Can MRAM be a factor for HPC ?

1. Introduction
2. Can MRAM help ?
3. Which MRAM ?

High Performance Computing

Current HPC → Pétaloflops (10^15 flop/s)

> MW overall operating power consumption + Same amount for cooling
> 100m² area

Towards Exaflop (10^18 Flops/s) requires drastic increase of compactness and energy efficiency

IC Power Consumption

Memory Wall

Memory vs. CPU speed mismatch: Logic keeps awaiting Data!

Memory Hierarchy

→ Logic issue is becoming a memory issue …
Logic Power Losses

Static Loss: Current leakage
- Gate stack structure:
  - Poly-Si
  - Metal Gate
  - High Interfacial layer
- Source-Drain leakage (direct tunneling)
- Gate-Channel tunneling

Dynamic Loss: Interconnects capacitance and Joule heating

Technology
- 90nm: 10^7 # transistors, ~10km wire length
- 65nm: 10^8 # transistors, ~30km wire length
- 45nm: 10^9 # transistors, ~100km wire length
- 32nm: 3.10^9 # transistors, ~300km wire length
- 22nm: 10^10 # transistors, ...

The (Memory) Holy Grail

- Non-Volatile to save data while logic OFF
- Low active power
- Fast enough to match logic speed
- "Infinitely" endurant to act as cache
- Easy to embed within logic
- With minimal wire length to minimize dynamic (RC) loss
- Single technology to answer multiple needs (RAM, ROM, Store)

Can MRAM be of any help?

Can MRAM be of any help?

Why MRAM?

- Non-volatile like Flash 10^4 years retention
- Dense like DRAM 10^12, small overheads
- Fast like SRAM ~10ns in normal mode
- High Endurance like SRAM / DRAM 10^12 cycles, up to 10^14

MRAM is not the best but ...
- Can replace SRAM at 1/6th of size, zero leakage
- Can replace e-Flash at >10^4 speed, lower power
- Can replace DRAM (if running out of steam)
An easy to embed memory

- "End-of-back-end" process (above-IC)
- Cell R compatible with CMOS (~ kΩ)
- Vdd driven switching
- No charge pumps required
- No trade-off with logic process
- Cheap (only 3 add-masks)

MRAM Cache

Option 1: DRAM & L2, L3 cache replacement @ same overall architecture

- Reduced Static power, (NV cache, no DRAM refresh)

Option 2: Memory blocks distributed within (above) logic core(s)

- Faster memory-logic communication
- Reduced dynamic power
- Advanced power management
- High data resilience

"Janus" architecture

- Reduced silicon footprint
- Multiple / short interconnects
- Distributed memory within logic

Toshiba Proposes MRAM/SRAM Hybrid Cache for Normally-off Computers

MRAM Cell

- High data resilience
eVaderis STT-based MCU for the IoT

Connected Object

Performances, Intelligence
(computing, amount of data)

Autonomy
(battery life, CO2)

Battery
Wireless
Sensor
Capture

10 to 100X less energy translates into
extended lifetime, more intelligence, less
CO2

Non-volatile data-centric control processor

“everywhere”

Energy

On-Chip Processing-Storage

Near-zero standby

Option 3: Non-volatility inside logic blocks (NV-Flip-flop, NV-latch, ...)

Fast save / restore of logic states

⇒ "Normally-OFF / Instant-ON" Computing

User Case
19 Mb MRAM + 1Mb SRAM
• 10^6 scalar measures
• 32 uncompressed 320x240 grayscale pictures (security standard)

Divide by 10 to 100 (parallelized boot from distributed memory blocks)

Divide by 100 X faster (no more sleep/deep sleep states)

Instant on → full shutoff

Send Store Wake-up Sleep Standby

Divide by 3 to 100 (depends on data profile)

Balance of RF gain and on-chip process cost

Relative Energy

Hybrid CMOS-MRAM logic

NV Flip-Flop

NEC develops a nonvolatile magnetic flip flop that enables standby-power-free SoCs

Empowered by Innovation

5 Jan 2009
DRAM-based Configuration memory

Periodic refresh of DRAM using MRAM content (scrubbing)

Advantages:
- High density (DRAM)
- No redundancy required
- Shadowed reconfiguration
- Low power (non-volatile)

Implemented on hybrid TowerJazz 130nm CMOS / Crocus MRAM process

There Are Many MRAMS!

Thermally Assisted (TAS)

Field-driven Toggle

Perpendicular STT (SPRAM)

DW motion

SOT (Spin-Orbit Torque)

OST (Precessional)

The Magnetic Tunnel Junction (MTJ)

Giant (Tunneling) Magnetoresistance
Acting on current through magnetization

Parallel “0” - Low R

Antiparallel “1” - High R

Which MRAM?

1Kb Ferrite core Control Data Corp (1965)

1Mb Bubble memory Intel (1980)

16Mb AMR Honeywell (1984)

4Mb Toggle Everspin (2004)

64Mb STT Hynix (2010)

R = \frac{R_{11} - R_{1\bar{1}}}{R_{11}}

H

TMR

Hard mask Ta

Etch stop layer Ru

Capping layer Ta

Storage layer CoBi

Tunnel barrier (MgO)

Reference layer CoO

Spacer (Ta)

Pinning layer MgO

MgO (Pt/Co)

Seed layer (Pt/Co)

Smoothing layer Ta/CuN

Base electrode Ta
MRAM (Read)

1T-1R architecture
Logic state = Magnetization (resistance) state

Figure of merit is $\Delta R/\sigma$ not $\Delta R$

Field-Driven MRAM

Use current pulses to generate overlapping magnetic fields at word/bit line crosspoint

Large cell size (30F²)
High power (2x16 mA / bit)
Low speed (35ns R/W)
Not scalable

Why STT MRAM?

Smallest cell size
Lowest current
Full scalability

What is STT?

Current-only switching
(no field)

$dM/dt = -\gamma M \times (H_{\text{eff}} + bI.M) + \gamma aI.M \times (M \times M_\perp) + \alpha M \times dM/dt$

Spin & Torque & Field

Gilbert Damping
Spin torque
Field torque
Zeeman

Field-driven MRAM

STT MRAM

Write Current is $I_{\text{inw}} + I_{\text{inl}}$
Write Current vs. cell area
Write Current vs. TMR Device Width

Why STT MRAM?

Smallest cell size
Lowest current
Full scalability

What is STT?

Current-only switching
(no field)

$dM/dt = -\gamma M \times (H_{\text{eff}} + bI.M) + \gamma aI.M \times (M \times M_\perp) + \alpha M \times dM/dt$

Spin & Torque & Field

Gilbert Damping
Spin torque
Field torque
Zeeman
Why perpendicular?

\[ \tau = \tau_0 e^{\Delta \tau_{\text{B}}} \]

\[ \tau_0 = 10^{-9} \text{s} \]

Barrier to switching

Thermal activation

\[ \tau_0 = 10^{-9} \text{s} \]

Thermal stability factor

\[ \Delta = \frac{K_{\text{eff}} V}{\kappa B T} \]

\[ K_{\text{eff}} = K_x - 2\pi M_s^2 \]

\[ K_{\text{para}} = \left( \frac{K_x + K_{\text{eff}}}{} \right) + K_y \]

Critical current

\[ j_{\text{c,para}} = \left( \frac{4e}{h} \right) \frac{\alpha k_B T}{g(0)pA} \left( \Delta + \frac{\pi M_s^2 V}{K_{\text{eff}} T} \right) \]

\[ j_{\text{c,para}} = \left( \frac{4e}{h} \right) \frac{\alpha k_B T}{g(0)pA} \Delta \]

Switching probability

\[ P_s = 1 - e^{-\frac{\Delta}{\tau}} \]

P-STT MRAM Demo

IEEE MAGNETICS LETTERS, Volume 2 (2011)

Demonstration of Ultralow Bit Error Rates for Spin-Torque Magnetic Random-Access Memory With Perpendicular Magnetic Anisotropy

IBM-MagC MRAM Alliance, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

P-STT MRAM is now becoming real!

Q3, 2014

Everspin and Globalfoundries team up on MRAM

Non-volatile memory for the masses or at least mass production

From Q3, 2014

Move over flash? TDK reveals its first next-gen MRAM memory prototype

Q4, 2014

I_c - I_{c,0} \propto \frac{1}{\tau}

I_c - I_{c,0} = \left[ 1 - \frac{k_B T}{E} \ln \frac{\tau}{\tau_0} \right] \frac{1}{\tau}

How Fast Can STT-MRAM Be?

STT-MRAM is now becoming real!

Q3, 2014

Everspin to Demonstrate Spin-Torque MRAM with Altera FPGA at Flash Memory Summit 2014

Sendohan Americas Ultra Low Latency, Increased Reliability with High Cycling Endurance

Q3, 2014

STT-MRAM Memory

STT-MRAM is now becoming real!

Q3, 2014

Everspin and Globalfoundries team up on MRAM

Non-volatile memory for the masses or at least mass production

From Q3, 2014

Move over flash? TDK reveals its first next-gen MRAM memory prototype

Q4, 2014

I_c - I_{c,0} \propto \frac{1}{\tau}

I_c - I_{c,0} = \left[ 1 - \frac{k_B T}{E} \ln \frac{\tau}{\tau_0} \right] \frac{1}{\tau}

How Fast Can STT-MRAM Be?
Thermally Activated Switching

\[ \Gamma = a_j M \times (P \times M) \]

- Stochastic reversal
- Incubation time preceding a large thermal fluctuation

\[ \frac{d M}{dt} = -\gamma_0 M \times H_{eff} + \frac{1}{M_s} \left( M \times \frac{d M}{dt} \right) + a_j M \times (A \times M) + a_{ja} M \times (P \times M) \]

\[ \text{Precessional STT demo} \]

\[ \text{Spin Orbit Torque (SOT) MRAM} \]

- 3-Terminals
- Infinite Endurance / Reliability
- Independent Read and Write paths
  - Adjustable Impedance
  - Maximized TMR
  - No read disturb
- High speed?
SOT Fast Switching

Moore’s Law

SRAM cell size

Power density
New for Knights Landing
(Next Generation Intel® Xeon® Processor)

Compute: Intel® Silvermont Arch, dual “very Epyc”
- Low-Power Cores with HPC Enhancements®
- 3X Single Thread Performance!* on Pole Epyc
- Intel Xeon Processor Binary Compatability®

On-Package Memory: High Performance:
- up to 16GB @ 2133MHz
- 1/3X the Space®
- 5X Bandwidth vs DDR4®
- 5X Power Efficiency®

Jointly Developed with Micron Technology®

Platform Memory: DDR4 Bandwidth and Capacity Comparable to Intel® Xeon® Processor

Technology enabler: The STT-RAM

Endurance (cycles)

Write speed (sec)

Energy/Jbit

MRAM

PCRAM

STT-RAM

FeRAM

MRAM

PCRAM

STT-RAM

FeRAM

MCU / SoC Implementation

MRAM performances may be tuned by shape / size (same core technology)
⇒ Replace simultaneously multiple memory instances

(Almost) All Non-Volatile Data Path!
- Hybrid TowerJazz 130nm CMOS / Crocus-MRAM process

- Digital Test at Spintec
  - MRAM programming and input transferred to DRAM
  - All inputs combinations tested and corresponding output checked