Non Volatile Main Memory for Handheld Devices: An idea whose time has come

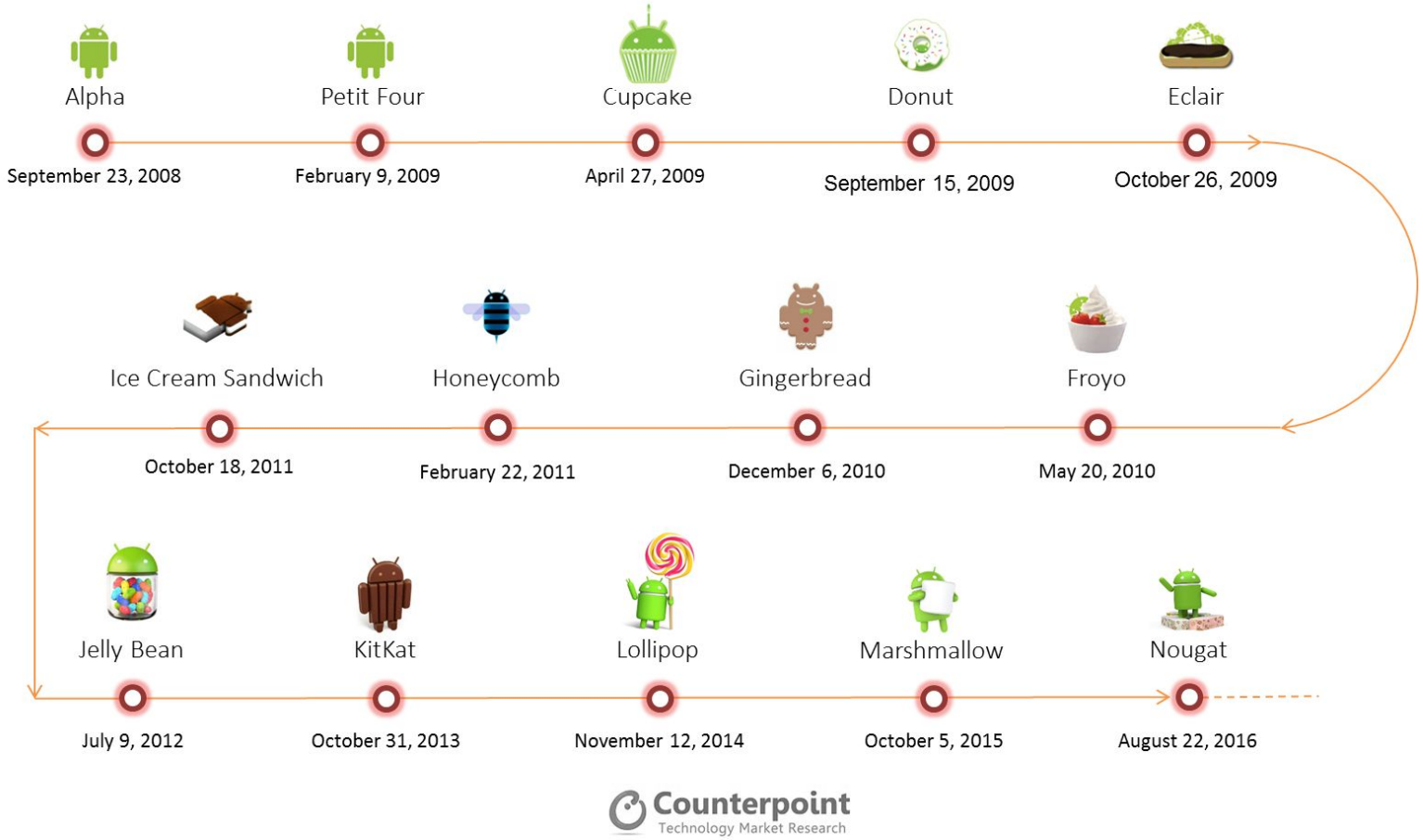
Manu Awasthi

Associate Professor

Ashoka University

Handheld Devices

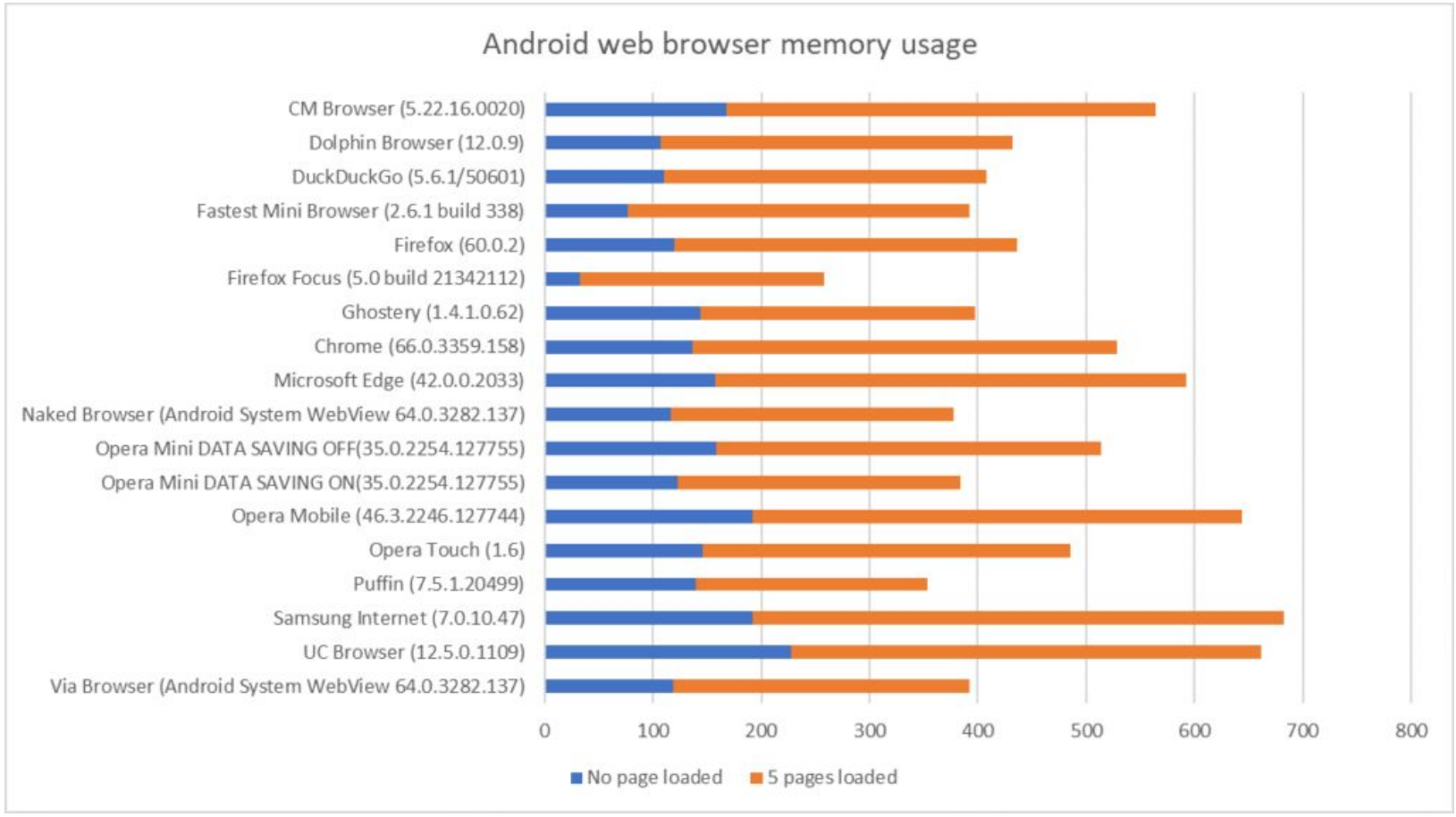
Android Versions



System software is evolving rapidly : Android has 1 release / year

https://www.counterpointresearch.com/can-android-o-de-fragment-android/

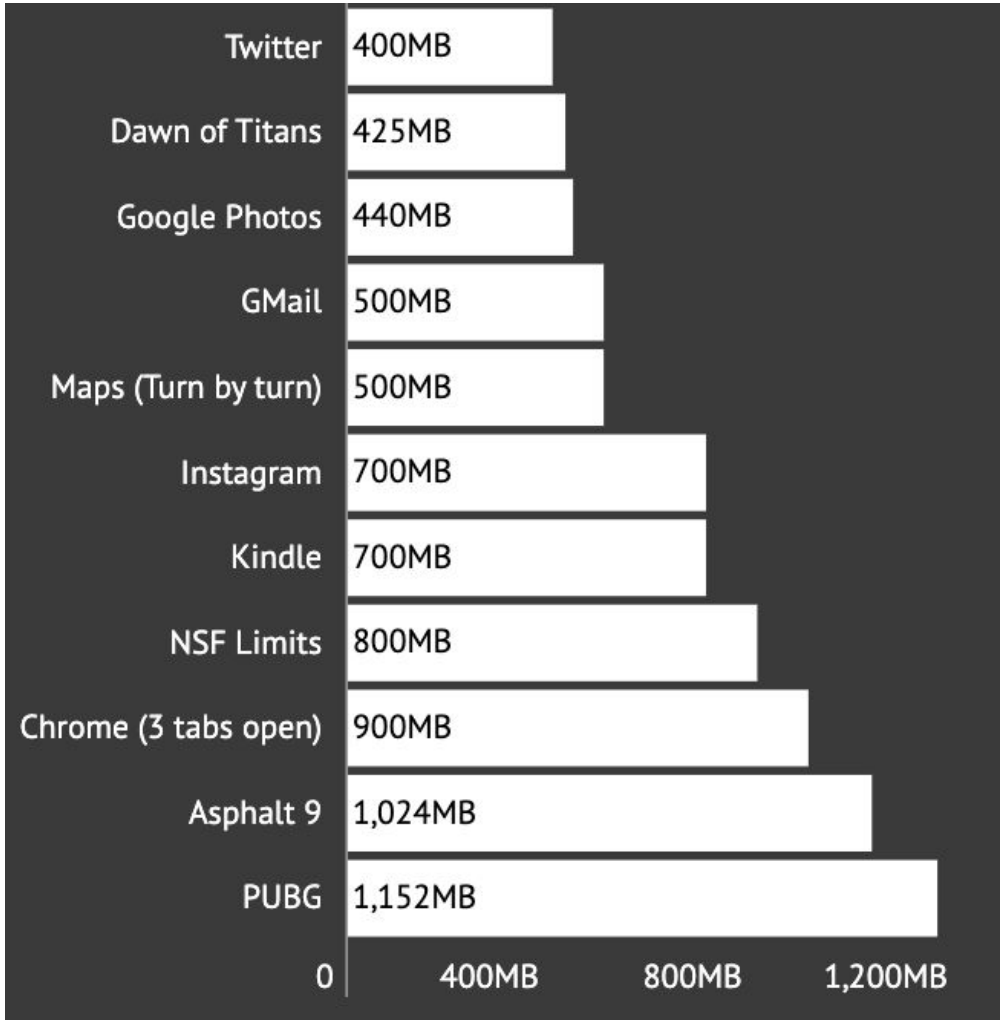
Memory Capacity : Individual Usage



Applications are becoming feature rich, with increasing memory capacity requirements

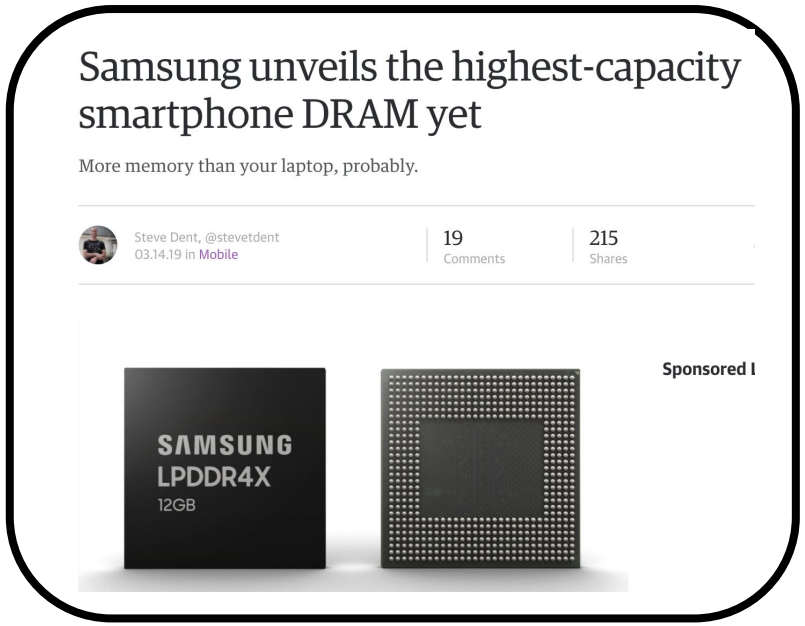
https://eitik.com/17-android-browser-tested-for-memory-usage-in-2018/

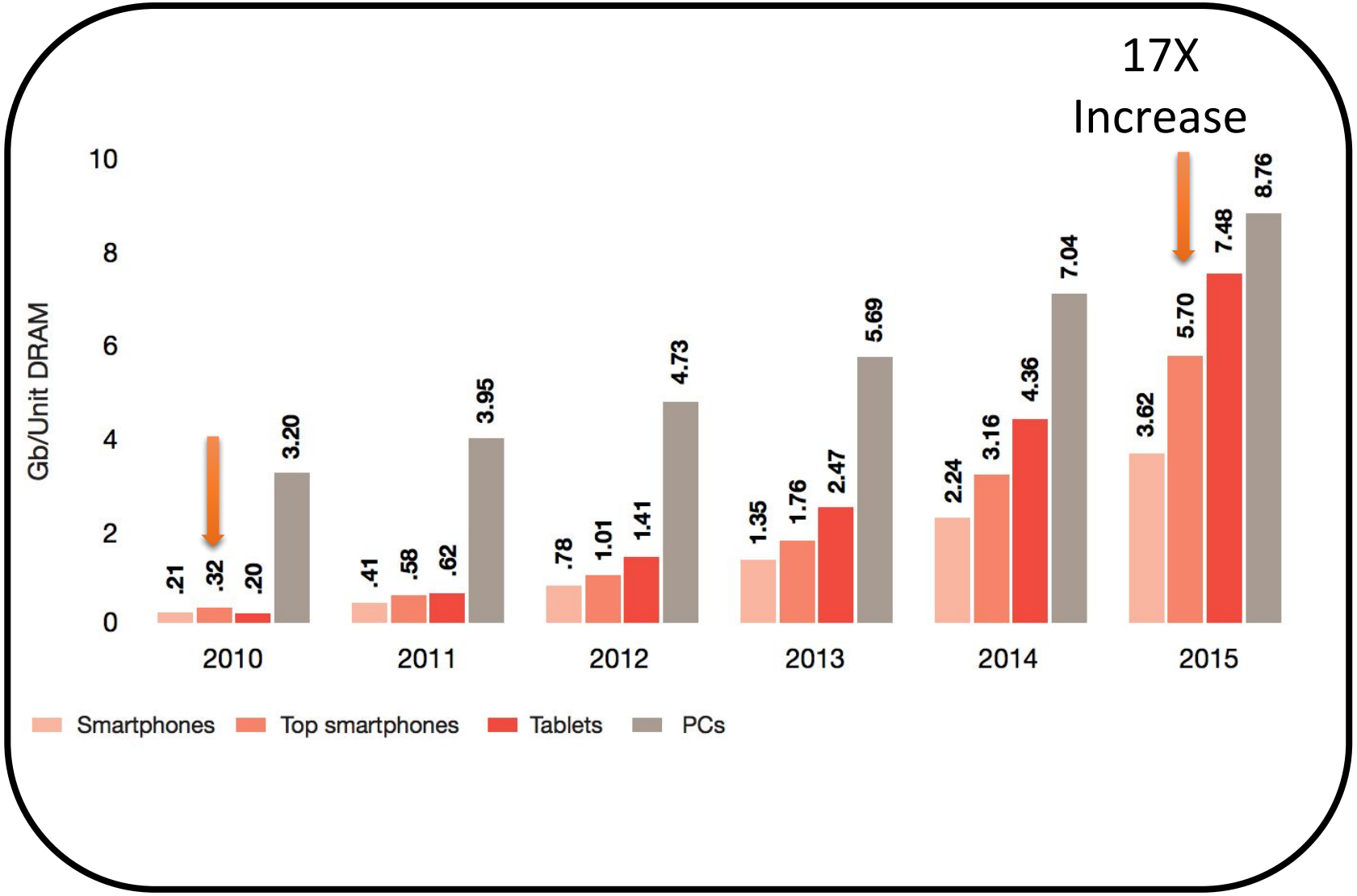
Memory Capacity : Individual Usage



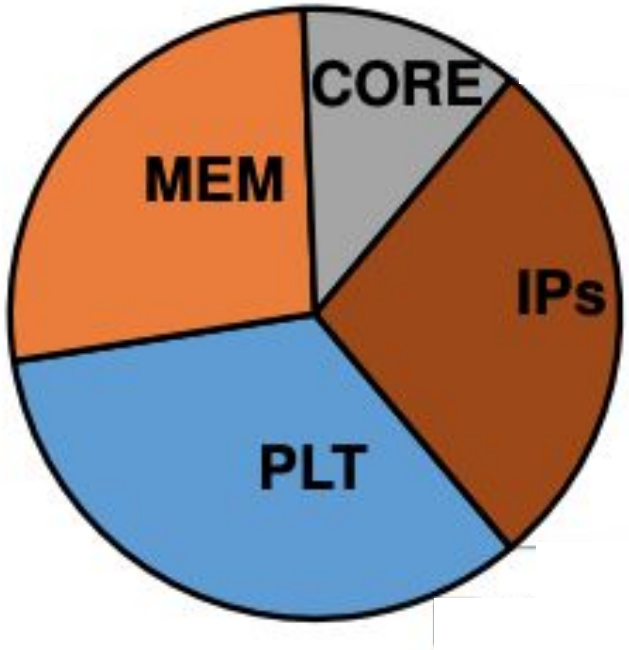
https://www.androidauthority.com/how-much-ram-do-you-need-in-smartphone-2019-944920/

Market Trends : Memory Capacity

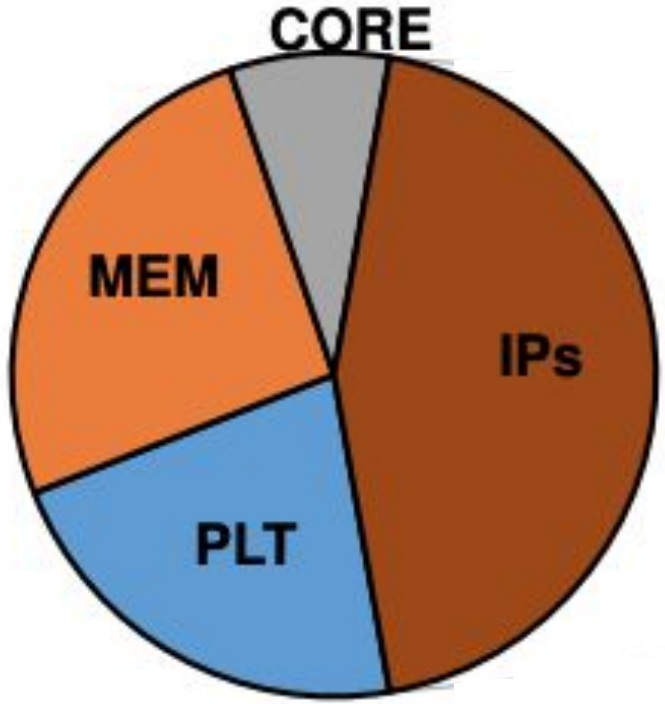




Energy Consumption



Skype



MP Game

Domain Knowledge Based Energy Management in Handhelds, Nachiappan et al. HPCA 2015

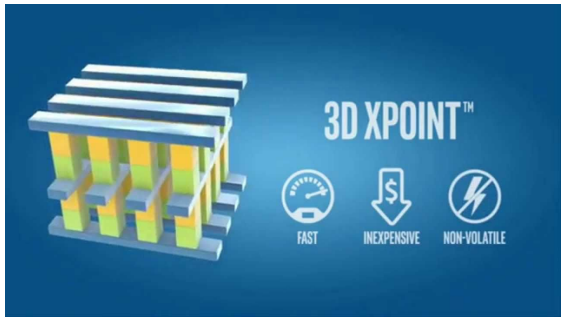
Summary of Trends

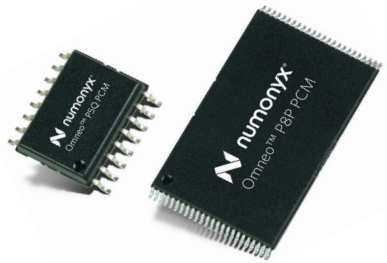
- Handheld applications are becoming complex and feature rich
  - Larger working sets
  - Much higher bandwidth capacity needs, especially when multiprogramming

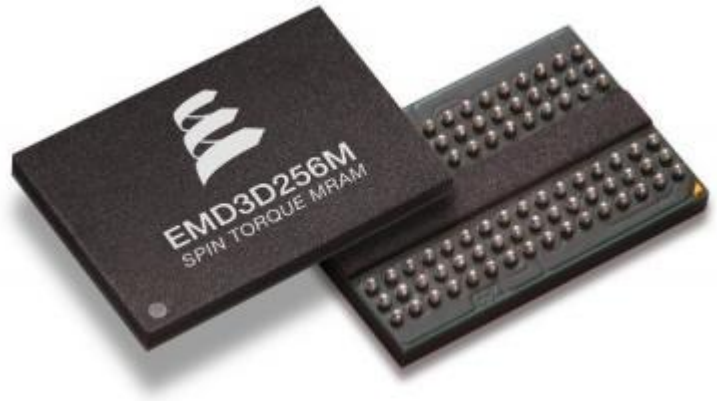
- Memory sub systems can consume less energy
  - Fraction of energy consumed by memory subsystem is growing

Both are (somewhat) contradictory goals

Non Volatile Memory Technologies







- Been around since 1960s, renewed interest with the projected decline of DRAM
- Many candidates : Phase Change Memory (PCM), Spin-Torque Transfer Memory (STT-RAM), 3D-Xpoint, Resistive RAM (ReRAM) etc.
- Vary based on underlying mechanism for storing information

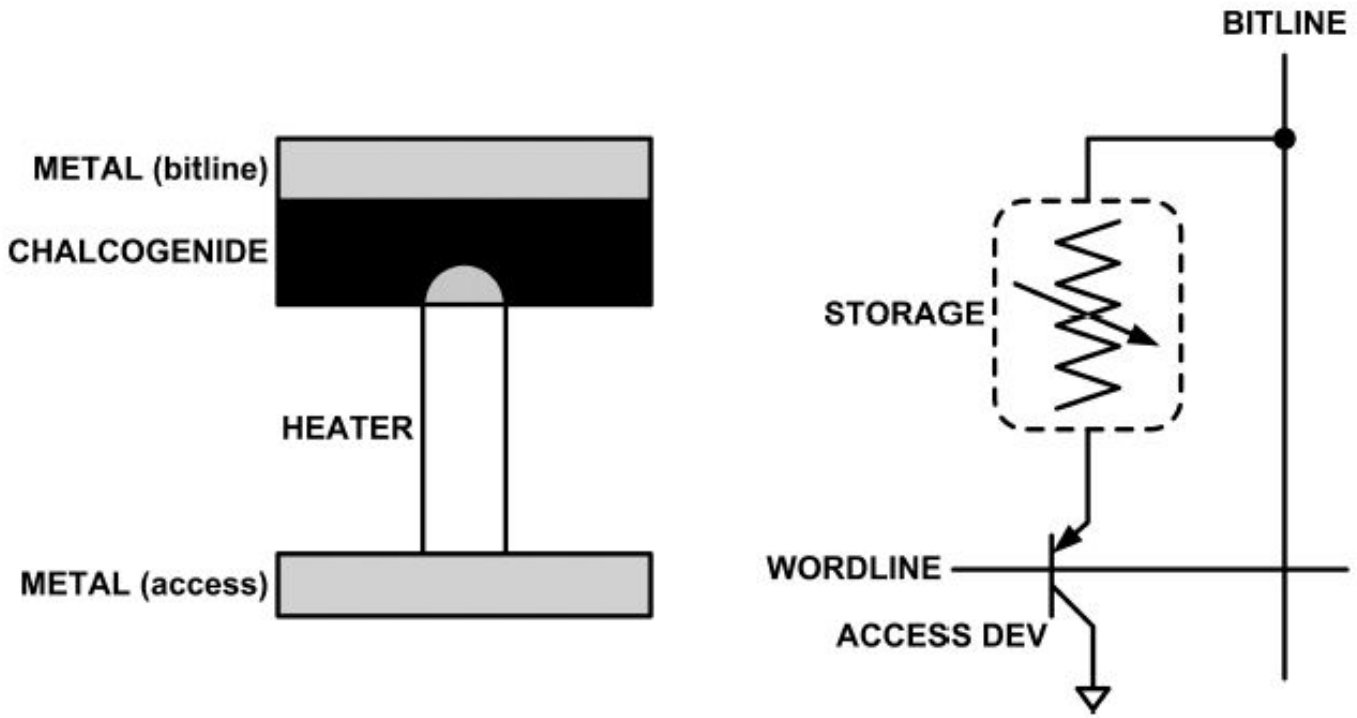
Non Volatile Memories

+ Many candidates: PCM, STT-MRAM, others
+ Higher areal density : 2x – 4x compared to DRAM
+ Lower access energies
+ No Refresh
- Higher access latencies
- Asymmetric read / write energies
- Reduced endurance

	Cell size	Access Granularity	Read Latency	Write Latency	Erase Latency	Endurance	Standby Power
HDD	N/A	512 B	5 ms	5 ms	N/A	$> 10^{15}$	1W
SLC Flash	$4–6F^2$	4 KB	25 μs	500 μs	2 ms	$10^4 - 10^5$	0
DRAM	$6–10F^2$	64 B	50 ns	50 ns	N/A	$> 10^{15}$	Refresh power
PCM	$4–12F2$	64 B	50 ns	500 ns	N/A	$10^8 - 10^9$	0
STT-RAM	$6–50F^2$	64 B	10 ns	50 ns	N/A	$> 10^{15}$	0
ReRAM	$4–10F^2$	64 B	10 ns	50 ns	N/A	10^{11}	0

A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems, Mittal et al., IEEE TPDS 2016

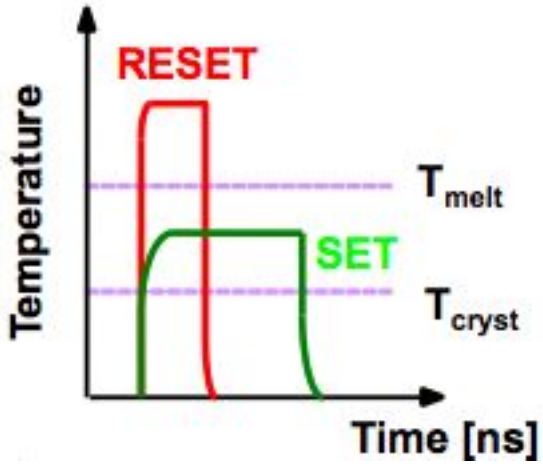
PCM Primer

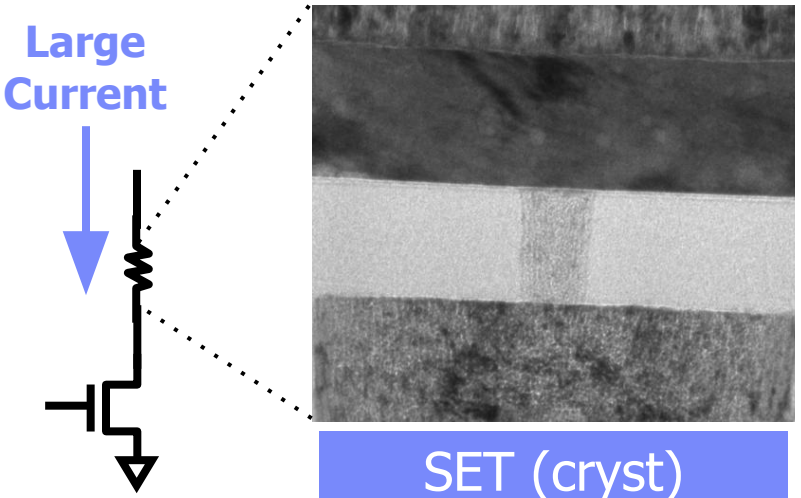


PCM is resistive memory: High resistance (0), Low resistance (1)
PCM cell can be switched between states reliably and quickly

PCM Working Example

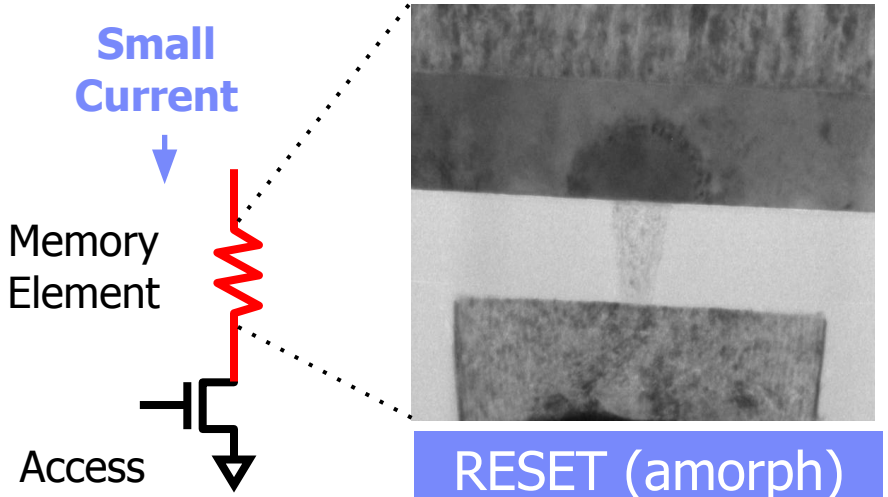
- Write: change phase via current injection
  - **SET**: sustained current to heat cell above T_{cryst}
  - **RESET**: cell heated above T_{melt} and quenched
- Read: detect phase via material resistance





SET (cryst)
Low resistance

10^3-10^4 Ω



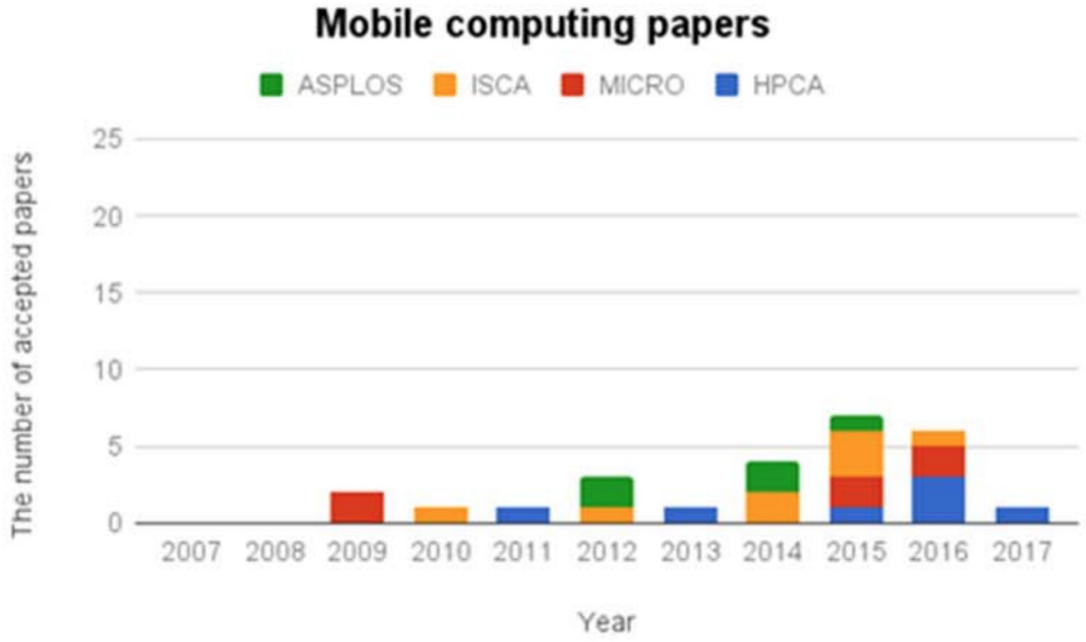
RESET (amorph)
High resistance

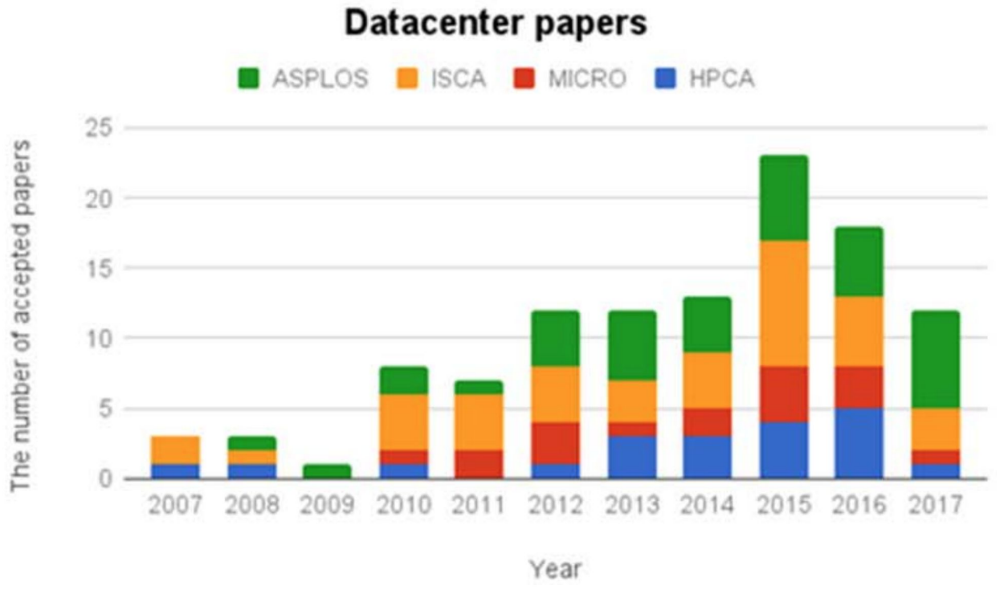
10^6-10^7 Ω

Slide Courtesy: Moinuddin Qureshi, IBM

Mobile Architecture Research

- Mobile computing research: 1% of research papers published each focus on mobile computing.
- **Lack of tools**

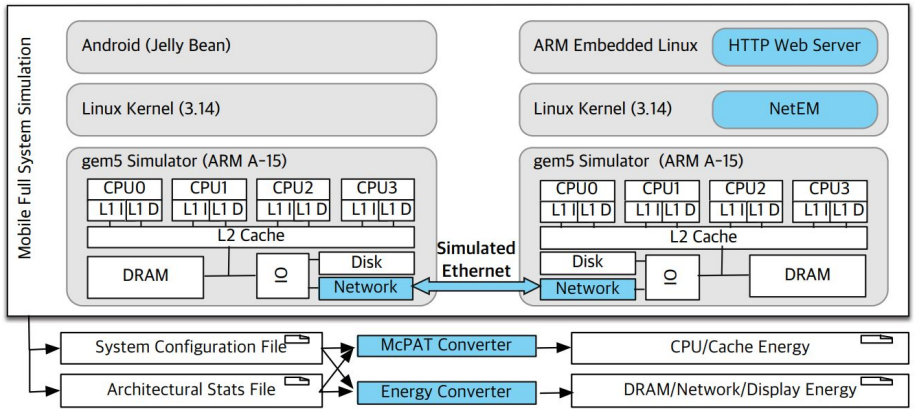


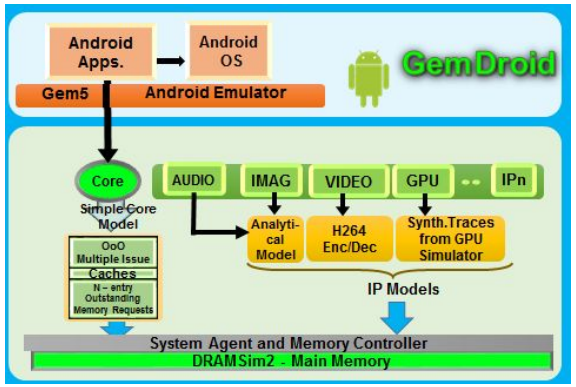


V. J. Reddi, H. Yoon, and A. Knies, “Two billion devices and counting,” IEEE Micro, vol. 38, no. 1, pp. 6–21, January/February 2018.

What is needed?

- Tool and Simulators





- Benchmarks

BBench
(Michigan)

AsimBench
(ICT, China)

Bench	Category
BBench*	Web Browser
K9Mail	Email
SinaWeibo	Social Network
NeteaseNews	News
KingsoftOffice	Document
AdobeReader	Document
BaiduMap	Map
MXPlayer	Video
TTPod	Audio
FrozenBubble	Game

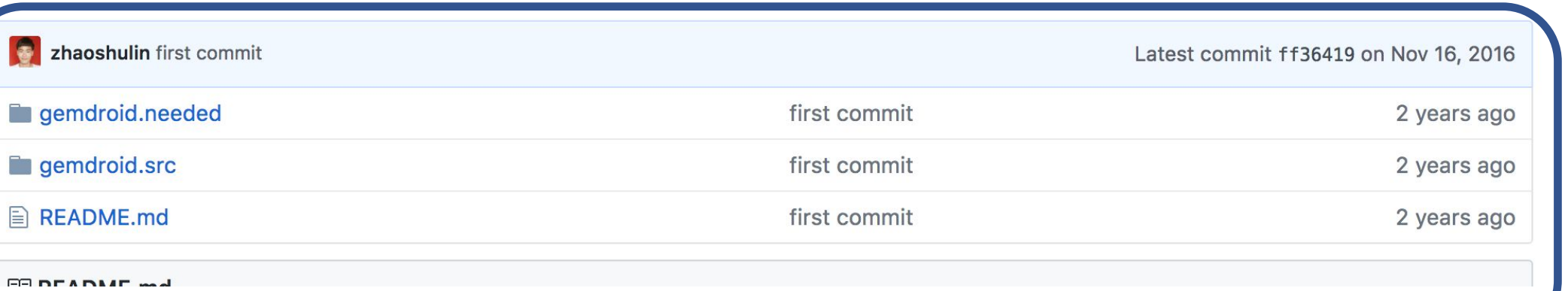
Current Status

Last updated	2013-04-19
Website	http://bbench.eecs.umic...
Language	JavaScript
Access level	Read

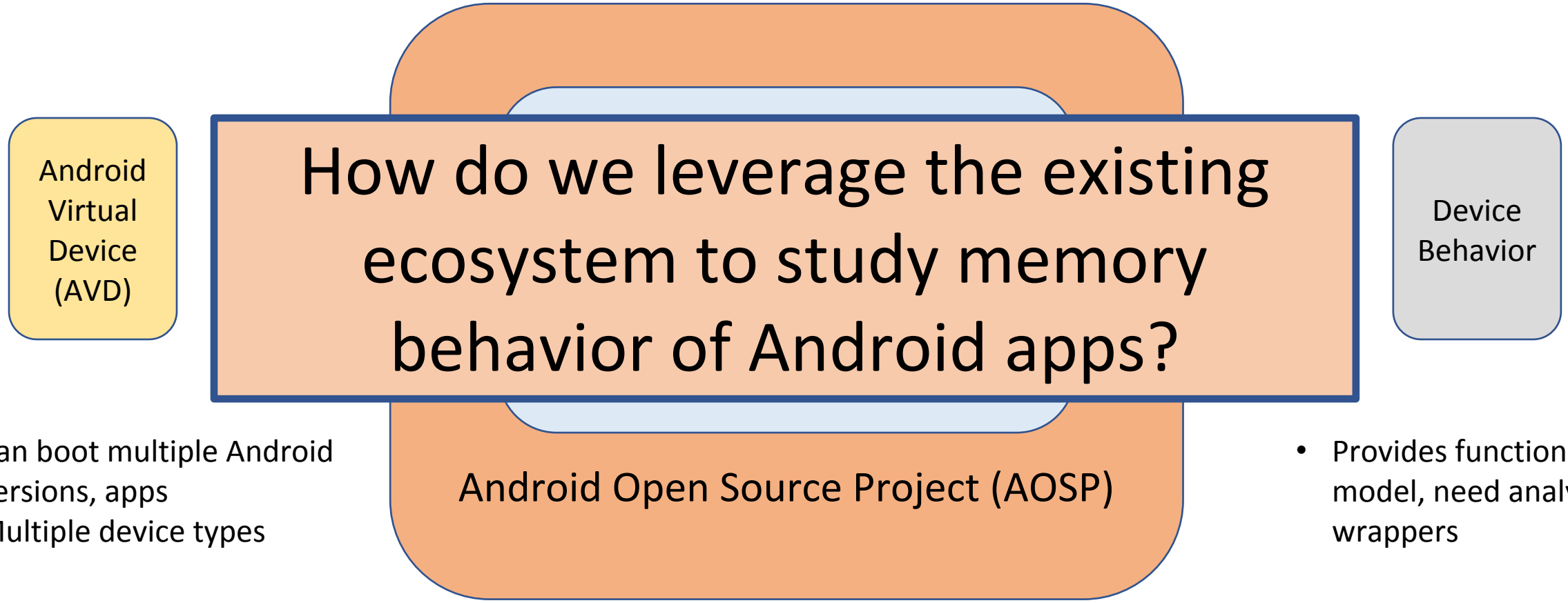
Android ICS Disk Image with BBench
Android Kernel 2.6.35

News

- 2013-12-16 AsimBench is renamed as Moby!
- 2013-12-11 Our paper introducing AsimBench is accepted by ISPASS'2014!
- 2013-08-21 AsimBench v2.0 is now released!



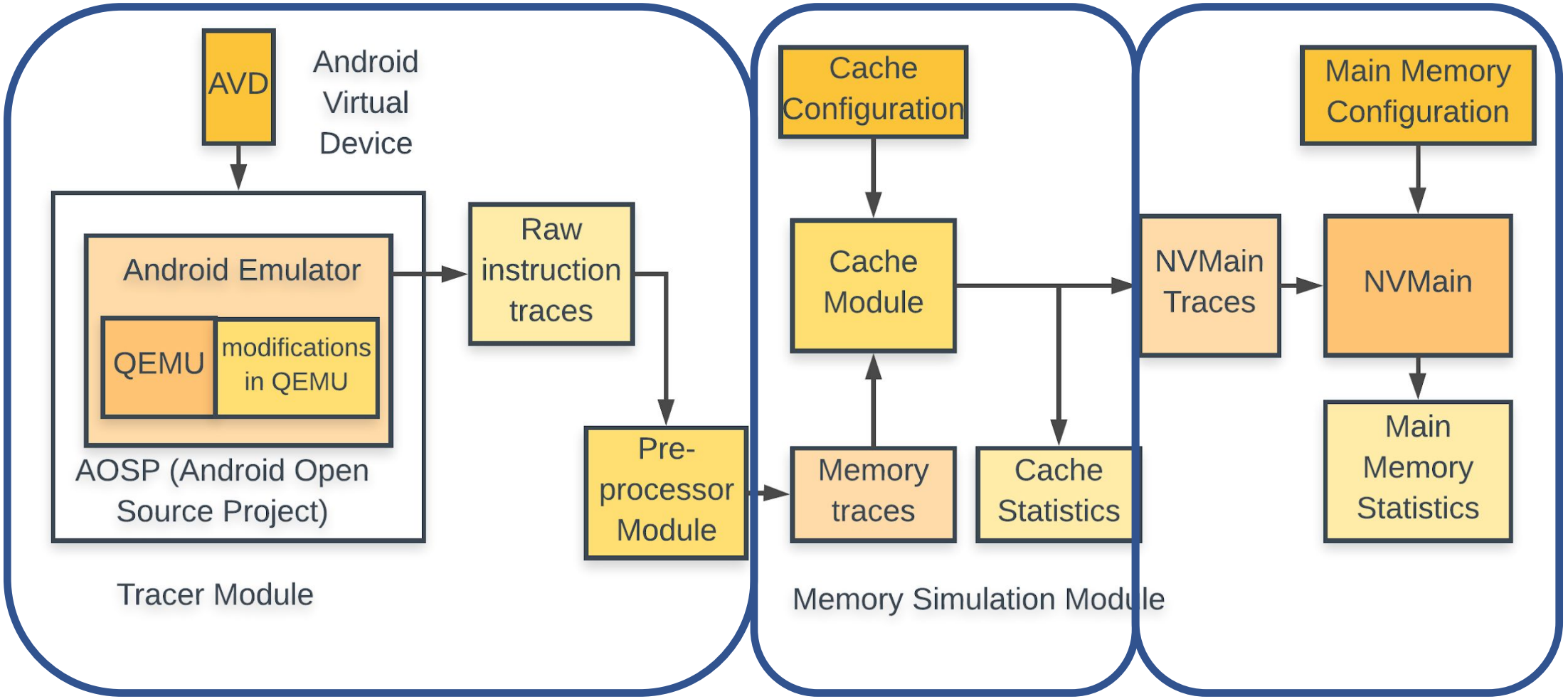
Android Emulator



- Can boot multiple Android versions, apps
- Multiple device types

- Provides functional model, need analysis wrappers

META: Tool Design



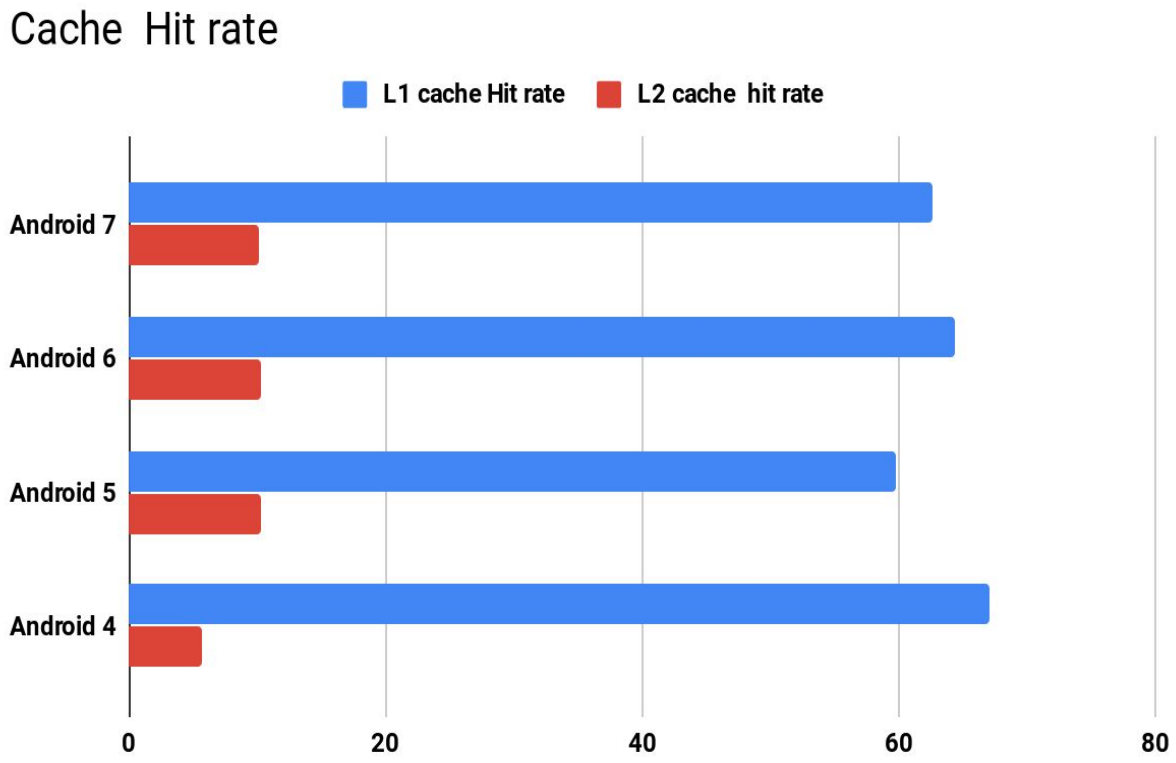
Raw Traces

```
##########
CPU ID = 1
Reg 0 = ee0762e8
Reg 1 = ee0762e8
Reg 2 = c054191c
Reg 3 = c053f280
Reg 4 = ee0762e8
Reg 5 = c054d6c0
Reg 6 = ee3c8000
Reg 7 = 101
Reg 8 = c02e9e48
Reg 9 = 200
Reg 10 = 0
Reg 11 = 0
Reg 12 = 0
Reg 13 = 9f0a5628
Reg 14 = ad314247
0xc02e9e78:  e59430ac ldr r3, [r4, #172]
0xc02e9e7c:  e3530000 cmp r3, #0    ; 0x0
0xc02e9e80:  0a000032 beq 0xc02e9f50
##########
```

Cache Simulation Module

```
[
{
    "level" : 1,
    "size"  : 32768,
    "associativity" : 8,
    "sets" : 64,
    "read_time" : 1,
    "write_time" : 2
},
{
    "level" : 2,
    "size"  : 262144,
    "associativity" : 4,
    "sets" : 1024,
    "read_time" : 10,
    "write_time" : 15
}
]
```

Cache hierarchy specification



L1/L2 Cache hit rates: Calculator App Android 4 (Kitkat) to Android 7 (Nougat).

Main Memory Simulation Module : NVMain

- NVMain : cycle-level main memory simulator
- Can simulate DRAM, emerging NVMs
  - DRAM variants : LPDDRx, DDRx
  - Emerging memory technologies: PCM, STT-RAM etc.
- Statistics on memory latencies, bandwidth, utilizations etc.

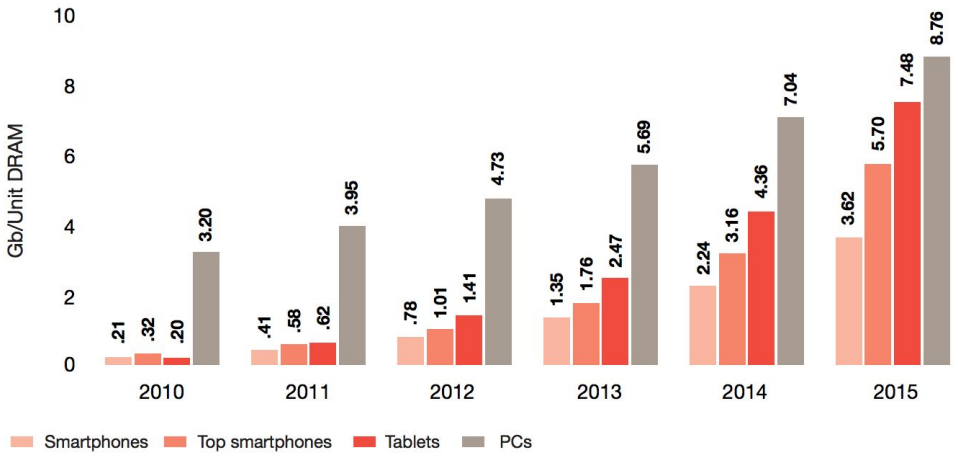
```
channel0.rank0.bank1.bankEnergy 165824mA*t
channel0.rank0.bank1.activeEnergy 81459mA*t
channel0.rank0.bank1.burstEnergy 35145mA*t
channel0.rank0.bank1.refreshEnergy 49220mA*t
channel0.rank0.bank1.bankPower 0.00409577W
channel0.rank0.bank1.activePower 0.002012W
channel0.rank0.bank1.burstPower 0.000868064W
channel0.rank0.bank1.refreshPower 0.00121571W
channel0.rank0.bank1.bandwidth 3239.79MB/s
channel0.rank0.bank1.dataCycles 2308
channel0.rank0.bank1.powerCycles 60730
channel0.rank0.bank1.utilization 0.0380043
channel0.rank0.bank1.reads 367
channel0.rank0.bank1.writes 210
channel0.rank0.bank1.activates 431
channel0.rank0.bank1.precharges 430
channel0.rank0.bank1.refreshes 23
channel0.rank0.bank1.activeCycles 51643
channel0.rank0.bank1.standbyCycles 9087
channel0.rank0.bank1.fastExitActiveCycles 0
channel0.rank0.bank1.fastExitPrechargeCycles 0
channel0.rank0.bank1.slowExitPrechargeCycles 0
channel0.rank0.bank1.actWaits 0
channel0.rank0.bank1.actWaitTotal 0
channel0.rank0.bank1.actWaitAverage -nan
channel0.rank0.bank1.averageEndurance 0
channel0.rank0.bank1.worstCaseEndurance 18446744073709551615
```

META : Potential Use Cases

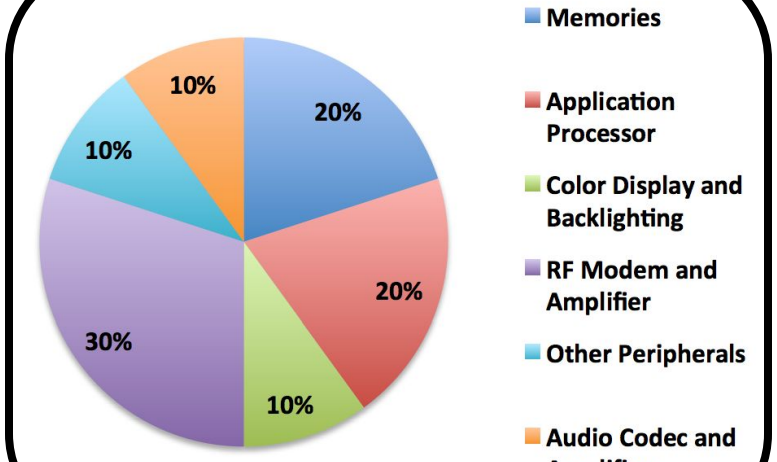
- Trace Generation
  - The traces can also be used to analyze instruction distribution profile.
  - Creation of synthetic inputs to models based on real instruction profiles
- Cache Hierarchy Modeling
  - A custom, N-level cache hierarchy
- DRAM, Non-volatile, Hybrid Memory Simulation
  - NVMain can model most technologies

Trends in Handheld Devices

https://thehackernews.com/2015/09/6gb-ram-smartphone.html



https://www.pwc.com/gx/en/technology/mobile-innovation/assets/pwc-dram-memory.pdf



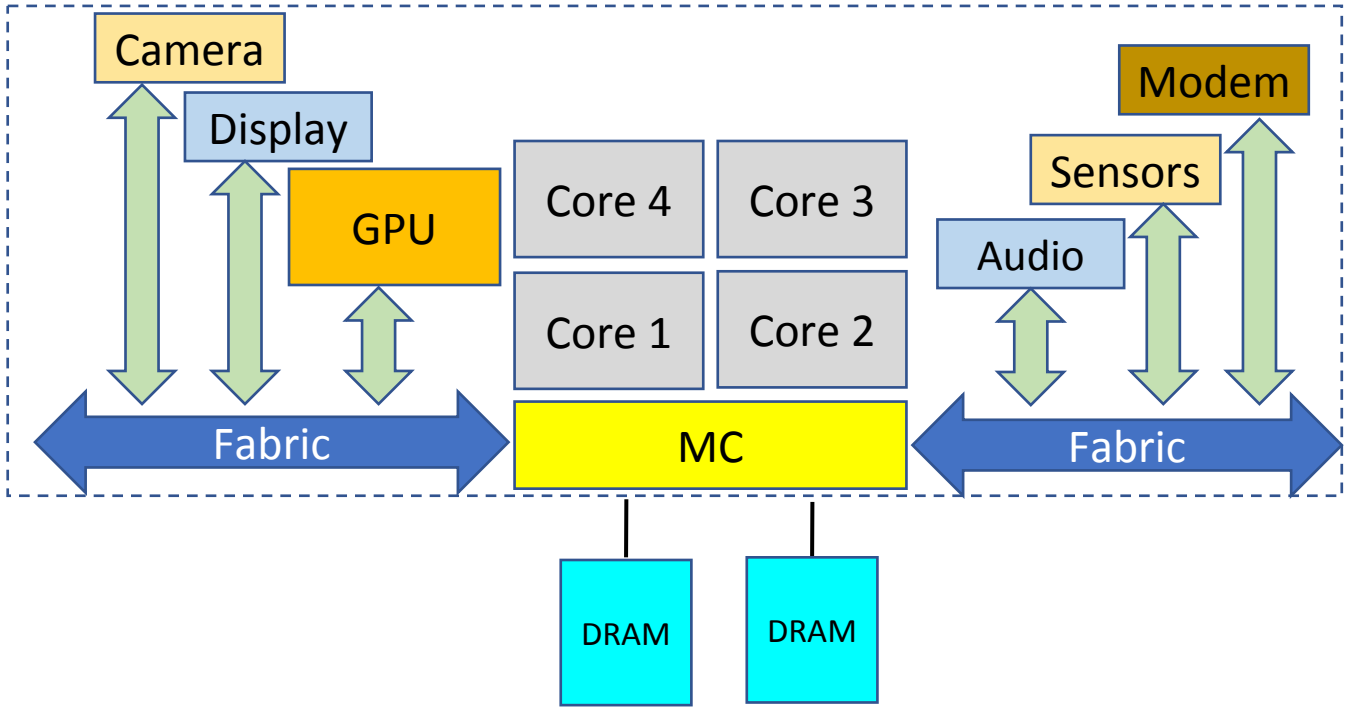
http://www.es.ele.tue.nl/~kgoossens/Chandrasekar14PHD.pdf

Requirements of Handheld Devices

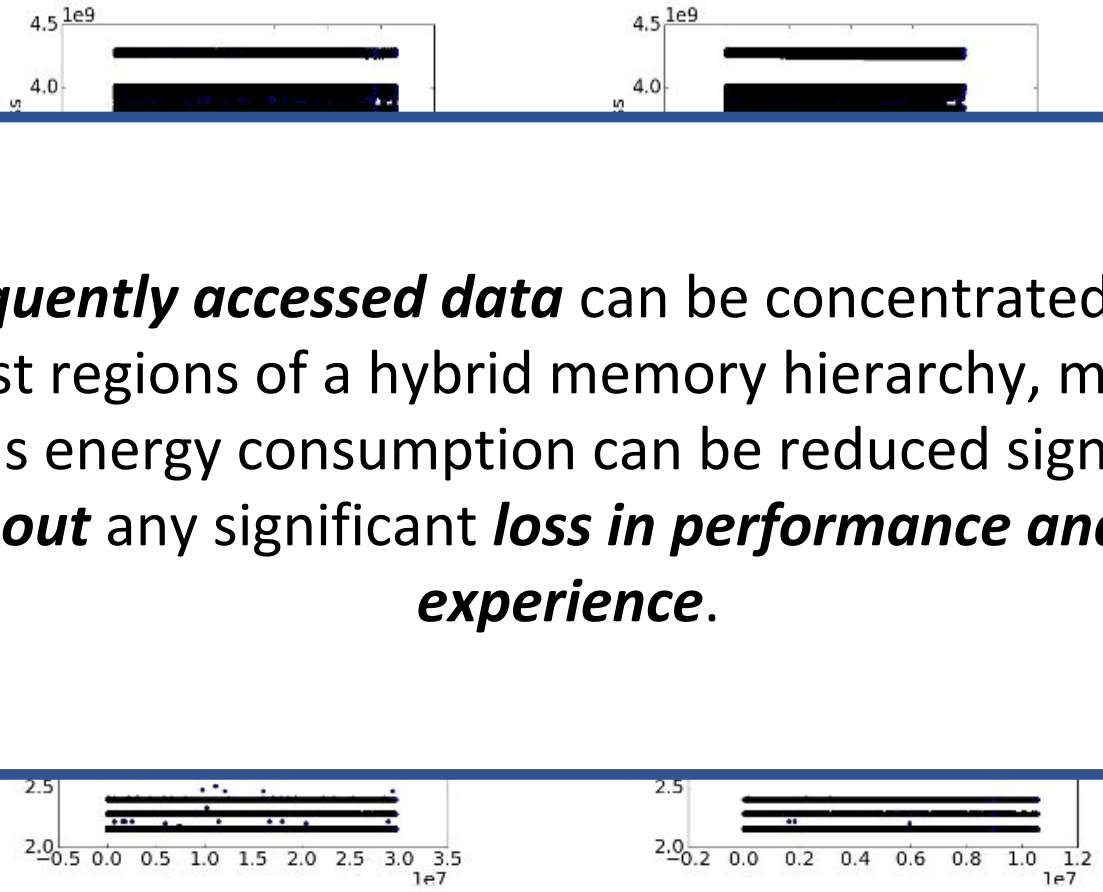
- Response Time
  - Most devices are for information consumption
  - Delays will hinder user engagement
- Energy Efficiency
  - Battery life is of paramount importance
- Increased need for memory capacity

- NVM Technologies cannot be used as is : Need architectural exploration of characteristic comparisons

Main Memory in Handhelds



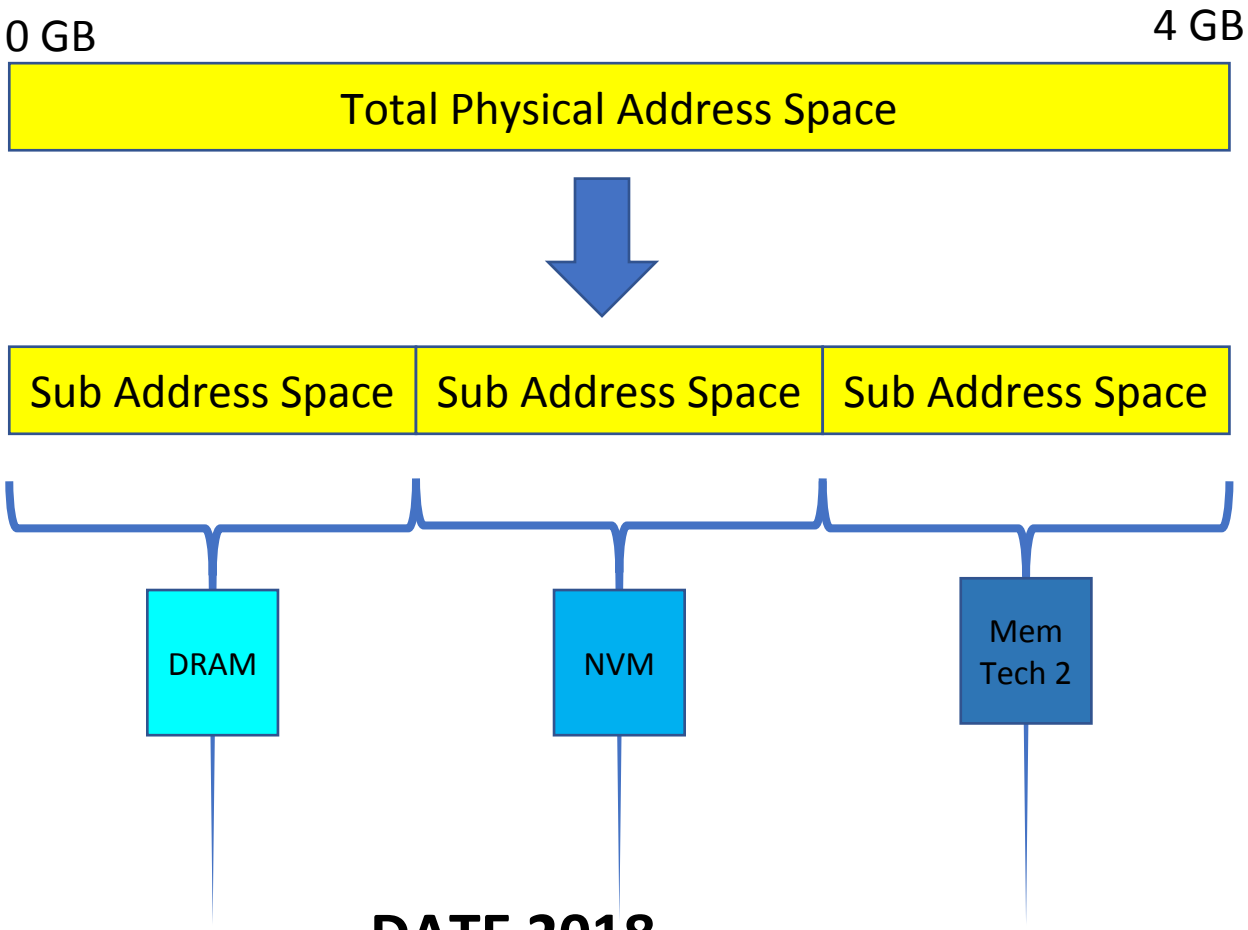
Handheld Applications



(c) Netease (d) Baidu

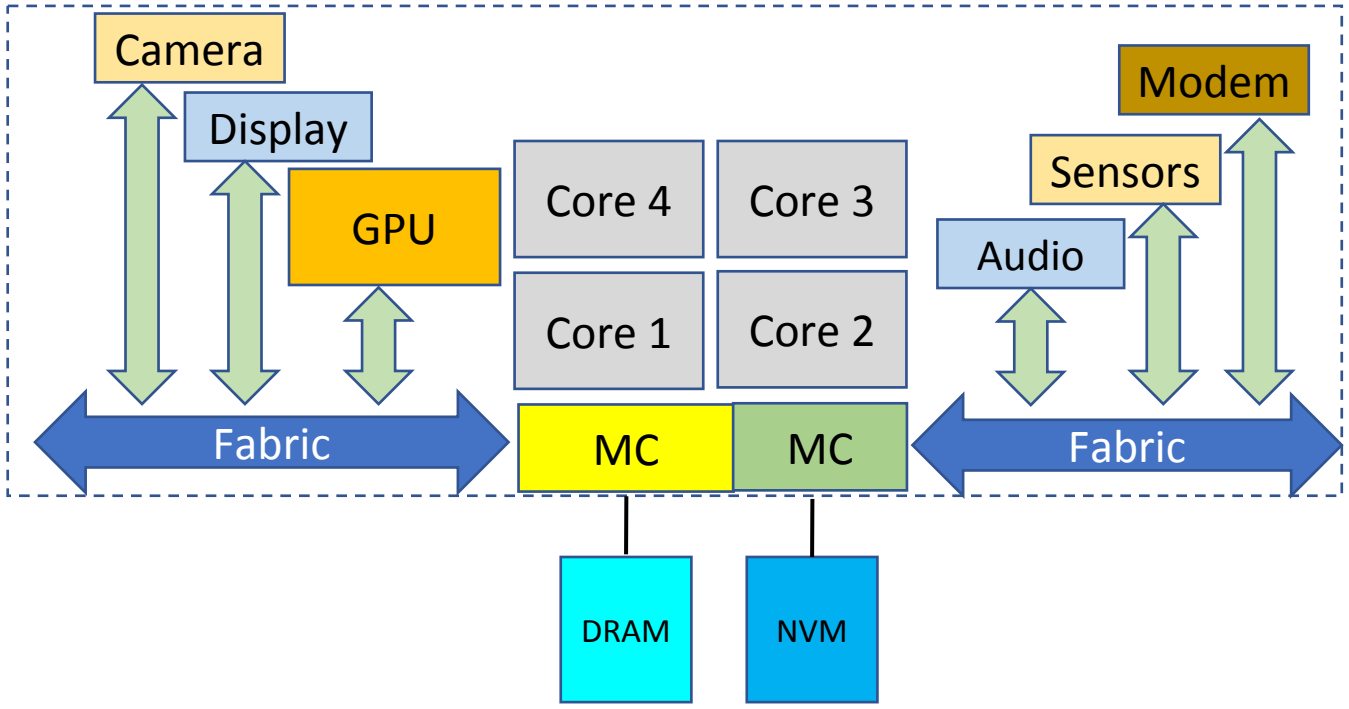
If ***frequently accessed data*** can be concentrated to the fastest regions of a hybrid memory hierarchy, memory system's energy consumption can be reduced significantly, ***without*** any significant ***loss in performance and user experience***.

Hybrid Memory Architectures for Handhelds

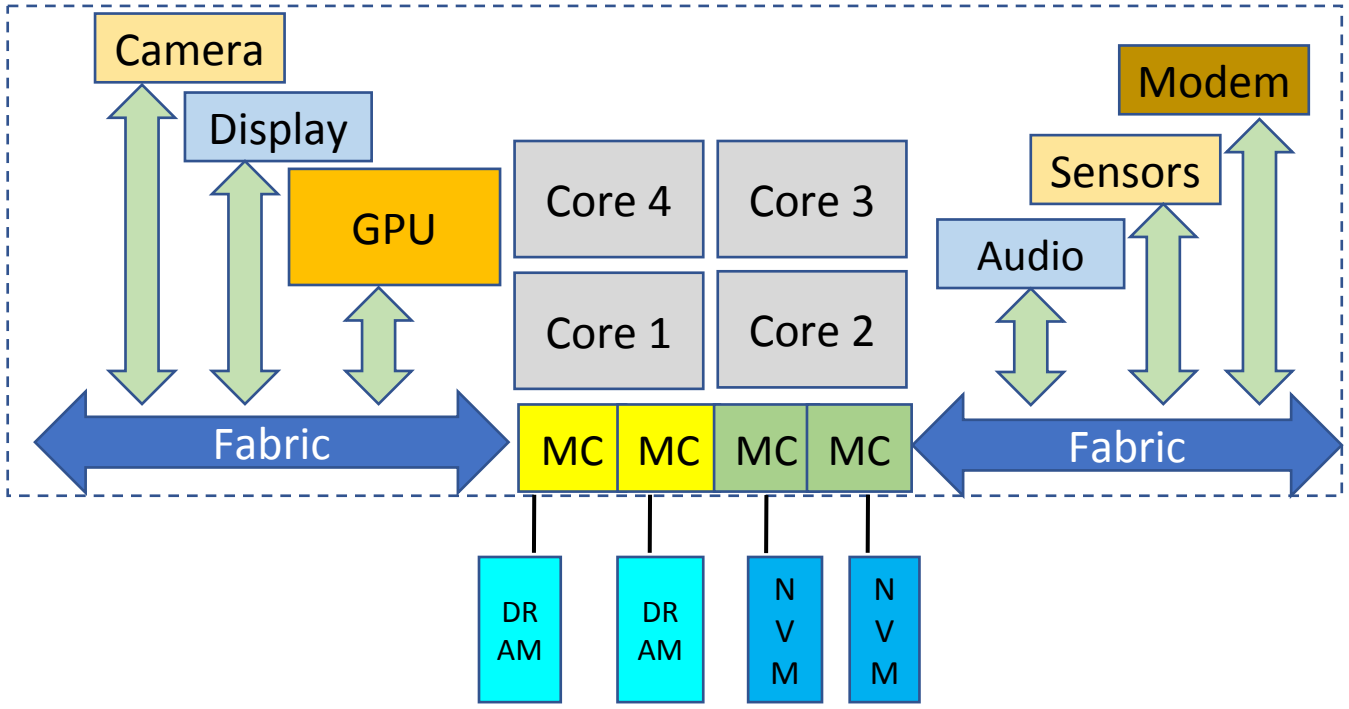


DATE 2018

Hybrid Main Memory in Handhelds

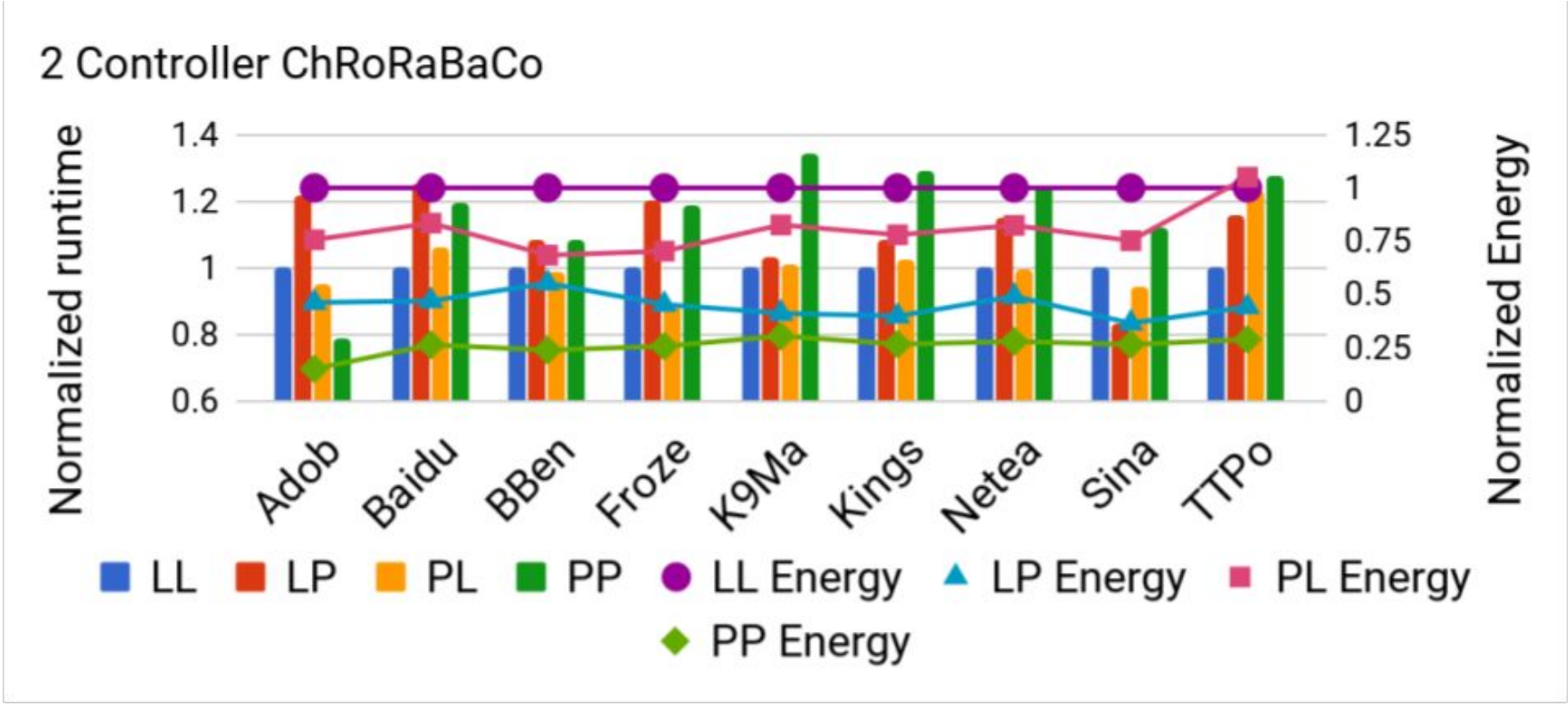


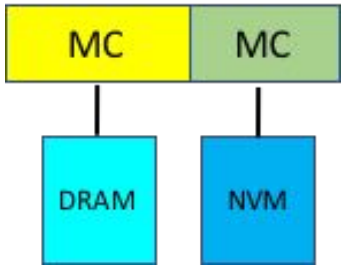
Hybrid Main Memory in Handhelds



DATE 2018

Results


2 Controller ChRoRaBaCo
Normalized runtime
1.4
1.2
1
0.8
0.6
1.25
1
0.75
0.5
0.25
0
Normalized Energy
Adob
Baidu
BBen
Froze
K9Ma
Kings
Netea
Sina
TTPo
LL
LP
PL
PP
LL Energy
LP Energy
PL Energy
PP Energy


MC
MC
DRAM
NVM

Summary and Key Takeaways

- Research in architectures of handheld devices architectures is important more so in the era of wearables
  - Memory sub-system is becoming increasingly important, even in handheld
  - Need tools, benchmarks to carry research forward
  - META – one step in that direction

- NVMs should be integrated into handheld memory hierarchy
  - Mechanisms to provide access to high capacity, low latency memories might require intelligent data management
  - H/W – S/W co-design is better than one or the other

Acknowledgements

Varun Gohil, Shreyas Singh, Sneha Ved (IIT Gandhinagar)

Nisarg Parikh (LD College of Engineering)

Sarabjeet Singh (Ashoka University)