The Semiconductor Memory Saga – 50 years

Betty Prince, PhD.
Memory Strategies International
<www.memorystrategies.com>

bprince@memorystrategies.com

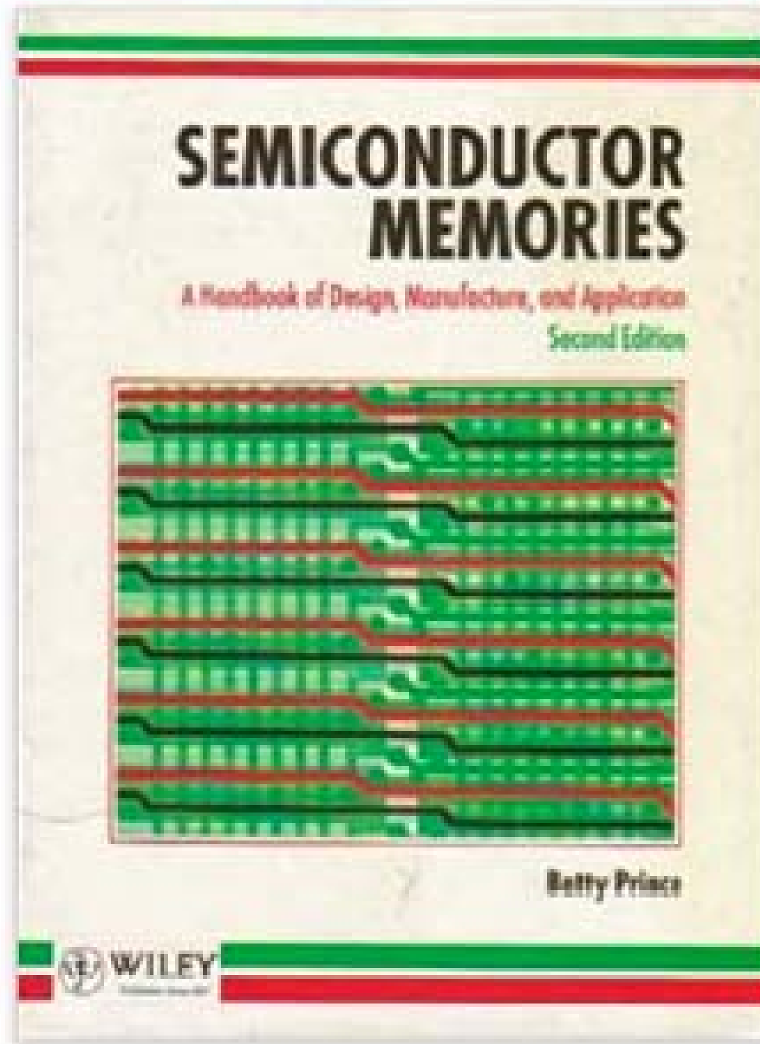
The Saga of Semiconductor Memories – 50 years

Overall - The Evolution of the Cache Hierarchy

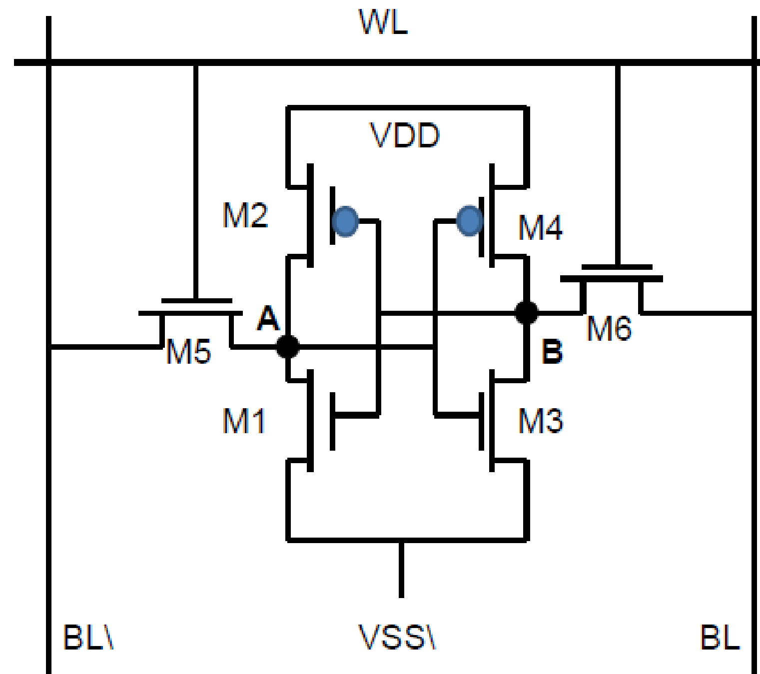
1. Main Memory for Computers – SRAMs and DRAMs
2. The Memory Wall – feeding enough data to the Computer
3. The Advent of Non-Volatile Memory (NOR and NAND)
4. Embedded Memories, Single Chip (MCU + SRAM + Flash)
5. NAND Flash Memory – Flash Passes DRAM in Volume
6. The New Moore's Law* - Vertical NAND, Vertical PCM
7. The End of Moore's Law. - Emerging Memories
8. Neuromorphic Memory, Neural Synapses, and AI

*Moore's Law - The number of transistors on a chip doubles every two years.

1. Main Memory for Computers – SRAMs and DRAMs



6T SRAM – First Main Memory



The original MOS memory was a 6T cell SRAM invented in 1964 at Fairchild as part of their CMOS logic portfolio.
A flipflop with transistors accessing the storage nodes.
It competed initially for Main Memory with a variety of memories: 4T SRAMs, 4T DRAMs and 3T DRAMs.

6T SRAM – Too Big for Main Memory

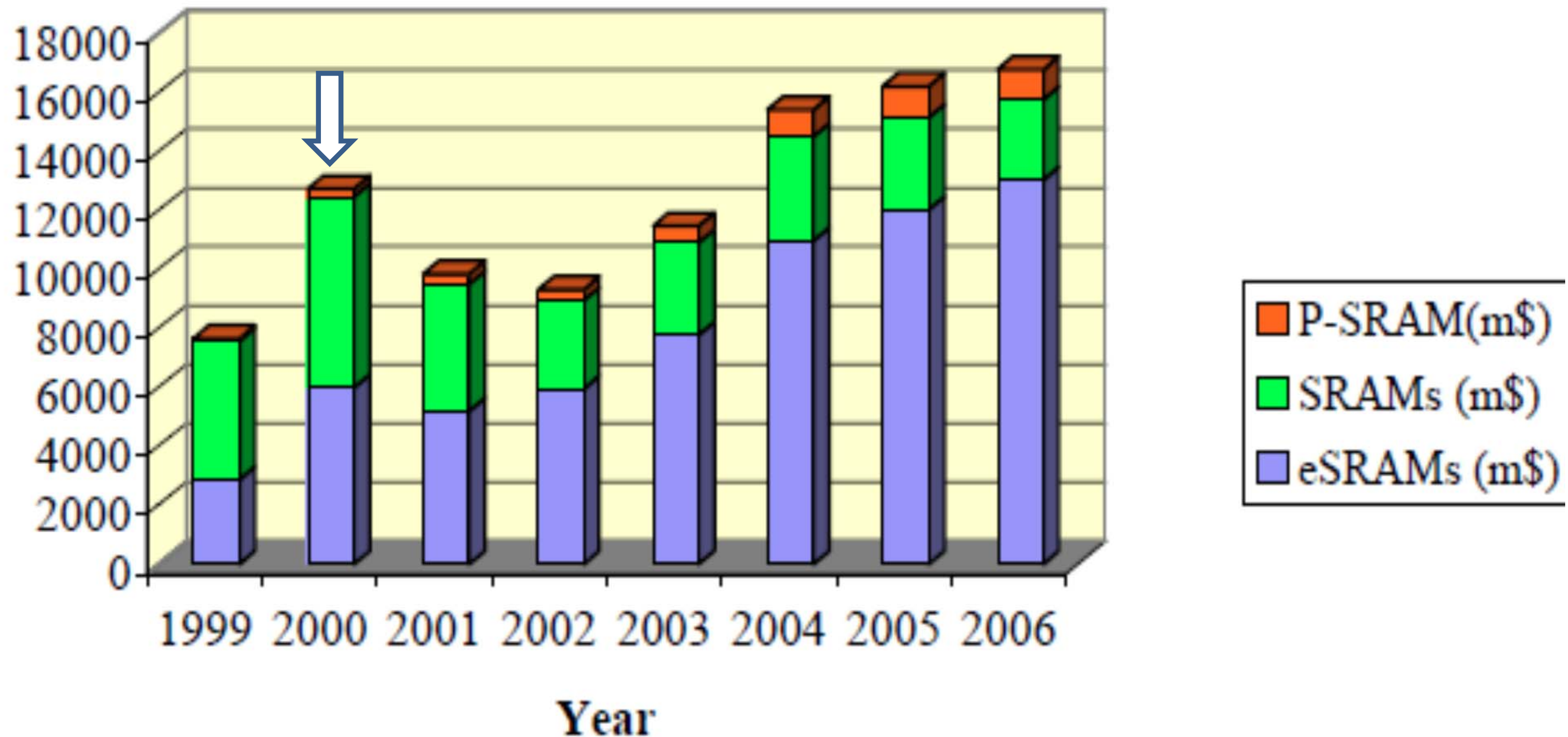
Advantages: Fast Access/Cycle Time
Logic Compatible
Low Standby Power
Random Access
Non-Destructive Read

Disadvantages: Dynamic (lost data when power off)
High active power
In Shrunk technology
- High Subthreshold leakage
- Standby power increased
- Process Variation increased

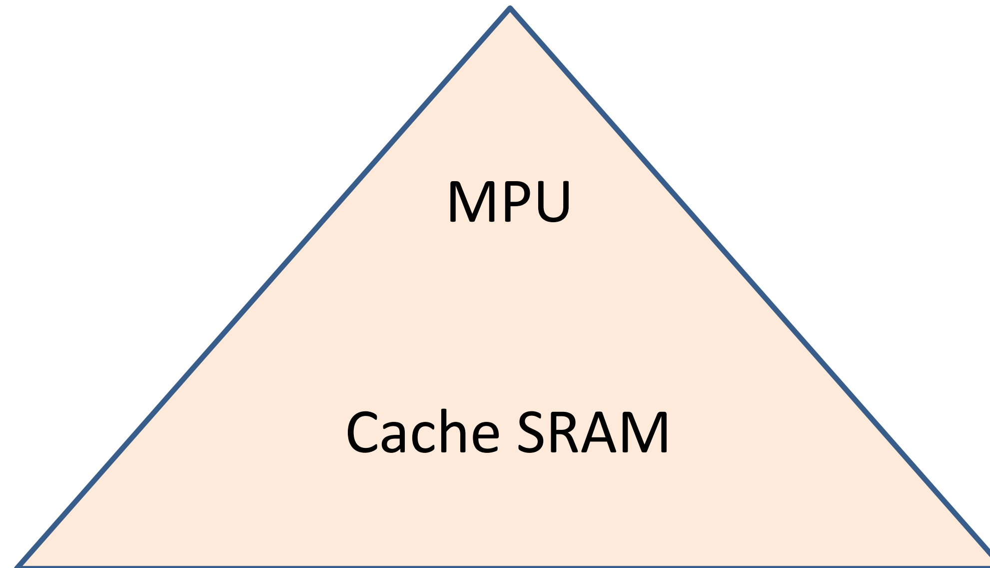
Standalone – large cell size increased cost
Embedded – large cell size (6 T) not as important

SRAM Market Split Standalone and Embedded in 2000.

By 2006, the SRAM was primarily an embedded memory as Cache in Processors and as Data and Code RAM in the Microcontroller.



SRAM and MPU form first cache hierarchy



The SRAM is made of transistors so it is compatible with a logic process. The MPU and SRAM can be on a single chip. This was the first embedded Memory.

Cache Hierarchy

The theory of the cache hierarchy is that the data that will be next requested by the processor can be contained in a small and very fast memory that sits next to the processor.

If this small fast cache can supply the data fast enough that the processor never, or seldom, needs to go to the main memory bank, then the main memory can use slower, cheaper memory and the average speed of the system still remain high.

In 1970 Computers Needed Working Memory

High Density	Megabytes
Low Cost	Small cell size /more chips on a wafer
High Speed	Random Access

Candidates:	Dynamic RAM (DRAM)
	Charge-Coupled Devices (CCD)

The 1T DRAM was patented in 1968 at IBM.
The Charge Coupled device (CCD) was invented at Bell Labs in 1969.

Intel 1103, A 3T Cell DRAM

First Commercial DRAM Chip – Intel 1103.

Presented at the 1970 ISSCC.

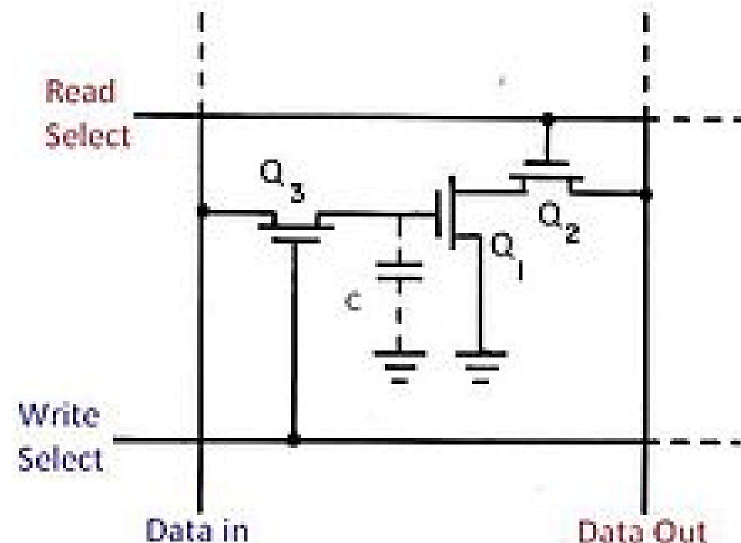
Best Selling Semiconductor Chip by 1972

Density – 1024 bits

Used in the HP 9800 Computer

Organization: 1Kx1

System: 64Kbytes (512 chips)



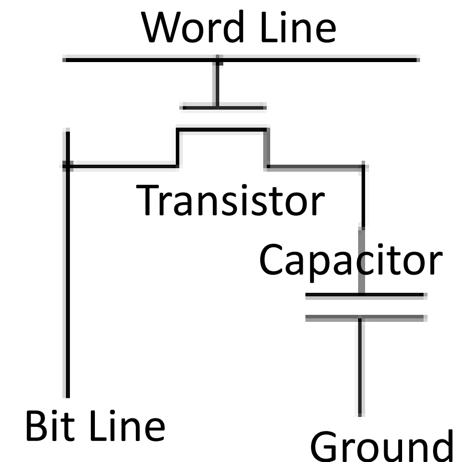
Advantages: **Small Cell Size**
Low Price, Random Access

Disadvantages: Not enough memory capacity, slow.

COMPUTERS NEEDED RAM. IT WAS IN HIGH DEMAND

1T1C DRAM

High memory demand meant volume production from beginning
Americans -> Japanese -> Korean -> (Taiwan, Europe small)
Today: Micron, Samsung, Hynix



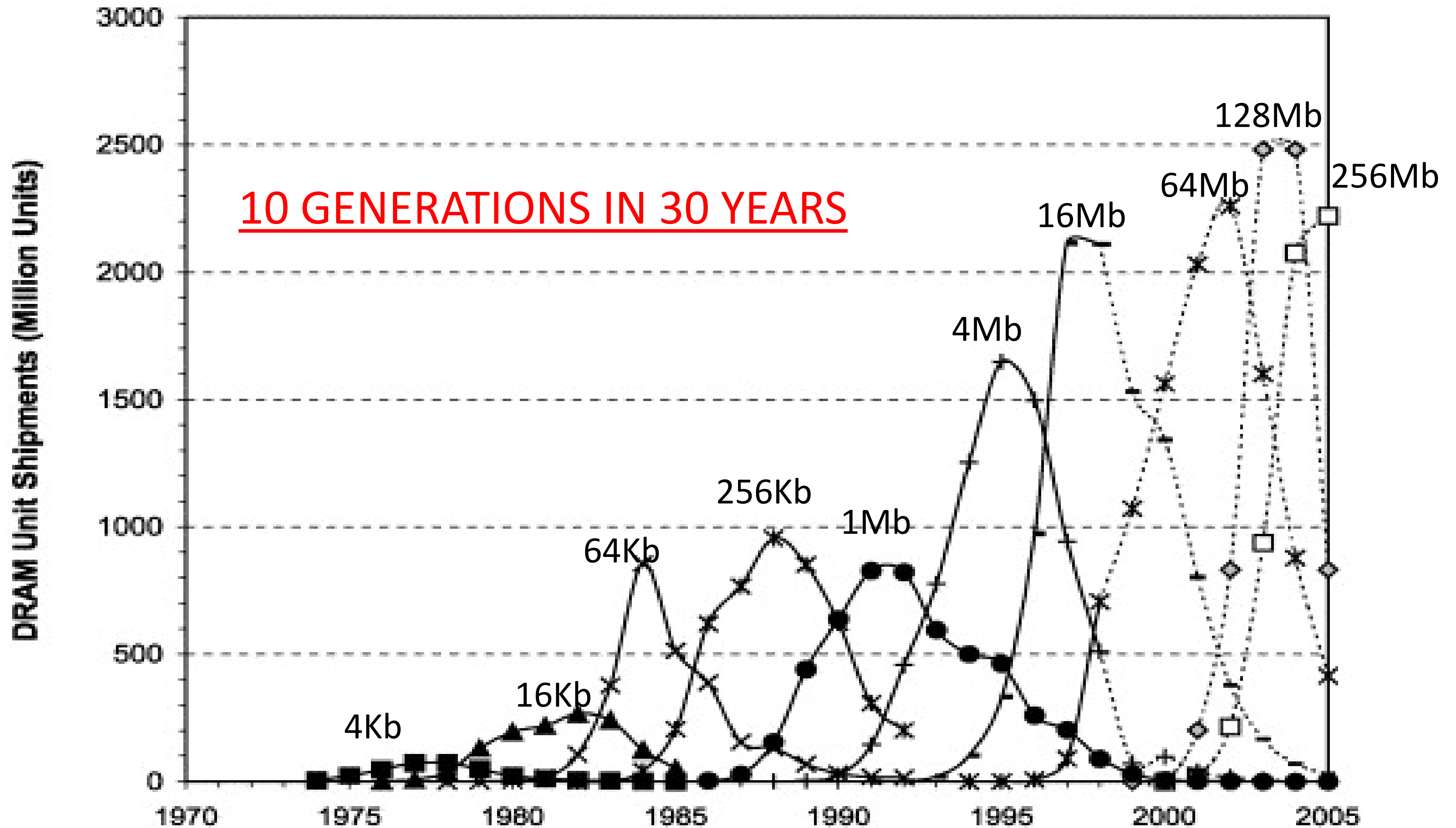
All the problems were dealt with in production:

- 12V, 5V → 5V only → 3.3V
- Soft Errors – Cosmic Rays
- Wide I/O – for smaller systems

JEDEC DRAM standardization driven by large users, set standards and followed them. The users didn't buy non-standard parts.

By 1980 the 1T DRAM was 27% of the total memory market.
By 1990 the DRAM dominated the main memory market.
The CCD Memory had essentially vanished.

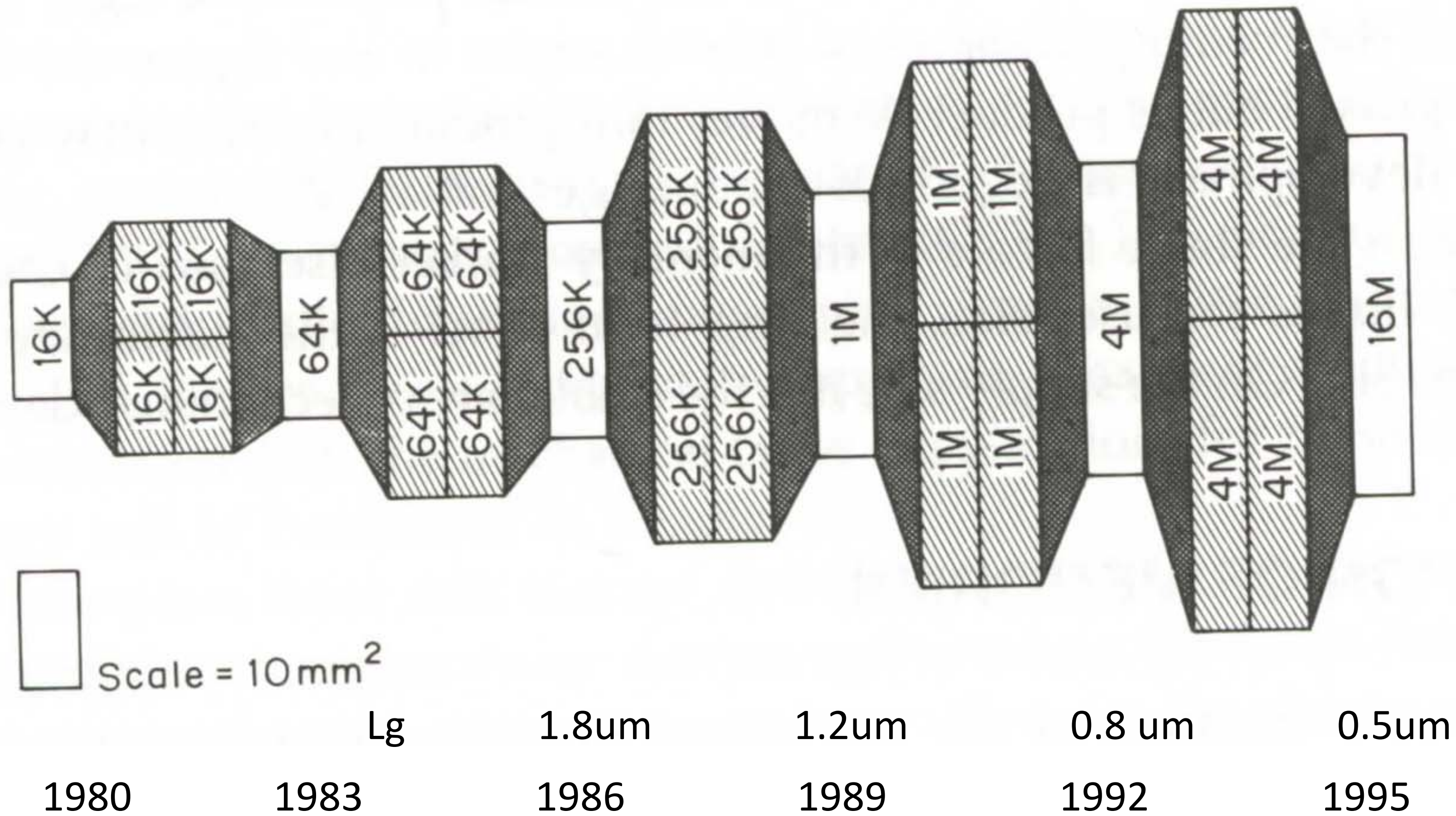
DRAM Shipments 1970-2005



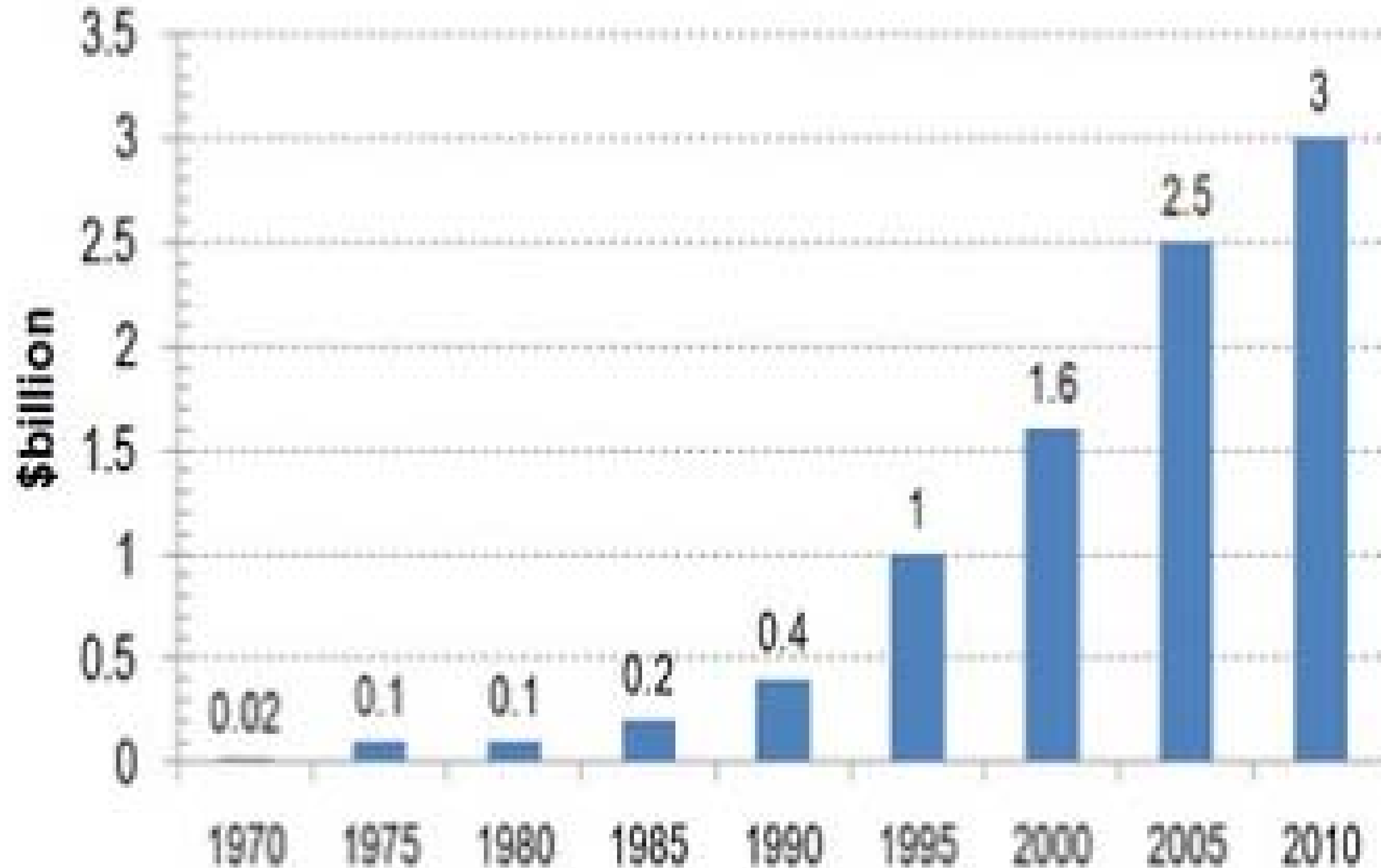
bprince@memorystrategies.com

N.M. Victor, J.H. Ausubel, Science Direct, April 2002

Size Reduction of successive generations of DRAM chips



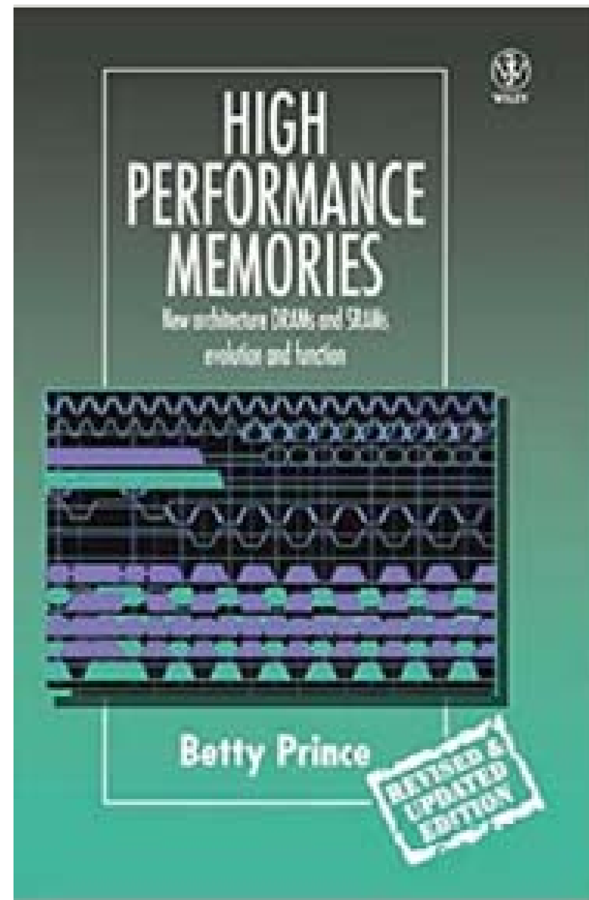
Cost of a DRAM Fab



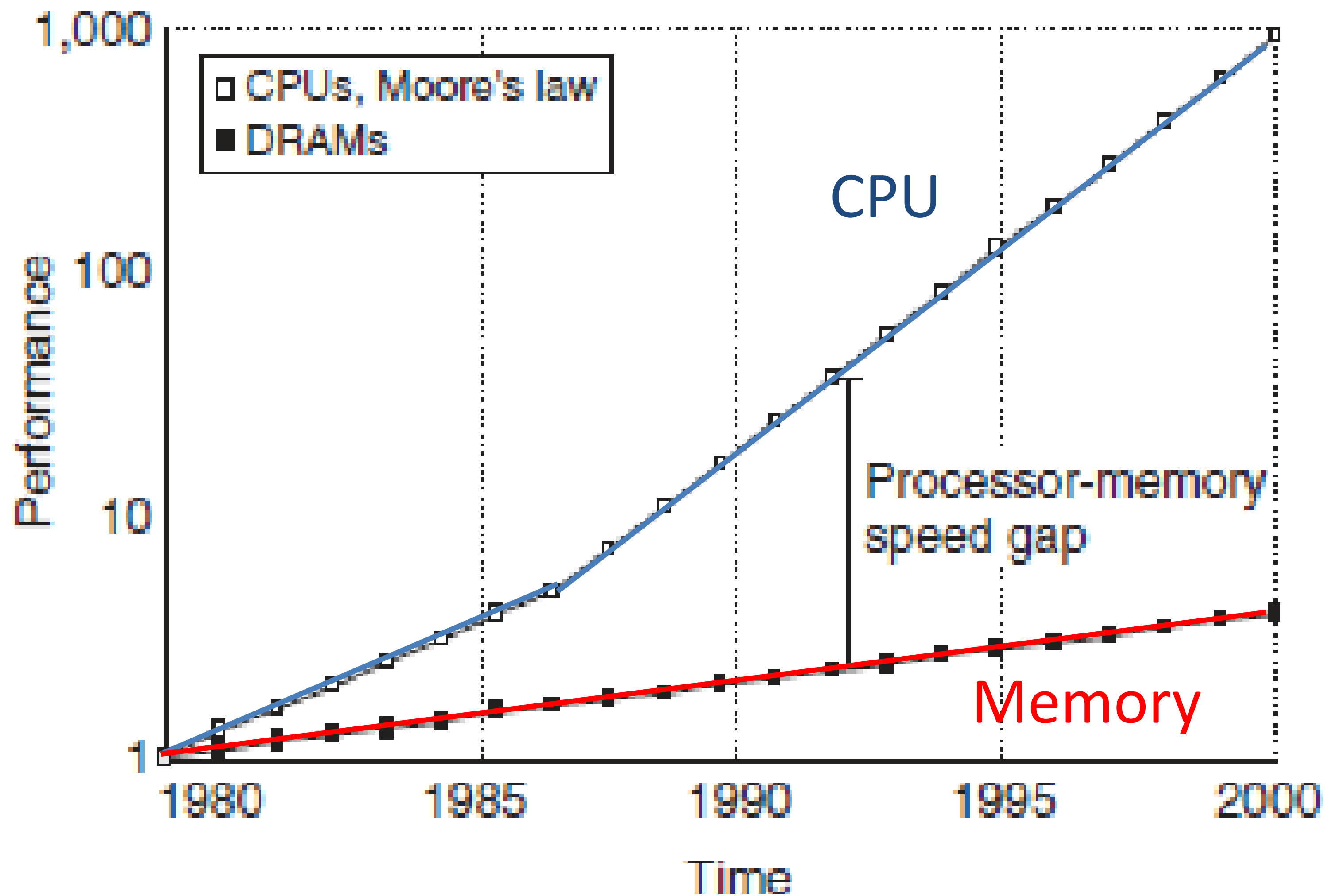
Source: Morgan Stanley

2. The Memory Wall

Feeding Enough Memory to the Computer



The Memory Wall (Processor-Memory Speed Gap)



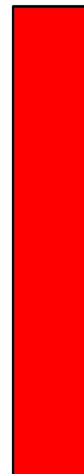
M.S. Parekh, P Thadesar, M. Kakir, ECTC, 2011

BANDWIDTH CONSIDERATIONS

WHAT MATTERS TO THE SYSTEM IS THE AMOUNT OF DATA THAT ARRIVES PER SECOND

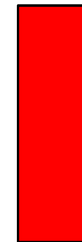
BANDWIDTH = SPEED OF BUS X WIDTH OF BUS

WIDE SLOW BUS **EQUALS** **NARROW FAST BUS**



X64
200 MHZ

1600 MB/SEC

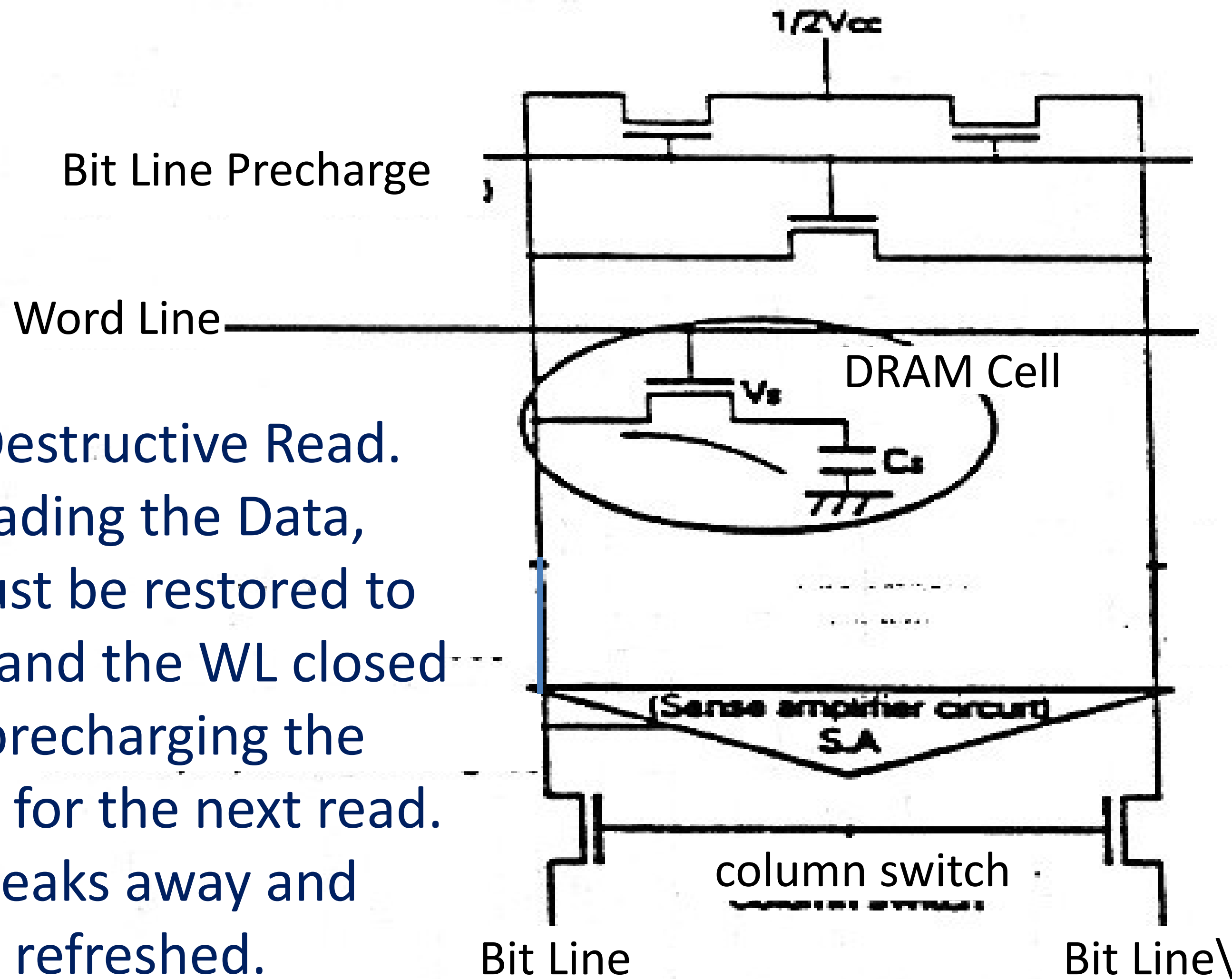


X16
800 MHz

NOTE: "DATARATE" USED TO MEAN BANDWIDTH
IT IS BEING USED NOW TO MEAN SPEED OF ONE BIT OF DATA ON THE BUS
I.E. THE SPEED OF THE BUS

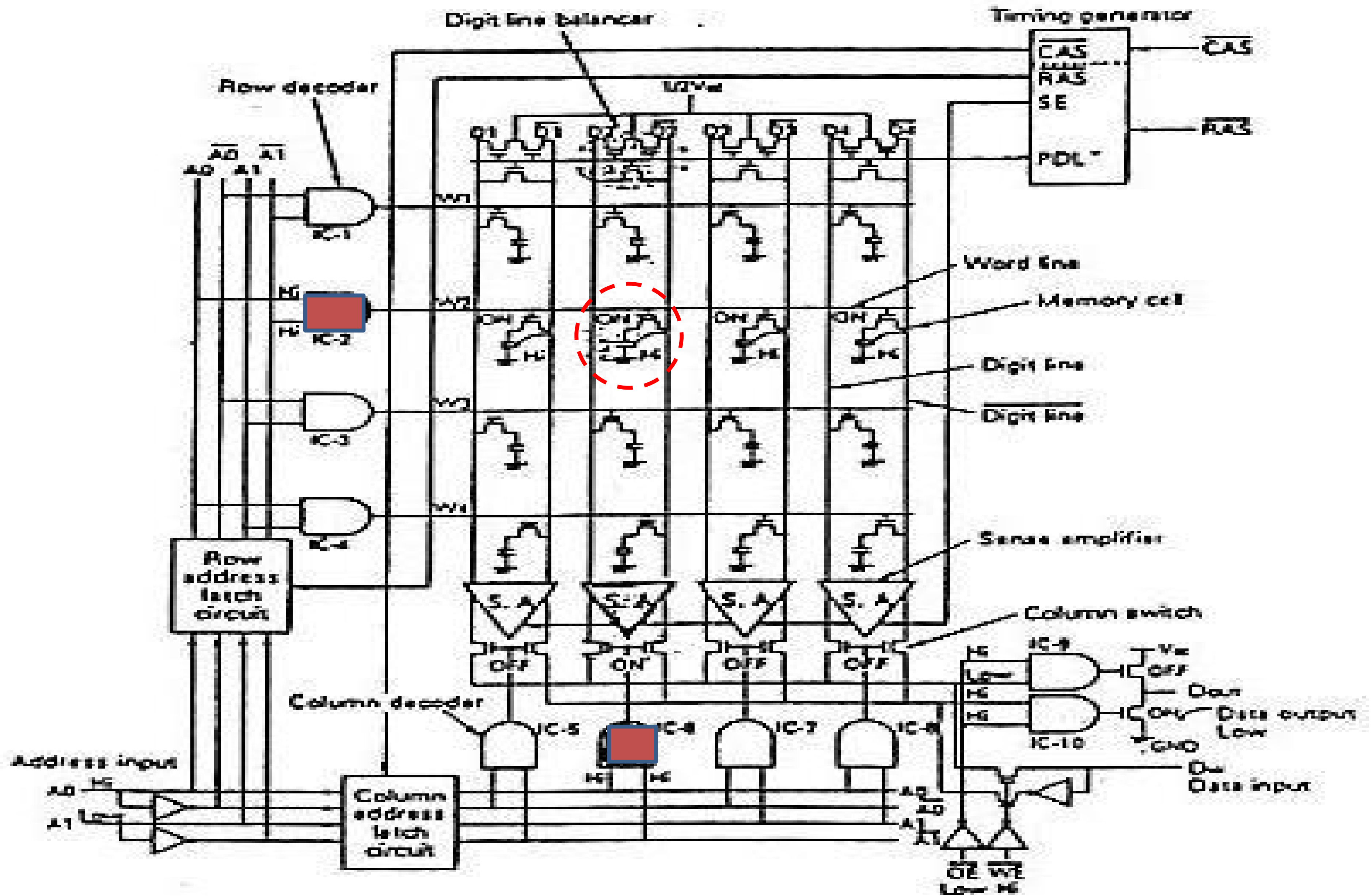
bprince@memorystrategies.com

Internal Function of the DRAM is Slow

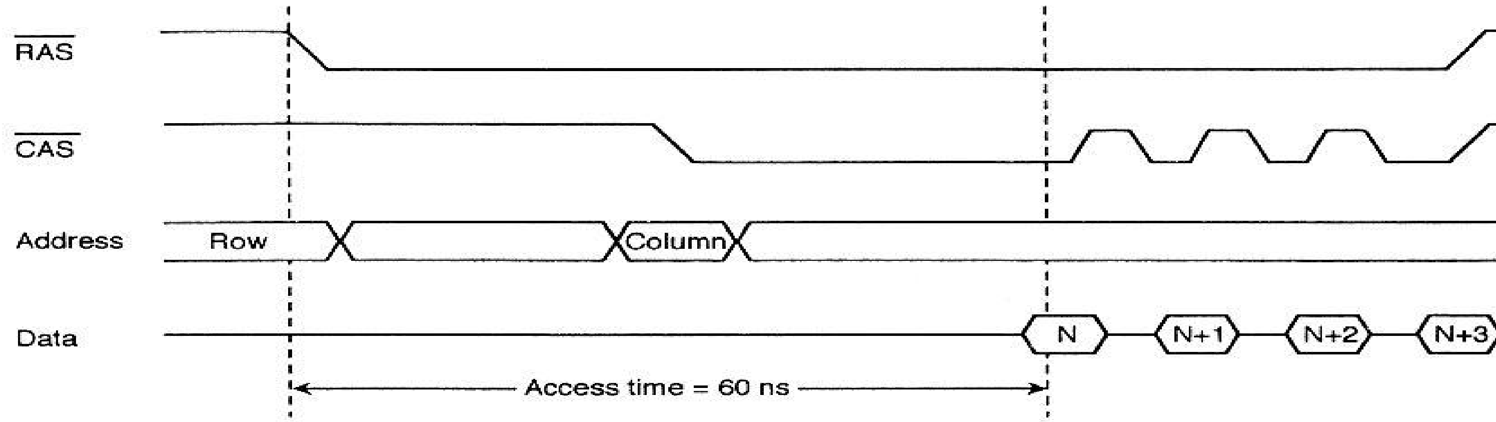


NOTE: Destructive Read. After reading the Data, Data must be restored to the cell and the WL closed before precharging the bit-lines for the next read. Charge leaks away and must be refreshed.

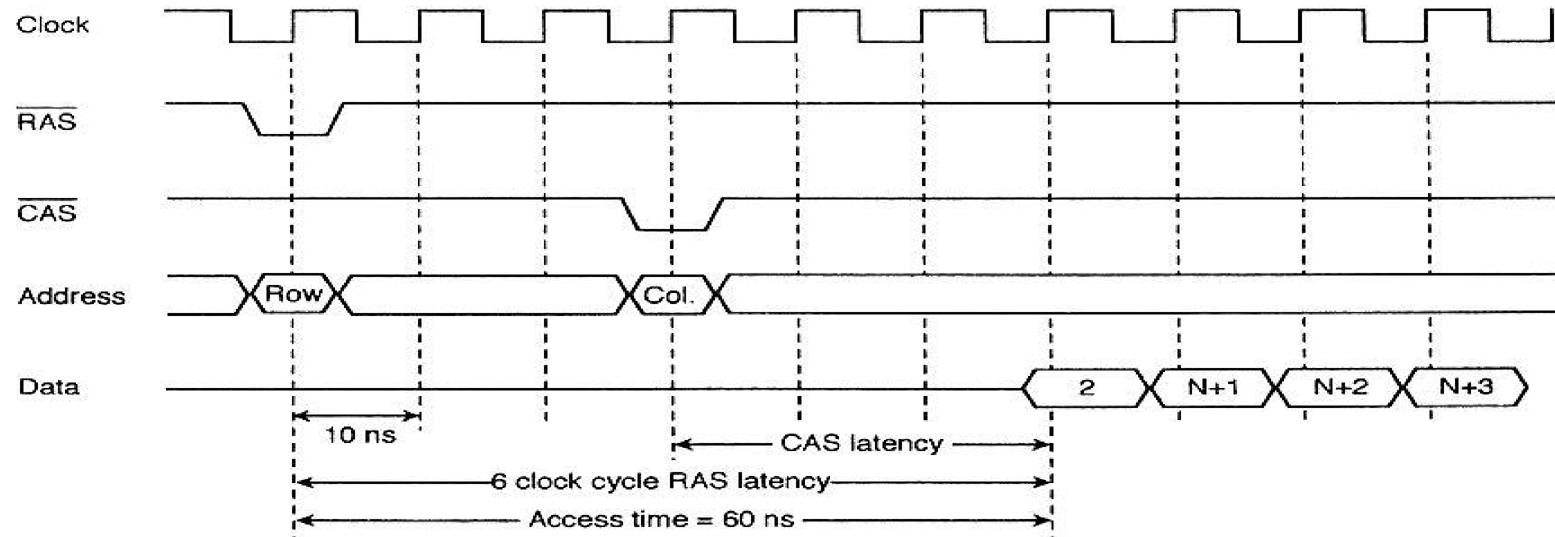
Simplified Circuit Diagram of 16-bit DRAM



SYNCHRONOUS(b) AND ASYNCHRONOUS(a) CONTROL

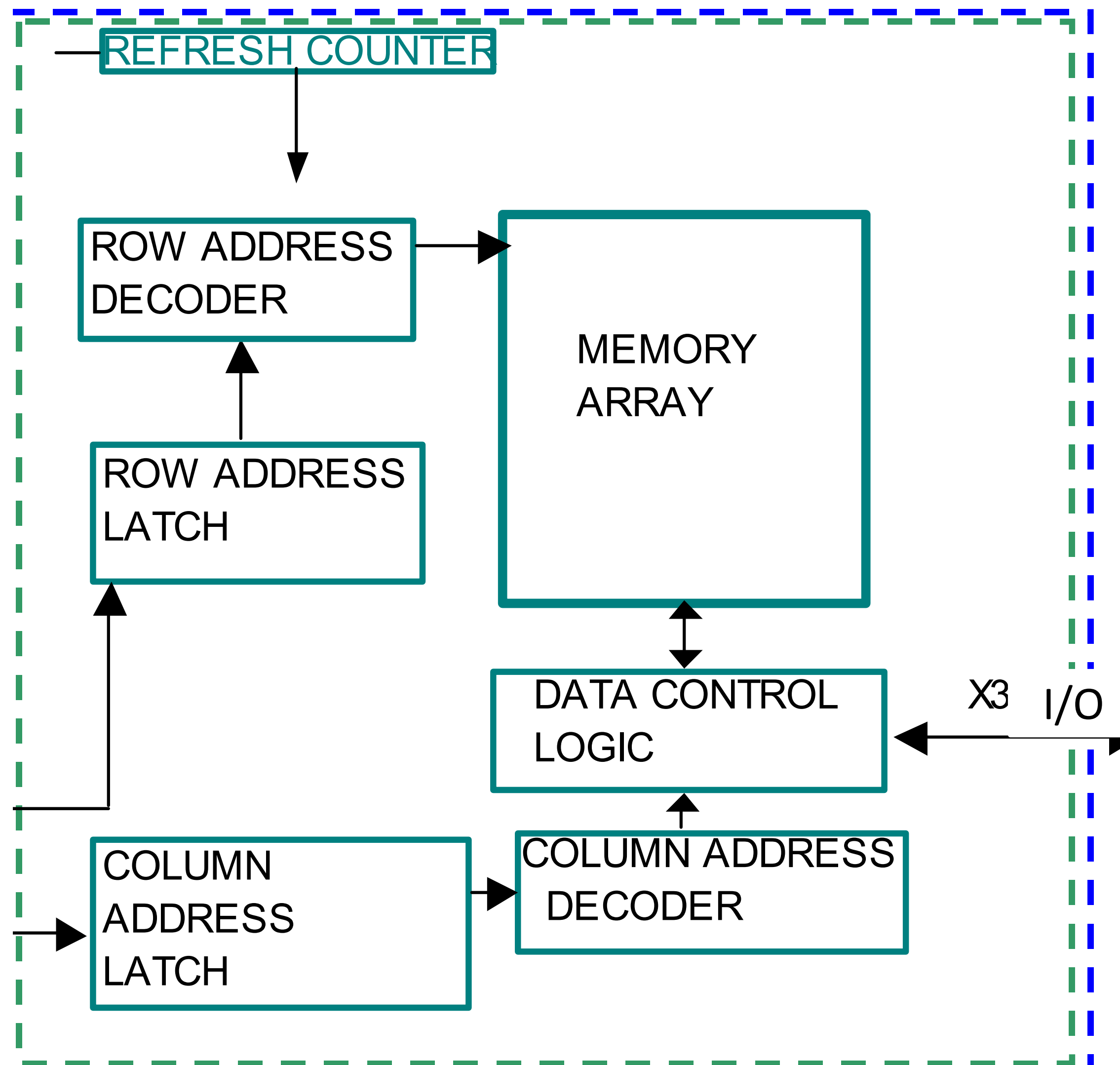


(a)



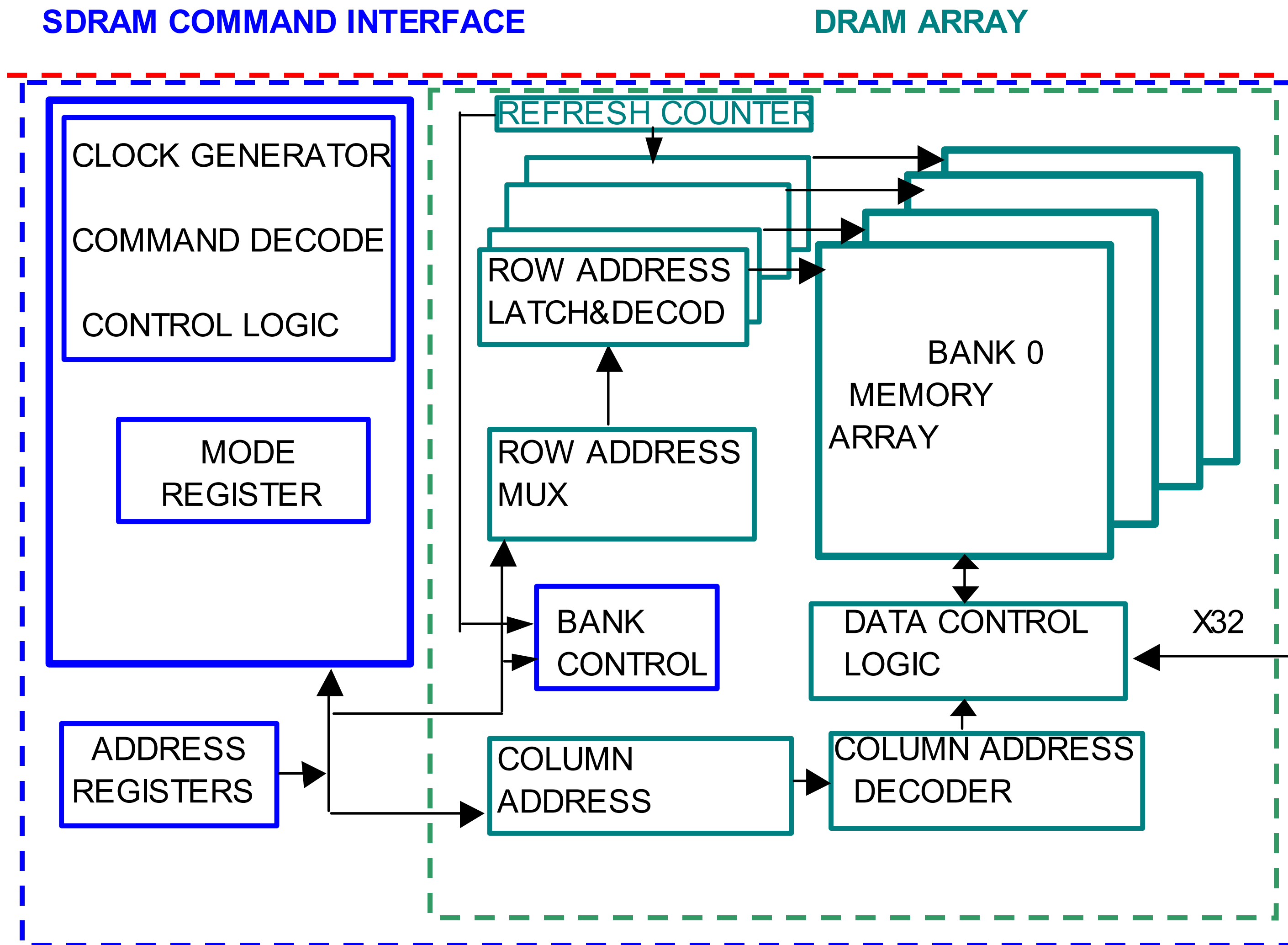
(b)

Basic Asynchronous DRAM



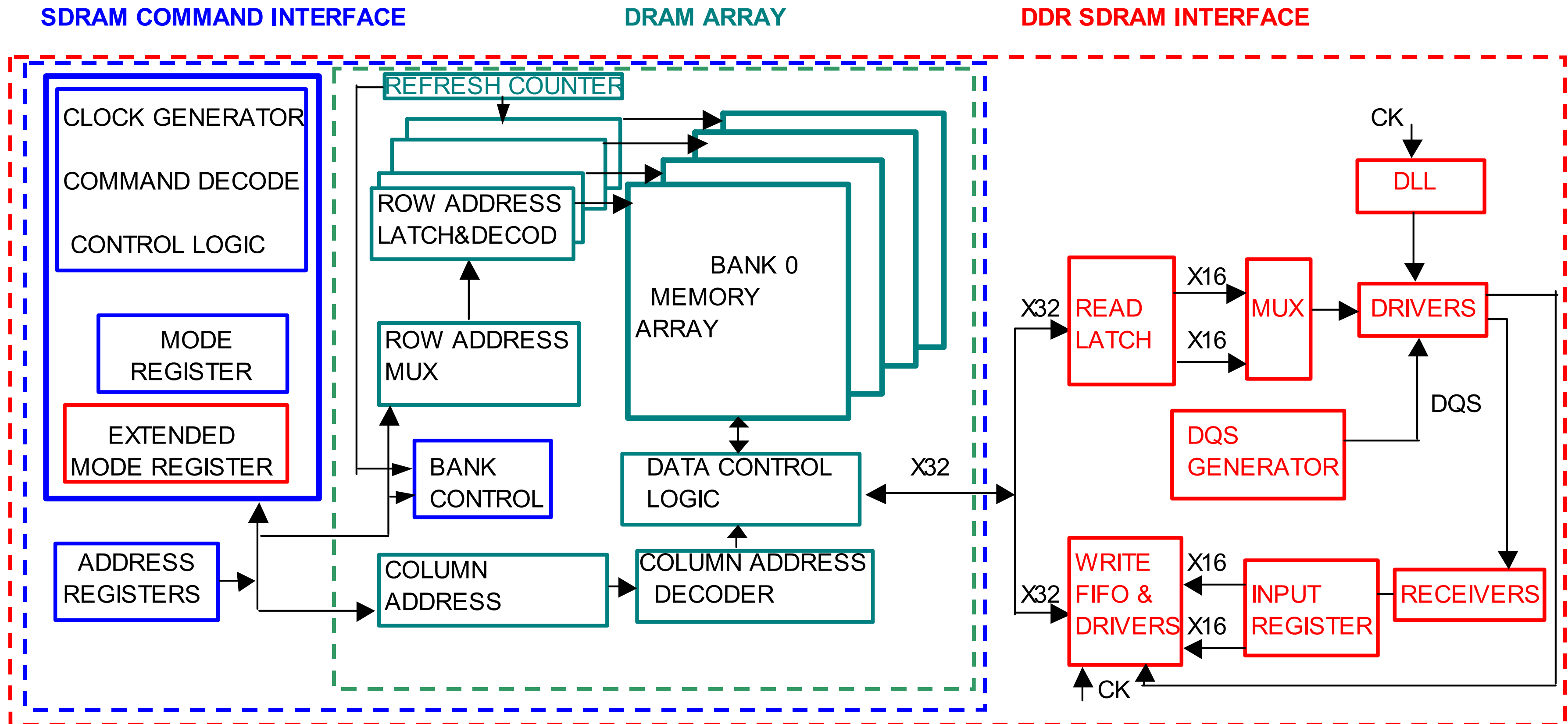
B. Prince, *"High Performance Memories", 1998, Wiley*

Add SDRAM Command Interface



B. Prince, *“High Performance Memories”, 1998, Wiley*

Add DDR SDRAM Interface



B. Prince, *"High Performance Memories", 1998, Wiley*

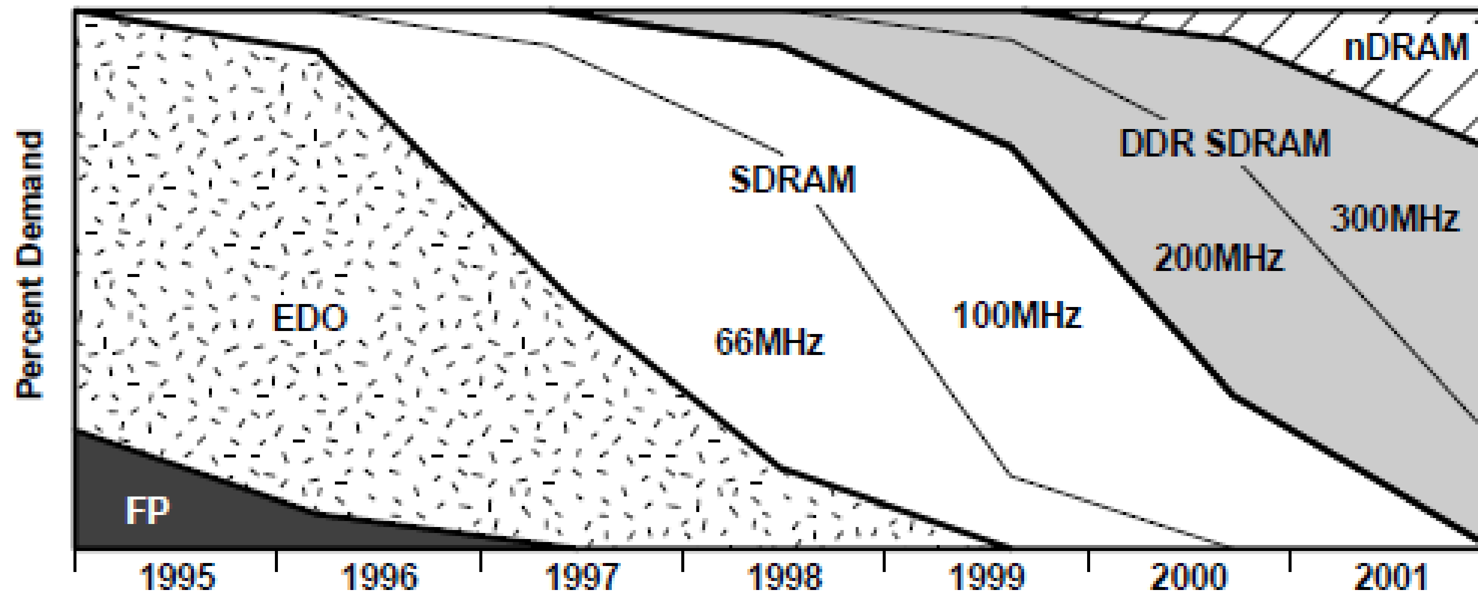
Memory Wall Solution – Clocked Memory

Issues: <u>Memory Wall</u> – Processors faster than DRAMS

Solutions:

1994 Synchronous Interface – ran from computer clock
Improved Data Rate, but not latency

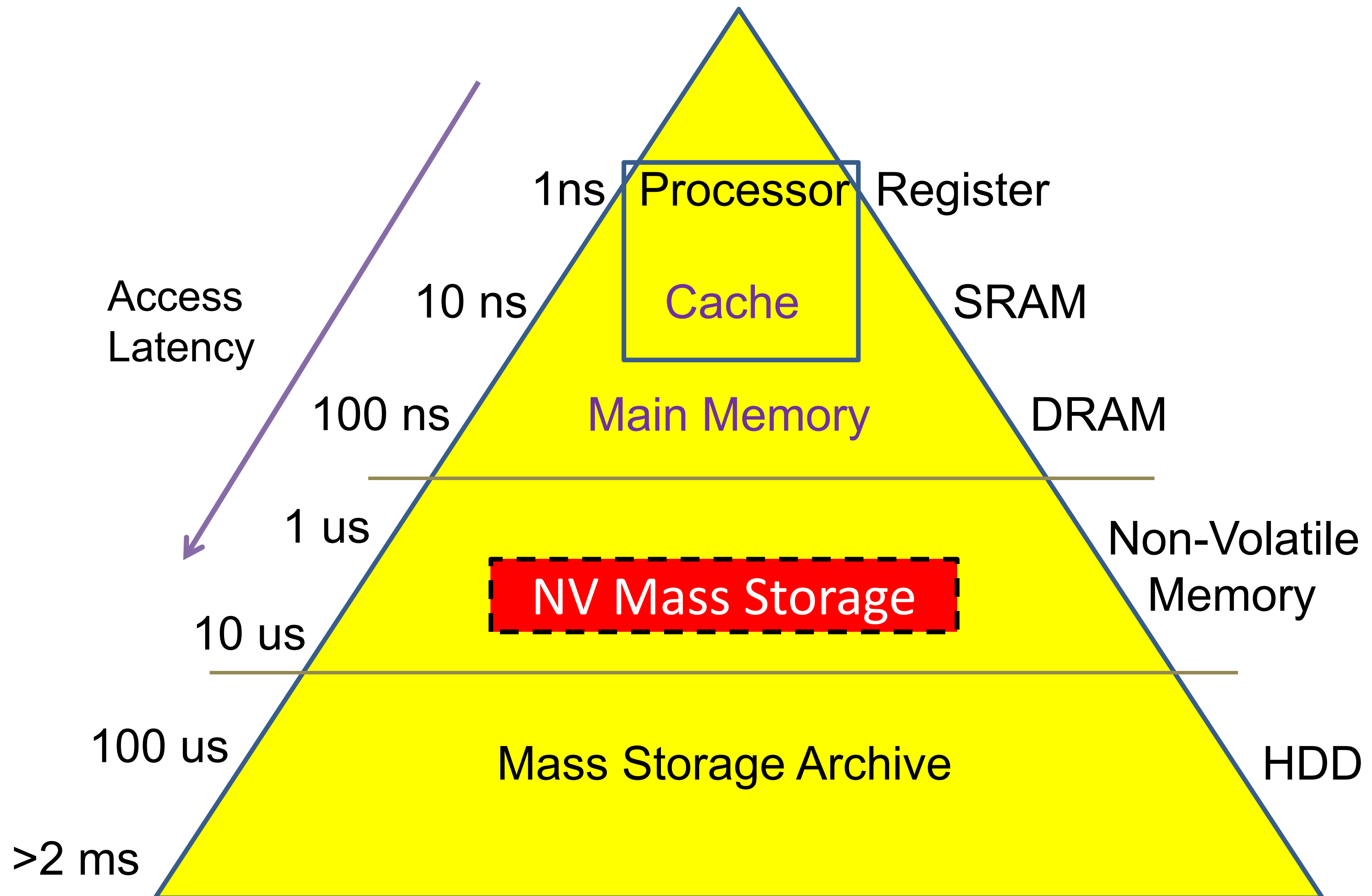
1998 DDR – doubled datarate , latency same.



DRAM APPLICATIONS (2000)

MARKET SEGMENT	APPLICATION	DRAM TYPE
COMPUTER	SERVER/WORKSTATION	SDRAM/DDR/DDR2
	DESKTOP	SDRAM/DDR/RDRAM
	ADD-ON MODULES	SDRAM/DDR
	PERIPHERALS	SDRAM
PERIPHERALS	PRINTER	WIDE SDRAM
	COPIER	STANDARD DIMMS
NETWORKING	SWITCHES/ROUTERS	AS-DRAM/LL SDRAM
	NETWORK PROCESSORS	EDRAM/LL SDRAM
	INTERNET APPLIANCES	LP-DRAM
	SET TOP BOX	SDRAM
CONSUMER	PDA/DIARIES	LP-DRAM
	HANDHELD	SDRAM
	GAMES	LL SDRAM/RDRAM

Cache Memory Hierarchy



3. Non-Volatile Memory

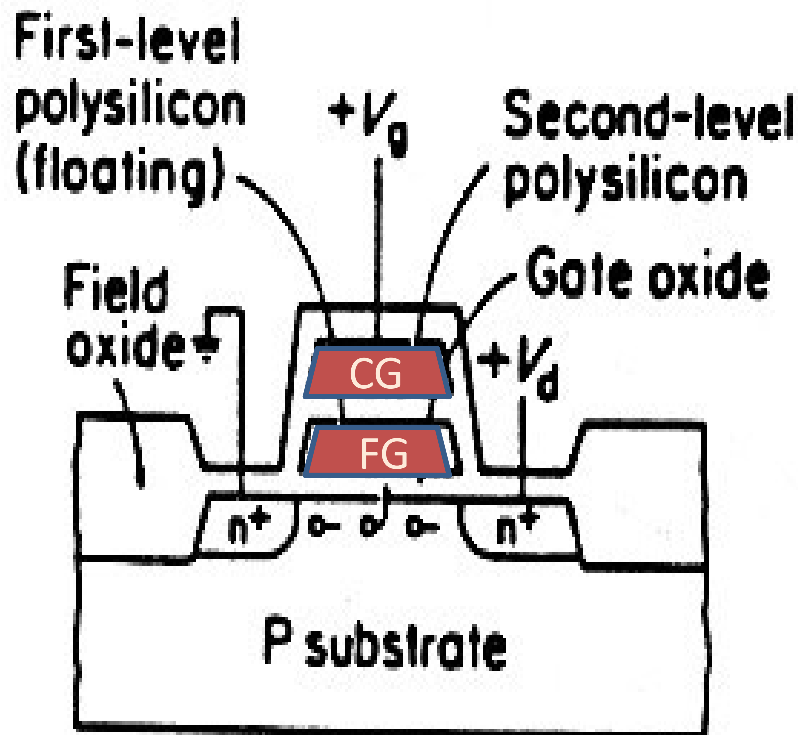
1970's		1980's		1990's
ROM →	EPROM →	EEPROM →	NOR Flash →	eFlash in MCU

1990's	2000	2010	2020
NAND Flash	Highest Volume	End of Scaling	3D Vertical NAND

EPROM – Erased by UV Light

bprince@memorystrategies.com

UV-EPROM Cell Structure

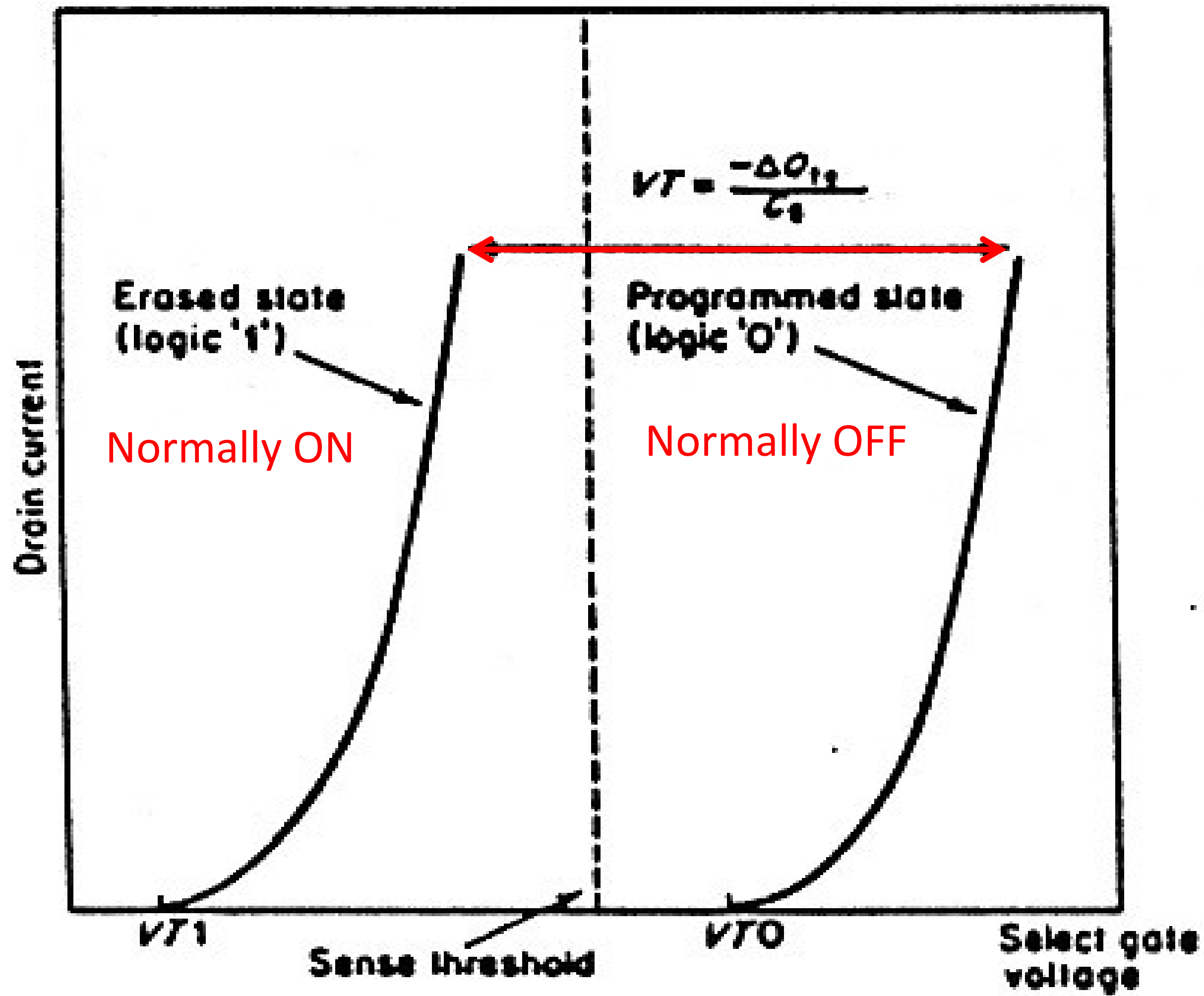


The UV-EPROM was electrically programmed, but erased by irradiation with ultra-violet (UV) light.

It had two polysilicon gates. The upper control gate is used for selection and is connected to the wordline. The isolated floating gate is capacitively coupled to the control gate and the substrate.

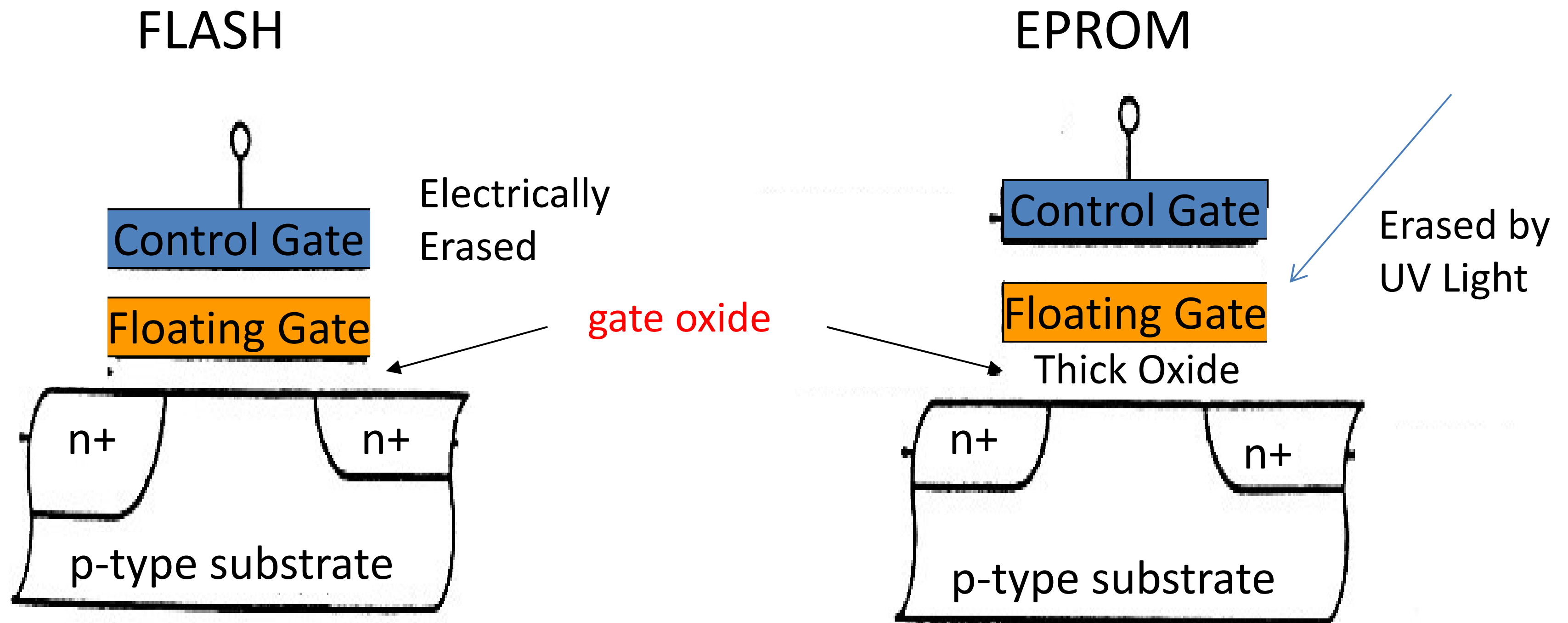
A charge on the floating gate alters the threshold voltage as seen from the control gate. When the floating gate is charged, the threshold voltage increases and a higher voltage is needed on the control gate to turn the transistor on.

EPROM PROGRAM/ERASE I-V CHARACTERISTIC



bprince@memorystrategies.com

Early FLASH Cell vs. EPROM Cell



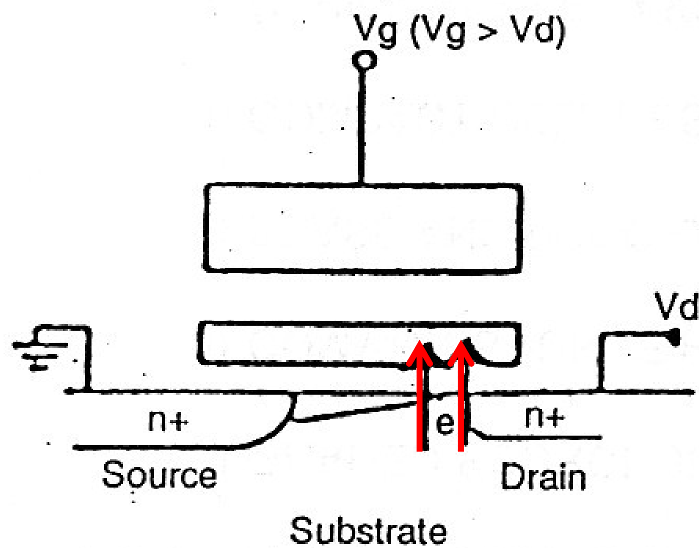
Flash Cell

EPROM Cell

Operation of the Stacked Flash EEPROM Cell

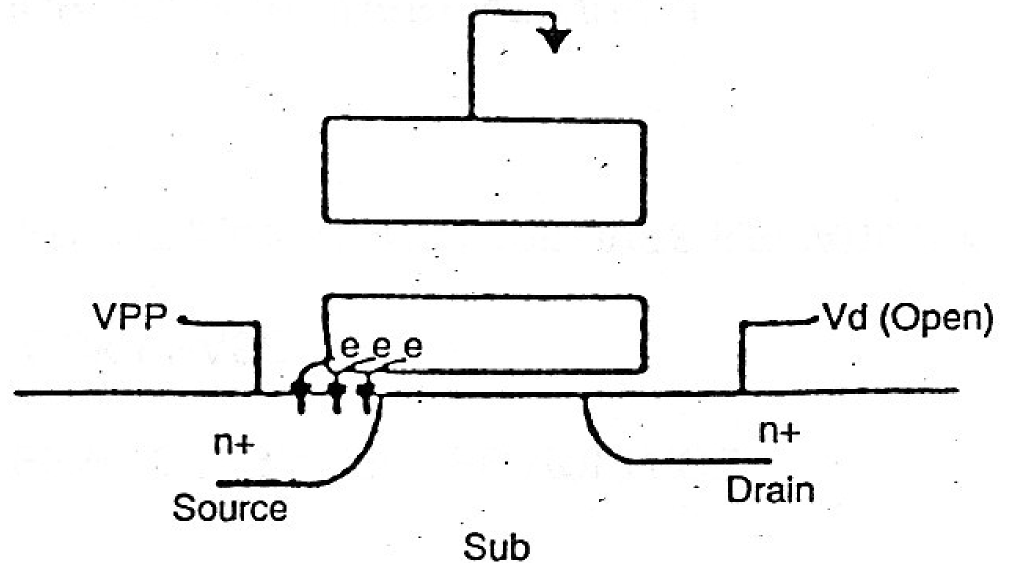
Programming

By CHE Injection



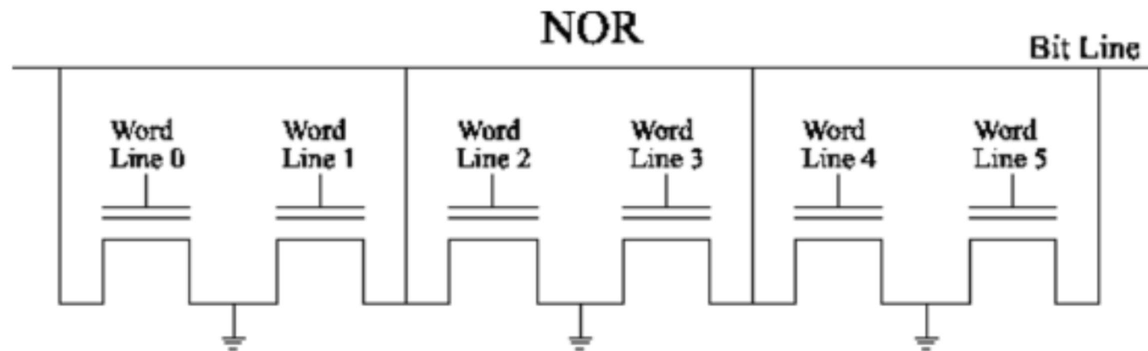
Erase

By FN Tunneling from Floating Gate to Source



The NOR Flash had three contacts to each cell: Source, Drain, Gate

NOR Flash Circuits



A NOR Flash array could be integrated onto the processor chip along with the SRAM. The era of the Microcontroller had begun

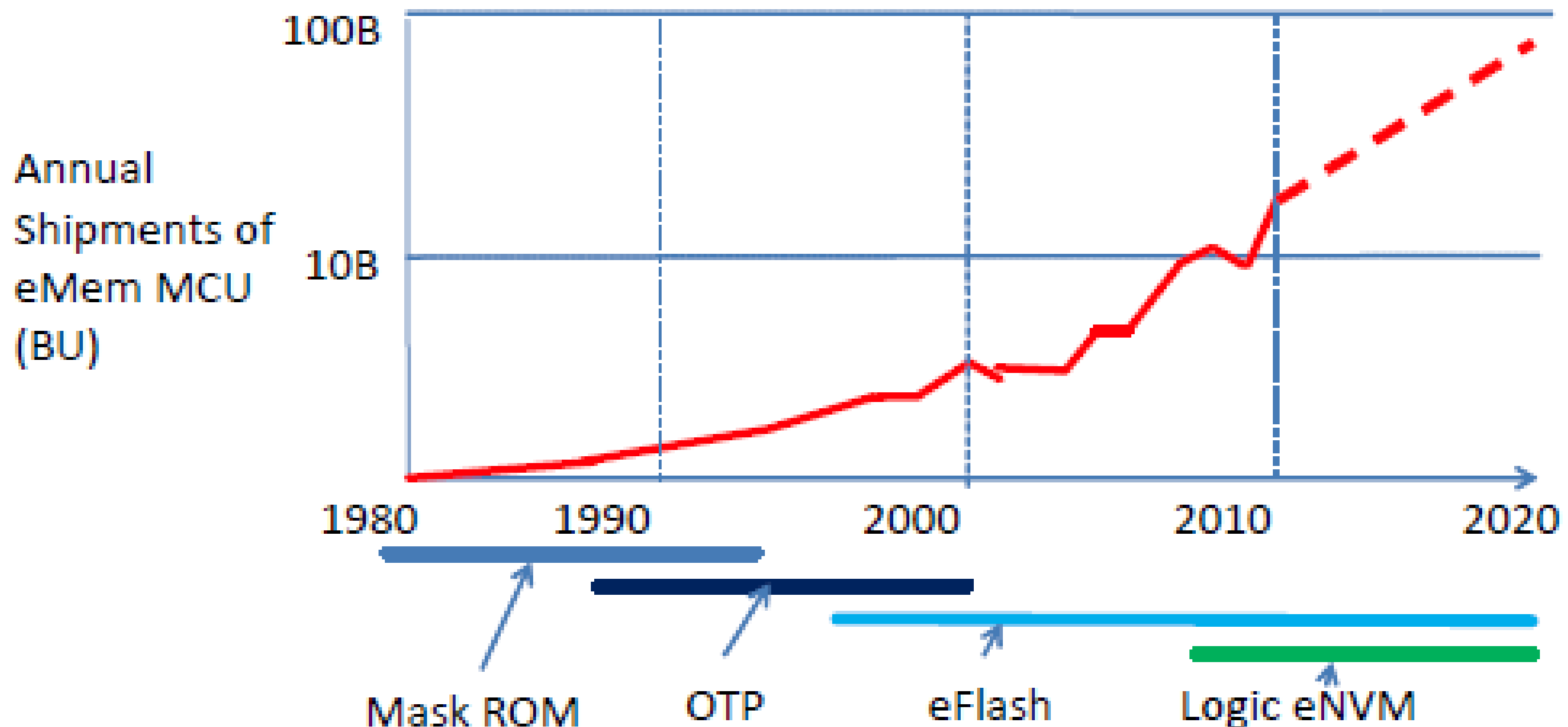
NOR FLASH APPLICATIONS

COMMUNICATIONS	**CELL PHONES**
	MODEMS
	NETWORKING EQUIPMENT
COMPUTER	**PC BIOS**
	FLASH CARDS
	DISK DRIVES
CONSUMER	**SET TOP BOXES**
	DIGITAL CAMERA'S
	WIRELESS
	PDA'S
INDUSTRIAL	**AUTOMOTIVE**
	DISTRIBUTED CONTROL
	CONFIGURATION STORAGE

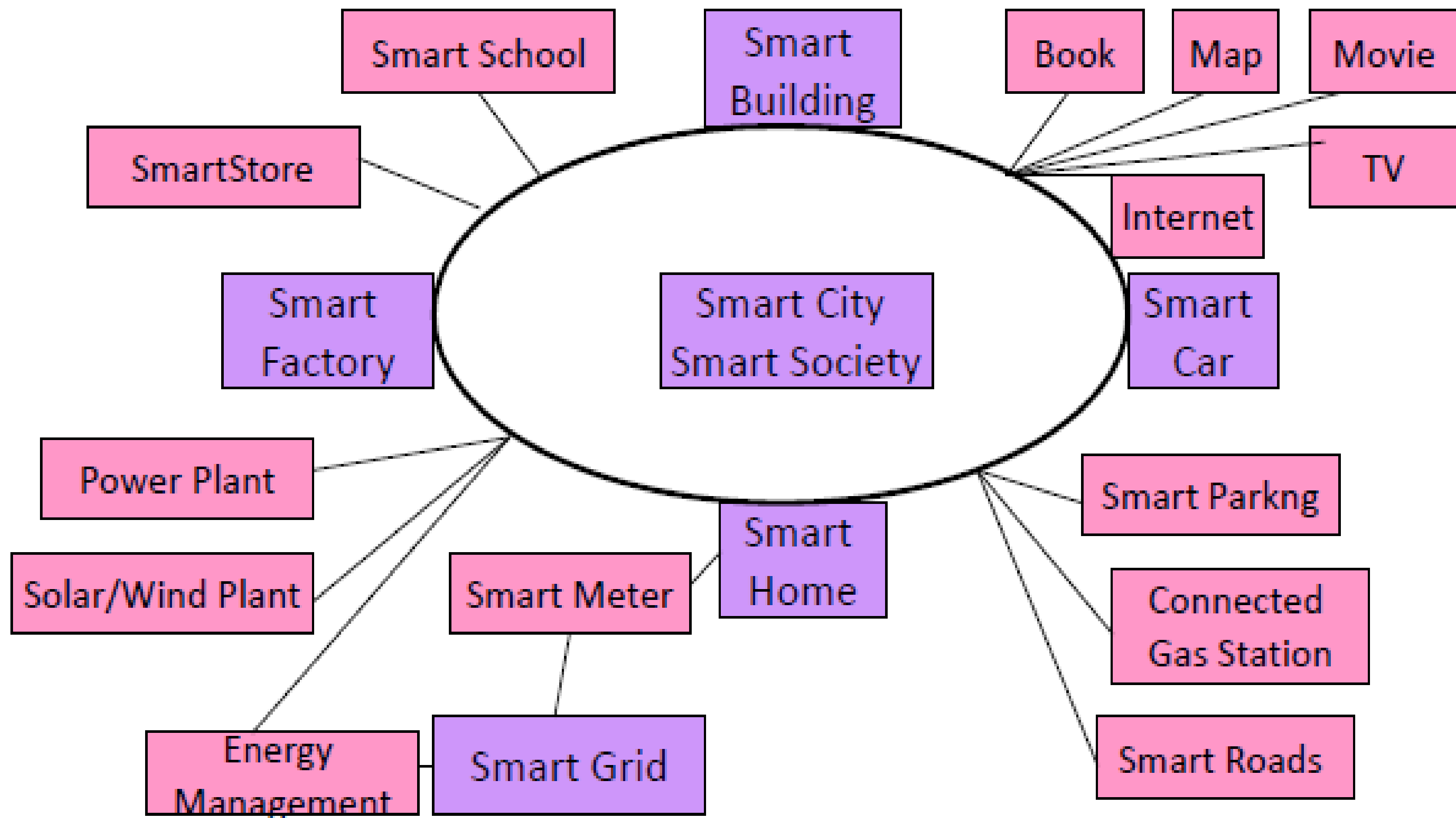
4. Embedded Memory

Single Chip Microcontrollers
(MCU + SRAM + Flash)

Microcontrollers (MCU + SRAM + Flash)



MCU Applications for Smart Society and Smart City

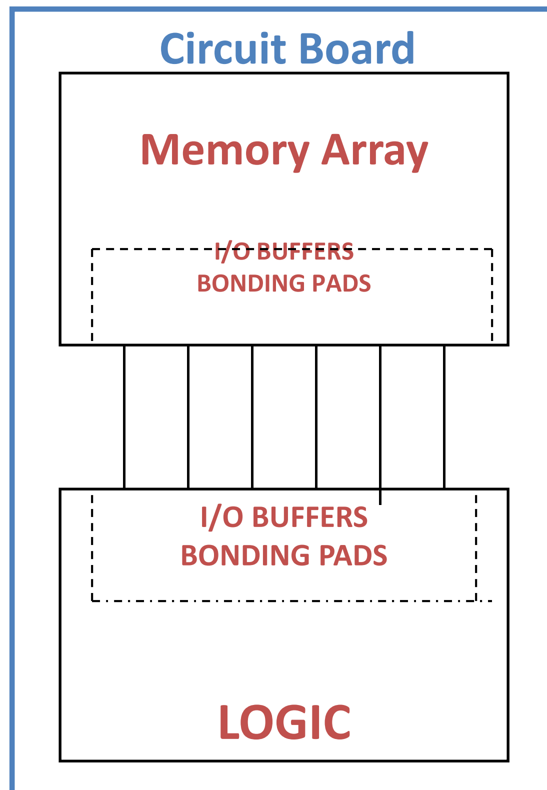


B.Prince, D. Prince, Adapted from *Memories for the Intelligent Internet of Things*

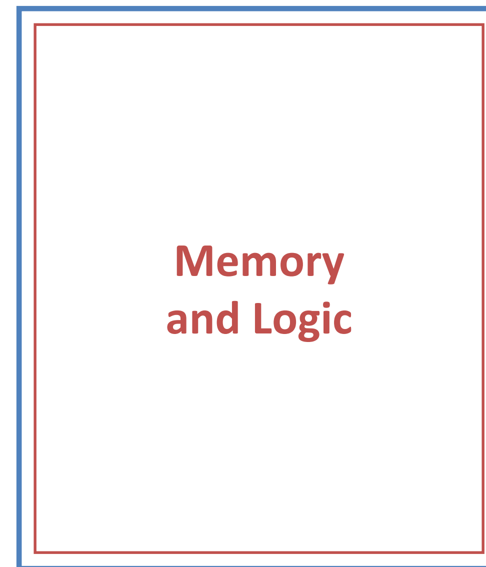
System Problems solved by integration

System Problem	On-Chip Solution
Tight System Form Factor	Reduces package count Reduces board size
Power Reduction	Reduces weight of battery Increases life of battery Reduces cost of cooling
Bandwidth limited	Permits wider internal bus Fewer external I/Os and wires Reduces System EMI Reduces Ground Bounce
Granularity	No silicon wasted on standard sizes Non standard configurations okay

Embedded Memory Reduces Silicon and Board Area Resulting in Smaller Formfactor



Separate Chips



Integrated Chip

bprince@memorystrategies.com

Power Reduction by Embedding Memory

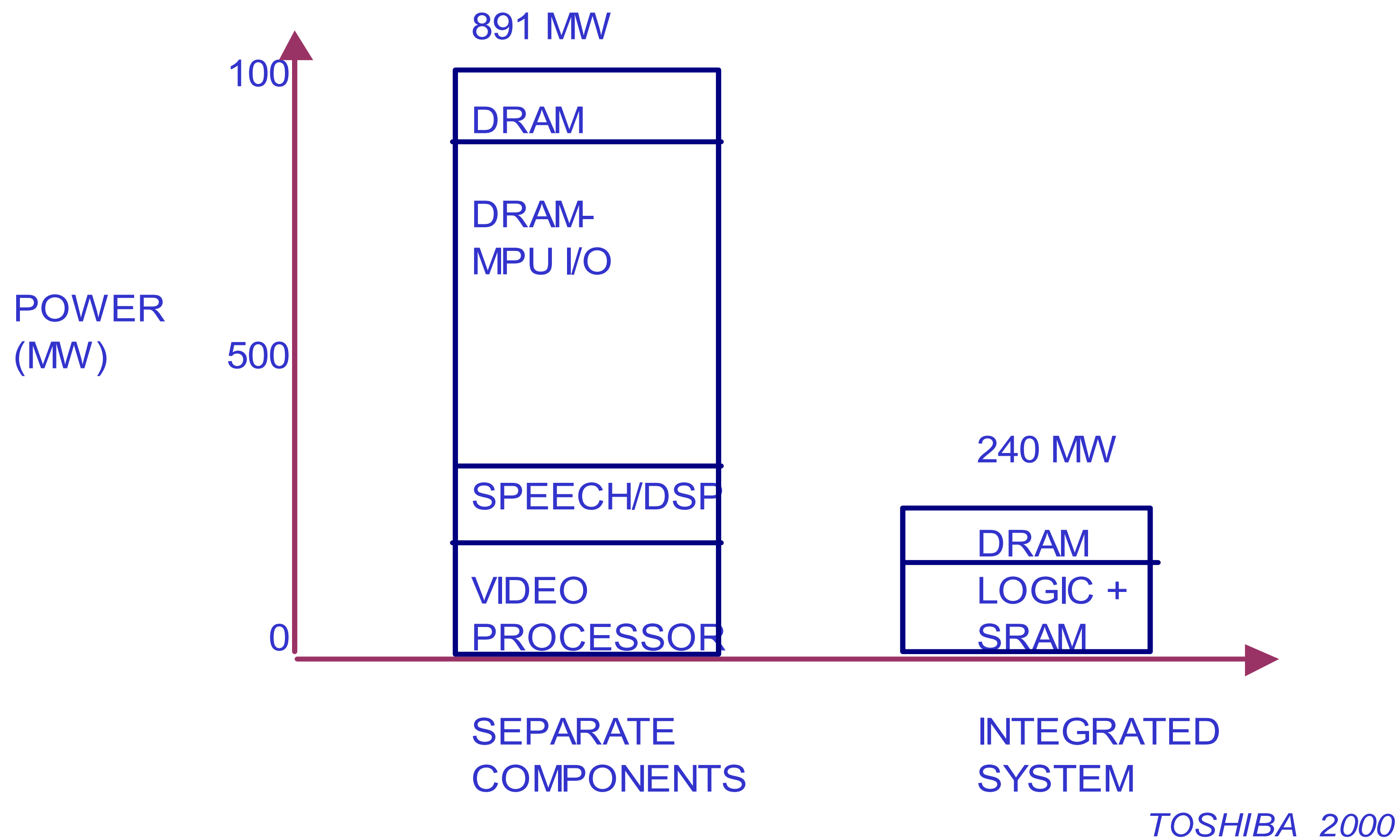
(Power = 1/2 Capacitance x Voltage2 x Frequency)

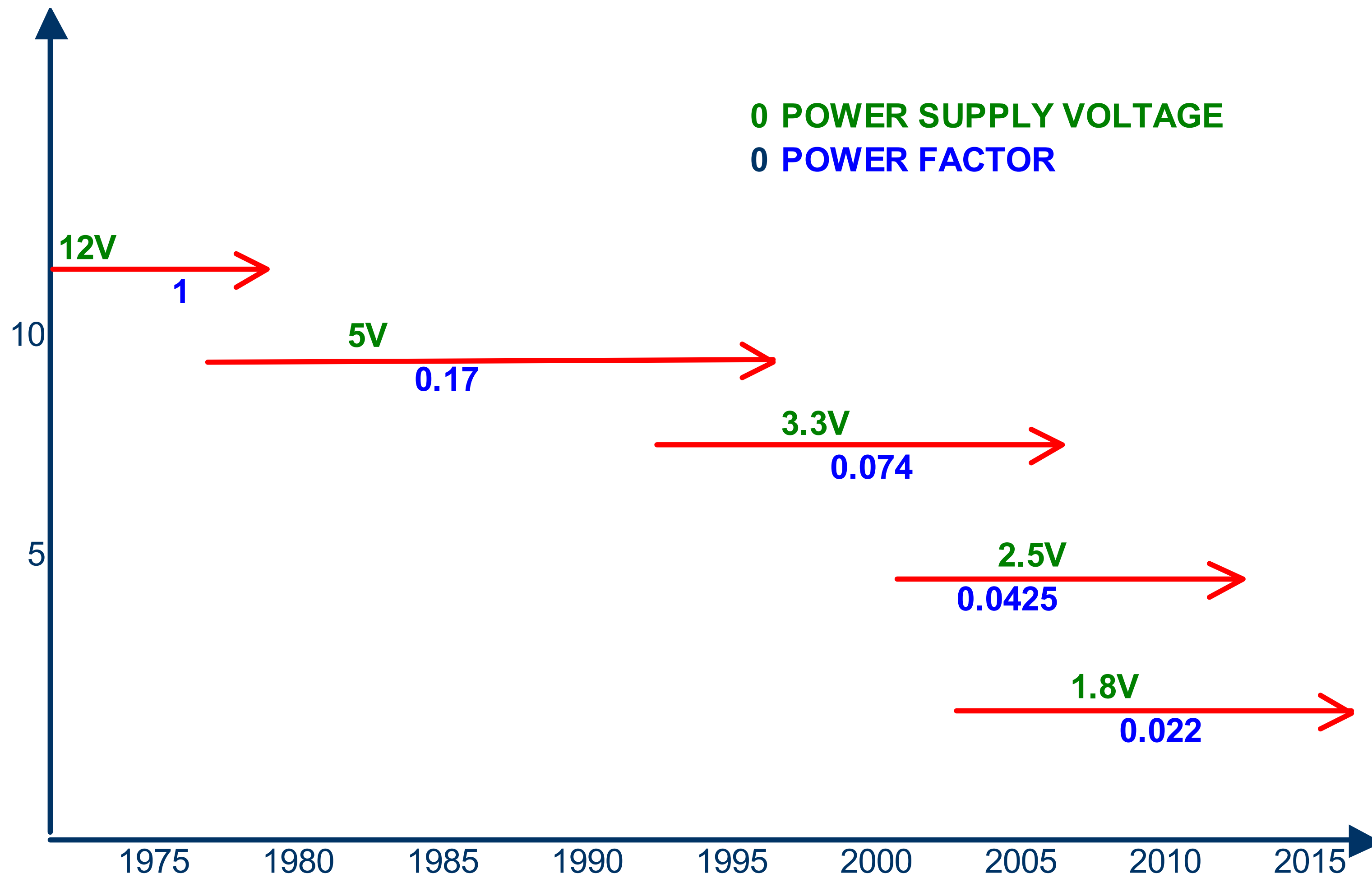
Embedding Memory:

- Reduces Capacitance of the Interface
- Reduces Required Speed of Operation (wider bus)
- May reduce voltage swing

VIDEO PHONE APPLICATION

EXAMPLE OF POWER SAVING IN EMBEDDED DRAM




POWER SUPPLY VOLTAGE INTERFACE TRENDS
0 POWER SUPPLY VOLTAGE
0 POWER FACTOR
12V
1
10
5V
0.17
3.3V
0.074
5
2.5V
0.0425
1.8V
0.022
1975
1980
1985
1990
1995
2000
2005
2010
2015

Integrating for Higher System Bandwidth

Higher Speed and Wider Bus on Chip

1. Reduce RAM – Processor Speed Mismatch

2. Solve Narrow RAM I/O Bandwidth Limitations

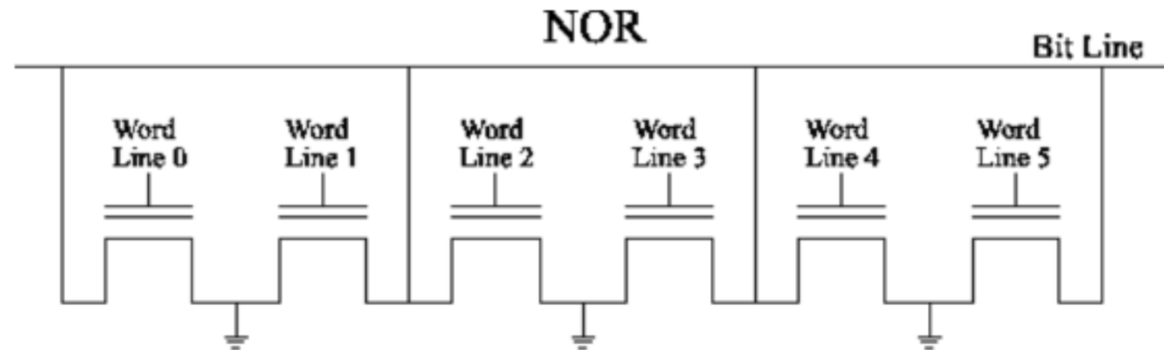
3. Eliminate Inter-Chip Transmission Line Effects
   - Ground Bounce (I/O's Switching) (L DI/DT Effect)
   - Reflections on System Wires

4. Reduce System Level EMI Radiation

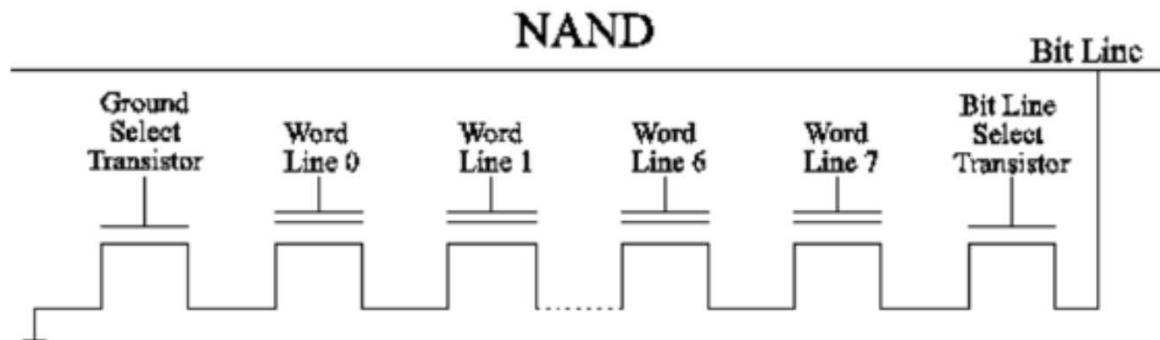
5. NAND Flash

NOR and NAND Flash Circuits

3 contacts to every cell

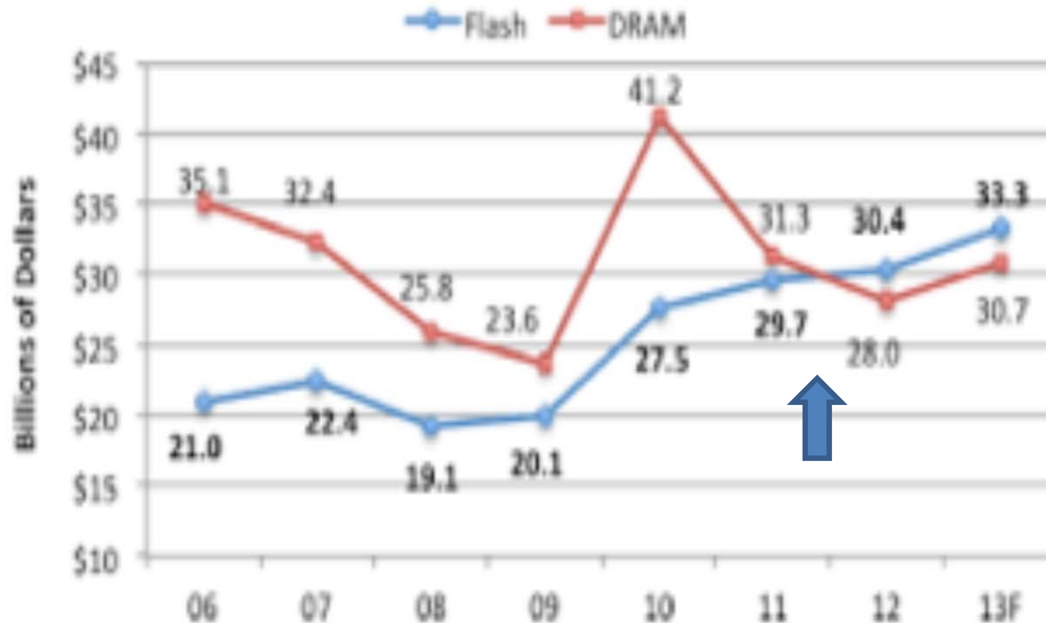


1 contact to every cell



NAND Flash

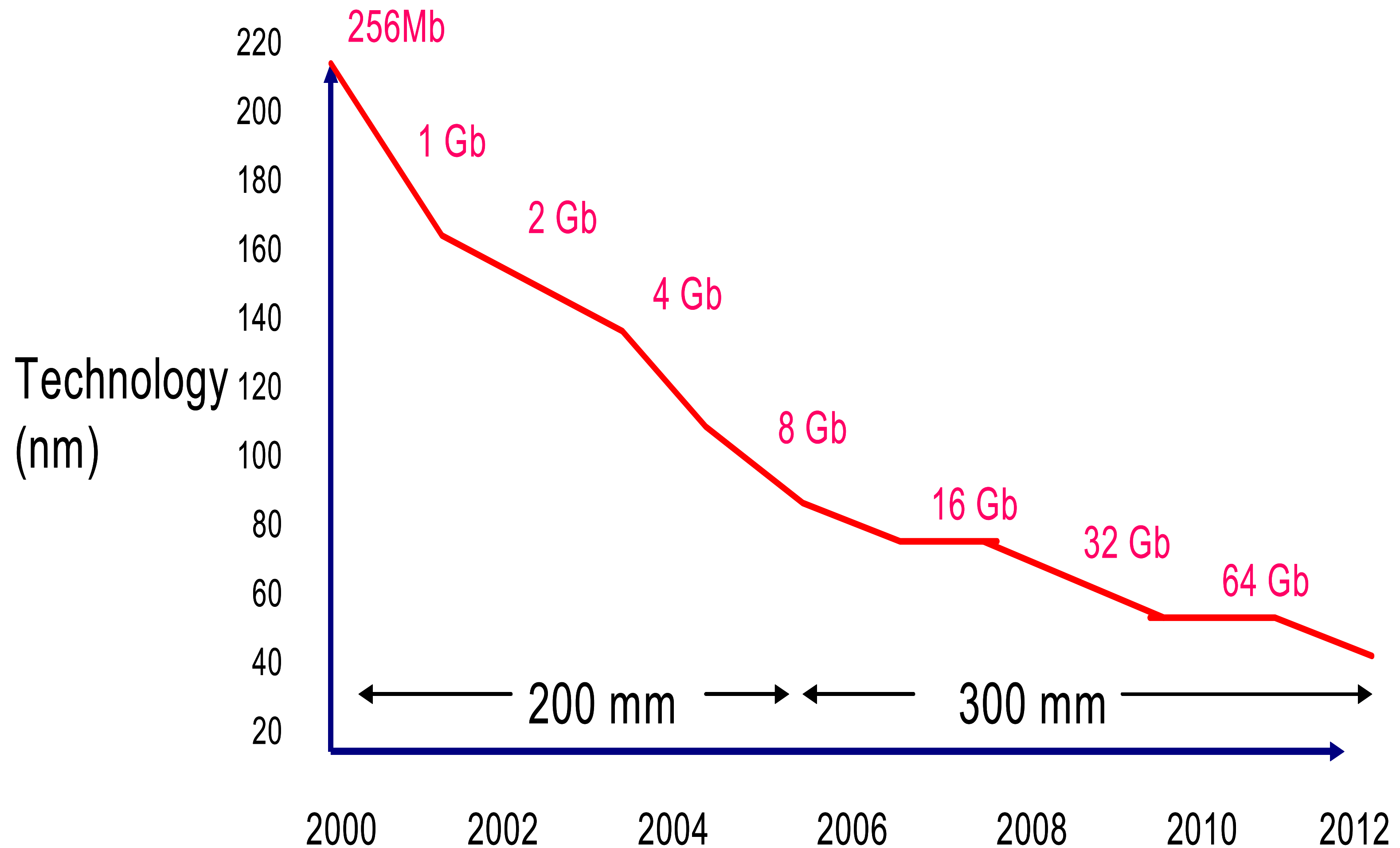
From Emerging Memory to Highest Volume Memory



WSTS, IC Insights

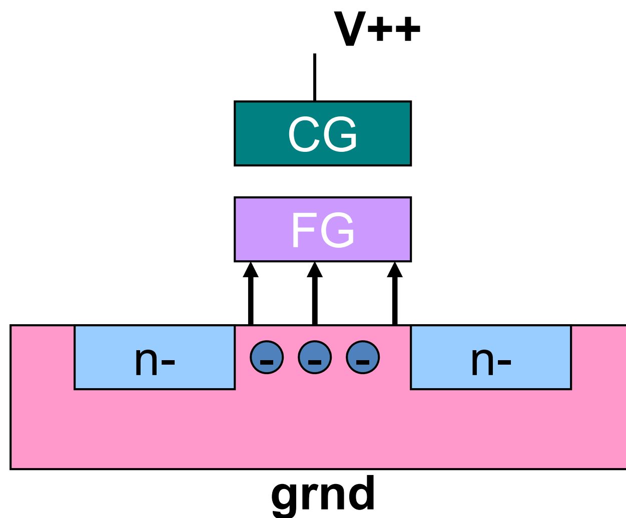
NAND flash was first shown at the IEDM in 1987 by Toshiba. It became the Largest Memory Market in 2011 when it passed the DRAM in Market Value, 24 years later.

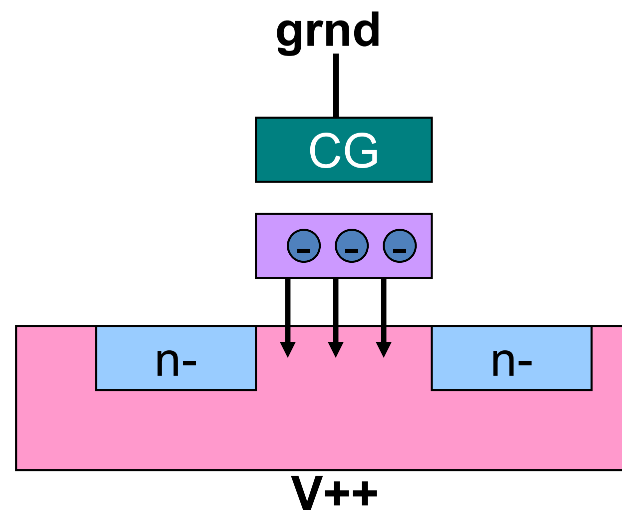
NAND Flash Bits per Chip by Technology and Year



Sandisk ISSCC 2012

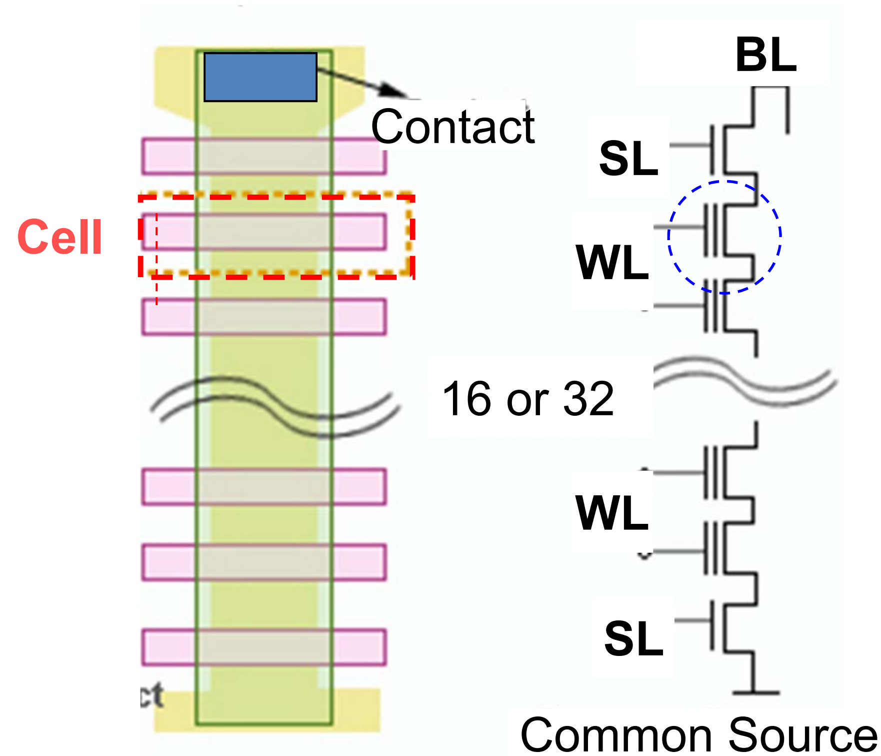
NAND Cell Operation





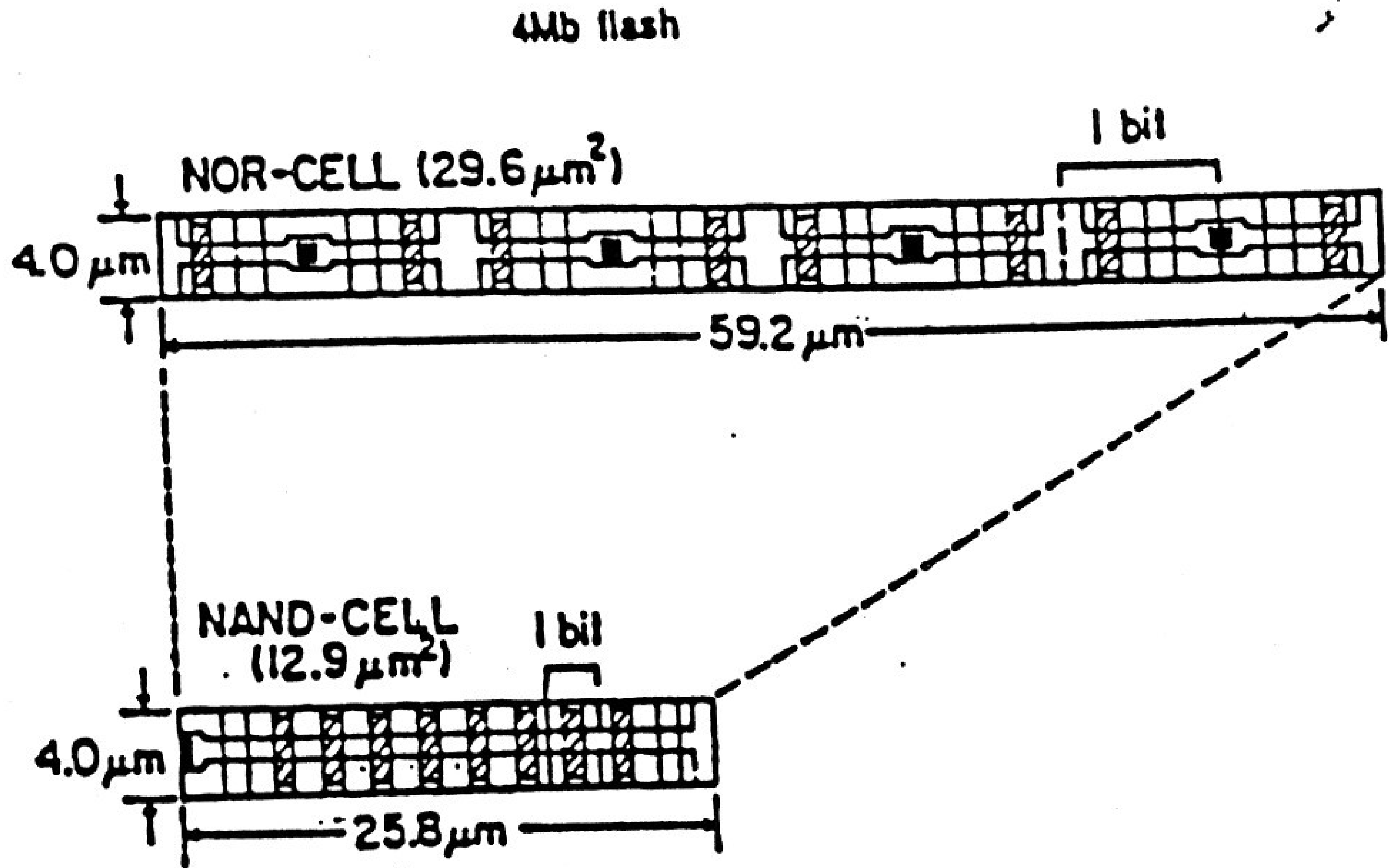
Double Poly Flash Memory Using Fowler Nordheim Channel Tunneling

NAND Flash String

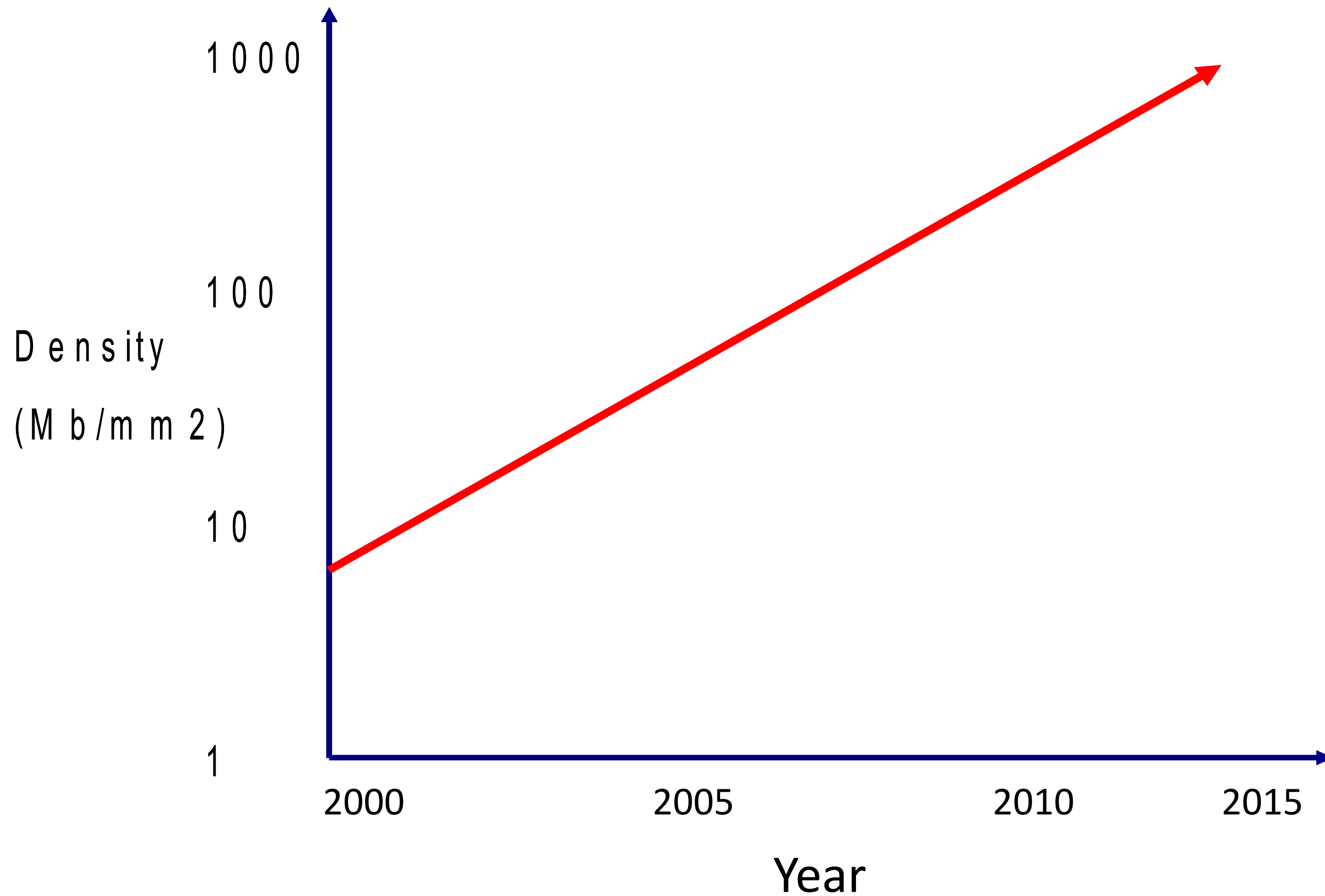


One Independant Contact per Cell

Comparison of conventional NOR cell with NAND cell layout in same technology



NAND Flash Density Trend by Year 2000 - 2013



Nearing End of Moore's Law*

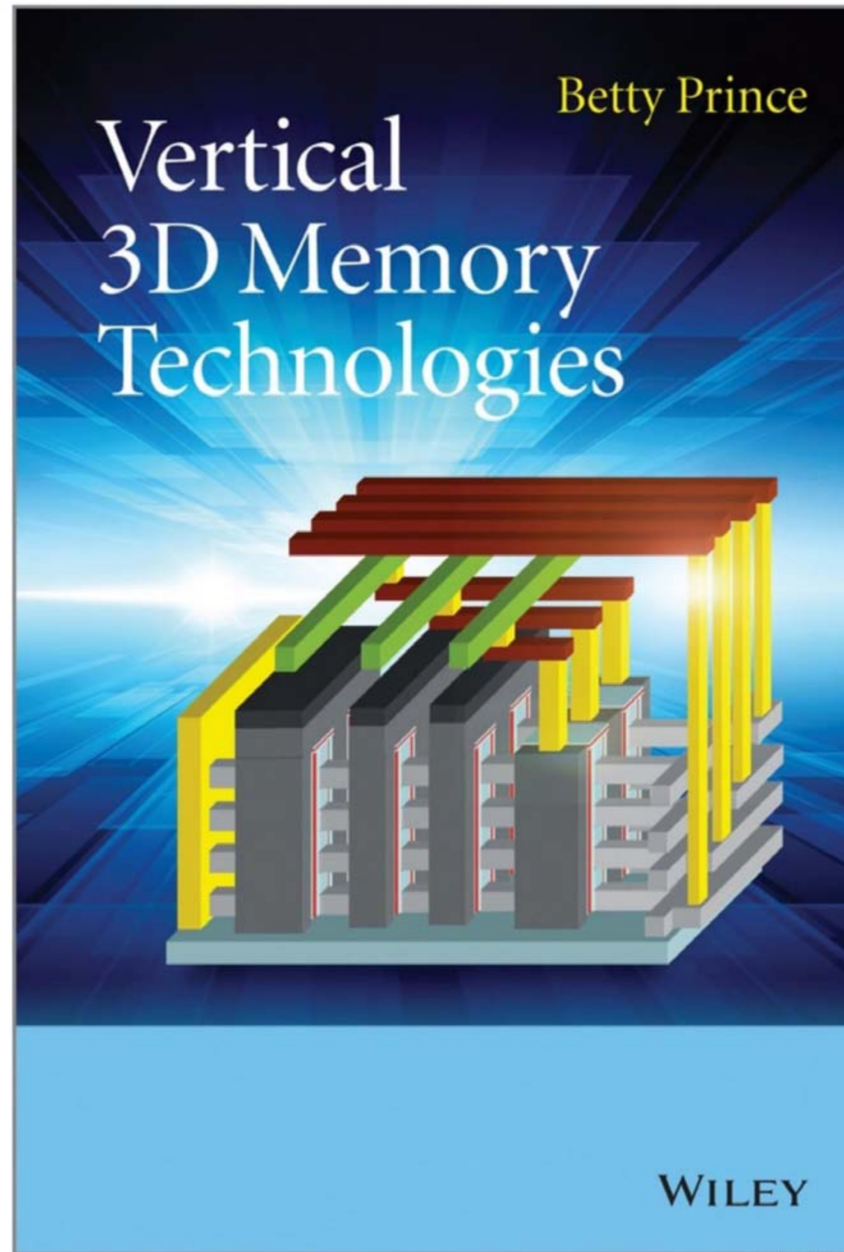
Options:

1. Vertical technology (3D Vertical NAND Flash)

2. New technology (MRAM, FeRAM, RRAM, PCM)

*Moore's Law - The number of transistors on a chip doubles every two years.

6. Vertical Technology



3D Vertical NAND Flash

The successor to the NAND Flash, the GAA Vertical NAND Flash was introduced in 2009 and by 2019 dominated the NV solid state storage market using four new technologies:

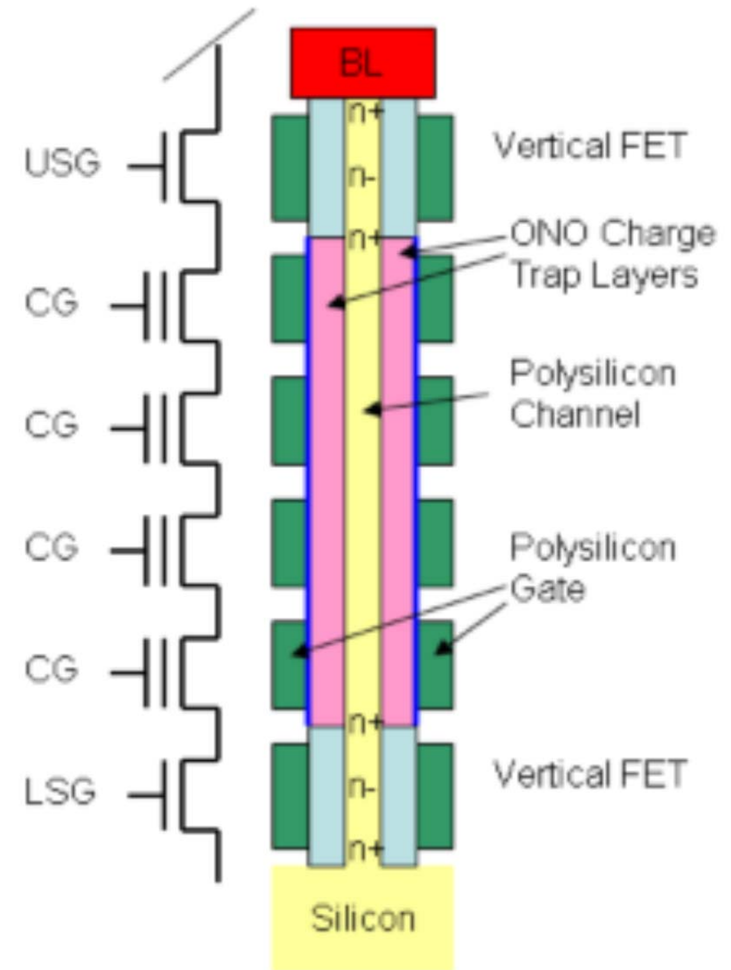
1. SONOS Charge Trapping Memory
2. Barrier Engineered SONOS
3. Gate All Around (GAA) Nanowire Technology
4. Junctionless Nanowire Channel

What it took was three large companies putting the device into production, standing behind it and continuing to engineer it. (similar to the introduction of the DRAM years before)

Vertical Channel NAND Flash - BICS

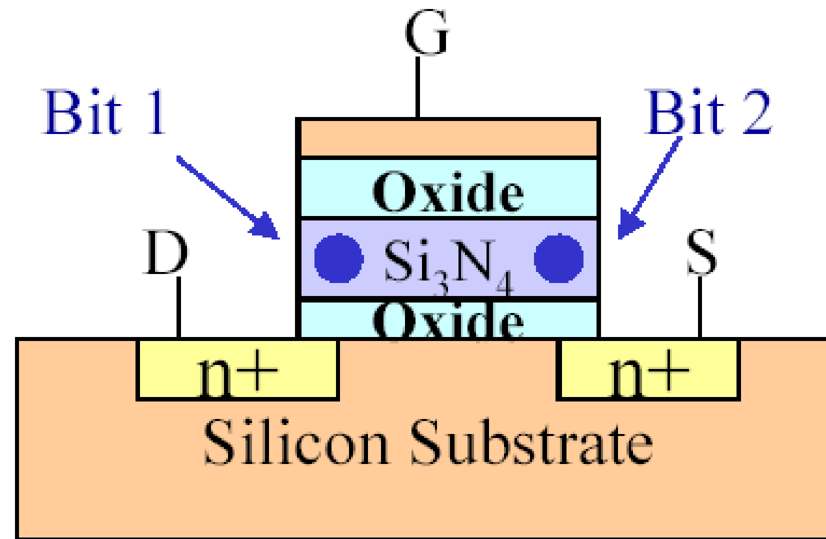
In the BiCS technology, an entire stack of electrode plates are punched through in one operation. A plug consisting of a string of charge trapping (SONOS) NAND flash transistors is built up in the via holes that are created.

This stacked CT memory array has a constant number of critical lithography steps regardless of the number of stacked layers used so the number of NAND transistors can be increased without increasing the number of process steps.



Toshiba,VLSI Symp. June 2007

Two Bit/Cell Charge Trapping Storage (SONOS)



-CHE Program / HH Erase

- Reverse Read

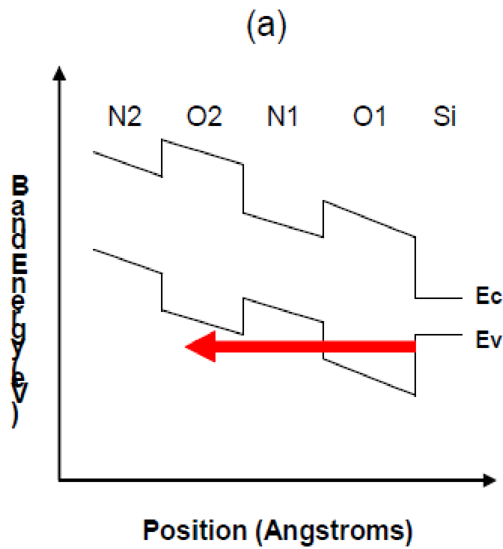
-Symmetric Operation

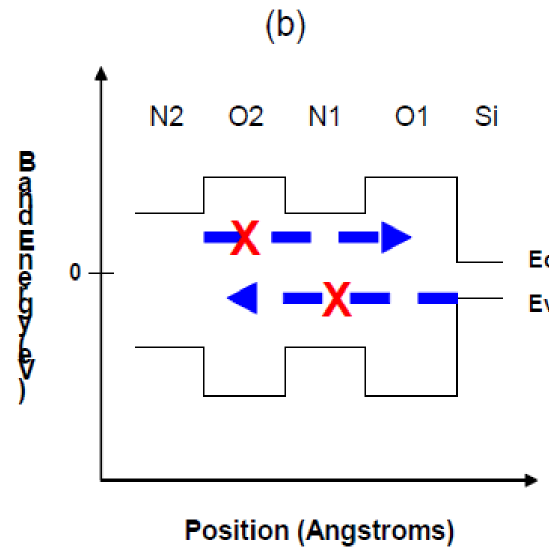
- Possible 2 bit storage

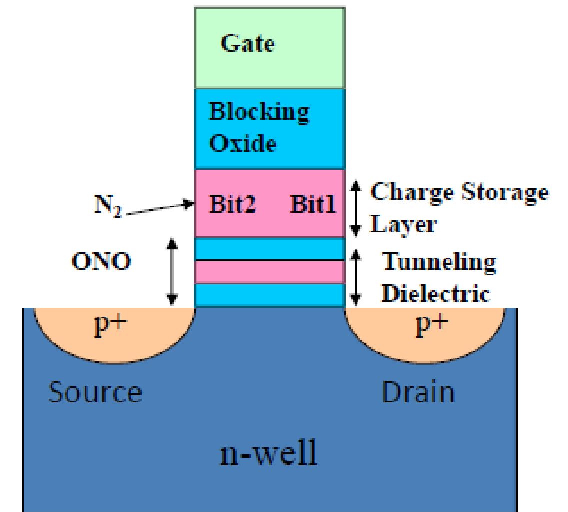
Early SONOS had data retention issues, through the tunnel oxide and back tunneling through the blocking oxide during erase.

Saifun/Spansion, 1999

TCAT Uses Barrier Engineered SONOS







(a)Band-Offset occurs at high electric field

(b) No Band-Offset in Data Retention Mode

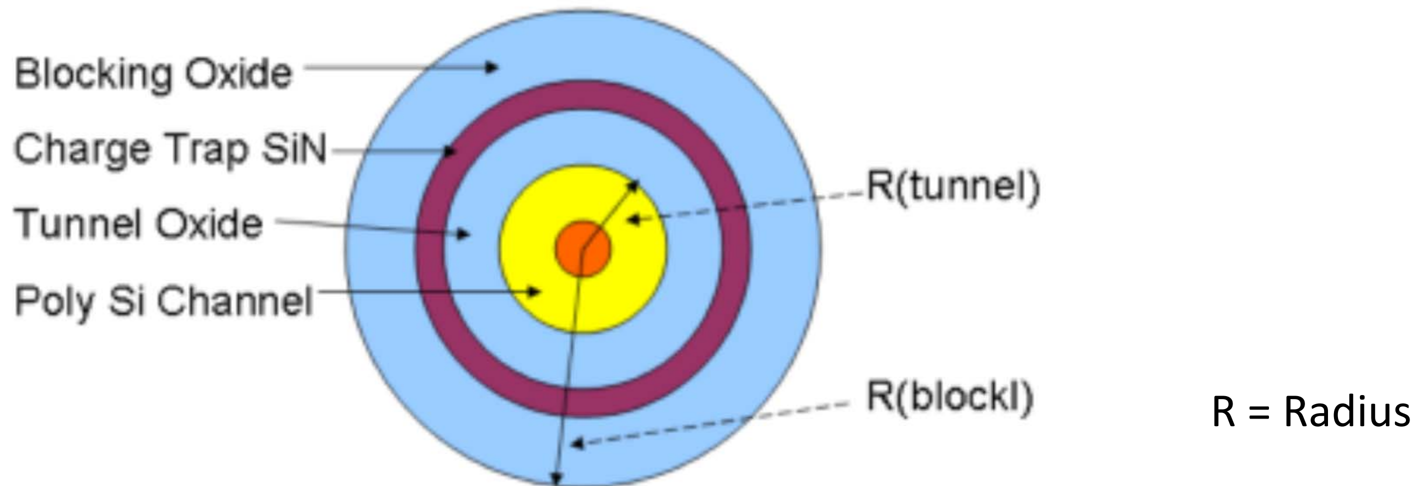
(c) BE-SONOS

With Barrier Engineered tunnel technology, at high electric field during erase, the band offset reduces the hole tunneling barrier so it is just that of the O1 layer. During data retention the full O1/N1/O2 barrier stack prohibits electron detrapping.

H.T. Lue, et al, (Macronix), IEEE IEDM, December 2005

3D Vertical Channel NAND Flash

GAA structure has faster erase performance with 1/R effect. (V across Blocking Oxide is 1/R of V across Tunnel Oxide which reduces the back- tunneling during erase)

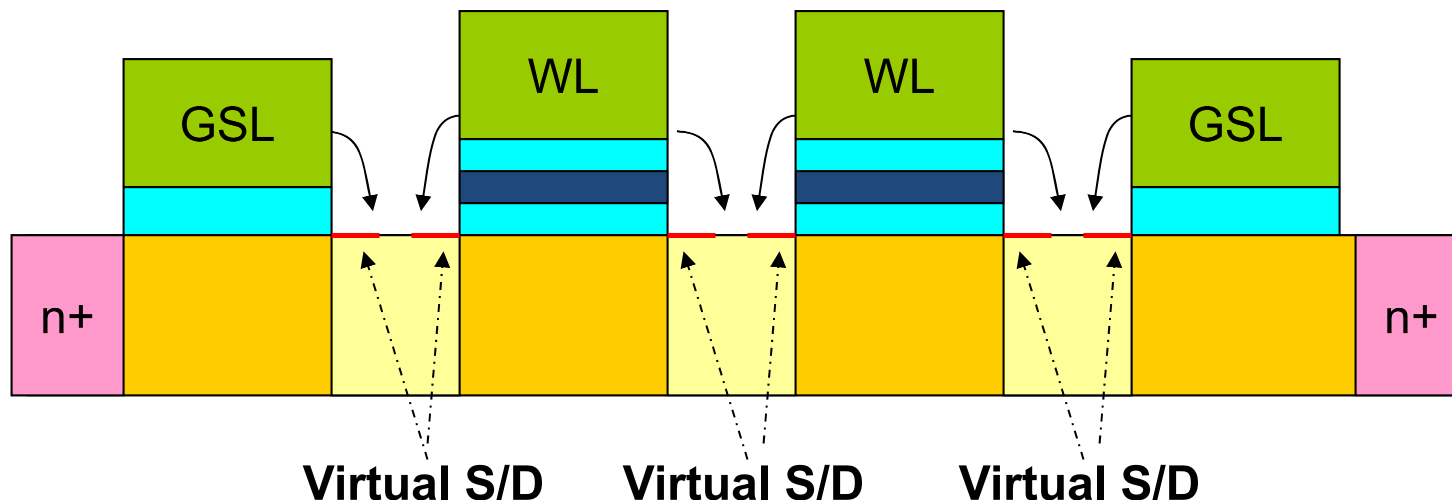


R = Radius

If the Tunnel Oxide is replaced with thin nitride ONO, then the Barrier Engineered SONOS improves data retention

K.T. Park, Samsung, ISSCC, Feb2014

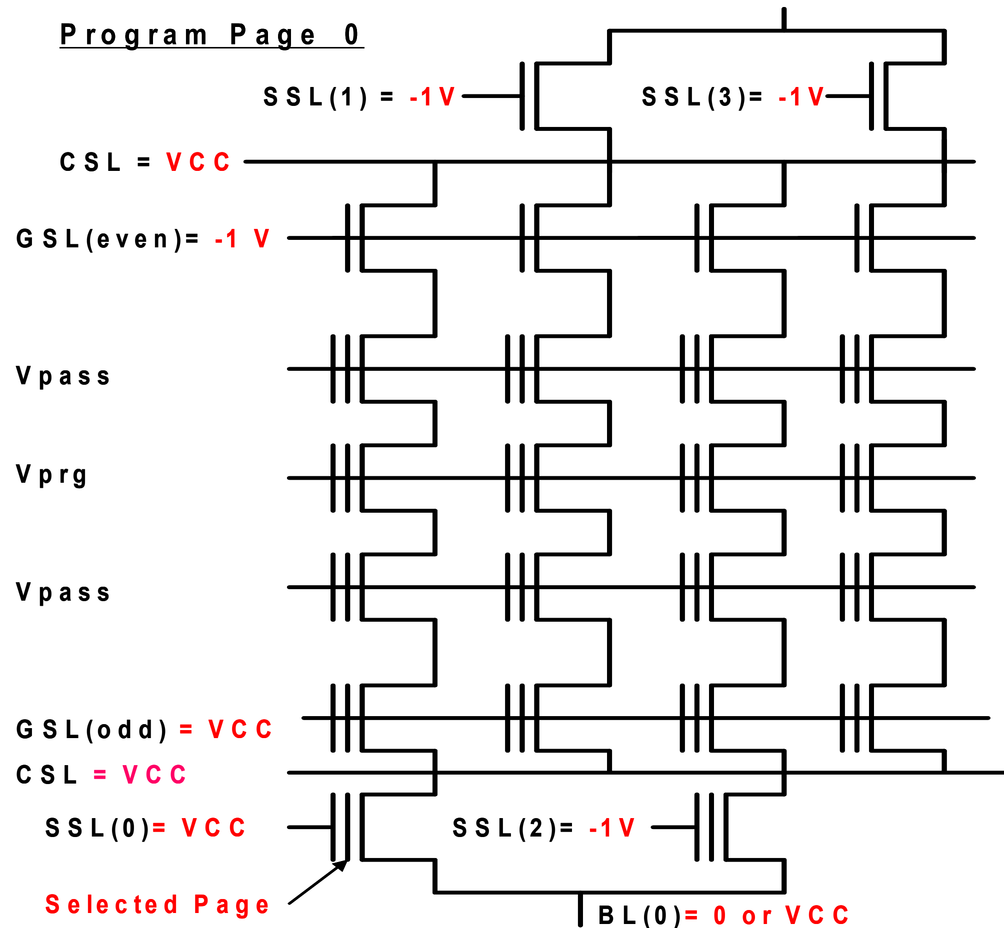
Junctionless Nanowire Channel



The field from the gates induces a virtual source and drain in the nanowire. This eliminates the need to diffuse a source drain in a 3D buried nanowire channel

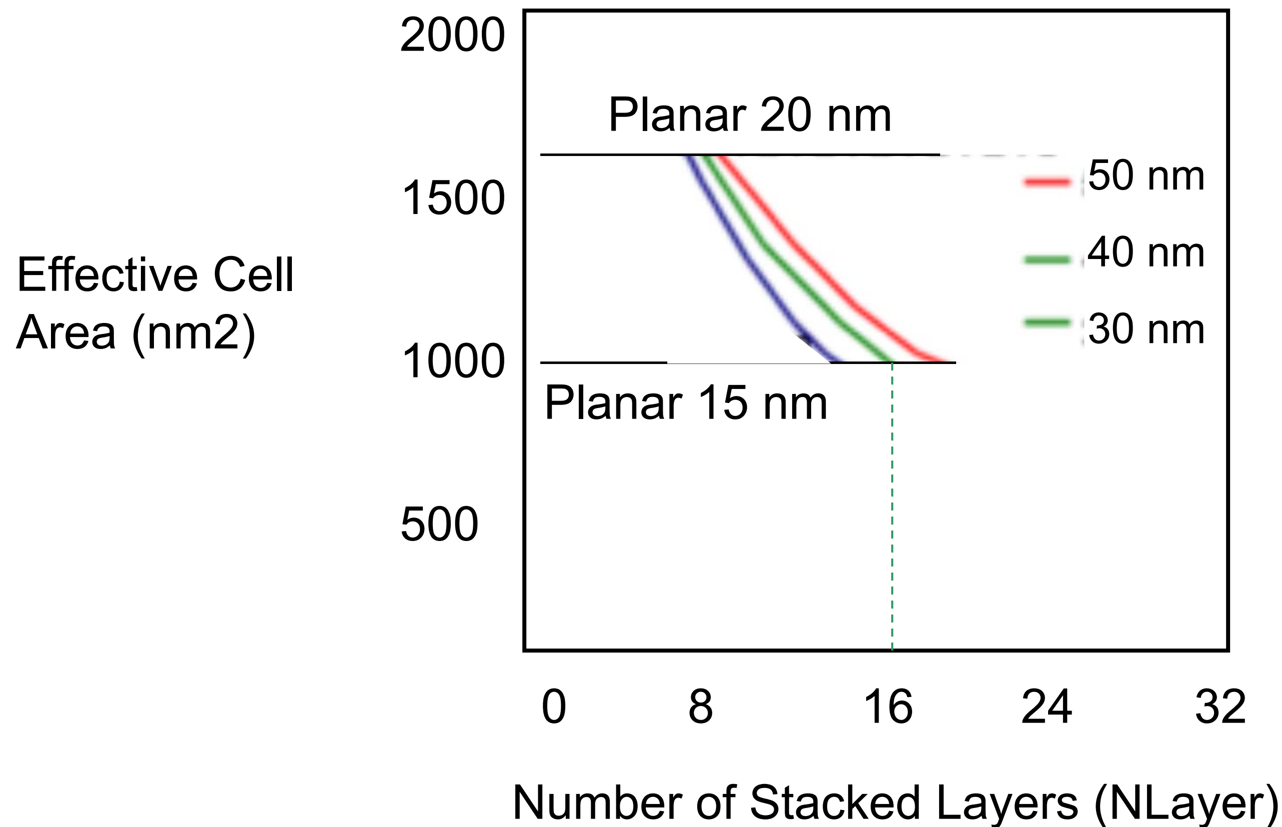
S.J. Choi, KAIST, VLSI 6/2011

One Page of Vertical NAND Flash



Macronix IEDM !2/2012

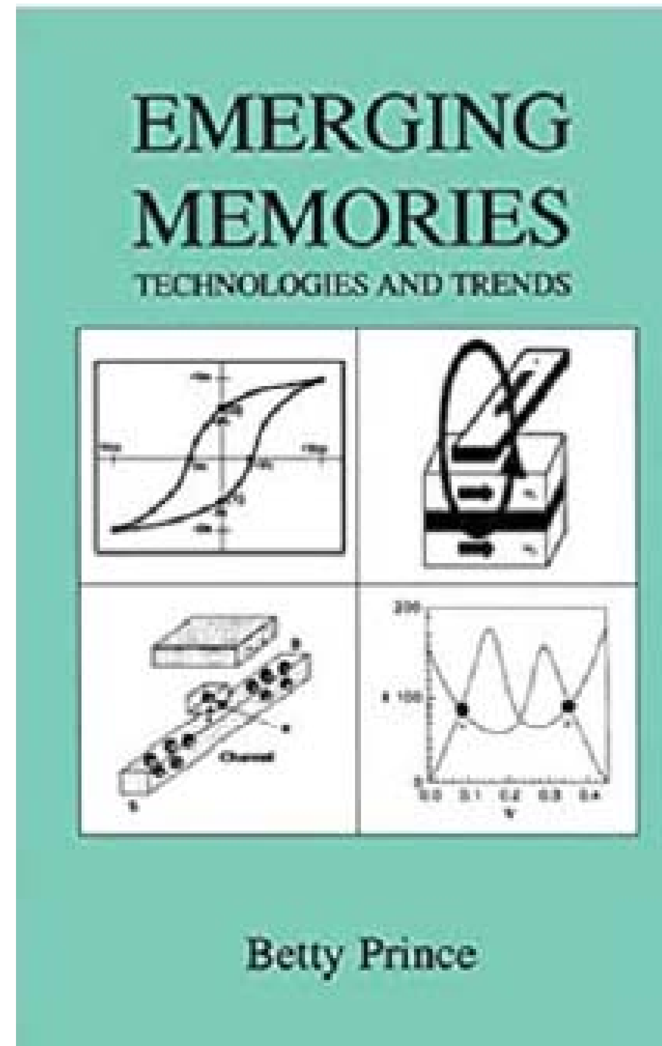
Effective Cell Area vs. Number of Stacked Layers



16 stacked layers in 30 nm technology same as planar 15 nm technology

(Based on Y. Yanagihara, et al, U. of Tokyo, IMW, May 20, 2013)

7. New Technology – Emerging Memories

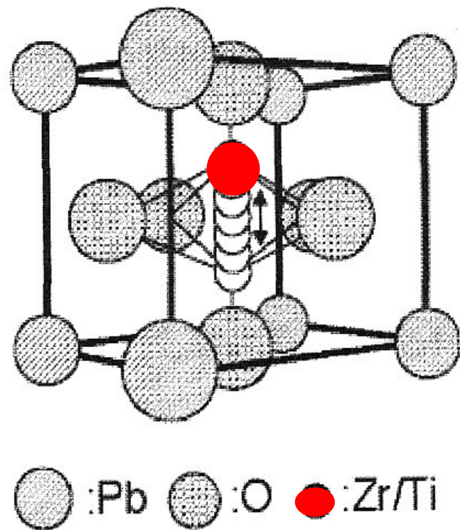


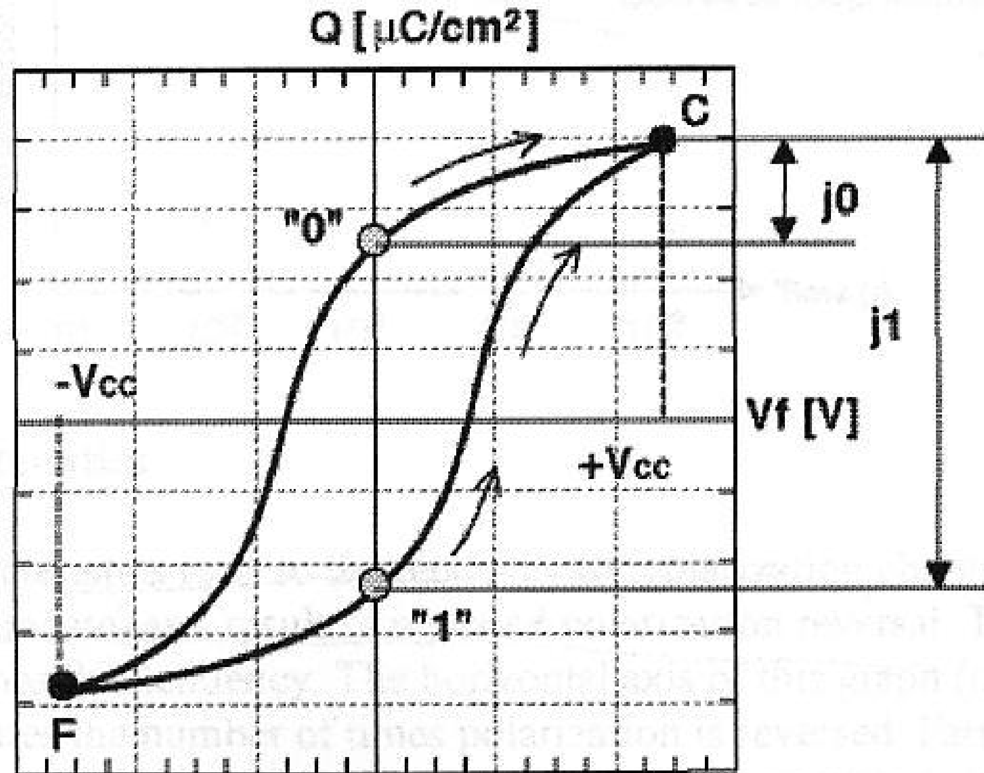
bprince@memorystrategies.com

New Technology – Emerging Memories

FeRAM	Still Emerging
MRAM	became new era embedded memory
PC-RAM	became new era high density memory
RRAM	Became the synapse in new era AI.

FeRAM Basics





Zr/Ti Ion displaced by external applied voltage gives rise to Hysteresis Curve.

Adapted from B. Prince, "Emerging Memories", 2002, Springer [17]

FeRAM

Advantage: Low Power Embedded Memory

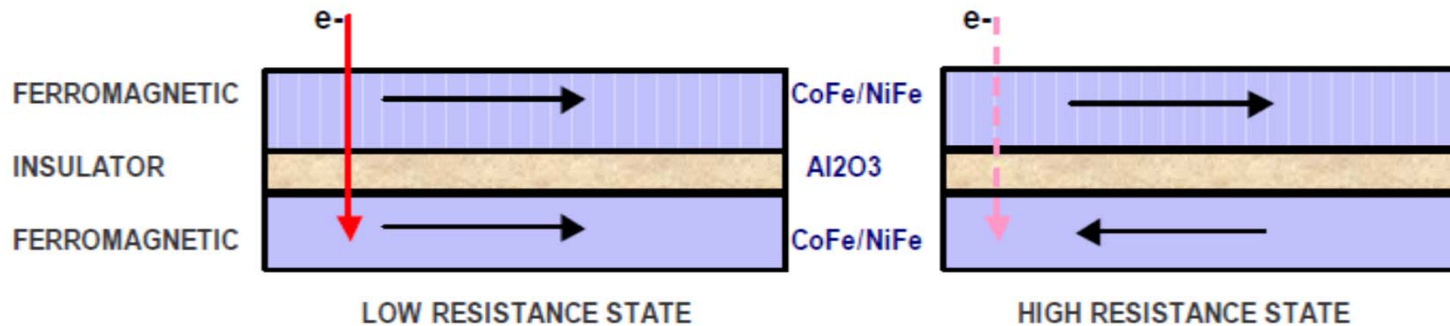
Disadvantages:

- Ferroelectric is contaminant in the Wafer Fab.
- Read is destructive and cell must be rewritten

FeRAM/FeFET has been “emerging” for 35 years.

The FeRAM Remains an Emerging Memory

Magnetic Tunnel Junction (MTJ) MRAM

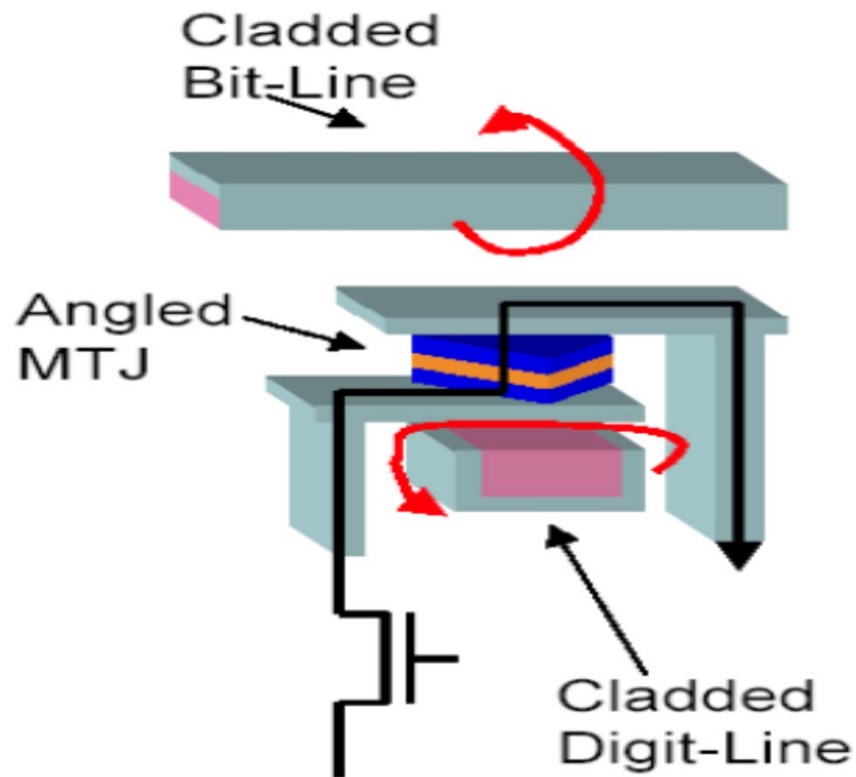


For the MTJ:
Current Tunnels through the Insulator

Advantages:
Signal Resistance change between states is 40% or more using a fixed FM layer and a free FM layer

Disadvantage:
If WRITE is by Field program with perpendicular external wires, The Write Current is very high about 4 mA/cell.

The MTJ Field Programmable MRAM



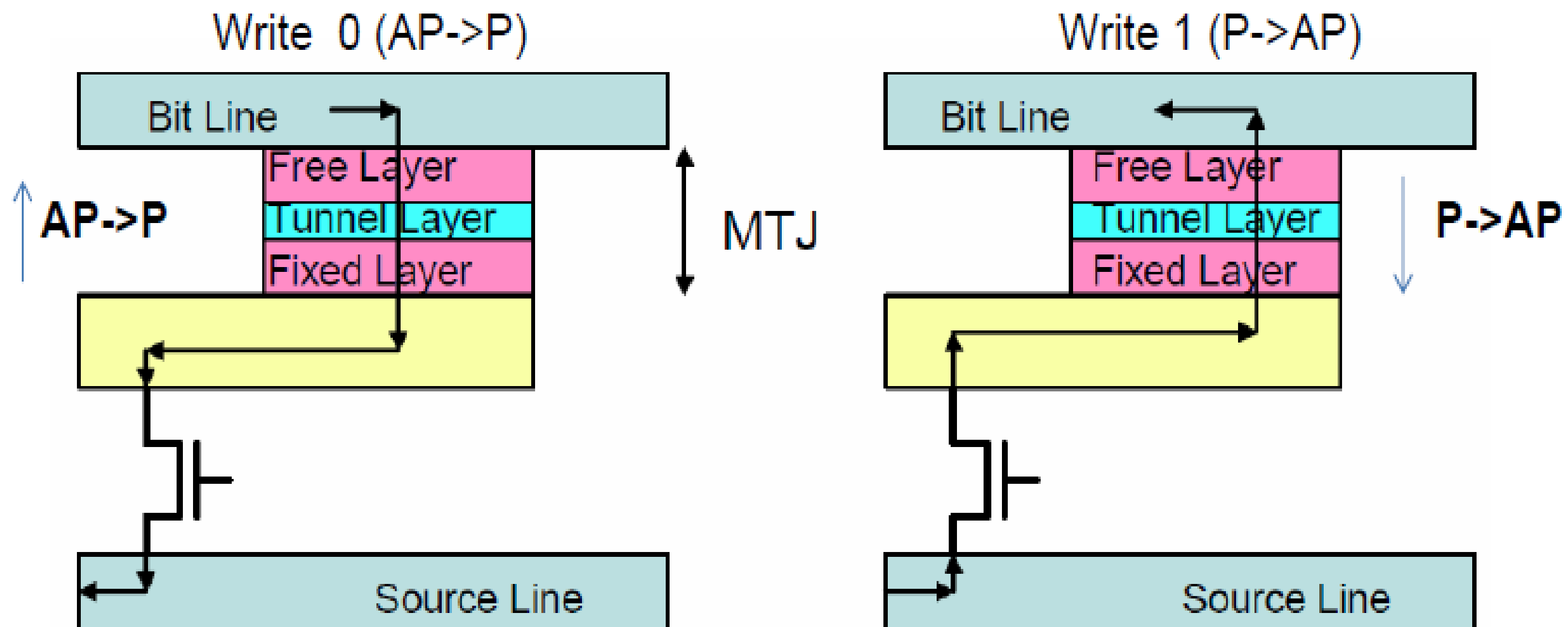
The free layer of the MTJ MRAM can be programmed by the magnetic fields of two currents running in perpendicular copper cladded external wires near the MTJ. The current per cell for field programming is about 4 mA per cell.

J.M. Slaughter, Freescale, IEDM, 2005

Spin Torque Transfer (STT) MRAM

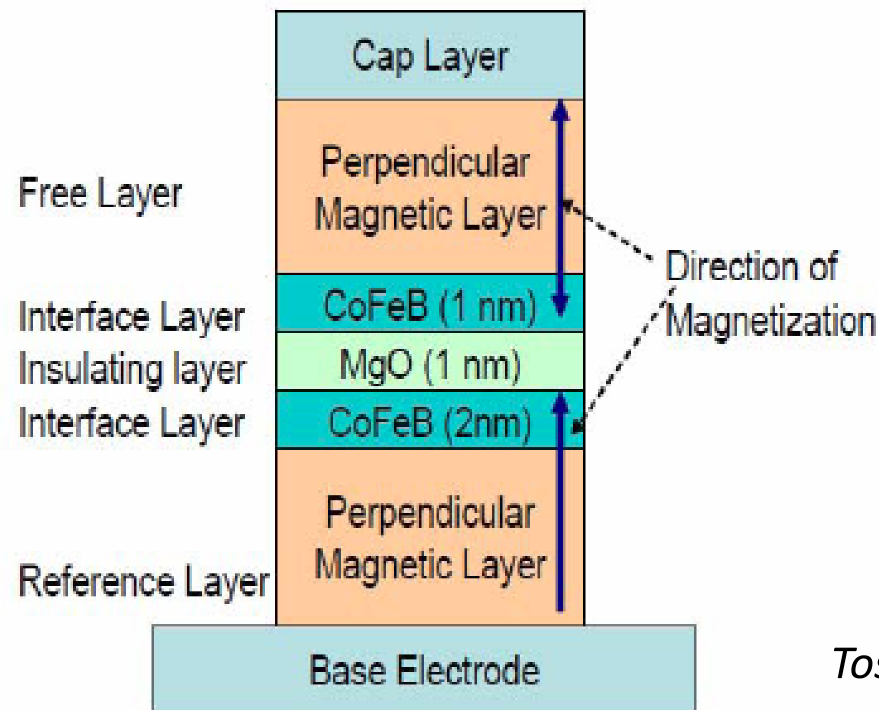
In 2005, Sony used the direction of current running through the MRAM to switch its polarization. This Spin Torque Transfer (STT) switching cell used spin-polarized electrons to flip the polarization of the cell. Power was reduced to less than 100 uA for a 45 nm technology.

Sony, IEDM, 2005



Source: Grandis/Samsung, 2007

Perpendicular (P)-STT MRAM Cell



Toshiba, VLSI-T June 2008

The p STT-MRAM has even lower power and can be used in L3 cache memory or as a microcontroller universal memory for Internet of Things (IoT) applications such as medical, industrial, mobile processor and automotive.
MRAM Foundries: TSMC, Global Foundries, Samsung, UMC

28 nm P-STT MRAM Macro for eMCU Applications

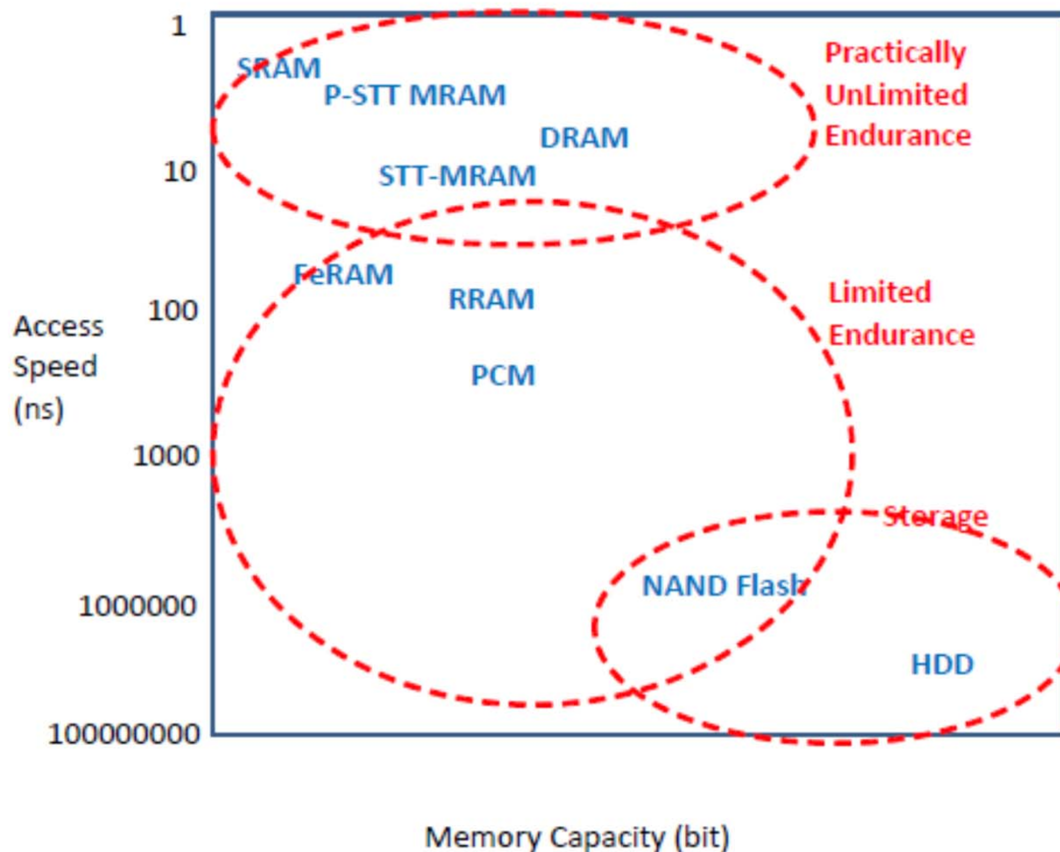
The MRAM Macro replaces both the NOR Flash and the SRAM in the eMCU

28 nm CMOS Logic
Cell Size – 0.364 um2
P-MTJ (MgOFeB)
Diameter: 38-45 nm
Clock Frequency–40 Mhz
V Core/IO=1.0V/1.8V
I/O – 32/64
Endurance 10^8 cycles 1.8V

Adapted from Samsung, IEDM, Dec. 2016

Using p-STT-eMRAM to Replace eSRAM Cache

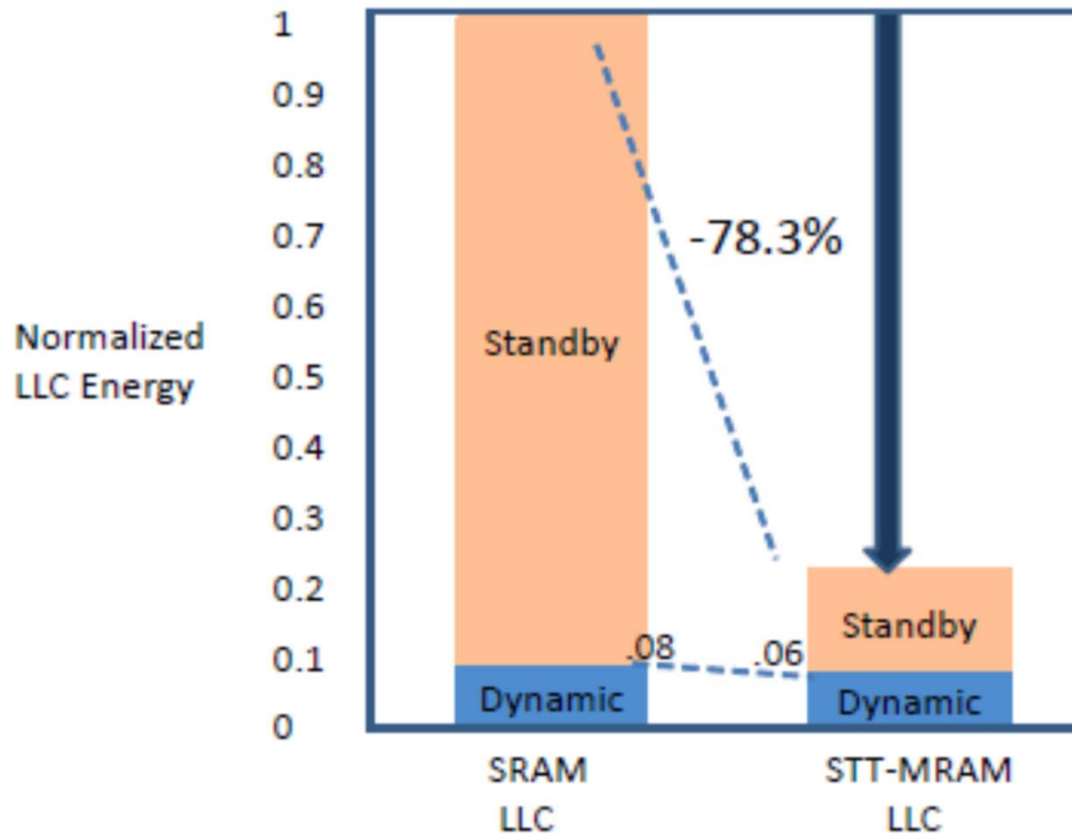
The P-STT MRAM has a comparable speed to the SRAM



Adapted from Fujita, IMW, May 2015

LLC Energy for eSRAM and eSTT-MRAM

The eSTT-MRAM LCC uses 78.3% less energy than the SRAM LLC



Adapted from Toshiba, VLSI-DAT April 2017

PCM 3D Xpoint Memory (Intel)

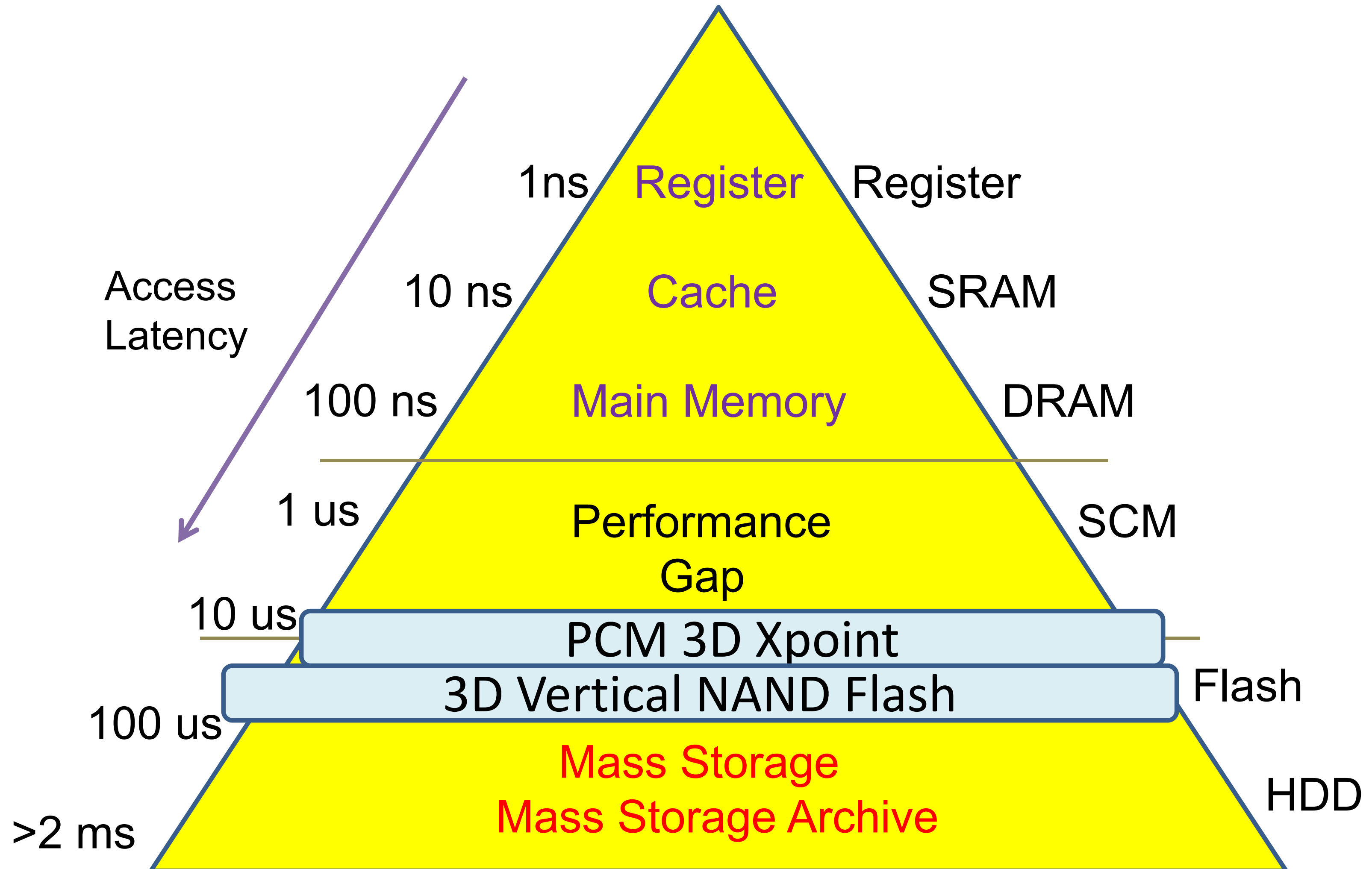
This Storage Class Memory (SCM) decreases Latency for storage and improves performance for Cloud applications

It is best when run with Vertical NAND SSD Storage which reduces access latency from the mass storage and in the Cache hierarchy is just below the PCM 3D XPoint in the SCM gap.

It is Non-Volatile Memory so it does not need to be reloaded.

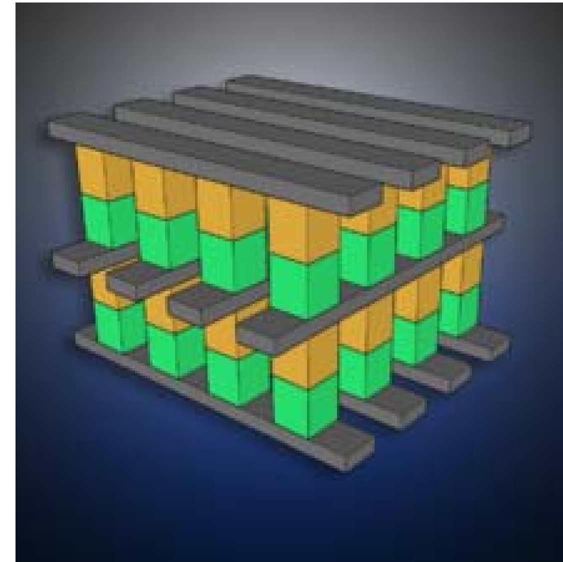
It is a Vertical 3D Cross-point Phase Change Memory

Cache Memory Hierarchy



Background of Phase Change Memory

Phase Change Memory (PCM) or Chalcogenide RAM has been around since the 1960's. Intel and IBM have both investigated PCM.

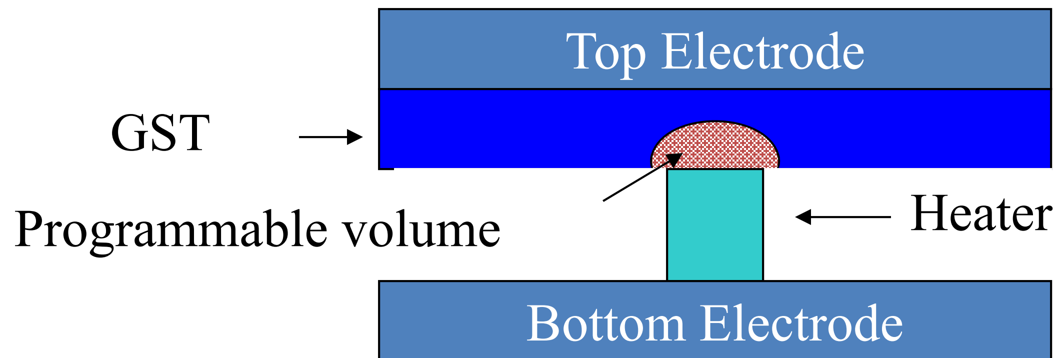


This non-volatile RAM uses the property of Chalcogenide glass to be In either an amorphous (high resistance) or a crystalline (low resistance) state.

Development of the Intel high density 3D Xpoint began around 2012. Both the selector and the storage part of the memory are made from chalcogenide materials

Simple PC-Memory Cell

Schematic Cross-Section



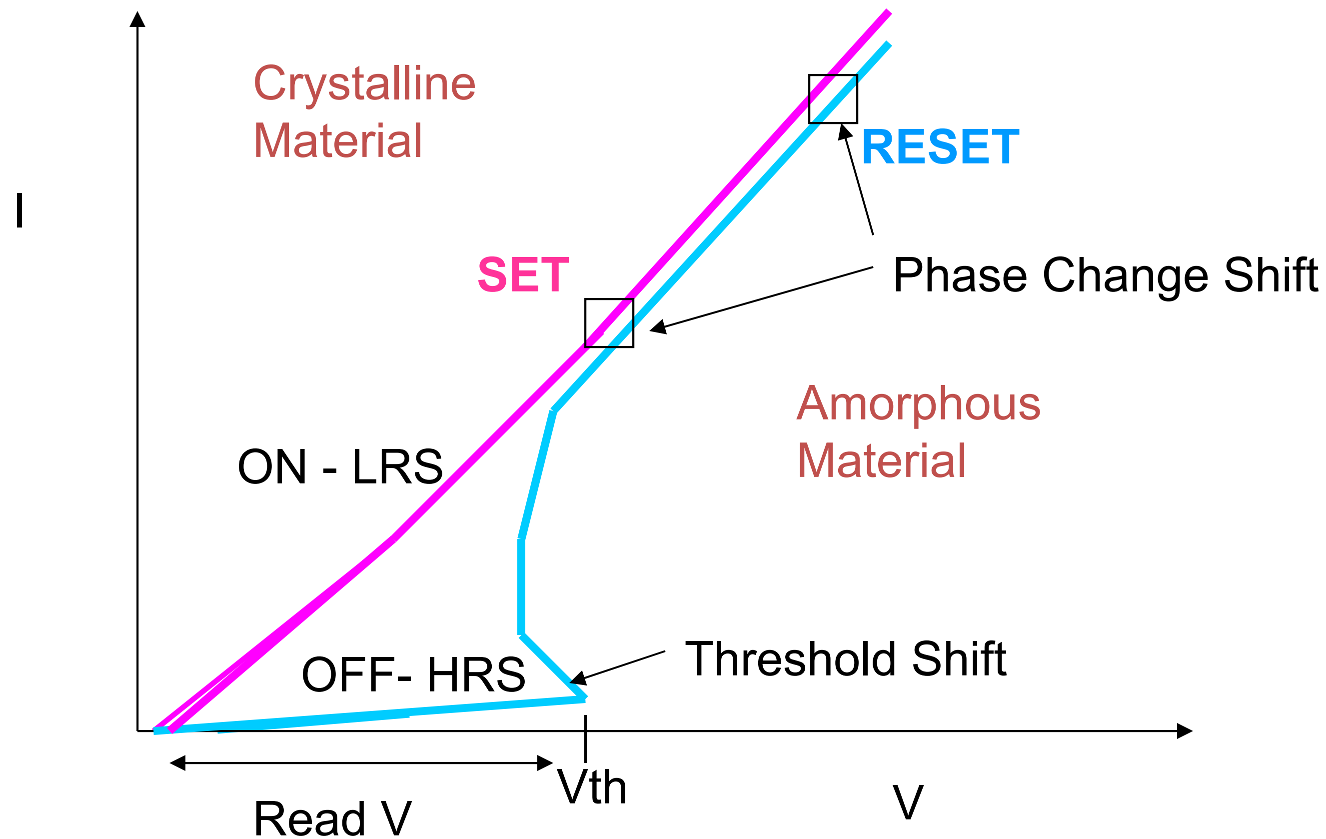
Data storage is by a thermally induced phase change between amorphous and polycrystalline states in a thin film chalcogenide alloy.

*GST = Ge2Sb2Te5 (Germanium-Antimony-Telurium

Ovonyx, Intel, ISSCC, 2002

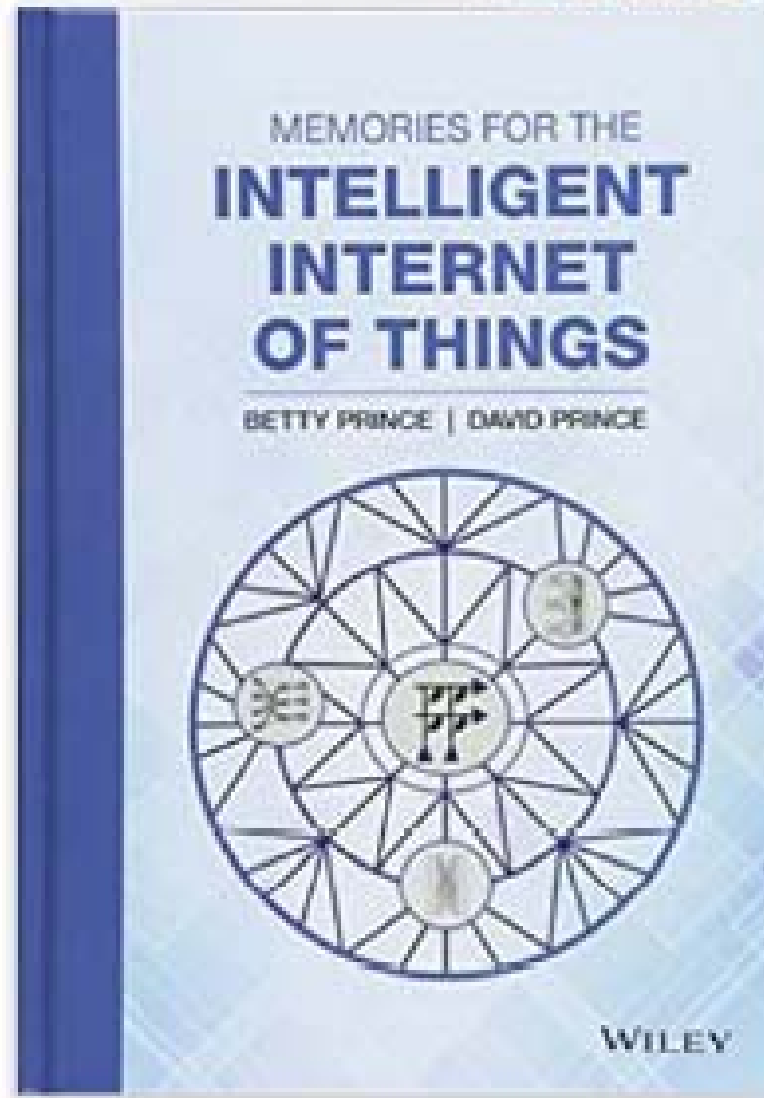
bprince@memorystrategies.com

Phase Change Memory I-V Curve



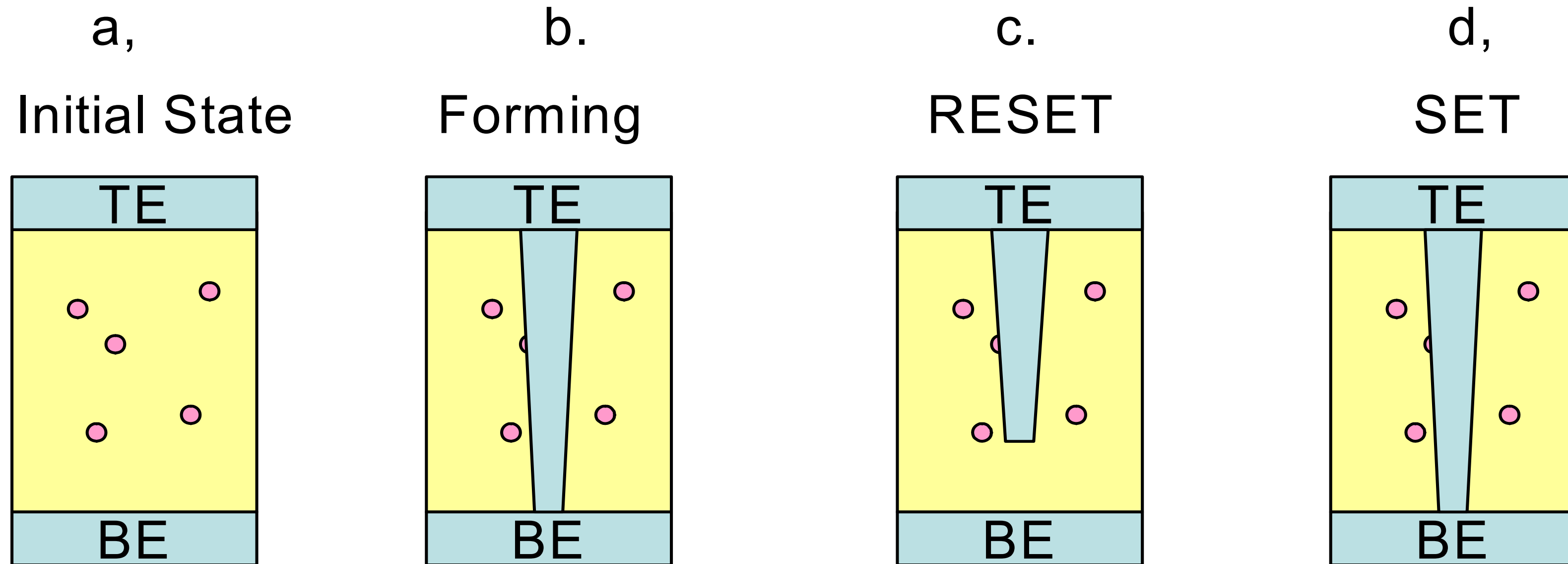
Ovonyx, Intel, ISSCC, 2002

8 Neural Synapses, Neuromorphic Memory and AI



Resistive RAM (ReRAM) Devices

Formation of the Conductive Filament



(a) starting defects in the dielectric enable filament formation.

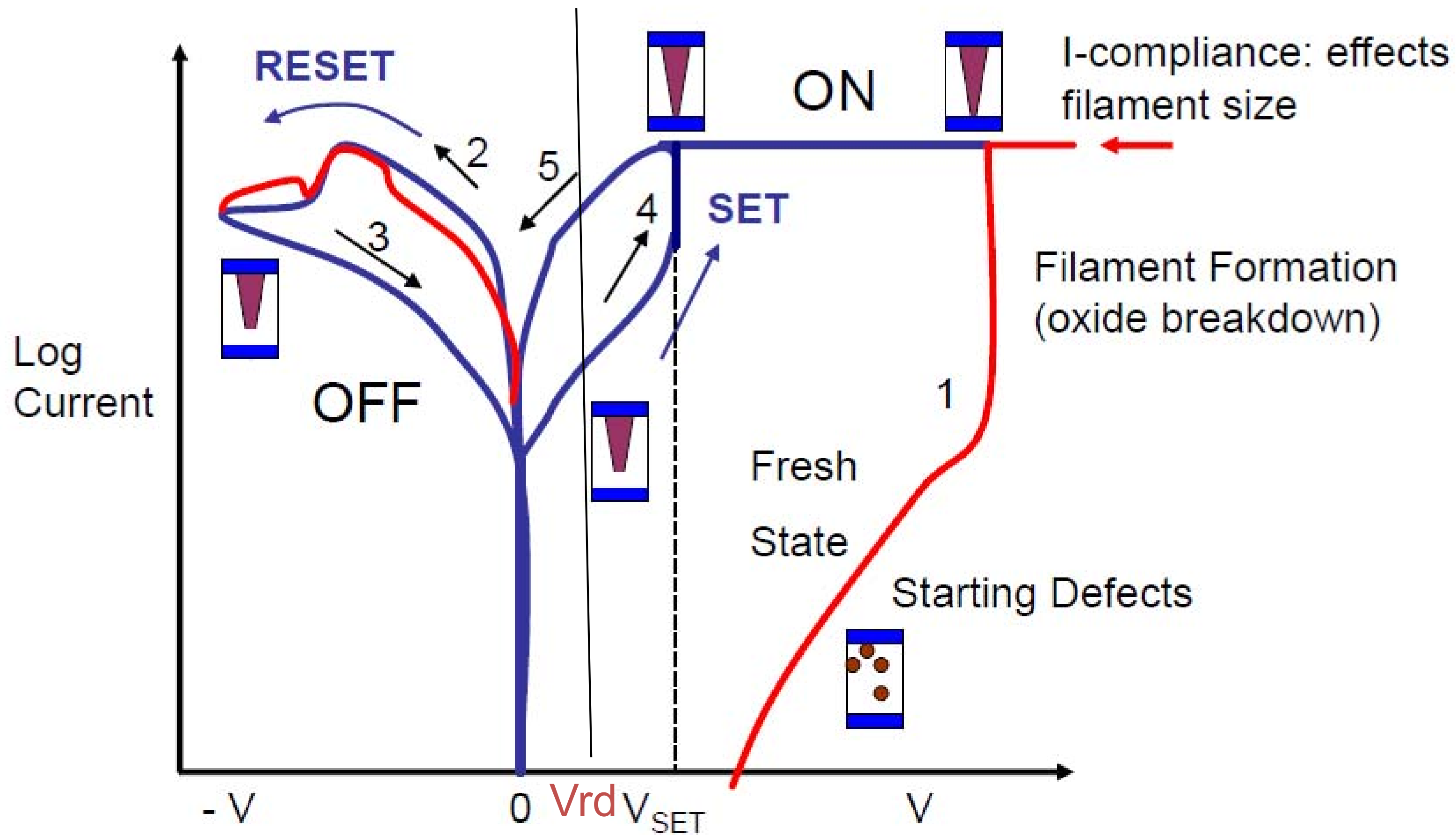
(b) voltage is applied across the element and a conductive filament forms with diameter determined by the compliance current limit.

(c) during RESET a gap forms in the filament. A HRS/OFF state results.

(d) during SET the filament reforms across the gap with a LRS/ON state

D.C. Gilmer, SEMATECH, Stanford, CNSE, U. of Albany, IMW, May 2012

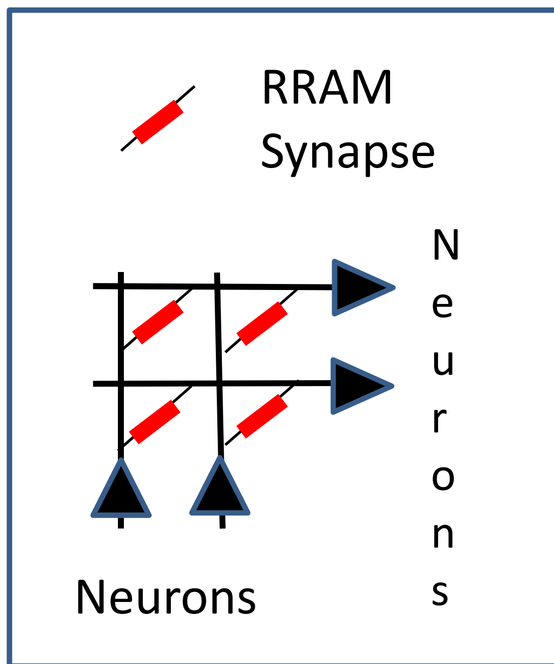
Typical Bipolar ReRAM Device IV Curve



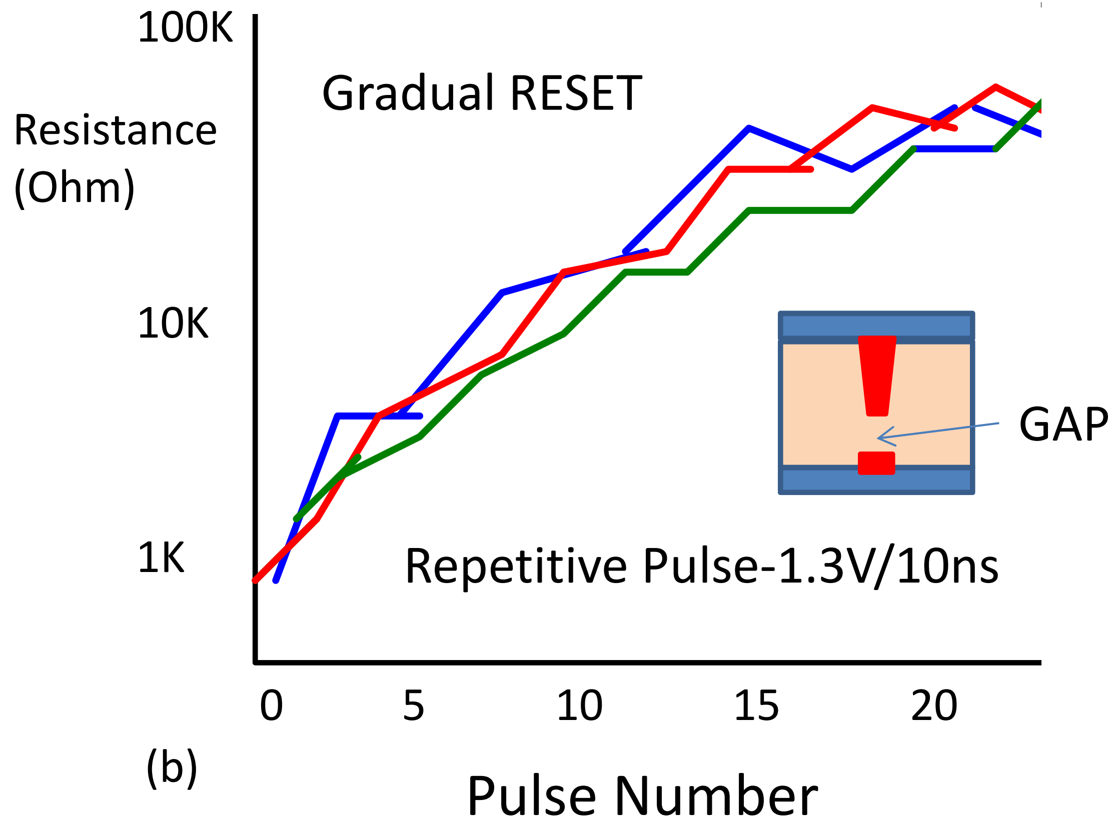
Bipolar ReRAM SETs at one voltage polarity and RESETs at the opposite polarity. The voltage can be changed for analog.

D.C. Gilmer, SEMATECH, Stanford, CNSE, U. of Albany, IMW, May 2012

An ReRAM Can be Used as an Analog Synapse



(a)

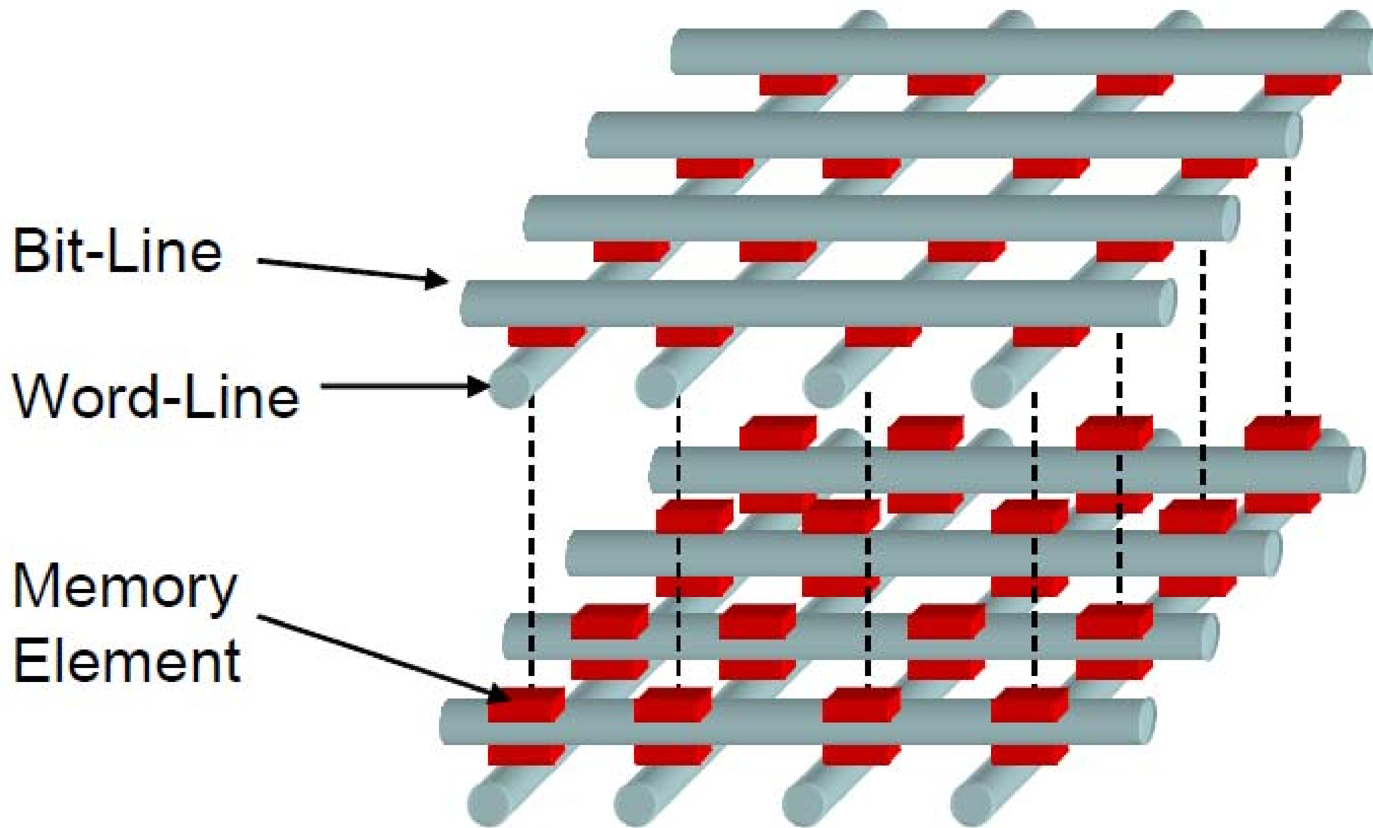


(b)

(a)Hybrid neuromorphic system with CMOS neurons & RRAM synapses. (b) During RESET a CF is ruptured and a variable tunnel gap forms between electrode and residual filament. Variation in tunnel gap results in multilevel resistance states.

Based on S. Yu, Arizona State, ISCAS, June 2014

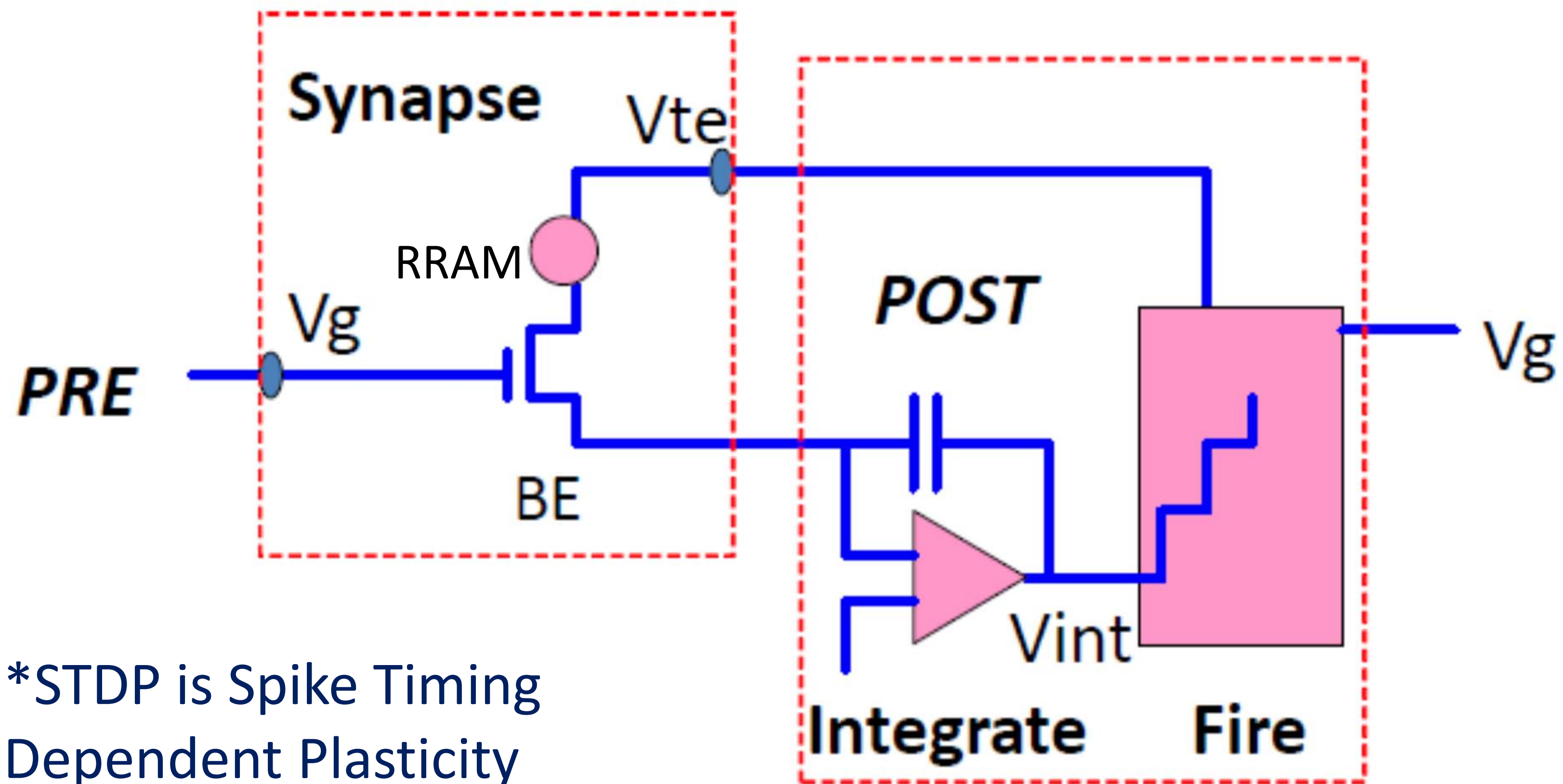
High Density Cross-Point Array for Neural Aps.



Theoretically this stack can be very high and the wire grid very small to store a high density of data.

B. Prince, 3D Vertical Memory Technology, (Wiley 2014)

PRE & POST Connections permit overlaps between spikes enabling STDP* with 1T1R Synapse



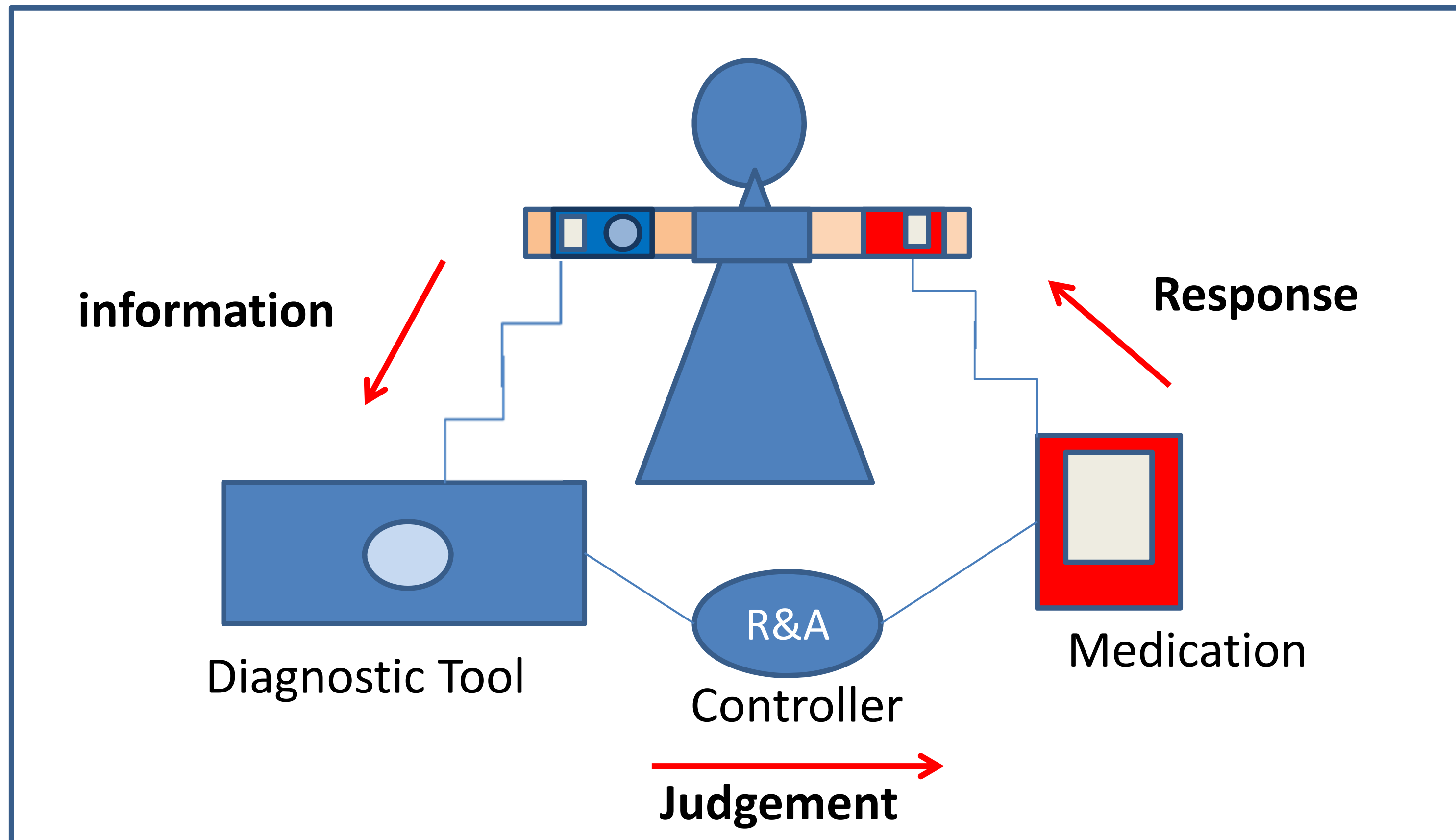
*STDP is Spike Timing Dependent Plasticity

Based on D. Ielmini, P. Di Milano, IUNET, ISCAS, May 2016

B. Prince, D. Prince, *Memories for the Intelligent Internet of Things*, 2018, (Wiley)

bprince@memorystrategies.com

Intelligent Personal Medical Controllers



*32b RISC MCU, coin battery

4KB eEEPROM, 12KB eSRAM
256KB eFlash

or **1MB eMRAM**

Neuromorphic
RRAM or PCM

Embedded Flash
Or
eMRAM

Adapted from "Memories for Intelligent IoT" B. Prince & D. Prince 2018 (Wiley)