Francesco Paterna and Shruti Patil

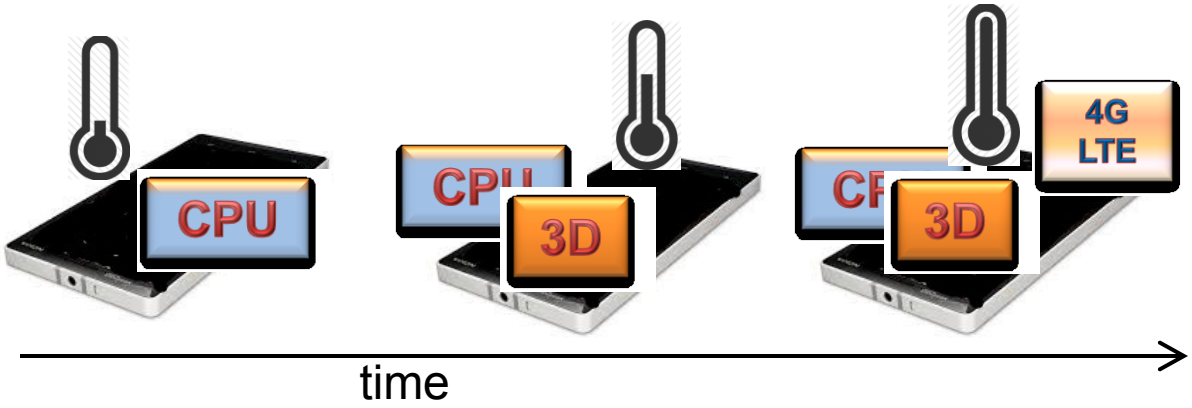
Postdoctoral Researchers at UC San Diego
Dept. of Computer Science and Engineering
System Energy Efficiency Lab

Online Characterization and Thermal Management of Mobile Phones

Motivation

- Mobile SoC have different types of units (i.e, CPU, GPU, DSP, and other accelerators)
- Hot topic: Characterizing how those units are used by apps
- Controlling SoC's efficiency to limit the impact on temperature
- Skin cover temperature
  - Human skin supports up to 45^{o}C and 41^{o}C for plastic and aluminum, respectively.

a) Mutual thermal dependency between CPU, GPU, 4G, etc.



electronista.com

icofinder.com

Motivation

- Mobile SoC have different types of units (i.e, CPU, GPU, DSP, and other accelerators)
- Hot topic: Characterizing how those units are used by apps
- Controlling SoC's efficiency to limit the impact on temperature
- Skin cover temperature
  - Human skin supports up to 45^{o}C and 41^{o}C for plastic and aluminum, respectively.

b) Ambient condition variations

orientation

contact surface

electronista.com howtouncle.com

Contribution

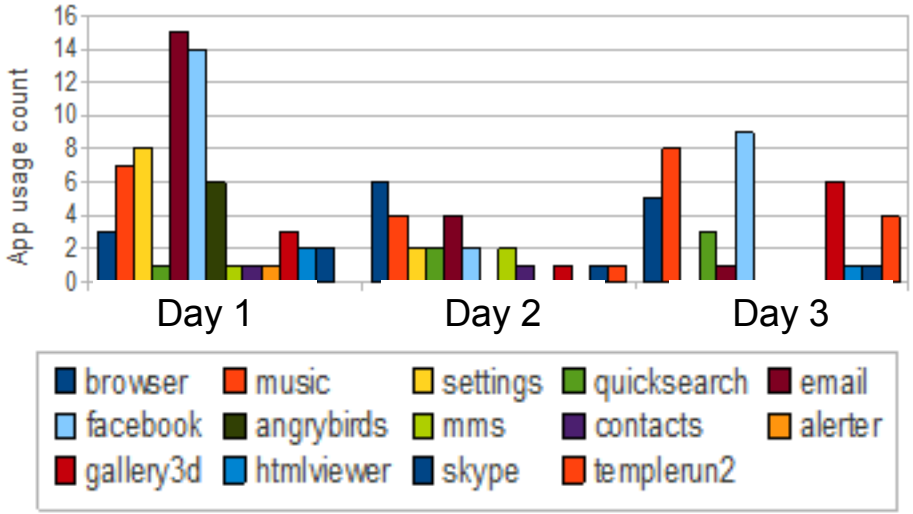
- Resources' usage app characterization
  - data from a group of users
  - utilization and power analysis

- Automatic thermal modeling of the device
  - Model leveraged by a proactive control to guarantee skin cover temperature under a threshold while operating smooth performance variations

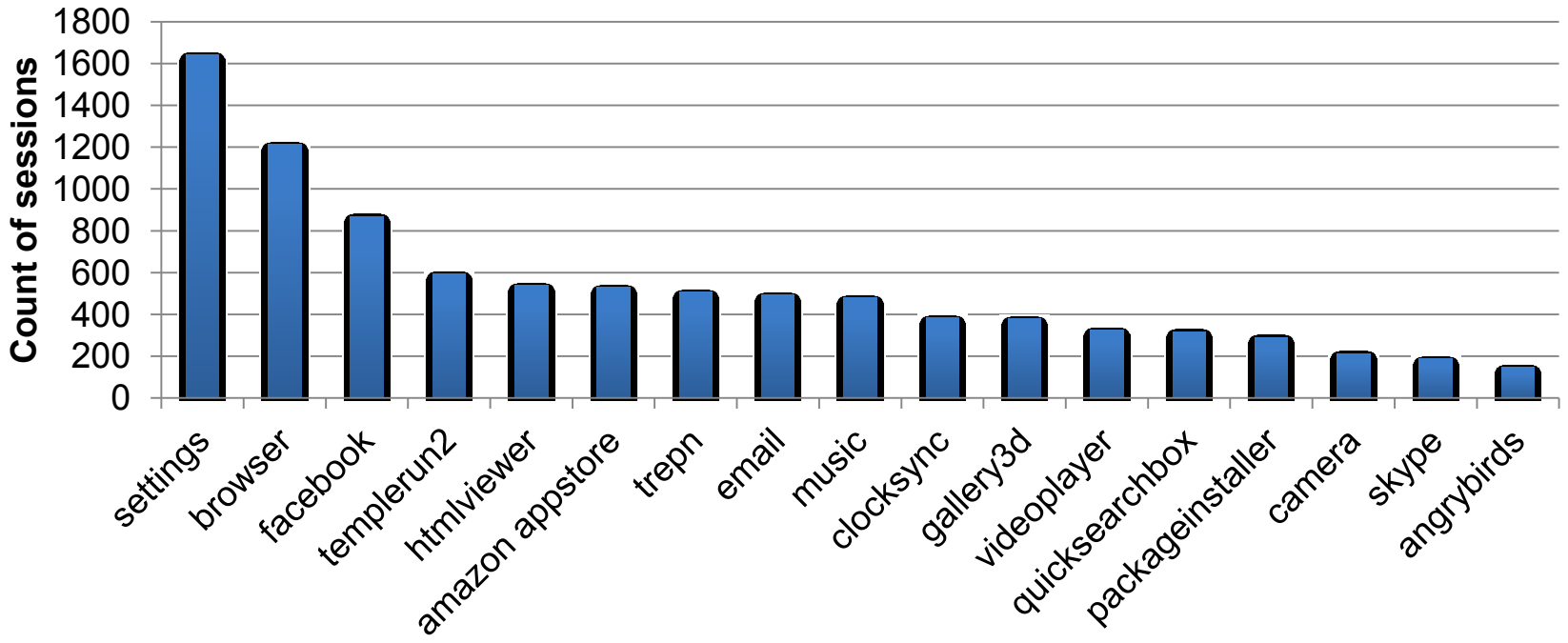
Detailed study of phone usage

- 33 UCSD students, 1 phone per students, 1 month duration
- Rooted Android phones running on Snapdragon MSM8660/8960 MDPs
  - Dual core CPUs; CPU0 master core -> higher utilization
  - Adreno 220/225 GPU
- Sampled performance counters, detailed power and memory bus measurements
- Access to Amazon Android App Store
  - Users ran 125 applications on their devices

Glimpse of user's 3 day app usage



App Usage Statistics



- Favorite Apps
  - Facebook, Email, Games, Browser, Camera, Skype, Music etc.

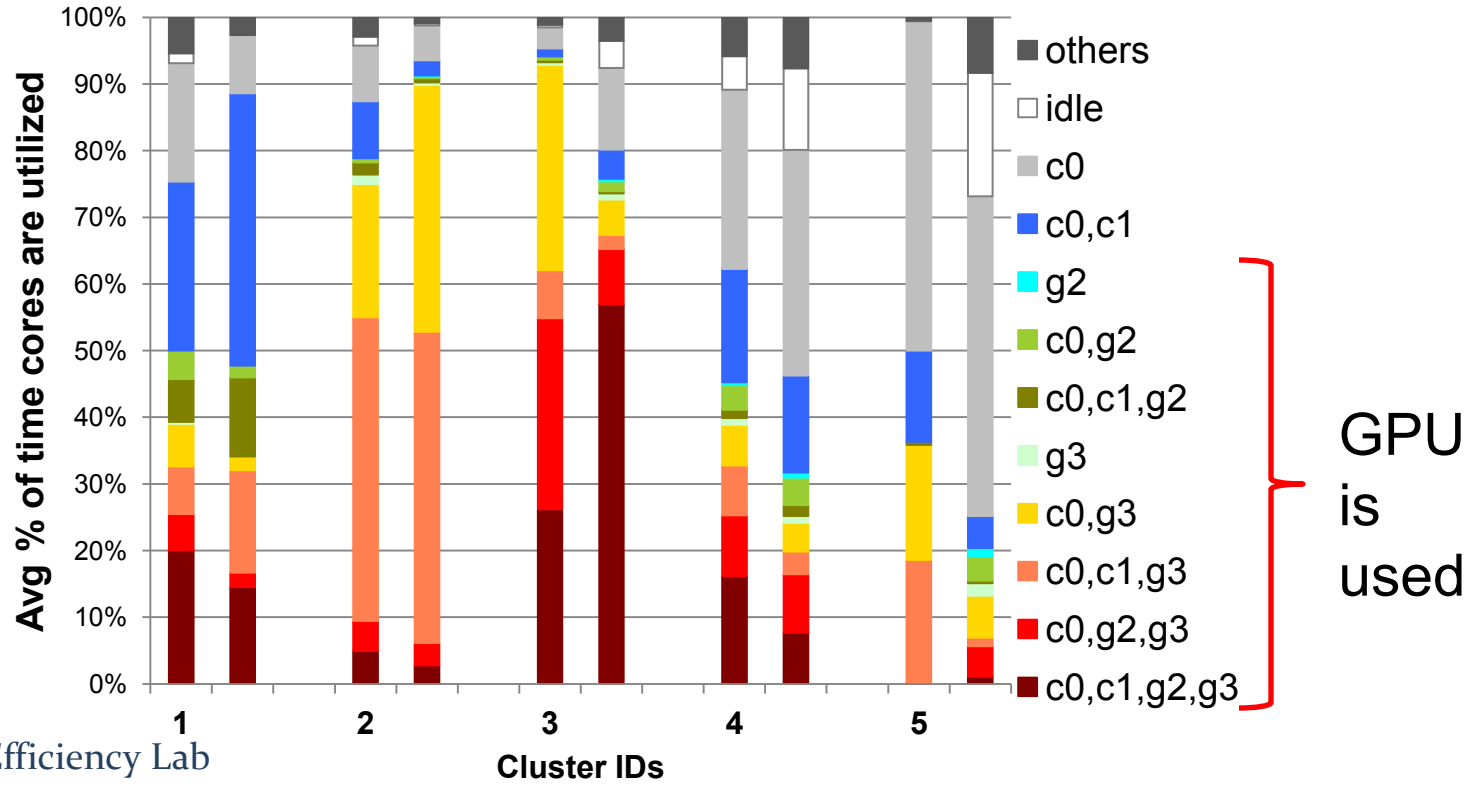
Study of CPU/GPU core usage per application

- We collectively analyzed all application sessions from a user study
- We pick 7 most frequently used applications for a more detailed study
- K-means clustering is used per application to analyze CPU/GPU usage patterns
- Select k so that goodness of fit for a cluster set is at min 80%:

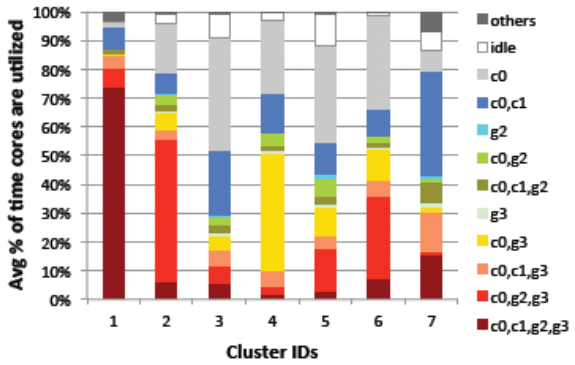
$$\text{Goodness} = \frac{\text{Variation between clusters}}{\text{Total Variation in Data}}$$

Clustering and feature extraction: Browser results

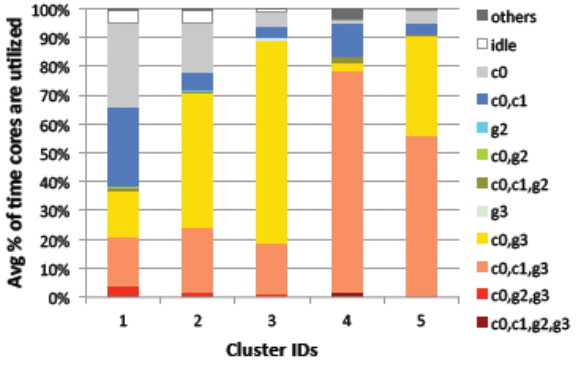
- Total 105 sessions grouped into 5 clusters
- c0,c1 => CPU0/CPU1; g2 => GPU2D; g3 => 3D GPU
- Browser sessions used GPU cores avg 46% of the time!



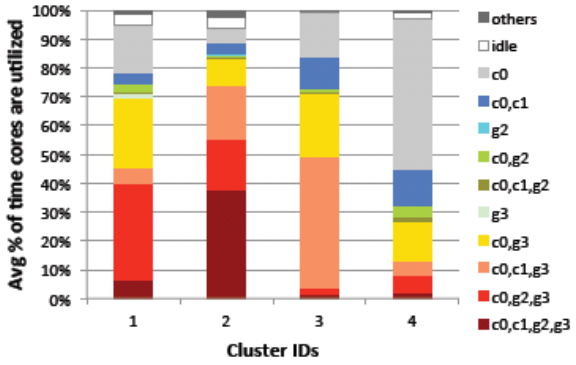
Clustering and feature extraction for 6 other applications



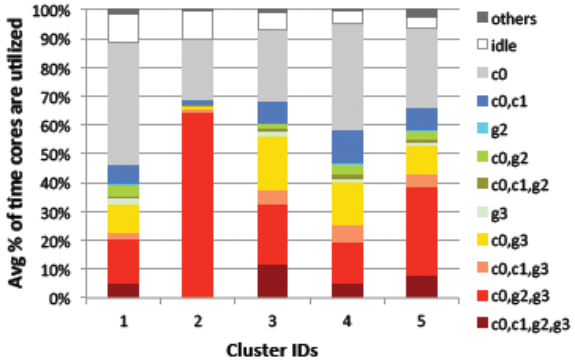
(a) Facebook



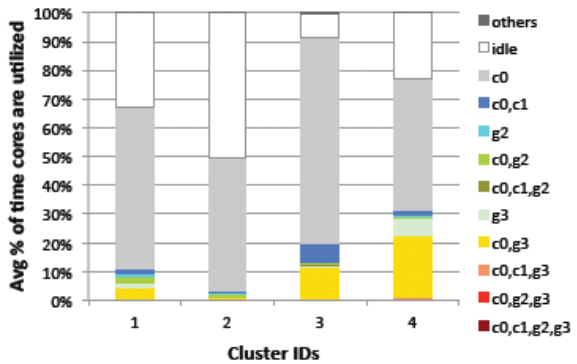
(b) Templerun



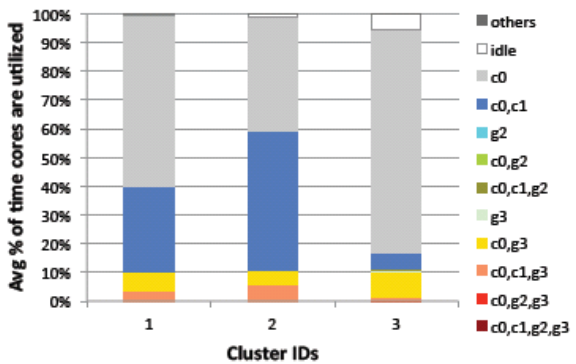
(c) Skype



(d) Email



(e) Music



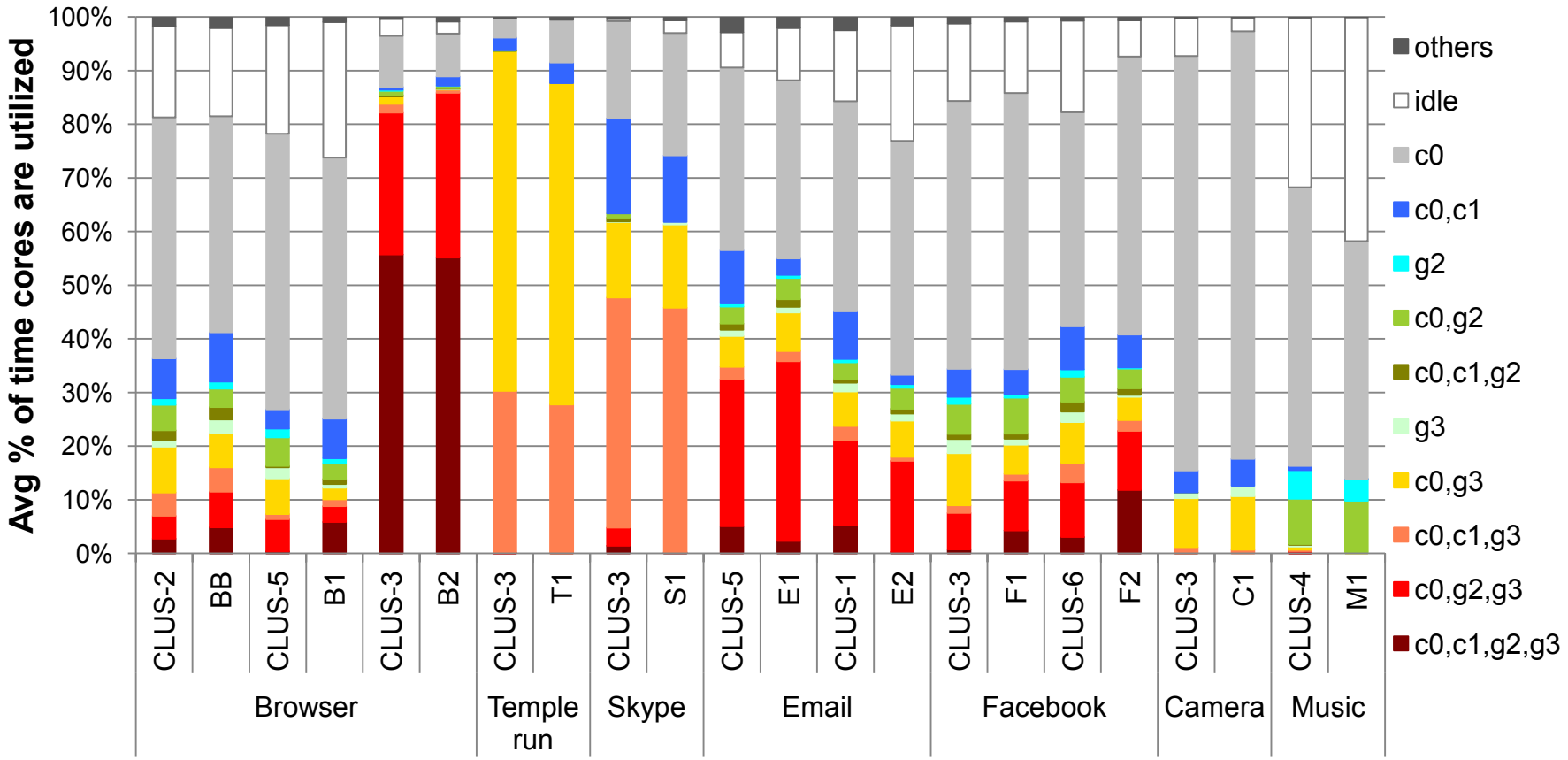
(f) Camera

- 4 apps use GPU cores 50% of the time on average
- Music and camera use accelerators instead of GPUs -> not shown in these results, so they appear CPU dominant

Power & memory BW measurements: Test scenarios

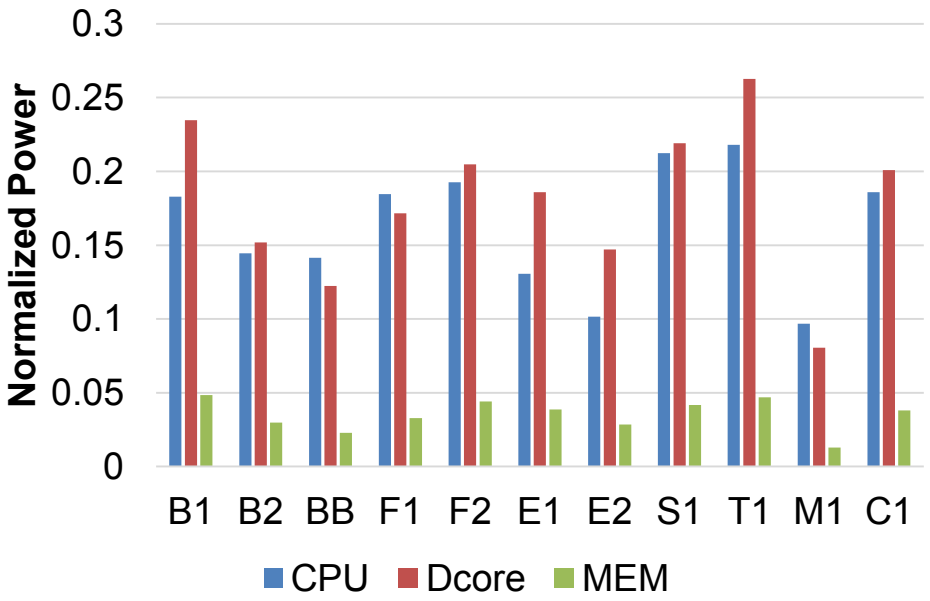
		Interaction
Browser	B1	Search for a Video, Play a Youtube video(A), Scroll while video is playing(B), Scroll when video is stopped(C)
	B2	Search for pictures, swipe through full-screen pictures from Google images(A), Scroll through the image results(B)
	BB	Display and scroll through webpages from 11 sites provided in BBench3.0, with 1s page delay, 0.5s scroll delay, 200px scroll size in 5 iterations.
Facebook	F1	View albums and pictures in a profile
	F2	View a video that plays within the app(A), view photos(B), view a video that plays on an external website(C)
Email	E1, E2	Quick scrolling through four emails multiple times (two text emails, one email with inline photos and one email with a single link that is viewed but not clicked).
Skype	S1	A 5min Skype call, with 1min guest video on, 2mins host front camera on, 1 min host back camera on.
Templerun	T1	A 4min play, with 4-5 instances of lost game lives.
Music	M1	Play music for 2mins, pause 2mins and play again for 2mins.
Camera	C1	Capture a 3min video with intermittent zooming in and out.

Matching sample tests/scenarios with clustered data



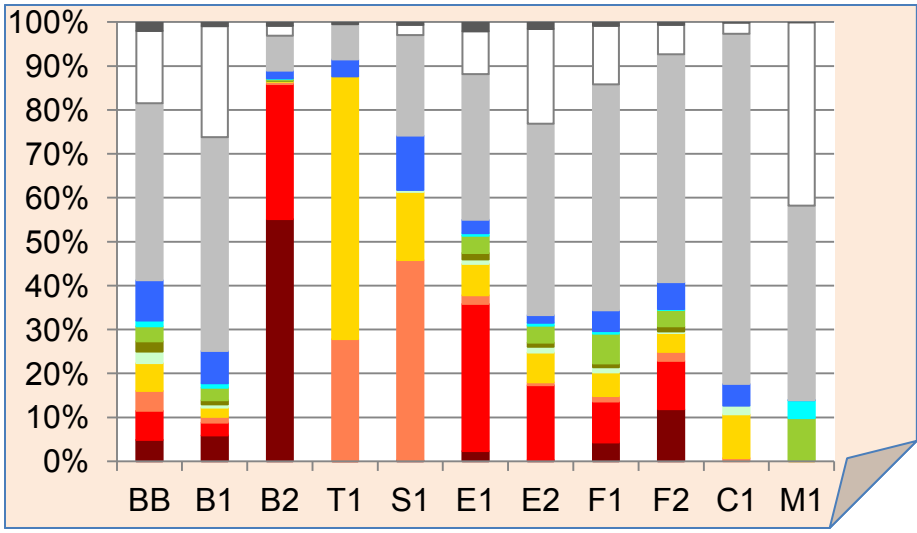
- Test runs and real data are very similar
- Record-and-replay utility enables more detailed data analysis

Power consumption: Representative scenarios



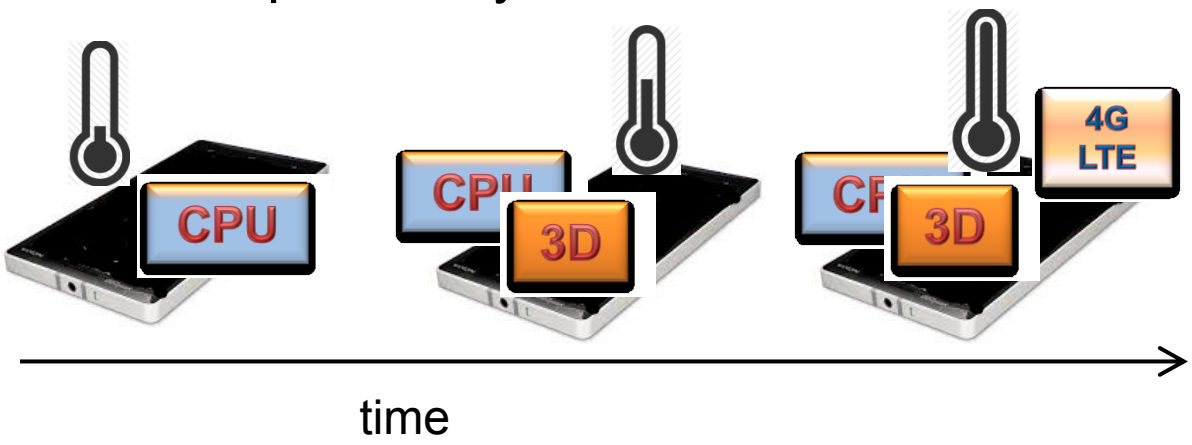
- CPUs, digital cores and memory consume ~50% of the average battery power overall when normalized to their respective average battery power
- Our CPU/GPUs utilization cluster analysis and power measurements generally correlate well, with a few exceptions

- Browser session B2 has higher GPU usage than B1 but
  - Total power of B1 is higher due to video decoder in Digital Core
- Facebook F1 test has less GPU usage than B2, but it consumes more power



Sources of Heat in Mobile Devices

a) Mutual thermal dependency between CPU, GPU, 4G, etc.



b) Ambient condition variations

orientation

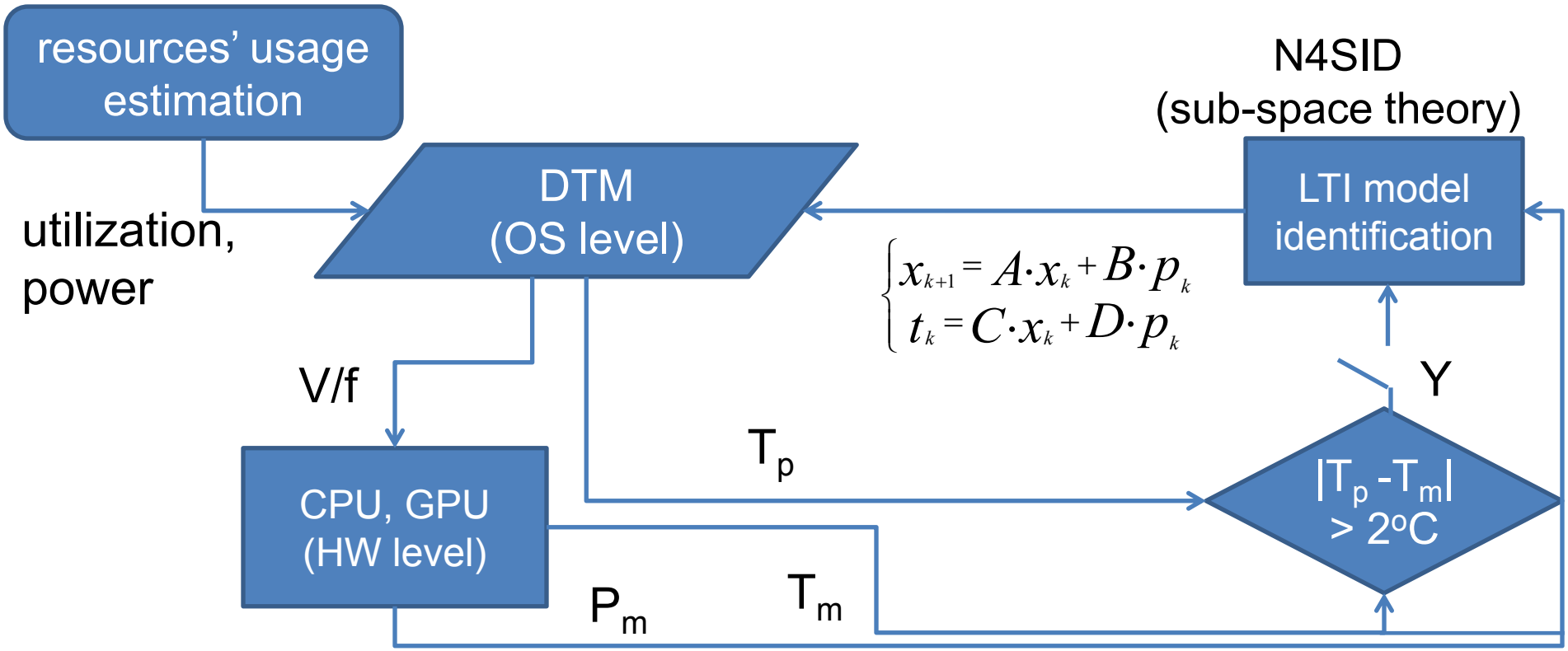
contact surface

electronista.com

electronista.com howtouncle.com

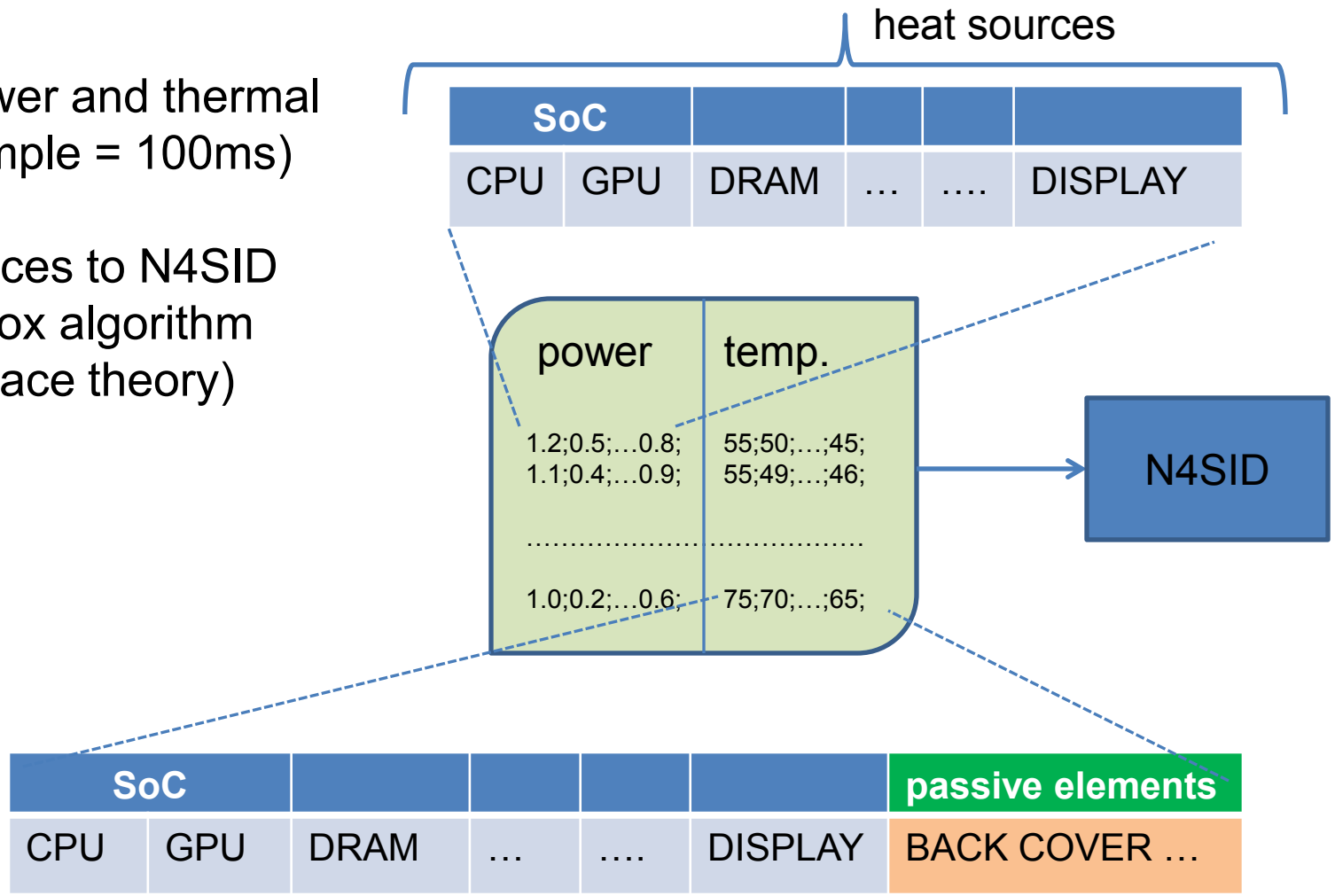
Dynamic Thermal Management

DTM: Dynamic Thermal Management
LTI: Linear Time-invariant Thermal Model
T_P: predicted temperature, T_m: measured temperature, P_m: measured Power



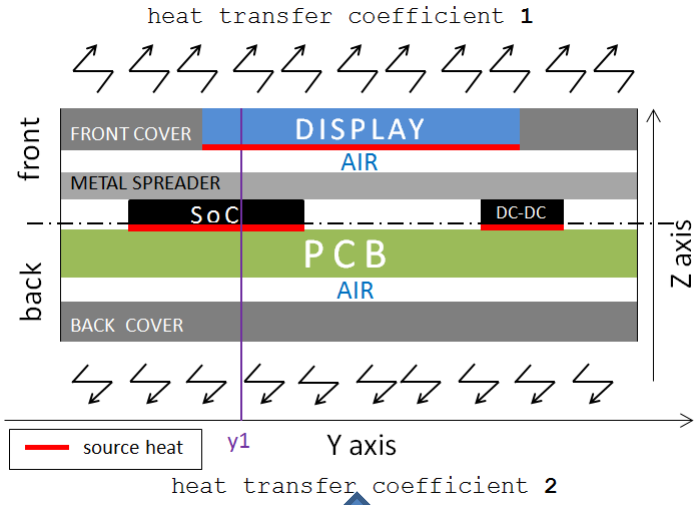
LTI Identification

- Collect power and thermal traces (sample = 100ms)
- Provide traces to N4SID
  - blind box algorithm (subspace theory)

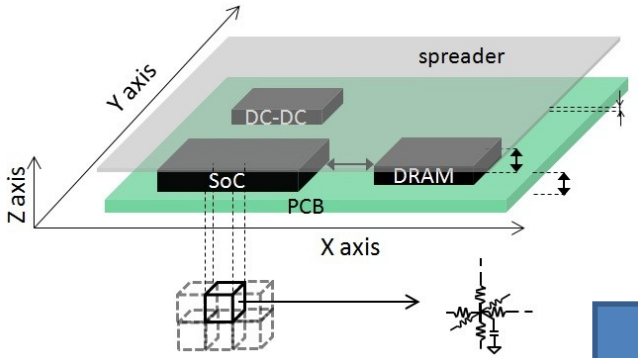


Experimental Setup

Using RC model of the target smartphone to have the extra thermal sensors needed

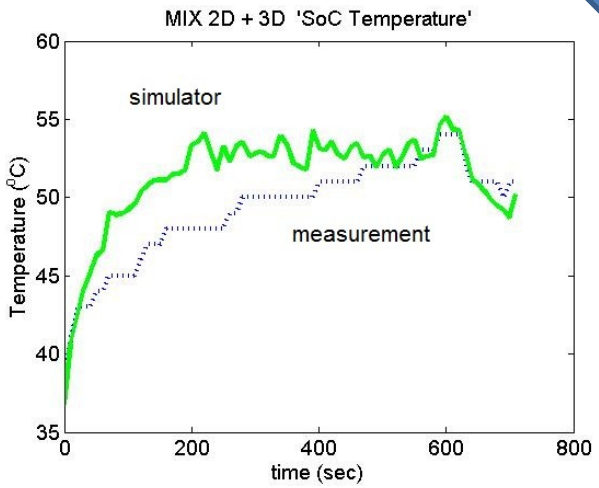


3D-ICE
simulations



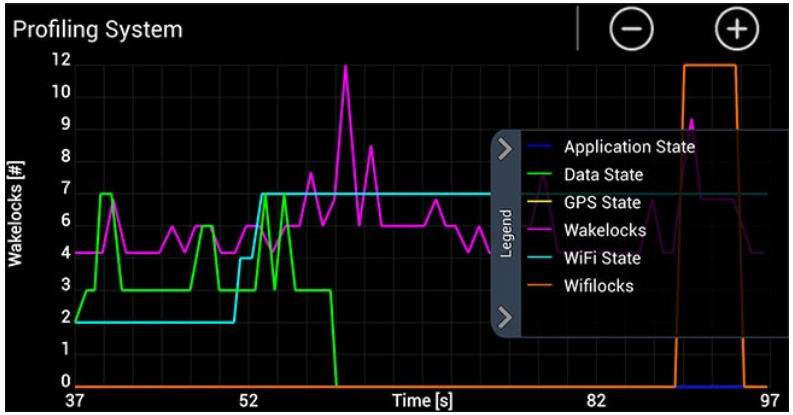
Snapdragon
MSM8660

measurements



Experimental Setup

- Four different suites:
  3D / 2D / MATH (0xBench) and Browser Benchmark
- Runs at different CPU and GPU frequency configurations
  - Evaluate cost $\text{COST}_{CPU}|_{f_{GPU}} = \text{UTIL} \times f_{CPU}$
    $\text{COST}_{GPU}|_{f_{CPU}} = \text{UTIL} \times f_{GPU}$
  - Measure power

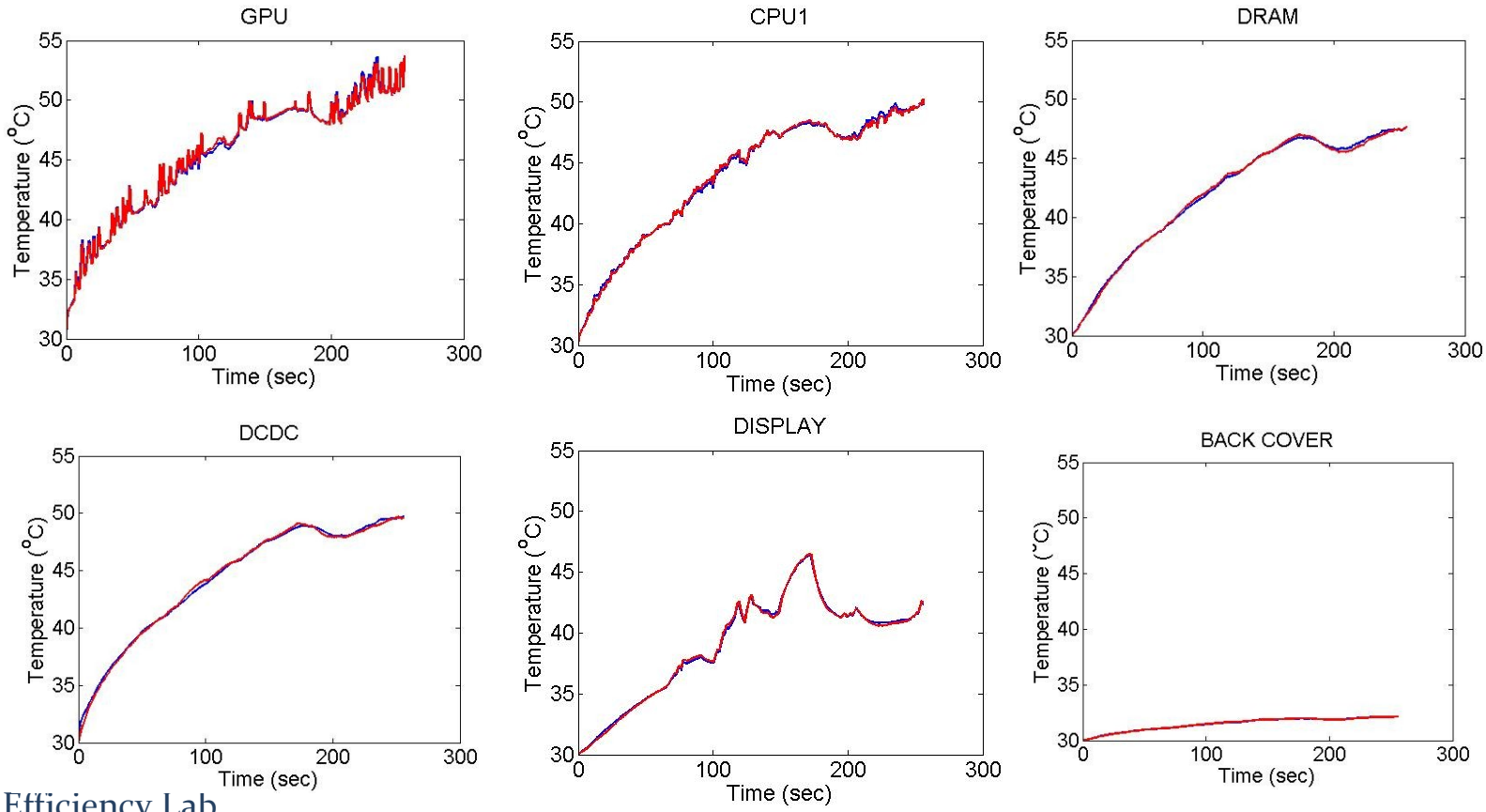


LTI Identification

- Average and Maximum errors are 0.3°C and 1.5°C
- Identification requires a few seconds on average

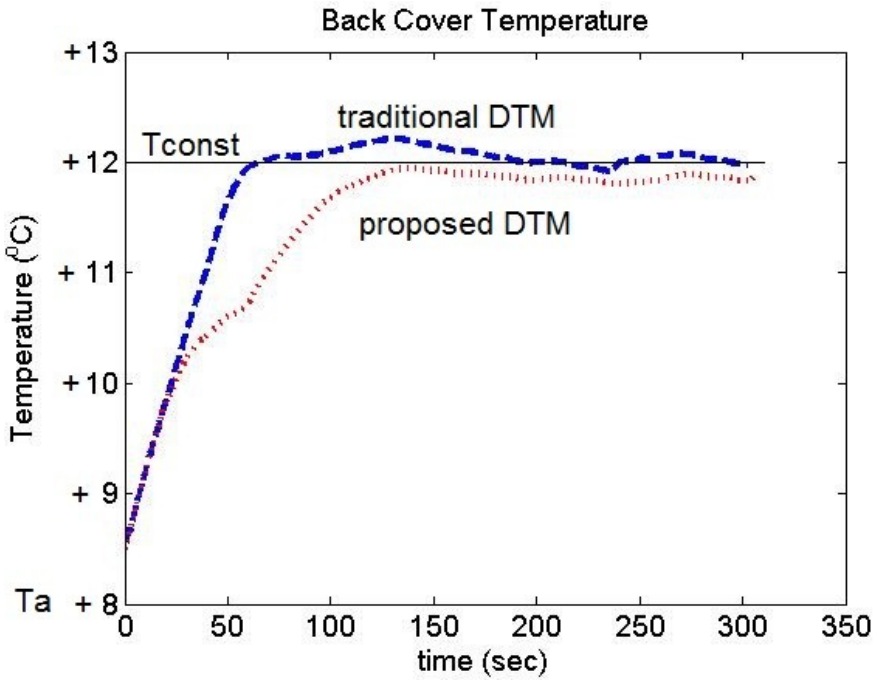
simulator

LTI



Ambient-aware DTM

- Thermal management leverages context of the phone, the LTI model, and detected application phases to adjust the management decisions as the environment changes



Summary

- We have illustrated important issues related to
  - Resources' usage versus apps
  - Resources' usage and ambient condition variations impact on the device temperature

- Our contributions can be summarized as follows:
  - Utilization and Power App Characterization Technique
    - Significant use of GPUs in real usage scenarios
  - Thermal/cooling management

Next Steps

- User study and characterization
  - generate use case models
  - define a Quality of Experience (QoE) metric
  - in-depth study of user-app-HW architectures

- Improve dynamic thermal control
  - realistic use case models
  - QoE constraint
  - maximize the energy-efficiency