Industry standard benchmarking of embedded systems

challenges, solutions, and opportunities

Shay Gal - On
Director of Software, EEMBC
shay@eembc.org

EEMBC Quick Background:

Industry-Standard Benchmarks for the Embedded Industry

- EEMBC formed in 1997 as non-profit consortium
- Defining and developing benchmarks
- Targeting processors and systems
- Expansive Industry Support
  - 43 members (silicon vendors, tool vendors and OEMs)
  - >80 commercial licensees
  - >200 university licensees

CELEBRATING 18 YEARS

WHAT IS A BENCHMARK?

- An established point of reference against which devices can be measured, comparing performance, reliability, efficiency etc.
- Benchmarks are being abused.
  - **Marketing tools**
  - **Sales tools**
  - **Inaccurate/biased measurements**
- **Benchmarks provide crucial data**

What makes a good embedded benchmark?

(and why do we need more then one?)

- Relevant to the target audience.
  - Who are the users? Marketing? Engineering? Consumer?
  - Represent real usage of the device?
- Repeatable (so we can trust the results).
- Impartial/Fair (compare platforms).
- Standardized ?
- Resistant to mistakes/cheating?
- Portable / available on many platforms?
- Easy to understand? Easy to compare? Other?

Unfortunately, one number cannot tell the whole story...

Worst Benchmark Pitfall?

- The "Magic Bullet" number
  - Easy for consumers and marketing people to understand
  - But, the devil is always in the details
- Worse yet, many times generated using flawed methodology!
  - Documented if in source form, could even seem reasonable
  - Mostly hidden otherwise
    - Though can be deduced with due diligence
  - Let me illustrate …

So, a benchmark expert entered a Store….

http://archive.constantcontact.com/fs042/1101916237075/archive/1102594461324.html

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a “benchmark expert” entered a Store….

$ 3,000.00

$ 200.00

They want $2,700 for the server and $100 for the iPod.

I will get both and pay only $2,240 altogether!

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a “benchmark expert” entered an Store….

Ma’am you are $560 short.

But the average of 10% and 50% is 30% and 70% of $3,200 is $2,240.

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

$ 200.00

50%
Discount
Buy Now!
10%
Discount
Buy Now!

$ 3,000.00

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a “benchmark expert entered” an Store….

Ma’am you cannot take the arithmetic average of percentages!

But… that is how Antutu calculates the score (not to mention academic research papers)!

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

50%
Discount
Buy Now!

$ 200.00

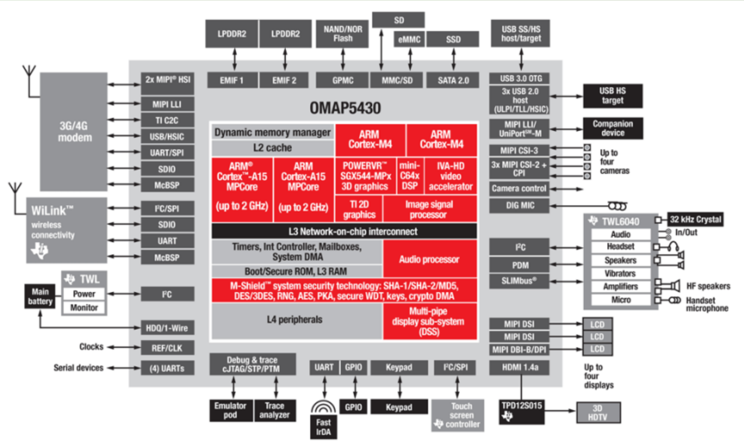
10%
Discount
Buy Now!

$ 3,000.00

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

What is unique for embedded benchmarking?

- Poor standards (except in few markets)
  - How do you apply a benchmark when the DUTs are inherently different in functionality?
- Energy consumption as important as (sometimes more important then) performance
  - Note energy and not power
- Duty cycles
  - Low power modes, and idle time part of normal operation and need to be factored.
- Specific workloads many times more important then generic indicators
  - motor control, printer, router etc.
- Non uniform systems
  - Master + DSP + GPU
  - Motor control + Safety
  - Etc…



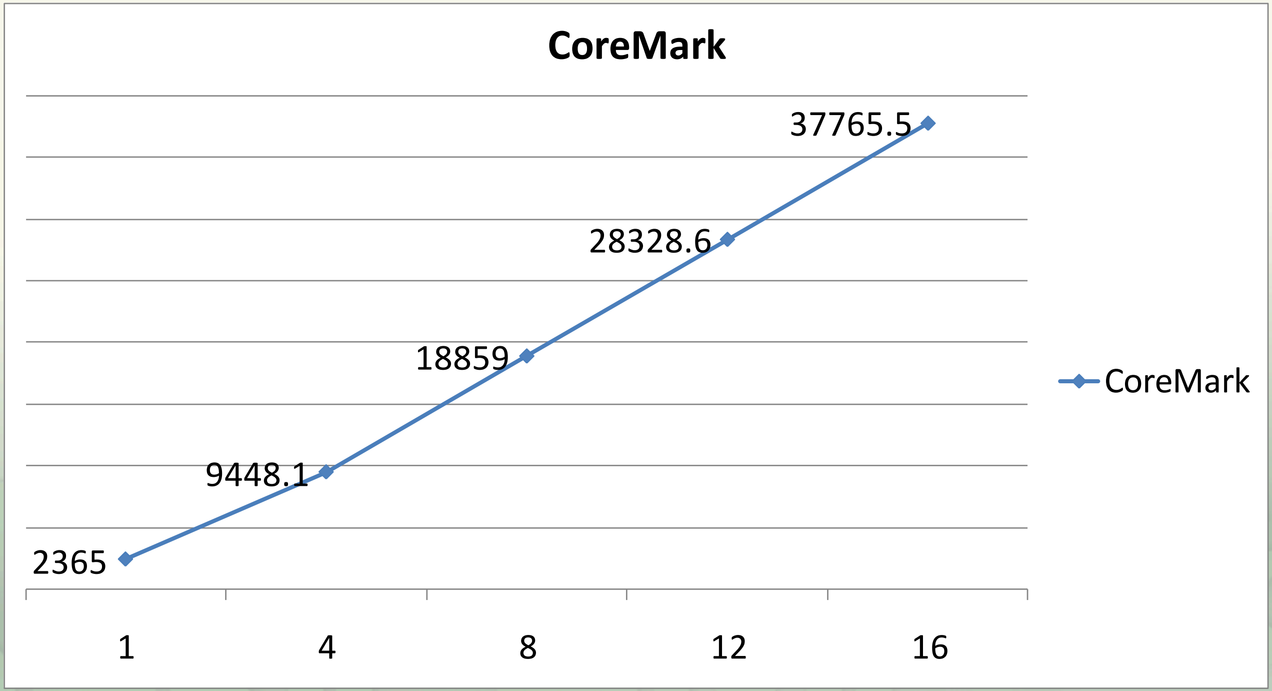
Benchmarking Solutions

- Generic benchmarks
  - CoreMark, Dhrystone, SPEC-CPU etc…
- Application / Platform specific solutions
  - BrowsingBench, ANDEBench, SPEC-JBB etc…
- Black box / data benchmarks
  - ULPBench, ETCPBench, DPIBench

GENERIC SOLUTIONS

- Commonly – throughput benchmarks
  - Easiest to develop
- How realistic is this?
  - Depends on the target (router vs. glucose meter vs. smartphone)
  - Predictions made based on this type of benchmark are better then MHz or number of cores, but for most embedded solutions can be misleading …
- How to account for multiple cores? (not necessarily all of the same capabilities)

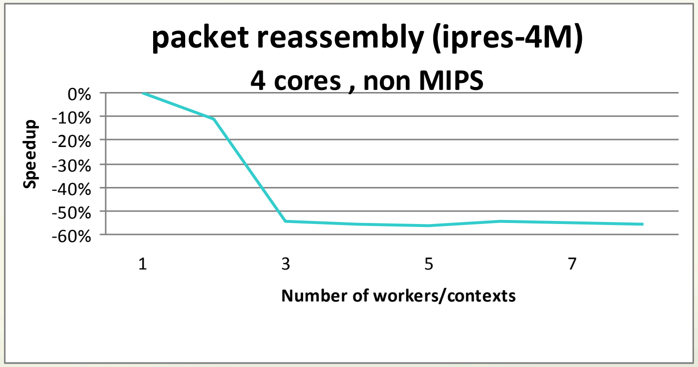
CoreMark
Core functionality for multiple cores?



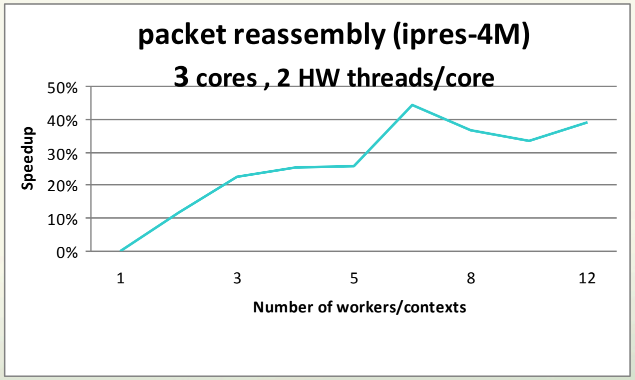
Information provided by Cavium for CN58XX

MULTIBENCH (IP REASSEMBLY)

Different ISA



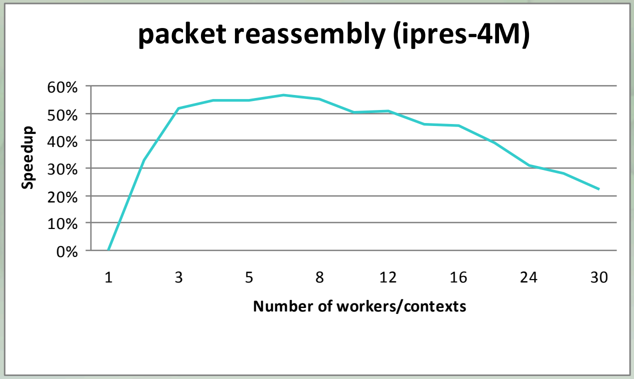
3 Core



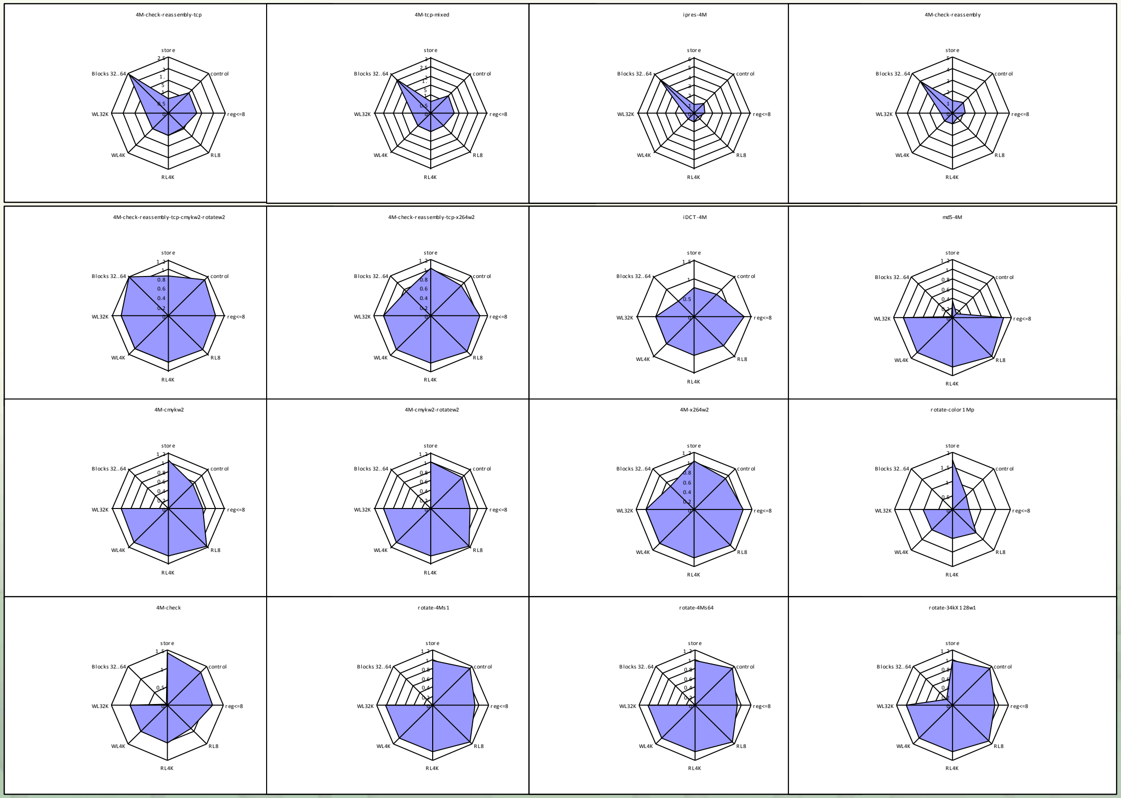
- IP-reassembly workload over 4M, one platform actually drops in performance!

- How do we design benchmarks that are relevant to the hardware being tested?

Many Core



Classification and performance prediction



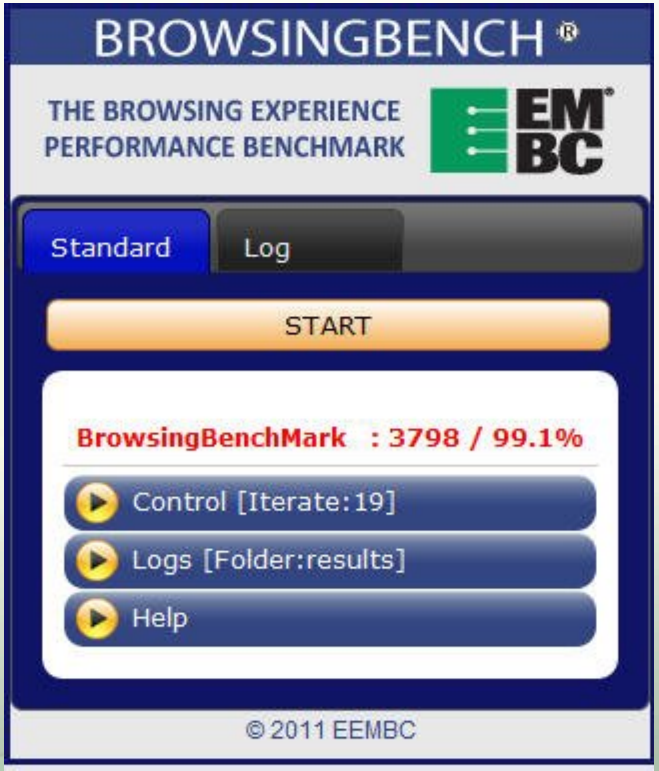
Correlation based feature subset selection + Genetic analysis.
8 data points for 80% accuracy in performance prediction.

APPLICATION/PLATFORM SPECIFIC

- Choose a problem that is relevant across a wide range of devices.
- Define in detail the methodology used to test the devices.
- Rely on the devices under test to already have a solution for the problem (since it is a relevant problem).
- Allows us to test every facet of the platform under test!
- But – requires a common problem, and a fully developed platform… Also tend to result in a benchmark that is “too complex” to be useful for anything but the particular application used.

BrowsingBench

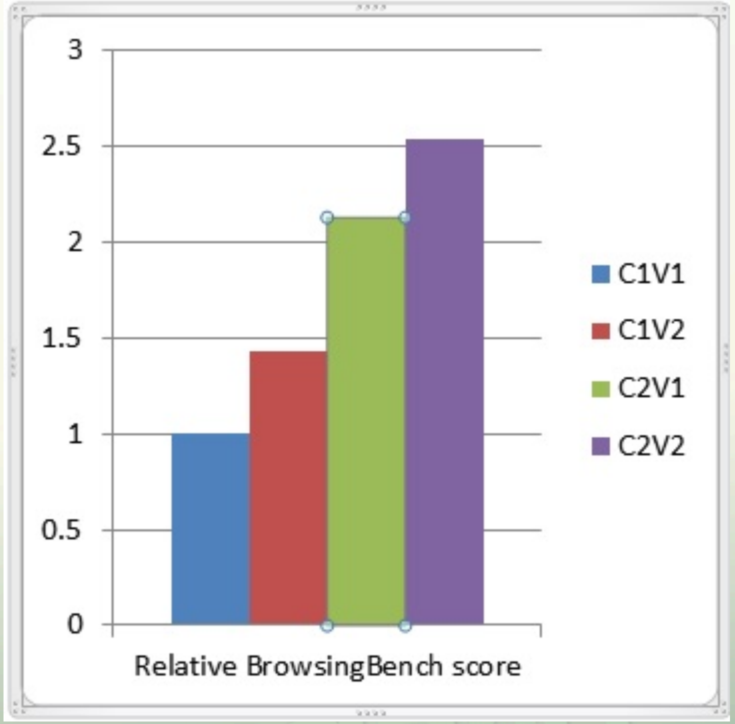
- The benchmark is a local web server.
- The target must have a browser.
- Use a common application that is available on target devices.
- Can test heterogeneous systems.

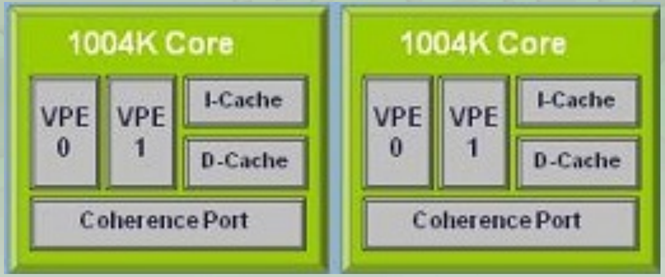


BrowsingBenchMark : 3798 / 99.1%

BrowsingBench multicore

- Legend:
  - C<N>V<K>
    - N: Number of physical cores
    - K: Number of virtual cores per physical core
- Full scale scenario testing a complex multicore system




1004K Core

VPE 0 | VPE 1 | I-Cache | D-Cache

Coherence Port

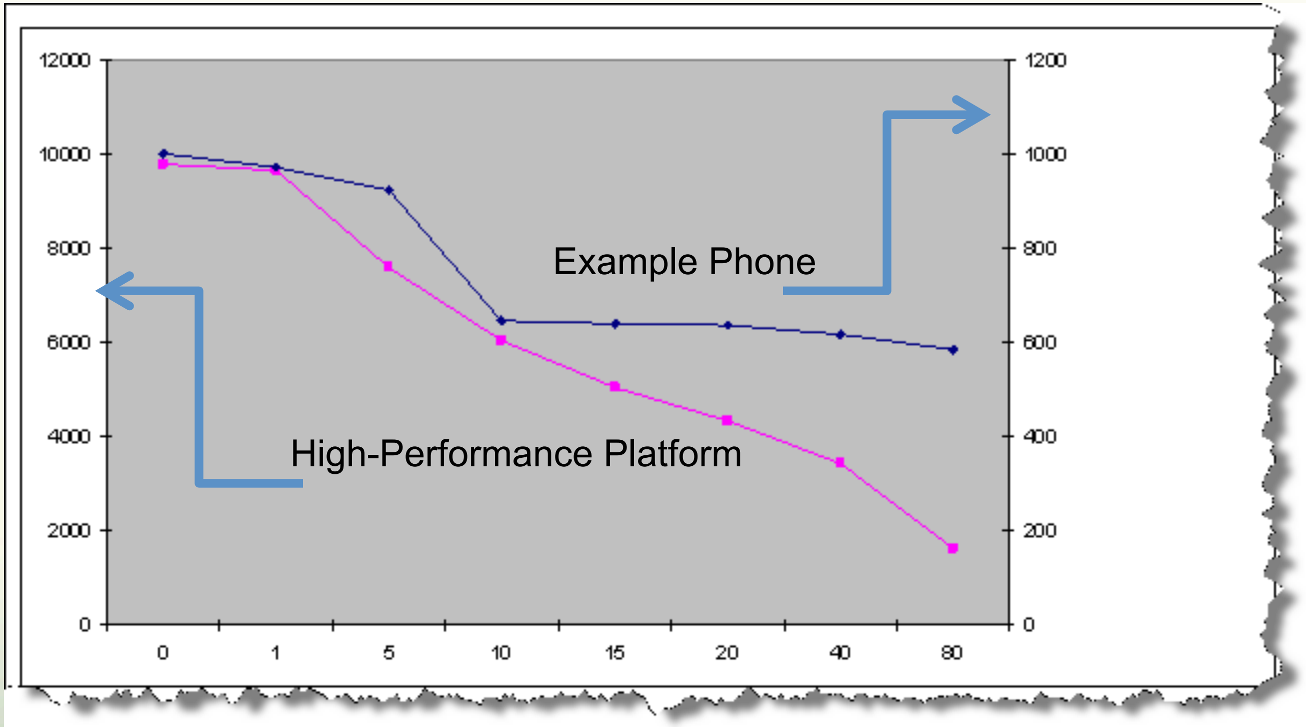
1004K Core

VPE 0 | VPE 1 | I-Cache | D-Cache

Coherence Port

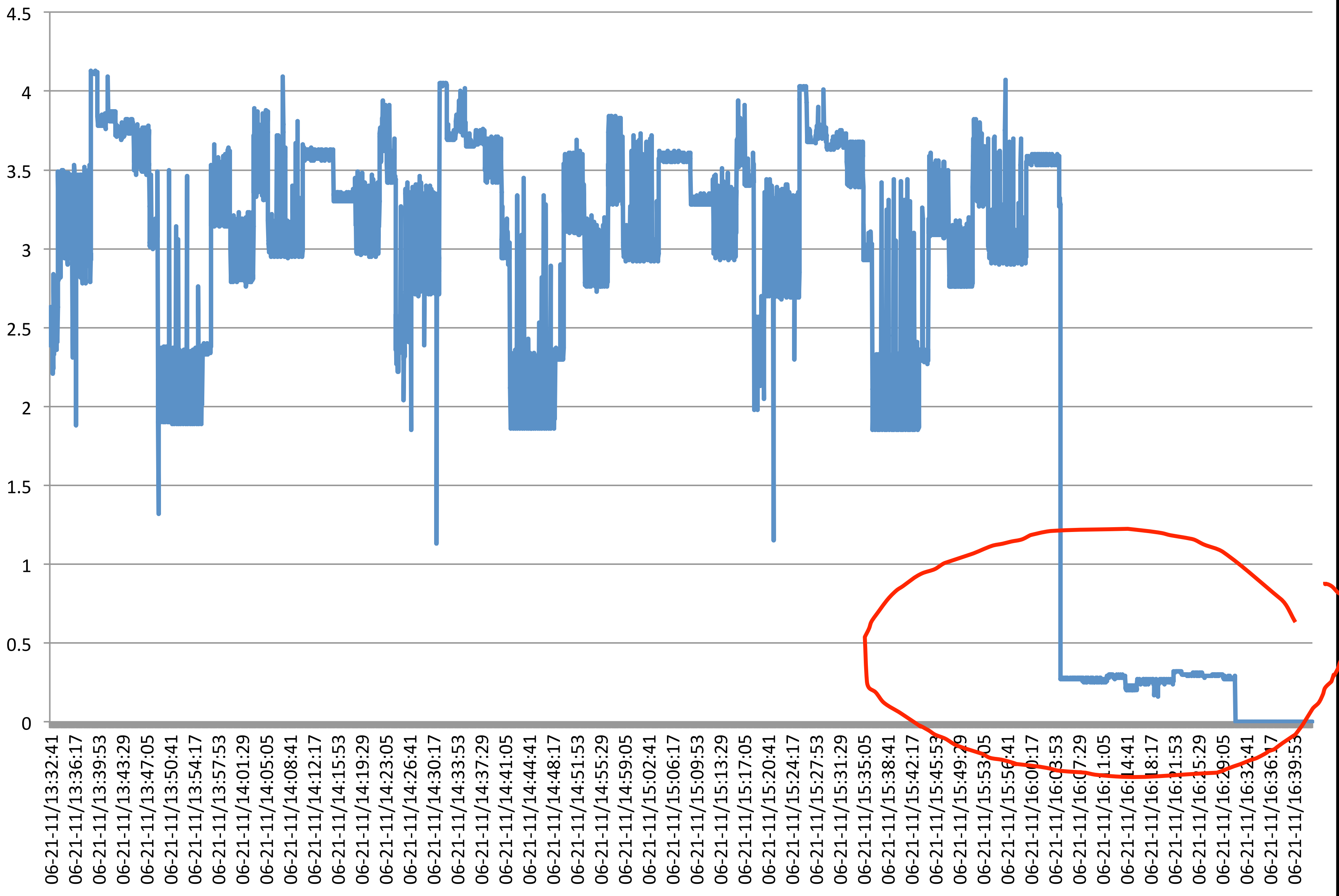
Data provided by Imagination Technologies

Latency Effects



- Latency is important since it is present in real world use case.
- Example Phone has an effective optimization for high latency connections
- Y-axis shows BrowsingBench score
  - Left axis is for high performance platform, right axis is phone

Light in FC, Droid-X, 2secs facetime, max brightness

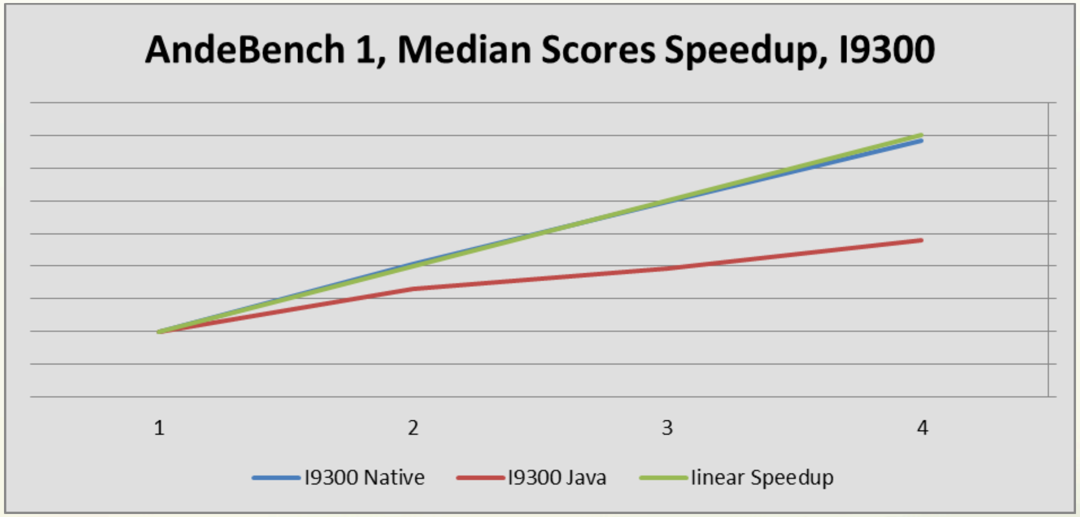


Platform Specific

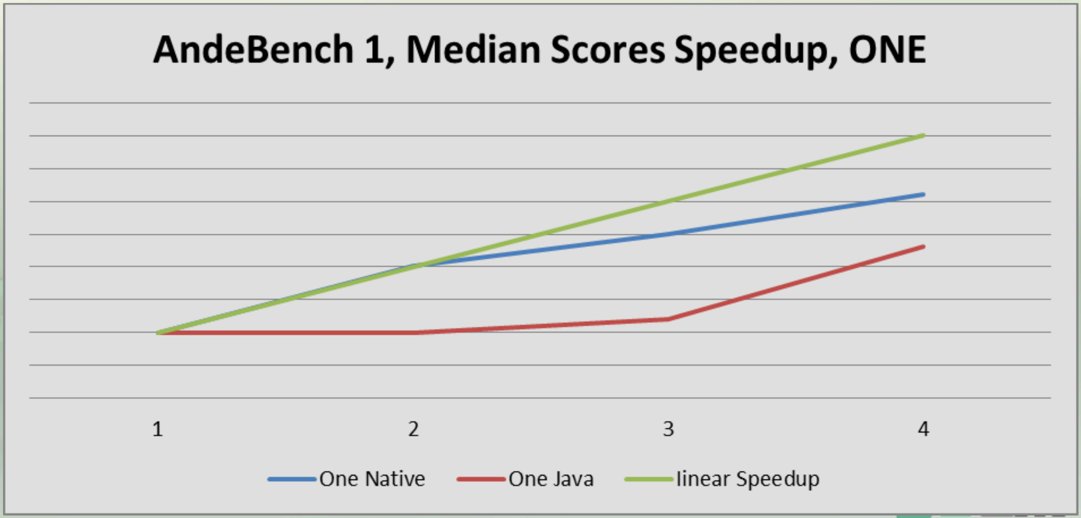
- Fixing on a platform allows using system APIs provided by that platform, and targeting important aspects of that platform.
- Examples:
  - LMBench – generic Linux functionality test distributed as source.
    - Pitfalls – people using it to compare different hardware platforms without understanding how it works.
      - Memory effects with SMP.
      - Memory latency with hardware assists. Etc.
  - ANDEBench (and other android benchmarks)
    - Pitfalls – distributed as binaries, used by consumers who do not understand what the benchmarks do…

ANDEBENCHV1 RESULTS

- Native scales
  - As expected
- Java does not
  - 3x scale for 4 core

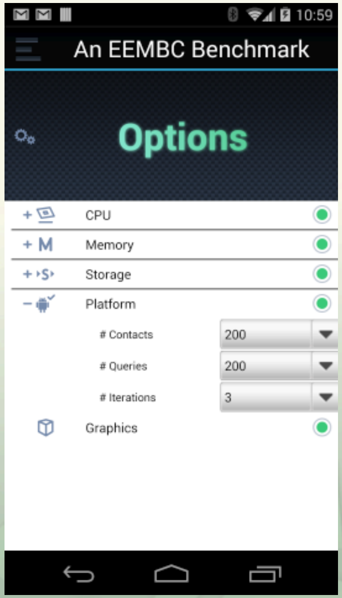


- Native scales to 2
  - But then OS effects
- Java does not
  - 2 core degrades



ANDEBENCH PRO

- Fixing on a platform such as android allows us also to call system APIs to perform complex tasks that are still common building blocks
  - Image filters and effects
  - Database API
  - XML parsing
  - Cryptography
  - Graphics
  - Populating GUI elements
- Is this a fair benchmark, considering that the services being called can be implemented differently on different platforms?



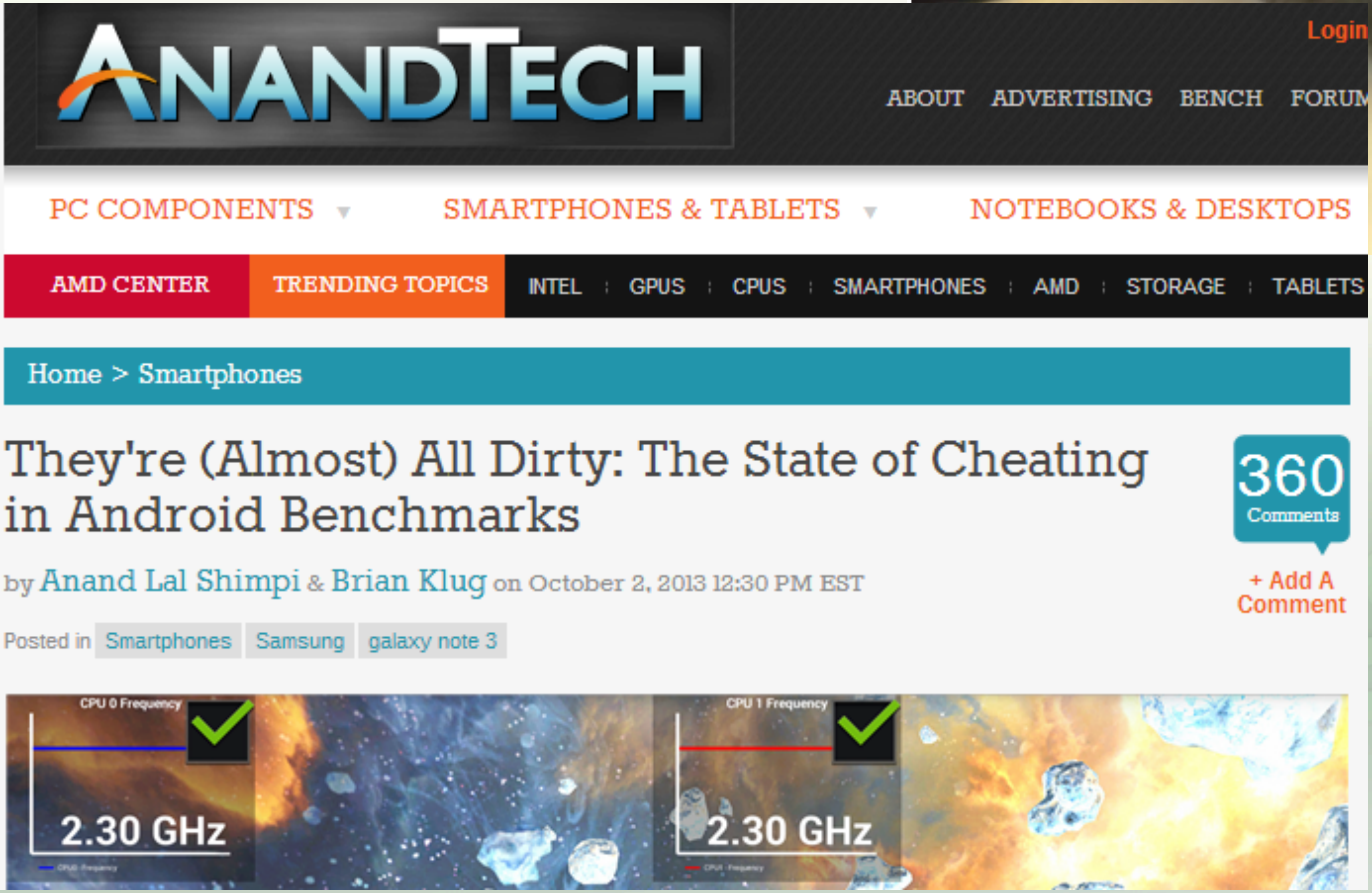
10:59

An EEMBC Benchmark

Options

- CPU
- Memory
- Storage
- Platform
  - # Contacts: 200
  - # Queries: 200
  - # Iterations: 3
- Graphics

And talk about benchmark abuse …



They're (Almost) All Dirty: The State of Cheating in Android Benchmarks

by Anand Lal Shimpi & Brian Klug on October 2, 2013 12:30 PM EST

360 Comments

+ Add A Comment

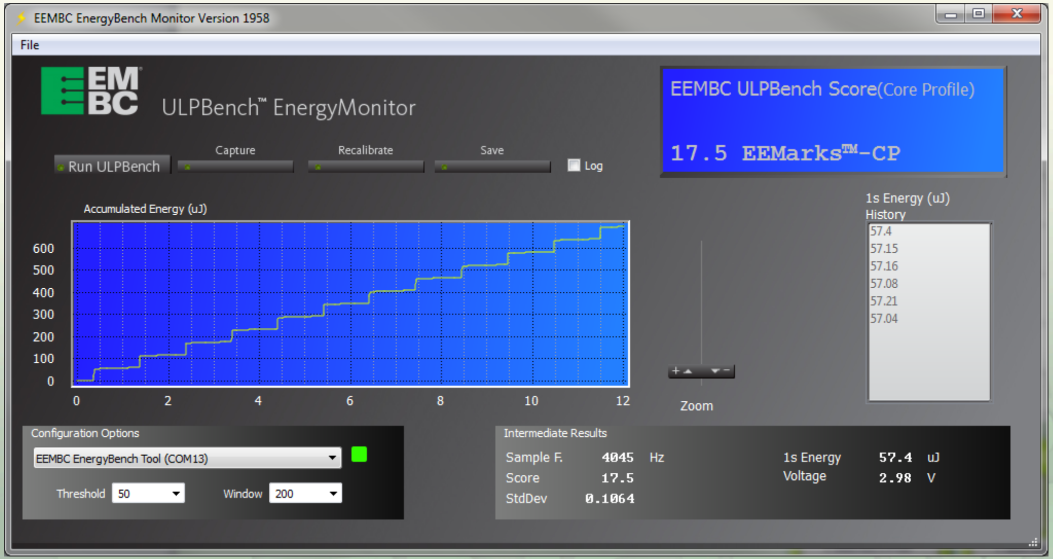
Posted in Smartphones Samsung galaxy note 3

DATA DRIVEN BENCHAMRKS

- Somewhat similar to scenario, these are even more loosely defined.
  - How fast can you compute an N node iteration of algorithm M with conditions X,Y,Z
    - E.g. 1200 pt FFT with SNR of 60dB or better
- Require (potentially) significant effort on each platform used.
- At times tests the engineer implementing the software more then the hardware. Unfortunately, that engineer does not come attached to the device under test…

ULPBench

- The workload is defined a unit of work to be done once per second. The metric is the average energy consumed per second (measured using specific hardware).



EEMBC EnergyBench Monitor Version 1958

File

ULPBench™ EnergyMonitor

Run ULPBench | Capture | Recalibrate | Save | □ Log

EEMBC ULPBench Score(Core Profile)

17.5 EEMarks™-CP

1s Energy (uJ) History

57.4
57.15
57.16
57.08
57.21
57.04

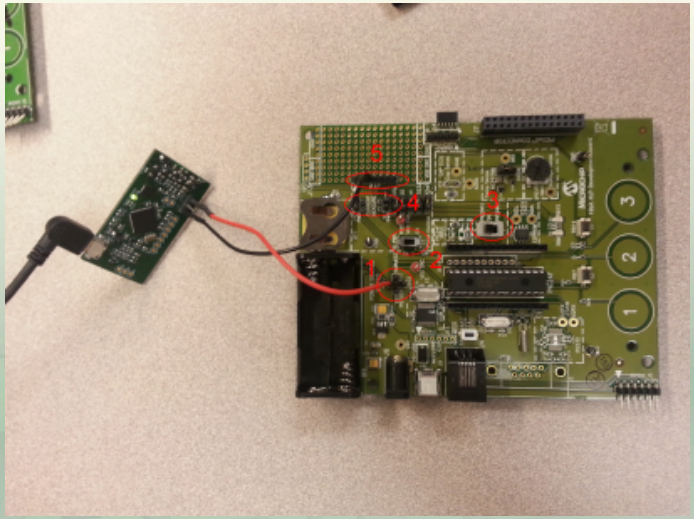
Configuration Options

EEMBC EnergyBench Tool (COM13)

Threshold 50 Window 200

Intermediate Results

Sample F.	4045	Hz	1s Energy	57.4	uJ
Score	17.5		Voltage	2.98	V
StdDev	0.1064				



How does EEMBC work?

- Industry consortium lets all vendors provide guidance during requirement definition, and feedback throughout the implementation process.
  - Open forum and open development
  - Democratic process (1 company, 1 vote)
- Content experts from companies of consultants used for each specific target benchmark, with benchmarking specific core expertise maintained by EEMBC.
  - Drawing on industry leaders for each benchmark
  - Avoid benchmarking pitfalls
- Unbiased certification available to members

SUMMARY

- Embedded devices in particular require great care in benchmark development.
- One benchmark will not resolve all questions about a device, thus we continue to develop new benchmarks.
- Creating good benchmarks is not easy, but working as an industry consortium helps.