Merging mmWave and Mobile Edge Computing in Future 5G Networks

Sergio Barbarossa

SAPIENZA
UNIVERSITÀ DI ROMA

Acknowledgments: 5G-MiEdge H2020 EU/Japan Project

Stefania Sardellitti, Paolo Di Lorenzo, Elena Ceci, Mattia Merluzzi

Acknowledgment

H2020 EU/Japan Project 5G-MiEdge:
Millimeter-wave Edge Cloud as an Enabler for 5G Ecosystem

List of participants

1 (EU coordinator) Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. Heinrich- Hertz-Institute, Germany

2 Commissariat à l'énergie atomique (CEA), France

3 Intel Deutschland GmbH Intel Germany

4 Telecom Italia, Italy

5 Sapienza University of Rome, Italy

6 (JP coordinator) Tokyo Institute of Technology, Japan

7 KDDI R&D Laboratories Inc., Japan

8 Panasonic AVC Networks Company, Japan

Summary

- Major thrusts of 5G:

  - Millimeter wave communications
  - Mobile Edge Computing

- The 3 Primary Colors of 5G: Communication-Computation-Caching (C^3)

  - communication vs. caching
  - communication vs. computing

- Computation offloading: Joint optimization of comm/comp resources

- Merging MEC & mmWave

- Conclusion

MEC: a key ingredient of 5G

Major players in 5G roadmap

- Millimeter wave links
- Dense deployment of small cell base stations
- Massive MIMO
- Network functionality virtualization (NFV)
- Application-centric architecture
- Mobile edge computing

How to achieve 1,000 increase of system capacity

$$10^3 = 10 \times 10 \times 10$$

How to achieve 1,000 increase of system capacity

$$\frac{\text{bit/sec}}{\text{Km}^2} = \frac{\text{bit/sec/Hz}}{\text{cell}} \times \frac{\text{cell}}{\text{Km}^2} \times \text{Hz}$$

spectral efficiency x 10

massive MIMO

dense deployment of access points x 10

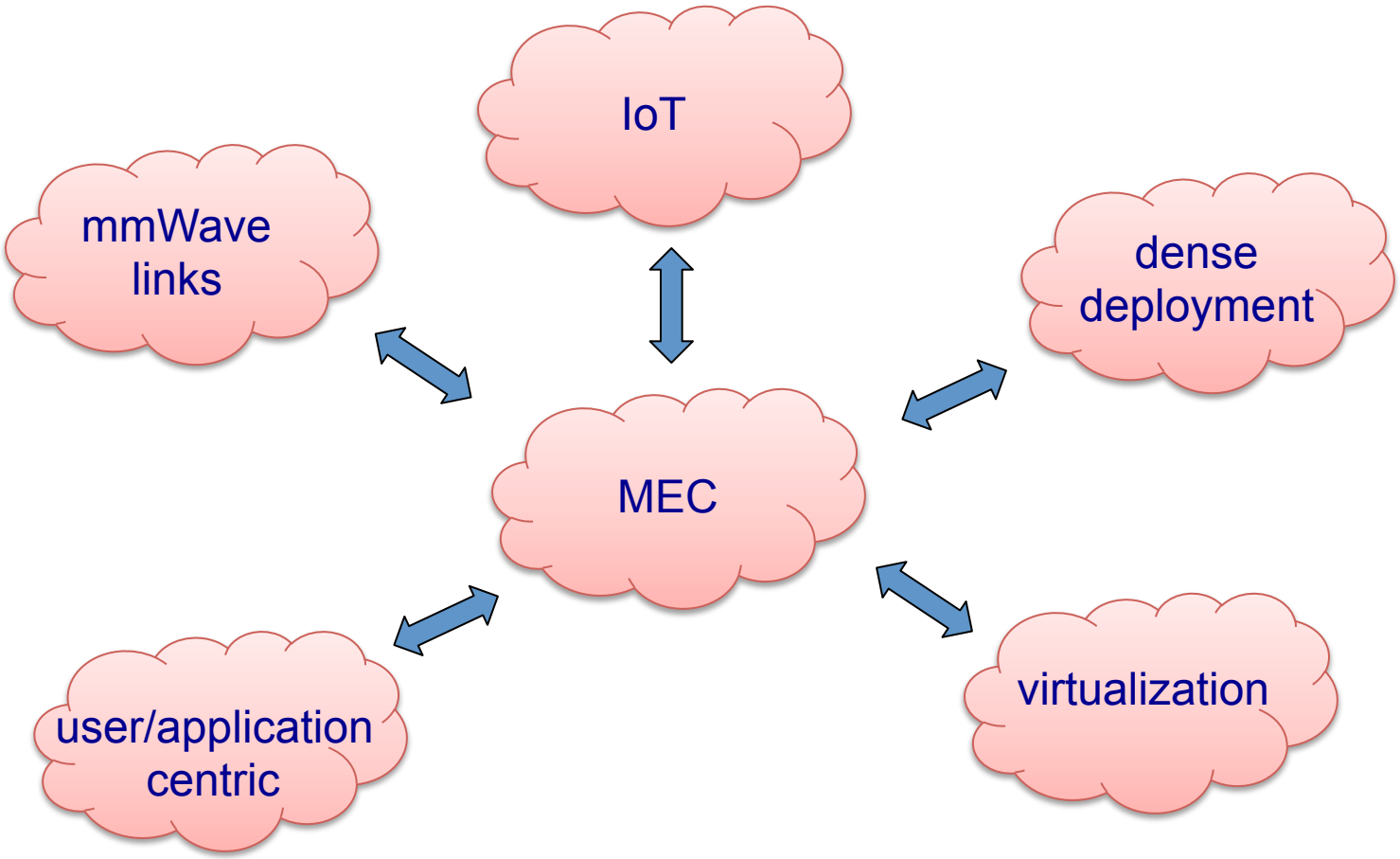
bandwidth x 10

increase carrier frequency

mmWave communications facilitate all three improvements

MEC: a key ingredient of 5G

MEC fits 5G roadmap perfectly well !



MEC vs. mmWaves

MEC ⟸ **mmWaves**

mmWaves provide high capacity radio access and wireless backhauling, facilitating low latency access to MEC services

MEC ⟹ **mmWaves**

MEC provides local computation power useful to optimize performance of mmWave communications

MEC vs. dense deployment

MEC ⟸ **dense deployment**

dense deployment of access points (AP) endowed with MEC functionalities bring IT services close to mobile end-user

MEC ⟹ **dense deployment**

MEC enables implementation of sophisticated interference mitigation techniques in a dense environment

MEC vs. virtualization

MEC ⟸ **virtualization**

virtualization enables deployment of virtual machines (VM) when/where needed

MEC ⟹ **virtualization**

MEC facilitates orchestration of VM's and their management taking into account user mobility

MEC vs. application-centric design

MEC ⟸ **application-centric design**

Application-centric design benefits from multi-tier service delivery where applications run as close as possible to end user

MEC ⟹ **application-centric design**

MEC promotes development of context-aware and RAN-aware applications

MEC vs. Internet of Things (IoT)

MEC ⟸ **IoT**

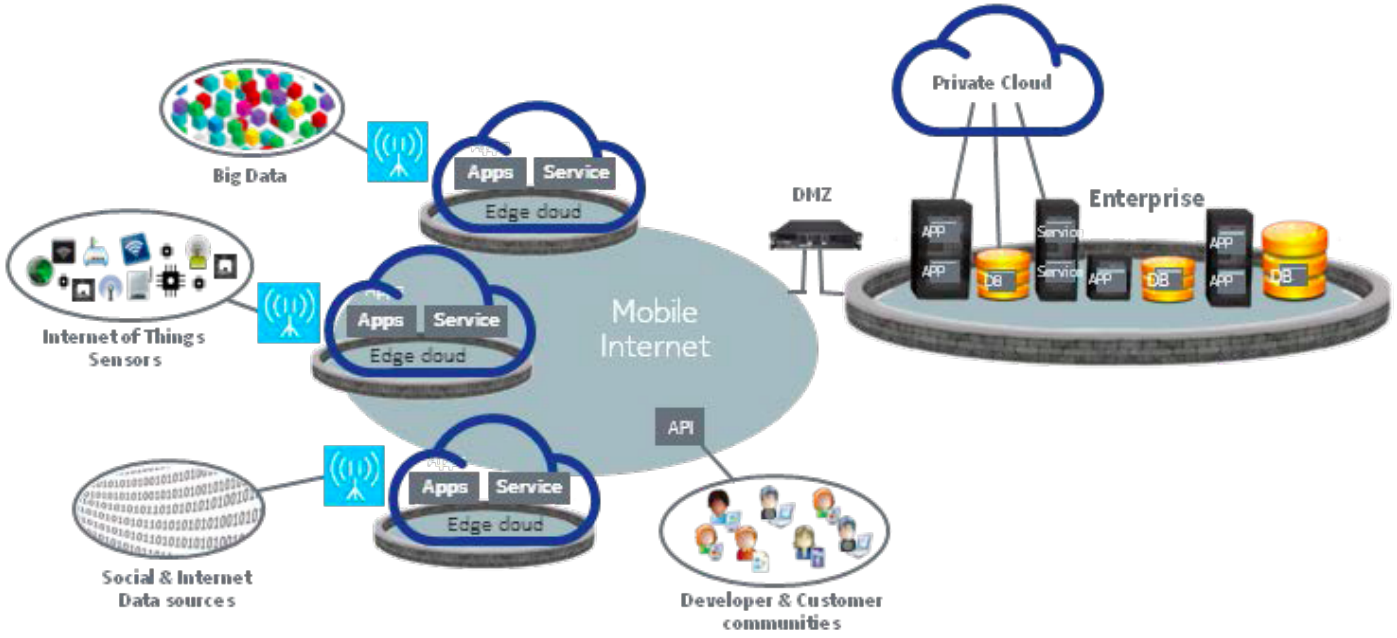
IoT expands MEC services to all “things”: sensors, actuators, …

MEC ⟹ **IoT**

MEC empowers tiny devices with significant additional computational capabilities through computation offloading

Mobile Edge Computing

Main idea: Offer application developers and content providers cloud-computing capabilities and IT services at the edge of the mobile network

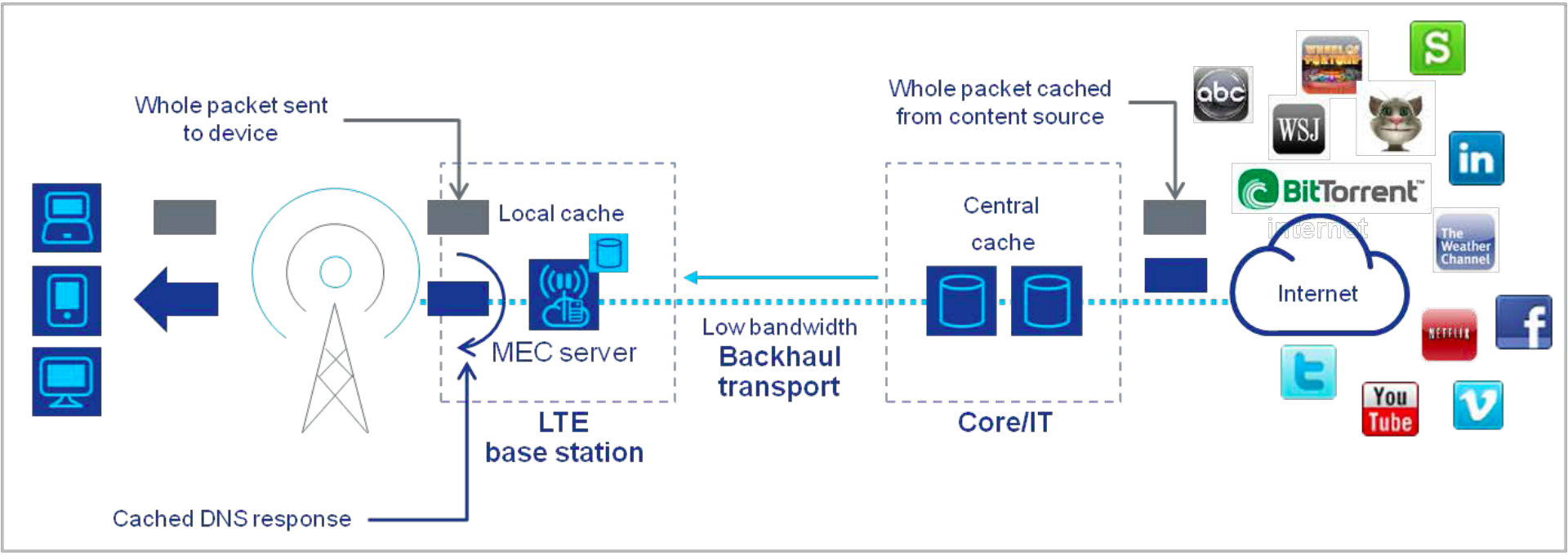


Specific features: - Proximity

- Ultra-low latency
- Real-time access to radio network information
- Location awareness

M. Patel et al.. "Mobile-Edge Computing – Introductory Technical White Paper", ETSI white paper, Sep 2014

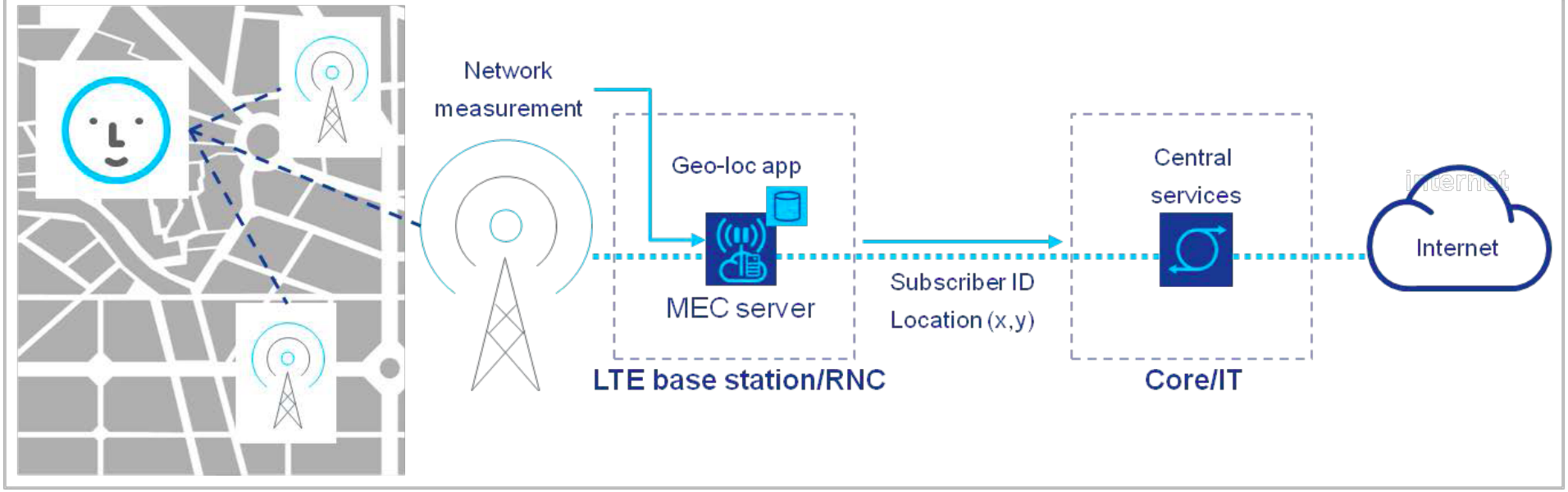
Use cases: Distributed Content and DNS Caching



- Popular content/data stored at the Base Station
- Backhaul and Transport savings (up to 35%)
- Improved QoE (20% improvement for loading a Web page)

M. Patel et al.. "Mobile-Edge Computing – Introductory Technical White Paper", ETSI white paper, Sep 2014

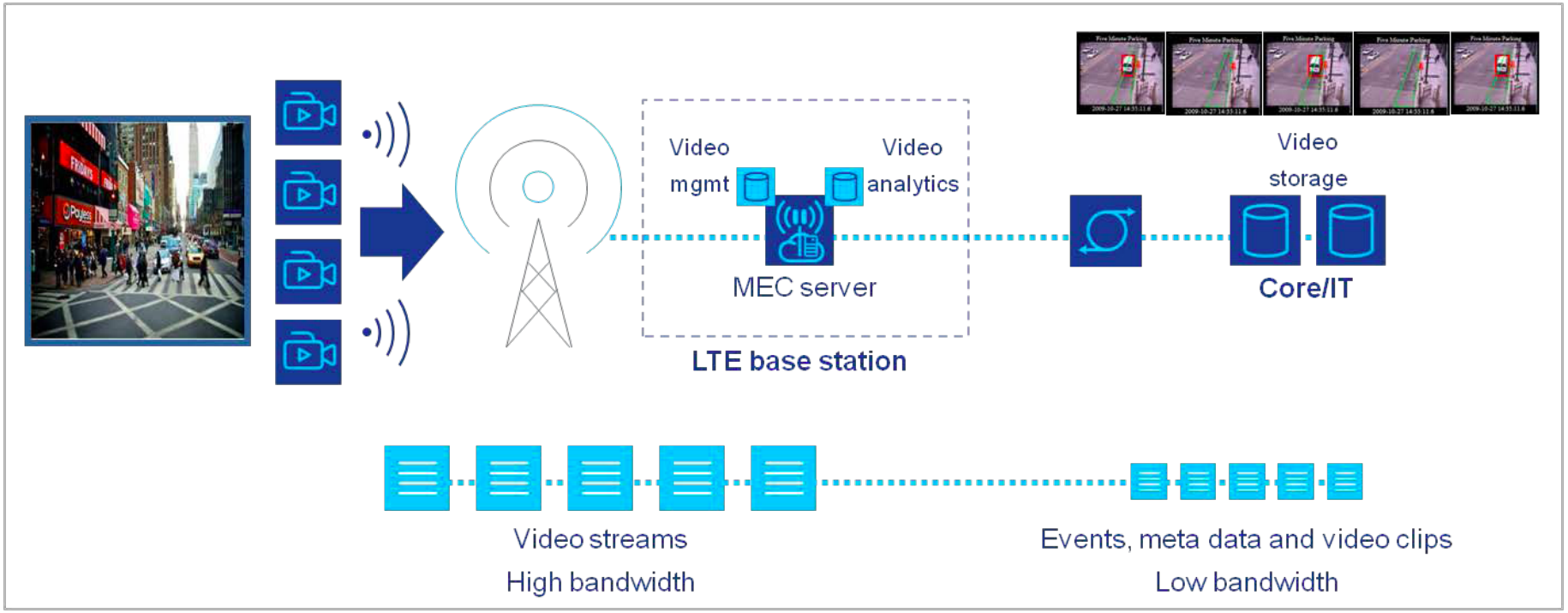
Use cases: Active Device Location Tracking



- Get Mobile Device location in real time and in a passive way (no GPS)
- Understand how the crowd is distributed or locate specific users
- Relevant in Smart City (Macro cells), retail (micro cells), and advertising

M. Patel et al.. "Mobile-Edge Computing – Introductory Technical White Paper", ETSI white paper, Sep 2014

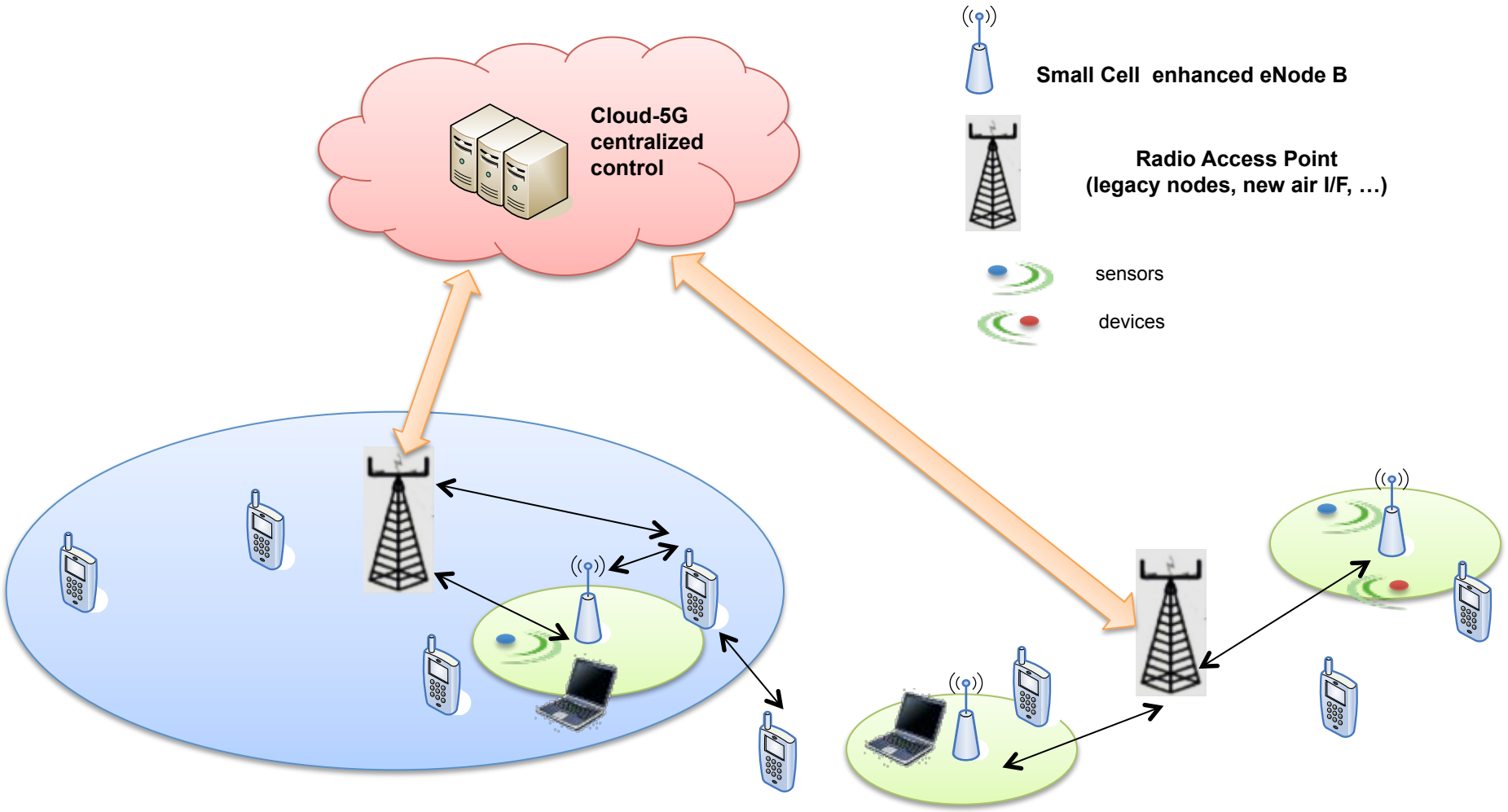
Use cases: Intelligent Video Analytics



- Analyze live video streams at the base station
- Trigger events automatically (e.g. abandoned bags, missing objects, crowd, etc.)
- Public safety, smart cities

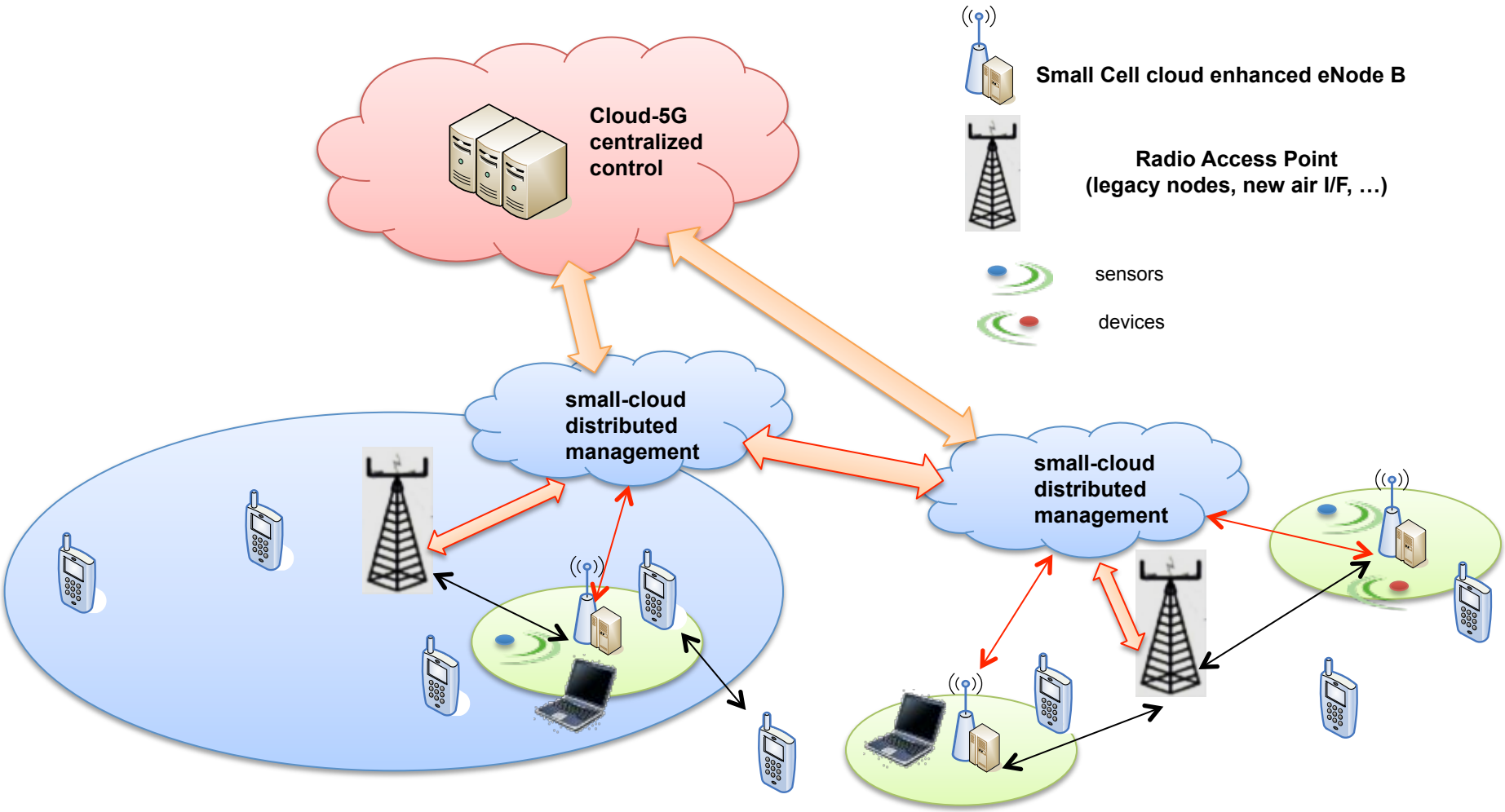
M. Patel et al.. "Mobile-Edge Computing – Introductory Technical White Paper", ETSI white paper, Sep 2014

Mobile Cloud Computing (MCC)



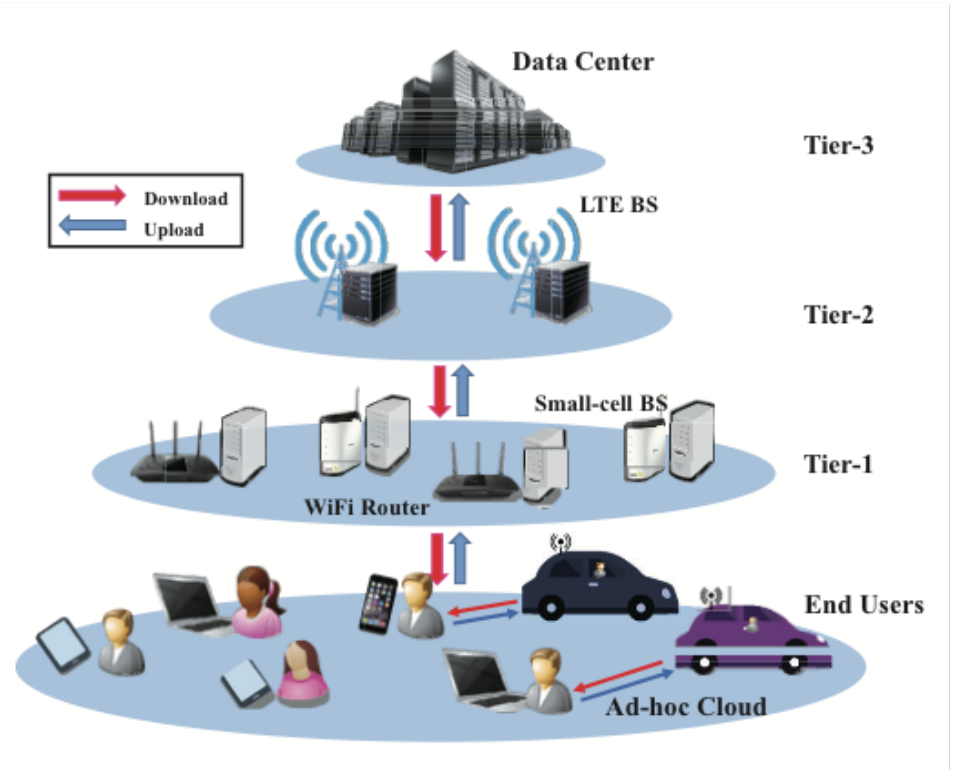
MCC is a centralized architecture: Intelligence is in the cloud

MEC: multi-tier IT service delivery



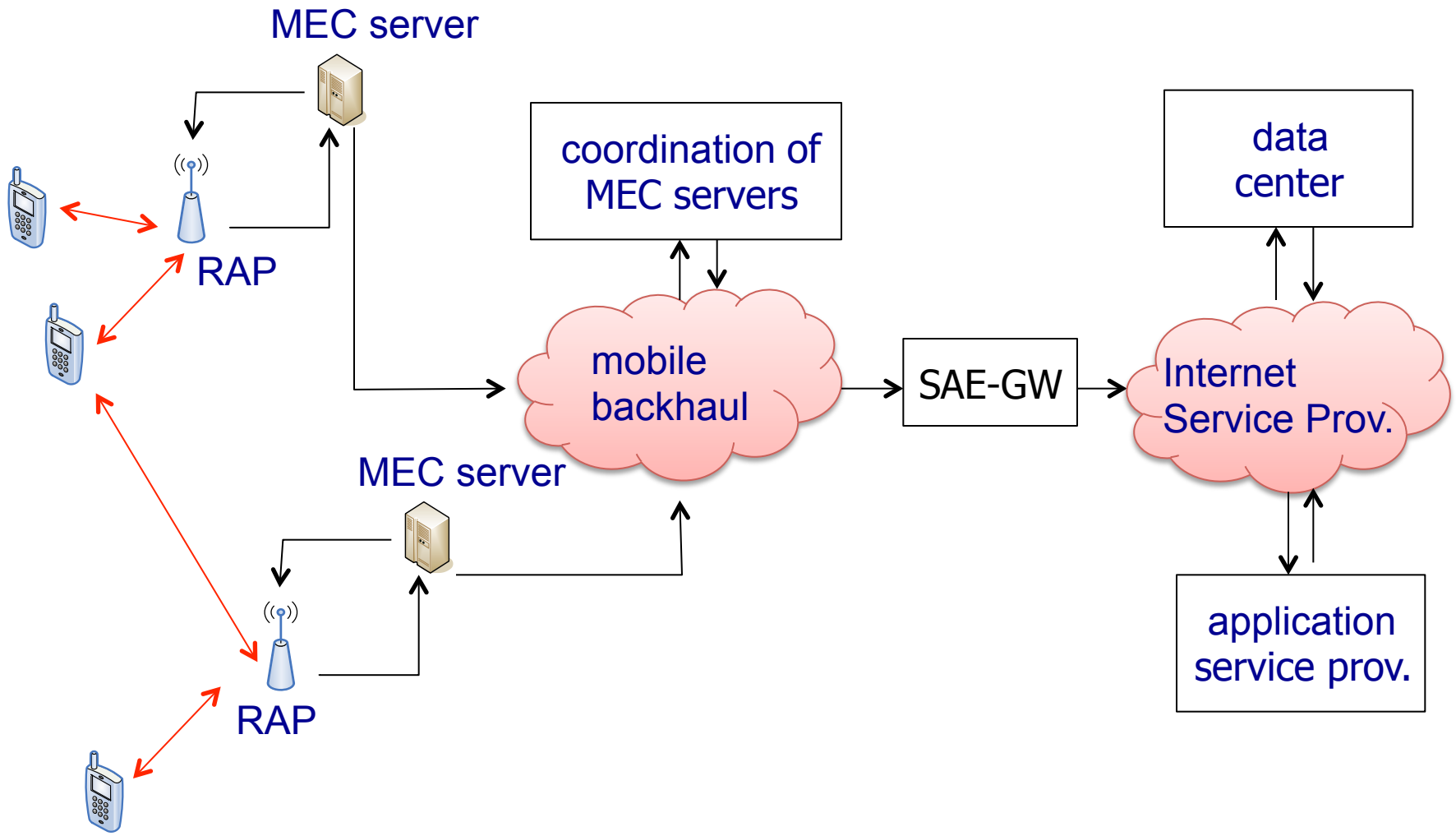
MEC is decentralized: Intelligence spreads towards the periphery

MEC: multi-tier IT service delivery ⇨ scalability



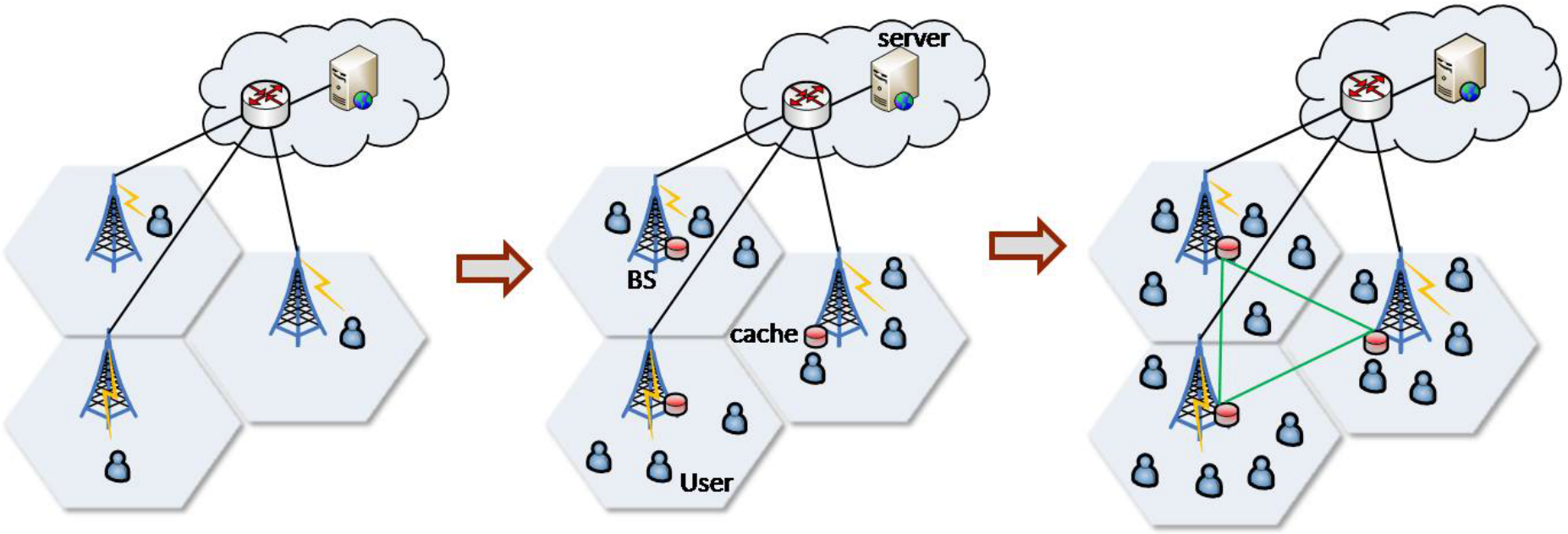
Applications run when/where more appropriate, given latency/reliability constraints and resource availability

MEC basic architecture



The 3 Primary Colors of 5G

Communication-Computation-Caching are three faces of a common framework



Communication

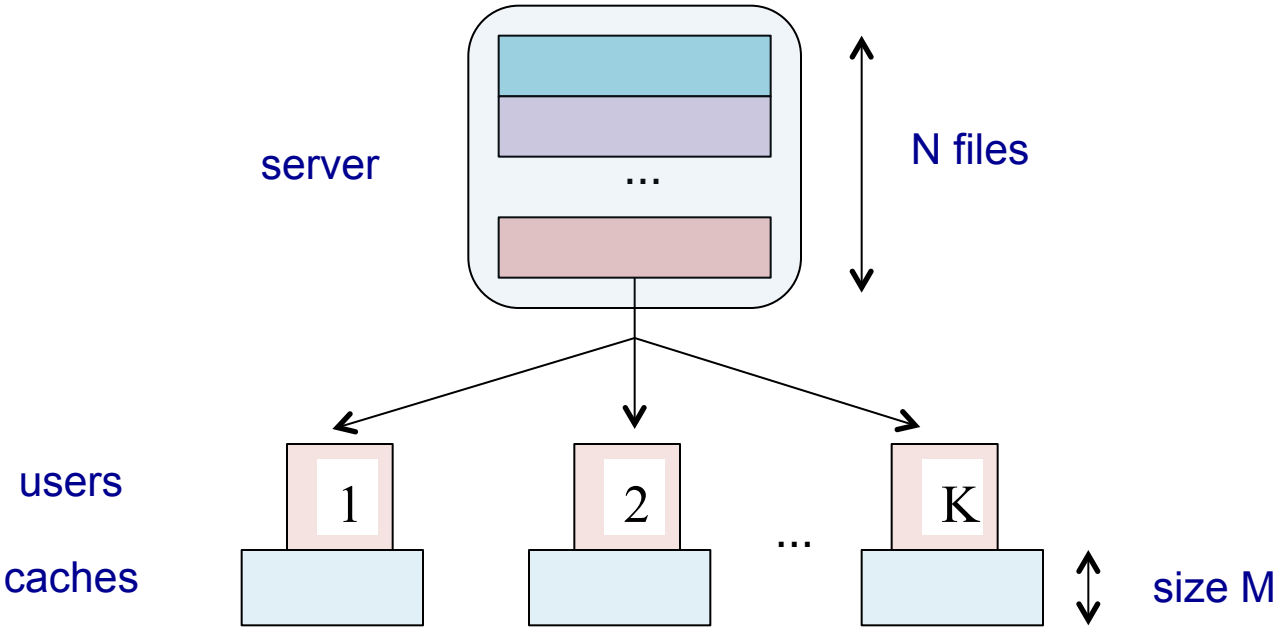
Communication + Caching

Communication + Caching + Computation

This holistic view suggests joint optimization of C^3 resource allocation

The 3 Primary C's of 5G

Fundamental trade-off between communication & caching

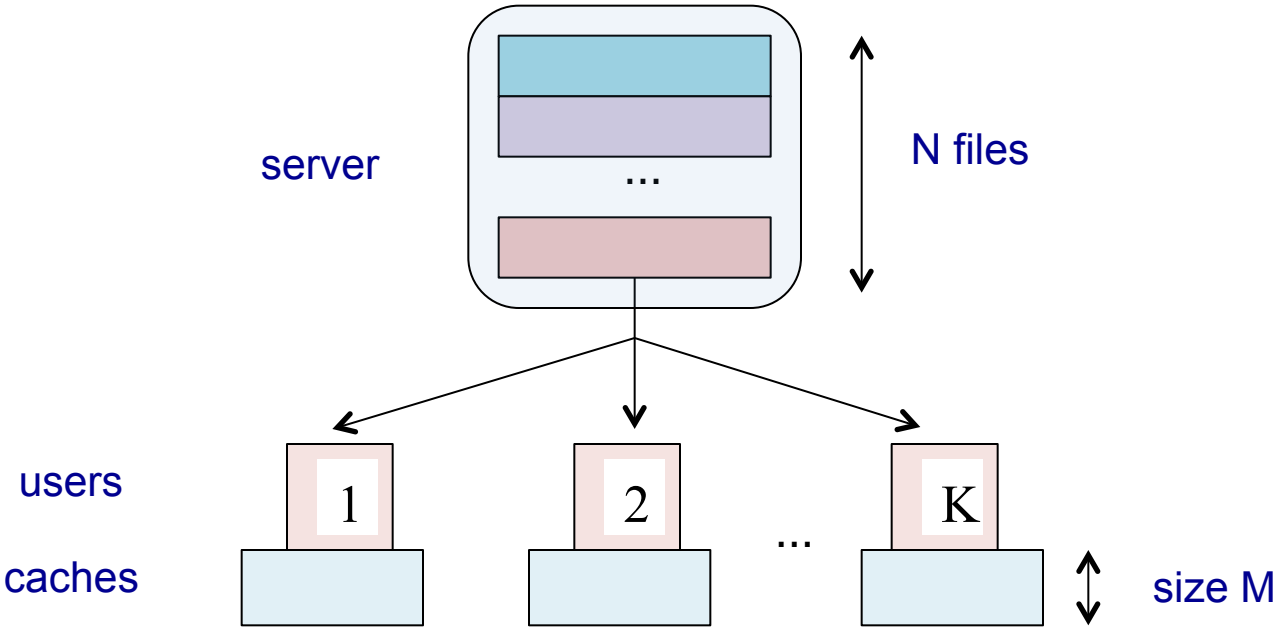


- Placement phase: caches are filled as a function of the database
- Delivery phase: each user may ask for any one of the N possible files

- Objective: design placement and delivery phases so that the traffic load of the shared link in the delivery phase is minimized

The 3 Primary C's of 5G

Fundamental trade-off between communication & caching



Load of shared link using **uncoded** caching: $K(1-M/N)$

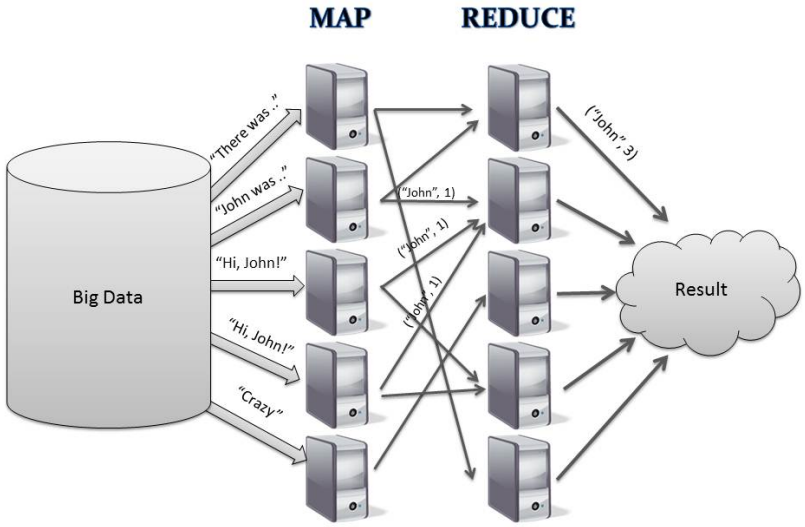
Load of shared link using **coded** caching: $K(1-M/N)\dfrac{1}{1+KM/N}$

MA Maddah-Ali, U Niesen, "Coding for caching: fundamental limits and practical challenges", IEEE Comm. Mag, 2016

The 3 Primary C's of 5G

Fundamental trade-off between communication & computation

Common architecture: MapReduce

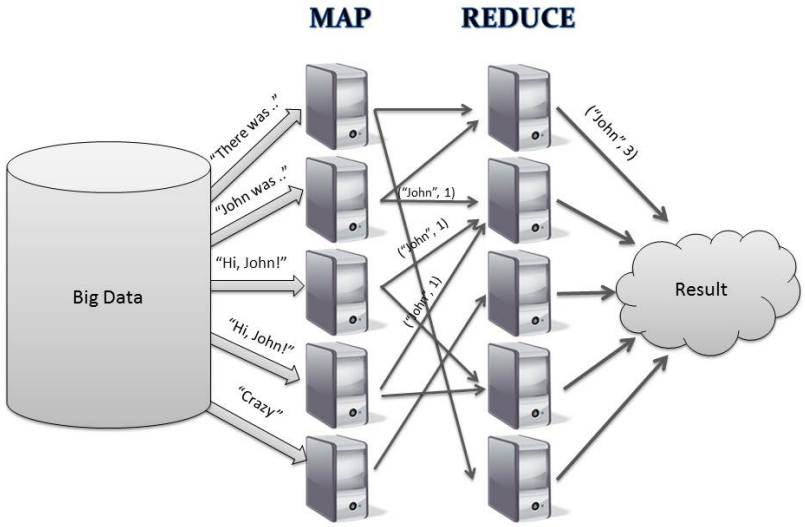


Question: Can coding help distributed computing in reducing the load of communication and speeding up the overall computation?

The 3 Primary C's of 5G

Fundamental trade-off between communication & computation

Common architecture: MapReduce



Answer: Coded Distributed Computing ⇨ $L(r) = \frac{1}{r}\left(1 - \frac{r}{K}\right)$

where K is the # of computing nodes and r is the computation load

S. Li, M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr, " A Fundamental Tradeoff between Computation and Communication in Distributed Computing", ISIT 2016

Computation offloading

Advantages offered by computation offloading using MEC

- Empower simple devices (e.g., tiny sensors in IoT) with augmented computational capabilities
- Prolong battery lifetime of mobile devices
- Reduce end-to-end latency associated to sophisticated applications
- Facilitate latency control with respect to MCC over wide area networks
- Enable RAN-aware content delivery

Computation offloading: Single-user case

Application parameters:

1. n_k = number of input bits transmitted by user k to offload program execution to the cloud
2. w_k = number of CPU cycles necessary to run user k application
3. f_k = number of CPU cycles/sec assigned to virtual machine associated to user k

Application classification features:

- Computational load vs. nr. of bits to be transferred

Computation offloading: Single-user case

Energy minimization under latency constraint

Overall latency: $\Delta = \Delta_T + \Delta_{exe} + \Delta_R$

where

1. Δ_T = time necessary to transfer the input bits b from MU to the serving SCeNB
2. Δ_{exe} = time for the server to run the application
3. Δ_R = time necessary to receive the result back

Note: Overall latency couples radio access and computational aspects

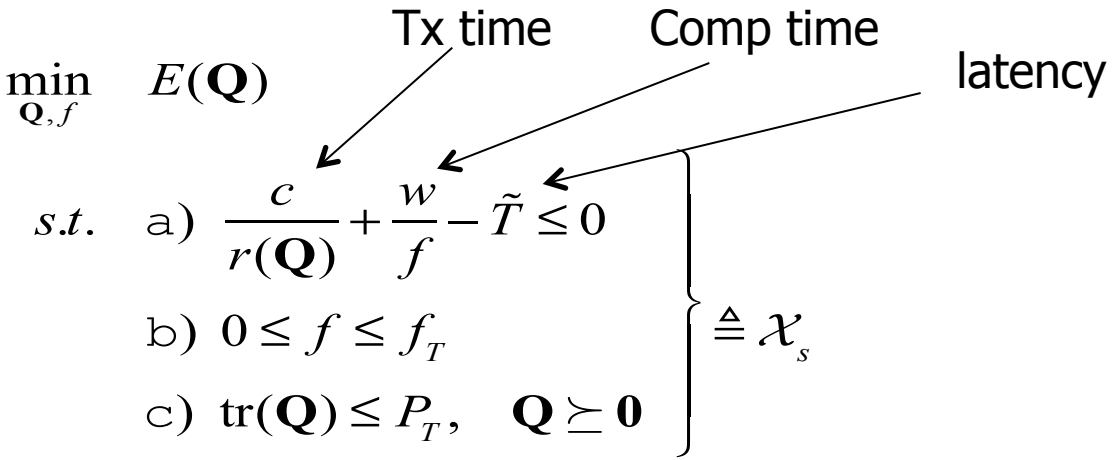
→ Joint optimization of radio / computational resources

S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Energy minimization under latency constraint

Problem formulation:

Find precoding matrices for MIMO transceivers and computing rate minimizing energy consumption @ mobile site, under latency constraint



Note: Exploit Adaptive Coding and Modulation (ACM) to adapt transmit rate

Remark: This problem is non-convex

S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Computation offloading: Single-user case

Energy minimization under latency constraint

Results:

1) Problem can be converted into a convex problem

2) Solution is found in closed form:

$$f^* = f_S \qquad \mathbf{Q}^* = \mathbf{U}(\alpha\mathbf{I} - \mathbf{D}^{-1})^+\mathbf{U}^H$$

where $\mathbf{H}^H\mathbf{R}_w^{-1}\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{U}^H$ and $\alpha > 0$ is chosen so that latency constraint is satisfied with equality

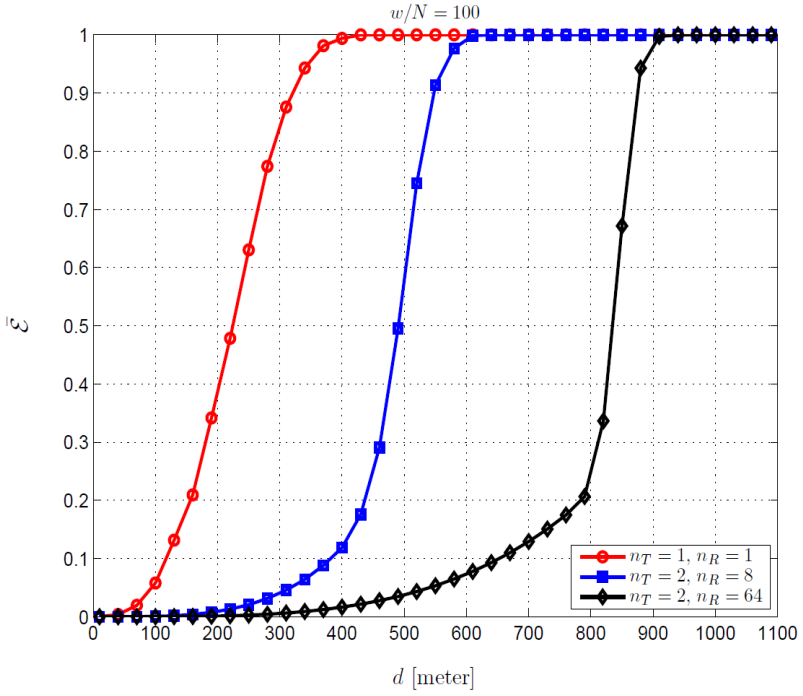
Note: Solution takes the well known water-filling form, where water-level depends on computational parameters and latency constraint

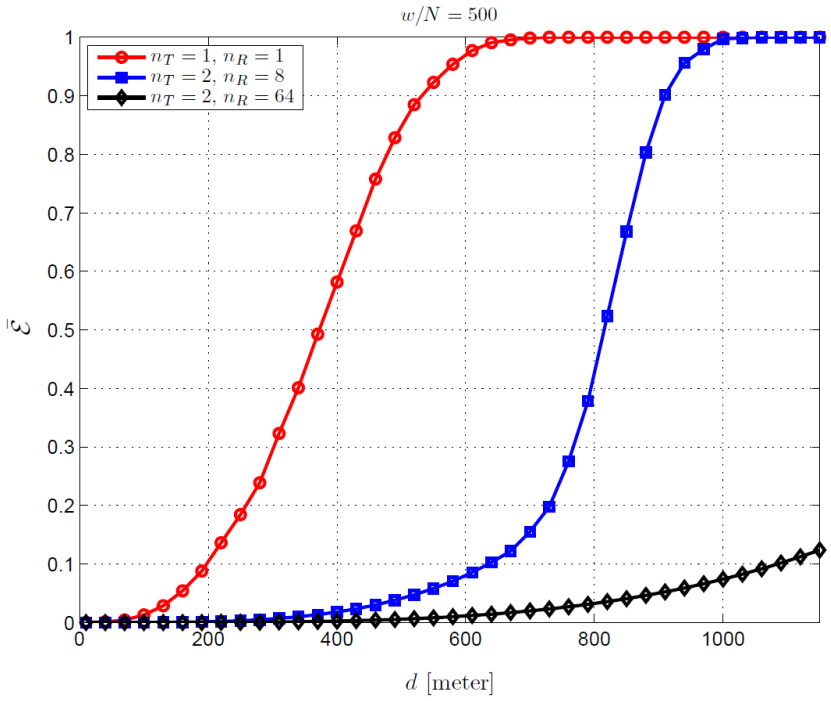
S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Energy minimization under latency constraint

Numerical results: normalized average energy vs. MU – SCeNB distance

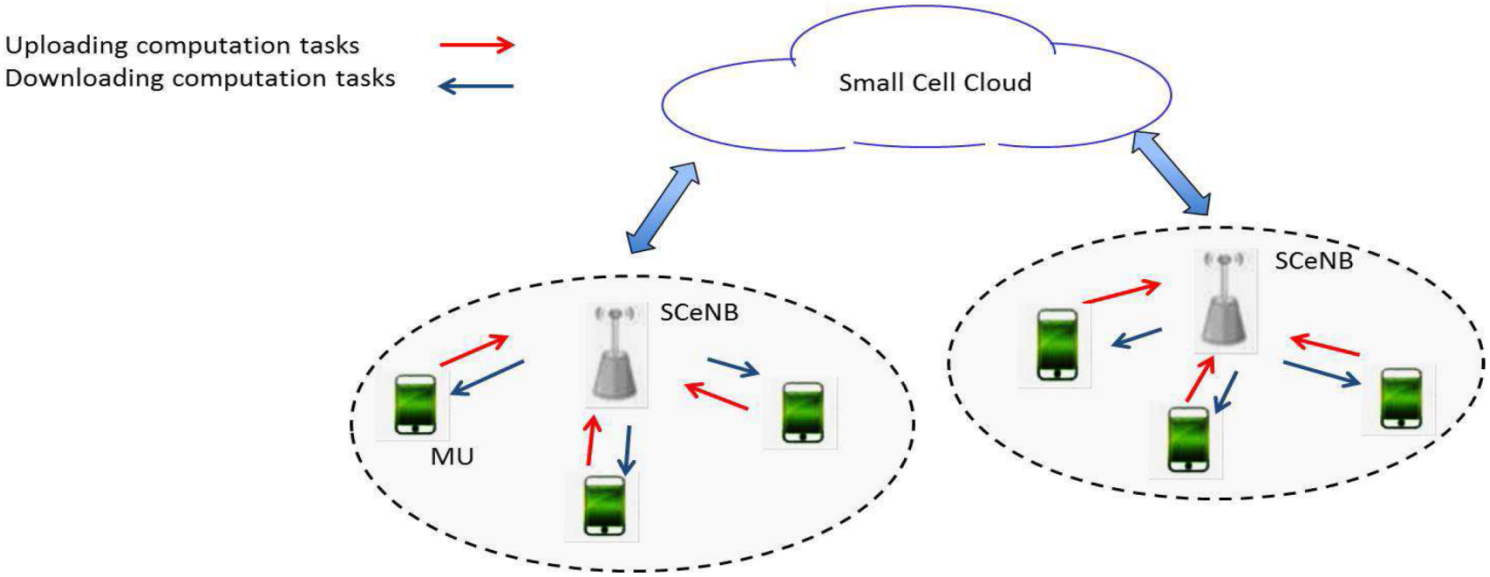
Computationally demanding applications





Single cloud serving multiple cells

MIMO multicell network



Goal: Joint allocation of computational and communication resources to minimize the transmit energy consumption under latency constraint in a dense deployment scenario

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Single cloud serving multiple cells

Degrees of freedom:

- Radio resources: precoding matrices or, equivalently, covariance matrices $\mathbf{Q} \triangleq (\mathbf{Q}_{i_n})_{i_n \in \mathcal{I}}$
- Computing resources: percentage of CPU cycles assigned to each user (VM) f_{i_n}

Constraints:

- Transmit power budget
- Computational capacity:

$$\sum_{i_n \in \mathcal{I}} f_{i_n} \leq f_S$$

Critical issues: Inter-cell interference, admission control incorporating QoE

Single cloud serving multiple cells

Problem formulation:

Find optimal $\mathbf{Q} \triangleq (\mathbf{Q}_{i_n})_{i_n \in \mathcal{I}}$ and $\boldsymbol{f} \triangleq (f_{i_n})_{i_n \in \mathcal{I}}$ minimizing overall energy spent by MUs

Tx time

Comp time

$$\min_{\mathbf{Q},\mathbf{f}} \; E(\mathbf{Q}) \triangleq \sum_{n,i} E_{i_n}(\mathbf{Q}_{i_n}, \mathbf{Q}_{-n})$$

$$\text{s.t.} \quad \text{a)} \; g_{i_n}(\mathbf{Q}, f_{i_n}) \triangleq \frac{c_{i_n}}{r_{i_n}(\mathbf{Q}_{i_n}, \mathbf{Q}_{-n})} + \frac{\omega_{i_n}}{f_{i_n}} - \tilde{T}_{i_n} \leq 0, \; \forall i_n \in \mathcal{I},$$

users' latency constraints

$$\text{b)} \sum_{i_n \in \mathcal{I}} f_{i_n} \leq f_T \quad \text{and} \quad f_{i_n} \geq 0, \quad \forall i_n \in \mathcal{I},$$

computing rate constraint

$$\text{c)} \; \mathbf{Q}_{i_n} \in \mathcal{Q}_{i_n}, \quad \forall i_n \in \mathcal{I},$$

Tx power constraint

where $r_{i_n}(\mathbf{Q}) = \log_2 \det \left(\mathbf{I} + \mathbf{H}_{i_n n}^H \mathbf{R}_{i_n}(\mathbf{Q}_{-n})^{-1} \mathbf{H}_{i_n n} \mathbf{Q}_{i_n} \right)$

$$\mathbf{R}_{i_n}(\mathbf{Q}_{-n}) \triangleq \mathbf{R}_n + \sum_{n \neq m = 1}^{N_c} \sum_{j=1}^{K_m} \mathbf{H}_{j_m n} \mathbf{Q}_{j_m} \mathbf{H}_{j_m n}^H$$

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Single cloud serving multiple cells

Challenges: Non-convex problem / admission control

- Numerical methods may be slow to converge and unreliable

- Depending on channel conditions, not all applications can be offloaded while respecting all constraints (power and latency)

Approach:

Step # 1: Admission control is achieved as solution of feasibility conditions:

$$\tilde{T}_{i_n} > \frac{c_{i_n}}{r_{i_n}(\overline{\mathbf{Q}})}, \forall i_n \in \mathcal{I}, \quad \text{and} \quad \sum_{i_n \in \mathcal{I}} \frac{\omega_{i_n}}{\tilde{T}_{i_n} - \frac{c_{i_n}}{r_{i_n}(\overline{\mathbf{Q}})}} \leq f_T$$

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Single cloud serving multiple cells

Challenges: Non-convex problem / admission control

Step # 2: Use Successive Convex Approximation (SCA) as a reliable and fast solution method

Original non-convex problem is replaced by a sequence of strongly convex problems exploiting the structure of the problem

Main idea: At each iterate, approximate the original nonconvex nonseparable objective function $E(\mathbf{Q})$ and constraint $g_{i_n}(\mathbf{Q}, f_{i_n})$ around the current iterate with a strongly convex function

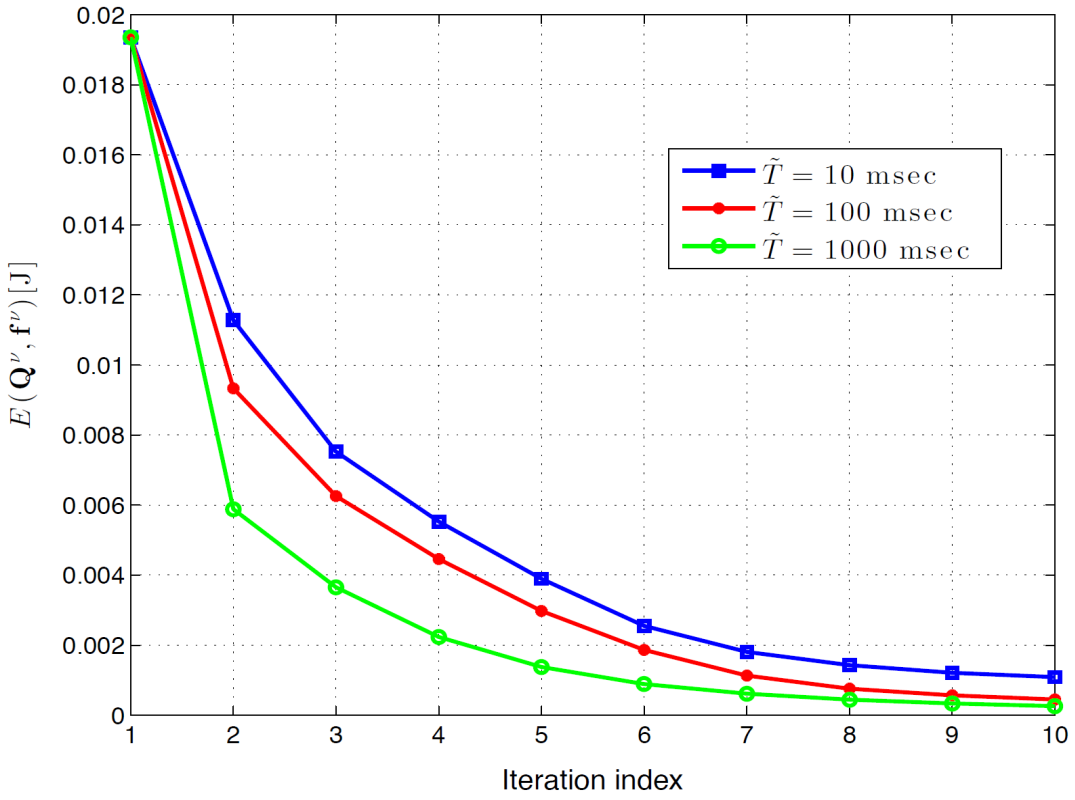
Remark: Different approximation choices (albeit appropriate) are available and lead to:

- Centralized algorithms
- Distributed algorithms with limited signaling overhead

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Single cloud serving multiple cells

Numerical results: Sum energy vs. iteration number

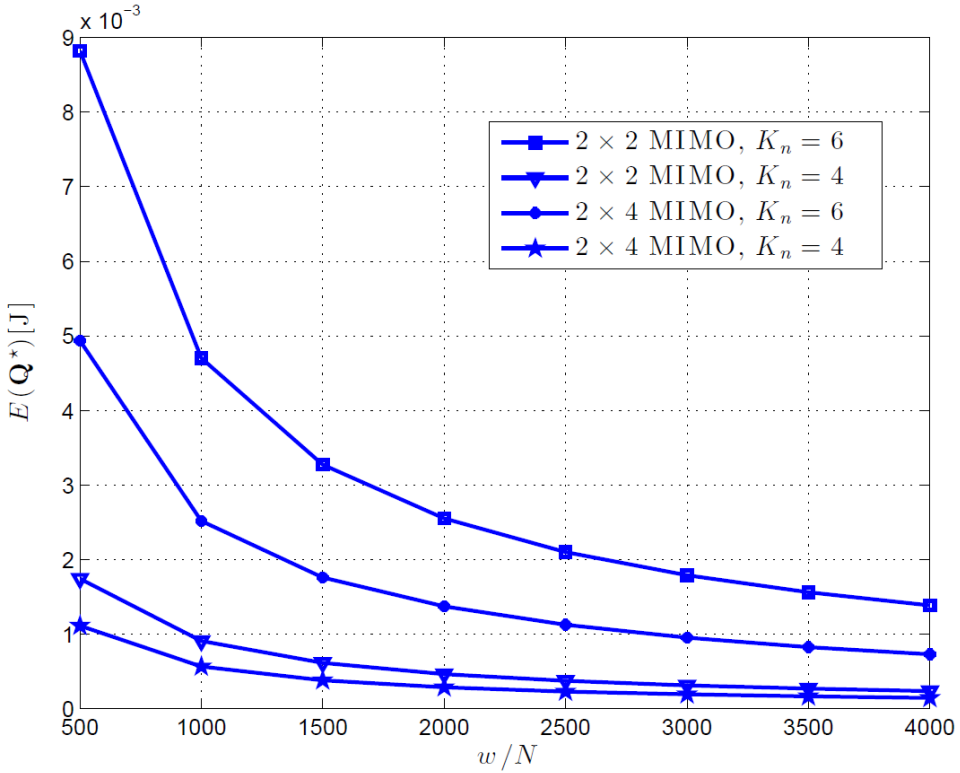


Proposed SCA algorithm converges in a very few iterations

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Single cloud serving multiple cells

Numerical results: Sum energy vs. computational load

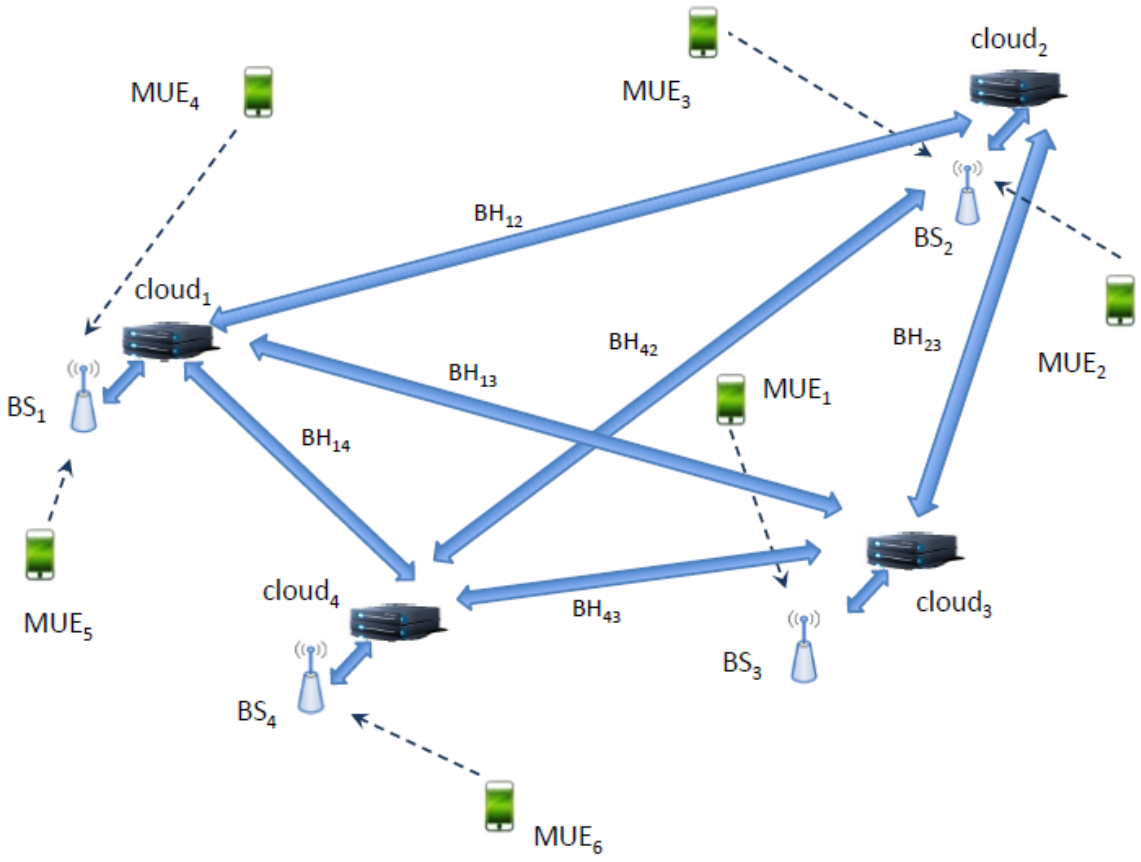


CPU cycles per Tx bits

S. Sardellitti, G. Scutari, S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing", IEEE Trans. on Signal and Information Processing over Networks, June 2015

Multiple clouds serving multiple cells

Goal: Joint optimization of radio/computational resources and mobile user-base station-cloud assignment



S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Multiple clouds serving multiple cells

Goal: Joint optimization of radio/computational resources and mobile user-base station-cloud assignment

Degrees of freedom:

- Precoding matrix $\mathbf{F}_k$ or, equivalently, covariance matrix $\mathbf{Q}_k$ of each user k

- Number f_{mk} of CPU cycles/sec assigned to k-th user virtual machine on the m-th cloud

- Assignment of each user to a base station and to a cloud: $a_{knm} \in \{0,1\}$

Each mobile user is served by a single base station and a single cloud, i.e.

$$\sum_{n=1}^{N_b} \sum_{m=1}^{N_c} a_{knm} = 1$$

but the method can be extended

S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Multiple clouds serving multiple cells

Problem formulation:

$$\min_{\mathbf{Q},\mathbf{f},\mathbf{a}} E(\mathbf{Q},\mathbf{a}) \triangleq \sum_{k=1}^{K} c_k E_k(\mathbf{Q},\mathbf{a}_k) \qquad (\mathrm{P})$$

weighted sum energy

subject to

i) $g_{knm}(\mathbf{Q}, f_{mk}, a_{knm}) \le L_k, \forall\, k,n,m$ — latency constraint

ii) $\mathrm{tr}(\mathbf{Q}_k) \le P_k, \quad \mathbf{Q}_k \succeq \mathbf{0}, \forall\, k \in \mathcal{I}$ — Tx power constraint

iii) $\mathbf{f} \ge \mathbf{0}, \quad \sum_{k=1}^{K}\sum_{n=1}^{N_b} a_{knm} f_{mk} \le F_m, \quad \forall\, m$ — computing constraint

iv) $\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} a_{knm} = 1,\ a_{knm} \in \{0,1\}, \quad \forall\, k,n,m$ — association constraint

where

$$g_{knm}(\mathbf{Q}, f_{mk}, a_{knm}) \triangleq a_{knm}\left(\frac{c_k}{r_{kn}(\mathbf{Q})} + \frac{w_k}{f_{mk}} + T_{Bmn}\right)$$

$$r_{kn}(\mathbf{Q}) = \log_2 \det\left(\mathbf{I} + \mathbf{H}_{kn}^H \mathbf{R}_{kn}(\mathbf{Q}_{-k})^{-1}\mathbf{Q}_k\mathbf{H}_{kn}\right)$$

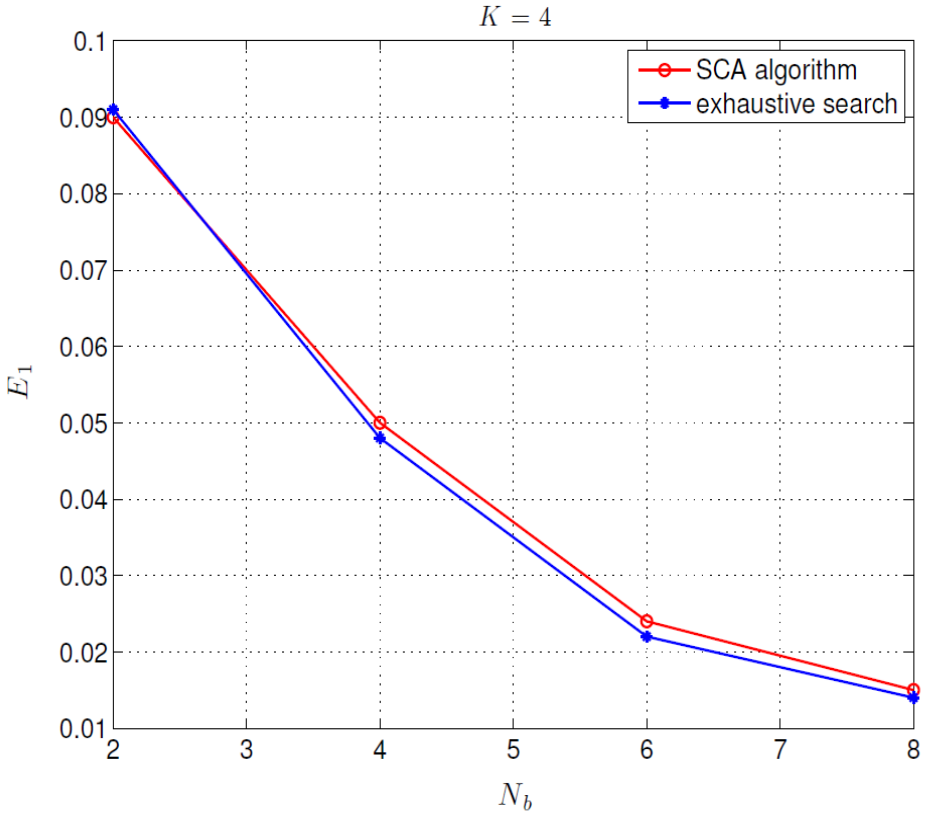
S. Barbarossa, S. Sardellitti, P. Di Lorenzo, "Communicating while Computing: Distributed Cloud Computing over 5G Heterogeneous Networks", IEEE Signal Processing Magazine, Nov. 2014

Multiple clouds serving multiple cells

Proposed solution: Successive Convex Approximation (SCA)

Comparison between relaxed problem and exhaustive search

Average energy consumption vs. number of base stations N_b



Note:

- Losses wrt exhaustive search are very small

- Dense deployment yields a considerable energy saving

Merge mmWaves and MEC

Shortcomings:

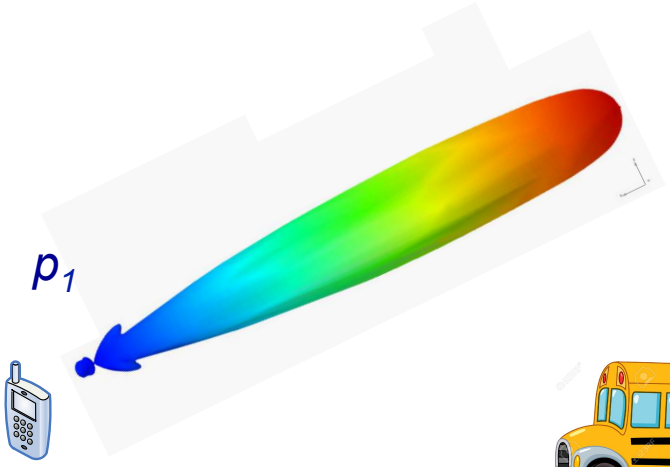
- Channel intermittency due to obstacles
- Interference from users transmitting from similar angles (e.g., canyon crowd)
- Large attenuation over long distance links

Proposed approach

- Overbook resources to meet latency constraints on average, depending on probabilities of blocking events
- Use multi-link communications whenever possible

S. Barbarossa, E. Ceci, M. Merluzzi, E. Calvanese-Strinati, "Enabling Effective Mobile Edge Computing Using Millimeter Wave Links", ICC 2017

Example: Single user / Multi-link to 2 radio access points

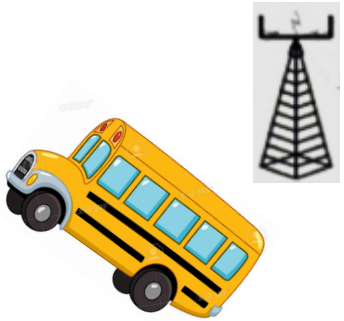


mobile user transmits to RAP_1

with probability $\mathbb{P}_1$

using power p_1

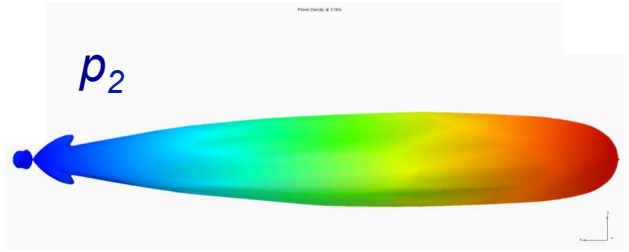
Example: Single user / Multi-link to 2 radio access points



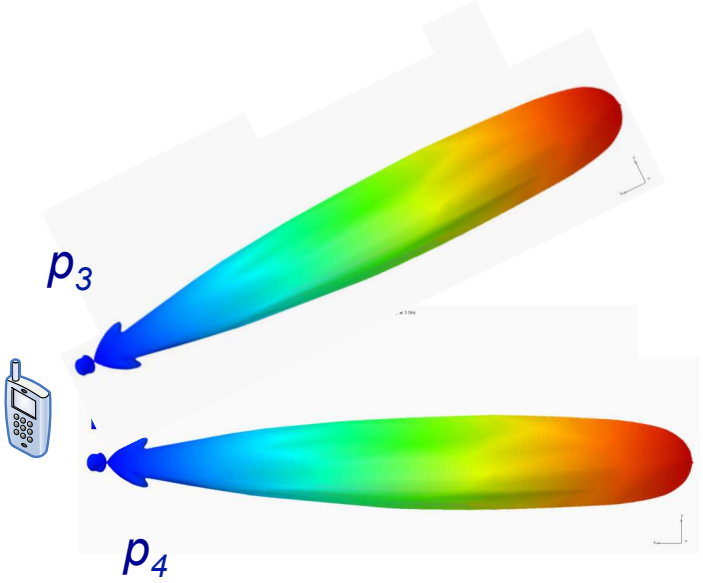
mobile user transmits to RAP_2

with probability $\mathbb{P}_2$

using power p_2



Example: Single user / Multi-link to 2 radio access points



mobile user transmits to both RAP_1 and RAP_2

with probability $\mathbb{P}_3$

using power p_3 and p_4

Problem formulation: min *average* transmit power consumption subject to *average* rate

Example: 2 access points

$$\min_{\boldsymbol{p}} \ \sum_{i=1}^{4} \mathbb{P}_i \, p_i, \ \text{s.}t.$$
$$i) \ \sum_{i=1}^{4} \mathbb{P}_i \log(1 + a_i p_i) \geq \bar{R}_{min}$$
$$ii) \ p_i \geq 0, i = 1, \ldots, 4;$$
$$iii) \ p_i \leq P_T, i = 1, 2; p_3 + p_4 \leq P_T,$$

where $\mathbb{P}_i$ is the probability of having link *i* on

Solution

$$p_i = \left[\beta - \frac{1}{a_i}\right]_0^{P_T}, i = 1, 2 \qquad p_i = \left[\frac{\beta \mathbb{P}_i}{\mathbb{P}_i + \nu_3} - \frac{1}{a_i}\right]_+, \ i = 3, 4$$

where

$$\beta = \exp\left(\frac{c}{2 - P_{I_1} - P_{I_2}}\right)$$

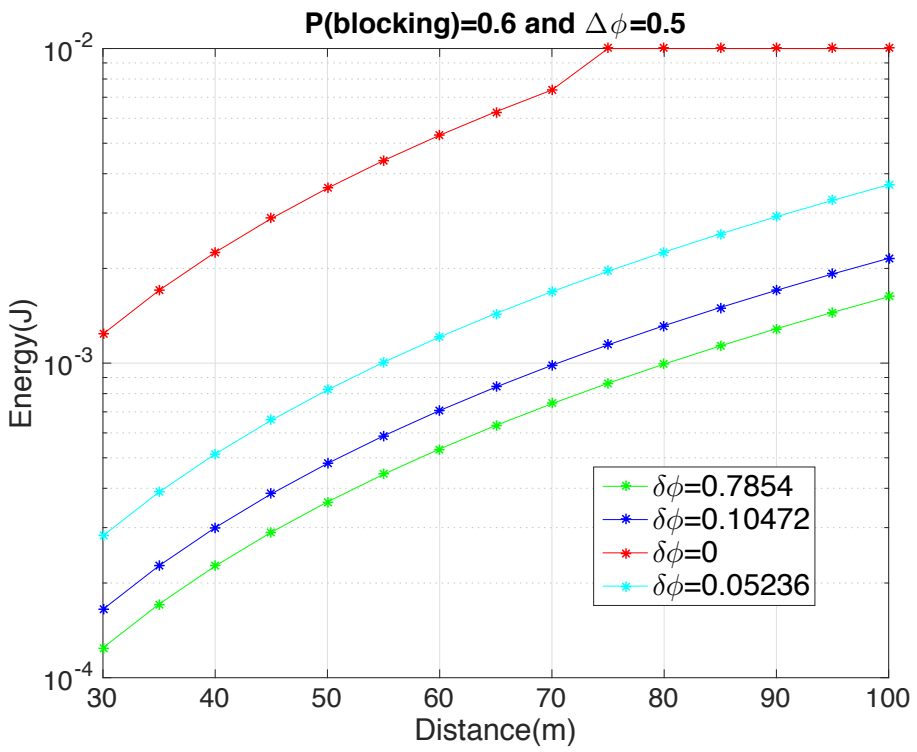
if optimal powers do not reach the power budget

Numerical results: Statistically dependent blocking events

Average energy consumption vs. distance

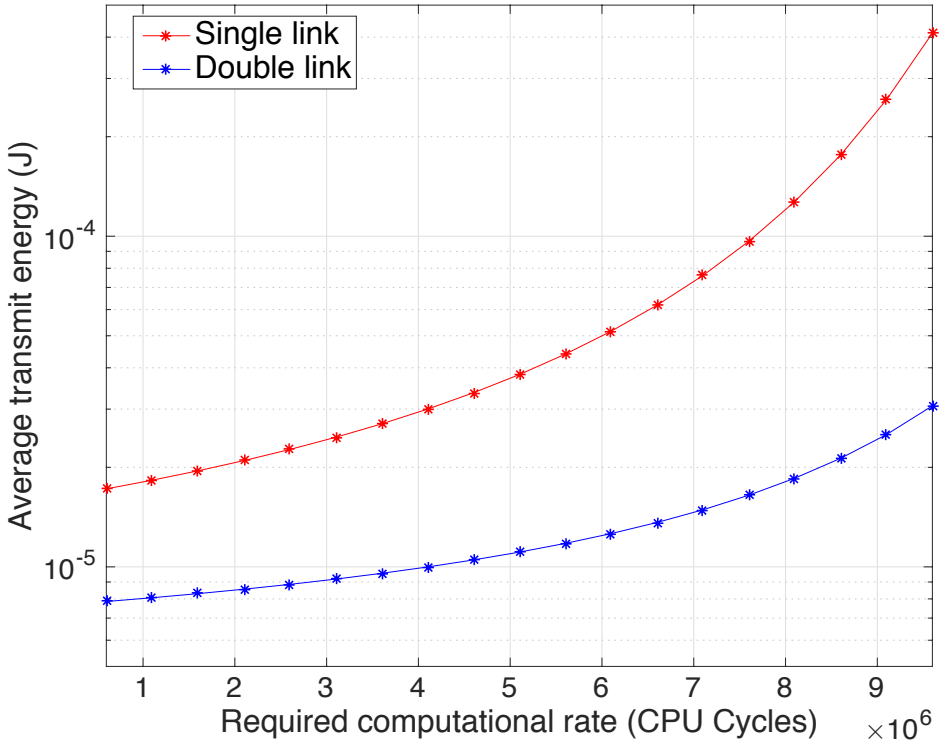
- RAP angular spread $\delta\phi$

- obstacle shadowing angle $\Delta\phi$



S. Barbarossa, E. Ceci, M. Merluzzi, "Overbooking Radio and Computation Resources in mmW-Mobile Edge Computing to Reduce Vulnerability to Channel Intermittency", EUCNC 2017

Numerical results: Average energy vs. computational load (independent blocking events)



S. Barbarossa, E. Ceci, M. Merluzzi, E. Calvanese-Strinati, "Enabling Effective Mobile Edge Computing Using Millimeter Wave Links", ICC 2017

Multi-user / Multilink

Problem formulation

$$\min_{\boldsymbol{R},\boldsymbol{f}} \sum_{k=1}^{K}(1-P_{I_k})\,\frac{1}{a_k}\,\left(2^{R_k}-1\right)$$

min average power

$$\text{s.t. } \frac{c_k}{(1-P_{I_k})\,R_k}+\frac{w_k}{f_k}\leq \Delta_k,\ \forall k$$

latency constraint

$$R_k \geq 0,\ \forall k$$

non-negative rate

$$R_k \leq \log_2(1+a_k P_T),\ \forall k$$

lower rate limit

$$\sum_{k=1}^{K} f_k \leq f_S;$$

overall computing rate

$$0 \leq f_k \leq f_S,\ \forall k$$

individual rate limits

Multi-user / Multilink

Solution:

Optimal computation rate: $f_k = \dfrac{\sqrt{w_k\,\gamma_k}}{\sum_{k=1}^{K}\sqrt{w_k\gamma_k}}\,f_S$

Optimal communication rate: $R_k = \dfrac{c_k}{(1-P_{I_k})\left(\Delta_k - \sqrt{\frac{w_k}{\gamma_k}}\sum_{k=1}^{K}\sqrt{w_k\gamma_k}/f_S\right)}$

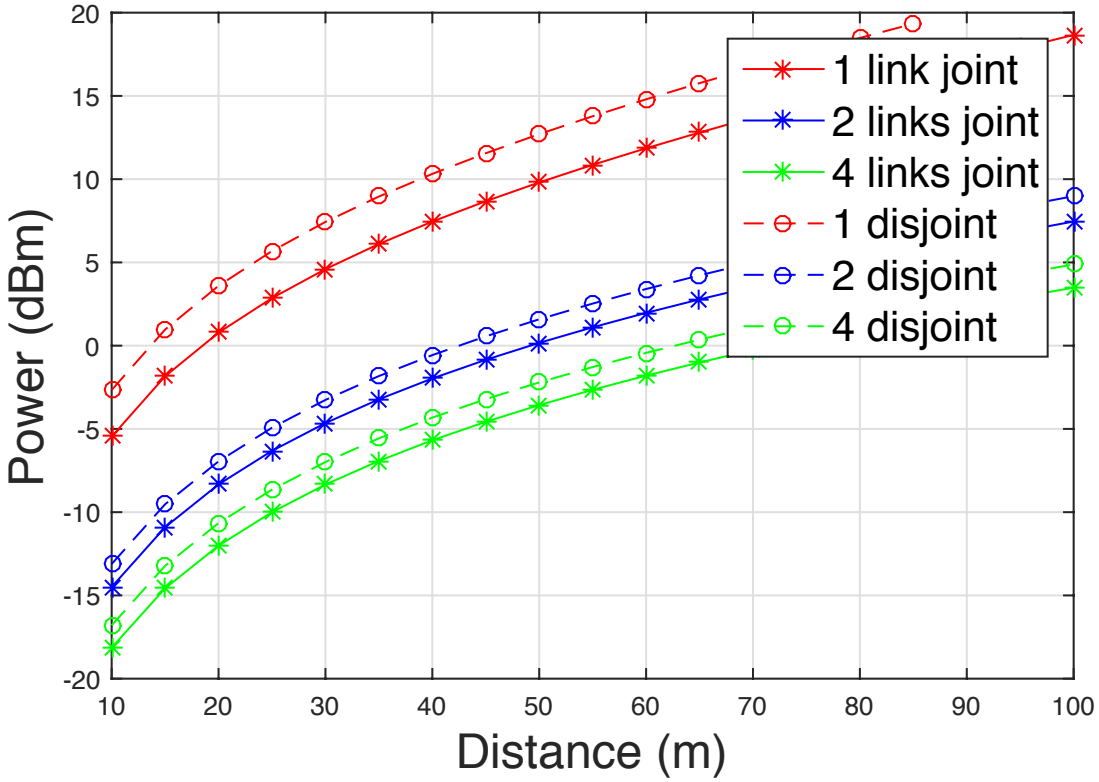
Note: Optimal computation rate does not coincide with proportional allocation

$$f_k = \frac{w_k}{\sum_{k=1}^{K} w_k}\,f_S$$

S. Barbarossa, E. Ceci, M. Merluzzi, "Overbooking Radio and Computation Resources in mmW-Mobile Edge Computing to Reduce Vulnerability to Channel Intermittency", EUCNC 2017

Multi-user / Multilink

Numerical results



S. Barbarossa, E. Ceci, M. Merluzzi, "Overbooking Radio and Computation Resources in mmW-Mobile Edge Computing to Reduce Vulnerability to Channel Intermittency", EUCNC 2017

Conclusions

- Mobile edge-computing enables joint optimization of radio and computing resources, depending on channel state, backhaul state, and application parameters
- Optimal association of mobile users to base stations and clouds provides a mechanism for optimal instantiation of virtual machines and represents a new way to handle handover
- Non-convex offloading problems have been efficiently solved by using a new class of SCA-based algorithms converging to stationary solutions of the original problem
- mmW channel intermittency can be counteracted by resource overbooking based on knowledge (estimation) of blocking probability
- Incorporation of online distributed learning mechanisms in mobile devices (task profilers) as well as in the edge of the network
- Extension to coded caching & coded computing

The research leading to these results are jointly funded by the European Commission (EC) H2020 and the Ministry of Internal affairs and Communications (MIC) in Japan under grant agreements N° 723171 5G MiEdge in EC and 0159-{0149, 0150, 0151} in MIC