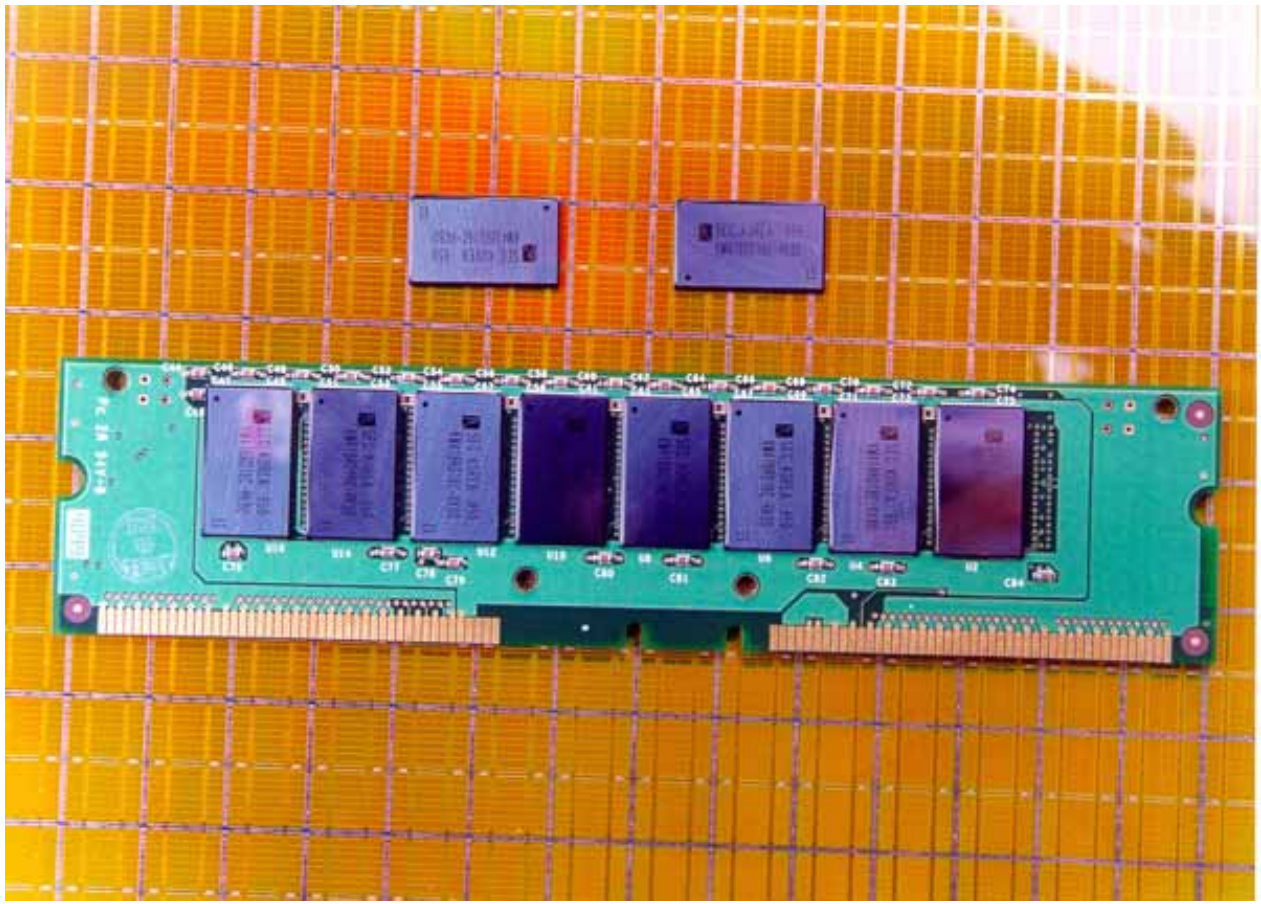
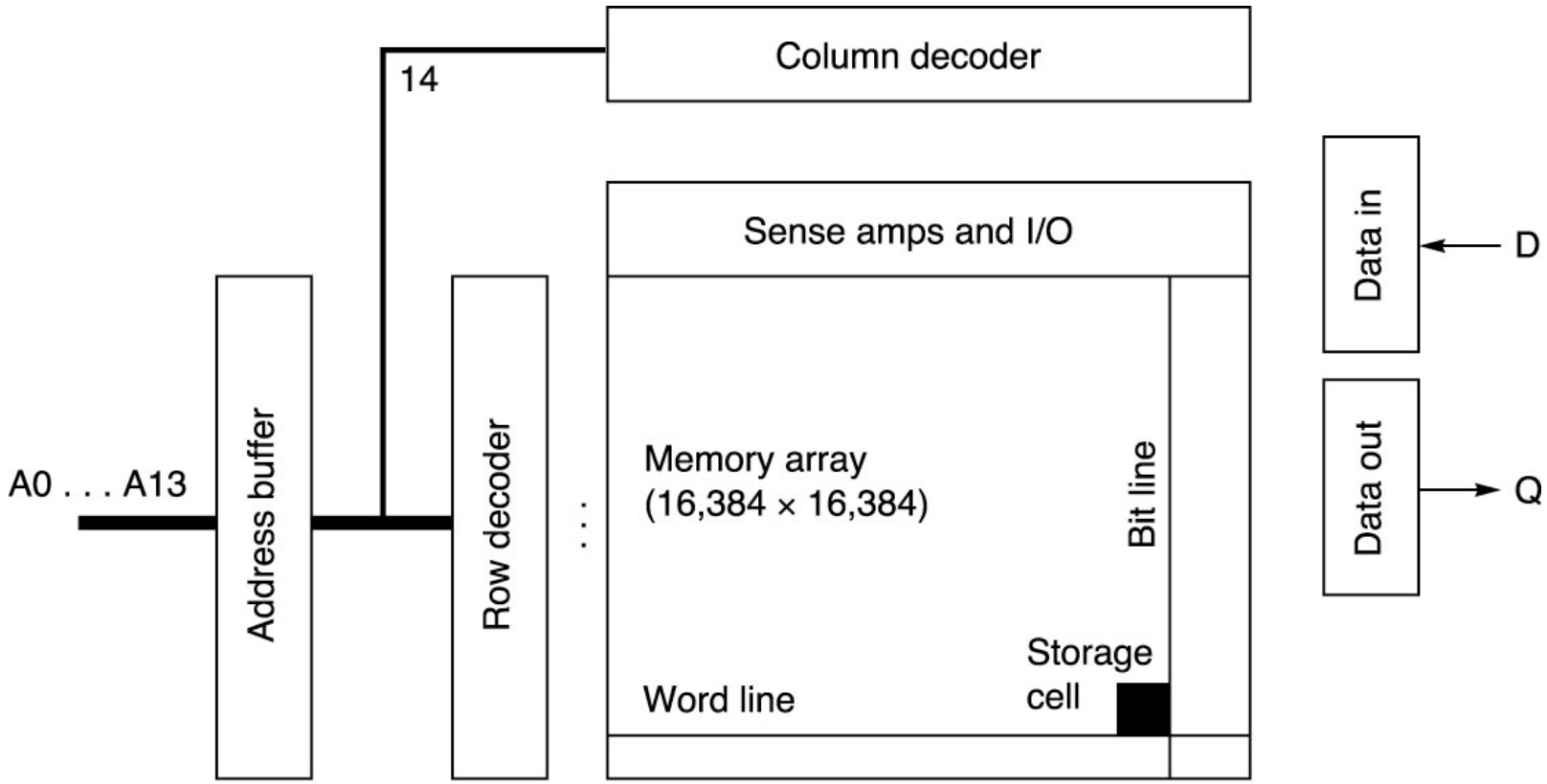
Lecture 18: DRAM Technologies

- Last Time:
  - Cache and Virtual Memory Review
- Today
  - DRAM organization
    or, “why is DRAM so slow???”

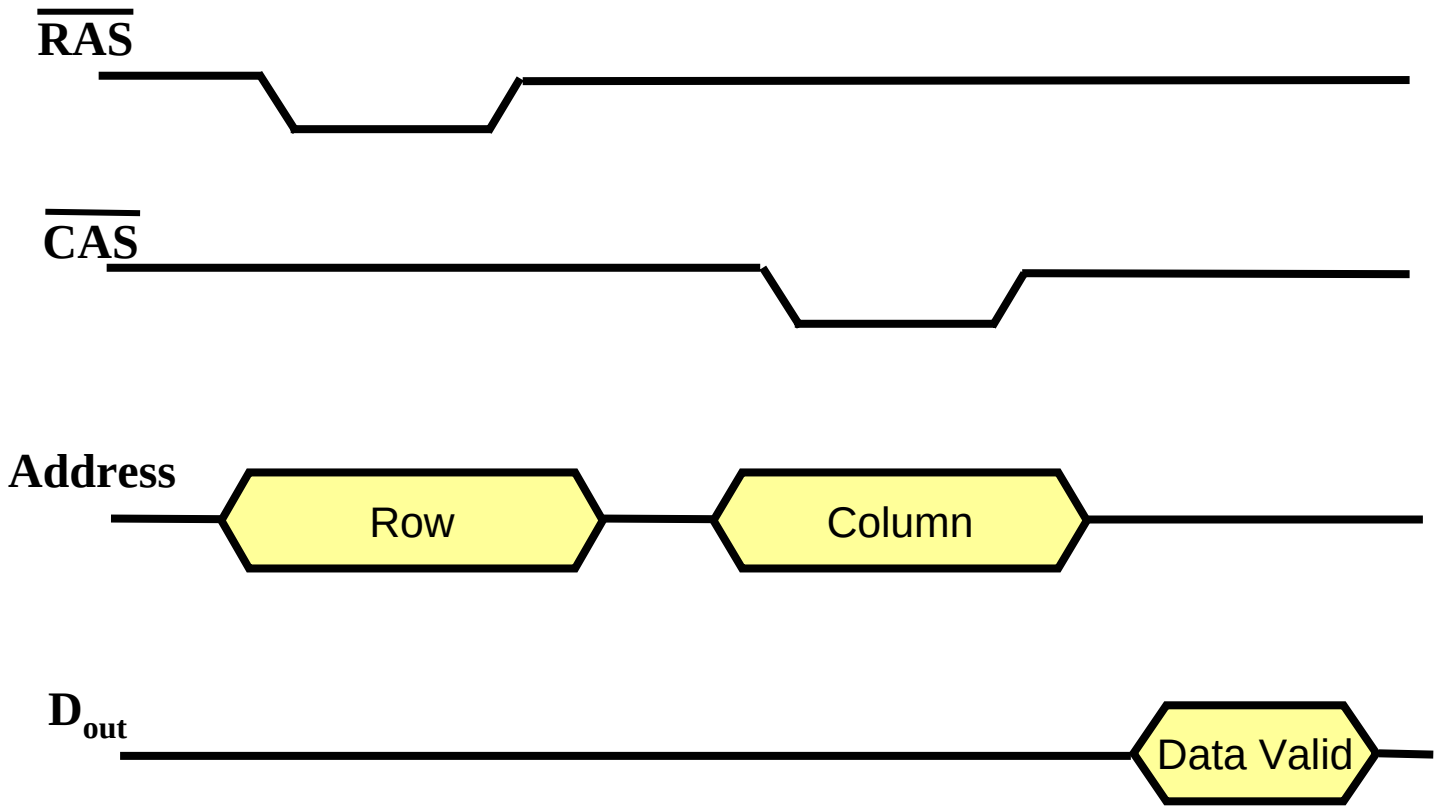
Main Memory = DRAM



Basic DRAM Architecture



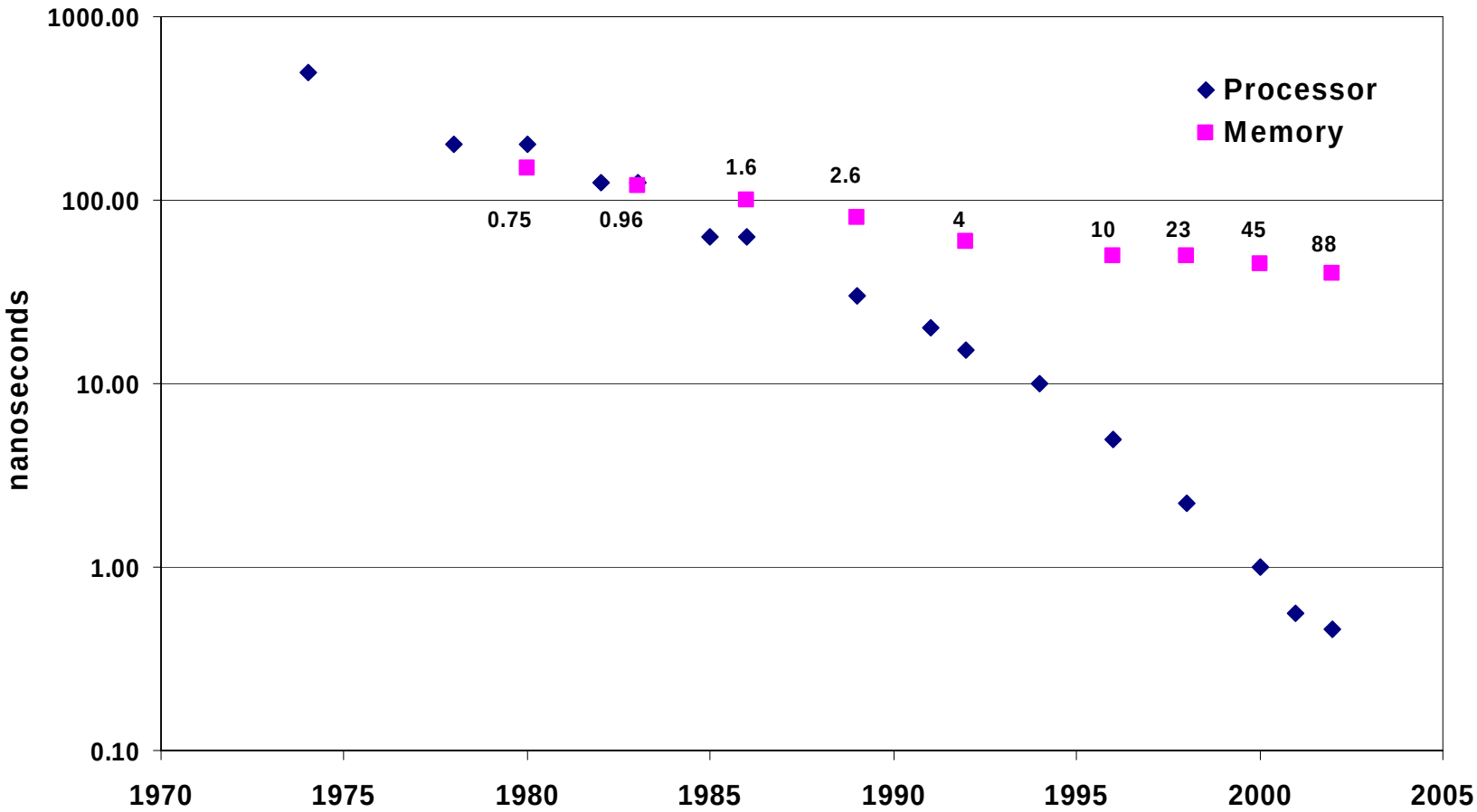
DRAM Access Time



Technology Trends

Year	Chip Size	Speeed	Cycle Time
1986	1Mbit	100ns	190ns
1989	4Mbit	80ns	165ns
1992	16Mbit	60ns	120ns
1996	64Mbit	50ns	110ns
1998	128Mbit	50ns	100ns
2000	156Mbit	45ns	90ns
2002	512Mbit	40ns	80ns

Processor/Memory Gap



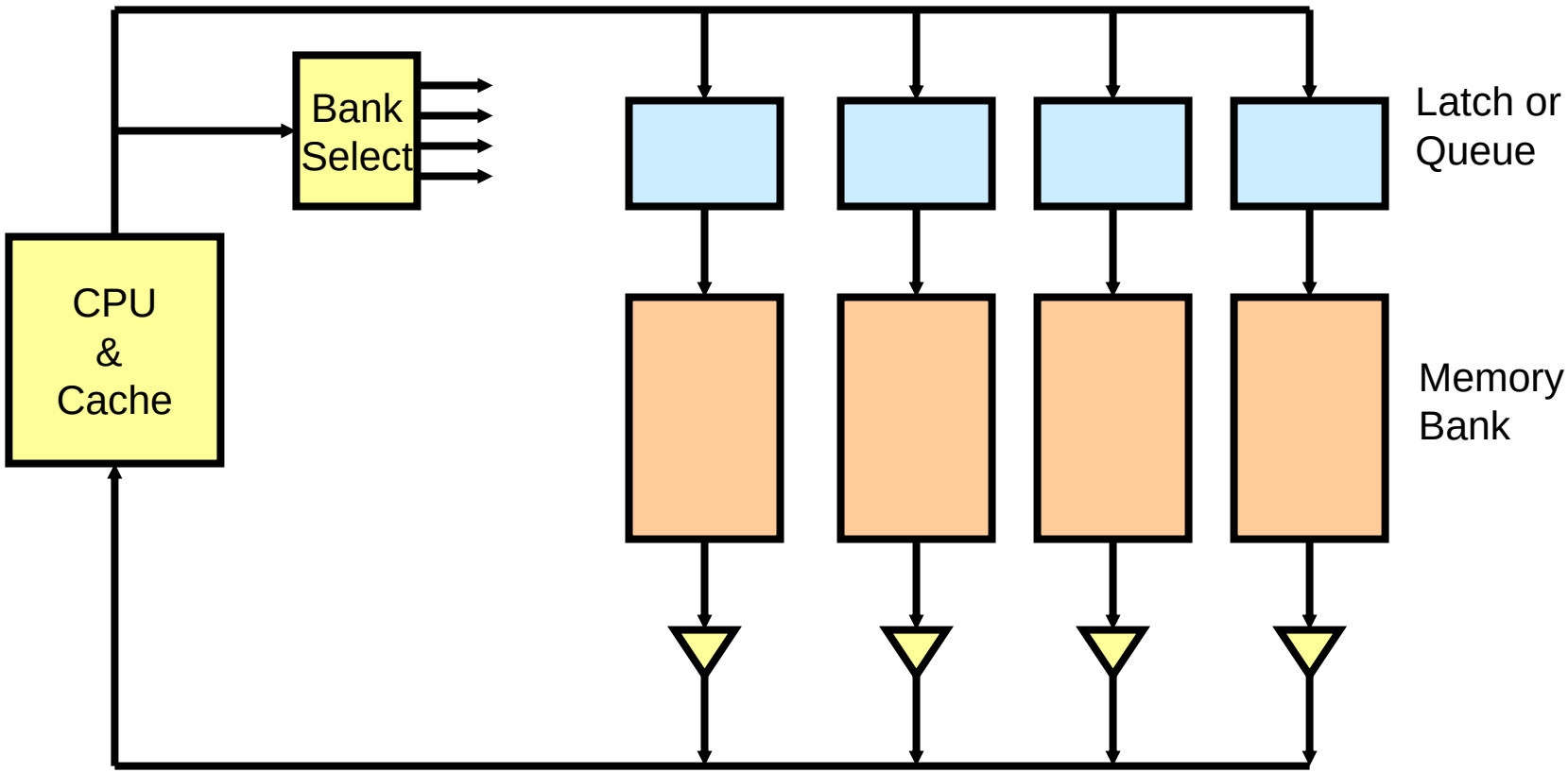
Improving External Memory System Performance

- Bandwidth vs. Latency
  - Bandwidth = #bits transferred per cycle
  - Latency = time to access DRAM
- Bandwidth
  - Memory bus width (16, 32, 64)
  - Multiple memory banks
    - Address interleaving
  - Multiple memory controllers (independent)
- Latency
  - Synchronous DRAM access modes
  - Faster interface (Rambus)

Memory Bus Width

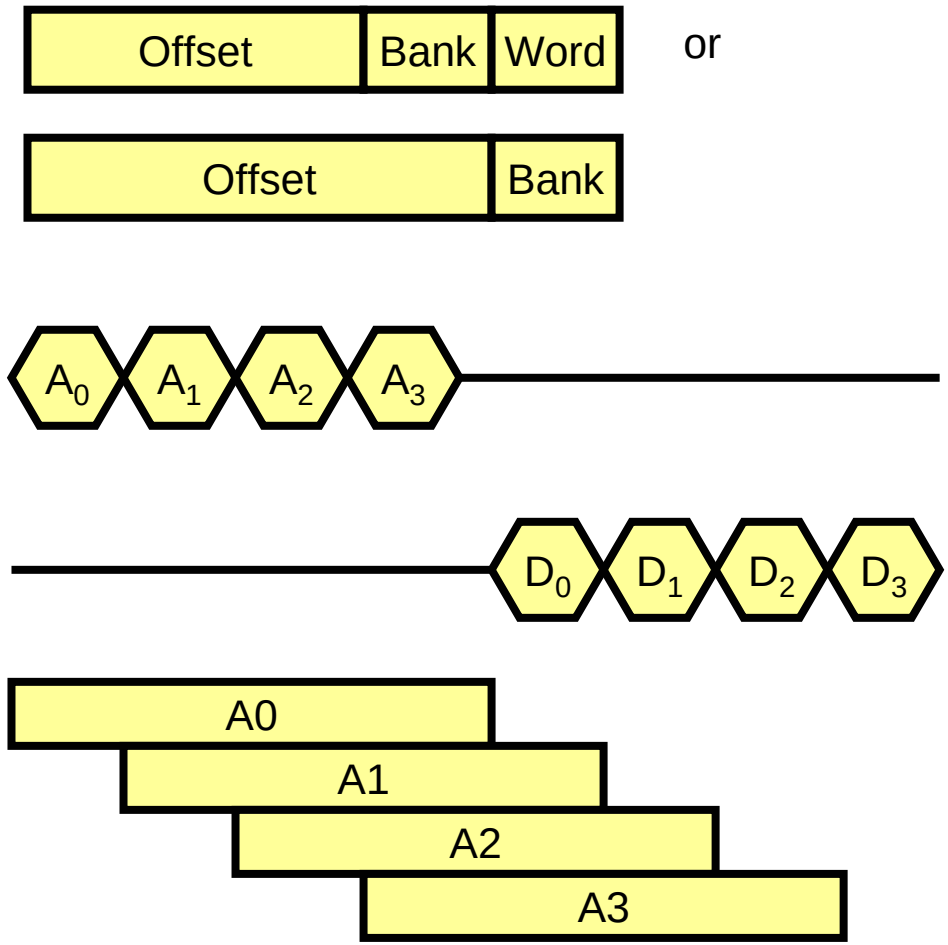
- Depends on microprocessor implementation
  - 386 = 32 bit external data bus
  - 386SX = 16 bit external data bus
  - Today = 64 bit data busses common, 128 bit soon
- Also interacts with external DRAM organization

Interleaved Memory Organization



Multiple Memory Banks (interleaving)

- Distribute memory address space across memory *banks*
- Route requests to banks based on low *block* address bits
- Allows memory accesses to go in parallel
- Two key issues
  - how are replies matched up with requesters?
  - how do we avoid *bank conflicts?*



Bank Conflicts

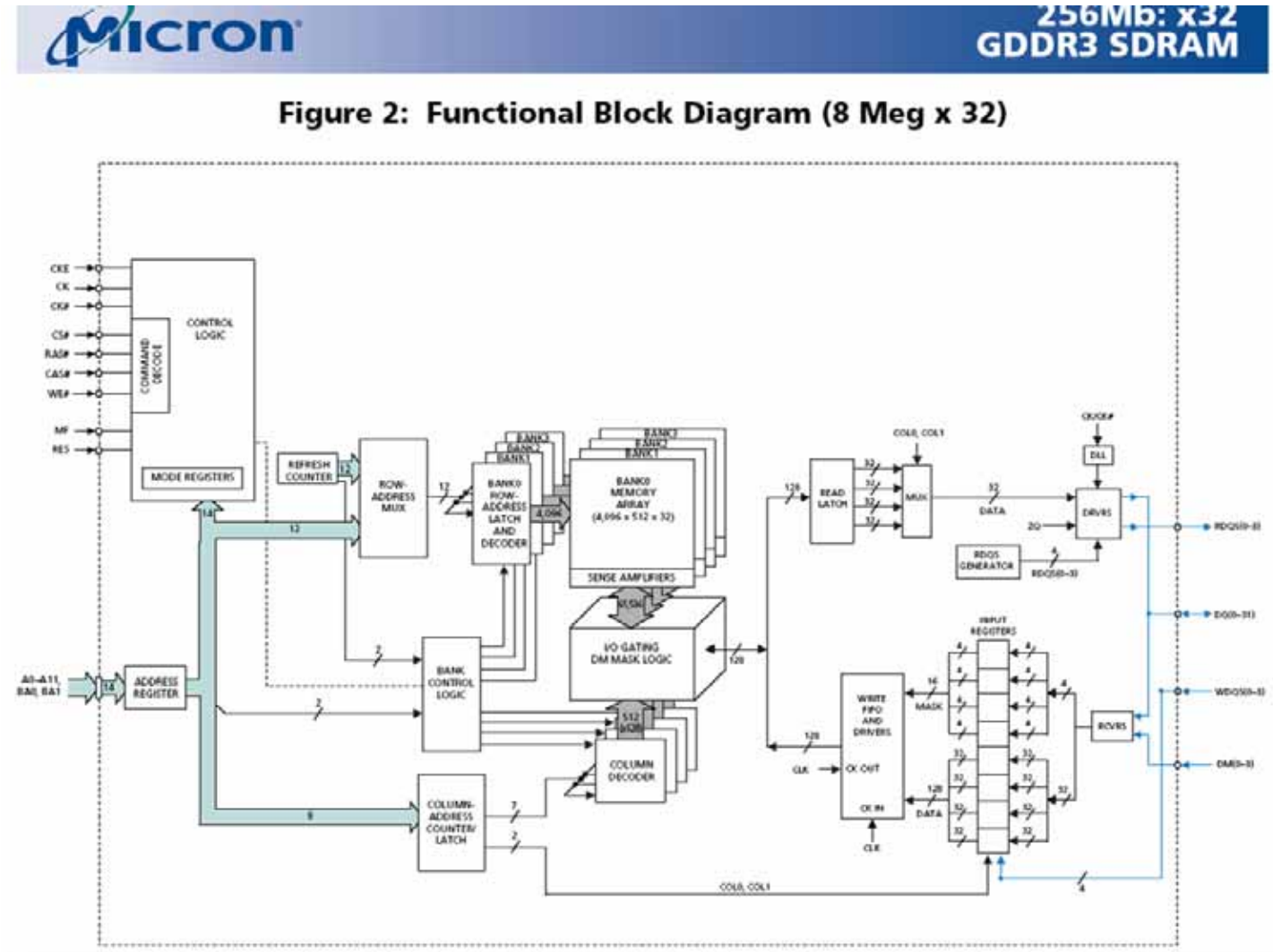
- Accesses may not reference banks evenly
- Consider
  - 0,1,2,3 … vs
  - 0,8,16,24 …
  - often caused by column access to a matrix
  - causes problems for large block size too
- Solutions
  - don't do that
    - number of columns in matrix not a power of 2
  - prime number of banks
    - number of active banks with stride s is lcd(s,b)
  - hash the banks
- These techniques now embedded inside a single DRAM chip

Improving per DRAM chip Performance: Synchronous DRAM

- Interface signals are clocked
  - Clock provided by microprocessor

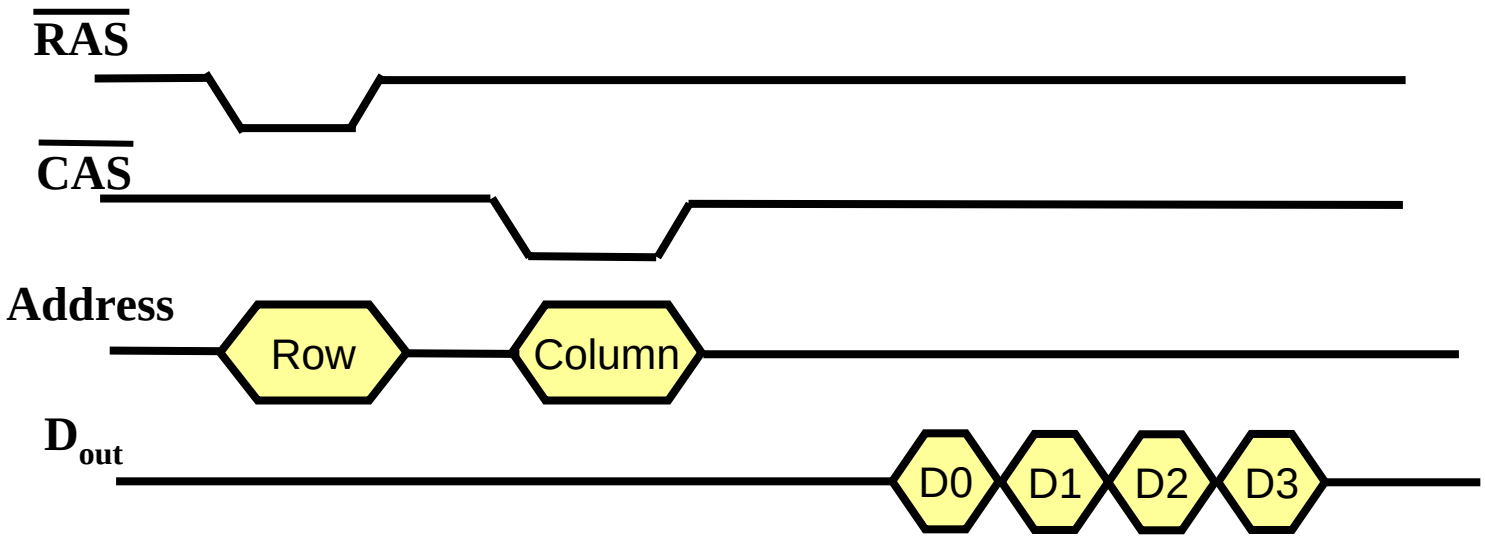
- Why?
  - Easier to designed timed protocols
    - "Data available 8 cycles after CAS"
  - Add intelligence to EMI (external memory interface) on CPU

Micron GDDR3 SDRAM - 2004

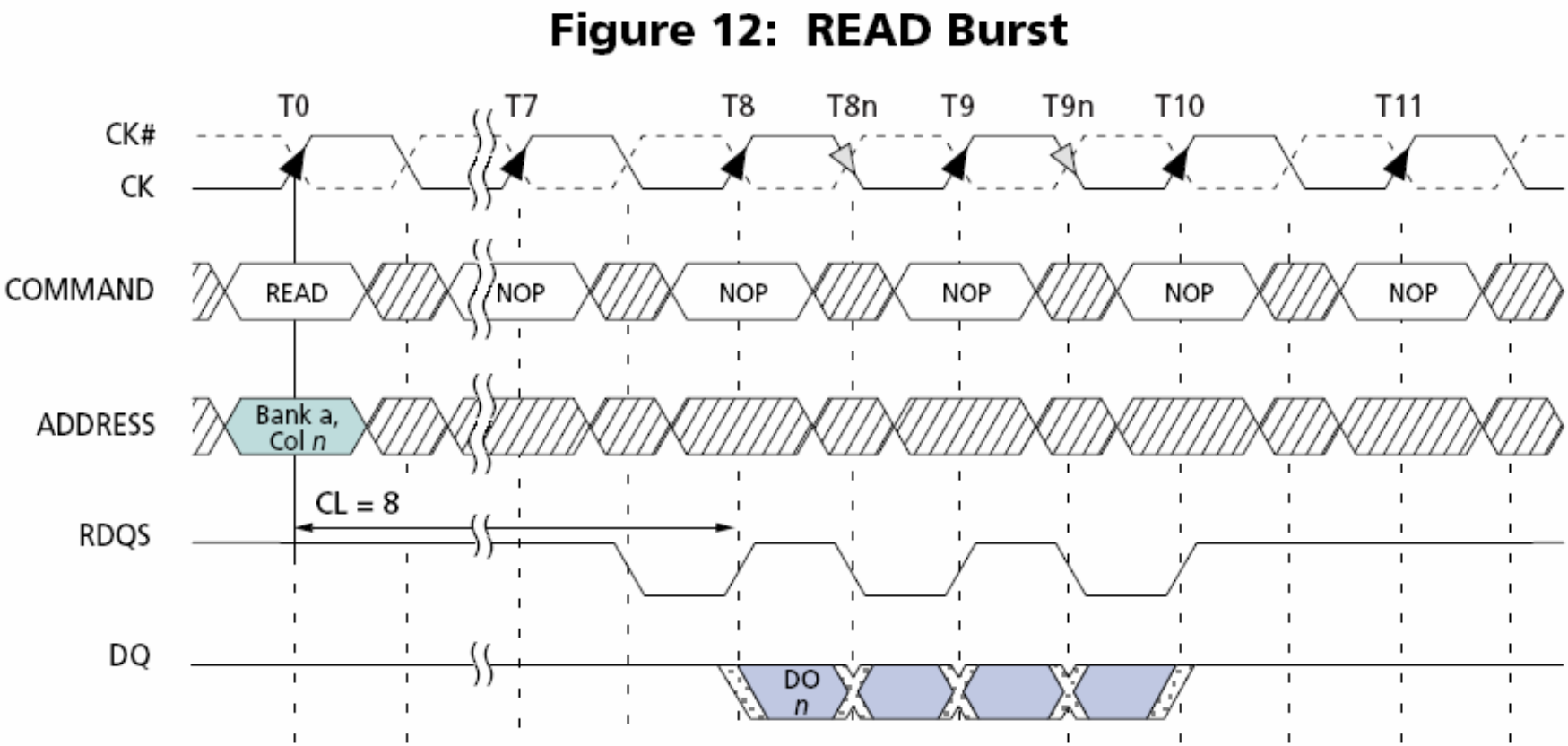


Burst Mode

- Provide one address and sequence of data comes out
  - Perfect for cache line reads and writes
  - Burst size programmable

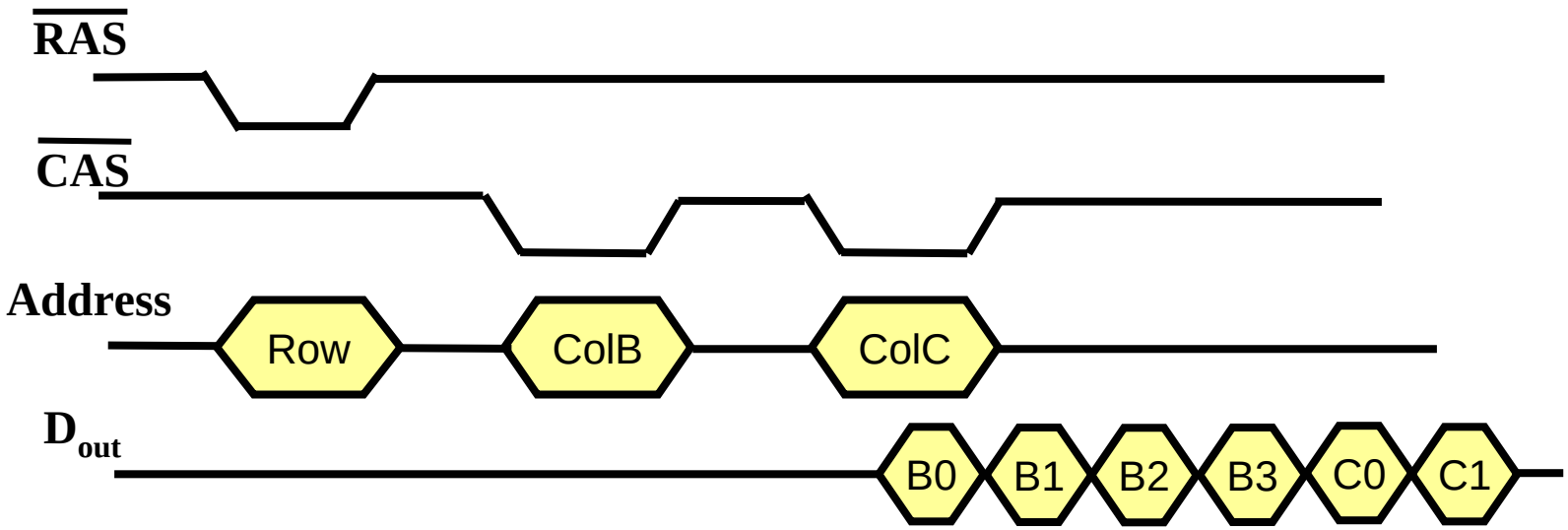


Synchronous DDR – approx 600 MHz clk



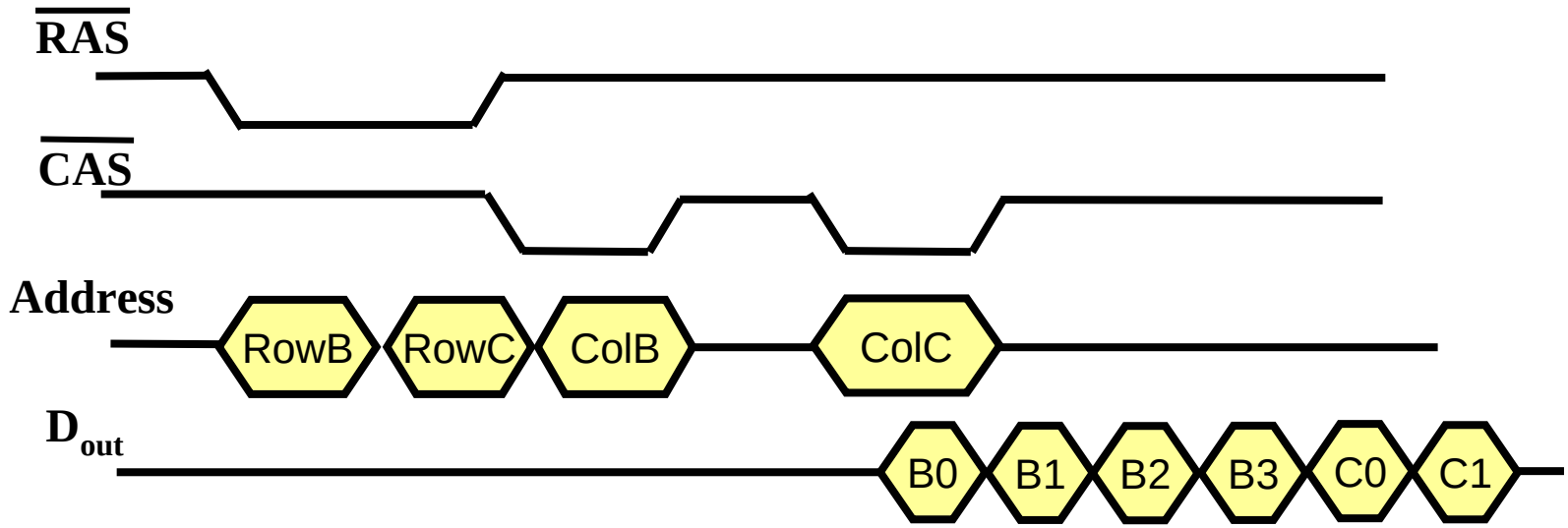
Page Mode Access

- One RAS (get whole row)
- Multiple CAS (different parts of the row)
- Exploits spatial locality (kind of like DRAM cache)



Pipelined Mode Access

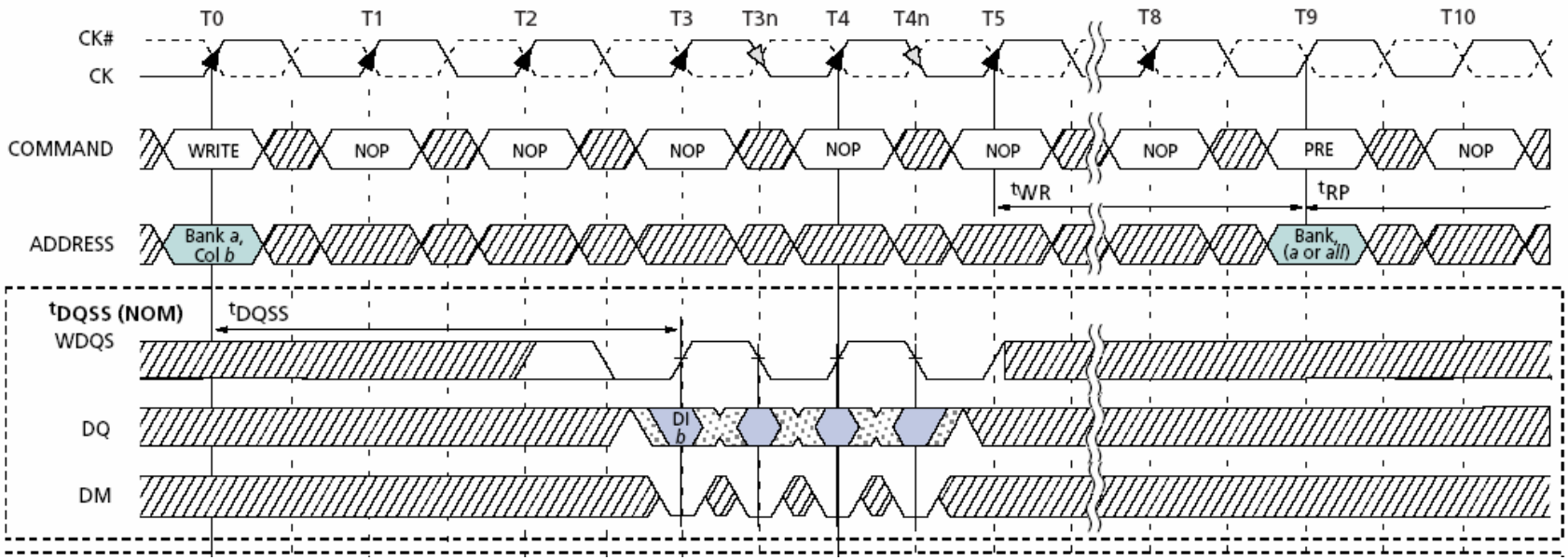
- Interleave access to multiple internal banks
- Lower latency for back-to-back access to different banks



- Can combine page/pipelined mode access
- Even better – schedule accesses to better use page/pipeline
  - Starting to be done in HW
- Can even put prefetching hardware at memory interface

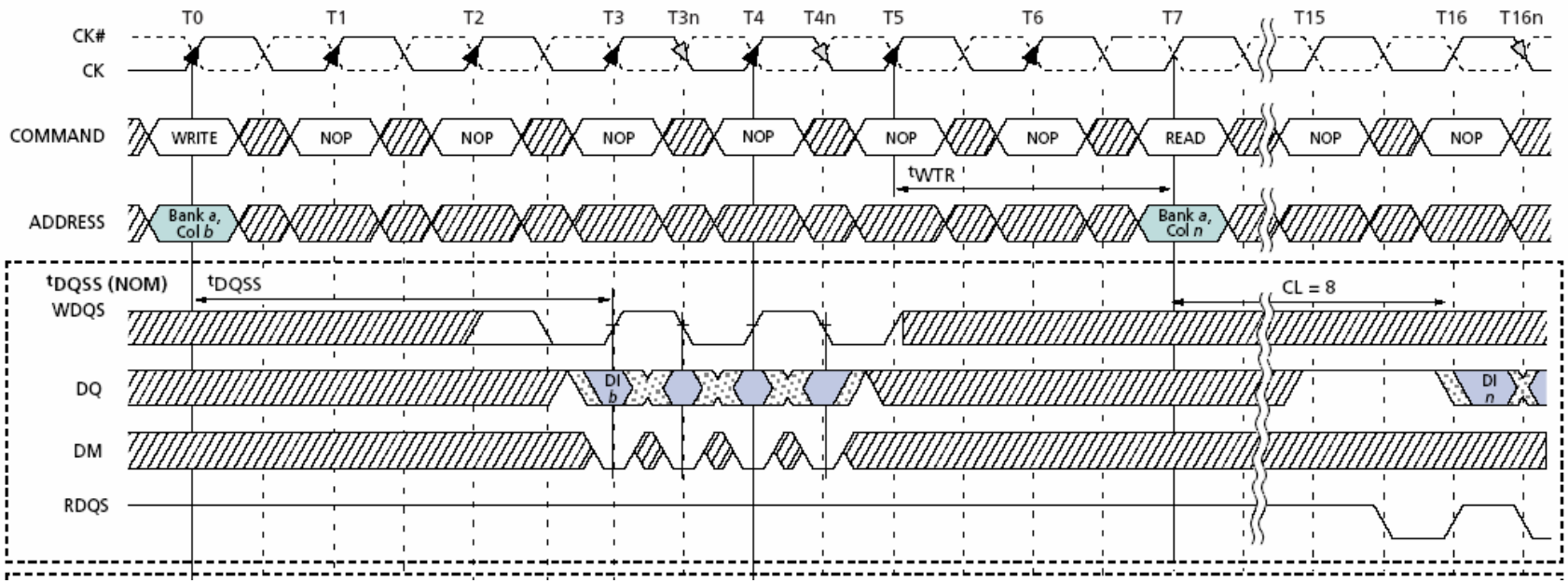
Changing the open page in a bank is slow

Figure 26: WRITE to PRECHARGE



Reads after writes are slow!

Figure 23: WRITE to READ



Newer DRAM Interfaces

- Double Data Rate (DDR) DRAM
  - Transfer data at rising and falling edge
  - Regular DRAM – 150MHz at 8byte width = 1.2GBytes/sec
  - Double data rate = 2.4GBytes/sec

- Rambus
  - RDRAM – split transaction bus, byte wide
    - More complicated electrical interface on DRAM and CPU
    - Restrictions on board level design
  - DRDRAM – Direct Rambus DRAM
    - 800 MHz interface at 18 bits = 1.6GB/sec per chip

Other Memory Technologies

- Embedded DRAM
  - DRAM built into processor chip (PIM-type architectures)
- ROM – read-only memory
  - Programmed at manufacture time
  - Arrays with programmable transistor sites
- Flash/EPROMs
  - Programmed by tunnelling current to gate
  - Floating gate store charge
  - Fast read, writes 10-100x slower than DRAM
  - Capacity 8x worse
- MRAM – magnetic RAM
  - Emerging technology – non-volatile storage
  - Slow reads, with higher power

Summary

- Two important characteristics of main memory
  - Latency – memory latency increasing – we're stuck with it
  - Bandwidth – typically limited by pin count on processor chip
    - Also a problem, but can at least throw money at it

- Variety of techniques to increase and make better use of BW
  - Increase transmission rates (DDR, Rambus)
  - Overlap multiple accesses
  - Prefetch data when memory bus is idle