MACHINE LEARNING MAKING SMARTPHONES SMARTER

BY SOMDIP DEY

CEO @ NOSH TECHNOLOGIES && LECTURER @ UNIVERSITY OF ESSEX

CAN YOU RELATE?

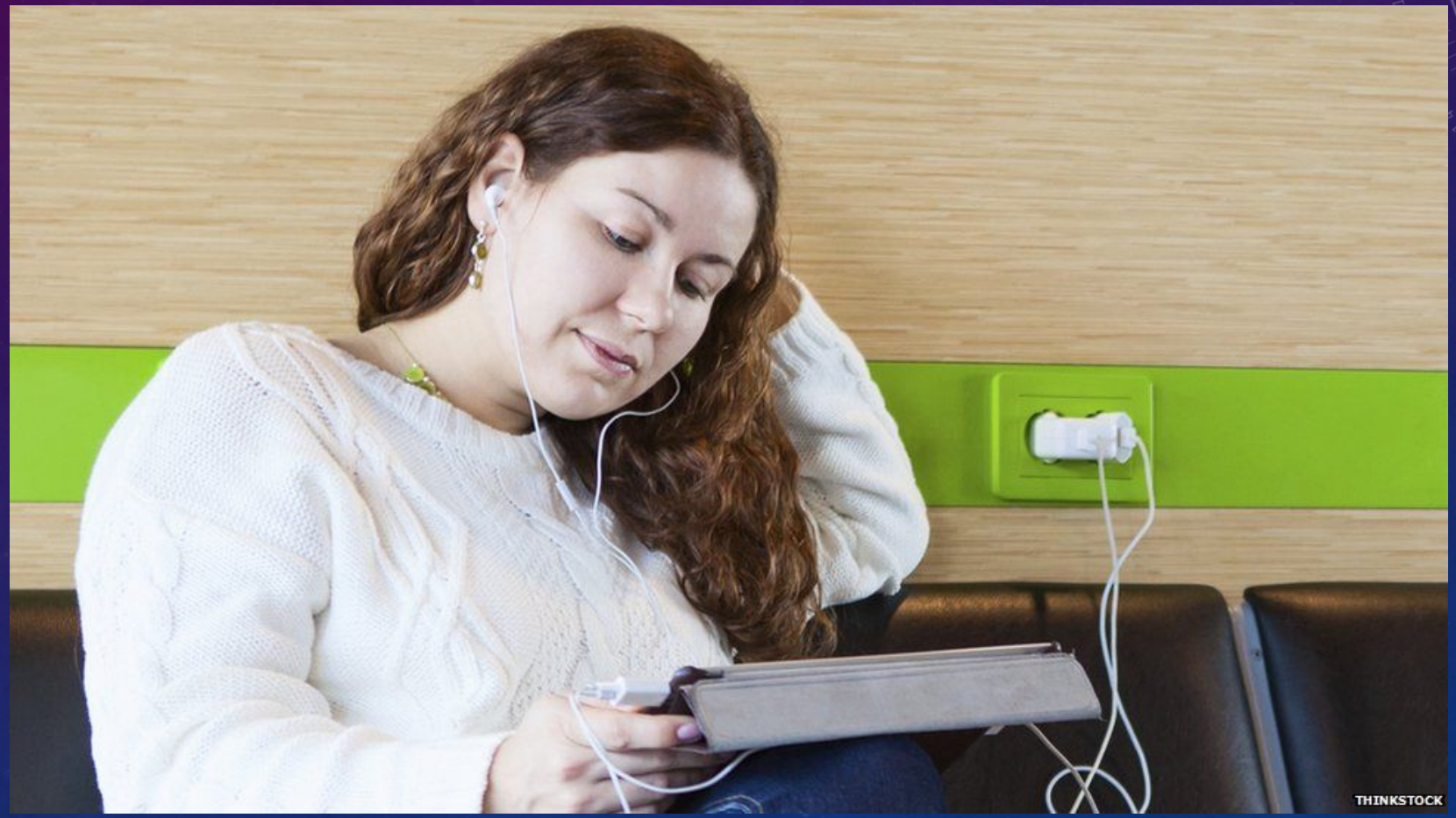


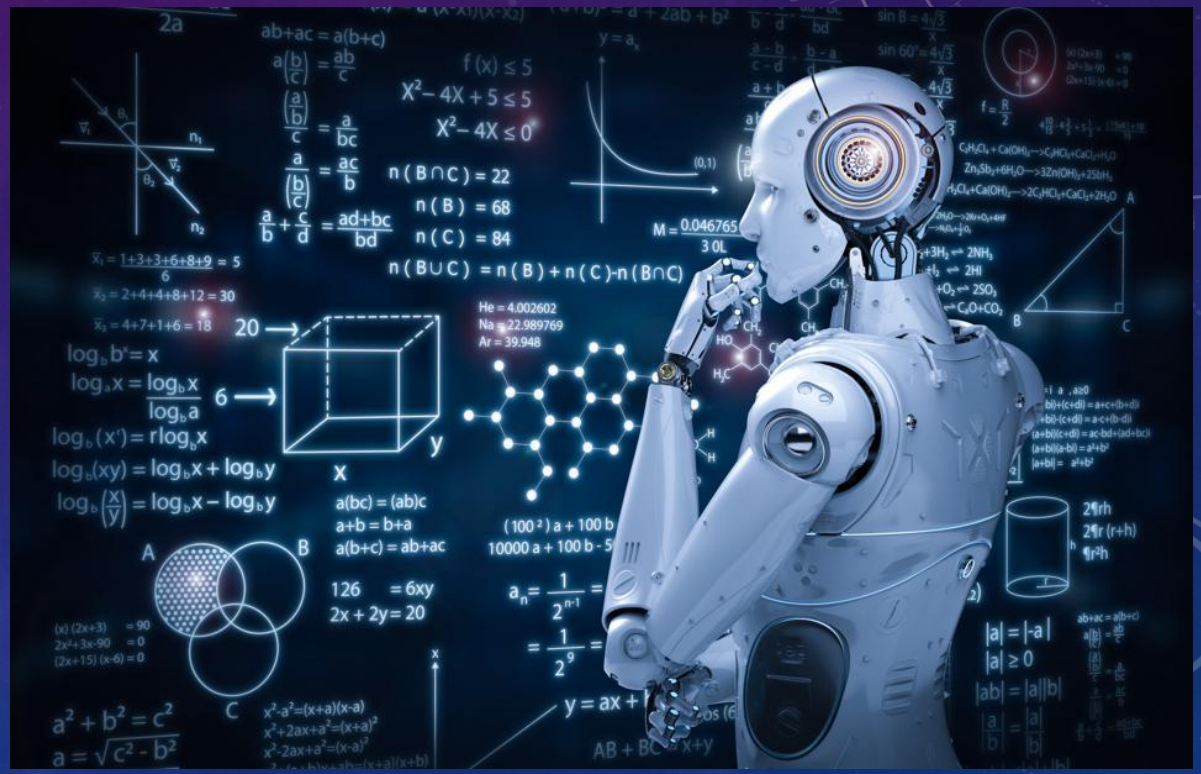
TABLE OF CONTENT

- Brief recap & advancement of AI and Machine Learning
- Power and thermal management in smartphones
- Next: Future smartphones can prolong battery life
- Future Works

AI AND MACHINE LEARNING

BRIEF RECAP OF AI AND MACHINE LEARNING

- Artificial Intelligence or AI is the field of study to enable computing machines to perform tasks that are commonly associated with intelligent beings like humans.
- Machine Learning or ML is a sub-field of AI where the computing machine can improve automatically through experience and by the use of data.



MACHINE LEARNING

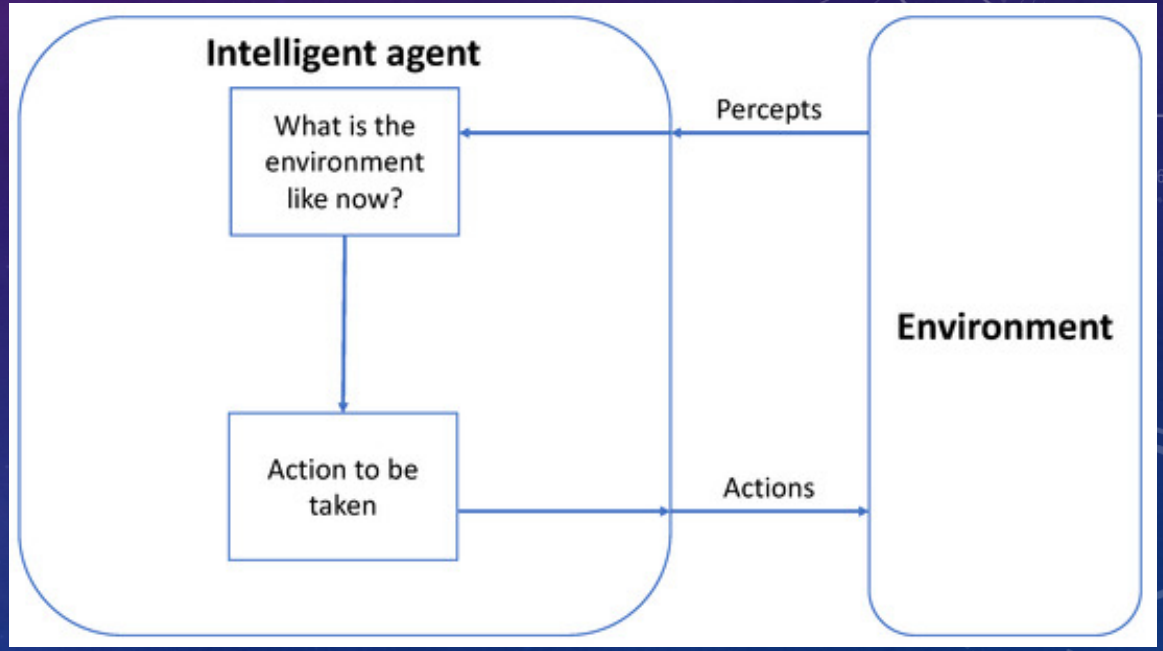
- Traditionally, ML can be broadly categorized into : **Supervised**, **Unsupervised** and **Reinforcement Learning (RL)**.
- Gave rise to other types of ML algorithms especially **Semi-Supervised Learning**.
- Two types of ML methodologies, which are important for this talk, are **Embedded Machine Learning (EML) and Reinforcement Learning (RL)**.

WHAT IS EMBEDDED MACHINE LEARNING

- **Embedded Machine Learning** (EML) is a sub-field of machine learning, where the machine learning model is run on embedded systems with limited computing resources such as wearables, edge devices and microcontrollers.
- Privacy and security of data is a key concern to many consumers with respect to AI/ML and embedded machine learning is best suited in this scenario as the ML models trains and infer on the embedded systems.

WHAT IS REINFORCEMENT LEARNING

- Reinforcement learning (RL) is a type of machine learning algorithm, where an intelligent agent, which is a computing system that perceives its environment to take actions autonomously in order to achieve cumulative rewards based on the knowledge gathered from the environment.



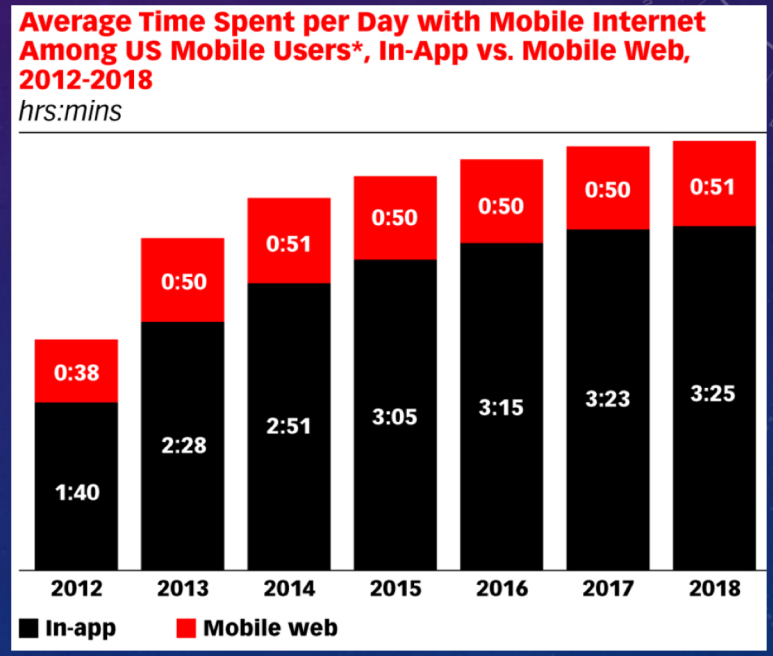
SMARTPHONES HAVE BECOME AN INTEGRAL PART OF OUR LIVES

SMARTPHONE USAGE

- In 2020, there were 3.5 billion smartphone users in the world.
- In 2019, 56% of all website traffic worldwide was generated through mobile phones.
- According to eMarketer, in 2018 mobile users spent 4 hours 16 minutes on an average on in-apps and mobile web.

SMARTPHONE USAGE STATISTICS

- Stats published by eMarketer shows an increase of 15.32% in time spent on in-app and mobile web by mobile users in 5 years.
- According to Mediakix, an average of 1 hour 56 minutes was spent on top 5 social media platforms: Youtube, Facebook, Snapchat, Instagram and Twitter.



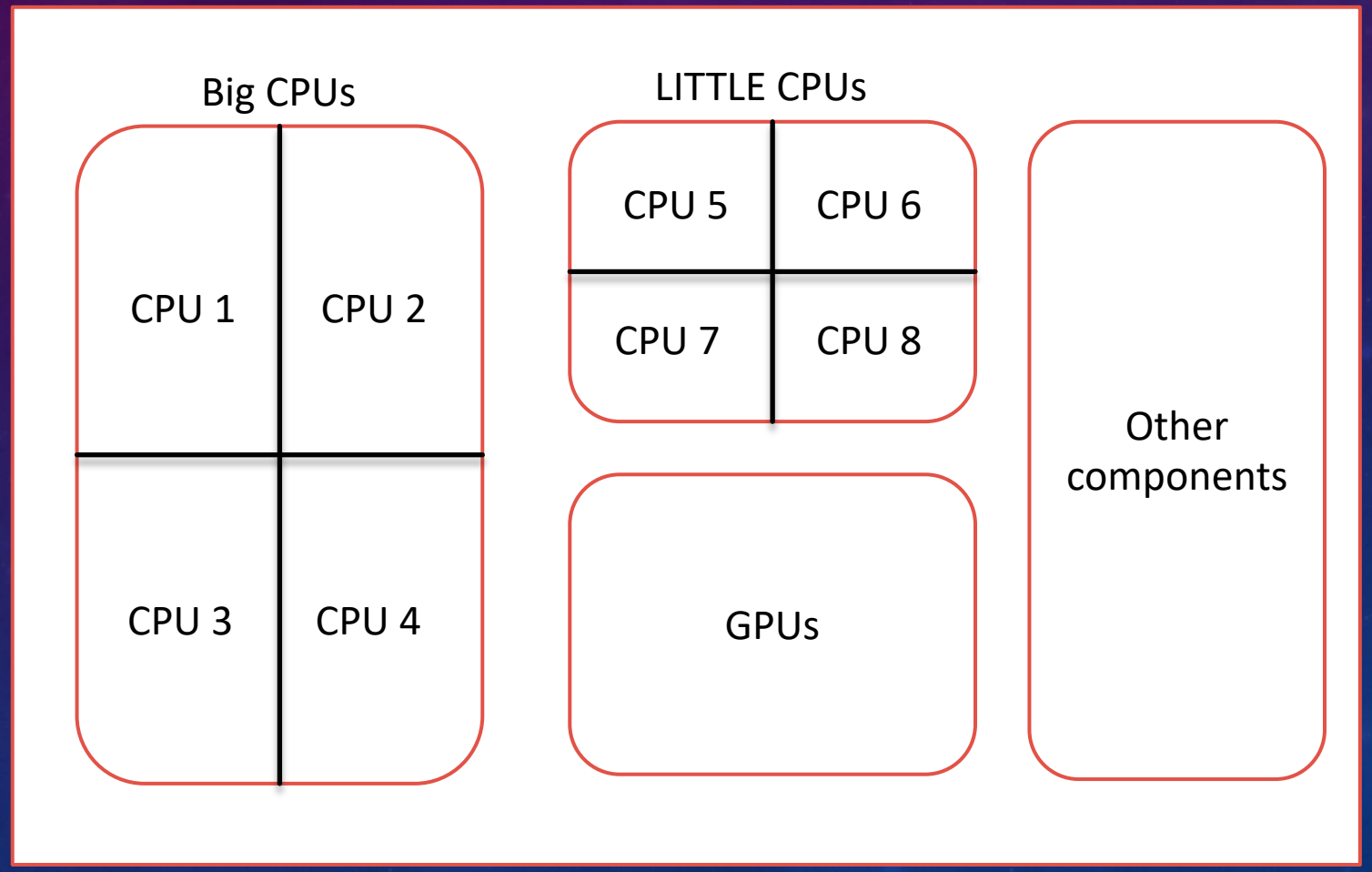
SMARTPHONE USAGE STATISTICS

- According to Rescue Time, an average person picks-up/look at their phones 52 times during their workday.
- 70% of the sessions are less than 2 minutes, 25% of the sessions lasting between 2 to 10 minutes and 5% of the sessions prolonged more than 10 minutes.
- Even the duration of the user picking-up/looking at their Edge device every time varies from user to user and hence, making the sessions stochastic in nature.
- **Conclusion**: Mobile phone users are increasing and the interaction session varies inter-session as well as from user to user.

MODERN STANDARD OF SMARTPHONES

- Most modern smartphones are equipped with heterogeneous multiprocessor systems-on-chips (MPSoCs).
- Example of devices using heterogeneous MPSoCs are Samsung S9, Note 9, Huawei P20, iPhone X, etc.

MPSOC (MULTI-PROCESSOR SYSTEM ON A CHIP)



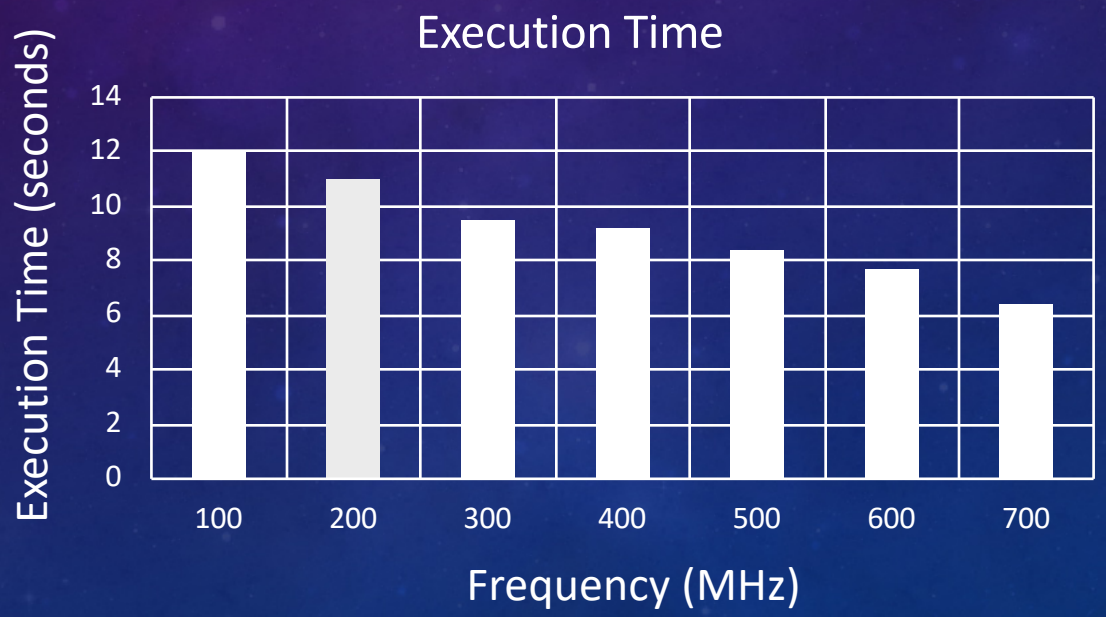
POWER AND THERMAL MANAGEMENT IN SMARTPHONES

- *Dynamic Power Management (DPM)* allows idle processing elements (PEs) or other idle components of the system to be suspended if required in order to reduce energy consumption.
- *Dynamic Voltage Frequency Scaling (DVFS)* allows processors to operate at variable voltage-frequency (v-f) levels.
- Customization of processors to match the processing needed of a task on an MPSoC.
- Customizing cache based memory access.
- Mapping tasks of an application to the processors so that workload could be balanced across all processors in an MPSoC. This improves utilization of processing elements (PEs) effectively and reduces energy consumption.

Reference:
Dey, Somdip, et al. "Rewardprofiler: A reward based design space profiler on dvfs enabled mpsocs." *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*. IEEE, 2019.

DYNAMIC VOLTAGE FREQUENCY SCALING (DVFS)

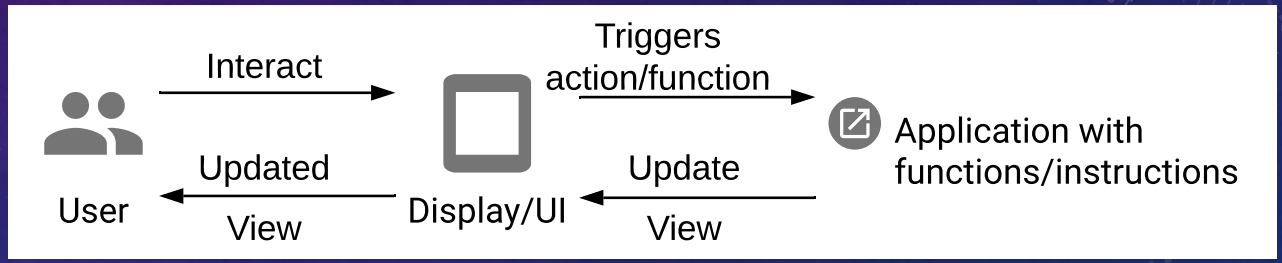
- Dynamic Voltage Frequency Scaling (DVFS) is used to reduce dynamic power consumption ($P \propto V^2f$).



OBSERVATIONS LEADING TO THE RESEARCH

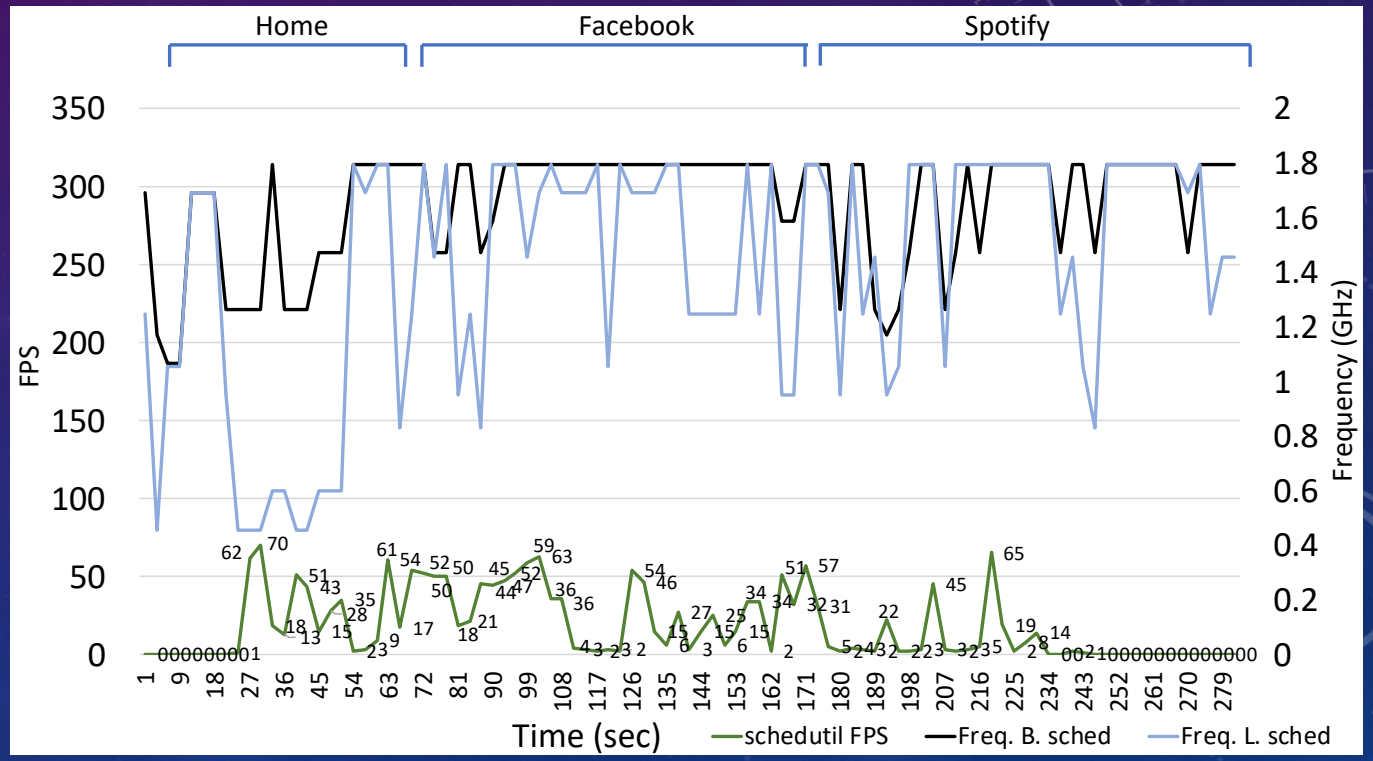
OBSERVATION NO. 1

- User’s interaction with the mobile device with touch screen.



OBSERVATION NO. 2

- Session of 5 minutes using Home screen, Facebook and Spotify apps on a mobile device (Samsung Galaxy Note 9 utilizing Exynos 9810 MPSoC).
- Higher operating frequency does not generate higher FPS.



RESEARCH FOCUS

- Main focus of this work is to meet QoS while optimizing power consumption and peak thermal behaviour based on user's interaction with the application.
- A new metric to incorporate power consumption and peak temperature to evaluate the performance at a given time period: *performance per degree watt* (*PPDW*).

$$PPDW_i = \frac{FPS_i}{\Delta T \times P_i}, \text{where } \Delta T = T_i - T_a$$

- T_i, P_i and FPS_i are the Temperature, Power consumption and FPS at that time period. T_a is the ambient temperature.

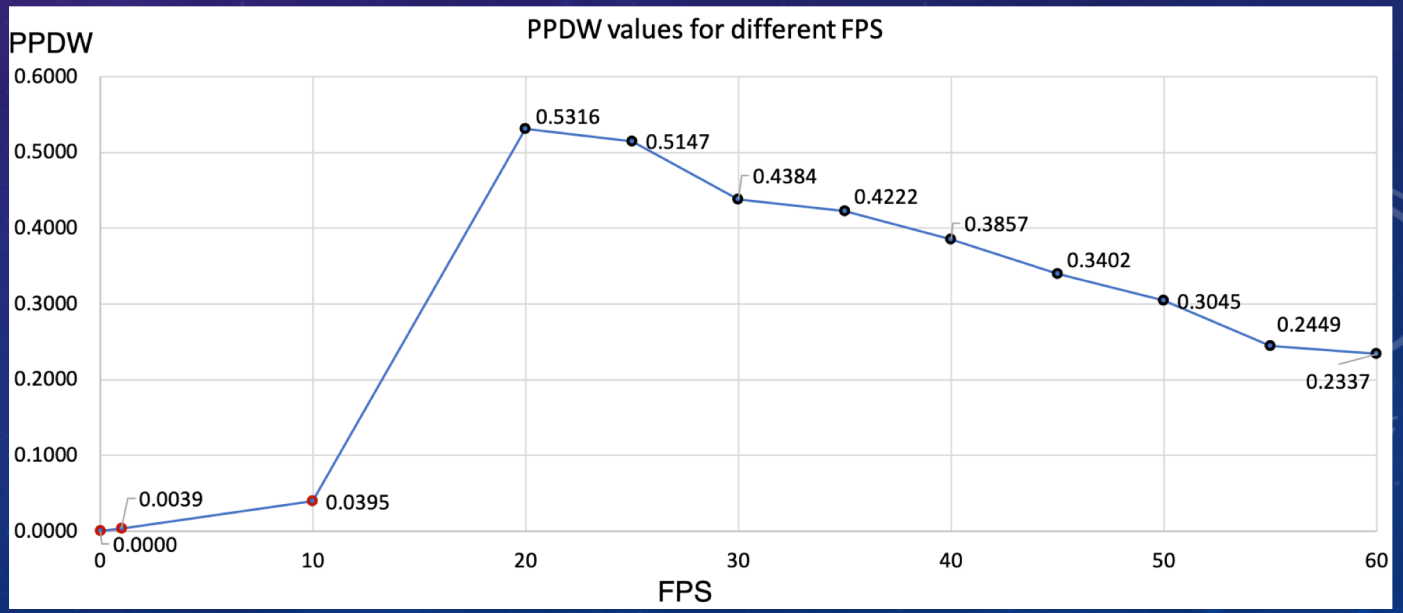
Reference:
Dey, Somdip, et al. "User interaction aware reinforcement learning for power and thermal efficiency of CPU-GPU mobile MPSoCs." *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2020.

RESEARCH FOCUS: REWARD FUNCTION CREATION

- Main objective is to minimize the value of PPDW, while the optimal minimal value, $PPDW_{desired}$, needs to be between $PPDW_{worst}$ and $PPDW_{best}$.
- Where, $PPDW_{worst} = FPS_{least}/((T_{max} - Ta) \times P_{max})$ &
- $PPDW_{best} = FPS_{max}/((T_{least} - T_a) \times P_{least})$
- General trend of PPDW

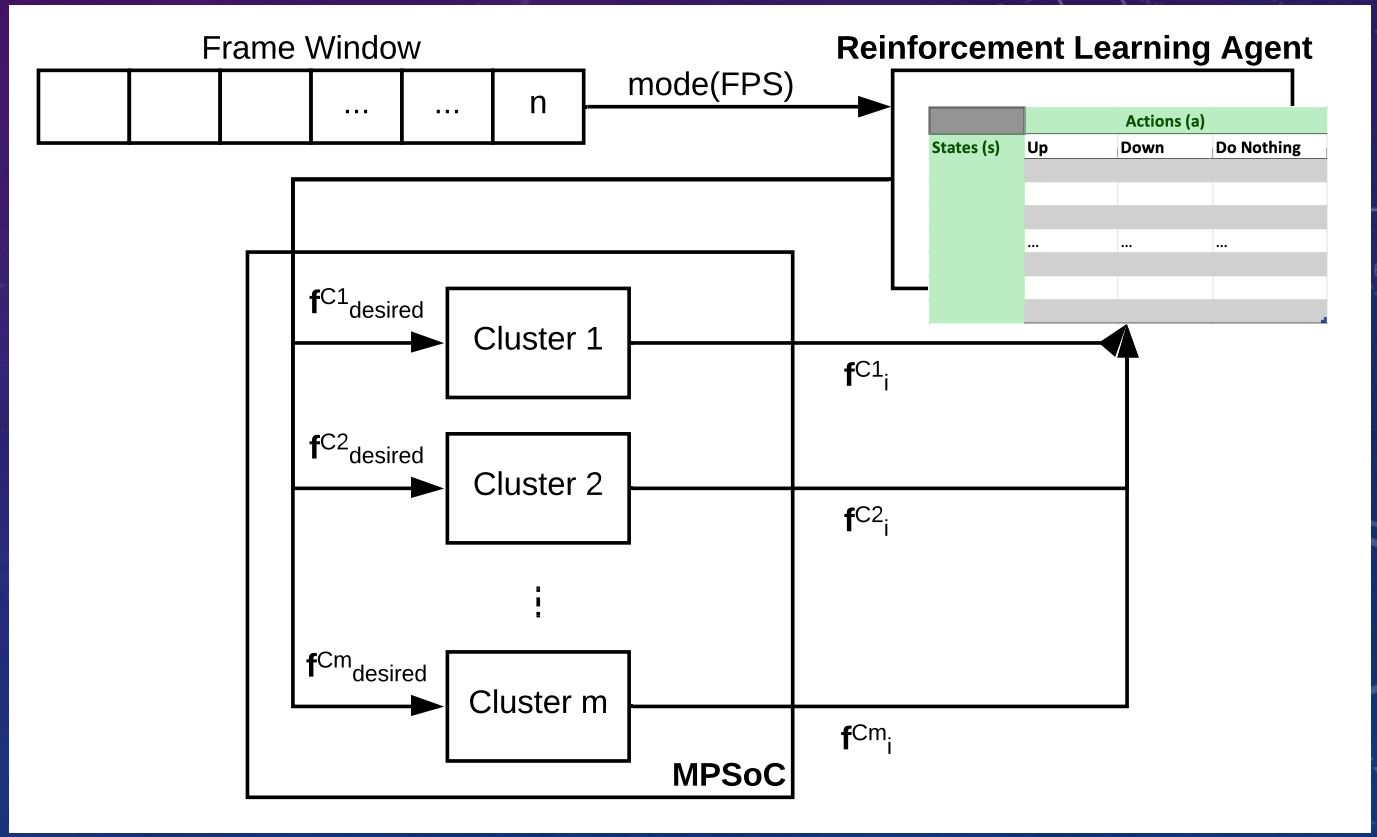
while executing Lineage 2

in the figure.



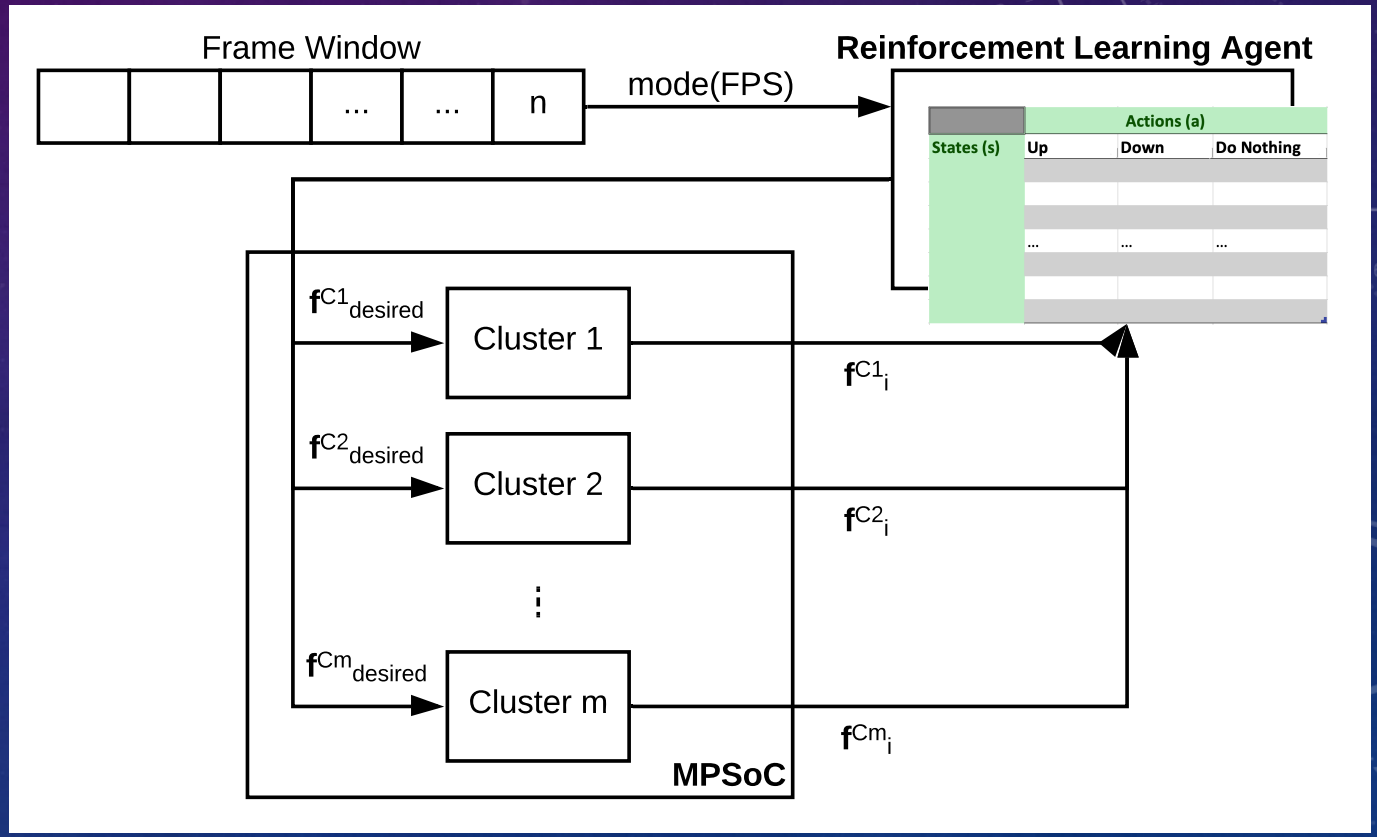
PROPOSED METHOD: NEXT

- Continuously monitors the frame rate every 25 ms for a window of *n* seconds.
- Compute mathematical mode of frame rates collected over *n* seconds to determine the desired FPS (Target FPS) for the executing app during that session.



PROPOSED METHOD: NEXT (CONTINUED)

- Target FPS is fed to the Reinforcement Learning (RL) agent.
- RL agent then selects either of the 3 actions: frequency up, frequency down, do nothing; for each of the *m* clusters of PE so that Target FPS could be met.
- Once the training is complete *maxfreq* for each PE cluster is to the appropriate frequency.



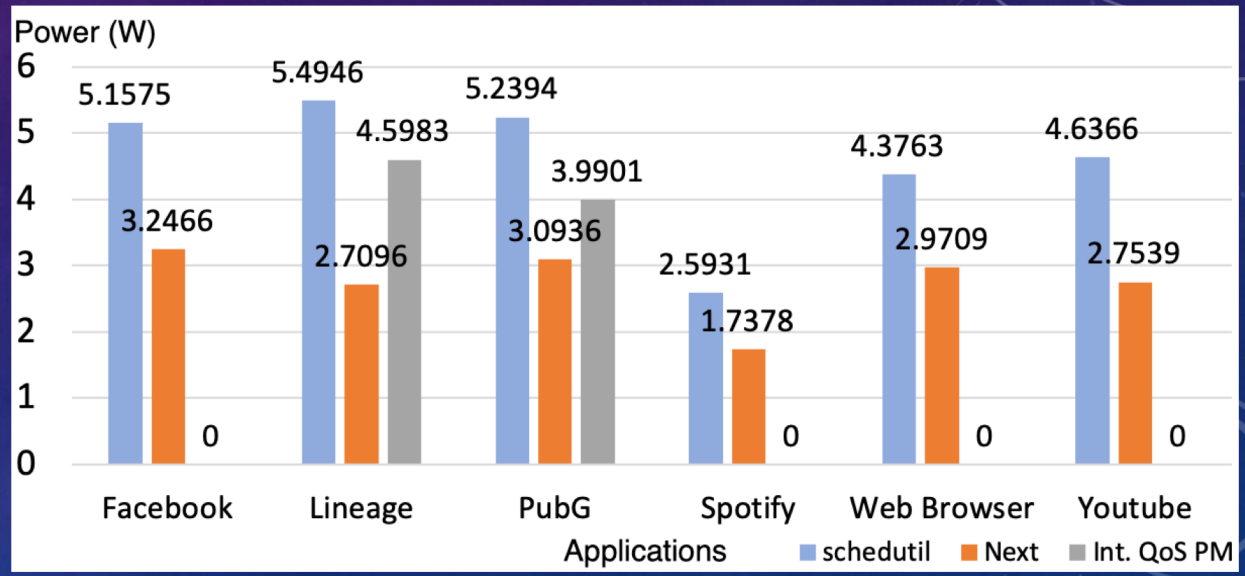
PROPOSED METHOD: NEXT (CONTINUED)

- The reward function for the RL agent: $R(s_i,a_i) = PPDW_i$ where s_i is the state of the agent and a_i is action taken by the agent.
- The agent's goal is to maximize reward, which means the agent has to optimize *PPDW* and achieve $FPS_{current} = \text{Target FPS}$

EXPERIMENTAL RESULTS: POWER OPTIMIZATION

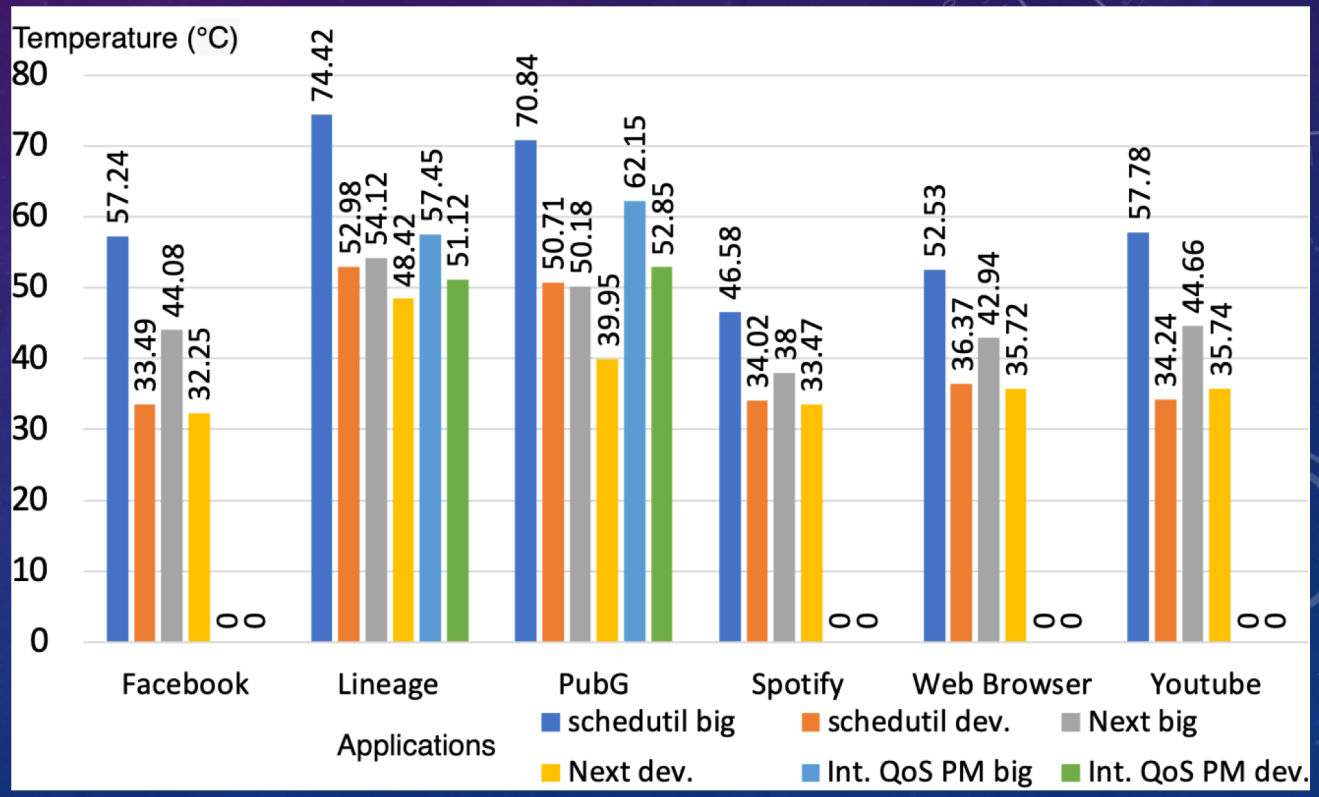
- For Next methodology the power savings for Facebook, Lineage, PubG, Spotify, Web Browser and Youtube compared to schedutil are 37.05%, 50.68%, 40.95%, 32.98%, 32.11% and 40.6% respectively.

*Int. QoS PM: QoS- aware power management methodology for gaming apps proposed by Pathania et al. (Pathania 2014)



EXPERIMENTAL RESULTS: TEMPERATURE OPTIMIZATION

- In the figure, the average peak temperature of big CPU cluster and the Samsung Note 9 device are shown.
- Compared to schedutil Next is capable of reducing peak temperature by 29.16% (maximum) for big CPUs and 21.21% (maximum) for the device in general.



FUTURE WORKS AND CONCLUSION

- Machine Learning is capable of understanding our human behavior with computing machines like smartphones.
- Understanding such patterns could greatly benefit how such a device could be charged or battery performance and performance of the device could be optimized.
- Currently, my team has developed a mechanism to also understand how we charge our smartphones and using such patterns the ML is capable of prolonging the battery of the smartphone without the need to constantly charge it.
- AI/ML is still in its infancy comparatively what we can achieve in the next 10 years and we have a long way ahead.

Reference:
Isuwa, S., Dey, S., Ortega, A. P., Singh, A. K., Al-Hashimi, B. M., & Merrett, G. V. (2022). QUAREM: Maximising QoE through Adaptive Resource Management in Mobile MPSoC Platforms. *ACM Transactions on Embedded Computing Systems (TECS)*.

Thank you! Any Questions?

- Email: somdip.dey@essex.ac.uk or dey@nosh.tech