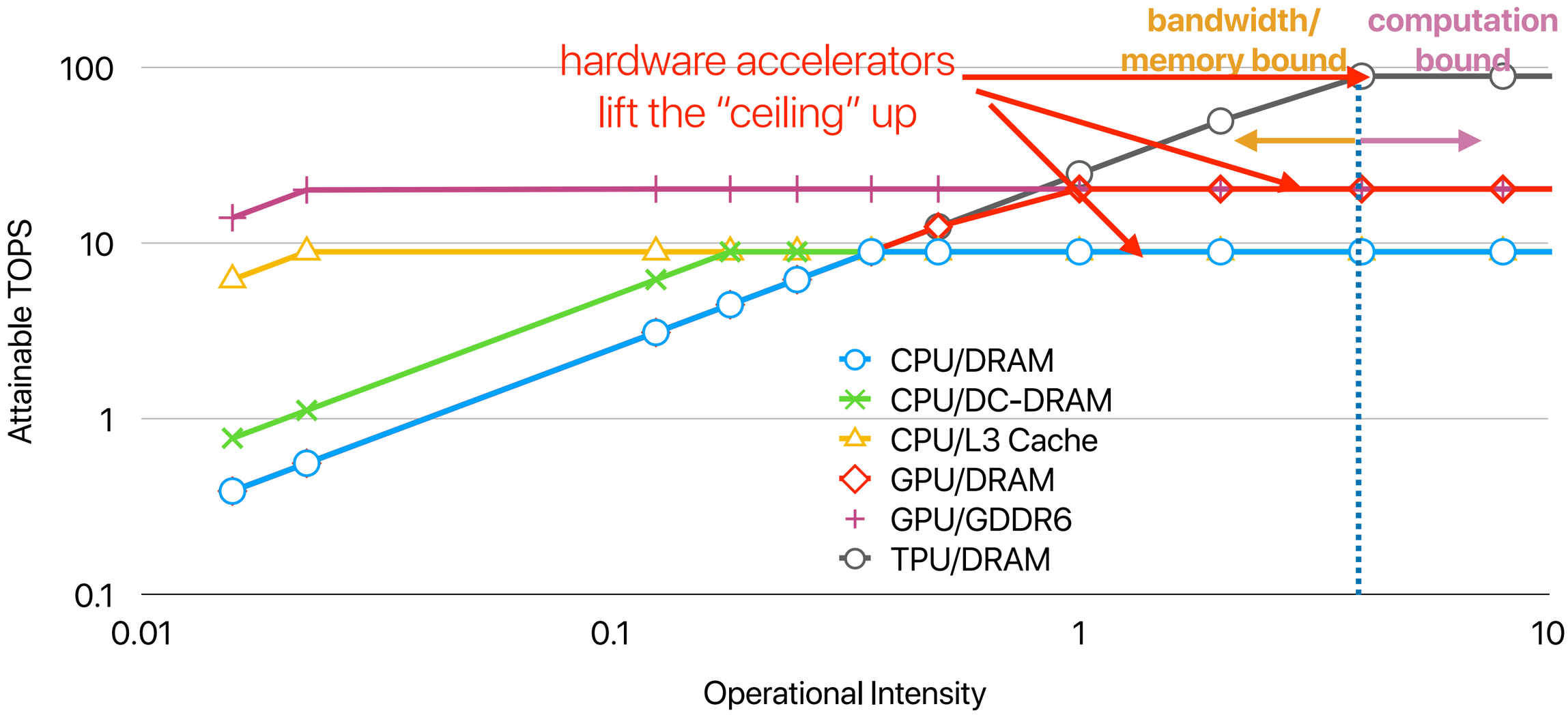
Modern Heterogeneous Computers: (3) Memory Components

Hung-Wei Tseng

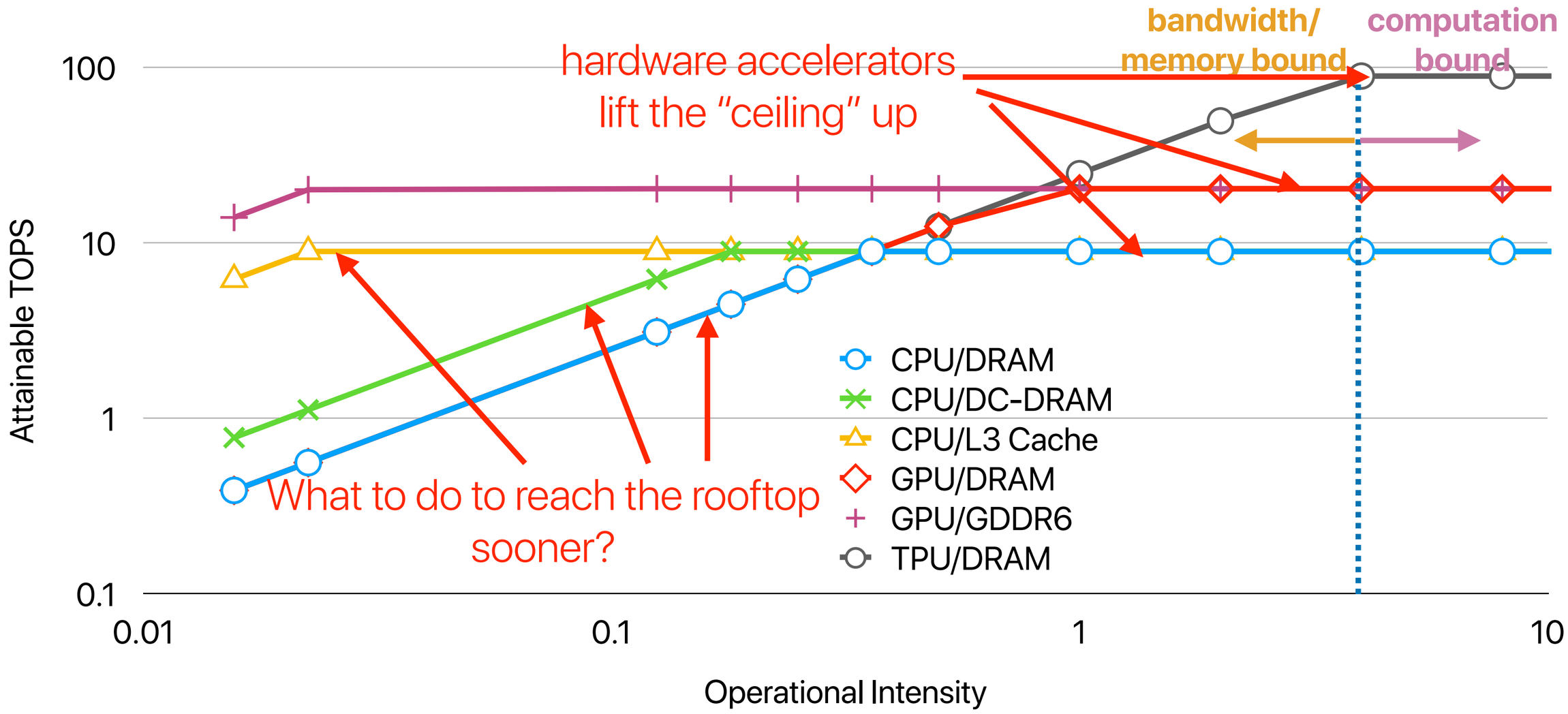
Recap: the roofline after using hardware accelerators



Lessons learned from accelerators?

- Accelerators “lift up” the roofline
  - Applications/compute kernels with higher arithmetic densities may be feasible
    - NN is feasible after GPGPU
    - Trade “complexity” with parallelism
  - Applications are more likely to be **memory**-bound
  - Your software should try to avoid frequent **memory** access
    - Try to use memory closer to the processing elements
  - The hardware design must not ignore the importance of **memory** bandwidth
- The most “efficient” system design must land on the “turning point” of your roofline model
  - TPU’s 167GB/sec **memory** bandwidth is an example

Recap: the roofline after using hardware accelerators



3:00

How can you shift roofline?

Example — matrix multiplications (cont.)

- Remember that we have a CPU core supporting 9G operations per second (OPS)
- If data are currently stored in SSD
  - We can supply data in 4GB/sec
  - The supplied data per second needs
    $$4GB \times \frac{1\ OPS}{8\ B} = 0.5\ GOPS < 9G$$
- If data are currently stored in DRAM
  - We can supply data in 25GB/sec per module
  - The supplied data per second needs per module
    $$25GB \times \frac{1\ OPS}{8\ B} = 3.12500\ GOPS < 9G$$
  - If we have 2 modules
    $$50GB \times \frac{1\ OPS}{8\ B} = 7.25\ GOPS < 9G$$

JEDEC standard DDR4 module [edit]

CAS latency (CL)

Clock cycles between sending a column address to the memory and the beginning of the data in response

tRCD

Clock cycles between row activate and reads/writes

tRP

Clock cycles between row precharge and activate

DDR4-xxxx denotes per-bit data transfer rate, and is normally used to describe DDR chips. PC4-xxxxx denotes overall transfer rate, in megabytes per second, and applies only to modules (assembled DIMMs). Because DDR4 memory modules transfer data on a bus that is 8 bytes (64 data bits) wide, module peak transfer rate is calculated by taking transfers per second and multiplying by eight.[60]

Size, Latency, and Bandwidth of Memory Subsystem Components

Assuming you have a large processor (about 16 cores), the following summarizes, for 2016, approximate data totals present in and moving through the system.

Memory	Size	Latency	Bandwidth
L1 cache	32 KB	1 nanosecond	1 TB/second
L2 cache	256 KB	4 nanoseconds	1 TB/second Sometimes shared by two cores
L3 cache	8 MB or more	10x slower than L2	>400 GB/second
MCDRAM		2x slower than L3	400 GB/second
Main memory on DDR DIMMs	4 GB-1 TB	Similar to MCDRAM	100 GB/second
Main memory on Cornelis* Omni-Path Fabric	Limited only by cost	Depends on distance	Depends on distance and hardware
I/O devices on memory bus	6 TB	100x-1000x slower than memory	25 GB/second
I/O devices on PCIe bus	Limited only by cost	From less than milliseconds to minutes	GB-TB/hour Depends on distance and hardware

https://www.intel.com/content/www/us/en/developer/articles/technical/memory-performance-in-a-nutshell.html

Example — matrix multiplications (cont.)

- Remember that we have a CPU core supporting 9G operations per second (OPS)
- If data are currently stored in SSD
  - We can supply data in 4GB/sec
  - The supplied data per second needs
    $4GB/s \times \frac{1\ OPS}{8\ B} = 0.5\ GOPS < 9G$
- If data are currently stored in DRAM
  - We can supply data in 25GB/sec per module
  - The supplied data per second needs per module
    $25GB/s \times \frac{1\ OPS}{8\ B} = 3.12500\ GOPS < 9G$
  - If we have 2 modules
    $50GB/s \times \frac{1\ OPS}{8\ B} = 7.25\ GOPS < 9G$
- If we can use cache "perfectly"
  - $400GB/s \times \frac{1\ OPS}{8\ B} = 50\ GOPS > 9G$

JEDEC standard DDR4 module [edit]

CAS latency (CL)

Clock cycles between sending a column address to the memory and the beginning of the data in response

tRCD

Clock cycles between row activate and reads/writes

tRP

Clock cycles between row precharge and activate

DDR4-xxxx denotes per-bit data transfer rate, and is normally used to describe DDR chips. PC4-xxxxx denotes overall transfer rate, in megabytes per second, and applies only to modules (assembled DIMMs). **Because DDR4 memory modules transfer data on a bus that is 8 bytes (64 data bits) wide, module peak transfer rate is calculated by taking transfers per second and multiplying by eight.**[60]

Shifting the roofline

Shifting the roofline

- Higher memory bandwidth
- Lower data volume

3:00

What kinds of memory technologies are presented in modern computer systems? Strength? Weakness?

Volatile v.s. Non-volatile

- Volatile memory
  - The stored bits will vanish if the cell is not supplied with electricity
  - Register, SRAM, DRAM
- Non-volatile memory
  - The stored bits will not vanish “immediately” when it’s out of electricity — usually can last years
  - Flash memory, PCM, MRAM, STTRAM

Memory technologies we have today

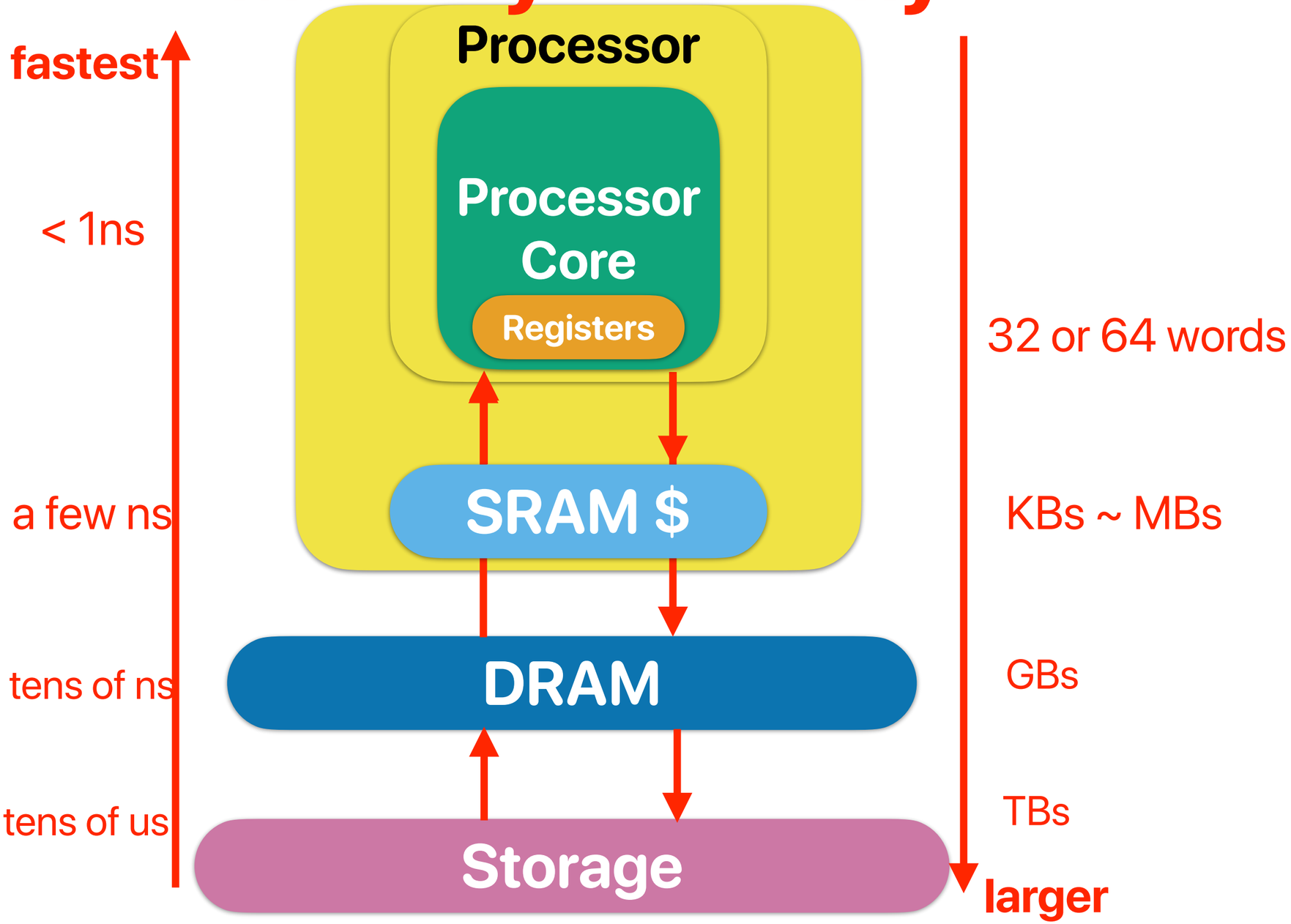
Memory technologies we have today

- SRAM
- DRAM
- Registers
- PCM
- 3DXPoint
- RRAM
- Flash memory

Memory technologies we have today

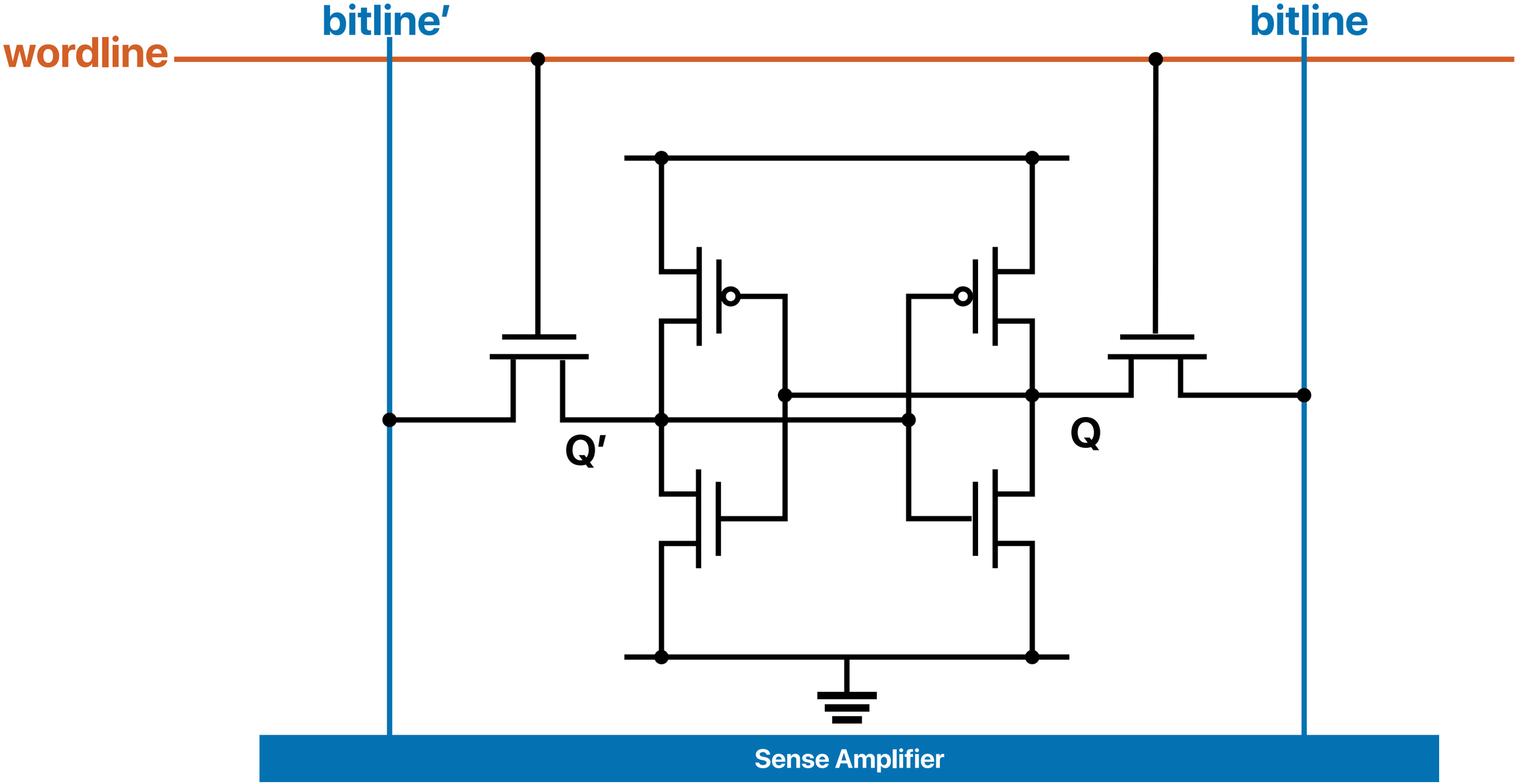
	Volatile Memory	Non-Volatile Memory
100ps	SRAM	
ns	DRAM	RRAM
us		PCM 3DXPoint Flash memory
ms		Hard Disk Drives

Memory Hierarchy

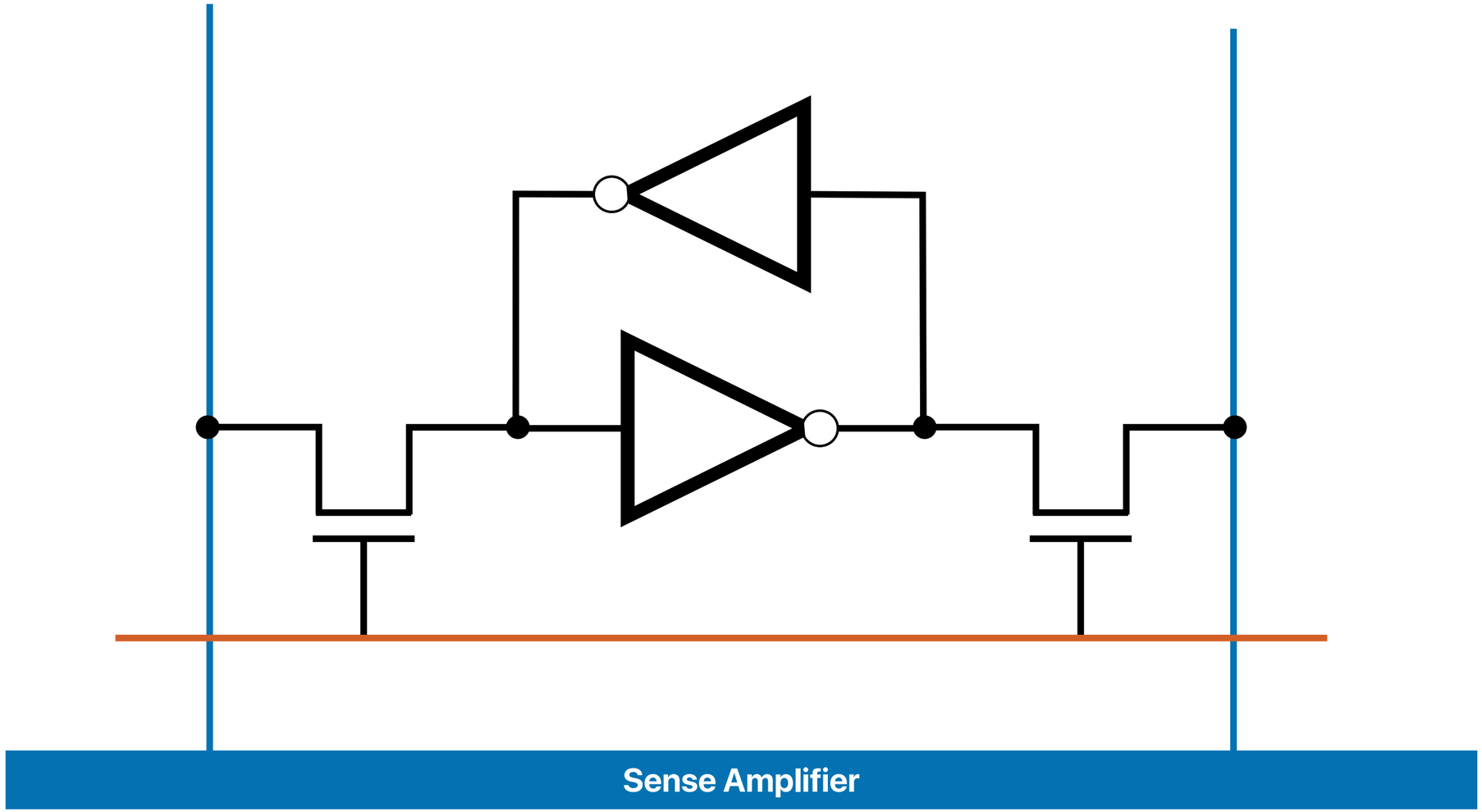


Static Random Access Memory (SRAM)

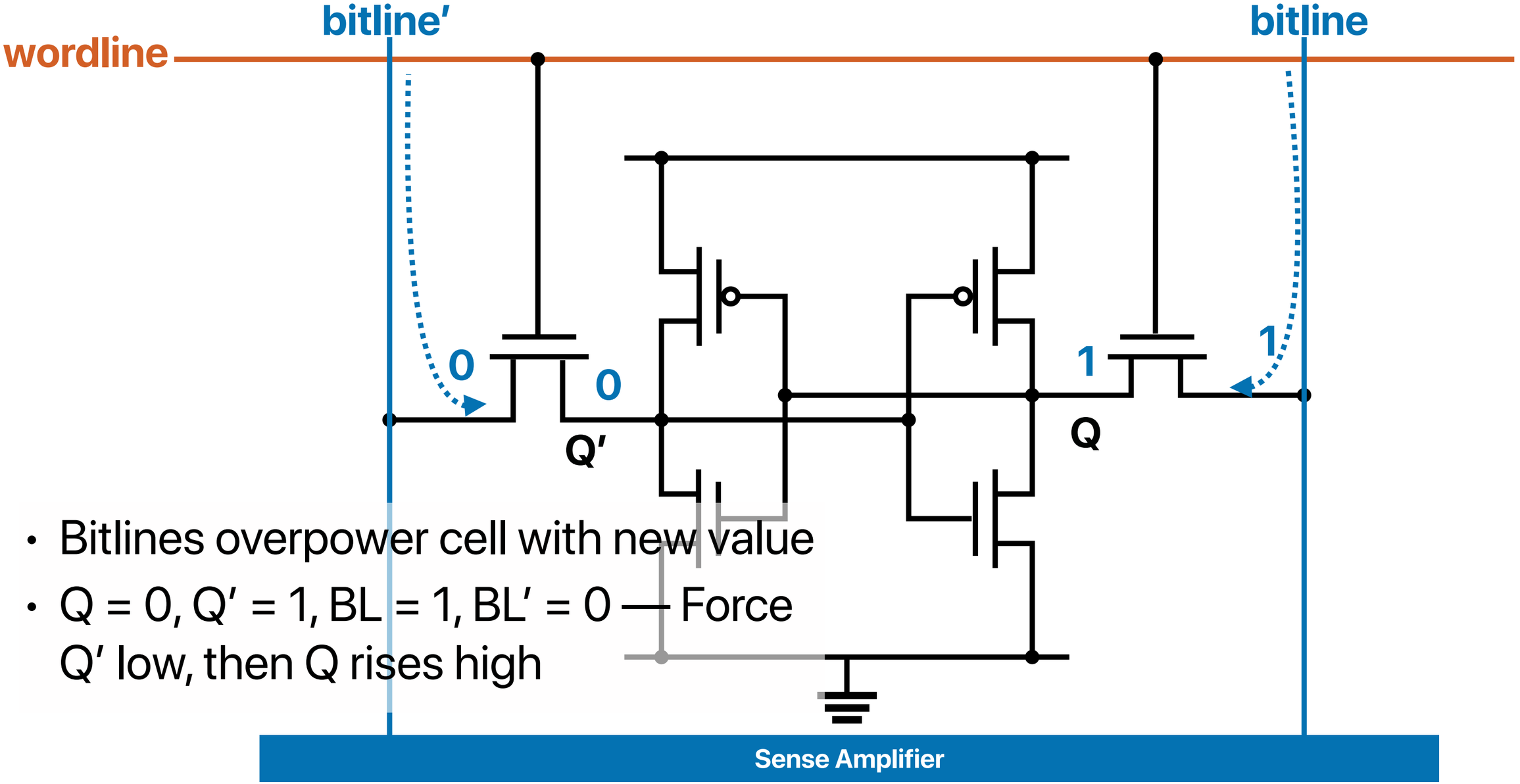
A Classical 6-T SRAM Cell



A Classical 6-T SRAM Cell

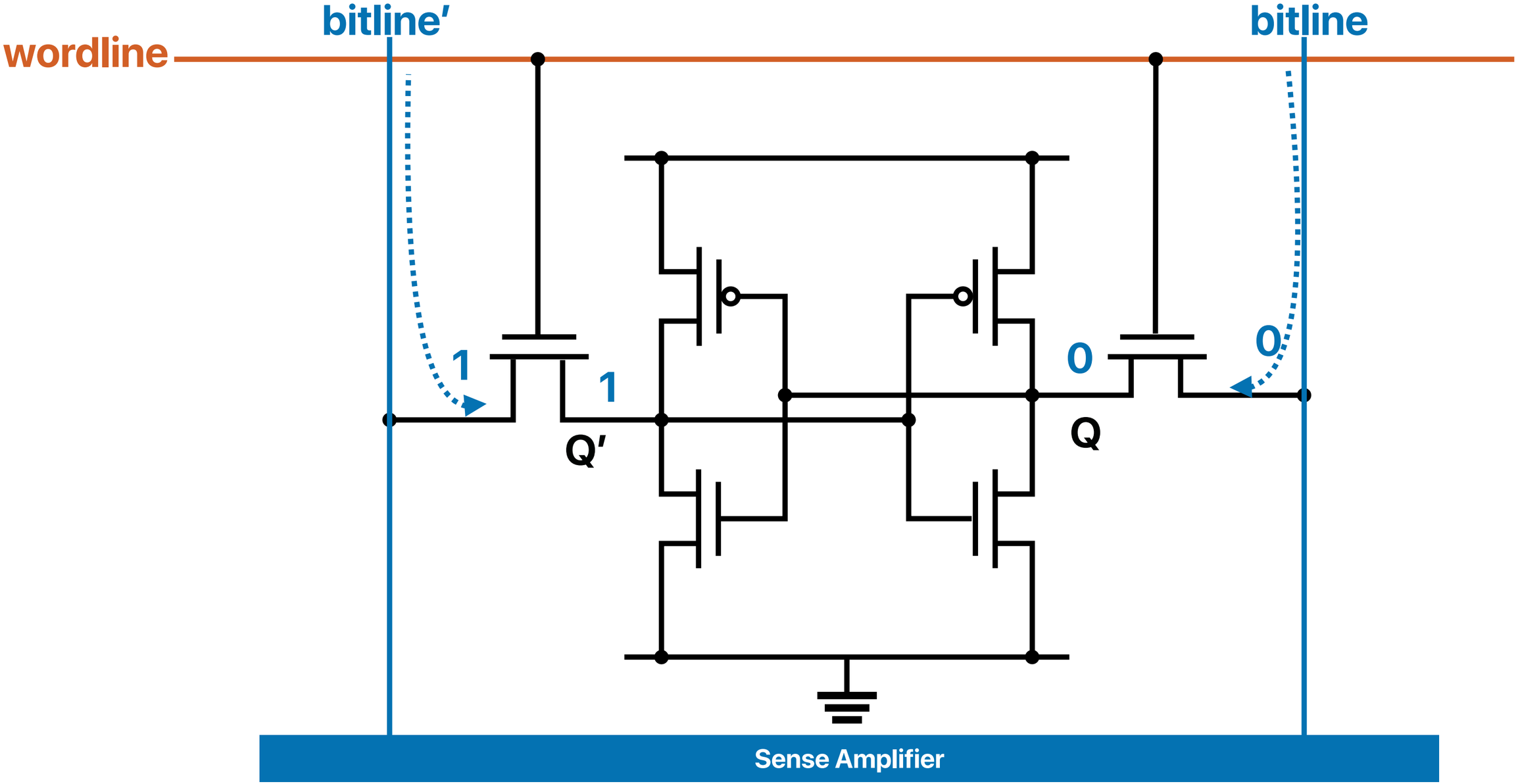


Write “1” to an SRAM Cell

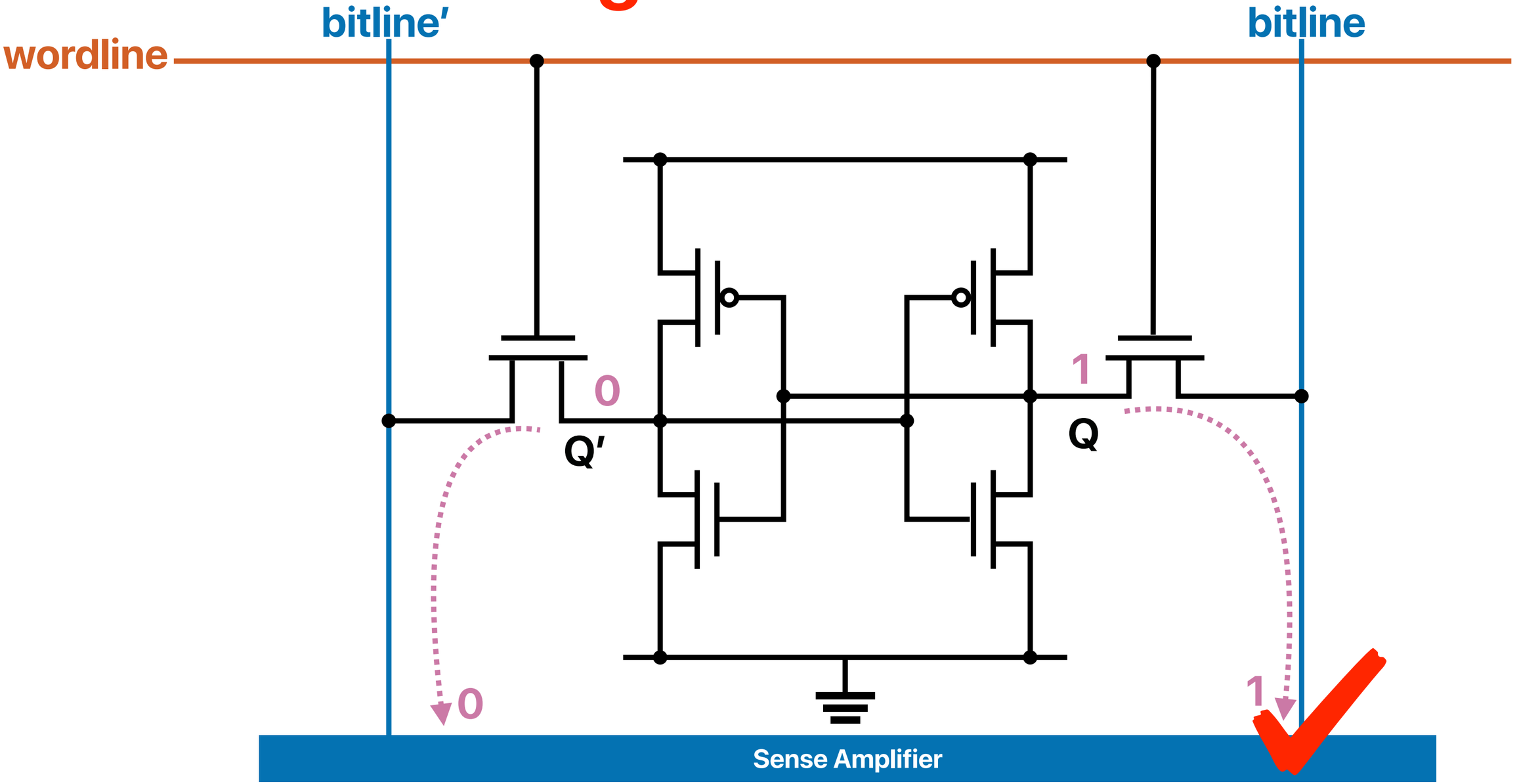


- Bitlines overpower cell with new value
- Q = 0, Q’ = 1, BL = 1, BL’ = 0 — Force Q’ low, then Q rises high

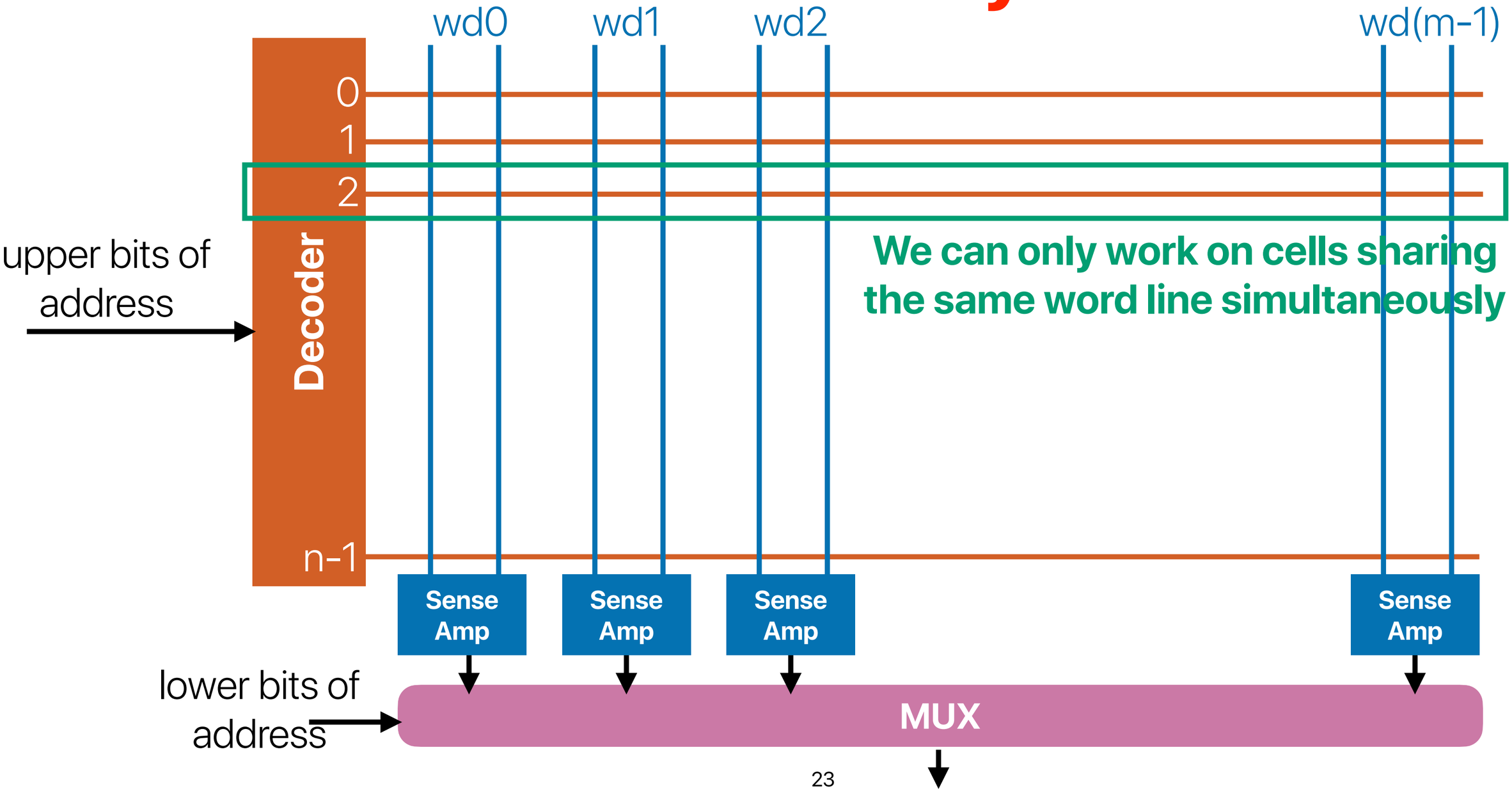
Write “0” to an SRAM Cell



Reading from an SRAM Cell

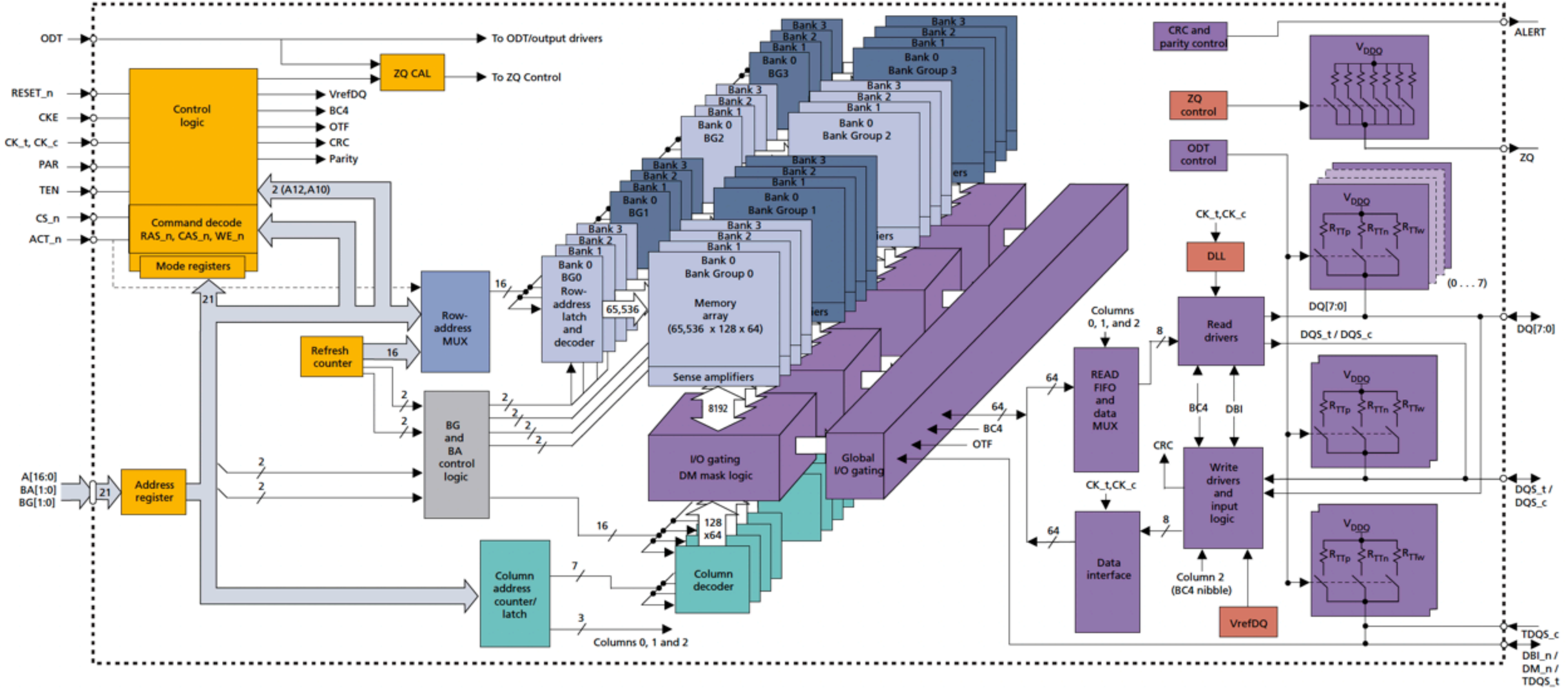


SRAM array

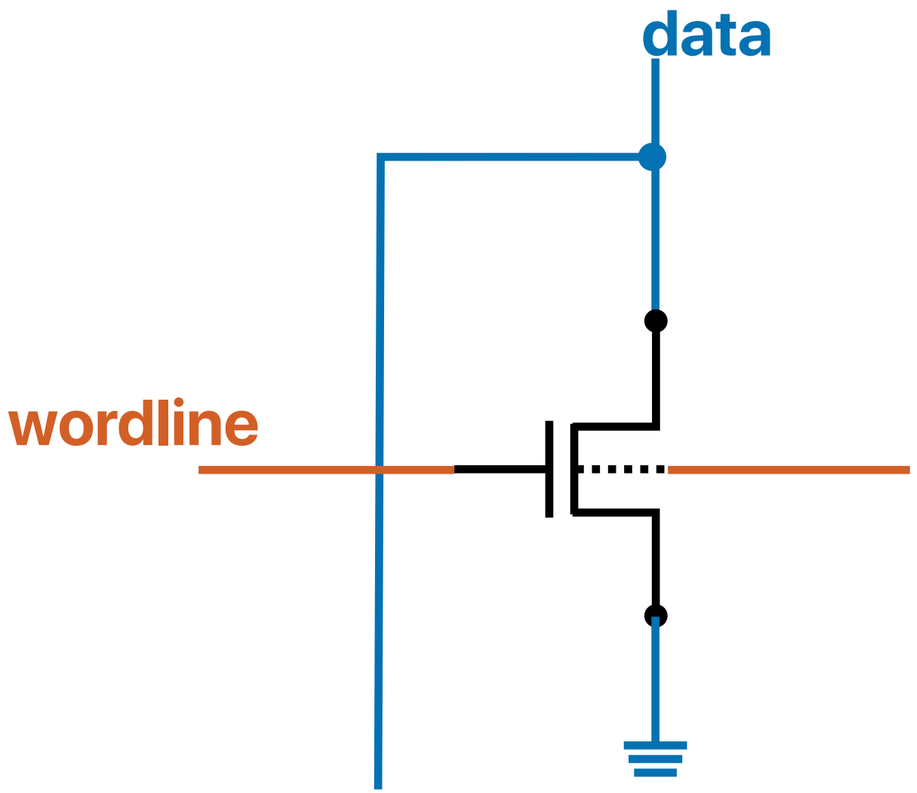

wd0
wd1
wd2
wd(m-1)
0
1
2
n-1
Decoder
upper bits of address
We can only work on cells sharing the same word line simultaneously
Sense Amp
Sense Amp
Sense Amp
Sense Amp
lower bits of address
MUX

3:00

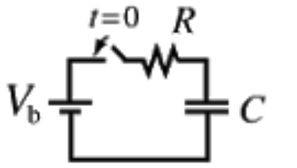
How is DRAM structured?


ODT
To ODT/output drivers
ZQ CAL
To ZQ Control
RESET_n
CKE
CK_t, CK_c
PAR
TEN
CS_n
ACT_n
Control logic
VrefDQ
BC4
OTF
CRC
Parity
2 (A12,A10)
Command decode RAS_n, CAS_n, WE_n
Mode registers
21
Row-address MUX
16
Refresh counter
16
2
2
BG and BA control logic
2
2
A[16:0]
BA[1:0]
BG[1:0]
21
Address register
Bank 0 BG0 Row-address latch and decoder
65,536
Bank 0 Bank Group 0
Memory array (65,536 x 128 x 64)
Sense amplifiers
8192
I/O gating DM mask logic
Global I/O gating
16
128 x64
Column decoder
Column address counter/ latch
7
3
Columns 0, 1 and 2
Bank 0 BG1
Bank 0 Bank Group 1
Bank 0 BG2
Bank 0 Bank Group 2
Bank 0 BG3
Bank 0 Bank Group 3
Bank 3
Bank 2
Bank 1
Bank 0
64
BC4
OTF
Columns 0, 1, and 2
READ FIFO and data MUX
64
CK_t,CK_c
64
Data interface
8
CRC and parity control
ZQ control
ODT control
ALERT
ZQ
CK_t,CK_c
DLL
8
Read drivers
BC4
DBI
CRC
Write drivers and input logic
8
Column 2 (BC4 nibble)
VrefDQ
V_{DDQ}
R_{TTp}
R_{TTn}
R_{TTw}
(0 . . . 7)
DQ[7:0]
DQS_t / DQS_c
DQS_t / DQS_c
TDQS_c
DBI_n / DM_n / TDQS_t

An DRAM cell



- 1 transistor (rather than 6)
- Relies on large capacitor to store bit
  - Write: transistor conducts, data voltage level gets stored on top plate of capacitor
  - Read: look at the value of data voltage



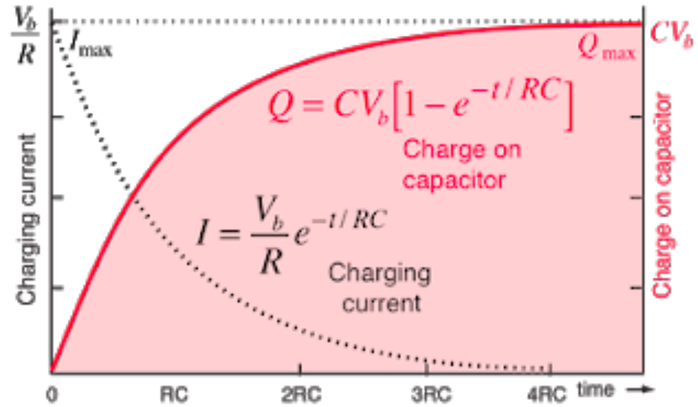
$$V_b = V_R + V_C$$

$$V_b = IR + \frac{Q}{C}$$

As charging progresses,

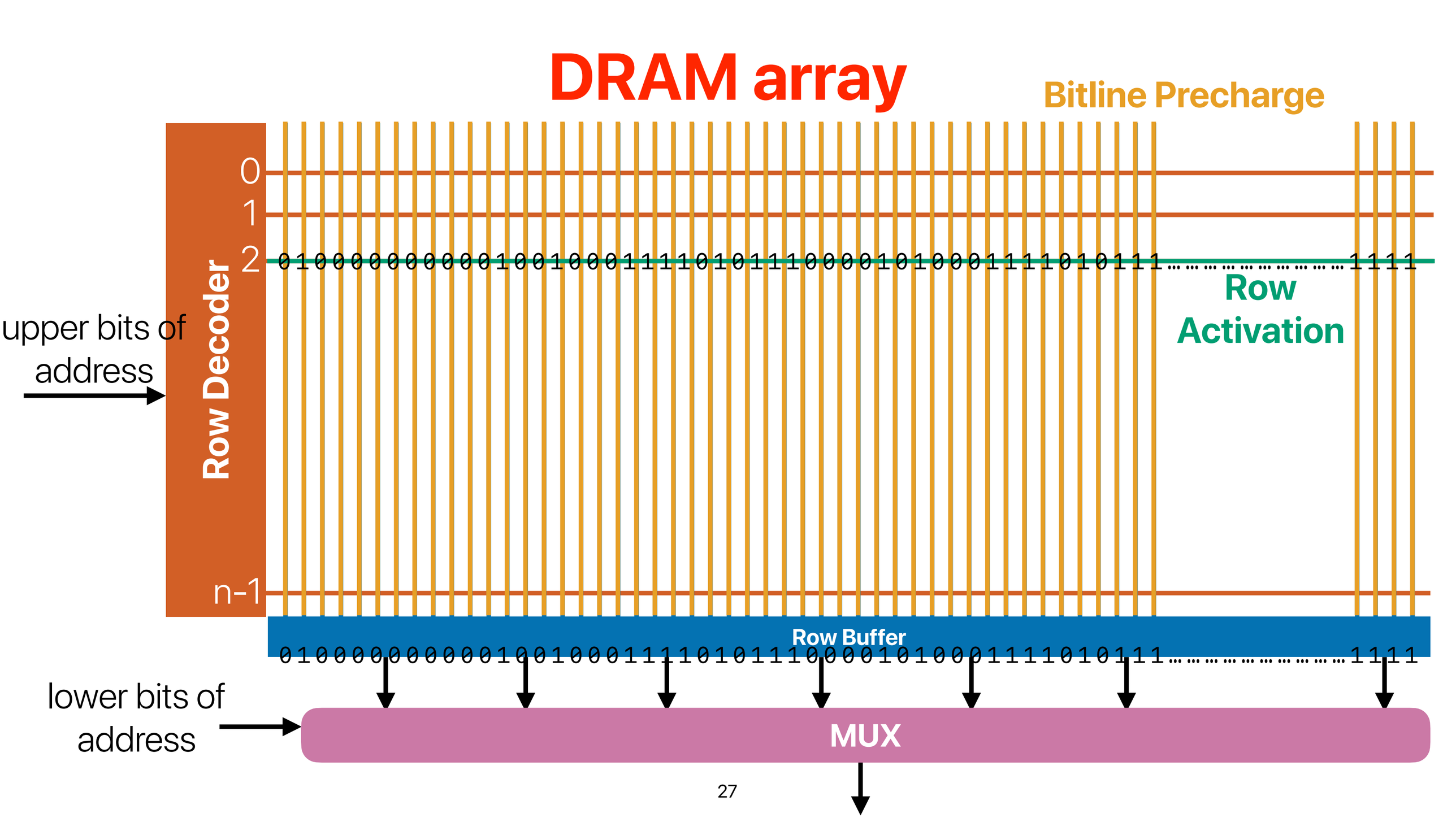
$$V_b = IR + \frac{Q}{C}$$

current decreases and charge increases.

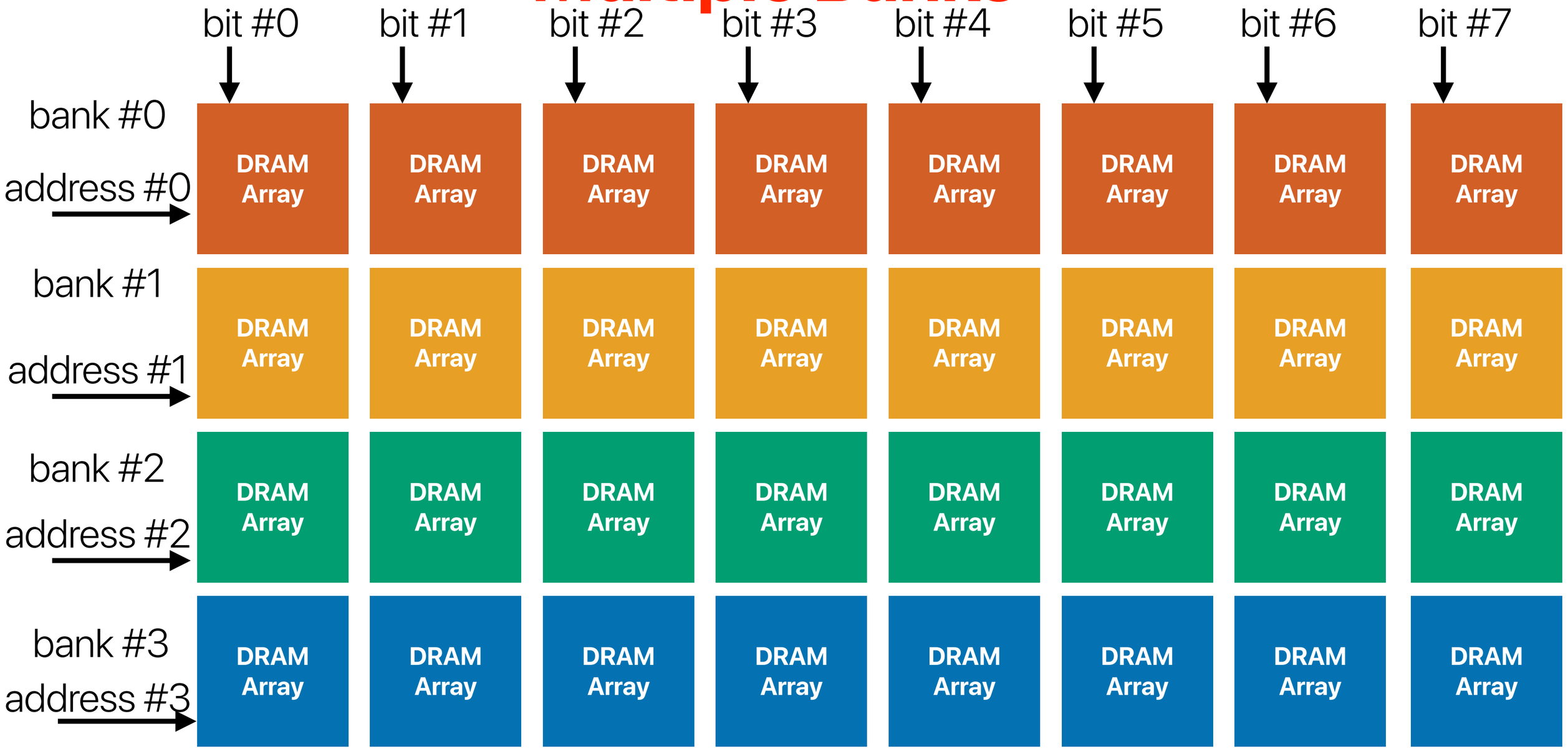


At $t = 0$
$Q = 0$
$V_C = 0$
$I = \frac{V_b}{R}$

As $t \to \infty$
$Q \to CV_b$
$V_C \to V_b$
$I \to 0$

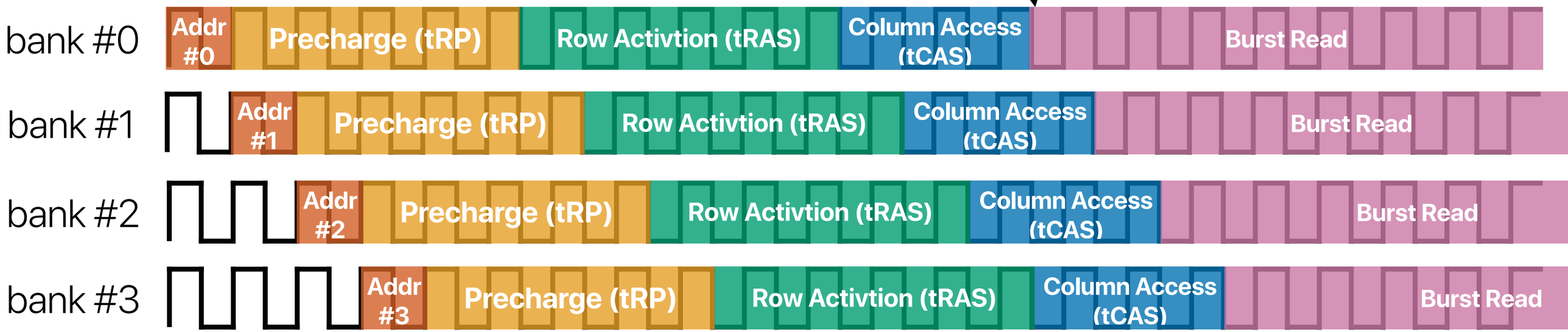

DRAM array
Bitline Precharge
Row Decoder
0
1
2
n-1
Row
Activation
upper bits of
address
Row Buffer
lower bits of
address
MUX

Multiple Banks



Multi-bank access

we can start output a “byte” from every 8 chips each cycle after this

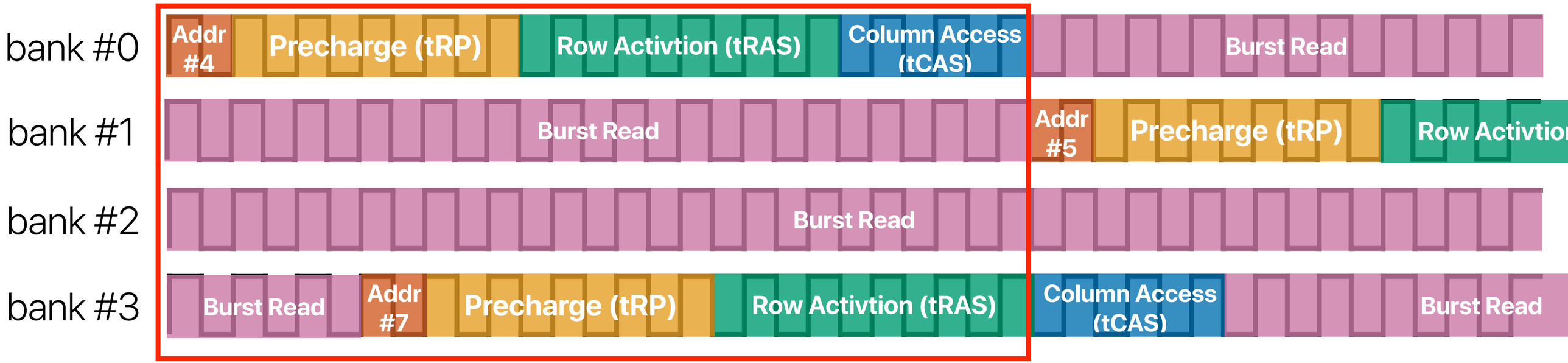


only one bank can accept request each cycle

the memory bandwidth can be fully utilized after this

Multi-bank access

The latency of pre-charge, row/column accesses is fully covered!



DRAM Performance

- Latency per "8-bit" — 0.75 ns (if it's row-buffered)

- Bandwidth per die = $\frac{1}{0.75ns} = 1.33GB/sec$

- 16 chips = $16 \times \frac{1}{0.75ns} = 21.33GB/sec$

2. Key Features

[Table 2] 8Gb DDR4 C-die Speed bins

Speed	DDR4-1600	DDR4-1866	DDR4-2133	DDR4-2400	DDR4-2666	Unit
	11-11-11	13-13-13	15-15-15	17-17-17	19-19-19	
tCK(min)	1.25	1.071	0.937	0.833	0.75	ns
CAS Latency	11	13	15	17	19	nCK
tRCD(min)	13.75	13.92	14.06	14.16	14.25	ns
tRP(min)	13.75	13.92	14.06	14.16	14.25	ns
tRAS(min)	35	34	33	32	32	ns
tRC(min)	48.75	47.92	47.06	46.16	46.25	ns

- JEDEC standard 1.2V (1.14V~1.26V)
- V_{DDQ} = 1.2V (1.14V~1.26V)
- V_{PP} = 2.5V (2.375V~2.75V)
- 800 MHz f_{CK} for 1600Mb/sec/pin, 933 MHz f_{CK} for 1866Mb/sec/pin, 1067MHz f_{CK} for 2133Mb/sec/pin, 1200MHz f_{CK} for 2400Mb/sec/pin, 1333MHz f_{CK} for 2666Mb/sec/pin
- 8 Banks (2 Bank Groups)
- Programmable CAS Latency (posted CAS): 10,11,12,13,14,15,16,17,18,19,20
- Programmable CAS Write Latency (CWL) = 9,11 (DDR4-1600), 10,12 (DDR4-1866),11,14 (DDR4-2133),12,16 (DDR4-2400) and 14,18 (DDR4-2666)
- 8-bit pre-fetch
- Burst Length: 8, 4 with tCCD = 4 which does not allow seamless read or write [either On the fly using A12 or MRS]
- Bi-directional Differential Data-Strobe
- Internal (self) calibration: Internal self calibration through ZQ pin (RZQ: 240 ohm ± 1%)
- On Die Termination using ODT pin
- Average Refresh Period 7.8us at lower than T_{CASE} 85°C, 3.9us at 85°C < $T_{CASE} \leq$ 95 °C
- Connectivity Test Mode (TEN) is Supported

The 8Gb DDR4 SDRAM C-die is organized as a 64Mbit x 16 I/Os x 8banks device. This synchronous device achieves high speed double-data-rate transfer rates of up to 2666Mb/sec/pin (DDR4-2666) for general applications.

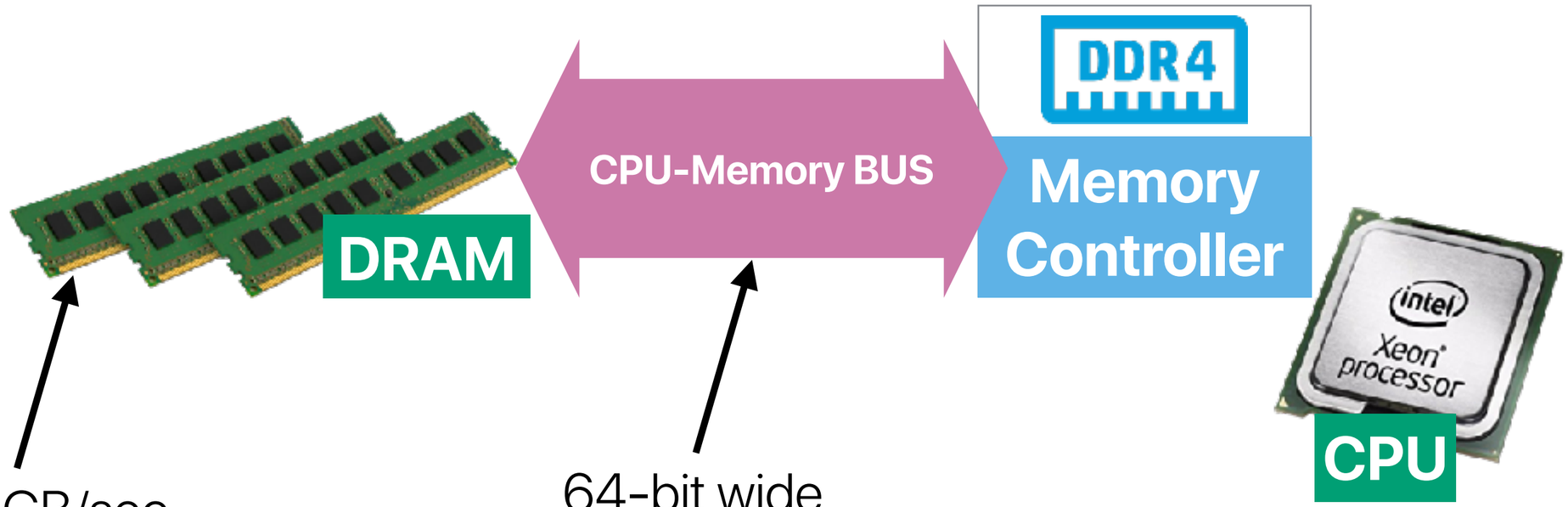
The chip is designed to comply with the following key DDR4 SDRAM features such as posted CAS, Programmable CWL, Internal (Self) Calibration, On Die Termination using ODT pin and Asynchronous Reset.

All of the control and address inputs are synchronized with a pair of externally supplied differential clocks. Inputs are latched at the crosspoint of differential clocks (CK rising and $\overline{CK}$ falling). All I/Os are synchronized with a pair of bidirectional strobes (DQS and $\overline{DQS}$) in a source synchronous fashion. The address bus is used to convey row, column, and bank address information in a $\overline{RAS}/\overline{CAS}$ multiplexing style. The DDR4 device operates with a single 1.2V (1.14V~1.26V) power supply, 1.2V(1.14V~1.26V) V_{DDQ} and 2.5V (2.375V~2.75V) V_{PP}.

The 8Gb DDR4 C-die device is available in 96ball FBGAs(x16).

https://semiconductor.samsung.com/resources/user-manual/x16%20only_8G_C_DDR4_Samsung_Spec_Rev1.5_Apr.17.pdf

How fast is the memory controller frequency?



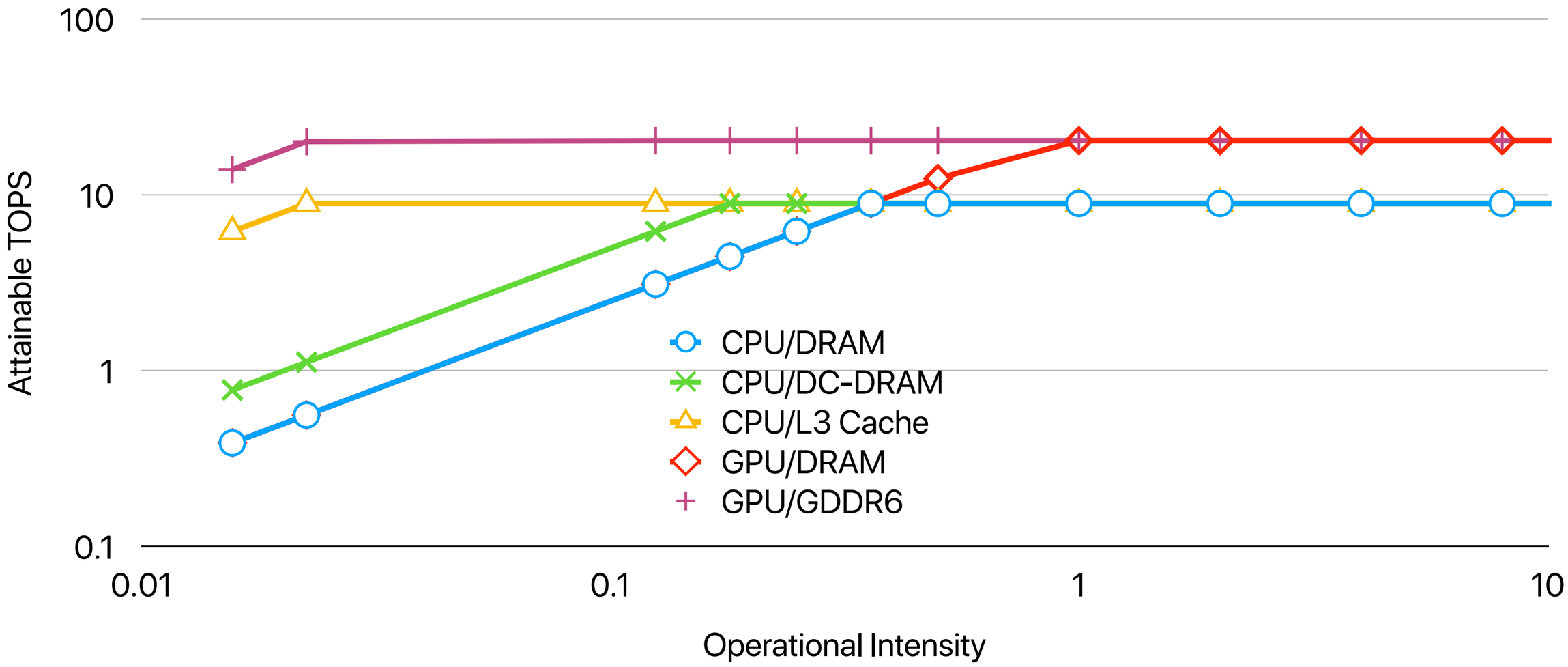
21.33 GB/sec each module

64-bit wide

To support such a data rate, the controller needs to operates at

$$\frac{21.33GB/sec}{64-bit} = \frac{21.33GB}{8B} = 2.667GHz$$

Recap: the roofline after using hardware accelerators



3:00

How can you increase the DRAM bandwidth?

Ideas of increasing bandwidth

- More parallel bits
- More banks
- More channels
- Widen the memory-processor bus

Fill the project/paper presentation proposal form

- Any topic related to the class is welcome
  - Any software technique/design related to the use of modern accelerators or memory architecture to accelerate applications processing data (I think that means every application now)
  - Any innovation in hardware architecture
- You can group in 3 for project
- You can group in 2 for paper presentations
  - But your group has to present a “topic” with a few papers and potentially compare them
- You may use your existing research project if it’s related to us

Electrical
Computer Science
Engineering

277

つづく