Energy-efficient execution of Federated learning tasks on mobile phones: An exploratory study.

Presented by Patrick Wapet, Post Doc at LIRIS Laboratory, INSA Lyon
In collaboration with Dr. Tran Giang Son, University of Science and Technology of Hanoi
and Dr. Boris Teabe, INP Toulouse
Supervised by By Vlad Nitu,

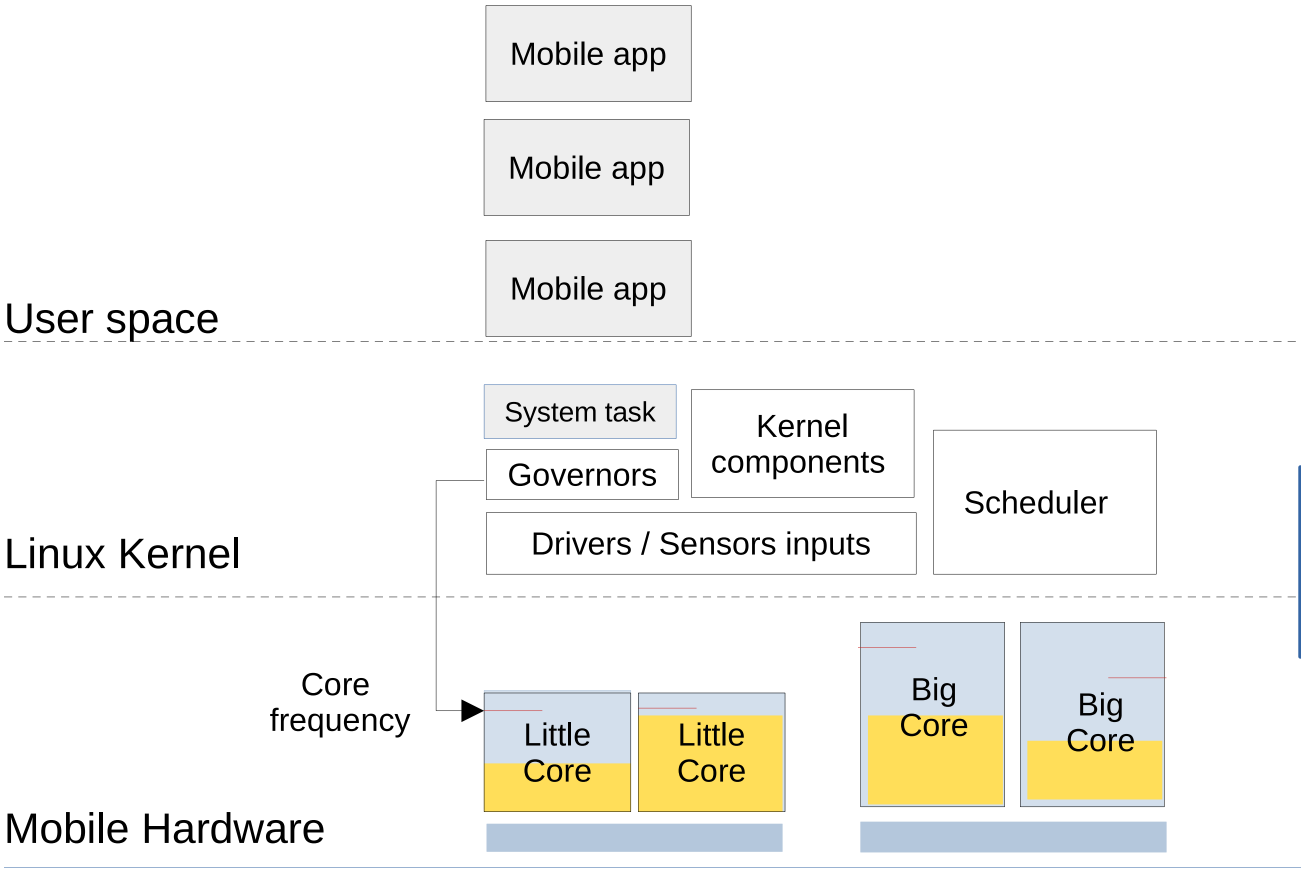
Summary

1. **Context:** Federated Learning and mobile phones.
2. **Problem definition:** Global scheme
3. **Challenges:** Parameters, metrics, approach and measurement tools.
4. **Experiments and observations:** reported according to the parameters, graphs and partial conclusions.
5. **Next steps:** next experiments, possibly implementations and submissions.

1. Context: Federated Learning

- Artificial Intelligence is more and more used in everyday life.
- By default it is a system that centralizes data.
- Posing the problem of privacy.
- A solution: keep the data with the users.
- On their devices : **Mobile phones**
- Do the processing on these phones: **Federated Learning**

2. Let state the problem: general scheme



Mobile app

Mobile app

Mobile app

User space

System task

Kernel components

Governors

Scheduler

Drivers / Sensors inputs

Linux Kernel

Core frequency

Little Core

Little Core

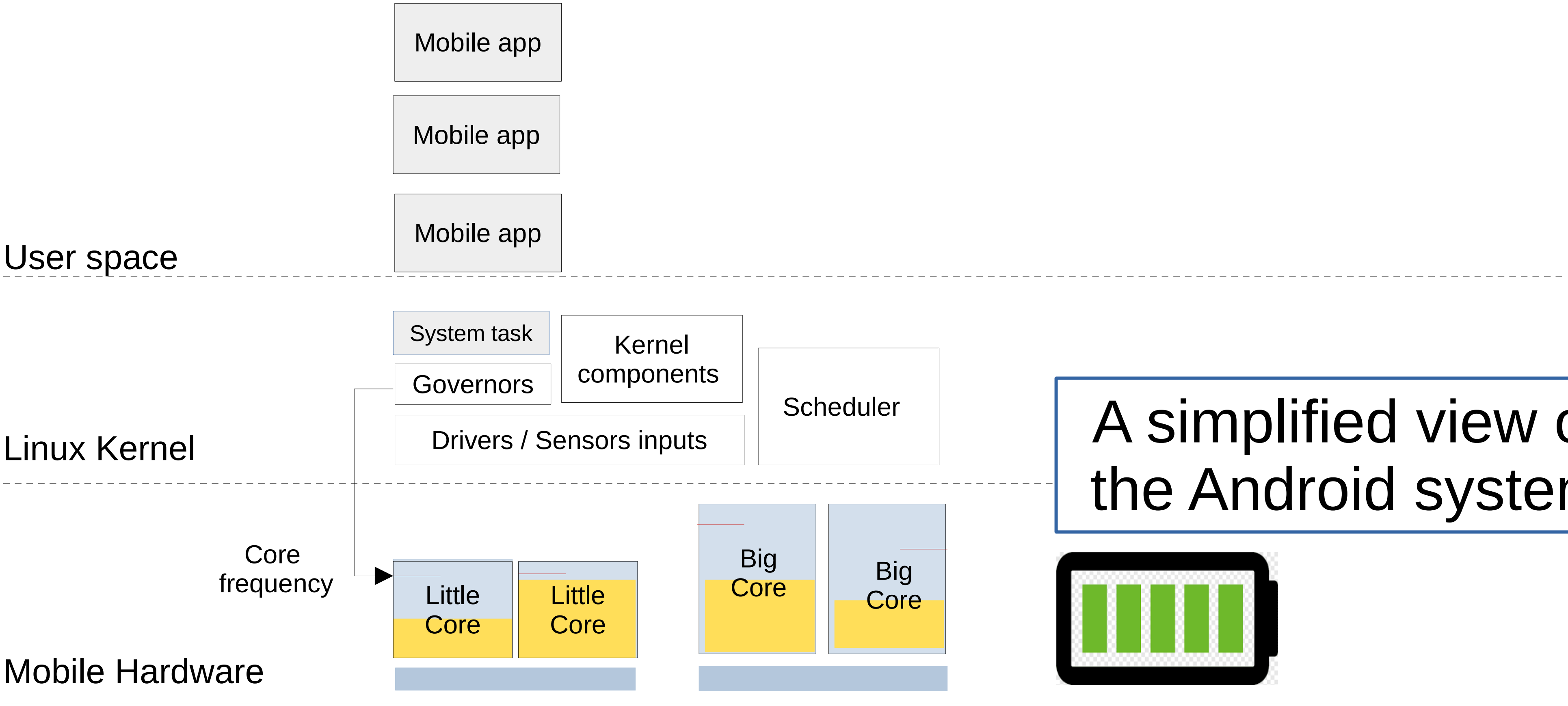
Big Core

Big Core

Mobile Hardware

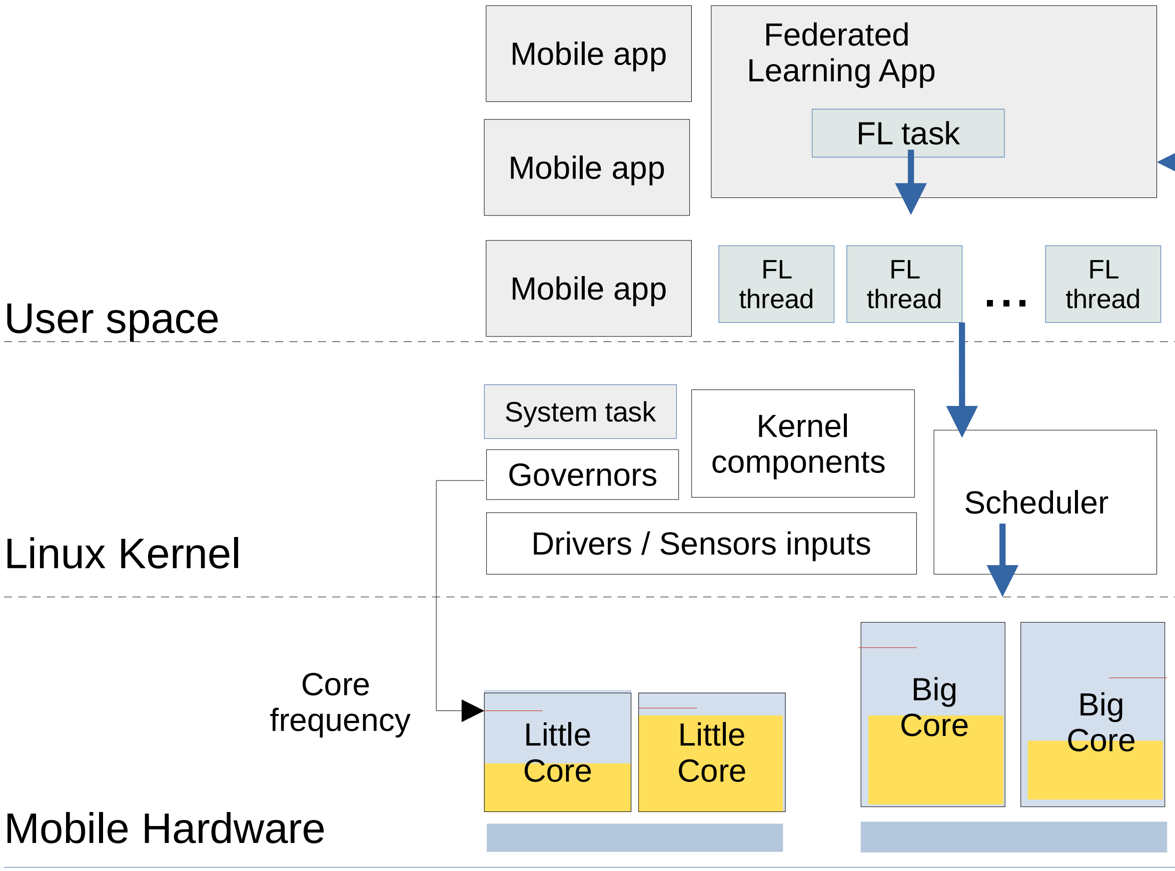
A simplified view of the Android system

2. Let state the problem: general scheme



A simplified view of the Android system

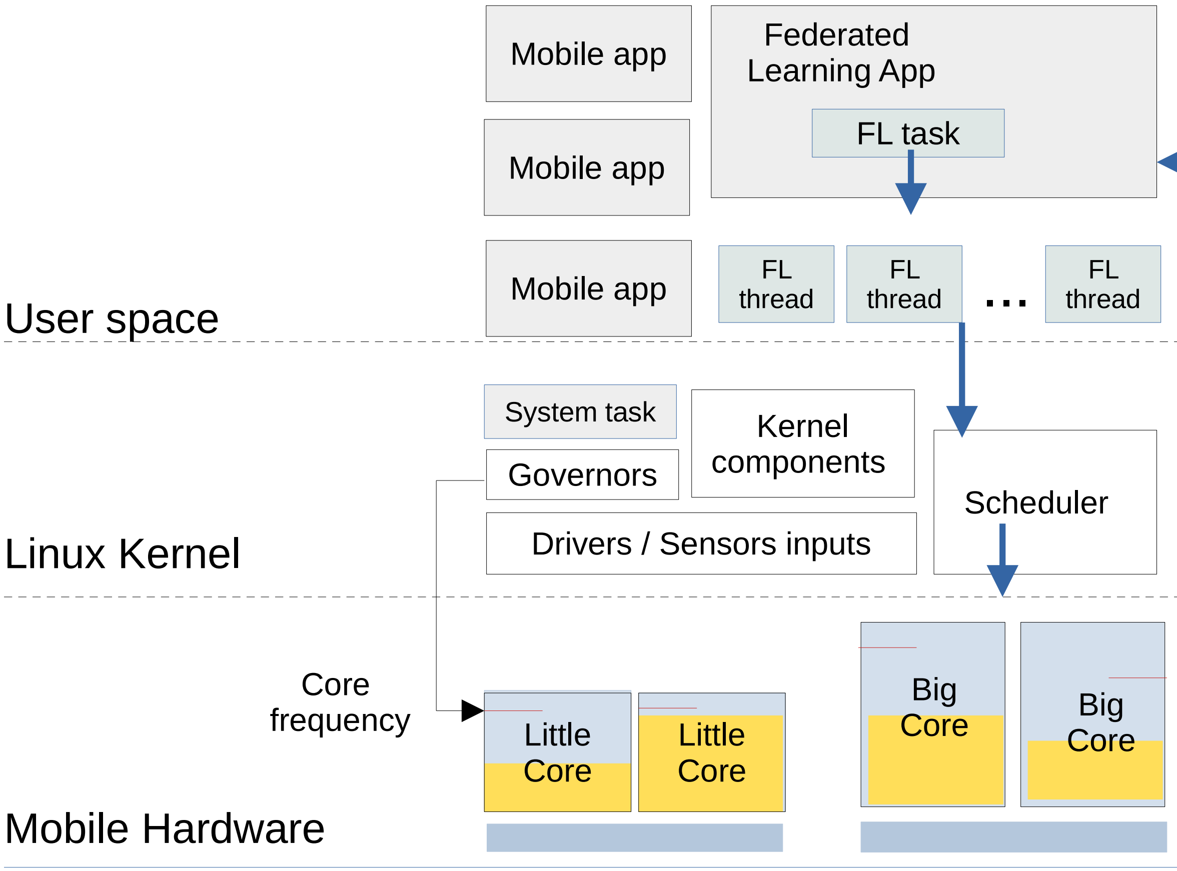
2. Let state the problem: general scheme



A new federated learning application

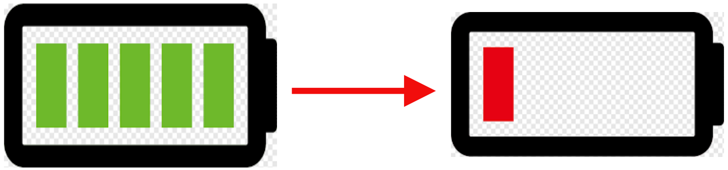
A simplified view of the Android system

2. Let state the problem: general scheme



A new federated learning application

A simplified view of the Android system



3.a. Let us define the **metric** to optimize

- The metric should reflect both:
  - **Computing power** of the FL task execution
  - **Electrical power absorption** of the phone.
- To compute this metric we have:
  - The **workload** of the FL task: number of CPU operations.
  - The **energy** consumed by the system: obtained by measurements.
- Metric adopted for the project: **energy efficiency**

$$energy_{eff} = \frac{Energy\,consumed}{workload\,computed} = \frac{Power\,absorbed}{Computing\,power}$$

3.b. What influences the energy efficiency

- The type of cores executing the task
  - Intuitively Big cores consumed high amount of Energy
  - Some research experiments prove that it can be a factor. [1]
- The task already present of the cores.
  - Energy discounted approach [2].
- The core frequency.

[1] Full-System Simulation of big.LITTLE Multicore Architecture for Performance and Energy Exploration. *Anastasiia Butko et al*
[2] Energy Discounted Computing on Multicore Smartphones, *Meng Zhu Kai Shen University of Rochester*
[3] Machine Learning-Based Approaches for Energy-Efficiency Prediction and Scheduling in Composite Cores Architectures
Hossein Sayadi et al.
[4] Temperature-Aware Scheduler Based on Thermal Behavior Grouping in Multicore Systems Inchoon Yeo and Eun Jung Kim

3.b. What influences the energy efficiency

- The type of cores executing the task
  - Intuitively Big cores consumed high amount of Energy
  - Some research experiments prove that it can be a factor. [1]
- The task already present of the cores.
  - Energy discounted approach [2].
- The core frequency.

We wanted to limit ourselves to the previous parameters, but it was with no real basis.

[1] Full-System Simulation of big.LITTLE Multicore Architecture for Performance and Energy Exploration. *Anastasiia Butko et al*
[2] Energy Discounted Computing on Multicore Smartphones, *Meng Zhu Kai Shen University of Rochester*
[3] Machine Learning-Based Approaches for Energy-Efficiency Prediction and Scheduling in Composite Cores Architectures Hossein Sayadi et al.
[4] Temperature-Aware Scheduler Based on Thermal Behavior Grouping in Multicore Systems Inchoon Yeo and Eun Jung Kim

3.b. What influences the energy efficiency

- The type of cores executing the task
  - Intuitively Big cores consumed high amount of Energy
  - Some research experiments prove that it can be a factor. [1]
- The task already present of the cores.
  - Energy discounted approach [2].
- The core frequency.
- The Number of threads of the best effort task [3].
- Core temperature [4].

We wanted to limit ourselves to the previous parameters, but it was with no real basis.

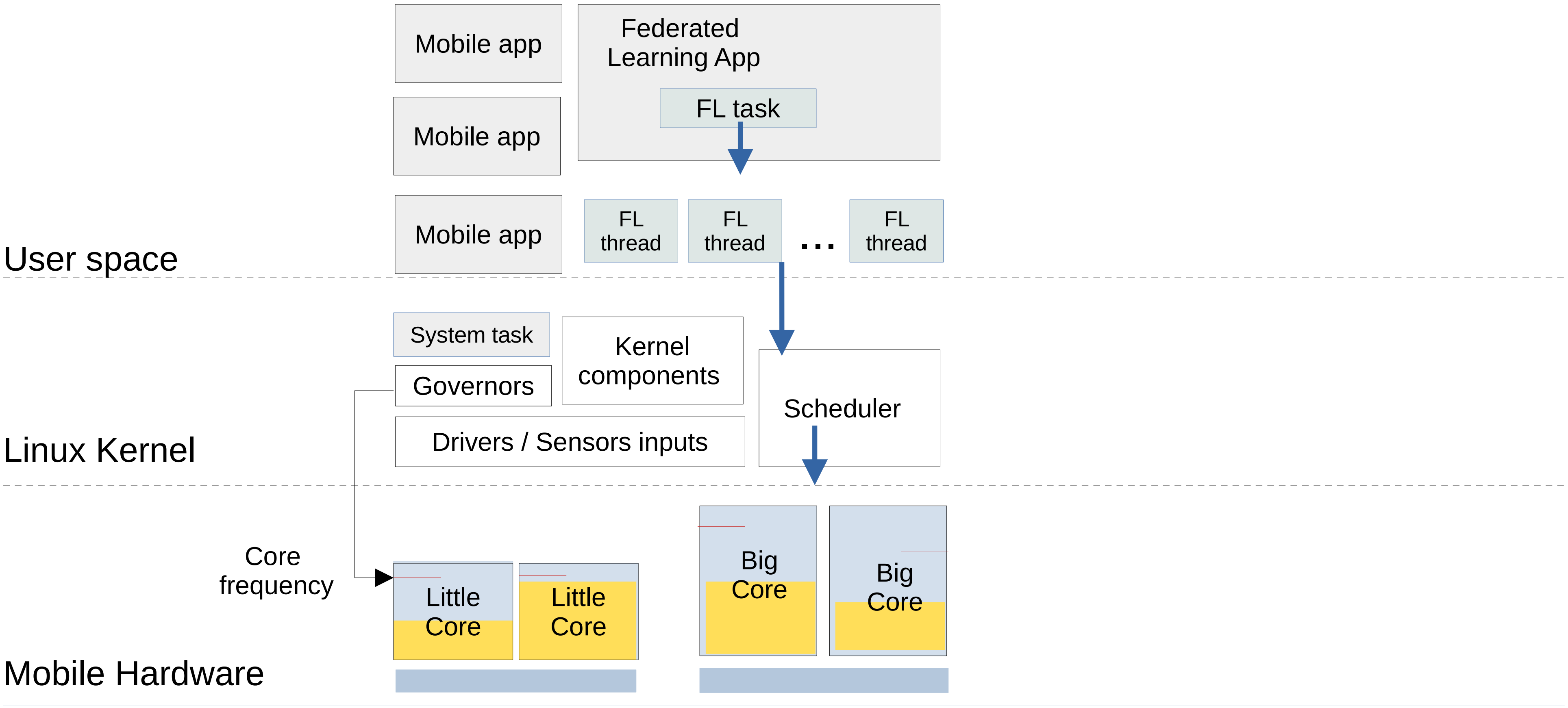
[1] Full-System Simulation of big.LITTLE Multicore Architecture for Performance and Energy Exploration. *Anastasiia Butko et al*

[2] Energy Discounted Computing on Multicore Smartphones, *Meng Zhu Kai Shen University of Rochester*

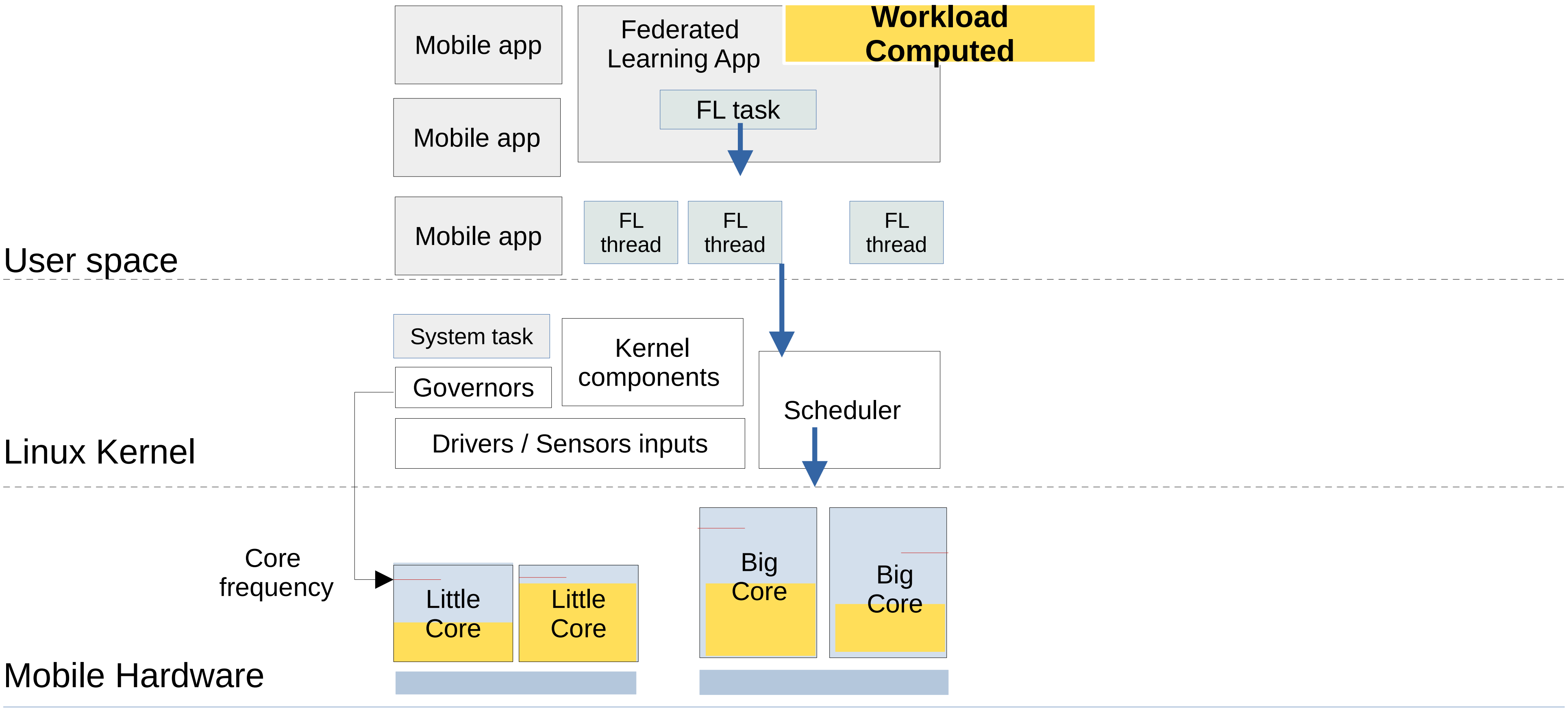
[3] Machine Learning-Based Approaches for Energy-Efficiency Prediction and Scheduling in Composite Cores Architectures Hossein Sayadi et al.

[4] Temperature-Aware Scheduler Based on Thermal Behavior Grouping in Multicore Systems Inchoon Yeo and Eun Jung Kim

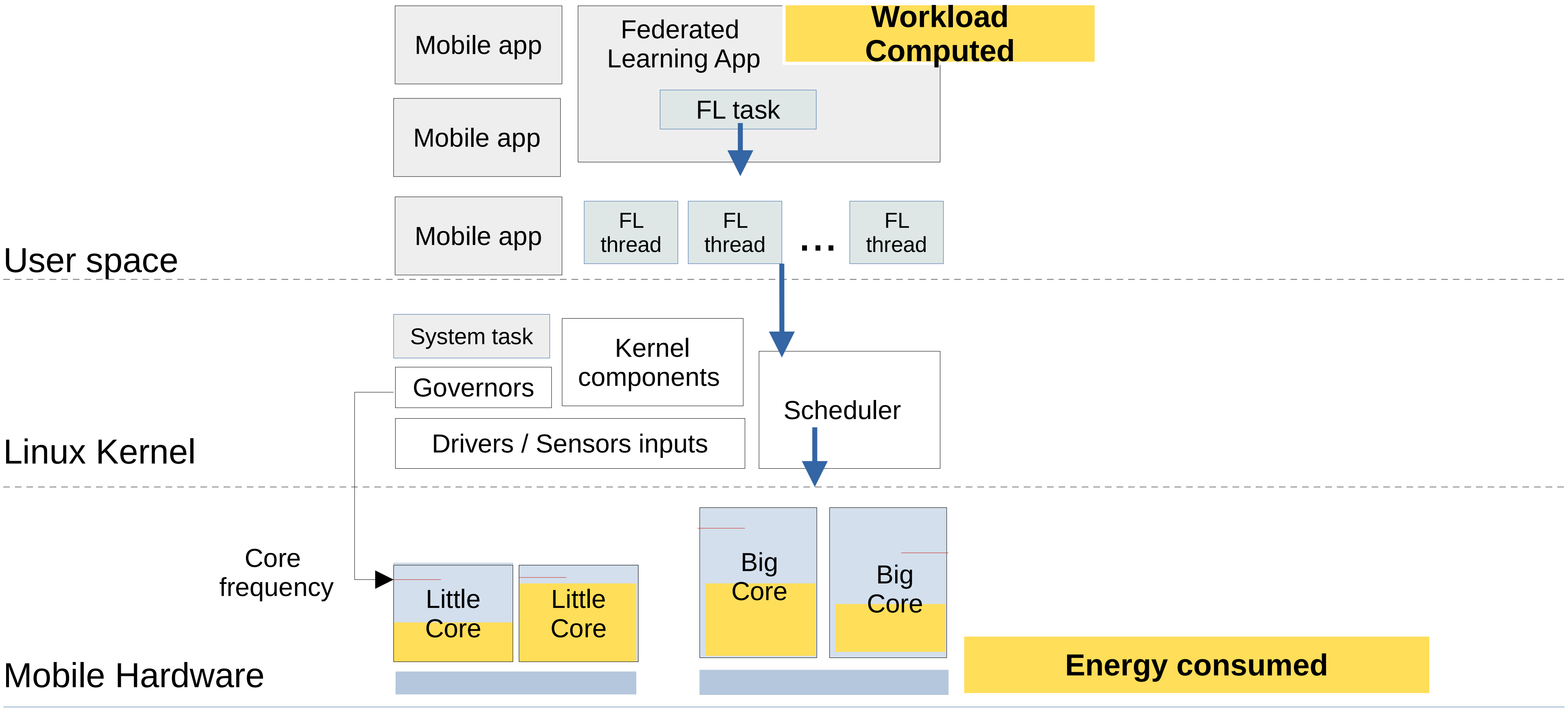
2. Let state the problem: general scheme



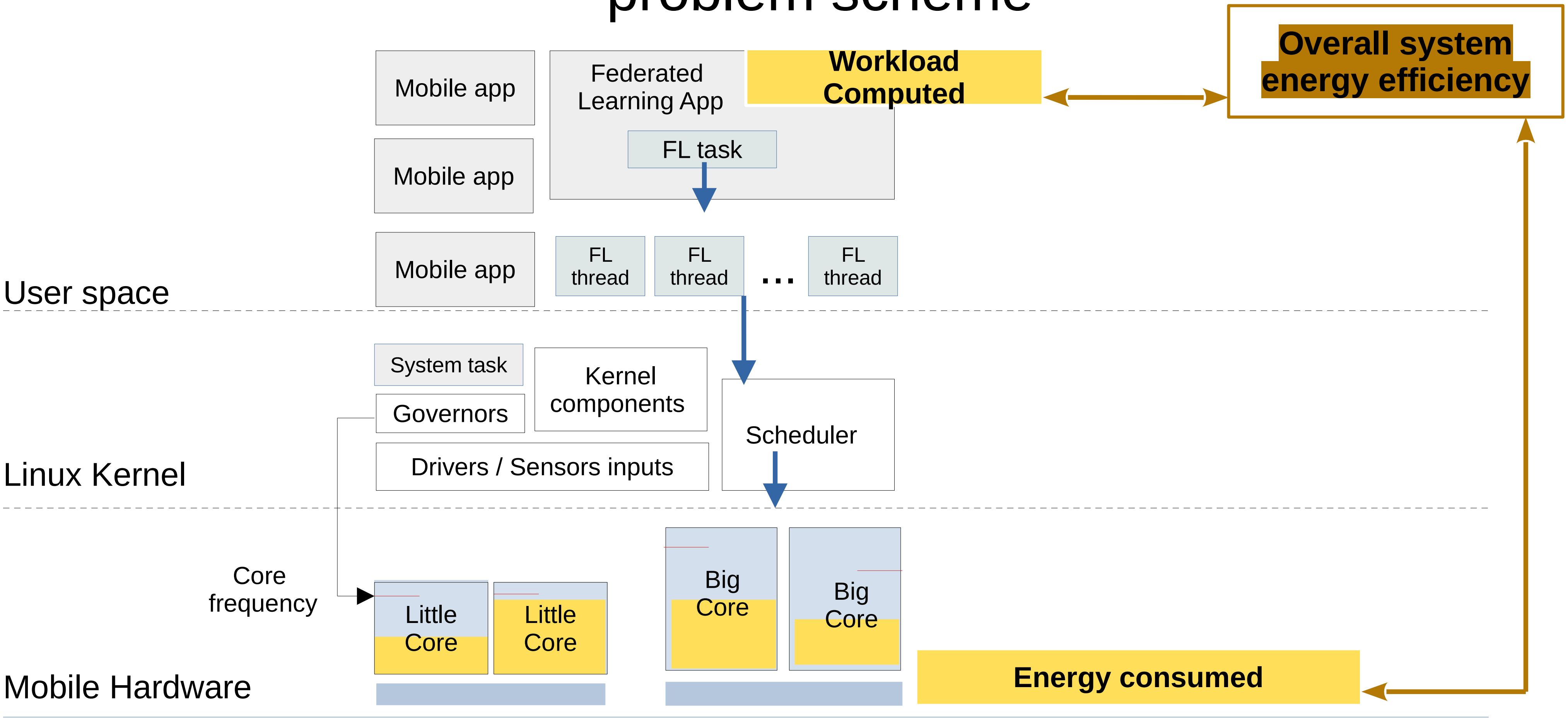
3.a. Let us define the **metric** to optimize on our problem scheme



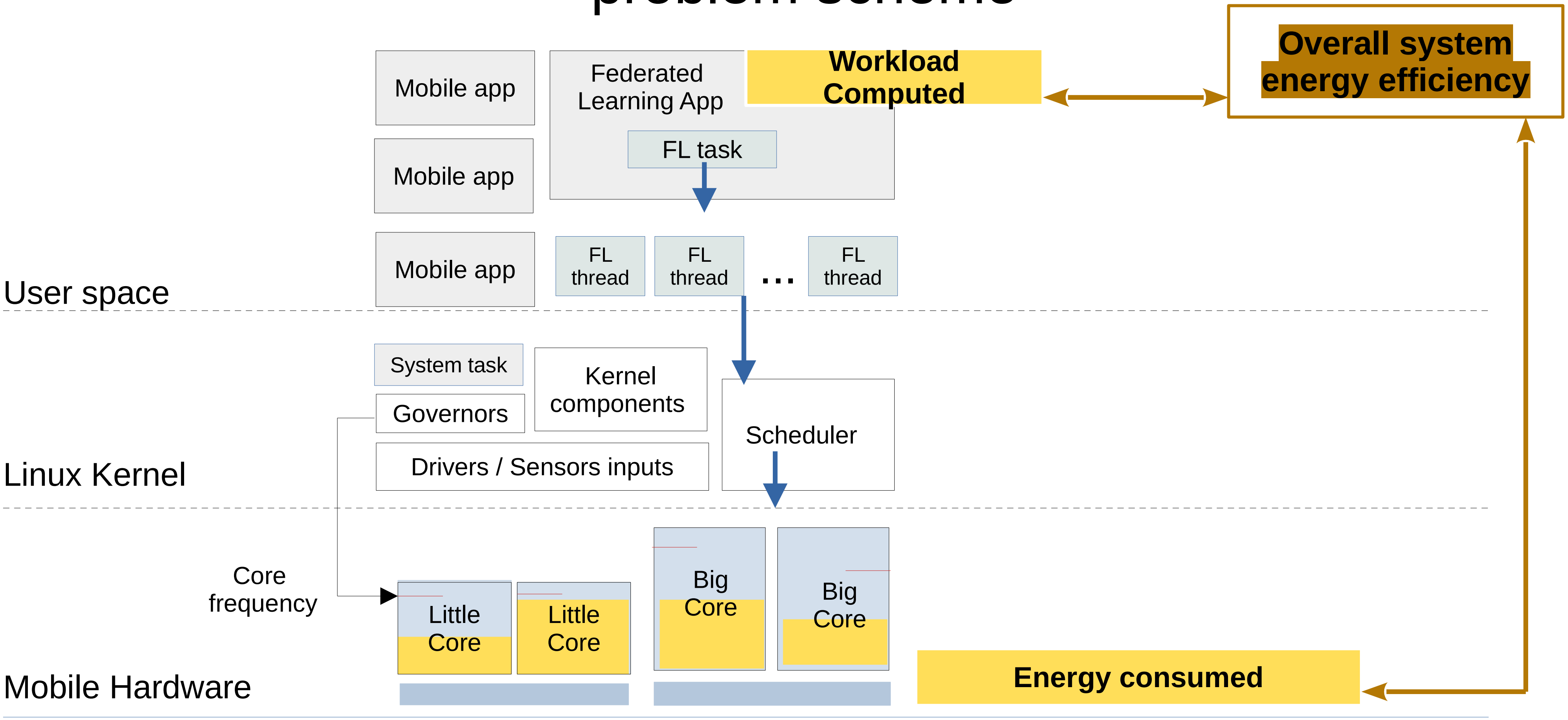
3.a. Let us define the **metric** to optimize on our problem scheme



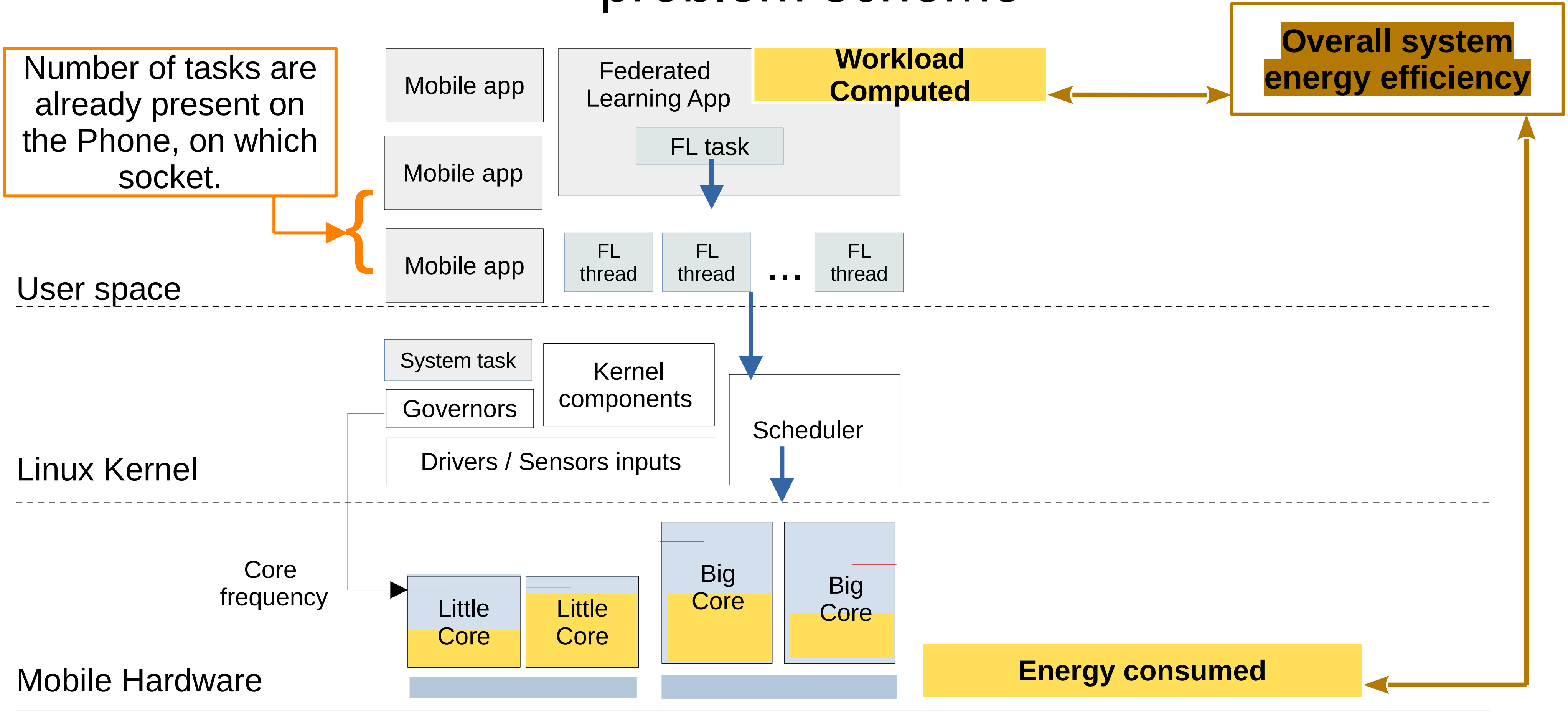
3.a. Let us define the **metric** to optimize on our problem scheme



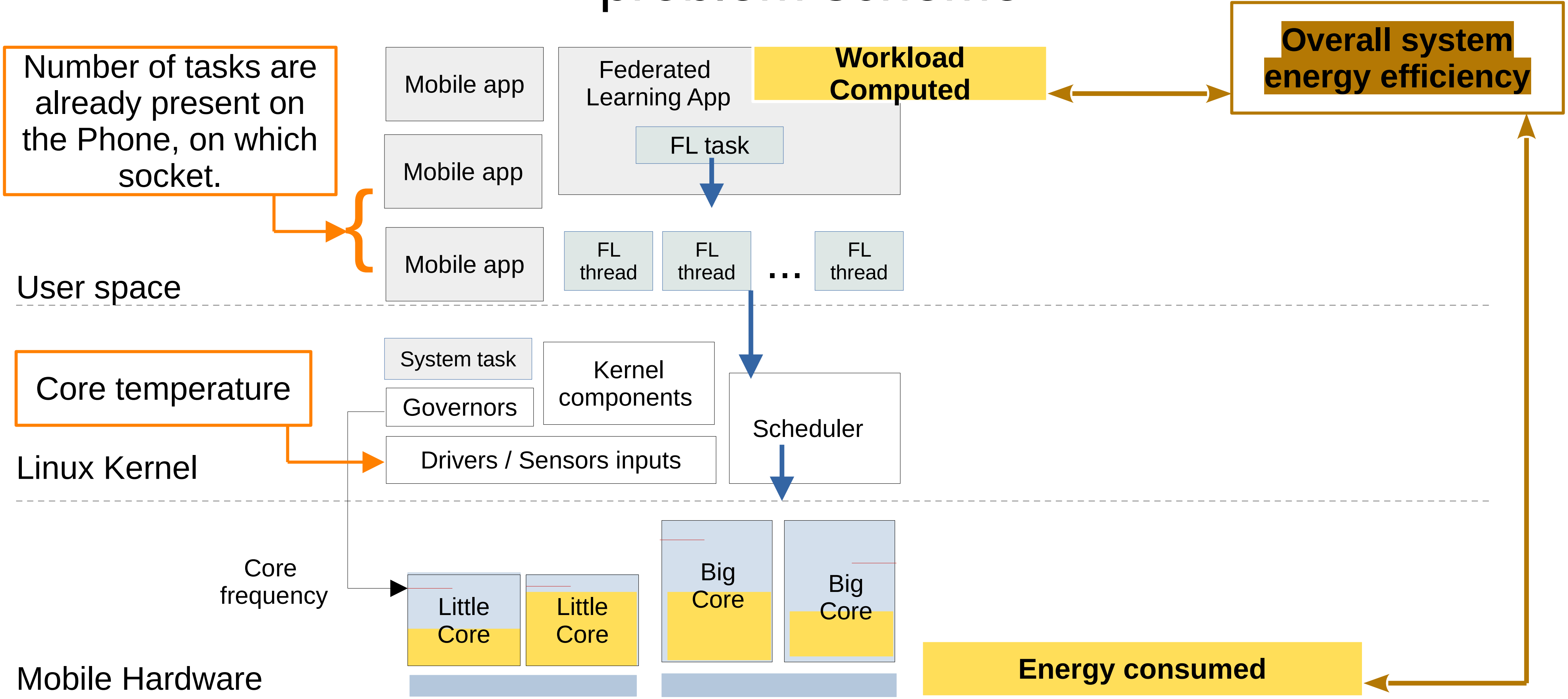
3.b. What **influences** the energy efficiency on our problem scheme



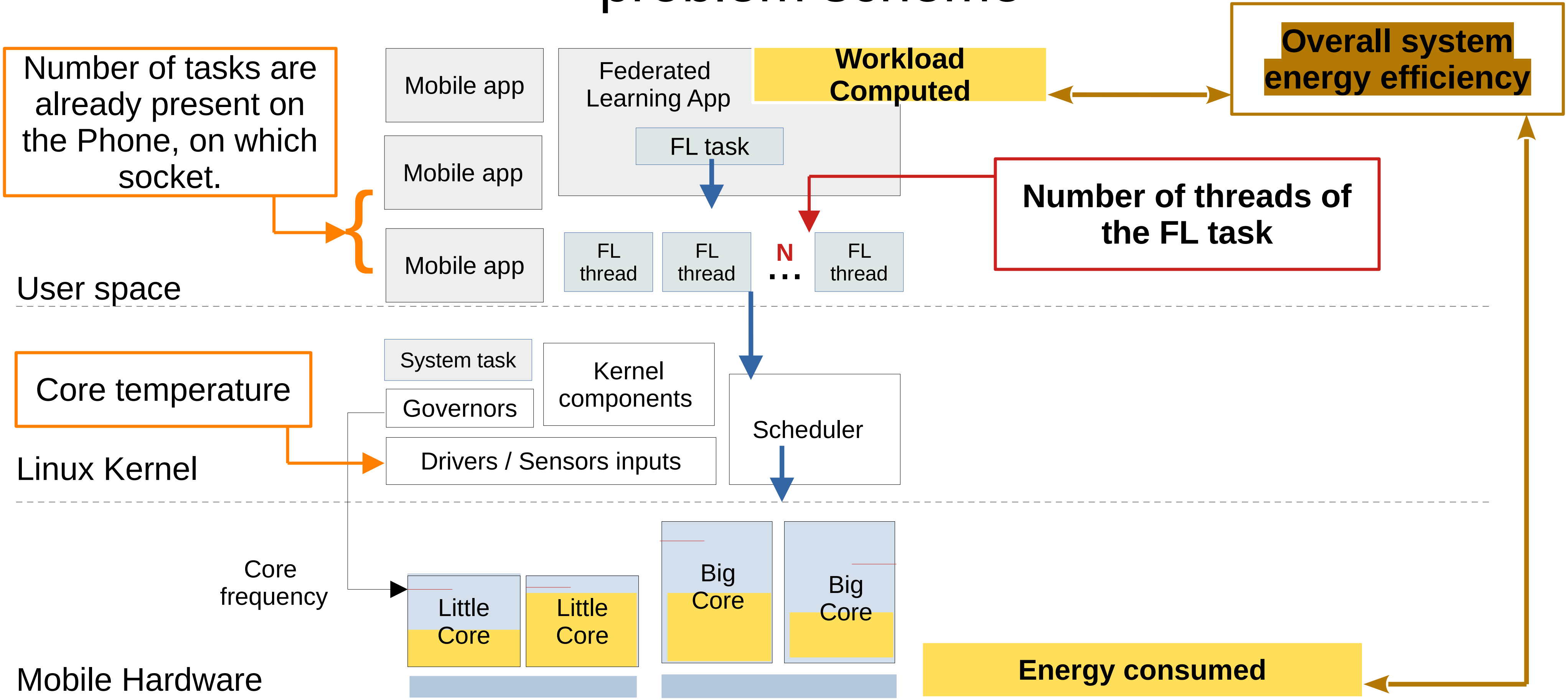
3.b. What **influences** the energy efficiency on our problem scheme



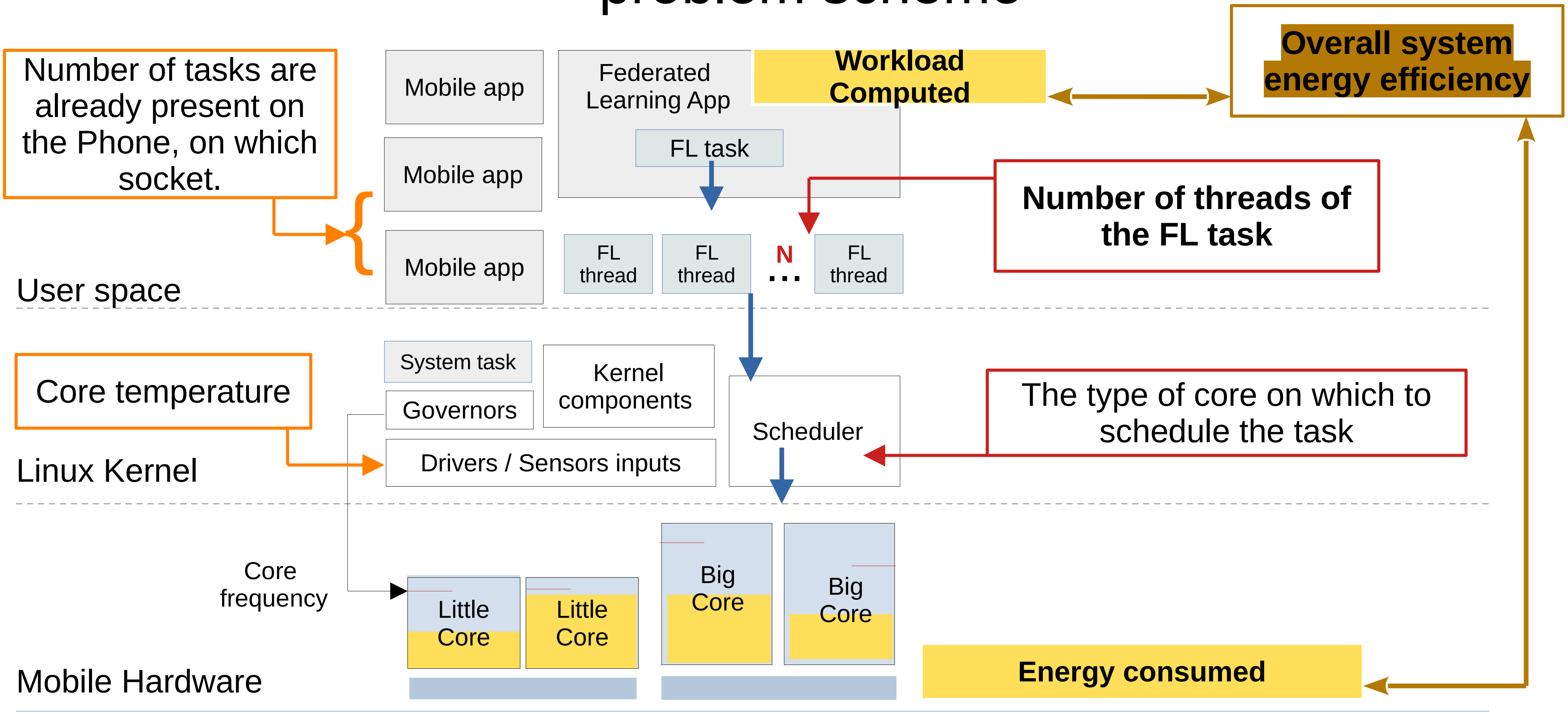
3.b. What **influences** the energy efficiency on our problem scheme



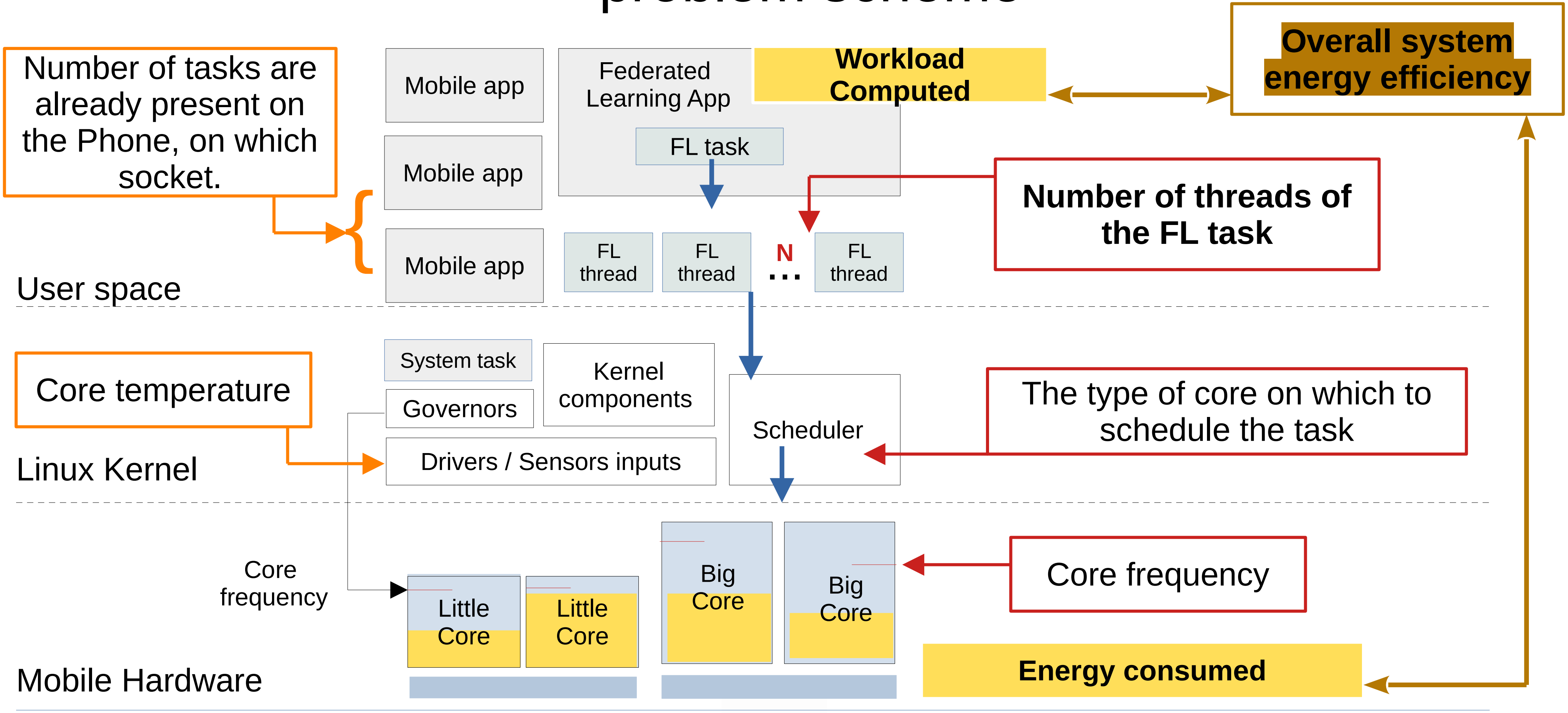
3.b. What **influences** the energy efficiency on our problem scheme



3.b. What **influences** the energy efficiency on our problem scheme



3.b. What **influences** the energy efficiency on our problem scheme



Number of tasks are already present on the Phone, on which socket.

Mobile app

Mobile app

Mobile app

Federated Learning App

FL task

Workload Computed

Overall system energy efficiency

Number of threads of the FL task

FL thread

FL thread

N

...

FL thread

User space

Core temperature

System task

Governors

Kernel components

Drivers / Sensors inputs

Scheduler

The type of core on which to schedule the task

Linux Kernel

Core frequency

Little Core

Little Core

Big Core

Big Core

Core frequency

Energy consumed

Mobile Hardware

3.d Approach to resolve the problem

I. Make in-lab experiments by varying scenarios parameters:

- Number of interactive task present on phones
- Number of threads of the FL task
- Type of cores
- Core frequencies
- Temperature

II. Bringing out the lessons learned about HOW those parameters influence energy efficiency.

III. Apply these lessons learned in the FL task scheduling decision:

- At user space Level
- At kernel Level (Scheduler, governor).

3.e Workload measurement

- Benchmarks (Newly added FL task)
  - **Prime number computation** (to quickly get an overview of cores energy efficiency) [1]
  - Tensor Flow Lite model on Mobile Device [2] (to have ML-like task behavior)
  - Federated Learning Tool from FLEET (for FL-like experiments) [3]
- Interactive apps (Other apps)
  - **Interactive app simulation** (with interruptions to quickly get an overview) [1]
  - Widely used mobile apps (YouTube, Instagram ...)

- Phone 1: Google Pixel 4A 5G:
  - 3 sockets: CPUs 0-5: 1.8048 GHz; CPU 6: 2.208 GHz;CPU7: 2.4 GHz
  - Memory: 6GB RAM
- Phone 2: Samsung galaxy S8
  - 2 sockets CPUs 0-3 : 1.69 GHz , CPUs 4-7: 2.314 GHz
  - Memory: 4GB RAM

[1] Prime number computation source code https://gitlab.liris.cnrs.fr/plwapet/benchmarking_app_to_test_big_cores/-/blob/main/app/src/main/java/com/opportunistask/scheduling/benchmarking_app_to_test_big_cores/PrimeNumberThread.java

[2] On-Device Training with TensorFlow Lite https://www.tensorflow.org/lite/examples/on_device_training/overview

[3] FLeet: Online Federated Learning via Staleness Awareness and Performance Prediction, Georgios Damaskinos, Rachid Guerraoui, Vlad Nitu et al. Source code https://github.com/gdamaskinos/fleet/

3.f Energy consumption measurement:

system API “*dumpsys batterystats*” from Android OS

- Widely used in research [1]
- We have used it for more than 7 months.
- Confirms the influence of the above-mentioned parameters on the energy efficiency
- But some results incompatible with reality

Power-meter tool

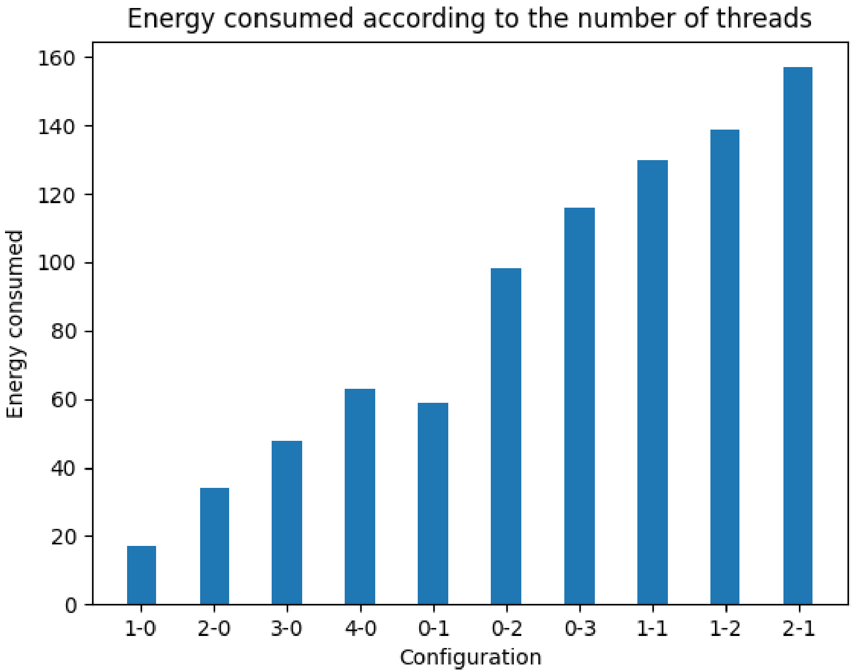
- Also widely used in research [2][3]
- The common installation required is **expensive**
  - Its makes phone battery no longer usable.
- **Alternative 1:** Software simulation of battery shutdown (Google Pixel 4A, 5G).
  - Modifying internal system file : *“charge_stop_level”*, *“charge_limit”*
  - USB mode power supply
  - Retrieving data from power-meter
- **Alternative 2 :** Full battery charging (Samsung)
  - Retriving data from system file “cc_info”
  - Retrieving data form power-meter

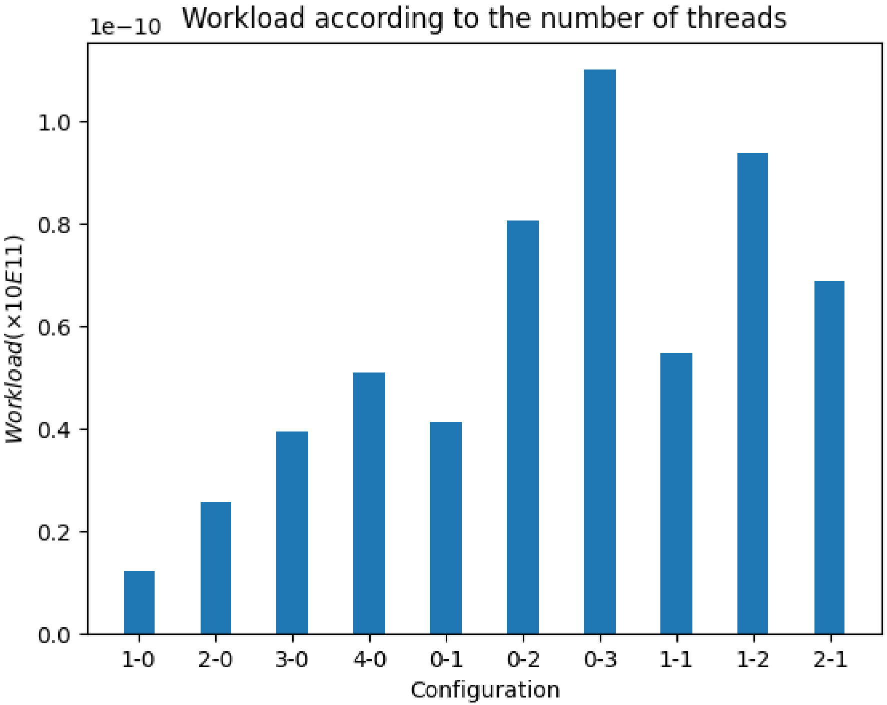
[1] Resource utilization and per formance, A comparative study on mobile crossplatform tools, Lucas Arvidsson, Max Bekkhus
[2]"Energy Consumption and Conservation in WiFi Based Phones: A Measurement-Based Study By Ashima Gupta and Prasant Mohapatra"
[3] Energy-Efficient Collaborative Sensing with Mobile Phones Xiang Sheng

4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Type or Core**
Experiments duration: 10 min
Legend: Configuration 0-1 means

- 0 thread on Little sockets
- 1 Thread on Big Socket



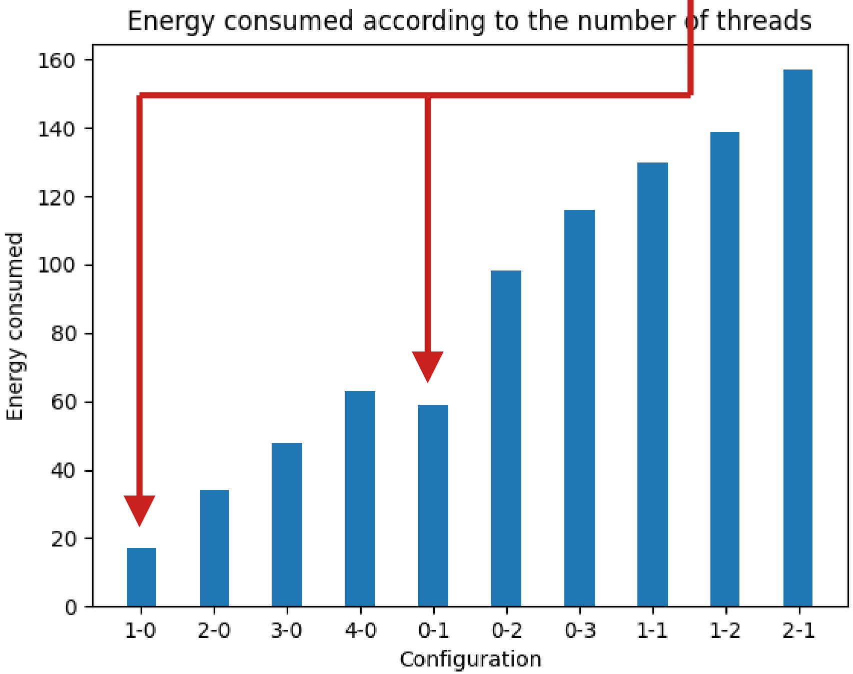


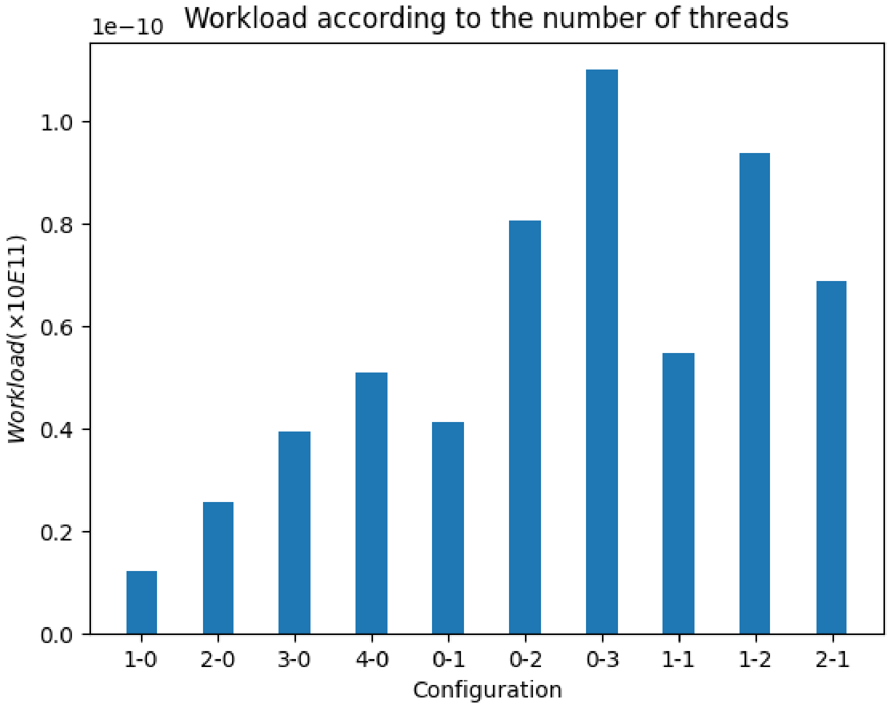
4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Type or Core**
Experiments duration: 10 min
Legend: Configuration 0-1 means

- 0 thread on Little sockets
- 1 Thread on Big Socket

Big Cores consume a lot of energy compared to little cores





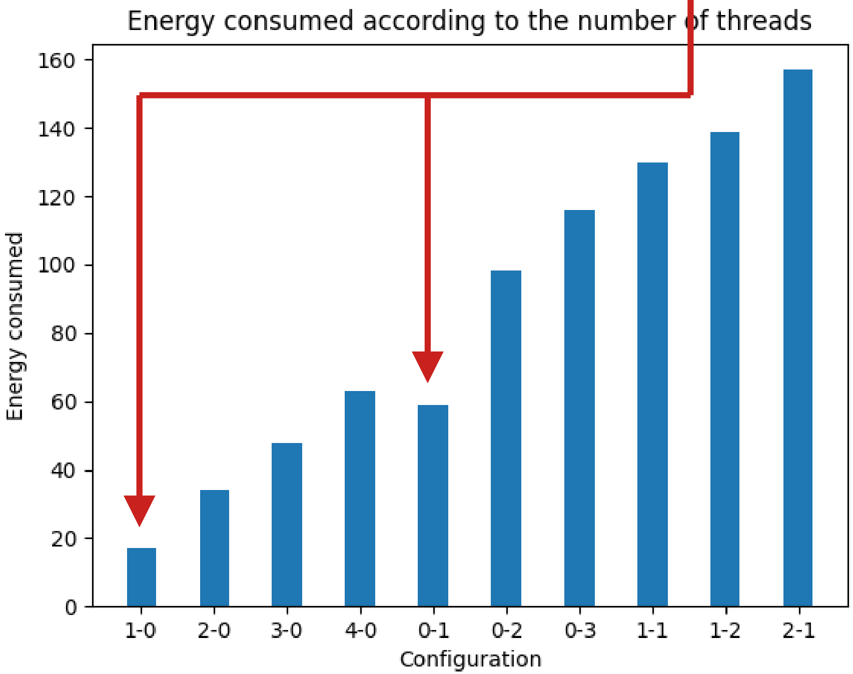
4. Experiments and observations (made using APIs)

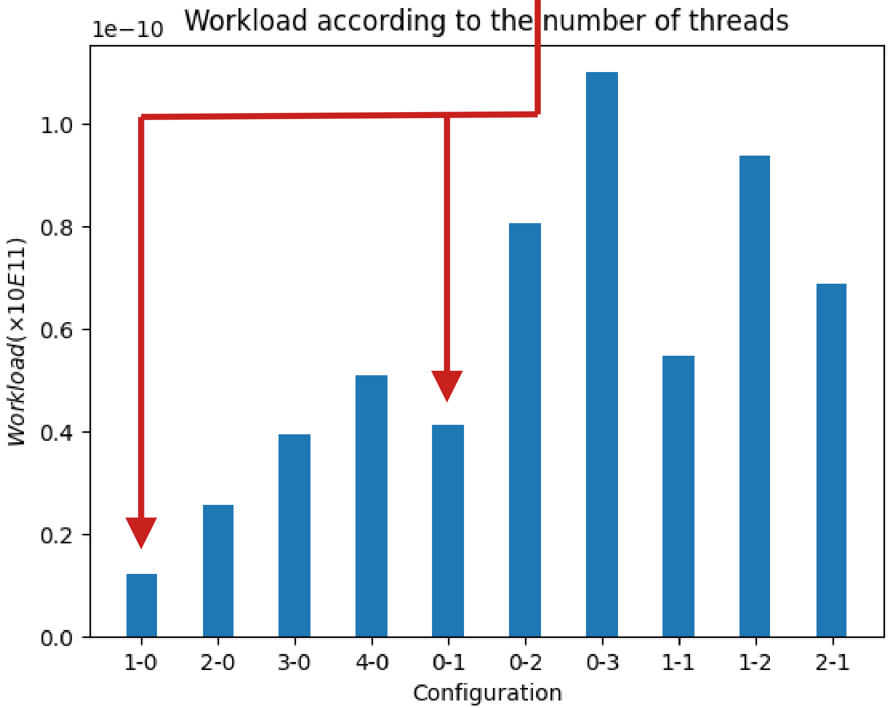
Phone: Samsung S8
Impact of: **Type or Core**
Experiments duration: 10 min
Legend: Configuration 0-1 means

- 0 thread on Little sockets
- 1 Thread on Big Socket

Big Cores consume a lot of energy compared to little cores

Big cores are very fast in computation

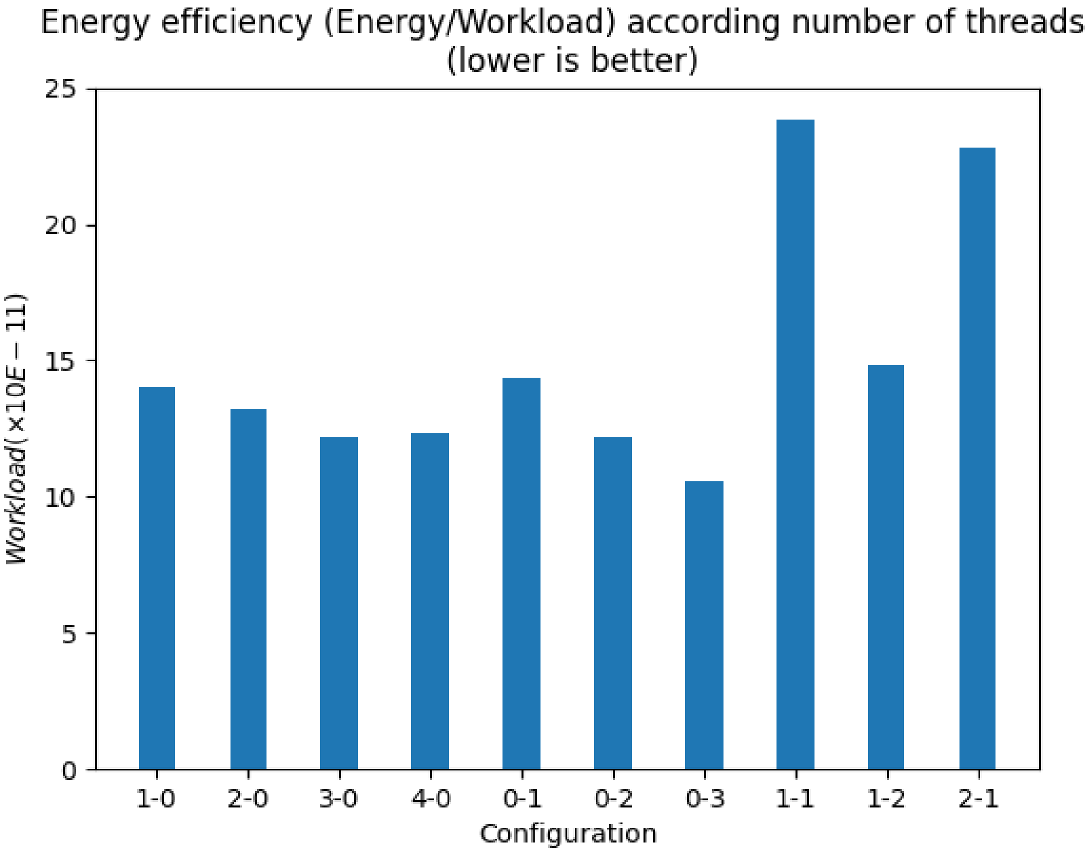




4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Number of Threads**
Experiments duration: 10 min
Legend: Configuration 0-1 means

- 0 thread on Little sockets
- 1 Thread on Big Socket

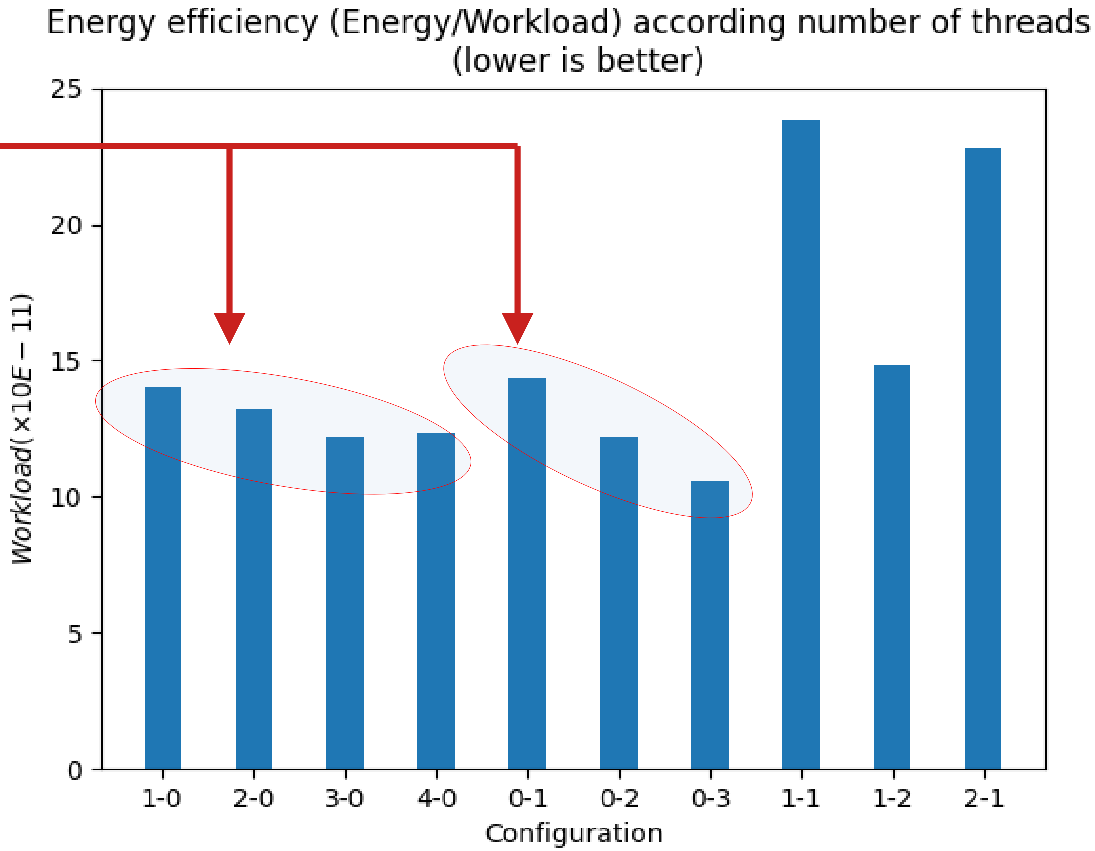


4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Number of Threads**
Experiments duration: 10 min
Legend: Configuration 0-1 means

- 0 thread on Little sockets
- 1 Thread on Big Socket

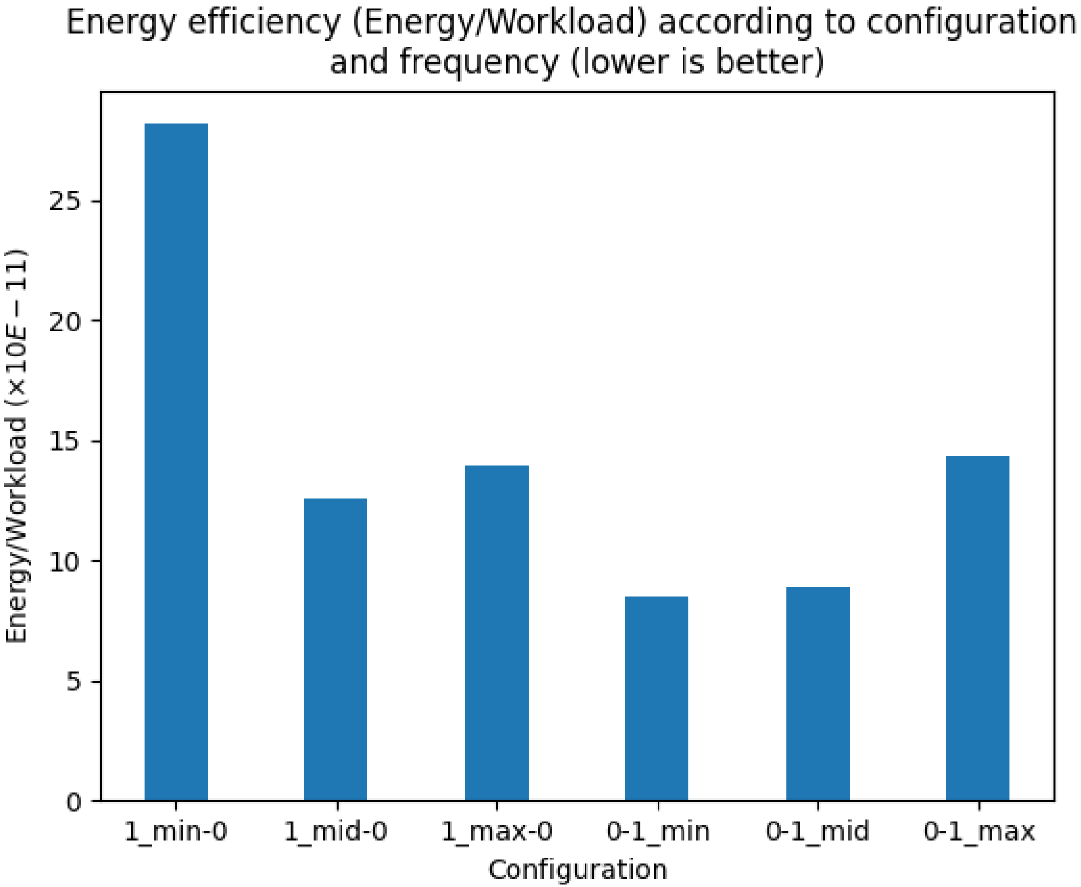
On the same socket the number of threads slightly increases with the efficiency



4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Frequency**
Experiments duration: 5 min
Legend: Configuration 0-1_mid means

- 0 thread on Little sockets
- 1 Thread on Big Socket
- Big socket runs with frequency at middle level.
- Mid = middle level, min = minimum level
  - Max = maximum frequency

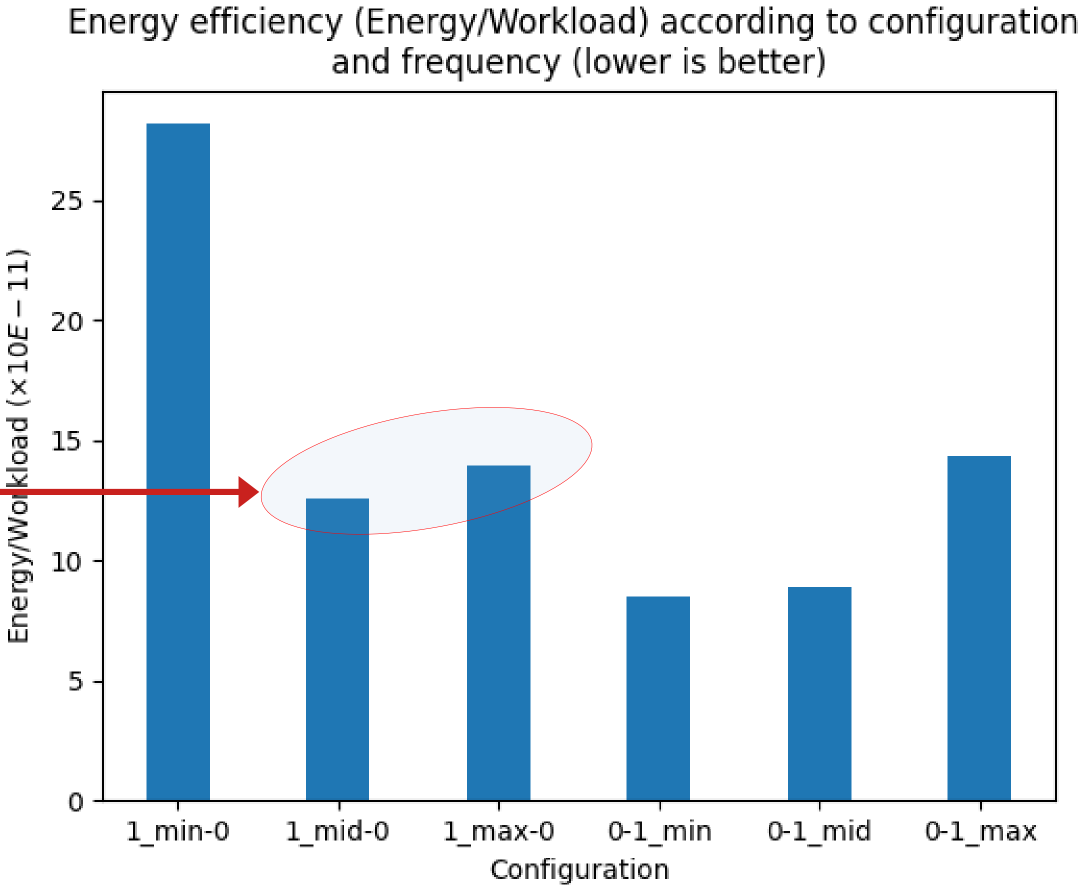


4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Frequency**
Experiments duration: 5 min
Legend: Configuration 0-1_mid means

- 0 thread on Little sockets
- 1 Thread on Big Socket
- Big socket runs with frequency at middle level.
- Mid = middle level, min = minimum level
  - Max = maximum frequency

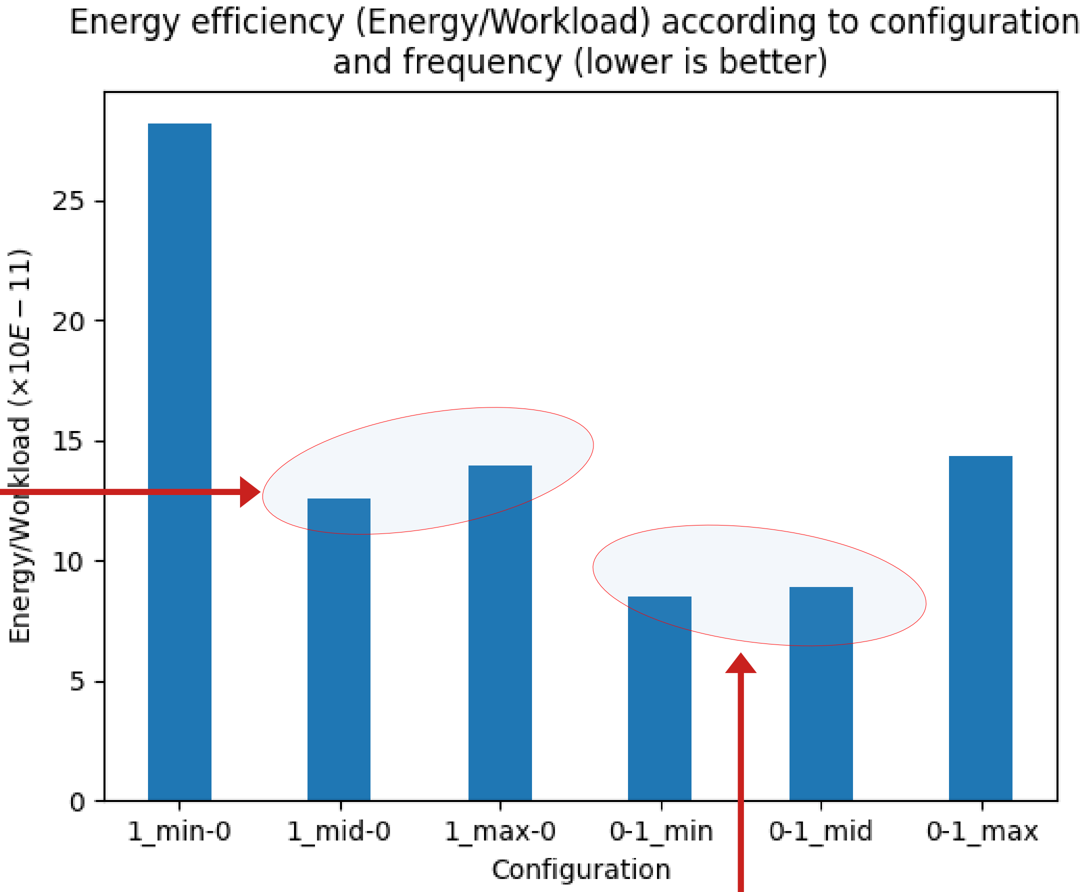
At slightly reduced frequency the Little cores are efficient



4. Experiments and observations (made using APIs)

Phone: Samsung S8
Impact of: **Frequency**
Experiments duration: 5 min
Legend: Configuration 0-1_mid means

- 0 thread on Little sockets
- 1 Thread on Big Socket
- Big socket runs with frequency at middle level.
- Mid = middle level, min = minimum level
  - Max = maximum frequency



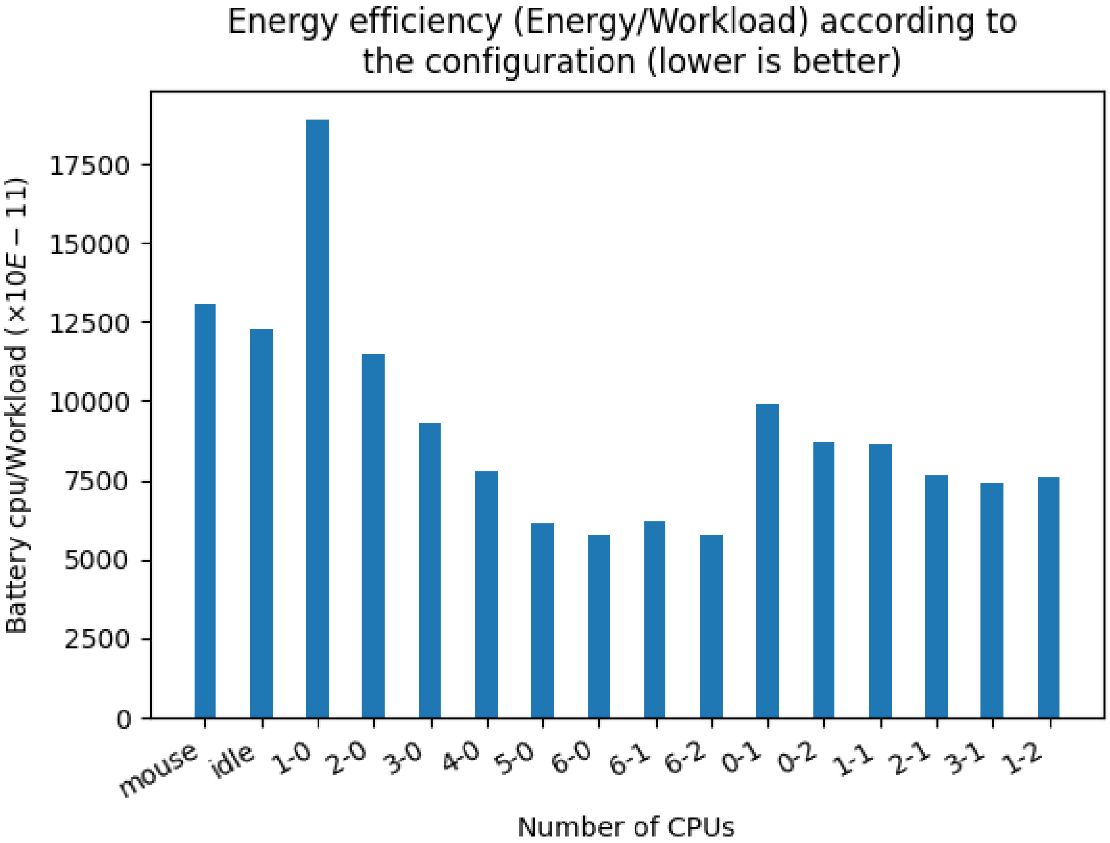
At slightly reduced frequency the Little cores are efficient

It is more efficient to reduced frequency on the Big cores as much as possible for one task.

4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Number of threads**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file
charge_stop_level
Legend: Configuration 0-1 means:

- 0 thread on Little core
- 1 thread on Big core

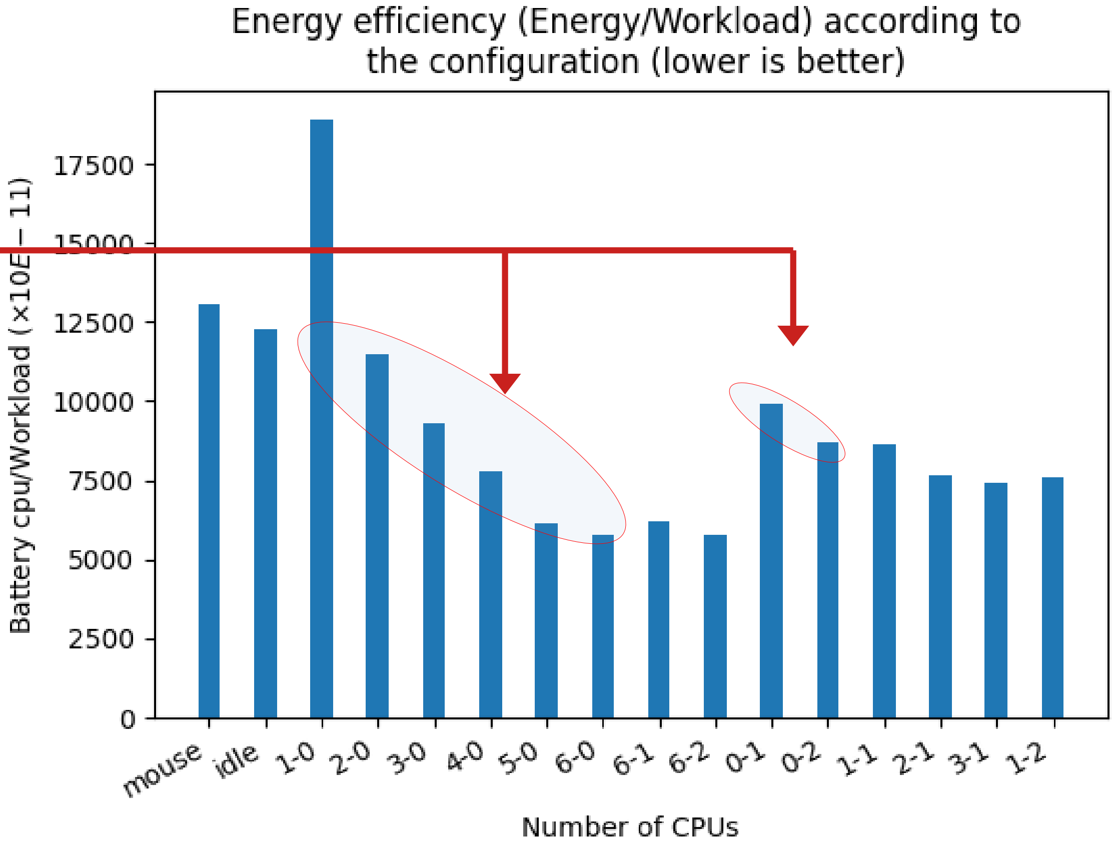

Energy efficiency (Energy/Workload) according to the configuration (lower is better)

4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Number of threads**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1 means:

- 0 thread on Little core
- 1 thread on Big core

On the same socket the number of threads slightly increases with the efficiency

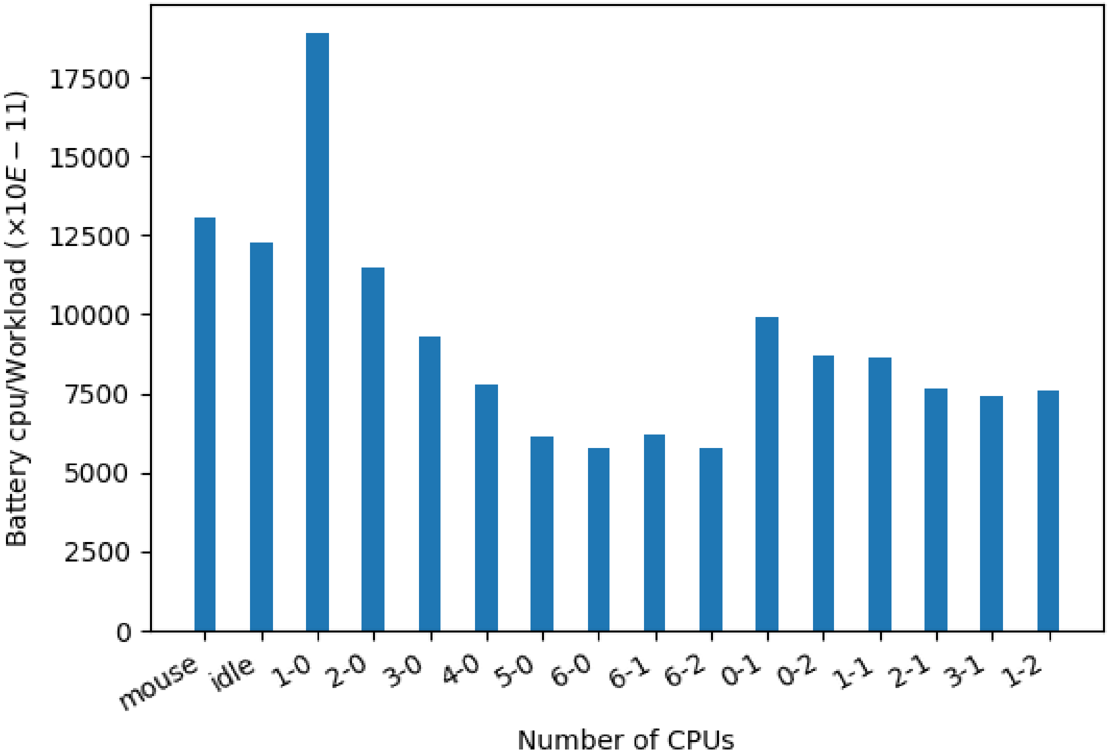


4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: Type of Cores
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1 means:

- 0 thread on Little core
- 1 thread on Big core

Energy efficiency (Energy/Workload) according to the configuration (lower is better)

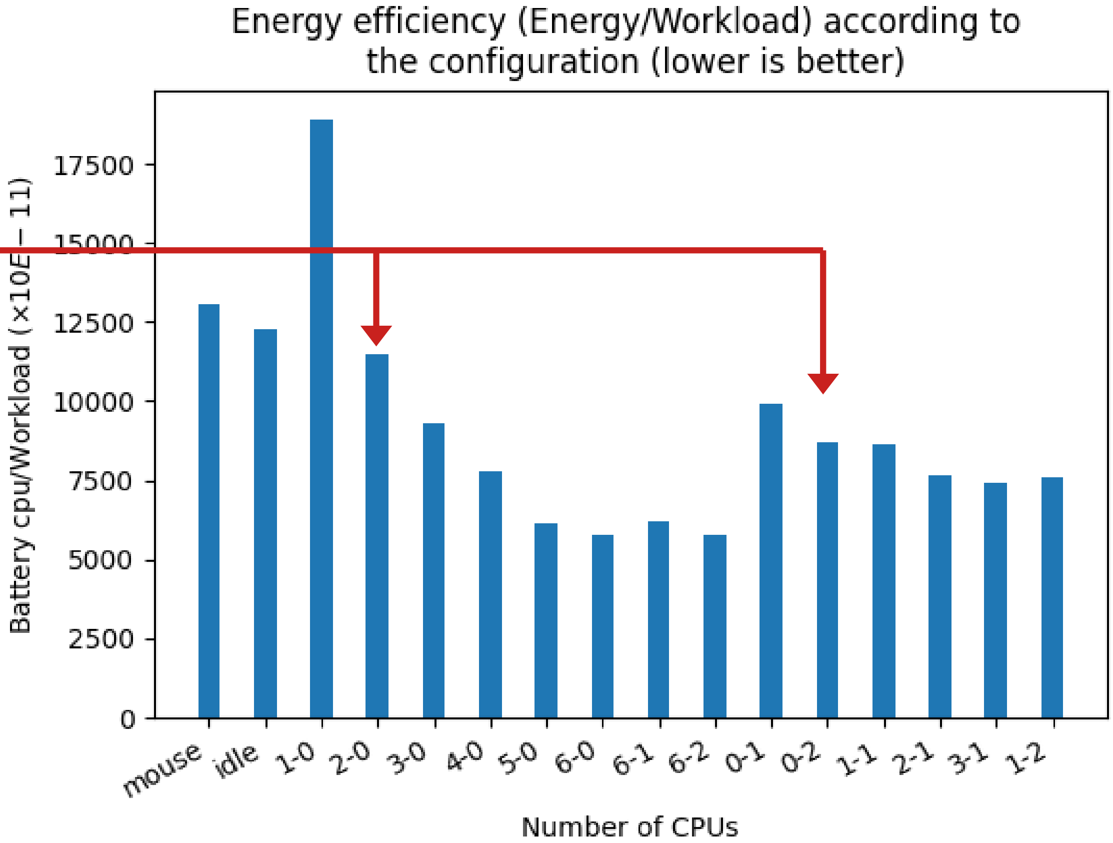


4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Type of Cores**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1 means:

- 0 thread on Little core
- 1 thread on Big core

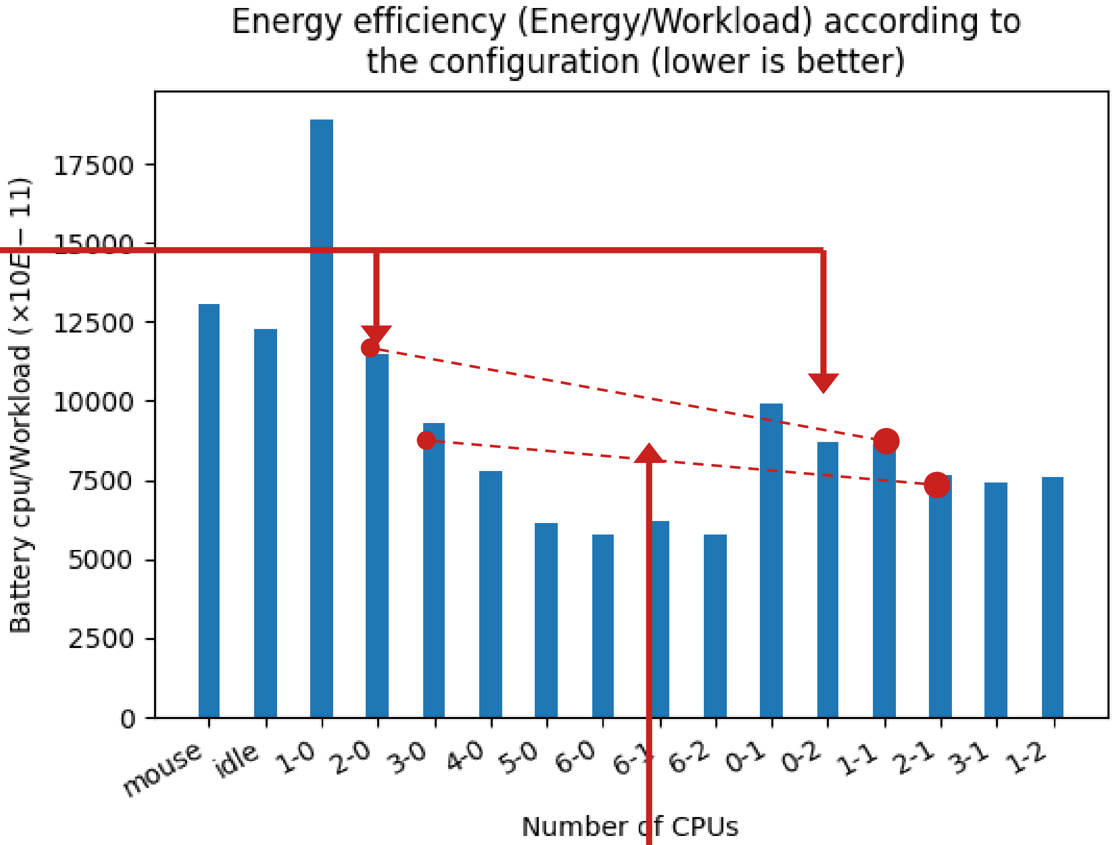
Big cores are much more efficient than little cores



4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Type of Cores**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1 means:

- 0 thread on Little core
- 1 thread on Big core



Big cores are much more efficient than little cores

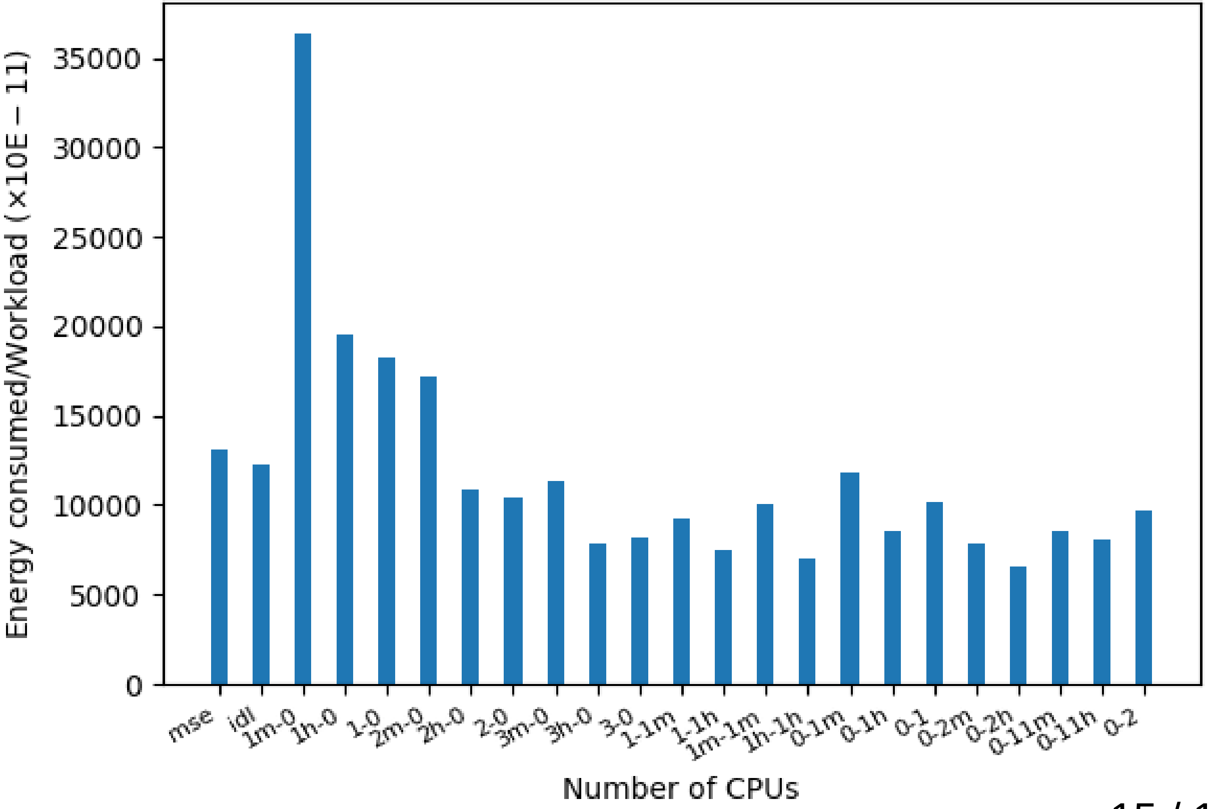
The efficiency of the big cores influences the overall efficiency of the configuration

4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Frequency**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1m means:

- 0 thread on Little core
- 1 thread on Big core
- The Big core has the min frequency
- H = half frequency, nothing = max frequency

Energy/ Workload according to the number of CPUs
m = idle (minimum) frequency, h = half frequence

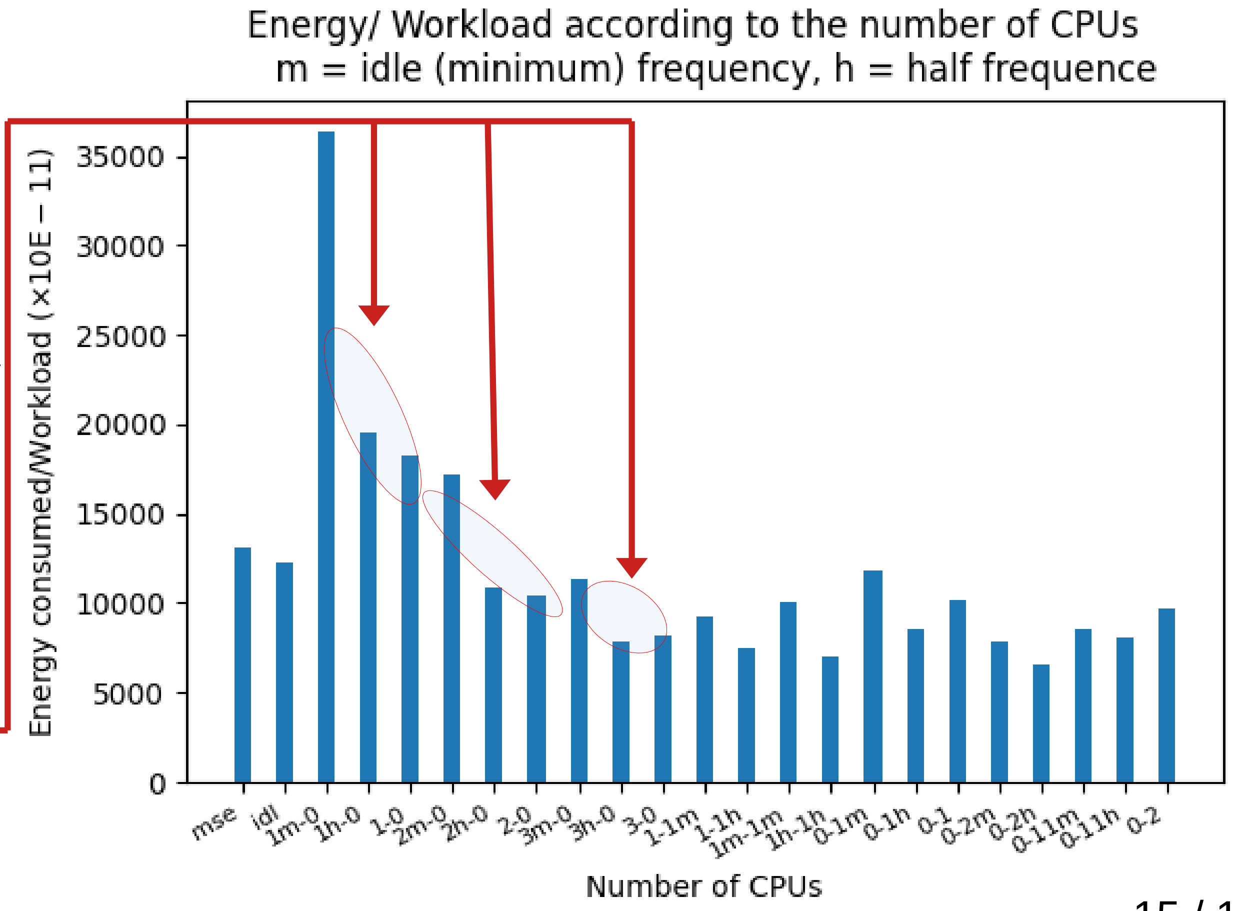


4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Frequency**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1m means:

- 0 thread on Little core
- 1 thread on Big core
- The Big core has the min frequency
- H = half frequency, nothing = max frequency

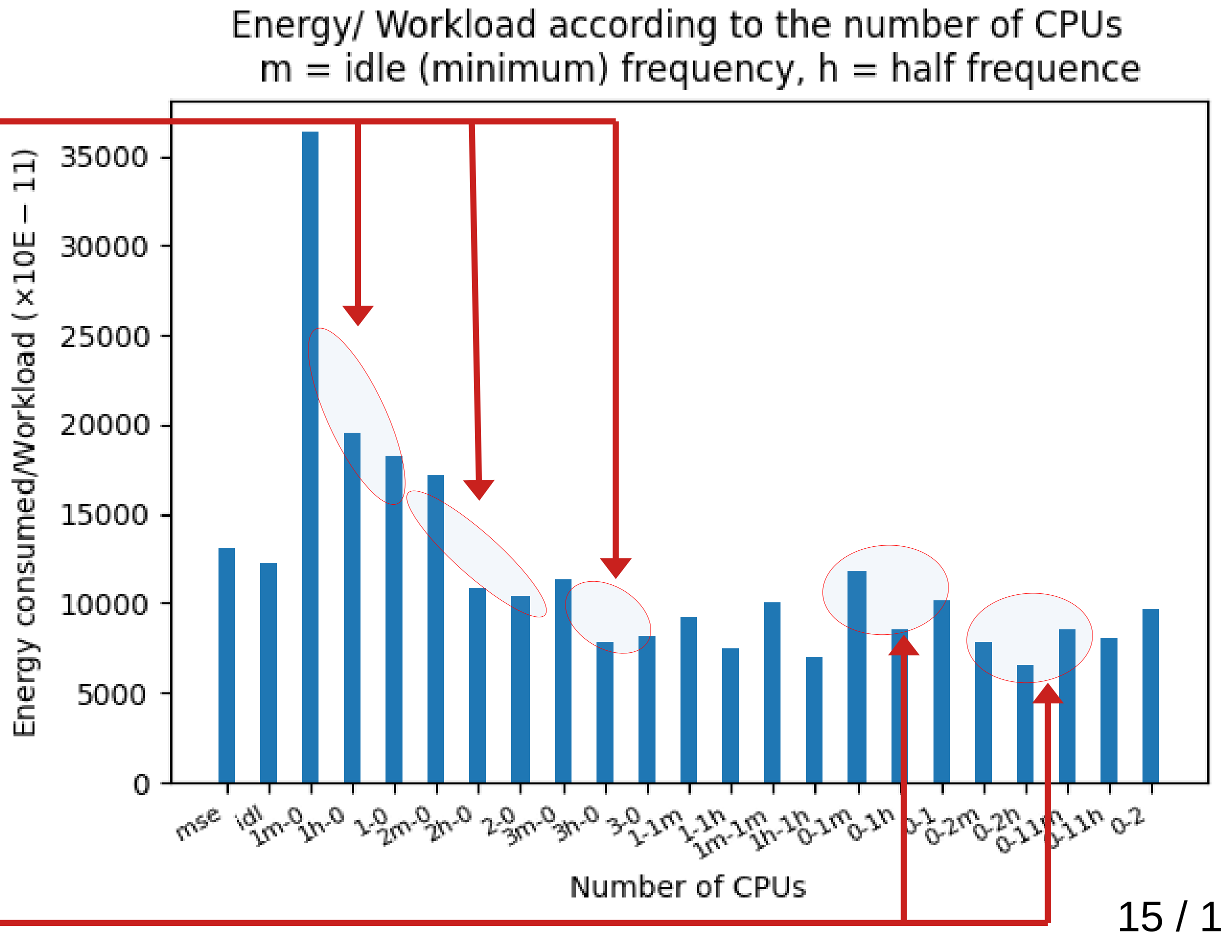
On the Little cores we are much more efficient with the maximum frequency



4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Frequency**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1m means:

- 0 thread on Little core
- 1 thread on Big core
- The Big core has the min frequency
- H = half frequency, nothing = max frequency



Energy/ Workload according to the number of CPUs
m = idle (minimum) frequency, h = half frequence

On the Little cores we are much more efficient with the maximum frequency

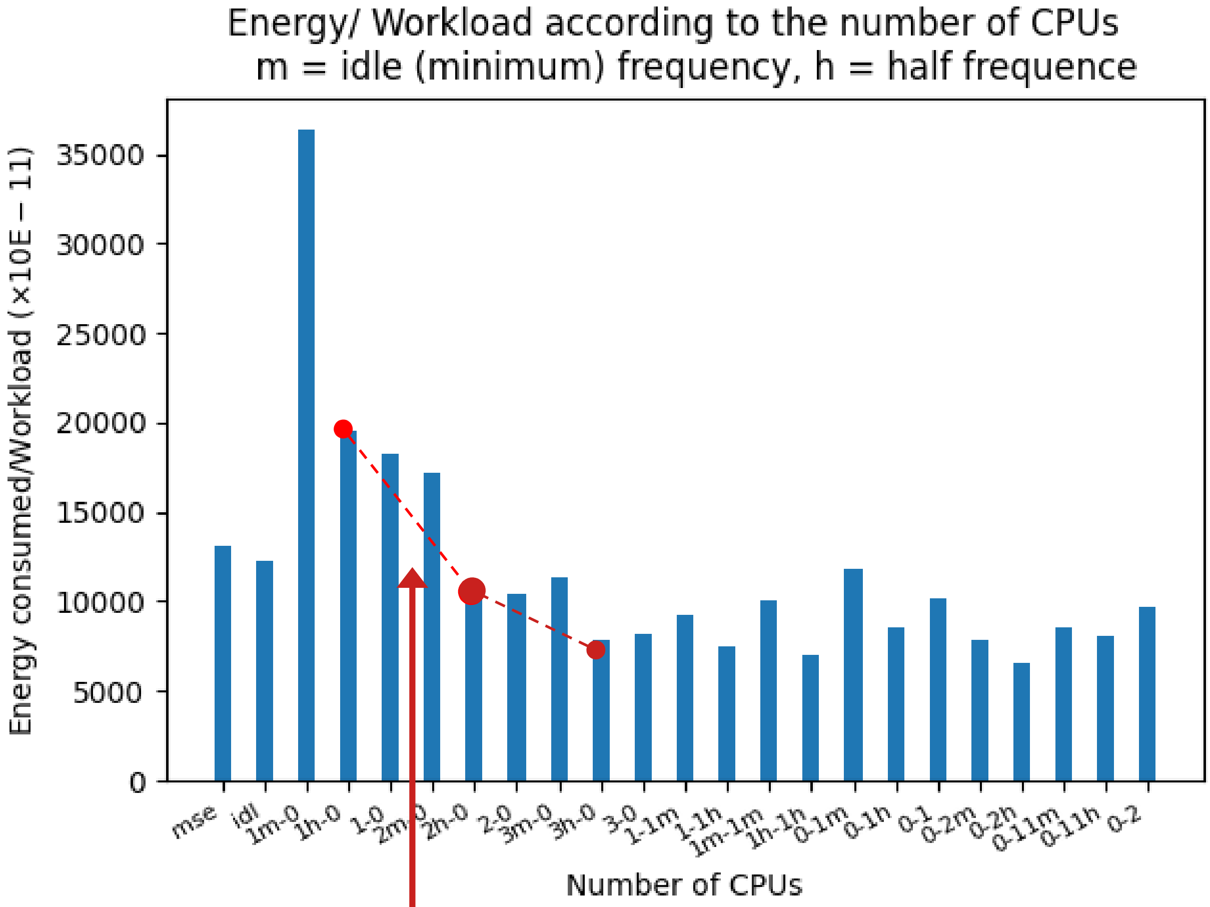
On the Big cores we are much more efficient with the mid frequency

4. Experiments and observations (made using the power-meter)

Phone: Google Pixel
Impact of: **Frequency and number of Threads**
Experiments duration: 10 min
Battery level: 50
No charging: Yes by the file *charge_stop_level*
Legend: Configuration 0-1m means:

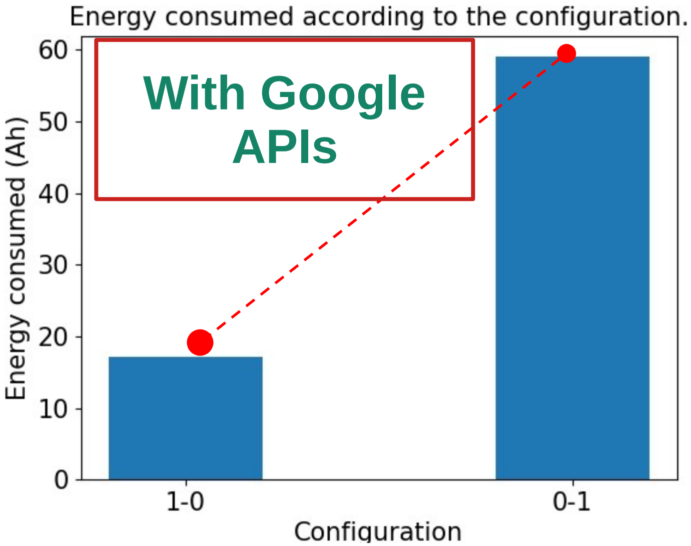
- 0 thread on Little core
- 1 thread on Big core
- The Big core has the min frequency
- H = half frequency, nothing = max frequency

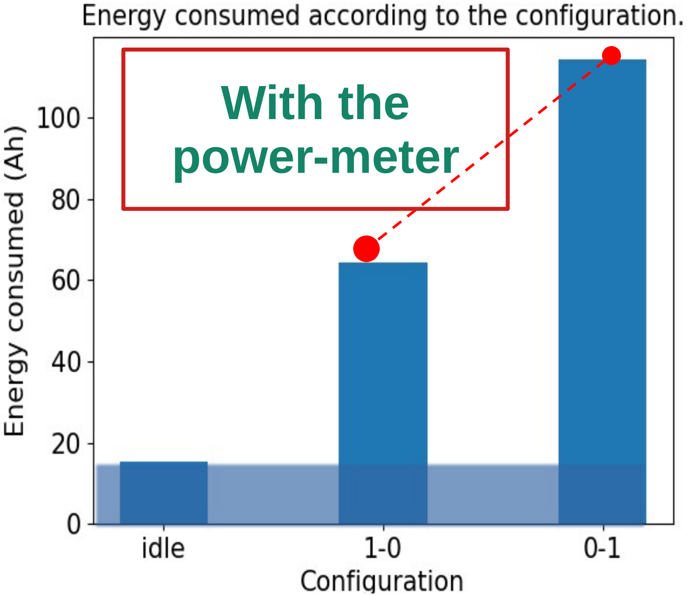
Fixing the frequency at mid level and increasing the number of threads increases the efficiency drastically



4. Next steps

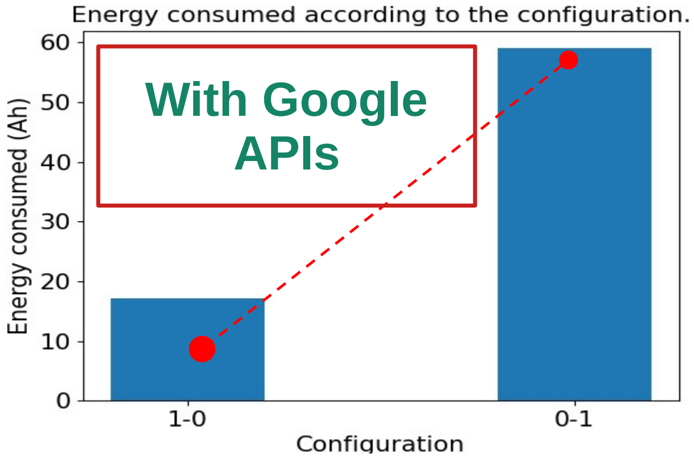
- Same experiments on Samsung
  - Good news: No limitations on the number of configurations as with APIs.
    We use *cc_info* file and the *power-meter*
  - We suspect that APIs on samsung was not far from reality in term of energy ratio.

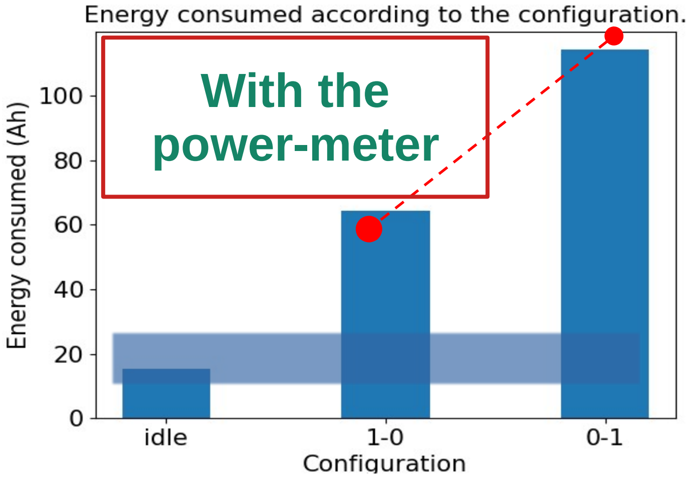




4. Next steps

- Made same experiments on Samsung
  - Good news: No limitations on the number of configurations as with APIs.
    We use *cc_info* file and the *power-meter*
  - We suspect that APIs on samsung was not far from reality in term of energy ratio.

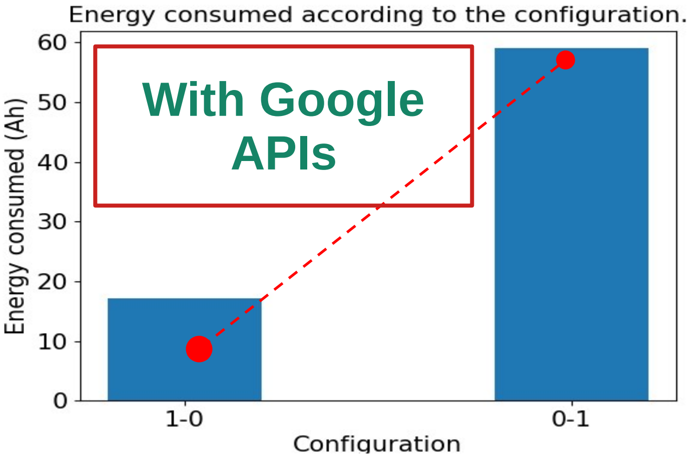


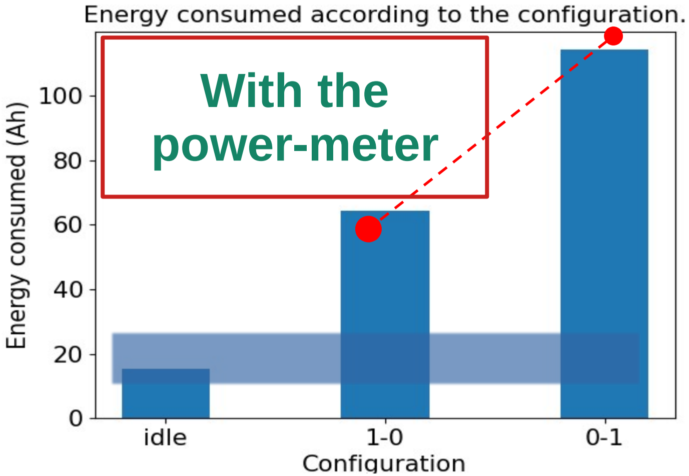


- **Validate observations made with other Benchmaks**

4. Next steps

- Made same experiments on Samsung
  - Good news: No limitations on the number of configurations as with APIs.
    We use *cc_info* file and the *power-meter*
  - We suspect that APIs on samsung was not far from reality in term of energy ratio.

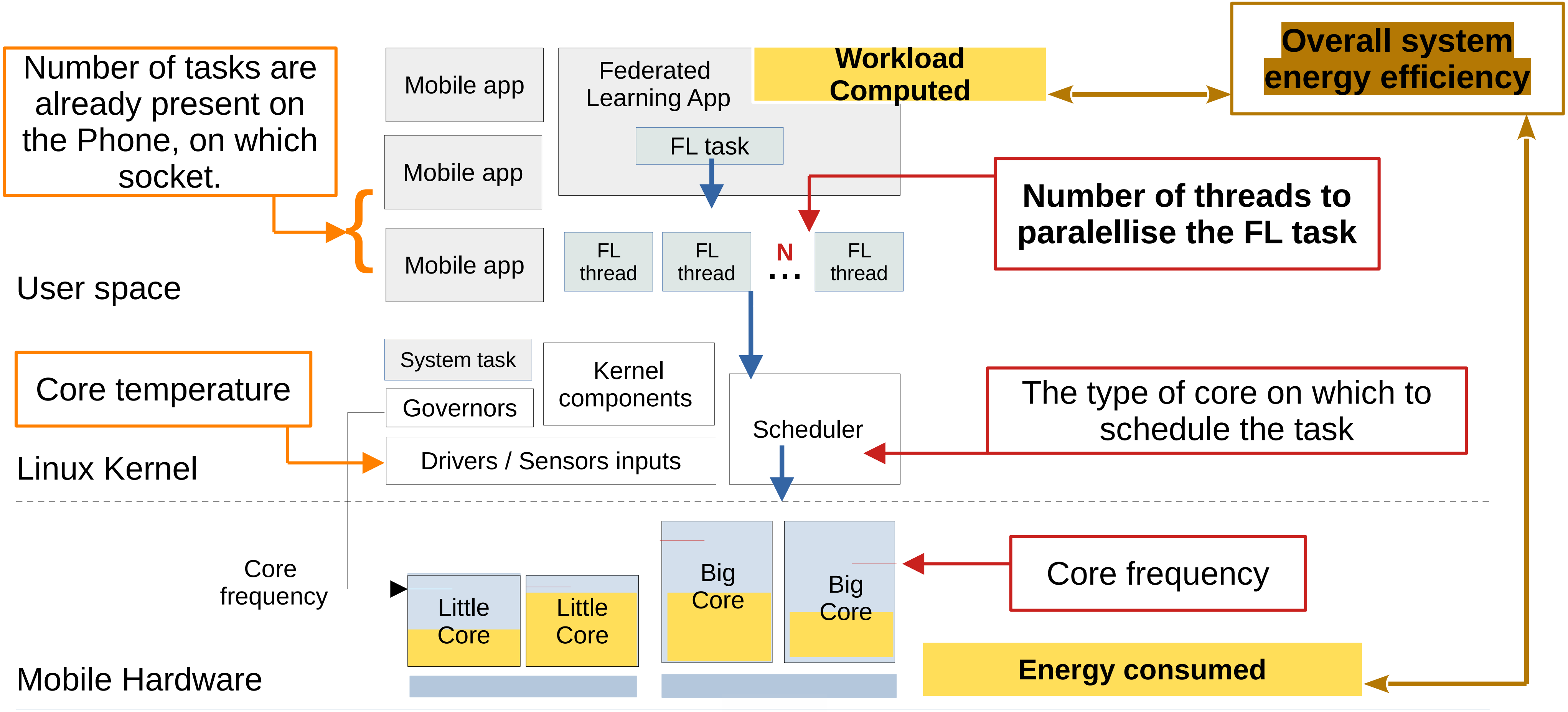




- **Validate observations made with other Benchmaks.**
- **Valorise lessons learned and observations (publication, solution..).**

Tank you for your attention.

General Problem Scheme



Number of tasks are already present on the Phone, on which socket.
Mobile app
Mobile app
Mobile app
Federated Learning App
FL task
Workload Computed
Overall system energy efficiency
Number of threads to paralellise the FL task
FL thread
FL thread
N
...
FL thread
User space
System task
Kernel components
Governors
Core temperature
Drivers / Sensors inputs
Scheduler
The type of core on which to schedule the task
Linux Kernel
Core frequency
Little Core
Little Core
Big Core
Big Core
Core frequency
Energy consumed
Mobile Hardware