Optimal dynamic proactive caching via reinforcement learning

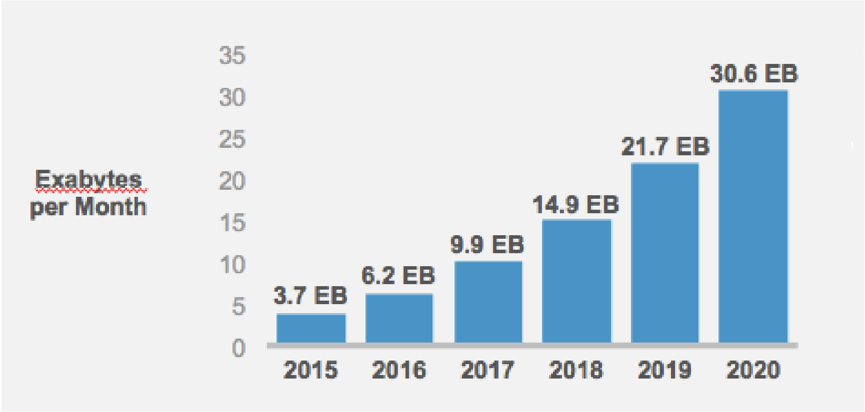
Alireza Sadeghi, Fatemeh Sheikholeslami, and Georgios Giannakis

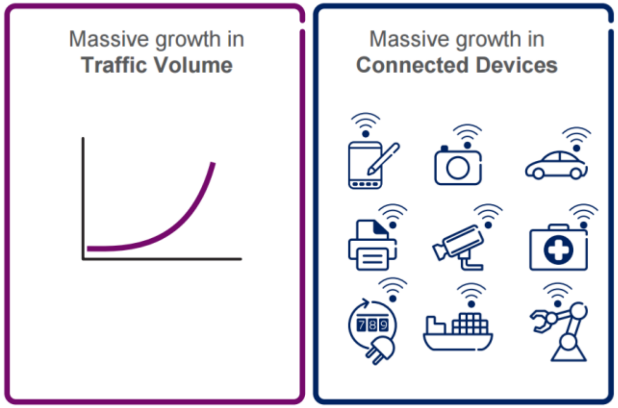
April 14, 2017

Acknowledgment: ARO W911NF-15-1-0492 and NSF-EARS 1343248.

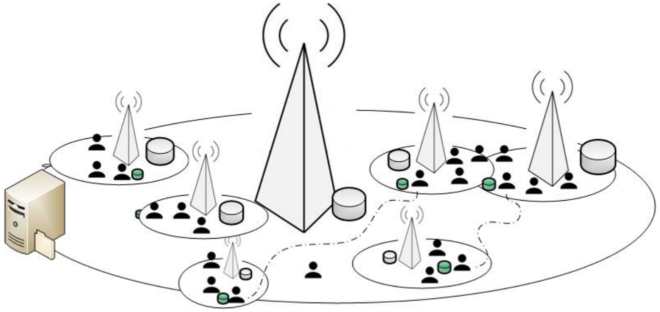
Evolution of wireless networks

- 8-fold growth of global mobile data traffic between 2015 and 2020





- Heterogeneous network architecture (HetNet)
  - Access technologies in heterogeneous subnets
- 60% of data is reusable, a.k.a. ***contents***
  - Utilization of storage units at small base stations (SBs)
  - Challenge: *what* and *when* to cache?

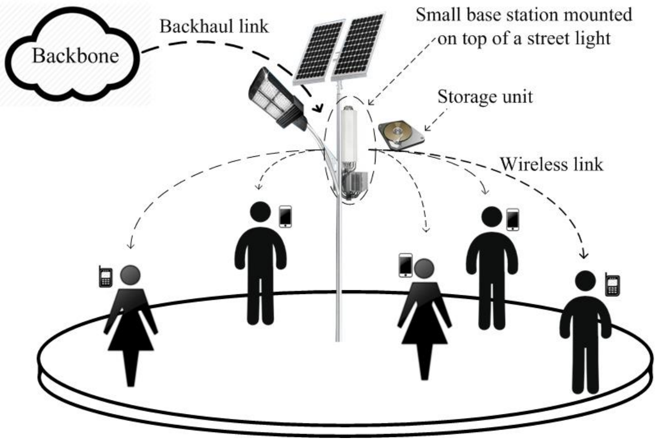


G. Paschos, E. Bastug, I. Land, G. Caire and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16-22, August 2016.

Caching in wireless networks

❑ Memory-enabled SBs

- Cache during off-peak hours
- Reduce load on backhauls during peak traffic periods
- Reduce cost for providing service with high QoS



❑ Generally unknown content popularity profiles

❑ Prior art

- Multi-armed bandit (MAB) formulation [D. Belasco et al'14]
- Distributed and convexified MAB [A. Sengupta et al'14]

$\Longrightarrow$ Unknown static popularity profile

- Dynamic user demand [Kim et al'17]

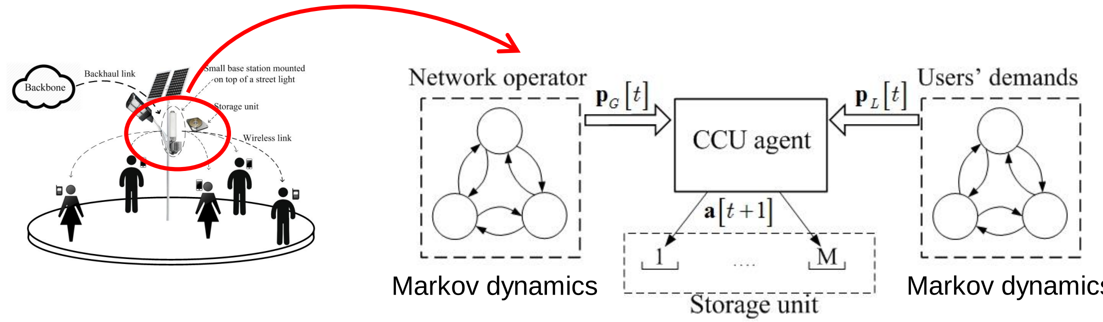
$\Longrightarrow$ Unknown dynamic popularity profile

Proposed approach

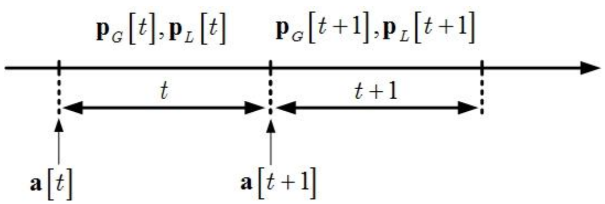
✓ Caching via <u>reinforcement learning</u> over files with <u>spatio-temporally</u> dynamic popularities

System model

❑ Discrete-time network



- SB with caching control unit (CCU)
- Total number of F contents w/ unit size in backbone
- Storage capacity of M files in SBs

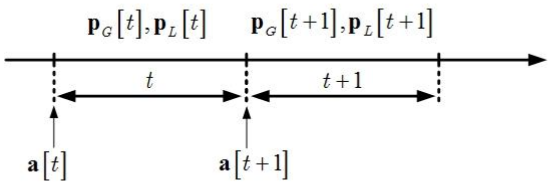


- Local popularity profile $\left[\mathbf{p}_{\mathrm{L}}\left[t\right]\right]_f = \dfrac{\#\text{ of local requests for file } f \text{ at time interval } t}{\text{Total }\#\text{ of local requests at time interval } t};$
- Global popularity profiles $\mathbf{p}_{\mathrm{G}}\left[t\right]$
- Action vector $\mathbf{a}\left[t\right] \in \{0,1\}^F$: SB caches f^{th} content if $\mathbf{a}_f\left[t\right] = 1$
- State vector $\mathbf{s}\left[t\right] := \left[\mathbf{p}_G^T\left[t\right], \mathbf{p}_L^T\left[t\right], \mathbf{a}^T\left[t\right]\right]^\top$
- Policy $\pi(\cdot)$ is a mapping from state space to action space $\Rightarrow$ $\mathbf{a}[t+1] = \pi(\mathbf{s}[t])$

Goal: Given $\{\mathrm{s}[\tau]\}_{\tau=0}^{\tau=t}$ and observed costs, optimize policy $\pi(.)$

Problem formulation

- Recall



- Costs
  - Refreshing the cached contents $h\left(\mathbf{a}[t], \mathbf{a}[t-1]\right) := \lambda_1 \mathbf{a}^\top[t](\mathbf{1} - \mathbf{a}[t-1])$
  - Fetching requested non-cached files $g\left(\mathbf{s}[t]\right) := \lambda_2 (\mathbf{1} - \mathbf{a}[t])^\top \mathbf{p}_L[t]$
  - Tracking global popularities $f\left(\mathbf{s}[t]\right) := \lambda_3 (\mathbf{1} - \mathbf{a}[t])^\top \mathbf{p}_G[t]$

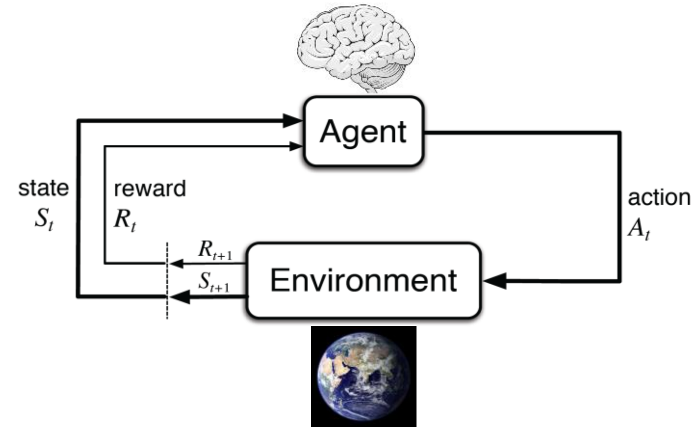
$$\Longrightarrow \quad C\left(\mathbf{s}[t-1], \mathbf{a}[t] \middle| \mathbf{p}_{\mathrm{G}}[t], \mathbf{p}_{\mathrm{L}}[t]\right) := h\left(\mathbf{a}[t], \mathbf{a}[t-1]\right) + g\left(\mathbf{s}[t]\right) + f(\mathbf{s}[t])$$

- Expected discounted cost $V_\pi\left(\mathbf{s}[\tau]\right) := \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=\tau}^{T} \gamma^{t-\tau} C\left(\mathbf{s}[t], \pi\left(\mathbf{s}[t]\right)\right)\right]$

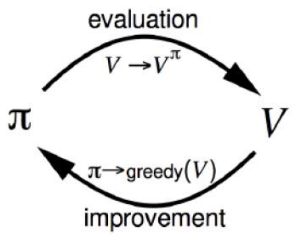
- **Goal**: Find the optimal policy $\pi^* = \arg\min_{\pi \in \Pi} V_\pi\left(\mathbf{s}_0\right)$

- Viable approaches
  - ✓ Adaptive dynamic programming
  - ✓ Q-learning
  - ✓ SARSA

Reinforcement learning (RL)



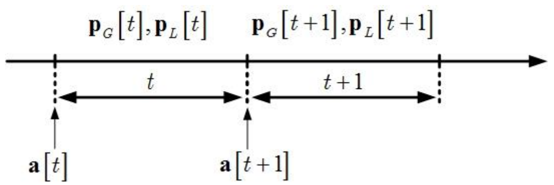
- Agent-environment interactions
  - **State**: mathematical representation of environment
  - **Action**: decision made by the agent
  - **Reward**: scalar feedback, how well agent is doing
- State value function (under policy π)
  - Immediate reward + discounted future rewards
- Objective
  - ✓ Find optimal policy $\pi(\cdot)$ such that $V_\pi(\mathbf{s})$ is maximized for all states



- Bellman equation $$V_\pi(\mathbf{s}) = \mathbb{E}\left[C(\mathbf{s}, \pi(\mathbf{s}))\right] + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathbf{T}(\mathbf{s}'; \mathbf{s}, \pi(\mathbf{s})) V_\pi(\mathbf{s}') \;\; , \forall \mathbf{s}, \mathbf{s}'.$$
- Q-learning
  - State-action value function $$Q^\pi(\mathbf{s}, \mathbf{a}) := \mathbb{E}\left[C(\mathbf{s}, \mathbf{a})\right] + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathbf{T}(\mathbf{s}'; \mathbf{s}, \mathbf{a}) V^\pi(\mathbf{s}')$$
- Optimality

$$\pi^*(\mathbf{s}) = \arg\min_{\mathbf{a}} Q^*(\mathbf{s}, \mathbf{a}), \quad \forall \mathbf{s} \in \mathcal{S},$$

Q-learning for proactive caching



Algorithm 1: Proactive-caching via Q-learning at CCU

Initialize $\mathbf{s}[0]$ randomly and $\hat{Q}(\mathbf{s}, \mathbf{a}) = 0 \; \forall \mathbf{s}, \mathbf{a}$
for t=1,2,... **do**
  Take action $\mathbf{a}[t]$ chosen probabilistically by

$$\mathbf{a}[t] = \begin{cases} \arg\min_{\mathbf{a}} \hat{Q}\left(\mathbf{s}[t-1], \mathbf{a}\right) & \text{w.p.} \quad 1-\epsilon_t \\ \text{random } \mathbf{a} \in \mathcal{A} & \text{w.p.} \quad \epsilon_t \end{cases}$$

Exploration-exploitation tradeoff

  $\mathbf{p}_L[t]$ and $\mathbf{p}_G[t]$ are revealed
  Set $\mathbf{s}[t] = \left[\mathbf{p}_L[t], \mathbf{p}_G[t], \mathbf{a}[t]\right]$
  Incur cost $C\left(\mathbf{s}[t-1], \mathbf{a}[t] \,|\mathbf{p}_G[t], \mathbf{p}_L[t]\right)$

Observations are revealed

  Update

$$\hat{Q}_t\left(\mathbf{s}[t-1], \mathbf{a}[t]\right) \leftarrow (1-\beta_t)\hat{Q}_{t-1}\left(\mathbf{s}[t-1], \mathbf{a}[t]\right) + \beta_t \times \left[C\left(\mathbf{s}[t-1], \mathbf{a}[t] |\, \mathbf{p}_G[t], \mathbf{p}_L[t]\right) + \gamma \min_{\mathbf{a}} \hat{Q}_{t-1}\left(\mathbf{s}[t], \mathbf{a}\right)\right]$$

Stochastic update of Q-table

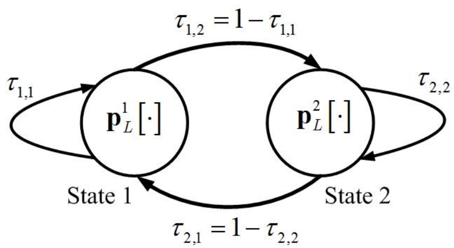
Convergence

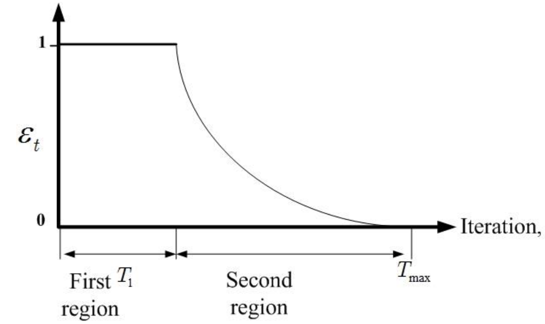
✓ If
1. All state-action pairs are continuously visited,
2. Step-size β_t satisfies $\sum_{t=1}^{\infty} \beta_t = \infty$ and $\sum_{t=1}^{\infty} \beta_t^2 < \infty$

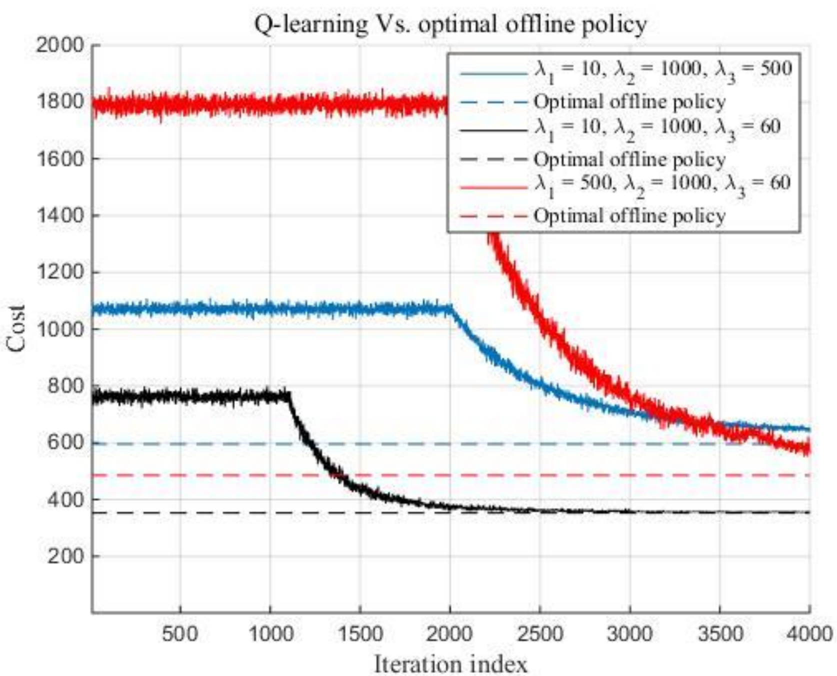
then policy will converge to the optimal policy , i.e., $\pi \to \pi^*$ w.p. 1.

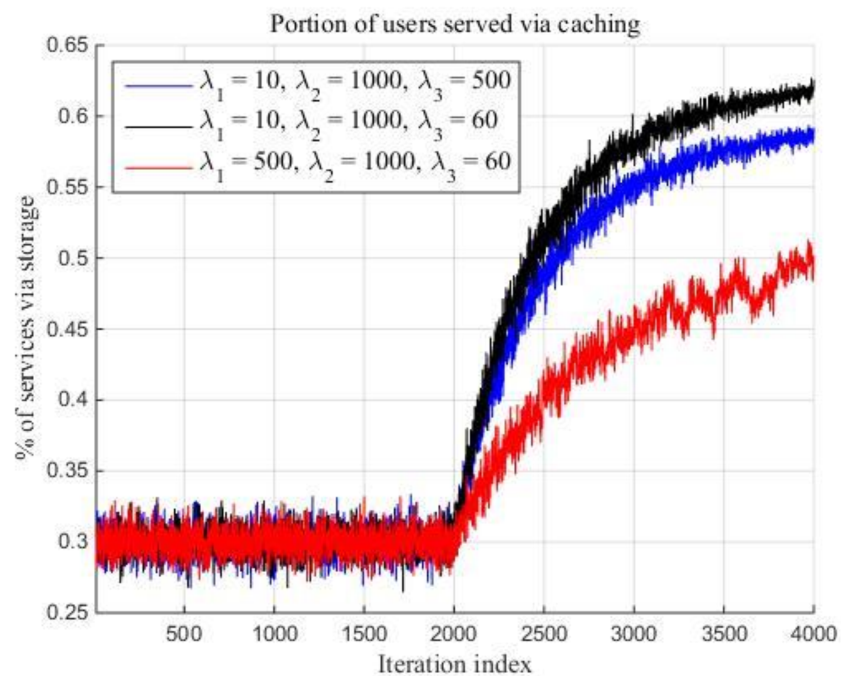
Simulation tests

- Consider a total of *F=10*, and *M=3*
  - Two-state Markov chain for modeling $\mathbf{p}_{\mathrm{L}}[t]$
  - State transition probabilities $\boldsymbol{\tau} := \begin{bmatrix} 0.35 & 0.65 \\ 0.75 & 0.25 \end{bmatrix}$
  - Similarly for $\mathbf{p}_{\mathrm{G}}[t]$ with $\boldsymbol{\tau}' := \begin{bmatrix} 0.6 & 0.4 \\ 0.45 & 0.55 \end{bmatrix}$
  - $\beta_t = 0.3$ and $\gamma = 0.6$









Thank you!