Samsung M3 Processor

Jeff Rupley
Sr Principal Engineer
Samsung Austin R&D Center – SARC

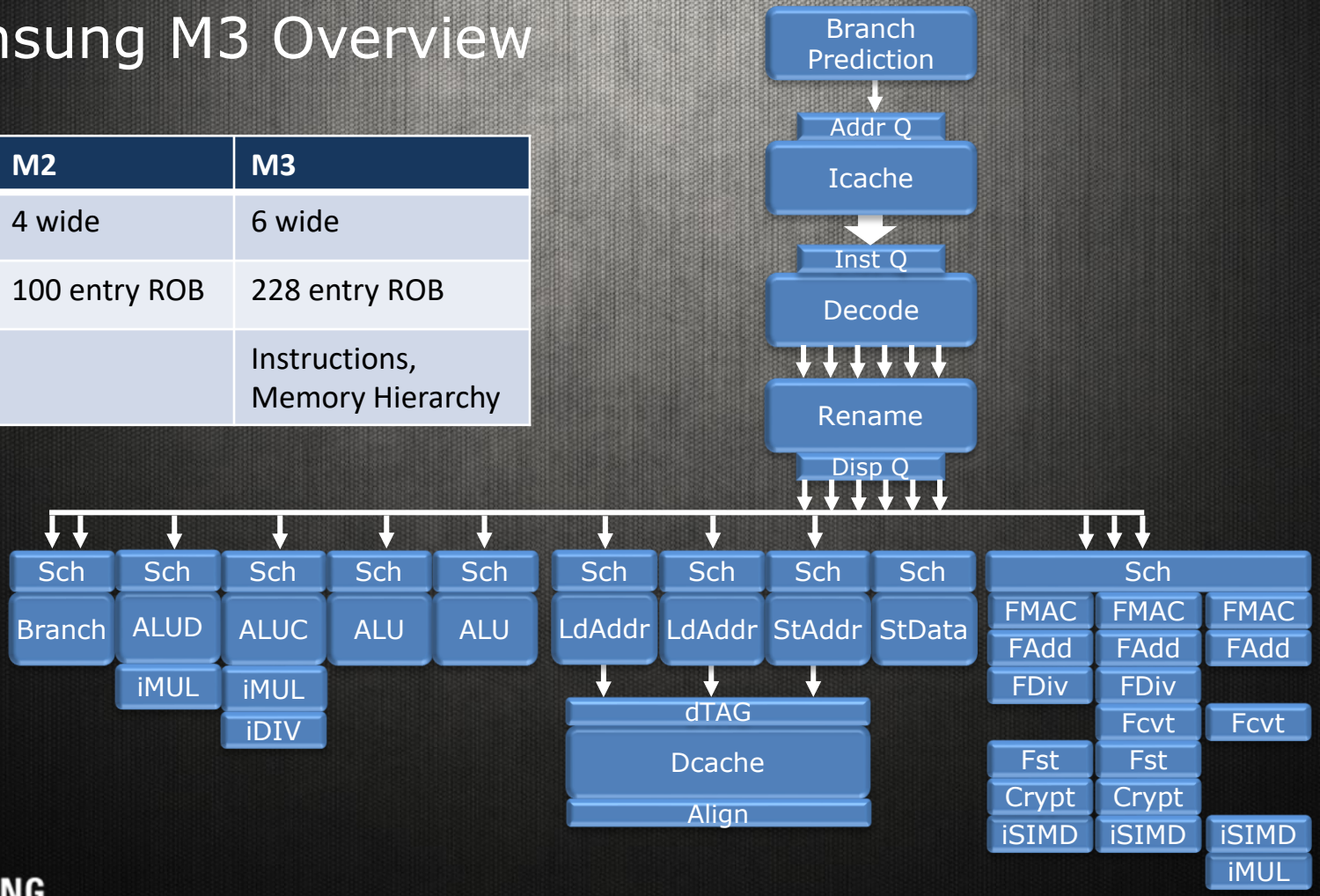
HotChips 2018

Samsung M3 Processor

- ISA - ARM v8.0, 64-bit/32-bit compliant.
- Leveraged database from M1 starting RTL back in 2015
- Goal:
  - ~~Incremental improvement~~
  - Much improved design: Wider, Deeper, Faster
- Challenge:
  - Smartphone launch cycle is relentless: must hit schedule
- Q1 2018:
  - Productized 2.7Ghz in Samsung 10nm LPP
  - Established new standard of performance for Android smartphones

Samsung M3 Overview

	M2	M3
Wider	4 wide	6 wide
Deeper	100 entry ROB	228 entry ROB
Faster		Instructions, Memory Hierarchy



Samsung M3 Front End

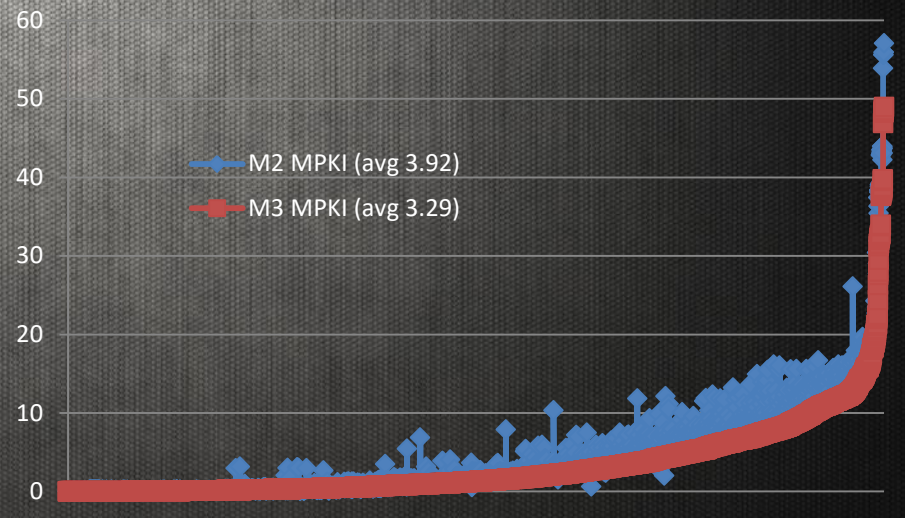
M3 Branch Prediction:

- 128-entry microBTB (2x)
- 4K-entry mainBTB: improved branch-taken latency
- 16K capacity L2 BTB (2x capacity, 2x bandwidth)
- Conditional predictor improvements including more weights for Neural Net
- Improvements to the Indirect Predictor

=> Net of above led to average MPKI reduction ~15%

M3 Instruction Fetch:

- 64KB/4-way
- Read up to 48 Bytes / cycle (2x fetch width)
- Decoupling Instruction Queue (nearly 2x deeper)
- 512 entry ITLB (2x)



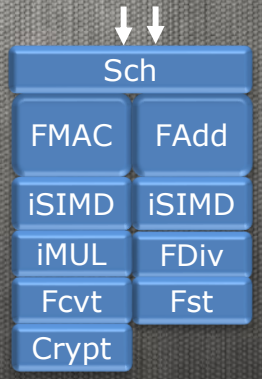
MPKI comparison across ~4800 traces sorted by M3

Samsung M3 Middle Machine

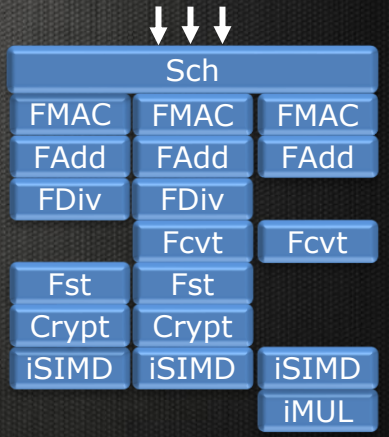
- Wider:
  - Decode up to 6 inst/cycle (1.5x wider than prior)
    - Several fusion idioms supported
  - Rename, Dispatch, Retire: up to 6 uops/cycle (1.5x wider than prior)
  - Up to 9 integer ops issued/cycle (versus 7 in prior)
    - 4th ALU including a 2nd integer multiplier
    - 2nd Load AGU – part of doubling load bandwidth
- Deeper:
  - 228-entry ROB (>2x deeper program window than prior)
  - 126-entry distributed integer scheduler (>2x deeper than M1)
- Faster Instructions:
  - Additional 1-cycle latency ops
  - Some ops optimized to 0-cycle latency
  - Integer Divider now radix 16 (4 bits/cycle) versus prior radix 4 (2 bits/cycle)

Samsung M3 FPU

- Wider:
  - 3rd dispatch and issue ports (1.5x)
  - 3x 128b FMAC/FADD
    (versus M1 1 128b FMAC + 1 128b FADD)
    => 2x maximum FLOPS
  - 2nd 128b Load port
    => critical to feed the FP "beast"
- Deeper Out-of-Order
  - 62-entry FP scheduler (nearly 2x versus prior)
  - 192-entry FP PRF (2x versus prior)
- Faster Instructions:
  - FMAC : 4-cycle MAC (was 5)
    3-cycle Mul (was 4)
  - FADD: 2-cycle (was 3)
  - FDIV: radix64 (was radix4)
    => 6 bits/cycle versus 2



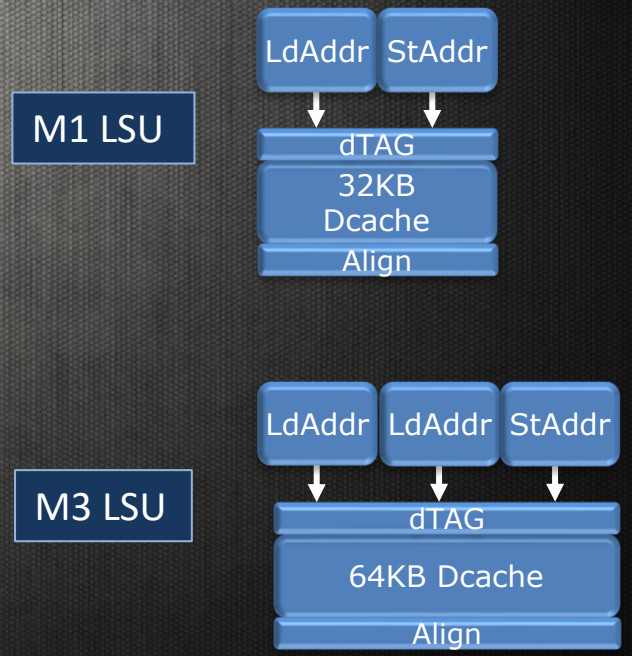
M1 FPU



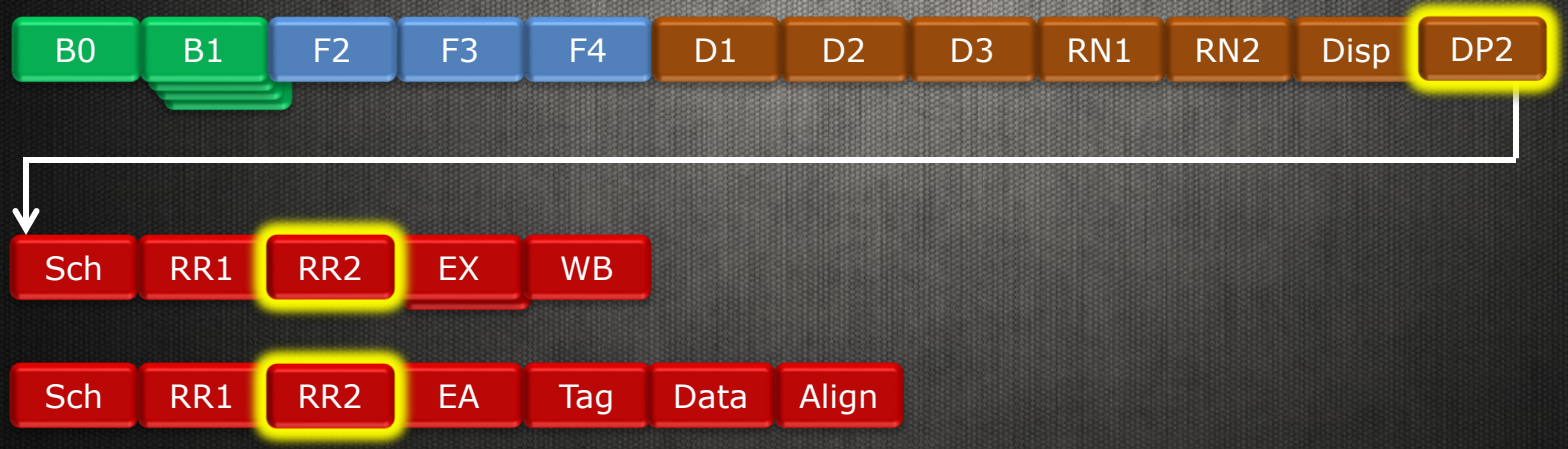
M3 FPU

Samsung M3 Load/Store Unit

- Bandwidths:
  - 2-Load/cycle (2x read bandwidth vs. prior)
  - 1-Store/cycle
  - Additional Stream/Copy Optimizations
- Depth:
  - Larger schedulers
  - Doubled Store Buffer
  - 12 outstanding misses (8 prior)
- Latencies:
  - 64KB/8-way D$ for 4 cycle (integer)
    - twice the former capacity @ same latency
  - Enhanced and Hybridized Prefetcher
  - TLBs
    - New mid-level 512-entry DTLB
    - Enhanced unified L2TLB – 4K entry (vs 1K)



Samsung M3 Core Pipeline



Deeper and Wider were not free. Versus M1:
1: A second stage of dispatch was added
2: A second stage for PRF read was added

Samsung M3 Cache Hierarchy

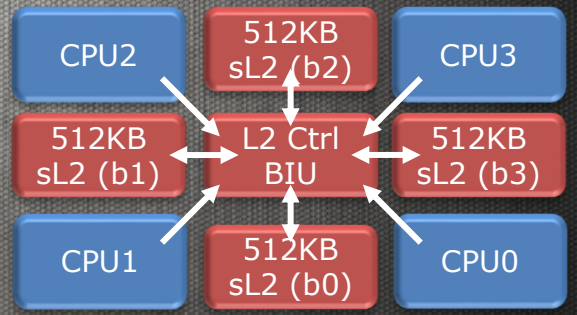
M1/M2: 16B/cycle/CPU
shared L2 (inclusive of D$)
- 2MB, 16-way, 22c

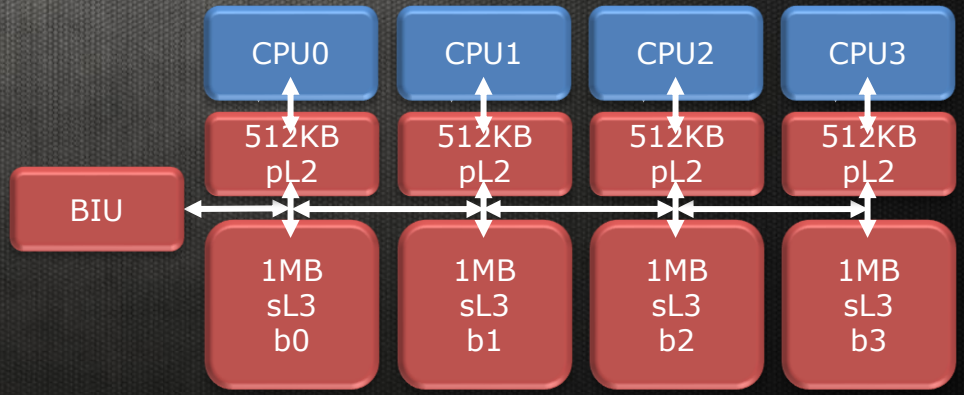
L2/BIU: 56 outstanding transactions

M3: 32B/cycle/CPU (2x bandwidth)
Private L2 (inclusive of D$)
- 512KB, 8w, 12c

SL3 (exclusive of L2$)
- 4MB, 16/way, ~37c typical (NUCA)
- Slice design - 1MB per slice
  => Goal: configurability

BIU – 80 outstanding transactions



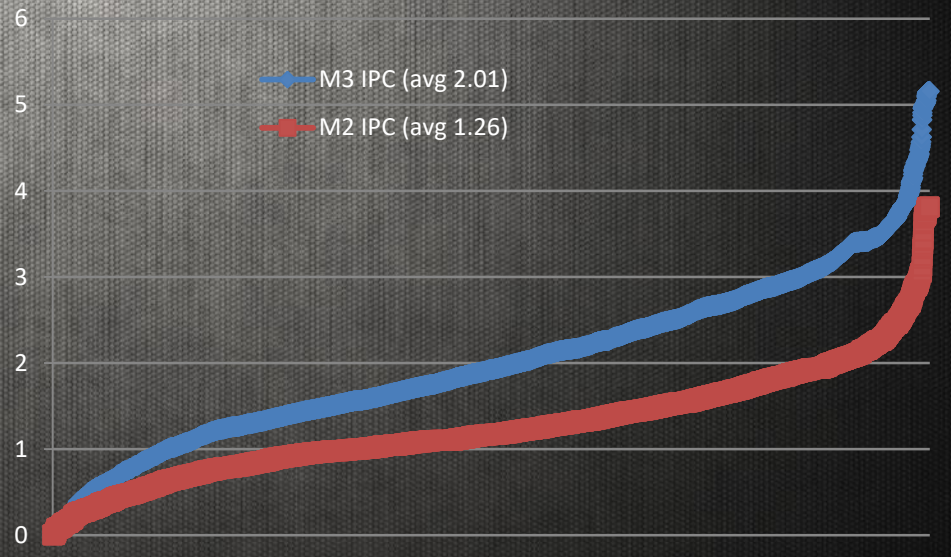


Performance Infrastructure

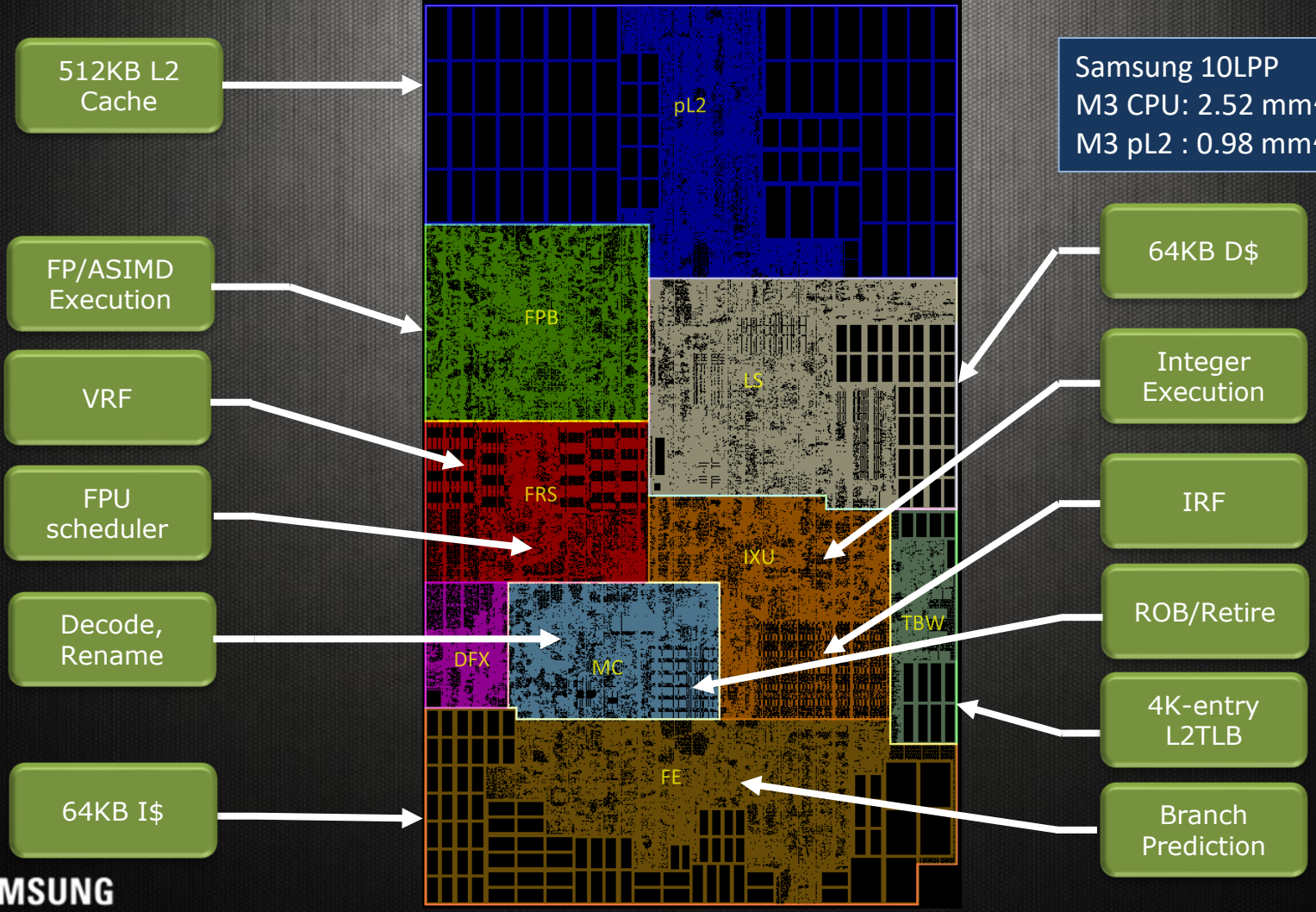
Dedicated performance team ran comprehensive simulations to guide tradeoffs across the design.

~4800 traces including:
Spec (2K/2K6, Int/FP), GBv4, Antutu, Octane, Sunspider, Bbench, browsermark, and more.

Correlation team ran hundreds of execution snippets across RTL and model to find design mistakes and improve prediction accuracy. Emulator team provided additional support by running long term simulations – found branch predictor "leak" this way.



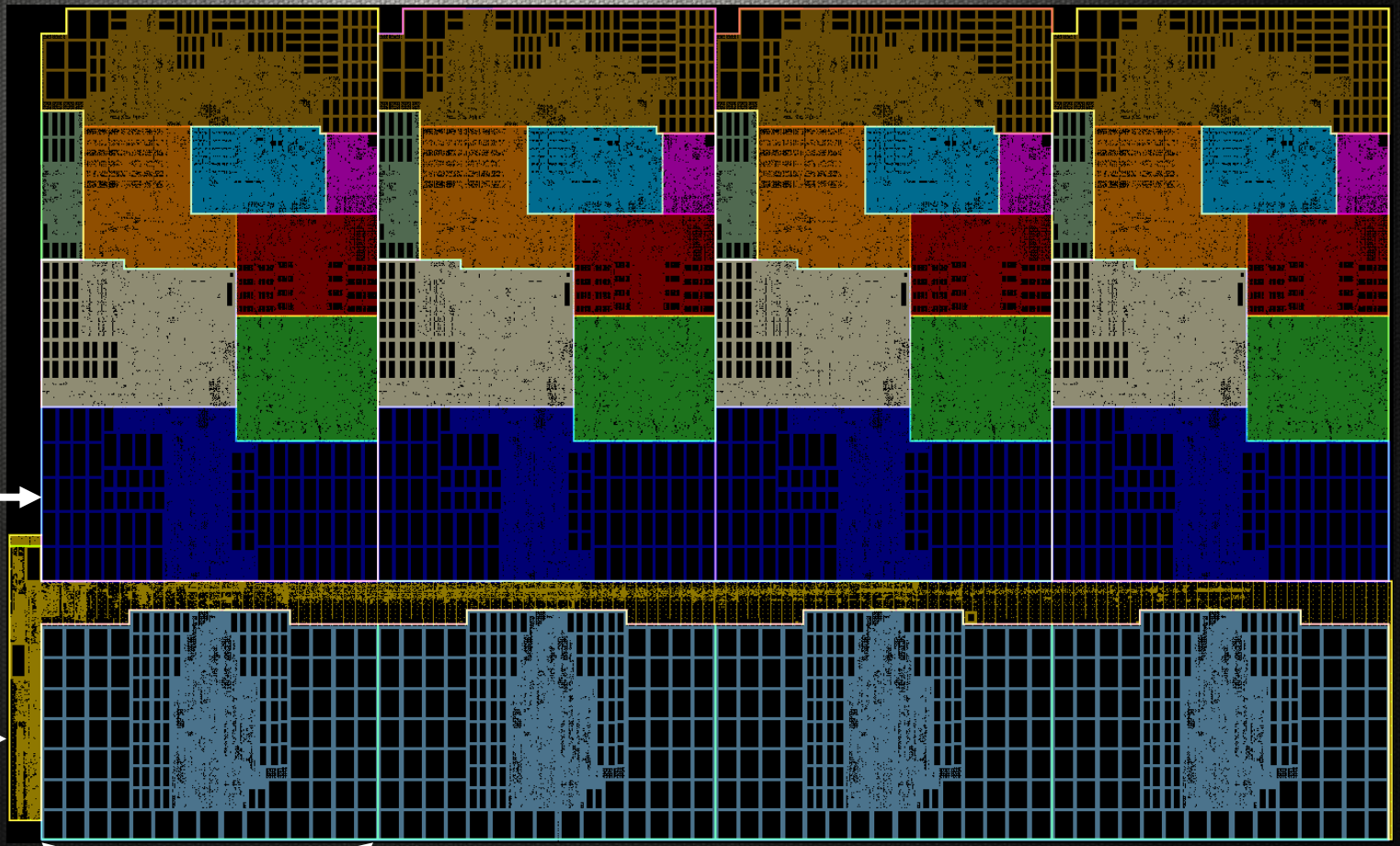
Simulated IPC comparison across ~4800 unweighted traces;
Both M2 and M3 sorted here;
IPC will vary across applications/benchmarks



Samsung 10LPP
M3 CPU: 2.52 mm^2
M3 pL2 : 0.98 mm^2

SAMSUNG

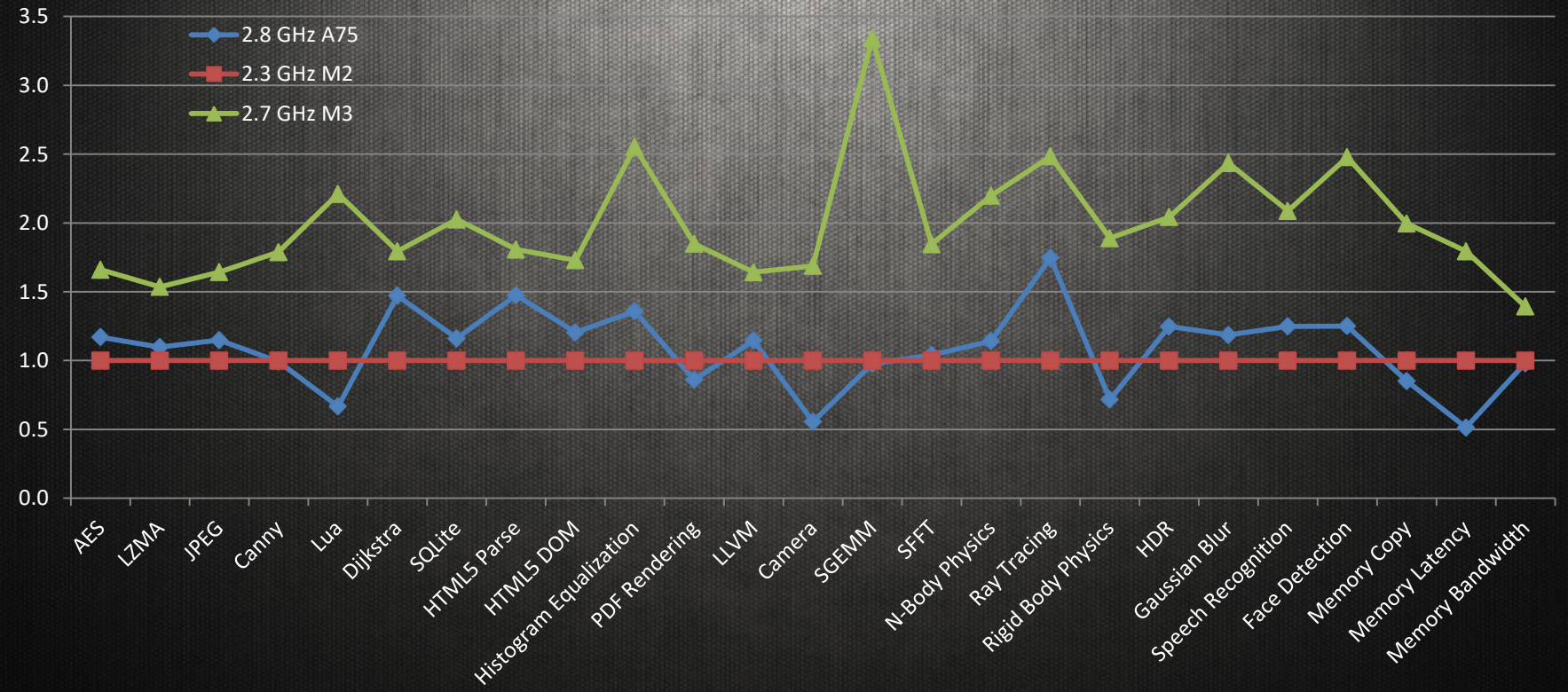
Samsung 10LPP
M3 4P Cluster:
20.9 mm^2



L3 slices: design once, replicate 4 times

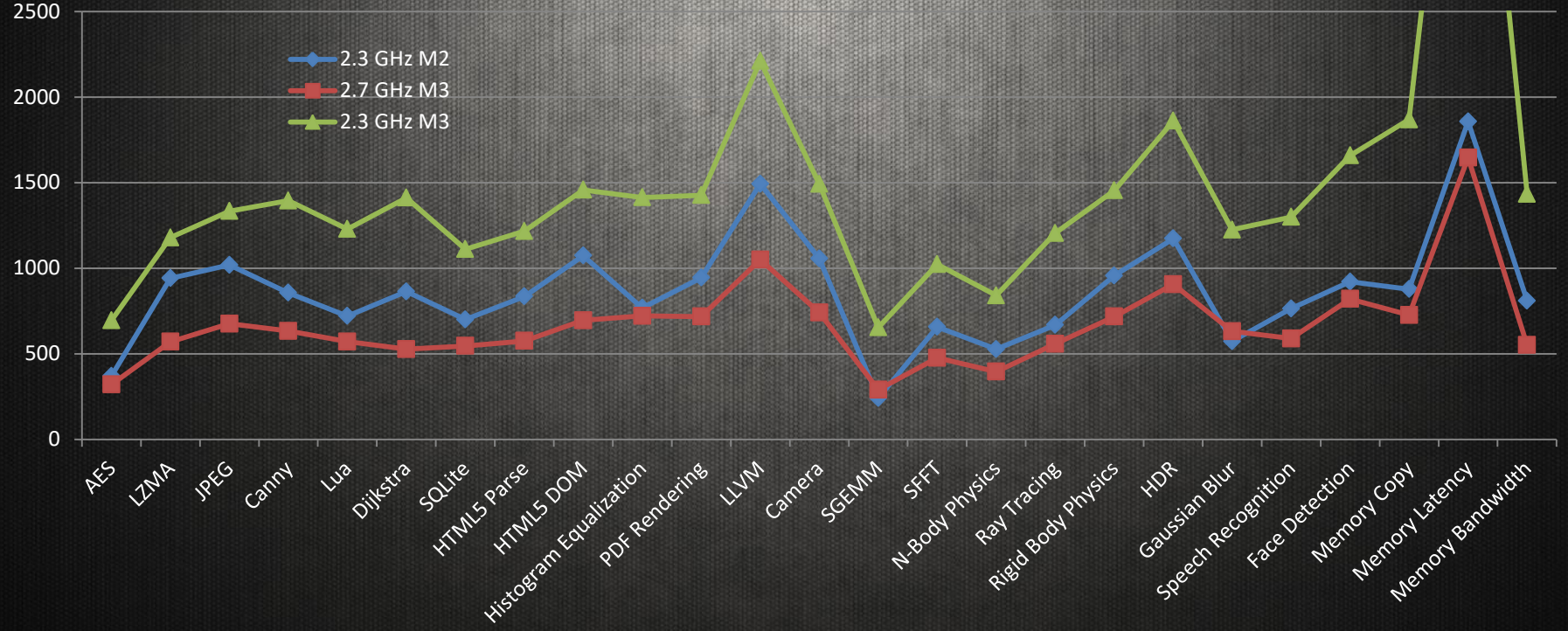
SAMSUNG

Silicon Performance Comparison – GB4 single relative to M2



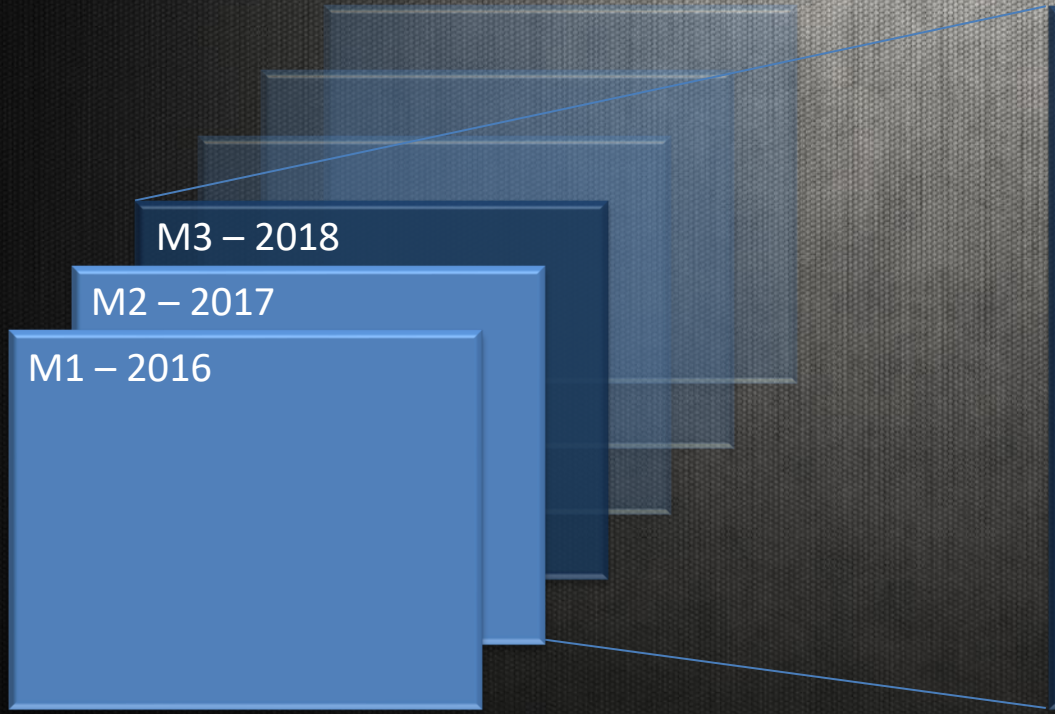
New level of performance established for the Android eco-system

Silicon Perf/W Comparison – GB4 single



Perf/Power Efficiency superior for M3 at iso-frequency with M2;
Efficiency in 1P frequency boost mode in-range

New CPU every year



Samsung M3

- Next Gen Planning Started - Q2 2014
- RTL Start – Q1 2015
- Fork features for incremental M2 - Q3 2015
- Replan for a bigger M3 push - Q1 2016
- Tapeout EVT0 – Q1 2017
- Product Launch: Q1 2018

Team now on strong annual cadence: expect more improvements every year

Conclusion

- Samsung's 3rd generation design for ARMv8
- On schedule production launch
- Wider, Deeper, Faster …

  Best CPU for Android smartphones
- More designs to come