RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

Future Memory Technologies

Seminar WS2012/13

Benjamin Klenk

2013/02/08

Supervisor: Prof. Dr. Holger Fröning
Department of Computer Engineering
University of Heidelberg

Amdahls rule of thumb

1 byte of memory *and 1 byte per second of I/O are required* ***for each instruction per second*** *supported by a computer.*

Gene Myron Amdahl

#	System	Performance	Memory	B/FLOPs
1	Titan Cray XK7 (Oak Ridge, USA)	17,590 TFLOP/s	710 TB	4.0 %
2	Sequoia BlueGene/Q (Livermore, USA)	16,325 TFLOP/s	1,572 TB	9.6 %
3	K computer (Kobe, Japan)	10,510 TFLOP/s	1,410 TB	13.4 %
4	Mira BlueGene/Q (Argonne, USA)	8,162 TFLOP/s	768 TB	9.4 %
5	JUQUEEN BlueGene/Q (Juelich, GER)	4,141 TFLOP/s	393 TB	9.4 %

[www.top500.org] November 2012

Outline

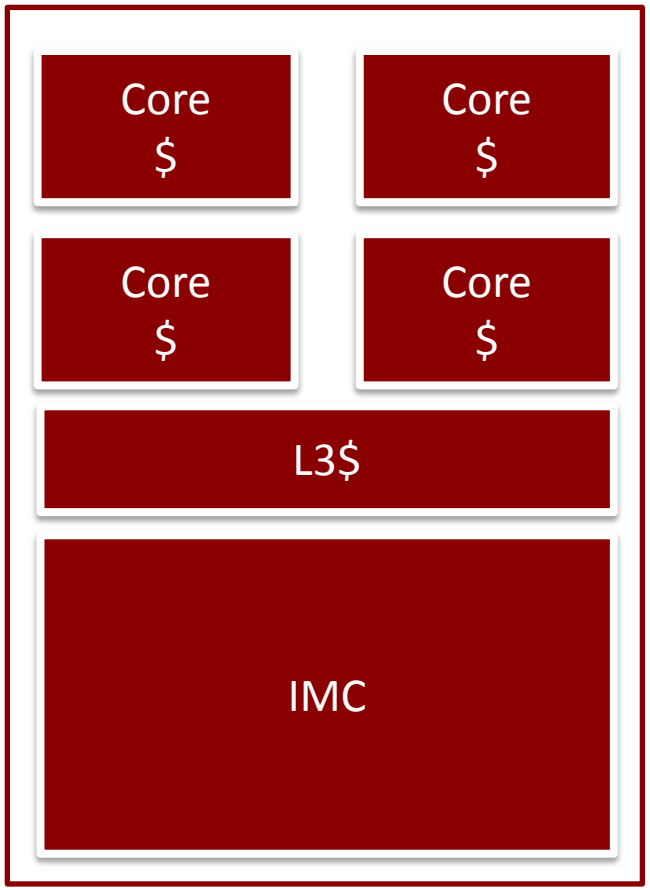
- Motivation
- State of the art
  - RAM
  - FLASH
- Alternative technologies
  - PCM
  - HMC
  - Racetrack
  - STTRAM
- Conclusion

Motivation

Why do we need other technologies?

The memory system

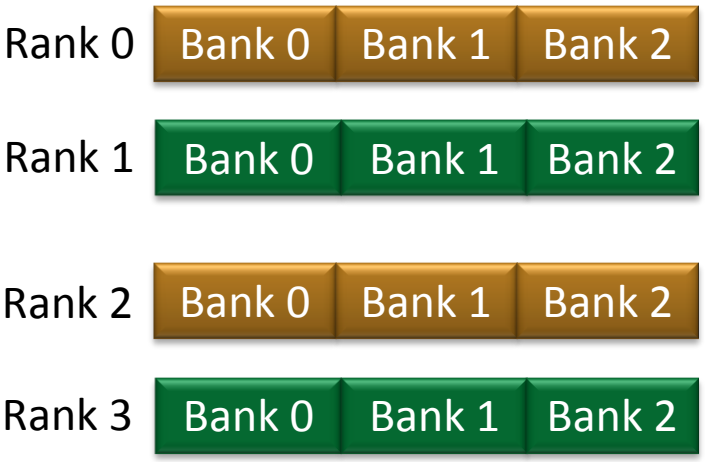
Intel i7-3770



- Modern processors integrate memory controller (IMC)
- Problem: Pin limitation

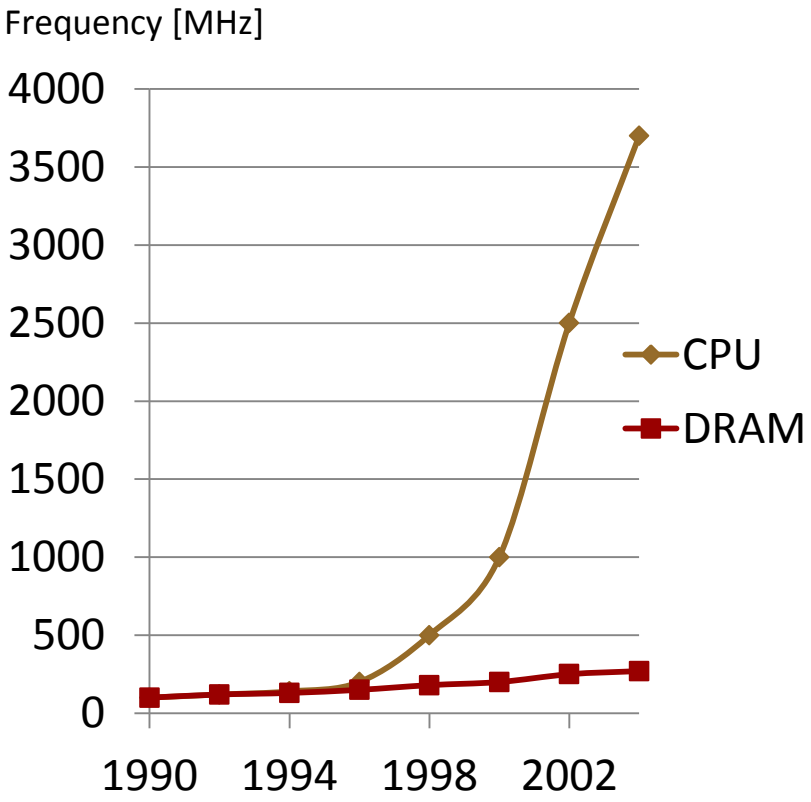
2 x DDR3 Channel
Max 25.6 GB/s

e.g.: 4x8GB = 32 GB
(typical one rank per module)



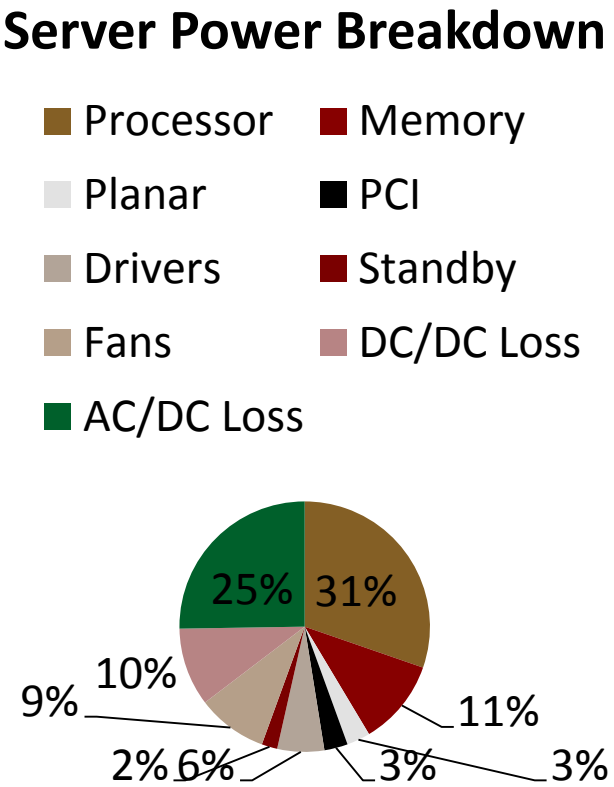
Performance and Power limitations

Memory Wall



[1]

Power Wall

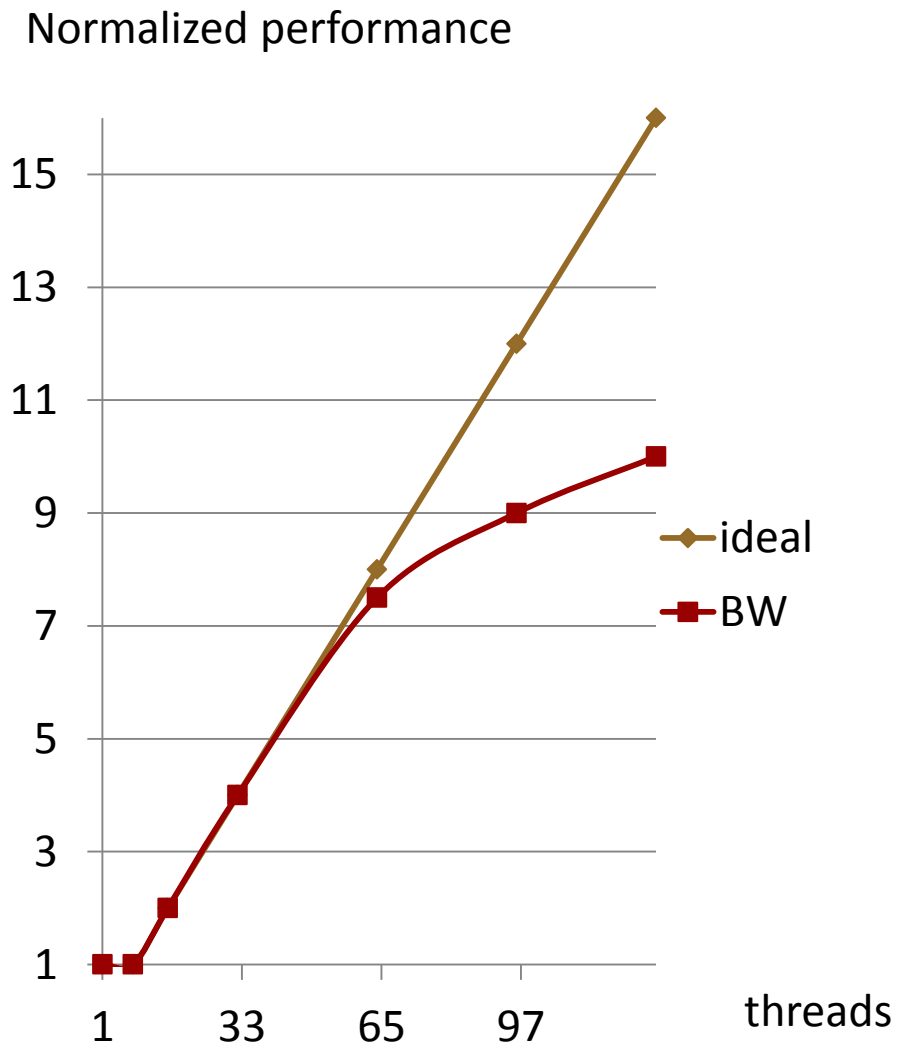


[Intel Whitepaper: Power Management in Intel Architecture Servers, April 2009]

Memory bandwidth is limited

- The demand of working sets increases by the number of cores
- Bandwidth and capacity must scale linearly
- 1 GB/s memory bandwidth per thread [1]

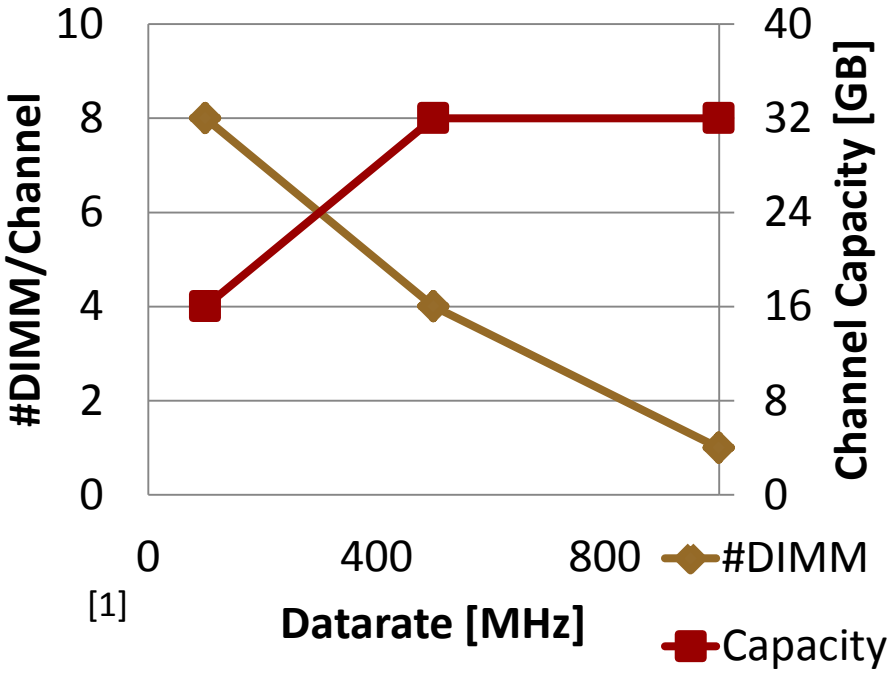
→ Adding more cores doesn't make sense unless there is enough memory bandwidth!



[1]

DIMM count per channel is limited

- Channel capacity does not increase
- Higher data rates result in less DIMMs per channel (to maintain signal integrity)
- High capacity DIMMs are pretty expensive



[1]

Motivation

- What are the problems?
  - Memory Wall
  - Power Wall
  - DIMM count per channel decreases
  - Capacity per DIMM grows pretty slow
- What do we need?
  - High memory bandwidth
  - High bank count (concurrent execution of several threads)
  - High capacity (less page faults and less swapping)
  - Low latency (less stalls and less time waiting for data)
  - And at long last: Low power consumption

State of the art

What are current memory technologies?

Random Access Memory

SRAM

- Fast access and no need of frequent refreshes
- Consists of six transistors
- Low density results in bigger chips with less capacity than DRAM

→ Caches

DRAM

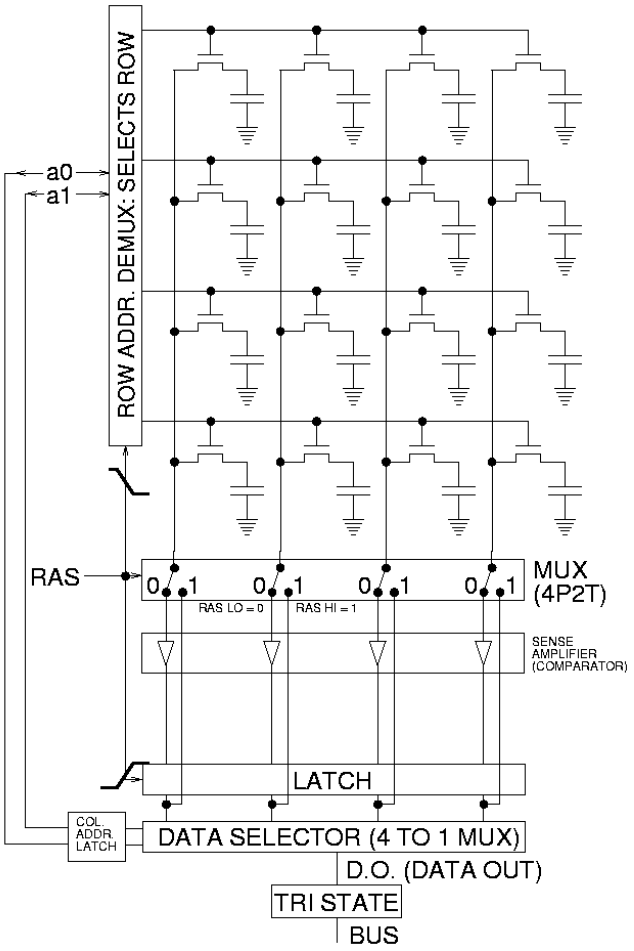
- Consists merely of one transistor and a capacitor (high density)
- Needs to be refreshed frequently (leak current)
- Slower access than SRAM
- Higher power consumption

→ Main Memory

DRAM

- Organized like an array (example 4x4)
- Horizontal Line: Word Line
- Vertical Line: Bit Line

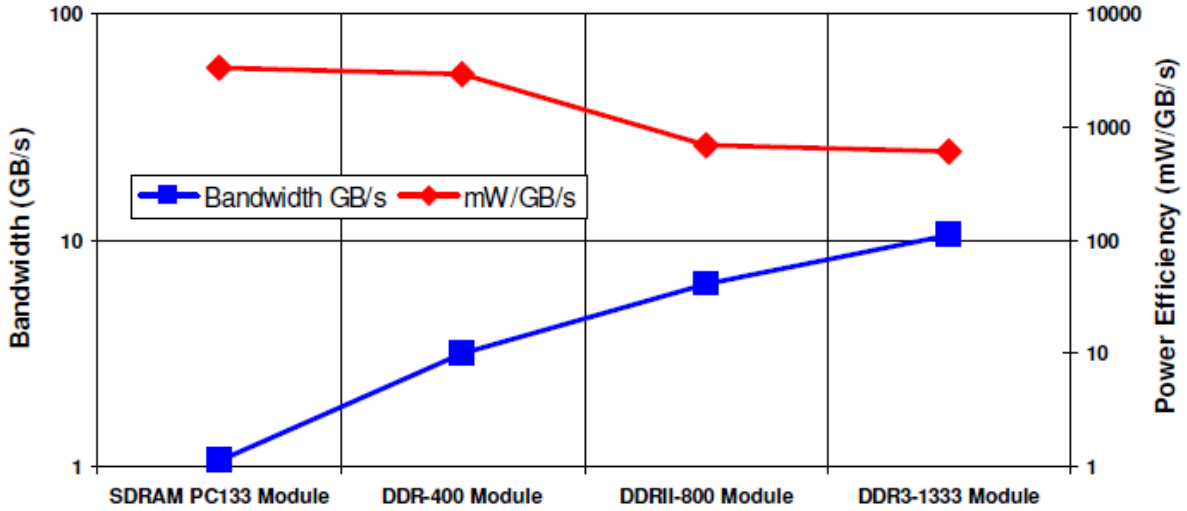
- Refresh every 64ms
- Refresh logic is integrated in DRAM controller



www.wikipedia.com

The history of DDR-DRAM

- DDR SDRAM is state of the art for main memory
- There are several versions of DDR SDRAM:

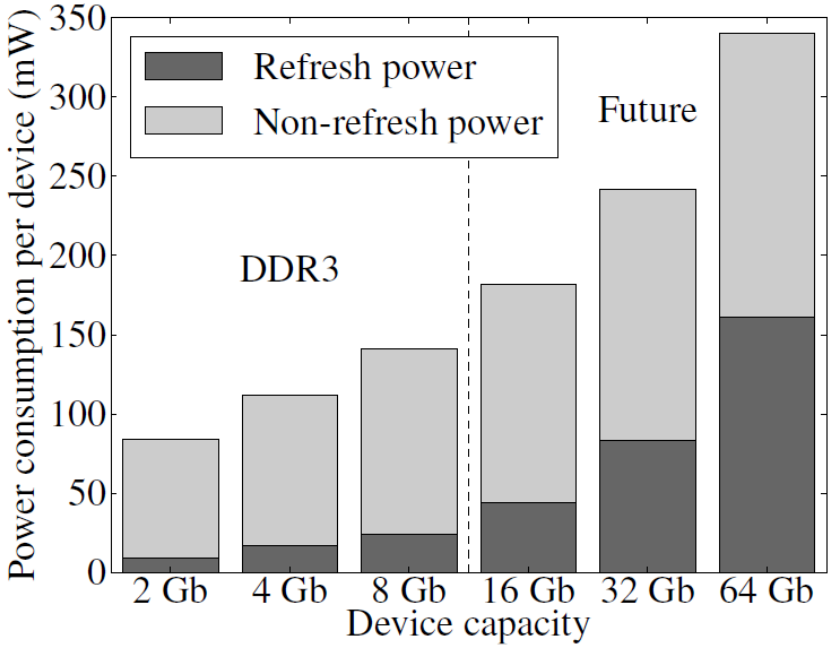


[9] ExaScale Computing Study

Version	Clock [MHz]	Transfer Rate [MT/s]	Voltage [V]	DIMM pins
DDR1	100-200	200-400	2.5/2.6	184
DDR2	200-533	400-1066	1.8	240
DDR3	400-1066	800-2133	1.5	240
DDR4	1066-2133	2133–4266	1.2	284

Power consumption and the impact of refreshes

- Refresh takes 7.8µs (<85°C) / 3.9µs (<95°C)
- Refresh every 64ms
- Multiple banks enable concurrent refreshes
- Commands flood command bus



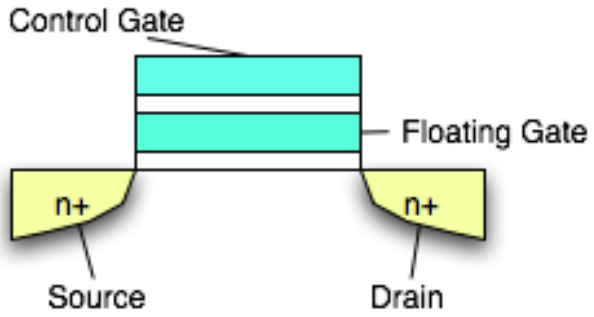
RAIDR: Retention-Aware Intelligent DRAM Refresh, Jamie Liu et al.

	1990	Today
Bits/row	4096	8192
Capacity	Tens of MB	Tens of GB
Refreshes	10 per ms	10.000 per ms

[1]

Flash

- FLASH memory cells are based on floating gate transistors
- MOSFET with two gates: Control (CG) & Floating Gate (FG)
- FG is electrically isolated and electrons are trapped there (only capacitive connected)
- Programming by hot-electron injection
- Erasing by quantum tunneling



http://en.wikipedia.org/wiki/Floating-gate_transistor

Problems to solve

- DRAM
  - Limited DIMM count → limits capacity for main memory
  - Unnecessary power consumption of refreshes
  - Low bandwidth
- FLASH
  - Slow access time
  - Limited write cycles
  - Pretty low bandwidth

Alternative technologies

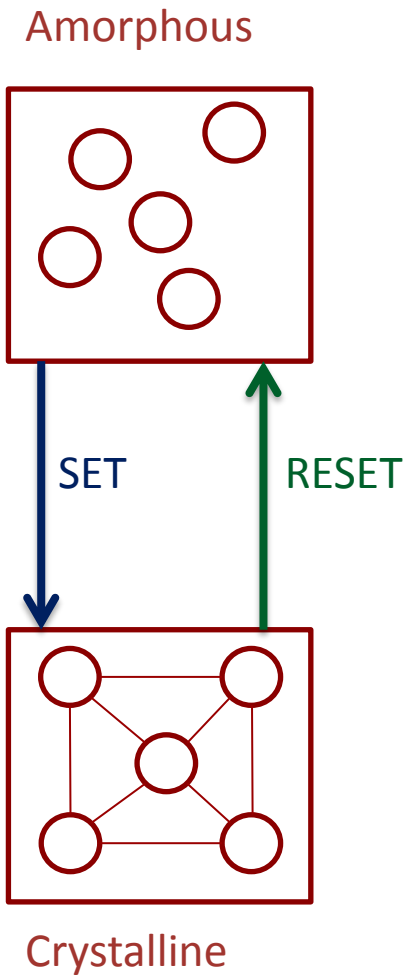
Which technologies show promise for the future?

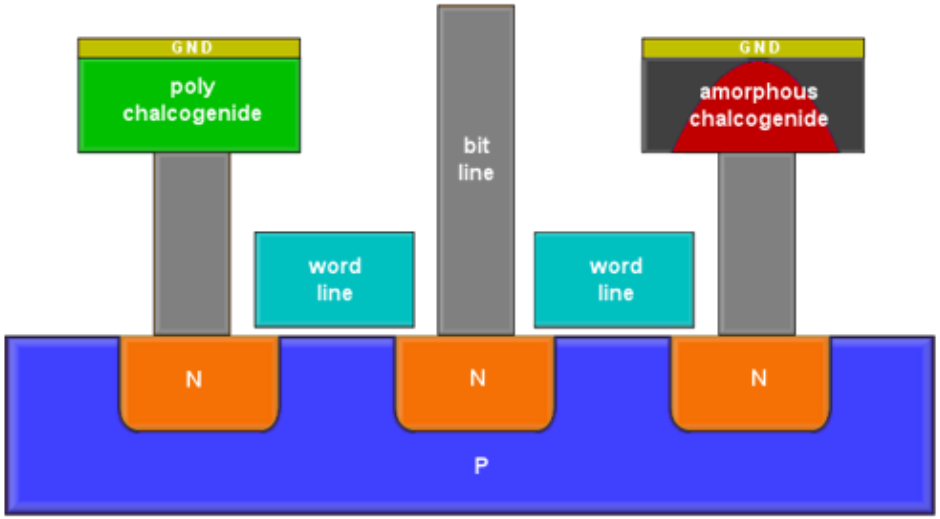
Outline

- Phase Change Memory (PCM, PRAM, PCRAM)
- Hybrid Memory Cube (HMC)
- Racetrack Memory
- Spin-Torque Transfer RAM (STTRAM)

Phase Change Memory (PCM)

- Based on chalcogenide glasses (also used for CD-ROMs)
- PCM lost competition with FLASH and DRAM because of power issues
- PCM cells become smaller and smaller and hence the power consumption decreases

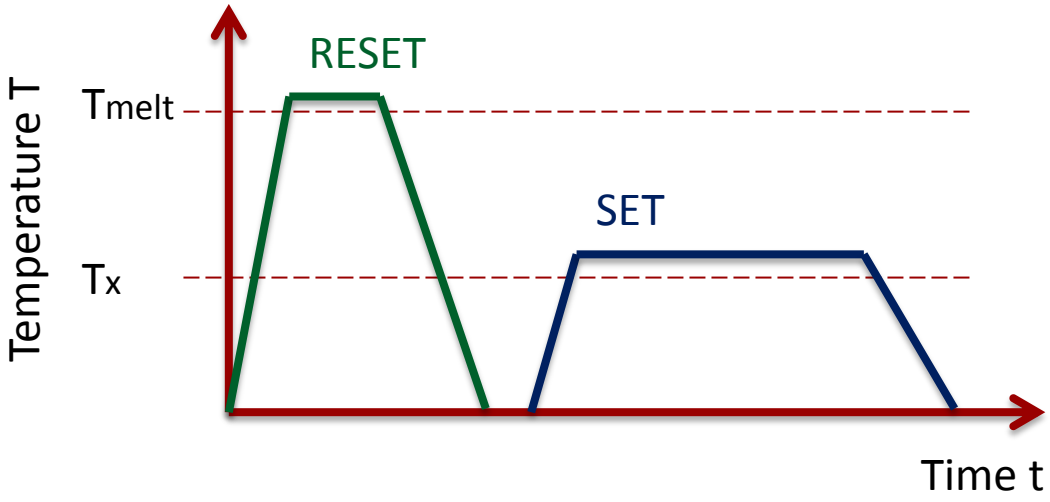


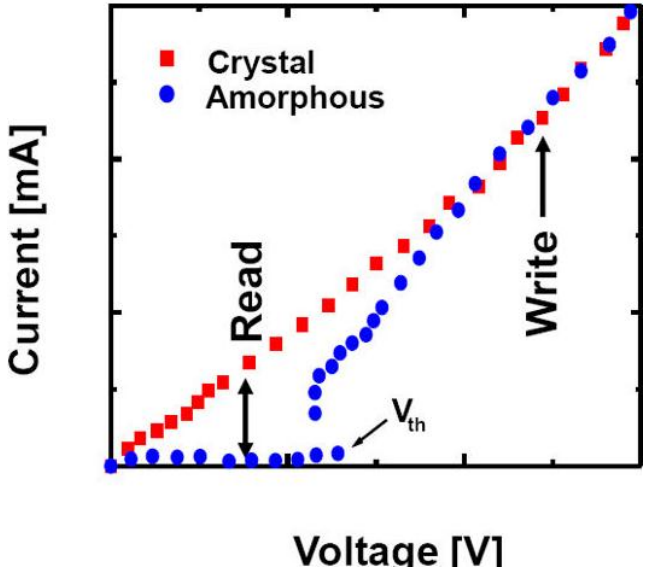


[http://www.nano-ou.net/Applications/PRAM.aspx]

How to read and write

- Resistance changes with state (amorphous, crystalline)
- Transition can be forced by optical or electrical impulses

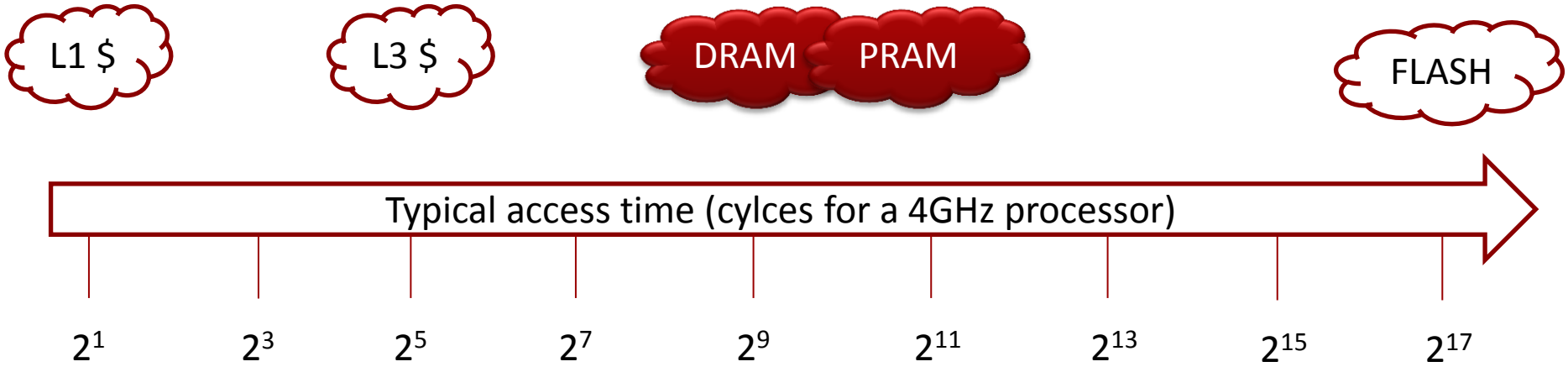




http://agigatech.com/blog/pcm-phase-change-memory-basics-and-technology-advances/

Access time of common memory techniques

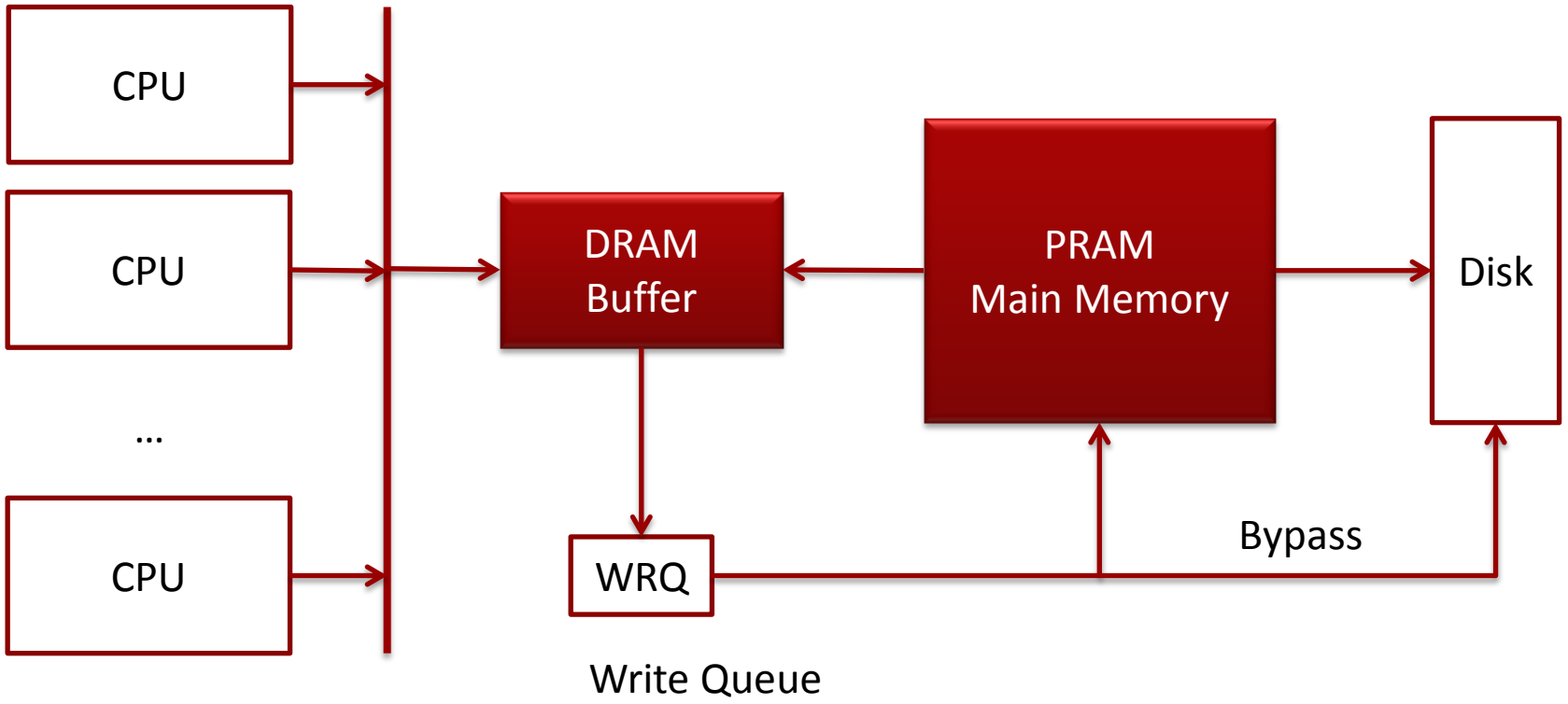
- PRAM still “slower“ than DRAM
- Only PRAM would perform worse (access time 2-10x slower)
- But: Density much better! (4-5F^2 compared to 6F^2 of DRAM)
- We need to find a tradeoff



[6]

Hybrid Memory: DRAM and PRAM

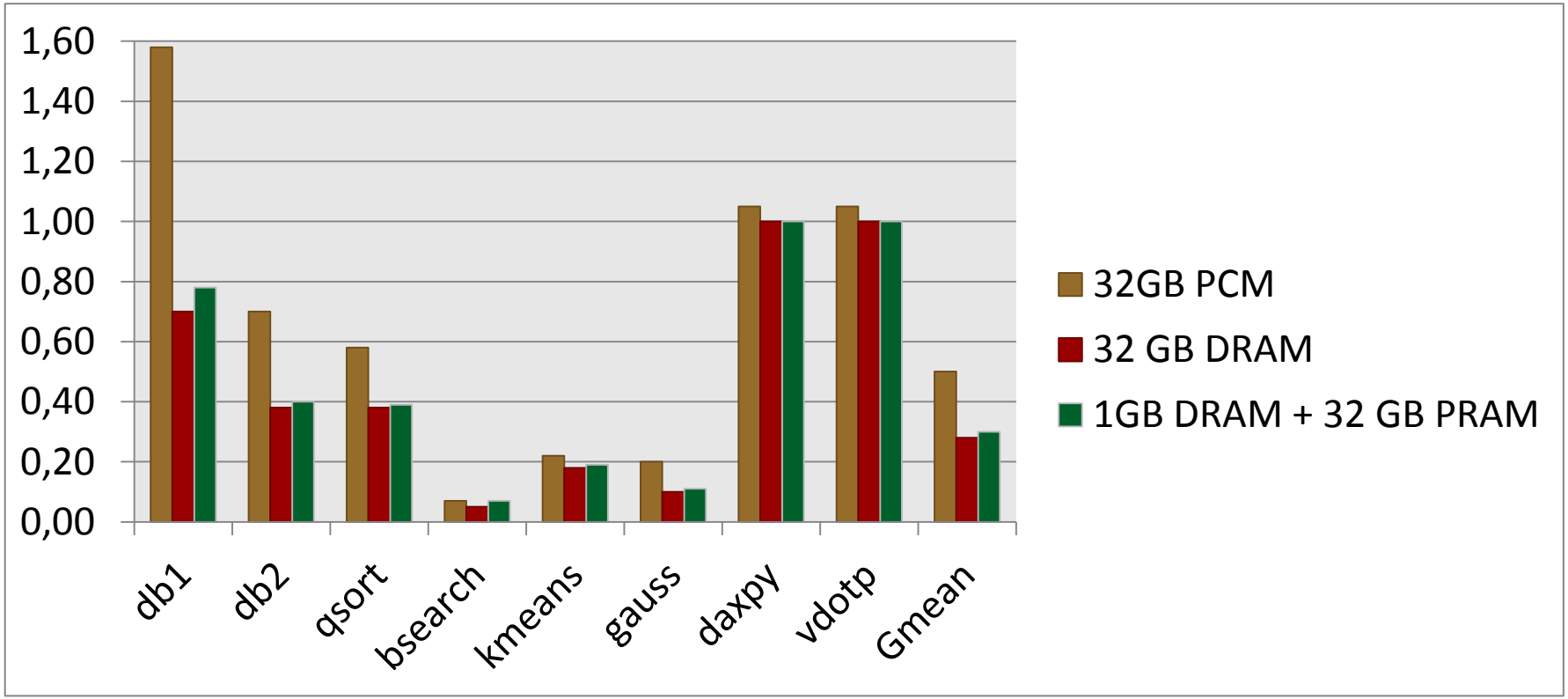
- We still use DRAM as buffer / cache
- Technique to hide higher latency of PRAM



[6]

Performance of a hybrid memory approach

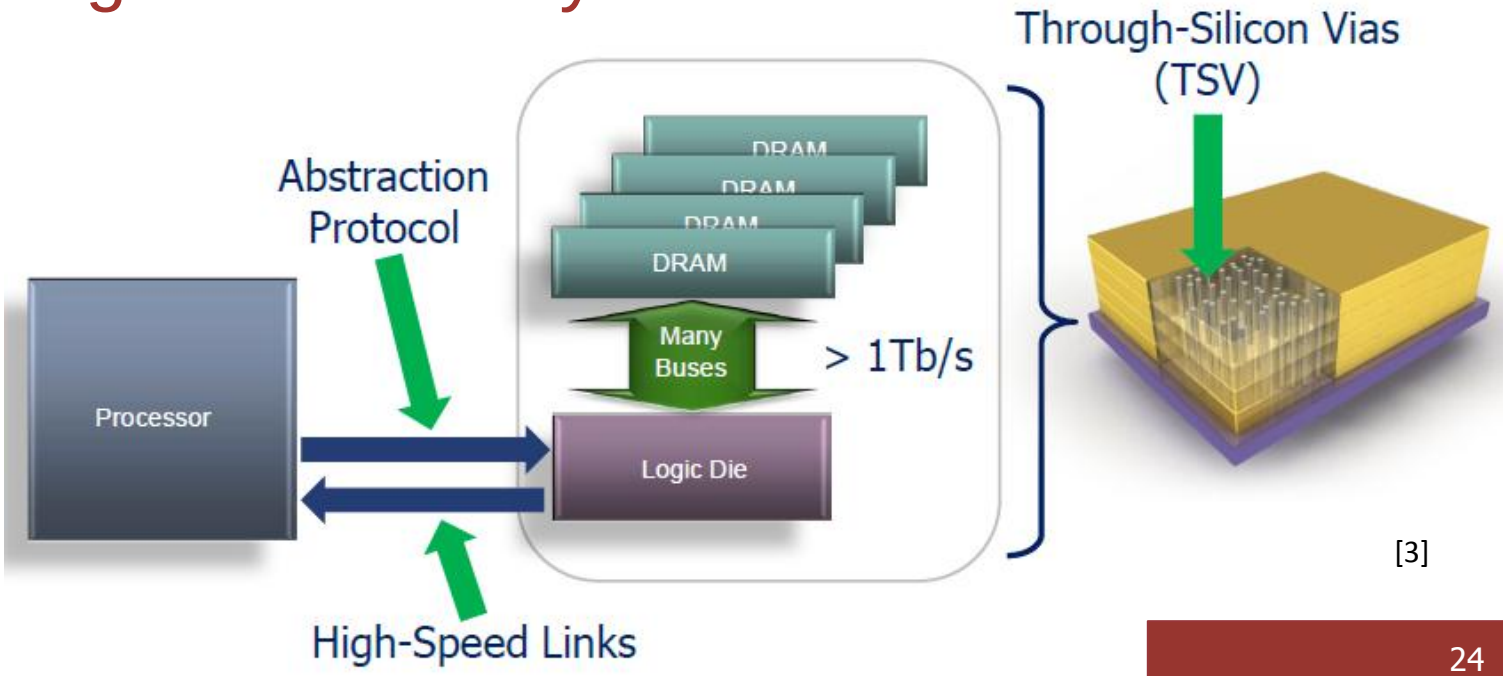
- Assume: Density: 4x higher, Latency: 4x slower (in-house simulator of IBM)
- Normalized to 8GB DRAM



[Scalable High Performance Main Memory System Using Phase-Change Memory Technology, Qureshi et al.]

Hybrid Memory Cube

- Promising memory technology
- Leading companies: Micron, Samsung, Intel
- 3D disposal of DRAM modules
- Enables high concurrency



[3]

What has changed?

Former

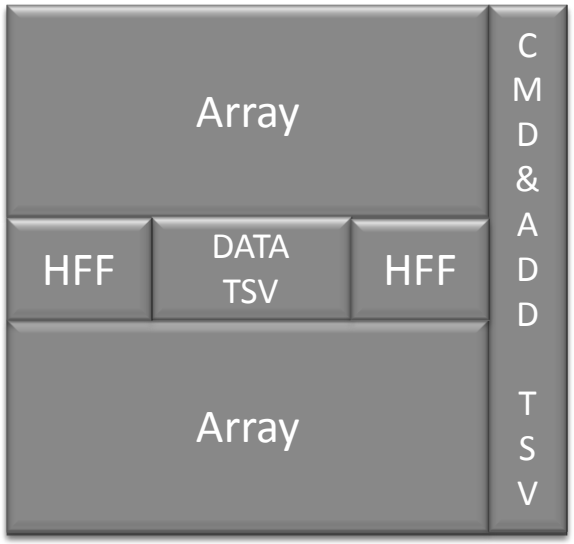
- CPU is directly connected to DRAM (Memory Controller)
- Complex scheduler (queues, reordering)
- DRAM timing parameter standardized across vendors
- Slow performance growth

HMC

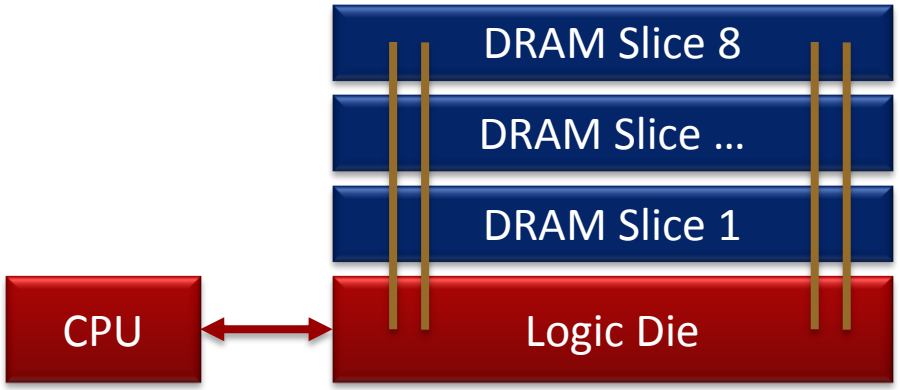
- Abstracted high speed interface
- Only abstracted protocol, no timing constraints (packet based protocol)
- Innovation inside HMC
- HMC takes requests and delivers results in most advantageous order

HMC architecture

- DRAM logic is stripped away
- Common logic on the Logic Die
- Vertical Connection through TSV
- High speed processor interface



[4]



[3]

High speed interface
(packet based protocol)

More concurrency and bandwidth

- Conventional DRAM:
  - 8 devices and 8 banks/device results in 64 banks
- HMC gen1:
  - 4 DRAMs * 16 slices * 2 banks results in 128 banks
  - If 8 DRAMs are used: 256 banks
- Processor Interface:
  - 16 Transmit and16 Receive lanes: 32 x 10Gbps per link
  - 40 GBps per Link
  - 8 links per cube: 320 GBps per cube (compared to about 25.6 GBps of recent memory channels)

[3]

Performance comparison

Technology	VDD	IDD	BW GB/s	Power W	mW/GBps	pj/bit	Real pj/bit
SDRAM PC133 1GB	3.3	1.50	1.06	4.96	4664.97	583.12	762.0
DDR 333 1GB	2.5	2.19	2.66	5.48	2057.06	257.13	245.0
DDR 2 667 2GB	1.8	2.88	5.34	5.18	971.51	121.44	139.0
DDR 3 1333 2GB	1.5	3.68	10.66	5.52	517.63	64.70	52.0
DDR 4 2667 4 GB	1.2	5.50	21.34	6.60	309.34	38.67	39.0
HMCgen1	1.2	9.23	128.00	11.08	86.53	10.82	13.7

HMC is costly because of TSV and 3D stacking!

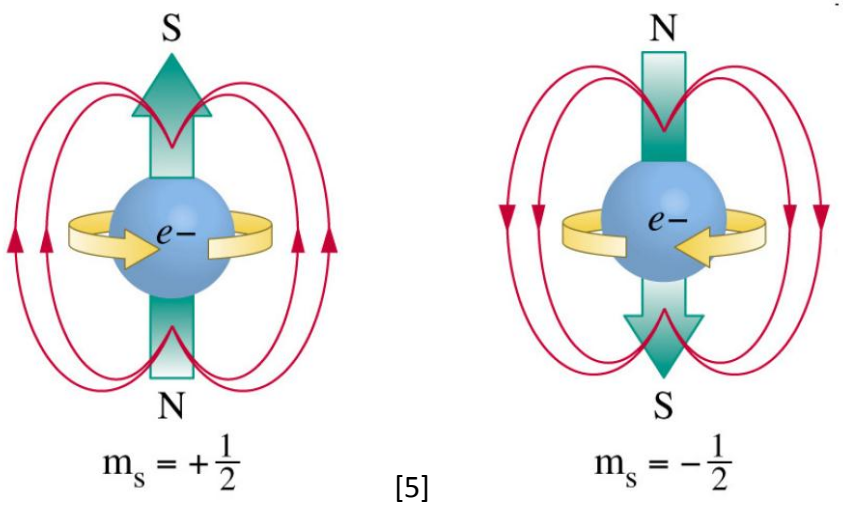
[3]

Further features of HMCgen1:

- 1GB 50nm DRAM Array
- 512 MB total DRAM cube
- 128 GB/s Bandwidth

Electron spin and polarized current

- Spin another property of particles (like mass, charge)
- Spin is either “up“ or “down“
- Normal materials consist of equally populated spin-up and down electrons
- Ferromagnetic materials consist of an unequally population

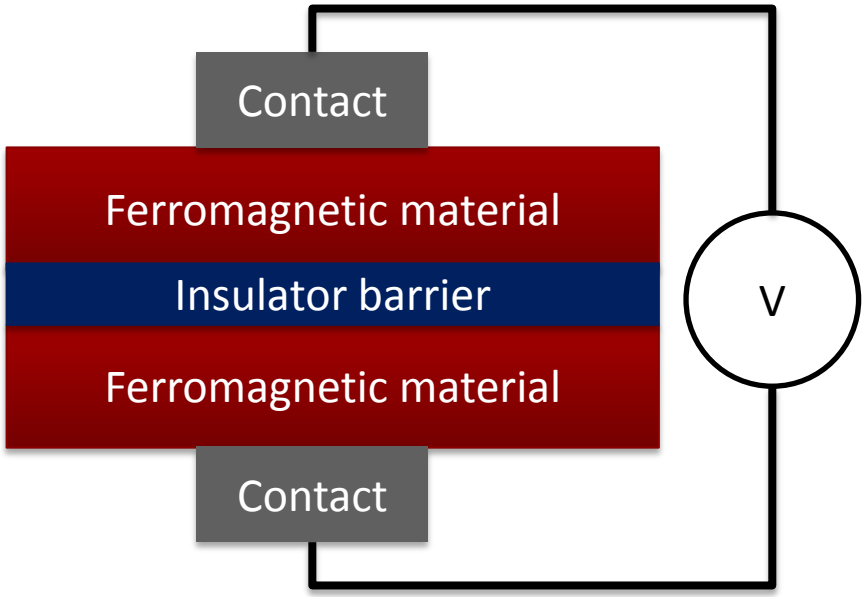


[5]

Magnetic Tunnel Junction (MTJ)

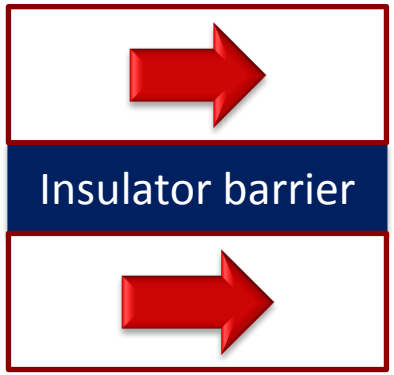
- Discovered in 1975 by M.Julliére
- Electrons become spin-polarized by the first magnetic electrode

- Two phenomena:
  - Tunnel Magneto-Resistance
  - Spin Torque Transfer



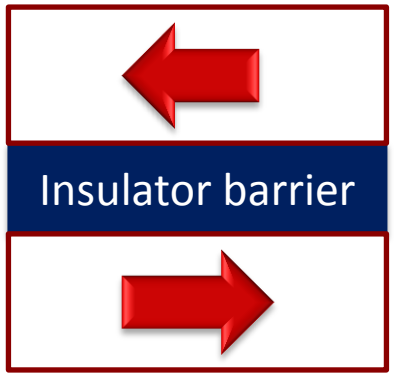
Tunneling Magneto-Resistance (TMR)

- Magnetic moments parallel: Low resistance
- Otherwise: High resistance



Low resistance

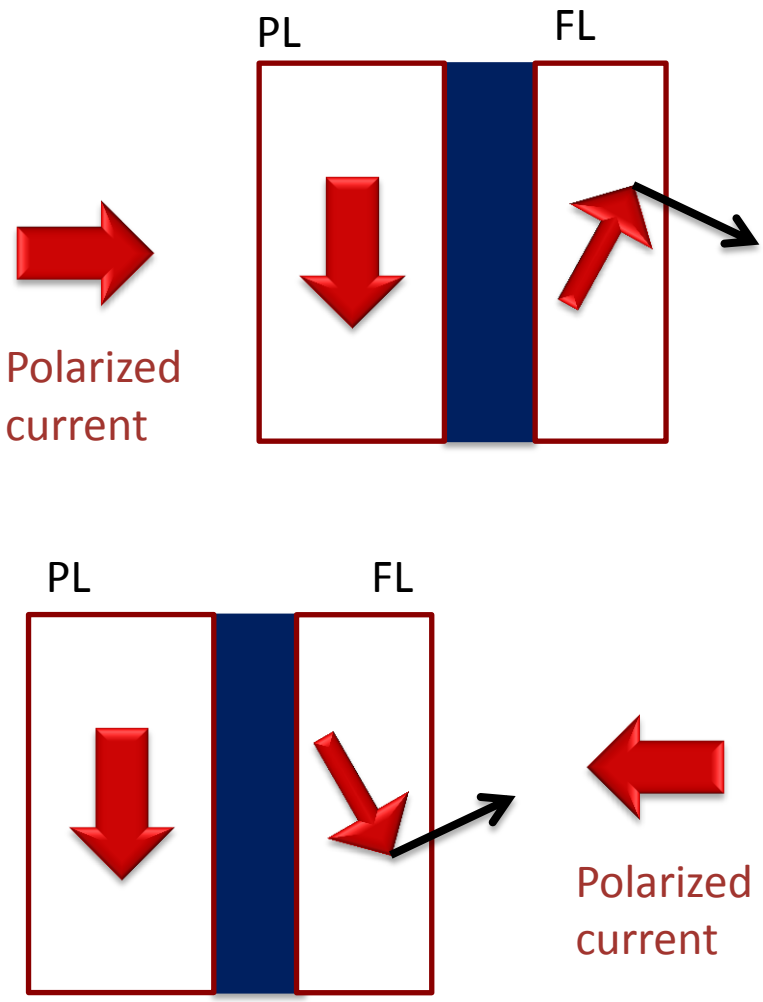
- 1995: Resistance difference of 18% at room temperature
- Nowadays: 70% can be fabricated with reproducible characteristics



High resistance

Spin Torque Transfer (STT)

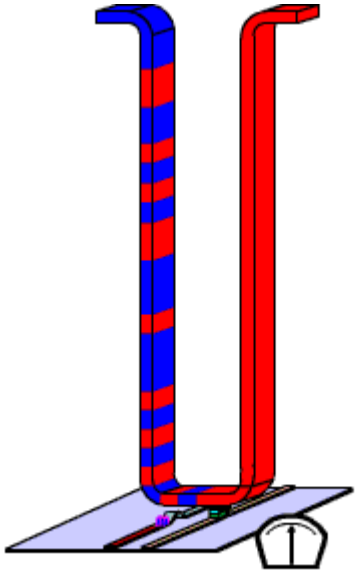
- Thick and pinned layer (PL) → can not be changed
- Thin and free layer (FL) → can be changed
- FL magnetic structure needs to be smaller than 100-200nm

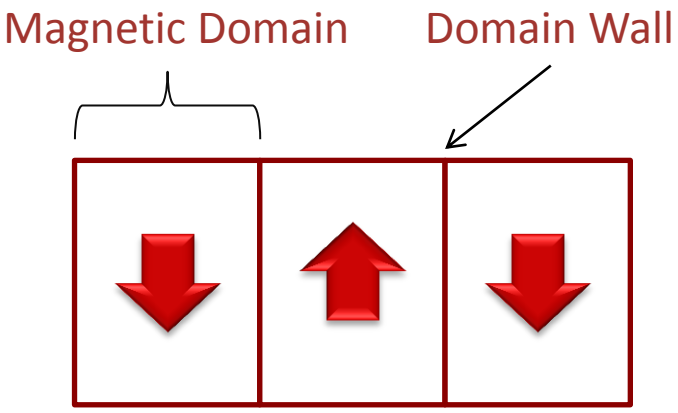


Racetrack Memory

http://researcher.watson.ibm.com/researcher/view_project_subpage.php?id=3811

- Ferromagnetic nanowire (racetrack)
- Plenty of magnetic domain walls (DW)
- DW are magnetized either “up“ or “down“
- Racetrack operates like a shift register

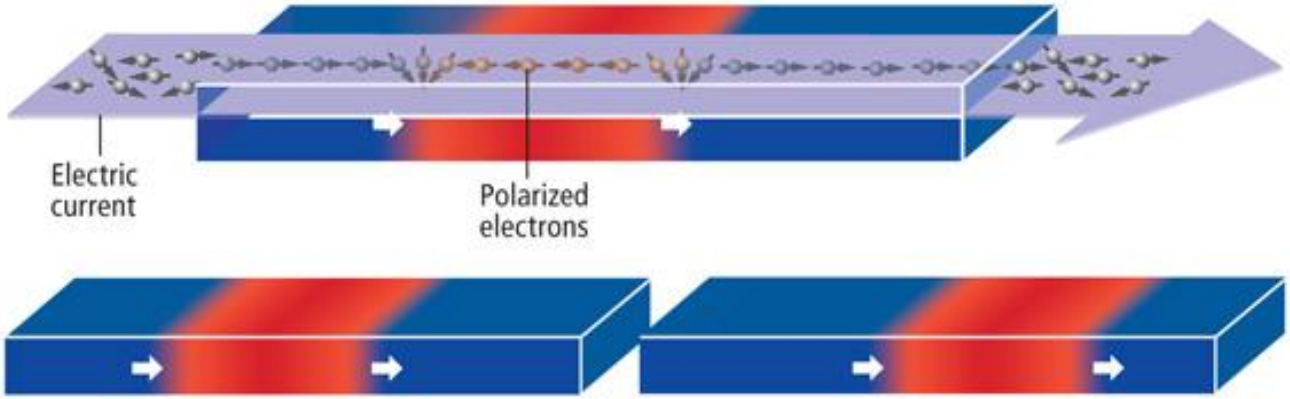




http://www.tf.uni-kiel.de/matwis/amat/elmat_en/kap_4/backbone/r4_3_3.html

Racetrack

- DW are shifted along the track by current pulses (~100m/s)
- Principle of spin-momentum transfer

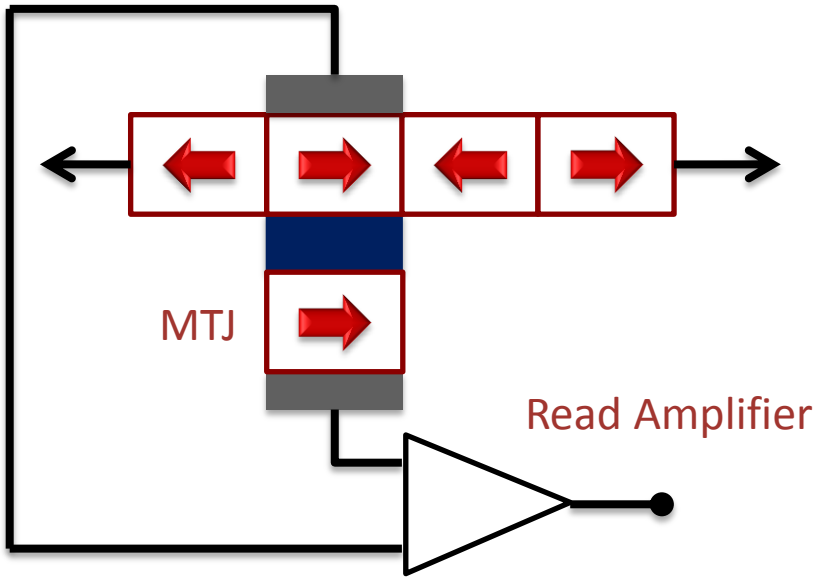


[Scientific American 300 (2009), Data in the Fast Lanes of Racetrack Memory]

Read & Write

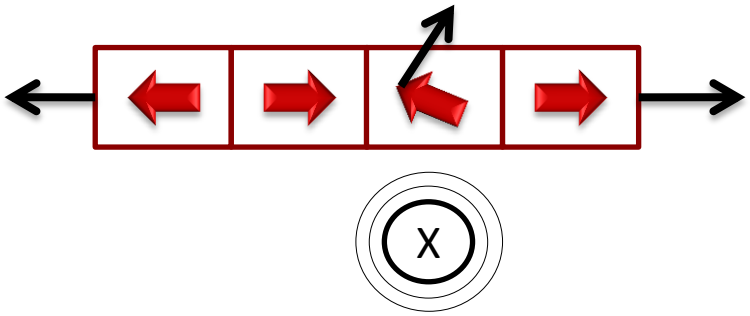
Read

- Resistance depends on magnetic momentum of magnetic domain (TMR effect)



Write

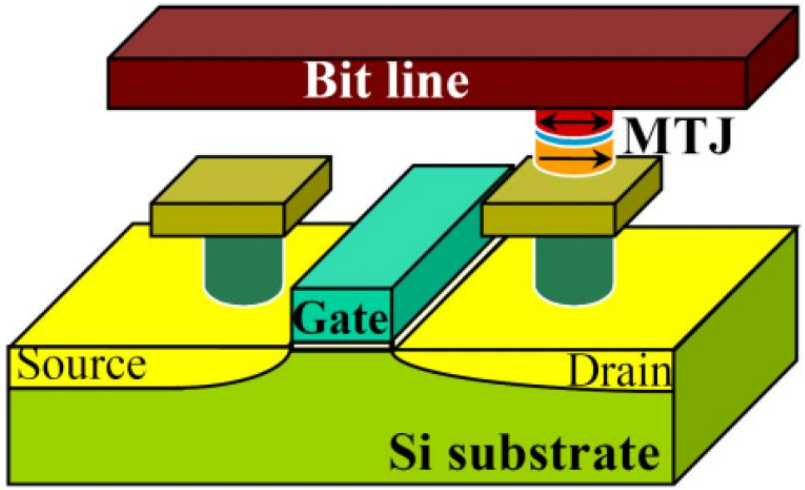
- Multiple possibilities:
  - Self field of current from metallic neighbor elements
  - Spin momentum transfer torque from magnetic Nano elements

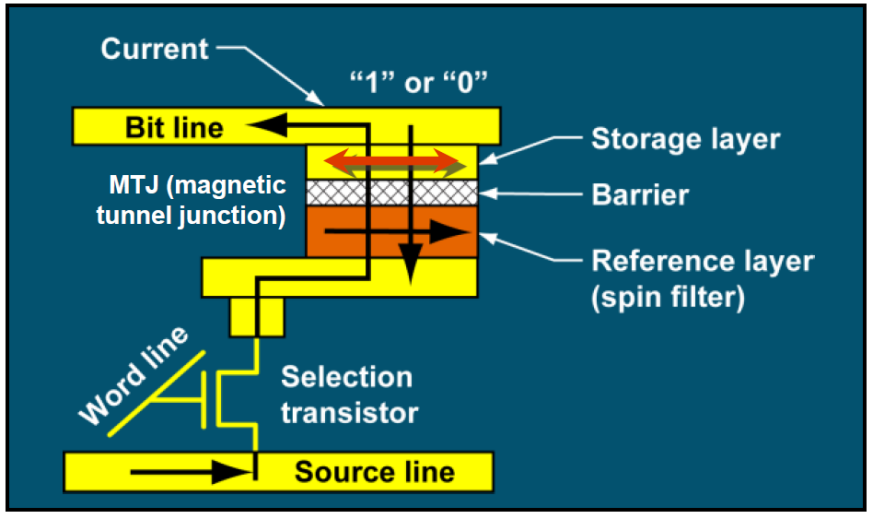


Magnetic field of current

STTRAM

- Memory cell based on MTJ
- Resistance changed because of TMR
- Spin-polarized current instead of magnetic field to program cell

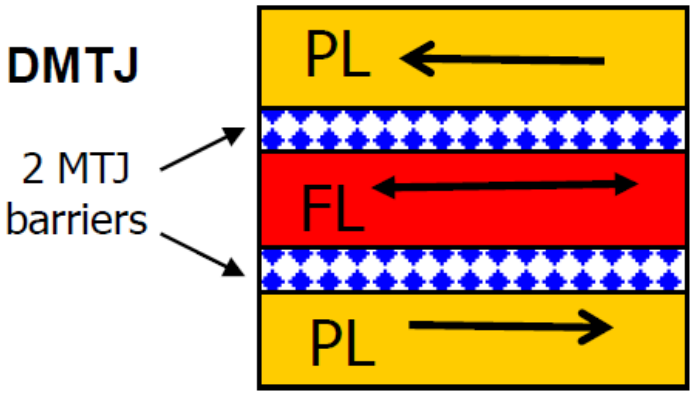




[7]

STTRAM provides…

- High scalability because write current scales with cell size
  - 90nm: 150µA, 45nm: 40µA
- Write current about 100µA and therefore low power consumption
- Nearly unlimited endurance (>10^{16})
- Uses CMOS technology
  - less than 3% more costs
- TMR about 100%
- Dual MTJ
  - less write current density
  - higher TMR



[7]

Conclusion

What have we learned and what can we expect?

Characteristics

Technology	Cell size	State	Access Time (W/R)	Energy/Bit	Retention
DRAM	$6F^2$	Product	10/10 ns	2pJ/bit	64 ms
PRAM	4-5F^2	Prototype	100/20 ns	100 pJ/bit	years
Racetrack	$\frac{20F^2}{DWs} \sim 5\ F^2$	Research	20-30 ns	2 pJ/bit	years
STTRAM	$4F^2$	Prototype	2-10 ns	0.02 pJ/bit	years

[3,6,7,10,11]

- HMC improves the architecture but still rely on DRAM as memory technology
- Energy/Bit is unequal to power consumption! (Interface and control also need power)
- e.g. DRAM cells are very efficient but the interface is power hungry!
- Access time means access to the cell! Latency also depends on access and control logic

Glance into the crystal ball

Technology	Benefits	Biggest challenges	Prediction
PRAM	• High Capacity	• Access Time • Power	Only as hybrid approach or mass storage
HMC	• Huge bandwidth • High capacity	• Fabrication costs	Good chances in near future
Racetrack	• High capacity	• Fabrication • Access time depends on density	Still a lot of research necessary
STTRAM	• Fast access • High density	• Tradoff between Thermal stabiltiy and write current density	Needs also more research

- Prediction is pretty hard
- DRAM will certainly remain as memory technology within this decade
- Every technology has its own challenges

[…] There is no holy grail of memory that encapsulates every desired attribute […]

Dean Klein, VP of Micron's Memory System Development, 2012

[http://www.hpcwire.com/hpcwire/2012-07-10/hybrid_memory_cube_angles_for_exascale.html]

Thank you for your attention!

Questions?

References I

[1] Jacob, Bruce (2009): The Memory System: Morgan & Claypool Publishers

[2] Minas, Lauri (2012): The Problem of Power Consumption in Servers: Intel Inc.

[3] Pawlowski, J.Thomas (2011) Hybrid Memory Cube (HMC): Micron Technology, Inc

[4] Jeddeloh, Joe and Keeth, Brent (2012): Hybrid Memory Cube: New DRAM Architecture Increases Density and Performance: IEEE Symposium on VLSI Technology Digest of Technical Papers

[5] Gao, Li (2009): Spin Polarized Current Phenomena In Magnetic Tunnel Junctions: Dissertation, Stanford University

[6] Qureshi, Moinuddin K. and Gurumurthi, Sudhanva and Rajendran, Bipin (2012): Phase Change Memory: Morgan & Claypool Publishers

References II

[7] Krounbi, Mohamad T. (2010): Status and Challenges for Non-Volatile Spin-Transfer Torque RAM (STT-RAM): International Symposium on Advanced Gate Stack Techology, Albany, NY

[8] Bez, Roberto et al. (2003): Introduction to Flash Memory: Invited Paper, Proceedings of the IEEE Vol 91, No4

[9] Kogge, Peter et al. (2008): ExaScale Computing Study: Public Report

[10] Kryder, Mark and Chang Soo, Kim (2009): After Hard Drives – What comes next?: IEEE Transactions On Magnetics Vol 45, No 10

[11] Parkin, Stewart (2011): magnetic Domain-Wall Racetrack Memory: Scientific Magazine January 14, 2011