EE-429
Fundamentals of VLSI Design

DRAM

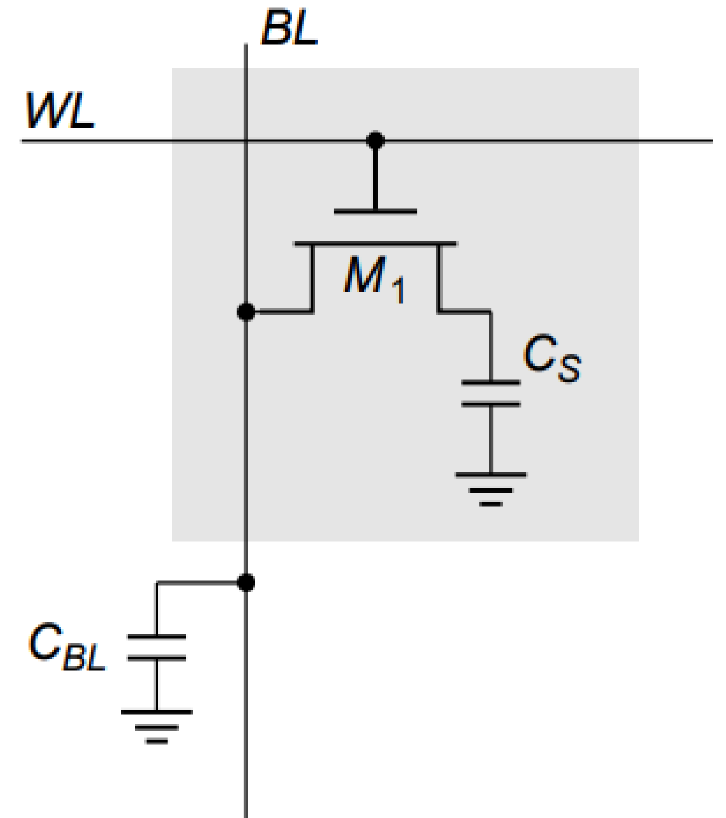
Andreas Burg

Dynamic Random Access Memory (DRAM)

- **SRAM typically provides not enough density due to the large bit-cell**
  - Requires 6 transistors for each bit to implement a bi-stable storage and 2 access transistors

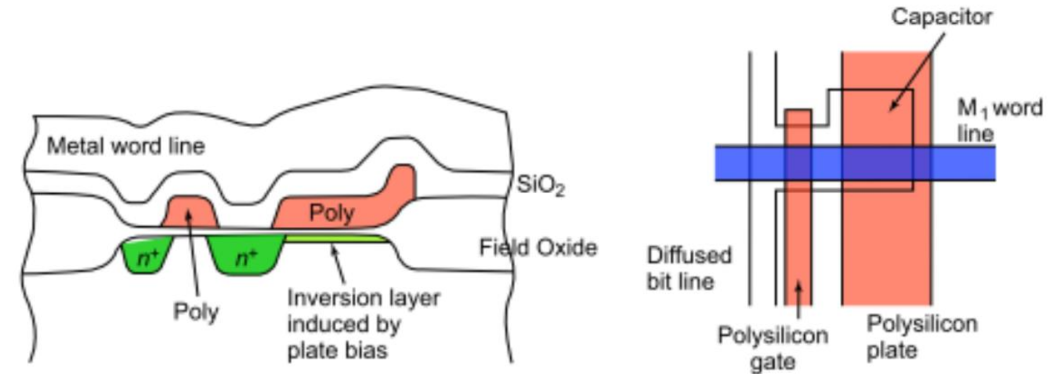
- **Dynamic storage: significantly more compact bit-cell**
  - Data is stored as charge on some form of capacitor C_S
  - A single transistor (M1), controlled by the word-line (WL) is used to access the storage capacitor from the BL

- **Typical DRAM cell size in a special DRAM process is around 6-8 F^2** (F: feature size of the process)
  - For comparison: SRAM cell size in standard CMOS is around 120-150 F^2 (>32nm) and >200 below 16nm

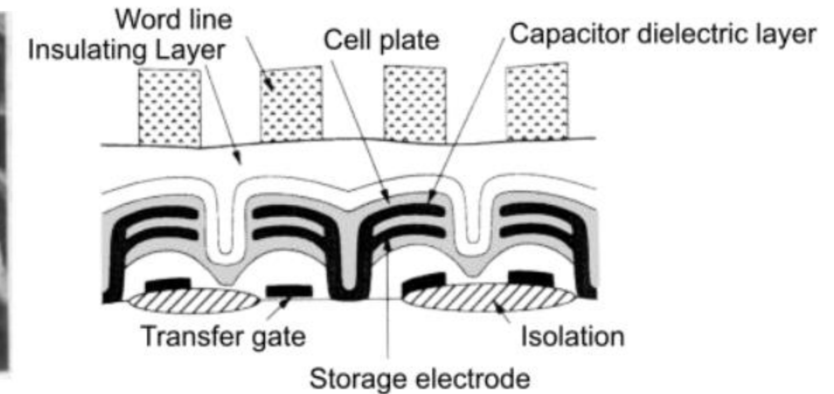


1T-1C DRAM Fabrication

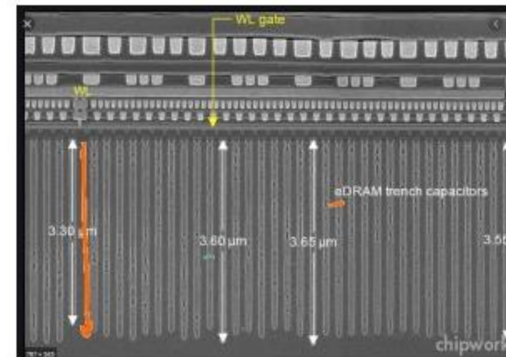
- **Standard process**
  - Capacitor made from poly/diffusion
  - Large area required for the capacitor
  - Used mostly in 1970s and 1980s

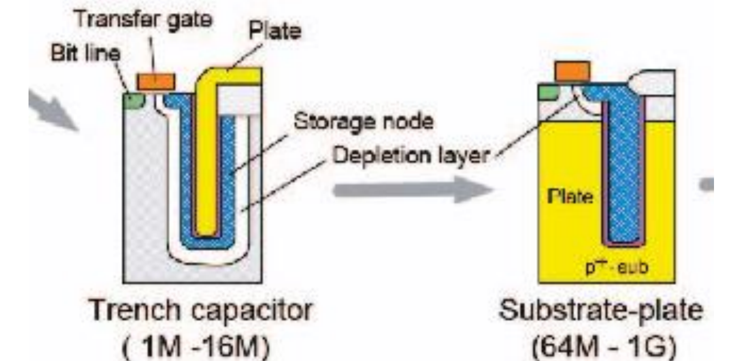


- **Stacked capacitor**
  - Capacitor based on a special plate capacitor on top of the bit cell



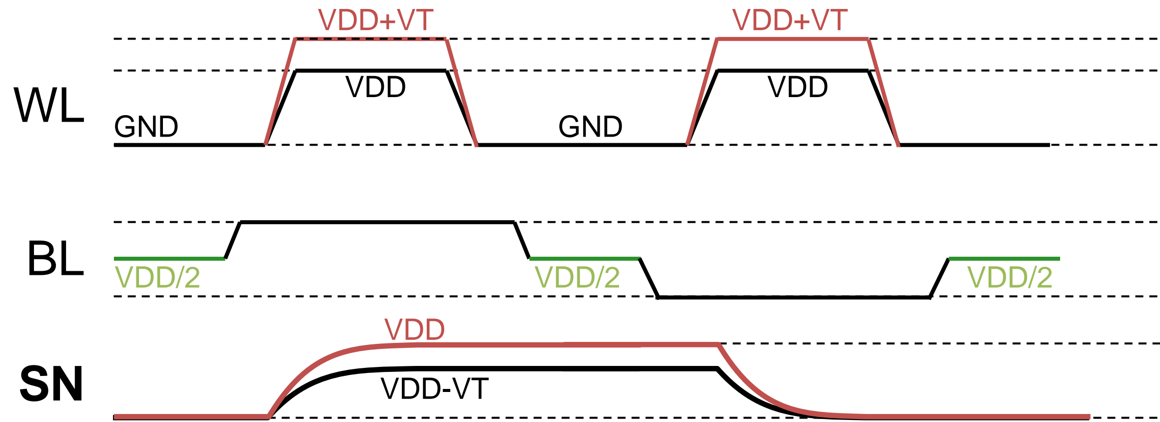
- **Trench capacitor**
  - 3D capacitor located under the bit-cell
  - Requires deep trenches which are tricky to fabricate
  - Most often used today in dense DRAMs





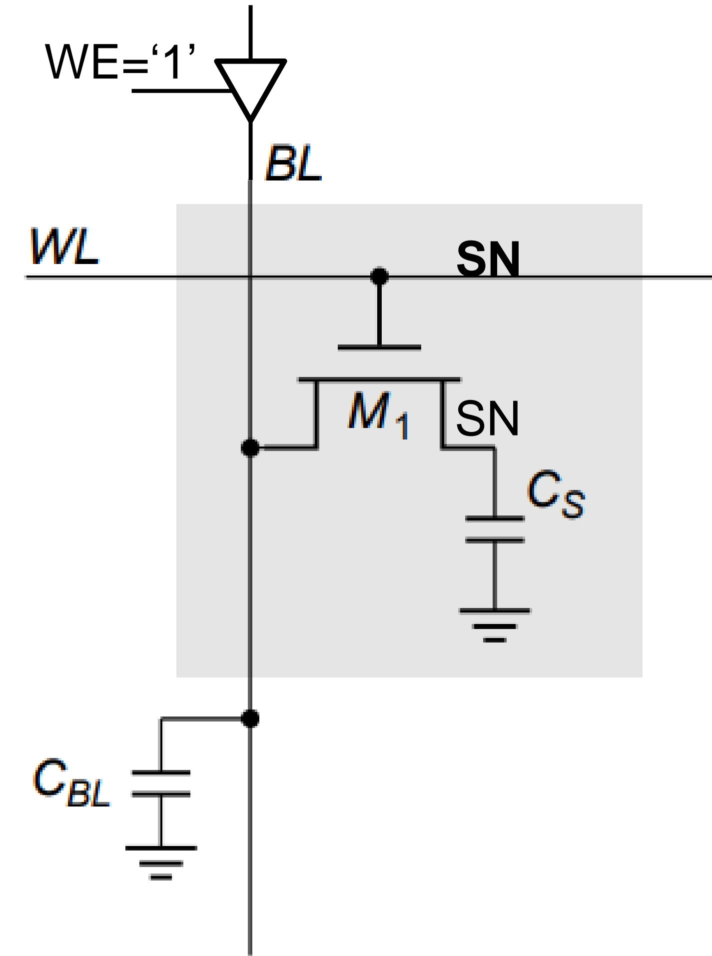
DRAM Read & Write Access

- **Write and read through same access transistor**
- **Write access:**



  - Storage node (SN) capacitor C_S is charged/dis-charged according to BL when WL is asserted
  - Writing a strong '1' can be achieved by over-driving the WL

  - Note: between write access cycles, WL returns to an idle (VDD/2) state



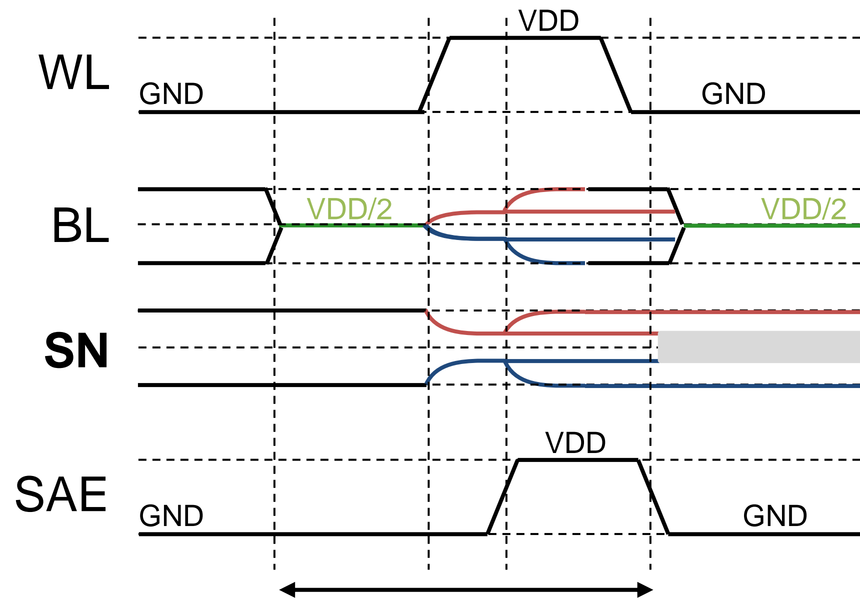
DRAM Read & Write Access

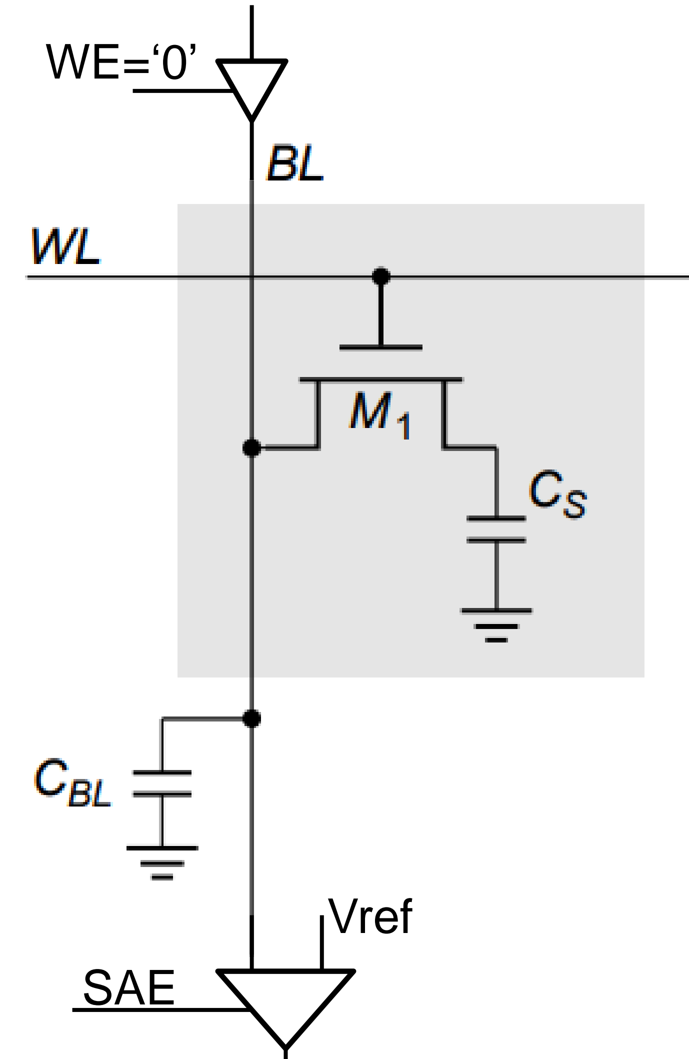
- **Read and write through same access transistor**
- **Read access:**
  - Pre-charge BL to VDD/2
  - WL is activated connecting the SN to the bit line BL
  - Charge-sharing between C_{BL} and C_S

$$\Delta V_{BL} = (V_{SN} - V_{BL})\frac{C_S}{(C_S + C_{BL})}$$

$$|\Delta V_{BL}| < \frac{(VDD/2)\cdot C_S}{(C_S + C_{BL})}$$

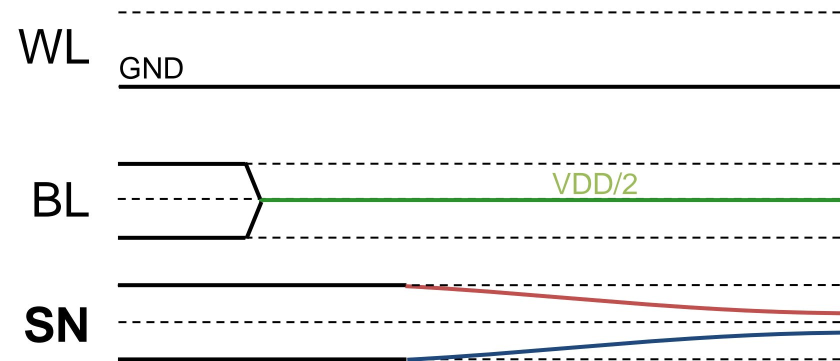
  - Once BL develops a sufficient offset, the sense amplifier (SA) is enabled with SAE
    - Depending on the SA type, the SA feedback may pull the BL and the SN to ‘1’ or ‘0’ (restore the SN voltage)

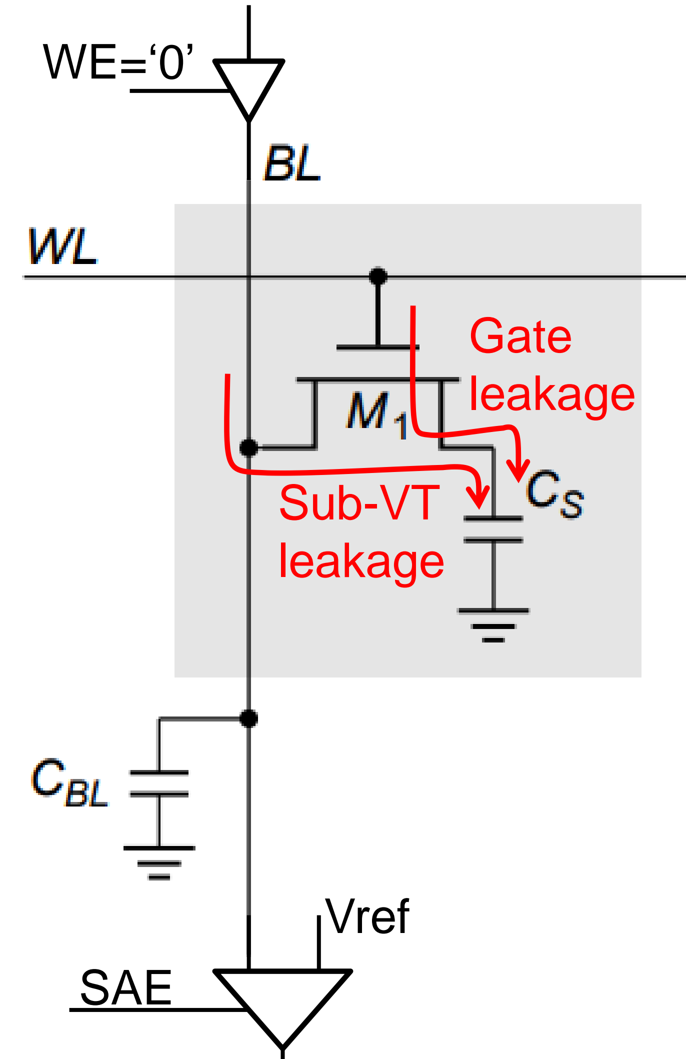




DRAM Retention

- **Between read and write accesses the DRAM is in retention mode**
  - Leakage currents through M1 degrade the SN level over time
- **Ideal retention: no read/no write**
  - BL is biased to balance both 0/1 retention leakage optimally
    - Not necessarily VDD/2, depending on transistor characteristics
- **Non-ideal retention:**
  - BL is biased only between read- and write-access cycles
  - During read and write: BL is temporarily at VDD or GND, leading to an overall slightly increased leakage



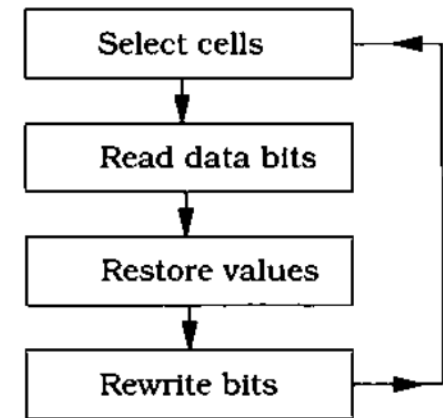


DRAM NEEDS REFRESH!!

DRAM Refresh

- **Data retention time (DRT) is determined by storage node capacitor and leakage currents**
  - Depends on technology and temperature
  - Typical DRT for dedicated DRAM chips are in the order of tens of milliseconds

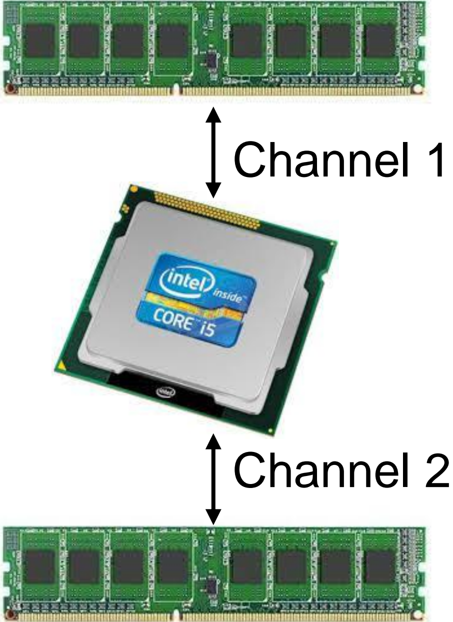
- **Dynamic memories (DRAM) need refresh**
  - Refresh = reading data, restoring logic levels and writing data back to the storage cells
  - Refresh rate depends on the DRT
  - Every refresh cycle refreshes one entire memory row

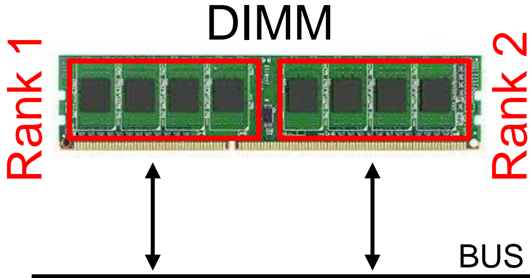


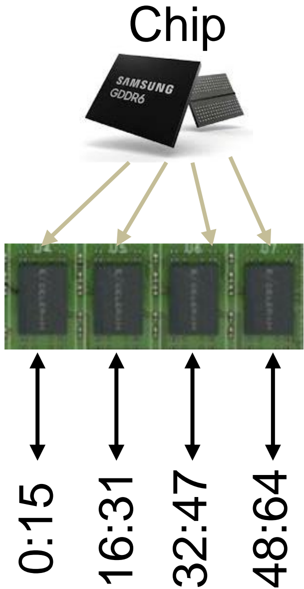
- **Refresh overhead depends on the retention time and the number of rows**

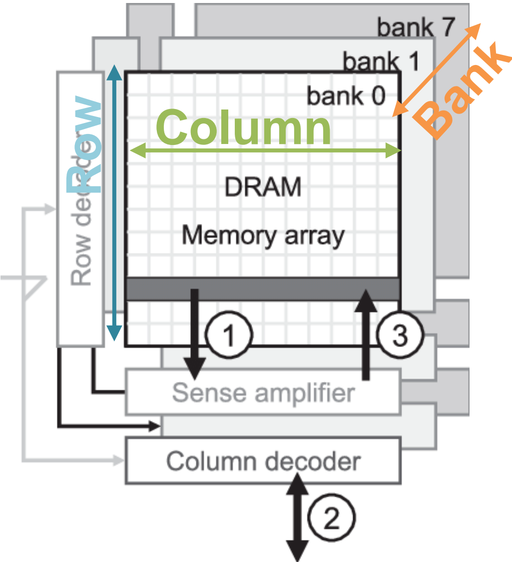
DRAM Organization

- **Large DRAM memories are organized hierarchically**









Channels

DIMMs

Ranks

Chips

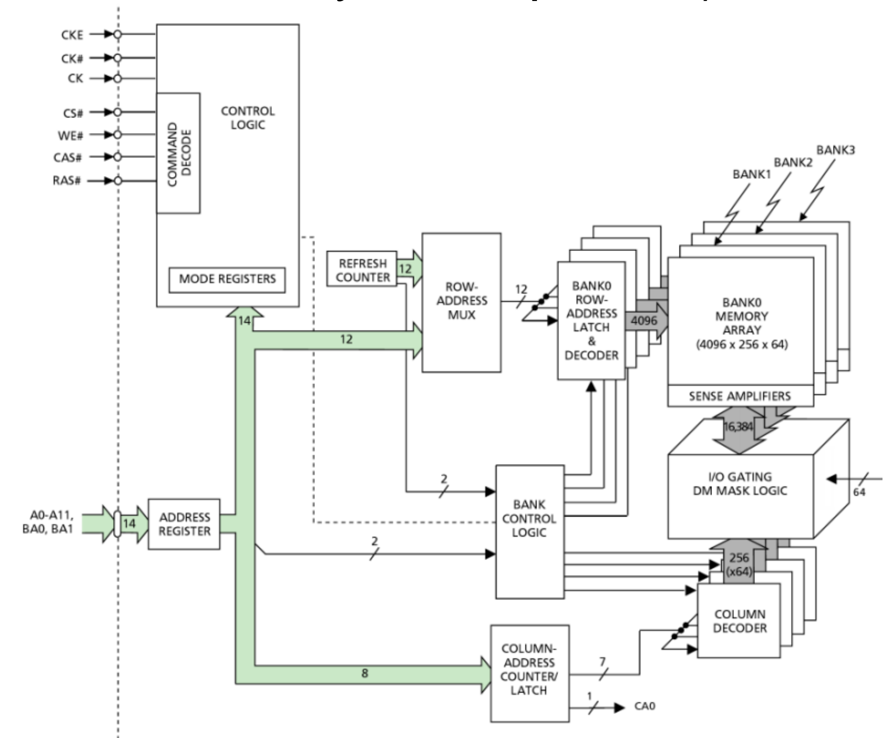
Banks

Rows

Columns

DRAM Access

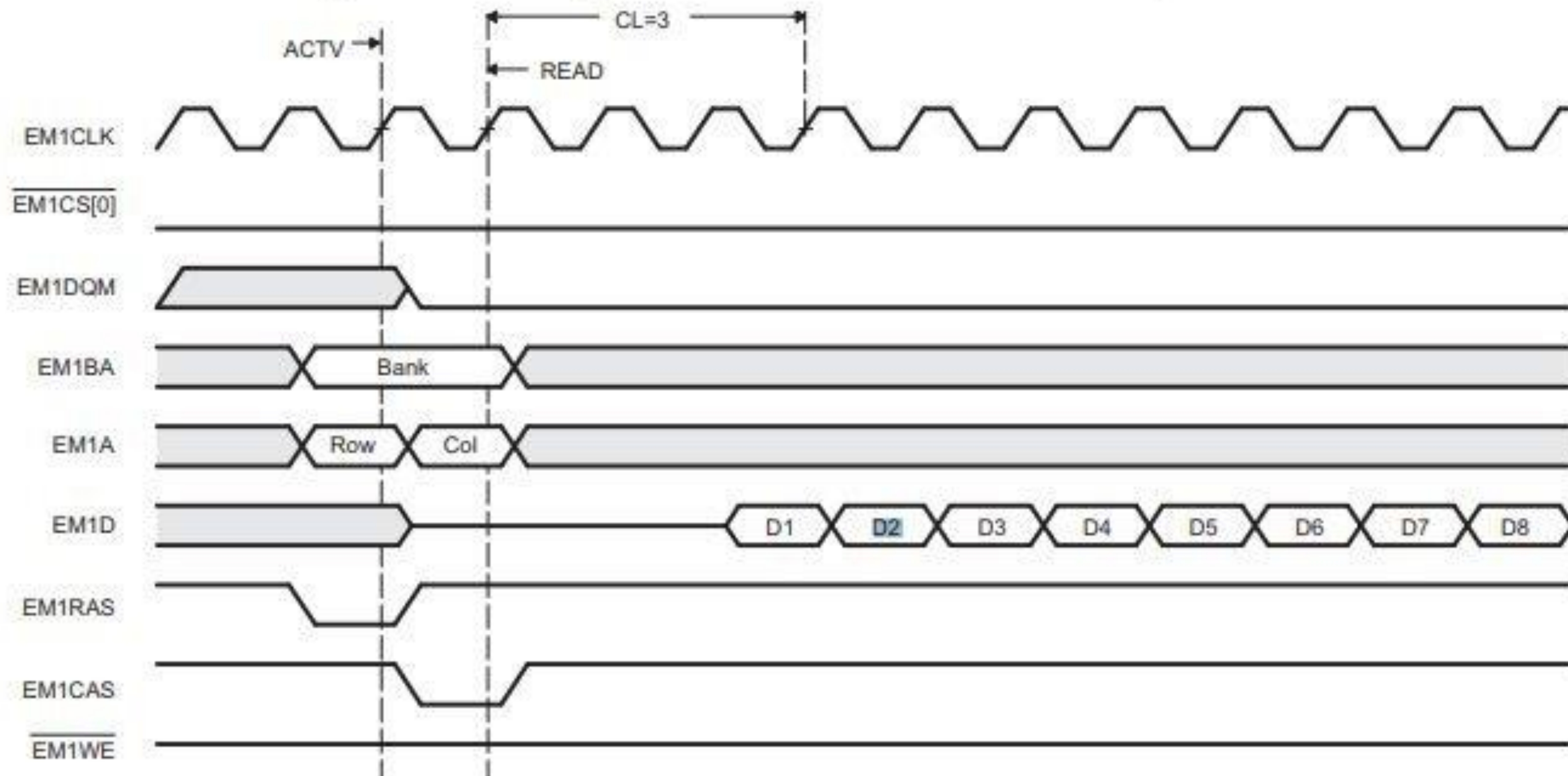
- **DRAM access is broken into several steps, based on the structure**
  - Row access is slow due to the difficult sensing procedure
  - Each row contains a large number of bits (fewer rows allow to refresh many bits in parallel)
  - Destructive read-access: each row access requires restore



- **Access procedure:**
  - Row access: provide row address and read data from storage array into the row buffer (latches)
  - Column access: provide column address and read or write data from/to row buffer
  - Precharge: write row buffer back to storage array and prepare for next read by precharging the BLs

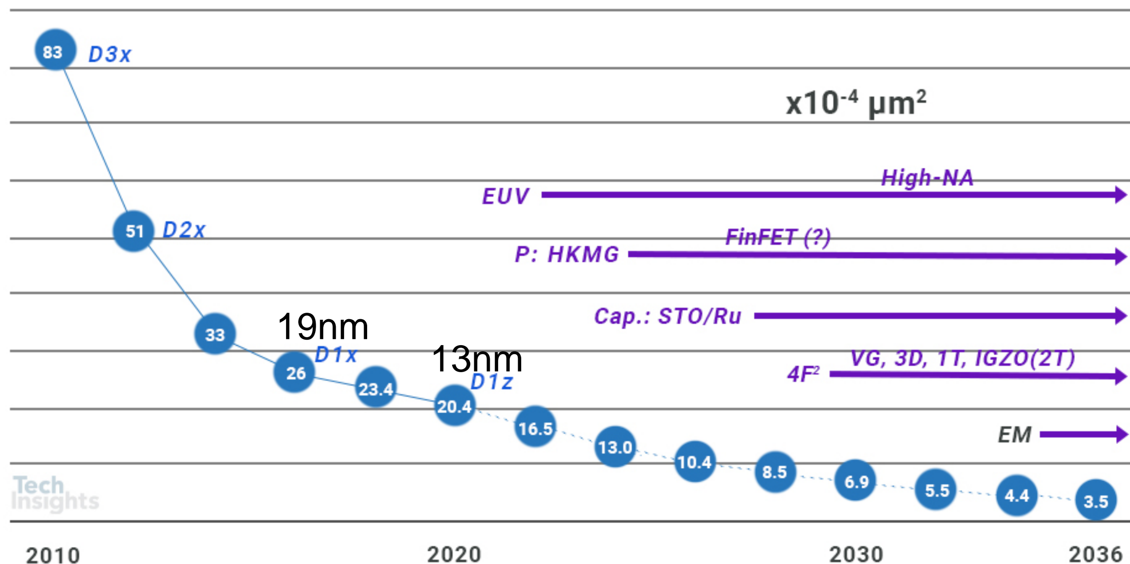
- **Row access is expensive: often followed by bursts from the row buffer**

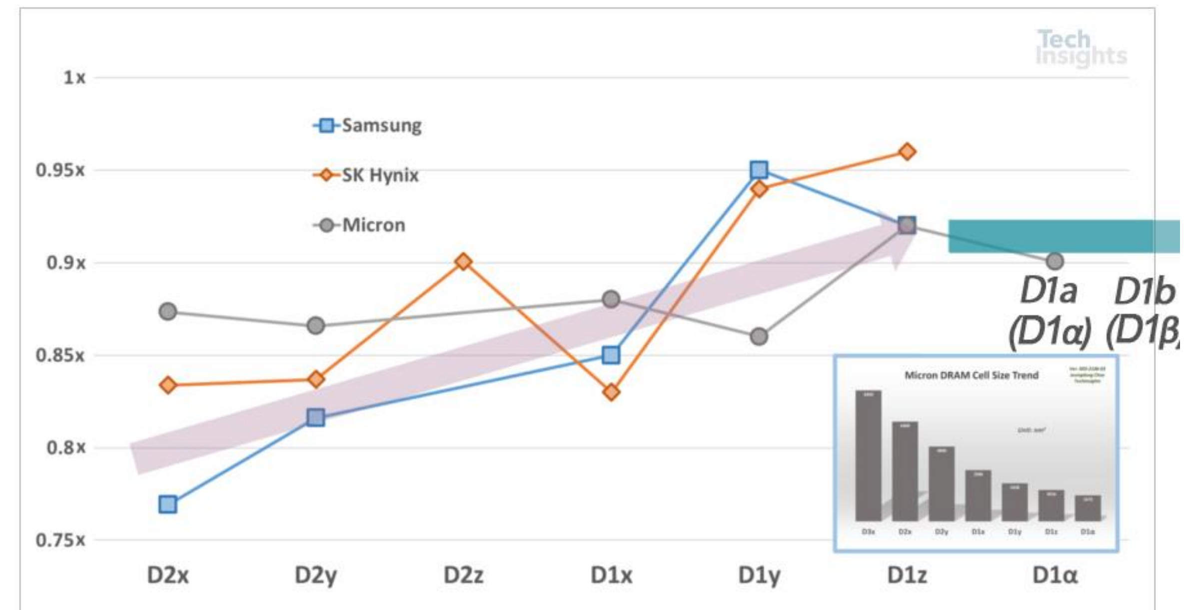
DRAM Access Example (Burst Read)



DRAM Scaling Trends

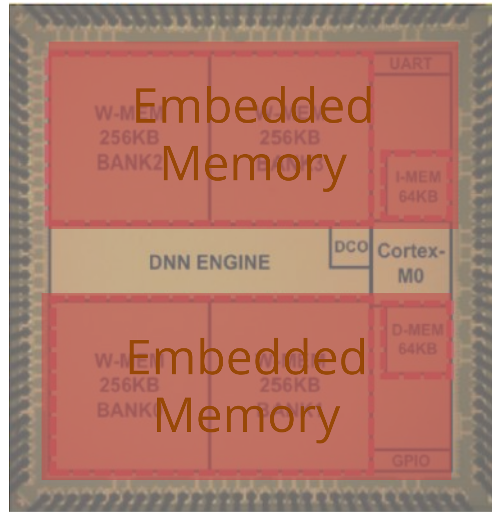
- **DRAM technologies behind CMOS in feature size (today: ~10-12nm, BULK) & scaling is slowing down significantly**
  - Main issue: shrinking cell capacitor area while maintaining its capacitance (~7f F)
    - Use of isolation materials with very high dielectric constants (K>50)
  - Other concern: keeping access transistor leakage low and compatibility with trench capacitors





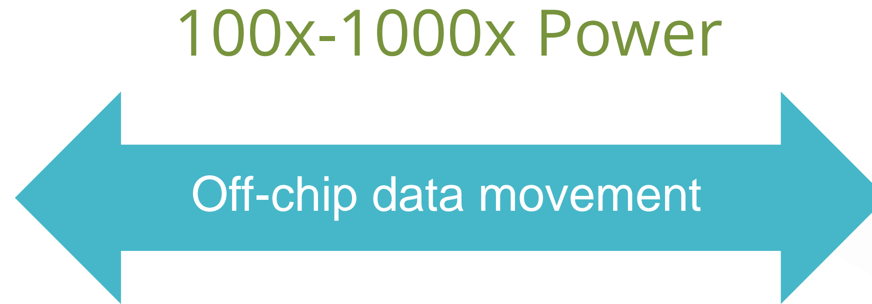
https://www.techinsights.com/blog/dram-scaling-trend-and-beyond

External DRAM is a Bottleneck



SoC

100x-1000x Power



100x-1000x Lower Bandwidth

Significantly higher BOM and
3rd party dependencies

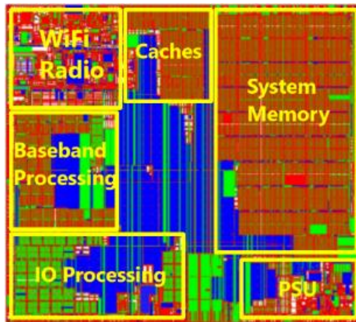
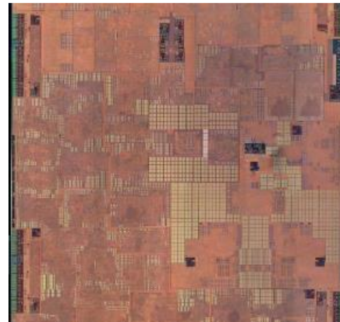
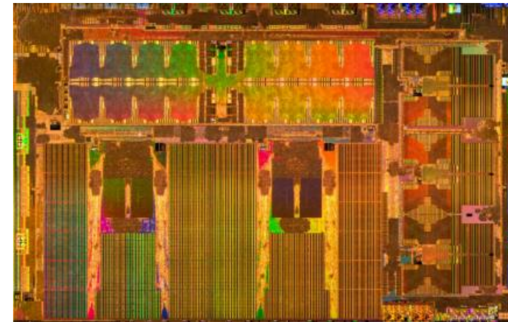
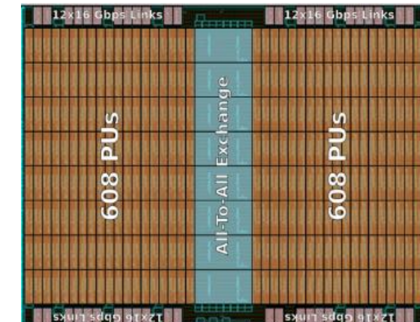
External DRAM

External memory should be avoided at all costs and if needed, access should be minimized

Memory is the Limiting Factor

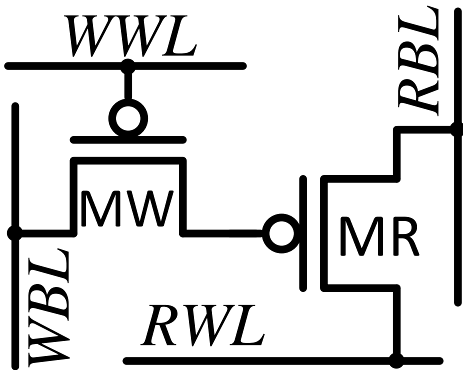
Memories are the limiting factor for cost and energy

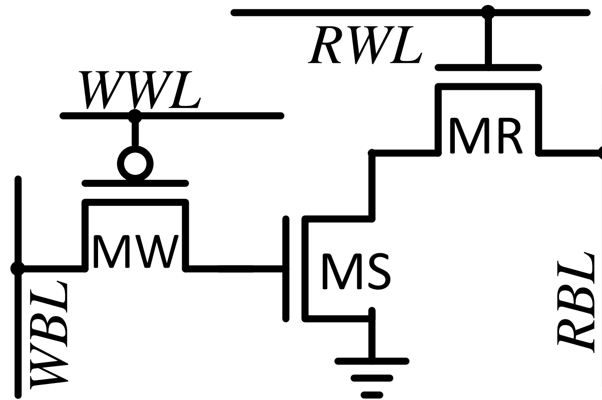
- On-chip memories have a poor area density and often dominate chip area and cost in many computing systems
- Memory often accounts for >50% of the active power and for 100% of the power during sleep/standby periods in low-power systems

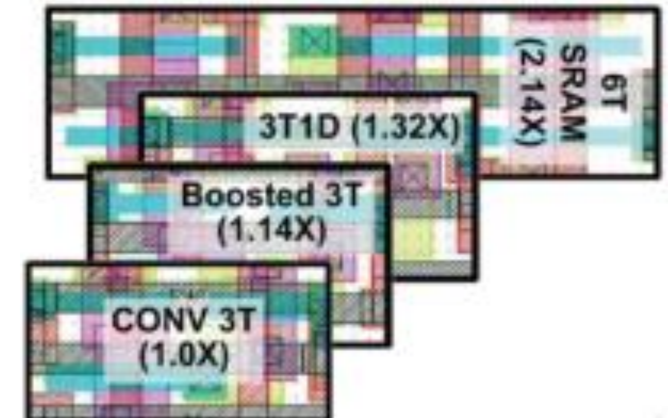
	IoT & MCU	Mobile	Automotive	ML/AI & Server
				
	MediaTek MT3620, 40nm	Apple A11, 10nm	Tesla FSD, 14nm	Graphcore IPU, 16nm
SRAM Area [%]	35%	31%	36%	75%

Gain Cell Embedded DRAM (eDRAM)

- **DRAM on a standard CMOS process**
  - Storage capacitor is a parasitic capacitor (gate capacitance + other parasitics)
  - 1 access transistor for write
  - 1-2 access transistors for read



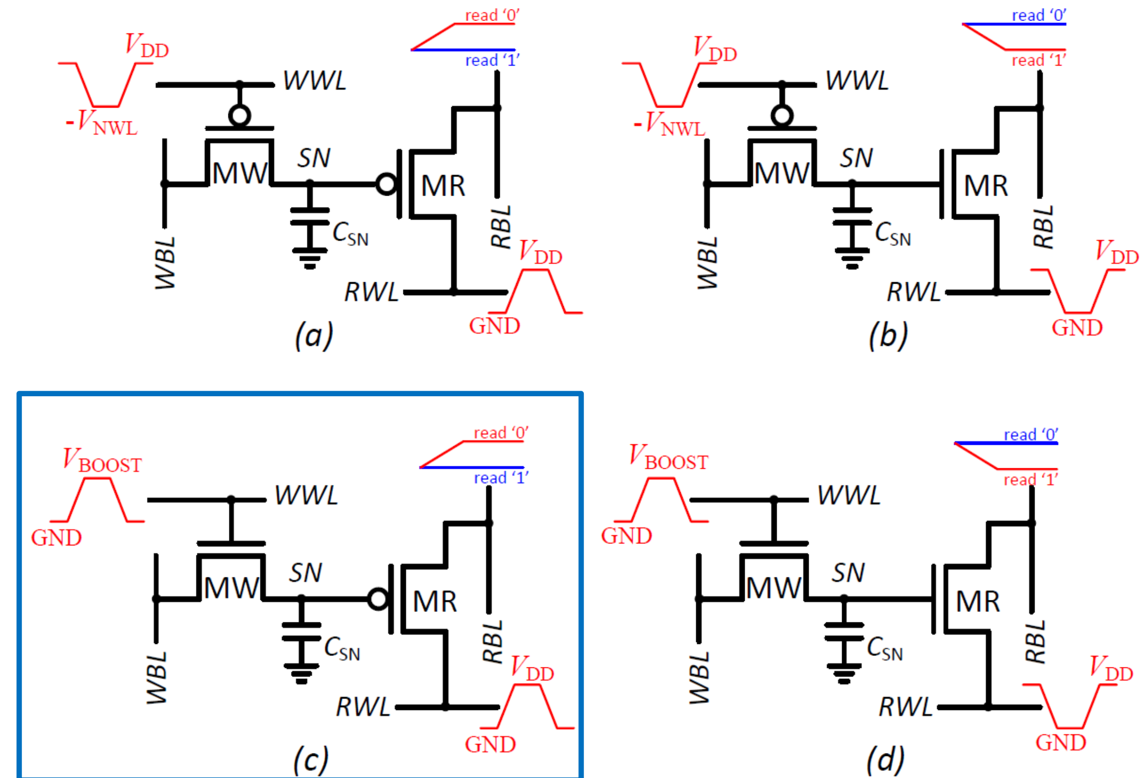




2T and 3T Gain Cell with **~70F² per bit**

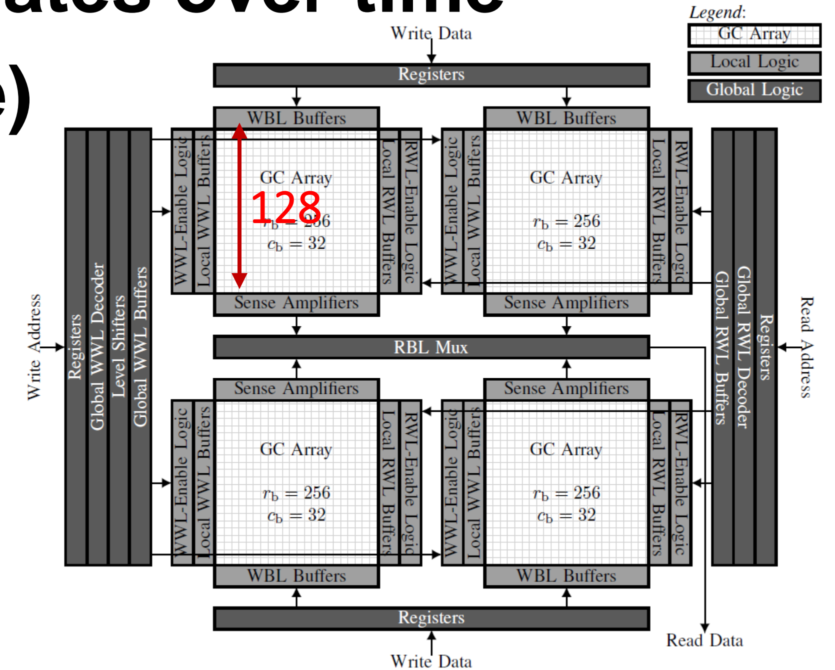
2T Gain Cell eDRAM: Basic Operating Principle

- **Write port (WWL & WBL), storage cap, and read port (RWL & RBL)**
  - Different combinations of PMOS and NMOS transistors
  - Use of different threshold options
- **Write operation:**
  - Boosted WWL, above VDD for NMOS, below V_{SS} for PMOS
- **Read:**
  - PMOS MR: Pre-discharge RBL, raise RWL -> SN=‘0’: RBL rises
  - NMOS MR: Precharge RBL, lower RWL



Gain Cell eDRAM Periodic Refresh

- **Dynamic storage mechanism: data deteriorates over time**
- **Need for periodic refresh cycles (read/write)**
  - Data arranged in sub-arrays
  - Parallel refresh in all sub-arrays



- **Array availability**

$$Availability\,[\%] = 1 - \frac{T_{\mathrm{clk}}}{T_{\mathrm{ret}}} N_{\mathrm{r}}$$

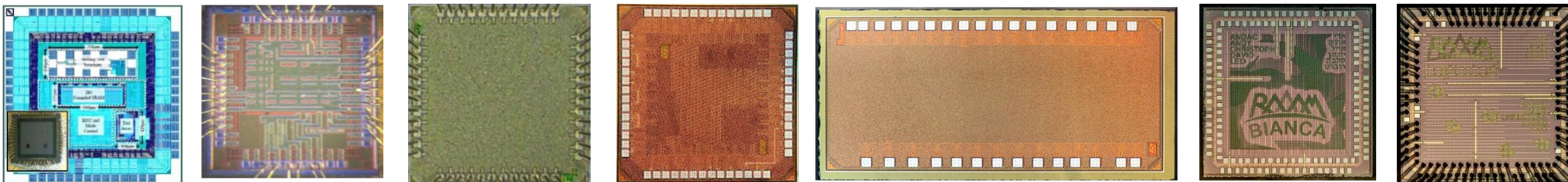
  - Typical retention times: $T_{\mathrm{ret}} =$ 100us – 1ms
  - Typical access/refresh cycle-time: $T_{\mathrm{clk}} =$ 10ns
  - Typical sub-array size N_r =128-256 rows

Typical array availability: ~98%

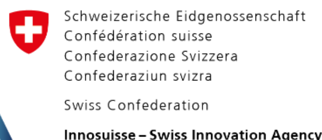
Delivering the Highest Density Volatile Embedded Memories in Standard CMOS

Reduced Cost | Longer Battery-Life | Better Performance

Looking for Engineers and Interns



Microphotographs of RAAAM's GCRAM Technology Implementations in 16nm – 180nm Processes



EPFL INNOGRANTS
Vice-Présidence pour l'innovation et la valorisation

BRIDGE

>>venture>>
Companies for tomorrow