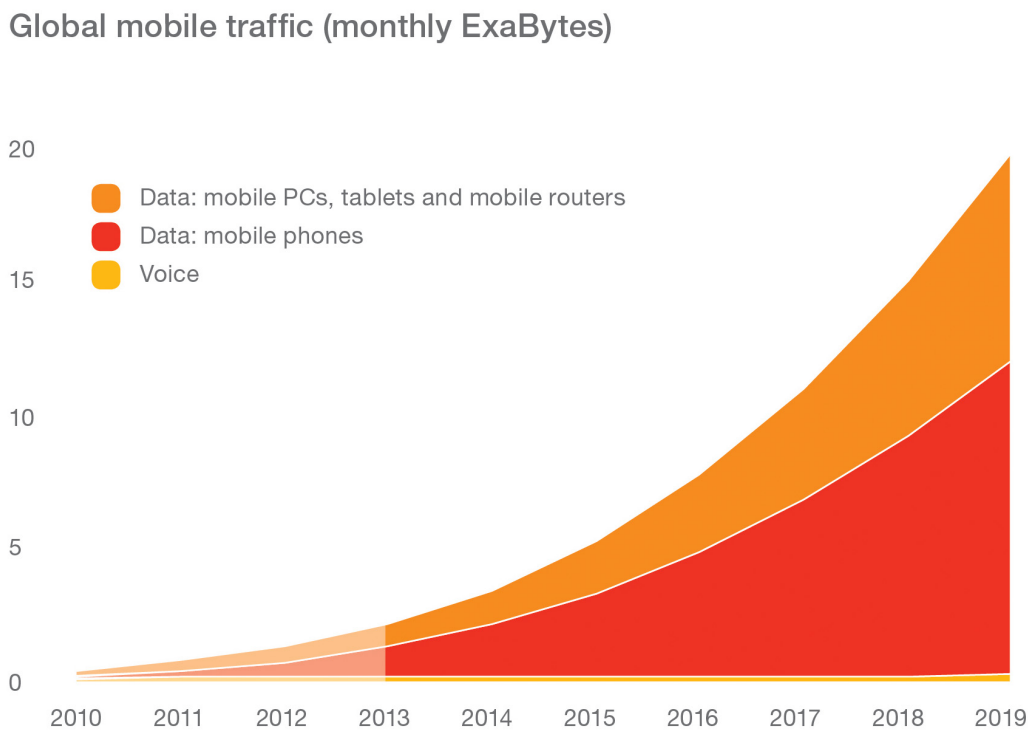
Predictive Delay-Aware Network Selection in Data Offloading

Haoran Yu, Man Hon Cheung, Longbo Huang, Jianwei Huang

Network Communications and Economics Lab (NCEL)
The Chinese University of Hong Kong (CUHK), Hong Kong

Institute for Interdisciplinary Information Sciences (IIIS)
Tsinghua University (THU), China

Mobile Traffic Explosion



Global Mobile Data Traffic Growth till 2019 (©Ericsson)

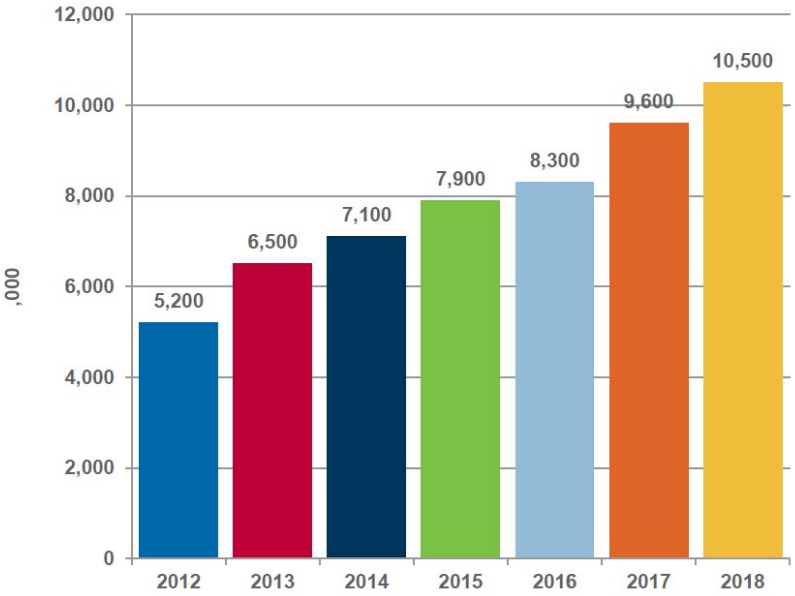
- Annual grow rate $\sim$ 45%
  - A 10-fold increase between 2013 and 2019

Mobile Traffic Explosion

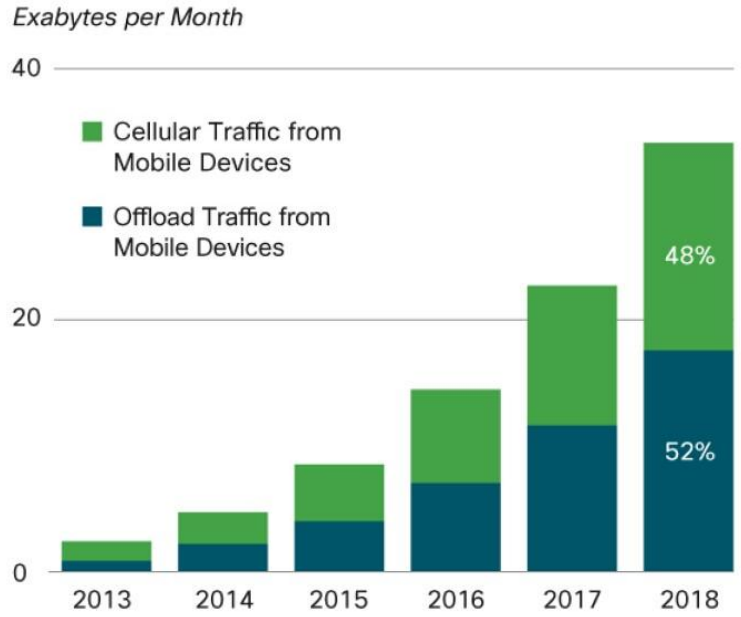
- Traditional ways to expand the network capacity
  - Acquiring new spectrum bands
  - Building more cell sites
  - Upgrading technologies (*e.g.*, 3G $\rightarrow$ 4G)
  - ...
- Problems: time-consuming and costly
- Question: How to expand the network capacity in a cost and time-efficient manner?
- One answer: Data offloading

Mobile Data Offloading

- Data offloading: deliver cellular traffic to other complementary networks (*e.g.*, Wi-Fi, femtocell)
  - Wi-Fi deployment rate is increasing (10.5 million by 2018)
  - More than half traffic will be offloaded (52% by 2018)



New Carrier-Grade Wi-Fi Per Year (©WBA)



Percentage of Offloaded Traffic (©Cisco)

Mobile Data Offloading

- Type 1: User-initiated offloading
  - Users decide which network (*e.g.*, cellular or Wi-Fi) to connect
- Type 2: Operator-initiated offloading (this work)
  - Mobile operator makes the network selection decision
  - Advantages: seamless switch, optimize revenue and QoE

Problem Description

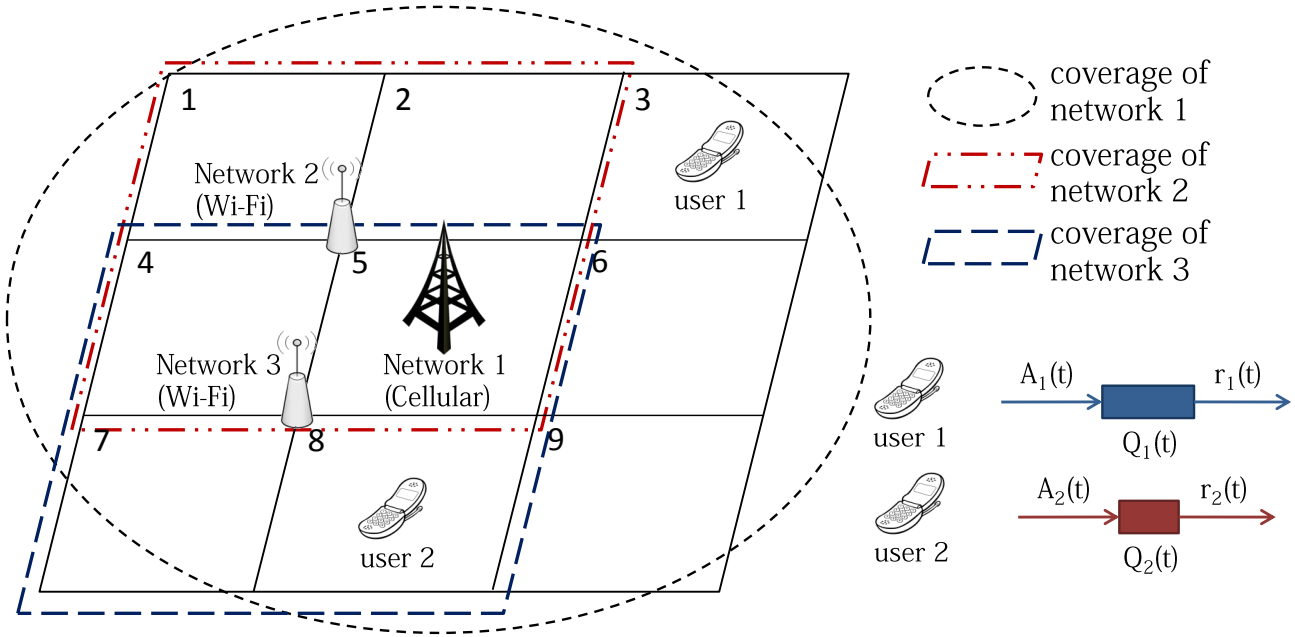
- Major concerns
  - Operation cost: networks have heterogeneous operation costs
  - Traffic delay: delaying traffic causes users' dissatisfaction

- Question: How does operator dynamically select networks for users so that the long-term operation cost and traffic delay is well balanced?

Problem Description

- Challenge: limited information on system randomness
  - Users' traffic demand and network availability
- Case 1: Only have current slot information
  - Algorithm: Lyapunov optimization $\rightarrow$ **DNS**
- Case 2: Have both current and predicted future information
  - Algorithms: a novel frame-based Ly. opt. $\rightarrow$ **P-DNS, GP-DNS**

System Model



- Single operator
- Multiple networks, locations, and users
  - Network availability is location-dependent
  - Users randomly move across the locations with random traffic arrivals

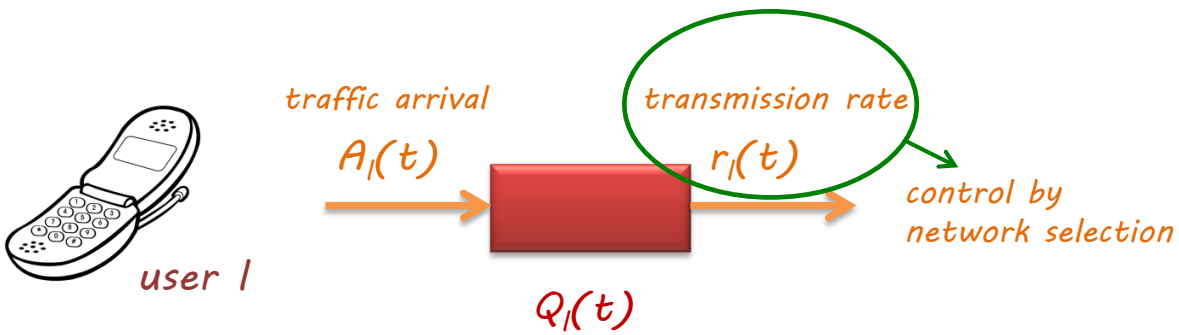
Notations

- System settings
  - Slotted system, $t \in \{0, 1, 2, \cdots\}$
  - Set of locations, $\mathcal{S} = \{1, 2, \cdots, S\}$
  - Set of users, $\mathcal{L} = \{1, 2, \cdots, L\}$
  - Set of networks, $\mathcal{N} = \{1, 2, \cdots, N\}$
  - Location-dependent availability: $\mathcal{N}_s \subseteq \mathcal{N}$, networks available at $s \in \mathcal{S}$
- System randomness
  - User $l \in \mathcal{L}$'s traffic arrival at t, $A_l(t)$
  - User $l \in \mathcal{L}$'s location at t, $S_l(t)$
- Operator's online decision
  - Network selection for l at t, $\alpha_l(t)$
  - Determine: User l's transmission rate at t, $r_l(\boldsymbol{\alpha}(t))$
  - Determine: Total operation cost at t, $c(\boldsymbol{\alpha}(t))$

Queueing Dynamics

- Users' data queues
  - User l's unserved traffic at t, $Q_l(t)$

$$Q_l(t+1) = \max\left[Q_l(t) - r_l\left(\boldsymbol{\alpha}(t)\right), 0\right] + A_l(t)$$



Case 1: Only Current Network Information

- At time t, operator only observes
  - Traffic arrivals, $\boldsymbol{A}(t)$ (random variable)
  - Users' locations, $\boldsymbol{S}(t)$ (random variable)
  - Data queues, $\boldsymbol{Q}(t)$ (determined by historical arrival and transmission)
- Propose **DNS** algorithm to make online decision on $\boldsymbol{\alpha}(t)$

Delay-Aware Network Selection (DNS)

Delay-Aware Network Selection (DNS) Algorithm

At each time slot t, the operator:

- Chooses the network selection vector $\boldsymbol{\alpha}^*(t)$ that solves

$$\textbf{minimize} \quad \left[-\sum_{l=1}^{L} Q_l(t) r_l(\boldsymbol{\alpha}(t))\right] + V c(\boldsymbol{\alpha}(t))$$

$$\textbf{variables} \quad \alpha_l(t) \in \mathcal{N}_{S_l(t)} \cup \{0\}, \ \forall l \in \mathcal{L}.$$

- Updates the queueing vector $\boldsymbol{Q}(t+1)$ accordingly.

- Intuition:
  - When $Q_l(t)$ is large, suspending service incurs severe delay.
    Strategy: serve user l immediately even with a high operation cost
  - When $Q_l(t)$ is small, suspending service does not lead to severe delay.
    Strategy: wait till enter Wi-Fi area

Performance of DNS

Performance of DNS

Under mild assumption on capacity region, for i.i.d. randomness:

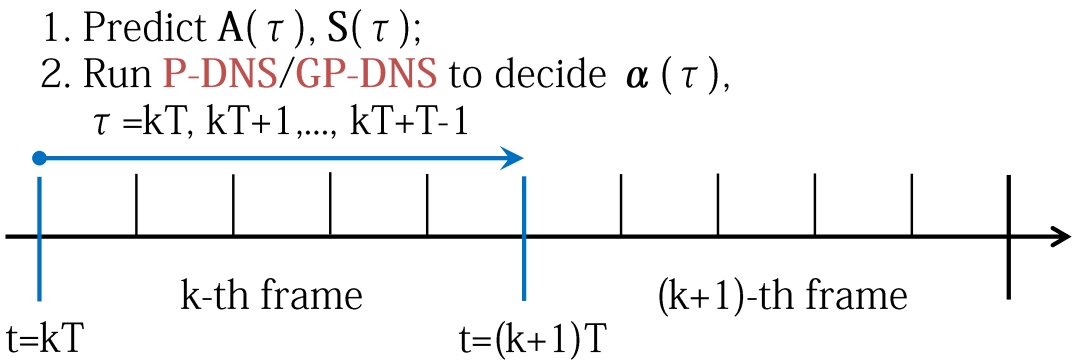
$$c_{av}^{\mathrm{DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{c(\boldsymbol{\alpha}(\tau))\} \leq c_{av}^{*} + \frac{B}{V},$$

$$Q_{av}^{\mathrm{DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{l=1}^{L} \mathbb{E}\{Q_l(\tau)\} \leq \frac{B+Vc_{\max}}{\eta}.$$

- $[O(1/V), O(V)]$ cost-delay tradeoff (V: control parameter)
  - Time average operation cost is within $O(1/V)$ of the optimality
  - Time average traffic delay is bounded by $O(V)$
  - Conclusion: The operation cost can be pushed arbitrarily close to the optimal value, but at the expense of an increase in the traffic delay
- **DNS** has similar cost-delay tradeoff under Markovian randomness

Case 2: Current and Future Information

- Question: how to improve **DNS** with predictable future information?
- Frame-based prediction and network selection structure
  - Frame size, T (prediction capability)
  - k-th frame, $\mathcal{T}_k \triangleq \{kT, kT+1, \ldots, kT+T-1\}$



Predictive Delay-Aware Network Selection (P-DNS)

Predictive Delay-Aware Network Selection (P-DNS) Algorithm

At time slot $t = kT$, $k \in \{0, 1, \ldots\}$, the operator:

- Chooses the network selection vectors $\{\boldsymbol{\alpha}^*(\tau)\}$, $\tau \in \mathcal{T}_k$, that solve

$$\textbf{minimize} \quad \sum_{\tau=kT}^{kT+T-1} \left(\sum_{l=1}^{L} Q_l(\tau)\left(A_l(\tau) - r_l\left(\boldsymbol{\alpha}(\tau)\right) + \theta\right) + V c\left(\boldsymbol{\alpha}(\tau)\right) \right)$$

$$\textbf{subject to} \quad Q_l(\tau), \tau \in \mathcal{T}_k, \text{ follows queueing dynamics}$$

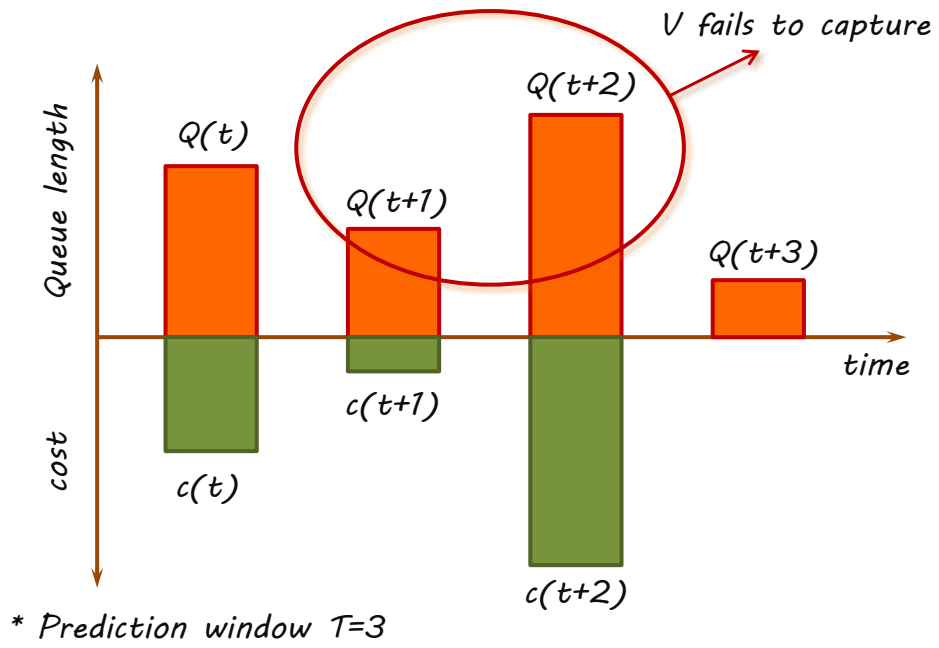
$$\textbf{variables} \quad \alpha_l(\tau) \in \mathcal{N}_{S_l(\tau)} \cup \{0\}, \ \forall l \in \mathcal{L}, \tau \in \mathcal{T}_k.$$

- Updates the queueing vector $\boldsymbol{Q}(kT + T)$ accordingly.

- Besides V, we add another positive control parameter θ

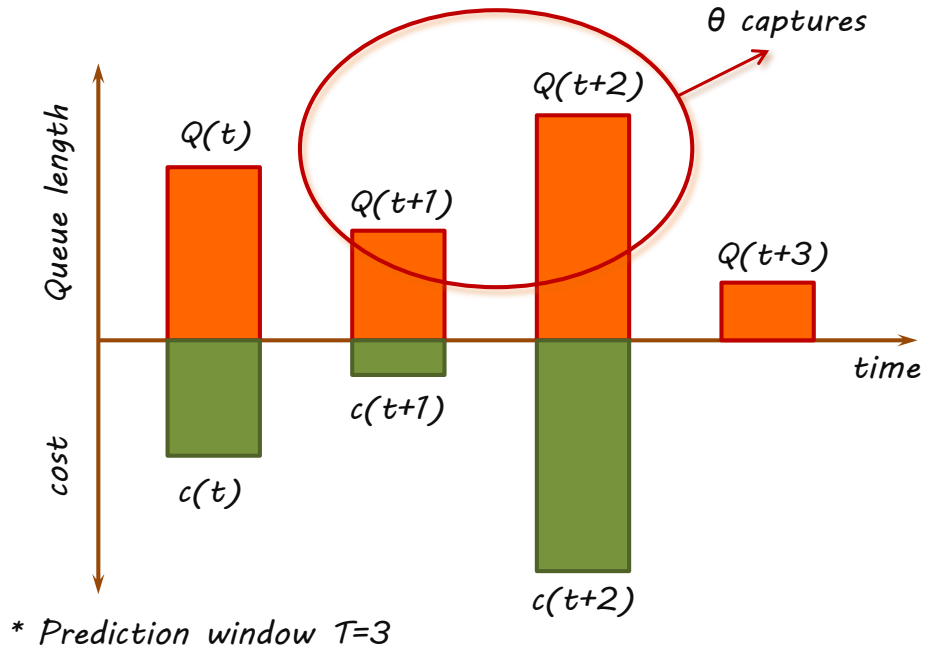
Predictive Delay-Aware Network Selection (P-DNS)

- V balances:
  - Total variance of queue length of the frame, *i.e.*,
    $$\sum_{\tau=t}^{t+2} \left(Q\left(\tau+1\right)-Q\left(\tau\right)\right)=Q\left(t+3\right)-Q\left(t\right)$$
  - Total operation cost of the frame, *i.e.*, $\sum_{\tau=t}^{t+2} c\left(\tau\right)$



Predictive Delay-Aware Network Selection (P-DNS)

- θ balances the queue variance within the frame
  - $\theta > 0$ assigns larger weights to the earlier slots of the frame $\rightarrow$ Serve users earlier rather than later $\rightarrow$ Reduce queue length of middle slots



Performance of P-DNS

- Consider perfect prediction

Performance of P-DNS

Under mild assumption on capacity region, for i.i.d. randomness:

$$c_{av}^{\text{P-DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{c(\alpha(\tau))\} \leq C(\theta) + \frac{B}{V},$$

$$Q_{av}^{\text{P-DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{l=1}^{L} \mathbb{E}\{Q_l(\tau)\} \leq \frac{B+VC(\theta)}{\theta}.$$

- Question Reduce the complexity of **P-DNS**?

Greedy Predictive Delay-Aware Network Selection (GP-DNS)

- Key idea: Solve the optimization in **P-DNS** approximately and iteratively (details omitted)
- Use ξ to describe the degree of approximation
  - $\xi = 1$: solve the optimization problem optimally
  - $\xi > 1$: a larger ξ implies a worse approximation

Performance of GP-DNS

Under i.i.d. randomness, **GP-DNS** with efficiency ξ achieves:

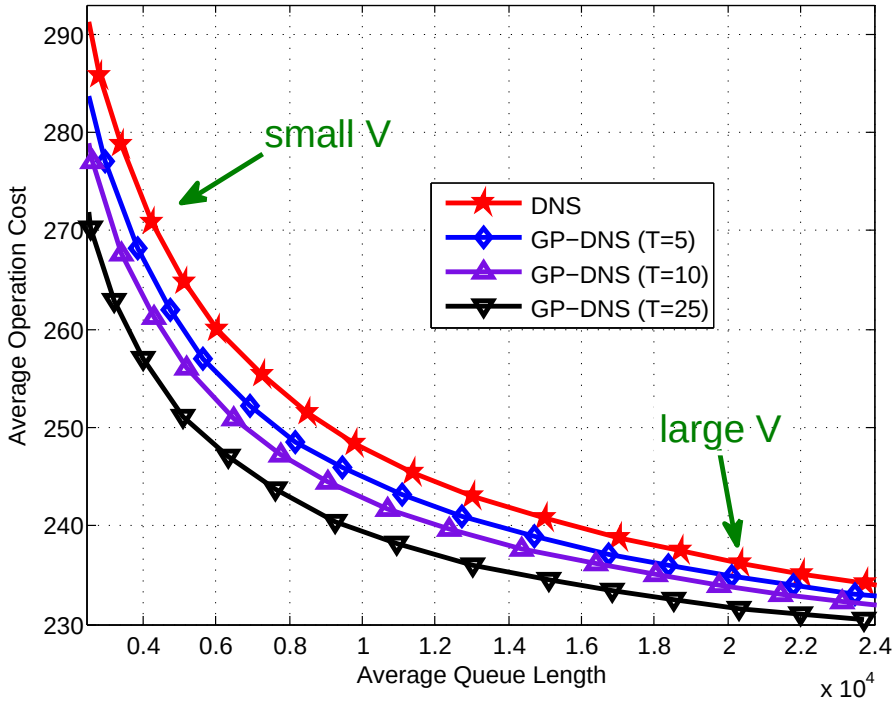
$$c_{av}^{\text{GP-DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{c(\boldsymbol{\alpha}(\tau))\} \leq \xi C(\theta) + \frac{B}{V},$$

$$Q_{av}^{\text{GP-DNS}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{l=1}^{L} \mathbb{E}\{Q_l(\tau)\} \leq \frac{B+V\xi C(\theta)}{\theta}.$$

Numerical Results

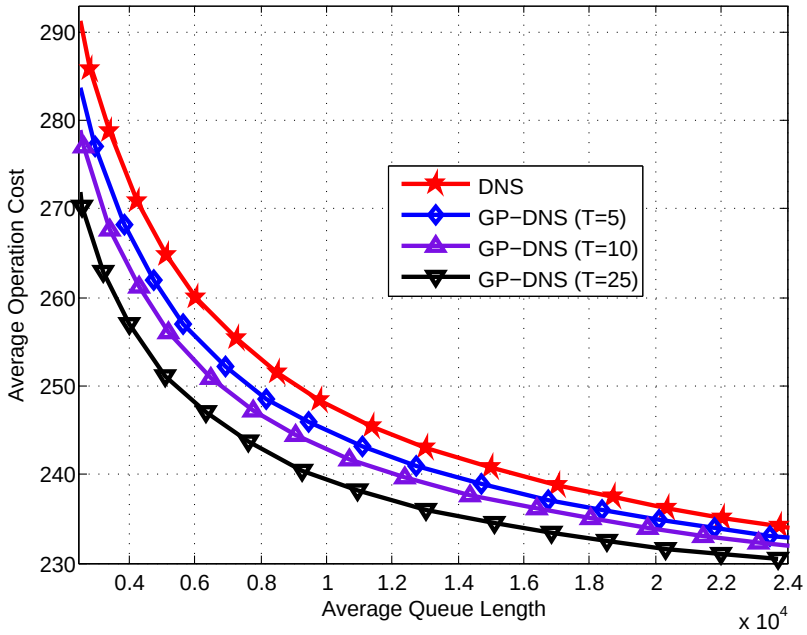
- Simulate **DNS** and **GP-DNS**
- Basic settings
  - 4 users, 8 networks, 64 locations
  - 1 cellular with full coverage, 672 Mbps (4G HSPA+)
  - 7 Wi-Fi networks, 150 Mbps (IEEE 802.11n)
- $c(\boldsymbol{\alpha}(t))$ and $r_l(\boldsymbol{\alpha}(t))$
  - Linear operation cost
  - Even data rate sharing
- Markovian randomness
- 100,000 slots

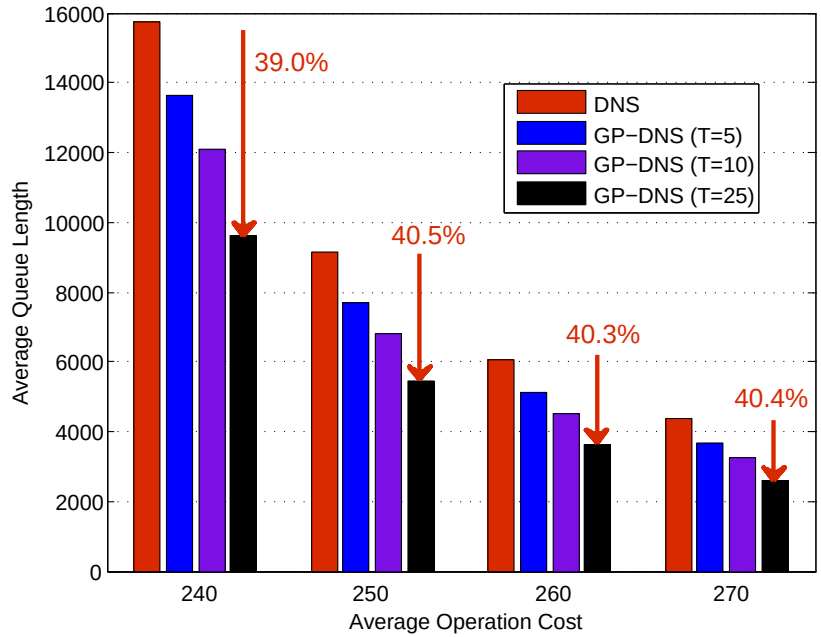
Observation 1: Cost-Delay Tradeoff



- As control parameter V increases,
  - Operation costs (of **DNS** or **GP-DNS**) approach the minimum value
  - Queue lengths *or* traffic delay (of **DNS** or **GP-DNS**) become larger

Observation 2: Prediction Improves Performance





- Cost-delay tradeoff increases with the prediction capability
  - If the operator pursues an operation cost of 250, **GP-DNS** with $T = 25$ (prediction window size) saves 40.5% traffic delay over **DNS**

Conclusion and Future Work

- Conclusion
  - Online network selection with cost-delay tradeoff
  - Current info.: **DNS**; current & future info.: **P-DNS**, **GP-DNS**
  - A novel frame-based Lyapunov analysis
- Future work
  - Heterogenous QoS requirement; inelastic traffic demand

THANK YOU

Network Communications and Economics Lab

http://ncel.ie.cuhk.edu.hk