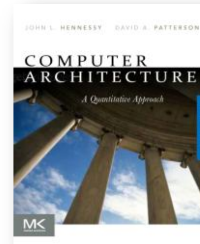
ECE 554 Computer Architecture
Lecture 5

Main Memory
Spring 2013



Sudeep Pasricha
Department of Electrical and Computer Engineering
Colorado State University

© Pasricha; portions: Kubiatowicz, Patterson, Mutlu, Binkert, Elsevier

1

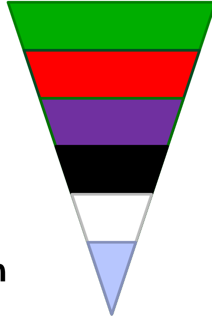
Main Memory Background

- **Performance of Main Memory:**
  - **Latency**: Cache Miss Penalty
    - » ***Access Time***: time between request and word arrives
    - » ***Cycle Time***: time between requests
  - **Bandwidth**: I/O & Large Block Miss Penalty (L2)
- **Main Memory is *DRAM*: Dynamic Random Access Memory**
  - Dynamic since needs to be refreshed periodically (8 ms, 1% time)
  - Addresses divided into 2 halves (Memory as a 2D matrix):
    - » ***RAS*** or ***Row Address Strobe***
    - » ***CAS*** or ***Column Address Strobe***
- **Cache uses *SRAM*: Static Random Access Memory**
  - No refresh (6 transistors per bit vs. 1 transistor + 1 capacitor per bit)
    Size: SRAM/DRAM *4-8*,
    Cycle time: DRAM/SRAM *8-16*

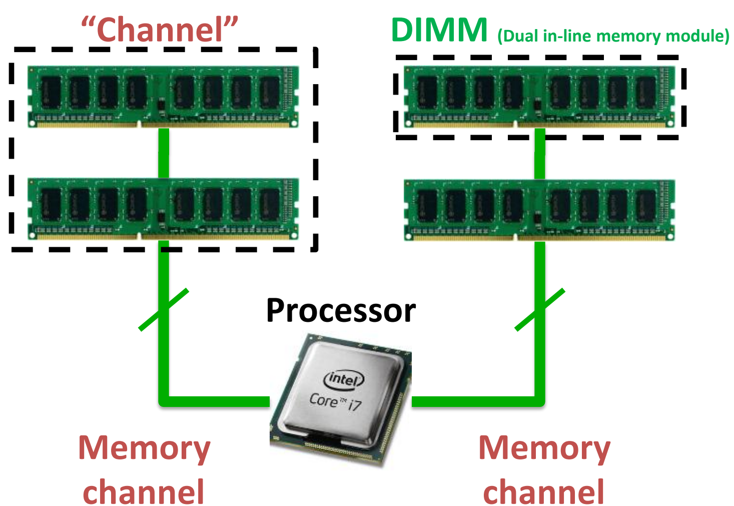
2

Memory subsystem organization

- Memory subsystem organization
  - Channel
  - DIMM
  - Rank
  - Chip
  - Bank
  - Row/Column



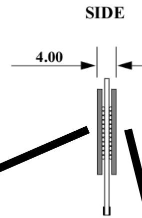
Memory subsystem



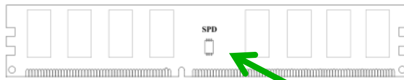
Breaking down a DIMM

DIMM (Dual in-line memory module)

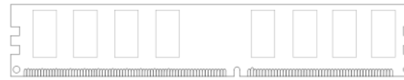
Side view



Front of DIMM



Back of DIMM

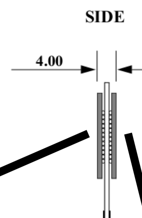


Serial presence detect (SPD)
- Stored in EEPROM on module
- has info to configure mem controllers

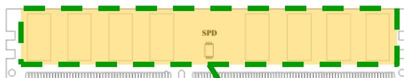
Breaking down a DIMM

DIMM (Dual in-line memory module)

Side view



Front of DIMM

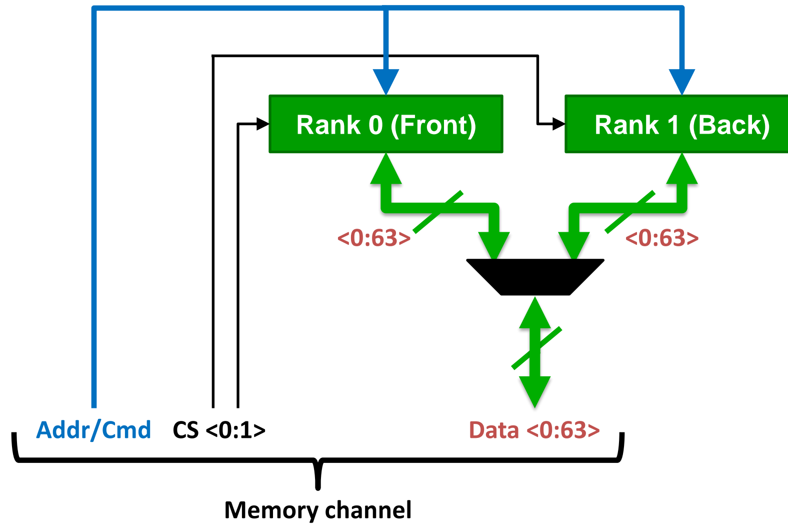


Back of DIMM

Rank 0: collection of 8 chips

Rank 1

Rank

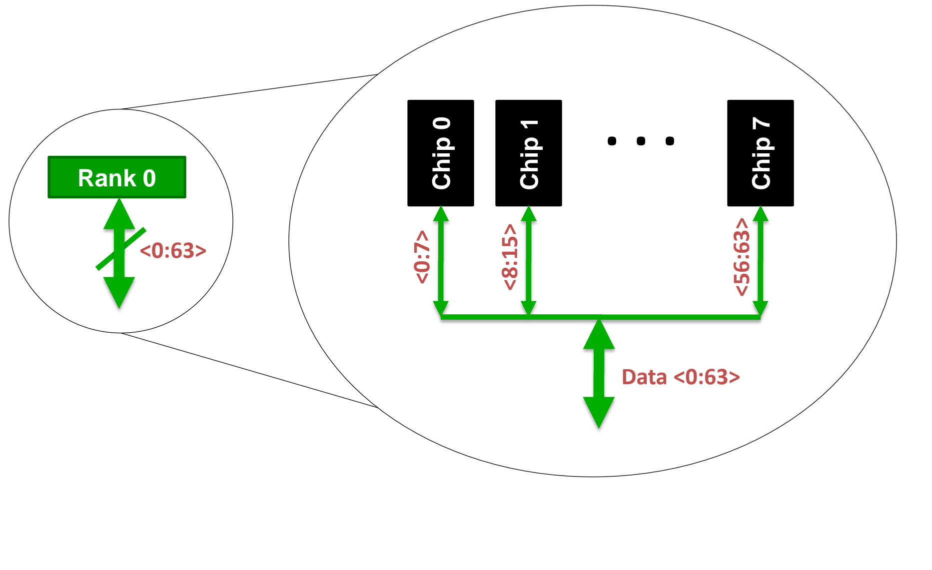


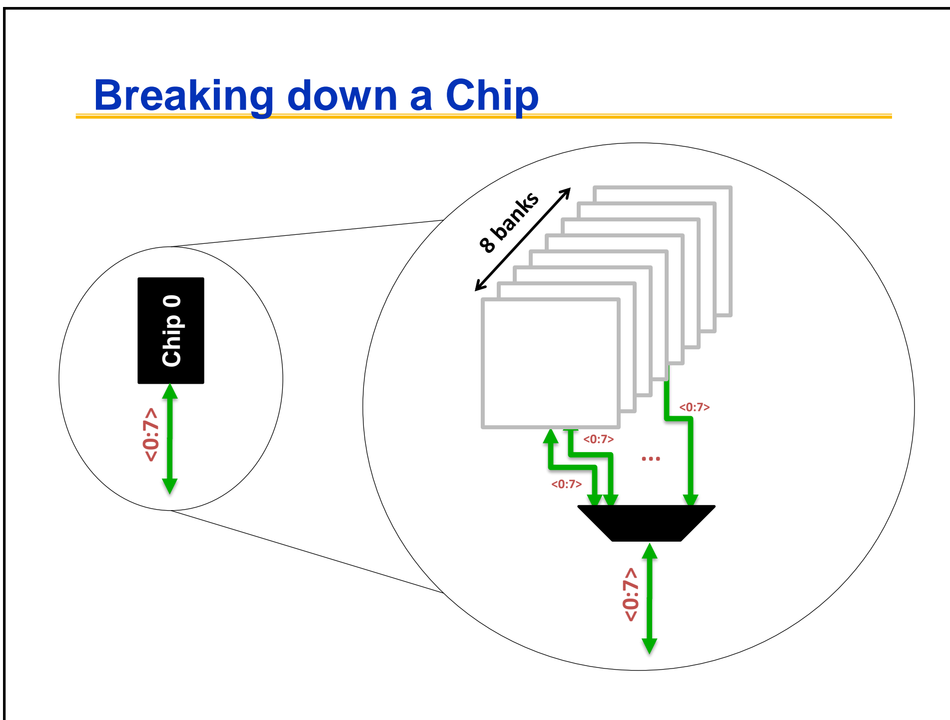
DIMM & Rank (from JEDEC)

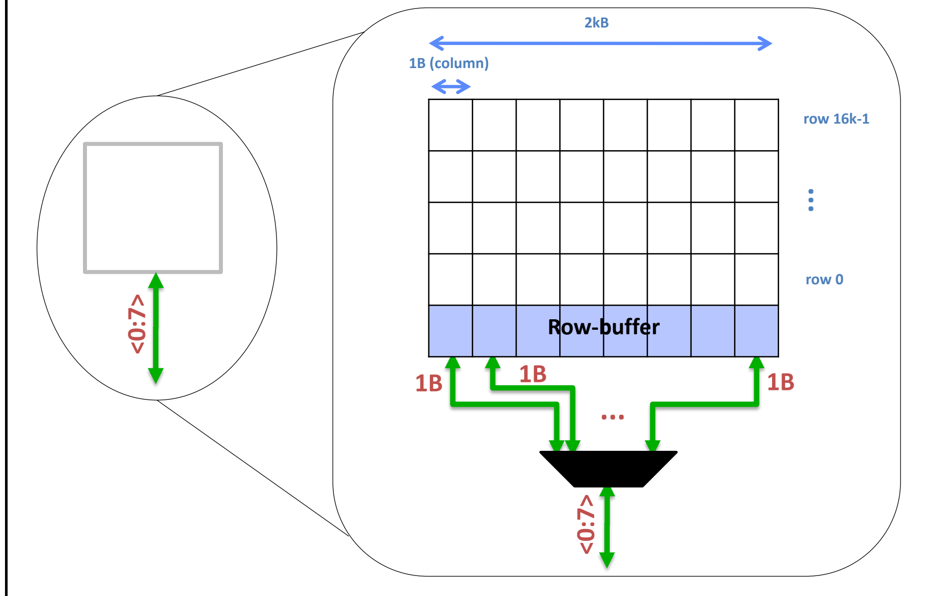
5 Unbuffered DIMM Details

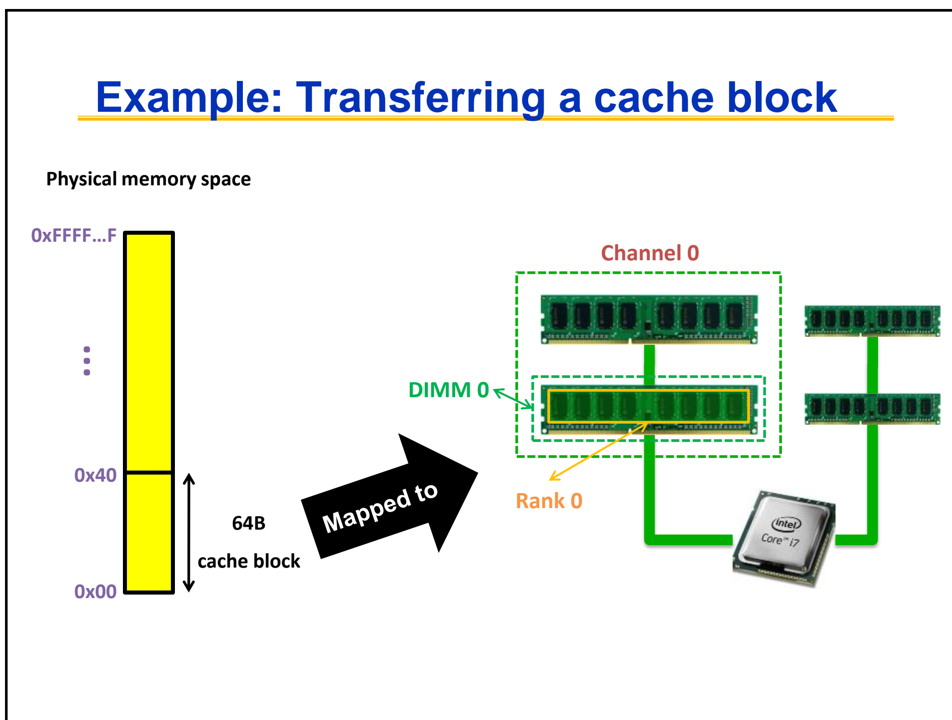
Table 8 — SDRAM Module Configurations (Reference Designs)

Raw Card Version	DIMM Capacity	DIMM Organization	SDRAM Density	SDRAM Organization	Number of SDRAMs	Number of Physical Ranks	Number of Banks in SDRAM	Number of Address Bits Row/Column
A	512MB	64 Meg x 64	512 Megabit	64 Meg x 8	8	1	8	13/10
	1GB	128 Meg x 64	1 Gigabit	128 Meg x 8	8	1	8	14/10
	2GB	256 Meg x 64	2 Gigabit	256 Meg x 8	8	1	8	15/10
	4GB	512 Meg x 64	4 Gigabit	512 Meg x 8	8	1	8	16/10
	8GB	1 Gig x 64	8 Gigabit	1 Gig x 8	8	1	8	16/11
B	1GB	128 Meg x 64	512 Megabit	64 Meg x 8	16	2	8	13/10
	2GB	256 Meg x 64	1 Gigabit	128 Meg x 8	16	2	8	14/10
	4GB	512 Meg x 64	2 Gigabit	256 Meg x 8	16	2	8	15/10
	8GB	1 Gig x 64	4 Gigabit	512 Meg x 8	16	2	8	16/10
	16GB	2 Gig x 64	8 Gigabit	1 Gig x 8	16	2	8	16/11
C[1]	256MB	32 Meg x 64	512 Megabit	32 Meg x 16	4	1	8	12/10
	512MB	64 Meg x 64	1 Gigabit	64 Meg x 16	4	1	8	13/10
	1GB	128 Meg x 64	2 Gigabit	128 Meg x 16	4	1	8	14/10
	2GB	256 Meg x 64	4 Gigabit	256 Meg x 16	4	1	8	15/10

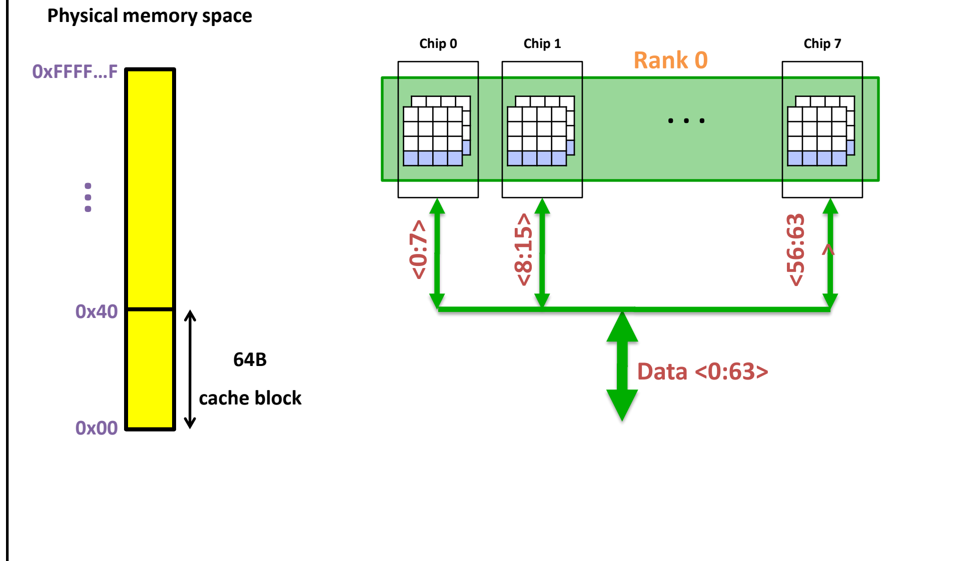

Breaking down a Rank
Rank 0
<0:63>
Chip 0
Chip 1
Chip 7
<0:7>
<8:15>
<56:63>
Data <0:63>


Breaking down a Chip
Chip 0
<0:7>
8 banks
<0:7>
<0:7>
<0:7>
<0:7>

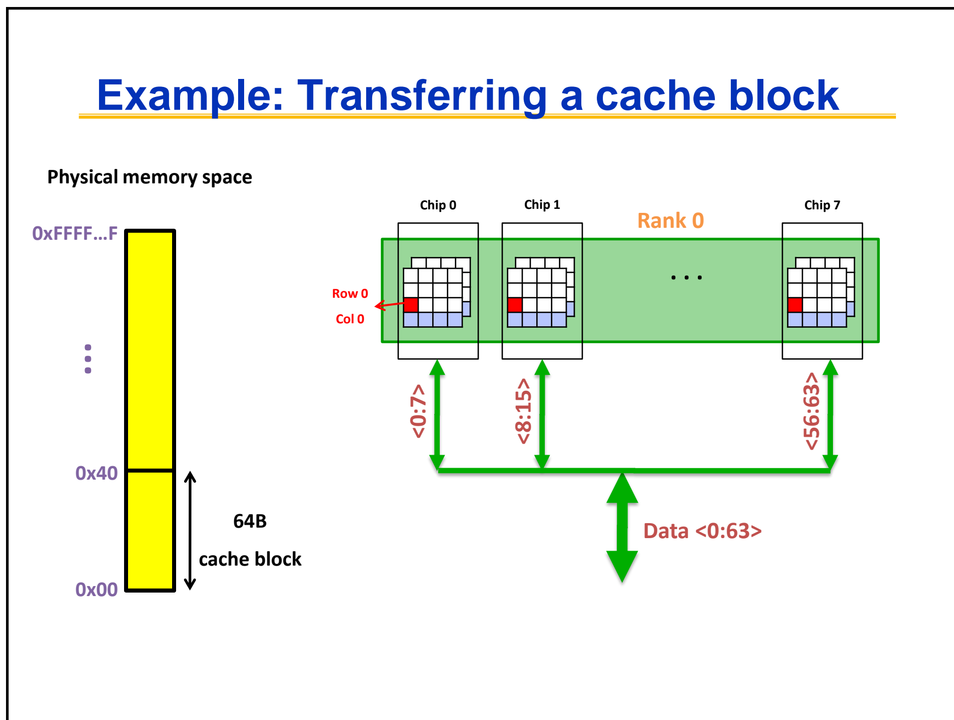

Breaking down a Bank
2kB
1B (column)
row 16k-1
row 0
Row-buffer
1B
1B
1B
<0:7>
<0:7>


Example: Transferring a cache block
Physical memory space
0xFFFF…F
0x40
0x00
64B
cache block
Mapped to
Channel 0
DIMM 0
Rank 0

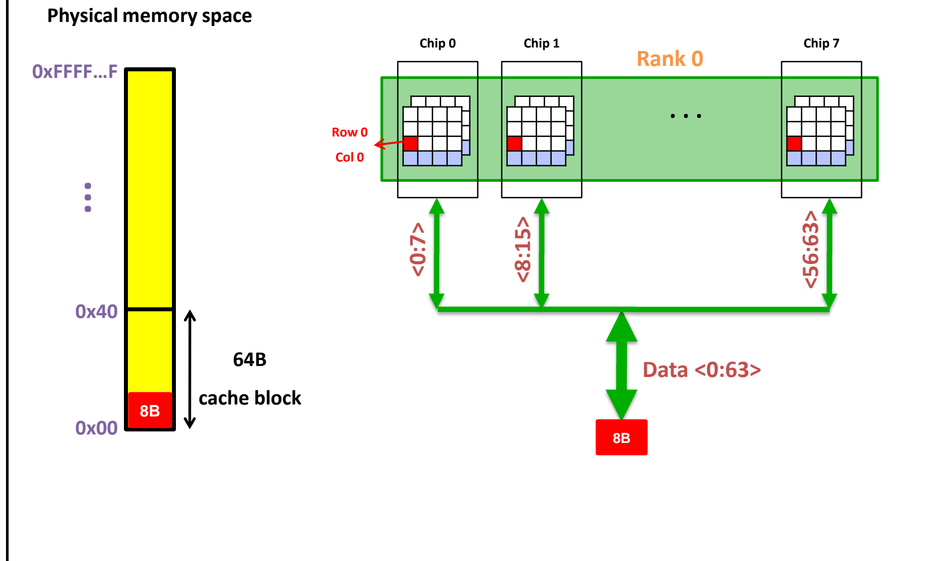
Example: Transferring a cache block



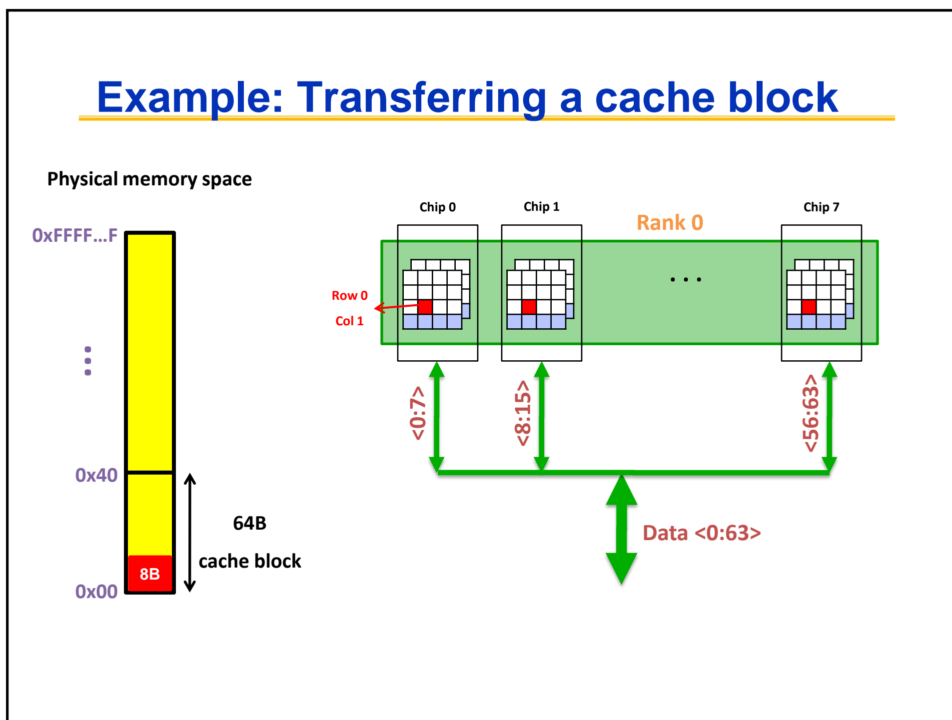
Example: Transferring a cache block



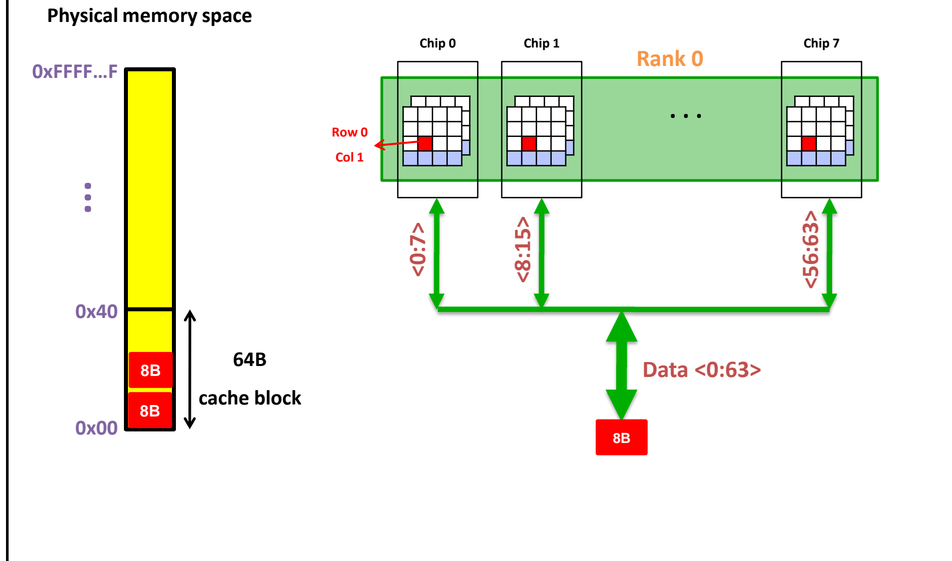
Example: Transferring a cache block

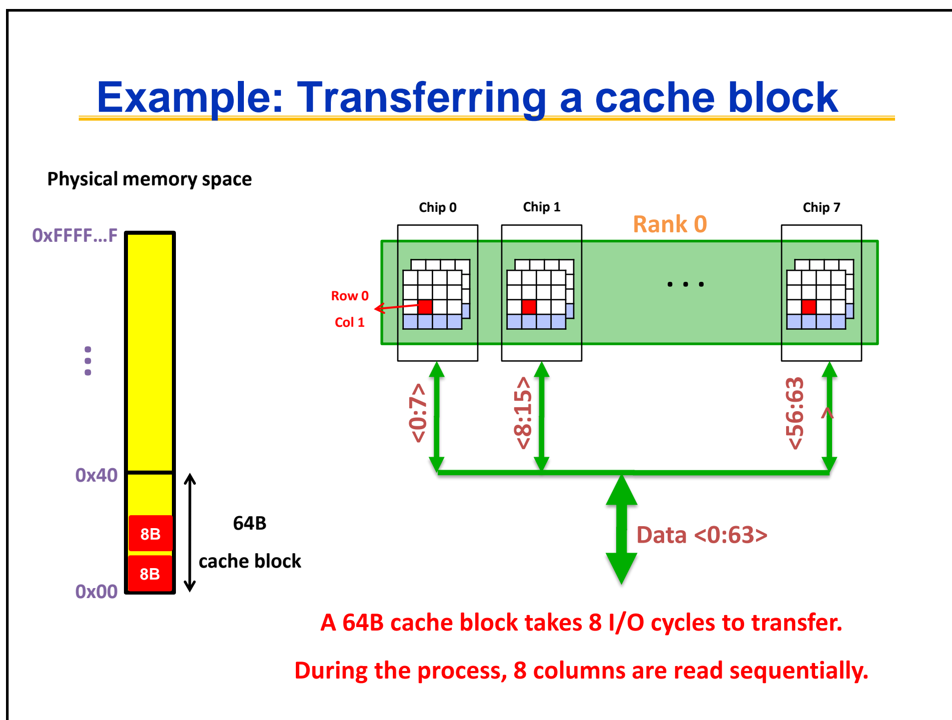


Example: Transferring a cache block

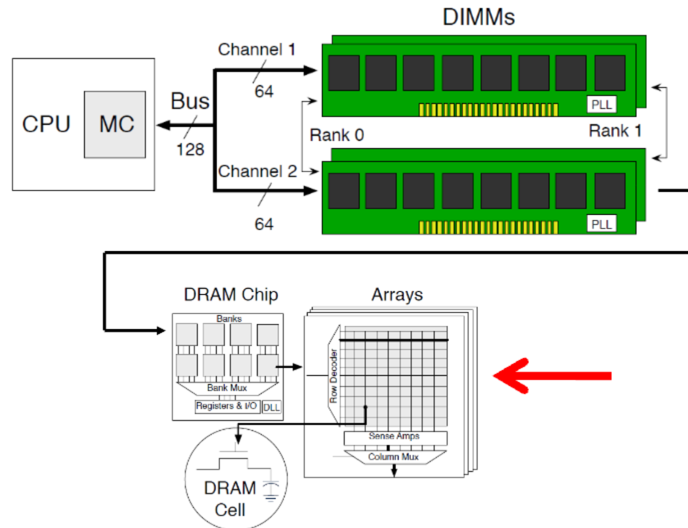


Example: Transferring a cache block

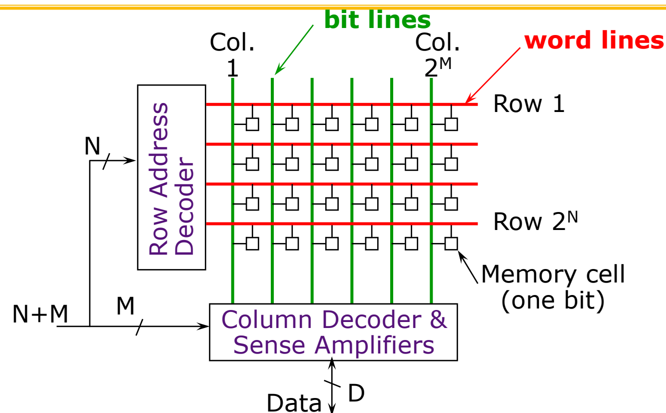




DRAM Overview

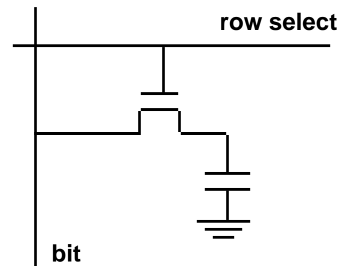


DRAM Architecture



- **Bits stored in 2-dimensional arrays on chip**
- **Modern chips have around 4 logical banks on each chip**
  - **each logical bank physically implemented as many smaller arrays**

1-T Memory Cell (DRAM)

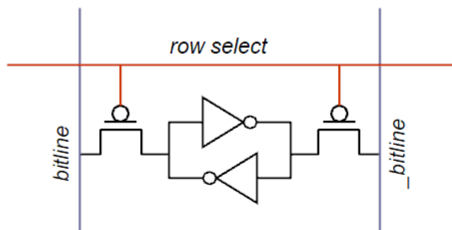


- **Write:**
  - 1. Drive bit line
  - 2.. Select row
- **Read:**
  - 1. Precharge bit line to Vdd/2
  - 2. Select row
  - 3. Storage cell shares charge with bitlines
    - » Very small voltage changes on the bit line
  - 4. Sense (fancy sense amp)
    - » Can detect changes of ~1 million electrons
  - 5. Write: restore the value
- **Refresh**
  - 1. Just do a dummy read to every cell.

21

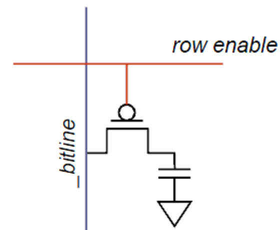
SRAM vs. DRAM

Static Random Access Mem.



- 6T vs. 1T1C
  - Large (~6-10x)
- Bitlines driven by transistors
  - Fast (~10x)

Dynamic Random Access Mem.



- Bits stored as charges on node capacitance (non-restorative)
  - Bit cell loses charge when read
  - Bit cell loses charge over time
- Must periodically refresh
  - Once every 10s of ms

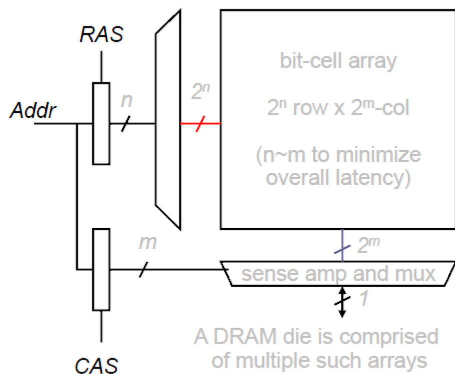
22

DRAM Operation: Three Steps

- **Precharge**
  - charges bit lines to known value, required before next row access
- **Row access (RAS)**
  - decode row address, enable addressed row (often multiple Kb in row)
  - Contents of storage cell share charge with bitlines
  - small change in voltage detected by sense amplifiers which latch whole row of bits
  - sense amplifiers drive bitlines full rail to recharge storage cells
- **Column access (CAS)**
  - decode column address to select small number of sense amplifier latches (4, 8, 16, or 32 bits depending on DRAM package)
  - on read, send latched bits out to chip pins
  - on write, change sense amplifier latches. which then charge storage cells to required value
  - can perform multiple column accesses on same row without another row access (burst mode)

23

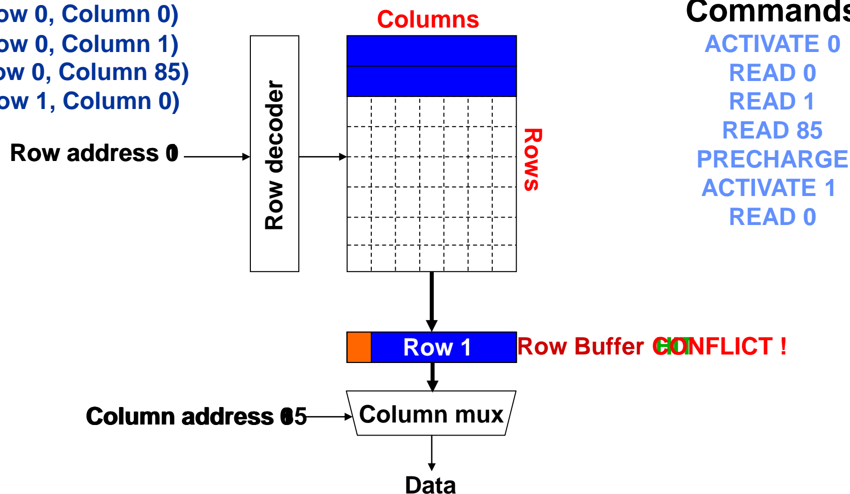
DRAM: Memory-Access Protocol



- 5 basic commands
  - ACTIVATE
  - READ
  - WRITE
  - PRECHARGE
  - REFRESH
- To reduce pin count, row and column share same address pins
  - RAS = Row Address Strobe
  - CAS = Column Address Strobe

24

DRAM Bank Operation

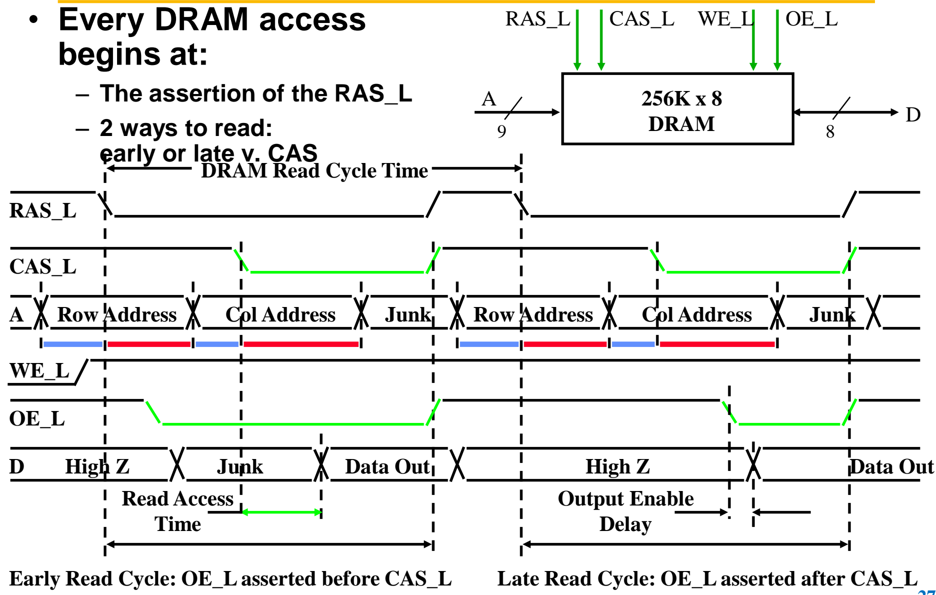


DRAM: Basic Operation

- Access to an "open row"
  - No need for ACTIVATE command
  - READ/WRITE to access row buffer
- Access to a "closed row"
  - If another row already active, must first issue PRECHARGE
  - ACTIVATE to open new row
  - READ/WRITE to access row buffer
  - Optional: PRECHARGE after READ/WRITEs finished

26

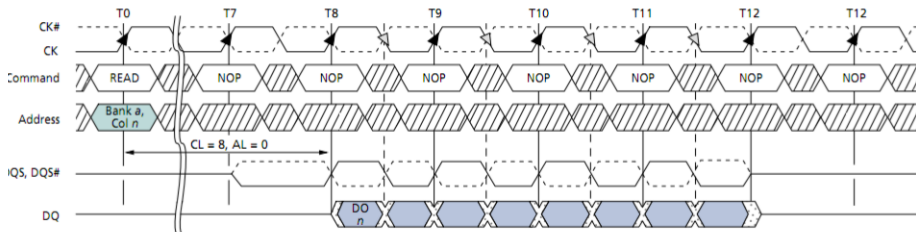
DRAM Read Timing (Example)



27

DRAM: Burst

- Each READ/WRITE command can transfer multiple words (8 in DDR3)
- DRAM channel clocked faster than DRAM core



- Critical word first?

28

DRAM: Banks

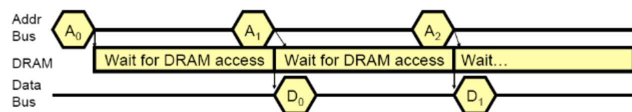
- Modern DRAM chips consist of multiple **banks**
  - Address = (Bank x, Row y, Column z)

- Banks operate independently, but share command/address/data pins
  - Each can have a different row active
  - Can overlap ACTIVATE and PRECHARGE latencies! (i.e. READ to bank 0 while ACTIVATING bank 1)

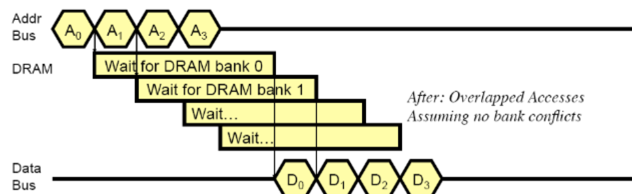
29

DRAM: Banks

- Enable concurrent DRAM accesses (overlapping)



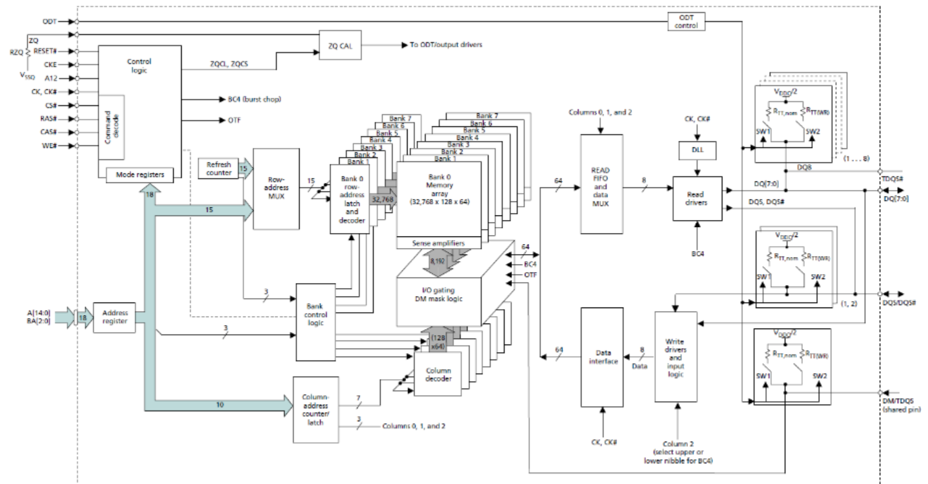
Before: No Overlapping
Assuming accesses to different DRAM rows



After: Overlapped Accesses
Assuming no bank conflicts

30

2Gb x8 DDR3 Chip [Micron]



Observe: bank organization

31

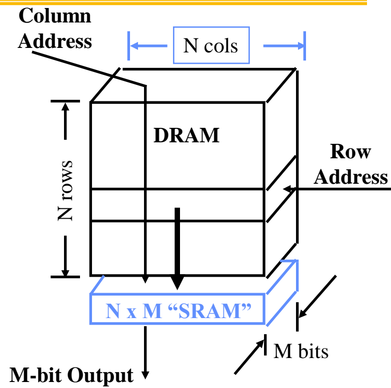
Quest for DRAM Performance

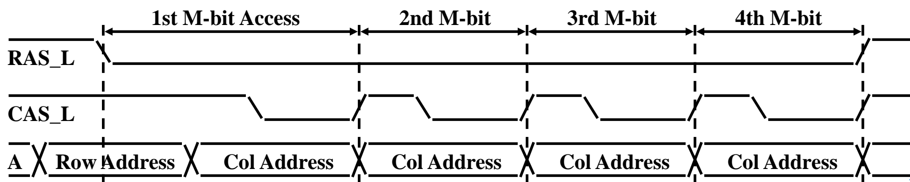
1. **Fast Page mode**
   - Add timing signals that allow repeated accesses to row buffer without another row access time
   - Such a buffer comes naturally, as each array will buffer 1024 to 2048 bits for each access
2. **Synchronous DRAM (SDRAM)**
   - Add a clock signal to DRAM interface, so that the repeated transfers would not bear overhead to synchronize with DRAM controller
3. **Double Data Rate (DDR SDRAM)**
   - Transfer data on both the rising edge and falling edge of the DRAM clock signal ⇒ doubling the peak data rate
   - DDR2 lowers power by dropping the voltage from 2.5 to 1.8 volts + offers higher clock rates: up to 400 MHz
   - DDR3 drops to 1.5 volts + higher clock rates: up to 800 MHz
   - DDR4 drops to 1-1.2 volts + higher clock rates: up to 1600 MHz

32

1. Fast Page Mode Operation

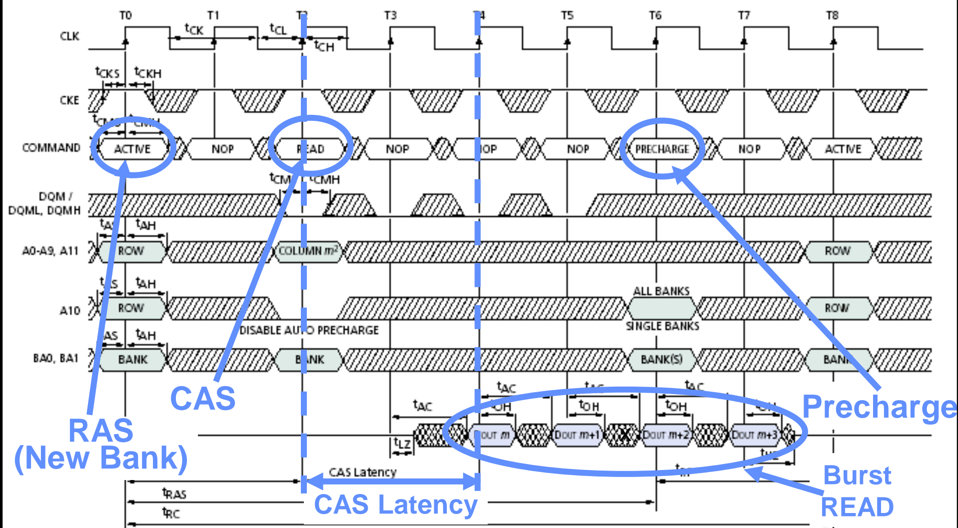
- **Regular DRAM Organization:**
  - N rows x N column x M-bit
  - Read & Write M-bit at a time
  - Each M-bit access requires a RAS / CAS cycle
- **Fast Page Mode DRAM**
  - N x M "SRAM" to save a row
- **After a row is read into the register**
  - Only CAS is needed to access other M-bit blocks on that row
  - RAS_L remains asserted while CAS_L is toggled





34

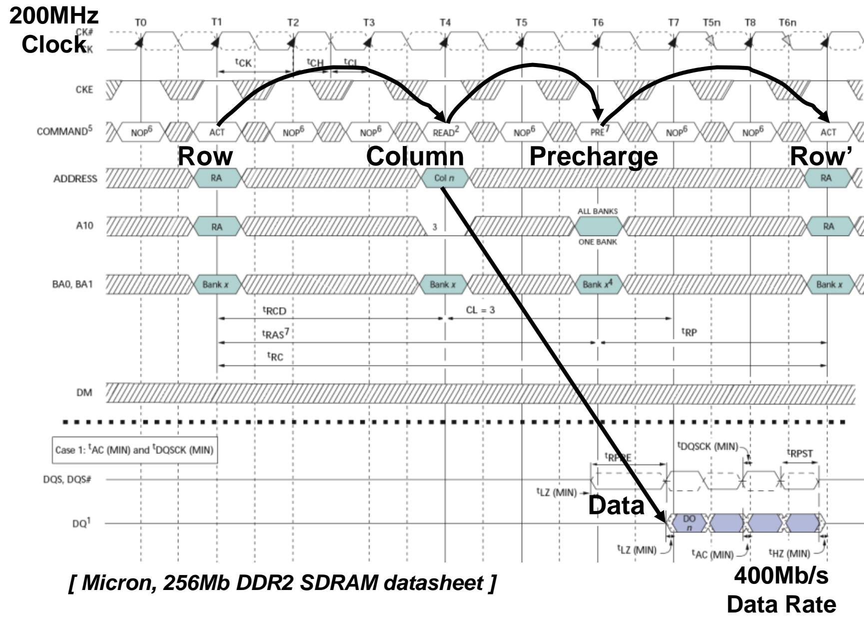
2. SDRAM timing (Single Data Rate)



- Micron 128M-bit dram (using 2Meg×16bit×4bank ver)
  - Row (12 bits), bank (2 bits), column (9 bits)

35

3. Double-Data Rate (DDR2) DRAM



[Micron, 256Mb DDR2 SDRAM datasheet]

36

Memory Organizations

Production year	Chip size	DRAM Type	Row access strobe (RAS) Slowest DRAM (ns)	Row access strobe (RAS) Fastest DRAM (ns)	Column access strobe (CAS)/ data transfer time (ns)	Cycle time (ns)
1980	64K bit	DRAM	180	150	75	250
1983	256K bit	DRAM	150	120	50	220
1986	1M bit	DRAM	120	100	25	190
1989	4M bit	DRAM	100	80	20	165
1992	16M bit	DRAM	80	60	15	120
1996	64M bit	SDRAM	70	50	12	110
1998	128M bit	SDRAM	70	50	10	100
2000	256M bit	DDR1	65	45	7	90
2002	512M bit	DDR1	60	40	5	80
2004	1G bit	DDR2	55	35	5	70
2006	2G bit	DDR2	50	30	2.5	60
2010	4G bit	DDR3	36	28	1	37
2012	8G bit	DDR3	30	24	0.5	31

Figure 2.13 Times of fast and slow DRAMs vary with each generation. (Cycle time is defined on page 95.) Performance improvement of row access time is about 5% per year. The improvement by a factor of 2 in column access in 1986 accompanied the switch from NMOS DRAMs to CMOS DRAMs. The introduction of various burst transfer modes in the mid-1990s and SDRAMs in the late 1990s has significantly complicated the calculation of access time for blocks of data; we discuss this later in this section when we talk about SDRAM access time and power. The DDR4 designs are due for introduction in mid- to late 2012. We discuss these various forms of DRAMs in the next few pages.

Memory Organizations

Standard	Clock rate (MHz)	M transfers per second	DRAM name	MB/sec /DIMM	DIMM name
DDR	133	266	DDR266	2128	PC2100
DDR	150	300	DDR300	2400	PC2400
DDR	200	400	DDR400	3200	PC3200
DDR2	266	533	DDR2-533	4264	PC4300
DDR2	333	667	DDR2-667	5336	PC5300
DDR2	400	800	DDR2-800	6400	PC6400
DDR3	533	1066	DDR3-1066	8528	PC8500
DDR3	666	1333	DDR3-1333	10,664	PC10700
DDR3	800	1600	DDR3-1600	12,800	PC12800
DDR4	1066–1600	2133–3200	DDR4-3200	17,056–25,600	PC25600

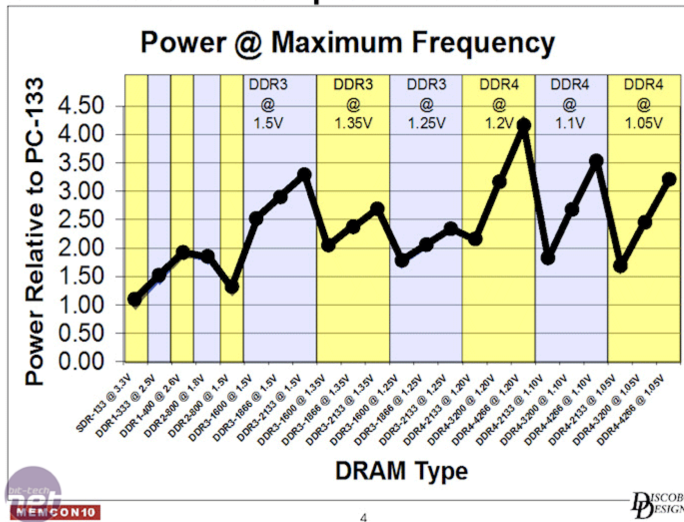
Figure 2.14 Clock rates, bandwidth, and names of DDR DRAMS and DIMMs in 2010. Note the numerical relationship between the columns. The third column is twice the second, and the fourth uses the number from the third column in the name of the DRAM chip. The fifth column is eight times the third column, and a rounded version of this number is used in the name of the DIMM. Although not shown in this figure, DDRs also specify latency in clock cycles as four numbers, which are specified by the DDR standard. For example, DDR3-2000 CL 9 has latencies of 9-9-9-28. What does this mean? With a 1 ns clock (clock cycle is one-half the transfer rate), this indicate 9 ns for row to columns address (RAS time), 9 ns for column access to data (CAS time), and a minimum read time of 28 ns. Closing the row takes 9 ns for precharge but happens only when the reads from that row are finished. In burst mode, transfers occur on every clock on both edges, when the first RAS and CAS times have elapsed. Furthermore, the precharge in not needed until the entire row is read. DDR4 will be produced in 2012 and is expected to reach clock rates of 1600 MHz in 2014, when DDR5 is expected to take over. The exercises explore these details further.

Graphics Memory

Product	Density	Banks	Part Num.	PKG & Speed	Org.	Interf.	Ref.	Voltage(V)	PKG.	PKG Type	Status
gDDR3 SDRAM	1Gb G-die	8Banks	K4W1G1646G	BC08/1A 11/12/15	64Mx16	SSTL_15	8K/64ms	1.5V ± 0.075V	96ball FBGA	Halogen-Free, Lead-Free & Flip-Chip	Mass Production
	2Gb C-die		K4W2G1646C	HC1A/11 12/15	128Mx16					Halogen-Free & Lead-Free	Mass Production
	DDP 4Gb D-die		K4W4G1646D	BC12	256Mx16						CS Jan.'11
GDDR3 SGRAM	512Mb I-die	8Banks	K4J52324KI	HC7A/08 1A/12/14	16Mx32	POD_18	8K/32ms	1.8V ± 0.1V	136ball FBGA	Halogen-Free & Lead-Free	Mass Production
	1Gb G-die		K4J10324KG	HC1A/14	32Mx32						CS Aug.'11
GDDR5 SGRAM	1Gb G-die	16Banks	K4G10325FG	HC03/04/05	32Mx32	POD_15	8K/32ms	1.5V ± 0.045V	170ball FBGA	Halogen-Free & Lead-Free	Mass Production
	2Gb C-die		K4G20325FC	HC03/04/05	64Mx32	POD_15	16K/32ms	1.5V ± 0.045V	170ball FBGA	Halogen-Free & Lead-Free	Mass Production

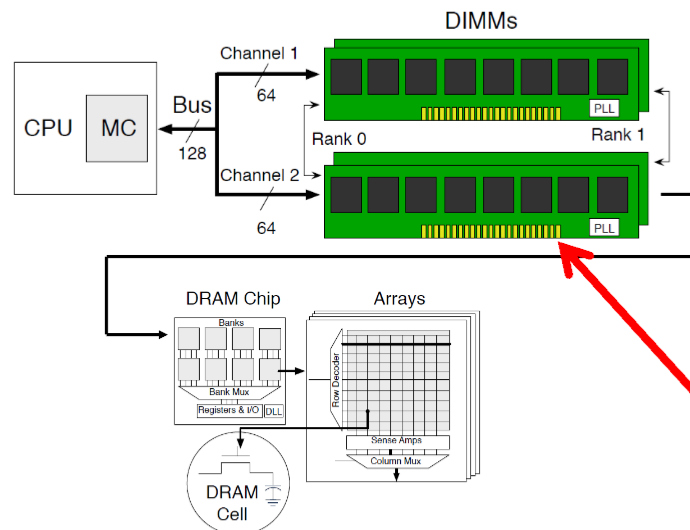
- **Achieve 2-5 X bandwidth per DRAM vs. DDR3**
  - **Wider interfaces (32 vs. 16 bit)**
  - **Higher clock rate**
    - » **Possible because they are attached via soldering instead of socketted DIMM modules**
  - **E.g. Samsung GDDR5**
    - » **2.5GHz, 20 GBps bandwidth**

DRAM Power: Not always up, but…



41

DRAM Modules



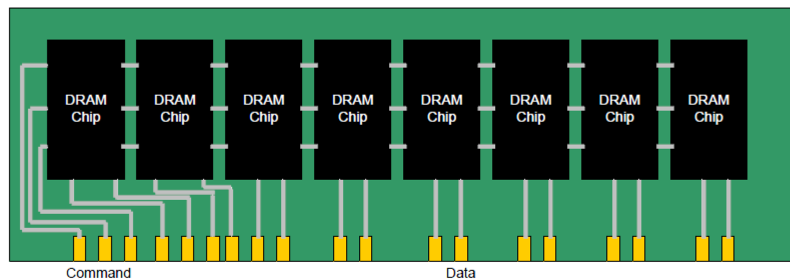
42

DRAM Modules

- DRAM chips have narrow interface (typically x4, x8, x16)
- Multiple chips are put together to form a wide interface
  - DIMM: Dual Inline Memory Module
  - To get a 64-bit DIMM, we need to access 8 chips with 8-bit interfaces
  - Share command/address lines, but not data
- Advantages
  - Acts like a high-capacity DRAM chip with a wide interface
    - 8x capacity, 8x bandwidth, same latency
- Disadvantages
  - Granularity: Accesses cannot be smaller than the interface width
    - 8x power

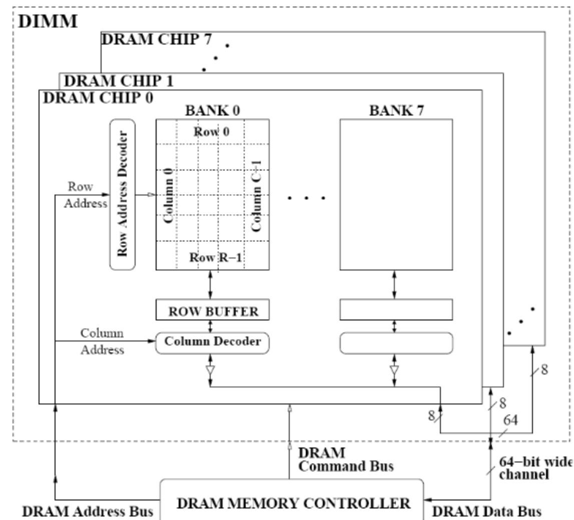
43

A 64-bit Wide DIMM (physical view)



44

A 64-bit Wide DIMM (logical view)



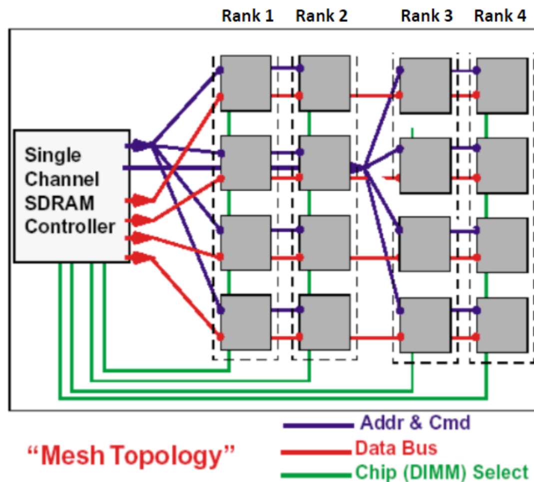
45

DRAM Ranks

- A DIMM may include multiple Ranks
  - A 64-bit DIMM using 8 chips with x16 interfaces has 2 ranks
- Each 64-bit group of chips is called a rank
  - All chips in a rank respond to a single command
  - Different ranks share command/address/data lines
    - Select between ranks with “Chip Select” signal
  - Ranks provide more “banks” across multiple chips (but don’t confuse rank and bank!)

46

Multiple DIMMs on a Channel



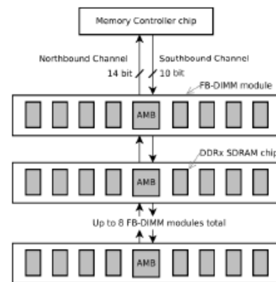
"Mesh Topology"

- Advantages:
  - Enables even higher capacity
- Disadvantages:
  - Interconnect latency, complexity, and energy get higher
  - Addr/Cmd signal integrity is a challenge

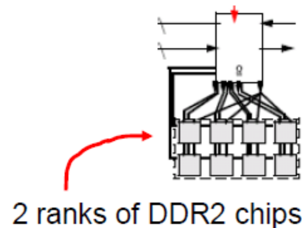
47

Fully Buffered DIMM (FB-DIMM)

- **DDR Problem**
  - **Higher capacity → more DIMMs → lower data rate (multidrop bus)**
- **FB-DIMM approach: use point to point links**
  - **introduces an *advanced memory buffer* (AMB) between memory controller and memory module**
  - **Serial interface between mem controller and AMB**
  - **enables an increase to the width of the memory without increasing the pin count of the memory controller**



Advanced Memory Buffer



2 ranks of DDR2 chips

48

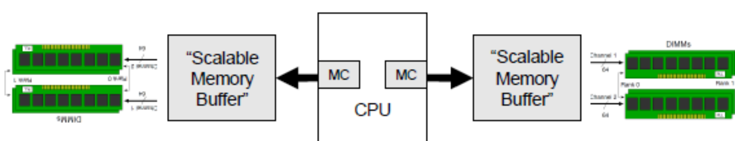
FB-DIMM challenges

- AMB is big, expensive, and power hungry
  - Low volume, so FB-DIMM modules are expensive
- Daisy chain adds significant latency
  - 8 slot FB-DIMM channels are slow
- Requires FB-DIMM memory controller
  - Incompatible with on-chip DDR3 controllers!

- **As of Sep 2006, AMD has taken FB-DIMM off their roadmap**
- **In 2007 it was revealed that major memory manufacturers have no plans to extend FB-DIMM to support DDR3 SDRAM**
  - **Instead, only registered DIMM for DDR3 SDRAM had been demonstrated**
  - **In normal registered/buffered memory, only the control lines are buffered whereas in fully buffered memory, the data lines are buffered as well**
  - **Both FB and "registered" options increase latency and are costly**

49

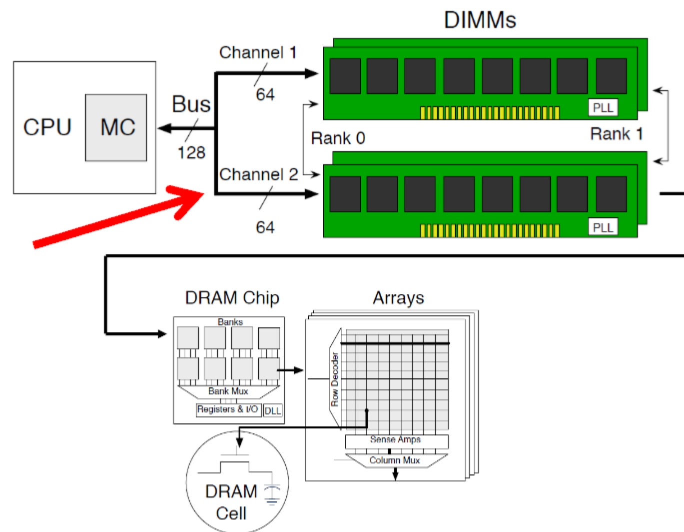
Intel Scalable Memory Buffer



- High-speed serial link from CPU to SMB
  - Two DDR3 channels behind SMB (2 slots per channel)
  - Can use commodity DDR3 modules
  - Mitigates pin-count on CPU
- On-chip MC manages DRAM access protocol
- Jury still out on the right design

50

DRAM Channels



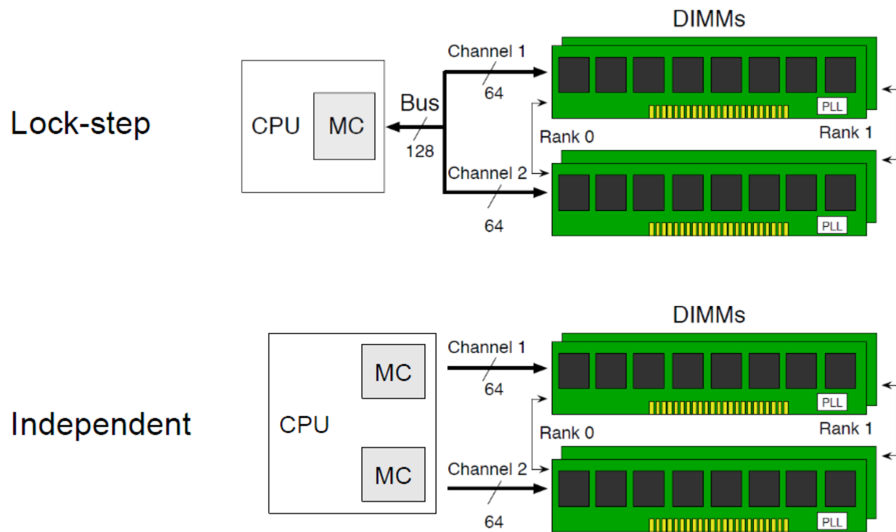
52

DRAM Channels

- Channel: a set of DIMMs in series
  - All DIMMs get the same command, one of the ranks replies
- System options
  - Single channel system
  - Multiple dependent (lock-step) channels
    - Single controller with wider interface (faster cache line refill!)
    - Sometimes called “Gang Mode”
    - Only works if DIMMs are identical (organization, timing)
  - Multiple independent channels
    - Requires multiple controllers
- Tradeoffs
  - Cost: pins, wires, controller
  - Benefit: higher bandwidth, capacity, flexibility

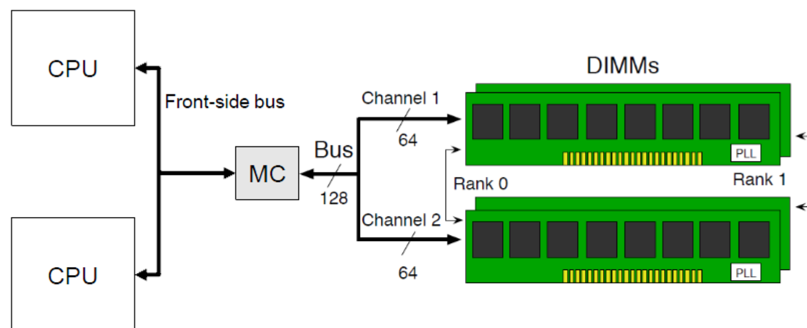
53

DRAM Channel Options



54

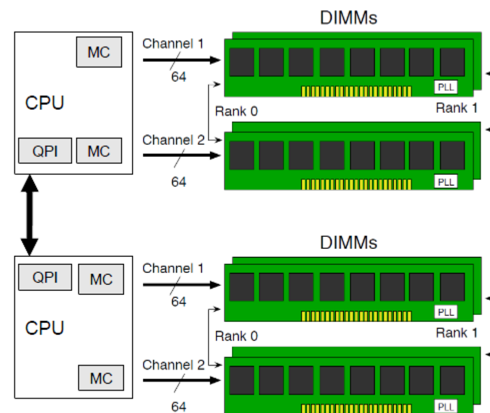
Multi-CPU (old school)



- External MC adds latency
- Capacity doesn't grow w/ # of CPUs

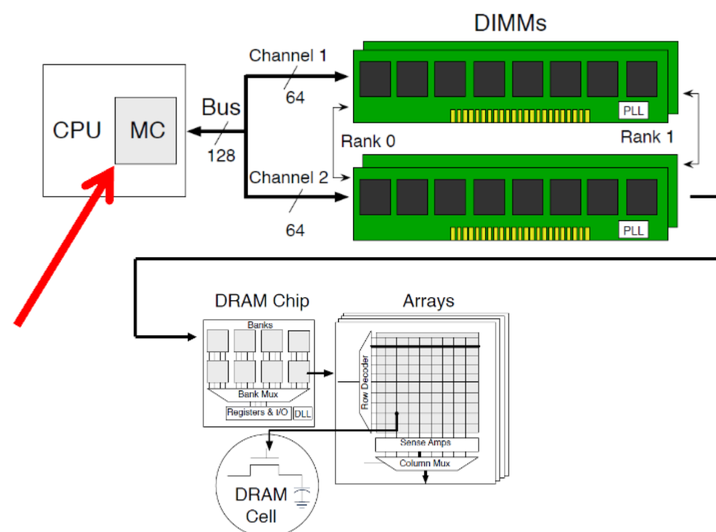
55

NUMA Topology (modern)



- Capacity grows w/ # of CPUs
- NUMA: "Non-uniform Memory Access"

Memory Controller



DRAM: Timing Constraints

- Memory controller must respect physical device characteristics
  - **tRCD** = Row to Column command delay
    - How long it takes row to get to sense amps
  - **tCAS** = Time between column command and data out
  - **tCCD** = Time between column commands
    - Rate that you can pipeline column commands
  - **tRP** = Time to precharge DRAM array
  - **tRAS** = Time between RAS and data restoration in DRAM array (minimum time a row must be open)
  - **tRC** = tRAS + tRP = Row "cycle" time
    - Minimum time between accesses to different rows

58

DRAM: Timing Constraints

- There are dozens of these…
  - tWTR = Write to read delay
  - tWR = Time from end of last write to PRECHARGE
  - tFAW = Four ACTIVATE window (limits current surge)
- Makes performance analysis, memory controller design difficult
- Datasheets for DRAM devices freely available
  - http://download.micron.com/pdf/datasheets/dram/ddr3/2Gb_DDR3_SDRAM.pdf

59

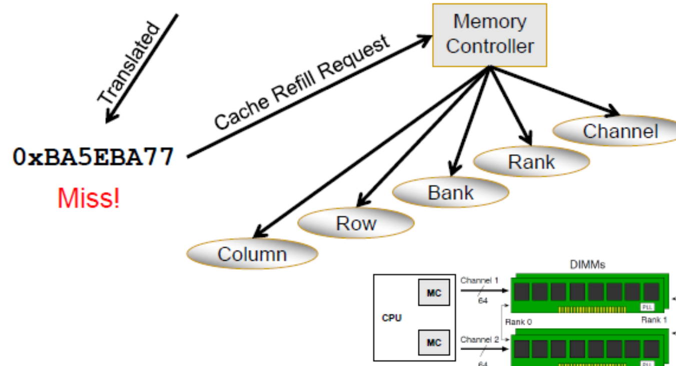
Latency Components: Basic DRAM Operation

- CPU → controller transfer time
- Controller latency
  - Queuing & scheduling delay at the controller
  - Access converted to basic commands
- DRAM bank latency
  - tCAS is row is "open" OR
  - tRCD + tCAS if array precharged OR
  - tRP + tRCD + tCAS (worst case: tRC + tRCD + tCAS)
- DRAM data transfer time
  - BurstLen / (MT/s) ← **500 MHz DDR = 1000 MT/s**
- Controller → CPU transfer time

60

DRAM Addressing

LD R1, Mem[foo]



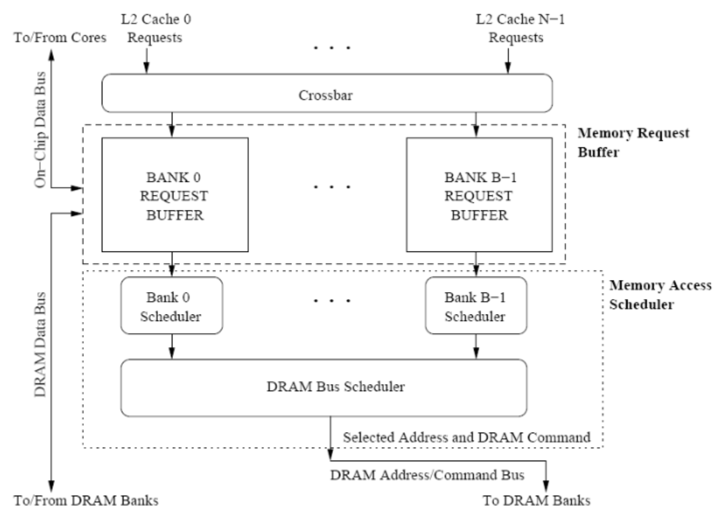
61

DRAM Controller Functionality

- Translate memory requests into DRAM command sequences
  - Map "Physical Address" to DRAM Address
  - Obey timing constraints of DRAM, arbitrate resource conflicts (i.e. bank, channel)
- Buffer and schedule requests to improve performance
  - Row-buffer management and re-ordering
- Ensure correct operation of DRAM (refresh)
- Manage power consumption and thermals in DRAM
  - Turn on/off DRAM chips, manage power modes

62

A Modern DRAM Controller



63

Row Buffer Management Policies

- Open row
  - Keep the row open after an access
  - Pro: Next access might need the same row → row hit
  - Con: Next access might need a different row → row conflict, wasted energy
- Closed row
  - Close the row after an access (if no other requests already in the request buffer need the same row)
  - Pro: Next access might need a different row → avoid a row conflict
  - Con: Next access might need the same row → extra activate latency
- Adaptive policies
  - Predict whether or not the next access to the bank will be to the same row

68

DRAM Controller Scheduling Policies (I)

- FCFS (first come first served)
  - Oldest request first
- FR-FCFS (first ready, first come first served)
  - 1. Row-hit first
  - 2. Oldest first
  - Goal: Maximize row buffer hit rate → maximize DRAM throughput

70

DRAM Controller Scheduling Policies (II)

- A scheduling policy is a prioritization order

- Prioritization can be based on
  - Request age
  - Row buffer hit/miss status
  - Request type (prefetch, read, write)
  - Requestor type (load miss or store miss)
  - Request criticality
    - Oldest miss in the core?
    - How many instructions in core are dependent on it?

71

DRAM Refresh (I)

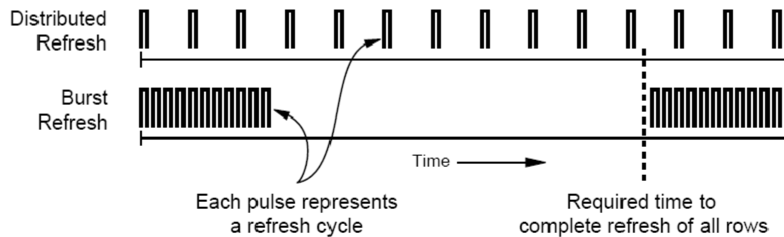
- DRAM capacitor charge leaks over time

- The memory controller needs to read each row periodically to restore the charge
  - Activate + precharge each row every N ms
  - Typical N = 64 ms

- Implications on performance?
  - DRAM bank unavailable while refreshed
  - Long pause times: If we refresh all rows in burst, every 64ms the DRAM will be unavailable until refresh ends

72

DRAM Refresh (II)



- Distributed refresh eliminates long pause times
- How else we can reduce the effect of refresh on performance?
  - Can we reduce the number of refreshes?

73

DRAM Controllers are Difficult to Design

- Need to obey DRAM timing constraints for correctness
  - There are many (50+) timing constraints in DRAM
- Need to keep track of many resources to prevent conflicts
  - Channels, banks, ranks, data bus, address bus, row buffers
- Need to handle DRAM refresh
- Need to optimize for performance (in the presence of constraints)
  - Reordering is not simple
  - Predicting the future?

74

DRAM Power Management

- DRAM chips have power modes
- Idea: When not accessing a chip power it down

- Power states
  - Active (highest power)
  - All banks idle (i.e. precharged)
  - Power-down
  - Self-refresh (lowest power)

- State transitions incur latency during which the chip cannot be accessed

75

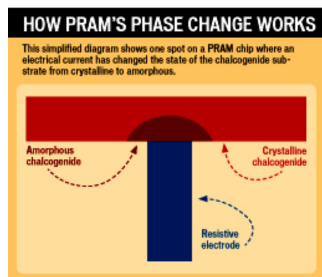
DRAM Reliability

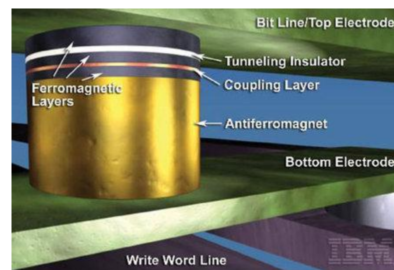
- DRAMs are susceptible to soft and hard errors
- Dynamic errors can be
  - detected by parity bits
    - usually 1 parity bit per 8 bits of data
  - detected and fixed by the use of Error Correcting Codes (ECCs)
    - E.g. SECDED Hamming code can detect two errors and correct a single error with a cost of 8 bits of overhead per 64 data bits
- In very large systems, the possibility of multiple errors as well as complete failure of a single memory chip becomes significant
  - Chipkill was introduced by IBM to solve this problem
  - Similar in nature to the RAID approach used for disks
  - Chipkill distributes data and ECC information, so that the complete failure of a single memory chip can be handled by supporting the reconstruction of the missing data from the remaining memory chips
  - IBM and SUN servers and Google Clusters use it
  - Intel calls their version SDDC

76

Looking Forward

- Continued slowdown in both density and access time of DRAMs → new DRAM that does not require a capacitor?
  - Z-RAM prototype from Hynix
- MRAMs → use magnetic storage of data; nonvolatile
- PRAMs → phase change RAMs (aka PCRAM, PCME)
  - use a glass that can be changed between amorphous and crystalline states; nonvolatile





77