บทที่ 5
หน่วยความจำ (MEMORY)

หน่วยความจำ (Computer memory

หน่วยความจำ คือ อุปกรณ์เก็บสถานะข้อมูลและชุดคำสั่ง เพื่อการประมวลผลของคอมพิวเตอร์ มีหลายประเภทโดยมีหลักการแบ่ง ดังนี้

- ตำแหน่งของหน่วยความจำ
- ความจุ
- หน่วยของการขนย้ายข้อมูล
- วิธีการเข้าถึงข้อมูล
- ประสิทธิภาพ
- ลักษณะทางกายภาพ

ตำแหน่งของหน่วยความจำ (Location)

มี 2 แบบ คือ หน่วยความจำภายใน และหน่วยความจำภายนอก

- **หน่วยความจำภายใน (Internal memory)** โดยทั่วไปจะหมายถึง หน่วยความจำหลัก (Main memory) ได้แก่ รอม (ROM) แรม (RAM) และแคช (Cache)
- **หน่วยความจำภายนอก (External memory)** หรือหน่วยความจำสำรอง (Secondary memory) หมายถึง สื่อบันทึกข้อมูลที่ไม่รวมอยู่บนเมนบอร์ด เช่น ฮาร์ดดิสก์ SSD Handy Drive สื่อจัดเก็บข้อมูลต่างๆ ซึ่งโปรเซสเซอร์สามารถเรียกใช้อุปกรณ์เหล่านี้ได้โดยผ่านหน่วยควบคุมไอโอ (I/O controller)

ความจุ (Capacity)

- **หน่วยความจำภายในหรือหน่วยความจำหลัก** ในปัจจุบันจะใช้หน่วยวัดความความจุเป็นไบต์ และอาจกำหนดขนาดของหน่วยวัดเป็นอย่างอื่นอีกได้เช่น เวิร์ด โดยข้อมูลขนาด 1 เวิร์ด อาจมีจำนวน 8, 16 หรือ 32 บิต ตามความต้องการของระบบปฏิบัติการ
- **หน่วยความจำภายนอก** มักใช้หน่วยวัดความจุเป็นไบต์ซึ่งแตกต่างกับการรับส่งข้อมูลซึ่งใช้หน่วยเป็นบิต

หน่วยของการขนย้ายข้อมูล (Unit of transfer)

- **หน่วยความจำภายใน**จะมีหน่วยของการขนย้ายข้อมูลเท่ากับจำนวนสายสัญญาณข้อมูลที่เข้าออกหน่วยความจำหลักหรือขนาดของบัสข้อมูล โดยปกติจะมีขนาดเท่ากับเวิร์ด แต่ในคอมพิวเตอร์บางระบบอาจมีความแตกต่างกันออกไป
- **หน่วยความจำภายนอก**จะมีการรับ/ส่งข้อมูลจำนวนมากครั้งละหลายเวิร์ด เรียกว่าบล็อก (Blocks)

วิธีการเข้าถึงข้อมูล (Method of accessing units of data)

- **การเข้าถึงแบบตามลำดับ (Sequential access)** ข้อมูลจะถูกจัดเก็บในรูปของระเบียน (Record) การเข้าถึงระเบียนจะเรียงตามลำดับก่อนหลัง เวลาที่ใช้ในการเข้าถึงจึงไม่แน่นอนขึ้นอยู่กับลำดับของระเบียนที่ต้องการเข้าถึง เช่น เทปแม่เหล็ก
- **การเข้าถึงโดยตรง (Direct access)** แต่ละระเบียนหรือบล็อกจะมีตำแหน่งที่ชัดเจนบนหน่วยความจำ ทำให้สามารถเข้าถึง ณ ตำแหน่งที่ระบุได้โดยตรง เวลาที่ใช้ในการเข้าถึงไม่แน่นอน เนื่องจากเวลาที่ใช้ในการค้นหาแต่ละตำแหน่งไม่เท่ากัน ขึ้นอยู่กับตำแหน่งปัจจุบันของหัวอ่านเขียนและตำแหน่งที่ต้องการเข้าถึง เช่น ดิสก์

วิธีการเข้าถึงข้อมูล (Method of accessing units of data)

- **การเข้าถึงแบบสุ่ม (Random access)** ตำแหน่งของหน่วยความจำประเภทนี้จะกำหนดไว้ชัดเจนโดยลักษณะทางกายภาพ เวลาในการเข้าถึง ณ ตำแหน่งใด ๆ จะเท่ากันเสมอ เช่น ROM, RAM หน่วยความจำแคชในบางระบบก็ใช้การเข้าถึงแบบสุ่มเช่นกัน
- **การเข้าถึงแบบร่วม (Associative)** การเข้าถึงข้อมูลแบบนี้คล้ายกับการเข้าถึงข้อมูลแบบสุ่ม แต่ใช้การเปรียบเทียบบิตข้อมูลในทุกเวิร์ดพร้อมกัน เช่น หน่วยความจำแคช

ประสิทธิภาพ (Performance)

- **เวลาที่ใช้ในการเข้าถึงข้อมูล (Access time หรือ Latency)** ในกรณีของแรมจะเป็นเวลาที่ใช้ในการอ่านหรือเขียนข้อมูล โดยจะเริ่มต้นนับจากเวลาที่ระบุตำแหน่งให้กับหน่วยความจำ จนกระทั่งถึงเวลาที่ข้อมูลพร้อมใช้ สำหรับหน่วยความจำประเภทอื่นเวลาในการเข้าถึงจะหมายถึง ระยะเวลาที่ใช้เคลื่อนย้ายกลไกในการอ่านเขียนข้อมูลไปยังตำแหน่งที่ต้องการ
- **วงรอบของหน่วยความจำ (Memory cycle time)** ใช้กับกรณีของแรมเท่านั้น โดยระยะเวลาหนึ่งวงรอบจะประกอบด้วย ระยะเวลาในการเข้าถึงหน่วยความจำ (Access time) บวกกับระยะเวลาอื่น ๆ ที่จำเป็นต้องใช้ เช่น ระยะเวลาที่รอให้สัญญาณต่าง ๆ เข้าสู่สถานะที่พร้อมใช้งาน ระยะเวลาที่รอสายสัญญาณเข้าสู่สถานะว่าง หรือระยะเวลาที่ใช้ในการสร้างสัญญาณข้อมูลขึ้นมาใหม่ ซึ่งระยะเวลาเหล่านี้มักจะเกี่ยวข้องกับบัสหลักของระบบ

ประสิทธิภาพ (Performance)

- **อัตราเร็วในการขนถ่ายข้อมูล (Transfer rate)** หมายถึง เวลาที่ใช้ในการรับส่งข้อมูลของหน่วยความจำ โดยมากจะมีหน่วยเป็นบิตต่อวินาที (bps – bit per second) กรณีของแรมระยะเวลานี้จะมีค่าเท่ากับ 1/(cycle time)

ลักษณะทางกายภาพ (Physical Type)

คุณสมบัติทางกายภาพของอุปกรณ์จัดเก็บข้อมูล (Physical characteristics of data storage) จะแบ่งตามสองคุณสมบัติหลัก ดังนี้

ความคงทนถาวรของข้อมูลที่จัดเก็บ

- **แบบชั่วคราว (Volatile)** ซึ่งข้อมูลจะสูญหายหากขาดกระแสไฟฟ้ามาหล่อเลี้ยง ได้แก่ แรม
- **แบบถาวร (Non-volatile)** จะเก็บข้อมูลไว้ได้ตลอดเวลาโดยไม่ต้องมีกระแสไฟฟ้ามาหล่อเลี้ยง เช่น รอม หน่วยความจำประเภทแม่เหล็ก และประเภทแสง

ลักษณะทางกายภาพ (Physical Type)

ความสามารถของการลบข้อมูลในหน่วยความจำ

- **แบบลบได้ (Erasable)** คือสามารถบันทึกลงได้หลาย ๆ ครั้ง เช่น แรม ดิสก์ เทป แม่เหล็ก
- **แบบลบไม่ได้ (Nonerasable)** เมื่อบันทึกข้อมูลลงไปแล้วจะไม่สามารถแก้ไขหรือลบได้สามารถอ่านได้อย่างเดียว เช่น รอม

ประเภทของหน่วยความจำหลัก

หน่วยความจำหลักของคอมพิวเตอร์แบ่งได้ 2 ประเภท คือ

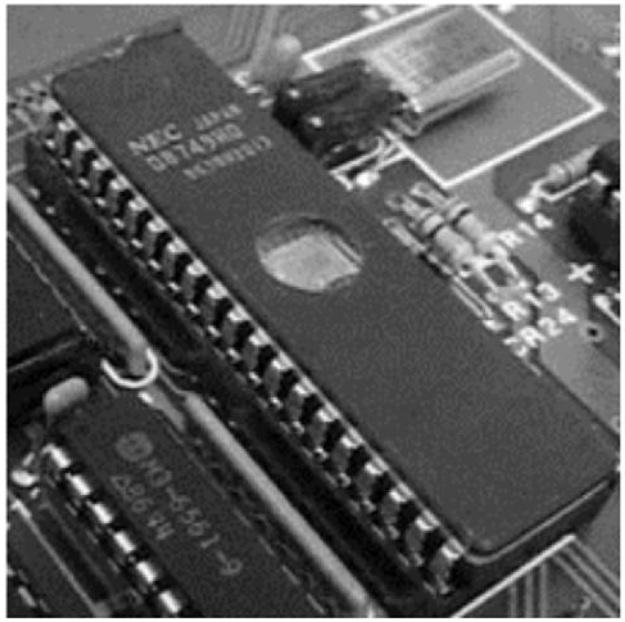
- ROM (Read-Only-Memory)
- RAM (Random Access Memory)

หน่วยความจำชนิดรอม (Read Only Memory : ROM)

- **รอม เรียกอีกอย่างหนึ่งว่า หน่วยความจำที่ข้อมูลไม่สูญหาย (Nonvolatile Memory)** เก็บข้อมูลหรือโปรแกรมไว้อย่างถาวร ไม่สามารถเปลี่ยนแปลงได้ สิ่งที่เก็บไว้ประกอบด้วยข้อมูลที่จำเป็นสำหรับการเริ่มสตาร์ทเครื่องคอมพิวเตอร์ และใช้เก็บโปรแกรมไบออส (BIOS) หรือโปรแกรมของคอมพิวเตอร์ที่ฝังอยู่ในฮาร์ดแวร์ของเครื่องที่ทำหน้าที่ตรวจสอบฮาร์ดแวร์และอุปกรณ์ต่าง ๆ ของคอมพิวเตอร์ สามารถแบ่งได้หลายชนิด ได้แก่
  - PROM (Programmable Read-only Memory)
  - EPROM (Erasable Programmable Read-only Memory)
  - EAROM (Electrically Alterable Read-only Memory)

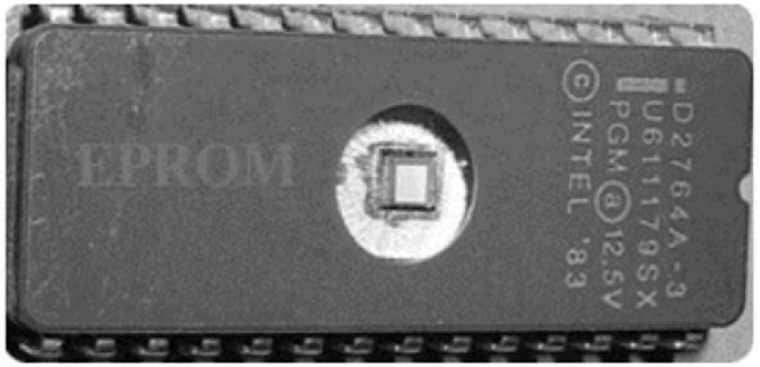
พีรอม PROM (Programmable Read-only Memory)

- ข้อมูลที่ต้องการโปรแกรมจะถูกโปรแกรมโดยผู้ใช้เอง โดยป้อนพัลส์แรงดันสูง (High Voltage Pulse) ทำให้ Metal Stripe หรือ Poly crystalline silicon ที่อยู่ในตัว IC ขาดออกจากกัน ทำให้เกิดเป็นลอจิก 1 หรือ 0 ตามตำแหน่งที่กำหนดใน หน่วยความจำนั้น ๆ เมื่อ PROM ถูกโปรแกรมแล้วข้อมูลภายในจะไม่สามารถ เปลี่ยนแปลงได้อีก มีความเร็วสูงกว่ารอมชนิดอื่น



อีพีรอม EPROM (Erasable Programmable Read-only Memory)

- ข้อมูลจะถูกโปรแกรมโดยผู้ใช้โดยการให้สัญญาณที่มีแรงดันสูง (High Voltage Signal) ผ่านเข้าไปในตัว EPROM ซึ่งเป็นวิธีเดียวกับที่ใช้ใน PROM แต่ข้อมูลที่อยู่ใน EPROM สามารถเปลี่ยนแปลงได้ โดยการลบข้อมูลเดิมที่อยู่ใน EPROM ออกก่อนแล้วค่อยโปรแกรมเข้าไปใหม่ การลบข้อมูลนั้นทำได้โดยการฉายแสงอัลตราไวโอเลตเข้าไปในตัว IC โดยผ่านทางกระจกใสที่อยู่บนตัว IC เมื่อฉายแสงประมาณ 5-10 นาที ข้อมูลที่อยู่ภายในก็จะถูกลบทิ้ง ซึ่งช่วงเวลาที่ฉายแสงนี้สามารถดูได้จากข้อมูลที่กำหนดมากับตัว EPROM เหมาะสำหรับงานของระบบที่ต้องปรับปรุงแก้ไขข้อมูลได้



อีเอรอม EAROM (Electrically Alterable Read-only Memory)

- EAROM หรือ EEPROM (Electrical Erasable EPROM) สามารถใช้ไฟฟ้าในการลบข้อมูลในรอมเพื่อเขียนใหม่โดยใช้เวลาน้อยกว่า EPROM ซึ่ง EAROM ใช้เทคโนโลยีแบบ NMOS การใช้งานโดยทั่วไปจะใช้ EPROM เพราะสามารถหามาใช้และทดลองได้ง่าย มีราคาถูก วงจรต่อง่ายไม่ยุ่งยาก และสามารถเปลี่ยนแปลงโปรแกรมได้ นอกจากระบบที่ทำการค้าเป็นจำนวนมากจึงจะใช้รอมประเภทโปรแกรมสำเร็จ

แรม (Random Access Memory)

- RAM แบ่งตามลักษณะการผลิตออกเป็น 2 ประเภท คือ
- **DRAM** ประกอบด้วยตัวเก็บประจุ (capacitors) ที่สามารถเก็บข้อมูลได้ระยะเวลาหนึ่ง ดังนั้นจึงต้องมีการรีเฟรชใหม่ทุก ๆ ช่วงเวลาเสมอเพื่อป้องกันข้อมูลสูญหาย มีราคาถูก ออกแบบและผลิตได้ง่าย
- **SRAM** ประกอบด้วยวงจรดี ฟลิปฟลอป (D flip-flop) ที่สามารถคงสถานะของสัญญาณไว้ได้จนกว่าจะมีการเปลี่ยนแปลงสัญญาณจึงไม่ต้องรีเฟรช มีความเร็วสูงและมีราคาสูง นิยมใช้ผลิตหน่วยความจำแคช

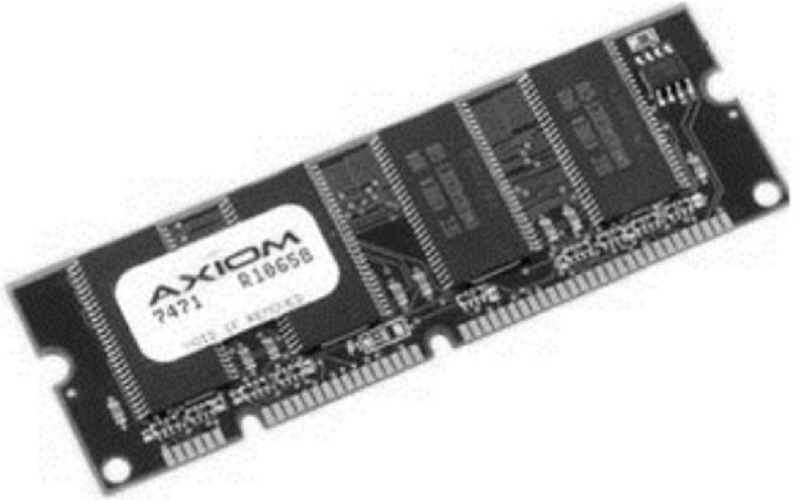
ชนิดของแรม

- หน่วยความจำแรมได้รับการพัฒนาอย่างต่อเนื่อง เพื่อให้ทำงานทันกับความเร็วของซีพียู สามารถจำแนกชนิดของแรมที่ใช้กันตั้งแต่อดีตจนถึงปัจจุบันได้ดังนี้
  - FPM DRAM (Fast Page Mode DRAM)
  - EDO RAM (Extended Data Output RAM)
  - BEDO DRAM (Burst EDO RAM)
  - SDRAM (Synchronous DRAM)
  - DDR SDRAM (SDRAM II)
  - RDRAM (Rambus DRAM)
  - SGRAM (Synchronous Graphic RAM)
  - Video Ram (VRAM)
  - Windows RAM (WRAM)

FPM DRAM (Fast Page Mode DRAM)

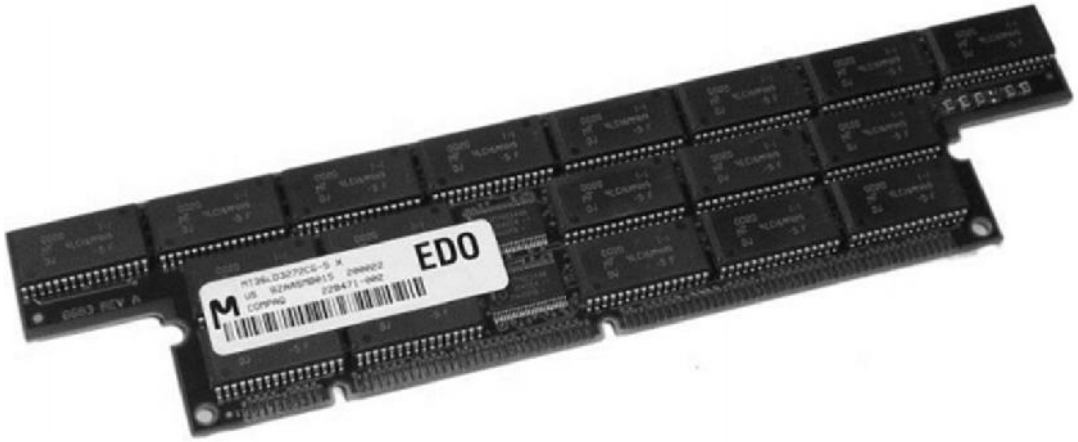
เป็นแรมรุ่นเก่าแก่ที่ใช้เทคโนโลยีที่ถูกผลิตขึ้นมาหลายปีแล้ว พบได้ในเครื่องรุ่น 286 มีโมดูลแบบ SIMM ขนาด 30 ขา และ 72 ขา ซึ่งใช้กับเครื่องรุ่นเก่า ในระยะแรกแรมแบบโมดูล FPM ออกมามีขนาด 2 4 8 16 และ 32 MB ความเร็ว 50 และ 70 ns การใช้แรมแบบนี้ต้องใช้แรมที่มีความเร็วเดียวกันเท่านั้น ไม่สามารถใช้แรมที่มีความเร็วต่างกันบนเมนบอร์ดเดียวกันได้

FPM เหมือนกับ DRAM เพียงแต่ว่าลดช่วงเวลาการหน่วงเวลาขณะเข้าถึงข้อมูลลง ทำให้เข้าถึงข้อมูลได้เร็วกว่า DRAM ในระบบ 32 bit จะมีอัตราการส่งถ่ายข้อมูล 100 MB ต่อวินาที ส่วนระบบ 64 bit จะมีอัตราการส่งถ่ายข้อมูลที่ 200 MB ต่อวินาที ปัจจุบันนี้แรมชนิดนี้แทบจะหมดไปจากท้องตลาดแล้ว



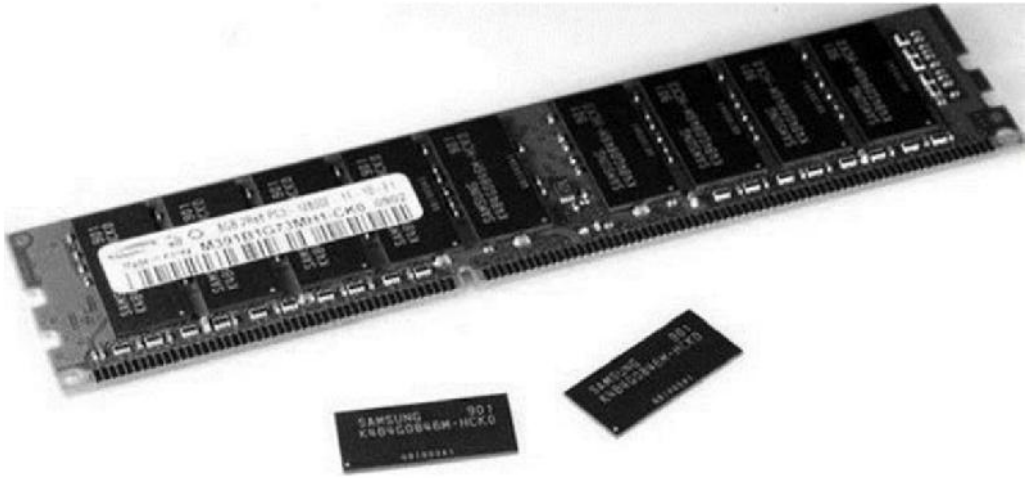
EDO RAM (Extended Data Output RAM)

ใช้ในเครื่องตั้งแต่รุ่น Pentium ขึ้นไป เป็นแรมที่มีความเร็วสูงกว่าแรมชนิด FPM โดยมีความเร็วอยู่ระหว่าง 70-50 ns (ตัวเลขยิ่งน้อยยิ่งเร็ว) EDO RAM ทำงานได้ดีที่ความเร็ว 66 MHz ขนาด 32 bit มีขาอยู่ 72 ขา มีโมดูลแบบ SIMM นอกจากนี้ EDO RAM ถูกออกแบบให้เป็นแบบ ECC (Error Correcting Code) ซึ่งมีคุณสมบัติที่สามารถตรวจสอบข้อผิดพลาดในการทำงานของแรมได้ แต่เมื่อความเร็วในการทำงานของเครื่องคอมพิวเตอร์สูงขึ้นโดยเฉพาะเมื่อความเร็วของบัสระบบอยู่ที่ระดับสูงกว่า 66 MHz ขึ้นไปหน่วยความจำแบบ EDO จึงอาจกลายเป็นตัวหน่วงให้ความเร็วในการทำงานของคอมพิวเตอร์ เนื่องจากซีพียูต้องเสียเวลาในการรอข้อมูลจากหน่วยความจำ



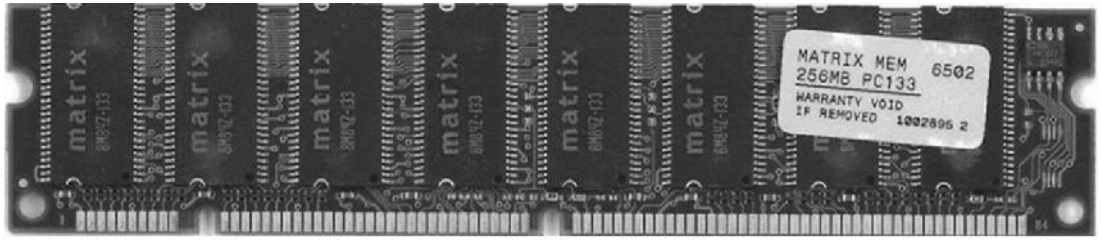
BEDO DRAM (Burst EDO RAM)

BEDO ได้เพิ่มความสามารถขึ้นมาจาก EDO RAM เดิมคือ Burst Mode โดยหลังจากที่ได้แอดเดรสแรกที่ต้องการแล้วมันจะทำการ generate อีก 3 แอดเดรสขึ้นมาทันทีภายใน 1 สัญญาณนาฬิกา ดังนั้นจึงตัดช่วงเวลาในการรับแอดเดรสต่อไป BEDO ไม่เป็นที่แพร่หลายและได้รับความนิยมเพียงระยะเวลาสั้น ๆ เนื่องจากทาง Intel ตัดสินใจใช้ SRAM แทน และไม่ได้ใช้ BEDO เป็นส่วนประกอบในการพัฒนาชิปเซตของตน ทำให้บริษัทผู้ผลิตต่าง ๆ หันมาพัฒนา SRAM แทน



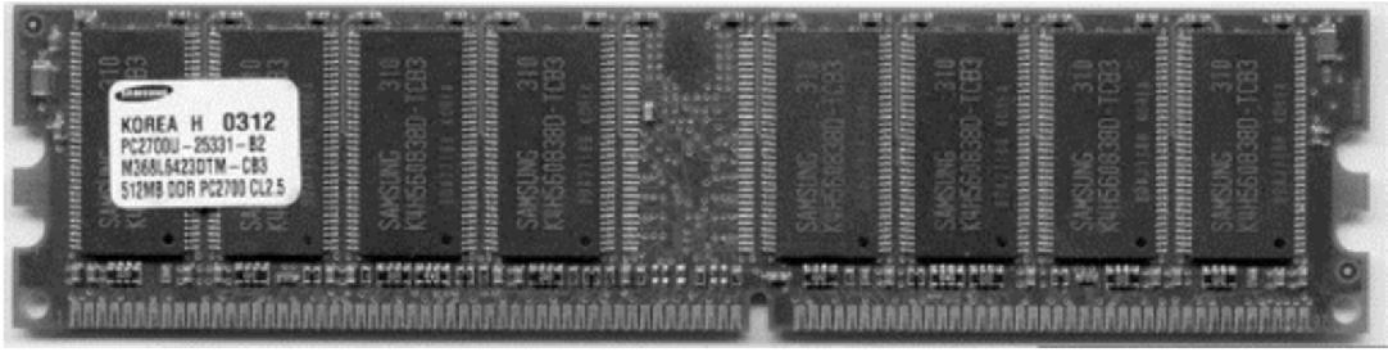
SDRAM (Synchronous DRAM)

จะแตกต่างจาก DRAM เดิมโดยจะทำงานสอดคล้องกับสัญญาณนาฬิกา ซึ่งมีความเร็วพอกันกับ BEDO RAM แต่ว่ามันสามารถทำงานได้ที่ 100 MHz หรือมากกว่า และมีอัตราการส่งถ่ายข้อมูลสูงสุดที่ 528 MB ต่อวินาที



DDR SDRAM (SDRAM II)

DDR SDRAM เป็นแรมที่แยกมาจาก SDRAM โดยจุดที่ต่างกันหลัก ๆ ของแรมทั้งสองชนิดนี้ คือ DDR SDRAM สามารถใช้งานได้ทั้งขาขึ้นและขาลงของสัญญาณนาฬิกาเพื่อส่งถ่ายข้อมูล ทำให้อัตราการส่งถ่ายเพิ่มขึ้นได้ถึงหนึ่งเท่าตัว โดยอัตราการส่งถ่ายข้อมูลสูงสุดอยู่ที่ 1 GB ต่อวินาที

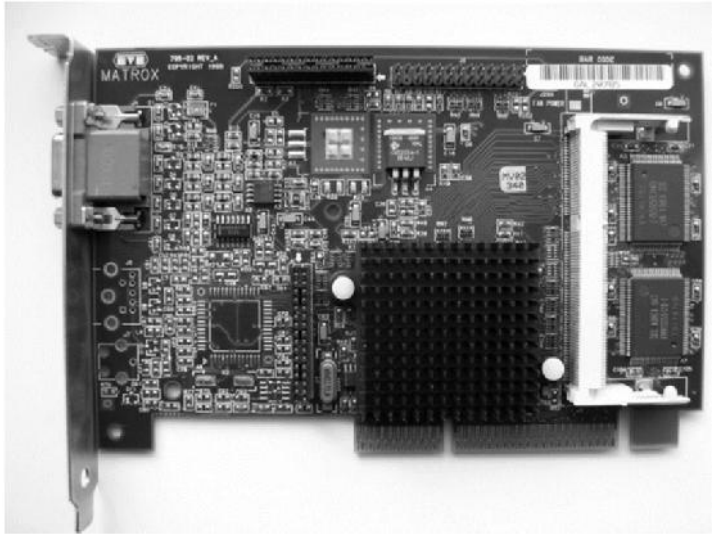


RDRAM (Rambus DRAM)

ชื่อของ Rambus เป็นเครื่องหมายการค้าของบริษัท RAMBUS Inc. ซึ่งก่อตั้งมาตั้งแต่ยุค 80 มี performance มากกว่าเป็นสามเท่าจาก SDRAM และมีเพียงแค่ช่องสัญญาณเดียวก็มีอัตราการส่งถ่ายข้อมูลสูงถึง 1.6 GB ต่อวินาที ถึงแม้ว่าเวลาในการเข้าถึงข้อมูลแบบสุ่มของแรมชนิดนี้จะช้า แต่การเข้าถึงข้อมูลแบบต่อเนื่องจะเร็วมากRDRAM มีการพัฒนาอินเตอร์เฟส และมี PCB (Printed Circuit Board) ที่ดี ๆ และรวมถึงคอนโทรลเลอร์ของอินเตอร์เฟสให้สามารถใช้งานได้ถึง 2 ช่องสัญญาณ และมีอัตราการส่งถ่ายข้อมูลเพิ่มเป็น 3.2 GB ต่อวินาที และถ้าหากใช้ได้ถึง 4 ช่องสัญญาณก็มีอัตราการส่งถ่ายข้อมูลเพิ่มขึ้นถึง 6.4 GB ต่อวินาที

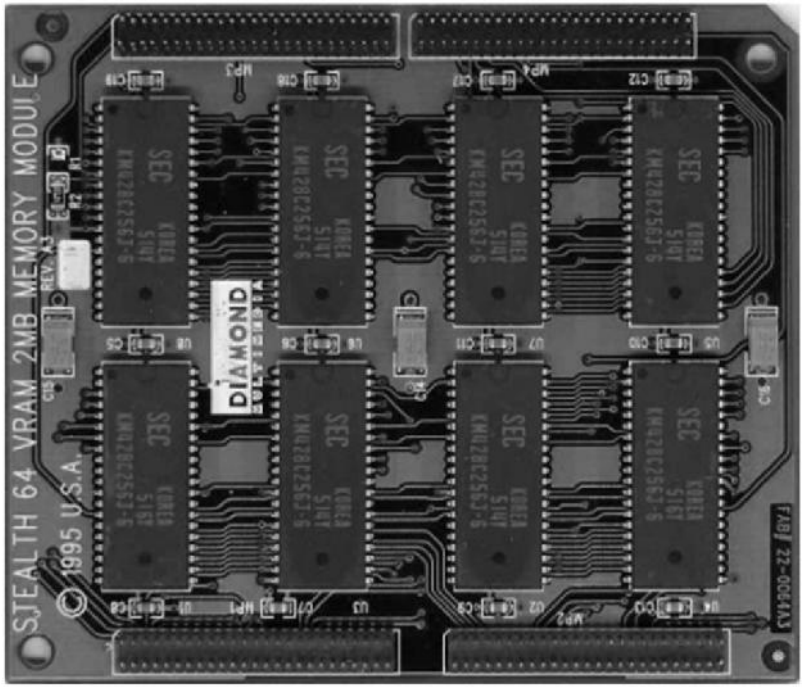
SGRAM (Synchronous Graphic RAM)

SGRAM แยกออกมาจาก SDRAM โดยถูกปรับแต่งมาสำหรับงานด้านกราฟิกเป็นพิเศษ แต่โดยโครงสร้างของฮาร์ดแวร์แล้วแทบไม่มีอะไรต่างจาก SDRAM เลย ใน graphic card ที่เป็นรุ่นเดียวกันบ้างก็ใช้ SDRAM บ้างก็ใช้ SGRAM เช่น Matrox G200 แต่จุดที่แตกต่างกันคือ ฟังก์ชันที่ใช้โดย page register ซึ่ง SGRAM สามารถเขียนข้อมูลได้หลาย ๆ ตำแหน่งในหนึ่งสัญญาณนาฬิกา ทำให้ความเร็วในการแสดงผล และ clear screen ทำได้เร็วมาก และยังสามารถเขียนแค่บาง bit ในการแก้ไข word ได้โดยไม่ต้องเขียนข้อมูลใหม่ทั้งหมด โดยการใช้ bit mask ในการเลือก bit ที่จะเขียนใหม่ SGRAM เหมาะกับงานด้านกราฟิกมากกว่างานปกติทั่ว ๆ ไป



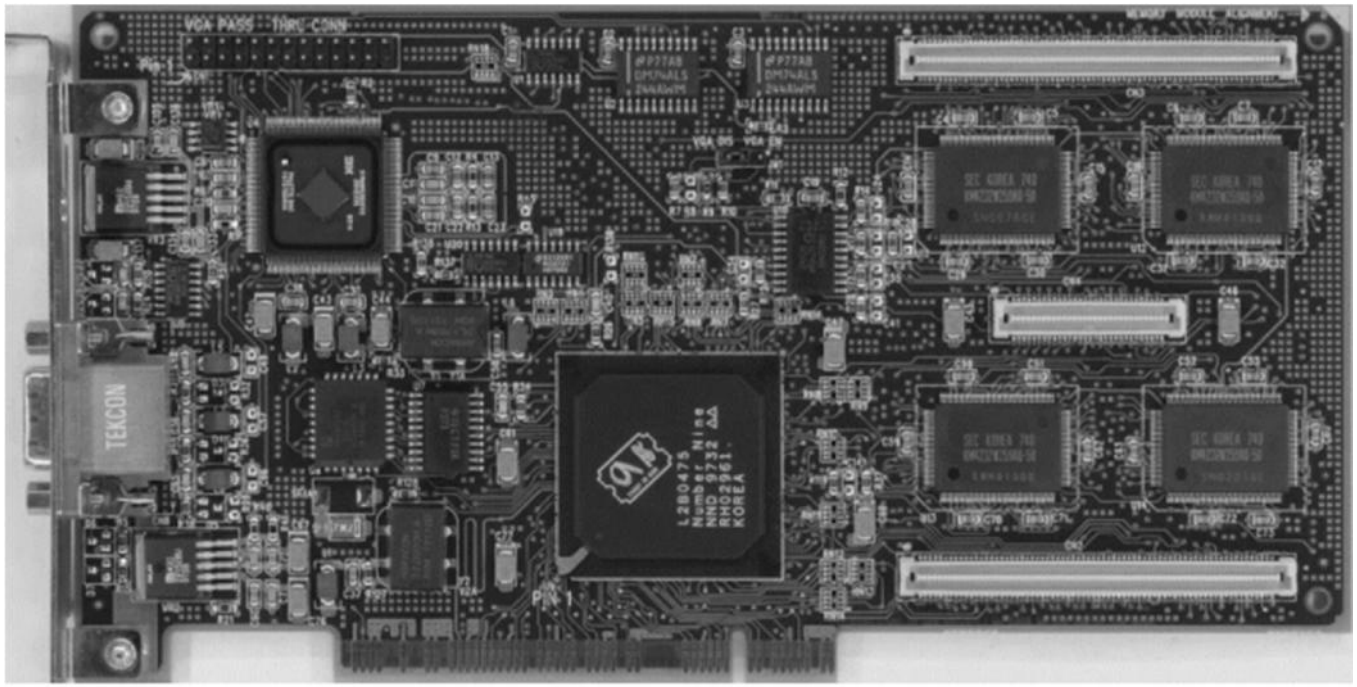
Video Ram (VRAM)

ได้รับการออกแบบให้ใช้บน Display card ซึ่ง VRAM พัฒนามาจาก DRAM แต่แตกต่างกันด้วยกลไกการทำงานบางอย่างที่เพิ่มขึ้นมาคือ VRAM มี serial port พิเศษเพิ่มขึ้นอีก 1-2 port ทำให้ผู้ใช้มองว่าเป็นแรมแบบพอร์ทคู่ (Dual port หรือ Triple port) นอกจากนี้ในส่วนของ parallel port ซึ่งเป็น standard interface จะถูกใช้ในการติดต่อกับ host processor เพื่อสั่งให้ทำการรีเฟรชภาพขึ้นมาใหม่ และ serial port จะส่งข้อมูลภาพออกสู่จอภาพ



Windows RAM (WRAM)

เป็นแรมที่ใช้บน graphic card ตระกูล Millennium, Millennium II และในรุ่น Number 9 Revolution IV ที่ถูกพัฒนาโดย Matrox ซึ่งคุณสมบัติโดยรวมของ WRAM ก็เหมือนกับ VRAM แต่แตกต่างกันตรงที่ WRAM สามารถรองรับ bandwidth ที่สูงกว่าและใช้ระบบ double buffer ทำให้สามารถทำงานได้เร็วกว่า VRAM



หน่วยความจำแคช

- ซีพียูที่ถูกผลิตออกมามีความเร็วสูงขึ้นเรื่อย ๆ จึงทำให้เกิดปัญหาด้านความเร็วที่แตกต่างกันระหว่างซีพียูกับแรม ทำให้เกิดแนวความคิดในการสร้างแคช (Cache) ขึ้นไว้ภายในซีพียู ซึ่งแคชก็คือหน่วยความจำขนาดเล็กที่มีความเร็วและประสิทธิภาพในการทำงานสูง
- หน้าที่ของแคชคือ จดจำคำสั่งและผลลัพธ์ที่ถูกเรียกใช้บ่อย ๆ เพื่อใช้ในการประมวลผลในครั้งต่อ ๆ ไป เพื่อซีพียูไม่ต้องเสียเวลารอเรียกคำสั่งนั้น ๆ อีก

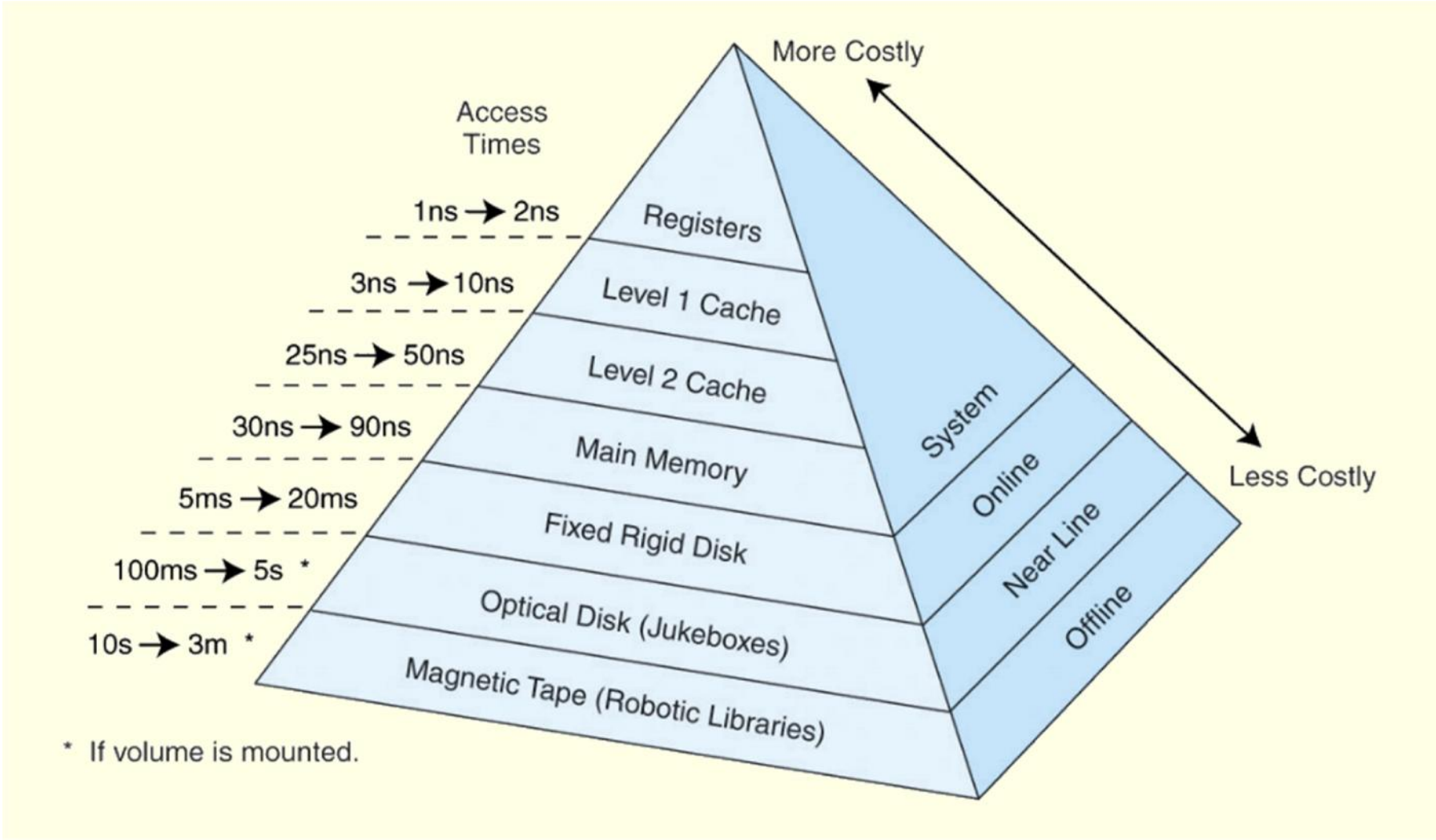
ระดับของแคช (Cache Level)

- แคชระดับที่ 1 (L1 Cache หรือ Cache level 1) เป็นแคชที่สร้างลงบนชิพของซีพียู หรือเรียกอีกชื่อหนึ่งว่า internal cache จะทำงานที่ความเร็วเดียวกับซีพียู L1 Cache มีขนาดความจุไม่ใหญ่มากนัก เช่น สำหรับชิพ Pentium MMX และ Pentium II มีขนาด 32 KB แต่ในปัจจุบันสำหรับซีพียูรุ่นใหม่ ๆ ความจุของแคชก็มีขนาดใหญ่ขึ้นตามไปด้วย
- แคชระดับที่ 2 (L2 Cache หรือ Cache level 2) มี 2 แบบคือ แบบที่อยู่บนซีพียู (Internal Cache) และแบบที่อยู่ภายนอกซีพียู (External Cache) แคชแบบที่อยู่ภายนอกซีพียูจะแบ่งเป็นอีก 2 แบบคือ แบบที่ติดตั้งอยู่บนแผ่นวงจรเดียวกับซีพียู เช่น แบบที่อยู่บนซีพียู Pentium กับ Pentium III รุ่นแรก ๆ (ลักษณะเป็นการ์ดใส่อยู่ในตลับ)
- แคชระดับที่ 3 (L3 Cache หรือ Cache level 3) เป็นแคชที่ออกแบบมาเฉพาะซีพียูของค่าย AMD รุ่น K6-III ซึ่งเป็นแคชแบบที่ติดตั้งมาบนเมนบอร์ด เนื่องจากซีพียูรุ่น K6-III มี L2 Cache ติดตั้งอยู่บนซีพียูแล้ว และเมนบอร์ดที่ใช้กับ K6-III ก็เป็นเมนบอร์ดรุ่นที่มีการใส่แคชมาให้แล้วบนเมนบอร์ด

ลำดับชั้นของหน่วยความจำ (The Memory Hierarchy)

- หน่วยความจำที่เร็วกว่ามักจะแพงกว่า
- การทำให้เครื่องคอมพิวเตอร์มีประสิทธิภาพดีที่สุดในราคาที่ถูกที่สุดจึงจำเป็นต้องมีการจัดลำดับชั้นในการทำงานของหน่วยความจำ
- หน่วยความจำขนาดเล็ก และทำงานได้เร็วจะบรรจุลงอยู่ในซีพียู
- หน่วยความจำขนาดใหญ่ ที่ทำงานได้ช้าจะถูกเข้าถึงได้ผ่านระบบบัส
- หน่วยความจำขนาดใหญ่ที่เก็บข้อมูลแบบถาวร ได้แก่ ดิสก์ เทป ซีดีรอม จะถูกวางตำแหน่งไว้ห่างออกไปจากซีพียู

ลำดับชั้นของหน่วยความจำ (The Memory Hierarchy)



ลำดับชั้นของหน่วยความจำ (The Memory Hierarchy)

- การเข้าถึงข้อมูลของซีพียู จะส่งการร้องขอไปยังหน่วยความจำที่อยู่ใกล้ที่สุดก่อน โดยทั่วไปก็คือหน่วยความจำแคช
- ถ้าไม่พบข้อมูลในแคช จะร้องขอจากหน่วยความจำหลัก และหน่วยความจำสำรองตามลำดับ
- เมื่อมีการร้องขอข้อมูลในหน่วยความจำ จะมีการดึง(Fetch) ข้อมูลนั้น และข้อมูลที่อยู่ใกล้เคียงไปเก็บที่หน่วยความจำแคชจนเต็ม
- เกณฑ์ในการเลือกข้อมูลใกล้เคียงมี 3 กรณี คือ
  - ข้อมูลที่มีการเข้าถึงล่าสุด และมีแนวโน้มที่จะเข้าถึงอีก
  - ข้อมูลที่อยู่ในคลัสเตอร์เดียวกัน
  - มีคำสั่งเข้าถึงข้อมูลในลักษณะแบบลำดับ จะเลือกข้อมูลที่อยู่ติดกันตามลำดับ

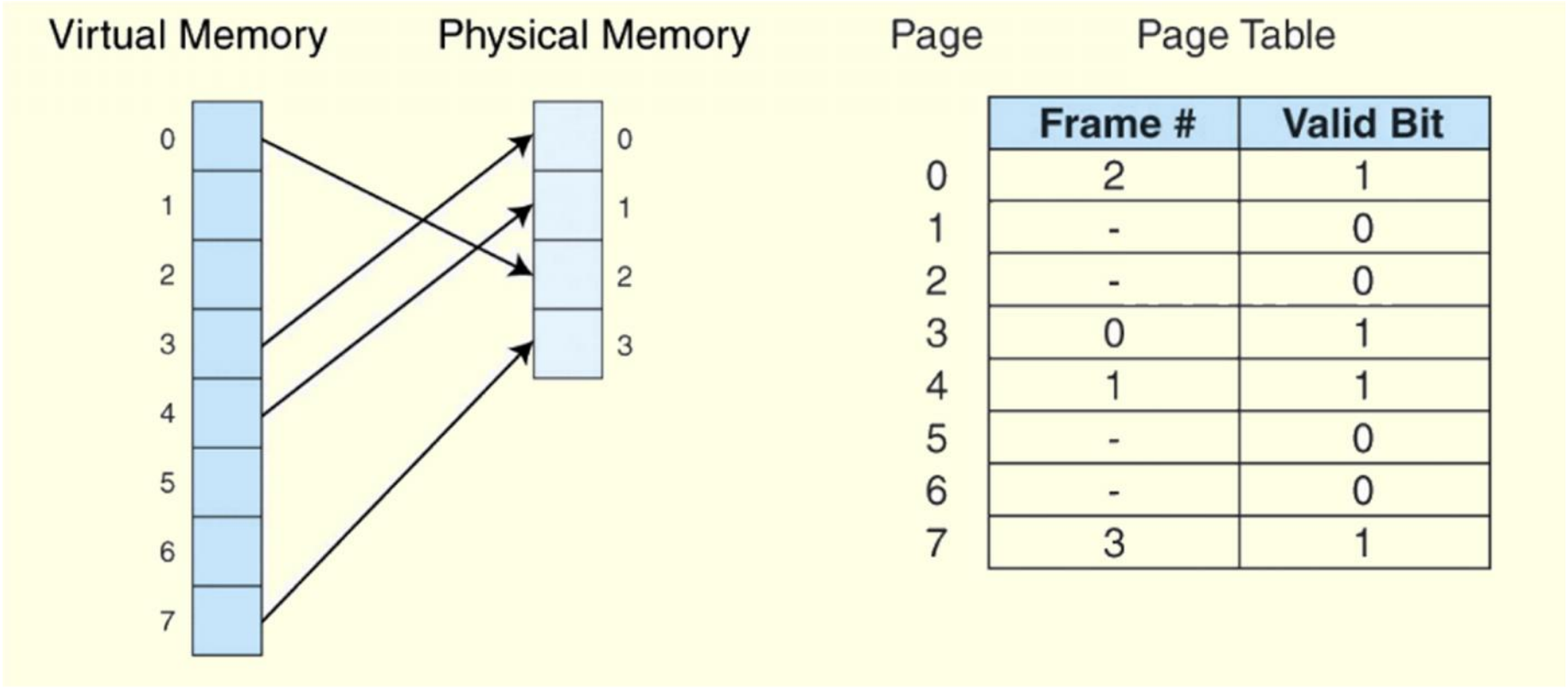
หน่วยความจำเสมือน (Virtual Memory)

- หน่วยความจำแคชจะช่วยเพิ่มประสิทธิภาพของระบบ เพราะเป็นหน่วยความจำที่สามารถเข้าถึงข้อมูลได้เร็วที่สุด แต่มีขนาดเล็ก
- หน่วยความจำเสมือนจะช่วยเพิ่มประสิทธิภาพของระบบ โดยการเพิ่มความจุของหน่วยความจำให้มีขนาดใหญ่ขึ้น โดยไม่ต้องเพิ่มขนาดของหน่วยความจำหลัก โดยการใช้เนื้อที่ในดิสก์แทนเนื้อที่ในหน่วยความจำหลัก
- ในระบบการแบ่งหน่วยความจำแบบหน้า (Paging) ข้อมูลในแต่ละหน้าจะถูกเก็บลงในดิสก์แทนในกรณีที่ข้อมูลมีปริมาณมากกว่าพื้นที่เก็บในหน่วยความจำหลัก และเป็นข้อมูลที่ยังไม่จำเป็นต้องใช้ในขณะนั้น

หน่วยความจำเสมือน (Virtual Memory)

- มีการกำหนดแอดเดรสจริง (Physical address) ซึ่งเป็นแอดเดรสในหน่วยความจำหลัก กับแอดเดรสเสมือน (Virtual address) ซึ่งเป็นแอดเดรสในดิสก์ขึ้นมา
- โปรแกรมจะสร้างแอดเดรสเสมือนจับคู่กับแอดเดรสจริง โดยตัวจัดการหน่วยความจำ (Memory Manager) ของระบบปฏิบัติการ
- เมื่อมีการอ้างอิงแอดเดรสที่ต้องใช้และยังไม่มีข้อมูลนั้นในหน่วยความจำหลัก จะโหลดข้อมูลจากดิสก์ในแอดเดรสที่จับคู่ไว้ไปยังหน่วยความจำหลัก เพื่อประมวลผลต่อไปโดยการโหลดทั้งหน้า

หน่วยความจำเสมือน (Virtual Memory)



Page	Frame #	Valid Bit
0	2	1
1	-	0
2	-	0
3	0	1
4	1	1
5	-	0
6	-	0
7	3	1

ลักษณะของหน่วยความจำในปัจจุบัน