Offloading on the Edge

Flow-level Performance Modeling and Optimization for Mobile Data Offloading

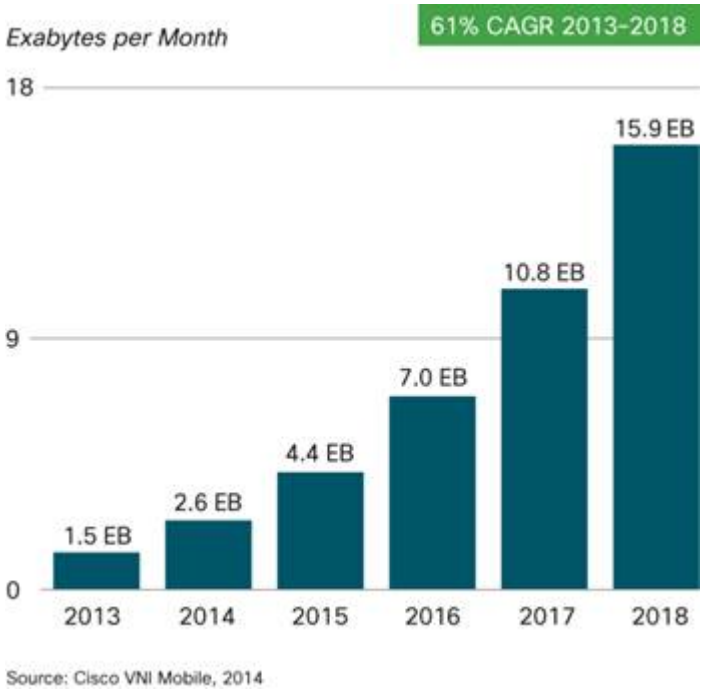
Thrasyvoulos Spyropoulos

(joint work with P. Sermpezis, F. Mehmeti, L. Vigneri, Delia Ciullo, Navid Nikaein)

Mobile Communications Department

Eurecom – Sophia Antipolis, France

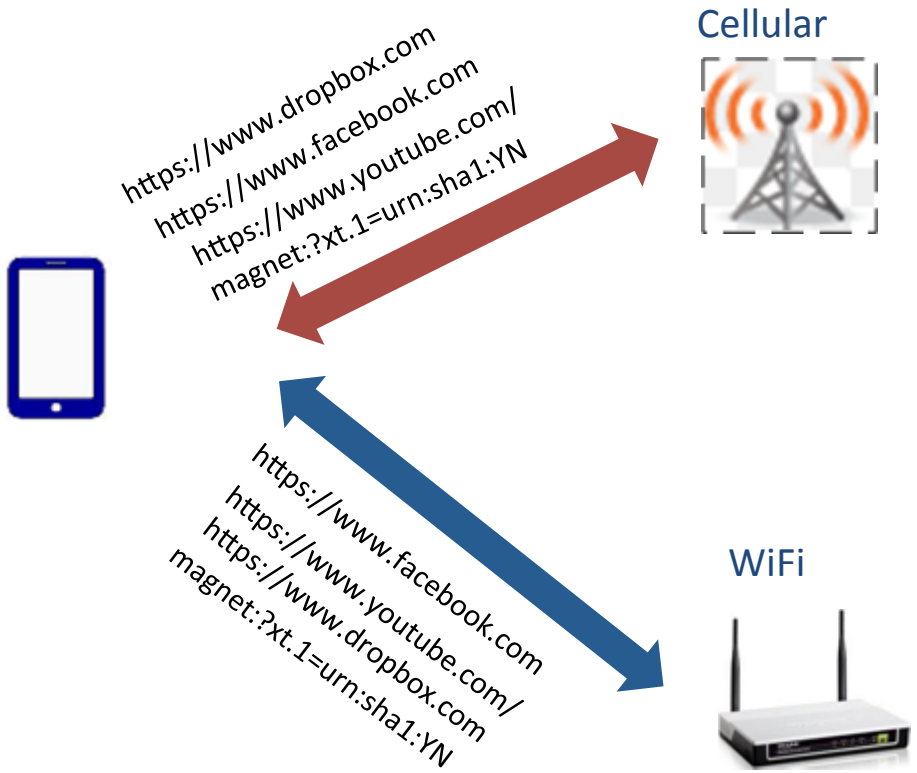
Why Offload?



Source: Cisco VNI Mobile, 2014

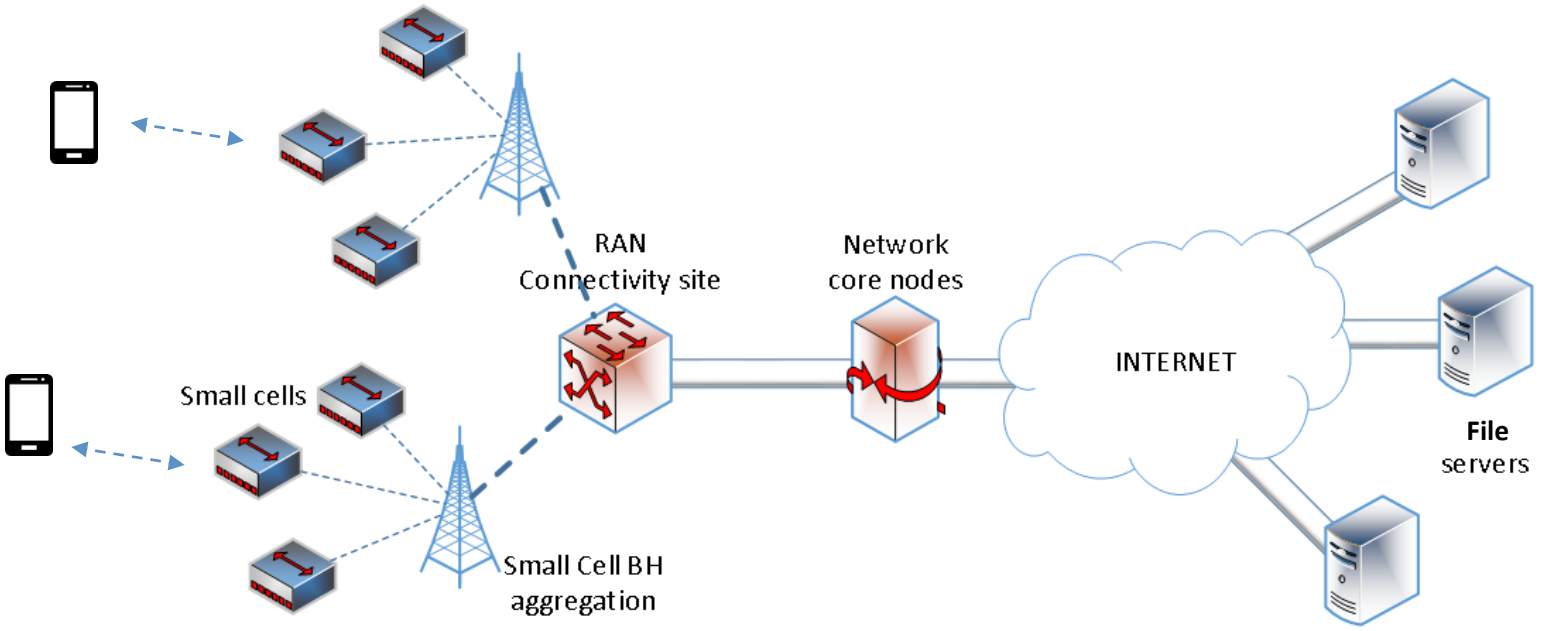
- Radio Access Improvements (LTE/ LTE-A) predicted to be surpassed already by 2016.
- Complete upgrade is costly
- Solution? “Dump” as much data (transmissions) elsewhere

How to Offload? WiFi-based



- ✓ Switch *all* traffic to WiFi → up to 40% offloaded today ☺
- ✓ Hotspot WiFi might have performance issues ☹
- ✓ Sporadic coverage ☹

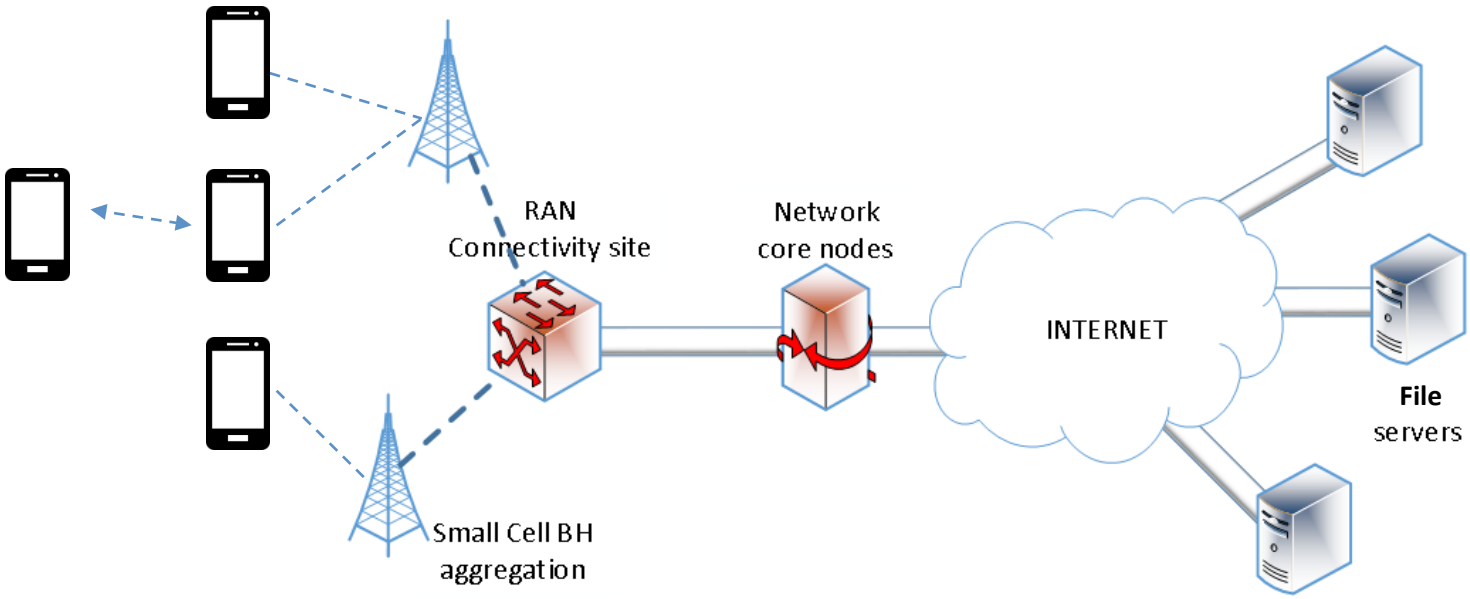
How to Offload? Small Cells



Heterogeneous Cellular Networks (**HetNets**) where small cells overlap with the main macro-net:

- ✓ Micro, pico, femto
- ✓ Requires a large investment ☹
- ✓ Moves the bottleneck to the backhaul ☹

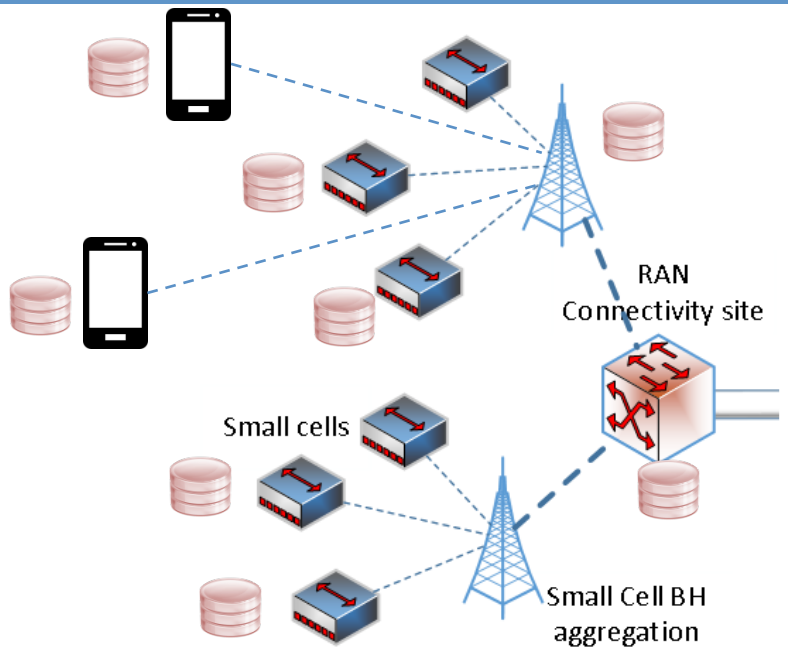
How to offload? D2D



UEs can transmit the content directly to other UEs

- ✓ No extra infrastructure (incentives...?)
- ✓ Shorter Tx distance: power↓ interference↓ capacity↑ ☺
- ✓ If used as relays, backhaul still a problem ☹

“On the Edge” Storage



Cache (popular) contents at the **edge** of the network:

- ✓ **Minimize duplicate transmissions on backhaul links (or even radio links)**

Our Goals

- **Performance Modeling**
  - Which **models**? Queuing Theory + Mean Field Analysis
  - User **metrics**? Flow-level and Content Access perf.
  - Operator metrics? Offloaded volume and cost
- **Optimization**
  - User centric: own cost, energy, etc.
  - Operator centric: total cost, congestion avoidance subject to QoS constraints
- **New Dimension: Delayed (Opportunistic) Access**
  - **Trade off some delay** (to access video, web page, cloud) for performance (user or operator costs).

Outline

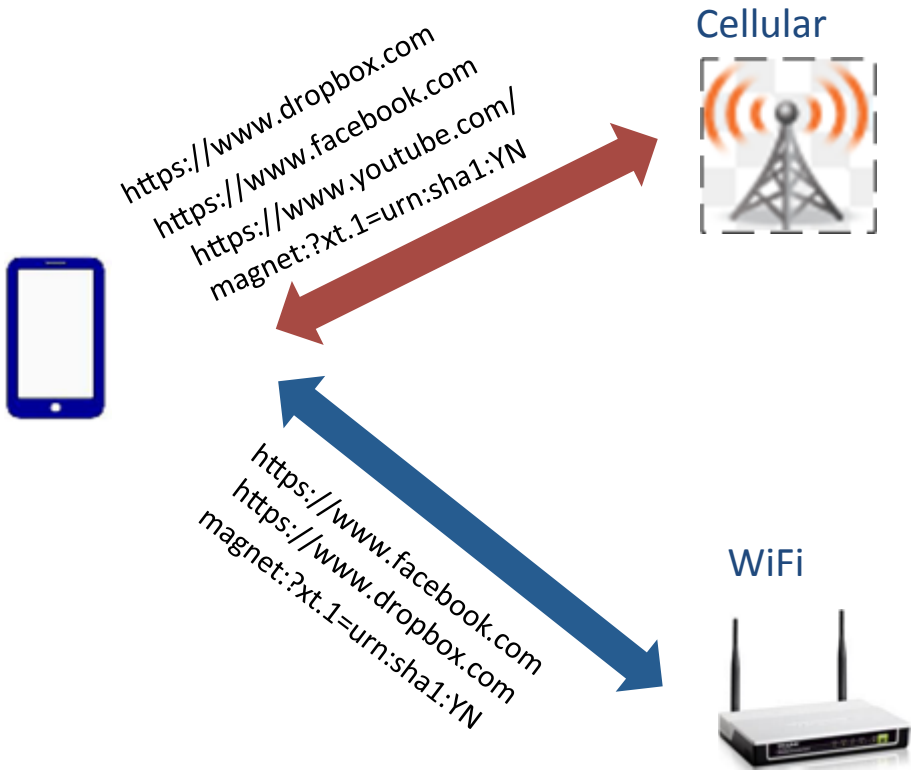
Part I: Opportunistic offloading of **flows** over WiFi

- An analytical model
- Size-based offloading

Part II: **Content storage** and **access** on the edge

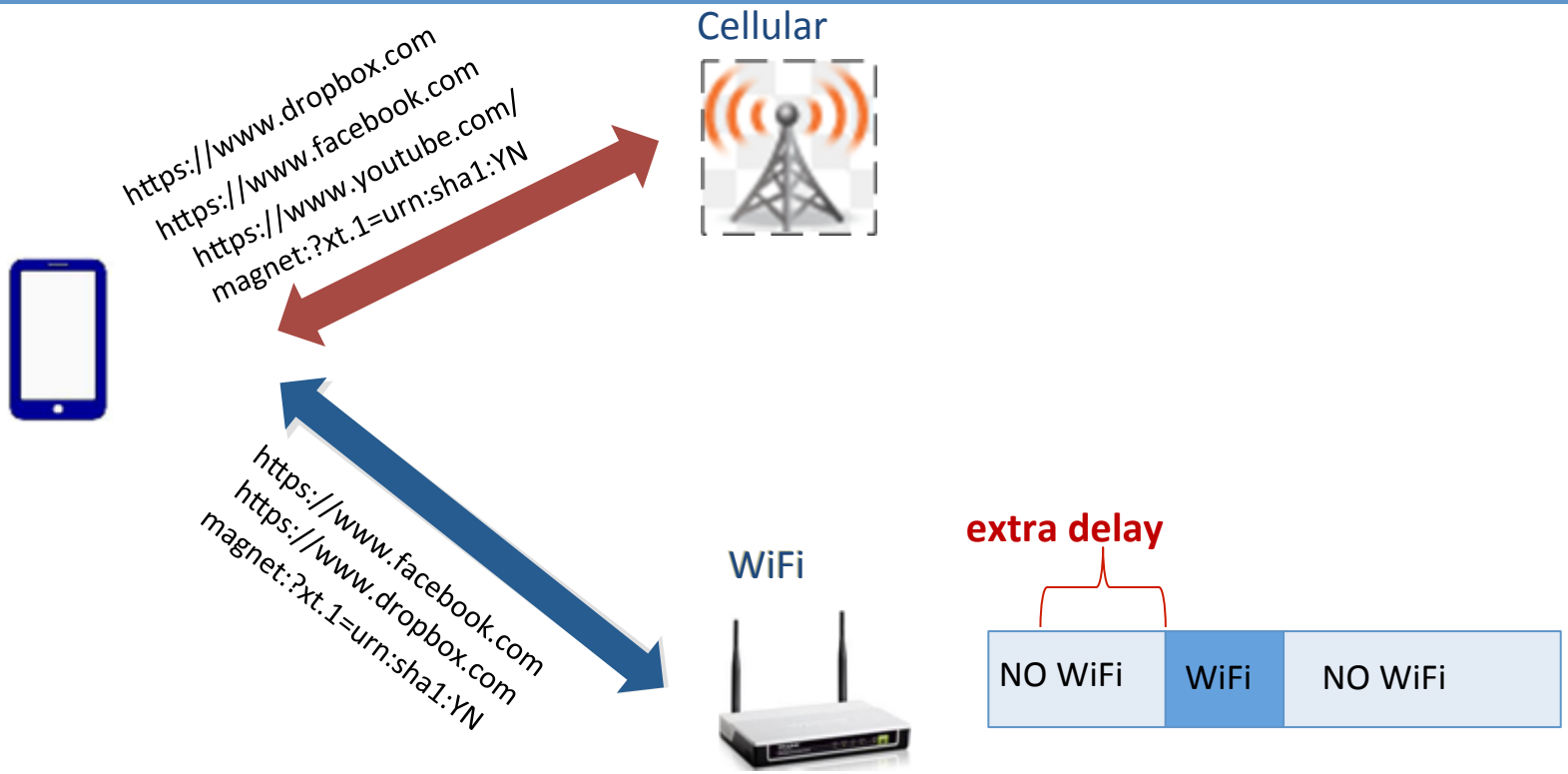
- An analytical model
- Cost-optimal caching strategies

Per Flow ("On the Spot") Offloading



+ Use both interfaces in parallel
+ Optimize which flows to offload (e.g. delay-insensitive)
– Current phones don't allow this (to change soon)

Delayed Offloading: wait for WiFi



DE…LAY???

- Users ARE willing to wait (from minutes to hour(s))
- If there is something to gain (money, energy, …)
- Depends on user, country, application, …

"TUBE: time-dependent pricing for mobile data," ACM Sigcomm 2012

"Practicalizing Delay-Tolerant Mobile Apps with Cedos," ACM MobiSys 2015

Flow Offloading: Key Questions

1. What is the performance of offloading

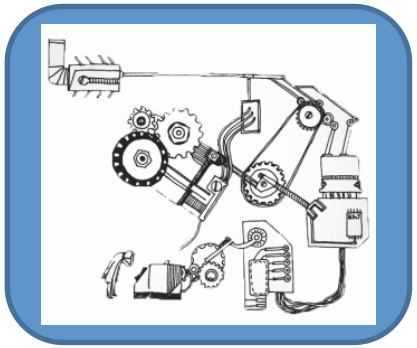
Offloading Policy

+

Network Conditions

- WiFi availability,
- Cellular/WiFi rates,
- Traffic load

model/analysis



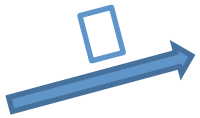
- ✓ # data offloaded
- ✓ average flow delay

2. How to optimize offloading policy

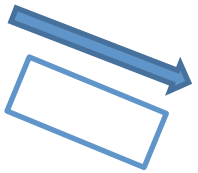
Incoming flows

assign flow to network

e.g. to
minimize **COST**,
while **E[Delay] < D_{MAX}**



Cellular



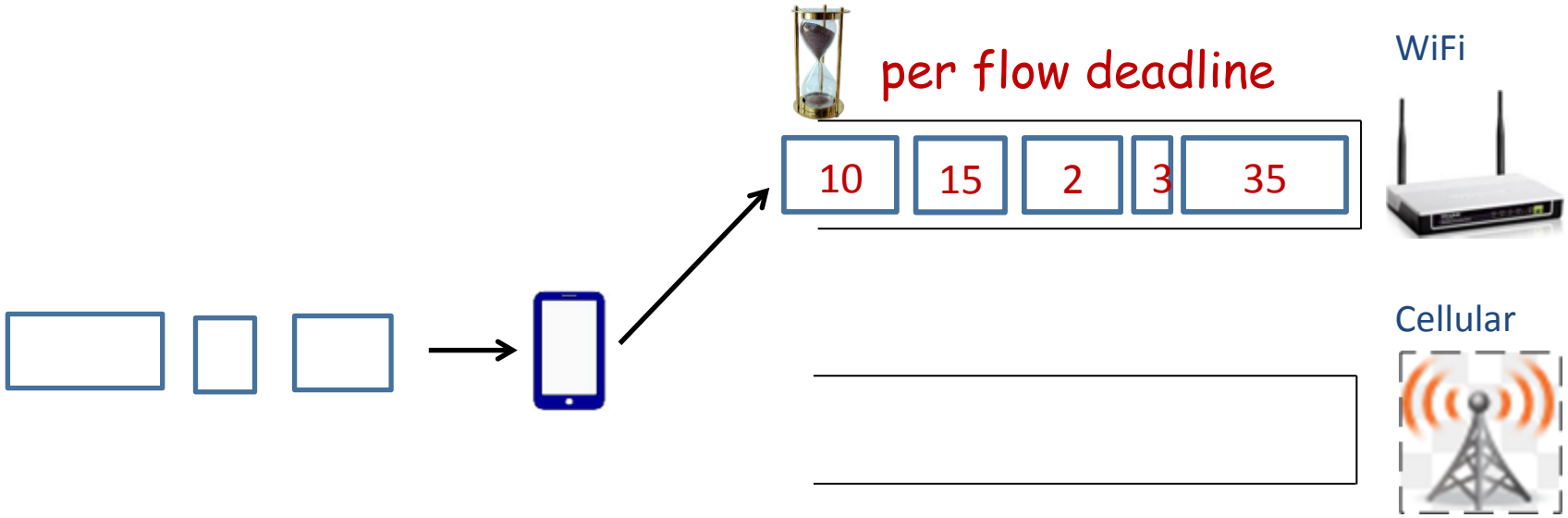
WiFi

Contribution: Analysis of (Delayed) Offloading

A Simple Policy (aggressive offloading)**:

Step 1) Send every data flow to WiFi queue by default

"flow": all packets in the same app request (e.g. file download)



Step 2) When deadline expires → transmit flow on cellular interface

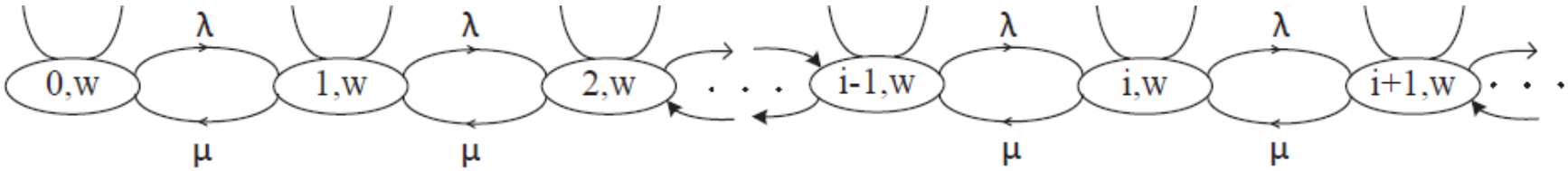
- Deadline only counts when no WiFi connectivity

*** K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: how much can WiFi deliver?" in ACM Conext 2010*
A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in ACM MobiSys 2010

WiFi Queue Model

- Queueing model with:
  - (i) abandonments/reneging
  - (ii) intermittent service

→ # of assumptions (relaxed in sims) → 2D Markov Chain

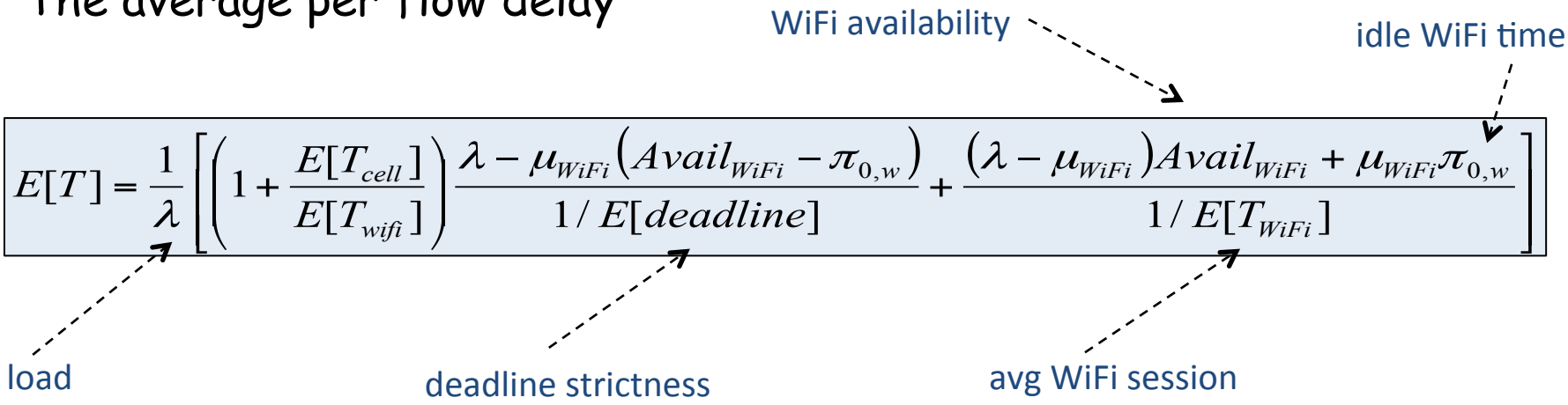


- Usually: Matrix – analytic methods (only numerically ☹)

- Structure ➔ Probability generating function (PGF) method ➔ system of ODE and linear equation ➔ closed form resuls ☺
  - Model valid for FCFS and PS scheduling!

Delayed Offloading Performance Formulas

- The average per flow delay



- The expected amount of offloaded data

$$p_r = 1 - \frac{\lambda - \mu_{WiFi}\left(Avail_{WiFi} - \pi_{0,w}\right)}{\lambda}$$

- F. Mehmeti, T. Spyropoulos, "Performance analysis of on-the-spot mobile data offloading," IEEE Globecom 2013

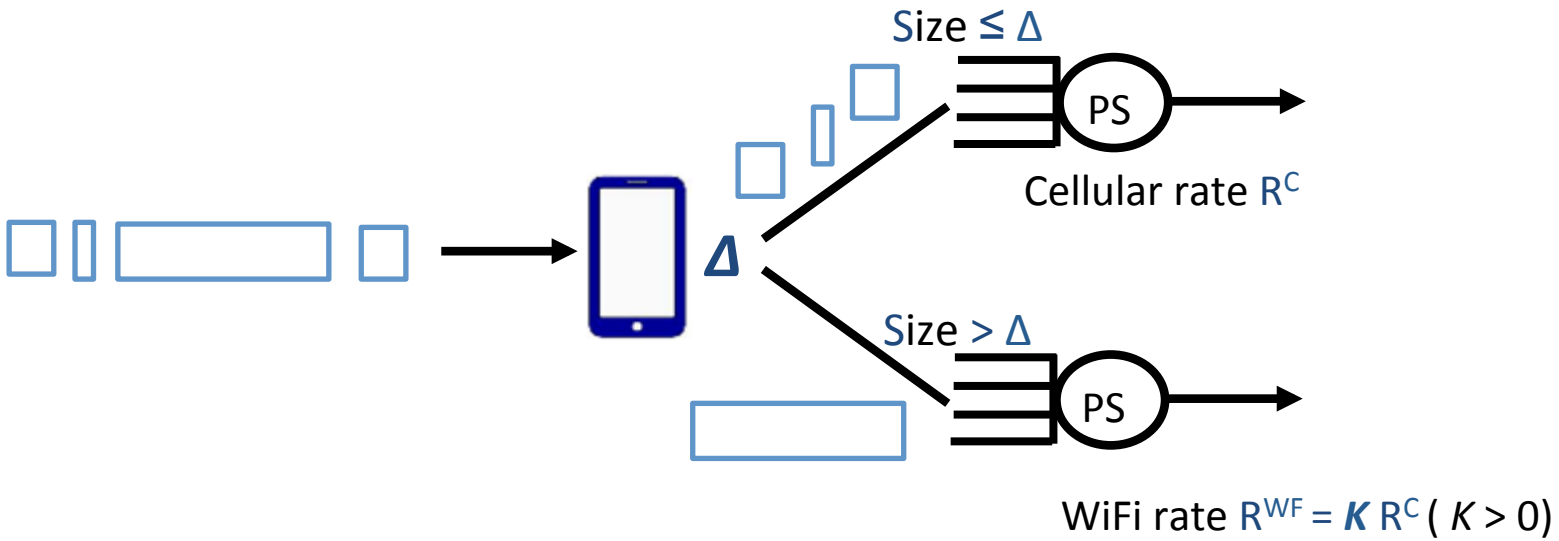
- F. Mehmeti, T. Spyropoulos , "Is it worth to be patient? Analysis and optimization of delayed mobile data offloading," IEEE Infocom 2014

Optimal Flow Assignment Policy**

Goal: Minimize (average) cost per flow, subject to delay constraint

- Processor Sharing (PS) model for queues (more realistic)
- Cost proportional to # of bits (monetary, simple energy model, etc.)

Result 1: **Size-based policy** is optimal

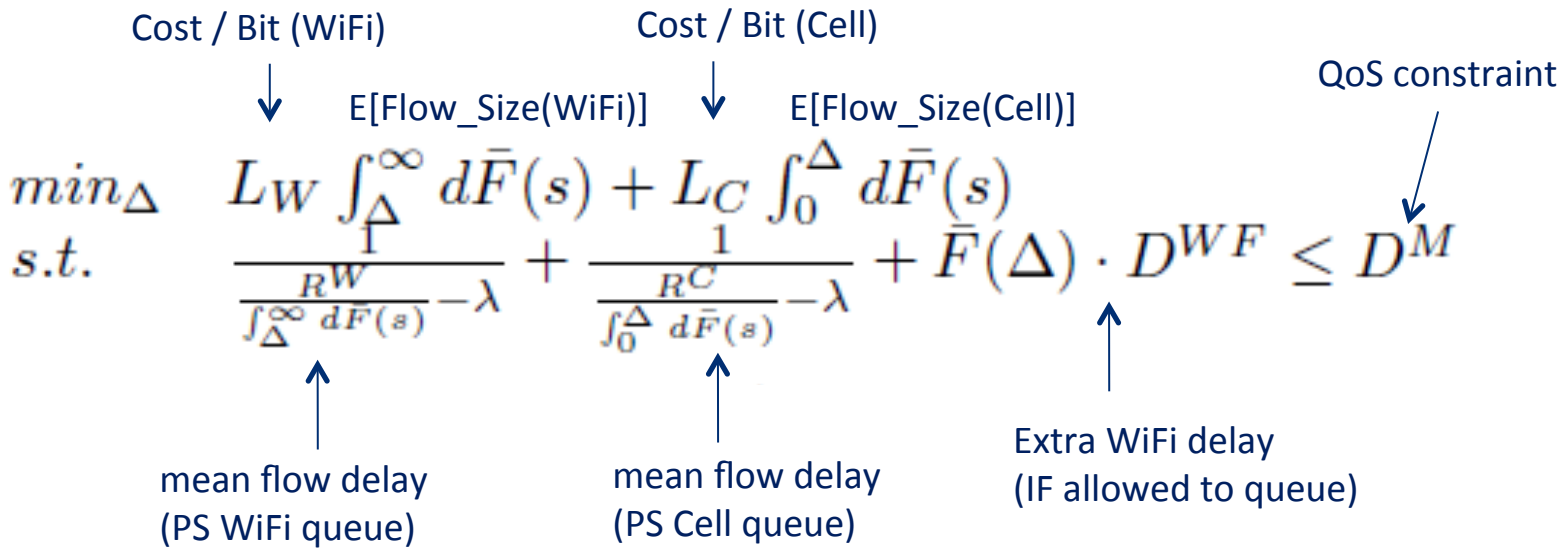


New Goal: Find optimal threshold Δ

***D.Ciullo, T. Spyropoulos, N. Nikaein, B. Jechoux "Sizing Up User Traffic: Smart Flow Assignment for Mobile Data Offloading," Eurecom Tech. Report, patent submitted*

Threshold Policy (TP) Optimality

- **Claim 1** *Among all the flow-assignment policies, the Threshold Policy (with Δ given by the optimization problem below), gives the minimum possible cost subject to an average delay constraint of D_M.*



Quasi-convex in general ☹

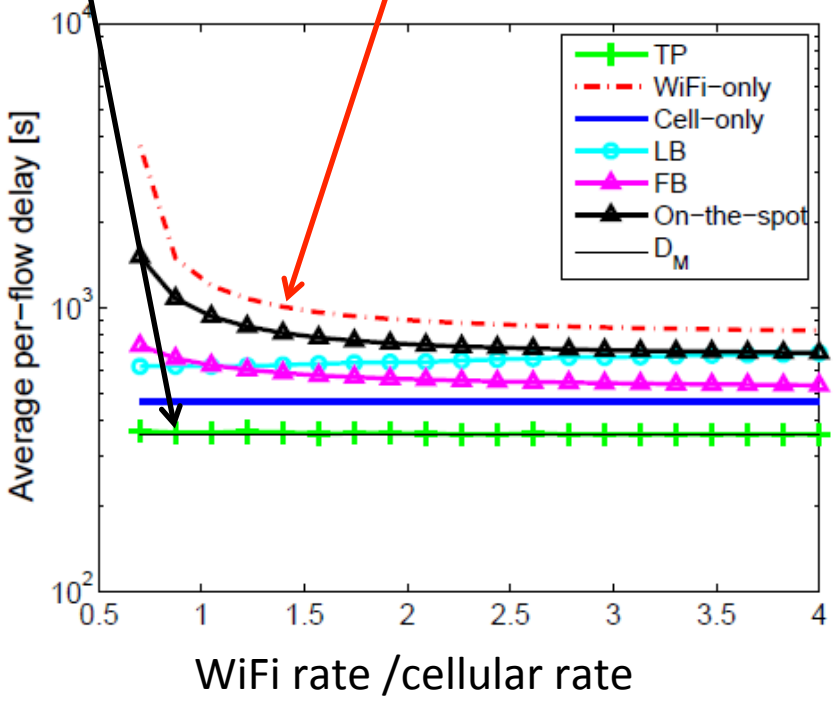
But, structure allows for simple, closed form ☺

Flow size variability plays a KEY role (through F(s))!

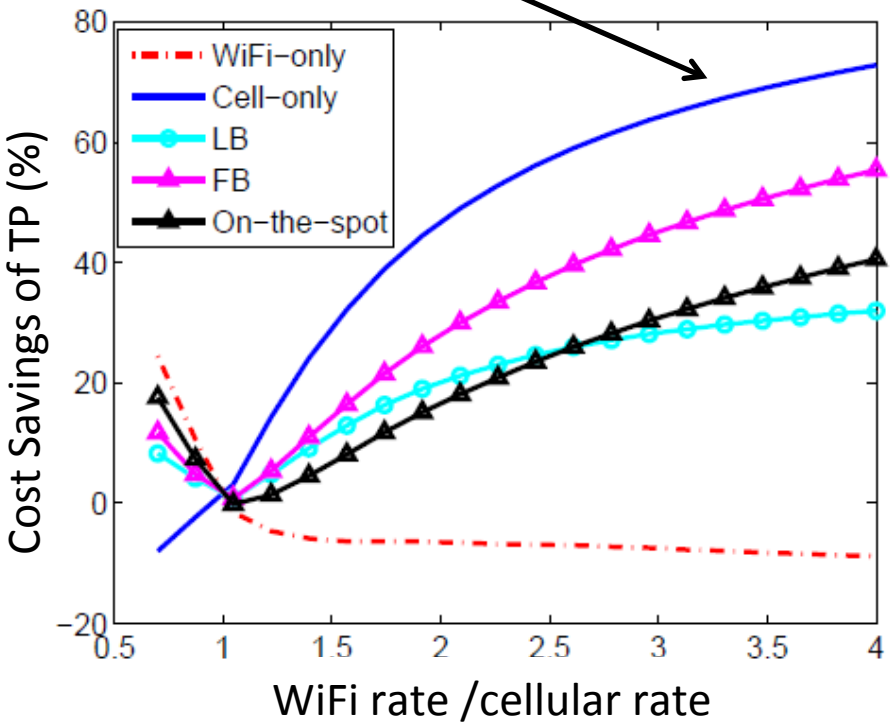
TP gains wrt other policies

TP is the only that satisfies the delay constraint!

WiFi-only: delay very high (log-scale!)

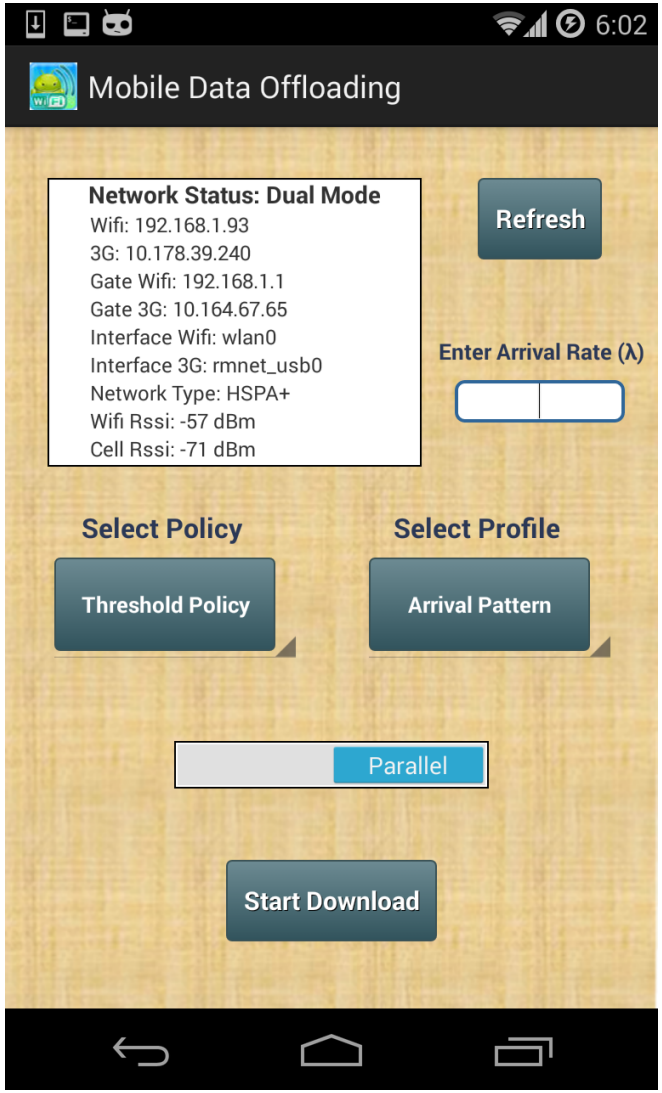


TP is much better than the other policies (**70**% of savings wrt Cell-only!)



✓ TP achieves the minimum cost among all policies that do not violate the per-flow delay constraint!

Android-based Implementation



- runs on Android-based mobile OS, *CyanogenMod* (v10.1), on rooted mobile phones
- It enables the **simultaneous** usage of WiFi and Cellular interfaces (modified Connectivity Service)
- Flow routing based on **IPTABLES**

Experimental Results

Input Params

R_{wifi} = 11.33 Mbps

R_{cell} = 7.27 Mbps

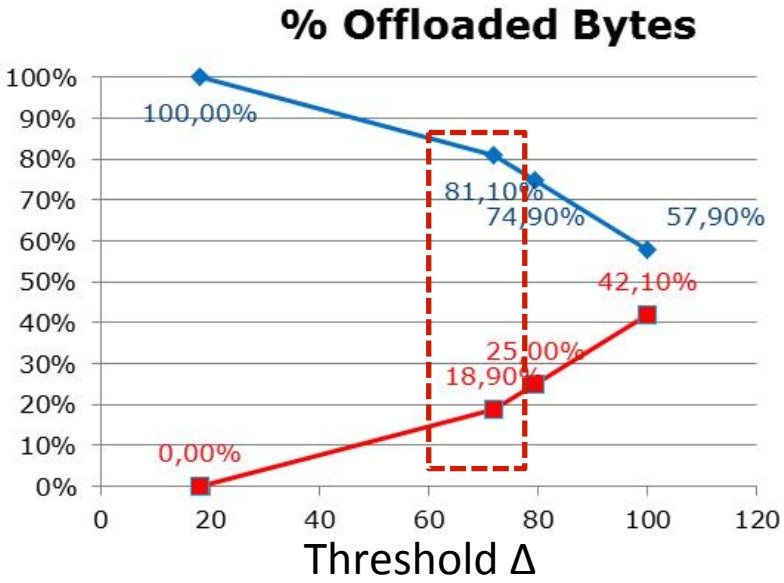
ρ = 0.9

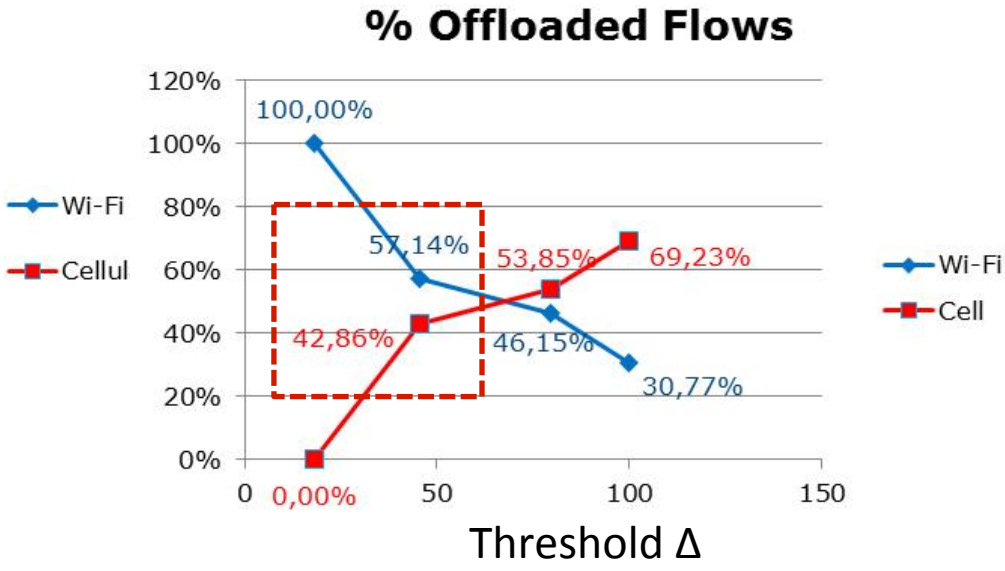
E[S] = 91.44 Mbits

Total data = 400 MB

S_{min} = 18.19 Mbits

S_{MAX} = 358.45 Mbits





Outline

Part I: Opportunistic offloading of **flows** over WiFi

- An analytical model
- Size-based offloading

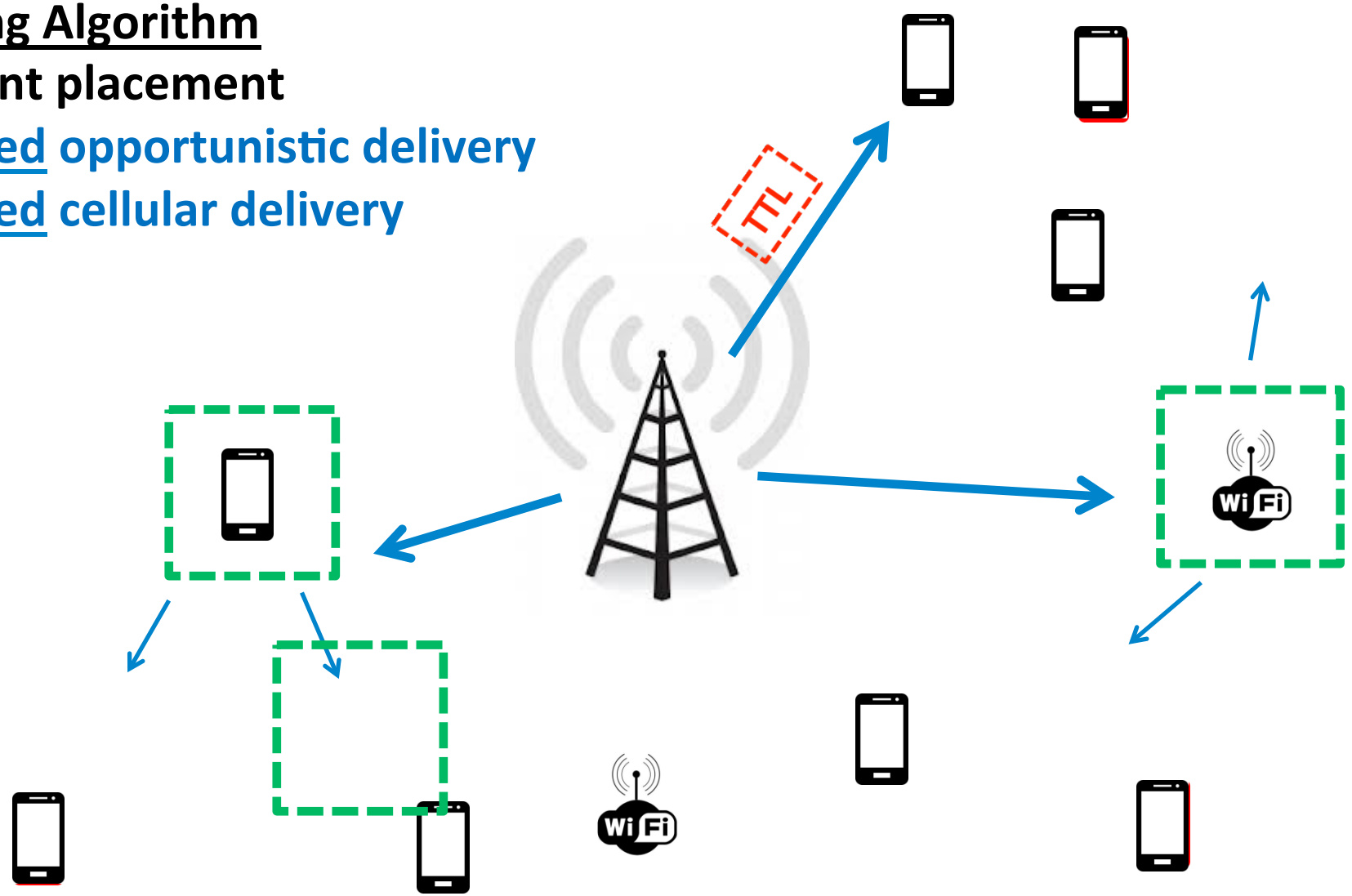
Part II: **Content storage** and **access** on the edge

- An analytical model
- Cost-optimal caching strategies
- Offloading through a vehicular cloud

Why “Opportunistic”? Allow Delayed Delivery

Offloading Algorithm

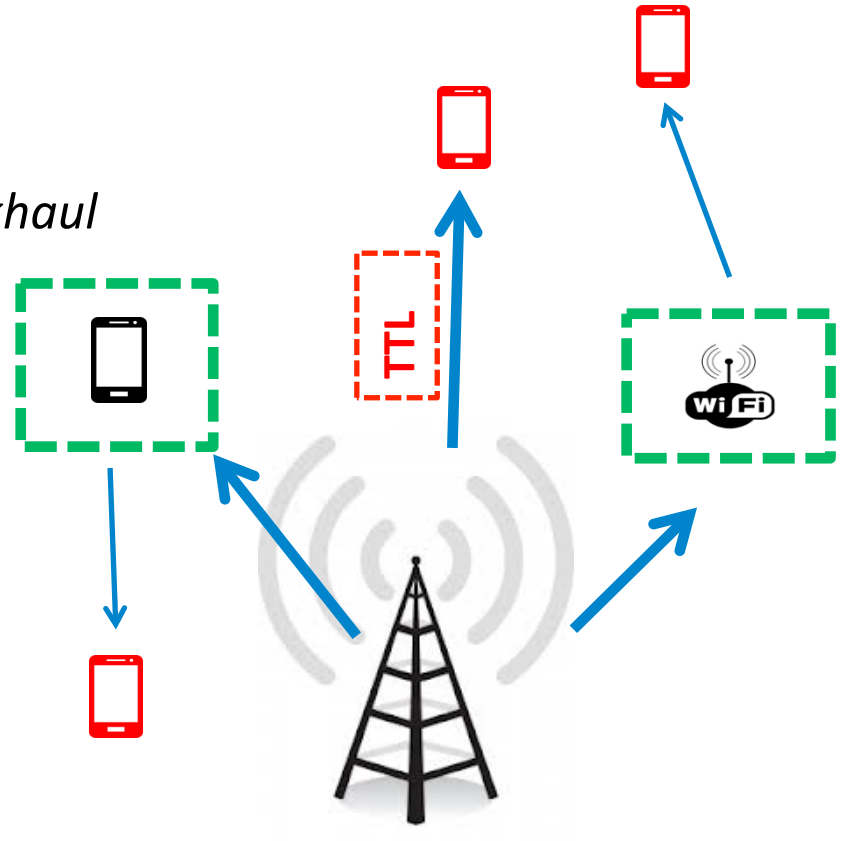
1) content placement
2) delayed opportunistic delivery
3) delayed cellular delivery



Optimal Content Placement

Various costs to consider

1) content placement (to caches):
   - $\mathbf{C_{BH}}$: *to small cells (SCs), from the backhaul*
   - $\mathbf{C_{BS}}$: *to user devices, cellular transmission from BS*
2) opportunistic delivery:
   - $\mathbf{C_{SC}}$: *from SC to user*
   - $\mathbf{C_{D2D}}$: *from user to user*
3) delayed cellular delivery:
   - $\mathbf{C_{BS}^{(TTL)}}$: *to user devices, cellular transmission from BS*



Optimization Problem

- **Objective**: minimize total cost
  - contents $\{k_1, k_2, \ldots\}$ → $min \{\sum_i C_i\}$

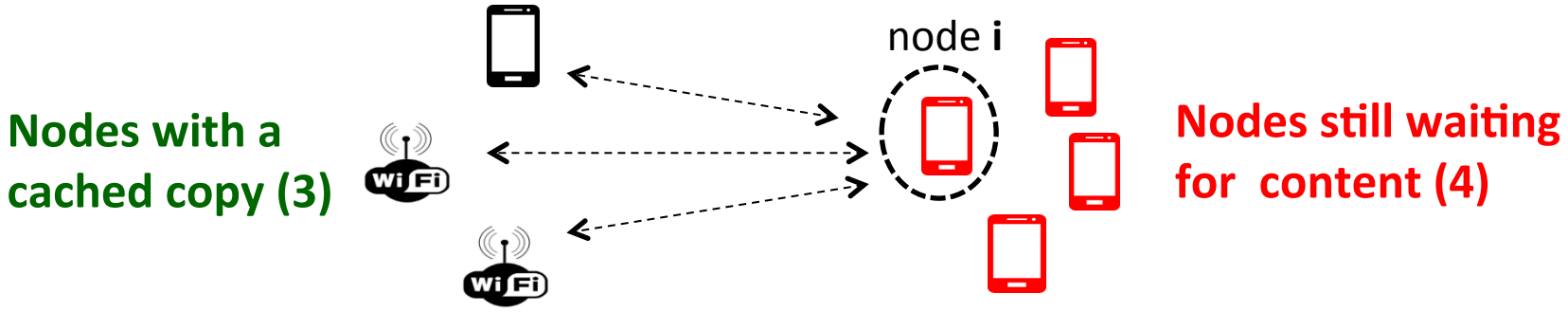
$$C_i = C_{BH} \cdot H_{SC}(0) + C_{BS} \cdot H_{MN}(0) + (C_{SC} \cdot q + C_{D2D} \cdot (1-q)) \cdot \Phi(i) \cdot P\{T \le TTL\} + C_{BS} \cdot (TTL) \cdot \Phi(i) \cdot (1 - P\{T \le TTL\})$$

Costs
H(0): #copies cached
$\Phi(i)$: popularity
$P\{T \le TTL\}$ and **q** depend on mobility

- **Optimization Variables**:
  - number of (initial) cached copies per content
  $H_{SC}(0)$ and $H_{MN}(0)$
- **Constraints**:
  1) # of cache replicas for content i < than # of caches
  2) total # of cached contents < total storage capacity

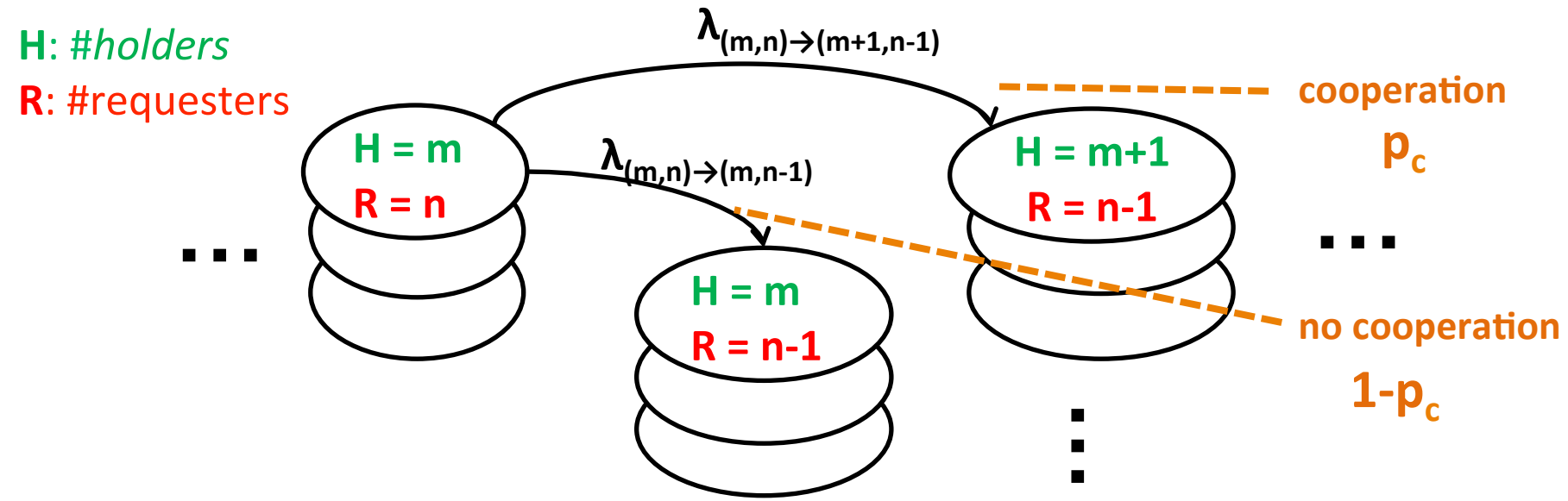
Key New "Ingredient": Performance of Delayed Access

- **Need these performance metrics to proceed:**
  - **P(T ≤ TTL):** delivery probability by TTL
  - **q:** ratio of requests served from SCs /D2D



- **Next Cache Hit**: When a red node meets a node (SC or UE) with cached copy ➔ depends on **mobility** and **availability**
- **Cache miss -- P(T > TTL):** a red node does not meet a cache with copy by TTL

Track Evolution of Cache Hits and New Holders



⟹ ***Mean Field – Fluid Model*** approximations

$$\lambda_{(m,n)\to(m+1,n-1)} \approx p_c \cdot H \cdot R \cdot \mu_\lambda$$

$$\lambda_{(m,n)\to(m,n-1)} \approx (1-p_c) \cdot H \cdot R \cdot \mu_\lambda$$

H(t): *#holders* at time *t*

$$dH(t)/dt = p_c \cdot H(t) \cdot R(t) \cdot \mu_\lambda$$

R(t): #requesters at time *t*

$$dR(t)/dt = -H(t) \cdot R(t) \cdot \mu_\lambda$$

Performance Prediction → Optimization

Delivery Probability: $P\{T_d \leq TTL\} = 1 - e^{-\mu_\lambda \cdot \int_0^{TTL} H(\tau)d\tau}$

Expected Delivery Delay: $E[T_d] = \int_0^{\infty} e^{-\mu_\lambda \cdot \int_0^{t} H(\tau)d\tau} \cdot dt$

$$min\downarrow \; H\downarrow SC \;,\; H\downarrow MN \quad \{\sum \theta=1\ldots M\uparrow \blacksquare C\uparrow \theta \;\}$$

$$s.t.\ \forall \theta:\ 0 \leq H\downarrow SC\uparrow \theta \leq N\downarrow SC$$

$$0 \leq H\downarrow MN\uparrow \theta \leq R\downarrow 0$$

$$\text{and} \quad \sum \theta=1\ldots M\uparrow \blacksquare H\downarrow SC\uparrow\uparrow \theta \leq \sum i=1\ldots N\downarrow SC\uparrow \blacksquare C$$

Total nb of SCs

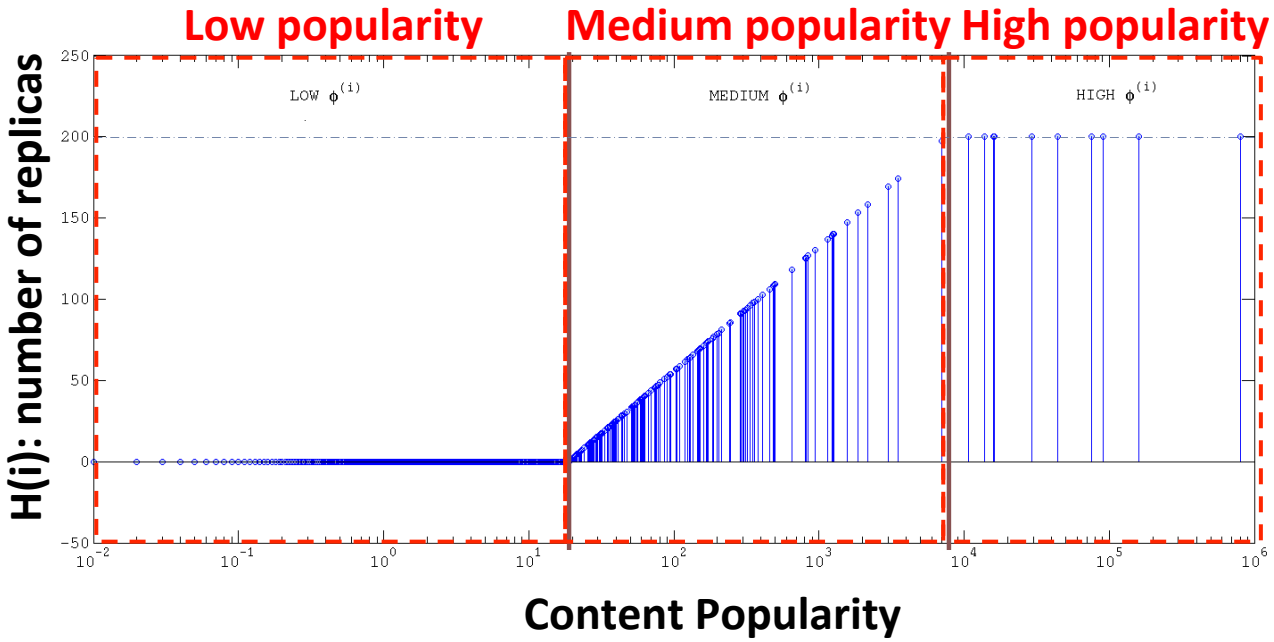
Capacity constraint

Simple Example: Only Initial Caching

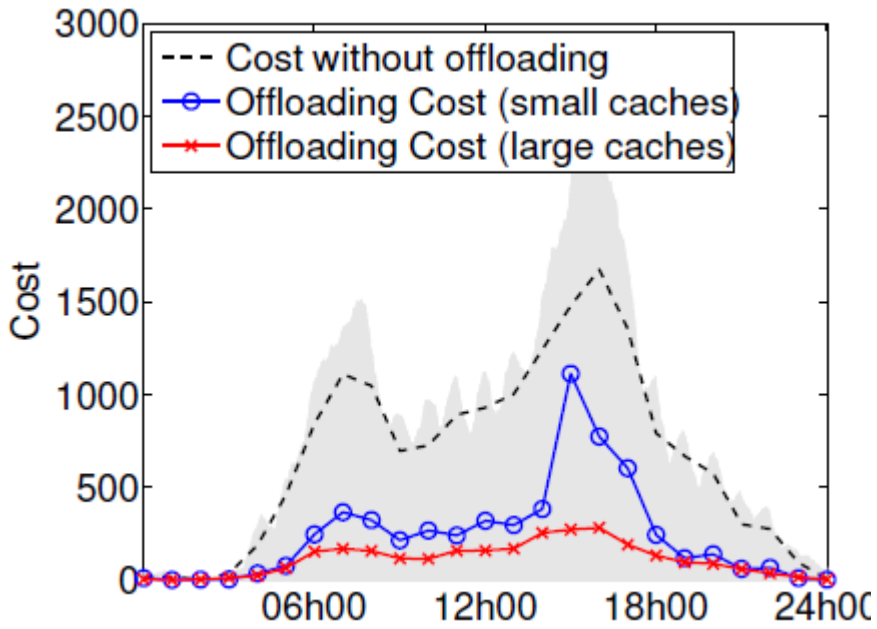
Solving using Lagrange multipliers (**convex** problem) gives:

$$H(i) = 1/\mu\lambda TTL \ln(\Delta T\mu\lambda TTL \Phi(i)/1 + \lambda(i) - \nu(i) + \rho(i))$$

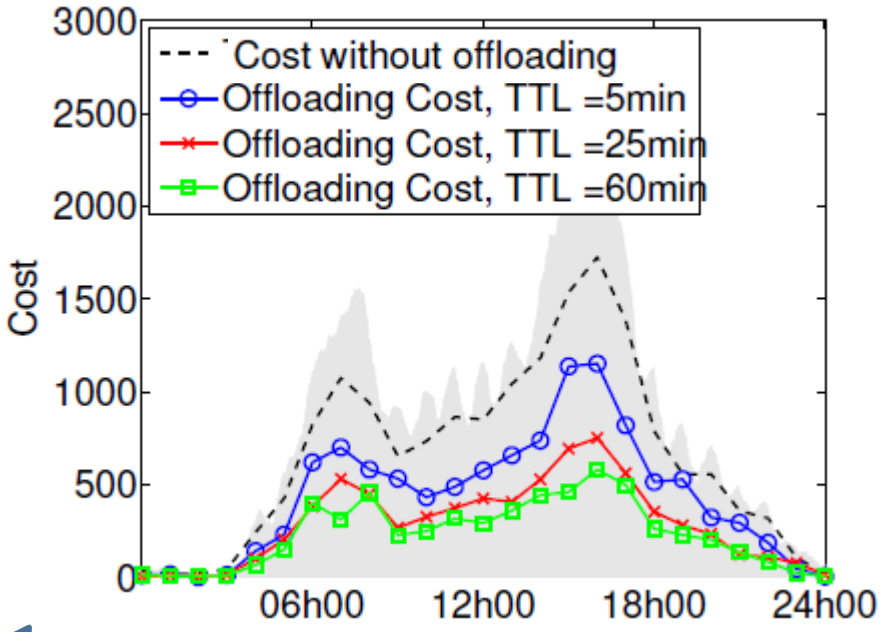
where λ, ν and ρ are Lagrangian multipliers



Performance Evaluation



Case 1: Offloading only through SCs



Case 2: Offloading only through MNs

- ***Significant cost decrease***
- ***Smoothen / Flatten traffic peaks***
  → avoid over-provision of network capacity
- ***Increase SCs caches (cheap) or TTL (incentives)***
  → lower & smoother cost

Publications Related to Part II

Pavlos Sermpezis, Thrasyvoulos Spyropoulos, ***"Not all content is created equal: Effect of popularity and availability for content-centric opportunistic networking"****, Proc.* ***ACM MobiHoc****, August 2014*

*Pavlos Sermpezis, Luigi Vigneri, Thrasyvoulos Spyropoulos, "****Offloading on the Edge: Analysis and optimization of local data storage and offloading in HetNets****", ArXiv* 1503.00648*, March 2015.*

Key Messages for 5G Research

- Performance **at flow or content level** is key
  - Application QoS ➔ (E2E) time to access content
  - Instantaneous throughput of BS maybe not best metric
- **Queueing** analysis to understand impact of: scheduler, network switching, etc.
  - Processor Sharing queues
  - Variable or intermittent service rate
- **Per flow decisions**
  - Offloading, Carrier aggregation, Routing/Association on Radio Access and Backhaul
  - Based on flow characteristics and network load
  - Facilitated by SDN
- **Delayed Access**
  - Need to understand impact of mobility and topology
  - Can improve network-wide performance (with reasonable impact on user QoE

Interesting Open Issues

- Understanding/modeling costs (incentives, congestion)
- Real-time conditions estimation and update
  - UE side: WiFi/cellular performance
  - BS side (popularity estimation)
- Understanding (local) content access patterns
- Joint scheduling + storage
- Cross-layer (PHY + Network interaction)