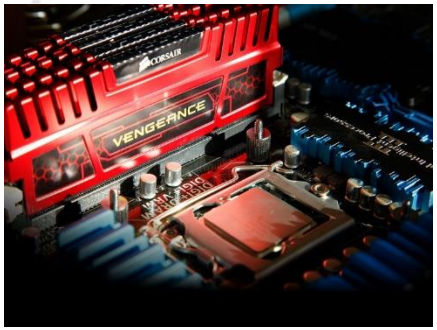
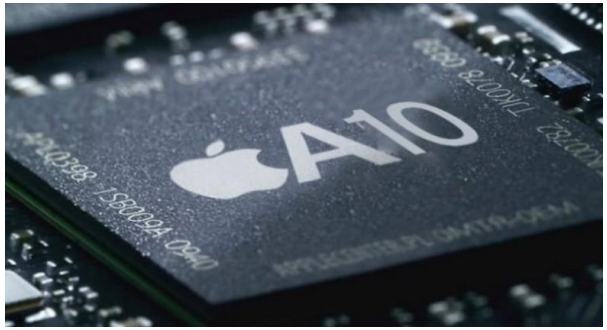
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

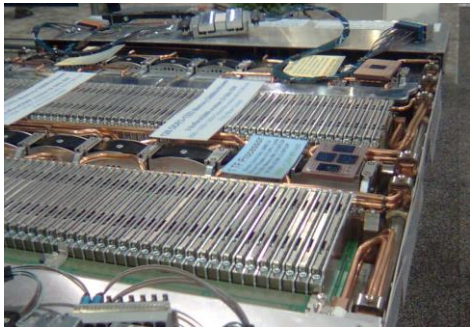
Efficient Memory Controller Architectures and their Integration in Transprecision Compute Platforms

Christian Weis

weis@eit.uni-kl.de







TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Content

- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Modes
- The Memory Controller
- Conclusion

Content

- **The Transprecision Concept**
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Modes
- The Memory Controller

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

The Transprecision Computing Paradigm

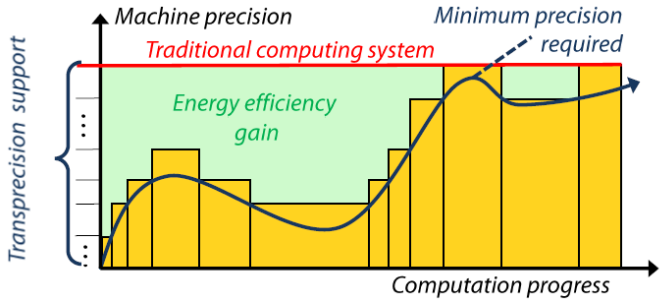
Traditional computing systems and applications

- Conservative assumption: each calculation must be accurate
- Hardware must 100% accurate
- Computation is more and more limited by power wall

⇒ Shift from traditional computing paradigms becomes mandatory

Transprecision compute paradigm

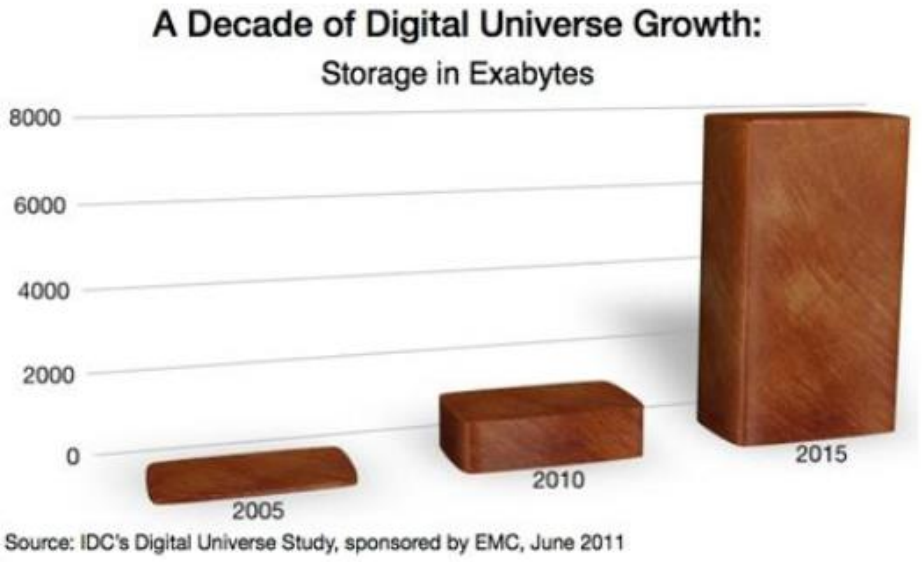
- **Spend just the right amount of energy** required for a particular operation
- **Just enough precision**, i.e. dynamic adoption to less precise arithmetic and even embrace errors incurred by underpowered or new circuit technologies
- **Off-load computation** to low power HW units



Focus in this talk on Memory (DRAMs)

Flooding of Data

Our world is creating more bits than ever: in 2011, more than 1.8 zettabytes (1.8×10^{21}) , growing by a factor of 9 in 5 years



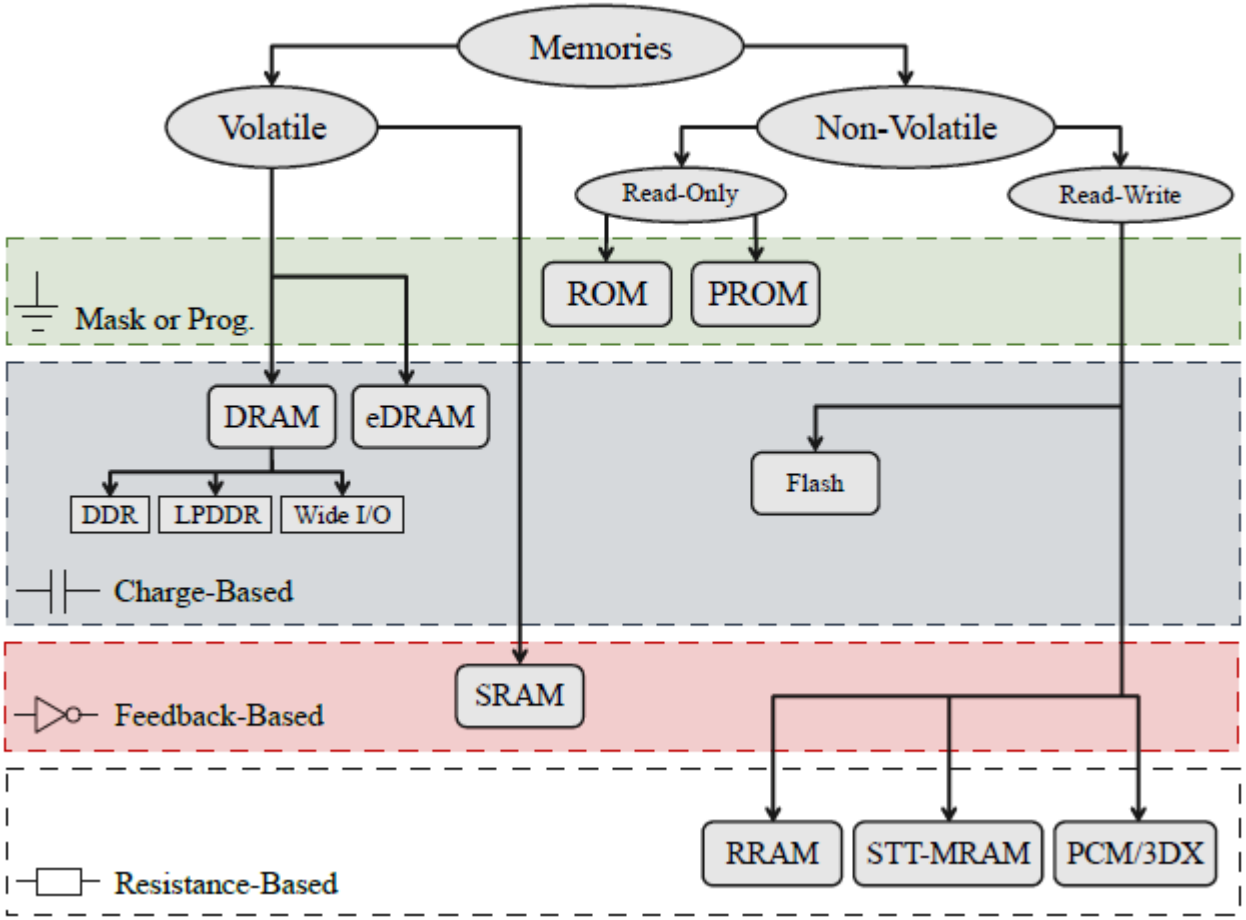
New computing paradigms ⇒ near-memory, in-memory computing

Content

- The Transprecision Concept
- **Overview on Memory Types**
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Modes
- The Memory Controller

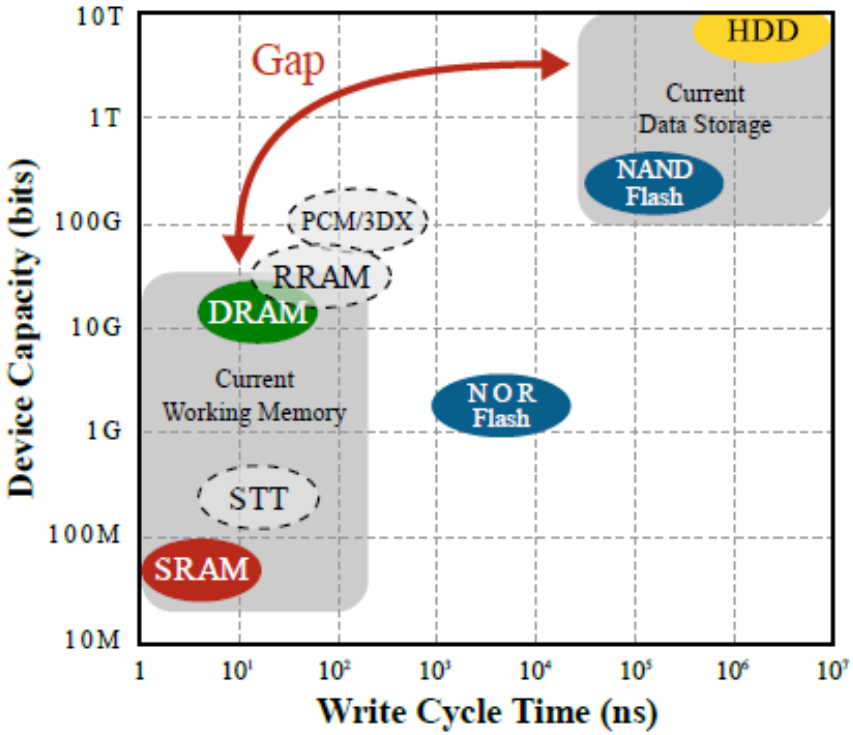
Memory Technologies

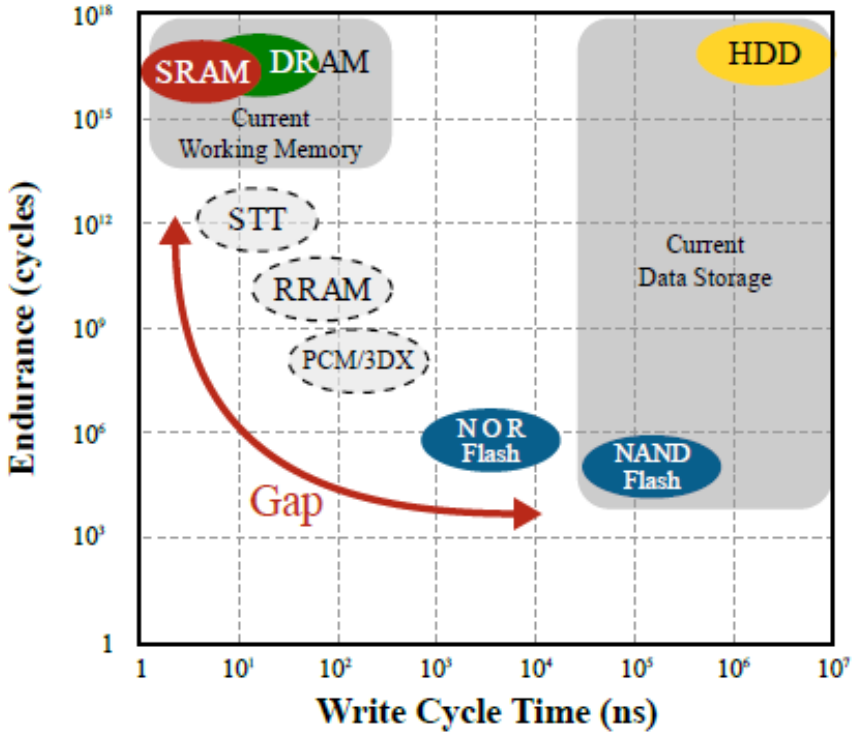
- Different memory types
- Different storage mechanisms



Memory Properties

- Trade-offs: capacity, latency, reliability

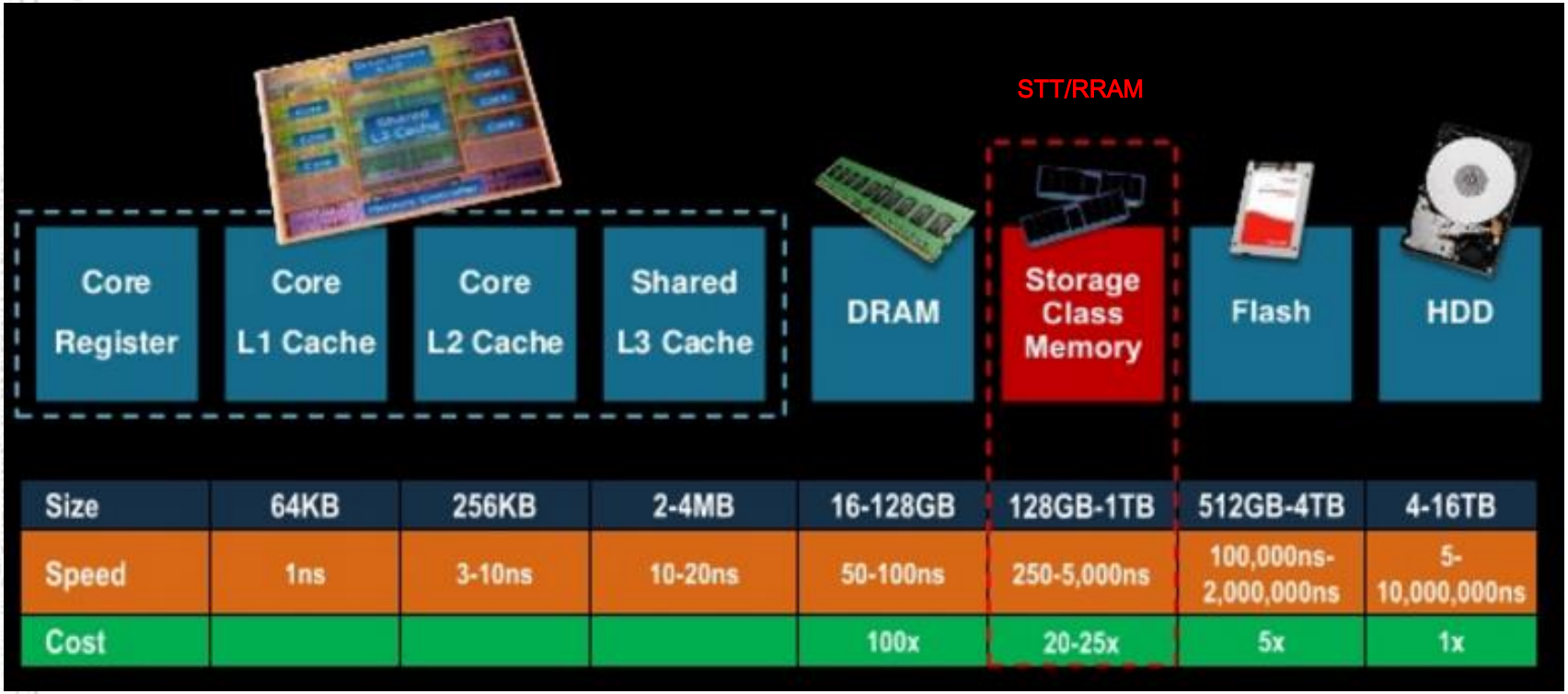




MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Memory Hierarchy in Computing

- Capacity, latency, cost



Source: Western Digital

Content

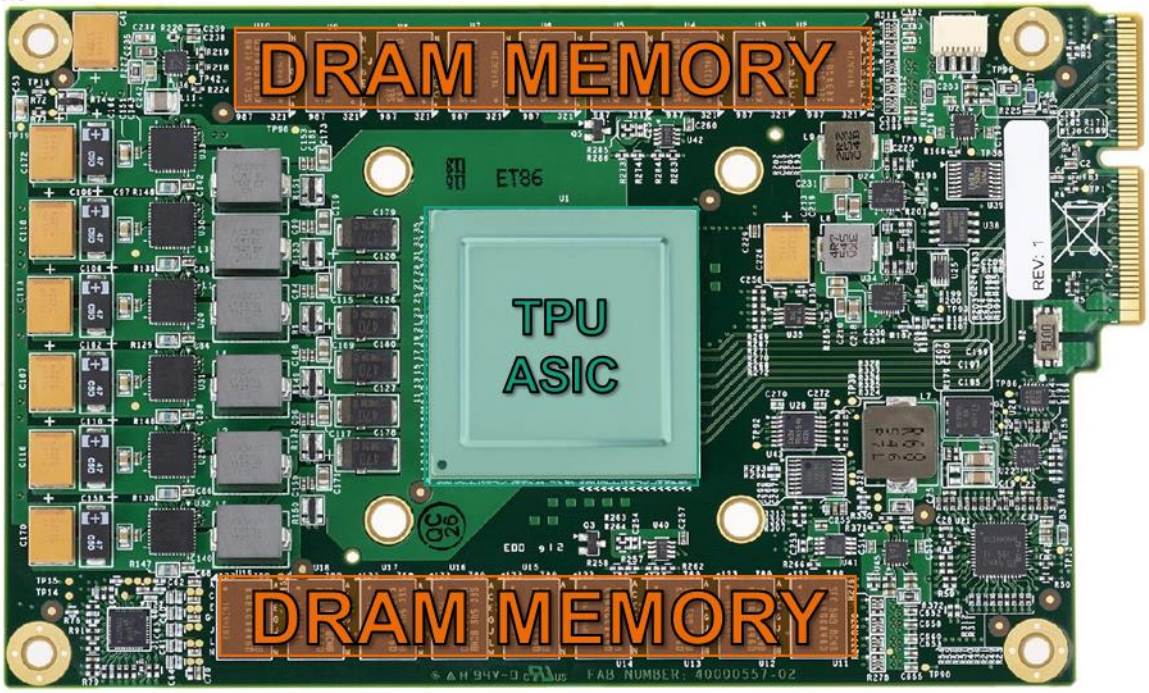
- The Transprecision Concept
- Overview on Memory Types
- **The Importance/Challenges of Memories in Computing**
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Modes
- The Memory Controller

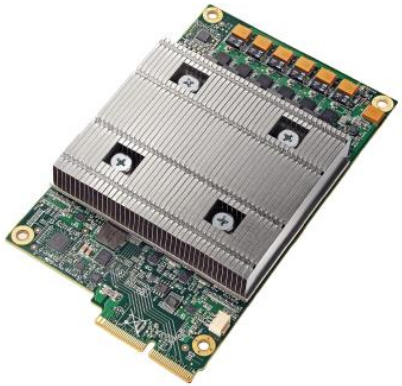
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

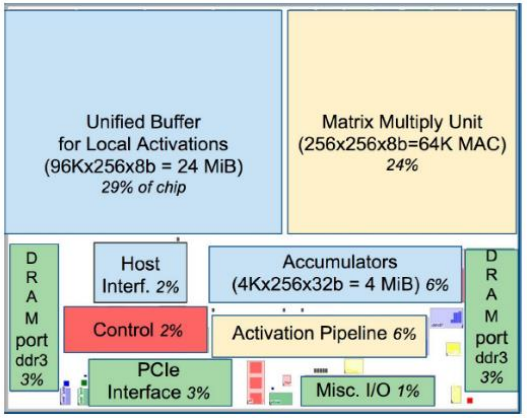
Importance Memory e.g. Google Tensor ASIC

ASIC in use in data centers since 2015

- 65.536 8-bit MAC matrix multiply units -> 92 TeraOps/second

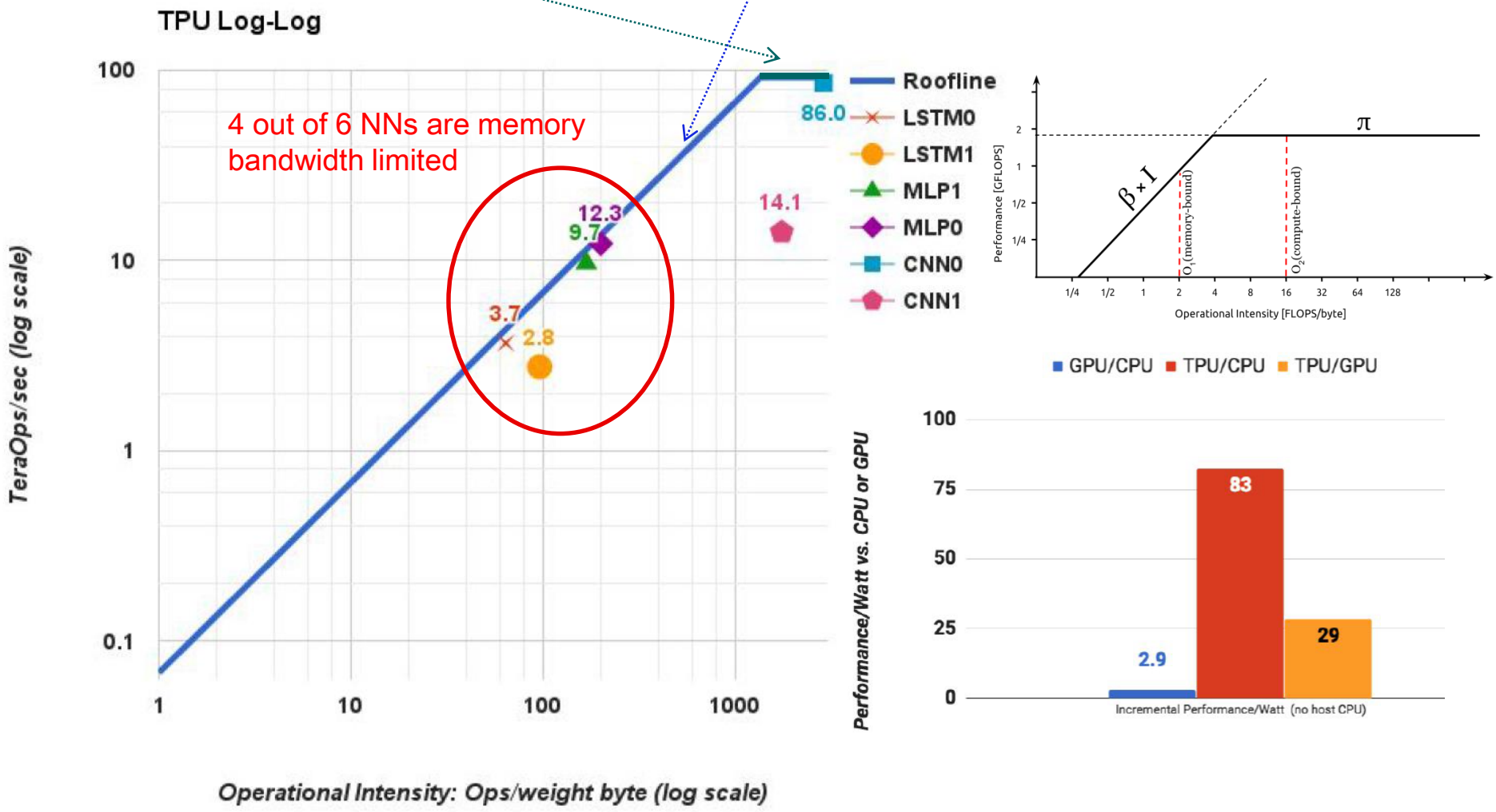






Roof Line Model Googles Tensor ASIC

Flop/s = min { peak Flop/s, DRAM_bandwidth * (Flop/s) / DRAM_byte_access }



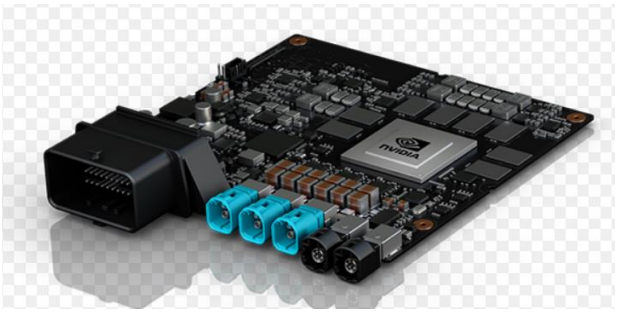
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Importance Memory e.g. Autonomous Driving

- Bandwidth ADAS autonomous level 4 & 5 : 400 – 1024 GB/s
- Memory Capacity autonomous level 4 & 5 : 18 – 24 GB
- Reliability: temperature 105°C – 125°C
- Mixed criticality system: guaranteed latency
- Security e.g. row hammer

DRAM	Peak Bandwidth	Low Energy	Reliability	Automotive Grade (Temp.)
LPDDR4	o	++	+	yes
GDDR5X/6	+	o	+	yes
HBM	++	+	o	no

marginal (o), medium (+), high (++)



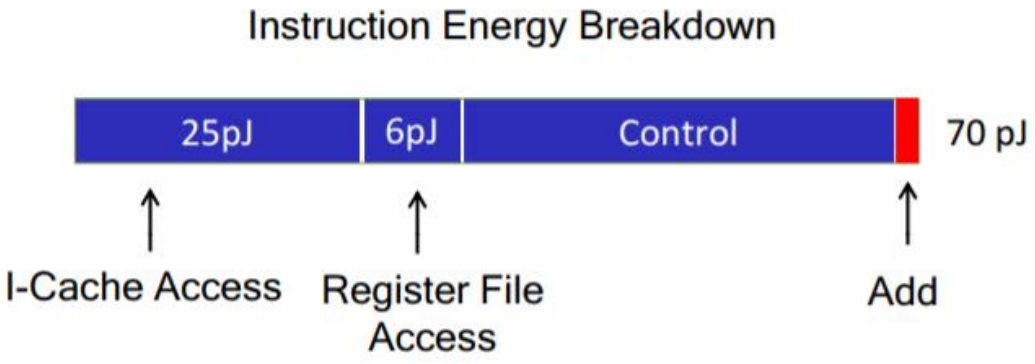
Driving Into the Memory Wall: The Role of Memory for Advanced Driver Assistance Systems and Autonomous Driving
M. Jung, S. A. McKee, C. Sudarshan, C. Dropmann, C. Weis, N. Wehn. *ACM International Symposium on Memory Systems (MEMSYS 2018)*, October, 2018, Washington, DC, USA.

Energy Cost Computing @ 45nm

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

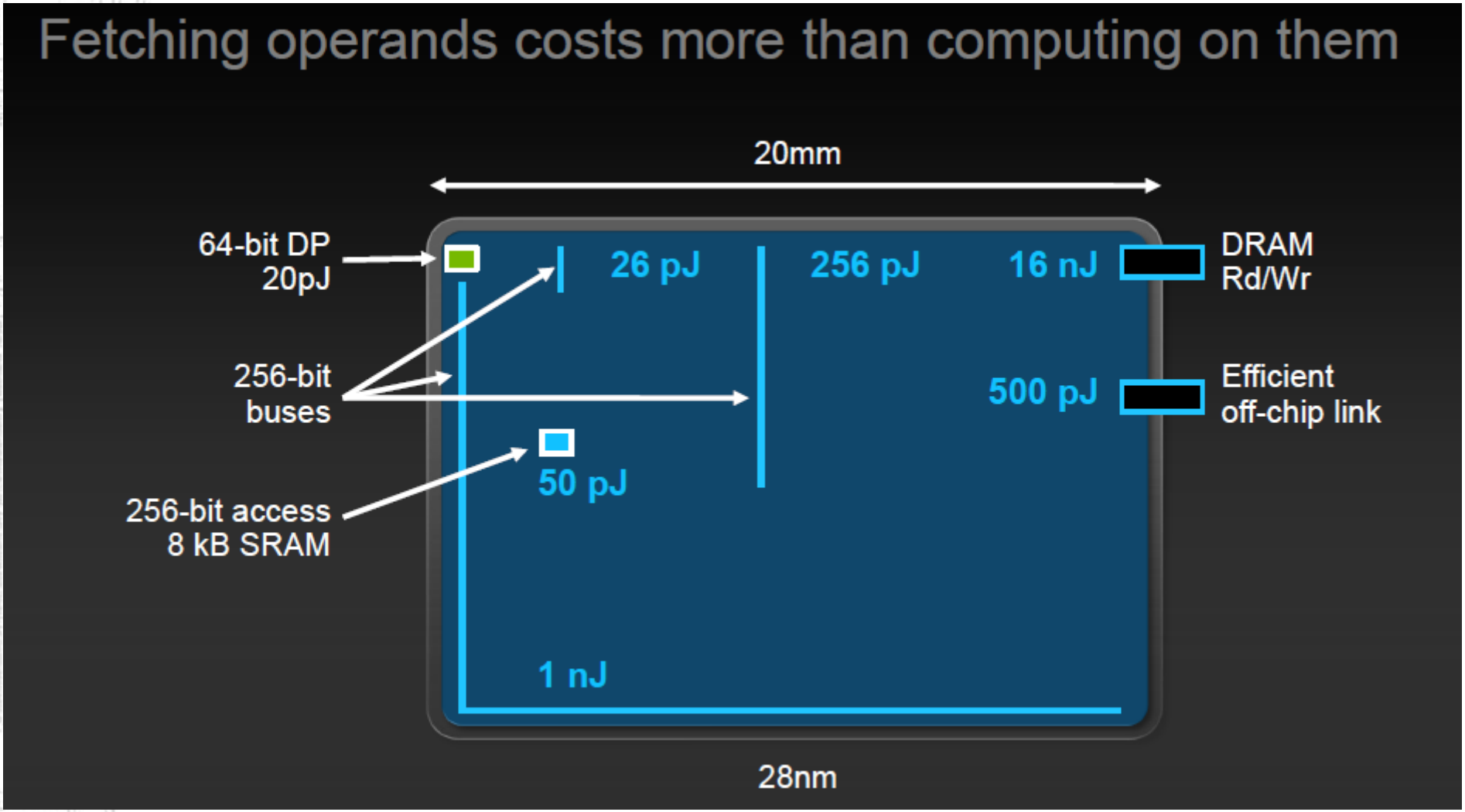
FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ



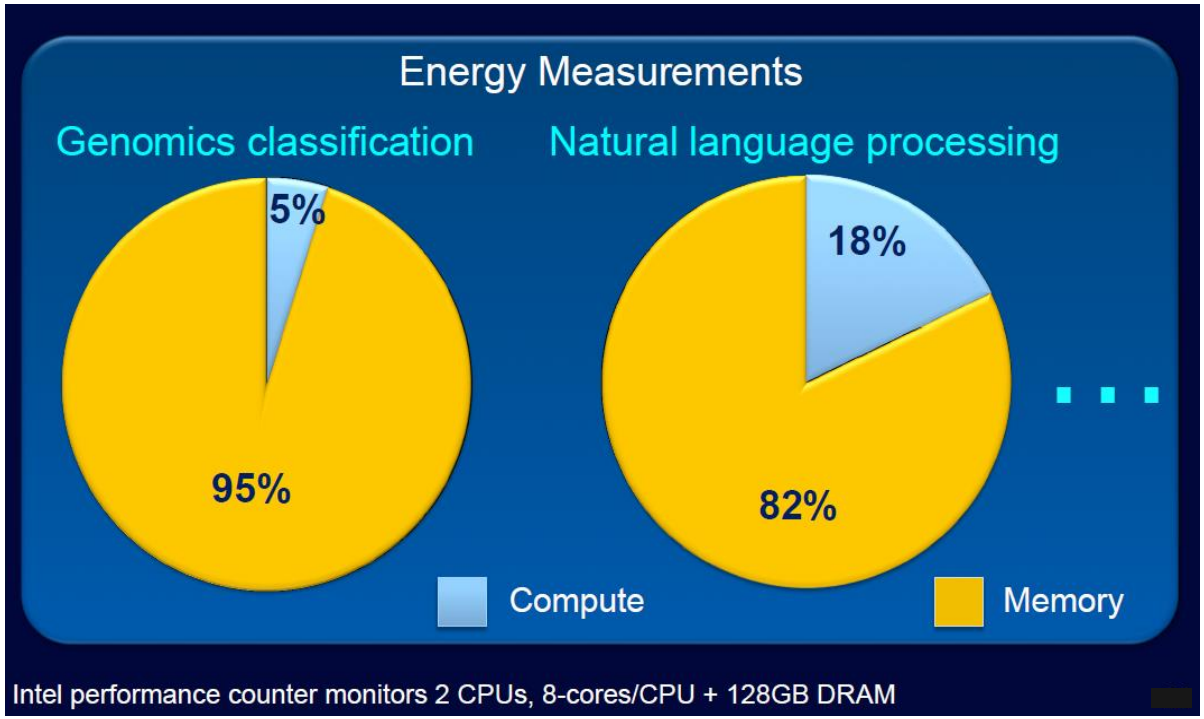
45nm, 0.9V, Horowitz ISSCC 2014

Energy Cost Computing @ 28nm



Source: Nvidia

Memory Energy Contribution



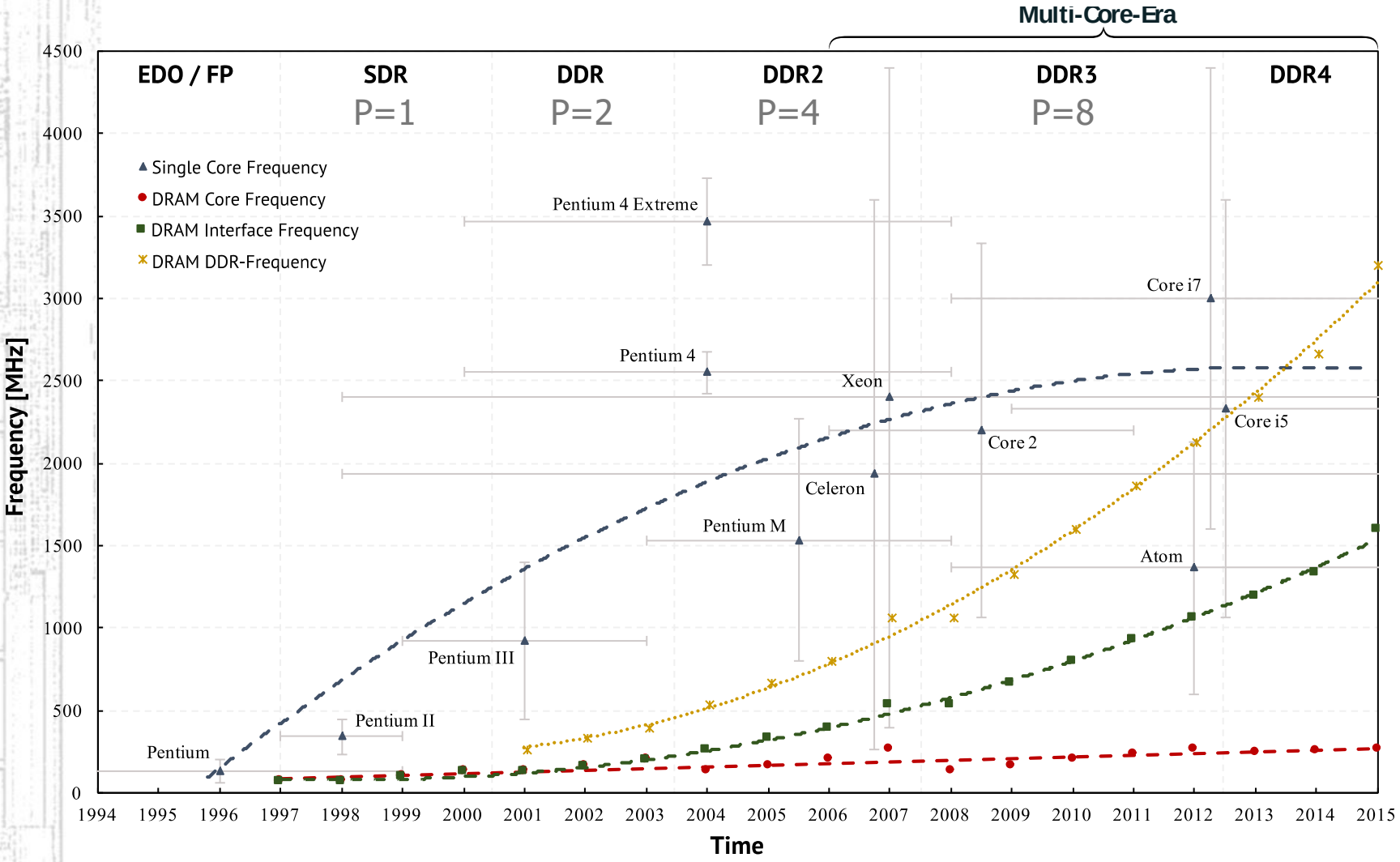
Source Subhasish Mitra

Relative energy cost

- SRAM vs. compute ~ 10x – 100x
- DRAM vs compute ~ 500x
- Flash vs. compute ~ 1000x

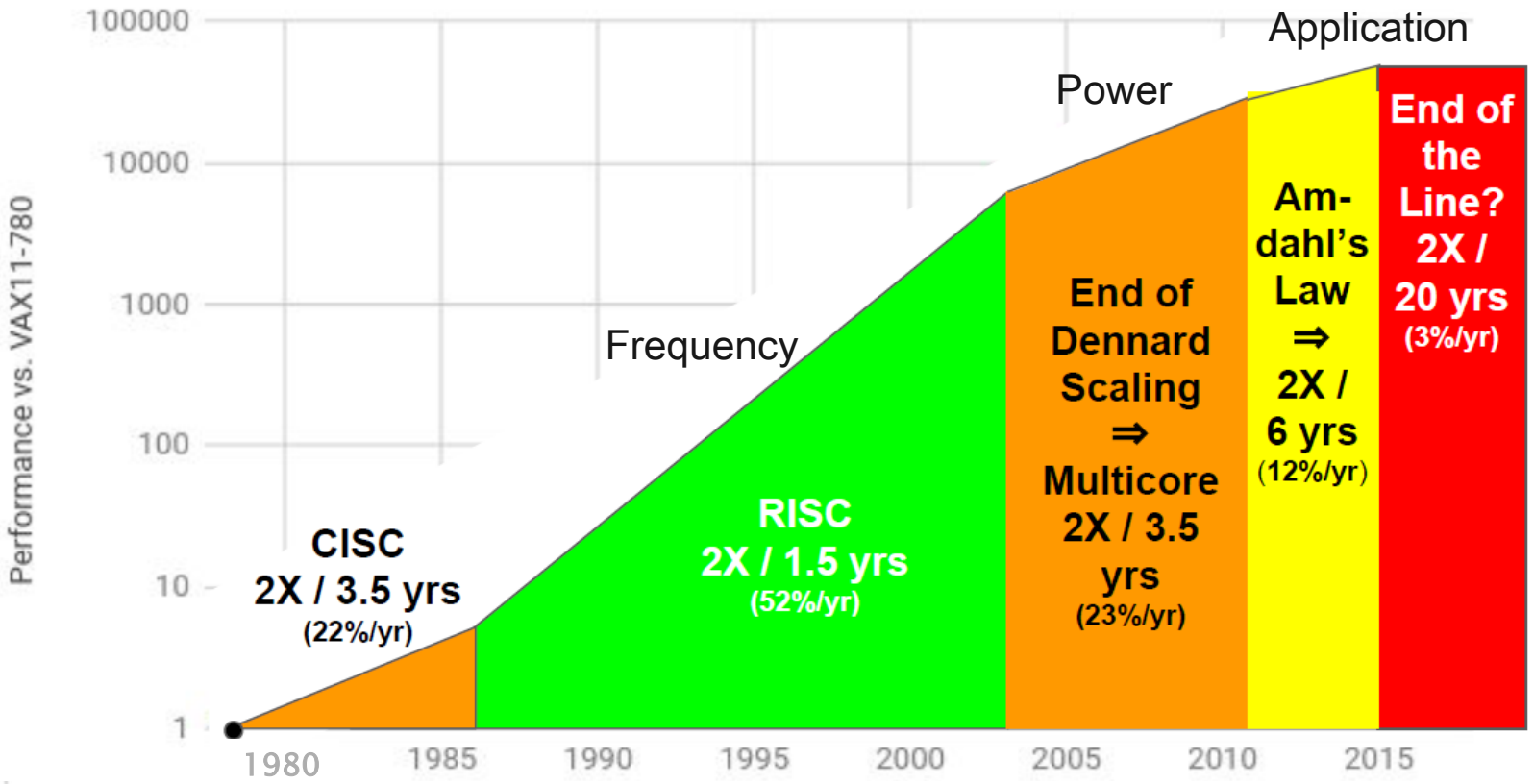
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Processor/DRAM Frequency Evolution



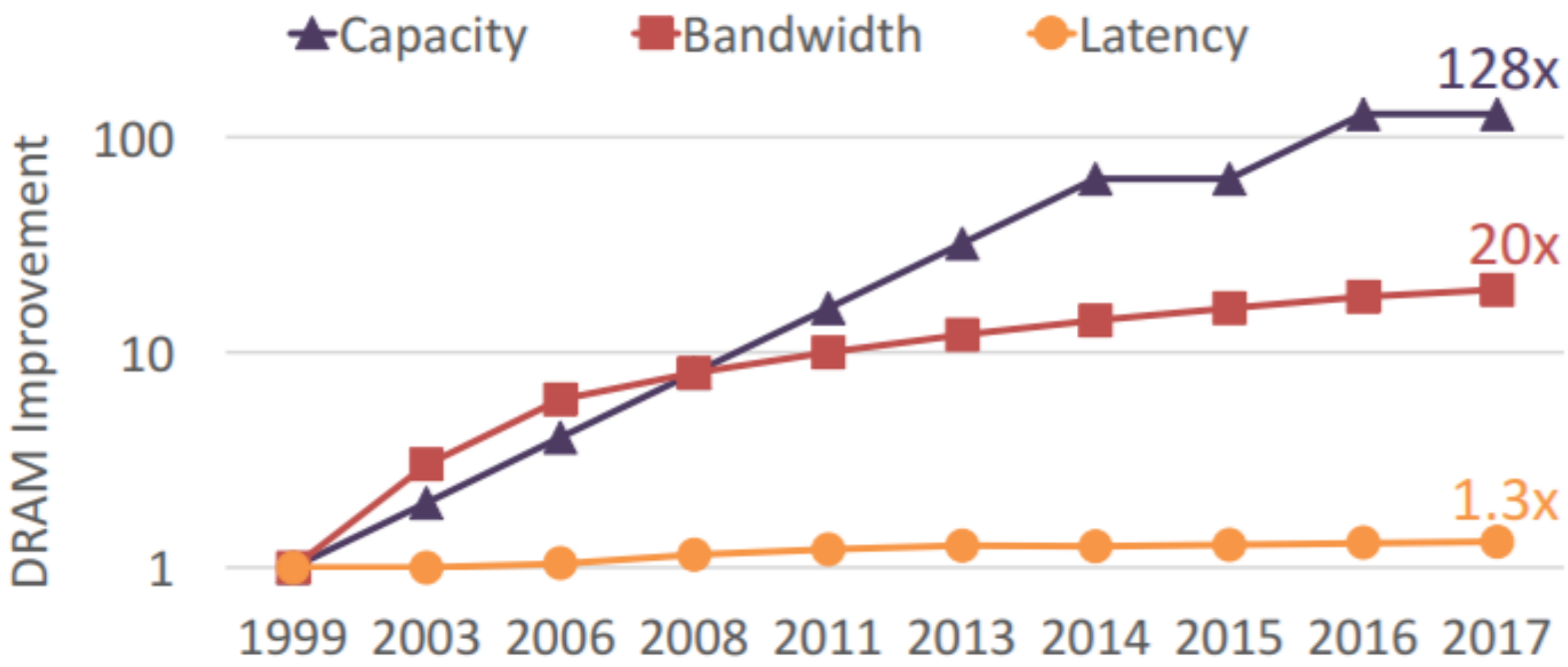
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Processor Performance Evolution



Source: Hennessy, Patterson

DRAM Performance Evolution



Source: K. Chang

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAM Chip Failures

DRAM's Damning Defects—and How They Cripple Computers

An investigation into dynamic random-access memory chip failure reveals surprising hardware vulnerabilities

By Ioan Stefanovici, Andy Hwang & Bianca Schroeder
Posted 23 Nov 2015 | 16:00 GMT

IEEE SPECTRUM

IEEE Spectrum Nov. 2015

Analysis of dozens high-performance computing clusters

- 300 terabyte-years of DRAM usage
- System failures: not software problems, but more than 60% of machine outage results from hardware issues
- Most common hardware problem was faulty DRAM

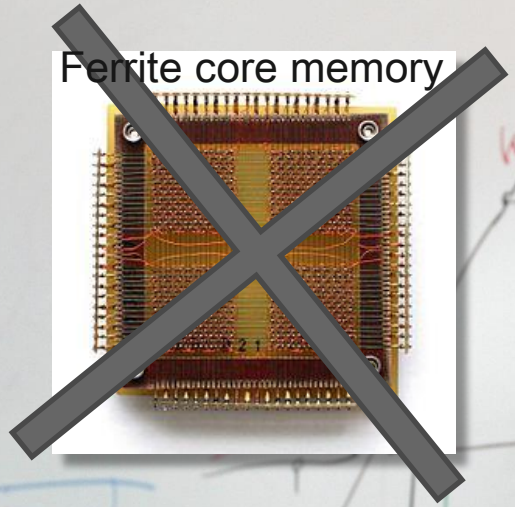
DRAM: permanent errors, soft errors, retention errors

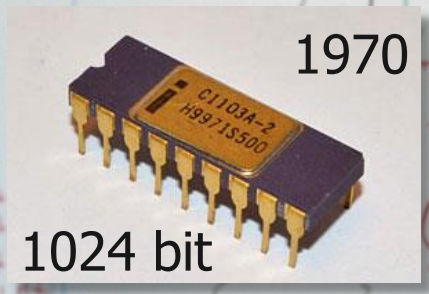
- Redundancy, ECC, refresh

Content

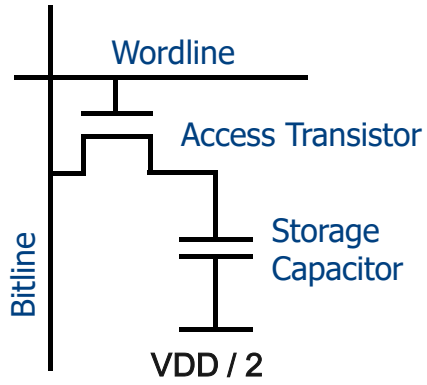
- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- **The DRAM Device**
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Modes
- The Memory Controller

DRAM: Invented by Robert H. Dennard in 1966





The DRAM Cell – A simple Model



- Data is stored by capacity
- Cell is selected with access transistor
- Charged capacitor represents a '1'
- Discharged capacitor represents a '0'
- Memory is volatile
- Cell is leaky: refresh needed

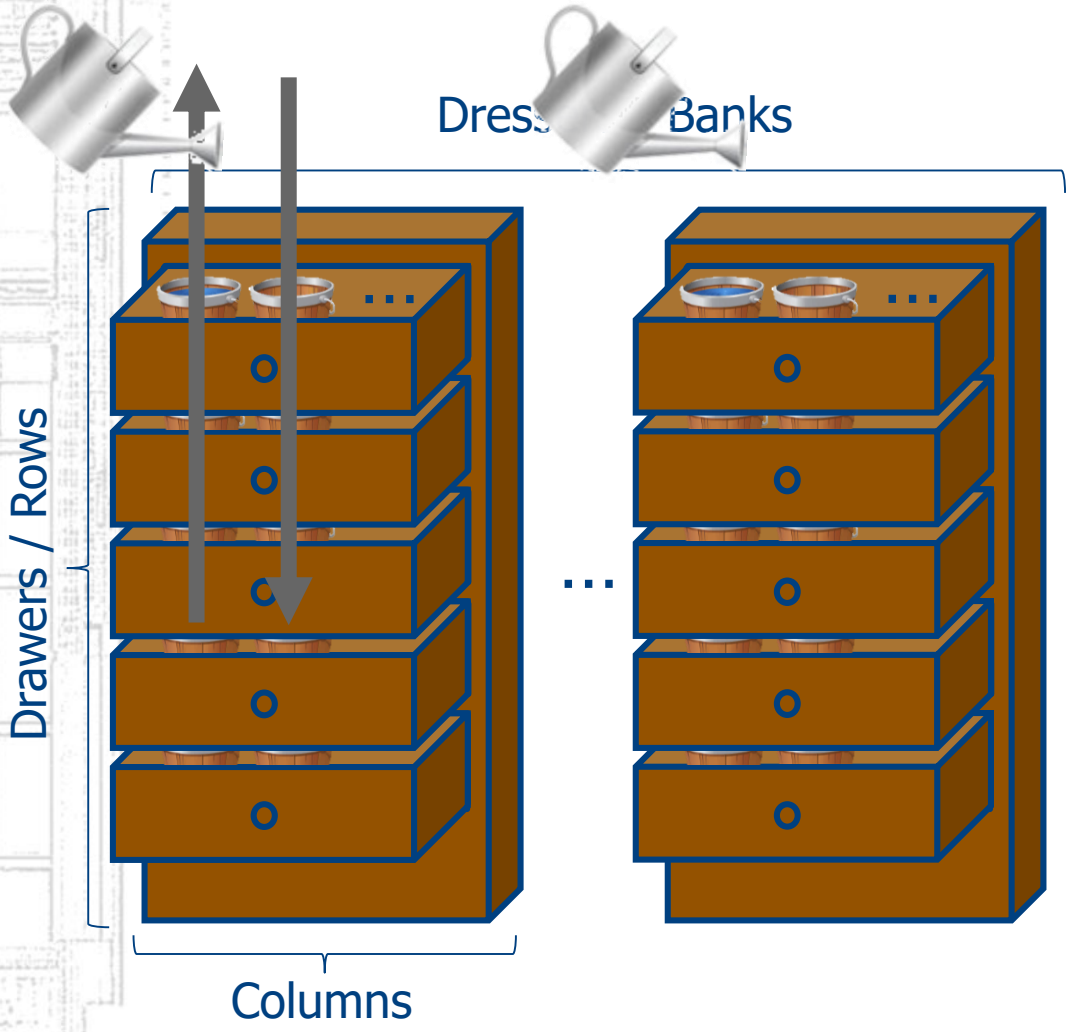
'1'

'0'

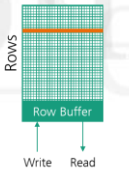
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAMs Operations

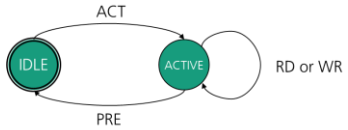
DRAM composed of banks/rows/columns



Finite State Machine Example – A Single Memory Bank



- In the beginning we are IDLE state
- If we want to read or write we have to be in the ACTIVE state.
- A specific row is stored in the bank's row buffer (ACT)
- Then we can perform read (RD) or write (WR) operations
- If we want to read data from another row we have to close the current row (PRE)



DRAM Commands

Activate (ACT)

- Only one row / bank
- Bank parallelism

Read (RD)

Write (WR)

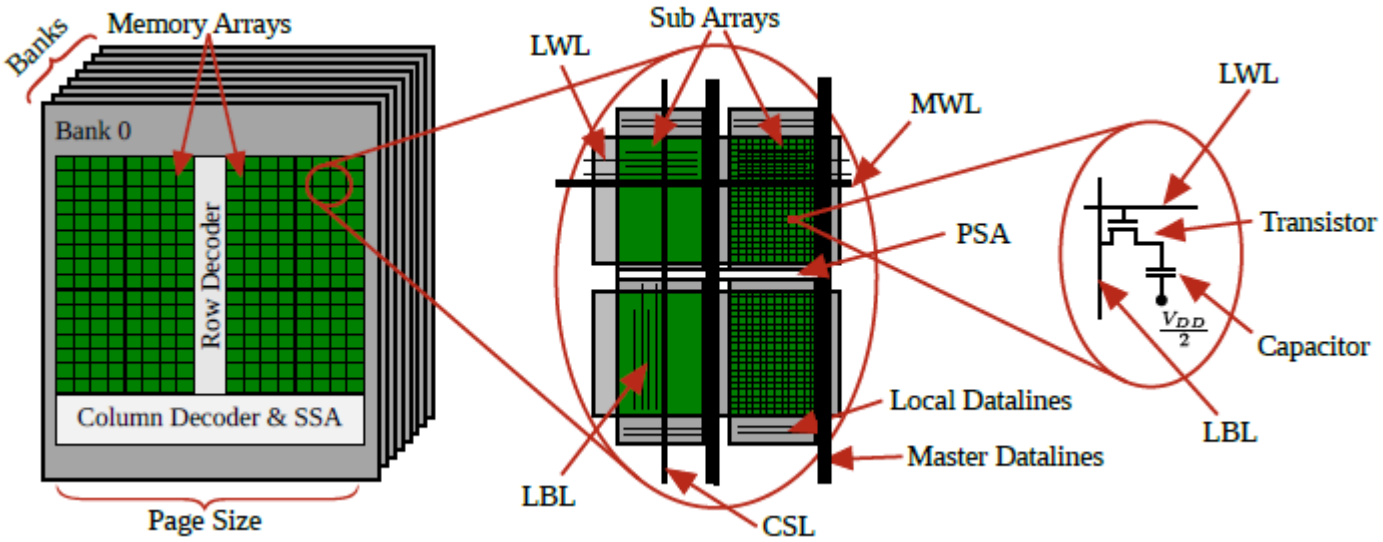
Precharge (PRE)

- Row miss

Precharge All (PREA)

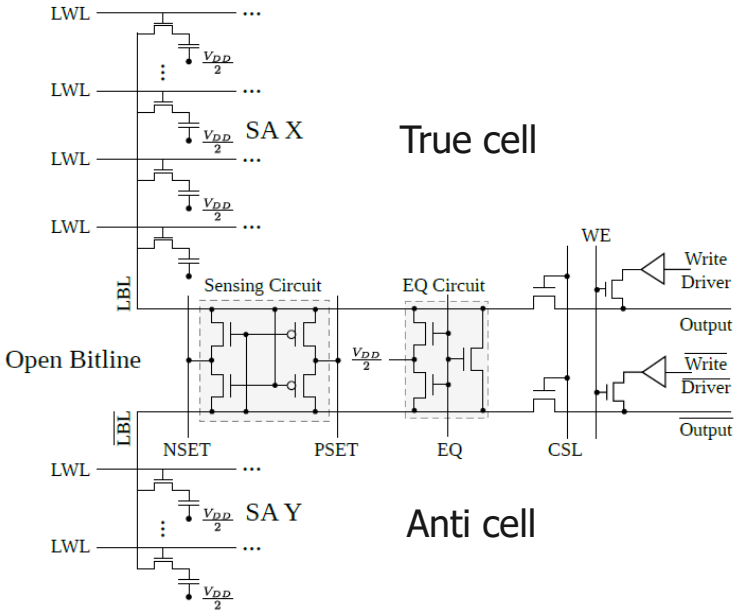
Refresh (REFA)

The DRAM Device

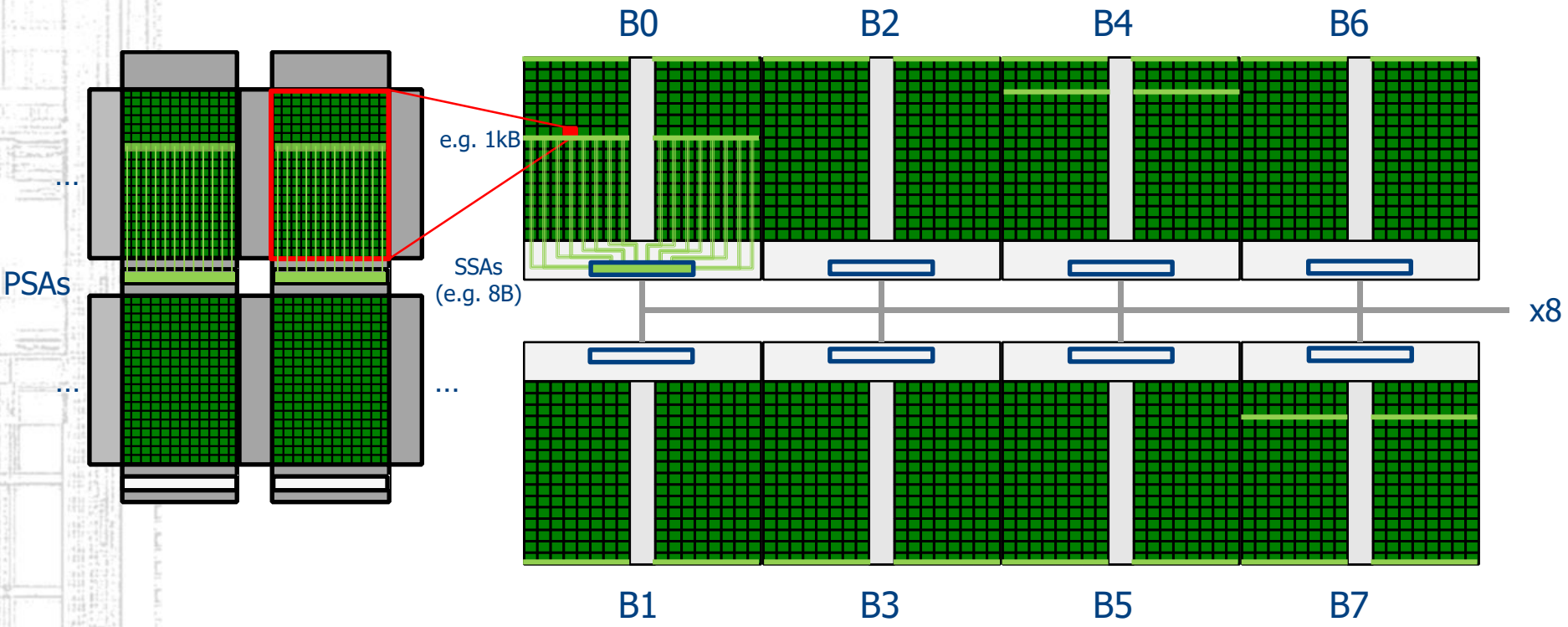


DRAM commands

- Read, Write
- Precharge, Activate
- Refresh



DRAM Device Operations

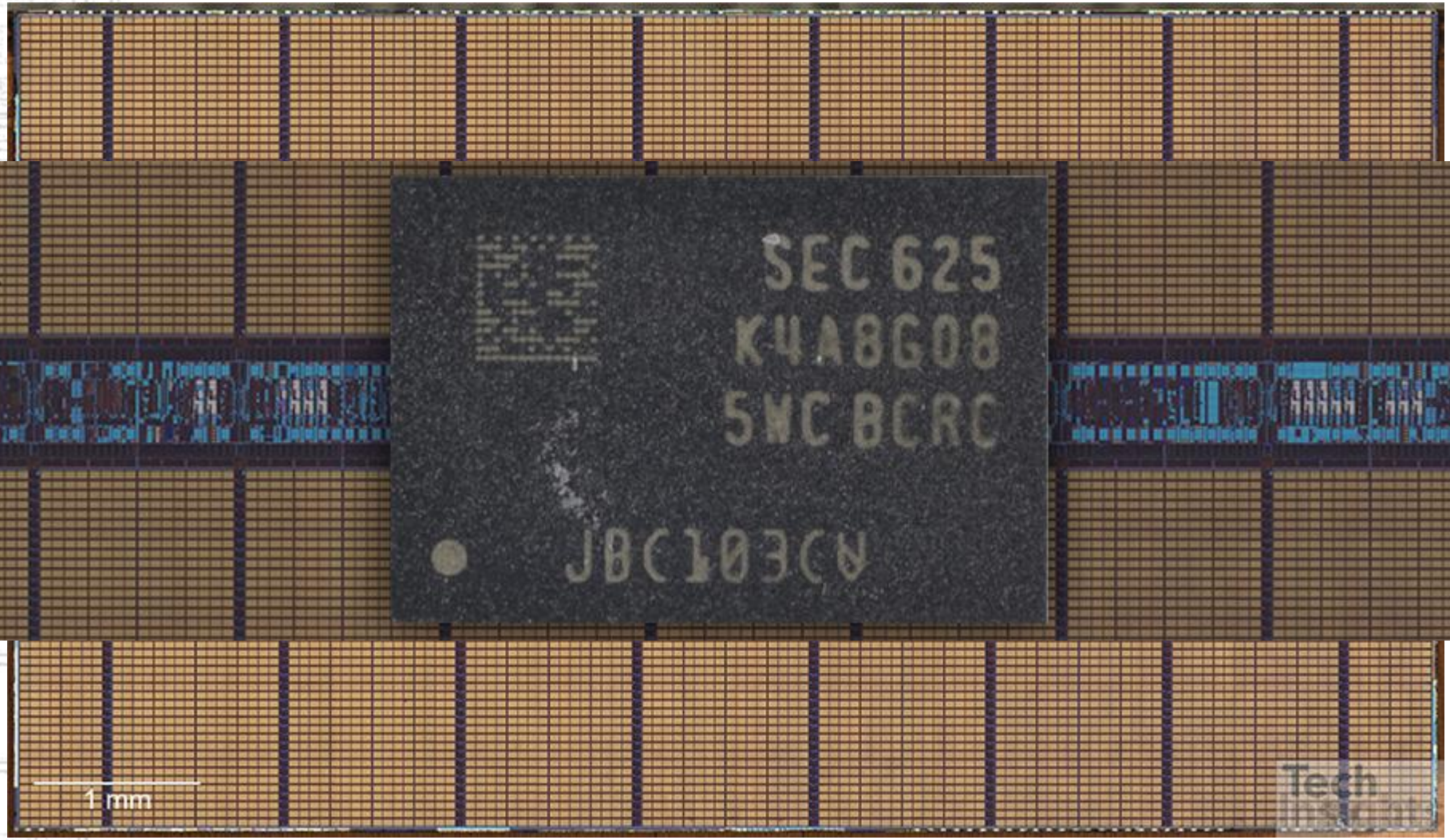


DRAM Commands

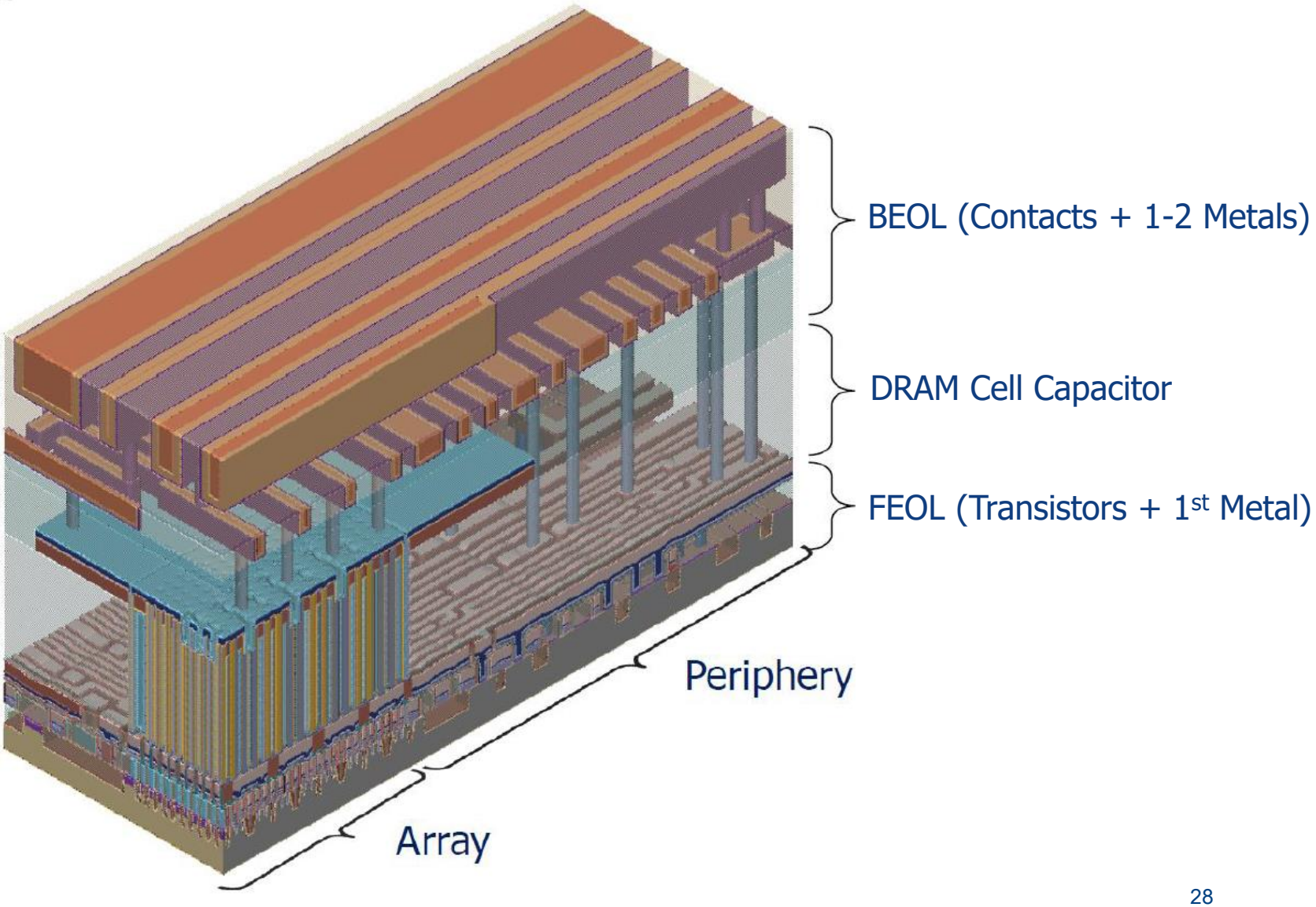
- ACT: Activates a specific row in a specific bank (sensing into PSA) *tRCD*
- RD: Read from activated row (prefetch from PSA to SSA and burst out) *tCL + tBURST*
- PRE: Precharges set LWL=0 set LBL=VDD/2 *tRP*
- REFA: DRAM cells are leaky and have to be refreshed *tREFI & tREF*

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

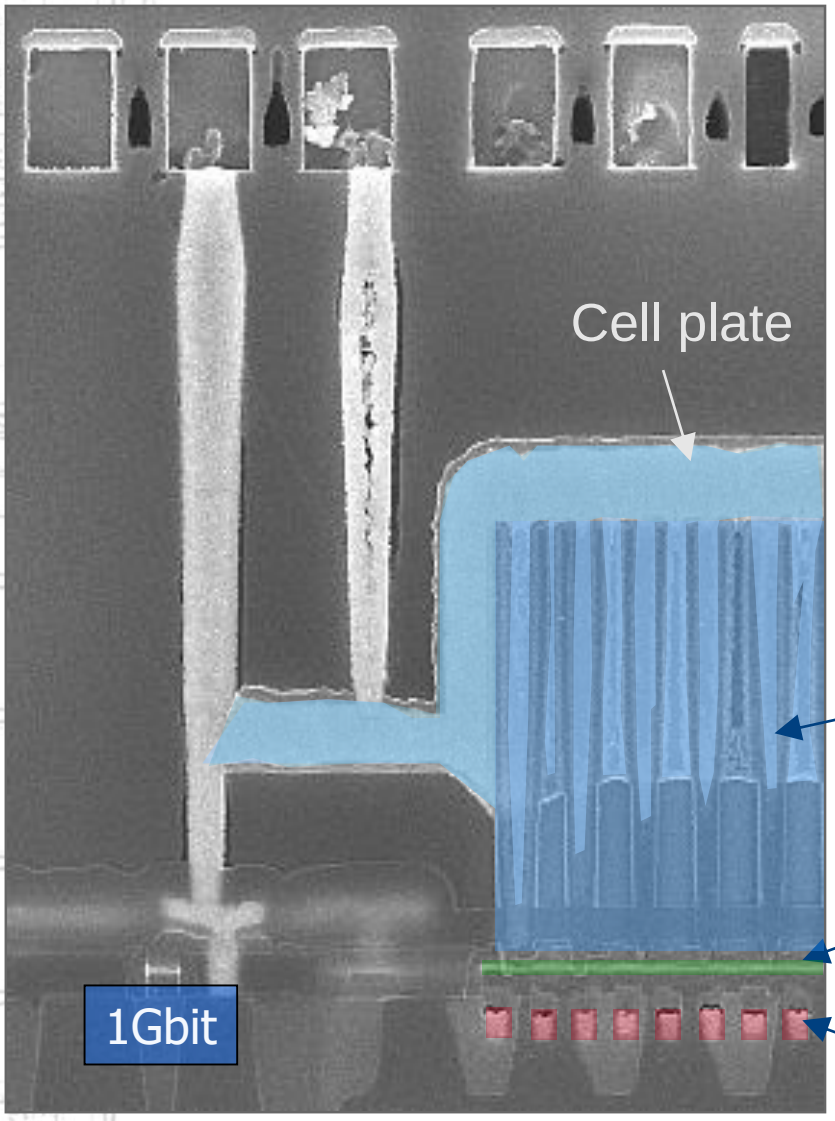
Samsungs 8Gb 20nm DDR4 Die



Isometric DRAM Technology View

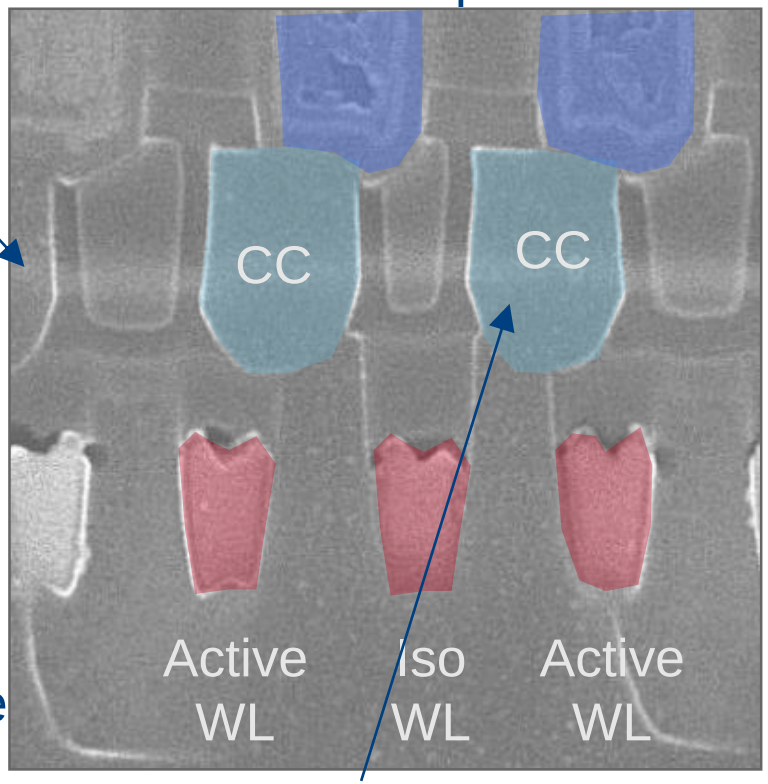


Evolution in DRAM Generations



65 nm Burried Wordline

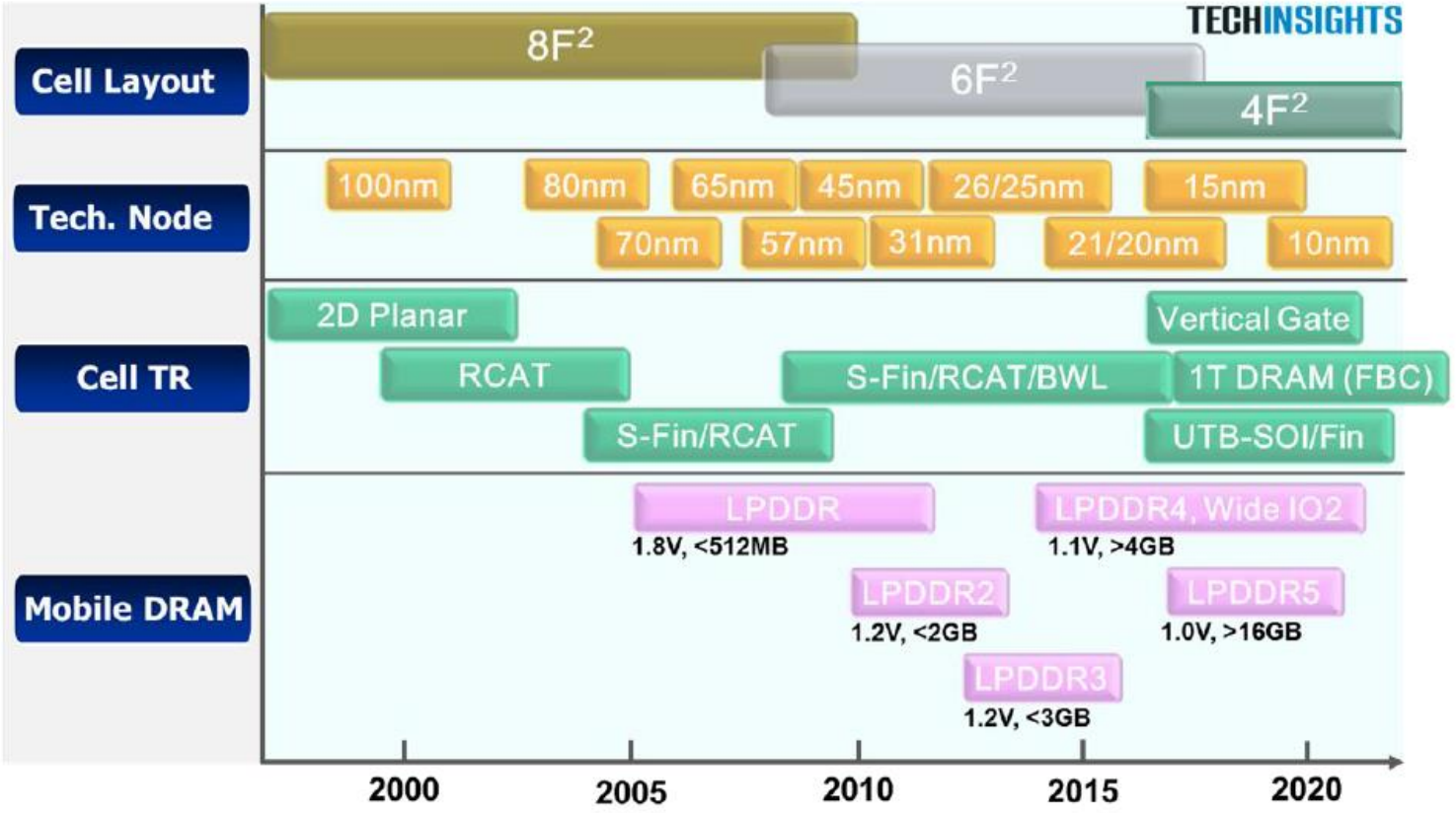
Lower part of cells



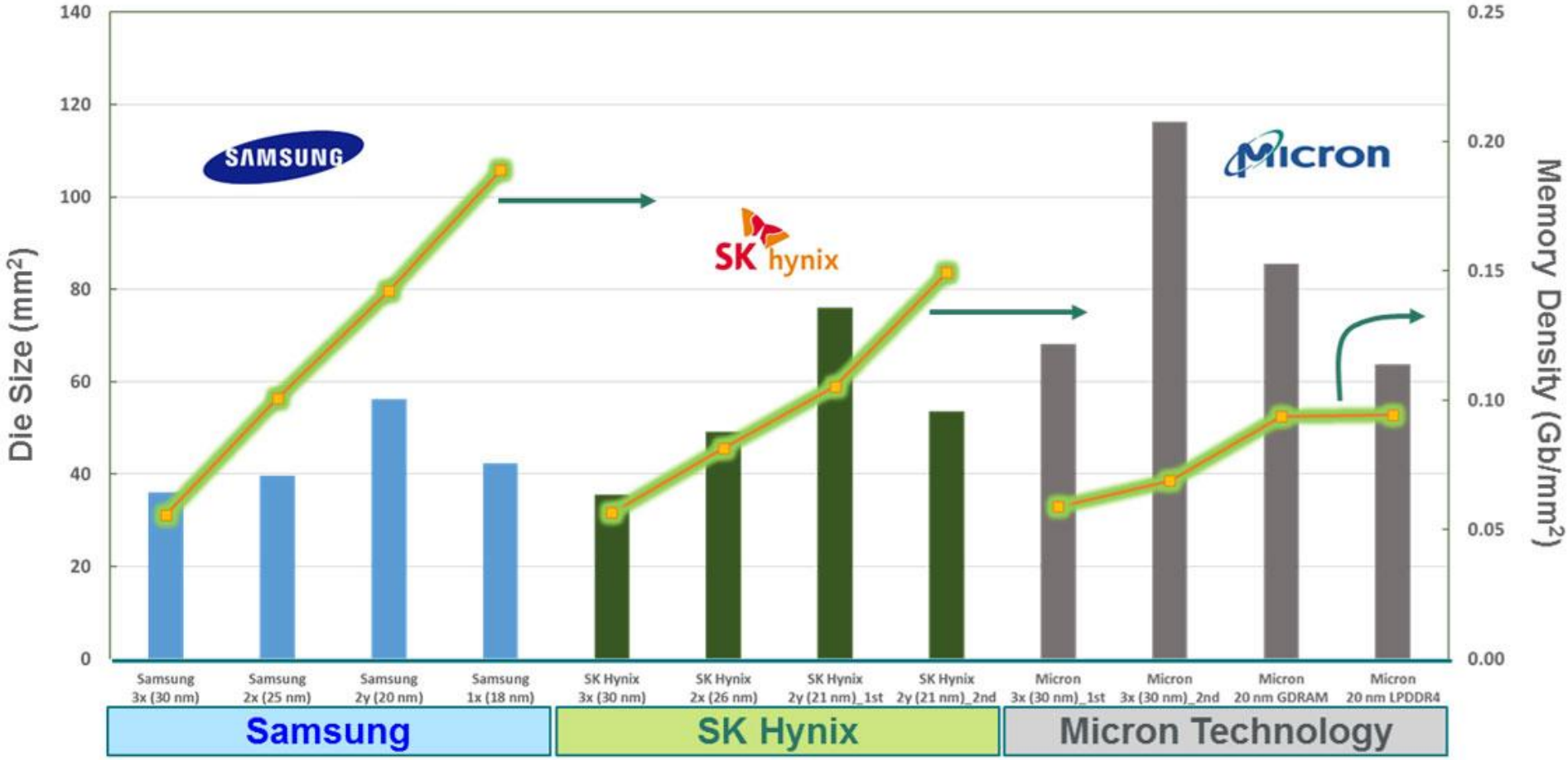
source: Qimonda

Design and Technology Trends

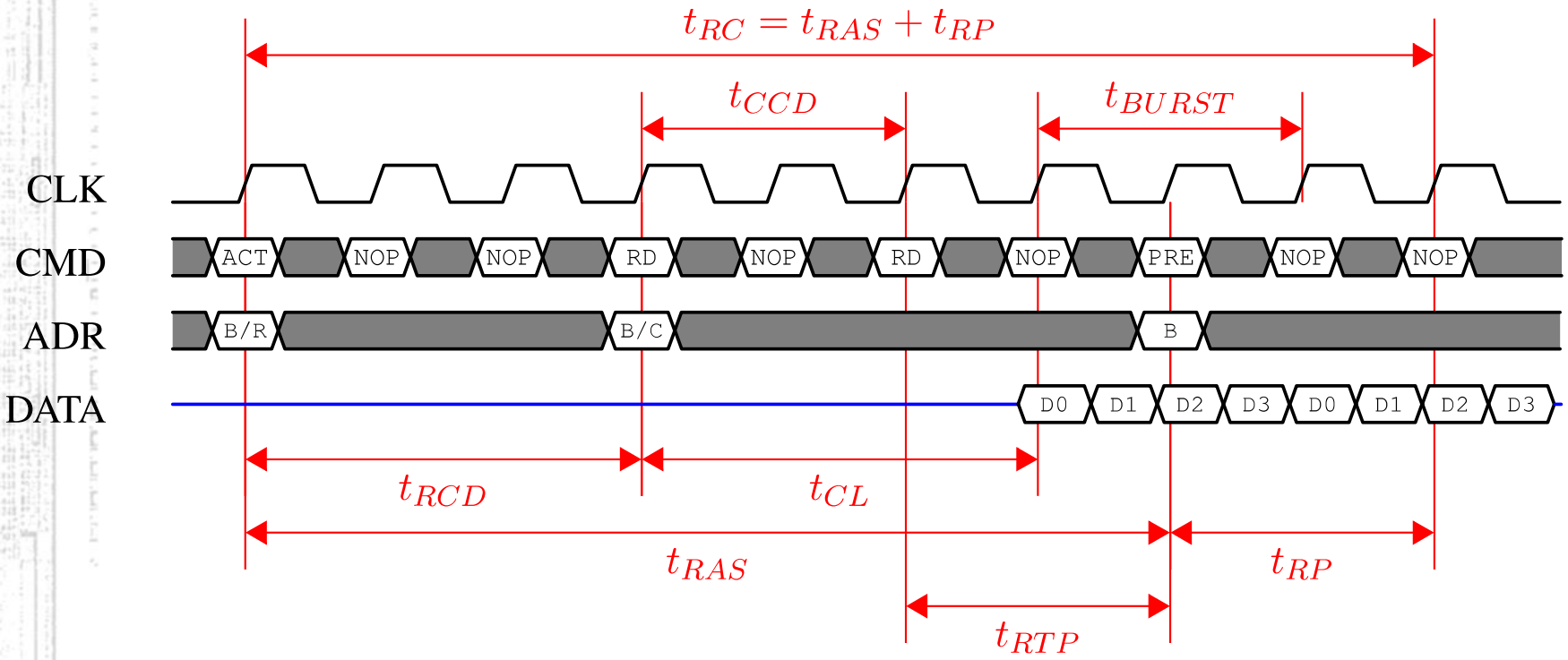
DRAM Technology Roadmap



DRAM Die Size & Density Trend



JEDEC Standard: e.g. Timing Dependencies



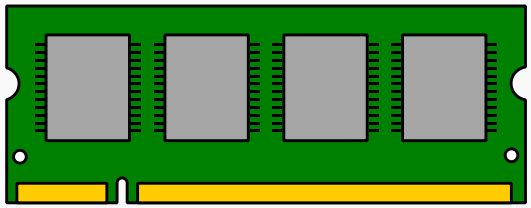
Timing dependencies must be fulfilled by the **DRAM controller**

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAM Subsystem Evolution

DIMM Based:

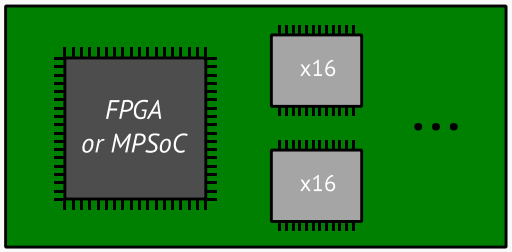
General Purpose Computers
e.g. DDR3, DDR4



Computational Units

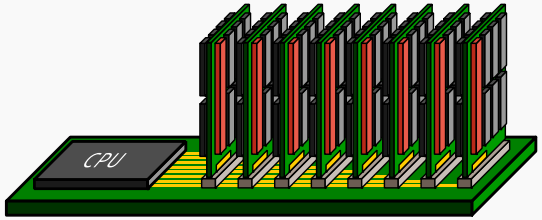
Device Based:

Embedded / Tablets / Graphic Cards
e.g. LPDDR3, GDDR5



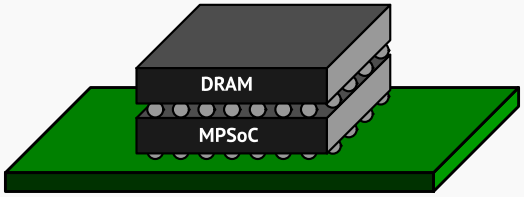
Buffer on Board:

Memory Controller on Buffer Chip,
Serial Connection
e.g. FBDIMM, IBM CDIMM, Intel SMI/SMB



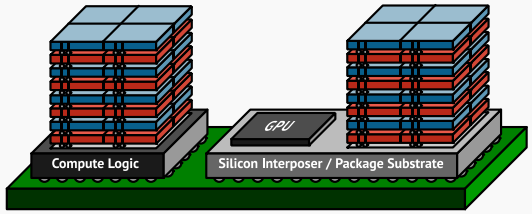
Package on Package (PoP):

Soldered on top of the MPSoC.
Smartphones
e.g. LPDDR3, LPDDR4



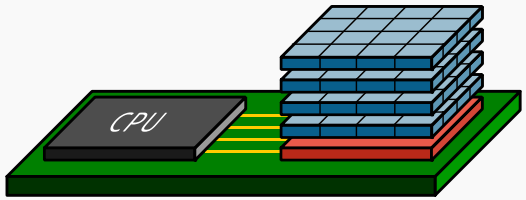
3D/2.5D-Integrated:

Stacked on Logic or Silicon Interposer
by means of TSVs
e.g. Wide I/O, HBM



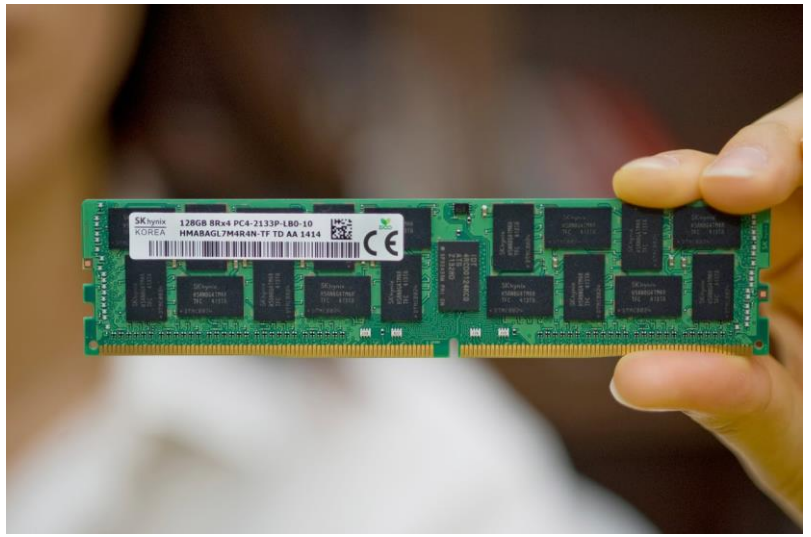
Memory Cube:

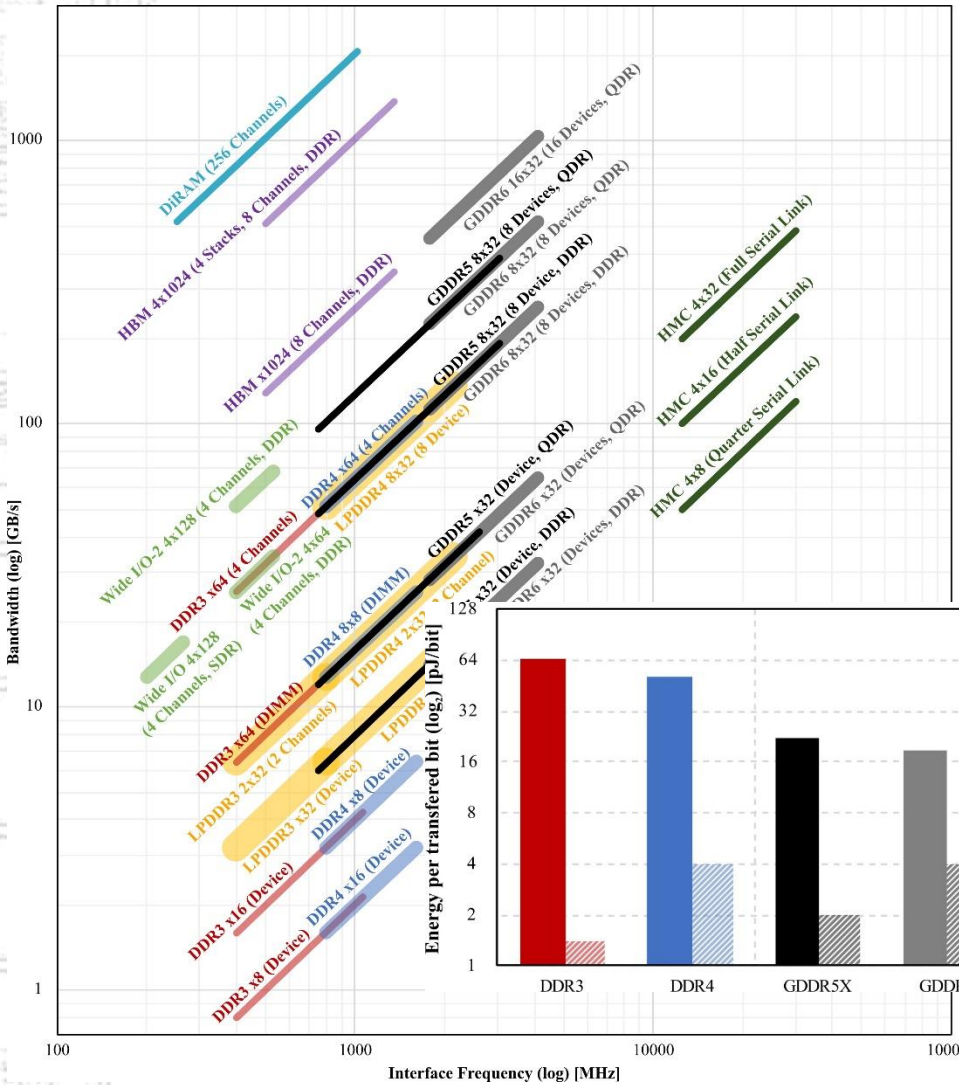
3D-Stacked, Memory Controller on
Bottom Layer, Serial Interconnect (SerDes)
e.g. HMC, SMC

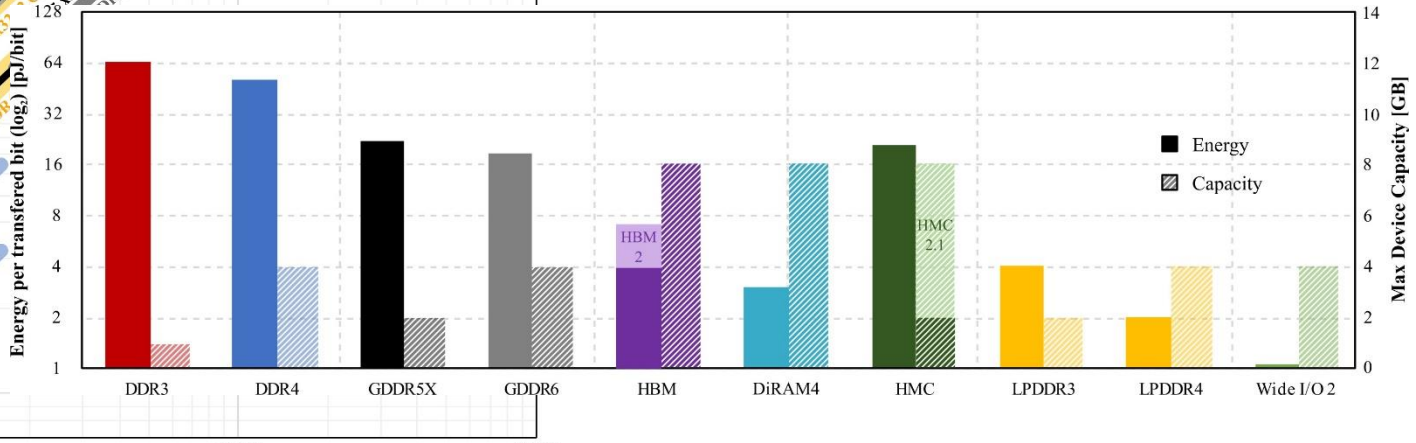


MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAM Subsystem Performance

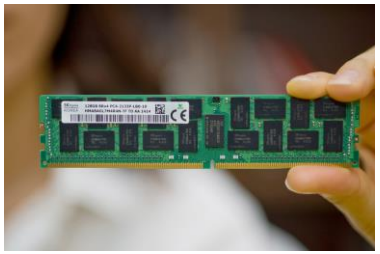






Detailed DRAM Energy Distribution

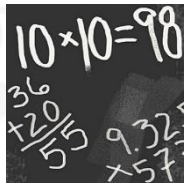
- DRAM Power Breakdown for Twitter Memcached Application*
- 2GB DDR3



(*) A High-Level DRAM Timing, Power and Area Exploration Tool, O. Naji, A. Hansson, C. Weis, M. Jung, N. Wehn
IEEE International Conference on Embedded Computer Systems Architectures Modeling and Simulation (SAMOS), July 2015

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAM in Transprecision Computing



Minimize energy/DRAM access

Reduced precision implies

- Reduced DRAM power/energy/latency, increased DRAM bandwidth

Minimize Refresh

- Impacts energy, latency and reliability
- Temperature aware refresh, bank wise refresh, optimized row-granular refresh, approximate DRAM
- ECC

Minimize row misses (Activate, Precharge)

- Impacts latency, bandwidth, energy
- Scheduling in memory controller
- Address mapping

Sophisticated power-down modes

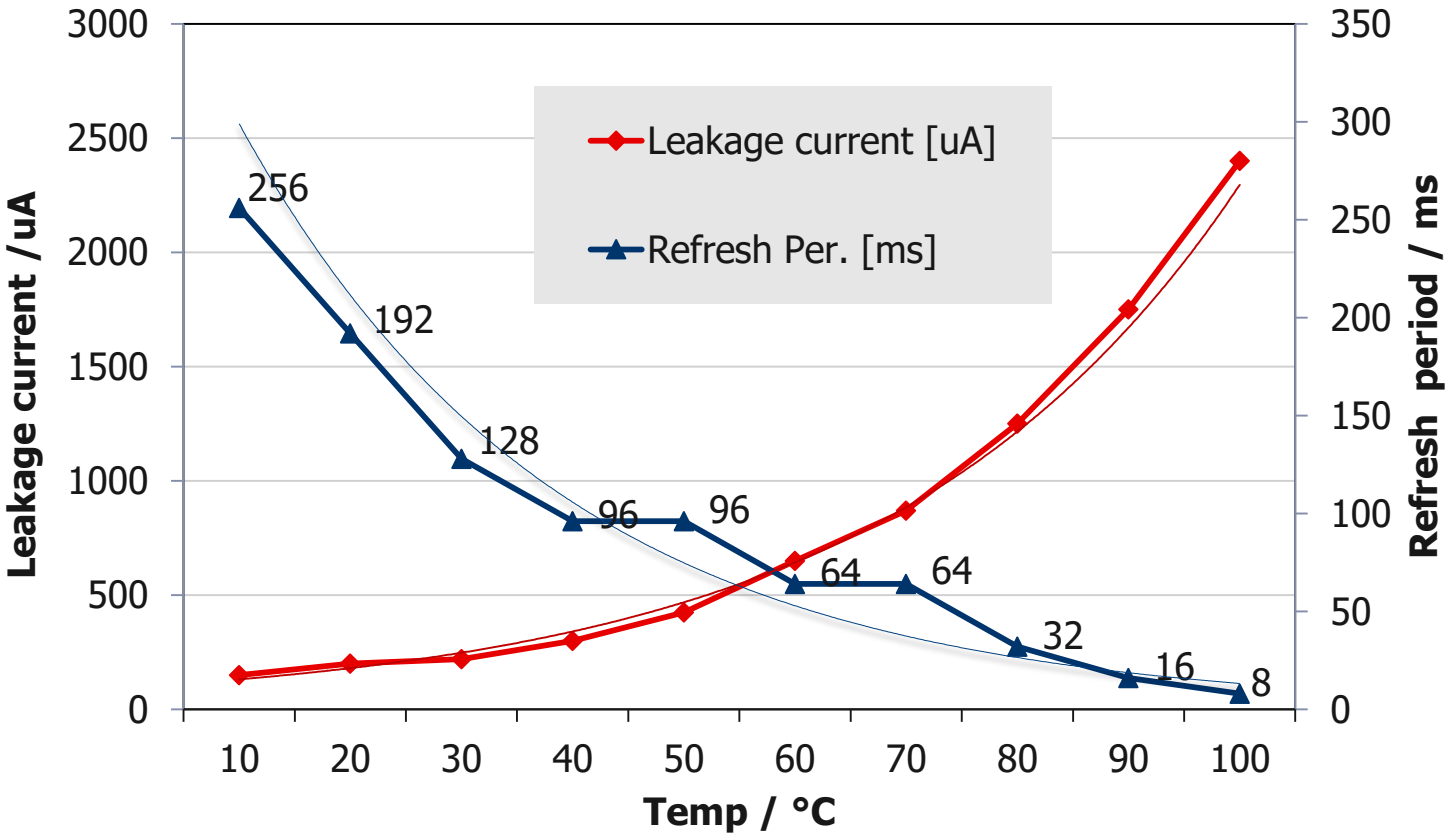
Near-Memory or In-Memory Computing

Content

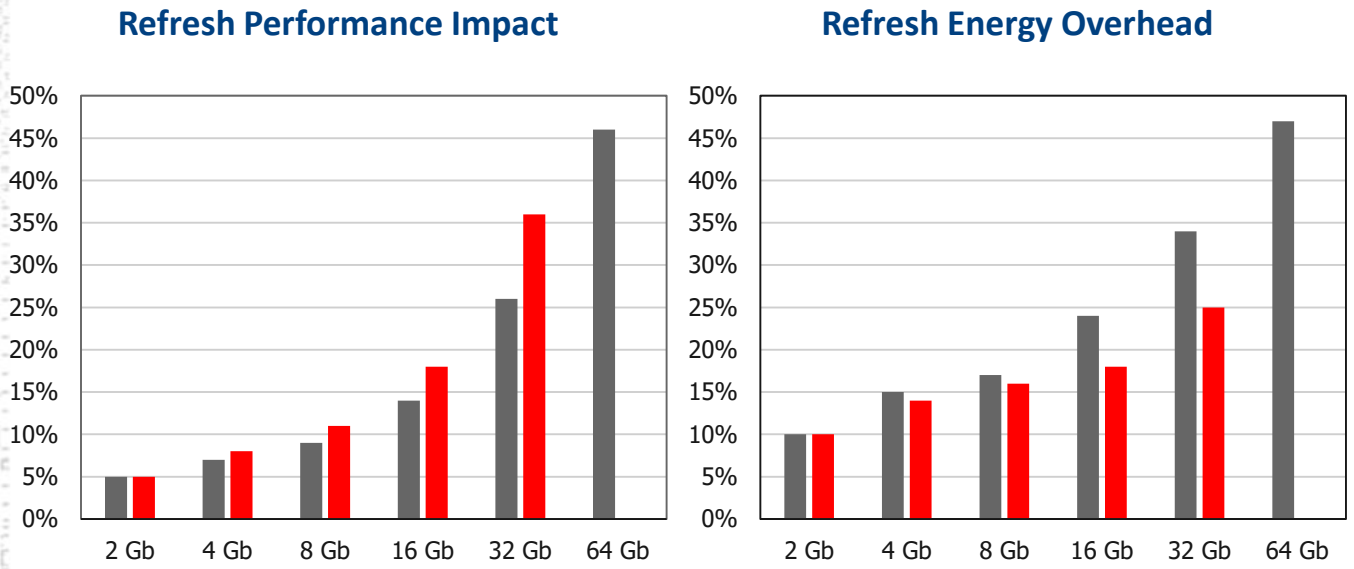
- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- **The Refresh Challenge and Optimizations**
- Improvement of Reliability
- Power Down Modi
- The Memory Controller

Refresh/Temperature Challenge

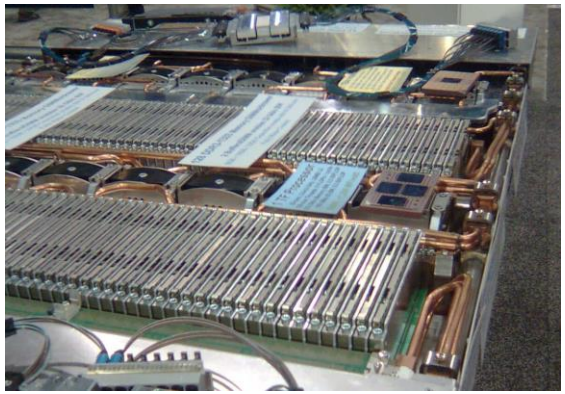
Exponential temperature/leakage current behavior → shorter refresh periods



Refresh/Temperature Challenge



- *J. Liu, et al. RAIDR: Retention-Aware Intelligent DRAM Refresh, ISCA 2012*
- *I. Bhati, et al. DRAM Refresh Mechanisms, Trade-offs and Penalties, IEEE Trans. 2015*



4 TB DDR3 DRAM
Stand-by 300W

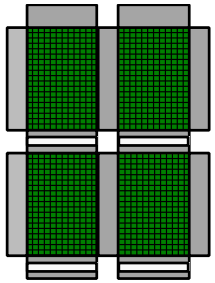
Paul Rosenfeld (IBM Server on display at Supercomputing)

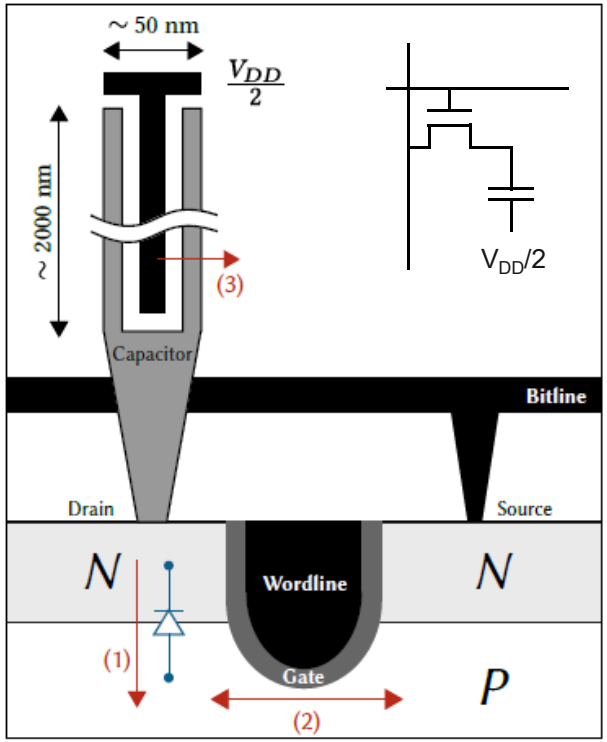
Retention Errors

1. Drain Leakage
2. Sub-Threshold leakage
3. Cell Capacitor Leakage

Cross-Talk/Coupling

- Bitline to bitline
- Wordline to wordline
- Wordline to bitline
- Bitline to cell





Challenge

- Characterization of retention error behavior of DRAMs
- Vendor: over pessimistic

Refresh Optimizations

Following optimizations will be addressed:

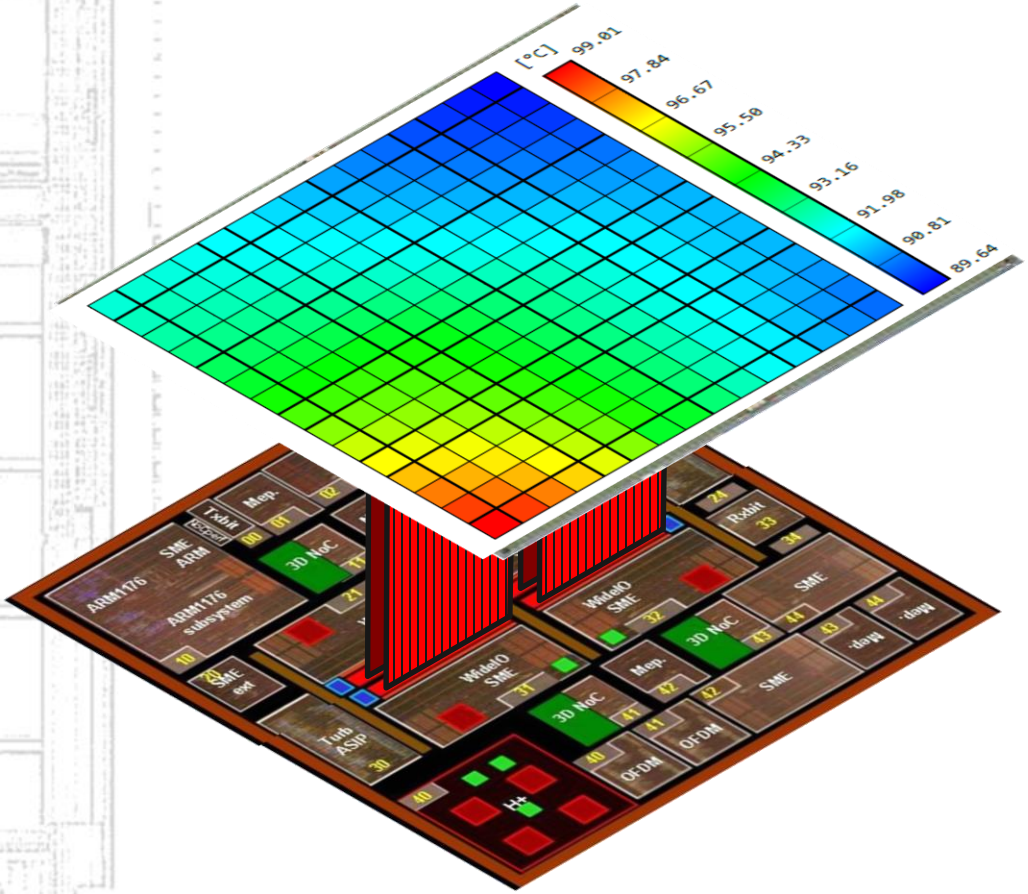
- Temperature-aware Fine-granular Refresh
  - e.g. per bank/per layer
- Approximate DRAM
  - disable or reduce Refresh-rate
- Optimized Row-Granular Refresh (ORGR)
  - not relying on Auto-refresh

DRAMs in 3D MPSoC's

Memory bandwidth/energy bottleneck ⇒ Wide I/O DRAM, HBM, HMC

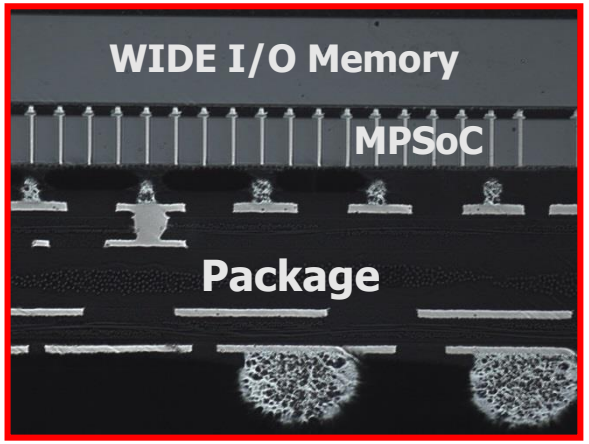
E.g. WIOMING 3D Magali (LETI): baseband processing, accelerators, CPUs, DRAM

- 65nm, 72mm², 1250 TSV, heaters/sensors



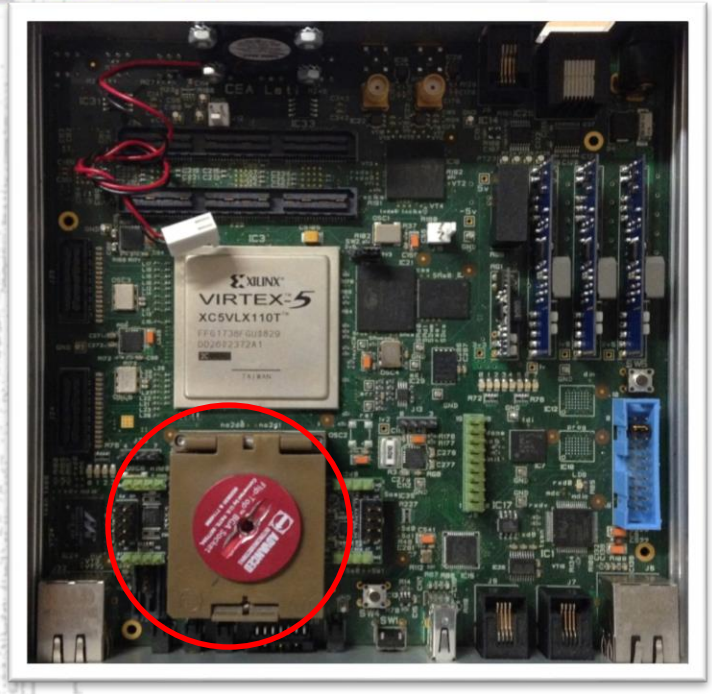
Wide I/O DRAM (50nm):

256Mb, 512 I/Os, 1Gb, 4 Channels, SDR@200MHz, 12.8 GBps



MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

3D-DRAM Retention Error Model



Measurements@WIOMING

On-chip heaters and sensors

Different refresh periods

Test a: Disable refresh for 1s

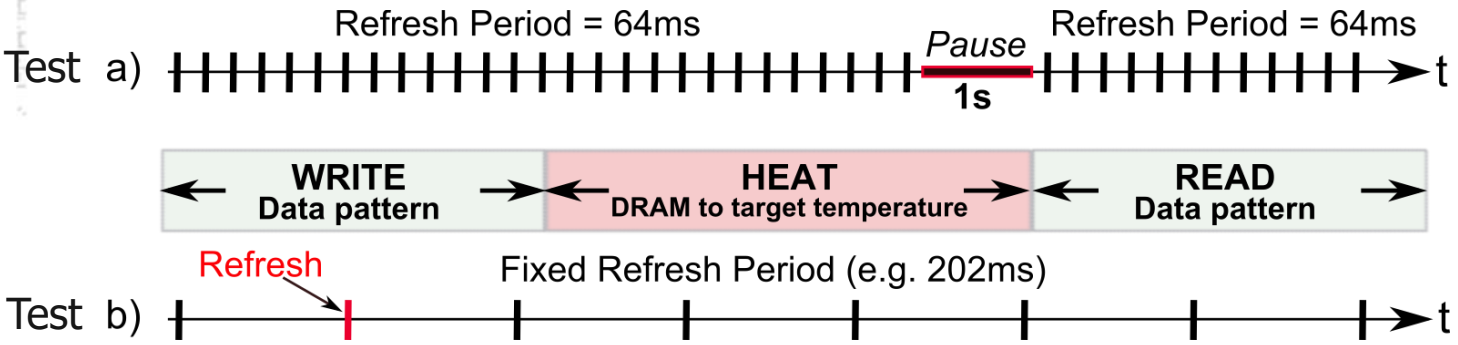
Test b: 128ms – 202ms

Different temperatures

80°C – 104°C

Different data patterns

0xFF…FF, 0xAA…AA, 0x55..55, random

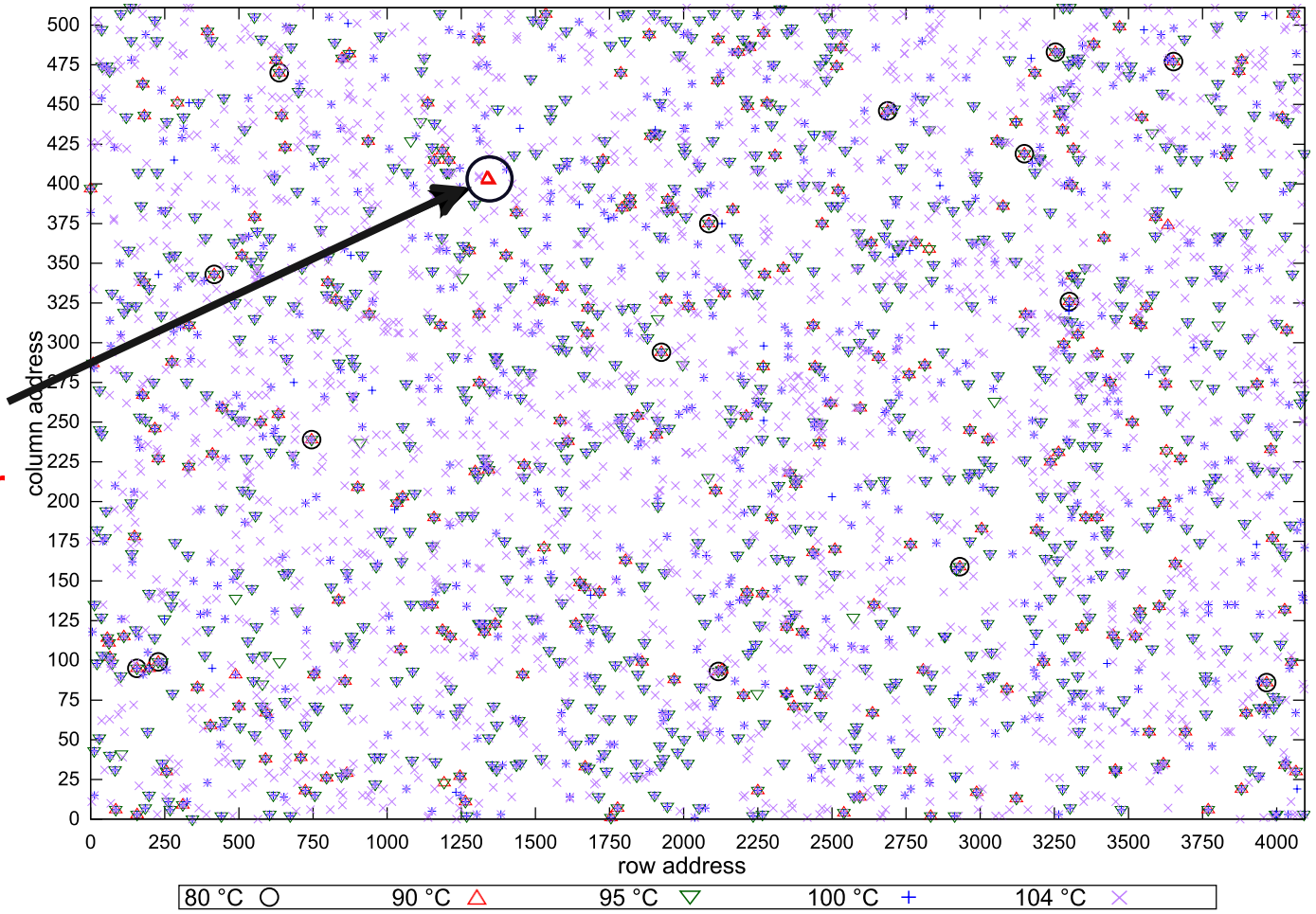


MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

3D-DRAM Retention Error Model

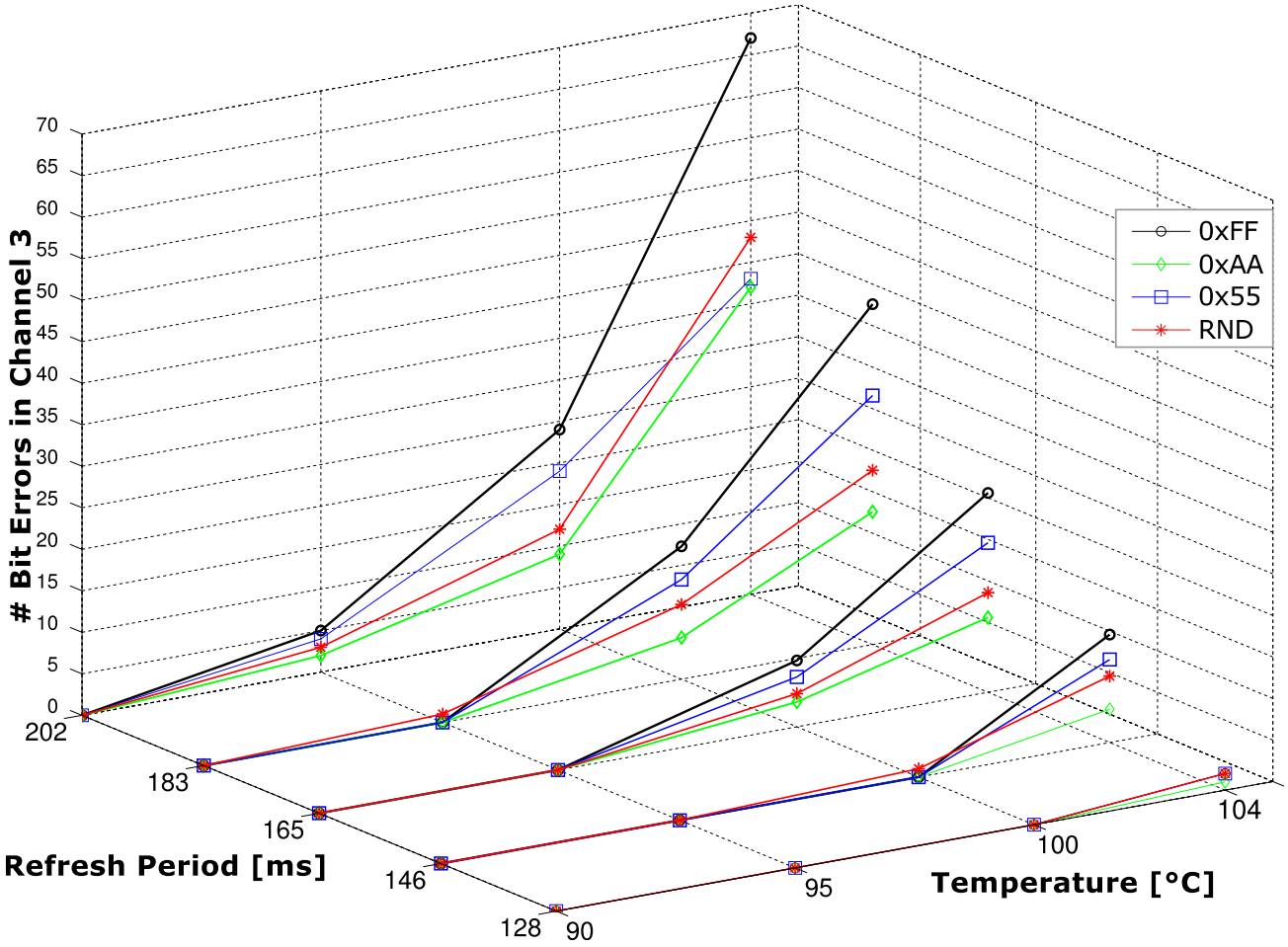
Variable Retention Times (VRT)

Unique bit error at 90°C



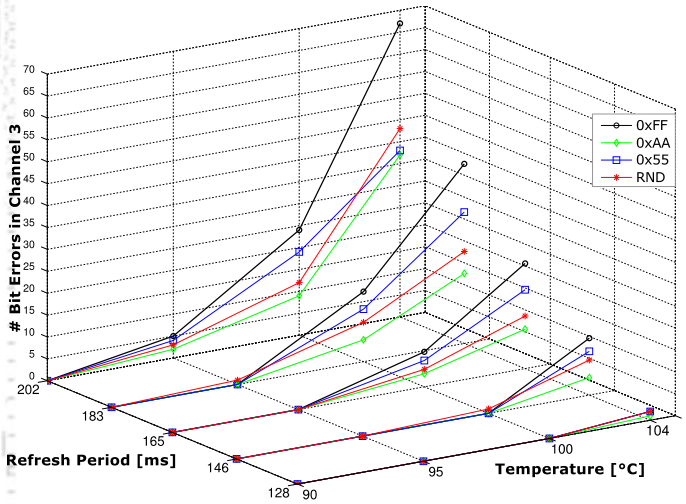
3D-DRAM Retention Error Model

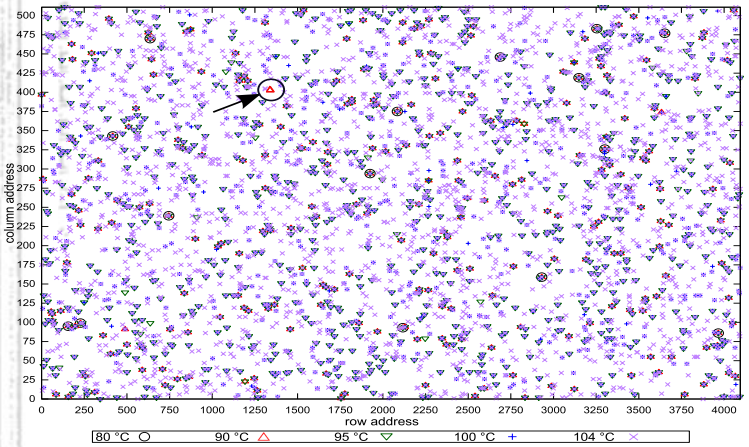
Data Pattern Dependency (DPD)



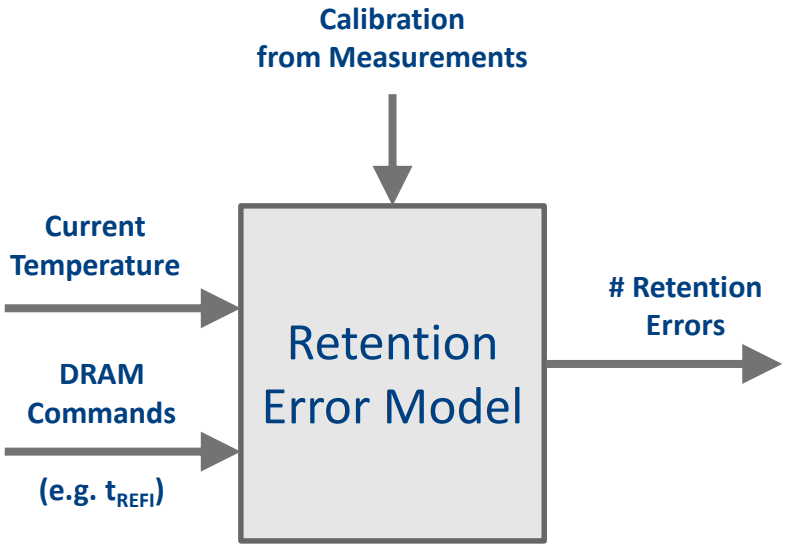
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

3D-DRAM Retention Error Model





- Modelling of Data Pattern Dependency
- Modelling of Variable Retention Times



DRAMSys

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Simulation Flow 3D MPSoC's

System Behavior
& Retention Error Modeling

Cores: gem5

DRAM: DRAMSys

Power Analysis

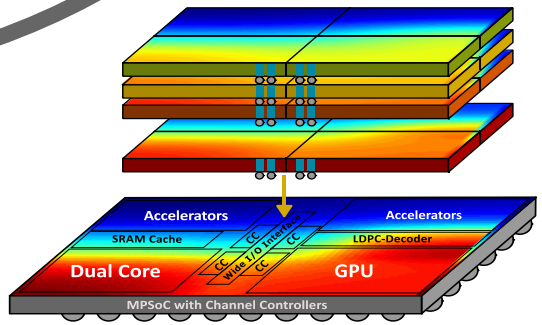
DRAM: DRAMPower

Refresh vs. Errors vs. Power vs. Performance

Impact on the Application

Thermal Analysis

3D



Partner Universities

UNIVERSITA' DI BOLOGNA · TU/e Eindhoven University of Technology · TU Delft Delft University of Technology

Industrial Partners

SYNOPSYS Synopsys Inc. · ARM ARM Limited · XILINX Xilinx Inc.

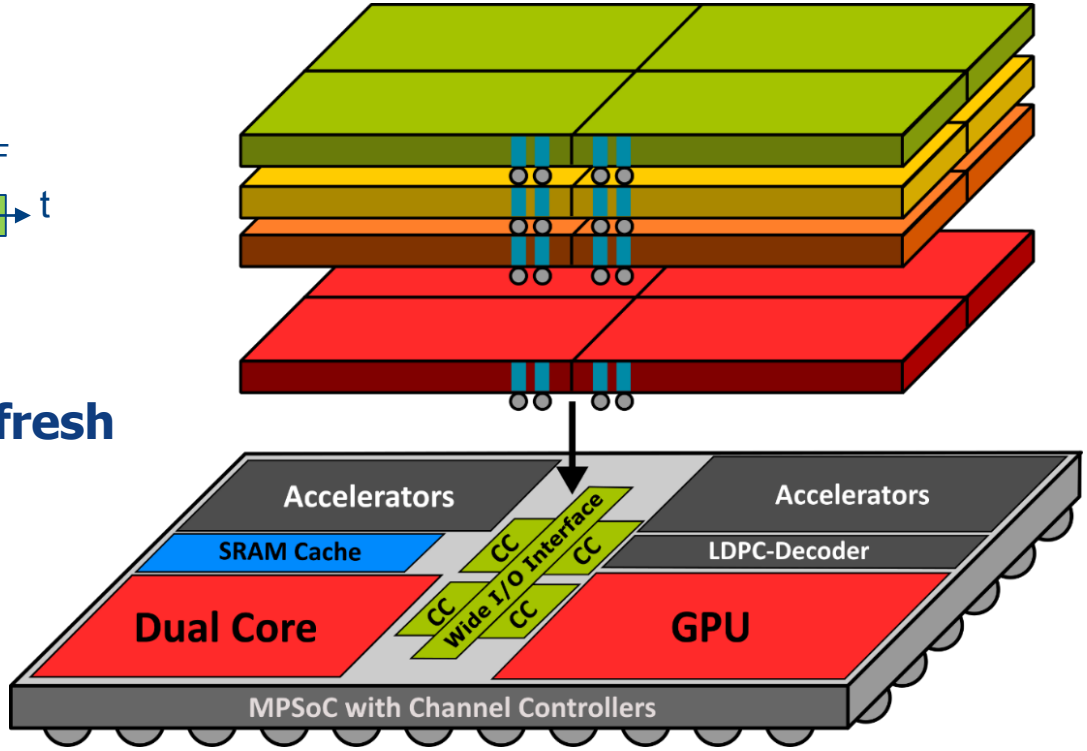
Research Institutes

Leti CEA Leti

Refresh in Stacked MPSoC

- Temperature = 100°C: refresh period every 8 ms
- Refresh command issued every: 488ns
- Refresh duration: 130ns

25 % of time spend in refresh

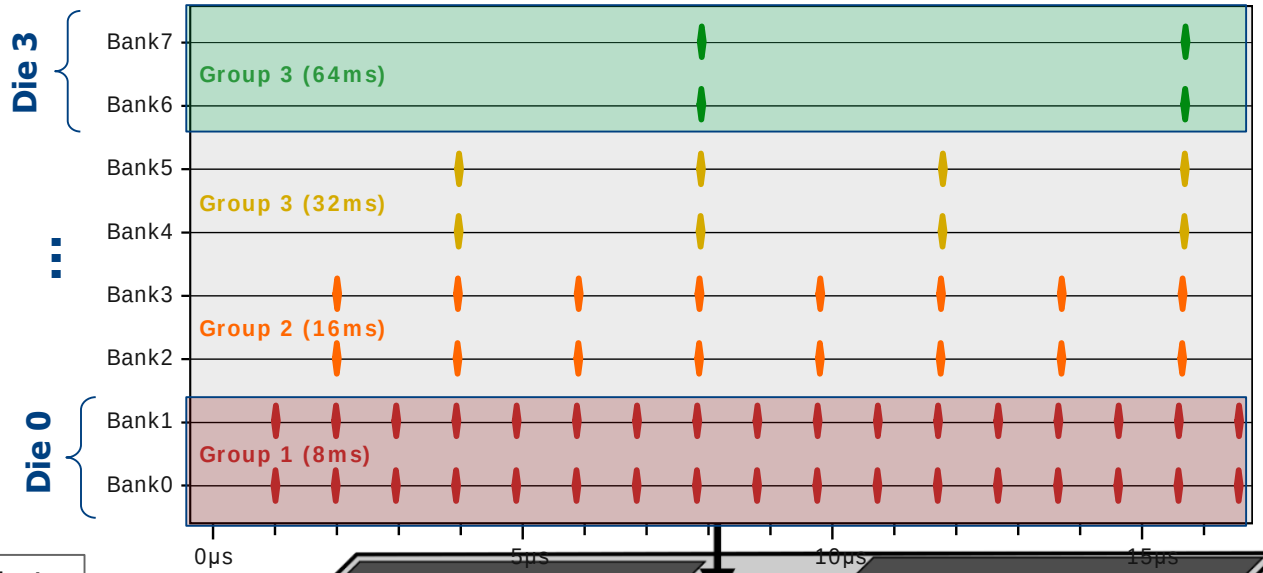


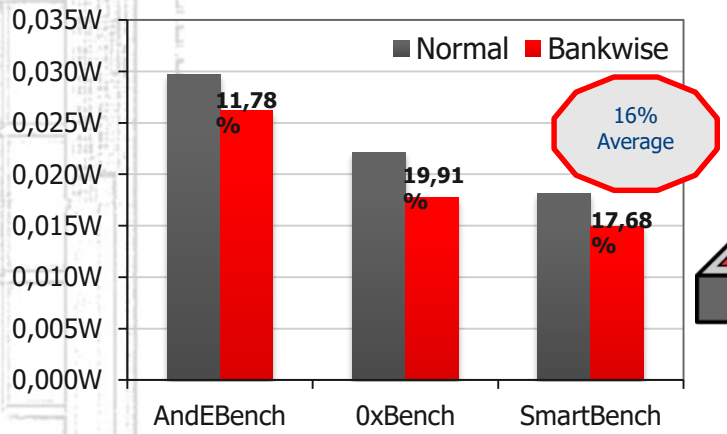
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

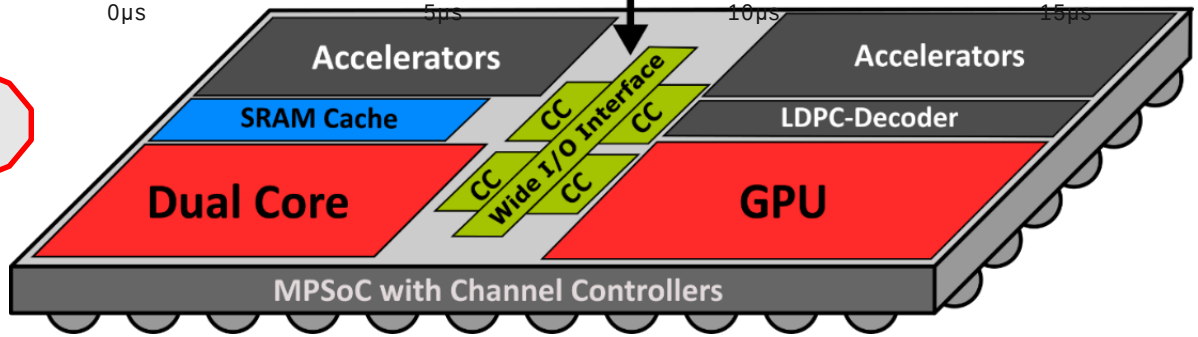
Temperature-Aware Bankwise Refresh

Using **Temperature-aware Refresh** in the DRAM stack

Different refresh rates on different **dice**, according to the their temperatures (individual for each die)







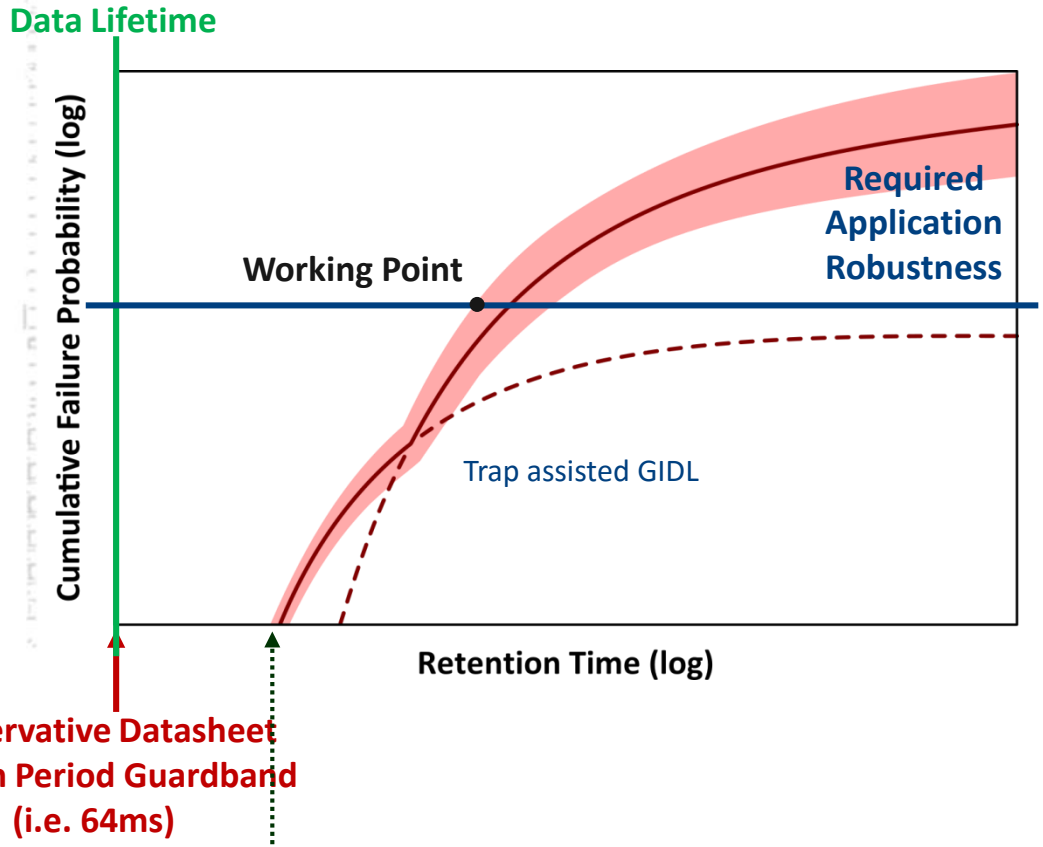
M. Sadri, M. Jung, C. Weis, N. Wehn and L. Benini, "Energy optimization in 3D MPSoCs with Wide-I/O DRAM using temperature variation aware bank-wise refresh, DATE'14

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Approximate DRAM

Lowering the refresh-rate or completely **switching off refresh**

- Consider DRAM device as a stochastic model that includes process variations



Switch Off Refresh

- If data lifetime is smaller than required refresh period
- If data lifetime is larger than required refresh period AND application has some robustness w.r.t. errors

Frame-buffer (GPUs)

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Cross Layer Refresh Policy

Separation of DRAM Stack into **unreliable** and **reliable** regions

- Unreliable region: bottom DRAM layer with disabled refresh
- Reliable regions: higher DRAM layers with temperature aware refresh
- Access unreliable region while reliable region is refreshed

Data stored on unreliable layers

- Data lifetime < refresh period
- Applications with inherent error resilience

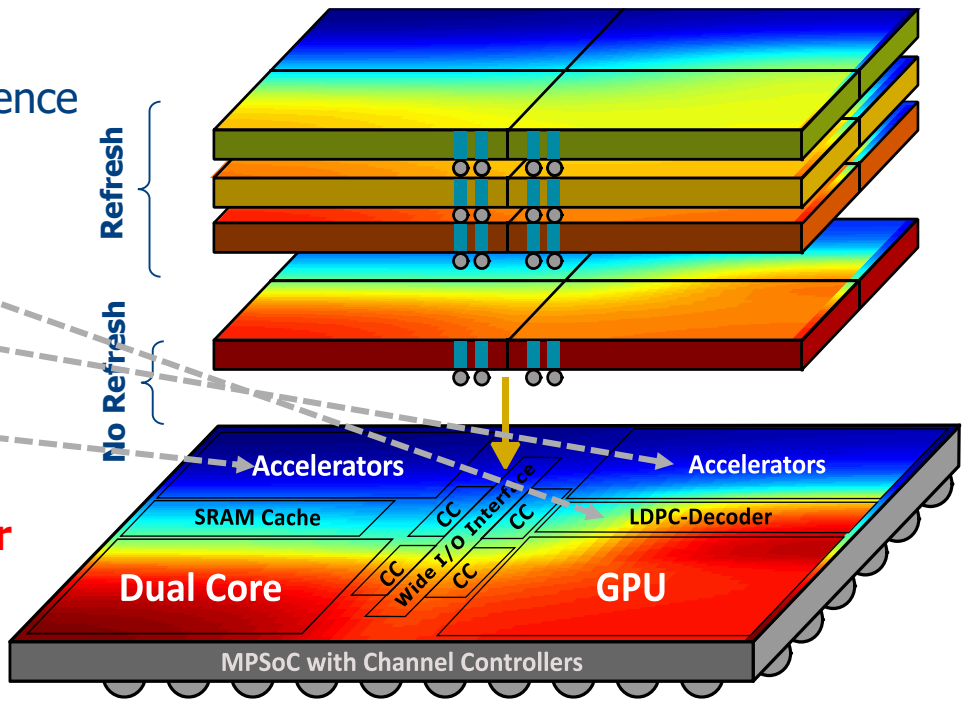
Investigated different applications

- Baseband processing (LDPC)
- Graph processing/Link assessment
- Image processing

⇒ **Saves 100% refresh power/layer**

⇒ **Increases bandwidth**

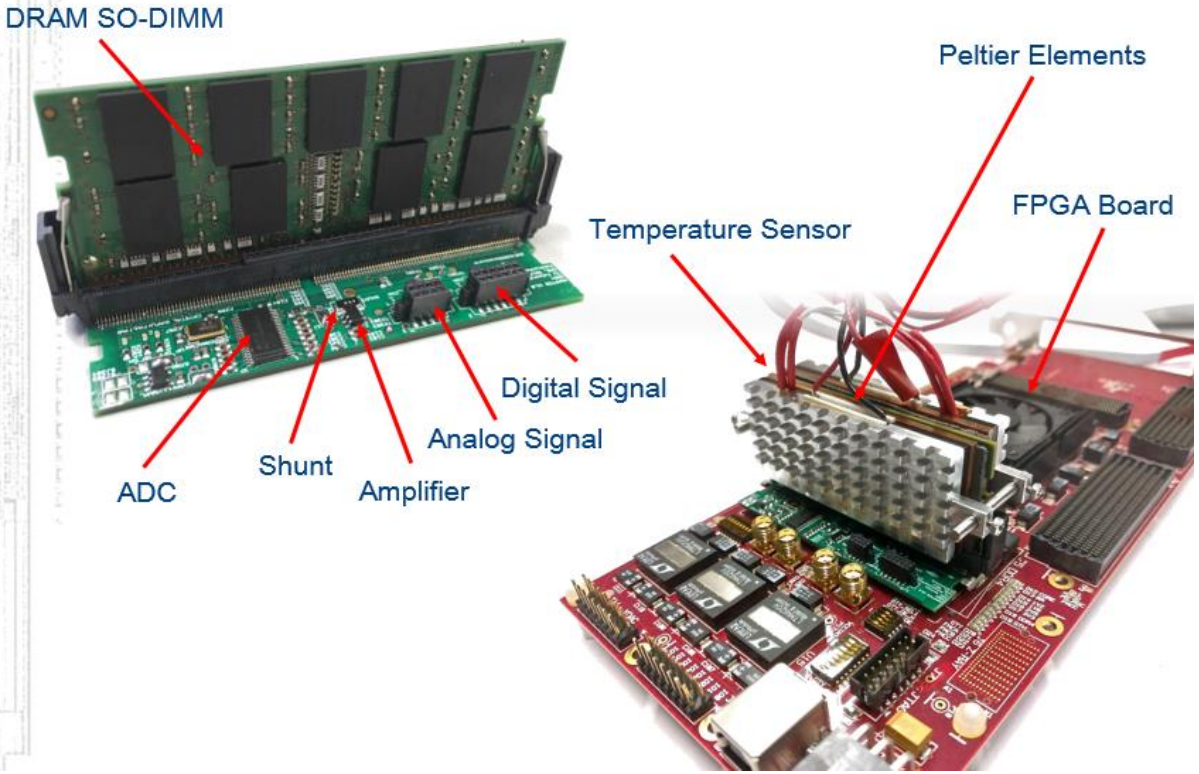
Simulation results



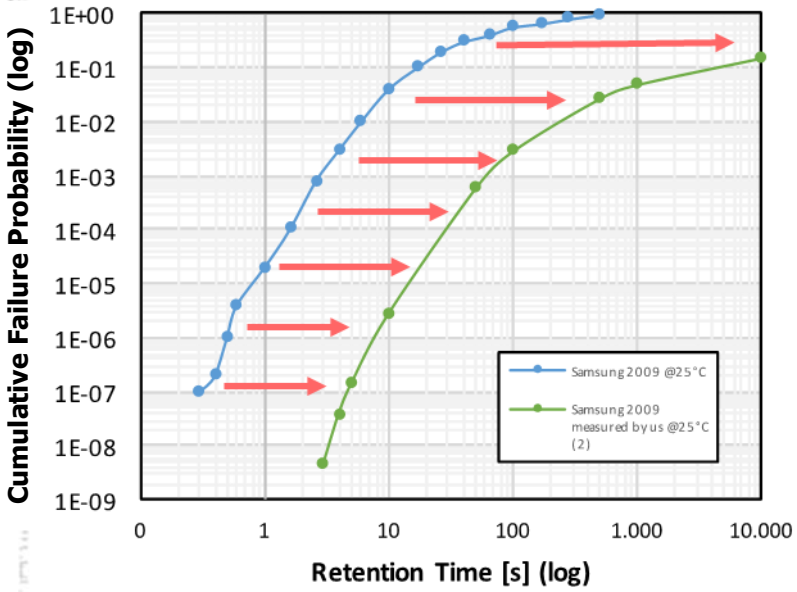
Matthias Jung, Éder Zulian, Deepak M. Mathew, Matthias Herrmann, Christian Brugger, Christian Weis, and Norbert Wehn. 2015. Omitting Refresh: A Case Study for Commodity and Wide I/O DRAMs MEMSYS '15

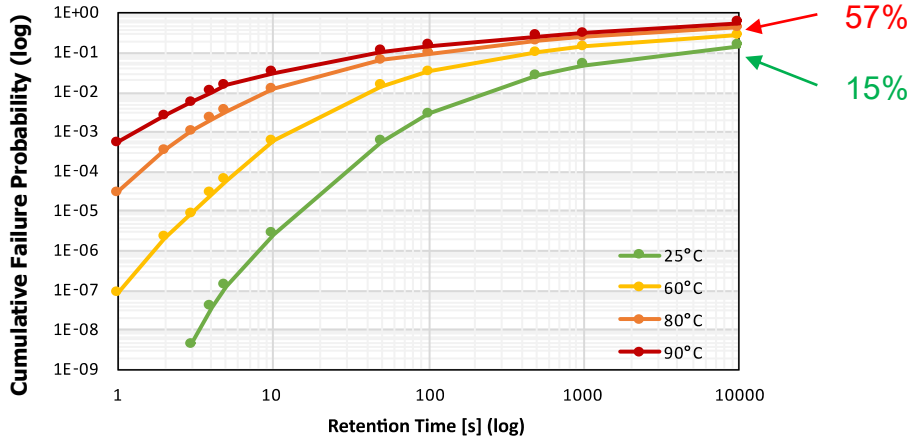
Commodity DRAMs - Measurement Platform

- Precise control of temperature (heating and cooling) of DRAM SO-DIMMs with JEDEC conform adapter
- Temperature range: 25-90°C, accuracy +/-2°C
- Precise current measuring for V_{dd} (1.2V), V_{PP} (2.5V) domain, +/-0.5mA accuracy
- Interface frequency up to 1.2 GHz



Retention Time Measurements[1]





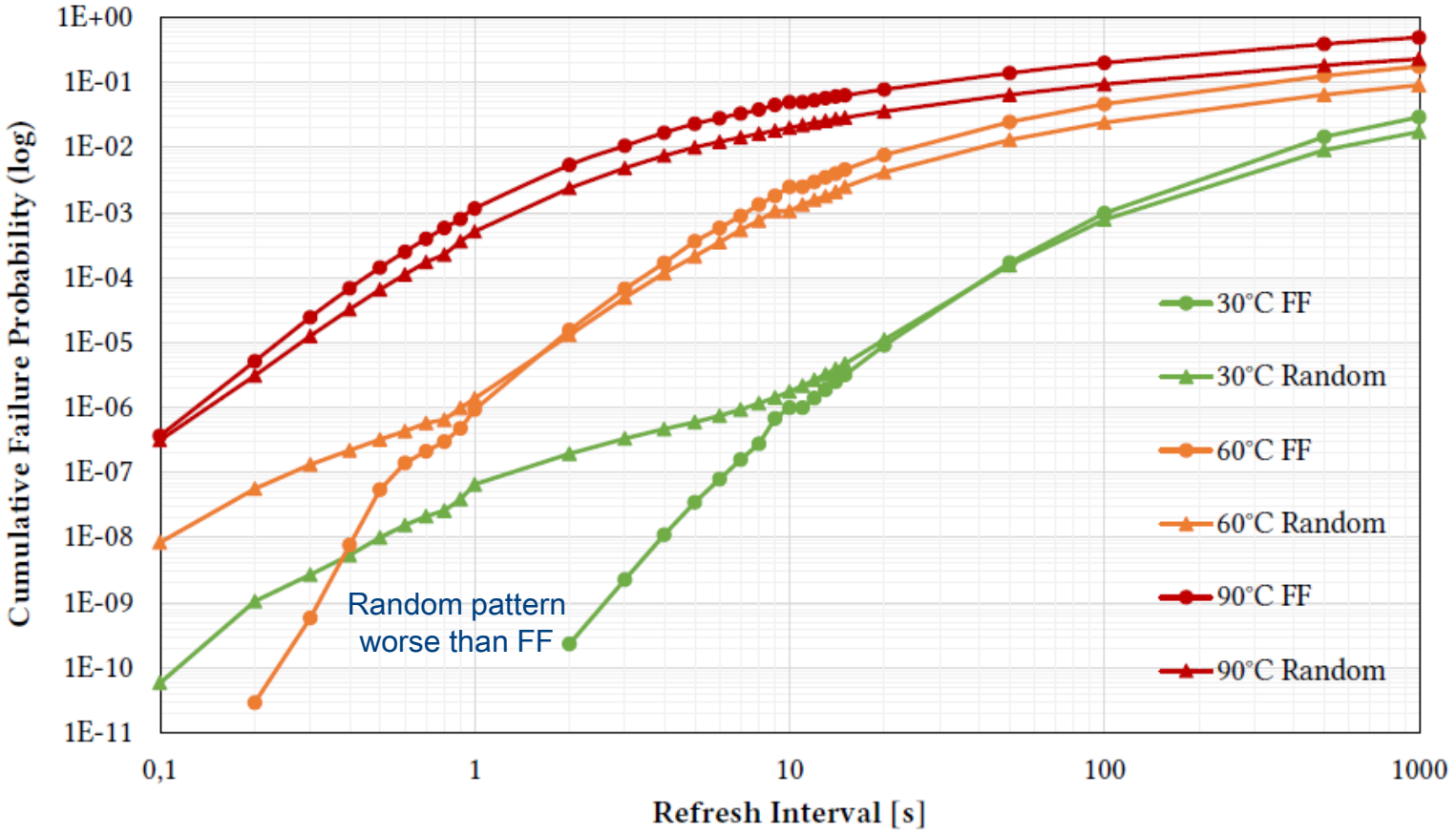
- Main reference in literature about retention errors published by Samsung[2]
- Measurements: 1-3 orders of magnitude better retention error behaviour
- DRAM can hold data much longer than specified, even at high temperatures

[1] Values normalized to total DDR3 DRAM size: 512 MB (Total number of DRAM cells: 4.294.967.296)

[2] Kim and Lee, *A New Investigation of Data Retention Time in Truly Nanoscaled DRAMs*, 2009

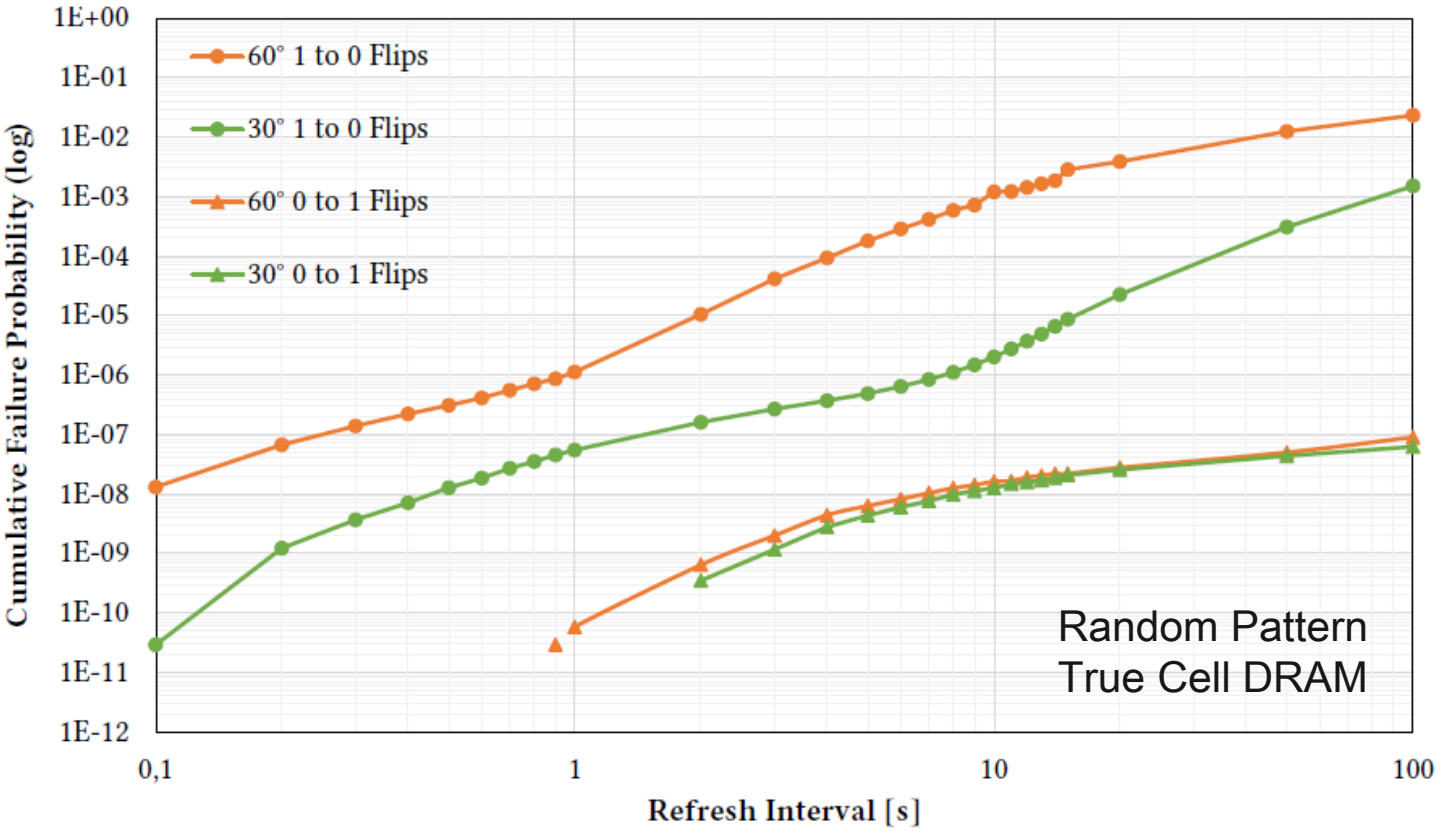
DDR4 Retention Time Measurements I

Retention behavior depends on cell leakage, cross talk, process variations, temperature, cell type, data pattern



DDR4 Retention Time Measurements II

Unsymmetrical error behavior dependent on cell type (true-cell, anti-cell)



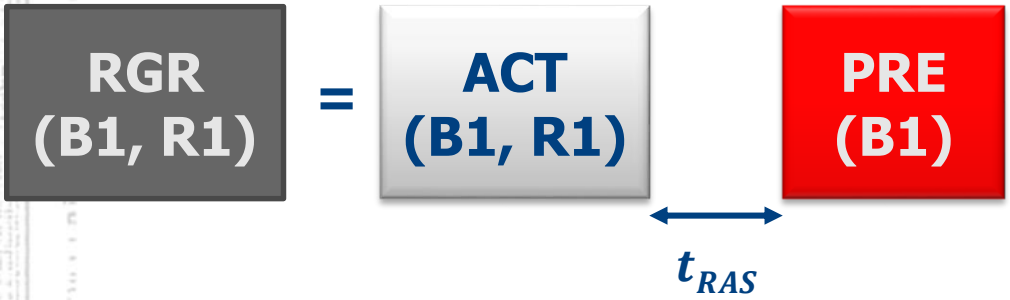
1 ⇒ 0 flip much more likely than 0 ⇒ 1 flip

Drawbacks of Auto-Refresh

- AREF lacks flexibility
- No access to internal refresh row counter
- No rows can't be skipped
- The complete DRAM has to be refreshed in the same rate

- Sele[illegible]efresh
- Reten[illegible]Aware Refresh
- Appro[illegible]DRAM

Row-Granular Refresh (RGR)

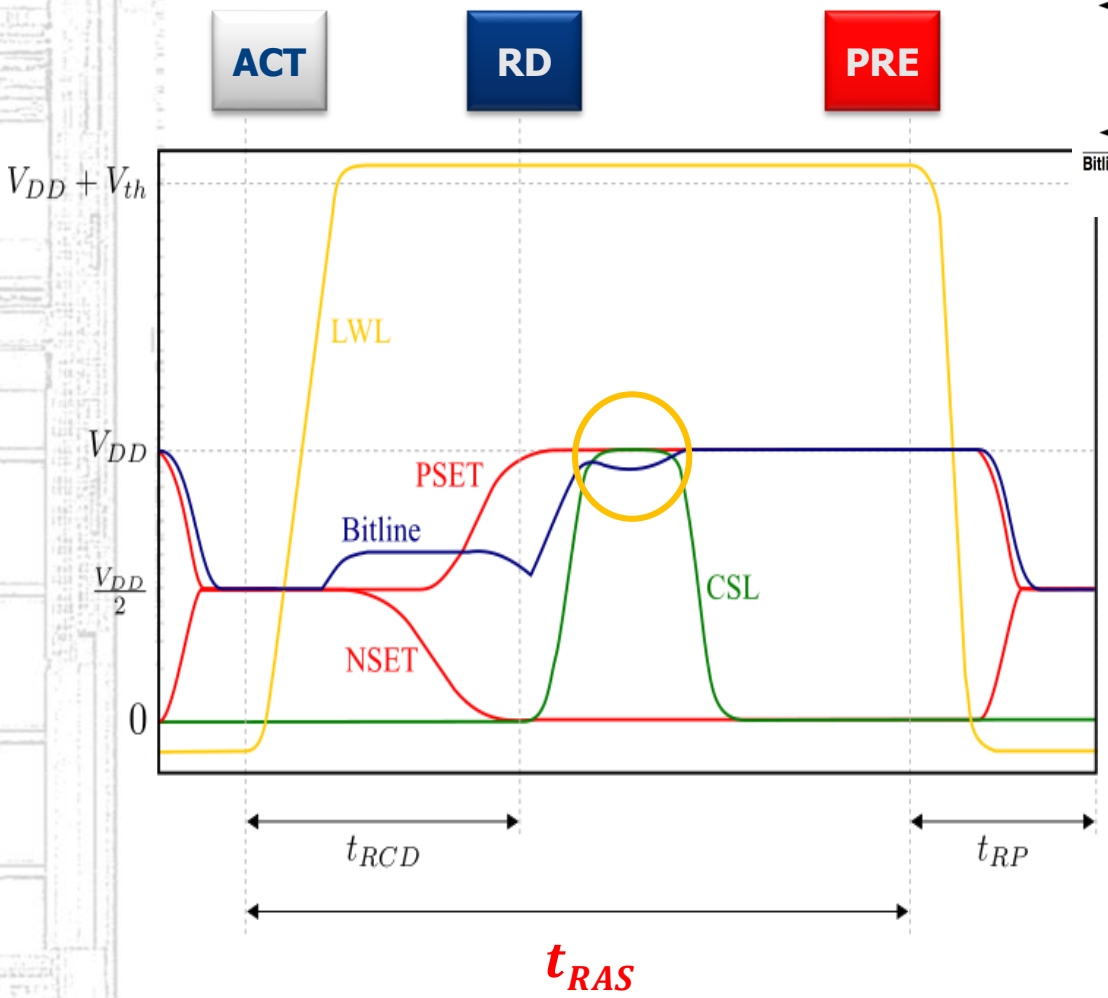


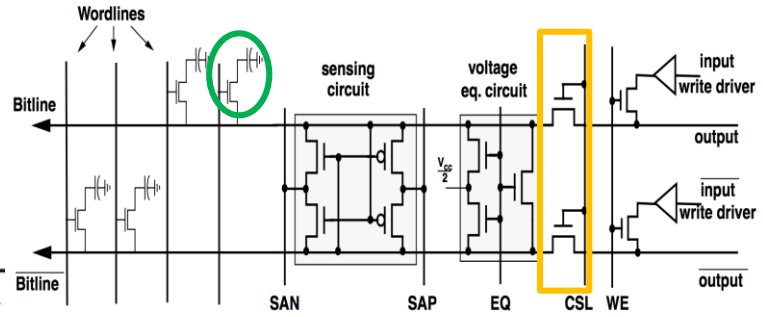
- t_{RAS} limits the single refresh execution time
- Violate t_{RAS} ??

- Mimics AREF by sending ACT and PRE commands to specific rows
- Performance and Energy savings by skipping unncessary refreshes
- Follow strict **JEDEC timings** for t_{RAS}
- Proven to be inefficient compared to AREF **

***Ishwar Bhati, et al. Flexible Auto-Refresh, ISCA 2015*

DRAM Sensing



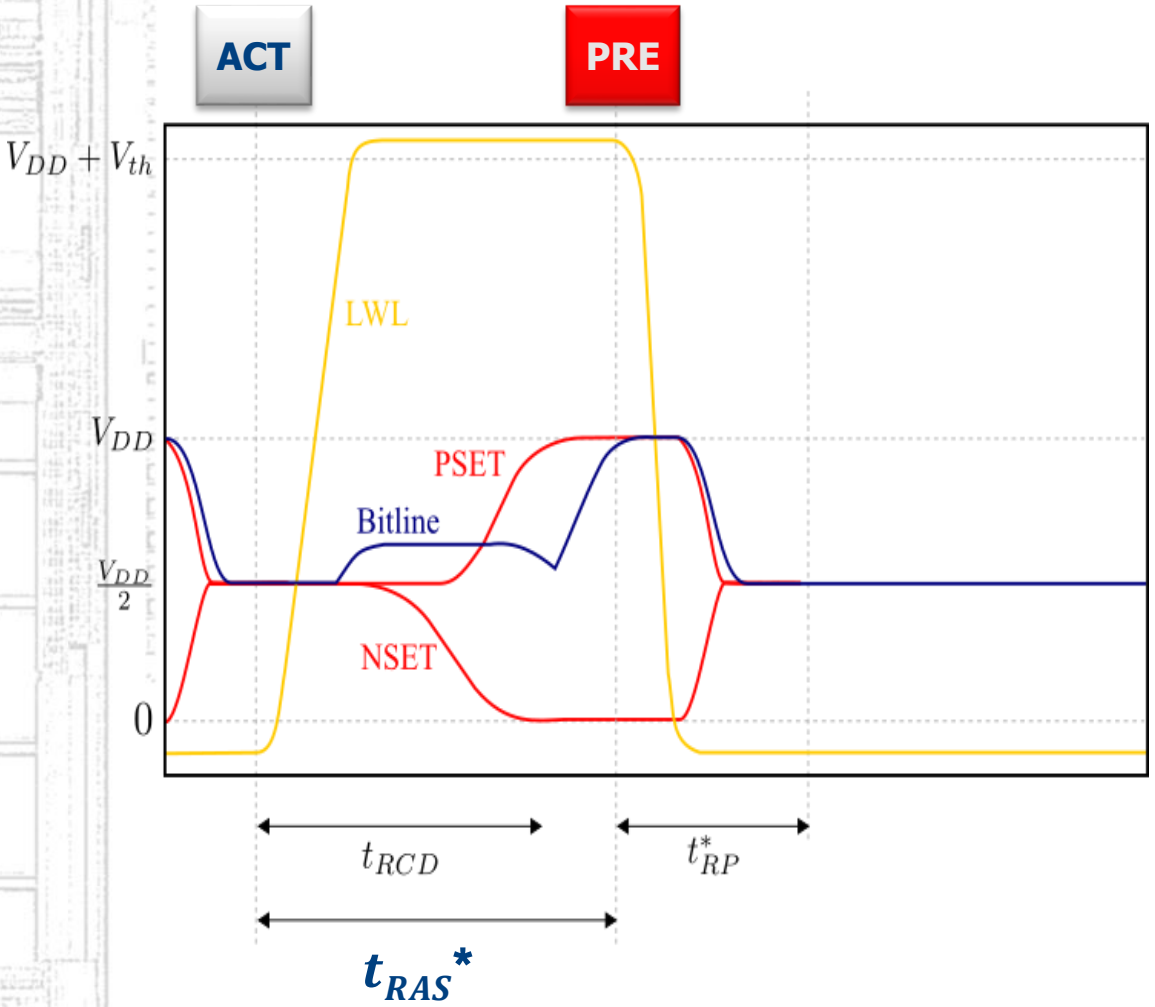


- Large t_{RAS} timing to compensate for **bitline** voltage drop (blue signal)

Patent is granted!

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

DRAM Sensing: Refresh

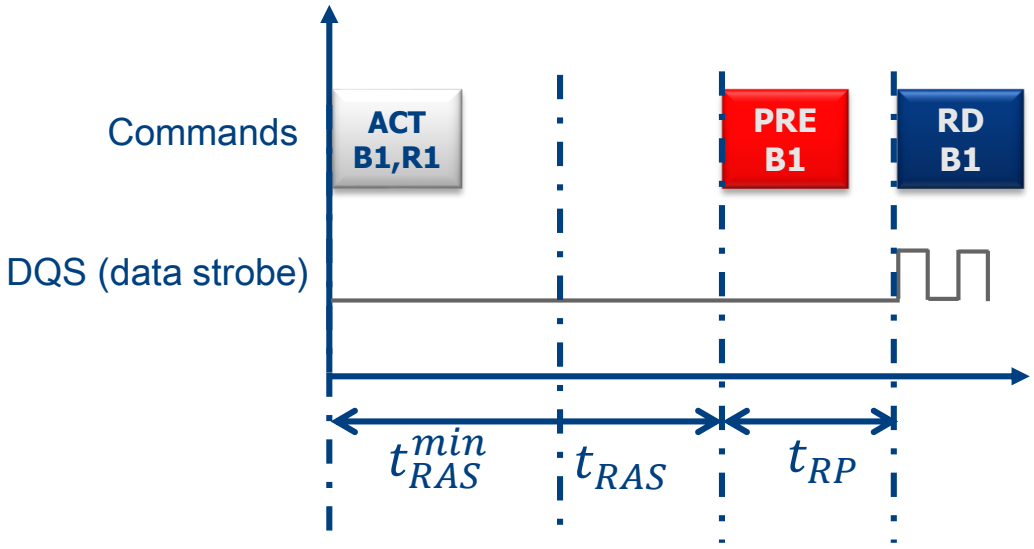


- No **voltage drop** due to Read/ Write. Therefore t_{RAS} can be reduced for performing RGR
- But, how much reduction possible?
- We found **two** different methods for 3 vendors to measure t_{RASmin}

Patent is granted!

Minimum t_{RAS} for Vendor A

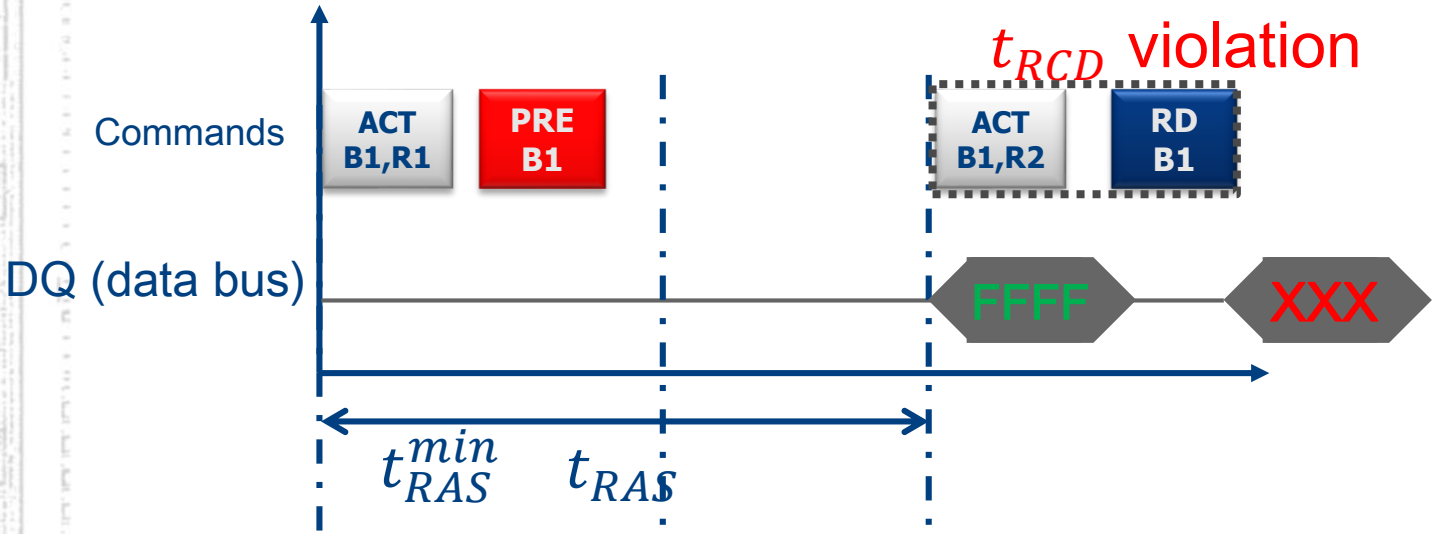
- $t_{RAS\ min}$ **timer** is present for additional saftey and for auto-precharge operations (RDA/WRA)
- Vendor specific implementations
- Reverse engineering technique performed during DRAM initialization, or during normal operation



t_{RAS} = 37.5 ns
t_{RAS}^{min} = 20.7 ns

Patent is granted!

Minimum t_{RAS} for Vendor B & C



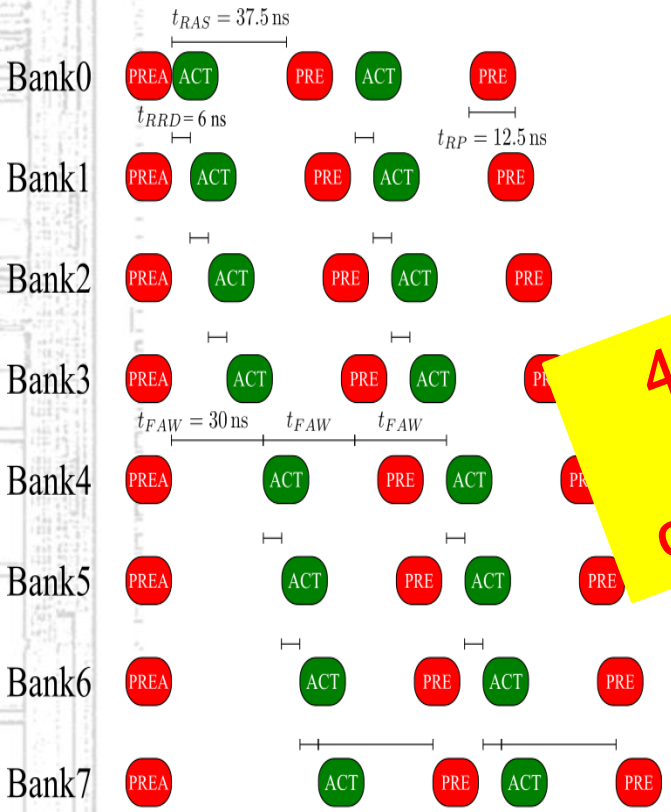
t_{RAS} = 37.5 ns

t_{RAS}^{min} = 22 ns

Patent is granted!

RGR vs ORGR (2Gb x16 DDR3)

RGR



t_{RFC} = 158 ns

ORGR

t_{RFC} = 86.25 ns

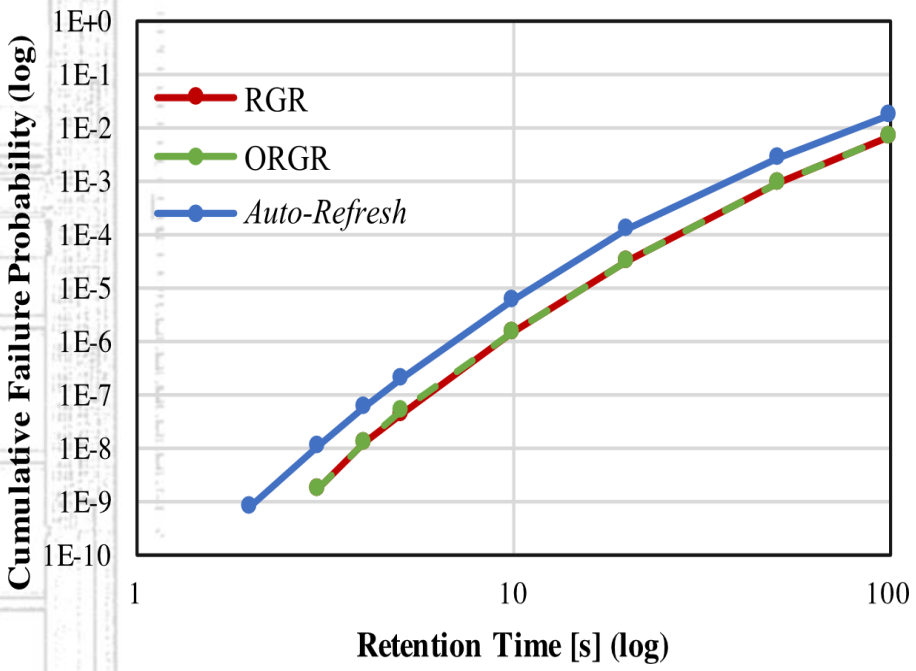
45% Savings in t_{RFC} when the complete DRAM is refreshed

Patent Pending

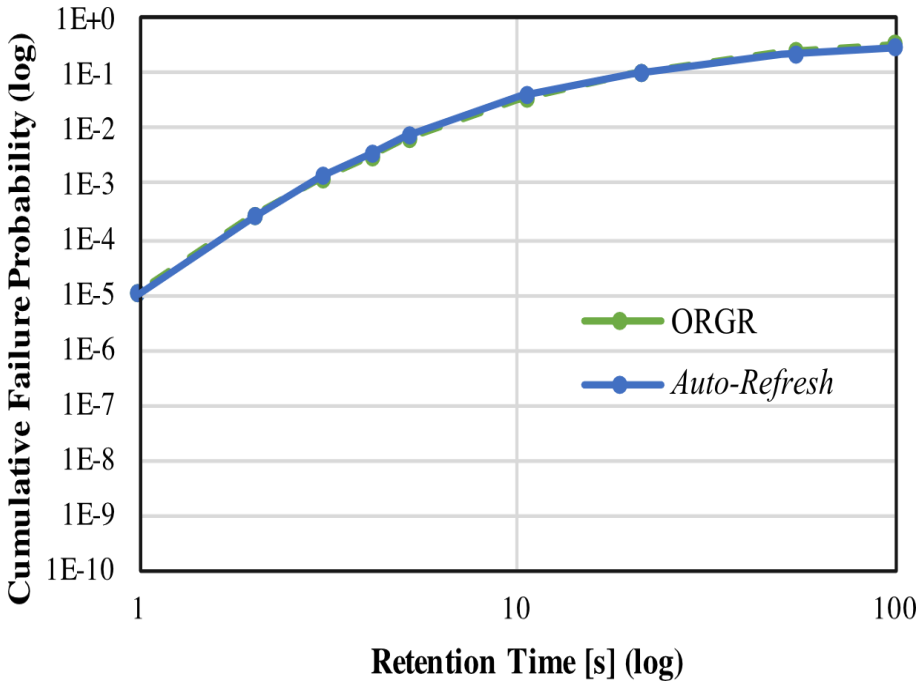
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Effect/Impact on the Retention Behavior

T=30°C



T=90°C



MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Performance and Energy Savings

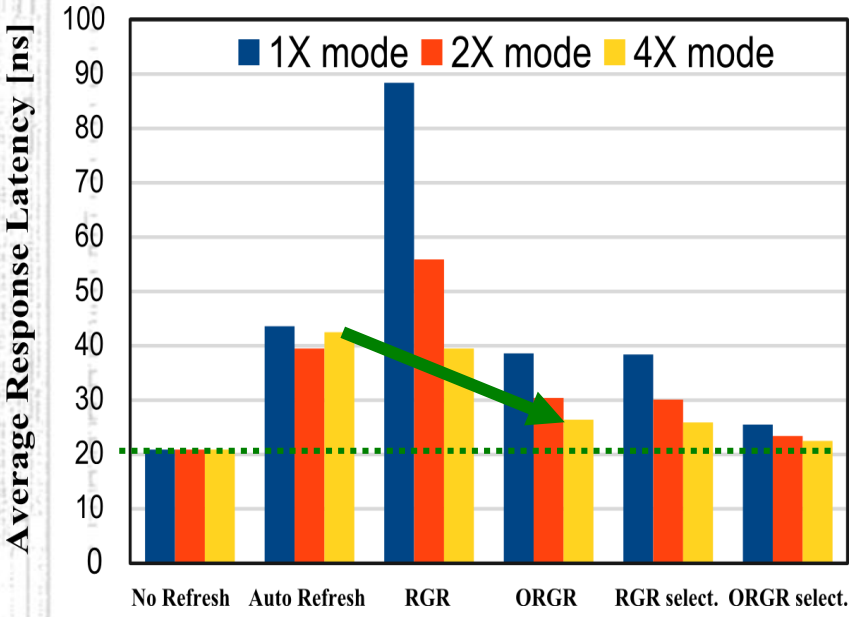
Refresh Technique	t_{RFC}	Refresh Energy
Auto Refresh	262.5	186.24
RGR	292.5	230.48
ORGR	146.25	209.72

- Measured for **4Gb** x16 DDR3 DRAM from Vendor A
- Refreshing the complete DRAM

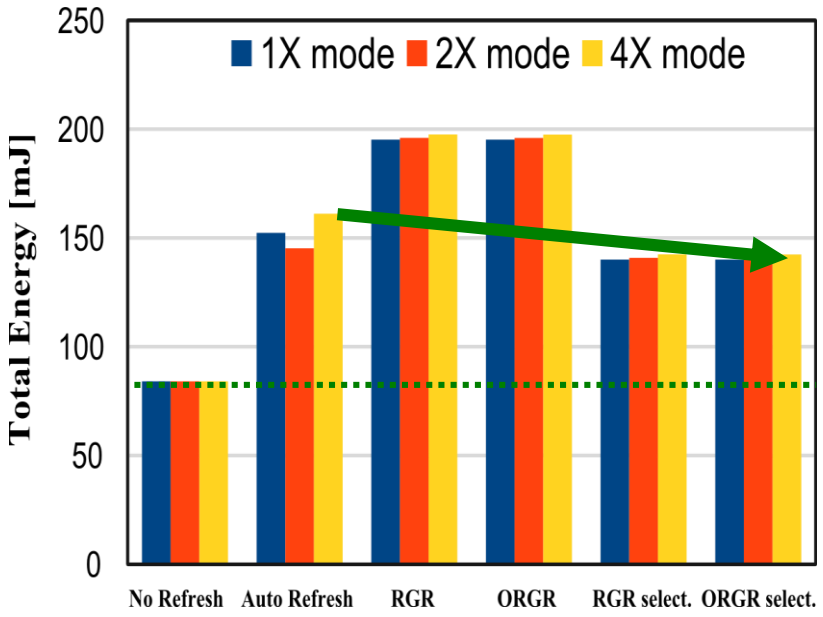
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Simulation Results for a 16Gb DRAM

Performance



Energy



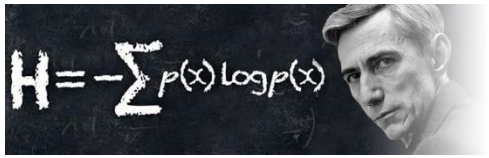
Application with a Sparse Access Pattern using AREF and ORGR

Content

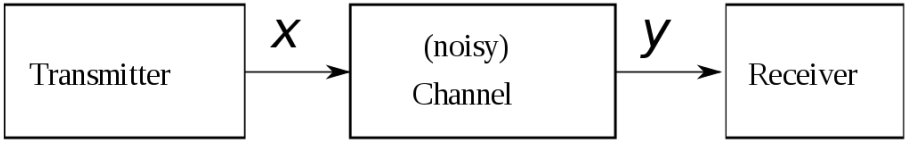
- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- **Improvement of Reliability**
- Power Down Modes
- The Memory Controller

Asymmetric Retention Behavior

Can we exploit asymmetry to increase DRAM's reliability



- Consider memory as noisy communication channel



- Theoretical measure for the quality of a noisy channel: channel capacity C

$$C = I(X;Y) = H(X) - H(X|Y)$$

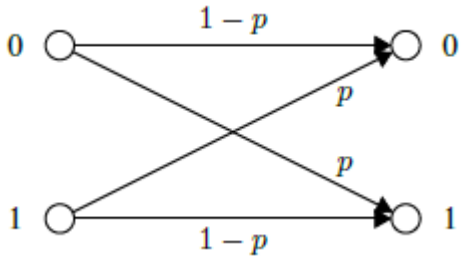
with H(X) being the entropy of a random variable X e.g. binary entropy function

$$H(p) = -p\ log_2 p\ - (1-p) log_2 (1-p)$$

- C is highest information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability

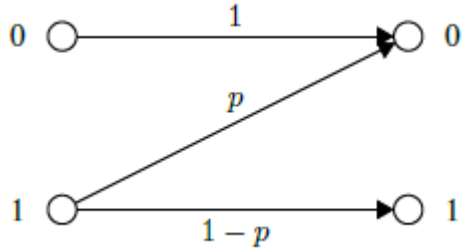
Information Theory

Symmetric retention behavior:
Binary Symmetric Channel (BSC)



$$C_{BSC} = 1 - H(p)$$

Asymmetric retention behavior:
Z-Channel



$$C_Z = \log_2\left(1 + (1-p)\cdot p^{\frac{p}{1-p}}\right) \approx 1 - \frac{1}{2}H(p)$$

Larger Channel Capacity

- Larger reliability if internal cell structure (true-cell, anti-cell) is known
- More efficient ECC techniques possible
- Appropriate data representation: e.g. small dynamic range C2 versus sign/magnitude

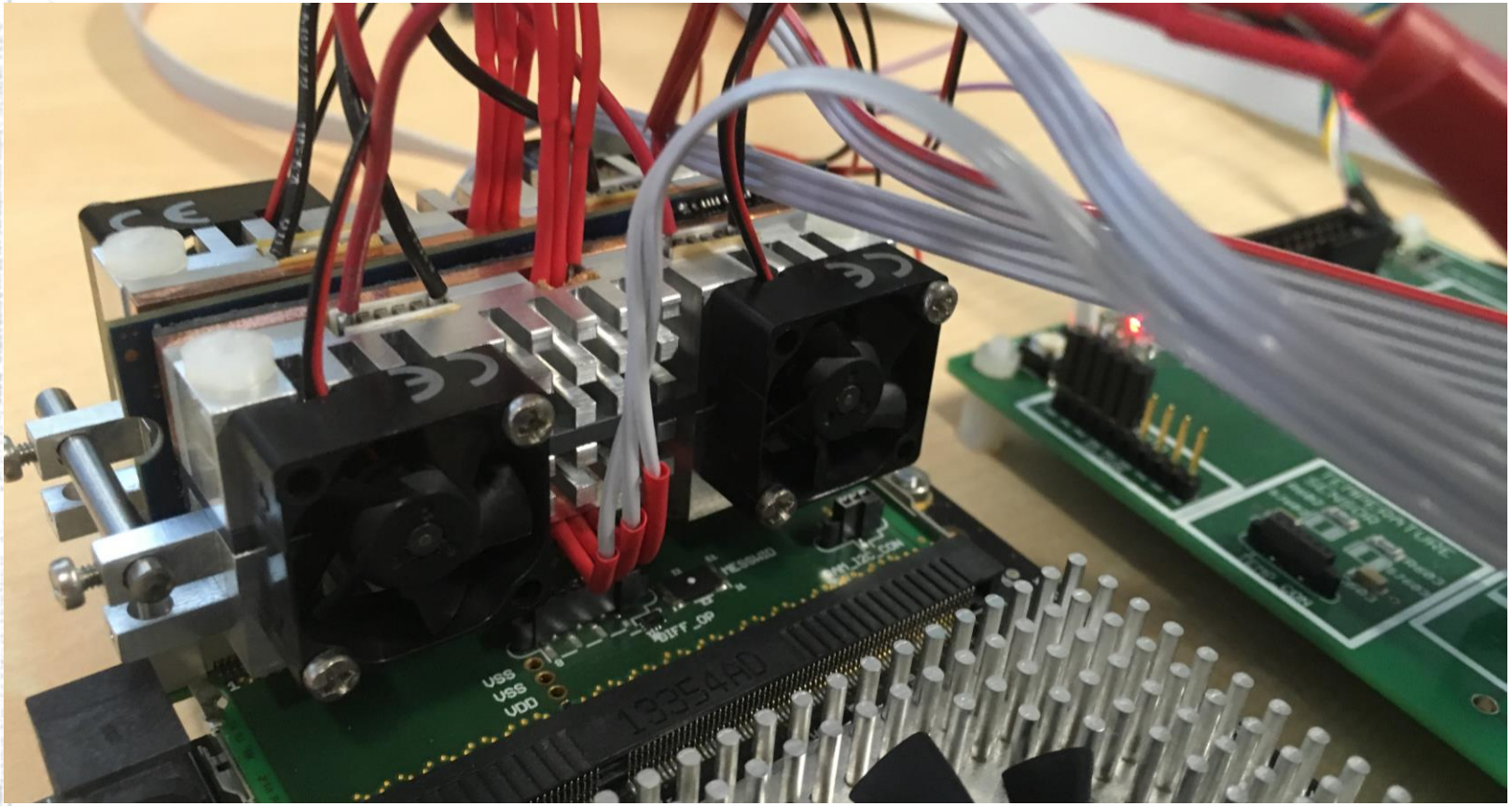
Challenge

- Internal structure normally not known (vendor secret)

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Reverse Engineering

- Fill DRAM with 1s
- Disable refresh for 20 hours, additionally heat up to 80ºC



Vendor Specific Storage

DRAM with true-cells only

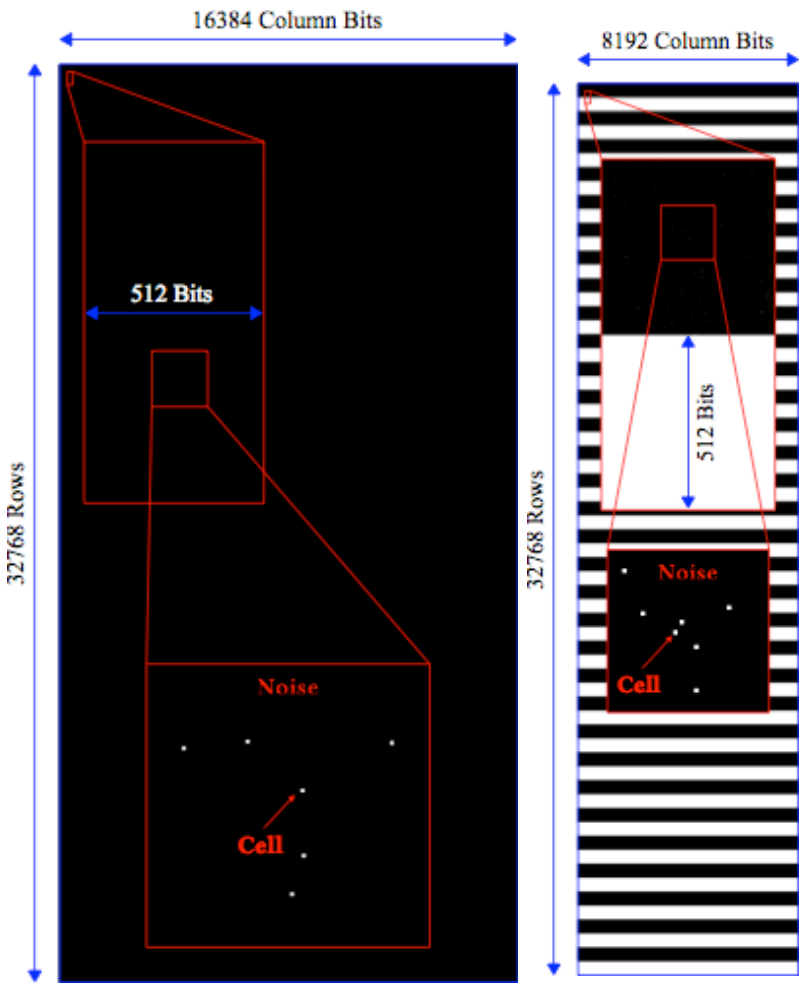
- Stored 1 leaks
- Bitmap is black

DRAM with anti-cells only

- Stored 1 can not leak since it's stored as 0
- Bitmap is white

DRAM with mixed-cells

- Bitmap is black/white

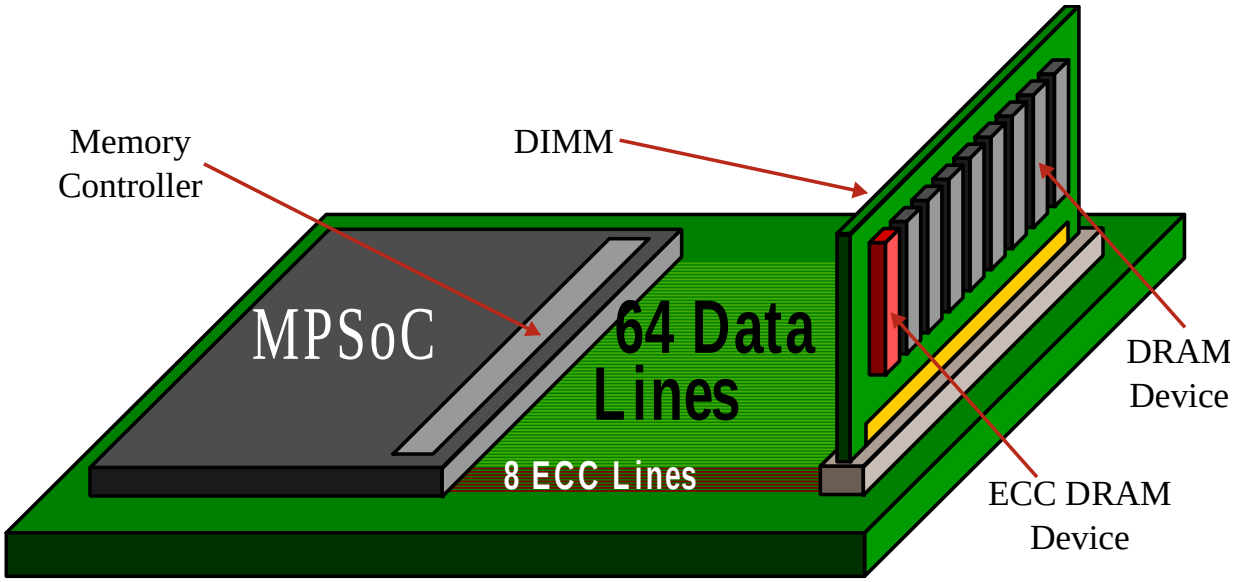


Increase DRAM's Reliability

Use of ECC techniques: e.g. (72, 64) SEC/DED hamming code

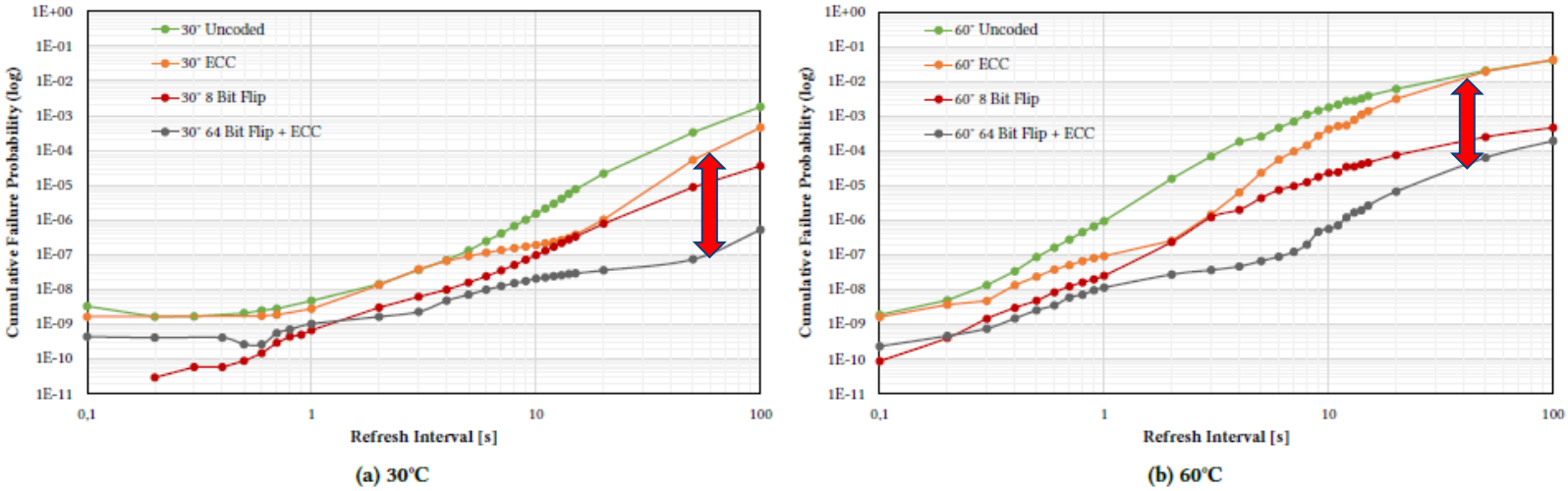
Simple but efficient technique to increase reliability:

- Example true-cell: flip data of DRAM burst if there are more ones than zeros
- Reduce Hamming to SEC Hamming (72,64) i.e. 64 bit data, 7 bit parity and 1 bit for storing flip information

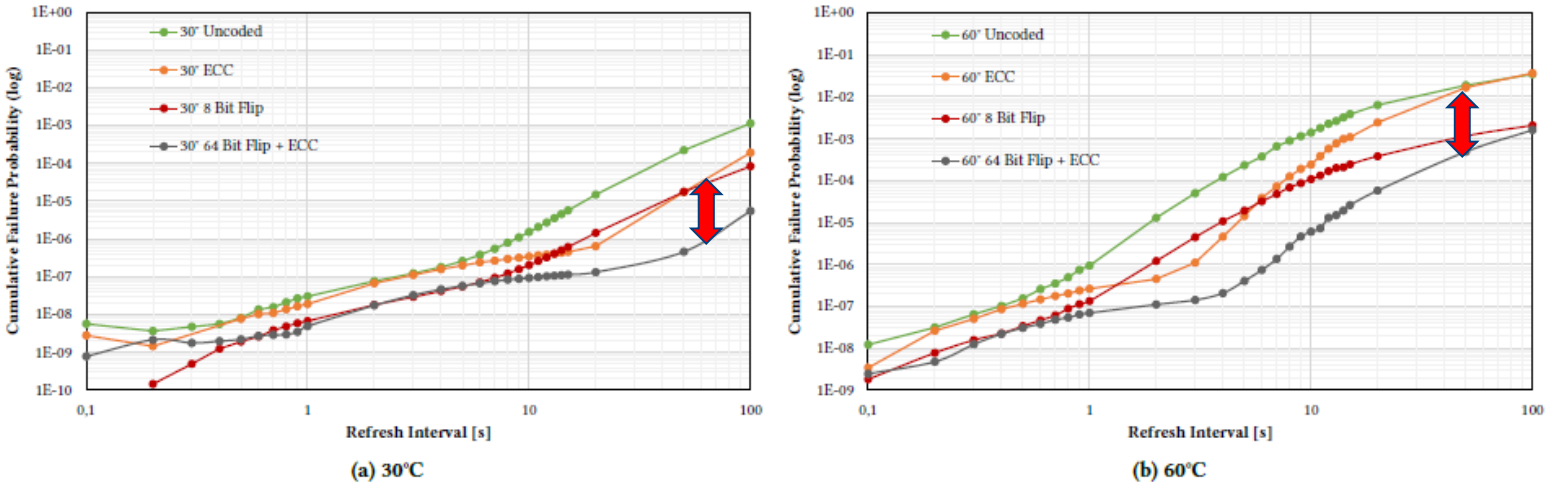


MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Results



64 bit: 8 bit dynamic range (2s complement)



64 bit: 16 bit dynamic range (2s complement)

Large improvement in retention behavior

Content

- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- **Power Down Modes**
- The Memory Controller

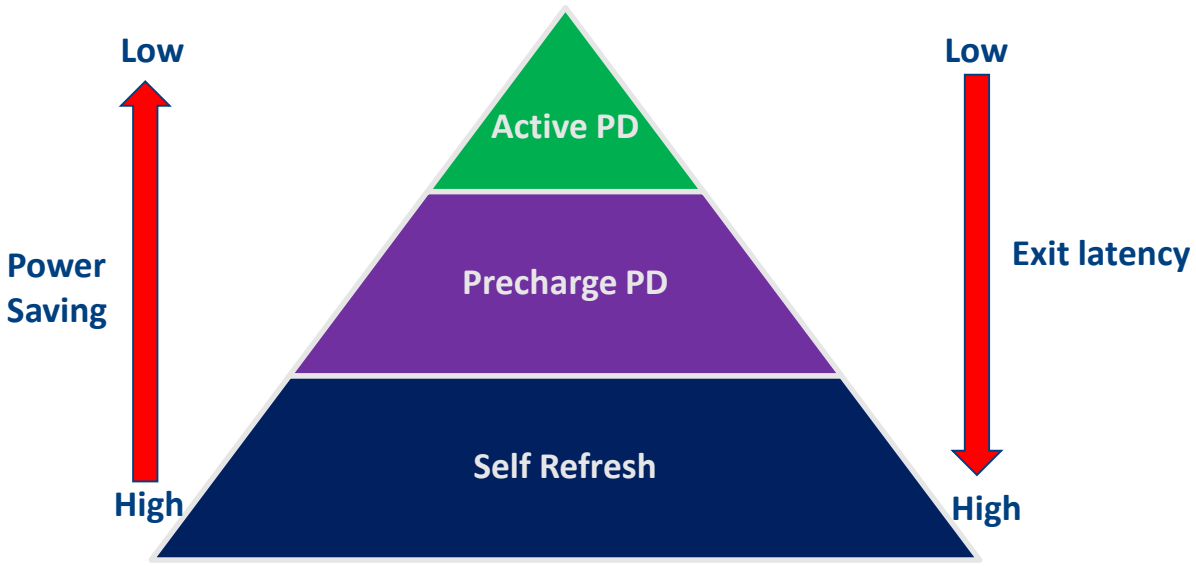
DRAM Power-Down Modes

Three power-down modes

- Active Powerdown (PDNA)
- Precharge Powerdown (PDNP)
- Self-Refresh (SREF)

When to enter which of these modes?

- E.g. Timeouts

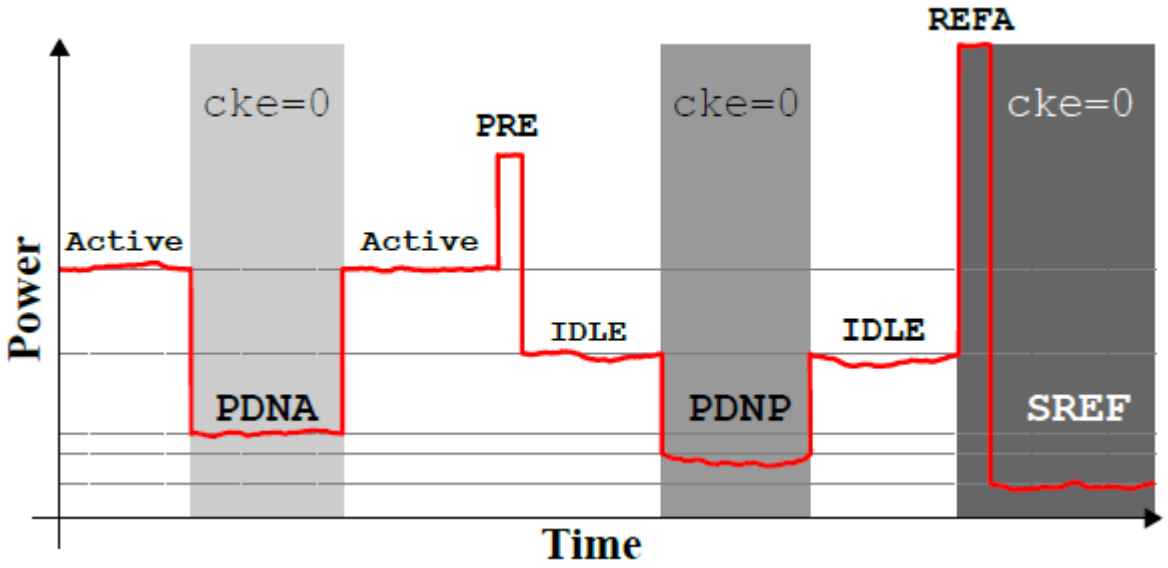


Power-Down Modes

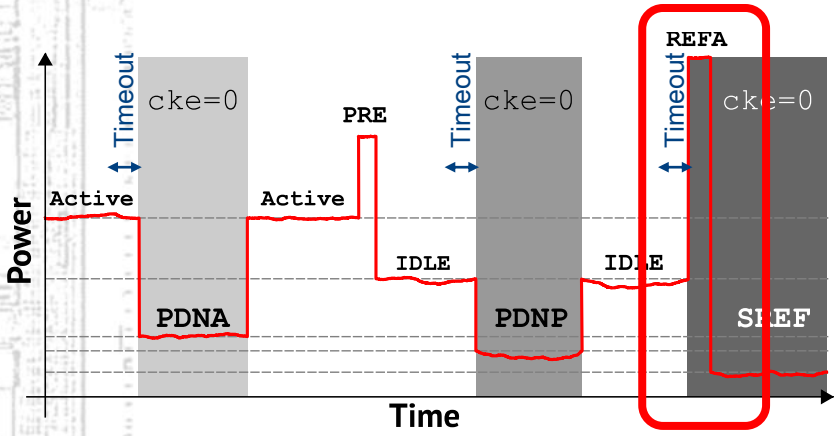
Active: at minimum one bank is active, no power down (cke=1), the controller has to schedule the refresh

Idle: all banks are closed and precharged, no power-down (cke=1), no internal refresh. The DRAM changes the state from Active to idle by issuing a precharge command

- Active Power-Down: at minimum one bank is active and no internal refresh
- Precharge Power-Down: all banks are closed and precharged and no internal refresh
- Self-Refresh: all banks are precharged and closed, the DRAM internal self-timed refresh is triggered



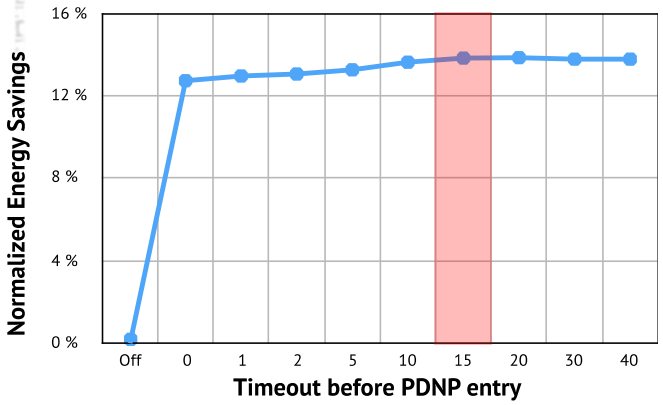
Power-Down Optimization

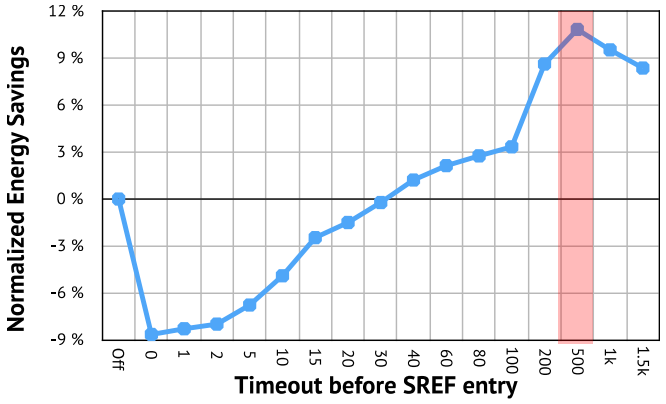


Timeouts are usually used to enter power down

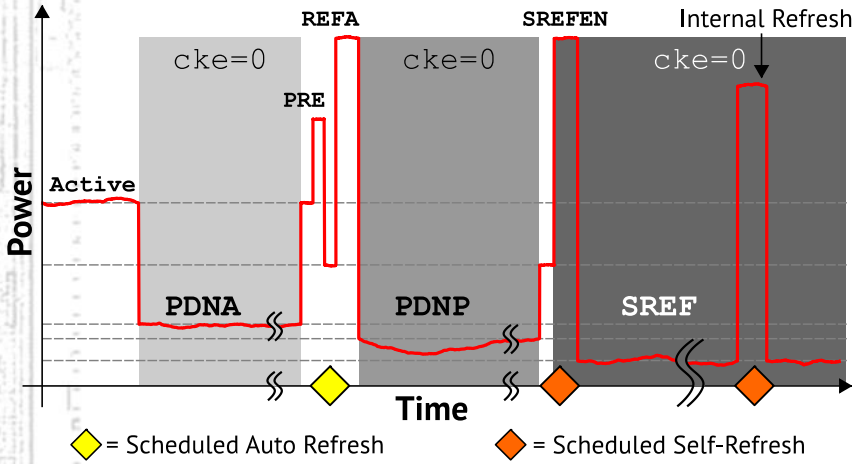
- SREF entry provokes refresh: high power
- Already observed by Schmidt [ICSD'09]

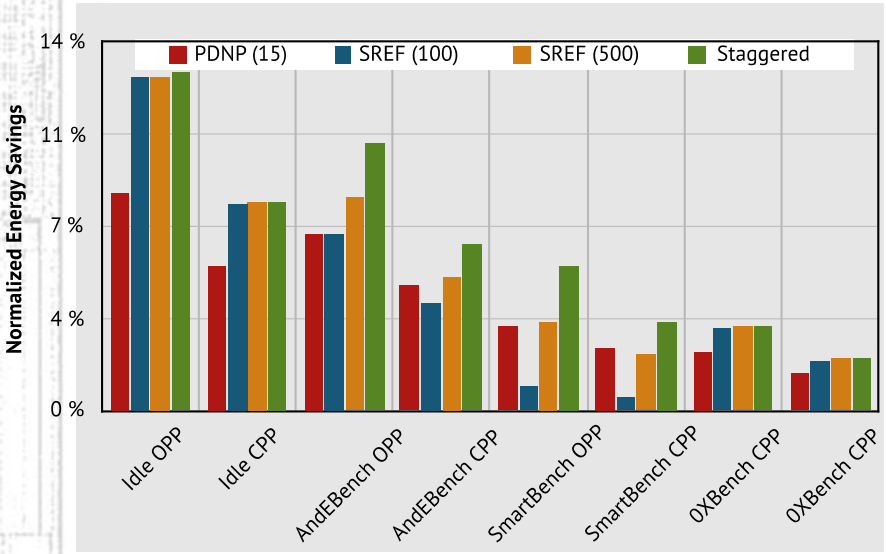
Avg. of 22 MediaBench and CHStone benchmarks





Power-Down Optimization





Staggered Power-Down[1]

- Instead of timers we use triggering of refresh
- Go to PDNA ASAP
- After a refresh go to the next better power down state (PDNP, SREF)

Up to 13% energy reduction

Works with today's DRAMs

Adopted by ARM

Bank-Wise Staggered Power-Down

- Each bank can go into power-down individually

Further improvement:

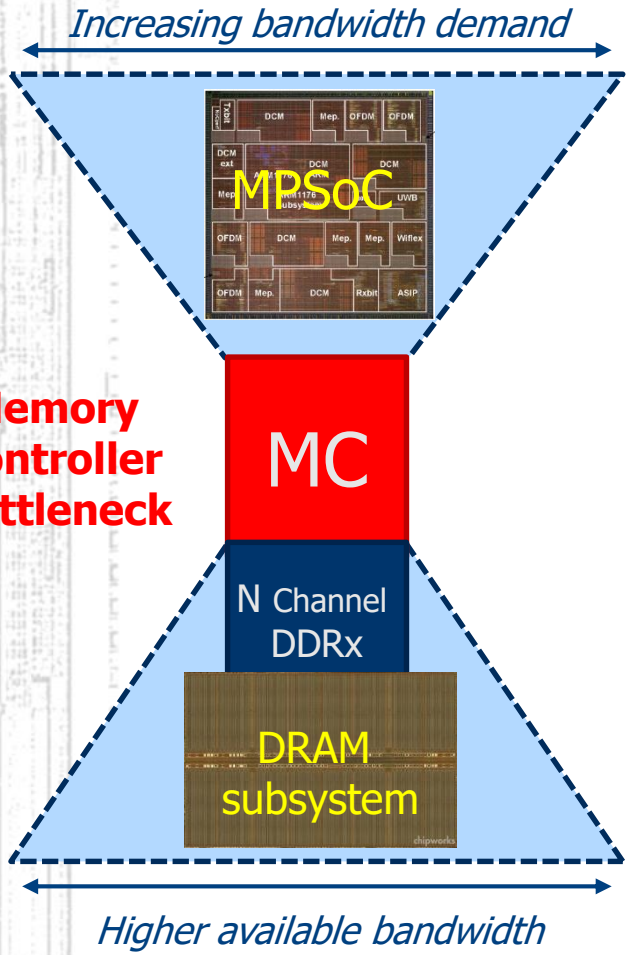
- -9% energy
- -10% latency

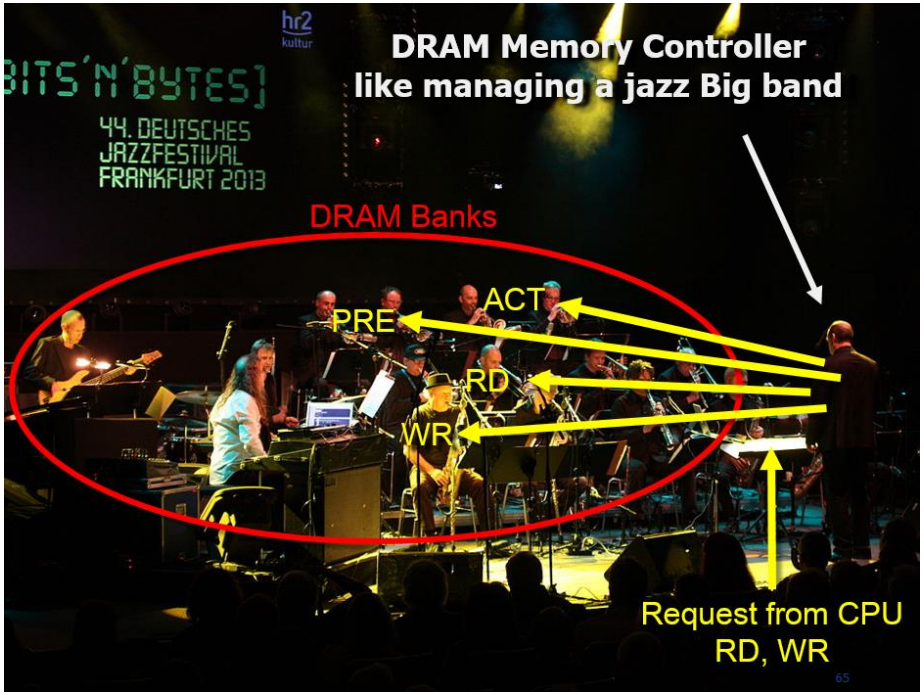
[1] M. Jung, C. Weis, N. Wehn, M. Sadri and L. Benini, "Optimized active and power-down mode refresh control in 3D-DRAMs," 2014 22nd International Conference on Very Large Scale Integration (VLSI-SoC)

Content

- The Transprecision Concept
- Overview on Memory Types
- The Importance/Challenges of Memories in Computing
- The DRAM Device
- The Refresh Challenge and Optimizations
- Improvement of Reliability
- Power Down Mode
- **The Memory Controller**

DRAM Memory Controller

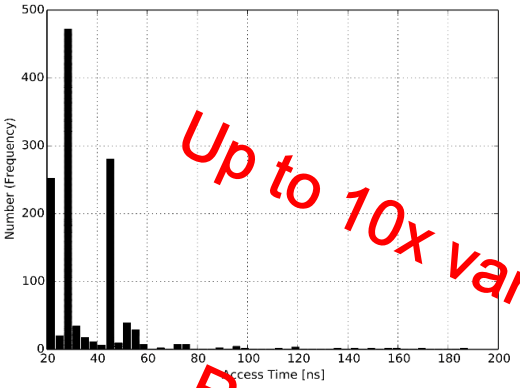




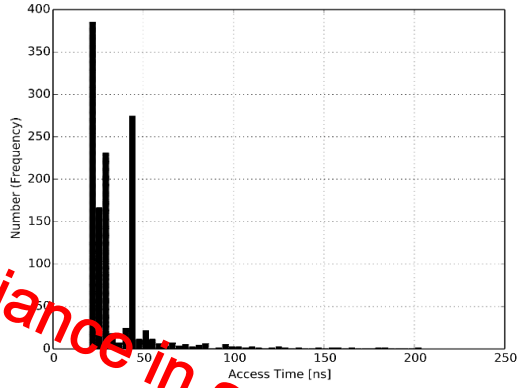
DRAM controllers must provide higher bandwidth and **lower latency**

DRAM Latency Variation

Chstone ADPCM / DDR3 / **BRC** / FCFS



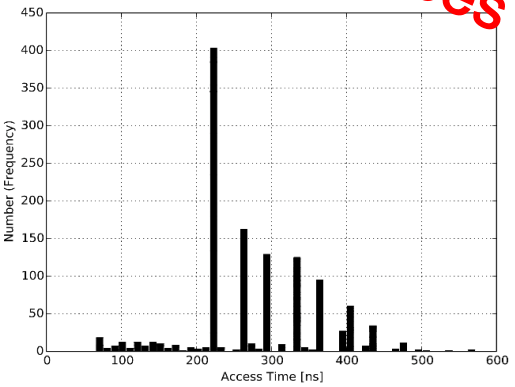
Chstone ADPCM / DDR3 / **RBC** /



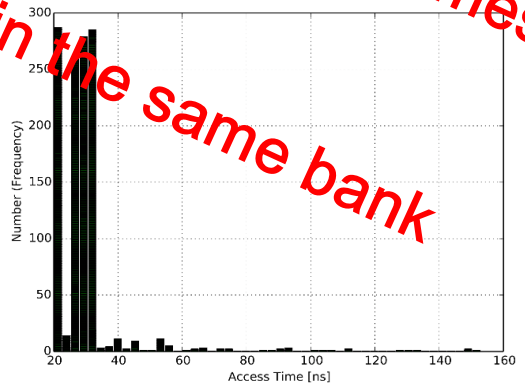
Up to 10x variance in access times

Row misses in the same bank

Chstone ADPCM / **WIDE IO** / BRC /



Chstone GSM / DDR3 / **RBC** / FCFS

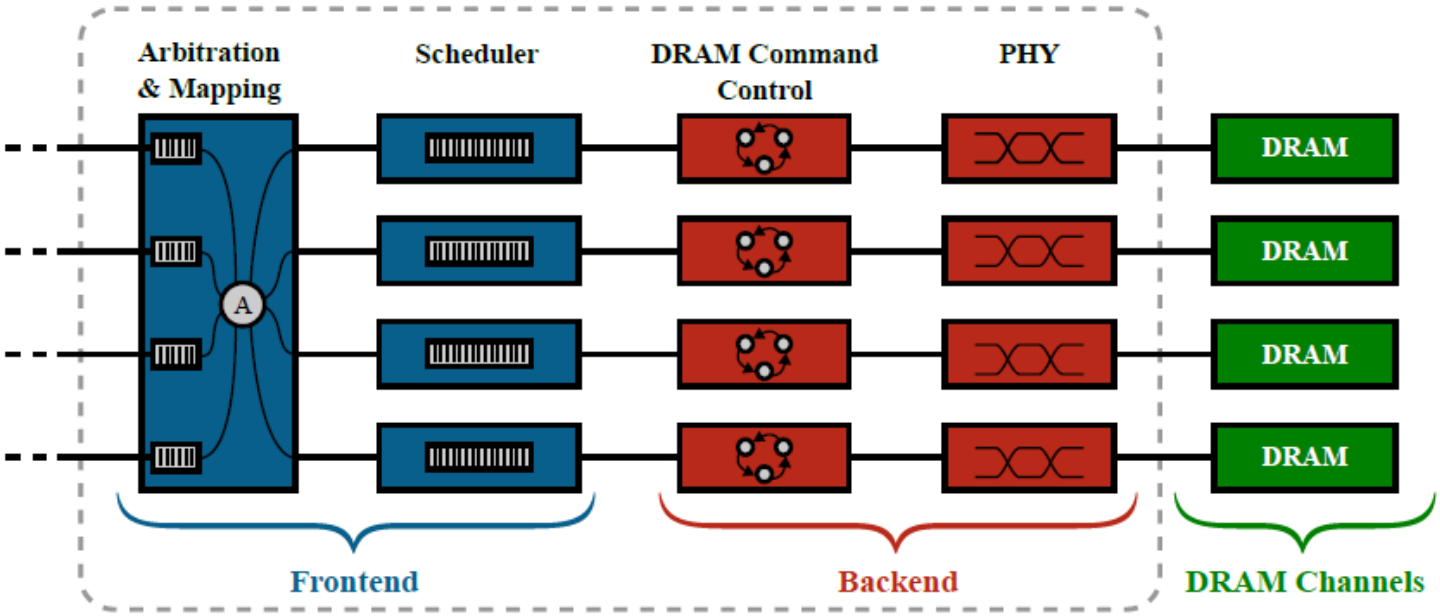


- DRAM latency varies largely
- Depending on
  - **Application**
  - **Address Mapping**
  - **DRAM**
  - **Memory Controller**
  - …

- Row miss: introduces latency (ACT/tRCD, PRE/tRP) and additional energy

The Memory Controller

Multichannel Memory Controller

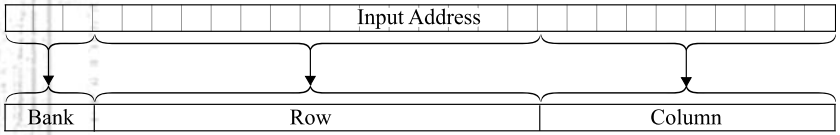


- Arbitration & Mapping: maps transactions to channels, maps addresses to bank/row/column
- Scheduling: minimizes latency e.g. First-Ready-First-Come -Served, open page policy, closed page policy, adaptive page policy
- DRAM Command Control: generates the DRAM commands
- Controller manages power down modi and refresh

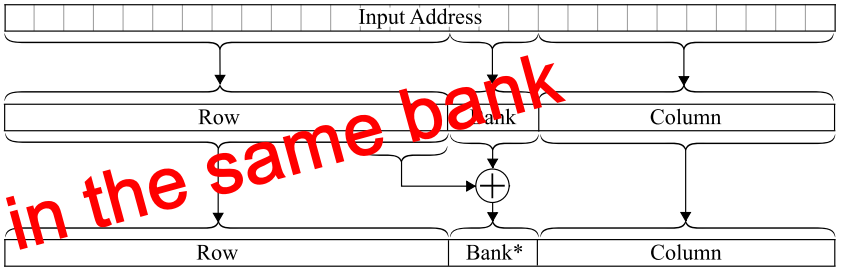
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Application Aware Address Mapping

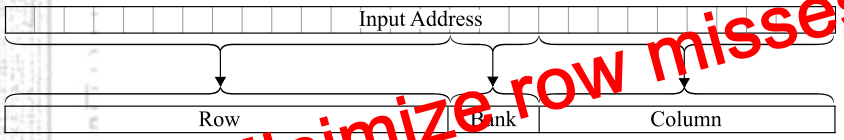
Standard Mapping (BRC)



Permutation-Based Page Interleaving

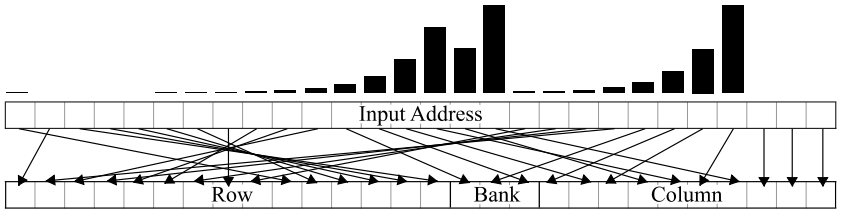


Bank Interleaving (RBC)

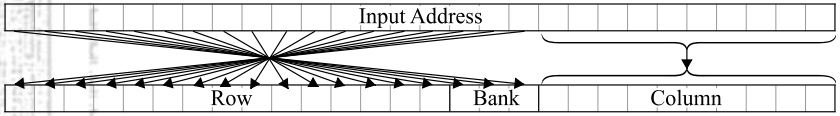


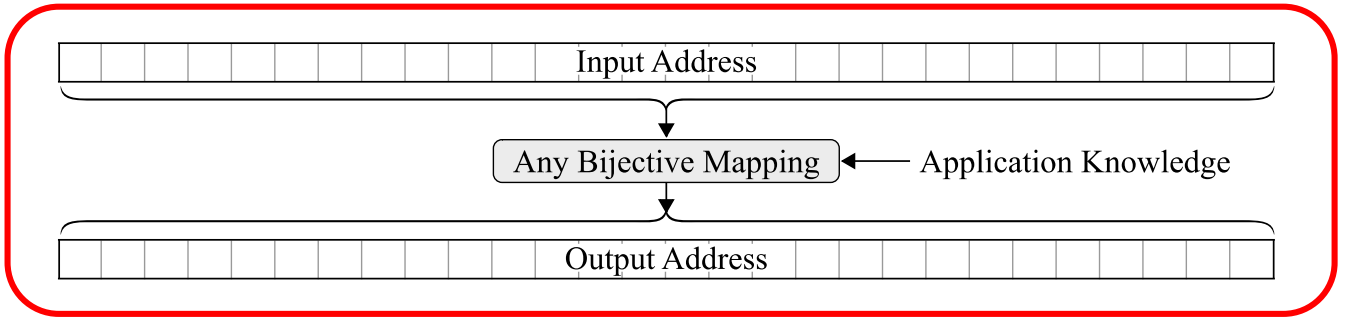
Minimize row misses in the same bank

Toggling Rate Analysis



Bit Reversal Address Mapping





Optimum Address mapping: ILP

$$\min \sum_{t=1}^{T} \sum_{b=1}^{B} y_b(t)$$

Minimize number of row misses for all banks over the access trace

$$s.t. \sum_{r=1}^{R} \sum_{b=1}^{B} x_{rb}(i) = 1 \qquad \forall i = 1,\ldots,n$$

x_{rb} = 1 iff address i is in row r and bank b (each address i is assigned to exactly one row and one bank)

$$\sum_{i=1}^{n} x_{rb}(i) \leq C \qquad \forall r = 1,\ldots,R \quad \forall b = 1,\ldots,B$$

Every row r in bank b has at most C addresses assigned to columns

$$\sum_{r=1}^{R} s_{rb}(t) \leq 1 \qquad \forall t = 1,\ldots,T \quad \forall b = 1,\ldots,B$$

$s_{rb}(t)$ = 1 iff row r in bank b is opened at time t (only one row per bank can be active)

$$s_{rb}(t) \geq x_{rb}(f(t)) \qquad \forall t = 1,\ldots,T \quad \forall r = 1,\ldots,R \quad \forall b = 1,\ldots,B$$

The row r in bank b must be open at time t if it is accessed at time t

$$y_b(t) \geq |s_{rb}(t+1) - s_{rb}(t)| \qquad \forall t = 1,\ldots,T\text{-}1 \quad \forall r = 1,\ldots,R \quad \forall b = 1,\ldots,B$$

1: if row miss, 0: if row hit

$$s_{rb}(0) = 0 \qquad \forall r = 1,\ldots,R \quad \forall b = 1,\ldots,B$$

All rows are closed at the beginning

We have shown if B=1 eq. Min-k-Cut $\Rightarrow$ **NP-Hard Problem**

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

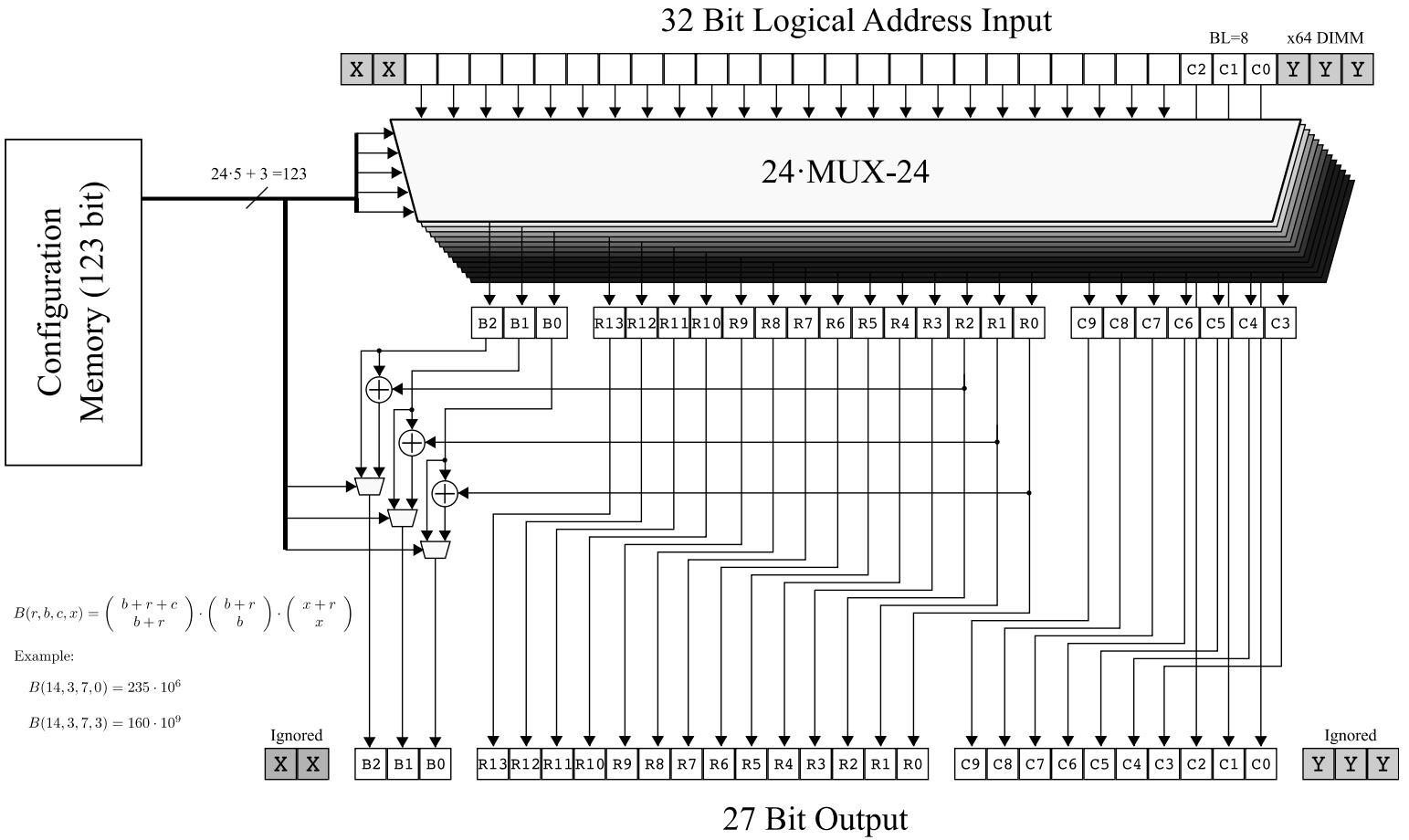
Architectural Template

In general: $2^{32}!$ factorial mappings

- Consider only 1 to 1 mapping of logical to physical address bits: 32! factorials

Use configurable hardware template composed of multiplexers

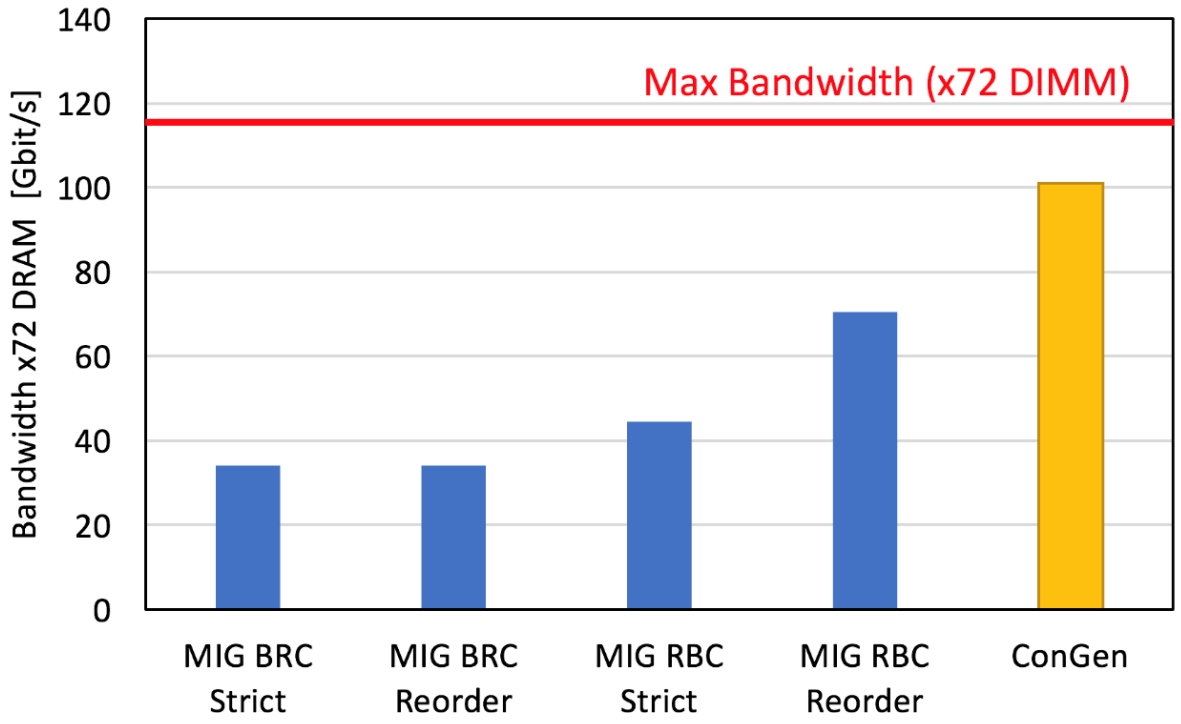
- Find exact solution with minimal row misses under this HW constraints



Example

Using Xilinx MIG Controller

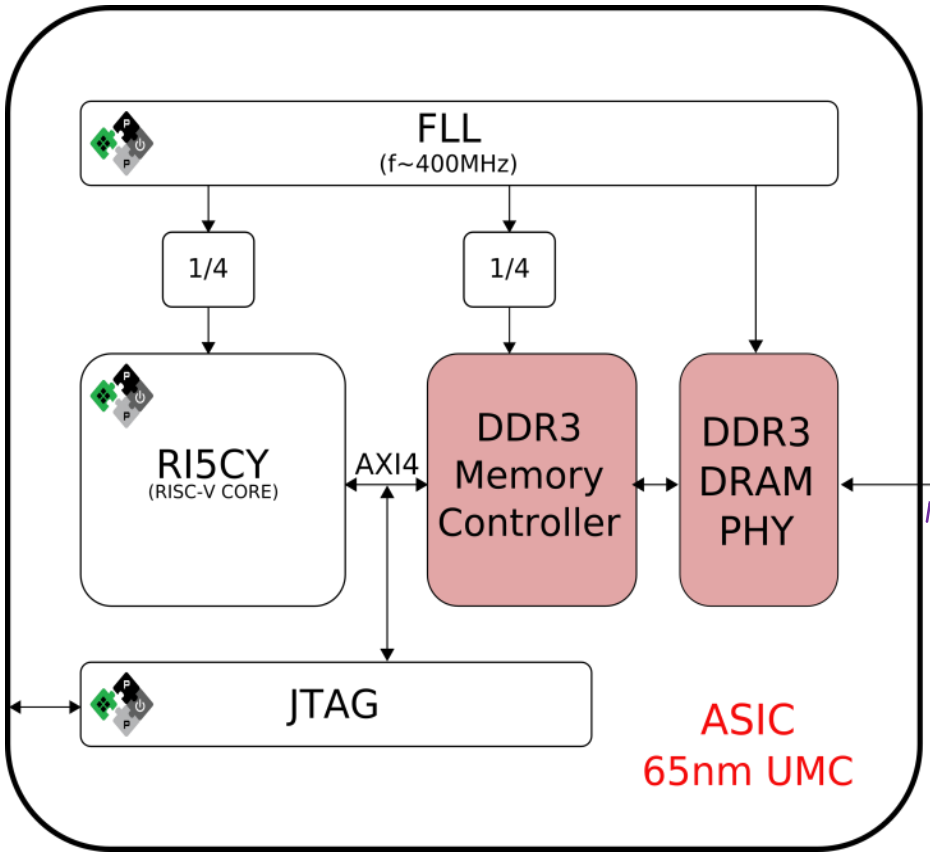
- Industrial holographic 3D image processing application (10 MIG Controller)



Only 9 LUTs necessary

Integration of Memory Controllers (1)

Memory Controller + PHY in RISC-V based SoC in a 65nm UMC technology



PULP
Parallel Ultra Low Power

DDR3/LP
DRAM
x8

64bit data per request!

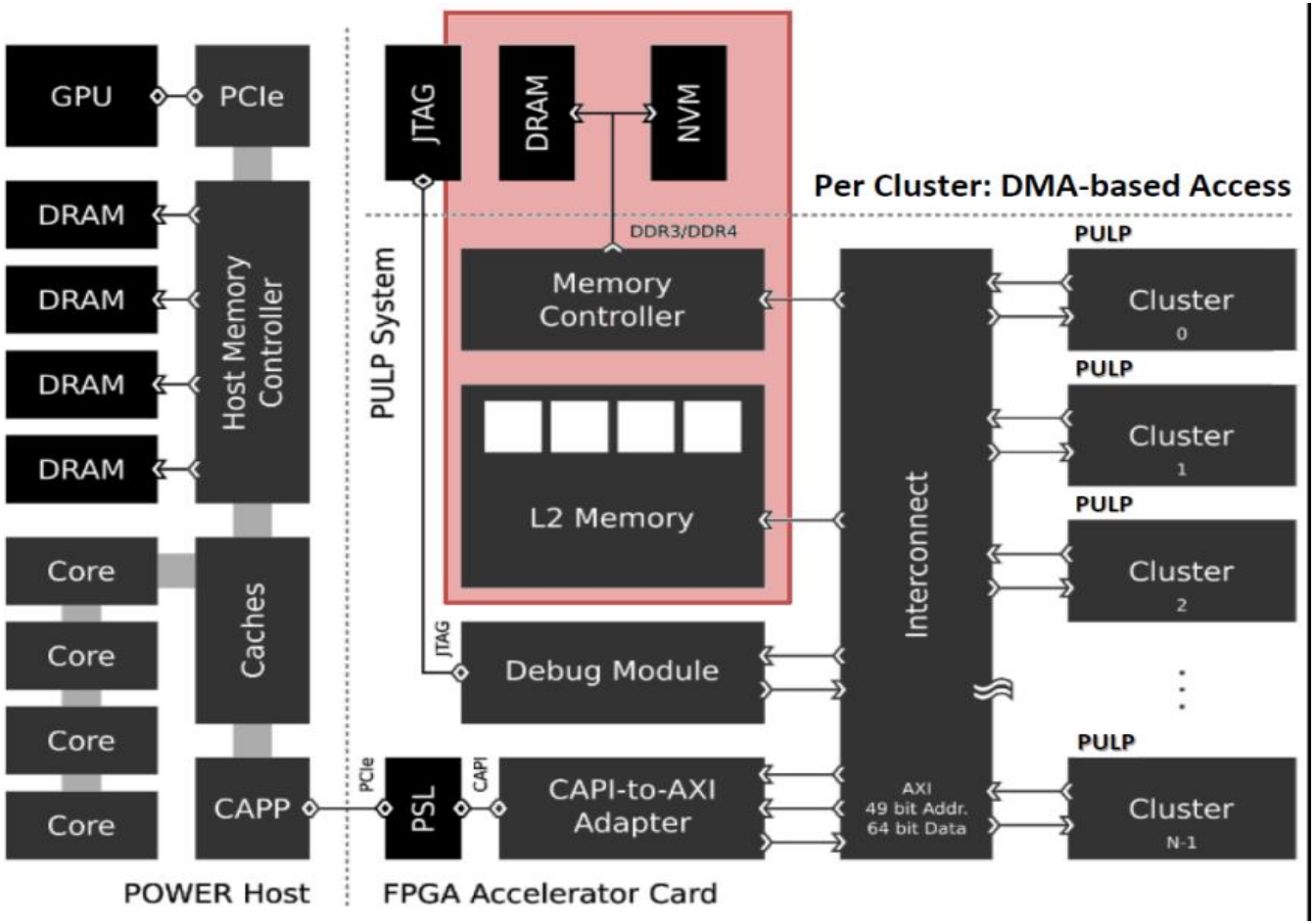
Embedding **DDR3 Controller** (AXI4) and **PHY**

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Integration of Memory Controllers (2)

DDR3/4 DRAM controller

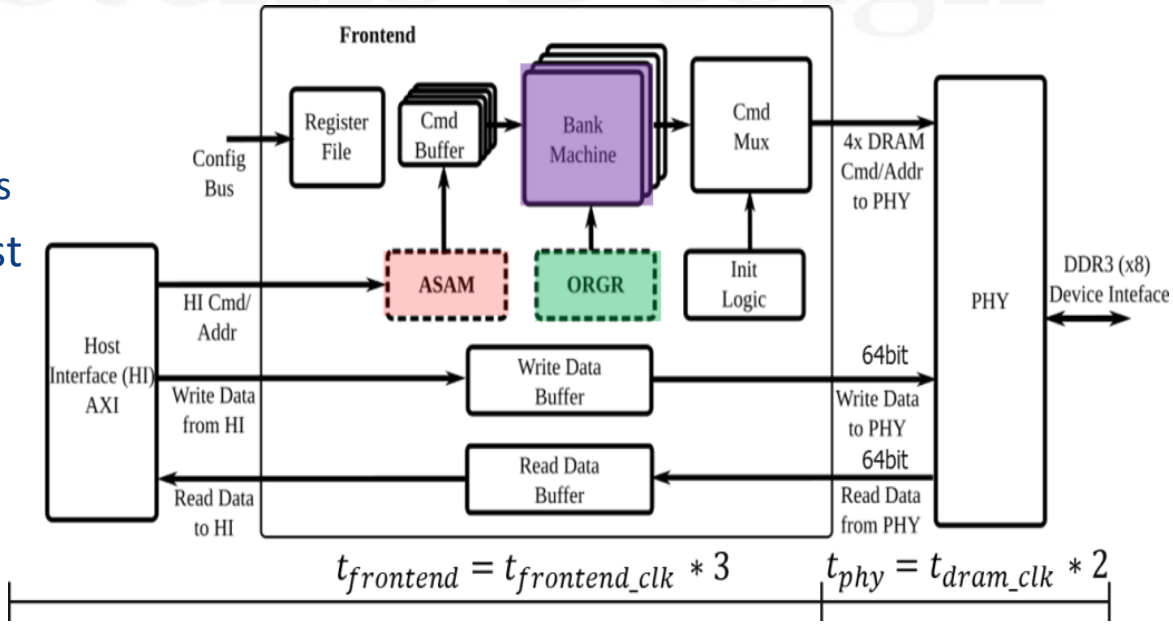
for **kW** Platforms (XILINX FPGA)



→ Controller will support **DDR4/DDR3** DRAMs and KUS + KUS+ **FPGAs**

Features of the DDR Controller and PHY

- 1:4 clocking scheme
  - Smart MUXes, slot based entries
- Low latency (3 clks) read request
- Per Bank architecture (bank machines)
- Optimized buffer depth
- Low complexity and flexible architecture
- Out-of-order scheduling for ACT and PRE commands, FCFS-FR
- Strict-order of CAS (R/W) commands
- Configuration via REG-Bus interface provided by ETH
- Unique feature: support of application knowledge by optimized mapping of logical to physical addresses (ASAM)



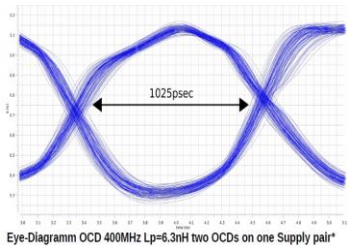
- Extra Address Watermark + Row Granular Refresh in Bank Machine FSM (**Approximate DRAM**)
- Extra Timer for Optimized Row Granular Refresh (**ORGR**)
- Application specific address decoder (**ASAM**) replaces traditional address decoder
- =>Optimal BW without Complex Schedulers, uses First Come First Serve
- **Lean, Low Power, Low Latency**

MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

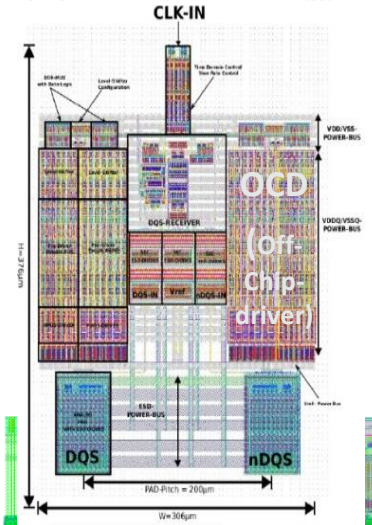
Floorplan/Area x8 Memory Channel

DDR3 PHY Features:

- Designed for 1066 Mbit/s/pin (f=533MHz)
  - – limited by the package (QFN-64) to 400MHz
- Independent calibration of the delays (DQ data, DQS strobe)
  - – 5ps resolution
- All-digital delay locked loop (DLL)
  - – separate placed and routed as macro block
- Off-chip calibration for match impedance drivers (off-chip driver - OCD)
  - – 7 slices based design … 240 - 34Ω
- Local BIAS of Receiver circuits
  - – no analog lines to be routed
  - – independent placement of the PADs



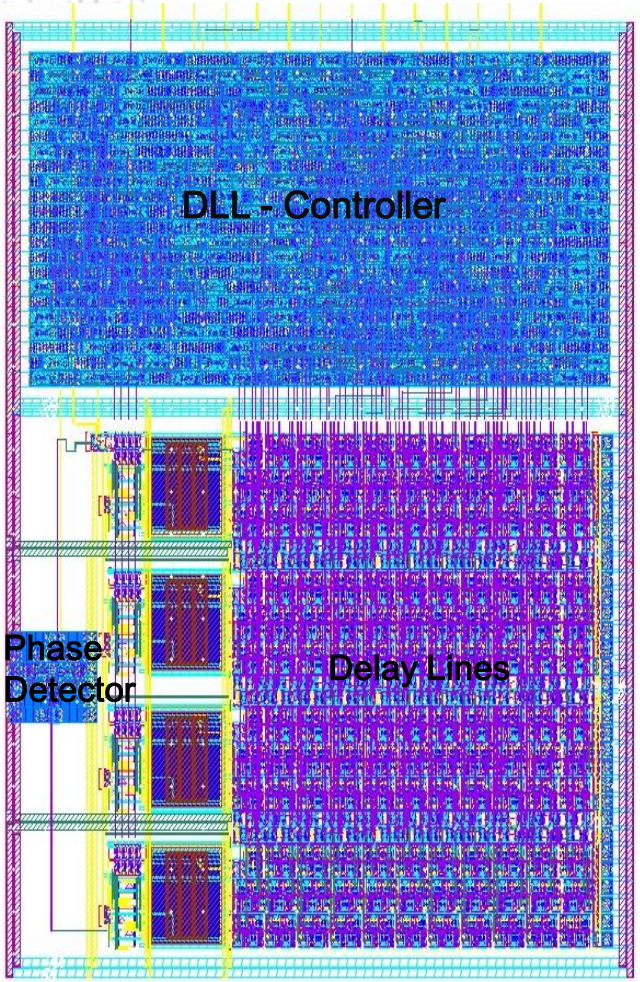
Eye-Diagramm OCD 400MHz Lp=6.3nH two OCDs on one Supply pair*



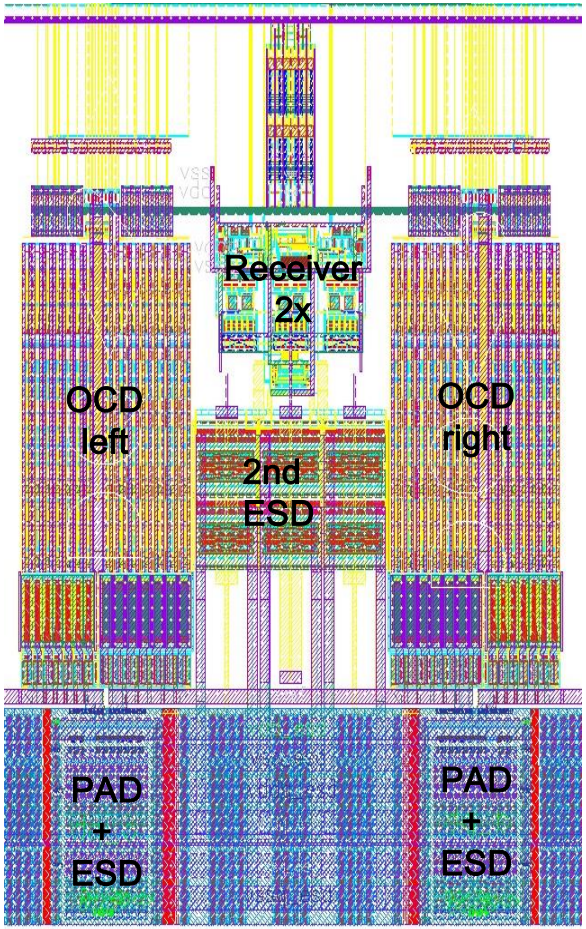
MICROELECTRONIC SYSTEMS DESIGN RESEARCH GROUP

Mix-signal design in the DDR3 PHY

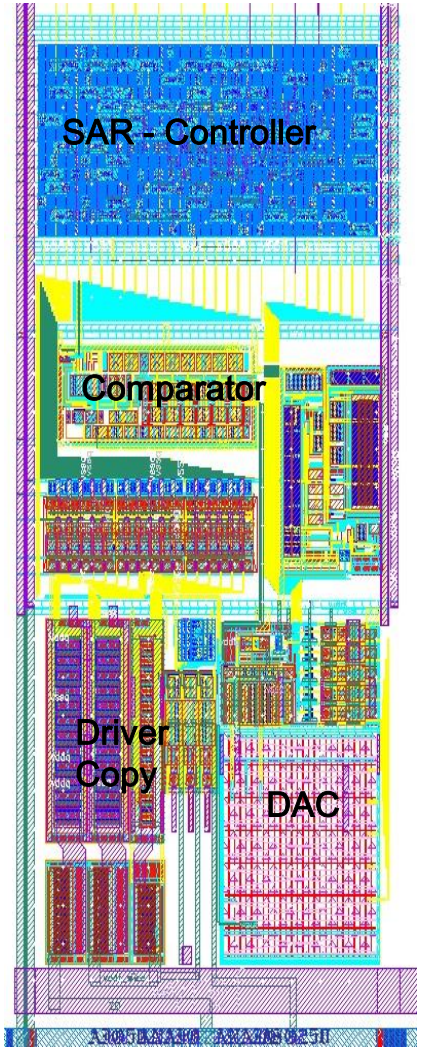
- All-digital delay locked loop (DLL)



- Final DDR3 transceiver design (OCD + receiver)



- Impedance calibration for the drivers (ocdcal)



Summary

- Transprecision computing targets minimum energy consumption per operation taking application into account
- Memories play an important role in computing systems (performance, energy)
- There are many optimizations to improve memory energy efficiency
  - Optimized refresh strategies (ORGR, bank-wise)
  - Use concept of approximate DRAM
  - Optimize access schemes (memory controller)
  - Use sophisticated power down modes
  - Use 3D integration
  - Use new memory/heterogeneous memory types

Thanks for your attention

Further information on
http://ems.eit.uni-kl.de
http://oprecomp.eu/

This project is co-funded by the European Union's H2020-EU.1.2.2. - FET Proactive research and innovation programme under grant agreement #732631.