คอร์” (Tensor Core) เข้าไปในตัวชิป บล็อกนี้ทำหน้าที่คำนวณคณิตศาสตร์แบบลดทอนความแม่นยำ (Reduced Precision) เช่น การเปลี่ยนรูปแบบของข้อมูลตัวเลขเพื่อเร่งความเร็วในการประมวลผลให้สูงขึ้นหลายเท่าโดยไม่สูญเสียความแม่นยำของโมเดล และการใช้หน่วยความจำแบนด์วิดท์สูง ร่วมกับช่องทางเชื่อมต่อความเร็วสูงอย่าง NVLink เพื่อมัดรวมจีพียูนับหมื่นตัวให้ทำงานร่วมกันเสมือนเป็นชิปยักษ์ตัวเดียวในศูนย์ข้อมูล

2. ทีพียู (Tensor Processing Unit, TPU)

ทีพียู ถูกกำหนดให้เป็น ชิปสถาปัตยกรรมเฉพาะทางสำหรับคลาวด์ จากกูเกิ้ล แตกต่างจากจีพียูที่มีรากฐานมาจากชิปกราฟิก ทีพียู ของกูเกิ้ลเป็นชิปประเภท เอสิซ (Application-Specific Integrated Circuit, ASIC) ที่ถูกออกแบบขึ้นมาใหม่ทั้งหมดเพื่อทำงานด้านการเรียนรู้โดยเครื่องโดยเฉพาะ ชิปประเภทนี้เปิดตัวครั้งแรกในปี ค.ศ. 2016 และปัจจุบันได้พัฒนาต่อเนื่องมาหลายยุค โดยยุคล่าสุด หัวใจหลักของทีพียูคือสถาปัตยกรรมคำนวณแบบ ซิสโตลิกอะเรย์ (Systolic Array) ซึ่งเลียนแบบระบบสูบฉีดเลือดของหัวใจ แทนที่จะส่งข้อมูลกลับไปกลับมาระหว่างหน่วยประมวลผลและหน่วยความจำหลังจากเสร็จสิ้นทุกๆ การคำนวณ (ซึ่งทำให้สิ้นเปลืองพลังงานและเกิดคอขวด) สถาปัตยกรรมซิสโตลิกจะยอมให้ข้อมูลไหลผ่านตารางเอแอลยูที่เชื่อมต่อกันโดยตรง ข้อมูลคูณและบวกจะถูกส่งต่อไปยังคอร์ข้างๆ ทันทีเป็นทอด ๆ เทคโนโลยีนี้ช่วยให้เครื่องมัลติเพลเยอร์ในหน่วยคำนวณเมทริกซ์ สามารถประมวลผลชุดข้อมูลนับแสนชุดได้พร้อมกันภายในหนึ่งสัญญาณนาฬิกา โดยปัจจุบันมีเชื่อมต่อชิปทีพียูเหล่านี้เข้าด้วยกันเป็นระบบซูเปอร์คอมพิวเตอร์ขนาดใหญ่ผ่านเครือข่ายวงจรสวิตช์ทางแสงเพื่อใช้ฝึกฝนโมเดลตระกูลเจมิไน

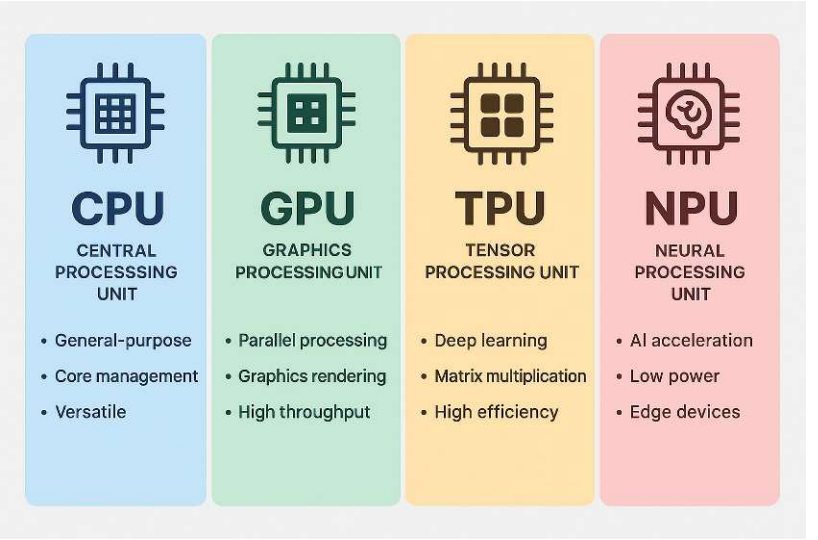


รูปที่ 20 อัลกอริทึมอาเรย์แบบซิสโตลิกที่ทำการสะสมค่าผลลัพธ์ภายในหน่วยประมวลผลข้อมูล (ที่มา https://en.wikipedia.org/wiki/Systolic_array)

3. เอ็นพียู (Neural Processing Unit, NPU) – พลังสมองกลเอไอในอุปกรณ์พกพา

ในขณะที่ จีพียู และ ทีพียู ขับเคลื่อนสงครามเอไออยู่บนระบบคลาวด์และศูนย์ข้อมูลขนาดใหญ่ เอ็นพียู คือตัวแทนของ เอไอชิป ที่ถูกย่อส่วนลงมาฝังอยู่ในอุปกรณ์พกพาในชีวิตประจำวันของเรา เช่น สมาร์ตโฟน แท็บเล็ต และคอมพิวเตอร์ยุคใหม่ (เช่น ชิปแอปเปิ้ลซิลิคอน (Apple Silicon) ควอลคอม สแนปดราก้อน (Qualcomm Snapdragon) และหน่วยประมวลผลตระกูลอินเทลรุ่นล่าสุด) เป้าหมายสูงสุดของเอ็นพียู ไม่ใช่ความแรงดิบระดับซูเปอร์คอมพิวเตอร์ แต่เป็นประสิทธิภาพต่อพลังงาน (Efficiency) และสถาปัตยกรรมที่ประหยัดไฟขั้นสุด เอ็นพียู ถูกออกแบบมาให้ทำหน้าที่รันโมเดลเอไอขนาดเล็กที่อยู่บนเครื่อง (On-Device AI) เช่น ระบบรู้จำใบหน้า (Face ID) การประมวลผลภาพถ่ายแบบเวลาจริง (Real-Time Computational Photography) และการถอดความเสียงเป็นตัวอักษร สถาปัตยกรรมของ เอ็นพียู จะมุ่งเน้นสตรีมข้อมูลประเภทจำนวนเต็มความละเอียดต่ำเป็นหลัก ซึ่งช่วยลดการใช้

พื้นที่แคชและการเคลื่อนย้ายข้อมูล ส่งผลให้อุปกรณ์สามารถทำงานประมวลผลเอไอได้ด้วยความเร็วหลายสิบล้านล้านปฏิบัติการต่อวินาที โดยที่ไม่ทำให้แบตเตอรี่โทรศัพท์มือถือหมดลงภายในไม่กี่นาที



รูปที่ 21 รูปเปรียบเทียบคุณสมบัติของซีพียู จีพียู ทีพียู และ เอ็นพียู (ที่มา https://resources.l-p.com/knowledge-center/cpu-vs-gpu-vs-tpu-vs-npu-architecture-comparison-explained)

3.3 อนาคตของ เอไอชิป (Future of AI Chips)

แม้ว่าในปัจจุบัน เอไอชิป สถาปัตยกรรมล่าสุด เช่น จีพียู ประสิทธิภาพสูง หรือ เทนเซอร์คอร์ที่มีระบบหน่วยความจำแบนด์วิดท์สูง จะสามารถขับเคลื่อนโมเดลปัญญาประดิษฐ์ระดับล้านล้านพารามิเตอร์ได้อย่างน่าทึ่ง ทว่าในมุมมองของนักออกแบบฮาร์ดแวร์และวิศวกรระบบ เทคโนโลยีเหล่านี้กำลังเดินหน้าเข้าสู่ “ทางตัน” ในระยะเวลาอันใกล้ ปัญหาหลักเกิดจากแนวคิดที่อุตสาหกรรมคอมพิวเตอร์ติดหล่มมานานกว่าครึ่งศตวรรษ นั่นคือ สถาปัตยกรรมฟอนนอยมันน์ (Von Neumann Architecture) ซึ่งกำหนดให้หน่วยประมวลผลและหน่วยความจำ แยกออกจากกันเด็ดขาดโดยมีบัสเป็นช่องทางเชื่อมตรงกลางในงานประมวลผลเอไอ ข้อมูลค่าถ่วงน้ำหนัก นับแสนล้านตัวต้องถูกส่งผ่านบัสเหล่านี้ไปกลับตลอดเวลา ส่งผลให้เกิด คอขวดฟอนนอยมันน์ (ดูบทที่ 1 หัวข้อ 1.3) ที่ต่อให้หน่วยประมวลผลคำนวณได้เร็วขึ้นเพียงใด แต่ก็ต้องนั่งรอข้อมูลจากหน่วยความจำอยู่ดี ยิ่งไปกว่านั้น พลังงานไฟฟ้ากว่า 80% บนชิปปัจจุบันถูกสูญเสียไปกับความต้านทานไฟฟ้าและความร้อนในขณะเคลื่อนย้ายข้อมูลนี้เอง จนทำให้เกิดวิกฤตการณ์พลังงานและขีดจำกัดด้านความร้อนในระบบคลาวด์ศูนย์ข้อมูลทั่วโลกเพื่อตอบสนองต่อวิกฤตนี้ วงการวิศวกรรม (ทั้งวิศวกรรมไฟฟ้า วิศวกรรมคอมพิวเตอร์ และวิศวกรรมเซมิคอนดักเตอร์) จึงกำลังปฏิวัติสถาปัตยกรรมเอไอชิป ไปสู่กระบวนทัศน์ใหม่ 3 แนวทางหลักที่จะพลิกโฉมอนาคตอย่างสิ้นเชิง

1. การคำนวณแบบเลียนแบบสมองมนุษย์ (Neuromorphic Computing)

มนุษย์เรามีสมองที่มีเซลล์ประสาท (Neurons) ประมาณ 8.6 หมื่นล้านเซลล์ และจุดประสานประสาท (Synapses) นับร้อยล้านล้านจุด ซึ่งสามารถคิดวิเคราะห์ จำแนกใบหน้า และเรียนรู้สิ่งซับซ้อนได้โดยบริโภคพลังงานเทียบเท่ากับหลอดไฟขนาดเล็กเพียง 20 วัตต์เท่านั้น ในขณะที่ซูเปอร์คอมพิวเตอร์ที่รันโมเดลเอไอระดับ

เดียวกันต้องกินไฟหลายเมกะวัตต์ (เทียบเท่าไฟฟ้าที่ใช้ในเมืองทั้งเมือง) แนวคิดวิศวกรรมนิวรอมอร์ฟิก (Neuromorphic Engineering) จึงถือกำเนิดขึ้นเพื่อลอกเลียนแบบความอัจฉริยะนี้ชิปนิวรอมอร์ฟิกจะสลัดทิ้งระบบดิจิทัลแบบเดิมที่ต้องพึ่งพาสัญญาณนาฬิกา (Clock-driven) ซึ่งคอยส่งสัญญาณกระตุ้นทรานซิสเตอร์ทุกตัวให้ทำงานพร้อมกันตลอดเวลาแม้ว่าจะไม่มีข้อมูลวิ่งเข้ามา เปลี่ยนไปสู่สถาปัตยกรรมโครงข่ายประสาทแบบส่งกระแสไฟฟ้า หรือ Spiking Neural Networks (SNNs) โครงสร้างภายในจะถูกออกแบบให้ทรานซิสเตอร์และตัวเก็บประจุทำหน้าที่จำลองพฤติกรรมเซลล์ประสาททางชีวภาพ โดยสวิตช์ภายในชิปจะอยู่ในสภาวะหลับใหลและไม่บริโภคกระแสไฟฟ้าเลย จนกว่าจะมีสัญญาณข้อมูลหรือ “กระแสไฟฟ้าสไปก์” (Spike) วิ่งเข้ามากระตุ้นจนถึงระดับขีดเริ่มเปลี่ยน (Threshold) เท่านั้นกลไกนี้เรียกว่า การคำนวณตามเหตุการณ์ (Event-driven Computing) ซึ่งนอกจากจะช่วยลดการสิ้นเปลืองพลังงานไฟฟ้าลงได้หลายพันเท่าแล้ว ยังเปิดโอกาสให้ชิปสามารถทำหน้าที่เรียนรู้และปรับตัวได้เองตลอดเวลาบนฮาร์ดแวร์ (On-chip Learning) ผ่านกลไกที่เลียนแบบความยืดหยุ่นของสมอง ทำให้ชิปประเภทนี้เหมาะอย่างยิ่งสำหรับระบบสมองกลฝังตัว หุ่นยนต์อัตโนมัติล้ำยุค และโดรนสำรวจที่ต้องประมวลผลข้อมูลปริมาณมากในพื้นที่จำกัดโดยไม่มีสายปลั๊กไฟเลี้ยง ตัวอย่างความสำเร็จในปัจจุบัน เช่น ชิปทดลองระบบประสาทตระกูลโลอิฮิของอินเทล (Intel Loihi) และชิปทรูนอร์ทของไอบีเอ็ม (IBM TrueNorth)

2. ชิปประมวลผลเอไอด้วยแสง (Optical / Photonic AI Chips)

เมื่ออิเล็กตรอนซึ่งเป็นประจุไฟฟ้าวิ่งผ่านเส้นลวดโลหะทองแดงหรืออลูมิเนียมขนาดจิ๋วระดับนาโนเมตรภายในชิป มันจะปะทะกับโครงสร้างอะตอมของโลหะ เกิดเป็นความต้านทานไฟฟ้าและเปลี่ยนพลังงานบางส่วนกลายเป็นความร้อนสะสม ยิ่งต้องการส่งข้อมูลให้เร็วขึ้น ความร้อนก็ยิ่งพุ่งสูงขึ้นเป็นเงาตามตัว อนาคตของ เอไอชิป

จึงมุ่งหวังที่จะปลดแอกตัวเองออกจากพันธนาการของอิเล็กตรอน โดยการเปลี่ยนมาใช้ โฟตอน (Photon) หรืออนุภาคของแสง วิ่งในสารกึ่งตัวนำแทน ซึ่งเรียกศาสตร์นี้ว่า ซิลิคอนโฟโตนิกส์ (Silicon Photonics)

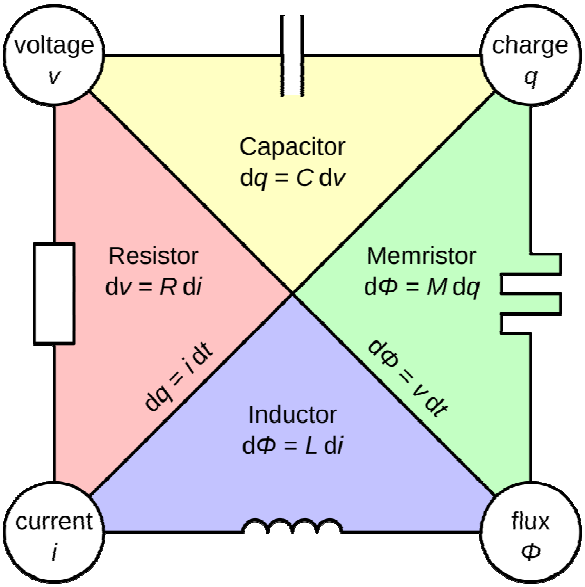
ชิปประมวลผลเอไอด้วยแสง (Optical AI Chip) จะเปลี่ยนผังวงจรทรานซิสเตอร์แบบเดิมให้กลายเป็นเครือข่ายของท่อนำคลื่นแสงขนาดไมโครเมตร (Optical Waveguides) ตัวแยกสัญญาณแสง (Splitters) และตัวเลื่อนเฟสแสง (Phase Shifters) ในการคำนวณคณิตศาสตร์พีชคณิตเชิงเส้นที่เป็นหัวใจของเอไอ ชิปประมวลผลด้วยแสงนี้จะใช้คุณสมบัติทางกายภาพของคลื่นแสงในการหาคำตอบทันที เช่น การนำลำแสงสองเส้นที่มีความเข้มข้นต่างกันมาแทรกสอดกัน (Interference) เพื่อแทนค่าการบวกและการคูณเมทริกซ์

เนื่องจากโฟตอนเคลื่อนที่ด้วยความเร็วแสงและไม่มีมวล มันจึงไม่เกิดแรงต้านทานทางไฟฟ้าและไม่สร้างความร้อนสะสมใดๆ ทำให้เราสามารถคำนวณคูณเมทริกซ์ขนาดแสนล้านชุดเสร็จสิ้นได้ในพริบตาเดียวตั้งแต่วินาทีที่แสงเดินทางผ่านวงจร (คือมีเวลาแฝงเกือบเป็นศูนย์) และข้ามผ่านกระบวนการในระดับความเร็วเฟมโตวินาที ชิปแสงจึงเป็นความหวังสูงสุดในการทะลายกำแพงความเร็วสำหรับการฝึกฝนแบบจำลองภาษาขนาดใหญ่และโครงข่ายประสาทเทียมในอนาคตอันใกล้

3. การคำนวณในหน่วยความจำและการปฏิวัติด้วยเมมริสเตอร์ (In-Memory Computing & Memristor)

หากไม่สามารถเปลี่ยนไปใช้แสงหรือระบบเลียนแบบสมองได้ อีกหนึ่งทางรอดที่อุตสาหกรรมกำลังลงมือทำคือการทุบทำลายกำแพงที่กั้นระหว่างหน่วยประมวลผลและหน่วยความจำทิ้งไป แล้วหลอมรวมพวกมันเข้าเป็นเนื้อเดียวกัน เรียกว่า การคำนวณในหน่วยความจำ (In-Memory Computing) หัวใจหลักที่ขับเคลื่อนนวัตกรรมนี้คืออุปกรณ์ฟิสิกส์วัสดุศาสตร์ตัวที่ 4 ของวงจรอิเล็กทรอนิกส์ (ต่อจากตัว

ต้านทาน ตัวเก็บประจุ และตัวเหนี่ยวนำ) นั่นคือ เมมริสเตอร์ (Memristor) หรือตัวต้านทานที่สามารถจดจำอดีตได้ เมมริสเตอร์มีความสามารถพิเศษคือ มันสามารถเปลี่ยนค่าความต้านทานไฟฟ้าภายในตัวเองได้ตามปริมาณกระแสไฟที่เคยไหลผ่าน และที่สำคัญคือมันจะ “จดจำ” ค่าความต้านทานนั้นไว้ตลอดกาลแม้ว่าจะปิดสวิตช์ไฟเลี้ยงไปแล้วก็ตาม ซึ่งเป็นคุณสมบัติแบบหน่วยความจำแบบไม่ลบเลือน (Non-volatile)



รูปที่ 22 ภาพแสดงความสมมาตรเชิงแนวคิดของตัวต้านทาน ตัวเก็บประจุ ตัวเหนี่ยวนำ และตัวเมมริสเตอร์

(ที่มา https://en.wikipedia.org/wiki/Memristor)

วิศวกรได้นำเมมริสเตอร์เหล่านี้มาต่อกันเป็นตารางกริดขนาดใหญ่ เรียกว่า ครอสบาร์อาเรย์ (Crossbar Array) เพื่อใช้เก็บค่าถ่วงน้ำหนัก (Weights) ของโมเดลเอไอในรูปแบบของค่าความนำไฟฟ้า G เมื่อต้องการส่งข้อมูลนำเข้า เข้ามาคำนวณ เราจะป้อนข้อมูลนั้นในรูปแบบของแรงดันไฟฟ้า วิ่งเข้าไปในโครงสร้างตารางนี้ จากกฎทางฟิสิกส์พื้นฐานอย่าง กฎของโอห์ม I = V G) แรงดันไฟฟ้าที่วิ่งผ่านความนำไฟฟ้าจะกลายเป็นกระแสไฟฟ้า (I) ซึ่งทำหน้าที่แทนผลคูณคณิตศาสตร์โดยอัตโนมัติ และกระแสไฟฟ้าจากทุกเส้นทางจะไหลมารวมกันที่ปลายสายตาม กฎกระแสไฟฟ้าของเคอร์ชอฟฟ์ (Kirchhoff's Current Law) กลายเป็นผลรวมสะสมทันที กระบวนการนี้ทำให้ เอไอชิป สามารถคำนวณคำสั่งคูณและบวกจำนวนนับล้านตัวแปรเสร็จสิ้นได้ภายในพื้นที่เซลล์หน่วยความจำของตัวเองโดยตรง โดยไม่ต้องเปลืองพลังงานและเวลาในการลากข้อมูลข้ามสายบัสคอมพิวเตอร์ออกไปข้างนอกเลยแม้แต่น้อย ช่วยยกระดับประสิทธิภาพการประมวลผลต่อวัตต์ขึ้นจากเดิมหลายสิบเท่า

การเดินทางผ่านสถาปัตยกรรมของเอไอชิป ในบทนี้ ตั้งแต่รากฐานความต้องการคำนวณขนานขนาดใหญ่ของอัลกอริทึม การปะทะกันของยักษ์ใหญ่ซีพียู จีพียู ทีพียู และเอ็นพียู ไปจนถึงเทคโนโลยีพลิกโลกในอนาคตอย่างชิปเลียนแบบสมอง ชิปแสง และการประมวลในหน่วยความจำ สะท้อนให้เห็นว่ามนุษยชาติกำลังอยู่ในยุค “ตื่นทอง” ของพลังคำนวณ สมรภูมินี้ไม่ได้จำกัดอยู่เพียงแค่ในห้องวิจัยของบริษัทยักษ์ใหญ่ไม่กี่แห่งอีกต่อไป แต่ได้กลายเป็นวาระแห่งชาติและวัฒนธรรมร่วมสมัยที่ทุกประเทศต่างเร่งขับเคลื่อนเพื่อชี้ชะตาเอกราชทางเทคโนโลยี ตัวอย่างที่เด่นชัดที่สุดในปัจจุบันคือประเทศเกาหลีใต้ ที่ซึ่งอุตสาหกรรม เอไอชิป เติบโตอย่างบ้าคลั่งจนกลายเป็นแกนหลักขับเคลื่อนเศรษฐกิจ

มีการเปิดพื้นที่สร้างนวัตกรรมผ่านเวทีประกวดระดับชาติ เช่น งาน K-AI Semiconductor Growth Forum เพื่อผลักดันกลุ่มสตาร์ตอัปสายแฟบเลต ให้เร่ง

พัฒนาชิปสถาปัตยกรรมชิปล้ำหน้าออกมาแข่งขันในเวทีโลก คลื่นพลังความตื่นตัวระดับมหาชนนี้ ถึงขั้นส่งผลให้สายงานวิศวกรรม เอไอชิป ในเกาหลีใต้กลายเป็นอาชีพยอดนิยมระดับประเทศ ภาพการแข่งขันที่เกิดขึ้นนี้เป็นสิ่งยืนยันว่า เอไอชิป ไม่ใช่เพียงแค่เรื่องของแผงวงจรอิเล็กทรอนิกส์ แต่คือโครงสร้างพื้นฐานอันทรงทรงอิทธิพลที่จะกำหนดทิศทาง มอบอำนาจการคำนวณ และขับเคลื่อนสติปัญญาของโลกยุคถัดไปอย่างแท้จริง สำหรับผู้ที่สนใจเจาะลึกภาพรวมการแข่งขันระดับมหภาคในเอเชีย สามารถรับข้อมูลเพิ่มเติมได้จากวิดีโอจำนวนมากใน Youtube เช่น Can Samsung Overtake TSMC? Taiwan vs. South Korea AI Chip Rivalry Explained ซึ่งมีการวิเคราะห์เจาะลึกกลยุทธ์นโยบายรัฐบาล คอขวดห่วงโซ่อุปทานหน่วยความจำและความเชื่อมโยงระดับโลกที่กำลังขับเคลื่อนอุตสาหกรรมเอไอชิปของเกาหลีใต้ในปัจจุบันได้อย่างเห็นภาพชัดเจน

แหล่งข้อมูลเพิ่มเติมและเอกสารอ้างอิง

สถาปัตยกรรมและการคำนวณของอัลกอริทึมการเรียนรู้เชิงลึก:

- Wikipedia contributors. "Deep learning." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Deep_learning
- Wikipedia contributors. "Large language model." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Large_language_model

โครงสร้างทางสถาปัตยกรรมคอมพิวเตอร์และหน่วยเร่งความเร็ว (Hardware Accelerators):

- Wikipedia contributors. "Hardware acceleration." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Hardware_acceleration
- Wikipedia contributors. "Graphics processing unit" (Section on Tensor Cores & AI). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Graphics_processing_unit
- Wikipedia contributors. "Tensor Processing Unit" (Systolic Array Architecture). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Tensor_Processing_Unit
- Wikipedia contributors. "Neural processing unit." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Neural_processing_unit

ข้อจำกัดและเทคโนโลยีเอไอชิปแห่งอนาคต:

- Wikipedia contributors. "Von Neumann architecture" (Von Neumann Bottleneck section). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Von_Neumann_architecture

- Wikipedia contributors. "Neuromorphic engineering" & "Spiking neural network." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Neuromorphic_engineering

- Wikipedia contributors. "Silicon photonics." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Silicon_photonics

- Wikipedia contributors. "Memristor" & "In-memory processing." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Memristor

ข้อมูลอุตสาหกรรมและสถานการณ์ปัจจุบันในเกาหลีใต้ (K-AI Semiconductor)

- นโยบายและการขับเคลื่อน NPU ของรัฐบาลเกาหลีใต้: Ministry of Science and ICT (MSIT), Republic of Korea. "K-AI Semiconductor Growth Forum" (รายงานผลลัพธ์การประชุมและการประเมินดัชนีประสิทธิภาพ K-Perf)

- ChosunBiz / The Asia Business Daily / MK English News. (June 2026). Korea touts AI chip gains as exports hit $30 million and rivals Nvidia (รายงานข่าวการเปิดตัวชิปสถาปัตยกรรม NPU ของบริษัทสตาร์ตอัป Rebellions, FuriosaAI และ Mobilint ร่วมกับความร่วมมือทางธุรกิจกับ Samsung SDS และ SK Telecom).