Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading

Josep Vidal, Olga Muñoz, Antonio Pascual

Universitat Politècnica de Catalunya

Barcelona

www.ict-tropic.eu

Education / Research Institutes

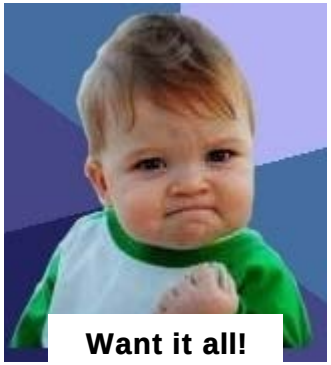
Technology providers

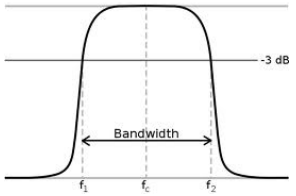
Manufacturer

Network operator

Increasing demands in the wireless world

The growing sophistication of smartphones allows to embrace an array of mobile of applications …





How can we modernize applications to improve user experience with today terminals?

Small-cell cloud computing

Cloud computing: allows terminals to have access to large resources

Small-cells: network capacity increase, energy savings, and new services

How about equipping each small cell base stations with computational and storage capabilities, so as to have a pervasive infrastructure of communication & computing devices?

Bring the cloud to the edge of the network! → Fog computing

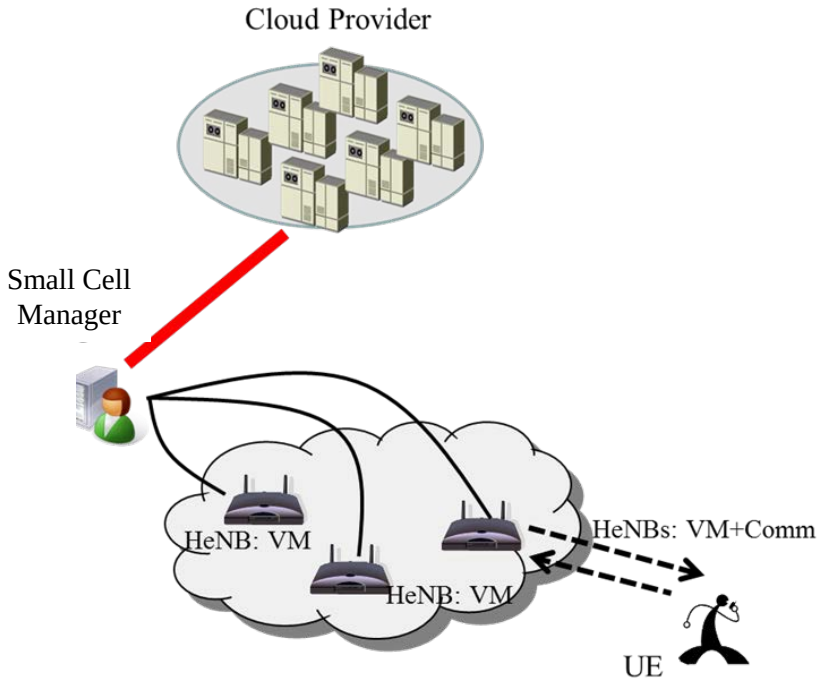
Offloading to small cell eNB: pros and cons

Running apps in empowered small cell eNB instead of external cloud

+ Reduced latency
+ Reduce usage of backhaul
- Management of virtual machines

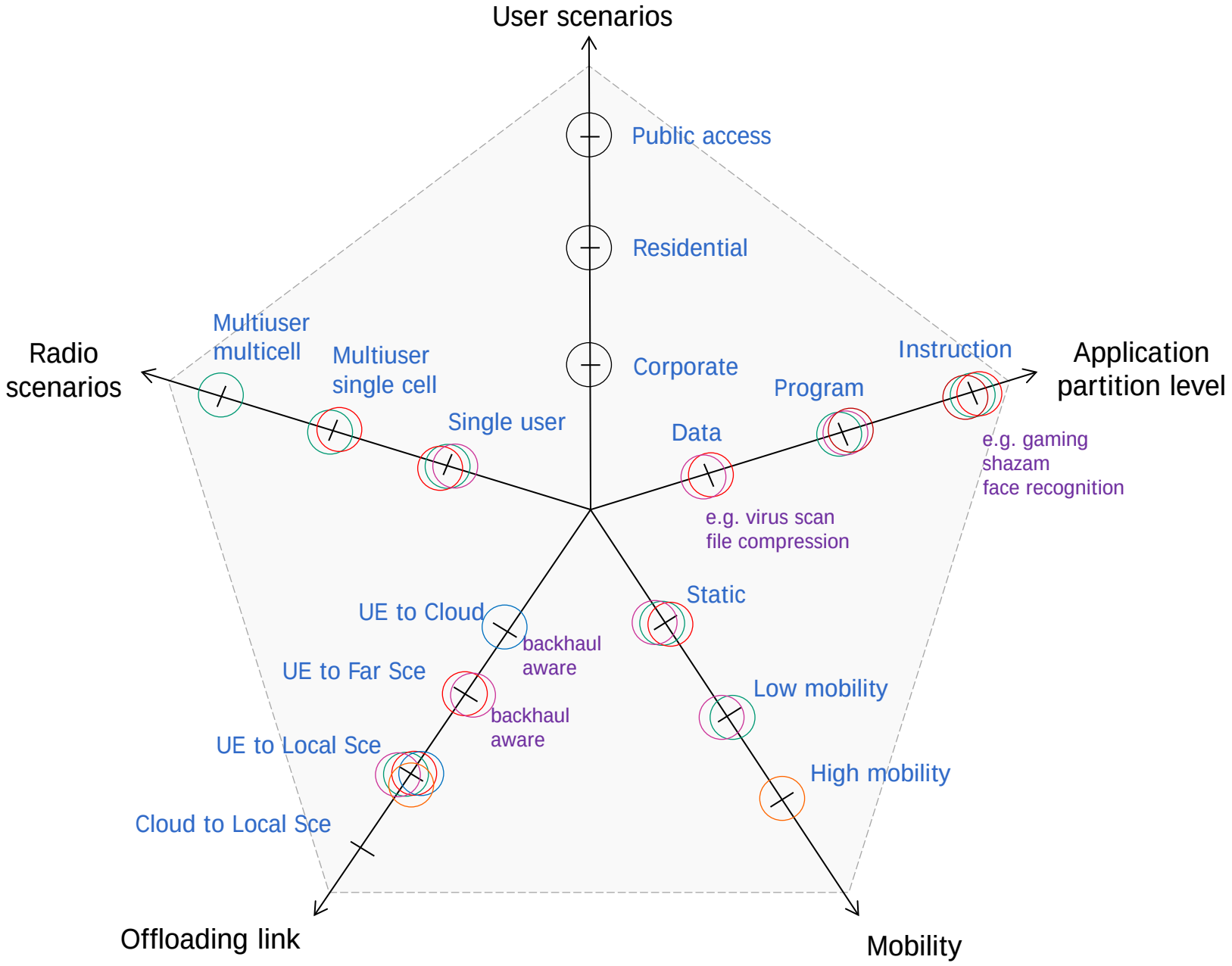
Running apps in empowered small cell eNB instead of MT

+ Improved latency and computation speed
+ Reduce battery consumption
- Management of parallelization
- Increased PHY utilization

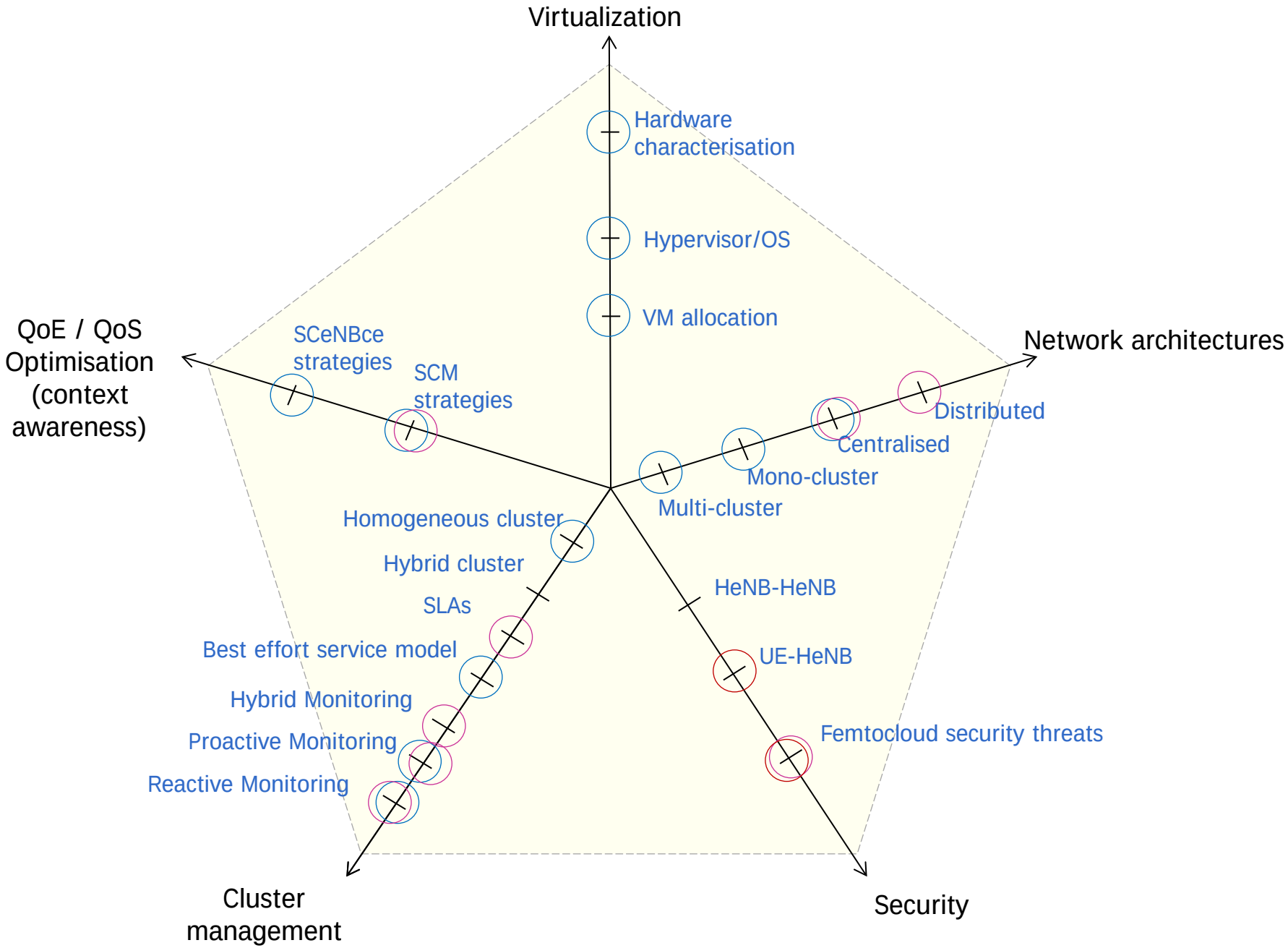


Small cell manager (SCM) is needed to allocate computational resources

Multidimensionality of the application offloading problem



Cloud and network support to application offloading



Joint optimization of the computation and radio resources

An energy-limited MT wants to launch a demanding application…

a) Shall I run the application locally?
b) Shall I run the application totally at the small cell cloud?
c) Shall I run the application partially at the small cell cloud?

What is the tradeoff in terms of energy and/or total computational time?

- The MT has less processing capabilities than the small cell cloud
- Additional energy and time for the exchange of input/output data

Joint optimization of the computation and radio resources

Our goal: Generate a **propagation channel-aware** method for the joint optimization of the computational and radio resources

Key performance indicators:

- **Energy consumption** for communication and computation
- **Total response time** for communication and computation

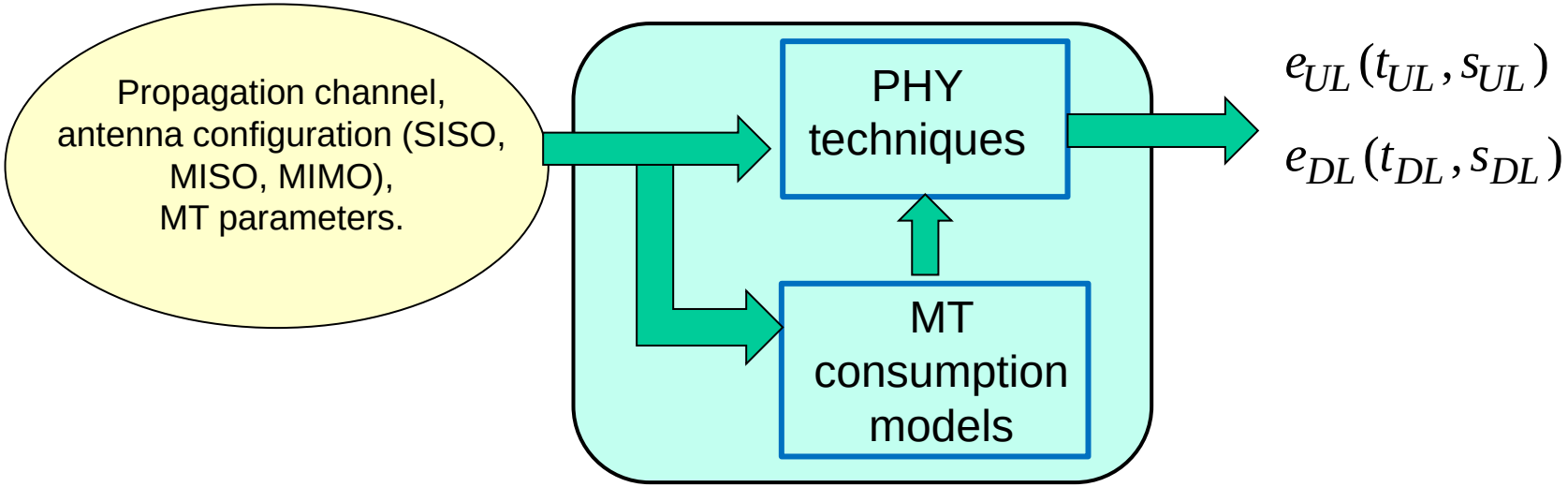
Two results from the optimization...

- Offloading decision (processing to be done locally/remotely),
- Optimal transmission strategy: transmission power, rate, MIMO precoders.

Communication abstraction models

We need closed form expression curves that relate:

- the energy spent (e_{UL}, e_{DL})
- the time in the transmission/reception (t_{UL}, t_{DL}),
- and the packet size (s_{UL}, s_{DL}).



From these curves one can optimize the latency and energy-trade off in the communication stage

Energy consumption at the MT

Power consumed by the MT when transmitting (UL)

$$p_{UL} \approx k_{tx,1} + k_{tx,2} \cdot p_{radiated}$$

Shannon's law

UL energy consumption:

$$e_{UL} \approx k_{tx,1} t_{UL} + k_{tx,2} t_{UL} \frac{2^{\frac{s_{UL}}{W_{UL} \cdot t_{UL}}} - 1}{\gamma_{UL}}$$

What is the optimal strategy for minimizing the energy spent in the communication?

Power consumed by the MT when receiving (DL)

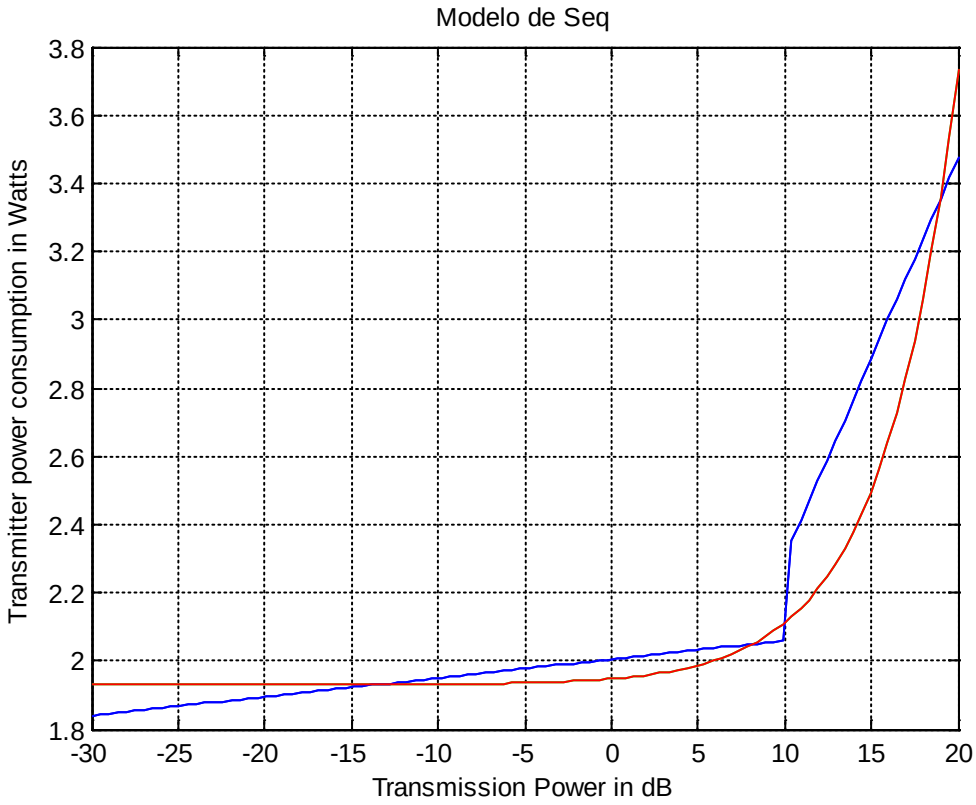
$$p_{DL} \approx k_{rx,1} + k_{rx,2} \cdot r_{DL}$$

DL energy consumption:

$$e_{DL} \approx k_{rx,1} t_{DL} + k_{rx,2} s_{DL}$$

For a given block size the energy in DL is minimized using the highest rate possible

Energy consumption at the MT



$k_{tx,1} = 0.4,\ k_{tx,2} = 18,$

$W_{UL} = 10MHz,\ s_{UL} = 5Mbyte$

Ack to Sequans Communications

UL energy consumption

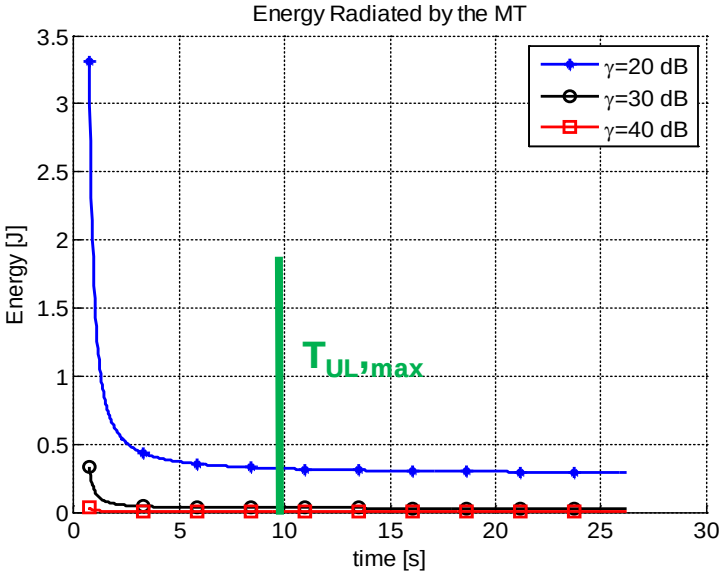
$$e_{UL}(t_{UL}, s_{UL}) = k_{tx,1} \cdot t_{UL} + k_{tx,2} \cdot t_{UL} \frac{2^{\frac{s_{UL}}{W_{UL} \cdot t_{UL}}} - 1}{\gamma_{UL}}$$

Radiated energy (SISO, MISO, SIMO)

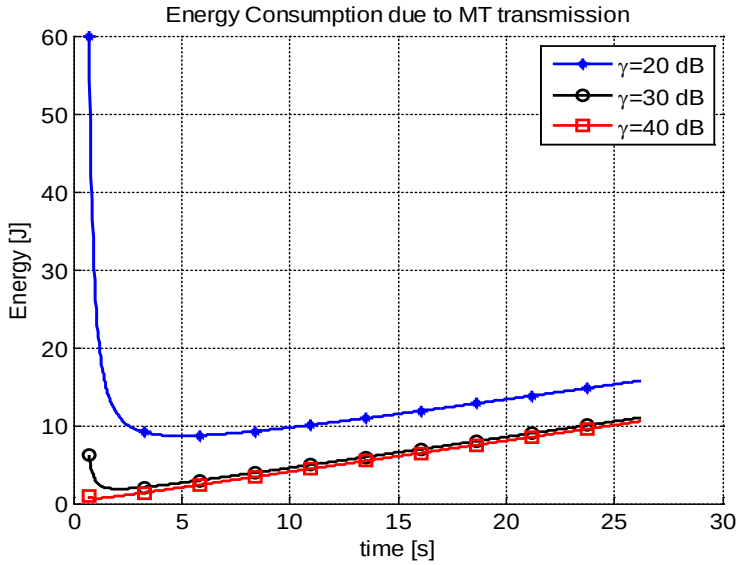
$k_{tx,1} = 0.4, \; k_{tx,2} = 18,$

$W_{UL} = 10MHz, \; s_{UL} = 5Mbyte$

UL communication: the optimal strategy in terms of **radiated energy** is to use the longest time possible (lowest rate)



UL communication: using the longest transmission time may not be optimal in terms of **total energy**



There is a trade-off between radiated energy and latency!

UL Energy consumption for the MIMO case

- The **UL transmission energy consumption** is minimized transmitting through the channel eigenmodes
- The # active eigenmodes and the waterlevel depends on t_{UL} and s_{UL}.

$$e_{UL}(t_{UL}, s_{UL}) = k_{tx,1} t_{UL} + k_{tx,2} t_{UL} \sum_{i=1}^{K(t_{UL},s_{UL})} \left(c(t_{UL}, s_{UL}) - \frac{1}{\lambda_i} \right)$$

$p_{radiated}$

$$c(t_{UL}, s_{UL}) = \frac{2^{\frac{s_{UL}}{W_{UL} t_{UL} K(t_{UL},s_{UL})}}}{\left(\prod_{k=1}^{K(t_{UL},s_{UL})} \lambda_k \right)^{\frac{1}{K(t_{UL},s_{UL})}}},$$

$waterlevel$

$e_{UL}(t_{UL}, s_{UL})$ is jointly convex in t_{UL} and s_{UL}, as it can be obtained as the solution of a convex problem with t_{UL} and s_{UL} as parameters

UL Energy consumption for the MIMO case

UL communication abstraction model for the general MIMO case:

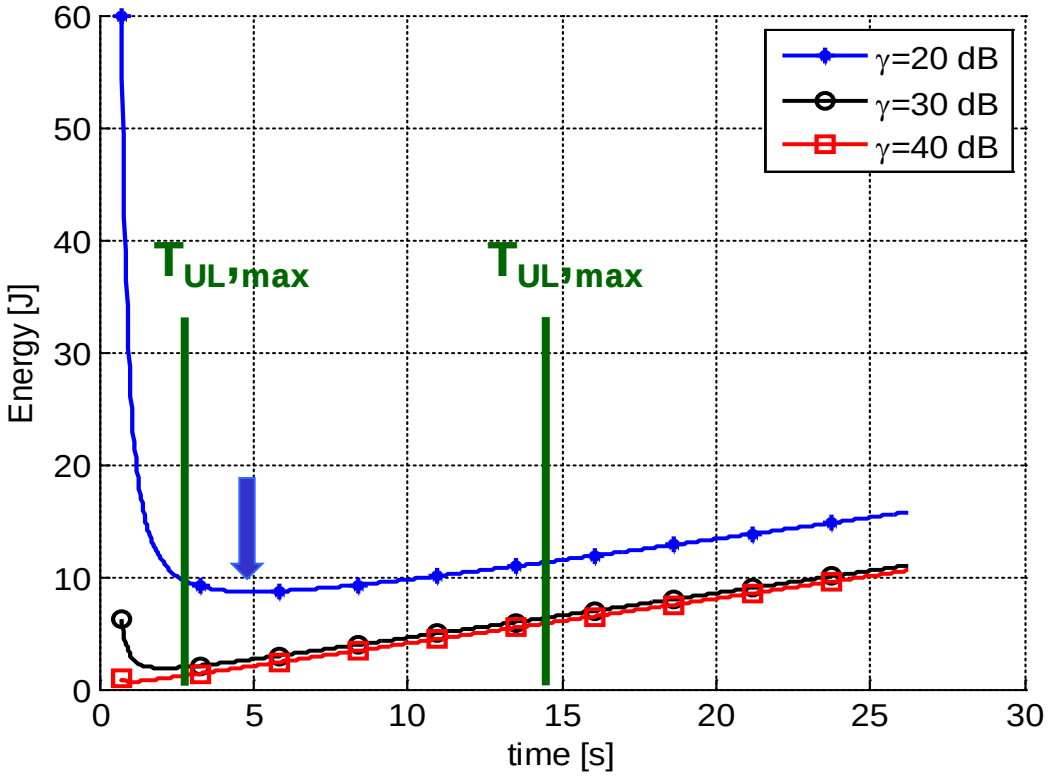
In the general MIMO case, $e_{UL}(t_{UL}, s_{UL})$ is jointly convex in t_{UL} and s_{UL}, as it can be obtained as the solution of a convex problem where t_{UL} and s_{UL} are parameters

$$
\begin{aligned}
\underset{b_{\text{UL}},\tau_{\text{UL}},\mathbf{Q}}{\text{minimize}} \quad & k_{\text{tx},1}\tau_{\text{UL}} + k_{\text{tx},2}\text{Tr}(\mathbf{Q}) \\
\text{subject to} \quad & C1: b_{\text{UL}} \leq W_{\text{UL}}\tau_{\text{UL}}\log_2\left|\mathbf{I} + \frac{\mathbf{HQH}^H}{\tau_{\text{UL}}}\right|, \\
& C2: \tau_{\text{UL}} = t_{\text{UL}}, \\
& C3: \mathbf{Q} \succeq 0, \\
& C4: b_{\text{UL}} \geq s_{\text{UL}}.
\end{aligned}
$$

The solution of this problem provides also the optimum energy covariance matrix, **Q***

UL energy consumption

Total energy spent in the UL vs. transmission time



Total energy is quasi convex on t_{UL}:

- it is either monotonic or,
- it has a single minimum

We still do not know how much time we can allocate to $T_{UL,max}$. For a total latency constraint, it depends on:

- The time required for computation
- The maximum latency constraint

Global resource allocation problem

Basic variables:

- Bits processed locally and remotely s_{P0}, s_{P1}
- Transmission time t_{UL}, t_{DL}

Objective: Minimize the energy consumption at the MT under a latency constraint imposed by the application

$$\underset{s_{P0}, s_{P1}, t_{UL}, t_{DL}}{\text{minimize}} \quad e_{UL}\left(t_{UL}, \beta_{UL} s_{P1}\right) + \varepsilon_{P0} s_{P0} + e_{DL}\left(t_{DL}, \beta_{DL} s_{P1}\right)$$

subject to

$$\max\left\{\tau_{P0} s_{P0}, t_{UL} + \tau_{P1} s_{P1} + t_{DL}\right\} \le L_{\max}, \quad \text{Latency constraint}$$

$$s_{P0} + s_{P1} = S_{app}, \quad \text{Full granularity}$$

$$e_{UL}\left(t_{UL}, \beta_{UL} s_{P1}\right) - k_{tx,1} t_{UL} \le k_{tx,2} t_{UL} P_{tx,MT}, \quad \text{Max. tx power}$$

$$\beta_{DL} s_{P1} \le t_{DL} R_{DL}^{\max}. \quad \text{Max. DL rate}$$

The problem is convex!

Understanding through simplification…

Normalized energy per bit in the UL

$$e_{UL}\left(t_{UL}, s_{UL}\right) = s_{UL} \cdot \bar{e}_{UL}\left(r_{UL}\right) \Rightarrow \bar{e}_{UL}\left(r_{UL}\right) = k_{tx,1}\frac{1}{r_{UL}} + k_{tx,2}\frac{1}{r_{UL}}\sum_{i=1}^{K\left(r_{UL}\right)}\left(c\left(r_{UL}\right) - \frac{1}{\lambda_i}\right)$$

- Given the channel and the terminal consumption features, the normalized energy per bit is depends only on the UL data rate

- Although $\bar{e}(r_{UL})$ is not necessarily convex, it has a global minimum at $R_{UL}{}^{*}$

Understanding through simplification…

Normalized energy per bit in the UL

$$e_{UL}\left(t_{UL}, s_{UL}\right) = s_{UL} \cdot \bar{e}_{UL}\left(r_{UL}\right) \Rightarrow \bar{e}_{UL}\left(r_{UL}\right) = k_{tx,1}\frac{1}{r_{UL}} + k_{tx,2}\frac{1}{r_{UL}}\sum_{i=1}^{K\left(r_{UL}\right)}\left(c\left(r_{UL}\right) - \frac{1}{\lambda_i}\right)$$

- The normalized energy per bit is a function of the UL data rate, and depends on the channel and terminal features.
- It has a **global minimum** (even it is non-convex)

Normalized energy per bit in the DL

$$e_{DL}\left(t_{DL}, s_{DL}\right) = s_{DL} \cdot \bar{e}_{DL}\left(r_{DL}\right) \Rightarrow \bar{e}_{DL}\left(r_{DL}\right) = \frac{k_{rx,1}}{r_{DL}} + k_{rx,2}$$

- It depends only on the DL data rate !
- It is optimum to use the highest r_{DL} possible.

Simplification of the problem

$$\underset{s_{P1}, r_{UL}}{\text{minimize}} \quad \beta_{UL} s_{P1} \bar{e}_{UL}\left(r_{UL}\right) + \varepsilon_{P0}\left(S_{app} - s_{P1}\right) + \beta_{DL} s_{P1} \bar{e}_{DL}\left(R_{DL}^{\max}\right)$$

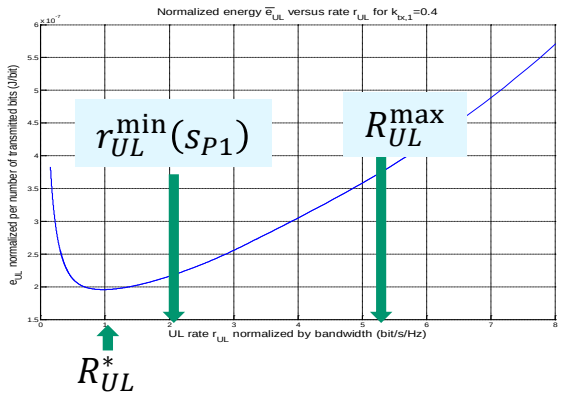
$$\text{s.t.} \quad S_{P1}^{\min} \le s_{P1} \le S_{P1}^{\max}$$

$$r_{UL}^{\min}\left(s_{P1}\right) \le r_{UL} \le R_{UL}^{\max}$$

The latency constraint imposes:

- A maximum and minimum value for s_{P1}
- A minimum value of r_{UL} for each value of s_{P1}

For each s_{P1}, there is an optimum r_{UL}



$$r_{UL}^{*}\left(s_{P1}\right) = \begin{cases} r_{UL}^{\min}\left(s_{P1}\right), & R_{UL}^{*} < r_{UL}^{\min}\left(s_{P1}\right) \\ R_{UL}^{*}, & r_{UL}^{\min}\left(s_{P1}\right) \le R_{UL}^{*} \le R_{UL}^{\max} \\ R_{UL}^{\max}, & R_{UL}^{*} > R_{UL}^{\max}. \end{cases}$$

The solution of the problem can be found as the solution of **one dimensional convex problem**

$$\underset{S_{P1}^{\min} \le s_{P1} \le S_{P1}^{\max}}{\text{minimize}} \; f_0\left(s_{P1}\right)$$

Optimality of total offloading / local computation

- **Total offloading is optimal if and only if** $\frac{df_0(s_{P1})}{ds_{P1}}|_{s_{P1}=\text{Sapp}} \leq 0$ **and** s_{P1}**=Sapp is feasible:**

$$\beta_{UL}\bar{e}_{UL}\left(r_{UL}^{*}\left(S_{app}\right)\right)+\beta_{DL}\bar{e}_{DL}\left(R_{DL}^{\max}\right)<\varepsilon_{P0}$$

$$S_{app}\left(\frac{\beta_{UL}}{r_{UL}^{*}\left(S_{app}\right)}+\tau_{P1}+\frac{\beta_{DL}}{R_{DL}^{\max}}\right)<L_{\max}$$

these expressions can be evaluated by the UE

⇒ less energy is spent in the communication than in the local execution

⇒ there is enough time for the UL/DL communication and remote computation

- **Local computation is optimal if and only if** $\frac{df_0(s_{P1})}{ds_{P1}}|_{s_{P1}=0} \geq 0$ **and** s_{P1}**=0 is feasible:**

$$\beta_{UL}\bar{e}_{UL}\left(r_{UL}^{*}\right)+\beta_{DL}\bar{e}_{DL}\left(R_{DL}^{\max}\right)>\varepsilon_{P0}$$

$$S_{app}\tau_{P0}<L_{\max}$$

these expressions can be evaluated by the UE

⇒ less energy is spent in the local execution than in the communication even for the rate spending least energy

⇒ there is enough time for local computation

.

Particular cases

- If the problem is not latency-constrained, we can use R_{UL}^{*}, it is optimum doing all the processing locally or remotely

$$s_{P_1}^{*} = S_{app}, \quad if \ \beta_{UL}\bar{e}_{UL}\left(R_{UL}^{*}\right) + \beta_{DL}\bar{e}_{DL}\left(R_{DL}^{\max}\right) < \varepsilon_{P0}$$

$$s_{P_1}^{*} = 0, \quad if \ \beta_{UL}\bar{e}_{UL}\left(R_{UL}^{*}\right) + \beta_{DL}\bar{e}_{DL}\left(R_{DL}^{\max}\right) > \varepsilon_{P0}$$

- If the goal is to minimize latency, then partial offloading is required, and the optimum partition depends on the maximum UL and DL data rate.

$$L_{\max} = L_o \quad \Rightarrow \quad \begin{cases} S_{P_0}^{*} = S_{\text{app}} \dfrac{\frac{\beta_{\text{UL}}}{R_{\text{UL}}^{\max}} + \tau P_1 + \frac{\beta_{\text{DL}}}{R_{\text{DL}}^{\max}}}{\frac{\beta_{\text{UL}}}{R_{\text{UL}}^{\max}} + \tau P_1 + \frac{\beta_{\text{DL}}}{R_{\text{DL}}^{\max}} + \tau P_0}, \\ S_{P_1}^{*} = S_{\text{app}} \dfrac{\tau P_0}{\frac{\beta_{\text{UL}}}{R_{\text{UL}}^{\max}} + \tau P_1 + \frac{\beta_{\text{DL}}}{R_{\text{DL}}^{\max}} + \tau P_0}, \end{cases}$$

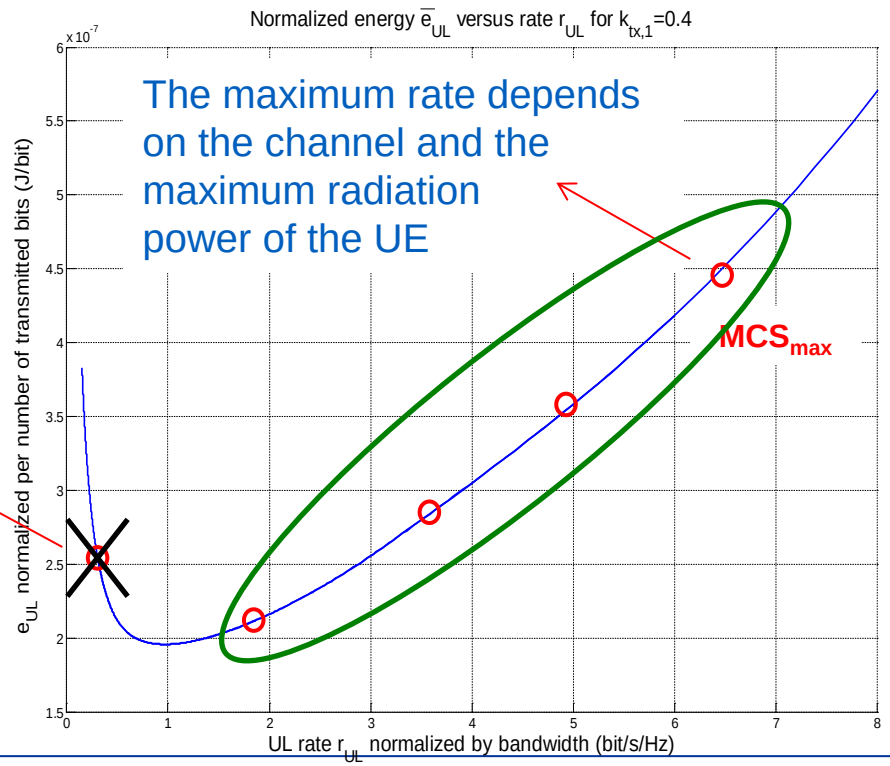
Practical estimation of the normalized energy per bit in UL

In practice, $\bar{e}_{UL}(r_{UL})$ can be estimated if the AP and the UE cooperate through "**training**":

- The UE ranges over different power levels and the AP indicates which is the MCS that can be supported
- The numerical values of the normalized energy are stored at the UE

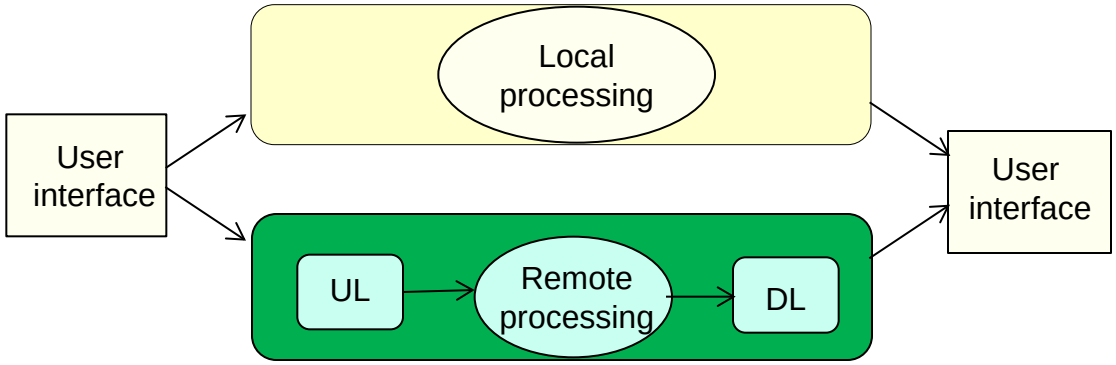
$$e_{UL}(t_{UL}, s_{UL}) = s_{UL}\bar{e}_{UL}(r_{UL})$$

This point is discarded because there are other points that are better both from the point of view of the rate and the energy



Application offloading: zip file compression

Gzip ASCII compress with a latency constraint: Some files compressed at the MT, some at the serving AP



Device/frequency	Power/W	Cycles/energy (C_{eff})
N810/400 MHz	0.8	480 MC/J
N810/330 MHz	0.7	480 MC/J
N810/266 MHz	0.5	540 MC/J
N810/165 MHz	0.3	510 MC/J
N900/600 MHz	0.9	650 MC/J
N900/550 MHz	0.8	690 MC/J
N900/500 MHz	0.7	730 MC/J
N900/250 MHz	0.4	700 MC/J

Table 1: Energy characterisitics of local compuing for Nokia N810 and N900 (MC=megacycle).

[Miettinen,2010]

Workload	Cycles/byte
`gzip` ASCII compress	330
`x264` VBR encode	1300
`x264` CBR encode	1900
`html2text` wikipedia.org	2100
`html2text` en.wikipedia.org	5900
`pdf2text` N900 data sheet	960
`pdf2text` E72 data sheet	8900

Table 3: Computation to data ratios for various workloads.

[Miettinen,2010]

Model for the processing at the MT

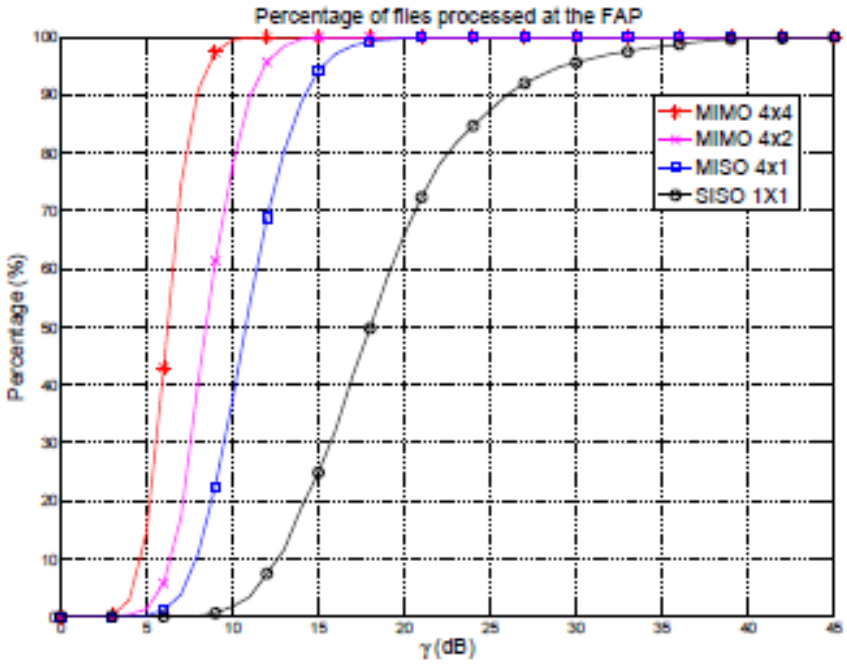
$$\varepsilon_{P_0} = 8.6 \cdot 10^{-8} \text{ J/bit}$$

$$\tau_{P_0} = 10^{-7} \text{ s/bit} \qquad \tau_{P_0} = 10^{-7} \text{ s/bit}$$

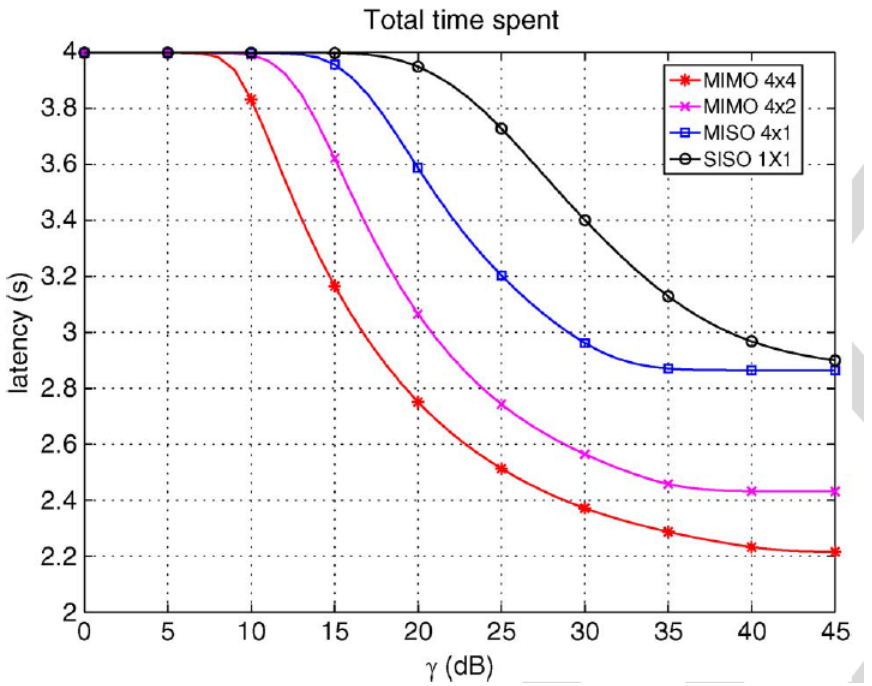
Numerical results

Application: gzip ASCII compress with a latency constraint (L_{max}=4s)

Total payload: 5 Mbytes



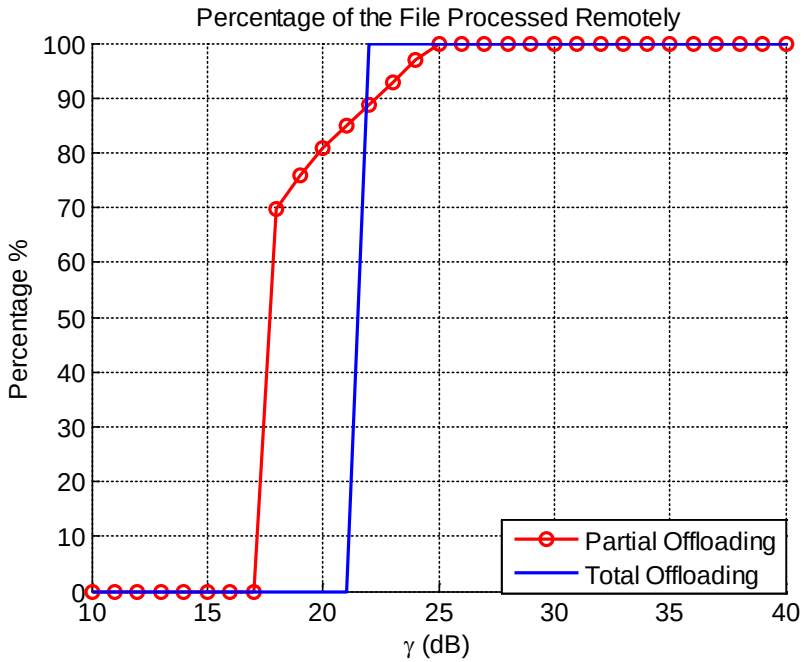
The offloading decision depends on the channel gain



For high channel gains, not all available time is used

Numerical results

Application: gzip ASCII compress with a latency constraint. Payload: 5 Mbytes



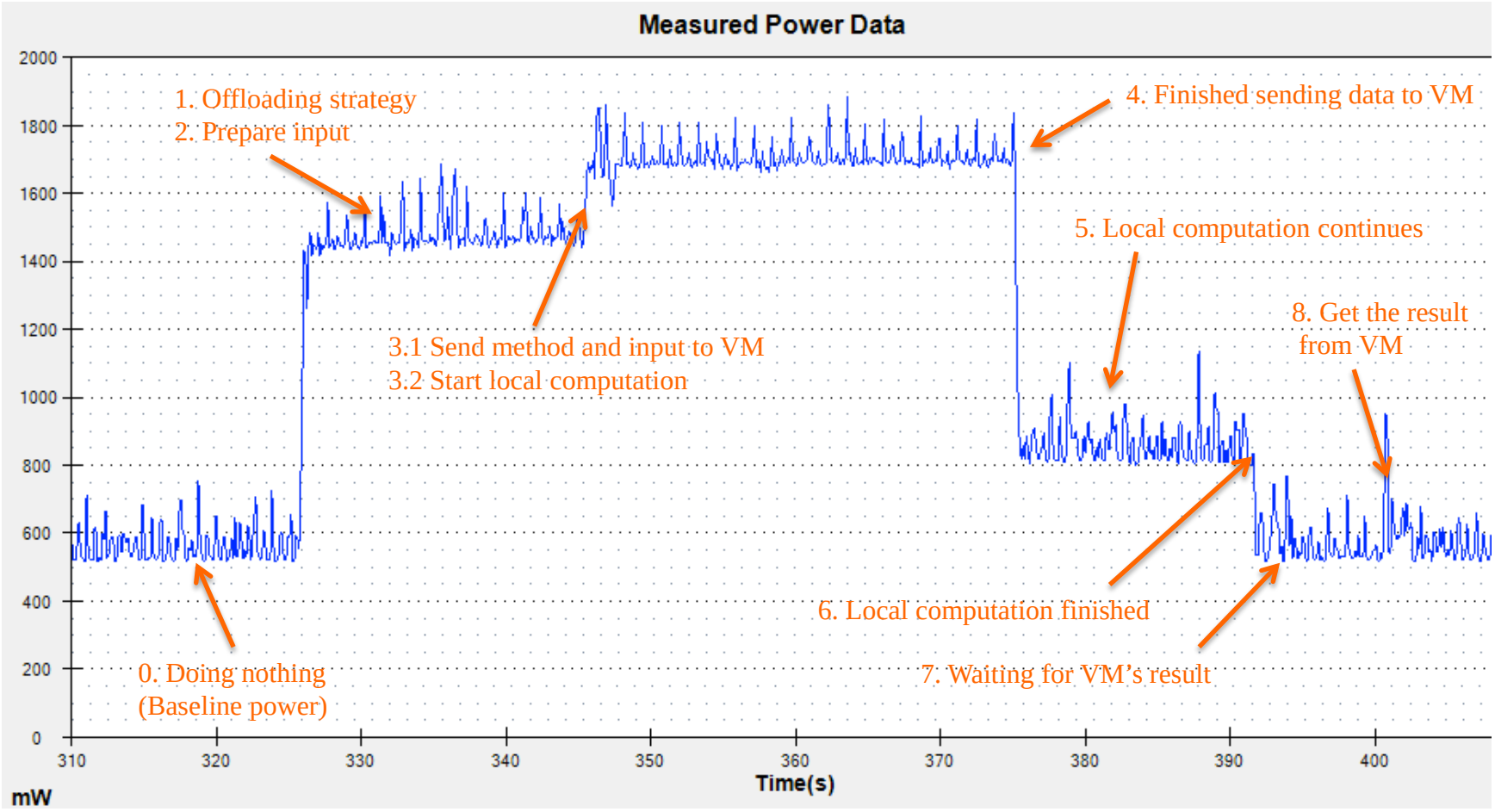
The actual offloading decision depends on the channel gain (L_{max}=4s):

- For low gain channels, it is better to do all the processing locally
- For high gain channels, it is better to do all the processing remotely

Partial offloading: Tasks can be distributed between the MT and the femto-cloud

Total/no offloading allowed: All tasks are forced to be done either locally or remotely

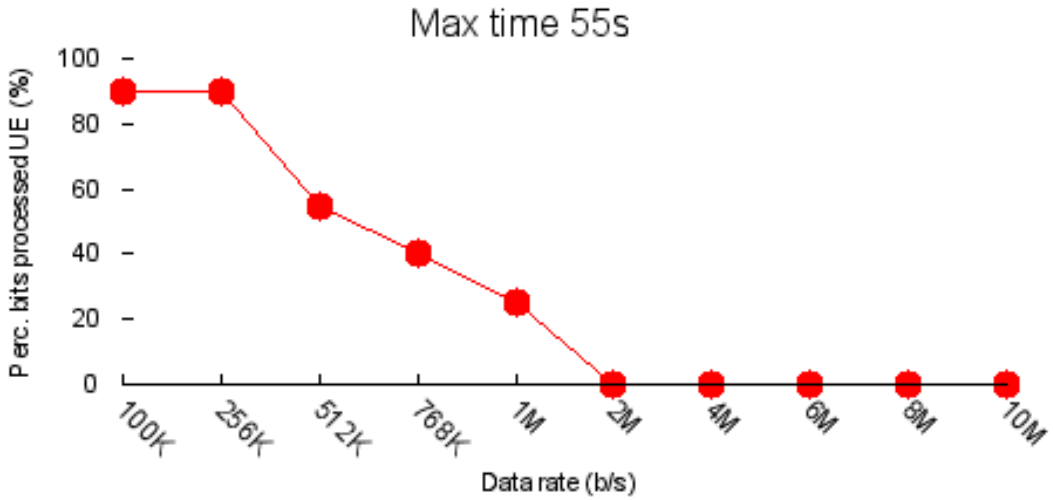
Proof of concept
(Ack. to J. Stefa, S. Kosta and A. Mei)

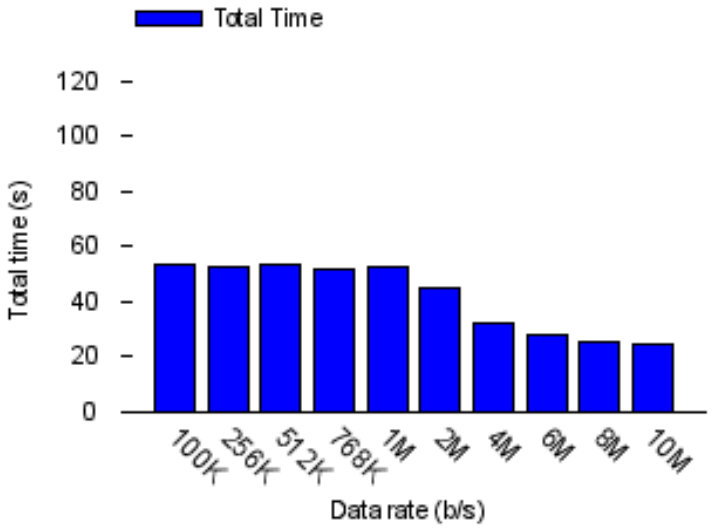


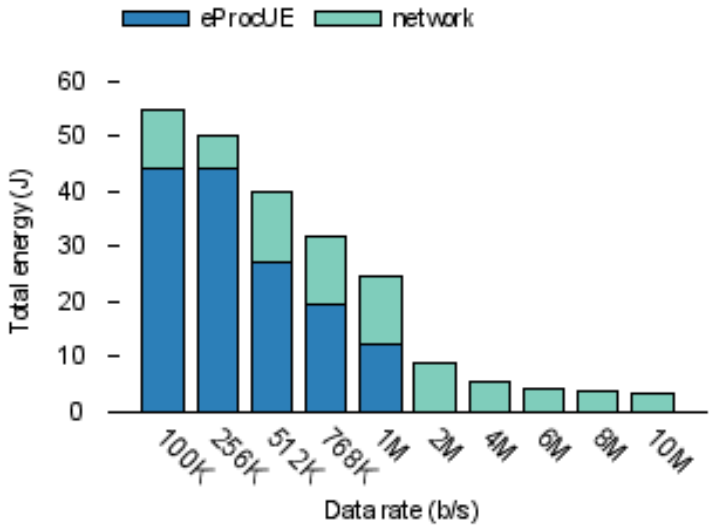
Power monitor: T_{max}=55 s, Data rate=512 kb/s

Proof of concept

(Ack. to J. Stefa, S. Kosta and A. Mei)

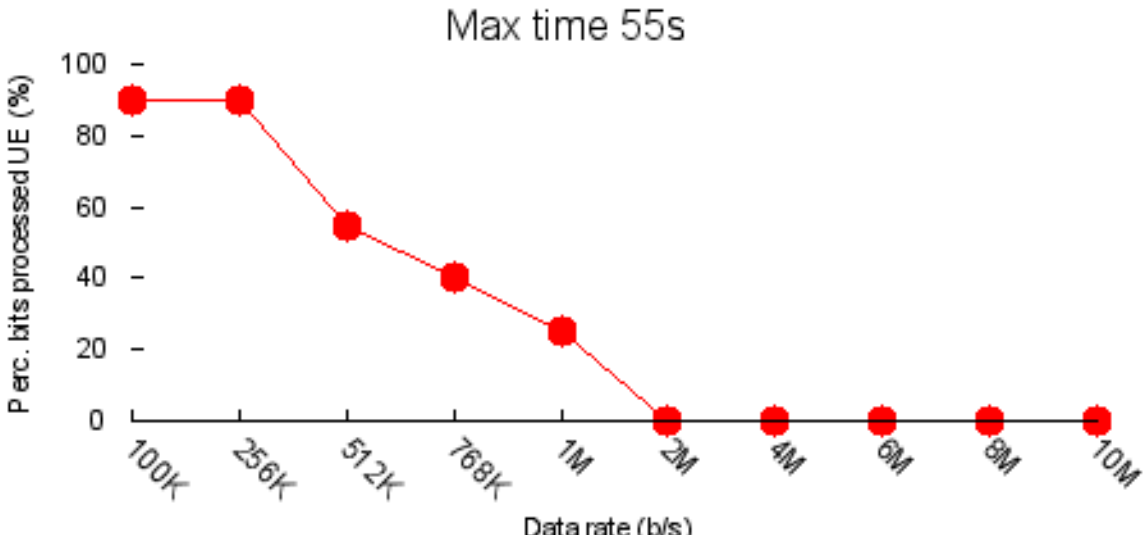


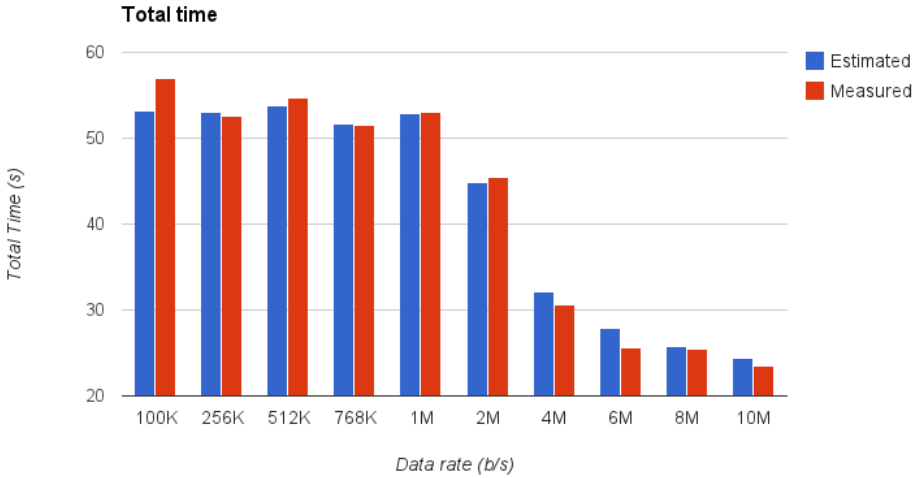


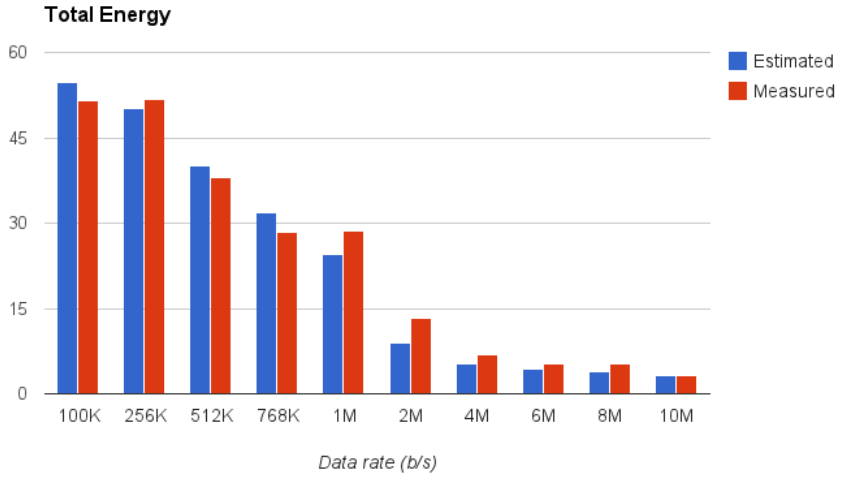


Proof of concept

(Ack. to J. Stefa, S. Kosta and A. Mei)







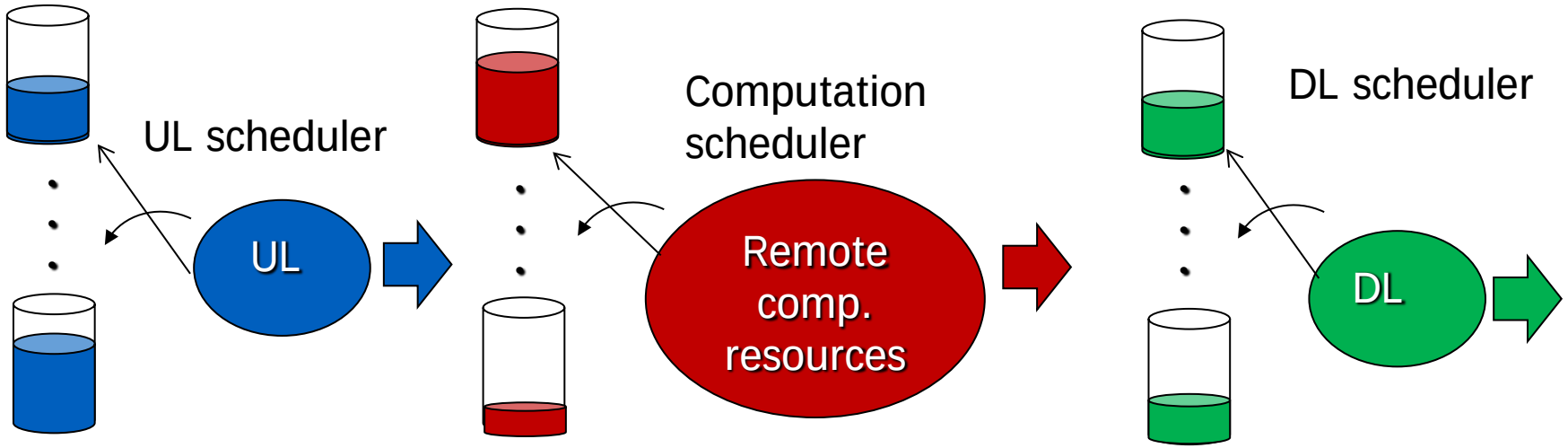
Estimated vs Measured for L_{max}=55s

Multiuser scenario

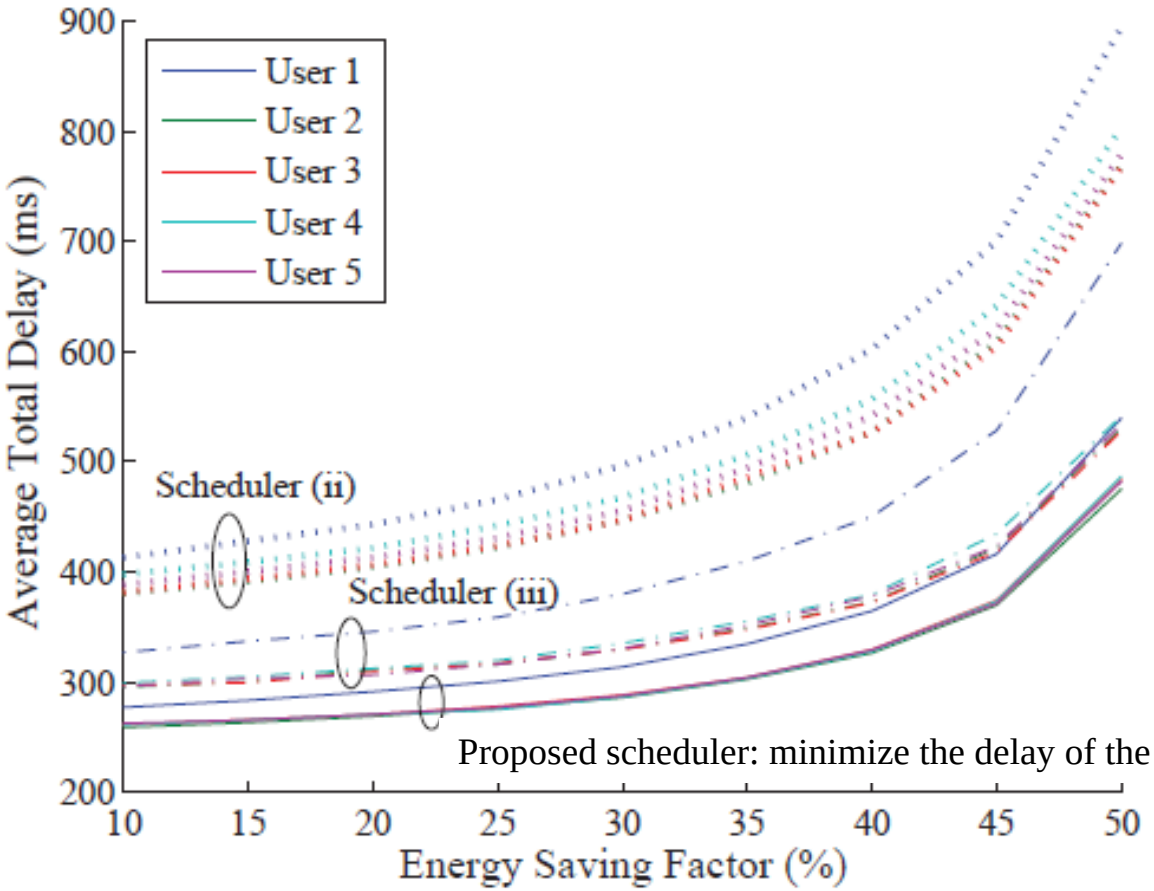
If multiple users are served by a single FAP, combine energy-latency resource allocation with scheduling policies.

Assumptions:

- Multiple users generate offloading petitions at a certain rate
- The offloading decision has already been taken for each UE involved
- A certain probability of latency performance can be guaranteed



Multiuser scenario



MT is a Nokia N900: 650 Mcycles/Joule when operating at 600 MHz

Proposed scheduler: minimize the delay of the the worst case user

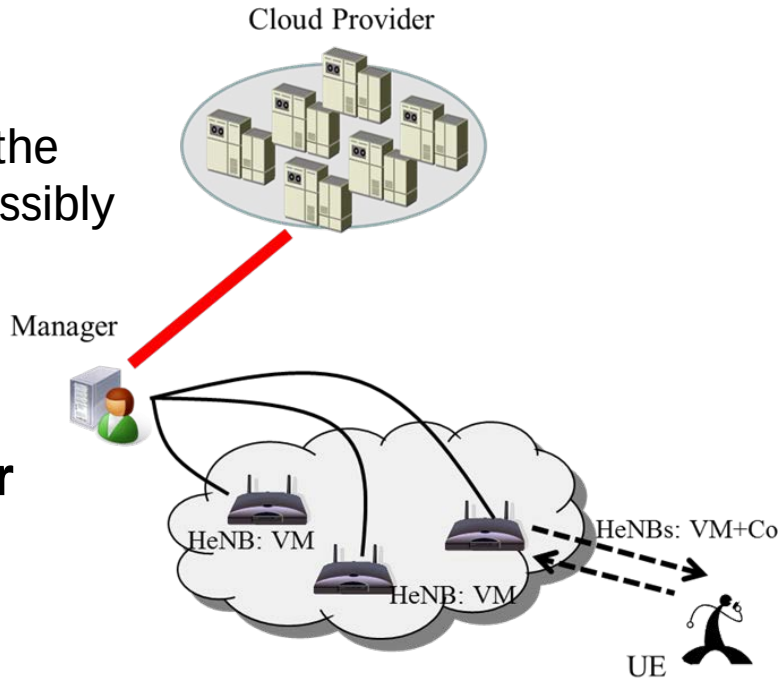
M. Molina, O. Muñoz, A. Pascual, J. Vidal, "Joint Scheduling of Communication and Computation Resources in Multiuser Wireless Application Offloading", IEEE PIMRC 2014, Washington, USA

What is next ?

Several APs available for computation:

The serving AP and other APs connected to the serving one with a non-ideal backhaul but possibly with higher computational capabilities.

What is best?
Do everything at the closest AP or distribute the processing?



The order of activation of the VMs does not depend on their computation capabilities but on the latency of the backhaul.

Once a VM is activated the amount of computation allocated to this VM depends on both computation capability and backhaul latency.

Thank you for your attention!

Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading

Universitat Politècnica de Catalunya (UPC)
Barcelona