# Slicing in 5G networks

Update 23/09/2020

Alessandro Spallina

#### Last Call

- Implemented smart conservative agent
- Normalization of alpha, beta and gamma (sum equal to 1)
- Plots of different mdp policies (discount factor)
- Plots of costs in detail (with and without multiplying factors alpha, beta, gamma)
- Bugfix

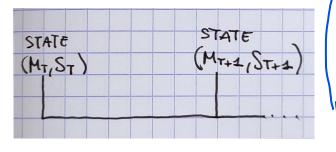
#### What's new?

- Support for finite horizon (mdp)
- Support for immediate action (mdp and simulator)
- Fix bad naming (conservative -> all on, smart conservative -> conservative)
- Fix formulation

#### Formulation - Assumption

#### **Delayed Action (timeslot view):**

- 1. State
- 2. Action chosen according to the state
- 3. Arrival phase (losses) 5,-4
- 4. Processing phase
- 5. Execution of the action chosen in (2)



Immediate Action (timeslot view):

- 1. State
- Action chosen according to the state
- 3. Execution of the action chosen in (2)
- 4. Arrival phase (losses)
- 5. Processing phase

SIMULATORA

#### Formulation - Assumptions of L. M. Bayati's Thesis

"We begin by serving the waiting jobs of the buffer, next we fill the free operational servers by the new jobs, then we fill the buffer." [1.3.1]

"At the beginning of each slot, and <u>based on the current state</u> of the system, an action  $\alpha \in A$  will be made to determine how many servers will be operational during the current slot." [4.1.2.1]

#### Immediate Action + phases exchange (timeslot view):

- State
- 2. Action chosen according to the state
- Execution of the action chosen in (2)
- 4. Processing phase
- 5. Arrival phase (losses)



#### Formulation - Transition Probability

This can be generalized as follows:

$$Q(m, s \to m', s') = \sum_{a=[m'-m]^+}^{\text{qsize}-m} P(\text{arr} = a) \cdot P(\text{proc} = m + a - m'|a + m)$$

+ 
$$\sum_{a=\text{qsize}-m+1}^{\infty} P(\text{arr} = a)P(\text{proc} = \text{qsize} - m'|\text{qsize})$$
 (3)

Where s' = s + action and

$$action = \begin{cases} 0 & \text{do nothing} \\ +1 & \text{allocate 1 server} \\ -1 & \text{deallocate 1 server} \end{cases}$$

- (2) non full queue
- (3) full queue but we have missing probabilities due the histograms

Where P(proc = x|y) is the probability of processing x jobs given that y jobs are found in the queue the instant when the processor starts to pick jobs from the queue. Observe that

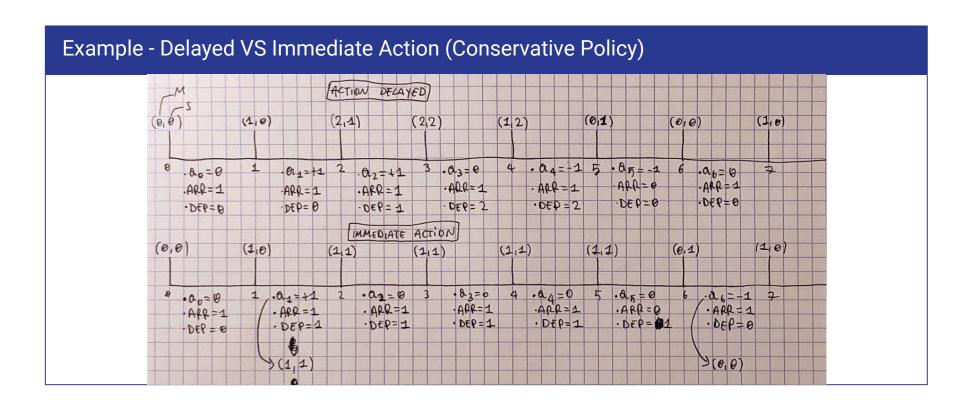
#### **Delayed Action**

$$P(\text{proc} = x|y) = \begin{cases} H_{\text{departures}}^{s}(x) & \text{if } x < y\\ \sum_{x=y}^{\infty} H_{\text{departures}}^{s}(x) & \text{if } x \ge y \end{cases}$$

#### Immediate Action

$$P(\text{proc} = x | y) = \begin{cases} H_{\text{departures}}^{s'}(x) & \text{if } x < y \\ \sum_{x=y}^{\infty} H_{\text{departures}}^{s'}(x) & \text{if } x \ge y \end{cases}$$

Notice that is the number of current servers s is equal to 0, then the departure histogram will be just  $\Delta_1([1.,0.,...,0.])$ 



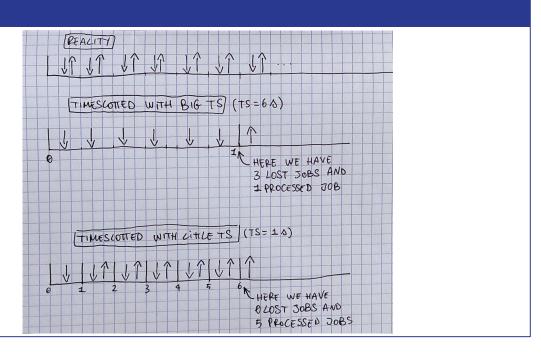
#### Timeslot Sizing - L. M. Bayati's Thesis

"We deal with time-homogeneous discrete time Markov chains. Actually it is important to assume the i.i.d.-ness of the arrivals. However to avoid to just keep that property as an assumption, we suggested to construct an arrival model which is i.i.d. In fact we do the following: the traffic trace must be sampled with a sampling period for which the arrival jobs can be considered as i.i.d. So, we sample, literally, the data with several sampling periods between a small sampling period (one second) and a big one (five minutes), then we applied the turning point test for each sampling periods in order to test the i.i.d.-ness of the sampled data. We found that **sampling period between 115 and 155**second has a p-value bigger than 0.05 thus at confidence level 0.95 we accept the null hypothesis, namely, the sampled data is i.i.d. [...] We consider frames of 136 second to sample the trace and construct empirical distributions for daytime, nighttime, and for each day of the week." [1.3.3]

#### **Example - Timeslot Sizing**

#### Assuming:

- QueueSize = 3
- 1 Server always On
- 1 job/sec



#### Timeslot Sizing - Questions

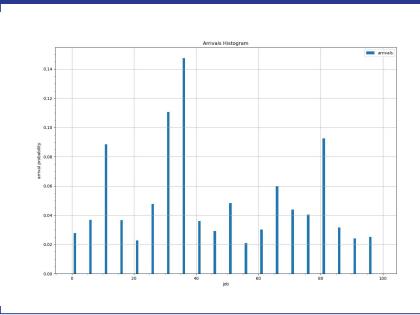
Since the behavior of the system is time-slotted (we see arrivals only at the beginning of a timeslot), the timeslot must be small enough to model what happens in continuous time as closely as possible.

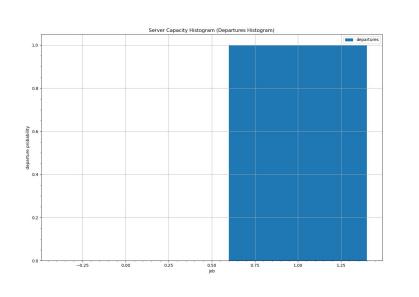
Can we scale down the timeslot after the construction of the arrivals histogram as in L. M. Bayati's Thesis?

# Simulation Results

**Common Parameters** 

#### Simulation Results - Common Parameters





#### Simulation Results - Common Parameters

- Queue size: 50
- Max allocated servers: 50
- C<sub>i</sub>: 1; alpha: 1
- C<sub>s</sub>: 1; beta: 1
- C<sub>i</sub>: 1; gamma: 1
- Number of simulations: 5
- Simulation Time: 1000 time slots
- MDP discount values: [0.8 0.99999]

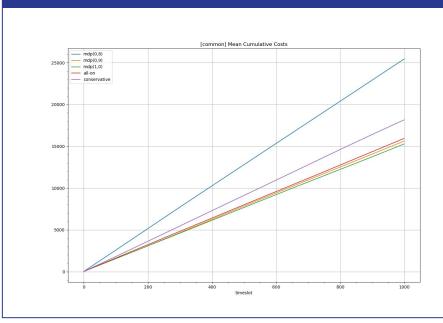
#### Simulation Results - Common Parameters

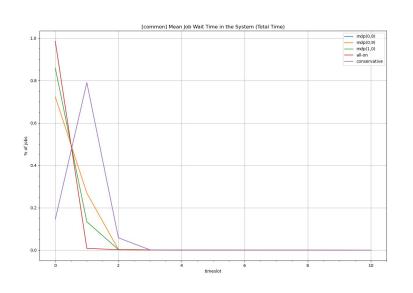
	Algorithm	Action Delayed	Sim. Time	RAM (max)
Scenario 1	Finite Horizon	Yes	94 min	approx. 700MB
Scenario 2	Finite Horizon	No	92 min	approx. 700MB
Scenario 3	Value Iteration	Yes	Stopped after 9h without results	??
Scenario 4	Value Iteration	No	Stopped after 9h without results	??

# Scenario 1

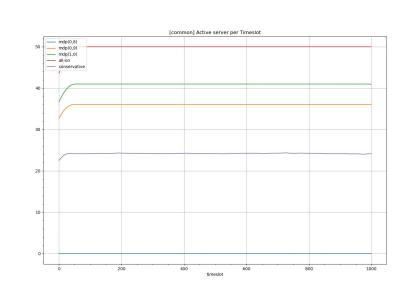
Finite Horizon, Action Delayed

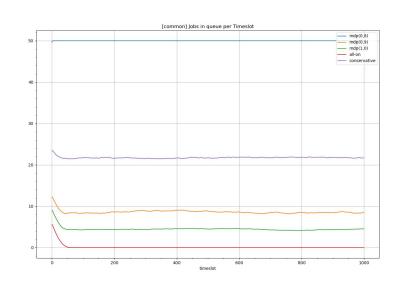
#### Scenario 1: Finite Horizon, Action Delayed



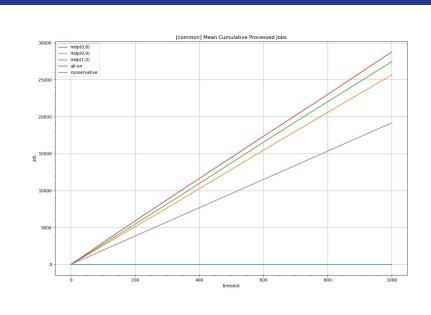


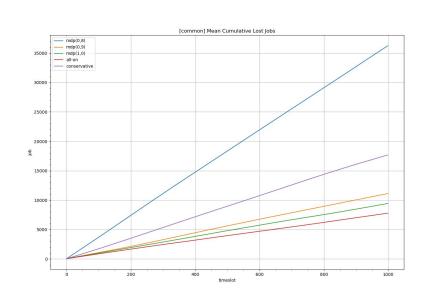
#### Scenario 1: Finite Horizon, Action Delayed



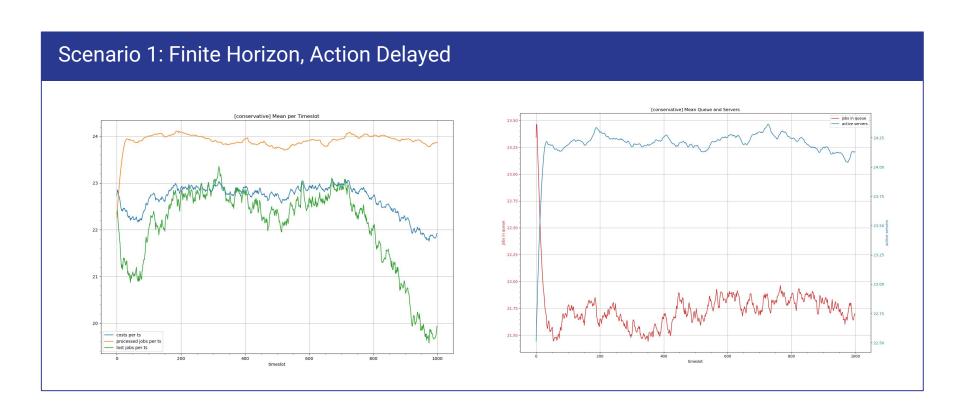


#### Scenario 1: Finite Horizon, Action Delayed





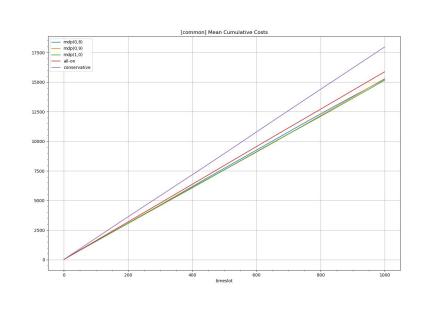


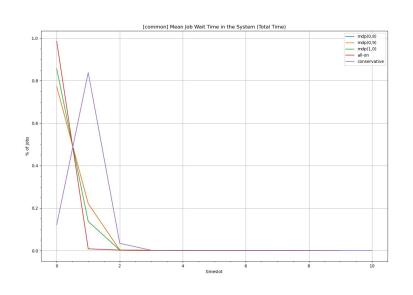


# Scenario 2

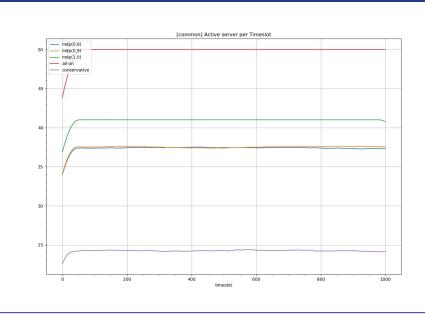
Finite Horizon, Immediate Action

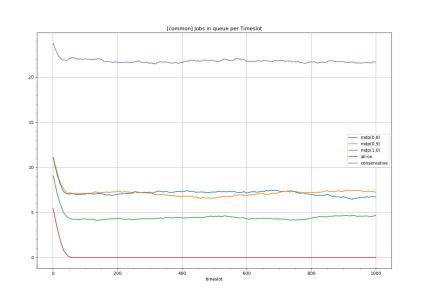
#### Scenario 2: Finite Horizon, Immediate Action





#### Scenario 2: Finite Horizon, Immediate Action





#### Scenario 2: Finite Horizon, Immediate Action

